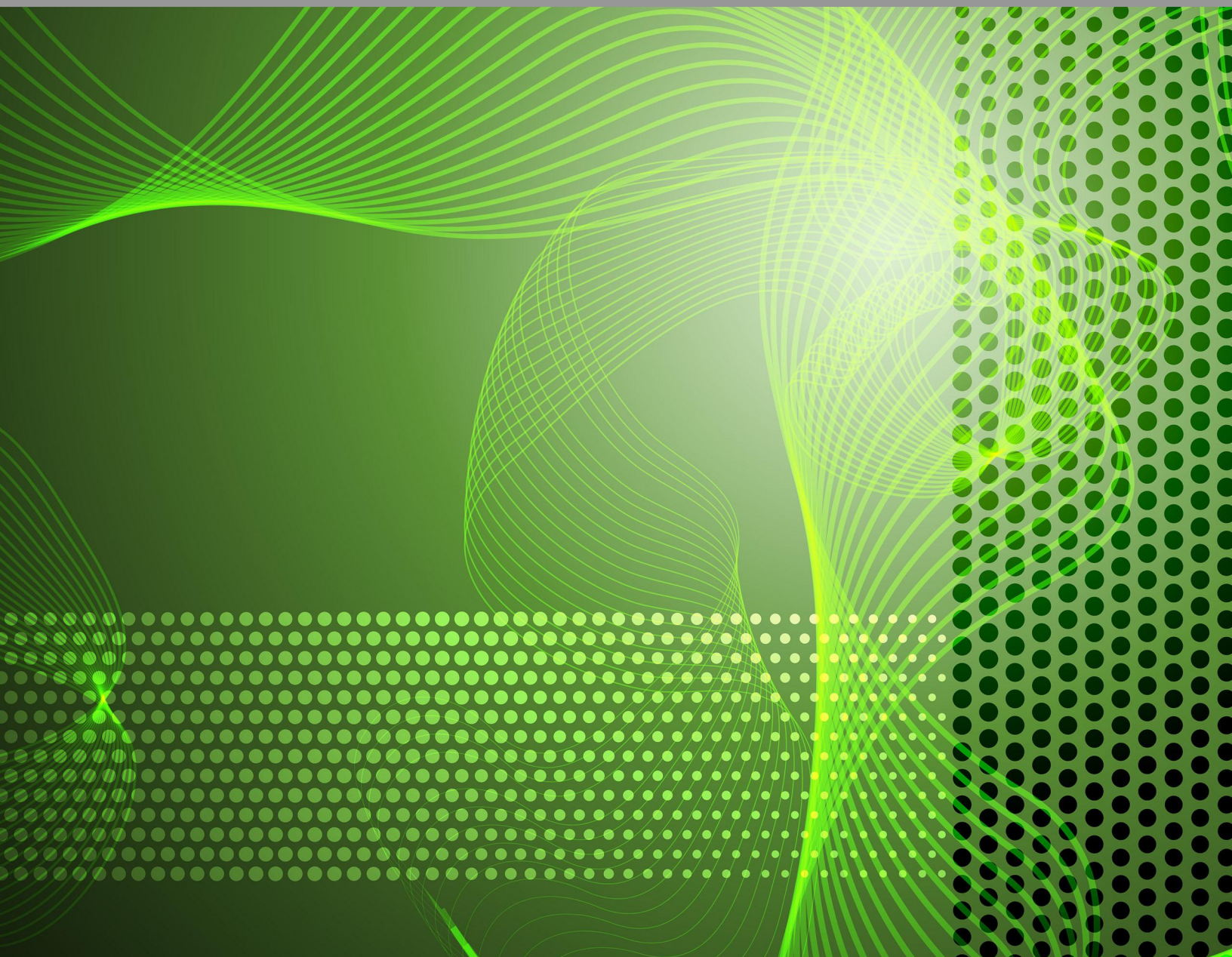


ck-12

flexbook
next generation textbooks

CK-12 Probability and Statistics

Concept Collection



CK-12 Probability and Statistics Concepts

Larame Spence

Say Thanks to the Authors

Click <http://www.ck12.org/saythanks>

(No sign in required)



To access a customizable version of this book, as well as other interactive content, visit www.ck12.org

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-source, collaborative, and web-based compilation model, CK-12 pioneers and promotes the creation and distribution of high-quality, adaptive online textbooks that can be mixed, modified and printed (i.e., the FlexBook® textbooks).

Copyright © 2016 CK-12 Foundation, www.ck12.org

The names “CK-12” and “CK12” and associated logos and the terms “**FlexBook®**” and “**FlexBook Platform®**” (collectively “CK-12 Marks”) are trademarks and service marks of CK-12 Foundation and are protected by federal, state, and international laws.

Any form of reproduction of this book in any format or medium, in whole or in sections must include the referral attribution link <http://www.ck12.org/saythanks> (placed in a visible location) in addition to the following terms.

Except as otherwise noted, all CK-12 Content (including CK-12 Curriculum Material) is made available to Users in accordance with the Creative Commons Attribution-Non-Commercial 3.0 Unported (CC BY-NC 3.0) License (<http://creativecommons.org/licenses/by-nc/3.0/>), as amended and updated by Creative Commons from time to time (the “CC License”), which is incorporated herein by this reference.

Complete terms can be found at <http://www.ck12.org/about/terms-of-use>.

Printed: April 2, 2016

flexbook
next generation textbooks



AUTHOR

Larame Spence

Contents

1	Why Study Statistics and Probability	1
1.1	Everyday Uses	2
1.2	Careers	6
1.3	Entertainment	10
1.4	References	14
2	Collecting Data for Study: Sampling	15
2.1	Introduction to Sampling	16
2.2	Population V.S. Sample	22
2.3	Distribution	27
2.4	Undercoverage	33
2.5	Response Rates	37
2.6	References	42
3	Common Types of Samples	43
3.1	Identifying Random Sampling	44
3.2	Implementing Random Sampling	48
3.3	Stratified Sampling	52
3.4	Cluster Sampling	56
3.5	Non-Probability Sampling	59
3.6	References	64
4	Evaluating and Displaying Data	65
4.1	Grouping Data	66
4.2	Analyzing Data	73
4.3	Relative Frequencies	79
4.4	Cumulative Frequencies	85
4.5	Creating Histograms	94
4.6	Interpreting Histograms	106
4.7	Frequency Polygons - Probability and Statistics	117
4.8	Creating Box-and-Whisker Plots	126
4.9	Interpreting Box-and-Whisker Plots	132
4.10	Creating Stem-and-Leaf Diagrams	138
4.11	Interpreting Stem-and-Leaf Plots	144
4.12	Creating Scatter Plots and Line Graphs	149
4.13	Interpreting Scatter Plots and Line Graphs	158
4.14	Creating Pie Charts	170
4.15	Interpreting Pie Charts	179
4.16	References	187
5	Central Tendency	190
5.1	Arithmetic Mean	191

5.2	Geometric Mean	197
5.3	Harmonic Mean	202
5.4	Median - Probability and Statistics	208
5.5	Mode - Probability and Statistics	212
5.6	Calculating Variance	216
5.7	Variance Practice	221
5.8	Calculating Standard Deviation	227
5.9	Coefficient of Variation	233
5.10	References	237
6	Probability	238
6.1	Basic Probability - Probability and Statistics	239
6.2	Union of Compound Events	243
6.3	Intersection of Compound Events	247
6.4	Multiplication Rule	252
6.5	Mutually Inclusive Events - Probability and Statistics	258
6.6	Calculating Conditional Probabilities	263
6.7	Identifying the Complement	270
6.8	Finding Probability by Finding the Complement	273
6.9	References	278
7	Probability Distribution	279
7.1	Understanding Discrete Random Variables	280
7.2	Understanding Continuous Random Variables	284
7.3	Probability Distribution	287
7.4	Visualizing Probability Distribution	293
7.5	Probability Density Function	299
7.6	Binomial Experiments	306
7.7	Expected Value	310
7.8	Random Variable Variance	317
7.9	Transforming Random Variables I	324
7.10	Transforming Random Variables II	329
7.11	References	333
8	Combinations and Permutations	334
8.1	Combinations and Permutations	335
8.2	Calculating Permutations	339
8.3	Permutations with Repeats	345
8.4	Permutations with Indistinguishable Members	348
8.5	Calculating Combinations	353
8.6	Calculating Combinations II	358
8.7	Using Technology	363
8.8	References	369
9	The Normal Distribution	370
9.1	Understanding Normal Distribution	371
9.2	The Empirical Rule	376
9.3	Z-Scores	381
9.4	Z Scores II	385
9.5	Z-scores III	390
9.6	The Mean of Means	396
9.7	Central Limit Theorem	402

9.8	Approximating the Binomial Distribution	409
9.9	References	416
10	Predicting and Testing	417
10.1	The Null Hypothesis	418
10.2	Critical Values	422
10.3	Tails	427
10.4	Confidence Intervals	432
10.5	The T-Test	438
10.6	Putting it Together	445
10.7	References	452
11	Linear Regression and Chi-Squared	453
11.1	Linear Relationships	454
11.2	Linear Correlation Coefficient	459
11.3	Least Squares	466
11.4	Contingency Tables	473
11.5	Chi Squared Statistic	479
11.6	Chi-Squared II - Testing for Independence	487
11.7	References	495
12	Reasoning	496
12.1	Inductive and Deducting Reasoning	497
12.2	Arguments	501
12.3	Euler Diagrams	505
12.4	Valid Forms	511
12.5	Hidden Premises	515
12.6	Structural Fallacies	520
12.7	Content Fallacies	525
12.8	References	530

CHAPTER

1

Why Study Statistics and Probability

Chapter Outline

- 1.1** **EVERYDAY USES**
 - 1.2** **CAREERS**
 - 1.3** **ENTERTAINMENT**
 - 1.4** **REFERENCES**
-

Here you will investigate several real-world applications of probability and statistics, hopefully coming to appreciate the very real value of understanding the concepts and how they relate to your own life.

1.1 Everyday Uses

Objective

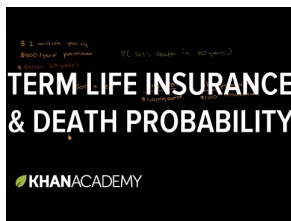
In this concept you will learn of some of the everyday uses of probability and statistics in the real world. By the end of the lesson, you should see that Probability and Statistics represents one of the most honestly useful math topics you are likely to study.

Concept

What does buying insurance or taking out a loan at the local bank have to do with statistics? What does predicting the weather have to do with probability? Why do boys generally pay more for car insurance than girls?

Stay tuned, after the concept and exercises below, we will return to these questions and discuss the answers.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62133>

<http://youtu.be/NSSoMafbBqQ> Khan Academy - Term Life Insurance and Death Probability

Guidance

Nearly every kind of business can benefit from an application of statistics. By knowing what product to sell at a particular time of year, or to a particular customer, a business can make the best use of product placement and advertising. Knowing what time(s) of the day, week, or year are the busiest, a manager can efficiently schedule her employees so as not to waste labor costs.



One type of business that makes extensive use of statistics is insurance sales. Insurance companies are just like most companies from the standpoint that they are in business to make a profit for their investors. That does not mean that buying insurance is a bad idea for individual people, or that the companies deliberately overcharge their customers, but it does mean that the companies are very careful to charge enough for each policy to 'insure' that the company makes money overall.

How can the companies know for certain how many people are going to make claims against their insurance policies? Or how big their claims will be? They can't know for *certain* since they don't have a way to see the future, but they can get a very reliable idea of the average number of claims from a specific population of people through the use of *sample* groups and the application of probability and statistics.

Example A

Suppose a particular insurance company has 100,000 clients, and research suggests that 1 out of each 25 people are likely to make a claim in a given year, for an average of \$15,000 per claim. Would a yearly premium of \$750 result in a profit for the company, given these statistics?

Solution: Essentially, since 1 out of each 25 people will cost the company \$15,000, the total premiums for each 25 people need to be greater than \$15,000 for the company to make a profit.

In other words, each policy is responsible for $\frac{1}{25}$ th of \$15,000, so each policy needs to cost *more than*:

$$\frac{\$15,000}{25} = \$600$$

Since the company plans to charge \$750 per year for each policy, that should result in a profit of \$150 per policy on average, for a total estimated profit of $\$150 \times 100,000 = \$15,000,000$ Not bad!

Of course, that level of profit is only likely if the original research and statistical calculation of \$15,000 per 25 people is correct (and the company has no other expenses - which is unlikely), so you can bet that huge multi-million dollar companies are *very* careful to make sure their calculations are as accurate as possible!

Example B

Predicting the weather is a tricky job. There are a nearly infinite number of possible *variables* that can affect the temperature and chance of precipitation for any given day. Of course, a weatherman cannot possibly take *all* of these variables into account every time he/she makes a prediction, so he/she must identify the most influential variables and just watch them closely for each prediction. Suppose that according to records, it has rained an average of 5 days during the month of April for each year over the last 15 years. If it is currently April 25th and there has been no rain, should the weatherman warn everyone to bring an umbrella to work for the next five days?



Solution: Maybe. However, the information regarding the average number of rainy days in April over the last 15 years probably won't have much to do with it. Although the history may be suggestive of a particular number of rainy days, it is certainly no guarantee of a specific result. If the weather conditions such as temperature, cloud cover,

relative humidity, etc., are all conducive to rain, then he/she is likely to predict rain, but the fact that there are only five days left is certainly no assurance that there *must* be rain all five final days so that the average will be fulfilled.

Example C

Suppose a car insurance company reviews the police records for thousands of speeding tickets and minor car accidents over a ten-year period, and notes the following:

TABLE 1.1:

	Speeding Tickets	“Fender Benders”
Boys ages 16 - 23	4,532	1,725
Girls ages 16 - 28	1,242	1,715

Would it make sense for the company to charge the same rates for boys and girls?

Solution: It certainly does not look like it.

According to the statistics, boys are nearly four times as likely to drive over the speed limit, and although there were slightly fewer recorded accidents for girls than boys, note that the age range for the girls was greater than for the boys. The greater age range suggests that there may have been more girls actually driving than boys, yet they ended up in nearly the same number of accidents!

However, it is extremely important to note that without data regarding the actual number of boys and girls in each group, we can't really get a good feel for the overall increased likelihood of boys making claims.

Concept Problem Revisited

What does buying insurance or taking out a loan at the local bank have to do with statistics?

It should make sense now to think that the interest rate you pay for a loan or the premium you pay for insurance is likely based to a great degree on what the statistics say about your likelihood to pay the loan off in a timely manner or make a claim against your policy.

What does predicting the weather have to do with probability?

Now we know that there are a number of different variables associated with the weather, including: historical weather patterns, current temperature and local trends, current humidity, regional weather patterns, and many more. The greater the number of variables taken into account and the more accurate the calculations, the more likely it will be that a particular weather prediction will be correct.

Why do boys generally pay more for car insurance than girls?

Statistically, boys drive faster than girls, and get into more accidents. That does not mean, of course, that any particular driver is more of a risk than any other just based on his or her sex, but overall it does mean that it makes logical sense for an insurance to charge a premium for male teen drivers.

Vocabulary

A **sample** is a smaller portion of all of the possible members of a given set. Chosen properly, a sample can provide a good approximation of the overall behavior of the entire group.

Data can be either singular or plural, and refers to pieces of information gathered from a sample for the use of statistical calculation.

In the context of statistics, a **variable** is a characteristic of the elements in your data set.

Guided Practice

- Which of these number(s) cannot represent a probability?
 - 0.00001
 - 0.5
 - 1.001
 - 0
 - 1
 - 20%
- Which types of studies below might a retail store make use of to improve sales?
 - The average amount of money spent by customers of various age groups
 - The type of products preferred by customers of various age groups
 - The best and worst selling products to female and male customers
 - The busiest season for selling a particular product
 - All of the above
- How could an understanding of statistics benefit *you* on university entrance exams, such as ACT or SAT?

Solutions:

- A probability can only be between 0% and 100% (or between 0 and 1, as a decimal). 0% means it will not happen, and 100% means it will happen, every number between represents some shade of “it *may* happen”. Choice “a” is negative, and choice “c” is greater than 100%, so neither is possible.
- “e” is correct. All of those bits of information would be very useful to a skilled retail manager, owner or salesperson (particularly if paid on commission).
- SAT/ACT preparation courses make extensive use of statistics to help students understand when to expect a problem to be easy or hard, when it is worth spending extra time solving a particular question, and when it is not.

Practice

- What is the difference between probability and statistics?
- What are three industries that make use of statistics?
- Why do girls generally pay less for car insurance?
- Why do retail stores start carrying holiday decorations and promoting gifts well before the holiday season?
- When an animated film is played in theaters, why is it often preceded by previews of children’s movies?
- Why are commercials for toys played during cartoons?
- Why are banner ads for beauty magazines often displayed on Hollywood rumor-type web pages?
- Why are children’s athletic shoes often promoted by professional sports players?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 1.1.

1.2 Careers

Objective

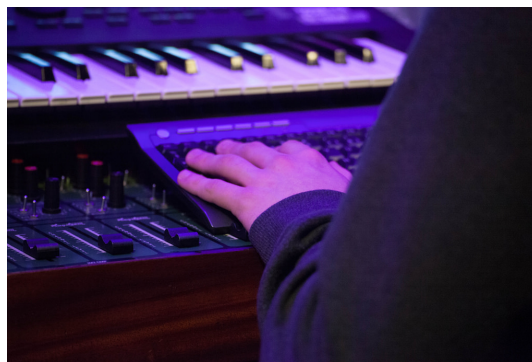
In this concept you will learn about some of the very many careers that make regular use of probability and statistics.

Concept

Are banking and weather forecasting the only careers that use statistics? If you plan to go into music or film production, or perhaps advertising, do you really have any use for statistics or probability?

Read on, in this lesson we will discuss careers that rely on statistics and probability - some may surprise you!

Guidance



Professional musicians and music producers depend on knowing what types of music are most popular among members of their *target audience*. A great jazz song will probably not be a platinum-seller if marketed to middle and high school students, and the latest rap sensation is probably going to be much less popular among the over-50 demographic than among teens! Even the right kind of music for a particular audience may not be ideal to release at specific times of the year (think Christmas music in June), and otherwise less popular songs may be 'hot' due to other factors (when an 80's song is used in a hit movie, sales of the original album can skyrocket).

Statistics can be remarkably helpful in determining when to record or produce a particular song or album based on how consumers have responded to similar situations in the past. Considering the number of people whose incomes depend on making the most of sales, it is no surprise that musicians and producers alike take the calculation of prime release dates and music types very seriously. For instance, if you release a Christmas song remake too soon, it's old news by the time the actual holiday comes around and nobody wants to buy it as a present. Release the same album too late, and people may have already purchased all of the holiday gifts they can afford.

Naturally the situation is similar for film production and actors/actresses. A movie about the world trying to avoid a giant asteroid may be reasonably popular in any season, but if it is released during the summer (when the biggest movie-going crowds are out of school) and right after a big news report about a comet visible in the sky... now the same movie is a multimillion dollar blockbuster!

In fact the concept holds true for pretty much any product that is sold to a wide audience: the more accurately a statistician can evaluate the data for a particular product or market, the more successful the sales may be if those statistics are properly put to use.

Example A

DJ Woodz, the popular artist, is hoping to release his next Platinum Record. He learns that a recent study suggests that songs about expensive streetcars are popular right now with 62% the 19-23 year-old male crowd. Woodz also stumbles across an article that hints that a big blockbuster movie about country music will be released soon. He knows from experience that songs about making it big as a rap singer have been rather consistently well received by about 45% of males ages 17-25.



How might these statistics help him to plan out his next album?

Solution: E.Z. can certainly make good use of the first and last statistics by recording a few songs about singers making it big and driving expensive cars! He may want to try to time the release of his album before the big movie about country music comes out, or perhaps delay it until the excitement of the film dies down. By factoring in as many accurate statistics as appropriate, Woodz can make the most of his next album.

Example B

Based on what you know about the application of statistics in the music and film business, can you identify some ways that statistics might benefit a retail-clothing store?

Solution: Successful clothes retailers continuously monitor statistics such as:

- Styles and colors that sold the best at the same time in previous years
- Popular trends in style, cut, and color in trend-setting areas
- Current preferences based on specific *demographics*
- Brands and styles currently popular at competing retailers

Example C

How might the application of statistics help a police officer, or an entire police force, make the most effective use of their limited funds to help keep people safe?

Solution: Police officers apply statistics in a number of ways. By monitoring which intersections have the most accidents, patrolmen can pay particular attention to ensuring that people drive carefully in those areas. By monitoring weather statistics, officers can be prepared for dangerous trends in the kinds of weather (snow, high winds, heavy rain), which may lead to greater need for men on duty.

Concept Problem Revisited

Are banking and weather forecasting the only careers that use statistics? If you plan to go into music or film production, or perhaps advertising, do you really have any use for statistics or probability?

It is probably pretty clear now that there are many, many careers that rely on an understanding and application of probability and statistics.

Certainly weather forecasters and bankers monitoring interest rates use statistics daily, but so do sales managers and advertisers in nearly every market from food to clothes to entertainment. Without understanding the appropriate

statistics, many (or all) of the biggest companies you know of would probably never have become successful in the first place.

Vocabulary

A **consumer** is anyone who purchases or uses a particular product or service.

A **demographic** is a specific group of consumers chosen by age, sex, religion, home country, place of residence, music preference, or any other specified characteristic.

A **target audience** is a particular demographic that is intended to be the main group of consumers of a particular product or service.

Guided Practice

1. What use would a travel agent have for statistics?
2. How might someone who has to prepare for hurricanes and tornadoes use statistics?
3. How might a stock broker use statistics?

Solutions:

1. Travel agents can be very helpful to travellers by knowing the statistically least expensive times to travel to a particular destination, or by knowing about locations that are statistically more likely to be rated highly by a particular kind of traveller, among many other things.
2. By studying the most active locales and seasons for dangerous weather, a person responsible for mitigating risk from such things can improve the chances that the potential victims have time to prepare.
3. Stock brokers monitor the statistically likely influencing factors for different kind of stocks, in order to make educated guesses about potential price increases and decreases. Because a good broker can more often predict when a price is about to increase, he or she is in a better position to buy the stock at a lower price and re-sell it shortly afterward for more money.

Practice

A June 2008 study by the U.S. Travel Association suggested that air travelers avoided an estimated 41 million trips over the previous 12 months because of frustration over the hassle of flying. The missed trips cost the U.S. economy more than \$26 billion. Also, nearly 50 percent of travellers polled said that they believe the air travel system is not likely to improve in the near future.

The extended effect of avoided trips was also costly to other industries; hotels lost almost \$6 billion and restaurants more than \$3 billion. Even federal, state and local governments lost more than \$4 billion in tax revenue because of reduced spending by travelers. (Source: Air Travel Survey, 2008)

1. How might a car salesman make good use of these statistics?
2. Would this information be useful to a potential restaurant owner?
3. How might a car-rental company make use of this information to boost rentals?
4. Based on the number of lost trips, how much does the loss of a single trip, on average, cost the US economy?
5. How might the government's loss of tax revenue negatively affect the travellers who choose to stay home and perhaps take a less expensive driving vacation closer to home?

Business travel in the U.S. is responsible for \$246 billion in spending and 2.3 million American jobs; \$100 billion of this spending and 1 million American jobs are linked directly to meetings and events. For every dollar produc-

tively invested in business travel, businesses experience an average \$12.50 in increased revenue and \$3.80 in new profits. (Source: [The Return on Investment of U.S. Business Travel](#))

6. What kinds of business would benefit from learning these statistics?
7. Do these numbers mean that every business should spend money on business travel?
8. Would these statistics be of interest to hotel owners?

The Hispanic/Latino population in the U.S. is expected to reach 47.8 million by 2010, representing 16 percent of the total population. By 2050, the Hispanic/Latino population is projected to total 102.6 million, comprising 24 percent of the U.S. population. In 2007, there were an estimated 16.2 million Hispanic adult leisure travelers who took a combined 50.4 million domestic and outbound trips and spent \$58.7 billion on their travels. (Source: Profile of Hispanic/Latino Leisure Travelers, 2008 Edition)

9. How might these statistics be of use to a travel agent?
10. Would these statistics benefit a restaurant owner?
11. How could the owner of a bed and breakfast inn make use of these statistics?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 1.2.

1.3 Entertainment

Objective

In this concept you will learn about some of the more entertaining applications of probability and statistics.

Concept

Do football players or coaches need to understand and use statistics? What about track and field competitors or swimmers? Can an understanding of probability make you more likely to win games of chance?

Read on to see how becoming an expert statistician can help you be more successful in ways you might never have imagined.

Guidance

Competitive sports are extremely popular in countries all over the world. In some countries, players can make huge salaries in competitive sports, and even countries where professional sports are outlawed encourage players to compete for national fame in the Olympics.



The obvious part of becoming a successful athlete is the vigorous physical training. Naturally a player will want to develop the instincts and physical prowess appropriate to his or her chosen sport. The less obvious, but no less important, part of becoming a champion athlete is developing and applying an understanding of the statistically important research about the sport.

Baseball pitchers need to know the particular types of pitches to throw at each hitter on an opposing team. By reviewing each hitter's past records of strikes, fouls, base hits, and home runs based on each type of pitch, the pitcher will know who is more susceptible to fast balls, curves, etc., and will be a much more valuable member of his or her team.

Football players are equally dependent on statistics. A quarterback who has no idea which teammates are most likely to catch a particular throw based on previous performance would not have much chance of leading his or her team

to victory. Coaches without a detailed understanding of what training methods are statistically most successful for developing the particular skills a team needs are not liable to lead their teams to victory.



Another pastime with eager participants all over the world is playing games of chance. Whether involving cards, dice, marbles, or anything else, any game of chance involves some level of probability by definition.

By becoming an expert on the likelihood of a particular card showing up in a random shuffle, or a specific roll turning up on a *fair die*, a clever player can significantly improve his or her chances of winning more often. Additionally, by studying and applying the statistical data of his competitors, a skilled player will know when to bet high on a bluff, or even hold an otherwise good play for a better time.

Example A

The Jets, a nationally competitive football team, are planning to play the #1 ranked team in their league in 2 weeks. What kind of statistics might they want to pay particular attention to as they research their future opponents?

Solution: Very generally, the Jets might want to know whether their opponents are more offensive or more defensive, whether they tend to 'run the ball' or pass more often. It would also be beneficial to identify the most dangerous players on the other team so that they can be specifically focused on. All of these things could be relatively easily calculated by a statistician reviewing past games, and would represent a significant advantage to the Jets as they go into the competition.

Example B

Xio is playing a card-based game of chance with two other people. The game is played in rounds. In each round each player is dealt one card from the deck, and each player bets that his or her cards have the highest point value in that round.

If the deck starts with 4 each of the cards numbered 1-10, and Xio knows that there have been two 9's, and three 10's out of the 15 cards already played, how confident should he be that the 9 he was just dealt will represent a positive outcome the next round?

Solution: There are 24 unknown cards remaining in the deck: 40(original deck) - 15(dealt previously) - 1(Xio's card), and only a 10 will beat Xio's 9. If there is only a single 10 remaining out of the 24 unknown cards, Xio can be pretty confident that his 9 will at least tie in this round.

The actual chance is $\frac{1}{24}$ or approximately 4% for the first player, and $\frac{1}{23}$ or also approximately 4% for the second:

$$4\% + 4\% = 8\% \text{ chance that Xio will lose this round, pretty good odds}$$

Example C

Jane is playing a game of chance with dice. The object is to roll two dice as many times in a row as possible *without* rolling a 7. If Jane has already rolled 5 times without getting a 7, how likely is it that she will roll a 7 on the next toss?

Solution: The first, and perhaps somewhat *counter-intuitive*, point to note is that dice do not have any memory! That means that Jane's chances of rolling a non-seven on her next roll are exactly the same as they were on her first

roll, so we need to get an idea of how likely a roll of seven is on any given toss.

Seven is a common key number when discussing two “fair” (each number has the same chance of appearing) six-sided dice. This is because there are more ways to roll a seven than any other combination. The chart below shows the number of ways to roll each possible value of 2-12:

TABLE 1.2: Roll:

2	1 + 1					
3	1 + 2	2 + 1				
4	1 + 3	3 + 1	2 + 2			
5	1 + 4	4 + 1	2 + 3	3 + 2		
6	1 + 5	5 + 1	2 + 4	4 + 2	3 + 3	
7	1 + 6	6 + 1	2 + 5	5 + 2	3 + 4	4 + 3
8	2 + 6	6 + 2	3 + 5	5 + 3	4 + 4	
9	3 + 6	6 + 3	5 + 4	4 + 5		
10	4 + 6	6 + 4	5 + 5			
11	5 + 6	6 + 5				
12	6 + 6					

We can see by counting on the chart that there are thirty-six possible combinations of two six-sided dice. There are six different combinations that result in a 7, thus, the chances of rolling a 7 would be:

$$\frac{6}{36} = \frac{1}{6} \text{ or } 17\% \text{ chance of rolling a } 7$$

It is clear that at the beginning of each roll, Jane’s chances of not getting the infamous 7 are really quite good. In later lessons we will discuss the overall probability of rolling a 7 six times in a row, which is quite different than the probability of rolling a 7 on a sixth toss.

Concept Problem Revisited

Do football players or coaches need to understand and use statistics? What about track and field competitors or swimmers? Can an understanding of probability make you more likely to win games of chance?

We can certainly see, based on the examples above, that statistics and probability calculations are integral to sports of all kinds, and that understanding the probabilities involved with any given outcome can help you make smart decisions in games of chance.

Vocabulary

A **fair die** is a die with an equal chance of landing on any side. An **unfair die** would be more likely to land on a particular number than the others.

A **sample space** is the set of all the possible outcomes of an event.

A **sample point** is just one of the possible outcomes

If something is **counterintuitive**, it is not what you might initially guess. For instance, it seems counterintuitive that a feather in a vacuum will drop like a stone, but it is a fact regardless.

An **event** is a particular occurrence in a series: one roll of a die, one flip of a card, etc.

Guided Practice

1. Suppose you have a single six-sided die, what is the probability of rolling an odd number for a game of chance?
2. Suppose you are playing a game where you flip two coins and try to guess how they will land. Would it be better to guess that both land with the same side up or that they would land with different sides showing?
3. If you draw a card at random from a single deck, what is the sample space of possible outcomes?

Solutions:

1. The sample space would be each of the possible numbers: 1, 2, 3, 4, 5, and 6. Since you can only roll one of them, and all are equally likely, the probability of any particular number is $\frac{1}{6}$.
2. If you flip two coins, the sample space is: $\{(H, T), (H, H), (T, T), (T, H)\}$. Since there are four possibilities, and two show the same side, while two show different sides, the probabilities are equal: 50%.
3. The sample space is the entire deck. Ace, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King of each of the four suits (The probability of pulling any single particular card is $\frac{1}{52}$).

Practice

1. Identify four kinds of entertainment or sports that benefit from statistics.
 - 1.
 - 2.
 - 3.
 - 4.
2. How could it be worthwhile for a pitcher to study the type(s) of pitch that a specific hitter is most likely to score a run on?
3. How could a strategy gamer benefit from studying the play style statistics of an opposing team?
4. How do players of online multiplayer games like World of Warcraft or Second Life use statistics?
5. Is there a benefit to learning statistics before playing games of chance? Give an example.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 1.3.

Various real-world uses of probability and statistics are discussed and explored. You will be given examples of multiple ways in which an understanding of probability and statistics could contribute to your lives and activities.

1.4 References

1. PublicDomainPictures. <http://pixabay.com/en/money-finance-wealth-currency-163502/> .
2. s58y. <https://www.flickr.com/photos/s58y/7719391630> .
3. Trausti Evans. <https://www.flickr.com/photos/114303073@N05/11934570484/> .
4. three6ohchris. <https://www.flickr.com/photos/three6ohchris/2133491020> .
5. Misty. <https://www.flickr.com/photos/mistybushell/7789626770/> .
6. Ella's Dad. <https://www.flickr.com/photos/mistybushell/7789626770/> .

CHAPTER

2

Collecting Data for Study: Sampling

Chapter Outline

- 2.1 INTRODUCTION TO SAMPLING
 - 2.2 POPULATION V.S. SAMPLE
 - 2.3 DISTRIBUTION
 - 2.4 UNDERCOVERAGE
 - 2.5 RESPONSE RATES
 - 2.6 REFERENCES
-

Here you will learn about some important considerations when collecting data to use for statistics and probability studies. This chapter is about the *concepts* involved with identifying and creating valid samples for data collection, later lessons will return to many of these topics and re-evaluate them from a more mathematically rigorous standpoint.



2.1 Introduction to Sampling

Objective

In this concept you will learn how to accurately gather data about a large population without needing to get responses from each and every member.

Concept

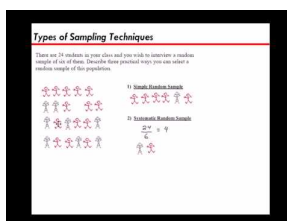
Suppose you were chosen to help pick out a theme for your school prom. Out of all of the initial suggestions offered by your team, you have narrowed the options down to 3: Famous Couples through the Ages, Romance Under the Sea, and Stairway to Heaven.



Since this is the Senior Prom, you feel that the Senior Class should make the final call. Unfortunately, there are over three hundred seniors in your school, and your deadline for a decision is in one hour! How could you get a good idea of the preference of the class as a whole in such a limited time?

By the end of this lesson, you should have no problem suggesting a good solution!

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62198>

<http://youtu.be/rASK8PpqakM> TheMathClips - Types of Survey Sampling Techniques

Guidance

There are many situations in life where we need to gather data on a very large or difficult to study *population*. Certainly it is ideal in most cases to be able to individually *poll* each and every member, but sometimes that just isn't feasible.

In such cases, the solution is to use a *sample* or *subset* that is carefully picked to accurately represent the full population. An experiment conducted on a well-chosen sample should provide an accurate representation of the results you would get by performing the same experiment on the population from which the sample was created.

There are many different ways to choose a sample, and all have applications for which they are more or less appropriate.

A few examples of sampling methods:

- Random Sampling (choosing representatives by rolling a die, for instance)
- Stratified Sampling (choosing a proportional number of representatives from each of a number of subgroups of the initial population) These divisions are chosen based on the belief that the subgroups differ significantly with respect to the variable that you are measuring. For example you might stratify by age or by income.
- Cluster Sampling (choosing representatives which are close to other representatives based on a particular factor such as location, age, color, size, etc.)
- Multi-Stage Sampling (narrowing down a field of representatives by successively applying multiple different sampling methods) For example you might stratify and then take a simple random sample from each stratum.

Example A

Would it be necessary to use a sample group to evaluate the effects of too much sugar on a group of 15 elementary-school children? What about a playground full of 300 children?

Solution: 15 children certainly seems like a manageable size group for study, so choosing a sample to represent the whole group is probably not necessary from that standpoint. However, this is the type of study where a *control group* would be an important consideration. If you just gave an extra handful of candy to every child, you would not know how much of the later energy actually came from the sugar, and how much was just a result of age. By pulling aside a control group of perhaps 6 students who would *not* get the extra sugar, you could better evaluate the *difference* in energy actually due to diet rather than age.

With 300 children all running around a playground, collecting them all together and attempting to organize a study might prove a daunting task. If you just chose a sample of perhaps 30 of them, some a little older, some younger, some boys, some girls, you could get an estimate of what would happen if you applied the study to the entire group.

Example B

Suppose you wanted to study the effect of rubbing marbles with candle wax before playing a classic game of marbles. After setting aside a control group, you are ready to choose a sample set of marbles to rub with the wax. Would a stratified sampling of the remaining marbles be a good choice in this situation?



Solution: Probably not. Marbles are generally created to be as alike as possible in every way other than appearance, and since appearance is unlikely to have an effect on the result of the wax experiment, it would not make sense to carefully attempt to represent each color or type of decoration. A random sample would be simpler and would very likely yield the same results.

Example C

The student council at your school has been given an assignment to find a good use for a grant that the school received to make school more enjoyable for the students. After a week or two of deliberation, the council announces that the studies they have conducted suggest that providing the cheerleading squad with new pom-poms is the #1 priority of 90% of the students in the school. Of course, the chess club members disagree and conduct their own study. If the chess team chooses a sample the same way the student council did, and their results suggest that 90% of respondents think that the money should go toward new chess clocks, what error do you think both groups committed in the choice of sample groups for study?

Solution: It would certainly appear that both groups were guilty of a process called 'cherry-picking', which means that they deliberately chose to question people who shared the same interests in order to get favorable results from their polls. Obviously neither group's results are likely to be representative of the entire student body, but rather only represent the views of the chess team and the cheerleading squad!

Concept Problem Revisited

Suppose you were chosen to help pick out a theme for your school prom. Out of all of the initial suggestions offered by your team, you have narrowed the options down to 3: Famous Couples through the Ages, Romance Under the Sea, and Stairway to Heaven.

Since this is the Senior Prom, you feel that the Senior Class should make the final call. Unfortunately there are over three hundred seniors in your school, and your deadline for a decision is in one hour! How could you get a good idea of the preference of the class as a whole in such a limited time?

This is an excellent case of the need for a representative sample of a population. Without having the time to poll all of the members of the senior class, you could get an idea of what the most popular theme would be by choosing a smaller number of seniors to represent the entire class. Just be careful to minimize the chance that your chosen representatives have any sort of **bias** that might keep them from properly representing the class as a whole.

Vocabulary

A **population** is the complete set (every single member) of a group of possible items to be studied.

To **poll** the members of a group means to question them regarding a specific topic.

A **sample** or **subset** is a smaller group of members chosen to represent a larger group. Properly chosen, a sample should provide the same results (on a smaller scale) as the population from which it was created.

A **control** group is a set of members deliberately kept as separate as possible from a particular study so as to provide an example of how the members should appear if unchanged.

Bias refers to a desire to achieve a specific result from a particular study, regardless of the data.

Random Sampling (choosing representatives by rolling a die, for instance)

Stratified Sampling (choosing a proportional number of representatives from each of a number of subgroups of the initial population) These divisions are chosen based on the belief that the subgroups differ significantly with respect to the variable that you are measuring. For example you might stratify by age or by income.

Cluster Sampling (choosing representatives which are close to other representatives based on a particular factor such as location, age, color, size, etc.)

Multi-Stage Sampling (narrowing down a field of representatives by successively applying multiple different sampling methods) For example you might stratify and then take a simple random sample from each stratum.

Guided Practice

1. What kind of sampling would you expect was used if the sample group was composed of 5 yellow, 3 green, 4 red, and 6 blue members, and the population included 48 blue, 32 red, 24 green, and 40 yellow members?
2. What type(s) of sampling method(s) might be most appropriate for approximating the number of cutthroat trout in a 25-mile section of river?
3. Would you reasonably expect bias to have affected a sample composed of 75% Toyota vehicles in a study of the most common cars in large U.S. cities?
4. Would a random sampling of students be the most appropriate method of sampling for a study of the most enjoyable after-school club in a large public school?
5. What might you conjecture about a study that claims 100% of respondents preferred “Super Sweet and Crunchy” cereal over “Super Duper Sweet” cereal?

Solutions:

1. Since the sample group contains exactly $\frac{1}{8}$ as many members of each color as the entire population, it is reasonable to suspect that a **stratified sampling** was used.
2. A 25-mile-long section of river is likely to include a number of different types of ecosystems that each would harbor a different density of fish. In order to get a good sample, a **multi-stage** sampling method comprised of a stratified sample of different ecosystems followed by a random sampling of fish in each ecosystem would probably be a good choice.
3. Although Toyota is a very popular vehicle manufacturer, 75% is an extremely high percentage of vehicles in a large city (reasonable estimates put Toyota somewhere between 25 and 30 percent). Such a huge number would definitely suggest sample bias.
4. Probably not, since a random sampling would likely include a large number of students who either have no opinion or have no experience with any after school clubs. More accurate results would be obtained by a multi-stage sample that first identified club members, and then randomly selected representatives from them.
5. There are a number of reasonable specific conjectures we might make, most related to inaccurate sampling methodology. Perhaps the sample was chosen from employees of the “Super Sweet and Crunchy” cereal company, perhaps respondents were offered a reward for choosing one option over the other, perhaps there was only a single member of the sample group or the “study” didn’t include milk for the other cereal, or didn’t offer samples of “Super Duper Sweet” to respondents at all

Practice

1. Margo collected 12 carrots in a bag. She drew 5 carrots out of the bag. Is this a random sample of the carrots in the bag?

2. Chris put some assorted colored kerchiefs into a box. He looks into the box and pulls out the blue kerchiefs. Is this a random sample of the kerchiefs in the bag?
3. Sue had red and white beans in a jar. She reached in and pulled out 10 beans, without looking in the jar. Is this a random sample of beans from the jar?

For questions 4-6, identify the population and the sample from each:

For example: In a class of 20 students, where each student is asked if they have gone to the movies in the past month, you would identify the population as 20 Students, and the sample as 20 students.

4. People aboard a plane who have aisle seats are asked if they travel more than 5000 miles per year.
 - a. Population:
 - b. Sample:
5. A team of marketing specialists survey every sixth child entering a park to find out how many rides they plan to go on while playing in the park.
 - a. Population:
 - b. Sample:
6. Every 15th adult at the exit door of the grocery store is questioned to find out if the store should increase its hours of operation.
 - a. Population:
 - b. Sample:
7. Luke wants to find out where most high school students buy their food for lunch. He surveys every fourth student he sees in the high school parking lot and asks them where they get food for lunch. Which would have been an improvement in Luke's experiment?
 - a. Survey all of the students in the school.
 - b. Survey all people in the parking lot.
 - c. Survey students in the lunch hall.
8. Sue is trying to determine the best location to sell snow cones. There are 4 locations in the city (on a side street, downtown, near a park and at a school. Sue observed that many people visit the downtown area and the park. Sue decided to sell snow cones in the downtown area where she saw the most people gather. What changes to Sue's sample would have given her a better understanding of where to sell snow cones?
9. Kerry collected shells from a visit to the ocean in a shoebox. She takes out a handful of shells from the box. Is this a random sample of shells in a box?
10. There are four dentists in a city. Their offices are located in four different parts of the city. Jake wants to attempt to figure out which dentist has the most patients. He observed that the Downtown and West Street areas have larger populations. He concurred that the dentists in those areas must have more patients. After comparing those two areas, he decided that the West Street dentist had the most patients because the area had more traffic. What changes to Jakes technique would have given him a better understanding of which doctor had the most patients?
11. Caroline wants to predict which restaurant will have less business during the Christmas season. There are three restaurants in the city. Two are on the outskirts of a city and one is in the city. She knows that two hotels situated on the outskirts are fully booked because one has Christmas show and one has a huge indoor pool. From this information she inferred that the restaurant in the city will have less business during the Christmas season. What could Caroline do to improve her experiment?
 - a. Ask people at the hotels if they like fast food.

- b. Survey all people to see which December holiday they celebrate.
- c. Look at the past holiday performance of the restaurants.

The table gives information about the number of girls in each of four schools.

TABLE 2.1:

School	A	B	C	D	Total
Number of Girls	126	82	201	52	461

12. Jenny did a survey of these girls. She used a stratified sample of exactly 80 girls according to school. Calculate the number of girls from each school that were in her sample of 80. Complete the table.

TABLE 2.2:

School	A	B	C	D	Total
Number of Girls					80

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 2.1.

2.2 Population V.S. Sample

Objective

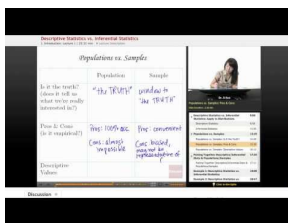
In this concept you will learn how to appropriately decide whether to gather a sample for study, or use the entire population in question.

Concept

Sometimes it can be a bit tricky to decide whether to conduct a particular study upon a sample group or on the entire population. Suppose you were attempting to put together a menu for a camping trip with a large group of friends and wanted to make sure nobody was allergic to peanuts before planning peanut-butter sandwiches for lunch. Would you need to question all 50+ friends individually? Would it make sense to choose a representative sample to poll instead? What if you wanted to pick a few popular types of soda to bring along, would that be a different situation?

At the end of the lesson, we'll return to this question to apply what we have discussed.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62200>

<http://youtu.be/vlbjxguaMrk> EducatorVids2: Statistics - Populations vs Samples

Guidance

Before you begin any particular study, you will need to decide whether you need to get data from the entire population in question, or just a representative sample of the population instead. For most studies, it makes much more sense to use a sample than to try to collect data on an entire population, but sometimes a sample is not enough. The most famous census is the U.S. Population Census, conducted once every 10 years.



According to the Constitution, the population of the United States is *enumerated* once every ten years by physical count, and estimated in the intervening years by statistical sample. Observing the incredible cost (the 2010 Census cost approximately 13 billion dollars!) and organizational effort required for the census makes it clear why there are so few census studies conducted on the U.S. population. However, smaller census studies are more common than you might think.

Some studies would make no sense at all to conduct on an entire population. In fact, one entire class of study comes to mind: *destructive study*. A destructive study requires that the sample be ruined for its intended use by the study itself. Vehicle manufacturers test the durability of different models by crashing sample vehicles into simulated walls or other cars. Obviously if such studies were conducted on the entire population, there would be no cars left to sell!

Example A

As a student you are most likely familiar with the most common census there is: the attendance count that takes place each morning. Each class is polled to identify any students who are not present, and the data is compiled in the administrative office.

What likely uses are there for this data, and why might it be collected as a census rather than using a representative sample?

Solution: The most apparent use of this data is to notify parents of students who did not make it to class, for safety and rule enforcement. Since the goal is to locate each and every possibly missing or late student, a statistical sample just isn't acceptable.

Example B

When insurance companies set auto insurance rates, they adjust them according to statistically relevant *demographic* differences among drivers. The process of determining which groups of drivers is the most likely to be involved in expensive accidents is a statistical analysis using police reports and accident claims as data sources.

It is widely accepted that teenage boys are the most expensive demographic to insure, would you expect this information to be based on the population of teenage boys, or of teenage drivers, or of a sample of the appropriate demographic(s), and why?

Solution: The information is based on a sample of the teenage male drivers demographic, compared to a sample of teen and adult drivers in general. It would be virtually impossible to conduct a true census of all accidents involving teen male drivers, as there are just too many and there is no real way to insure that all accidents are correctly documented.

Example C

Suppose your biology teacher wanted to encourage the students in her class to work together on a large project, so she promised the class a pizza party if every single student completed the assigned homework by the deadline. With the deadline fast approaching, you decide to make sure that everyone is on track to get the assignment done on time.



Is this a situation where it would be appropriate to conduct a sample poll of the students, or should you do a full census of all 32 students in the class?

Solution: If you want to be sure that everyone is really on track, you'd better complete a full *census*. A well-chosen sample would give you an idea of how far along the class is in general, but would not be effective at identifying all of the *outliers* which are really the most important data points in this particular study.

Concept Problem Revisited

Suppose you were attempting to put together a menu for a camping trip with a large group of friends and wanted to make sure nobody was allergic to peanuts before planning peanut-butter sandwiches for lunch. Would you need to question all 50+ friends individually? Would it make sense to choose a representative sample to poll instead? What if you wanted to pick a few popular types of soda to bring along, would that be a different situation?

As inconvenient as it might be, you would certainly be well advised to actually ask each and every one of the friends planning to attend the trip about possible peanut allergies. Since even a single person having a severe allergic reaction would probably ruin the trip for everyone, the time saving of a sample poll instead of the complete census would just not be worth the risk.

The soda choice would indeed be a very different situation. Since it is very unlikely that anyone is going to be more than a little inconvenienced by a particular set of drink choices, a quickly generated list of suggestions from a half-dozen people or so would probably be just fine.

Vocabulary

A *population* is the complete set (every single member) of a group of possible items to be studied.

To *poll* the members of a group means to question them regarding a specific topic.

A *sample* or *subset* is a smaller group of members chosen to represent a larger group. Properly chosen, a sample should provide the same results (on a smaller scale) as the population from which it was created.

A *control* group is a set of members deliberately kept as separate as possible from a particular study so as to provide an example of how the members should appear if unchanged.

Bias refers to a desire to achieve a specific result from a particular study, regardless of the data.

Guided Practice

1. A study is to be conducted on the psychological effects of personally witnessing a jewelry store theft from a local mall. Police records suggest that there were a total of 23 witnesses. Is this a situation that would suggest that the entire population be included in the study, why or why not?
2. A new medicine has been developed that the developer claims will stimulate hair growth in balding men. Would you expect there to be safety tests conducted on the population of men before release?
3. The Ford Explorer is a popular sport-utility vehicle sold in the U.S. originally equipped with Firestone tires. In May of 2000, Ford and Firestone were both accused of responsibility in hundreds of vehicle accidents caused by tire failure. Given that all vehicles sold in the U.S. undergo extensive safety testing, how could so many bad products have slipped through?
4. You and your team are conducting a study on the differences in the ability of students in your school to focus during different times throughout the day. Each day your team chooses every 3rd student to walk in the door, and you study 112 students on Monday, 78 on Tuesday, and 109 on Wednesday. If there are 299 students in the school, is this a sample or a population?
5. Why would it be virtually unarguable to state that a product claiming to be “Everyone’s Favorite Soda,” has not been properly evaluated from a statistical standpoint?

Solutions:

1. The relatively small population size in this example certainly suggests that a full census be taken. A shopping mall is likely to contain a rather broad range of demographics, and the 23 witnesses are therefore likely to have many differences in age, sex, background, profession, etc.. Any representative sample taken would probably not be able to accurately represent the full range of possible factors affecting the results of the study.
2. Read the question carefully! In statistics, “population” has a very specific meaning. It would be impossible to conduct safety tests on every man in the world, therefore any safety tests would have to be conducted on a representative sample, not on the population of male humans.
3. There are many ways that the problem could have gone unnoticed. This is a situation where a census study of every Explorer produced is just not feasible; much of the testing simply has to be conducted on a representative sample. Perhaps the sample vehicles used for safety testing just happened to be ones with good tires, or perhaps the safety tests weren’t extensive enough, or the results were incorrectly evaluated.
4. Even though your team collected samples equal to the population of the school, it would still be a representative sample rather than a true census since your random selection method almost certainly resulted in the observation of some students multiple times, and missed others entirely.
5. A population study on every single person in the world is impossible.

Practice

The local public library wants to know if it should increase its hours of operation.

1. How would you want to go about conducting your research? Would you collect a sample or take a census?
2. How would you collect your sample? What time of day would be best to collect the information? Why?

Some college students who were writing a research paper on whether people their age prefer vocal or instrumental music, decide to do so by sampling 100 people at a concert.

3. What is their population?
4. What is their sample?
5. What is wrong with their sample, based on the identified population?

Identify the Population and the Sample

6. In a survey of 1500 American households, it was found that 20% of the households own a computer.

7. In a recent survey of 2578 highschool students, it was found that 28% of them come from single parent homes.
8. The average height of every 6th person entering the movie theatre within a 3 hour period was 5'4".

Identify each scenario as either sampling or census, and identify it as either random or not random.

9. Only 12 tickets are available for over 30 candidates. All their names are thrown into a hat and 12 are pulled out.
10. A student wants to know how many students in school have ever worn a cast. Every student who comes to school that day is handed a short survey that they must turn in before they head to lunch.
11. You ask 30 people in a clothing store which clothing store is their favorite.

Identify the choice that best completes the statement or answers the question.

12. A local business owner wants to find out which benefits plan its employees would prefer. Which of the procedures listed below would be the best way to obtain a statistically unbiased sample?
 - a. Survey a random sample of employees from a list of all employees
 - b. Invite all employees to indicate their choices by email
 - c. Place suggestion boxes at random locations in the company's plant and offices
 - d. Assemble a group with one member from each department and ask them their preference.
13. A simple random sample of 300 people is selected from the 1650 male students in a university business course to take part in a business analysis test. The population being considered is:
 - a. 300
 - b. 1650
 - c. People taking part in the test
 - d. Male students enrolled in a university business course.
14. Which is the best example of an unbiased question?
 - a. Does the school board have the right to enforce a dress code?
 - b. Do you think the principal is doing a good job in spite of his questionable character?
 - c. Do you prefer a daytime or evening class schedule?
 - d. Do you think the government should be allowed to seize whatever property they want to build a new highway?
15. Which question is biased?
 - a. Do you prefer daytime or evening television programming?
 - b. Should there be a school dress code?
 - c. Do you prefer news or mindless sitcoms?
 - d. Do you think a new highway should be built?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 2.2.

2.3 Distribution

Objective

In this concept you will learn about two common uses of the term “distribution” in the study of probability and statistics: the concept of the *normal distribution* in a data set, and of the importance of proper distribution among members of a sample group.

Concept

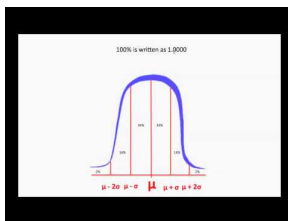
The local county fair is holding a raffle/competition, and the winner gets a \$100 gift card. You would love to win the card, but the competition seems impossible! To enter, you have to guess how many M&M candies of each color: red, blue, yellow, brown, and green, are in a huge jar of M&M’s.

There is certainly no way you can actually count them all, you can’t even see most of them since they are in the center and hidden by the candies on the outside. How could you use statistics to help you make an educated guess at the distribution of the colors? Would it help if you knew there were approximately 650 M&M’s in a pound, and about 5 pounds of candy in the jar?

The answers are found after the lesson.

Watch This

This video describes the *Normal Distribution*, watch it, but do not be concerned if he uses terminology you are not familiar with, we will review it in more detail later.



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62202>

http://youtu.be/GlcEKt9_Hes?t=1m12s The Normal Curve - 0a - Introduction

Guidance

One of the more important goals of a statistical data analysis is to determine the overall *distribution* of the data points. Are the values relatively close together? Do they conform to a specific pattern? Do values tend to occur in groups, or suggest a particular shape? By evaluating the distribution of the data, we not only improve our ability to predict future values, but can also determine how reliable the data is as a model of the real situation.

The most well-known and common distribution in statistics is the **normal distribution**, often referred to as a **bell-curve**. Normally distributed data follows a specific pattern of decreasing numbers of data points as values range further from the **arithmetic mean** (commonly known as the average) of the set. Specifically, in a normally distributed set of data, approximately 68% of all the data points are within 1 **standard deviation** of the mean, and 99.7% of the data lie within three standard deviations of the mean. Don't worry if these terms seem confusing at this point, in subsequent lessons, particularly in the Predicting Values and Normal Distribution chapters, we will be detailing a more rigorous mathematical evaluation of distribution and standard deviation. For now, it is enough to know that if data is normally distributed, approximately $\frac{2}{3}$ (68.2%) of your results should have values within 1 "step" of the average value, and nearly all of your results (95.4%) should be within 2 "steps."

In statistics, **distribution** may also refer to the differences among members of a sample or population. For instance, **demographic distribution** can be a major consideration when choosing subjects for a sample group. When many different members of a population are likely to respond differently to the same stimuli, it is usually important to attempt to maintain the same ratio of such differing responses as that of the entire population.

You will rarely or never collect data from a group made up of identical members, and differences in point of view or personal preference can have surprising effects on experimental results. The more precise you need your results to be, the more important it becomes to monitor the distribution of your sample.

There may be many differences among members of a sample group that influence responses. Common differences include age, size, sex, level of education, religion, culture, geographic location etc. Of course there are many less common characteristics that may affect the results of a study. The goal of a sample is to take into account as many such differences as possible and attempt to represent them in the same ratio as the entire population.

Example A

You are part of a student committee planning to install a vending machine beside the football field to sell snacks for the benefit of the football team. Naturally, you want to stock the machine with the most popular products, so you decide to conduct a poll of the likely consumers. A brainstorming session with the rest of your committee yields the following likely groups of consumers:

- Football players
- Cheerleaders
- Parents
- Coaches
- Male students in the audience
- Female students in the audience
- Reporters
- Sports Scouts

Your committee is split by the debate of how to best take the different demographics into account. Part of the committee believes you should ask equal numbers of each group to list their preferred snacks and drinks, and the other part thinks it would be best to give preference to the players and cheerleaders. Perhaps both groups are incorrect; might it not be most effective to buy items that the students in the audience would prefer?

What do you think? What *distribution* considerations will result in the most sales from the vending machine?

Solution: This problem illustrates the fact that simply representing the demographics of a population as closely as possible may not be the most effective distribution of a sample. Often it is important to identify which statistics are the most valuable to your particular study. In order to maximize sales from the vending machine, it would probably be much more valuable to stock it primarily with items most appealing to the male students in the audience, as they are the most likely to have the desire, freedom, and money to spend on snacks during a game. Additionally, they probably represent the largest single group, followed closely by the female students and then the players and cheerleaders.

Taking these considerations into account, you should identify a sample that is primarily composed of male students, with a smaller number of female students. The other groups are either unlikely to have statistically significant differences in preferences anyway, or are just not numerous enough to be significant.

Example B

In your economics class, you are studying shopping expenditures during the holiday season. The data indicates that the average household will spend approximately \$770 on gifts during the month of November. Assuming the data is normally distributed, and that approximately $\frac{2}{3}$ of all households will spend between \$700 and \$840:



- What is the likelihood that any given household will spend more than \$910?
- What is the chance that a household will spend less than \$700?

Solution: Recall that normally distributed data suggests that $\frac{2}{3}$ of the data points occur within 1 standard deviation of the average, and that 95% occur within 2 standard deviations. If $\frac{2}{3}$ of the households spent between \$700 and \$840, that would indicate that 1 standard deviation represents \$70 since \$700 and \$840 are each \$70 away from the average of \$770.

- Since \$910 is \$140 more than the mean expenditure of \$770, that means that it is $2 \times \$70$ or 2 standard deviations above the mean. We can assume that approximately 95% of *all* values are *less extreme* than \$910, meaning that only 5% will be further than \$140 away from the average. Since half of the remaining 5% of households (2.5%) would be made up of the families who will spend an extremely small amount (less than \$630), we can assume the other **2.5% to spend more than \$910**.
- \$700 is 1 standard deviation below the mean, so approximately $\frac{2}{3}$ of all values are less extreme, and $\frac{1}{3}$ are more extreme. $\frac{1}{6}$ of the values will be more than 1 standard deviation above the mean, and $\frac{1}{6}$ below, so we should expect approximately $\frac{1}{6}$ of the households in the study to spend less than \$700.

Example C

Suppose you are attempting to estimate the demographic distribution of a school football game. Given the size and constant motion of the crowds, you quickly realize that counting them all isn't going to work well. Deciding to

use a random sample instead, you pick a few different groups at random to calculate the average distribution of the crowd.

- If you observe a total of 50 people, and count 22 male students, 17 female students, 9 parents, and 2 others, what would the average demographic distribution of the crowd appear to be as a percentage?
- If sales records indicate a total of 475 tickets sold, what would you estimate the actual count of each demographic group to be?

Solution: a) To calculate the contribution of each group to the whole as a percentage, divide the number of members in each group by the total members you counted:

$$\frac{22}{50} = .44 \rightarrow 44\% \text{ male students}$$

$$\frac{17}{50} = .34 \rightarrow 34\% \text{ female students}$$

$$\frac{9}{50} = .18 \rightarrow 18\% \text{ parents}$$

$$\frac{2}{50} = .04 \rightarrow 4\% \text{ others}$$

b) To estimate the total distribution of the crowd, multiply each group's estimated percentage by the total number of tickets sold:

$$44\% \times 475 = 209 \text{ male students}$$

$$34\% \times 475 = 162 \text{ female students}$$

$$18\% \times 475 = 86 \text{ parents}$$

$$4\% \times 475 = 19 \text{ others}$$

Concept Problem Revisited

Guess how many MM candies of each color: red, blue, yellow, and brown, are in a huge jar.

There is certainly no way you can actually count them all, you can't even see most of them since they are in the center and hidden by the candies on the outside. How could you use statistics to help you make an educated guess at the distribution of the colors? Would it help if you knew there were approximately 650 MM's in a pound, and about 5 pounds of candy in the jar?

If you identify an average ratio of colors in a sample of the candies, you could apply that ratio to the estimated total number of candies in the jar.

If there are approximately 650 candies in a pound, and 5 pounds in the jar, we can estimate a total of approximately 3,250 total candies. To get an average distribution of colors, we could either use a sample of the candies we can see through the side of the jar and calculate the percentage of each, or we could research online to see what the company advertises: 24% blue, 13% red, 14% yellow, 14% brown (16% green and 20% orange, but the raffle doesn't ask about them).

$$24\% \text{ of } 3250 = 780 \text{ estimated blue}$$

$$13\% \text{ of } 3250 = 423 \text{ estimated red}$$

$$14\% \text{ of } 3250 = 455 \text{ estimated each yellow and brown}$$

Of course this is no guarantee of the actual numbers of each color, but given the relatively large sample size, these numbers are likely to be quite a bit more accurate than a simple guess.

Vocabulary

Demographic distribution describes the relative numbers of different types of members of a sample or group.

The **normal distribution (bell-curve)** is a specific type of distribution of data in which the number of data points becomes much fewer as the values stray further from the mean (average) value of the data. Important characteristics of the normal distribution are that it is continuous and symmetric and that the mean = median = mode

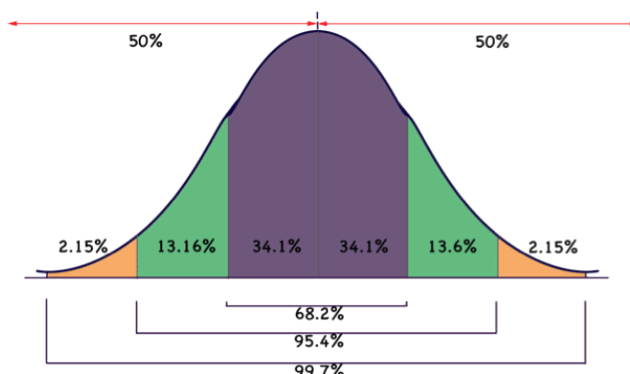
A **standard deviation** is a calculated value representing a specific difference from the mean of a data set. Consider the following: a standard deviation is a statistic that measures the average distance of each data point from the mean of the data set.

Guided Practice

1. Suppose a group of 150 students in a college English course take a final exam, and the instructor calculates that the mean score is 87%, with a standard deviation of 3%. If the scores are normally distributed, what is the approximate probability that a randomly selected score will be between 87% and 93%?
2. In the same class, what is the approximate probability that a randomly selected score will be between 84% and 87%?
3. If the rainfall in Denver during the month of May has a mean of 2.4" and a standard deviation of .4", what is the approximate probability that a randomly selected May will have more than 2" of rain?
4. Assuming the same statistics, what is the approximate probability of receiving between 2" and 3.2"?

Solutions:

1. Recall that normally distributed data indicates approximately 68.2% of values within 1 standard deviation of the mean, and 95.4% within 2 standard deviations. Also, recognize that if 68.2% of values are within 1 SD of the mean, then 34.1% are within 1 SD *above* the mean, and 34.1 are *below* the mean, as you can see in the graphic:



Since the standard deviation of the grades is 3%, there are two standard deviations between 87% and 93%. If we look at the percentages above 87% for the next two standard deviations, we see that the first incorporates 34.1% of the data, and the second incorporates another 13.6%. Therefore the likelihood that a given score will be between 87% and 93% is $34.1\% + 13.6\% = 47.7\%$.

2. 84% is 1 standard deviation below the mean, so the probability that a randomly selected score will be between 84% and 87% is 34.1%.

3. Because normally distributed data have the same mean and median, we can start by noting that only $\frac{1}{2}$ of all months will have a rainfall of less than the median: 2.4%. Additionally, *another* 34.1% will have between 2" and 2.4" of rain, since 2" is one standard deviation away from the mean. That means a total

of $50\% + 34.1\% = 84.1\%$ of months will have more than $2''$ of rain.

4. $2''$ of rain is 1 standard deviation below the mean, and $3.2''$ is 2 SD's above the mean. Since there are 68.2% of values within 1 SD above and below the mean, and 13.6% between 1 and 2 SD's above the mean, there would be $6.8\% + 13.6\% = 20.4\%$ of months with rainfalls between $2''$ and $3.2''$.

Practice

1. Carfax rates its cars annually on customer satisfaction. If Clara researches last years' Mazda, and discovers that it received a mean customer satisfaction rating of 85, with a standard deviation of 4. Assuming the data is normally distributed, what is the probability that Clara herself would give it a rating between 81 and 89?
2. Caleb will be taking a math test tomorrow to make up for the one he missed last week when he was sick. The scores of the students in the class who took it on time were normally distributed with a mean of 84% and a standard deviation of 3%. What is the probability that Caleb will get at most an 81 on the test?
3. Jonah is looking over the final exam scores of the previous year's graduates in the Engineering program from which he is about to graduate. The final exam scores of students were normally distributed with a mean of 70 and a standard deviation of 4. What percentile would Jonah be in if he scores a 78 on the final exam?
4. Scores of each of the previous winners in the state championships for "States Best Chili" were normally distributed with a mean of 74 and a standard deviation of 5. Sarah is competing tomorrow. What is the probability of her winning with a score of between 79 and 84 on her chili?
5. Scores on previous drivers tests taken by 16 year olds were normally distributed with a mean of 82 and a standard deviation of 3.1. George will be taking the driving test tomorrow, what is the probability that he will receive at least an 88.2 on the test?
6. Previous biology test scores were normally distributed with a mean of 76 and a standard deviation of 2.8. Peter will be taking the test tomorrow. What is the probability of Peter getting at most 78.8 on the test?
7. A correlation was found between previous winners of the Noble Peace Prize and their test scores on a standardized test. Every person scoring at least 2 standard deviations above the mean on the test went on to receive a Nobel Peace Prize, and no person with less than that did receive the prize. If the trend continues, and if the standardized test scores were normally distributed with a mean of 89 and standard deviation of 1.4, will Susan go on to win a Noble Peace Prize if she earned a 91.6 on the test?
8. Recent competitors in "Battle of the Bands" received competition scores that were normally distributed with a mean of 89 and a standard deviation of 3.5. "Heavy Metal Trash Cans" will be competing this weekend. What is the probability of the band scoring between 82 and 91.5 in the competition?
9. Tami wants to become a flight attendant but must take a test to do so. Applicants that took the test earned scores that were normally distributed with a mean of 80 and standard deviation of 2.1. Tami will be taking the test today. What is the probability of Tami getting at least 77.9 on the assessment?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 2.3.

2.4 Undercoverage

Objective

In this concept you will learn about two related types of sample bias: *undercoverage* and *self-selection bias*. Both of these common issues with sample selection can lead to very misleading results.

Concept

In 1936, a well-known and highly respected magazine called the *Literary Digest* announced the result of the poll it had conducted on who would be elected president. During prior election years, the magazine had demonstrated remarkable accuracy in predicting the election winner. This time, the magazine predicted that Republican Alfred Landon, Governor of Kansas, would win by a wide margin (57% v.s. 43%) over the incumbent Democrat, President Franklin D. Roosevelt.

Unfortunately for the *Literary Digest*, when the results of the actual election came in, Roosevelt was the victor by a landslide: 62% vs. 38%! Obviously there was a serious problem with the poll conducted by *Literary Digest*, given that the margin of error was an unheard-of nearly 20%. The irony is that the poll was also one of the most ambitious surveys of the type ever conducted. Nearly 10 million people chosen from telephone books, club memberships, magazine subscriptions and other resources had been mailed the survey card, and approximately 2.5 million people responded.

The error was almost entirely due to sample bias, specifically undercoverage of the less-wealthy democratic segments of the population. What caused the bias, and how could the magazine have improved the accuracy of their poll?

After we discuss undercoverage and self-selection bias, and work a few examples, we will return to this question. Can you figure out the answer on your own before then?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62204>

<http://youtu.be/R2vhjC5qCQk> Election: Biased sampling procedure

Guidance

There are many different types of *sample bias*, any of which can skew the results of an experiment or survey. *Undercoverage* is one common type, referring to a sample with too few examples of one or more segments of the population it is meant to represent. In some cases, particularly where the under-represented group is quite small in comparison to the others in the entire population, undercoverage may not have much of an effect. However, if the

undercovered segment is significant enough, the results of the sample may not accurately estimate the characteristic of the population.

Self-selection is related to undercoverage, and can actually be the cause of it. Self-selection refers to the policy of asking voters to submit responses on their own, rather than collecting the answers from them. The problem with self-selection is it limits the voters to those with the time and inclination to respond (known as **non-response bias**), which reduces the overall sample size, and also skews it toward the type of person who believes in the value of taking time to respond to polls!

Example A

You are assisting with a study attempting to determine the satisfaction of school communication with students who speak a second language at home. The plan is to send home a questionnaire to the parents of the students, asking them about their opinion.

What kind(s) of bias is this survey method particularly prone to? How might they be addressed?

Solution: This method of sampling is liable to result in both non-response and undercoverage bias. Non-response bias is an issue any time a sample population is expected to submit a questionnaire, as your results are going to include more input from the type of person who is willing and able to complete and submit your survey. In this case, undercoverage is a particular problem, since the population most affected by the study is also unusually liable to misinterpret the questions or the reason for them due to the language barrier.

One possible solution might be to conduct a phone survey conducted by a native speaker in the target language(s).

Example B

What type(s) of bias do the experiments below suggest?

- An experiment to determine the danger of mixing household chemicals is conducted by collecting samples of chemicals found under the experimenter's sink.
- Mall shoppers are asked to fill out and return a form rating their shopping experiences at each of the 26 stores to identify the most popular stores in each of 4 categories.
- A study of the average grades of mathematics students polls 16 Algebra I students, 14 Geometry students, 7 Calculus students, and 19 Statistics students.

Solution:

- Undercoverage bias - This experiment is a prime example of the problems associated with **convenience sampling**, since the only chemicals used were the ones conveniently found in one location, the results could not be assumed to be the same as with chemicals found under other sinks.
- Non-response bias - Since the results are dependent on the shoppers turning in a response form on their own, the results will be biased toward a specific type of personality, and will not reflect a true cross-section of shoppers' experiences.
- Undercoverage - The study only includes approximately $\frac{1}{2}$ as many Calculus students as the other subjects.

Example C

There is a commonly referenced story about the difficulties of marketing products internationally, related to the Chevy Nova automobile. According to the story, the Chevrolet motor company lost millions over an attempt to sell the popular U.S. vehicle in Mexico without noting that "No-Va" means "No-Go" in Spanish!

The truth is that the story is just an urban myth, and that the Nova sold well in Latin America, but the caution is valid nonetheless. If the situation had occurred as described, what sort of bias might have been the culprit in Chevy's market research that could have led to the misunderstanding?

Solution: It is certainly reasonable to suspect that undercoverage might have been a contributing factor here. Any studies or market research that Chevy conducted in the United States about the popularity of the name "Nova" would have included far more native English speakers than Spanish speakers.

Concept Problem Revisited

In 1936, the Literary Digest predicted that Republican Alfred Landon, Governor of Kansas, would win the presidential race by a wide margin (57% v.s. 43%) over the incumbent Democrat, President Franklin D. Roosevelt. When the results of the actual election came in, Roosevelt was the victor by a landslide: 62% v.s. 38%! The error was almost entirely due to sample bias, specifically undercoverage of the less-wealthy democratic segments of the population.

What caused the bias, and how could the magazine have improved the accuracy of their poll?

The bias was caused by the magazine's method of sampling. Choosing the voters by telephone listing (remember that phones were much more of a luxury in 1936!), club membership, and magazine subscribers resulted in a bias toward the wealthier members of the population. Perhaps a door-to-door poll in some of the lower-income areas of the country would have provided some valuable insight. At a minimum, the magazine could have at least issued a statement regarding the possible bias in the survey due to the limited range of incomes targeted.

Ironically, the uncommonly large size of the sample actually made the bias worse, since there was a huge number of responses from the wealthier demographic, overshadowing the limited number of other responses. Had the study been a bit more limited in size, the fewer other responses might not have been so drastically outnumbered, particularly if the smaller study were conducted in a more balanced area.

Vocabulary

Convenience sampling refers to the process of choosing a sample based on members who are easily accessible.

Self-selection is a sampling method that requires the subject to offer a response to an input.

Non-response bias is commonly caused by self-selection, subjects with a reason not to respond which may be unrelated to the actual study are not included, skewing the results.

Undercoverage describes a sample with too few members of a given group or demographic.

Sample bias refers to a sample with a non-random distribution of members.

Guided Practice

1. If a sample of 100 high school students indicated that 78% thought the most important class in a high school curriculum was "Woodworking", what might you suspect about the chosen sample?
2. If a study posted results indicating that only 1% of polled students liked football, what bias is likely to have affected the sample selection?
3. Suppose "Super-Sugar" cola company indicated that every person polled who preferred "Super-Sugar Cola" over all other brands of soda was a multi-millionaire. What type(s) of sample selection bias would you suspect that might prevent you from running right out to buy a case of "Super-Sugar" so you could become a multi-millionaire?

Solutions:

1. It would certainly appear that the sample was not a likely cross-section of the average public school. It is a good bet that the female population was undercovered during the sample selection process.
2. Obviously the athletic students were undercovered in this sample. Maybe this study was conducted using the students who weren't polled during the study referenced in question 1!
3. This is an example of "cherry-picking", a sampling technique where only very specific people are polled to insure a particular appearance for the results. If "Super-Sugar Cola" only sampled multi-millionaires, then *any* person who preferred their drink would be a multi-millionaire. Obviously this method would also create an undercoverage bias, since the less-wealthy soda drinkers were not included in the sample.

Practice

Discuss how undercoverage could be a source of bias in each of the following surveys:

1. A poll showed that 85% of respondents believe that teens make better drivers than adults.
2. The U.S. census of 1980 states that 32,194 Americans are 100 years old or older. However, Social Security figures show only 15, 258 adults of this advanced age (Los Angeles Times, Dec. 4, 1983)
3. In a census in Russia, 1.4 million more women than men reported that they were married (U.S. News & World Report, Aug. 30, 1976).
4. To find out how important the clothes of vice-presidential candidate might be, researchers ran a survey shortly after the 1984 Democratic convention in three locations: the Wall Street area of New York City, State Street in Chicago, and Crown Center in downtown Kansas City. The 347 respondents were shown pictures of women wearing three outfits, and the pictures did not show the women's faces. Then the respondents were asked several questions about how the outfits affected respondents' feelings of competence regarding the model serving in a public office (Los Angeles Times, Aug., 3, 1984). 310 respondents indicated that the color and fit of the outfit was important in creating feelings of competence.
5. One year after the Detroit race riots of 1967, interviewers asked a sample of residents in Detroit if they felt they could trust most of their neighbors, some of their neighbors, or none at all. In one sample, 35% answered "most"; in another sample, only 7% answered "most".
6. In a comment on deregulation of banking, "[the head of California's Security Pacific Bank] reckons the higher interest accounts, and all the other new financial services, are designed for the most affluent 15% to 20% of Security Pacific Bank's customers. By extension—as 2million customers are surely a sample of the general population—the new world of deregulated finance benefits the top-earning 15% to 20% of U.S. households" (Los Angeles Times, Dec. 4, 1983).

In the following scenarios, identify if we are dealing with a sampling or a nonsampling error. In each case, be as specific as possible about the source of error. Would this type of error result in bias?

7. In a telephone survey that randomly selects participants, we try to contact a person five times and he/she never picks up the phone.
8. An interviewer chooses people on the street to interview regarding their preference for walking v.s. driving.
9. The police department of Lexington would like to know more about people's opinion about their police force. They send an officer in uniform to randomly selected households, but many of the selected households refuse to participate.
10. A survey asks the question "Do you agree with the U.S. Supreme Court's decision that corporations are allowed to spend huge amounts of money to sway elections in their favor?"
11. In a survey that would like to measure the overall health of college students, including the prevalence of sexually transmitted diseases, some participants are not willing to admit that they have contracted such a disease.
12. In Fayette County, 53.8% of registered voters are registered as Democrats. However, in a SRS of 200 registered voters, only 45% of them are registered as Democrats.
13. An interviewer enters all the information into a database during the interview, and accidentally records that a person has 22 children, instead of 2.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 2.4.

2.5 Response Rates

Objective

In this concept you will learn about *response rates*, a measure of the relative number of actual responses received from a statistical evaluation.

Concept

It is commonly accepted that any survey conducted by mail, or over the internet, or by telephone will have a very low response rate. It is not unheard of for such surveys to have less than 5% of the chosen sample actually return usable results.

It certainly seems logical to attempt to get a response rate as high as possible, but does a low response mean that the experiment is invalid? The answer may surprise you, look for it after the example questions below.

Guidance

Response rate is expressed as the percentage of total attempts made at data collection divided by the number of successful responses. If you conduct a survey of 50 students, and 10 of them choose not to participate, then you would say that you had a *response rate* of $\frac{40}{50} = .80$ or 80%.

There is no actual widely accepted formula for determining an acceptable response rate. Although it is logical and generally accepted that a higher response rate is better, there is quite a bit of debate over the attention that is generally paid to response rate as a determining factor in the validity of a particular poll. In fact, during a study of the importance of a high response rate, a group of researchers compared the accuracy of predictions made from data collected over 15 years by telephone (with a response rate of 60%) and from data collected by mail (20% response rate). It turned out that the *mail* survey, even with the much lower response rate, was more than 3 times as accurate in predictions as the telephone survey was! *(Visser, Krosnick, Marquette, and Curtin - 1996) However, other studies have shown that the results may also go the other way, with a low response rate directly resulting in poor accuracy in the results.

What all this tells us is that it is important to keep track of response rates in order to have as much data as possible to use in determining the accuracy of your results, but that even a poor response may not mean that the data is incorrect. What is ultimately important is to limit or offset the difference that any missing responses may have on your results.

A common approach to evaluating the effects of non-responders is to return to a small sample of members who initially did not respond, and make a concerted effort to get responses from them to see if the results appear significantly different from the data collected from the main group.

Example A

What is the calculated response rate from a survey attempting to find the most popular dog food brand of a sample of 130 dogs, if 24 of them did not eat any of the brands at all?

Solution: Response rate is calculated as:

$$\frac{(\text{number of successful collections})}{(\text{total attempts at collection})} = \frac{106}{130} = .815 \text{ or } 81.5\%$$

Example B

If you are conducting a study of the effects of music on homework study, and you have determined that you will need a minimum 70% response rate in order to have enough data to accurately determine the results, how many responses will you need to receive out of the 53 students you originally study?

Solution: Multiply the response rate by the total number of attempts to find the minimum number of successful collections:

$$.70 \times 53 = 37.1$$

Since the minimum stated response is 70%, and since we cannot collect .1 results, round up to 38 successful data collections.

Example C

While conducting a study of the most popular foods to serve in the school cafeteria, you collect the following data:

- Students in the school: 425
- Students not present in the cafeteria for questioning: 225
- Students preferring pizza: 130
- Students preferring hamburgers: 60
- Students preferring pasta: 10

Use this data to answer the following:

- What is the calculated response rate of the study?
- What is the most popular food?
- Based on the *response rate*, would you consider the results reliable?
- Based on *all available information*, would you consider the results reliable?

Solutions:

- The calculated response rate is $\frac{200}{425} = .471 = 47.1\%$
- Pizza: $\frac{130}{200} = .65 = 65\%$ of responders
- The response rate suggests that less than half of the students in the school submitted a vote, so the results would seem suspect.
- The 47.1% response rate actually represents 100% of the students who eat in the cafeteria, since the students not available for questioning went elsewhere for lunch. Considering this, the results were likely quite reliable.

Concept Problem Revisited

It is commonly accepted that any survey conducted by mail, or over the internet, or by telephone will have a very low response rate. It is not unheard of for such surveys to have less than 5% of the chosen sample actually return usable results.

It certainly seems logical to attempt to get a response rate as high as possible, but does a low response mean that the experiment is invalid?

No, it doesn't. Certainly a very low response rate should be investigated, but the more important consideration is how well the results actually collected represent a random sampling of the population under study.

Vocabulary

A **population** is the complete set (every single member) of a group of possible items under study.

To **poll** the members of a group means to question them regarding a specific topic.

A **sample** or **subset** is a smaller group of members chosen to represent a larger group. Properly chosen, a sample should provide the same results (on a smaller scale) as the population that created it.

A **control** group is a set of members deliberately kept as separate as possible from a particular study to provide an example of how the members should appear if unaffected by the study in any manner.

Bias refers to a desire to achieve a specific result from a particular study, regardless of the data.

Guided Practice

1. What is the response rate of a survey of 368 persons, if 312 responses are collected?
2. How many responses would you need to collect to achieve an 82% response rate from a survey of 1250 people?
3. Is 428 responses sufficient to achieve an 80% response rate from a sample of 475 persons?
4. What range of responses would you need to achieve a response rate between 70% and 75% from a survey of 24,000?

Solutions:

$$1. \text{ Response Rate} = \frac{\text{Number of responses collected}}{\text{Number of responses attempted}} = \frac{312}{368} = .847 \text{ or } \approx 85\%$$

$$2. .82 = \frac{\text{Number of responses collected}}{1250} \rightarrow .82 \times 1250 = \text{Number of responses collected} \rightarrow$$

$$1025 = \text{Number of responses collected}$$

$$3. \frac{428}{475} = .90 \text{ or } 90\% \text{ so, yes, it is sufficient.}$$

$$4. 70 \times 24,000 = 16,800 \text{ and } .75 \times 24000 = 18,000. \text{ You would need between 16,800 and 18,000 responses.}$$

Practice

Look at the following table taken from a report located on the: <http://www.educause.edu> website.

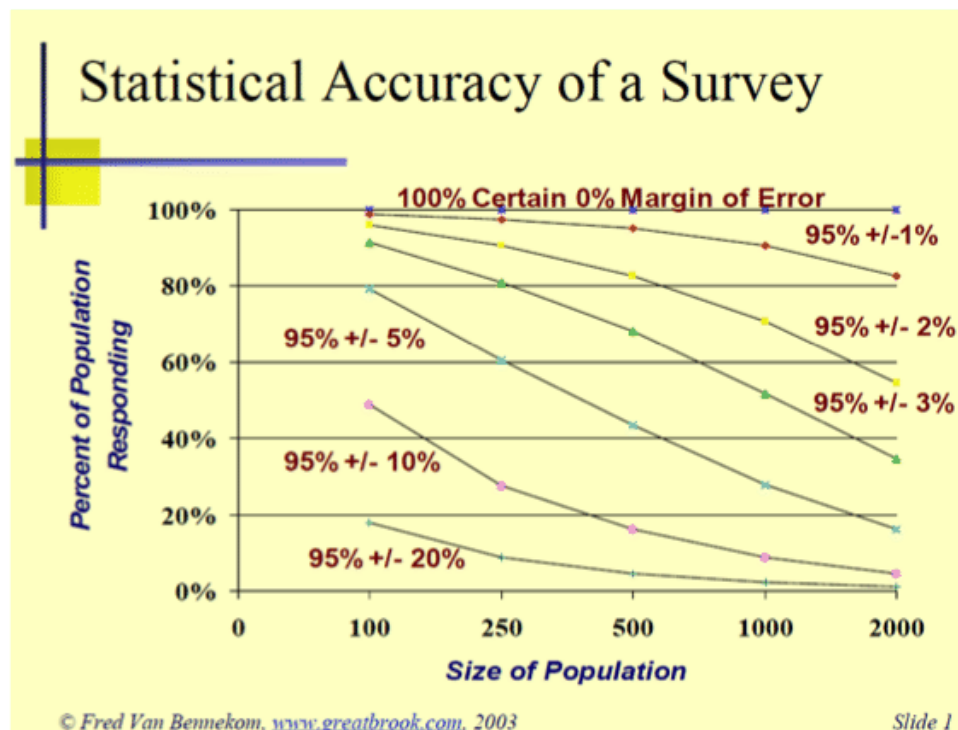
How satisfied are you with the number of places where you can sit and work on your laptop/netbook on campus?*		
	First 18% of Total Sample (284 of 1,598)	All Other Participants
Very dissatisfied and dissatisfied	45 (16%)**	58 (16%)
Neutral	67 (23.5%)	121(33.3%)
Satisfied and very satisfied	172 (60.5%)	184 (50.7%)
Total n for column	284 (100%)	363 (100%)

1. How many people total were polled?
2. How many responded in the first go around? What response rate does this represent?
3. How many responded in the second go around? What response rate does this represent?
4. What is the overall response rate?
5. Did the response rate appear to matter, if you compare column 1 and column 2 responses and percentages?
6. What are the satisfaction ratings, should the data from the two columns be combined?

1. Very Dissatisfied and dissatisfied:

2. Neutral:
 3. Satisfied and Very Satisfied:
7. Should there be an additional column representing the population that did not respond? Create a chart showing percentages based on total population polled including non-responsive people polled.
1. Very Dissatisfied and Dissatisfied:
 2. Neutral:
 3. Satisfied and Very Satisfied:
 4. No Response:
8. “One type of survey error arises from what is known as *nonresponse bias*. This is the idea that what one learns from those who did respond may differ significantly from what might be learned from those who did not respond.” Based on this quote and the intention of the survey, what are some reasons that so many people may have been unresponsive? What do you think the high percentage of “No Response” could tell us in this study?
9. Based on data gathered in question 6 and question 7, which information in this scenario do you think would be more important to the campus that conducted the research?

Use the Graph below to answer the questions that follow.



10. What increases as population size increases regardless of the percent of responding population?
11. If you expect to have a low response rate, what do you need to increase in order to improve the accuracy of your study?

An example of how to use the chart to determine accuracy:

You have a population of 1000, and you were able to contact via phone 500 of those people. You were able to get responses out of half of those you actually spoke to. So 25% of the total population participated in your study. Using the chart above, find the intersection of 1000 on the horizontal axis and 25% on the vertical axis. You would be approximately 95% certain of +/- 5% accuracy in your survey results.

12. You own a coffee shop. You have about 750 regulars that come in and drink coffee at your coffee shop. You want to know if you should increase the speed of your wifi connection. 500 of your customers agree to take the survey home and complete it. Only 300 of the customers returned their survey.
 1. What is your population size?
 2. What is your sample size?
 3. Based on participation, what can you expect the accuracy of your findings to be?
13. As the manager of a ski resort, you are always looking for new ways to improve your customer's experience. You are thinking of adding an inner tubing/sledding slope right outside your hotel. At peak season, you have 2000 people staying in the hotel. You ask everyone who checks in to please complete the survey, but 356 people decline. At the end of the survey period, you have received 693 surveys.
 1. What is your population size?
 2. What is your sample size?
 3. Based on participation, what can you expect the accuracy of your findings to be?

Standing outside a mall that reports a minimum of 500 people per day walking through the doors, you attempt to complete some research for school. Your report will require an accuracy of +/- 10%.

14. How many of the people walking through the door will you need to respond in order to achieve this accuracy?
15. If we also know that we need a 25% response rate, what would our sample size need to be?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 2.5.

This chapter reviewed the concept of using a sample set to represent an entire population when collecting statistical data. Students should have gained or improved an understanding of various potential difficulties caused by inexpertly chosen sample sets.

2.6 References

1. Vancouver Film School. <https://www.flickr.com/photos/vancouverfilmschool/4422241701/> .
2. Joe Shlabotnik. <https://www.flickr.com/photos/joeshlabotnik/2207112346/> .
3. alexkerhead. <https://www.flickr.com/photos/alexkerhead/3127788202/> .
4. Casey Fleser. <https://www.flickr.com/photos/somegeekintn/2577313610/> .
5. Pizza Masetti Craiova. <https://www.flickr.com/photos/pizzamasetti/4932649560/> .
6. Jorge Franganillo. <https://www.flickr.com/photos/franganillo/3678747186/> .
7. Jennifer C.. <https://www.flickr.com/photos/29638108@N06/5220659212/> .
8. CK-12 Foundation. . CCSA
9. . <http://www.educause.edu/ero/article/using-research-based-practices-increase-response-rates-web-based-surveys> . CCSA
10. CK-12 Foundation. . CCSA

CHAPTER **3** Common Types of Samples

Chapter Outline

- 3.1 IDENTIFYING RANDOM SAMPLING**
 - 3.2 IMPLEMENTING RANDOM SAMPLING**
 - 3.3 STRATIFIED SAMPLING**
 - 3.4 CLUSTER SAMPLING**
 - 3.5 NON-PROBABILITY SAMPLING**
 - 3.6 REFERENCES**
-

There are many different methods of sampling, each ideal for particular types of data collection. In this chapter, you will learn how to identify and implement the best method(s) for your own needs.

3.1 Identifying Random Sampling

Objective

Here we will build on our understanding of *random sampling* and how to apply it to choose members of an effective statistical sample.

Concept

Suppose you were handed a bag full of a large number of small, unknown items, and asked to try to identify the contents without looking in the bag. Naturally the easiest way to investigate would be to reach inside and grab a handful to see what turned up. If you reached in and then pulled out a handful of nothing but marbles of various colors off the top, you'd quite reasonably state that this must be a bag of marbles.



Is it possible that you might be wrong? How? What could you have done to improve your guess?

Watch This

Implications...

Definition of random sample of size n : every sample of size n is equally likely to be selected.

MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62206>

<http://youtu.be/YtGVor-pIE4> Teachertubemath - Random Sample

Guidance

A random sampling is easily the most commonly known method of choosing representatives of a population to comprise a statistical sample. Since the most likely error in developing an accurate statistical sample comes from some sort of personal *bias*, careful researchers try to allow as little influence from personal decisions to affect the choice of the sample members as possible. The easiest way to avoid conscious or subconscious bias is to use a random method of selection. However, true random sampling may be more difficult than it appears. A good definition of random sampling is: “A sample consisting of individuals each chosen entirely by chance, in such a way that, at every stage of the process, every potential member of the sample has the same probability of being chosen as every other member.”

In order to have a truly random sample, all of the factors involved with the choice of representatives must be entirely random processes, and out of the control of any influence which might have any sort of bias. A *simple random sample* is the process of assigning a number to each member of the population under study, and then using a random number generator to pick the samples. In subsequent lessons, we will learn how to individually identify and implement other types of random samples, including: Stratified Random Samples, Systematic Random Samples, and Clustered Random Samples.

Example A

In order to randomly pick 15 students out of a class of 210, a researcher enters the names into a spreadsheet and numbers them 1-210. Using a random number generator, he picks 15 numbers between 1 and 210, and uses the names associated with those 15 numbers.

Is this a random sample?

Solution: Yes, this is a random sample. In fact, this is probably the most common method of randomizing the choice of names from a list. By converting the names to numbers, the researcher minimizes the chance that he/she will be influenced by a subconscious preference for a particular name. Further, by using a random number generator to pick the 15 numbers, he keeps his own thoughts away from the actual number selection process.

Example B

A researcher is conducting a survey on how many students at a particular high school go off-campus for lunch. He decides to pick every 5th person in line at the school cafeteria and ask how often he/she goes out to get lunch instead of eating at the cafeteria.

Is this a random sample? Is it appropriate for the study?

Solution: This is a random sample, but not of the correct population. The concept of ‘every 5th person’ is convenient and effective as a method of choosing random students from the line, but choosing students from the lunch line at school excludes the students who are *already* off campus for lunch!

Example C

Evan is conducting a study of the types of seashells on the 2-mile stretch of beach near his home.

How might he choose a random sample to represent the population of seashells on the beach? Would simply walking around without aim and picking up seashells be effective? Why or why not?



Solution: Just walking around without direction will not actually result in a random sample, since Evan may have conscious or unconscious preferences that would affect where he would choose to walk. Also, 'just picking up' seashells would further influence the sample with his own preferences since he would be picking up the shells he chooses.

If Evan wanted to randomize the sample, he could do so with a set of dice. He could roll the dice before leaving, and use the number to set a distance to travel before stopping. He could pick up any seashells within reach at the stop, and roll the dice again. If he turned right on even rolls, and left on odd rolls, the results would be much more random.

Concept Problem Revisited

If you reached in and then pulled out a handful of nothing but marbles of various colors off the top, you'd quite reasonably state that this must be a bag of marbles.

Is it possible that you might be wrong? How? What could you have done to improve your guess?

It is entirely possible that you are correct, but it is also pretty clear that the results would be suspect at best. By only grabbing a single handful of items off the top of the bag, you are limiting your results to items that worked their way to the top. There may well have been other items at the bottom of the bag that you did not get a sample of by just grabbing off the top. You could certainly improve the reliability of the sample by first shaking the bag well so that your sample is appropriately randomized.

This is an example of *convenience sampling*, which is selecting a sample more because it is easily obtained than because it is random or representative.

Vocabulary

A **sample** is a smaller portion of all of the possible members of a given set. Chosen properly, a sample can provide a good approximation of the overall behavior of the entire group.

Bias is preference or favoritism, conscious or unconscious.

In a truly **random sample**, each and every member of the population under study must have the same chance of being selected.

Guided Practice

For each question below, first decide if the example describes a random sample, second describe why you believe it is/isn't random.

1. Several apples are selected from each bin of different types at the market.

- Each student is assigned a number and a die is rolled. Starting with that number each 4th student is chosen until the quota is met.
- To build a sample of students in his state, Brian first made a list of school districts and assigned each a number. He then randomly chose 4 districts and assigned each school in those districts to a number and randomly chose 3 schools from each district. Once he had the schools, he assigned numbers to each student in each school, and used a random number generator to pick 15 students from each school.
- Katrina closed her eyes and reached into the beverage stand, randomly selecting a soda.

Solutions:

- This is probably not a random sample. The question does not specify how the apples are chosen from each bin. You would have an unconscious preference for shinier apples if you just selected one from a bin.
- This is a valid random sampling, specifically a *systematic random sample*. Each student has the same opportunity to be chosen as any other.
- This is a valid random sampling method, specifically a *multi-stage cluster sample*. Each student in the state has the same opportunity to be chosen for the sample.
- This is not a random sample. Katrina would be more likely to pick one within reach than to select from the top or bottom, and would be unlikely to reach behind the front sodas to select one further back. Some sodas have a greater chance of being chosen than others.

Practice

State whether each example describes a random sample:

- Consumer Reports* is conducting a poll to evaluate consumer opinions on the latest console gaming platform. The publisher includes a postcard in an issue of *Consumer Reports*, and asks readers to fill out the form and return it.
- Kelly closes her eyes, opens the dictionary and randomly points to a word on the page. She repeats the process 10 times.
- Shake a bag of marbles well, reach in and select a marble with eyes closed.
- Number the students in your class 1-32, and roll six standard dice to choose a random student, discarding any number greater than 32.
- Walk along a pebbled path and randomly pick up a stone.
- Roll standard 4 dice and choose a name starting with the letter associated with that number if the alphabet is numbered 1-26.
- Walk around a park and randomly pick up fallen leaves to see what kind of trees are in the park.
- Search for “dog” on Google to get a random sample of popular dogs.
- Ask shoppers in a hardware store if they know how to replace a light switch to learn the percentage of people that know how.
- Spin a spinner numbered 1-25 to decide how much to charge for a glass of lemonade.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 3.1.

3.2 Implementing Random Sampling

Objective

Here you will learn a few methods of randomizing a sample to obtain a representative set that can be used to project results back to the full population.

Concept

You probably know by now that you need a truly random sample of a population if you want to use the results to predict the responses of the entire population, but do you know how to randomize a particular list of data? If you were given a list of student names or addresses, could you build a random sample from it?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62208>

<http://youtu.be/jUBtNBVy42M> Shawn Burke - Simple Random Sampling

Guidance

Obtaining a true random sample is more complex than most people think, but there are a number of viable ways to build a random sample from different types of populations. In general, the first and most important step is to organize the identities of your population into a set that can be numbered, since numbers are relatively easy to randomize.

Once each member of your population is assigned to a unique numeric identifier, there are a few ways to choose your actual sample:

- **Simple Random Sample:** Use a random number generator or a random number table to identify the members of your sample directly from your list. If you get a number from the random source that does not directly correspond to one on your list, pick another until you do. Continue the process until you have enough members of your sample. (Example A)
- **Systematic Random Sampling:** First divide the total number of members in the population to be sampled by the number of members you want in your sample. The result is your step size. Use a random generator to identify a starting number, then skip down from the starting number by your step size and pick the result, skip down again and pick, until you get as many results as you need for your sample. (Example B)
- **Stratified Random Sampling:** This method is most effective when sampling a population with a limited number of disparate sub-groups formed by members' shared attributes or characteristics. Start by dividing

the population into the recognized subgroups, and then use the *simple random sampling* method above to either:

- Pick the same percentage of your sample as each subgroup represents of the population (proportional random sampling). This method ensures that the sample space will have the same proportion of each subgroup as the population. OR
 - Pick an equal number of units from each subgroup until you fill your sample space (disproportionate stratified random sampling). This method may be used particularly if one or more subgroups are too small to result in a useable sample when the proportional method is used. To yield accurate results, you will need to apply proper *weighting* to your results.
- **Cluster Sampling (multistage):** This method is most applicable to very large populations spread over large geographic areas. For instance, if you wanted to sample individual classrooms randomly sampled from the entire Midwest United States, you might first collect a random sample of states in the area, then a random sample of districts from each sampled state, then a random sample of schools from each sampled district, and finally a random sample of classrooms(s) from each sampled school. It is important to use a random sampling method at each stage.

Example A

How would a simple random sample be implemented in the following situation?

Ciere wants to collect a random sample of 25 students out of the 146 members of the Junior Class at school.

Solution: To collect a *simple random sample*, Ciere should assign a unique number to each student in the class and use a random number generator to pick 25 numbers between 1 and 146. Each student associated with one of the chosen numbers becomes part of the sample.

Example B

Howard wants to collect a sample of 15 dogs from the local animal shelter for his study on most commonly surrendered dog breeds. If he decides to use a *systematic random sample*, what would the process be?

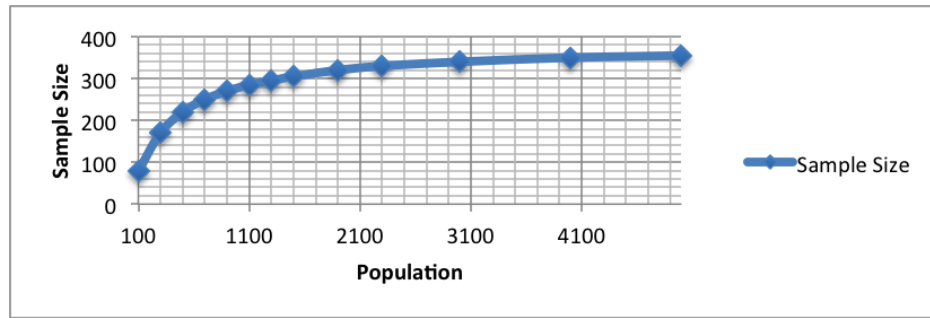


Solution: Howard should obtain a list of all the dogs in the shelter from the front office. He should identify the number of dogs, and assign a unique number to each one. He should next divide the total number of dogs in the shelter by 15 (the number he wants in his final sample) to get n (the step size). Finally, he should roll a couple of dice and start by counting down from the top of the numbered list of dogs until he reaches the number he rolled. From that point, he should take every n^{th} dog until he reaches 15 dogs.

Example C

Erina is considering the possibility of opening a small sporting goods store just down the street from the local WalMart. Her calculations suggest that if at least 13% of the estimated 4000 adult customers visiting WalMart each day spend money on sporting goods, then her store has a good chance of being profitable based on providing those

customers an alternative. Erina plans to use the graph below to help determine the sample size she will need. How might she arrange a random sampling of the customers to get the data for her survey?



Solution: This is another good situation for a systematic random sampling. Erina's population in this case is expected to be approximately 4000, which is the estimated daily number of customers at her WalMart. We can estimate based on the graph that for a population of 4,000, she will need a sample of 350 people.

To find the step size (n), we divide: $\frac{4000}{350} \approx 11$

Erina should roll a pair of dice, let that number of people go by, then question every 11th person out the door to see if they purchased sporting goods.

To increase the accuracy of her predictions, Erina should repeat the process on different days of the week, and ideally in different seasons (which might not be reasonable depending on her time frame), to see how the purchase types cycle.

Concept Problem Revisited

If you were given a list of student names or addresses, could you build a random sample from it?

Having worked through this lesson, you probably know that there are a number of ways to build a good, random sample. Likely the most efficient option since you already have a list would be either a simple random or systematic random sample.

Vocabulary

A **random sample** should be conducted such that each and every member of the population has the same chance to be selected as any other. At each stage of a multi-stage selection process, each subset of units should have the same probability of being chosen as any other subset.

Guided Practice

Which random sampling technique (Simple, Systematic, Cluster, Multi-Stage, or Stratified) would be most appropriate in each of the following situations?

1. Finding the ratio of girls vs. boys in a high school with 400 students.
2. Identifying the favorite soda of students at a football game.
3. The preferred biscuit of each of the different dog breeds at a Humane Society.
4. The average ACT or SAT score of 500 students from rural schools in the State of Colorado.

Solutions:

1. A simple random sample would work here. Number the students 1 - 400 and use a random number generator to find 50 or so students for your sample.

2. Students often like to drink the same as their friends, so a cluster sample of 2 or 3 students from each of several groups of students would be a good method. Don't forget to use a random number generator to choose the groups and the members of each group.
3. Since you want to specifically evaluate the favorite biscuit of *each breed*, you should do a stratified sample. Identify the number of breeds, and the number of dogs of each, then use a random method of choosing a number of representatives of each breed.
4. This would be a multi-stage sample. First you need to select only students from rural schools from the entire population of CO students (perhaps by randomly selecting a limited number of rural schools). Second, you should number the subset of students individually and divide the total by 500 to obtain your 'step size' n . Third, use a random generator to select a starting number and choose every n^{th} student until you reach 500.

Practice

Identify appropriate sampling methods to collect samples in the following situations:

1. Identifying the contents of an opaque bag.
2. Estimating the percentage of shoppers who use plastic bags at a specific store.
3. Estimating the percentage of students in your school who like vampire movies.
4. Estimating the percentage of dogs that bark at guests and passing cars.
5. Predicting the most popular book genre among students in your class.
6. Predicting the most popular book genre among U.S. elementary school students.
7. Estimating the percentage of defective incandescent bulbs produced at a factory.
8. Estimating the percentage of defective light bulbs produced at a factory.
9. Estimating the percentage of defective incandescent bulbs produced at U.S. factories.
10. Estimating the percentage of teachers at a school that work more than two Saturdays per month.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 3.2.

3.3 Stratified Sampling

Objective

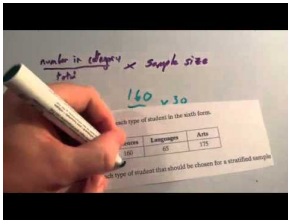
Here you will practice identifying and implementing *stratified random sampling*.

Concept

Suppose you wanted to find out if age influences the choice of classes for students at a particular university. You might divide the students up by age ranges such as: Under 18, 18 - 21, 21 - 25, 25 - 35, and 35 and over. How could you make sure a random sample of college students would have members of each age range?

Look to the end of the lesson for the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62212>

<http://youtu.be/fEaP-3ocMnQ> corbettmaths - Stratified Sampling

Guidance

Stratified random sampling is an excellent method of choosing members of a sample when there are clearly defined **subgroups** in the population you are studying. Each subgroup, called a **stratum** (strata if plural), should have a clearly defined characteristic that separates the members from the rest of the population.

To implement stratified sampling, first find the total number of members in the population, and then the number of members of each stratum. For each stratum, divide the number of members by the total number in the entire population to get the percentage of the population represented by that stratum. Finally, take the percentage and multiply by the number of units you want in your final sample group to see how many you need from each stratum. Always round any decimals *upto* whole units assuming you cannot take half of a sample.

As a formula, this process looks like:

$$\frac{\text{\# of members in the stratum}}{\text{total members in population}} \times (\text{desired sample size}) = (\text{\# of members from stratum})$$

Example A

How many Blue Heelers would you need for a stratified sampling of 50 dogs from a population consisting of:

- 247 Collies

- 138 Pit Bulls
- 96 English Mastiffs
- 172 Blue Heelers
- 222 Welsh Corgis

Solution:

First identify the total number of dogs in the population:

$$247 + 138 + 96 + 172 + 222 = 875 \text{ dogs}$$

Then divide the number of Blue Heelers by the population count:

$$\frac{172}{875} = .197 \text{ or } 19.7\%$$

Finally, multiply this number by the desired sample size:

$$.197 \times 50 = 9.85 \rightarrow \text{rounds to } 10 \text{ Blue Heelers}$$

Example B

How many members would you need from each age stratum to obtain a stratified sample of 350 from the following population?

TABLE 3.1:

Age	Count
15yrs to 18yrs	297
18yrs to 21yrs	349
21yrs to 27yrs	323
27yrs to 35yrs	240
35yrs to 42yrs	191

Solution:

First find the total population count:

$$297 + 349 + 323 + 240 + 191 = 1400$$

Then divide the count of each stratum by the total to get the percentage:

TABLE 3.2:

Age	Count	%
15yrs to	297	$\frac{297}{1400} = 21.2\%$
18yrs to	349	$\frac{349}{1400} = 24.9\%$
21yrs to	323	$\frac{323}{1400} = 23.0\%$

TABLE 3.2: (continued)

27yrs to	240	$\frac{240}{1400} = 17.1\%$
35yrs to	191	$\frac{191}{1400} = 13.6\%$

Finally, multiply the percentage of each stratum by the desired sample size:

- 15 - 18yrs: 21.2% of 350 = 74.2 → round to **74**
- 18 - 21yrs: 24.9% of 350 = 87.15 → round to **87**
- 21 - 27yrs: 23% of 350 = 80.5 → round to **80**
- 27 - 35yrs: 17.1% of 350 = 59.85 → round to **60**
- 35 - 42yrs: 13.6% of 350 = 47.6 → round to **48**

Example C

Would it be appropriate to use 42 samples of green and 78 samples of blue marbles for a stratified sample of 120 marbles from a population of 960 green and 1500 blue marbles?

Solution:

Just compare the ratios of each color:

- Green sample ratio: $\frac{42}{120} = .35$
- Green population ratio: $\frac{960}{2460} = .39$
- Blue sample ratio: $\frac{78}{120} = .65$
- Blue population ratio: $\frac{1500}{2460} = .61$

We can see by looking at the ratios that the actual population that they don't quite match. There should be 47 green and 73 blue in the sample. This may not seem like enough of a difference to pose a problem, but notice that the 5 too few green marbles is more than 10% of the sample, and the 5 too many blues is nearly 10% of the blue sample. That is enough to possibly skew the results.

Concept Problem Revisited

Suppose you wanted to find out if age influences the choice of classes for students at a particular university. You might divide the students up by age ranges such as: Under 18, 18 - 21, 21 - 25, 25 - 35, and 35 and over. How could you make sure a random sample of college students would have members of each age range?

By now, I'm sure you can see that a stratified sample would be perfect for this situation.

Vocabulary

A **stratum** is a single category or sub-population out of a larger population. The characteristics separating each of the strata should be clearly defined.

A **stratified sample** is a sample of a population chosen such that each of several strata is represented in the same ratio in the sample as in the population.

Guided Practice

1. Ivana wants to create a sample of the students in her school to see if it would be a good idea to put up posters of country music bands in each grade's locker hall. Is this a good situation to use a stratified sample?
2. If Laurana wants to create a stratified sample of the distance an arrow can be shot from each of several different types of bows in the population of bows from her tribe, will she need to get a complete count of every single bow owned by every tribe member?

3. If Tanis wants to investigate the waterproofing of Kitiara's 200 pairs of boots, should he first try to separate them into different groups by style or maker?

Solutions:

1. Yes, absolutely. Ivana will want to get a sample of the students in the school that is stratified by grade level to make sure each grade will appreciate the posters, since she plans to put them up in each hall.
2. Inconveniently, yes. If she does not get a full count, she will not be able to come up with an accurate ratio to 'aim for' in her bow sample. Since she wants to use her sample to make predictions about the entire population, she needs to be sure she has a true random sample. She needs to be certain that each bow has an equal chance of ending up in the sample.
3. It would be a good idea, yes. Different makers or styles are liable to be more similar to each other than to the entire population.

Practice

For questions 1-5, assume you intend to create a stratified sample of 250 from a population of 920 trucks, 1540 subcompact cars, 1320 sedans, 450 motorcycles, 110 R.V.'s, 550 luxury cars, and 780 sports cars.

1. What percentage of the population is represented by sedans?
2. How many motorcycles should you have in your sample?
3. How many subcompacts should you have in your sample?
4. Is 10 R.V.'s a good number for your sample?
5. Should you have more than 15% of your sample represented by trucks?

For questions 6-10, assume your stratified sample consists of 29 cats, 62 small dogs, 48 large dogs, 19 birds, 37 pot-bellied pigs, and 55 horses. Assume the total population of pets is 6474.

6. How many horses are there in the entire population?
7. What percentage of the population is represented by dogs?
8. Are there more than 1000 pot-bellied pigs in the population?
9. What would the total population be if there were no horses?
10. What percent of the sample is made up of cats?

For questions 11-15, decide whether a stratified sample is warranted and why.

11. The estimated mileage of U.S. automobiles compared to vehicle weight.
12. The average height of college basketball players.
13. The G.P.A. of students in various sports.
14. The number students in your school with access to the internet.
15. Homework grades of sports participants.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 3.3.

3.4 Cluster Sampling

Objective

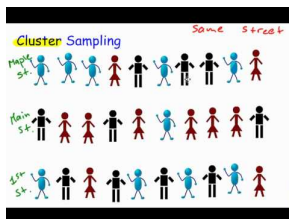
Here you will learn to recognize and implement *cluster sampling*.

Concept

Suppose you are hoping to predict the most popular favorite movie among U.S. high school students. Since your population is all high school students in the U.S.A., a simple random sample is just not feasible since you cannot possibly number each student individually. How then could you manage to get a representative sample to use for extrapolation?

Look to the end of the lesson for the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62210>

<http://youtu.be/QOxXy-I6ogs> Steve Mays - Cluster Sampling

Guidance

Cluster sampling is ideal for extremely large populations and/or populations distributed over a large geographic area. The concept of cluster sampling is that we use SRS (simple random sampling) to choose a limited number of groups or *clusters* of samples from a population, and then again apply SRS to the chosen clusters in order to identify specific samples.

Since you complete each step in the cluster sampling process using SRS, the results can be used for extrapolation. However, there is still a danger of ending up with a non-representative sample if the clusters you are choosing from are not each representative of the population. (See Example B)

The prime benefit of cluster sampling is that it can do an excellent job of reducing the size of a very large population down to something more manageable without ruining your ability to gather a *representative sample*.

Example A

A consumer report journalist wants to publish a blog about the most popular cars in the U.S. She has decided to use publicly available vehicle registration data to identify the most often registered car makes. How could she use cluster sampling to help her build a representative sample of U.S. car owners?



Solution: One way to get a representative sample of vehicle registrations across the whole country would be to number a list of all of the states in the U.S., and then use a **random number generator (RNG)** to pick out 4 or 5 states. From each state, she could then number the counties, use an RNG to pick a county or two, and then repeat to identify cities or towns. By narrowing down the extremely large initial population in this way, she can maintain the randomness of her sample without needing to number every car owner in the U.S.

Example B

Kevin is attempting to create a representative sample of students in his school for a poll asking students' opinions on shortening the school day by 1hr for students over 18yrs old. The results of his survey suggest that over 95% of students think it is a bad idea. Kevin is rather surprised that the results are so overwhelmingly negative, and he wonders if he did something wrong when selecting his sample.

If Kevin chose his sample with the cluster sampling method, and started by clustering the students by grade level, can you see why his results might be suspect?

Solution: Did you recall from the lesson that we mentioned that each cluster should be representative of the population? By clustering his samples by grade level, Kevin opened himself up to bias right away. Given the results he received, it is likely that he ended up with all of his samples being freshman who (approximately 15yrs old) thought it unfair that older students should have a shorter schedule!

Example C

How could you use a cluster sample to estimate the average density of various tree types in a large forest?

Solution: A common method for this type of study is to use a map. If you lay a virtual grid over a map of the forest, you can then number the squares and use an RNG to identify a number of square clusters of trees. You can then count the number of each type of tree in each cluster.

Concept Problem Revisited

Suppose you are hoping to predict the most popular favorite movie among U.S. high school students. Since your population is all high school students in the U.S.A., a simple random sample is just not feasible since you cannot possibly number each student individually. How then could you manage to get a representative sample to use for extrapolation?

This is an ideal opportunity to use a cluster sample. You could number each state and use an RNG to choose a few states, then repeat to choose a couple of school districts in each state, then a few schools from each district, and finally 1 or 2 classes from each school.

Vocabulary

A **representative sample** is a smaller number of members of a population whose responses to events model those of the entire population.

A **cluster** is a naturally occurring subgroup of a population.

A **cluster sample** is a sample created by randomly selecting members from each of several clusters that have a similar makeup to the population as a whole.

Guided Practice

For questions 1-3, describe why or why not each scenario describes a cluster sample.

1. Armand chooses 4 of the 10 busses in front of his school, and polls 10 students from each to see if they think busses are comfortable.
2. A cup of milk is selected from 10 of the 50 gallons being studied.
3. 5 dogs are chosen from each breed at the show.
4. How could you use the cluster method to select a representative sample of the types of energy drink carried by gas stations in Colorado?

Solutions:

1. This is a valid cluster sample because it is reasonable to assume that the students in each bus are representative of the population of bus riders.
2. This is not a cluster sample, it is merely an SRS, since each gallon can be considered a single unit, and the cup is just a smaller portion of the sample.
3. This is a stratified sample, not a cluster sample, since the groups are not each representative of the population of show dogs.
4. You might start with an overlay of a map of Colorado, and use an RNG to identify a few areas. Then sample the types of drink at one store of each gasoline brand located in the chosen areas (since different stores in the same geographical area from the same company usually carry the same inventory).

Practice

For questions 1-10, decide if each situation is an example of a properly selected cluster sample.

1. 150 light bulbs are evaluated from 1 randomly selected pallet every 30 minutes.
2. 5 light bulbs are evaluated from each case of light bulbs.
3. 10 cars are reviewed from each of 10 randomly selected used-car dealers.
4. 15 candy bars are tested from each shipment.
5. 150 laptops are tested from each company.
6. 100 laptops are evaluated from each of 5 randomly selected dealers.
7. 25 students from each grade were asked the names of their favorite bands.
8. 25 students from each school were asked the names of their favorite bands.
9. Gas prices were sampled from each gas station in town to find the cheapest.
10. 15 gas stations were sampled from each town to find the town with the cheapest.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 3.4.

3.5 Non-Probability Sampling

Objective

Here we will wrap-up our discussions on types of samples by discussing specialized sample collection methods that are deliberately not random.

Concept

Suppose your friend were to tell you “candy is sweet.” You would probably accept the statement without argument, since we generally think of candy as almost a synonym for sweet in the U.S. However, if you wanted to prove the claim wrong and demonstrate that candy is not always sweet, you could conduct an experiment to see if candy is sometimes *not* sweet.

Given the practically limitless different types of candy, how could you collect a useable sample if you can’t possibly give each and every type of candy an equal chance of being a part of a random selection?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62214>

http://youtu.be/Q_-Bts76Tff?t=11m7s mohdkhairieuum - Topic 5 Sampling

Guidance

There are a number of recognized non-probability sampling methods, including:

- Convenience Sampling - Choosing samples based on easy or convenient access
- Volunteer (or Snowball) Sampling - Asking for volunteers or for recommendations from other samples
- Judgment Sampling - Deliberately choosing samples based on a desired characteristic
- Quota Sampling - Choosing samples to fill a specific quota of each of several sub-populations of the original

Some studies do not lend themselves to the collection of a randomized sample. Although a random sample is necessary if the goal is to directly generalize results back to an entire population, sometimes that isn't the purpose of the study.

Common situations where non-probability sampling may be appropriate include:

- a. Qualitative Research - Studies with the general goal of identifying topics worthy of future more detailed (quantitative) study

- b. Studies particularly focused on specific portions of large population
- c. Studies with limited funding and/or time
- d. Studies with a goal of either disproving a particular theory, or of demonstrating the existence of a specific trait in a population.

Example A

A toy designer is looking to develop the next 'big thing' in toys for young children. A brainstorming session results in 15 possible new ideas across a wide range of types from puzzles to remote-controlled chickens. A full-scale prototype development and testing for each idea would not be cost effective, so the company decides to perform a preliminary study. One of the employees suggests drawing sketches of the ideas and taking them to his son's daycare to see which pictures get the most attention from the kids there.

What kind of sampling is this? Is it an appropriate choice for the application?



Solution: As a preliminary data collection resource, this *convenience sampling* of kids is a reasonable choice. Once the suggestions have been narrowed down a bit by judging the responses of the day-care kids, a more conventional random sampling may be used to pick out a specific design or two and then further refine it.

Example B

You are trying to convince your teacher that listening to music while doing homework improves scores. You decide to conduct a study of the effects of listening to music while doing homework and correlate it to student scores. Since your hypothesis is that listening to music improves scores on homework, would it be most effective to select a random sample of all the students in your school? If not, what type of non-probability sampling would be most appropriate and why?

Solution: If you are trying to find out if listening to music affects homework scores, then you would want a random selection of the entire population. However, since your very specific goal is to demonstrate music being associated with high scores, you might want to take a judgment sample of only students who listen to music while doing homework, and have high scores.

Example C

The student council at Cedar Valley Public School wants to gauge student opinion on the quality of their extracurricular activities. They decide to survey approximately 150 of the school's 1,000 students using the grade levels (7 to 12) as the sub-population.

The table below gives the number of students in each grade level. Fill in the missing values for the number of students that should be in the sample from each grade level.

TABLE 3.3:

TABLE 3.3: (continued)

Grade level	Number of students enrolled	Percentage of students (%)	Quota of students in sample of 150
7	150		
8	220	22	
9	160		
10	150	15	
11	200		
12	120	12	
Total	1,000	100	150

Solution:

First fill in the missing percentages by dividing the number of students in each grade by the total number of students in the school:

$$\begin{aligned} \text{Grade 7: } & \frac{150 \text{ students}}{1000 \text{ students}} = 15\% \\ \text{Grade 9: } & \frac{160 \text{ students}}{1000 \text{ students}} = 16\% \\ \text{Grade 11: } & \frac{200 \text{ students}}{1000 \text{ students}} = 20\% \end{aligned}$$

Now apply each grade's percentage of enrollment to the sample size of 150:

- Grade 7 sample should contain 15% of 150 or 22.5 students, rounded up to **23 students**
- Grade 8 sample should contain 22% of 150 or **33 students**
- Grade 9 sample should contain 16% of 150 or **24 students**
- Grade 10 sample should contain 15% of 150 or 22.5 students, rounded up to **23 students**
- Grade 11 sample should contain 20% of 150 or **30 students**
- Grade 12 sample should contain 12% of 150 students or **18 students**

Concept Problem Revisited

It is commonly accepted that any survey conducted by mail, or over the internet, or by telephone will have a very low response rate. It is not unheard of for such surveys to have less than 5% of the chosen sample actually return usable results.

It certainly seems logical to attempt to get a response rate as high as possible, but does a low response mean that the experiment is invalid?

No, it doesn't. Certainly a very low response rate should be investigated, but the more important consideration is how well the results actually collected represent a random sampling of the population under study.

Vocabulary

Convenience Sampling - This type of sampling involves selecting subjects on the ease and convenience of access.

Volunteer (Snowball) Sampling - This type of sampling requires that people volunteer themselves or their friends for a study.

Judgment Sampling - This type of sampling occurs when the investigator already has made an assumption about a characteristic of the population, and samples are selected accordingly.

Quota Sampling - In this type of sampling, sampling is done until a specific number of subjects for various sub-populations have been selected. However selection is not random, but relies on the interviewer to make the selection of the subject.

Guided Practice

1. Your mom says that sticking your tongue out every day will cause your face to get stuck that way. Would a non-probability sample be appropriate for a study attempting to prove that hypothesis wrong? What type would you recommend, and why?
2. Would a non-probability sample be appropriate for a study attempting to show that brand-name band-aids are superior? Which type would be appropriate and why?
3. How would a Snowball sampling method apply to a study of which flavor of gum has the longest-lasting flavor?
4. Would a non-probability sample be a good choice if you want to run a study to see if people who wear glasses are the best students? What type of sample would you recommend and why?

Solutions:

1. This is actually an appropriate use of a convenience sample. According to the hypothesis "... will cause it to get stuck", all you need is one example of it *not* getting stuck to disprove the statement. Since any example will work, you might as well try easy possibilities first.
2. No. Here you are hoping to generalize from your sample to the whole population of band-aid users. Extrapolation requires a true random sample.
3. Here your study is actually on the *gum* rather than on the chewer, so asking friends to ask friends what gum they have found has the longest flavor would be convenient and appropriate.
4. A judgment sample of the best students to see if they wear glasses would be a more efficient way to test your hypothesis than just a random sample of all students or all glasses-wearers.

Practice

1. A marketing company offers \$75.00 to the first 100 people who respond to their advertisement in a magazine and complete a questionnaire. This situation is an example of:
 - a. simple random sample
 - b. convenience sample
 - c. voluntary response sample
 - d. multistage cluster sample

The marketing class at a local high school wants to conduct a survey of the opinions of 60 students. Identify each type of sampling method they might use listed below.

2. Survey the first 60 students to walk through the doors at school in the morning.
3. Marketing class members each ask a friend for his/her opinion, and for the names of 3 other students to ask also.
4. Class members discuss who would be most appropriate to survey based on the results they want, then choose those persons.
5. Marketing class members decide to split students up into groups based on color of clothing, then choose a sample with the same ratio of colors as the whole school.

6. Number the students in the official school roster. Use a table of random numbers to choose 60 students from this roster for the survey.
7. A researcher plans a study to examine the depth of belief in an afterlife among the adult population of a small town. He obtains a simple random sample of 100 adults as they leave church one Sunday morning. All but one of them agree to participate in the survey. Which of the following is a true statement?
- Proper use of chance as evidenced by the simple random sample makes this a well-designed survey.
 - The high response rate makes this a well-designed survey.
 - Selection bias makes this a poorly designed survey
 - None of these statements is true.
8. Do any of the following use simple random sampling?
- Bingo game
 - Presidential elections
 - US Census

Identify the type or types of sampling used for the following.

9. Sarah went through a telephone directory and called every person with a name she liked.
10. Four people divided a telephone directory evenly and each called the first 10 numbers they found.
11. Every 5th block of 10 students walking past the classroom where the surveyors are working is exhaustively sampled about their faith in opinion polls.

Describe possible weaknesses of each of the following sampling procedures:

12. A sample of size 200 from a population of corporate personnel were asked to give their opinion about the federal government's affirmative action hiring program. Ninety-three percent expressed opposition to the program.
13. A TV show takes a survey by asking individuals to call in to identify whether they are 'FOR' or 'AGAINST' more restrictions on gun control.
14. We are interested in obtaining a sample representative of all males age 18 or older. We run an advertisement on the Internet, ask for volunteers, and then choose a random sample from the list of people who volunteer.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 3.5.

Various methods of statistical sampling were reviewed, and students were given example applications for each. Having completed this chapter, students should feel competent identifying appropriate sampling methods for various applications and properly applying those methods to collect useful samples of differing populations.

3.6 References

1. NEMO. <http://pixabay.com/en/brown-drive-paper-food-fruit-24550/?oq=bag> .
2. PDPhotos. <http://pixabay.com/en/shell-seashell-beach-old-remains-3234/?oq=SEA%20SHELL> .
3. Emery Way. <https://www.flickr.com/photos/emeryway/2906577602/> .
4. CK-12 Foundation. . CCSA
5. Axion23. <https://www.flickr.com/photos/gfreeman23/11613922755> .
6. Mike Mozart. <https://www.flickr.com/photos/jeepersmedia/13981016973> .

CHAPTER

4**Evaluating and Displaying
Data****Chapter Outline**

- 4.1 GROUPING DATA**
 - 4.2 ANALYZING DATA**
 - 4.3 RELATIVE FREQUENCIES**
 - 4.4 CUMULATIVE FREQUENCIES**
 - 4.5 CREATING HISTOGRAMS**
 - 4.6 INTERPRETING HISTOGRAMS**
 - 4.7 FREQUENCY POLYGONS - PROBABILITY AND STATISTICS**
 - 4.8 CREATING BOX-AND-WHISKER PLOTS**
 - 4.9 INTERPRETING BOX-AND-WHISKER PLOTS**
 - 4.10 CREATING STEM-AND-LEAF DIAGRAMS**
 - 4.11 INTERPRETING STEM-AND-LEAF PLOTS**
 - 4.12 CREATING SCATTER PLOTS AND LINE GRAPHS**
 - 4.13 INTERPRETING SCATTER PLOTS AND LINE GRAPHS**
 - 4.14 CREATING PIE CHARTS**
 - 4.15 INTERPRETING PIE CHARTS**
 - 4.16 REFERENCES**
-

Properly collecting and organizing data is important, but without a way to show the results to others in a meaningful format, data becomes much less useful. Visualizing data in the form of an image allows you to identify trends and comparisons that may be difficult to see in a list of values. Graphs, plots, and charts are the primary methods of data imaging.

Graphs and charts come in a bewildering array of shapes, sizes, and types, but there are a number of them that are used quite regularly and that form the basis for many others.

In this chapter, you will learn how to construct and interpret many common data visualizations.

4.1 Grouping Data

Objective

Here you will learn about different methods of organizing data for use in statistical research.

Concept

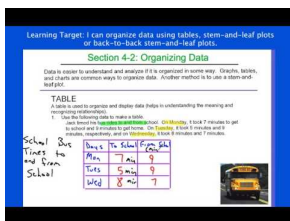
Suppose you were given the following data:

87, 72, 91, 91, 73, 83, 79, 81, 87, 72, 81, 91, 73, 73, 73

If you were told you were going to evaluate this data using common methods of central tendency and dispersion, how might you start by organizing the data in order to make the study as straightforward as possible?

Watch This

This video presents a couple of methods of organizing data:



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63578>

<http://youtu.be/NASBOFGMy2c?t=40s> FortLucasMath - Organizing Data

Guidance

Data in its original form, just a list of numbers, names, letters, colors, etc., is known as **raw data**, and is often not particularly useful without analysis of some kind of organization. The series of numbers in the concept question above, for instance, doesn't really mean anything at the moment. Without some sort of context and some level of organization, this is just a bunch of meaningless values.

Data can be classified into two general types, **quantitative** and **qualitative**. There are a number of ways to group or organize each type of data to make it more useful.



- **Quantitative** Data (data that may be conveniently described numerically):
 - Dates or times are commonly organized chronologically
 - Data with recurring values is often organized in a **frequency distribution**
 - **Univariate** data likely to be used for evaluating the mean or the range of a population is generally organized in increasing or decreasing magnitude or alphabetically.
 - **Bivariate** data is usually organized in a table showing how the two variables change in relation to each other.
 - Compare and contrast tables are excellent for evaluating two or more variables
- **Qualitative** Data (data that may be difficult to describe with numerical values)
 - Commonly grouped by category
 - Categories are often evaluated using a frequency distribution
 - Data may be organized in order of importance
 - **Inductive** organization orders information by increasing complexity, listing facts prior to conclusions and advancing from specific examples to general conclusions.
 - **Deductive** organization is the inverse of inductive, listing recommendations/conclusions followed by supporting facts/data.

Example A

Elaina is preparing to create a **histogram** to illustrate the data that she collected on average time spent taking a particular test in her Statistics class.

16 mins, 18.5 mins, 14.5 mins, 16 mins, 19 mins, 18 mins, 16.5 mins, 15 mins, 15 mins, 14.5 mins, 14 mins, 16 mins, 12.5 mins, 19.5 mins, 14 mins, 15 mins, 16.5 mins, 14 mins, 18 mins, 16 mins

- a. How should she organize the data to make the construction of the histogram as straightforward as possible?
- b. What will the data look like after it is organized?

Solution: A histogram is a graph that illustrates the **relative frequency** or **probability density** of a single variable.

- a. Since Elaina will need to identify the number of values in each category of the data, it would be ideal to organize the data in groups called classes or intervals. With the given data, intervals of 1 minute each would seem appropriate.
- b. Minutes Required to Complete the Test:

12.5 | 14, 14, 14, 14.5, 14.5 | 15, 15, 15 | 16, 16, 16, 16, 16.5, 16.5 | 18, 18, 18.5 | 19, 19.5

Example B

Orlando is planning to create a box-and-whisker plot to illustrate how much more popular dogs and cats are as pets than fish-tanks, reptiles, and birds. He has collected the data below from a randomized sample of homes in his town, using a survey questioning the number of pets each family has in each category

House 1: 2 dogs, 2 cats, 0 birds, 0 reptiles, 1 fish tank

House 2: 3 dogs, 2 cats, 1 birds, 0 reptiles, 0 fish tank

House 3: 0 dogs, 3 cats, 0 birds, 1 reptiles, 1 fish tank

House 4: 2 dogs, 1 cats, 2 birds, 0 reptiles, 1 fish tank

House 5: 2 dogs, 1 cats, 0 birds, 0 reptiles, 0 fish tank

House 6: 2 dogs, 2 cats, 0 birds, 0 reptiles, 0 fish tank

House 7: 3 dogs, 1 cats, 1 birds, 0 reptiles, 2 fish tank

House 8: 3 dogs, 2 cats, 0 birds, 0 reptiles, 1 fish tank

House 9: 2 dogs, 3 cats, 0 birds, 0 reptiles, 0 fish tank

House 10: 1 dogs, 3 cats, 0 birds, 0 reptiles, 0 fish tanks

How should Orlando organize the raw data to facilitate the creation of his box-and-whisker plot? What will the organized data look like?

Solution:

- a. Since Orlando's box-and-whisker plot is specifically meant to highlight the number of dogs and cats, it would be a good idea to organize the data in groups by importance, with dogs and cats first. Since he will need to identify the mean, range, and quartiles of the data, it would also be good to organize each group by increasing values.
- b. Dogs: 0, 1, 2, 2, 2, 2, 2, 3, 3, 3
 - a. Cats: 1, 1, 1, 2, 2, 2, 2, 3, 3, 3
 - b. Birds: 0, 0, 0, 0, 0, 0, 0, 1, 1, 2
 - c. Reptiles: 0, 0, 0, 0, 0, 0, 0, 0, 0, 1
 - d. Fish Tanks: 0, 0, 0, 0, 0, 1, 1, 1, 1, 2

Example C

Cheng is interested in the phenomenon of the changes in how fast time seems to pass to people as they age. He has collected data from 300 people between the ages of 10 and 70. Each person reported the time it *seemed* to take to complete three neutral (neither particularly liked nor disliked) activities, one 5mins, one 15mins, and one 60mins long. Now Cheng has a massive and somewhat intimidating list of numbers, and he needs to decide how to organize what he has into something useful.

- a. Identify at least 2 different ways that Cheng might organize the raw data that would illustrate changes in time perception as people age.
- b. How might Cheng consolidate the data so he doesn't end up needing to plot nearly 1000 values on a chart or graph?

Solution: With such a huge amount of raw data, Cheng's greatest challenge will be consolidating it into a useful and informative format.

- a. Cheng might choose to organize the data by increasing time in several age groups, sorting the values first by age, and then by perceived time for each activity. He might also wish to sort first by actual activity length, then by age or perceived time passage.

- b. Finding the mean perceived length for each of several age ranges would be a great way for Cheng to maintain the general integrity of his data while reducing the sheer volume.

Concept Problem Revisited

87, 72, 91, 91, 73, 83, 79, 81, 87, 72, 81, 91, 73, 73, 73

If you were told you were going to evaluate this data using common methods of central tendency and dispersion, what sort of preparation could you do in order to make the study as straightforward as possible?

Central tendency measurements are generally facilitated by organizing data in increasing value from left to right. Ideally, it would be convenient to also note the total number of values, along with their sum, as you are ordering them.

Vocabulary

Central tendency is the behavior of the 'main portion' of a set of data; the most common measure of central tendency is the mean or average.

Probability density is similar to **relative frequency** in that both are ways to evaluate how often a particular value or range appears in a set of data.

A **frequency distribution** is a depiction of the number of occurrences of each data point in the set.

A **histogram** is a visual way to show data, primarily used for continuous variables, that is much like a bar chart.

Deductive means that a conclusion was stated first, and then supported by statistics. **Inductive** means that statistical information was evaluated, and then used to postulate a generalization.

Bivariate data has two variables, whereas **univariate** data only has one.

Quantitative data is composed of numerical values, and **qualitative** data is generally not. Qualitative data is generally applied to "which of these..." questions, whereas quantitative data is commonly applied to "how many of these..." questions.

Guided Practice

A class of 40 students took a science exam. They earned the following percentages on their tests:

73, 45, 62, 34, 59, 20, 48, 50, 78, 38, 52, 91, 57, 82, 46, 51, 62, 58, 39, 50, 72, 73, 63, 52, 41, 37, 28, 46, 71, 75, 36, 28, 44, 90, 51, 28, 60, 18, 47, 40.

1. Describe or demonstrate a means of displaying the results more clearly.
2. The teacher wants to compare the student's scores with those of another class. Describe a means of organizing the data that would make it easy to compare the two sets of data.

3. The teacher gave grades as follows:

A grade: 90 and above

B grade: 80 to 89

C grade: 70 to 79

D grade: 60 to 69

F grade: 59 and below

Make a table to show how many students achieved each grade

4. Determine if the data is qualitative or quantitative.

- a. The majority of the people in Asia most often wear the color red.
- b. A survey was done among elementary age children to discover their favorite fruit.

5. These are the numbers of cars sold at a local dealer over the last 12 days. Create a Frequency Distribution Table.
3, 5, 1, 7, 3, 2, 8, 1, 3, 2, 6, 4.

Solution:

1. A good start would be to simply organize the numbers in increasing order:

18, 20, 28, 28, 34, 36, 37, 38, 39, 40, 41, 44, 45, 46, 46, 47, 48, 50, 50, 51, 51, 52, 52, 57, 58, 59, 60, 62, 62, 63, 71, 72, 73, 73, 75, 78, 82, 90, 91.

Now we can see at a glance that the numbers range from 18 to 91, with a greater frequency in the mid-range than at the extremes.

2. To compare the scores with another class, it would be convenient to have the number of scores in each range summarized. She might either tally the number of scores between 0 and 10, then 10 and 20, and so on, or just tally the number of A's, B's, etc.

3. The table would look like this:

TABLE 4.1:

A	B	C	D	F
2	1	6	4	26

(Either that was a frightfully difficult exam, or the students didn't study well!)

4. These are both qualitative. Neither A, nor B could be expressed as numerical data.

5. To create a frequency distribution table for 3, 5, 1, 4, 3, 2, 2, 1, 3, 2, 5, 4, simply label the values that occur in the set across the top, and the number of occurrences of each in a 2nd row beneath, either as numerals or as tally marks:

TABLE 4.2:

Value:	1	2	3	4	5
Frequency:	<i>II</i>	<i>III</i>	<i>III</i>	<i>II</i>	<i>II</i>
	2	3	3	2	2

Practice

For Q's 1-3, determine if the data is qualitative or quantitative.

1. The average temperature of a particular city is 23 degrees C.
2. Determine if the number of hours a person spends in front of a computer will affect their eye sight.
3. A random survey was done to find out the average speed of cars on a highway.
4. Which letter has the greatest frequency in the following sentence?

THE SUN ALWAYS SETS IN THE WEST.

5. Joe scored the following numbers of goals in their last twenty soccer games: 3, 0, 1, 5, 4, 3, 2, 6, 4, 2, 3, 3, 0, 7, 1, 1, 2, 3, 4, 3.

- a. Organize the values from smallest to greatest
- b. Which number had the greatest frequency?

6. The following number gives the first 31 digits of pi: 3141592653589793238462643383279
- Treating each digit as a separate unit of data, how might you organize the units to prepare for an evaluation of their frequency and range (spread)?
 - How would the units appear after the organization?
 - What is the frequency of the digits 3, 5, and 7?

7. A die was thrown 100 times. The frequency distribution is shown in the following table:

TABLE 4.3:

Roll	Frequency
1	21
2	11
3	15
4	19
5	16
6	18

- What is the total frequency of numbers less than 4?
- What percentage of throws of the die were higher than 5?
- How many throws scored greater than 2, but less than or equal to 5?

50 students took a test with a total of 10 possible points. The frequency distribution is shown in the following table:

TABLE 4.4:

Score	Frequency
0	1
1	2
2	1
3	3
4	1
5	4
6	9
7	8
8	7
9	10
10	4

- If 60% is a passing score, how many students passed the test?
- How many students scored above 80%?
- What percentage of the students had 5 or more questions correct?
- How many students scored greater than or equal to 4, but less than or equal to 7?

A spinner is in the shape of a regular heptagon marked with the numbers 1 to 7. Sue spun the spinner 50 times and recorded her results:

1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7

12. Create a frequency table with the data.

13. Which spin had the greatest frequency?

14. Which spin had the least frequency?

A teacher decided to survey the students in her class to determine the number of siblings each of them had. The following numbers are the total number of siblings reported by each student in the class: 2, 0, 1, 0, 1, 0, 4, 3, 4, 9, 2, 1, 3, 1, 5, 1, 2, 1, 2, 4, 3, 2, 2, 6, 3, 2, 4, 2, 3, 5

15. Organize the numbers in a manner conducive to the creation of a frequency table.

16. Create a Frequency Table.

17. How many students were surveyed to collect this data?

18. How many families have 4 children or less?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.1.

4.2 Analyzing Data

Objective

Here you will practice identifying the most appropriate data visualization for a particular set of data and the intended purpose of the study.

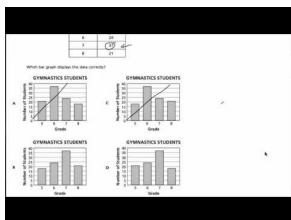
Concept

How do you decide what sort of graph is most appropriate for a particular application? If you know you have a set of average test scores from a number of different tests and are attempting to compare them, should you use a histogram, box-and-whiskers plot, or a bar chart? If you have 500 yes/no/maybe responses to a survey, should you visualize them with a pie chart, a bar chart, or a frequency polygon?

Choosing an effective data visualization can be a bit daunting; particularly at first, but with practice it will become much less difficult.

Watch This

This video is a rather detailed discussion of the use of a number of different types of graphs.



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63582>

<http://youtu.be/c3DBUzSIJR8> ShaunTeaches - Different Types of Graphs

Guidance

There are quite a number of common types of data visualizations, some of which have been in use for hundreds of years, and using modern technology there are even more. Some of the more modern types are animated and/or interactive, and may continually update themselves as new data is collected via the Internet. There are a number of other lessons in this course that individually detail the creation and interpretation of some of the more common graphing types, but in this lesson we will focus on some of the advantages and disadvantages of each. Having a good idea of the strengths of different methods will help you to choose an appropriate method for your own study.

TABLE 4.5:

Graphing Method	Strengths / Advantages	Weaknesses / Disadvantages
-----------------	------------------------	----------------------------

TABLE 4.5: (continued)

Histogram	Good for comparing multiple ranges of values, can be visually appealing, good for consolidating large amounts of data into a limited number of categories	Gets cluttered if there are too many categories of data, can take a while to construct - particularly by hand, not ideal for comparing a large number of categories, precise data is not easily displayed since values are grouped
Box-and-Whisker Plot	Quickly demonstrates a five-point summary of the data (mean, quartiles, range), can be used to compare multiple data sets, good for consolidating very large data sets	Individual data usually lost as data is grouped, not often considered as visually appealing as some other graphs, can be somewhat slow to construct
Stem and Leaf Diagram	Maintains accuracy of individual data values, can be easily evaluated to find ranges and notable clusters, good for medium to med-large data sets	Generally not particularly attractive, not convenient for identifying central tendencies for larger data sets, can be confusing for audiences not familiar with the idea.
Scatter Plot	Good for showing increasing/decreasing data trends, maintains accuracy of individual values, makes outliers immediately apparent, faster to construct than some, excellent for bivariate data.	May lead to incorrect generalizations about the data, may take a while to construct if there are many data points, limited flexibility for creating attractive designs, regression should only be applied to continuous variables.
Line Plot	Continuous variables can be represented with a few discrete points, handles bivariate data, moderate flexibility in design, easily interpreted by audiences	Requires continuous data, may not be visually striking
Frequency Polygon	Same as Line Plot above, but with greater visual flexibility for grabbing audience attention	Requires continuous data, anchors on base may be assumed as zero value data points
Pie Chart	Simple to create and interpret, easily visualize relative percentages, great for comparing values from multiple sources, can be very visually striking	Limited uses, requires discrete data, may be overly simplistic

Example A

Which of the following types of graphs is the most appropriate for displaying/evaluating the body weights of 300 different pet dogs? Why?

Bar Chart Scatter Plot Pie Chart Histogram

Solution: A histogram is the best option here because it allows you to break the data up into as many or as few categories or weight classes as you wish, and clearly shows the comparison between them. Also, a histogram would be a relatively efficient way to track the rather large data set.

Example B

Which one of the given graph types is most appropriate for evaluating the relationship between amount of sunlight and plant growth rate?

Bar Chart Scatter Plot Pie Chart Histogram

Solution: A scatter plot is the right tool here, since you are comparing two different variables. Since these are both *continuous variables*, you could also evaluate the trend of the data points. If the trend appears to be linear, you could use *linear regression* to identify the average comparison, and then add a linear plot to the graph to illustrate that average.

Example C

The relative distribution of several age ranges of trees in a forest: 0-5yrs, 6-15yrs, 16-30yrs, 31yrs and older.

Bar Chart Scatter Plot Pie Chart Histogram

Solution: This would be a great use for a pie chart, since you are dealing with the entire population of trees, and are interested in comparing the percentage of the forest represented by each.

Concept Problem Revisited

How do you decide what sort of graph is most appropriate for a particular application? If you know you have a set of average test scores from a number of different tests and are attempting to compare them, should you use a histogram, box-and-whiskers plot, or a bar chart? If you have 500 yes/no/maybe responses to a survey, should you visualize them with a pie chart, a bar chart, or a frequency polygon?

Choosing the best visual representation of a data set is a matter of identifying the purpose of your study and the type of data you intend to display.

A particular average representing each test could be pretty clearly displayed by a bar chart, and could be made quite striking with appropriate design.

500 results spread over only three categories would be a good use of a pie chart. You have the entire population (members of the survey), and your goal is to show the relative portions of each answer.

Vocabulary

Linear regression is the process of evaluating a single line best representing a number of individual plots on a scatter plot that shows a linear trend.

Continuous variables represent data that could be infinitely divided into smaller and smaller 'pieces'. In other words, data that can take on an infinite number of values in any specified interval. The most common continuous variable is time.

Guided Practice

Choose the most appropriate graphing method for each situation, and explain your reasoning:

1. The number of high school diplomas earned in Denver, CO for each year between 1980 and 1990.

Bar Chart Scatter Plot Pie Chart Histogram

2. The proportion of minnows in a pond that are damaged by chemical dumping.

Bar Chart Scatter Plot Pie Chart Histogram

3. The average number of puppies birthed by five breeds of dogs.

Bar Chart Scatter Plot Pie Chart Histogram

4. The number of students in the upper 10%, upper 50%, lower 50%, upper 25%, and lower 25% on finals at a particular university.

Bar Chart Scatter Plot Pie Chart Histogram

5. The number of yellow, red, and white roses found in each of 500 ten by ten foot plots in Central Park, New York.

Bar Chart Scatter Plot Pie Chart Histogram

Solutions:

1. Histogram, this is a comparative study of a limited number of categories of data. Note that although time is continuous, the groups of 1 year each allow us to graph with a histogram.
2. Pie Chart, this is a finite population and you are comparing proportions of a limited number of categories.
3. Bar Chart, since you are dealing with averages of discrete data in a limited number of categories.
4. Bar Chart, your categories overlap (upper 50% includes upper 10% for example), so a pie chart would not be appropriate.
5. Scatter Plot, there are WAY too many categories (individual flower garden plots) to try to use a histogram or bar chart.

Practice

What type of Graph should you use?

1. Track and compare the different amounts of time you spend playing video games versus doing your homework and practicing piano, over the period of a month.
2. You are trying to prove that your family spends too much money on groceries each month. Your families monthly budget looks like this:
 1. \$1,250 home mortgage
 2. \$500 utilities
 3. \$800 car payments
 4. \$300 entertainment
 5. \$800 groceries
3. You employ 4 sales people. You would like to track their sales over the last quarter.
4. You would like to compare the number of votes that 4 candidates received in the last student council election.
5. You would like to know if there is a relationship between the time you spend studying for a test and the test scores you receive in a class, knowing you study more for classes you enjoy.

6. Choose methods of graphing commonly used to “Compare”. For instance comparing the sales performance of one car to another.
 1. Bar Graphs
 2. Column Graphs
 3. Scatter Plots
 4. Pie Charts
 5. Line Graphs
 6. Data Tables
 7. Box Plots
 8. Histograms
7. Choose types of graphs commonly used to show “Distribution”. For instance, the waiting room times of patients in 5 different doctors’ offices.
 1. Bar Graphs
 2. Column Graphs
 3. Scatter Plots
 4. Pie Charts
 5. Line Graphs
 6. Data Tables
 7. Box Plots
 8. Histograms
8. Choose types of graphs commonly used to show “Parts of a Whole” For instance, number of female viewers of a new gaming website.
 1. Bar Graphs
 2. Column Graphs
 3. Scatter Plots
 4. Pie Charts
 5. Line Graphs
 6. Data Tables
 7. Box Plots
 8. Histograms
9. Choose types of graphs often used to show “Trends over Time”. For instance, number of umbrellas sold over a 365 day period.
 1. Bar Graphs
 2. Column Graphs
 3. Scatter Plots
 4. Pie Charts
 5. Line Graphs
 6. Data Tables
 7. Box Plots
 8. Histograms
10. Choose types of graphs used to show “Deviations”. For instance, sales numbers in an established business when a competitor opens across town.
 1. Bar Graphs
 2. Column Graphs
 3. Scatter Plots
 4. Pie Charts
 5. Line Graphs
 6. Data Tables

7. Box Plots
 8. Histograms
11. Choose types of charts that can be used to show “relationship” For instance, number of tutoring students obtained after report cards are released.
1. Bar Graphs
 2. Column Graphs
 3. Scatter Plots
 4. Pie Charts
 5. Line Graphs
 6. Data Tables
 7. Box Plots
 8. Histogram

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.2.

4.3 Relative Frequencies

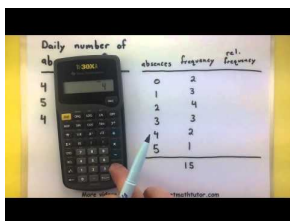
Objective

Here you will learn about comparing the relative number of times that different data values appear in a set.

Concept

If you were evaluating a set of data describing the numbers of “A’s”, “B’s”, “C’s”, and “D’s” that students earned on a particular test, and needed to display the data on a *relative frequency table*, how would you go about it?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63584>

<http://youtu.be/7jUIt39tUBM> MySecretMathTutor - How to make a relative frequency distribution

Guidance

Frequency tables are closely related to histograms and stem-and-leaf plots. A relative frequency table is specifically designed to display the ratio of each individual frequency to the total frequency of the data. To begin building a relative frequency table, start by grouping values into categories, classes, or intervals, depending on the type of data. You should try to limit the number of intervals or classes to less than a dozen in most cases, and you can use the square root of the number of actual data points as a guide if you wish.

Once you have all of your data separated into separate classes or categories, (often known as “*binning*”, since the data is divided up into multiple “bins,” one for each specified class, category, or interval), tally the number of values in each category and the total number of values all together.

To calculate the relative frequency of each category, divide the category, class, or interval frequency by the overall frequency. The decimal you get will represent the part of the entire sample that is represented by that category. Once you have calculated all of the relative frequencies for every category, add them up to make sure they total 1.0.

Note! If you are graphing the relative frequencies of a *continuous variable*, you will need to specify how to handle any values that fall right on the boundary of a category (also commonly called a class). Here are a couple of ways to do this:

- You can specify on your table that values equal to lower class limits are included in a class, but values equal to upper class limits are not (this is the conventional method). This means that a value of 5 would be considered part of a 5-10 class, but not part of a 1-5 class.

- You can also define each category so that there are no overlapping values:

1-4.99 5-9.99 10-14.99 15-20

Example A

You are given a bag of marbles in multiple colors, if there are 25 red, 22 yellow, 17 green, and 28 blue marbles, what are the relative frequencies of each color?

Solution: Start by totaling the number of marbles: $25 + 22 + 17 + 28 = 92$ total marbles



Divide the number of each color by the total number of marbles:

$$\frac{25 \text{ red marbles}}{92 \text{ total marbles}} = .272$$

$$\frac{22 \text{ yellow marbles}}{92 \text{ total marbles}} = .239$$

$$\frac{17 \text{ green marbles}}{92 \text{ total marbles}} = .185$$

$$\frac{28 \text{ blue marbles}}{92 \text{ total marbles}} = .304$$

Add your totals together to verify that they equal 1:

$$.272 + .239 + .185 + .304 = 1.0$$

Note that each of the relative frequencies can also be understood as percentages:

$$.272 = 27.2\% \text{ red marbles}$$

$$.239 = 23.9\% \text{ yellow marbles}$$

$$.185 = 18.5\% \text{ green marbles}$$

$$.304 = 30.4\% \text{ blue marbles}$$

$$27.2\% + 23.9\% + 18.5\% + 30.4\% = 100\%$$

Example B

A police officer is reviewing accident statistics for her city. She notes that there were a total of 23 incidents involving teen drivers between ages sixteen and twenty-one, 19 incidents involving drivers aged twenty-two through twenty-six, 19 involving twenty-seven to forty-year-olds, and 18 for ages above forty-one.

What are the relative frequencies for each age range?



Solution: The total number of accidents is:

$$23 + 19 + 19 + 18 = 79 \text{ total accidents}$$

The relative frequencies are:

$$\frac{23 \text{ in age range } 16-21}{79 \text{ total}} = .291$$

$$\frac{19 \text{ in age range } 22-26}{79 \text{ total}} = .241$$

$$\frac{19 \text{ in age range } 27-40}{79 \text{ total}} = .241$$

$$\frac{18 \text{ in age range } 41+}{79 \text{ total}} = .228$$

Verify that the relative frequencies total 1.0:

$$.291 + .241 + .241 + .228 = 1.001 \text{ (due to rounding)}$$

Example C

A local high school has 150 students who drive to school. Examining the parking lot, you note that there are 25 white cars, 35 red cars, 13 green cars, 19 blue cars, and 58 others.

What are the chances, expressed as percentages that randomly chosen students have each of the different colored cars?

Solution: We are given the total number of cars in the question: 150

Divide each of the individual colors by the total and convert the decimal answers to percentages:

$$\frac{25 \text{ white cars}}{150 \text{ total cars}} = .167 = 16.7\%$$

$$\frac{35 \text{ red cars}}{150 \text{ total cars}} = .233 = 23.3\%$$

$$\frac{13 \text{ green cars}}{150 \text{ total cars}} = .087 = 8.7\%$$

$$\frac{19 \text{ blue cars}}{150 \text{ total cars}} = .127 = 12.7\%$$

$$\frac{58 \text{ others}}{150 \text{ total cars}} = .387 = 38.7\%$$

Concept Problem Revisited

If you were evaluating a set of data describing the numbers of “A’s”, “B’s”, “C’s”, and “D’s” that students earned on a particular test, and needed to display the data on a relative frequency table, how would you go about it?

Add up the number of entries in each category, A, B, C, and D, to get the total number of data points. Divide the number of values in each category by the total to get the relative frequencies. Convert the decimal values to percentages if necessary.

Vocabulary

A **relative frequency table** compares the number of entries in each of several categories to the number of entries in the entire population.

Binning is the common term for the process of dividing data up into multiple categories, classes, or intervals in preparation for graphing.

A **continuous variable** is a variable that can represent *any* value between a given minimum and maximum. Age is a common continuous variable, since age can be measured in infinitely small increments. By contrast, a **discrete variable** can only represent *specific* values in a given range. The number rolled on a standard die is a discrete variable since it can only be one of the numbers 1 - 6, no partials or fractions.

Guided Practice

- The Sackmore and Headbut village football teams have played each other 50 times. Sackmore has won 10 times, Headbut has won 35 times, and the teams have drawn 5 times. Based on past performance, what is the probability that Sackmore will win the next match?
- Tony estimates that the probability that there will be an empty space in the car park when he arrives at work is $\frac{4}{5}$. His estimate is based on 50 observations. On how many of these 50 days was he *unable* to find an empty space in the car park?
- A pair of dice (one red, one green) is cast 30 times, and on 4 of these occasions, the sum of the numbers facing up is 7. What is the relative frequency that the sum is 7?
- The students in a class were asked what kind of music they liked. 18 liked rock, 11 liked pop, 5 liked hip hop, and 8 liked country. Create a frequency and relative frequency table using this information.
- In 1990, there were approximately 10,000 fast food outlets in the US that specialized in Mexican food. Of these, the largest were Taco Bell with 4809 outlets, Taco John's with 430 outlets and Del Taco with 275 outlets. The relative frequency that a fast food outlet that specializes in Mexican food is none of the above is:

Solutions:

- So far, Sackmore has won 35 out of the 50 matches. We can write this as a fraction, which (reduced) is: $\frac{7}{10}$. This fraction isn't really the probability of Sackmore winning, but it is an *estimate* of that probability. We say that the *relative frequency* of Sackmore winning is $\frac{7}{10}$.
- If Tony has figured that he *is* able to find a space 4 of every 5 times he arrives, then he *is not* able to find a space 1 in every 5 times. If we set the ratio: $\frac{1}{5} = \frac{x}{50}$, we can solve for x to find that he did not have a space 10 times.
- Out of thirty throws, four of them were 7's. The relative frequency is $\frac{4}{30}$ or $\frac{2}{15}$.
- To create the frequency table, we just need one column for each category:

TABLE 4.6:

Rock	Pop	Hip Hop	Country
18	11	5	8

To convert to a relative frequency table, just divide each frequency by the total:

TABLE 4.7:

Rock	Pop	Hip Hop	Country
$\frac{18}{42} = .43$	$\frac{11}{42} = .26$	$\frac{5}{42} = .12$	$\frac{8}{42} = .19$

- The likelihood that a restaurant is *not* one of the top three would equal the number of Mexican fast food restaurants

that are not one of the three: $10,000 - 4809 - 430 - 275 = 4486$, divided by the total number of Mexican fast food restaurants, **10,000**:

$$\frac{4,486}{10,000} = .4486 \text{ or } 44.86\%$$

Practice

30 Students in a class surveyed each other to find out their favorite movie series, and recorded the results in a table like the one shown below.

TABLE 4.8:

Movie Series	Number of Likes
Twilight	7
Lord of the Rings	5
Pirates of the Caribbean	9
Harry Potter	6
Narnia	2
High School Musical	1

1. What was the relative frequency for Narnia?
2. What was the relative frequency for Pirates of the Caribbean?
3. 100 people were asked whether they were left-handed. 8 people answered yes. What is the relative frequency of left-handed people in the survey?
4. The relative frequency of getting a white candy from a particular bag is 0.3. If the bag contains 100 candies, estimate the number of whites.
5. Kyle observed 80 cars as they drove by his bedroom window. 24 of them were red. What is the relative frequency of red cars?
6. The relative frequency of rain in April is .6. There are 30 days in April. Estimate the number of days of rain expected in April.

Use the table below listing the heights of 100 male semiprofessional soccer players.

TABLE 4.9:

HEIGHTS (INCHES)	FREQUENCY OF STUDENTS	RELATIVE FREQUENCY
59.95-61.95	5	
61.95-63.95	3	$\frac{3}{100} = 0.03$
63.95-65.95		$\frac{15}{100} = 0.15$
65.95-67.95	40	$\frac{40}{100} = 0.40$
67.95-69.95	17	
69.95-71.95	12	$\frac{12}{100} = 0.12$
71.95-73.95		$\frac{7}{100} = 0.07$
73.95-75.95	1	$\frac{1}{100} = 0.01$
	Total = 100	Total =

7. Fill in the blanks and check your answers.

8. The percentage of heights that are from 67.95 to 71.95 inches is:
9. The percentage of heights that are from 67.95 to 73.95 inches is:
10. The percentage of heights that are more than 65.95 inches is:
11. The number of players in the sample who are between 61.95 and 71.95 inches tall is:
12. What kind of data does this chart highlight, qualitative or quantitative?
13. What is the height interval for the players who fall under the frequency of .03?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.3.

4.4 Cumulative Frequencies

Objective

Here you will learn about organizing and graphing the running totals of multiple categorical frequencies.

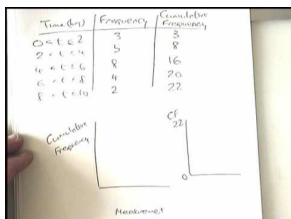
Concept

A school lunch break is divided into multiple sessions, the 1st group, consisting of 20 girls and 50 boys, begins lunch at 11:30 am, the 2nd group, with 35 girls and 30 boys begins at 11:40, and the 3rd group, 40 girls and 25 boys, at 11:50.

Assuming that none of the students from an early group leave the cafeteria before everyone goes back to class at 12:10, what is the relative frequency of girls in the cafeteria at 11:35? What about at 11:45? 11:55?

In this lesson we discuss *cumulative frequencies*, and by the end of the lesson, this sort of question will be no problem at all.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63600>

<http://youtu.be/D1s0XPMnHzo> Maths520 - Cumulative Frequency Part 1

Guidance

Cumulative frequencies describe a sort of 'running total' of frequencies in a frequency distribution. You will find cumulative frequencies in many real-world situations, since we often need to collect data for a larger study by conducting several smaller studies.

To calculate a cumulative frequency, simply create a frequency distribution table, then add the frequency from the first category to that of the second category, then add that total frequency to the third category, and so on. Because cumulative distributions don't generally show different combinations of categories, it is a good idea to arrange your categories in a suitable order so that the cumulative frequencies are most likely to be of use as the running total increases.

A *relative cumulative frequency* table shows how the cumulative frequency after each successive interval compares to the *total* frequency. To create a relative cumulative frequency table, calculate the relative frequency of each interval or category, and then add the relative frequency of each category to all the prior ones.

Example A

Create a *cumulative frequency table* showing the number of hours per week that Brian watches television, based on the given information.

Brian's T.V. Time



- Monday: 1.5 hrs
- Tuesday: 2.25 hrs
- Wednesday: 2 hrs
- Thursday: 1.75 hrs
- Friday: 1.5 hrs
- Saturday: 3.25 hrs
- Sunday: 2.5 hrs

Solution:

To find the cumulative frequency of hours of T.V. watched, add the frequency of each day to the total frequency of the day before:

TABLE 4.10:

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Cumulative Frequency	1.5 hrs	$1.5 + 2.25 = 3.75$ hrs	$3.75 + 2 = 5.75$ hrs	$5.75 + 1.75 = 7.5$ hrs	$7.5 + 1.5 = 9$ hrs	$9 + 3.25 = 12.25$ hrs	$12.25 + 2.5 = 15$ hrs

Example B

A car dealer is calculating the total sales for the past month, and wants to identify the percentage of monthly sales that have occurred after weeks 1, 2, 3, and 4. Create a *relative cumulative frequency* table to show the information the dealer wants.

TABLE 4.11:

Week Number	Cars Sold
1	21
2	12
3	17
4	24

Solution: First total up the sales for the entire month:

$$21 + 12 + 17 + 24 = 74 \text{ cars}$$

Then find the relative frequencies for each week by dividing the number of cars sold that week by the total:

- The relative frequency for the first week is: $\frac{21}{74} = .284$
- The relative frequency for the second week is: $\frac{12}{74} = .162$
- The relative frequency for the third week is: $\frac{17}{74} = .23$
- The relative frequency for the fourth week is: $\frac{24}{74} = .324$

To find the relative cumulative frequencies, start with the frequency for week 1 and for each successive week, total all of the previous frequencies:

TABLE 4.12:

Week Number	Cars Sold	Relative Frequency	Cumulative Frequency
1	21	.284	.284
2	12	.162	.284 + .162 = .446
3	17	.23	.446 + .23 = .676
4	24	.324	.676 + .324 = 1.0



Note that the first relative cumulative frequency is always the same as the first relative frequency, and the last relative cumulative frequency is always equal to 1.

Example C

Create a cumulative and relative cumulative frequency table for the following data:

Number of customers in store on Black Friday each year for the years 1995 - 2005:

47, 49, 48, 54, 57, 52, 61, 65, 67, 66, 70

Solution: Label a table with the years 1995 - 2005 across the top and frequency, cumulative frequency, and relative cumulative frequency down the side. Leave 3 blanks underneath each year, and input the frequency for each year:

TABLE 4.13:

Year:	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Frequency	47	49	48	54	57	52	61	65	67	66	70
Cumulative Frequency											
Relative Cumulative Frequency											

To calculate the cumulative frequencies, sum the total of all previous frequencies under the frequency for each year:

TABLE 4.14:

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
------	------	------	------	------	------	------	------	------	------	------	------

TABLE 4.14: (continued)

Frequency	47	49	48	54	57	52	61	65	67	66	70
Cumulative Frequency	47	96	144	198	255	307	368	433	500	566	636
Relative Cumulative Frequency											

To calculate the relative cumulative frequencies, divide the cumulative frequency for each year by the total cumulative frequency (636 customers, the same as the value for the last year, 2005).

TABLE 4.15:

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Frequency	47	49	48	54	57	52	61	65	67	66	70
Cumulative Frequency	47	96	144	198	255	307	368	433	500	566	636
Relative Cumulative Frequency	.074	.151	.226	.311	.401	.483	.579	.681	.786	.890	1.0

Concept Problem Revisited

A school lunch break is divided into multiple sessions, the 1st group, consisting of 20 girls and 50 boys, begins lunch at 11:30 am, the 2nd group, with 35 girls and 30 boys begins at 11:40, and the 3rd group, 40 girls and 25 boys, at 11:50.

Assuming that none of the students from an early group leave the cafeteria before everyone goes back to class at 12:10, what is the relative frequency of girls in the cafeteria at 11:35? What about at 11:45? 11:55?

First create a table showing each session and the number of boys and girls in each session:

TABLE 4.16:

Session 1 - 11:30		Session 2 - 11:40		Session 3 - 11:50	
Boys	Girls	Boys	Girls	Boys	Girls
50	20	30	35	25	40

Next, add a row to the bottom of your table for cumulative relative frequency of girls:

TABLE 4.17:

Session 1		Session 2		Session 3	
Boys	Girls	Boys	Girls	Boys	Girls
50	20	30	35	25	40
Girls: $\frac{20}{70} = .286$ or 28.6 %		Girls: $\frac{55}{135} =$.407 or 40.7%		Girls: $\frac{95}{200} =$.475 or 47.5%	

Vocabulary

Relative frequency compares the frequency of a particular value to the entire sample.

Relative cumulative frequency is a running total of *relative* frequencies of all values up to and including the current category.

Cumulative frequency is a running total of all *non-relative* frequencies up to and including the current category (*not* as compared to the total).

Guided Practice

For 25 days, the snow depth at Whiteout Mountain was measured (to the nearest inch) and recorded as follows:

242, 228, 217, 253, 239, 266, 242, 251, 240, 223, 219, 246, 260, 258, 225, 234, 230, 249, 245, 254, 243, 235, 231, 257

1. Determine a reasonable class interval
2. Set up a frequency distribution table
3. Find the frequency for each class interval
4. Calculate the cumulative frequencies progressively from left to right
5. Use the information gathered from the frequency distribution table to plot a cumulative frequency graph.

Solutions:

1. The values range 217 - 266, a class interval of 10 would break the data up into a reasonable 6 bins.
2. and 3: Create a table with the intervals noted across the top, and the frequency of each below:

TABLE 4.18:

210-220	220-230	230-240	240-250	250-260	260-270
2	3	5	7	5	2

4.

TABLE 4.19:

Interval:	210-220	220-230	230-240	240-250	250-260	260-270
Frequency:	2	3	5	7	5	2
Cumulative Frequency:	$\frac{2}{24}$ 8.33%	$\frac{5}{24}$ 20.83%	$\frac{10}{24}$ 41.66%	$\frac{17}{24}$ 70.83%	$\frac{22}{24}$ 91.66%	$\frac{24}{24}$ 100%

Practice Questions:

1. A weather forecaster highlights the lows over-night for the past 30 days in a small town in Wisconsin. The temperature readings are given in Fahrenheit, and are shown below. Use the data to complete the frequency table.

41°, 58°, 61°, 54°, 49°, 46°, 52°, 58°, 67°, 43°, 47°, 60°, 52°, 58°, 48°,
44°, 59°, 66°, 62°, 55°, 44°, 49°, 62°, 61°, 59°, 54°, 57°, 58°, 63°, 60°

TABLE 4.20:

Interval	Tally	Frequency
----------	-------	-----------

TABLE 4.20: (continued)

40 - 44		
45 - 49		
50 - 54		
55 - 59		
60 - 64		
65 - 69		

2. The following represents scores that a class received on their most recent Biology test. Complete the frequency table below.

58, 79, 81, 99, 68, 92, 76, 84, 53, 57, 81, 91, 77, 50, 65, 57, 51, 72, 84, 89

TABLE 4.21:

Interval	Tally	Frequency
50 - 59		
60 - 69		
70 - 79		
80 - 89		
90 - 99		

3. James received the following scores on his quizzes in US History over the course of 1 year. Complete the frequency table below using the scores:

85, 72, 97, 81, 77, 93, 100, 75, 86, 70, 96, and 80.

TABLE 4.22:

Interval/Grades	Tally	Frequency
61 - 70		
71 - 80		
81 - 90		
91 - 100		

4. Sue competed against 14 others in a time trial for the 400-meter run at the state finals. Their times have been recorded in the table below. What percent of the runners completed the time trial between 50 and 53.9 seconds?

TABLE 4.23: 400 Meter Run Time Trails

Interval	Tally	Frequency
50.0 - 50.9		0
51.0 - 51.9		2
52.0 - 52.9	/	6
53.0 - 53.9		3
54.0 - 54.9		4

5. The following data is the starting weights of 30 adults, in pounds, who participated in a study on weight loss. Use the data to create a cumulative frequency table. Determine appropriate intervals for the weights given.

195, 206, 100, 98, 150, 210, 195, 106, 195, 168, 180, 212, 104, 195, 100, 216, 195, 209, 112, 99, 206, 116, 195,

100, 142, 100, 135, 98, 160, 155

TABLE 4.24: Weight of Adults before diet program participation

Interval	Frequency	Cumulative Frequency

6. The following table shows the weights in pounds for students attending a “Get Fit” summer program. Use the data to find cumulative frequency.

TABLE 4.25: Weights of “Get Fit” camp students

Interval	Frequency	Cumulative Frequency
91-100	6	
101-110	3	
111-120	0	
121-130	3	
131-140	0	
141-150	2	
151-160	2	

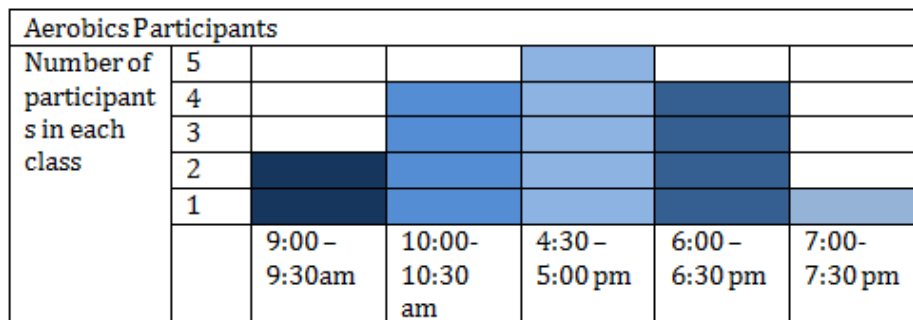
7. 20 students were asked how many meals their family eats out in one week. The results are listed below. Complete the frequency table using this data.

{6, 5, 4, 5, 0, 7, 1, 5, 4, 4, 3, 2, 2, 3, 2, 4, 3, 4, 0, 7}

TABLE 4.26: Number of Meals Out

Interval	Tally	Frequency	Cumulative Frequency
0-1			
2-3			
4-5			
6-7			

8. The following graph shows the participants in Jane’s aerobics classes. How many total students does she teach on the given day?



9. The following table shows a cumulative frequency distribution of the ages of tri-athletes. According to the table, how many tri-athletes are in their forties?

TABLE 4.27: Cumulative Frequency Distribution of Tri-Athlete Ages

Age Group	Total
20 - 29	8
20 - 39	18
20 - 49	25
20 - 59	31
20 - 69	35

10. The quiz scores for a job placement math assessment for the 10 applicants were: 61, 67, 81, 83, 87, 88, 89, 90, 98, and 100. Which frequency table is accurate for this set of data?

TABLE 4.28:

Interval	Freq	Interval	Freq	Interval	Freq	Interval	Freq
61 - 70	2	61 - 70	2	61 - 70	2	61 - 70	2
71 - 80	0	71 - 80	2	71 - 80	0	71 - 80	2
81 - 90	8	81 - 90	8	81 - 90	6	81 - 90	7
91 - 100	10	91 - 100	10	91 - 100	2	91 - 100	10

It is that time of year again. It is Easter, and it is time to hunt eggs. The kids have been divided up into age groups, and their eggs hidden in different areas of a large park. The adult leaders have gathered the data from the three age groups: “Kinders”, “Pee Wee”, and “Big Eggs”, and compiled the results into frequency tables. The problem is, each of the adult leaders reported their information differently.

For questions 11-13, use the data gathered by each adult leader and follow the instructions for interpreting the data given by each.

11. Team “Kinders” coach reported the following results for the number of eggs found by the kiddos in their group, convert them to a cumulative frequency chart.

TABLE 4.29:

Time (s) in minutes	Frequency of Eggs Found
0 - 5	20
5 - 10	35
10 - 15	15
15 - 20	22
20 - 25	8
25 - 30	9

12. Team “Pee Wee” also did a great job of finding eggs. Their results were recorded in a cumulative frequency table, convert to a frequency table.

TABLE 4.30:

Time (s) in minutes	Cumulative Frequency of Eggs Found
0 - 5	20
5 - 10	27

TABLE 4.30: (continued)

10 - 15	31
15 - 20	31
20 - 25	37
25 - 30	50

13. The “Big Eggs” were the champions the last three years running. They are known for always collecting the same amount of eggs in each interval. This year they collected a total of 120 eggs.

- a. Complete a frequency chart
- b. Complete a cumulative frequency chart

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.4.

4.5 Creating Histograms

Objective

Here you will learn how to create *histograms* from data.

Concept

Consider the data from a previous lesson regarding the number of cars sold during each week of a particular month:

TABLE 4.31:

Week Number	Cars Sold
1	21
2	12
3	17
4	24

How would you go about displaying this data as a histogram?

By the end of this lesson, I think you'll see that converting properly organized data into a visual format is really quite straightforward.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63626>

http://youtu.be/KCH_ZDygrm4 MySecretMathTutor - How to Make a Histogram

Guidance

A histogram is really just a rather specific form of a bar chart where the bars are designed to be proportional in area to the frequency and in width to the *interval* that is represented by the bar. The most immediately obvious difference is that there is also no gap between bars, allowing histograms to display continuous data, if it has been organized into intervals.

Creating a histogram is relatively easy once your data has been properly organized, so clearly laying out your data is an important first step. Organizing the data into a table before actually constructing your histogram may also help.

You will want to be sure that you have clearly separated your data into intervals or categories. The number of divisions may depend on your particular data set, but should be few enough not to be confusing to your audience. A

common rule of thumb is to aim for a number of intervals approximately equal to the square root of your number of observations or data points, and generally should be between 5 and 10 intervals. The width of each interval should be approximately the range of your data divided by the number of intervals. In other words, if you have 25 data points ranging from values of 1 to 100, you would expect to have perhaps 5 intervals of 20 units in width each.

Example A

Create a histogram from the following data:

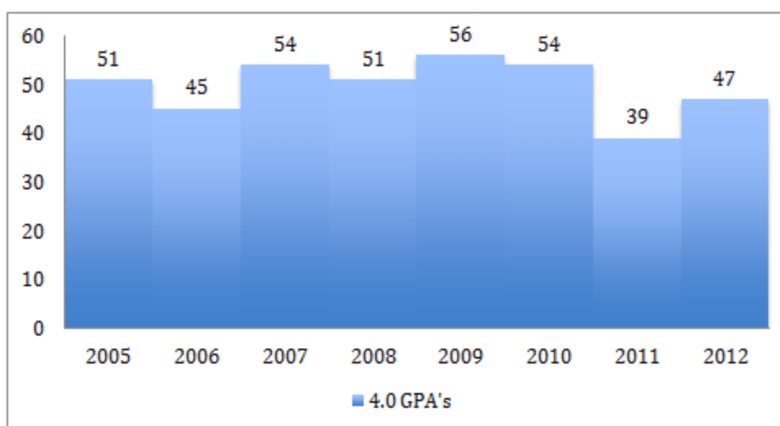
TABLE 4.32:

Year	Number of 4.0 G.P.A.'s
2005	51
2006	45
2007	54
2008	51
2009	56
2010	54
2011	39
2012	47

Solution: The data in this table is already conveniently arranged for display as a histogram. Note that there are only 8 intervals of 1 year each and a single frequency count for each interval. The only thing we need to identify before we actually place this data into a histogram is the frequency range.

The lowest frequency on the table is 39, in year 2011. The highest frequency is 56, in year 2009. That means we have a vertical range of $56 - 39 = 17$ units. Let's plan to label the frequencies from 0 to 60 up the left side of our histogram.

To convert this table into a histogram, all we need to do is create a chart with the intervals (the years 2005 through 2012) across the bottom along the x -axis, and the frequencies (the number of 4.0 G.P.A.'s each year) up the side.



Example B

Create a histogram demonstrating the number of prank calls reported over time using the data below.

TABLE 4.33:

Year	Number of Pranks	Year	Number of Pranks
1970	24	1985	24
1971	42	1986	28

TABLE 4.33: (continued)

1972	38	1987	38
1973	42	1988	40
1974	42	1989	42
1975	27	1990	26
1976	25	1991	39
1977	37	1992	32
1978	27	1993	26
1979	33	1994	31
1980	25	1995	36
1981	32	1996	37
1982	40	1997	38
1983	38	1998	28
1984	26	1999	33

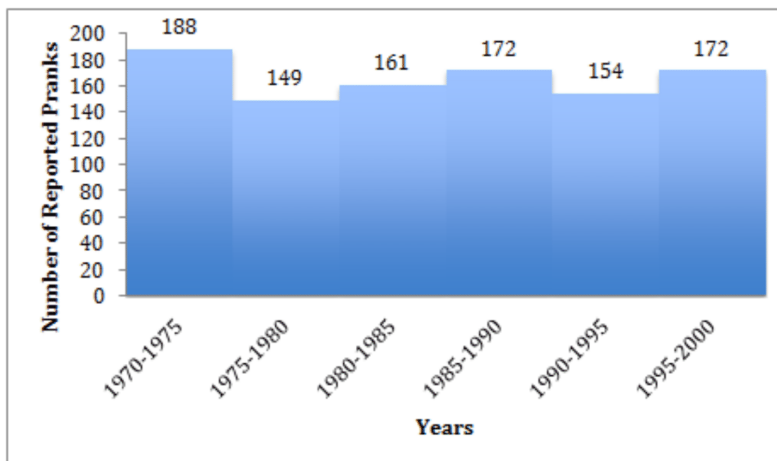
Solution: This table has much more data than Example A. If we were to follow the same procedure, we would end up with 30 intervals, which would make a rather crowded and confusing histogram. We can solve this by *binning* the data before we import it into our histogram.

Binning data means to group the given intervals or categories into broader ranges (bins) in order to limit the number of different intervals. In this case, we have a total of 30 data points, so our 'square-root rule of thumb' would suggest 5-6 bins. If we opt for 5 bins, we would have $\frac{30}{5} = 6$ years per bin, and if we opt for 6 bins, we would have $\frac{30}{6} = 5$ years per bin. 5-year intervals are common, and will be easy for your audience to understand. If we group our original data into 5-year intervals, it looks like this:

TABLE 4.34:

Interval	Number of Pranks
1970 - 1975	$24 + 42 + 38 + 42 + 42 = 188$
1975 - 1980	$27 + 25 + 37 + 27 + 33 = 149$
1980 - 1985	$25 + 32 + 40 + 38 + 26 = 161$
1985 - 1990	$24 + 28 + 38 + 40 + 42 = 172$
1990 - 1995	$26 + 39 + 32 + 26 + 31 = 154$
1995 - 2000	$36 + 37 + 38 + 28 + 33 = 172$

Now we can take our binned data and display it as a histogram with our 5-year intervals along the bottom and number of reported pranks (the frequencies) up the side:



Example C

Create a histogram to display the number of shoe types in several price ranges, using data from the table below:

TABLE 4.35:

Shoe Type	Price	Shoe Type	Type
sandals	\$25	loafers	\$44
sneakers	\$63	high heels	\$28
sneakers	\$43	boots	\$26
sneakers	\$38	high heels	\$43
loafers	\$62	sandals	\$62
sneakers	\$45	loafers	\$30
sneakers	\$56	high heels	\$51
sneakers	\$25	sneakers	\$58
loafers	\$44	high heels	\$63
boots	\$30	loafers	\$43
high heels	\$36	loafers	\$31
loafers	\$29	sneakers	\$56
loafers	\$62	boots	\$54
loafers	\$48	high heels	\$55
sneakers	\$45	sandals	\$40

Solution: The first step is to identify your categories or intervals. In this situation, we are expected to compare the frequency of shoe types in price range intervals, so we will need to decide what our intervals should be by dividing up the price ranges appropriately.

The lowest price in the data set is \$25, and the highest is \$63. This gives us a *range* of \$38, so we can round that up to \$40. \$40 divides conveniently by 5, so we could go with 8 intervals of \$5 each. That may be a bit more detailed than we really need since there are only 30 data points, which would suggest 5 or 6 intervals rather than 8, but we are still under 10, and the \$5 intervals will be easy for our audience to identify with.



Next we need to identify a range of frequencies for our selected intervals, which means we will need to count the number of entries in each interval using our question data. At this point we can either 'hunt through' the data, counting as we go, or we can re-organize the original data by price to simplify the counting process. Since we are not actually using the data regarding the names of each shoe type, only the frequency based on price, we can just organize the prices in dollars by ascending order:

25, 25, 26, 28, 29, 30, 30, 31, 36, 38, 40, 43, 43, 43, 44, 44, 45, 45, 48, 51, 54, 55, 56, 56, 58, 62, 62, 62, 63, 63

Then we can group the list according to our chosen interval of \$5:

25, 25, 26, 28, 29 | 30, 30, 31 | 36, 38 | 40, 43, 43, 43, 44, 44 | 45, 45, 48 | 51, 54 | 55, 56, 56, 58 | 62, 62, 62, 63, 63

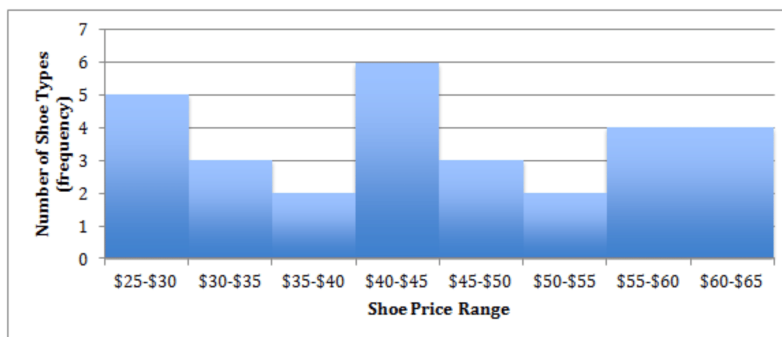
By convention, a value equal to the lower limit of an interval is included, but a value equal to the upper limit is not (upper and lower interval limits are also called *class limits*). For example, a \$40 entry is included in the \$40-\$45 interval since \$40 is the lower limit of that interval, but the two \$45 entries are included in the next interval up (\$45-\$50), since \$45 would be the upper limit of the \$40-\$45 interval.

Finally, we take the grouped data and summarize by frequency in each interval:

TABLE 4.36:

Interval (Price range)	Frequency (Number of shoe types)
\$25-\$30	5
\$30-\$35	3
\$35-\$40	2
\$40-\$45	6
\$45-\$50	3
\$50-\$55	2
\$55-\$60	4
\$60-\$65	4

Finally we construct our histogram frame with intervals along the bottom and frequencies up the side:



Concept Problem Revisited

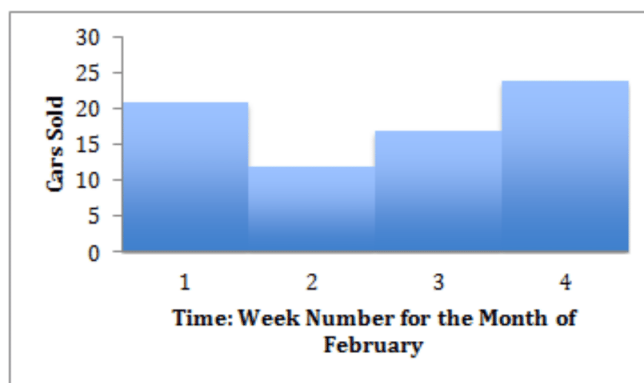
Consider the data from a previous lesson regarding the number of cars sold during each week of a particular February:

TABLE 4.37:

Week	Cars
1	21
2	12
3	17
4	24

How would you go about displaying this data as a histogram?

Compared to the examples you completed during the lesson, this is a piece of cake! The data is already organized into 4 convenient bins, and the number of cars sold has been totaled for each bin. Just create your histogram with weeks across the bottom and number of cars sold up the side:



Vocabulary

A **Histogram** is a specific form of a bar chart with zero space between each bar, a bar area that is proportional to the frequency it represents, and a bar width equal to the data interval.

An **Interval** is a range of data. Grouping data into intervals can be beneficial in a number of ways, including simplifying the appearance and minimizing the effect of individual measurement errors.

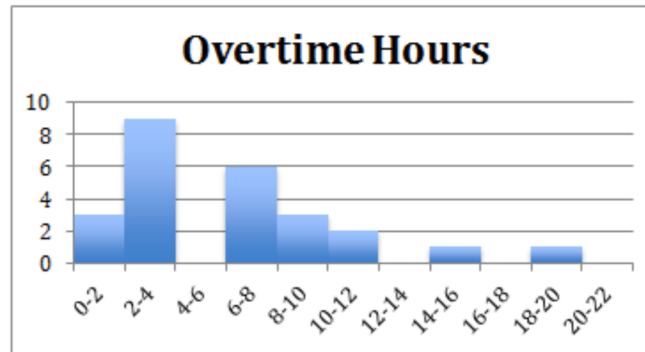
A **Range** is a value representing the difference between the least and greatest value in a data set. The range can be found by subtracting the smallest value from the largest.

Binning is the process of grouping data ranges into appropriate intervals. There is no all-around best number of bins, and different numbers of bins can reveal different things about a data set.

Class limits are, collectively, the upper and lower limit of an interval. A **class mark** is the middle value, or average of the class limits.

Guided Practice

The head of human resources wants to know how many of the department heads work overtime in a given month. The data is displayed in the following histogram. Use it to answer questions 1 - 3.



1. What do the numbers on the horizontal axis represent?
2. What do the numbers on the vertical axis represent?
3. What percent of department heads work 6 or more hours of overtime in a given month?
4. Here are the driving scores of 15 new drivers. Create a histogram from their scores:
88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73
5. Using the data from the created histogram, identify the percentage of people who did not pass the driving test if the minimum score was 70%.

Solutions:

1. The horizontal axis shows intervals of hours worked overtime.
2. The vertical axis displays the number of department heads falling within each interval.
3. 52%. The histogram displays information about 25 employees. Out of these 25 employees 13 worked more than 6 hours overtime. We divide $\frac{13}{25}$ and come up with .52, or 52% who work more than 6 hours of overtime.
4. First we should determine how to break the range of values into intervals. In this instance, since our data set consists of driving scores, it would make sense to choose intervals of 10 points: 40-50, 50-60, ... 90-100, since most tests are determined by a certain percentage. By counting how many of the 25 observations fall in each of the intervals, we get the following table:

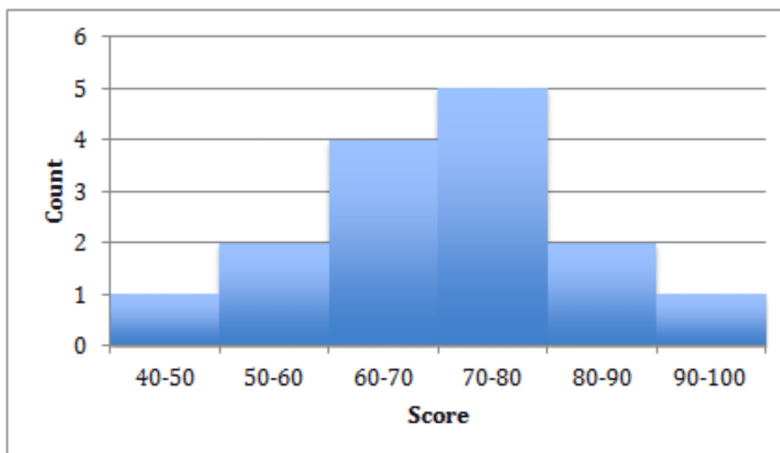
TABLE 4.38: Driving Scores

Score	Count
[40-50)	1
[50-60)	2
[60-70)	4
[70-80)	5

TABLE 4.38: (continued)

[80-90)	2
[90-100]	1

To construct the histogram from this table, we plot the intervals on the X -axis, and represent the number of observations in each interval (the frequency of the interval) on the Y -axis, by the height of a rectangle located above the interval:



5. 47% - First we must find the total number of people who took the driving test. It looks like this: $1 + 2 + 4 + 5 + 2 + 1 = 15$. 15 people took the test. Now we identify how many received below 70 %, $1 + 2 + 4 = 7$. To find out the percentage, we divide the number of people who earned less than 70%, or 7, by the total number of people who took the driver's test, 15. $\frac{7}{15} = .47$. We can say that 47% of the people who took the driver's test did not pass.

Practice

The following numbers need to be used in creating a histogram. Use this data to answer questions 1-4

9.3, 10.6, 10.6, 10.6, 10.2, 10.7, 9.9, 10.5, 10.4, 8.8, 9.6, 9.8, 9.2, 10.9, 10.0, 11.4, 10.8, 8.9, 10.4, 8.9, 9.6, 8.4, 11.2, 10.3, 10.1, 9.6, 11.1, 9.4, 9.4, 10.6, 8.9, 10.8, 9.2.

- Identify the range for the above data.
- Based on the data, what would be an appropriate number of Intervals?
 - 7-10
 - 11-15
 - 13-20
- If the width of an interval is .4, and the lowest reading is 8.4, it is the starting point. What will be the upper end of the first interval? The second, third and so on?
- Create a histogram for the data set above. Remember that numbers that fall on the high end of one interval and the low end of the next interval should be placed in the lower interval. For instance in the interval 8.4 - 8.8, only 8.4, 8.5, 8.6, and 8.7 belong, 8.8 would be the lowest value in the next interval.
- The following table shows a cumulative frequency distribution of the ages of tri-athletes. Convert the data to standard frequencies, and create a histogram to graph the absolute frequencies.

TABLE 4.39: Cumulative Frequency Distribution of Tri-Athlete Ages

Age Group	Total
20-29	8
20-39	18
20-49	25
20-59	31
20-69	35

6. The quiz scores for a job placement math assessment for the 10 applicants were: 61, 67, 81, 83, 87, 88, 89, 90, 98, and 100. Using the table below create a histogram to graph the data.

TABLE 4.40:

Interval	Frequency
61-70	2
71-80	0
81-90	6
91-100	2

7. Create a Histogram using the following list of overnight lows:

41°, 58°, 61°, 54°, 49°, 46°, 52°, 58°, 67°, 43°, 47°, 60°, 52°, 58°, 48°, 44°, 59°, 66°, 62°, 55°, 44°, 49°, 62°, 61°, 59°, 54°, 57°, 58°, 63°, 60°

And the following intervals:

40-44

45-49

50-54

55-59

60-64

65-69

8. The following represents scores that a class received on their most recent Biology test. Create a histogram from the scores.

58, 79, 81, 99, 68, 92, 76, 84, 53, 57, 81, 91, 77, 50, 65, 57, 51, 72, 84, and 89.

Use 10-point intervals starting with the interval 50-59.

9. James received the following scores on his quizzes in US History over the course of 1 year. Create a histogram from the scores.

85, 72, 97, 81, 77, 93, 100, 75, 86, 70, 96, and 80.

10. Sue competed in a time trial for the 400-meter run at the state finals. She ran against 14 others. Their times have been recorded in the table below. Complete the chart and create a histogram to represent the data.

TABLE 4.41: 400 Meter Run Time Trials

Interval	Tally	Frequency
50.0-50.9		
51.0-51.9		

TABLE 4.41: (continued)

52.0-52.9		
53.0-53.9		
54.0-54.9		

11. The following data is the weight of 30 adults, in pounds, who participated in a study on weight loss. Use the data to create a cumulative frequency table. Determine appropriate intervals for the weights given.

195, 206, 100, 98, 150, 210, 195, 106, 195, 168, 180, 212, 104, 195, 100, 216, 195, 209, 112, 99, 206, 116, 195, 100, 142, 100, 135, 98, 160, 155

12. The following table shows the weights in pounds for students attending a “Get Fit” summer program. Create a histogram to graph the data.

TABLE 4.42:

Interval	Frequency
91-100	6
101-110	3
111-120	0
121-130	3
131-140	0
141-150	2
151-160	2

13. The graph below shows the distribution of scores of 30 students on a history exam. Complete the frequency table below it using the data.

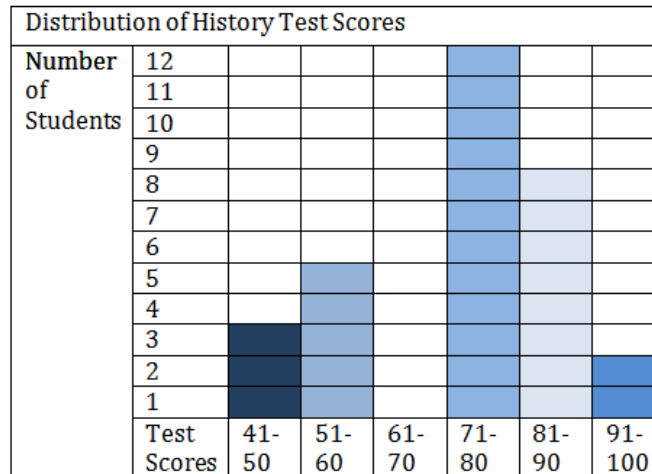


TABLE 4.43:

Test Scores	Frequency
91-100	
81-90	
71-80	
61-70	
51-60	

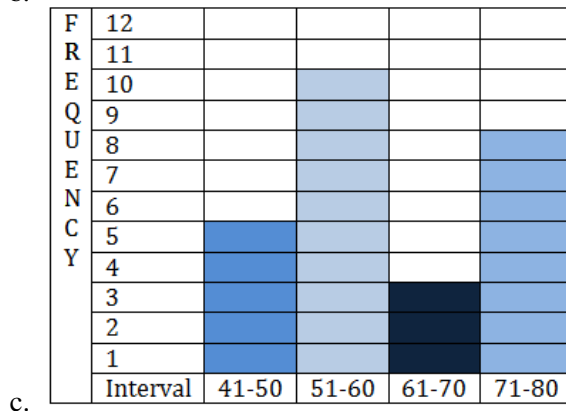
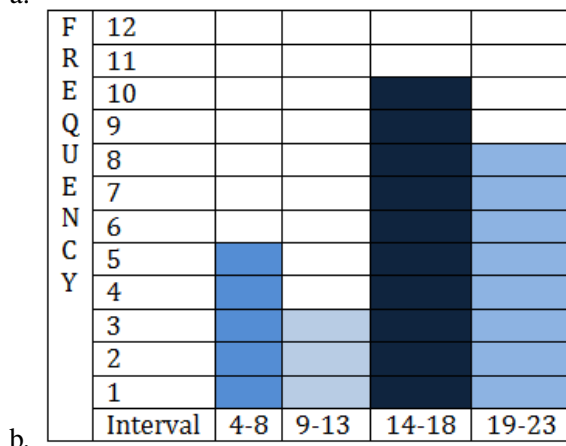
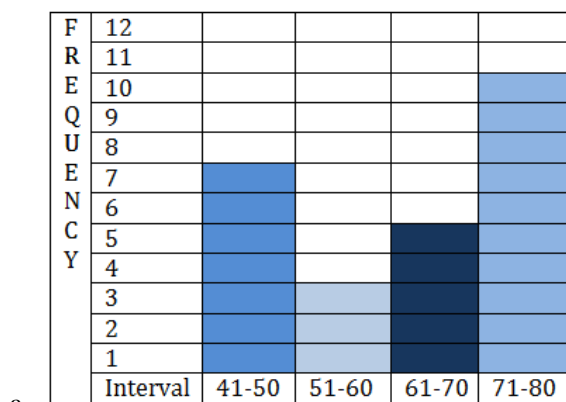
TABLE 4.43: (continued)

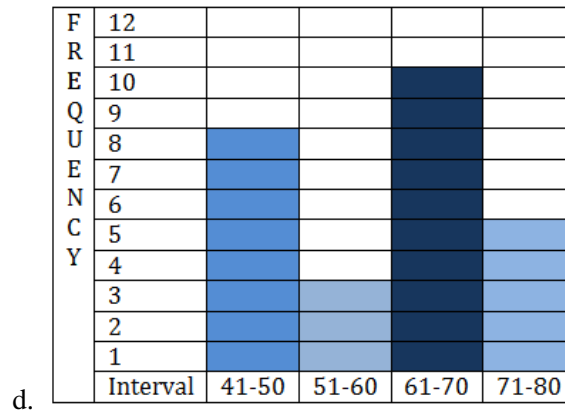
41-50	
-------	--

14. Which one of the following histograms represents the data in the table below?

TABLE 4.44:

Interval	Frequency
4-8	8
9-13	3
14-18	10
19-23	5





A teacher was asked how many students they had to ask to stop texting on their cell phones during each of their class periods on a given day.

The following data was collected.

TABLE 4.45:

Class Period	Frequency
1	16
2	11
3	5
4	3
5	1

15. Convert the data to a histogram.

16. How many class periods did a teacher have to ask less than 9 times?

17. How many times total did the teacher have to ask students to stop texting?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.5.

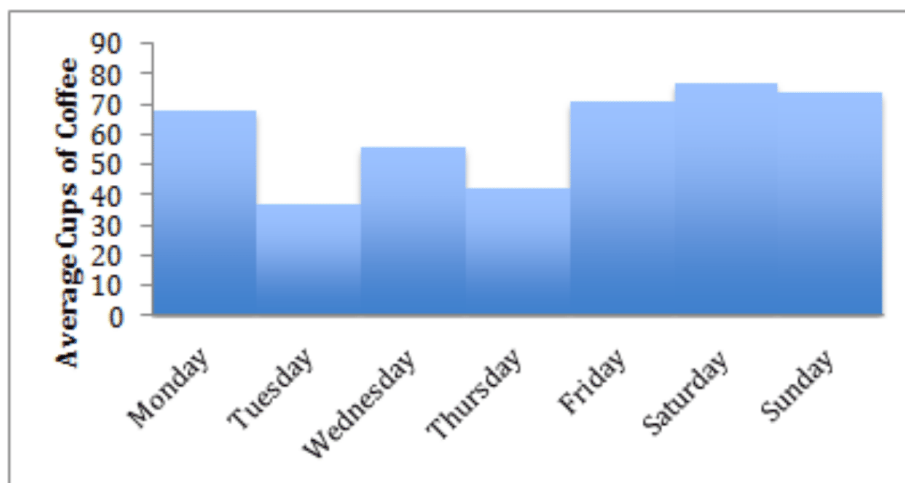
4.6 Interpreting Histograms

Objective

Here you will learn how to read a histogram to develop conclusions based on data.

Concept

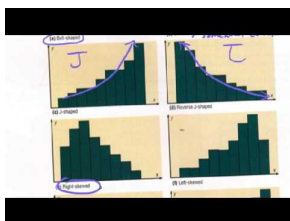
Brian runs a small business selling coffee to morning commuters. Like any other small business owner, he would like to increase his traffic. He decides to offer a “buy 10, get one free” punch-card promotion, and figures that if he gives double credit to customers for coffees purchased on days with generally slower sales, he can increase his overall traffic. If the histogram below represents the average number of coffees sold during a week, what can he conclude about the best and worst day(s) to offer “double punches”? How would you describe the shape of his histogram? Does the shape reveal any particularly useful data in this instance?



This lesson will help you become familiar with using histograms. At the end of the lesson, we will return to this question to apply your skills.

Watch This

This is a good video describing the different shapes of data distributions. The instructor spends quite a bit of time at the beginning on the “bell curve” of a normal distribution, that we will be discussing in some detail later, so you may start at apx 3:30 if you wish to save time.



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63707>

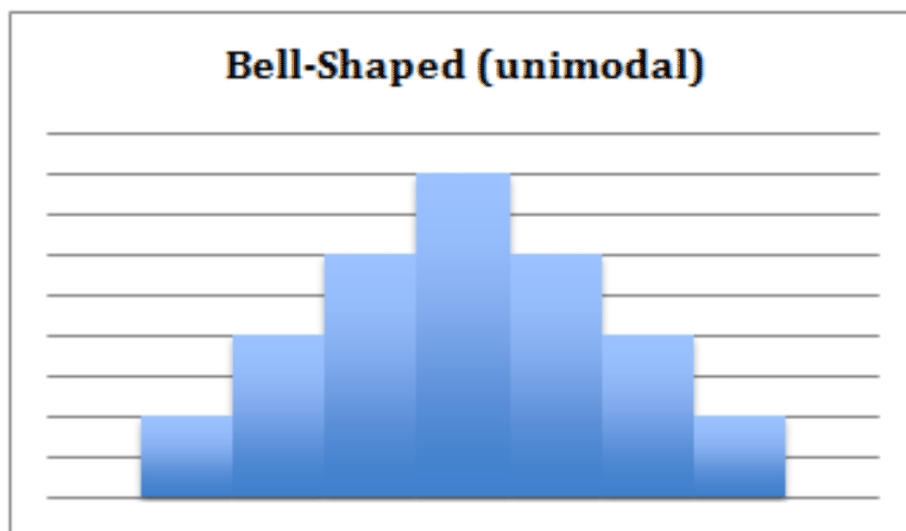
<http://youtu.be/LIT7qBmrBxA> mattemath - Types of Data Distributions (shapes and names)

Guidance

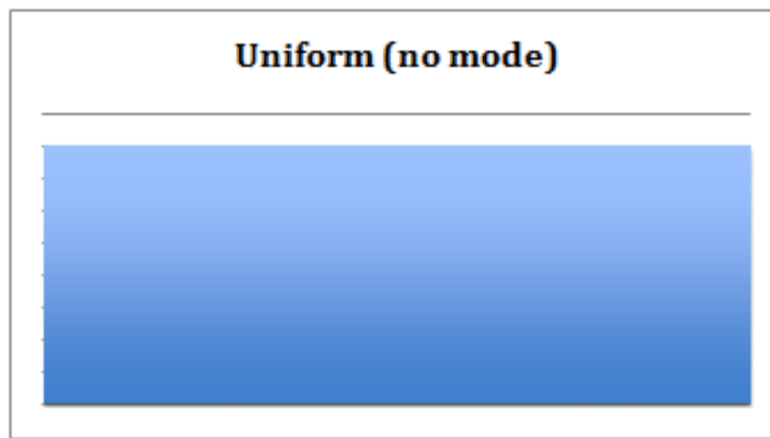
Histograms are a very common method of visualizing data, and that means that understanding how to interpret histograms is a valuable and important skill in virtually any career. There are a number of things to pay particular attention to when reading a histogram, including:

- **Range:** Recall that a range is a description of the difference between the greatest and least values in a given data set. On a histogram, this is important in two particular ways:
 - How widely dispersed are the frequencies of each bin? Extremely large frequency ranges (particularly as a percentage) may indicate data that is fundamentally unreliable.
 - How wide are the bins themselves? Specifically, how broad are the intervals or how descriptive are the classes? Unusually large or small intervals, or unusually broad or narrow categories may indicate important observations about the data as a whole.
- **Frequency Density:** The major difference between a bar graph and a histogram is the way in which the frequencies of each class or interval are represented. On a **bar graph**, the frequency is the *height* of the bar. On a **histogram**, the frequency is measured by the *area* of the bar. What that means is that you can use a histogram with different interval or class widths to represent data with varying densities. (See Example C)
- **Shape:** The shape of a histogram can lead to valuable conclusions about the trend(s) of the data. In fact, the shape of a histogram is something you should always note when evaluating the data the histogram represents. Some common shapes and their indications are:

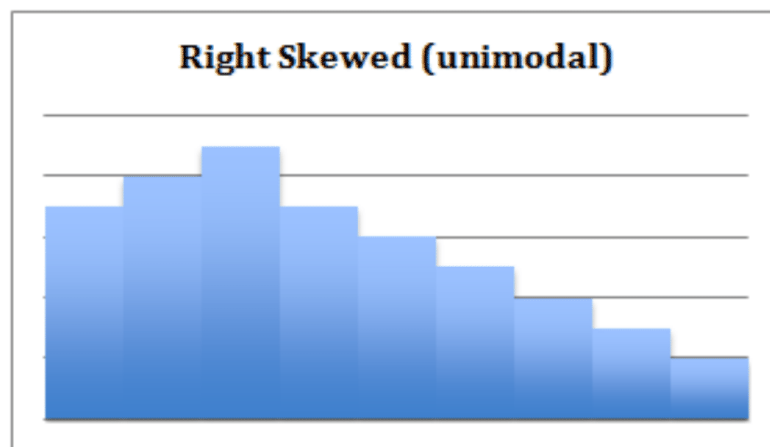
a. **Bell-Shaped:** A histogram with a prominent 'mound' in the center and similar tapering to the left and right. One indication of this shape is that the data is **unimodal** - meaning that the data has a single mode, identified by the 'peak' of the curve. If the shape is symmetrical, then the mean, median, and mode are all the same value. Note that a **normally distributed** data set creates a symmetric histogram that looks like a bell, leading to the common term for a normal distribution: a **bell curve**.



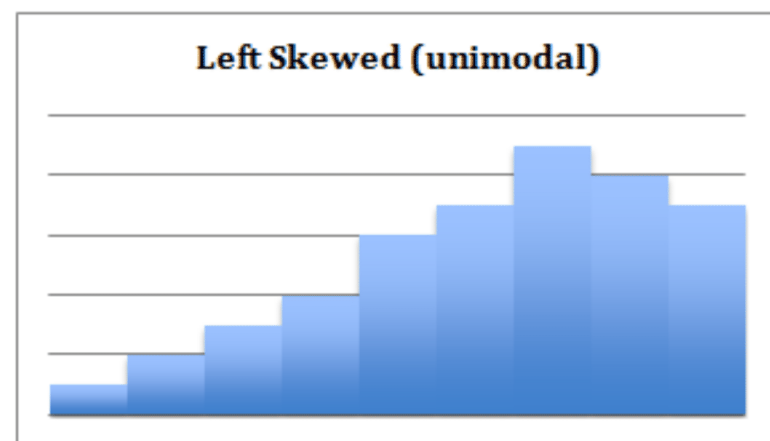
b. **Uniform:** A uniform shaped histogram indicates data that is very consistent; the frequency of each class is very similar to that of the others. A data set with a uniform-shaped histogram may be **multimodal** - the having multiple intervals with the maximum frequency. One indication of a uniform distribution is that the data may not be split into enough separate intervals or classes. Another possibility is that the scale of the histogram may need to be adjusted in order to offer meaningful observations.



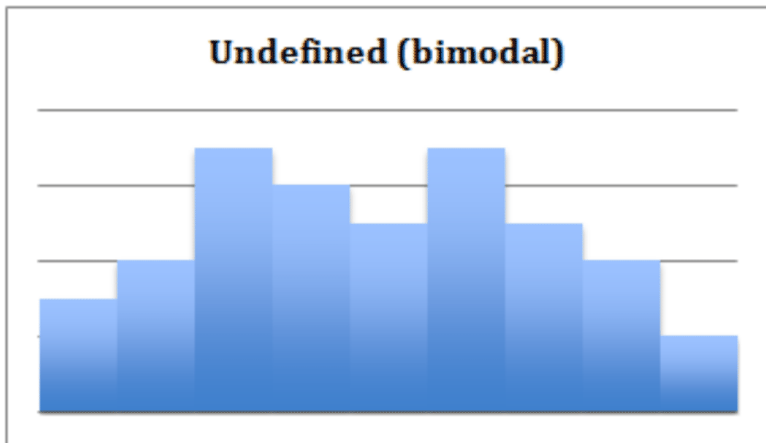
c. **Right-Skewed:** A right-skewed histogram has a peak that is left of center and a more gradual tapering to the right side of the graph. This is a unimodal data set, with the mode closer to the left of the graph and smaller than either the mean or the median. The mean of right-skewed data will be located to the right side of the graph and will be a greater value than either the median or the mode. This shape indicates that there are a number of data points, perhaps **outliers**, that are greater than the mode.



d. **Left-Skewed:** A left-skewed histogram has a peak to the right of center, more gradually tapering to the left side. It is unimodal, with the mode closer to the right and greater than either mean or median. The mean is closer to the left and is lesser than either median or mode. This shape indicates that the preponderance of any outliers is lesser than the mode.

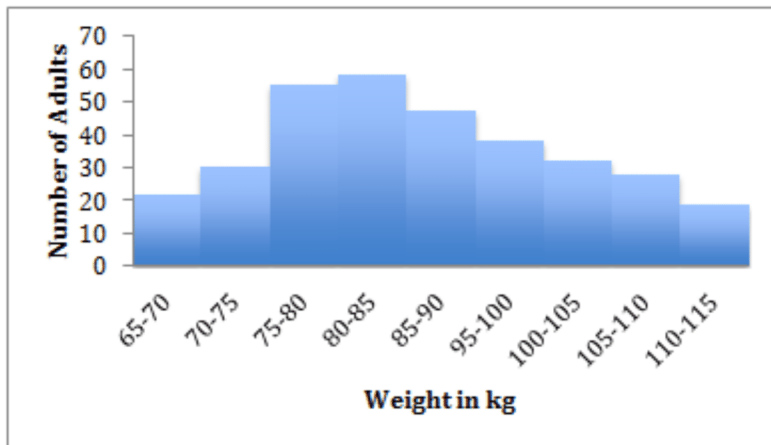


e. **Undefined Bimodal:** This shape is not specifically defined, but we can note regardless that it is bi-modal, having two separated classes or intervals equally representing the maximum frequency of the distribution.



Example A

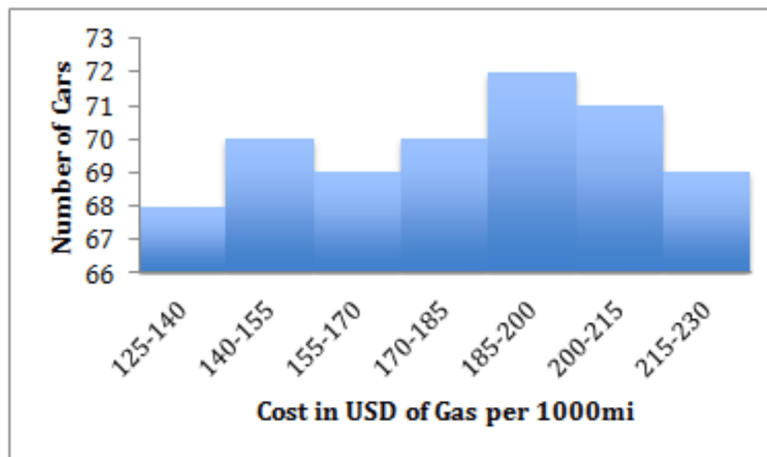
Describe the shape of the histogram and state a few notable characteristics:



Solution: This is a right-skewed distribution, indicating that there are a number of values greater than the mode. If the modal class of 80-85kg represents a healthy normal weight, this graph would suggest a sample that tended toward being overweight.

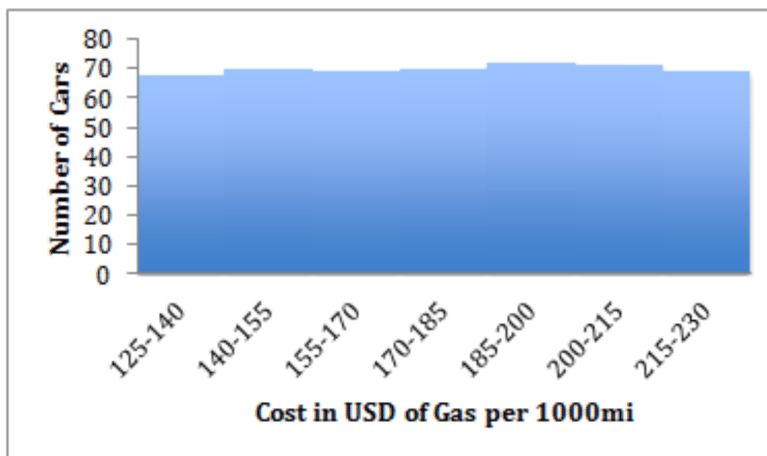
Example B

Identify the general shape of the histogram and what the shape indicates about the data:



Solution: This is a slightly tricky one. The overall shape appears somewhat left-skewed and obviously unimodal at first glance. However a closer look tells a different story, note that the overall range is $72 - 68 = 4$ cars. That is a very small range, only about 5% of the mean. The shape is deceiving in large part because the vertical axis does not start at 0, which exaggerates the differences between the classes.

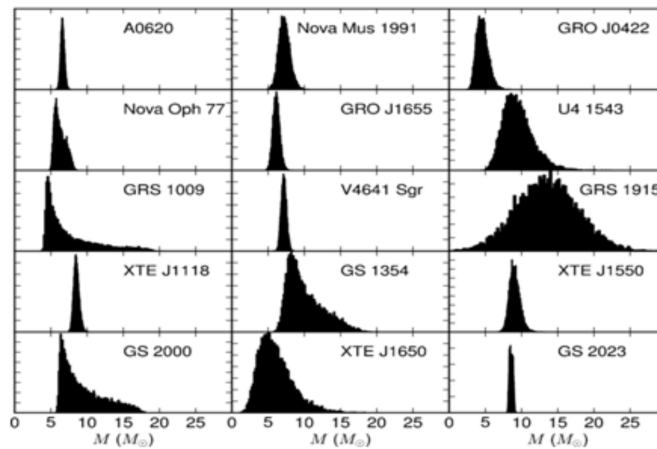
Look what happens if we re-draw the histogram *with the same data* but with the vertical axis at 0:



Pretty huge difference, isn't it? Now it is apparent that this is really a pretty uniform distribution, and that there is not a very meaningful difference in frequency between the classes.

Example C

The image below represents data on the relative masses of a number of sampled black holes.



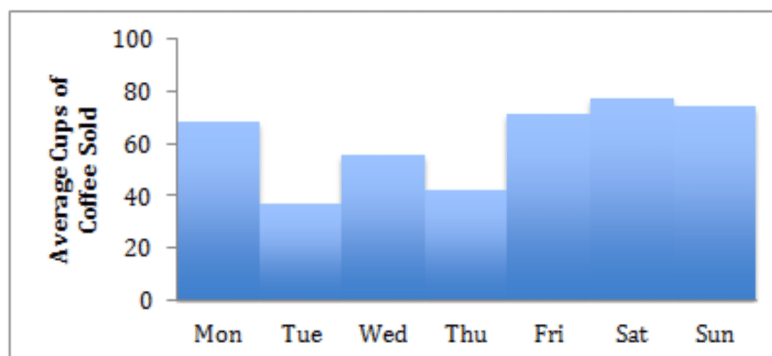
Evaluate the group of histograms as a whole; identify the common shapes and any notable features. Data source: <http://iopscience.iop.org>

Solution: Most of the individual histograms are clearly unimodal, and all are clustered rather closely around a single peak, with the exception of GRS 1915. Most of the graphs appear largely symmetrical, with the others being right-shifted. The sharp and narrow peaks in most of the plots suggest that the mass measurements are generally consistent. The location of the majority of the peaks at the same general location on the scale would suggest that the masses of the different black holes appear similar at this scale. The tendency of the non-symmetrical plots to be right-shifted suggests that it would be more reasonable to favor slightly greater mass estimates than slightly lesser ones.

The GRS 1915 plot is notably different, and the broad peak suggests that perhaps clear data on the mass of that particular black hole is difficult to come by.

Concept Problem Revisited

If the histogram below represents the average number of coffees sold during a week, what can he conclude about the best and worst day(s) to offer “double punches?” How would you describe the shape of his histogram? Does the shape reveal any particularly useful data in this instance?



Brian should note that he is currently receiving less traffic on Tuesdays and Thursdays than he is the rest of the week. Those two days would be ideal for his “double punches”. This particular histogram does not have a well-defined shape, and therefore no particular information is liable to be pulled from it.

We might note that Friday, Saturday, Sunday, and Monday all in fact occur subsequently, so it could be said that the *data* suggests a peak during those days. As drawn, however, the histogram does not.

Vocabulary

Multimodal histograms have more than one 'peak' in the data. Recall that the mode is the most common value, so a multimodal histogram represents data with multiple classes that have a frequency equal to the greatest single frequency in the data.

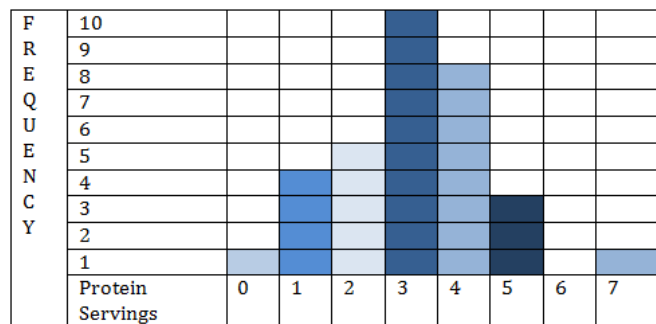
Unimodal histograms have a single peak, and represent data with a single most common frequency.

Outliers are uncommon frequencies occurring some distance from the peak. Specifically, data points that are more than 1.5 times the middle half of your data above the upper, or below the lower, quartiles may be considered **mild outliers** and points more than 3 times the middle half of your data are **extreme outliers**.

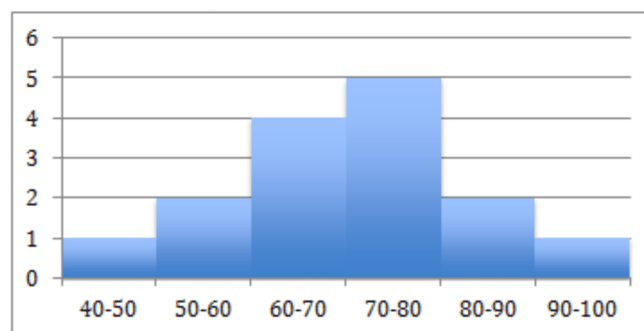
A **normal distribution** creates a histogram in the shape of a bell. This **bell curve** makes it clear that the majority of the data lies close to the mean.

Guided Practice

A random sampling was taken of pedestrians on a street corner walking to work one day. They were asked how many servings of protein they ate, on average, on a daily basis. The data collected is displayed in the histogram below:



1. How many people say that they eat at least 4 servings of protein daily?
2. What percentage of people surveyed eat no more than 3 servings of protein daily?
3. What proportion of the people surveyed eats exactly 5 servings of protein daily?
4. What type of distribution does the histogram below display?



- a. Symmetric, single peaked (unimodal) distribution
- b. Symmetric, double peaked (bimodal) distribution
- c. Skewed left distribution

d. Skewed right distribution

5. Using the image from question 4, determine the spread and any outliers of this graph.

Solutions:

1. 8 people claim four servings per day, 3 claim five servings, and 1 claims seven servings, for a total of 12

2. 32 people responded, and of them 20 people eat 3 servings or less.

To find the percentage, divide the number who eat 3 or fewer servings by the total number of responses: $\frac{20}{32} = 62\%$

3. 3 people claim five servings per day.

To find the percentage, divide: $\frac{3}{32} = .094$ or $\approx 10\%$

4. This is a symmetric, single peaked (unimodal) distribution.

5. No outliers.

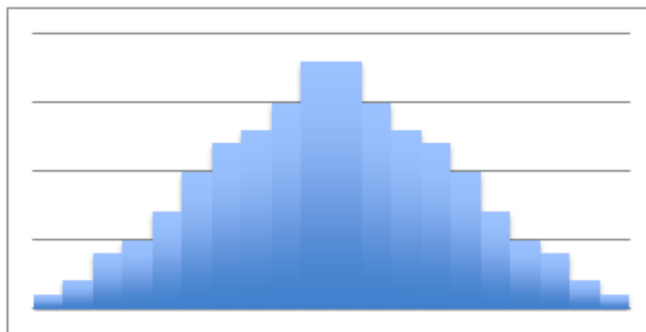
TABLE 4.46:

approximate min:	45 (the middle of the lowest interval of scores)
approximate max:	95 (the middle of the highest interval of scores)
approximate range:	$95 - 45 = 50$

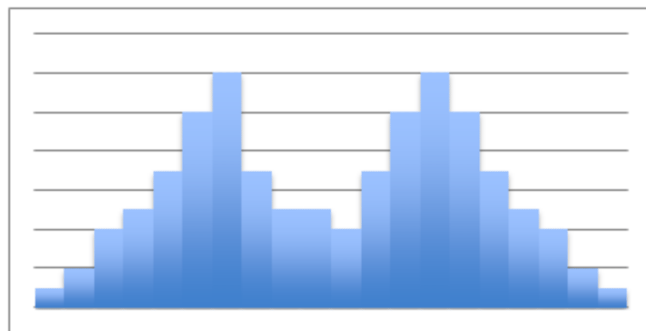
Practice

Identify which images show symmetric distributions and which show skewed distributions. Identify what type of symmetric or skewed distributions are displayed.

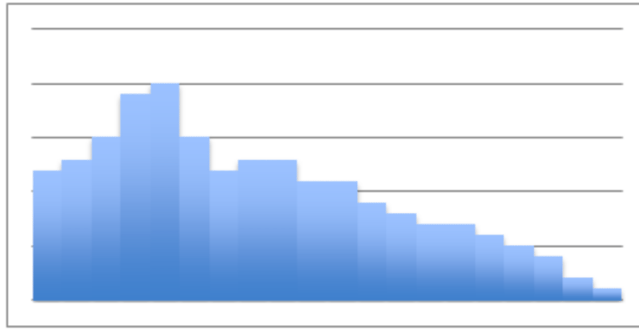
1.



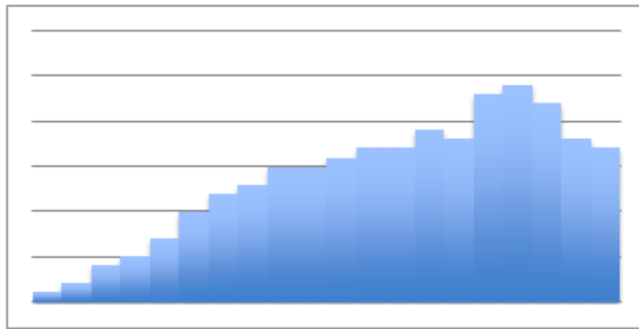
2.



3.



4.



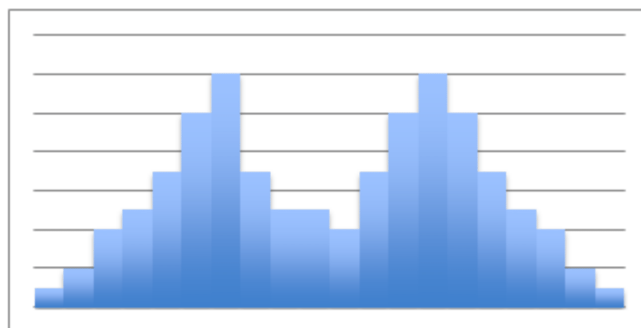
5. What do you think is the shape of the distribution of the age at which a child takes its first steps? Why?

- Symmetric - Uniform
- Skewed left
- Skewed right
- Symmetric - Unimodal
- Symmetric - Bimodal

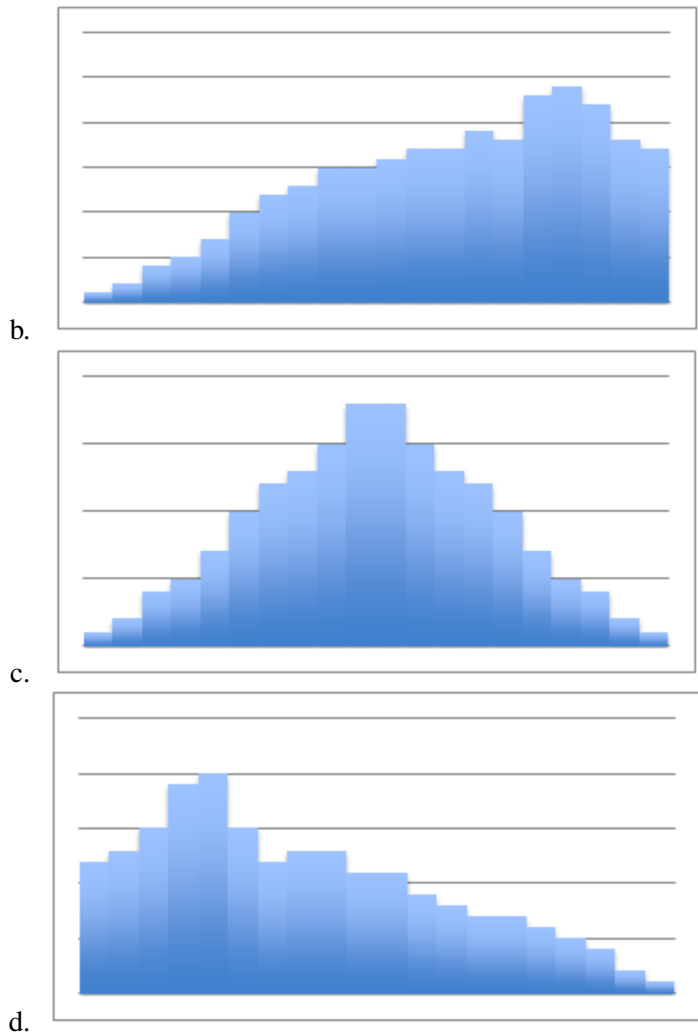
6. What do you think is the shape of the distribution of rolling a 6-sided die 1,000 times is? Why?

- Symmetric - Uniform
- Skewed left
- Skewed right
- Symmetric - Unimodal
- Symmetric Bimodal?

7. Match the graph with the data it most likely displays.



a.



SAT Math Scores of future doctors and engineers.

Prices of 1,000 homes within a given geographical area.

Cholesterol levels of 1000 adults.

Men’s & women’s clothing sizes.

The data below shows the number of surveyed people, and their respective ages, who enjoy riding roller coasters.

Use the histogram below to answer questions 8-11.

F R E Q U E N C Y	10											
	9											
	8											
	7											
	6											
	5											
	4											
	3											
	2											
	1											
	AGE	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-50	51-55	56-60

8. What is the shape of this histogram?

9. What is the center of this histogram?

10. What is the Spread of this histogram?

11. What are the outliers of this histogram?

Use the histogram below to answer questions 12-15.

70				█	
60				█	
50				█	
40			█	█	
30			█	█	
20		█	█	█	█
10	█	█	█	█	█
0	20	40	60	80	100

12. What is the shape of this histogram?

13. What is the center of this histogram?

14. What is the Spread of this histogram?

15. What are the outliers of this histogram?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.6.

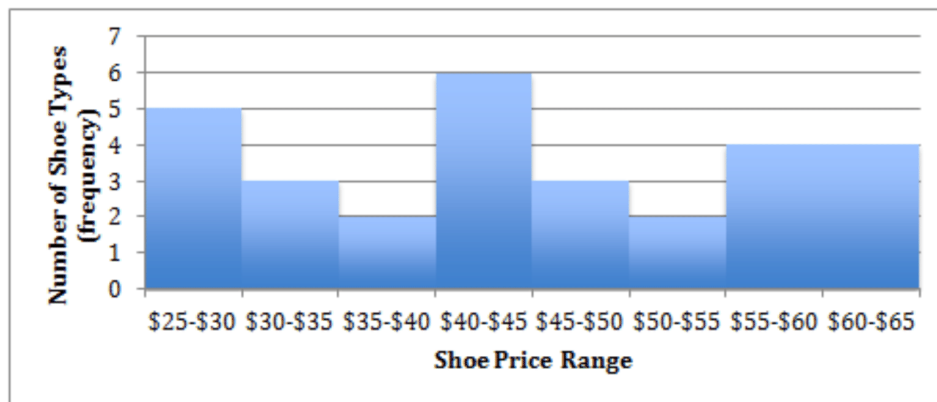
4.7 Frequency Polygons - Probability and Statistics

Objective

Here you will learn to create frequency polygons and you will learn about the differences between frequency polygons and histograms or bar charts.

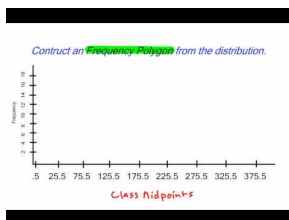
Concept

The histogram below is built from data given in the lesson on histograms. How would you convert the histogram into an absolute frequency polygon or a relative frequency polygon?



After the lesson below, we will return to this question and apply what we have learned.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63714>

<http://youtu.be/gy4dgzxmT4A> Steve Mays - Frequency Polygon

Guidance

A frequency polygon is closely related to a histogram. In fact, the difference is really only in the actual construction of the graph. Whereas a histogram is built of bins with a width representing the interval, and a height representing the quantity of data points in each interval, a frequency polygon is constructed by drawing a point to represent the

frequency of a particular interval and connecting that point to the one representing the frequency of the next interval. The result is a shape very much like a histogram constructed from the same data, but with points instead of columns.

The primary purpose of a frequency polygon is to allow histogram-like data representation of two sets of data on the same graph. Two histograms on the same graph tend to shroud each other and make comparison more difficult, but two frequency polygons can be graphed together with much less interference.

There are two common varieties of frequency polygon: *absolute frequency* and *relative frequency*. The difference between them is how the frequency is counted. An *absolute frequency* polygon has 'peaks' that represent the actual number of points in the associated interval. A *relative frequency* polygon has peaks that represent the percentage of total data points falling within the interval.

To create an absolute frequency polygon:

- Construct a frame just as you would for a histogram.
- Label the vertical axis with the range of frequencies to be graphed, and the horizontal axis with the intervals you have chosen. Make your horizontal axis long enough to include a full interval above and below your graphed data so that the finished polygon has a visible starting and ending point.
- Sum the number of points in each interval and mark a point representing the sum along the midline of the interval. The midline is on the arithmetic mean of the each interval, and can be calculated by adding the lower and upper limits of each interval and dividing the sum by two.
- Once all points have been accounted for, connect the points and color in the area under the line.
- If you are graphing a second set of data, repeat the process.

To construct a relative frequency polygon:

- Construct a frame just as you would for a histogram.
- Label the vertical axis from 0 - 100%, and the horizontal axis with the intervals you have chosen.
- Sum the number of points in each interval, divide the sum of each interval by the total number of data points, and multiply by 100. The result is the percentage of the total number of data points that is represented by each interval. Mark a point representing the percentage along the midline of the interval.
- Once all points have been accounted for, connect the points and color in the area under the line.
- If you are graphing a second set of data, repeat the process.

Example A

Construct an absolute frequency polygon to represent the data in the table.

TABLE 4.47: Time Spent Playing with Toys vs Unwrapping Presents and Eating on Christmas Day

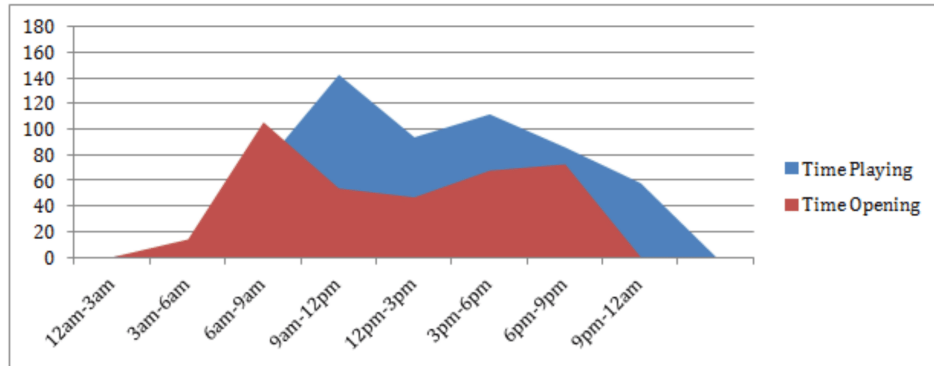
	12am - 3am	3am - 6am	6am - 9am	9am - 12pm	12pm - 3pm	3pm - 6pm	6pm - 9pm	9pm - 12am
Time Playing	0 min	10 min	72 min	143 min	94 min	112 min	86 min	58 min
Time Opening	0 min	14 min	106 min	54 min	47 min	68 min	73 min	0 min

Solution: First construct the frame for your graph; label the y -axis from 0 - 180 minutes, and the x -axis with each of the intervals described in the table (12am-3am, etc.).

Next, mark a single point along the midline of each bin, at the appropriate height to represent the frequency of minutes spent playing during that interval.

Finally, connect the points and color in the area between the x -axis and the line. Your completed frequency polygon

should look something like this:



Example B

Construct a relative frequency polygon using the psychology test score data in the table.

TABLE 4.48:

Interval		Frequency
29.5	39.5	0
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1
169.5	179.5	0

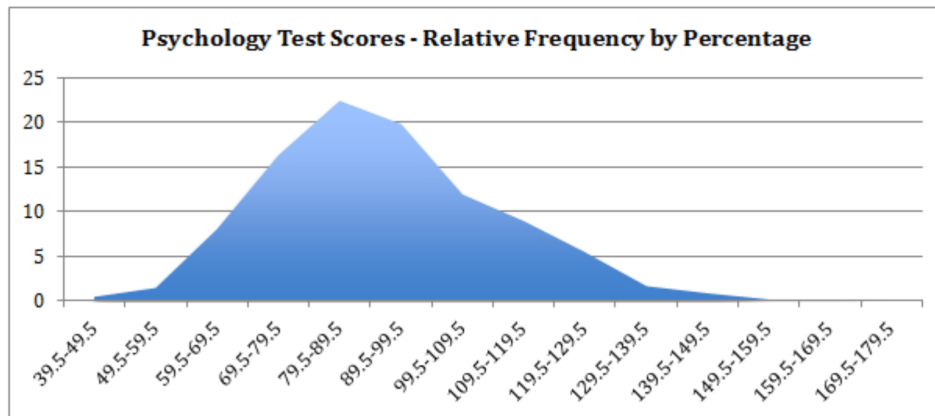
Solution: The data has been organized into intervals and the frequency for each has been noted. In order to convert the stated *absolute* frequencies into *relative* frequencies, first find the sum of the frequencies from all intervals and then divide each interval's frequency by the total frequency and multiply by 100 to get the percentage.

$$0 + 3 + 10 + 53 + 107 + 147 + 130 + 78 + 59 + 36 + 11 + 6 + 1 + 1 + 0 = 652 \text{ total scores}$$

- Relative frequency for the 1st interval, including scores between 29.5 and 39.5:
 - $\frac{0}{652} \times 100 = 0\%$
- Relative frequency for the 2nd interval, including scores between 39.5 and 49.5:
 - $\frac{3}{652} \times 100 \approx .5\%$ (round to 0%)
- Relative frequency for the 3rd interval, including scores between 49.5 and 59.5:
 - $\frac{10}{652} \times 100 \approx 1.5\%$

- Relative frequency for the 4th interval, including scores between 59.5 and 69.5:
 - $\frac{53}{652} \times 100 \approx 8\%$
- Repeat for the remaining values

Now you can create your graph as before, labeling the vertical axis from 1-25% (since none of the intervals are greater than 25% of the total) and the horizontal axis with the indicated intervals. The finished product should look rather like:



Example C

The table below contains data on the ages of male and female students at a school dance. Create a multiple relative frequency polygon to illustrate the data in the table.

TABLE 4.49:

Interval (Age in Years)	Male Frequency	Female Frequency
13-13.5	6	3
13.5-14	5	6
14-14.5	9	13
14.5-15	13	19
15-15.5	12	16
15.5-16	16	17
16-16.5	14	11

Solution: First we will need to find the sum of each frequency column and divide the frequency of each interval by the sum to get the relative percentage of each interval. If we add a row for the sum of each column, and two columns for relative percentages, our revised table will look like this:

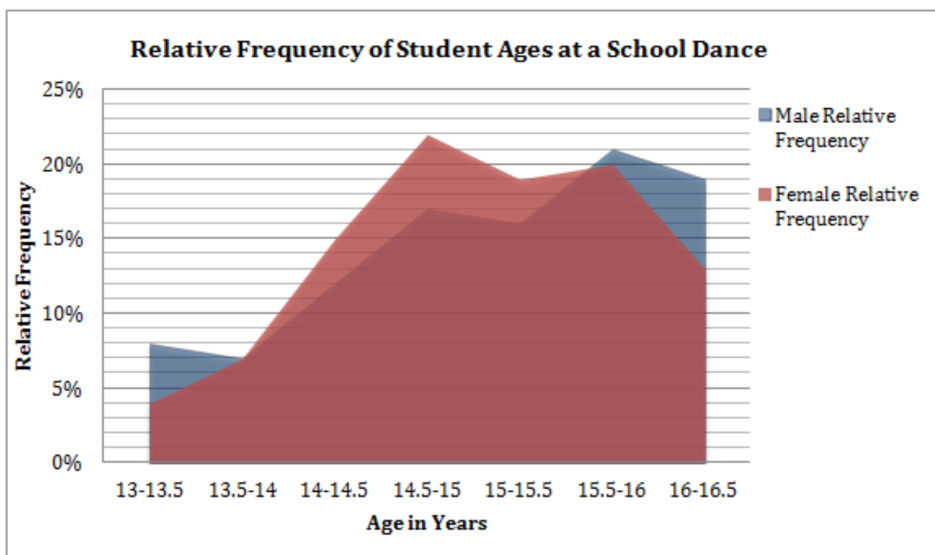
TABLE 4.50:

Interval (Age in Years)	Male Frequency	Male Relative Frequency	Female Frequency	Female Relative Frequency
13-13.5	6	8%	3	4%
13.5-14	5	7%	6	7%
14-14.5	9	12%	13	15%
14.5-15	13	17%	19	22%
15-15.5	12	16%	16	19%
15.5-16	16	21%	17	20%

TABLE 4.50: (continued)

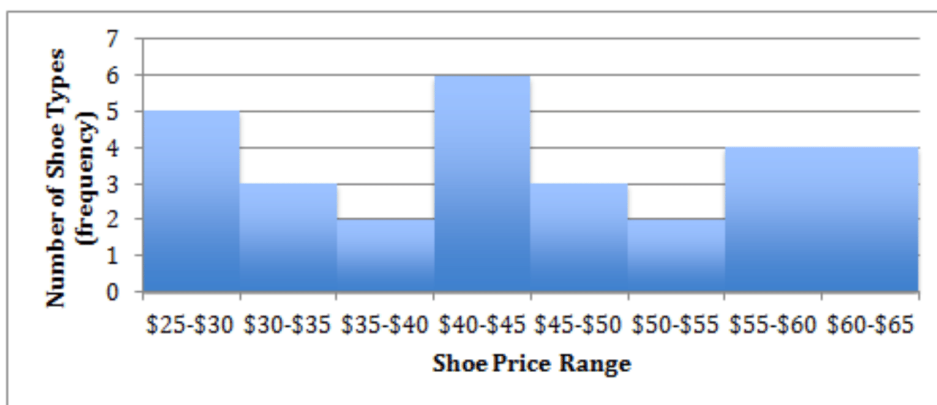
16-16.5	14	19%	11	13%
TOTAL	75	100%	85	100%

Then we need to graph both relative frequency polygons on the same graph so they can be easily compared.



Concept Problem Revisited

The histogram below is built from data given in the lesson on histograms. How would you convert the histogram into an absolute frequency polygon or a relative frequency polygon?



To convert the data into an absolute frequency polygon, plot a single point to represent the frequency of each interval at the midline of the interval, then connect the points with a line and color in the area between the line and the *x*-axis.

To convert into a relative frequency polygon, find the sum of frequencies of the entire sample, and divide the frequency of each interval by the sum, then multiply by 100 to get the relative percentage of each interval.

Vocabulary

Absolute frequency is an actual count of the number of data points that lie within a given interval.

Relative frequency is the ratio of the frequency of each interval to the frequency of the entire sample. May be expressed as a decimal between 0 and 1, or as a percentage.

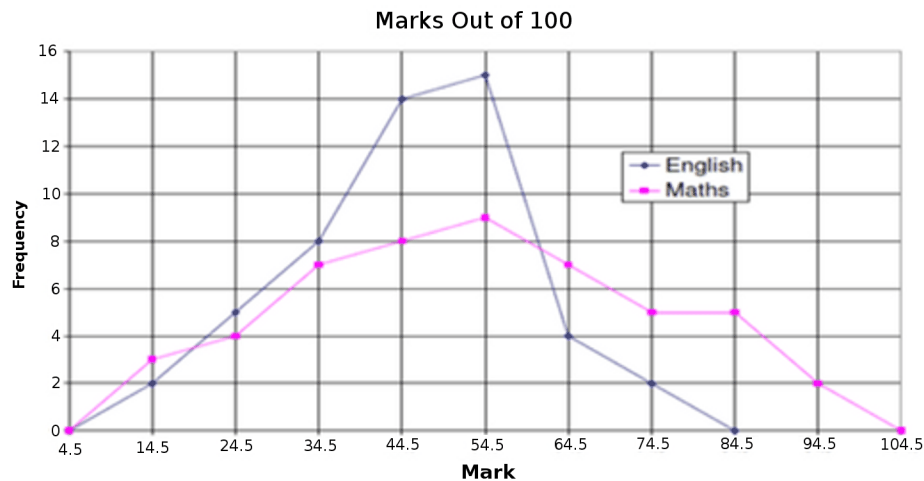
Guided Practice

The following table represents the average times it took for 25 men and 25 women to run a race.

TABLE 4.51:

Time in Minutes	30-40	40-50	50-60	60-70	70-80
Frequency Men	0	4	3	8	10
Frequency Women	2	6	8	6	3

1. Create a frequency polygon (remember to use the midpoints of the time on your graph) displaying the above data.
2. Is the data continuous or discrete?
3. Look at the graph below. If the “Mark” represents the midpoint of the represented data, list the classes that the marks were grouped into.



Use the following data for Q's 4 and 5:

126, 146, 161, 156, 134, 164, 157, 156, 154, 156, 154, 156, 138, 164, 154, 173, 159, 143, 152, 132, 164, 158, 170, 149, 178, 136, 177, 165, 128

4. Complete the Table

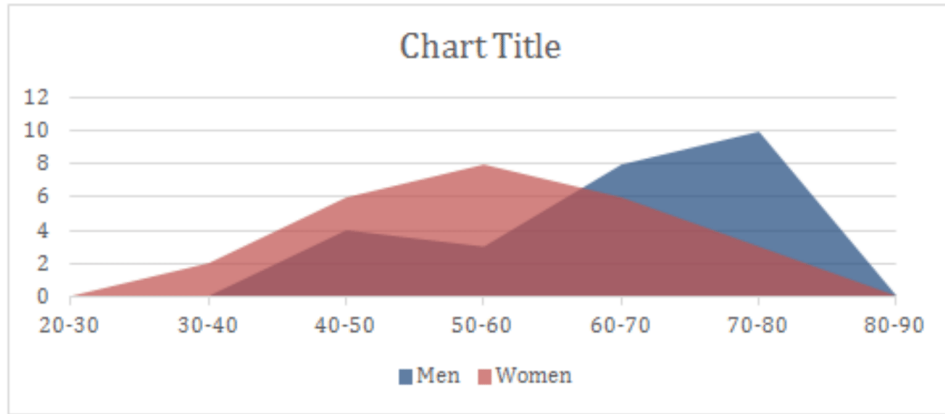
TABLE 4.52:

Class Interval	Midpoint	Frequency
120-129		

5. Create an Absolute Frequency Polygon for the data

Solutions:

1. Build a chart with intervals along the bottom (on the x-axis) matching those in the table with one category above and below, and with frequency values from 1- 10 up the side (y-axis):



2. The data is continuous, divided into intervals.

3. Remember that the data point is the *center* of the interval it represents, so the intervals are:

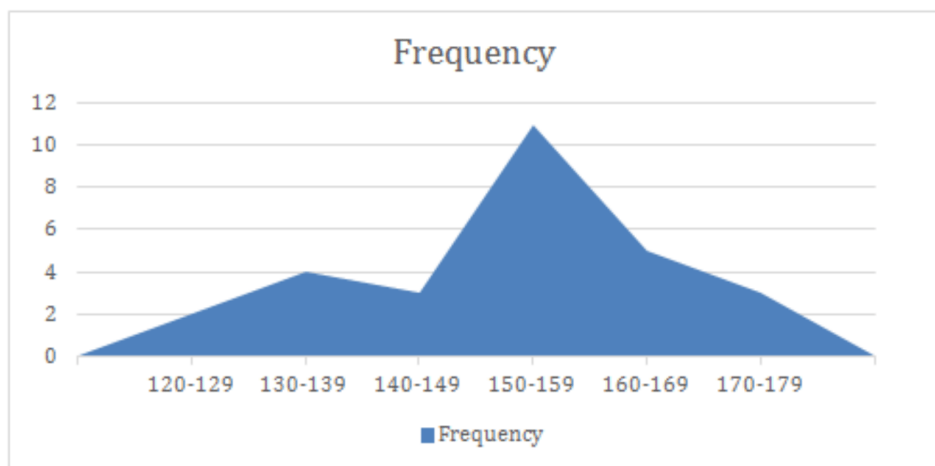
0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99, 100-109

4. The table shows an initial interval of 10 units, filling in the remainder of the intervals at the same range yields:

TABLE 4.53:

Class Interval	Midpoint	Frequency
120-129	124.5	II
130-139	134.5	III
140-149	144.5	III
150-159	154.5	IIIIIIIIII
160-169	164.5	IIII
170-179	174.5	III

5.



Practice Questions:

1. Why is a frequency polygon called a “polygon”?

2. What is the advantage of a frequency polygon?

A police radar unit measured the speed of 25 cars on a certain street during the week.

The resulting speeds were:

29, 23, 30, 30, 27, 24, 30, 25, 23, 28, 25, 24, 28, 30, 23, 30, 27, 25, 29, 24, 23, 26, 30, 28, 25

The policeman then completed the same procedure at the same time of day, only on a weekend.

The resulting speeds were:

30, 35, 25, 27, 30, 26, 25, 32, 29, 30, 28, 27, 27, 34, 32, 30, 27, 26, 31, 32, 35, 29, 33, 32, 28

3. Create a frequency table for both sets of data, assume an interval of 2 mph.

4. Create a combined absolute frequency polygon for the above data (both sets on the same chart).

5. What information can be derived from the two sets of data? What can you infer about weekday vs. weekend drivers?

6. If you had to take an educated guess, what would you say the speed limit on the street is?

Based on the following data:

11, 21, 30, 40, 57, 37, 51, 27, 35, 47, 33, 50, 14, 23, 63, 42, 66, 38, 65, 19, 43, 31, 32, 53, 56, 33, 25, 44, 52, 39

7. Complete the Table

TABLE 4.54:

Class Interval	Midpoint	Frequency
10-19		

8. Create an Absolute Frequency Polygon for the data.

Based on the following data:

45, 70, 58, 71, 63, 65, 88, 69, 71, 62, 75, 89, 75, 67, 70, 83, 73, 98, 77, 91, 52, 65, 87, 95, 43, 97, 53, 71, 66, 88

9. Complete the Table

TABLE 4.55:

Class Interval	Midpoint	Frequency
40-49		

10. Create an Absolute Frequency Polygon for the data.

Based on the following data:

108, 129, 133, 127, 117, 148, 128, 144, 130, 144, 120, 123, 103, 134, 152, 124, 134, 122, 147, 125, 117, 131, 148, 128, 134, 106, 159, 154, 121, 138

11. Complete the Table

TABLE 4.56:

Class Interval	Midpoint	Frequency
100-109		

12. Create an Absolute Frequency Polygon for the data.

13. Create a Relative Frequency Polygon for the data, and compare the appearance with that of the absolute frequency polygon.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.7.

4.8 Creating Box-and-Whisker Plots

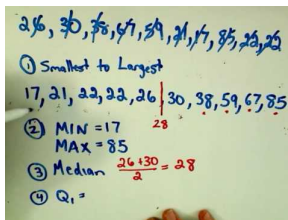
Objective

Here you will learn how box-and-whisker plots are created, and some common uses of them.

Concept

If you were asked to create a visual representation of the mean, upper and lower 25% (quartiles), and maximum and minimum (extremes) scores on the final test in your College Algebra class, how would you go about it? Would a box-and-whisker plot be appropriate? Why or why not? What would the plot look like if the mean was 82%, the lowest score was 59%, highest was 96%, and if a quarter of the class scored above 86% while another quarter scored below 70%?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63751>

<http://youtu.be/GMb6HaLXmjY> PatrickJMT - Box and Whisker Plot

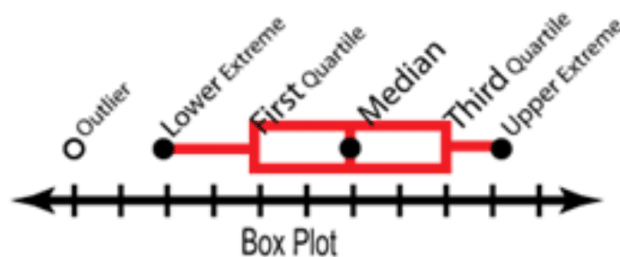
Guidance

Box-and-whisker plots (or **box plots**) are ideal for visually representing the *five number summary* of data.

First, *organize the data by increasing value*, then

The five number summary (or five statistical) is composed of:

- The minimum and maximum values - called the *extremes*
- The middle value - called the *median*
- The values halfway between each extreme and the median - called the *quartiles*.



It is important to recognize that a five number summary is more dependent on the *position* of each value in numerical order than on the value itself. A common confusion when gathering data for a box plot is to think that the plot is based on the *arithmetic mean* of the data rather than the median, don't fall into this trap! To create a box plot, which is based on the five number summary, you first need to organize your data in increasing numerical order, and then identify your five numbers *based on position in the ascending series*.

A sometimes tricky detail is the handling and identification of *outliers*. Once you have identified the median and quartiles of your data, you should review the values at the lower and upper limits to see if there are any that seem unusually extreme before considering them to be part of your 5 number summary. Specifically, data points that are more than 1.5 times the *inter-quartile range* (the range of values between the first and third quartiles - representing the middle half of your data), may be considered *mild outliers*. Any points more than 3 times the inter-quartile range may be considered *extreme outliers*. Outliers are commonly plotted as stars or asterisks (mild outliers) or open circles (extreme outliers), and are not a part of the actual box plot or the five number summary.

Once you have identified your five number summary, create a number line extending at least 10% past the upper and lower extremes of your data, and plot each of the five numbers above the appropriate locations on the line. Now create a rectangular box with sides on the 1st and 3rd quartiles. Draw a vertical line inside the box to represent the median, and draw horizontal lines from the sides of the box to the extremes. Finally, identify any mild outliers with asterisks/stars and extreme outliers with open circles.

Example A

Create a five number summary for the data below, and identify any outliers:

1, 5, 8, 2, 1, 7, 4, 4, 5, 6, 8, 2, 6, 5, 9

Solution: A five number summary includes the median, the upper and lower extremes, and the first and third quartiles. The first step to identifying them is to organize the data by ascending numerical value:

1, 1, 2, 2, 4, 4, 5, 5, 5, 6, 6, 7, 8, 8, 9

- Finding the median: Note that there are 15 values, an odd number, so the middle number in the series is the median. The value "5" has 7 values above and 7 below. **5 is the median.**
- The 1st quartile is the median of the lower half of the data. There are 7 values below the median, and the middle number of them is "2", with three values below and three above before the median. **The 1st quartile is 2.**
- The 3rd quartile is the median of the upper half of the data. There are 7 values above the median, and the middle value is "7", with three values above it and three below it before the median. **7 is the 3rd quartile.**
- Are there any outliers? The inter-quartile range is the difference between the 1st and 3rd quartiles: $7 - 2 = 5$. Recall that a value should be considered an outlier if it is unusually low in frequency and greater than 1.5 times the inter-quartile range from the median. In this case, that would mean any number more than 7.5 above the 3rd quartile, 7, or below the 1st quartile, 2. That would make any value less than -5.5 or greater than 14.5 be considered a mild outlier. There are no negative values and no values greater than 9, so **there are no outliers.**
- The minimum and maximum values are the least and greatest values, respectively. Since we have organized our data in ascending order, the minimum value is on the far left, "1", and the maximum value is on the far right, "9". **The minimum is 1 and the maximum is 9.**

Example B

Identify the five statistical summary and any outliers in the data below:

18, 16, 18, 17, 15, 2, 17, 20, 19, 18, 15, 16, 28, 18

Solution: Begin by ordering the data numerically:

2, 15, 15, 16, 16, 17, 17, 18, 18, 18, 18, 19, 20, 28

- Median: There are 14 values, an even number, so the median is the average (arithmetic mean) of the two

middle numbers, 17 and 18. **17.5 is the median.**

- **1st and 3rd quartiles:** The middle number in the lower 50% is 16, and the middle of the upper 50% is 18. **16 is the lower quartile and 18 is the upper quartile.**
- The inter-quartile range is $18 - 16 = 2$. Any value less than 13 or greater than 21 may be considered a mild outlier, and any value less than 10 or greater than 24 may be considered an extreme outlier. **2 and 28 are both extreme outliers.**
- The least value is 2 and the greatest value is 28. **2 is the minimum and 28 is the maximum.**

Example C

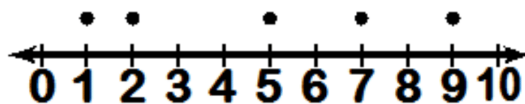
Create box plots representing the data from examples A and B.

Solution:

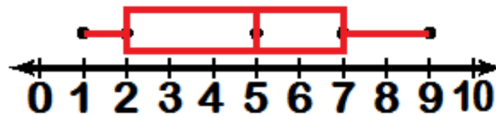
A. The data from example A was encapsulated in the five number summary:

Median: 5 1st quartile: 2 3rd quartile: 7 Minimum: 1 Maximum: 9

Draw a number line running from 0 to 10, and plot the five number summary above it:



Draw a rectangle including the first and third quartiles, and a vertical line for the median. Since there are no outliers, draw a “whisker” from each side of the box to the extremes:



B. The data from example B includes:

Median: 17.5

1st quartile: 16 3rd quartile: 18

Minimum: 2 Maximum: 28

Outliers (extreme): 2, 28

Draw a number line running from 0 - 30, and plot the five number summary:



Note that since “2” and “28” are both extreme outliers, the box-and-whiskers only extend to the greatest and least non-extreme values. This is sometimes called a **modified boxplot**.

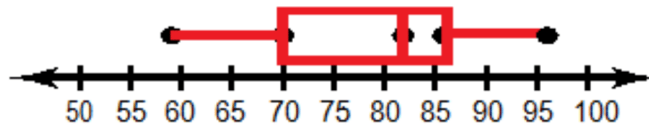
Concept Problem Revisited

If you were asked to create a visual representation of the mean, upper and lower 25% (quartiles), and maximum and minimum (extremes) scores on the final test in your College Algebra class, how would you go about it? Would a box-and-whisker plot be appropriate? Why or why not? What would the plot look like if the mean was 82%, the

lowest score was 59%, highest was 96%, and if a quarter of the class scored above 86% while another quarter scored below 70%?

This would be an excellent application for a box plot. In fact, this is just about the *best* use of one. You will find, if you haven't already, that the SAT and ACT college application exams report grades in just this manner. Colleges (and students themselves) inevitably wish to see how a particular score compares to others on the same test, and a box plot is ideal for that purpose.

If the data in the question were plotted as a box plot, it would appear like this:



Vocabulary

The **5 number summary** (or 5 statistic summary) is the collective term used to describe the minimum and maximum, middle, and 25% and 75% values in a data set.

The **extremes** are the minimum and maximum values in a set of data.

The **median** is the middle value in a set of data, when the data is organized in numerical order.

The values halfway between each extreme and the median are called the **quartiles**.

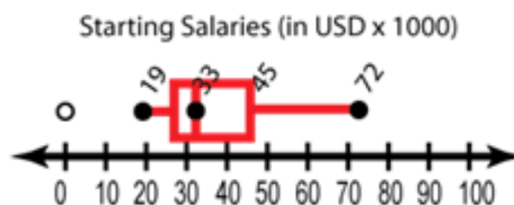
The **arithmetic mean** is a measure of central tendency calculated by finding the sum of the data, divided by the number of data entries. This value is referred to as the **average** in common language.

The **inter-quartile range** is the range of values between the first and third quartiles - representing the middle half of your data. In other words: the $IQR = Q3 - Q1$, and 50% of the data is in the box.

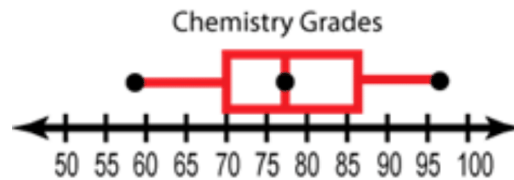
Outliers are values uncommonly distant from the mean. Mild outliers are determined as values at least 1.5 times the inter-quartile range above or below the 3rd or 1st quartiles. Extreme outliers are values greater than 3 times the inter-quartile range from the upper or lower quartiles.

Guided Practice

1. The box-and-whisker plot below shows the starting salaries for graduates of a small college. What is the range of the starting salaries?



2. Mr. Andrews made a box-and-whisker graph of the quiz grades in his chemistry class. What is the median quiz grade for the class?



3. Mr. Foreman grades on a curve in which the top 25% of the test scores earn A's, the middle 50% earn C's, and the bottom 25% earn F's. The box and whisker plot below shows the distribution of scores on the last test. What is the range of scores for people who earned C's?



Solutions:

1. The range is the difference between the maximum and minimum values, that is:

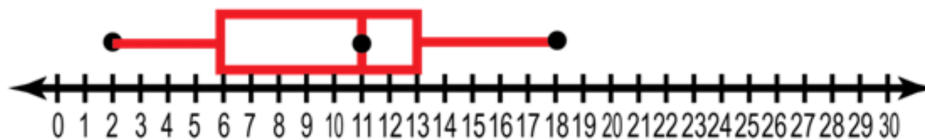
$$\$72,000 - \$19,000 = \$53,000$$

2. The median is denoted by a line in the center of a boxplot, in this case, that would be **77%**

3. If the middle 50% of Mr. Foreman's class earns C's, then all of the scores in the interquartile range, the area between Q1 and Q3, would be included. Since the "box" of a boxplot indicates the IQR, that would be **65% - 80%**.

Practice

Use the following boxplot to answer questions 1 - 5 below:



1. What is the median
2. What is the lower quartile
3. What is the upper quartile
4. What is the minimum value
5. What is the maximum value
6. What are the five values called?
7. What is the range of the data?
8. What percentage of the data is below the upper quartile?
9. What percentage of data is located between the lower quartile and the median?
10. What percentage of data is above the median?
11. What percentage of data is below the lower quartile?
12. Calculate the Range for the following data: 5, 21, 10, 9, 12, 12, 16, 16, 9, 6, 20, 8, 10, 26, 4, 26, and 14.
13. Calculate the First Quartile for the following data: 5, 21, 10, 9, 12, 14, 13, 16, 9, 6, 20, 8, 12, 24, 4, 26, and 14
14. State the five number summary of the following data set: 13, 14, 10, 4, 18, 17, 11, 10, 5, 7, 10 19, 13
15. Construct a box and whisker plot for the data set given in question 12

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.8.

4.9 Interpreting Box-and-Whisker Plots

Objective

Here you will learn to efficiently pull information from box plots.

Concept

If you were asked to evaluate a box plot to find the median, quartiles, extremes and outliers, would you know how? What does it mean if the 'box' in a box plot is unusually long or short? Does a long 'whisker' on one or both sides mean something important?

Review the lesson below and we'll return to answer these questions at the end.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62550>

<http://youtu.be/b2C9I8HuCe4> Khan Academy - Reading Box and Whisker Plots

Guidance

Box-and-whisker plots (or “box plots”) are commonly used to compare a single value or range of values for easier, more effective decision-making. Box and whisker plots are very effective and easy to read, and can summarize data from multiple sources and display the results in a single graph.

Use box and whisker plots when you have multiple data sets from independent sources that are related to each other in some way. Examples include comparing test scores between schools or classrooms, and exploring data from before and after a process change.

Remember that the line inside the box represents the middle value when the data points are arranged numerically. Because the median is only identified by location in a series, it can sometimes be very indicative of the trend or average of the data set as a whole, and sometimes is not useful for that purpose at all (see Example A).

Recall that skewed data appears as a longer “tail” in one direction on a histogram, it is similar on a box plot. If the box in a box plot is stretched in one direction or the other, then the data is skewed in that direction. Data skewed right indicates a closer concentration of values on the *left*, since the plot indicates values more “strung out” on the right side.

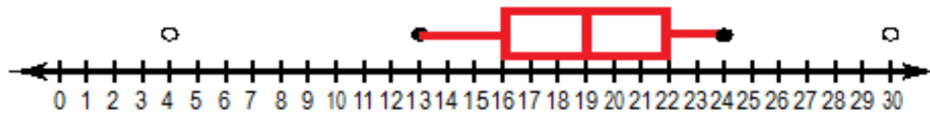
A longer box indicates a greater *interquartile range* since the sides of the box indicate the 1st and 3rd quartiles. A greater interquartile range is an indicator of data that may be somewhat unreliable. Since the interquartile range

represents the 50% of the data closest to the median, a greater range in this section of the plot suggests that the median may not be a great indicator of central tendency.

A plot with long whiskers represents a greater range for the overall sample than simply a longer box itself does. Data covering a greater range is naturally less reliable as an indicator of highly probable values, but given the option, longer whiskers are less of a concern than a long box. A broad range of possibilities but a strong likelihood of central values is more reliable to use for prediction than a moderate overall range with little concentration at the median.

Example A

Identify the 5 number summary and any outliers depicted in the box plot below:

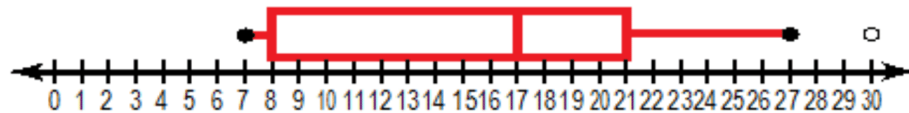


Solution: The 5 number summary is depicted by the vertical bars in the box and by the endpoints of the 'whiskers':

- Minimum: 13
- 1st Quartile: 16
- Median: 19
- 3rd Quartile: 22
- Maximum: 24
- Outliers (depicted by open circles disconnected from the box and whiskers): 4 and 30

Example B

What is indicated by the shape of the box plot below?

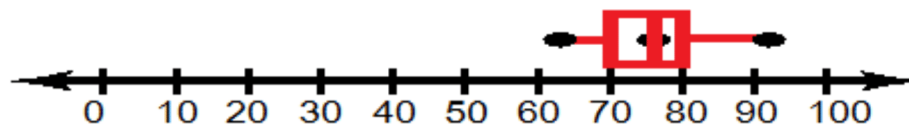


Solution: The box in the plot extends nearly to the lower extreme, indicating that the data less than the median is likely at least relatively consistent, since there is not a large jump between the lower 25% and the minimum. The longer whisker on the upper side suggests that there may be larger variance among the greater values, since there is a greater distance from the 3rd quartile to the upper extreme than from the median to the 3rd quartile.

Example C

A percentile box plot compares a particular value or range of values to an averaged reference point. The values on the scale represent the percentage of scores less than the plotted value. For instance, a score of 55% indicates that 55% of other values were less than the indicated score, and 45% were greater.

Maria recently completed a standardized test, and the box plot below describes her results. The median is her actual calculated percentile, and the rest of the 5 number summary suggests the range of percentiles that her score is expected to lie within once all scores are tabulated. Based on the information in the graph, would you expect Maria to be proud of her score? Why or why not?



Solution: Maria's score is expected to lie between the 62nd and 92nd percentile, with the most likely comparison being the 76th percentile. Since the 76th percentile indicates that her score was higher than that of 76% of all the students who took the test, and only 24% achieved a higher score than hers, yes, I would certainly say she has reason to be proud!

Concept Problem Revisited

If you were asked to evaluate a box plot to find the median, quartiles, extremes and outliers, would you know how? What does it mean if the 'box' in a box plot is unusually long or short? Does a long 'whisker' on one or both sides mean something important?

With the practice you have had now, these questions should be easy!

- Median: the center vertical line in the 'box'
- 1st and 3rd Quartiles: the leftmost and rightmost vertical lines of the 'box'
- Lower and Upper Extremes: the endpoints of the 'whiskers'

Vocabulary

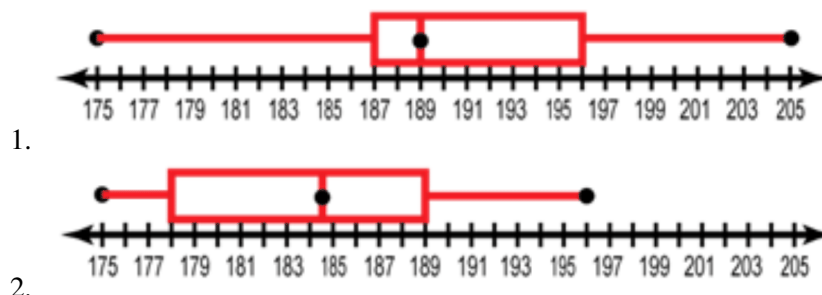
The *interquartile range* is calculated by subtracting the 1st quartile from the 3rd quartile and represents the middle 50% of the sample.

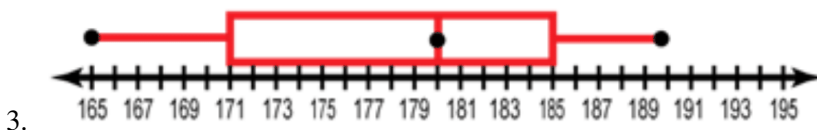
Guided Practice

1. Make a Box and Whisker plot from the following data sets.
 1. Initial weight (December) of 14 women in a weight loss study (pounds) 190, 175, 187, 199, 205, 187, 176, 180, 187, 191, 200, 193, 188, 196
 2. Weights of the same women one month later (January) 187, 174, 181, 189, 196, 178, 174, 176, 181, 186, 188, 191, 183, 191
 3. Weights of the same women in February. 181, 165, 176, 182, 190, 176, 171, 170, 171, 185, 187, 181, 179, 186
2. How do the data in a and c compare?
3. How did the median change?
4. How did the maximum weight change?
5. How did the minimum weight change?
6. How did the range change?
7. How would you judge the effectiveness of the weight loss method used in the study?

Solutions:

1. For all three sets, first organize the data by increasing numerical order and identify the five-number summary (FNS). Once you have the FNS, create the box plot for each just as in the examples above. The three plots should resemble the images below:

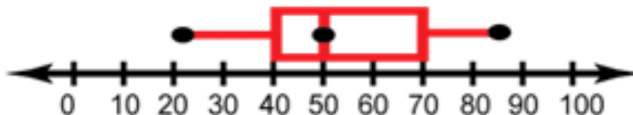




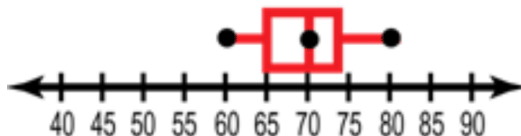
- 3.
2. If we compare the data between a and c, we can see that the overall weights of the women in the study did indeed go down. In fact, the minimum value at the start of the study was greater than the maximum two months later.
3. The median in December was 189, and in February it was 180.
4. The maximum in December was 205, and went down to 190 by February.
5. The minimum weight in December was 187, and it also went down, to 171 by February.
6. The range increased notably, from a mere 9 pounds in December, to more than 1.5 times that, 14, in January.
7. It would appear that the method was effective, at least in the short term. The increased range would indicate that it was somewhat more effective for some participants than others.

Practice

1. What is the five number summary of the following box and whisker plot?



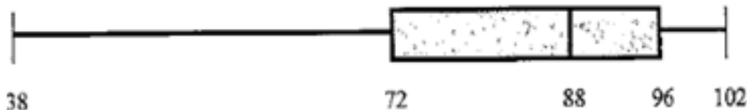
2. The box plot shows the heights in inches of boys on a High School Baseball Team. What is the 5 number summary of the plot?



3. Listed are the heights in inches of girls on a High School Ski Team. Make a plot of the girls' heights. 58, 59, 59, 60, 62, 65, 68, 69, 70, 70, 71

4. Comparing the heights between the two teams, which has the taller players on average? How do you know?

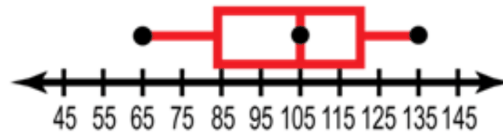
Use the box and whisker plot below to examine scores received on an English GED Test to answer questions 5-9



5. What was the high score on the test?
6. What percent of the class scored above a 72?
7. What was the median score on the test?
8. What percent of the class scored between 88 and 96?

9. Would you expect the mean to be above or below the median? Explain

Use the graph below that shows how much girls spent on average per month on clothes during August.



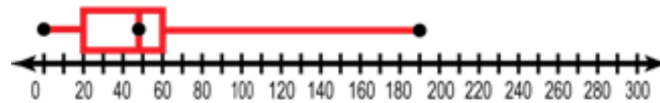
10. How many girls shop for clothes?

11. What percent of girls spent less than \$85.00 in August on clothes?

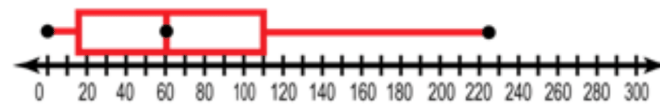
12. Would you expect the mean number of dollars spent to be higher or lower than the median? Explain

Use the graphs below to compare the amount of time a teenager spends in the bathroom getting ready for school and the amount of time they spend in the bathroom getting ready to go to a party.

TIME SPENT GETTING READY FOR SCHOOL:



TIME SPENT GETTING READY FOR A PARTY:



13. What percent of teenagers spend at least 15 minutes getting ready for a party?

14. What is the 3rd Quartile for the time spent getting ready for a party?

15. Is it more common for a teenager to spend more than 1 hour getting ready for school or between 1 and 2 hrs getting ready for a party? Explain

Answer True or False for questions 16-24.

16. _____ Some teenagers do not spend time getting ready for parties.
17. _____ The graph of time spent getting ready for a party contains more data than the getting ready for school graph.
18. _____ 25% of teenagers spend between 48 and 60 minutes getting ready for school.
19. _____ 15% of the teenagers did not go to parties that month
20. _____ In general teenagers spend more time getting ready for a party than getting ready for school.
21. _____ The Party data is more varied than the homework data
22. _____ The ratio of teenagers who spend more than 110 minutes getting ready for a party to those who spend less is about 2:1
23. _____ 225 Teenagers watch TV.
24. _____ Twice as many teenagers spend more than 1 hour on getting ready for school, than they do spending an hour getting ready for a party.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.9.

4.10 Creating Stem-and-Leaf Diagrams

Objective

Here you will learn how to create Stem-and-Leaf Diagrams.

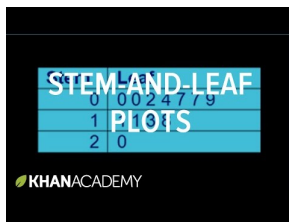
Concept

If you were asked to make a stem and leaf plot of the number of hours that Brian worked each month, given the data below, would you know or remember how? By the end of this lesson you will!

Make a stem plot of the number of hours that Brian worked each month:

Hours per month: 137, 149, 142, 121, 135, 133, 138, 137, 129, 139, 126, 139

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63757>

<http://youtu.be/LEFE1km5ROY> Khan Academy - Reading Box and Whisker Plots

Guidance

Stem-and-Leaf plots, also known as *stem plots*, are similar to sideways histograms, but are more descriptive because they are built by displaying the specific values of individual data points. Histograms are more visual and can be more striking and/or creative, but stem plots can be more accurate and can make further study of the data more efficient.

Creating a stem plot is not difficult, but, like most things, is easier and quicker with practice. To begin creating your stem plot, start by making two columns and labeling one *leaf* and the other *stem*. You can either create a full table, like this one:

TABLE 4.57:

Stem	Leaf
0	0
0	0
0	0

Or you can make one out of simple characters, like this:

Stem	Leaf
0	0
0	0
0	0

Just be sure that it is easy to see which column your values are placed in when the data is entered.

Depending on the size of your numbers, your stem column will either represent the single greatest place value of your data points, or perhaps the two or more greatest place values. The leaf column will represent the smallest place value(s) of the set. If you are listing multiple place values in either column, make a key such as the one below easily visible.

Key: 53,243

53 | 243

If you have two different related sets of data to compare (or *bivariate* data), you can use a *back-to-back* version of the plot. The back-to-back version places the stem in the center of three columns, and the different leaves on each side, as below. Note that ordering the data on the plot requires that you arrange the values so that they increase *from the center out*, rather than just from left to right.

TABLE 4.58:

Leaf	Stem	Leaf
3 2 1 0	0	0 1 2 3
3 2 1 0	1	0 1 2 3
3 2 1 0	2	0 1 2 3

Whatever the purpose of your stem plot, always be certain to *label the plot clearly!*

Example A

Make a stem plot of the money that Jelani spent this month:

Jelani's purchases: \$59, \$54, \$49, \$68, \$62, \$74, \$57, \$30, \$40, \$80, \$86

Solution: Start by creating your table with the heading, stem and leaf columns labeled:

TABLE 4.59: Jelani's Purchases for the Month

Stem (tens)	Leaf (ones)

Then begin to insert the data, creating a single row for each tens place:

TABLE 4.60: Jelani's Purchases for the Month

Stem (tens)	Leaf (ones)
3	
4	
5	
6	
7	
8	

Finally, fill in the leaves with the ones place values indicated by your data, listing them in increasing numerical order:

TABLE 4.61: Jelani's Purchases for the Month

Stem (tens)	Leaf (ones)
3	0
4	0 9
5	4 7 9
6	2 8
7	4
8	0 6

Example B

Beryl loves basketball, and she tracked the number of points her favorite pro player scored in the first 25 games of his last season. Create a stem plot of the data.

Points per game: 25, 41, 47, 42, 20, 37, 38, 40, 33, 21, 27, 46, 42, 28, 23, 35, 29, 42, 31, 45, 25, 27, 33, 37, 42

Solution: Notice that the data points are all between 20 and 50, so we will need only 3 stems in our plot, each of which will have quite a few leaves. Because there are so many points, it might be worth the time to first organize the data in increasing numerical order:

20, 21, 23, 25, 25, 27, 27, 28, 29, 31, 33, 33, 35, 37, 37, 38, 40, 41, 42, 42, 42, 42, 45, 45, 47

Now it is a simple matter to place all of the correct leaves in the proper column:

TABLE 4.62: Points Scored per Game

Stem (tens)	Leaf (ones)
2	0 1 3 5 5 7 7 8 9
3	1 3 3 5 7 7 8
4	0 1 2 2 2 5 5 7

Example C

Create a stem plot to compare the hours that Bruna and Giselle each worked per week during the same three months of last year:

Bruna hours worked: 40, 31, 40, 31, 36, 36, 20, 21, 40, 33, 31, 50, 47, 34

Giselle hours worked: 28, 23, 35, 29, 42, 31, 45, 25, 27, 34, 31, 40, 32, 41

Solution: First sort the data in ascending numerical order:

Bruna hours worked: 20, 21, 31, 31, 31, 33, 34, 36, 36, 40, 40, 40, 47, 50

Giselle hours worked: 23, 25, 27, 28, 29, 31, 31, 32, 34, 35, 40, 41, 42, 45

Next create a labeled three column table with the tens column in the center and the ones column for each person on the sides, note that all of the values are between 20 and 50, so we will need only 4 rows. List the values in increasing numerical order from the center out!

TABLE 4.63: Hours Worked per Week

Bruna		Giselle
1 0	2	3 5 7 8 9

TABLE 4.63: (continued)

6 6 4 3 1 1 1	3	1 1 2 4 5
7 0 0 0	4	0 1 2 5
0	5	

Concept Problem Revisited

Make a stem plot of the number of hours that Brian worked each month:

Hours per month: 137, 149, 142, 121, 135, 133, 138, 137, 129, 139, 126, 139

First arrange the values in ascending numerical order:

121, 126, 129, 133, 135, 137, 137, 138, 139, 139, 142, 149

Now create your table, listing the first *two* place values in the stem column (since these are all three-digit numbers):

Hours Brian Worked Each Month:

12		1	6	9			
13		3	5	7	7	8	9
14		2	9				

Key: 147 = 14 | 7

Vocabulary

Stem Plots are also known as stem-and-leaf plots because they are a method of arranging data by separating the greatest place value(s) (the stems) from the least place value(s) (the leaves).

A **Stem** in a stem plot is a values or column of values that represent the greatest place value(s) in a set of data.

A **Leaf** in a stem-and-leaf plot is a value or column of values that represent the smallest place values in a set of data.

A **Back-to-Back stem plot** is a modified stem-and-leaf plot with the stem in the center and the leaves on the sides, it is used to compare two different related sets of data (bivariate data).

Guided Practice

1. If you asked to make a stem and leaf plot from the following test scores, what would be the first step? 73, 99, 92, 91, 85, 82, 76, 92, 80, 70, 9, 59, 91, 95, 86, 77, 78, 71, 64, 82, 98, 65, 56
2. What is the range of the data? Once the range has been identified, what is going to be the range of the stem and why?
3. Create a Stem-and-Leaf plot for this data.
4. What type of graph would this be if it were turned vertically?
5. In which interval did the least number of students score? In which did the most students score?

Solutions:

1. The first step would be to put the scores in numerical order so that they looked like this: 56, 59, 64, 65, 69, 70, 71, 73, 76, 77, 78, 80, 82, 82, 85, 86, 91, 91, 92, 92, 95, 98, 99
2. The range of the data is from 56 to 99. Therefore the stems, identified by the 10's place value, will range from 5 to 9
3. Your plot should look like this:

TABLE 4.64:

Stem	Leaf
5	6 9
6	4 5 9
7	0 1 3 6 7 8
8	0 2 2 5 6
9	1 1 2 2 5 9

4. A Histogram

5. The first stem: 50 to 59, contains the least number of values, 2 of them. The 70 - 79, and 90 - 99 stems each had 6 scores, for the high.

Practice

Make a stem-and-leaf plot from the following data. The data represents the number of times in the last year that each polled student attended a school function.

Number of functions attended: 17, 16, 16, 16, 16, 15, 15, 19, 18, 17, 15, 13, 14, 15, 23, 12, 36, 48, 45, 28, 23, 24, 18, 17, 16, 15, 16, 13, 17, 19, 14, 15, 17, 12

1. Based on the data, how many total students were polled?
2. What interval represents the least number of functions attended by students?
3. What interval represents the greatest number of functions attended by students?

Create stem-and-leaf plots for the following sets of data.

4. Fifteen people were asked how often they bought flavored coffee on their way to work during the last 10 working days. The number of times each person bought flavored coffee was as follows: 5, 7, 9, 9, 3, 5, 1, 0, 0, 4, 3, 7, 2, 9, 8
5. A middle school teacher asked her students how many books they had read over the last year. Their reported answers were as follows: 13, 22, 20, 5, 11, 6, 14, 26, 20, and 13.
6. Joe is training for a swim meet, and keeps track of the number of laps he swims each day. Create a stem-and-leaf plot to organize the data: 21, 22, 23, 18, 28, 27, 24, 24, 28, 29, 24, 20, 27, 28, 21, 40, 21, 9, 27, 23, 28, 27, 25, 27, 19, 31, 30, 24, 30, 28.
7. Redraw the stem and leaf plot, using the above data using 5-unit intervals.
8. Comment on how the data looks different, and explain which chart you feel is a better representation, and why.

The weights to the nearest kilogram of 30 kids were recorded. 59.1, 61.6, 62.2, 61.5, 60.8, 59.9, 60.4, 59.1, 61.0, 60.8, 61.5, 56.2, 61.8, 65.6, 60.3, 58.8, 59.0, 61.3, 62.0, 61.3, 58.5, 60.9, 60.3, 62.6, 60.1, 59.2, 61.8, 61.8, 58.5, 62.1

9. What numbers will be represented by the stem? What numbers will be represented by the leaf?
10. What is the data range? Where will the stems start and finish?
11. Create a stem-and-leaf plot for the data.
12. Which weights would be considered outliers?

You are the football coach & you have two place-kickers. Manny has *made* field-goals of 23 yards, 34 yards, 30 yards, 23 yards, and 28 yards, but has *missed* field goals of 16, 16 and 23 yards. The other, Jose, has *made* field goals of 11 yards, 16 yards, 21 yards, 20 yards, but has *missed* field goals of 26 and 28 yards. There is only 15 seconds left on the play clock.

13. Create Stem-and-Leaf plots to represent the Manny's data.
14. Create Stem-and-Leaf plots to represent the Jose's data.
15. Your team can win the state championship if your kicker makes an 18 yard field-goal. Which kicker would you choose & why?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.10.

uniform with a stem only representing the hundreds place may show significantly more detail if the stem were set to represent the hundreds and tens places both, since this would create many more leaves.

Skewed: As with the horizontal skewing of a histogram, stem plots with a obvious skew toward one end or the other tend to indicate an increased number of outliers either lesser than the mode (skewed down - correlating to a left-skew in a histogram) or greater than the mode (skewed up - correlating to a right-skewed histogram). Plots with an upward skew will have a mode that is smaller than either the mean or the median, and a mean that is greater than either the median or the mode. Downward-skewed plots will have a mean lesser than median or mode and a mode greater than either mean or median.

Example A

Using the stem plot below, make a list of the number of hours that Isolde worked each week:

2	4 4 9
3	2 2 4 7 8
4	0 1 1 2 2 3 4
5	0 1 6 8

Solution: Use the values in the left column to represent the tens place, and the values in the right column to represent the ones, and write each number out as a list:

Isolde's work record (hours per week):

24, 24, 29, 32, 32, 34, 37, 38, 40, 41, 41, 42, 42, 43, 44, 50, 51, 56, 58

Example B

Describe the type and meaning of the shape of the stem plot, identify the extremes, and state the mode(s) of the data. What does the data describe, as stated by the stem plot?

TABLE 4.65: Hathor's Expenses for Jan 2011

1	2 4 5 5 8
2	3 3 6 7 7 9
3	0 0 3 3 3 7 8 8
4	1 4 4 7
5	0 2 3 3

Solution: The data is bell-shaped, indicating that the majority of the data is clustered around the median. The median is 33, since there are 13 values above and below that point, and the extremes are 12 and 53 since those are the greatest and least values in the sample. The mode is 33, since there are 3 points with that value and no other value has more than 2 entries.

The plot describes the money spent in the month of January 2011, by someone named Hathor. Hathor bought mainly items less than \$40, though there were a few purchases above the \$40 mark, none were greater than \$53.

Example C

Compare the weekly incomes of Sabah and Anaelle according to the stem plot below.

- What things stand out based on the shapes of the two plots, if any?
- What is the place value of the stem?
- What does the blank on Anaelle's side next to the "1" indicate?
- Who earned the most money overall?
- Who had the most consistent paychecks?

- f. Who had the single greatest and who the single least paycheck?
g. Did they both record the same number of paychecks?

TABLE 4.66: Week Income Comparison (USD)

Sabah		Anaelle
99 97 88	0	89 95
79 23	1	
93 68 42	2	09 17 22 58 77
97 58 12 00	3	01 29 33 51 67 82

Key: 147 = 1 | 47

Solution: Let's take each of the questions individually:

- Both plots are slightly up-skewed, indicating that the lower values are likely less indicative of the most common incomes. Sabah's plot is otherwise pretty uniform, whereas Anaelle's has a notably greater frequency of values in the \$200 to \$400 range.
- The stem represents the hundreds place.
- Anaelle did not record any paychecks between \$100 and \$200.
- Without even totaling the columns, we can see that Anaelle earned more, as her column is heavily weighted towards the greater values, and she has also recorded an additional check.
- Anaelle, as indicated by the greater number of checks clustered in the \$200 - \$400 range and few outliers.
- Sabah had the smallest paycheck at \$88, and also the greatest at \$397.
- No, Anaelle recorded 13, Sabah only 12.

Concept Problem Revisited

If you were given a stem plot depicting the money spent each week for a few months, how would you go about finding the various important parameters of the sample?

First look over the general shape of the plot to get an idea of the trend(s) of the data. Then identify the maximum and minimum extremes and the median and mode. Ask yourself if there are any particularly extreme outliers, or if the data is spread evenly, if there is there an obvious mode, if the median and the mode are close together, and if there are any other standout values. You can quickly get a good picture of the data by understanding how a stem plot works and applying your knowledge, but the data is useless if you don't know how to read the data or what questions to ask.

Vocabulary

Trends in data sets or samples are indicators found by reviewing the data from a general or overall standpoint. Values such as the 5 number summary can help to identify trends such as increasing or decreasing values over time or fluctuations in values caused by a particular factor.

Bins are the intervals or categories in histograms, the name comes from the idea that you review the data and toss each point into a corresponding 'bin' based on which range or category it fits.

The **range** of a set of data is the difference in value between the least (the minimum extreme) and greatest (the maximum extreme) values in the set. Investigating the range of a set can help to determine how widely spaced the data points are.

Guided Practice

The stem and leaf plot below shows the grade point averages of 18 students.

0	8
1	2 4 4 6 7
2	0 4 5 5 6 7 9
3	2 3 5 8
4	0

1. What is the range of the data in the stem and leaf plot?
2. How many students have a grade of 2 or more?
3. What is the mode of the grades?
4. What is the median of the grades?

Solutions:

1. Range = maximum value – minimum value = $4.0 - 0.8 = 3.2$
2. $7 + 4 + 1 = 12$ students
3. Two modes: 1.4 and 2.5
4. There are 18 data values and they are already ordered in the stem and leaf diagram.

$$\text{Median} = \frac{(\text{the 9th value} + \text{the 10th value})}{2} = \frac{(2.5 + 2.5)}{2} = 2.5.$$

Practice**True or False**

1. Some sets of data do not have a mean.
2. There is always a mode for each set of data
3. The median is always a number in the data set.

Based on the following stem-and-leaf plot of student heights in inches, how tall is tallest person in class?

4.

TABLE 4.67:

4	5 7 8
5	0 2 6 8 9
6	0 2 4 7 8

Taylor and some friends went crawdad hunting. The number of crawdads captured have been recorded below.

TABLE 4.68:

0	2 7 9
1	0 4 7
2	1 4 8
3	3 6

5. How many people went with Taylor?
6. What were the total number of crawdads captured?

The following stem- and-leaf plot shows a range of 10 numbers.

TABLE 4.69:

87	5
88	1 4 5
89	1 7 7
90	6 9

7. What is the highest number?
8. What is the median value of the data set?

The back to back stem and leaf plot below shows the exam grades (out of 100) of two different class periods. The digit in the stem represents the tens and the digit in the leaf represents the ones.

TABLE 4.70:

Class Period 1		Class Period 2
5 0	4	1 3 4 5
3 3 2 1	5	3 4 5 5 7 9
8 6 5 4 3 1	6	1 2 3 5 6 6 7 9
9 7 6 3 1 0 0	7	0 3 4 6 8 9
7 4 3 2 1	8	1 6
5 3 2 0	9	0 1

9. How many students scored higher than 60 in section 1?
10. How many students scored higher than 60 in section 2?
11. What are the minimum and maximum scores in section 1?
12. What are the minimum and maximum scores in section 2?
13. Without counting, which section has more students scoring 80 or more?
14. Without counting, which section has more students scoring 50 or less?
15. Describe the shape of the data set. Is it symmetric? Are there extreme values in the low or high numbers?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.11.

4.12 Creating Scatter Plots and Line Graphs

Objective

Here you will learn how to take raw or organized *bivariate* data and present it in a visual format with a scatter plot or line plot.

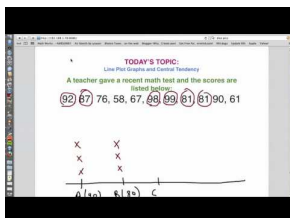
Concept

Scott's teacher was reviewing the research that Scott had conducted regarding the best car stereo systems to buy for college students on a limited budget. Scott had two long columns of numbers indicating the comparison between sound quality (which Scott had summarized with a 10-point scale for each stereo), and cost rounded to the nearest dollar before tax.

The teacher commended Scott on the detailed research, but pointed out that the list of numbers was kind of hard to make sense out of. He suggested that Scott plot the values on a scatter plot or line graph to see if there was a 'sweet spot' indicating the best compromise between quality and cost.

How should Scott decide which type of plot is best for his purpose? How would he go about taking the data from columnar form and converting it into the data visualization he decides to use?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63762>

<http://youtu.be/TZGOIeKp0fc> MathMasterD - Line Plot Graph Tutorial Video

Guidance

Line plots, followed closely by scatter plots, are by far the most common method of displaying *bivariate data*. By assigning one variable to each axis and plotting points by both horizontal and vertical location simultaneously, you can quickly and easily show the degree to which one set of data is influenced (or not influenced) by another.

There are two general types of bivariate data sets that are graphed on a line or scatter plot: observed (or experimental) data and calculated (or predicted) data.

- **Calculated Data:** To create a line or scatter plot of calculated data, you must first identify your two variables as either *dependent* or *independent*. A dependent (or input) variable may also be referred to as the *explanatory* variable, and has values that are assigned to it. An independent (or output) variable may also be called the *response* variable, and has values that are the result of computations performed on the input variable. By

convention, the independent variable is plotted on the horizontal, and the dependent variable is plotted on the vertical.

- **Observed Data:** The most common reason to graph two sets of data on the same graph is to evaluate the level of *statistical correlation*. By plotting the two sets of data on separate axes of the same graph, we can see a visual representation of possible related changes in values between the two sets. As with calculated data, you should plot the values of the variable that you expect is the explanatory variable on the horizontal and the expected response variable on the vertical.

When graphing observed data, you do not always know which value is the input and which the output, or even if the two values are indeed dependent at all! In later lessons, we will return to this concept to learn a number of methods to evaluate data and determine the degree of correlation between multiple value sets. For now, place the variable you think is most reasonably the input on the horizontal.

The first and most important step is to organize your data so that it is easy to see how a given input value relates to a given output value. By convention this is done with a 'T' chart or a two-column graph, with the input value on the left and the output value on the right, or vertically with the input on the top and output on the bottom.

Once you have the table constructed, start with the first pair of values and move across your horizontal axis to the first input value and up the vertical axis to the associated output value. Continue the process until all of your points have been graphed.

Once all of your points have been plotted, if you are creating a scatter plot, you're done! If you are creating a line plot, start at your minimum input value and connect the points as you move to the right on the input axis.

Example A

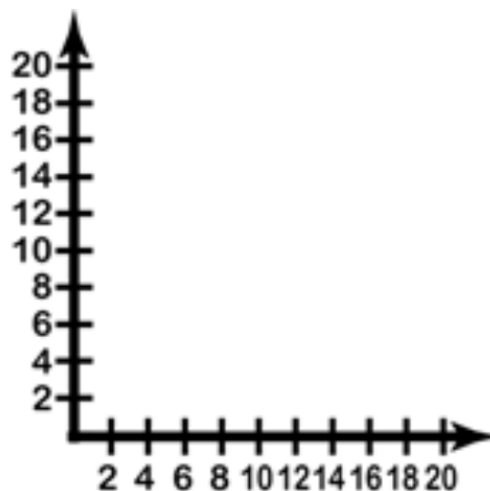
Construct a scatter plot from the given values.

TABLE 4.71:

Input	1	3	5	7	9	11	13	15	17
Output	2	4	6	8	10	12	14	16	18

Solution: The data here is already organized into associated input and output values, so you simply need to create a graph with a horizontal and vertical axis on which to plot the points.

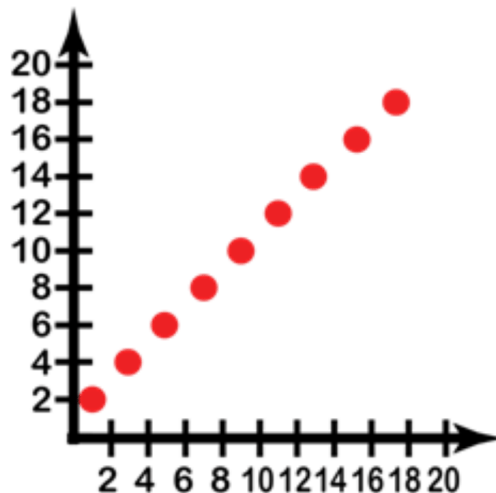
Notice that I have only created the positive values here, since the table of values was all positive.



Now we just plot the points from the table, starting with the first vertical pair: Input = 1, Output = 2. Incidentally, when describing a single point of bivariate data, the conventional method of writing it is in the form (input, output)

or (x,y) . So our first point would be $(1, 2)$, the second would be $(3, 4)$ and so on.

Now we fill in the values on the graph, starting with $(1, 2)$. Beginning at the lower-left corner, which represents $(0, 0)$, move 1 point to the right and 2 points up. The second point is 3 points to the right and 4 points up. Continue until all 10 points are graphed. Since the question asks specifically for a scatter plot, once the individual points are plotted, we are done.



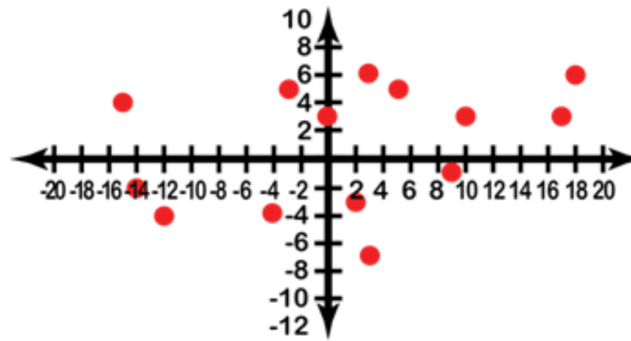
Example B

Romane loves jellybeans, and she eats an average of 20 each day. Worried about her weight, she decides to see if there is an obvious correlation between the number of jellybeans she eats and her weight. If she records the data below, which variable would be the input and which the output? Create a line plot from the data.

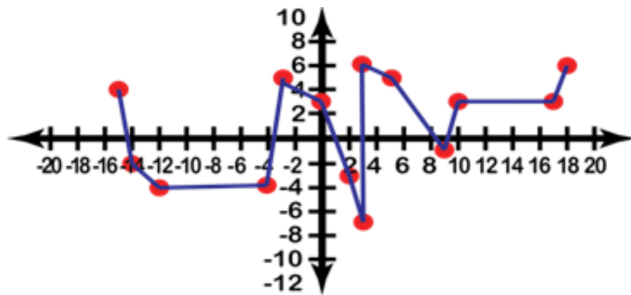
TABLE 4.72:

Increase or Decrease # of beans	Increase or Decrease in Weight
+2	-3
-3	+5
-7	+3
+5	+5
-12	-4
+18	+6
-14	-2
-15	+4
+17	+3
+0	+3
-4	-4
+3	+6
+9	-1
+10	+3

Solution: Since Romane can control the number of jellybeans she eats, that would be the input variable, and the increase or decrease in weight would be the output. If we create an (x,y) graph and plot the points, the result looks like this (note that this time the graph shows negative and positive values!):



Finally we connect the points from left to right, since the question specified a line graph:



Example C

Does more sleep consistently improve math grades?

Organize the data below by creating a 'T' table for the x and y values, then graph the data as either a scatter plot or line graph, whichever is most appropriate.

TABLE 4.73:

Math Homework Score (out of 20 points)		Hours of Sleep (the night before)	
Day 1	11	Day 1	7
Day 2	19	Day 2	8
Day 3	9.5	Day 3	6.5
Day 4	11	Day 4	7
Day 5	15	Day 5	5
Day 6	6	Day 6	3
Day 7	11.5	Day 7	8.5
Day 8	18.5	Day 8	7.5
Day 9	14	Day 9	6
Day 10	18	Day 10	8
Day 11	15.5	Day 11	5.5
Day 12	15.5	Day 12	4.5
Day 13	12	Day 13	9
Day 14	15.5	Day 14	5.5

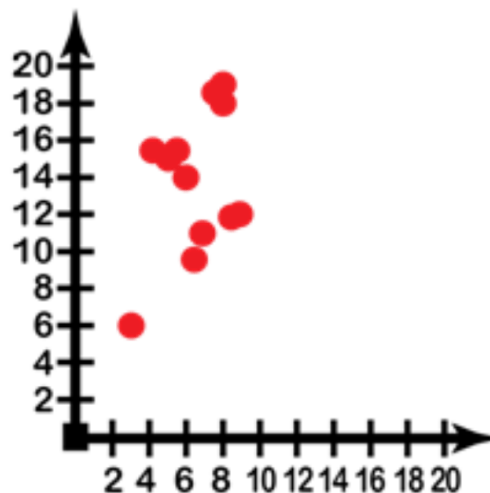
Solution: The data as given is organized by day in each table. In order to answer the question: “Does more sleep consistently improve math grades?” We need to correlate the data from each night’s sleep with the next day’s grades.

In this case, *which* day it is does not matter as much as the hours of sleep the night before, so we can pull the 'Day' column out of each table (being careful not to change the order of the values!) and make a new table with only the correlating scores and hours of sleep. This gives us:

TABLE 4.74:

Hours of Sleep	Math Score
7	11
8	19
6.5	9.5
7	11
5	15
3	6
8.5	11.5
7.5	18.5
6	14
8	18
5.5	15.5
4.5	15.5
9	12
5.5	15.5

Since the question is asking about the correlation between sleep the night before and grade the next day, the sleep becomes the *input variable* (or *independent variable*) and score becomes the *output variable* (*dependent variable*). Plotting the points on an (x,y) graph yields:



Given the significant scattering of points as we move left to right, it is appropriate to maintain the scatter plot layout.

In later lessons we will discuss *linear regression*, the process of identifying a *line of best fit*. A *line of fit* is a line drawn through a scatter plot that indicates a trend that the data follows, and the line of best fit is the mathematically derived most accurate indicator of that trend.

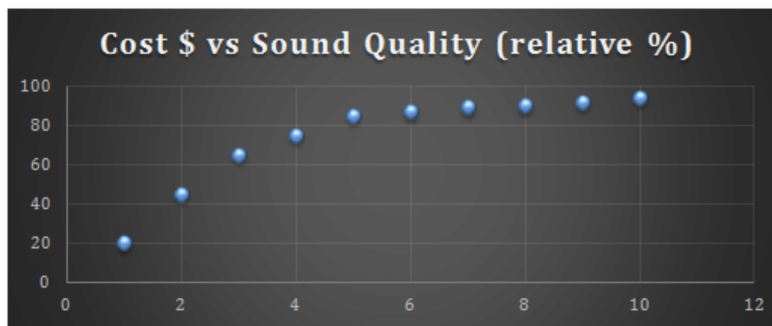
Concept Problem Revisited

Scott's teacher was reviewing the research that Scott had conducted regarding the best car stereo systems to buy for college students on a limited budget. Scott had two long columns of numbers indicating the comparison between

sound quality (which Scott had summarized with a 10-point scale for each stereo), and cost rounded to the nearest dollar before tax.

The teacher commended Scott on the detailed research, but pointed out that the list of numbers was kind of hard to make sense out of. He suggested that Scott plot the values on a scatter plot or line graph to see if there was a 'sweet spot' indicating the best compromise between quality and cost. How should Scott decide which type of plot is best for his purpose? How would he go about taking the data from columnar form and converting it into the data visualization he decides to use?

Scott should view cost as the independent variable (the input) and sound quality as the dependent variable (the output). By creating an (x,y) graph and plotting each of the corresponding cost/quality values, he will generate a much clearer comparison for the audience of his report. If he notes a particularly clear correlation (all or most of the points in a line) between increasing cost and improved quality, he may wish to indicate a line of fit to further illustrate the trade-off. A graph of his data might look something like this:



Vocabulary

Bivariate data is data composed of two changing sets of values. Distance travelled over differing time intervals, or weight compared to height, or income related to age are all examples of bivariate data.

Dependent variables, also called **output variables** or **response variables** and commonly represented by 'y', have values that depend on the value of another variable.

Independent variables, also called **input variables** or **explanatory variables** and commonly represented by 'x', have values that are not determined by another variable.

A **line of fit** is a straight or continuously curved line representing the trend of changes in the comparison of two data sets (or one set of bivariate data). **The line of best fit** is the mathematically calculated most accurate depiction of the trend(s). Note that a line of fit need not be straight, or even continuous.

Linear regression is the process of identifying a line of fit or the line of best fit for a given function.

Guided Practice

1. Construct a scatter plot to represent the data from the chart below indicating the number of birds killed by planes each year.

TABLE 4.75:

YEAR	PLANES REGISTERED	BIRDS KILLED
1978	6	13
1979	4	12
1980	7	14

TABLE 4.75: (continued)

1981	3	11
1982	7	14
1983	6	13
1984	3	12
1985	4	11
1986	1	9
1987	4	12

2. Construct a line graph to illustrate the data.

TABLE 4.76:

Height Change	4	2	8	13	16
Weight Change	3	3	3	4	3

3. Mike decided to see if the teenage drivers in his city were truly more likely to get into accidents, and he collected the data below. Graph the data appropriately for his study.

TABLE 4.77:

Year	Teen Drivers Registered	Auto Accidents
1980	90	80
1981	70	90
1982	60	90
1983	20	100
1984	30	90
1985	100	90
1986	70	100
1987	60	100
1988	70	100
1989	90	90

4. How do you determine which values to graph on the vertical axis and which on the horizontal?

5. Given the data below, which variable represents the explanatory variable? What is the related term for the other variable? How do you know which is which?

TABLE 4.78:

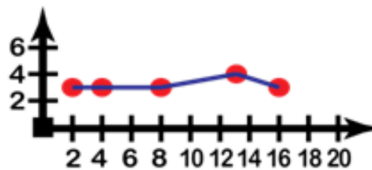
Sneezes	8	6	5	2	1	4	7	12
Tissues	18	13	7	5	1	9	19	43

Solutions:

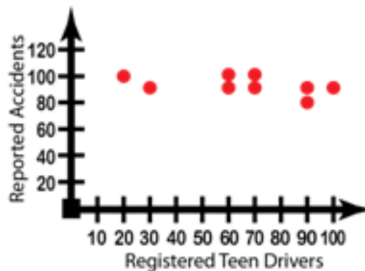
1. The number of planes registered is the input, and the number of birds killed is the output.



2. Change in height is the input, and change in weight the output. Connect the points to create a line graph instead of a scatter plot.



3. The number of teens registering to drive is the input, and the number of reported accidents the output. Note that the very little vertical change despite a significant horizontal change would indicate virtually no correlation between the two values.



4. The input value or the cause is the independent variable, and the output or the effect is the dependent variable. The independent variable goes on the horizontal and the dependent on the vertical.

5. Sneezes are the input variable, graphed on the horizontal 'X' axis, since sneezing is the cause related to the effect of using a tissue.

Practice

1. Create a scatter plot of the data shown below. Describe the relationship that exists within the data.

TABLE 4.79: Cost associated with raising a child

Child's Age	3	6	9	12	15
Annual Cost	11,800	12,800	13,700	16,000	17,800

2. Create a scatter plot from the data in the table below.

TABLE 4.80:

X	-3	-3	-2	-1	0	0	0	1	1	1	2	3
Y	3	4	3	2	1	2	0	0	-1	-1	-3	-4

3. Draw a reasonable line of fit.

4. What is the equation of the line of fit?

5. The data below shows the number of hours spent studying for a history quiz. Draw a scatter plot.

TABLE 4.81:

Study in Hours	4	3	6	2	1	5	4
Grade in Percent	85	78	93	71	61	91	76

6. Draw a reasonable line of fit.

7. What is the equation of the line of best fit?

8. Predict the grade for a student who studied 7 hours.

9. Could the line go on forever? Why or why not?

10. The table below shows the number of reported food poisoning cases at a local hospital. Create a scatter plot for the data.

TABLE 4.82: Reported Food Poisoning Gases

Year	2005	2006	2007	2008	2009
Cases	38	26	19	15	17

11. What relationship, if any exists in the data?

12. Draw a line of fit. Write the slope intercept form of an equation for the line of fit.

13. The table shows the average and maximum lifespan of animals that are kept in captivity. Create a scatter plot to represent the data.

TABLE 4.83: Lifespan in Years

Average	13	26	16	9	36	41	42	21
Maximum	48	51	41	21	71	78	62	55

14. Draw a reasonable line of fit, and write the slope intercept form of the equation.

15. Predict the maximum lifespan for an animal with an average age of 33 years.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.12.

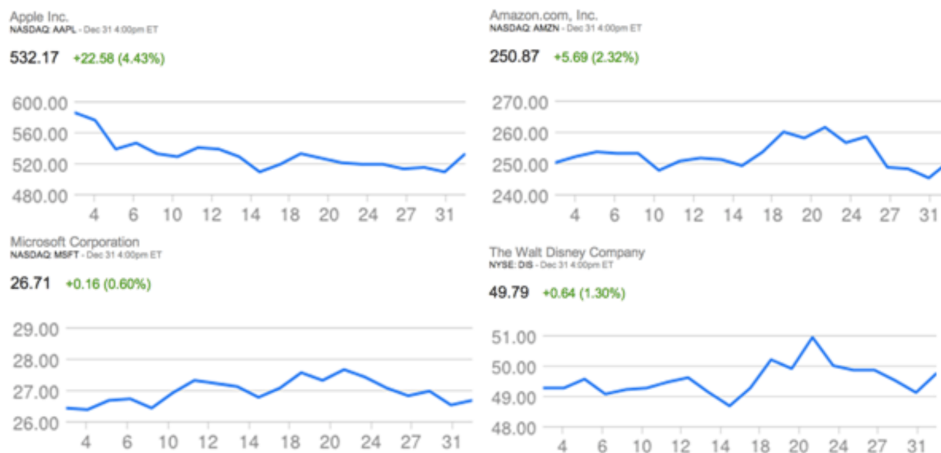
4.13 Interpreting Scatter Plots and Line Graphs

Objective

Here you will practice recognizing and using some of the primary information available from a scatter plot or line graph.

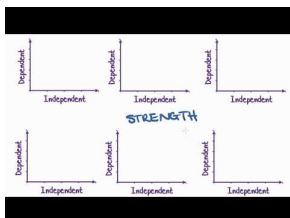
Concept

Steve is playing a stock market simulation game in his social studies class. He has chosen to invest in Apple Inc., Amazon.Com Inc., Walt Disney Company and Microsoft. He originally bought all four stocks in the 10th day of the month. Now he needs to choose one of them to sell. Based on the recent performance of each of them as shown in the line graphs below, which would you recommend he choose and why?



Seem a bit daunting? It certainly can be! We will return to this question at the end of the lesson to review the situation.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63798>

http://youtu.be/PE_BpXTyKCE vcefurthermaths - Maths Tutorial: Interpreting Scatterplots (statistics)

Guidance

The primary use for scatter plots and line graphs is to demonstrate or evaluate the correlation between two variables. Though the two are similar in many ways, there are distinct differences, and specific situations in which one is appropriate and the other is not.

- A **scatter plot** is generally used when displaying data from two variables that may or may not be directly related, and when neither of the variables is under the direct control of the researcher. The primary function of a scatter plot is to visualize the strength of correlation between the two plotted variables. The number of sunburned swimmers at the local pool each day for a month would be an example of a data set that would best be displayed as a scatter plot, since neither the weather nor the number of swimmers present is under the control of the researcher.
- A **line graph** is appropriate when comparing two variables that are believed to be related, and when one of the variables is under the direct control of the researcher. The primary use of a line graph is to determine the **trend** between the two graphed variables. The mileage of a particular car compared to speed of travel would be a good example, since the mileage is certainly correlated to the speed and the speed can be directly controlled by the researcher.

In later lessons we will discuss methods of quantifying the level of correlation between two variables and calculating a line of best fit, but for now we will focus on identifying specific examples of weak or strong correlation and identifying different types of trends.

- **SCATTER PLOTS:**

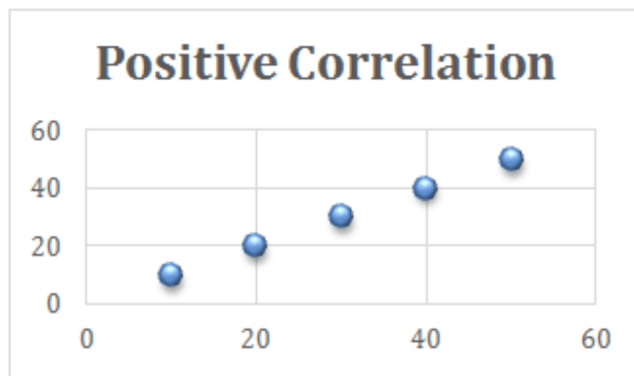
- Two variables with a **strong correlation** will appear as a number of points occurring in a clear and recognizable linear pattern. The line does not need to be straight, but it should be consistent and not exactly horizontal or vertical.
- Two variables with a **weak correlation** will appear as a much more scattered field of points, with only a little indication of points falling into a line of any sort.

- **LINE GRAPHS:**

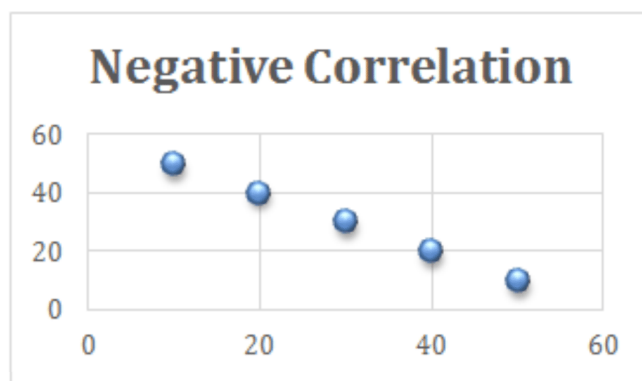
- A **linear relationship** appears as a straight line either rising or falling as the independent variable values increase. If the line rises to the right, it indicates a **direct relationship**. If the line falls to the right, it indicates an **inverse relationship**.
- A **non-linear relationship** may take the form of any number of curved lines, and may indicate a squared relationship (dependent variable is the square of the independent), a square root relationship (dependent variable is the square root of the independent), an inverse square (dependent variable is one divided by the square of the independent), or many other possibilities.

- **BOTH:**

- A **positive correlation** appears as a recognizable line with a positive **slope**. A line has a positive slope when an increase in the independent variable is accompanied by an increase in the dependent variable (the line rises as you move to the right).

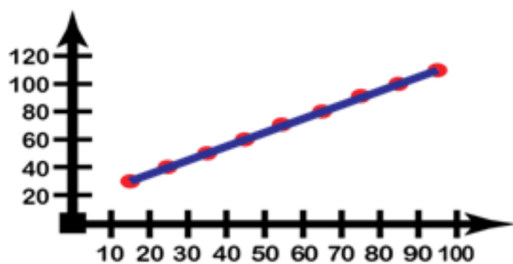


- A **negative correlation** appears as a recognizable line with a negative slope. As the independent variable increases, the dependent variable decreases (the line falls as you move to the right).



Example A

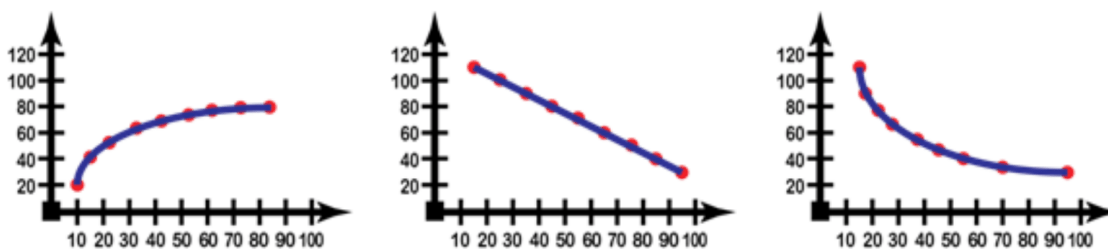
What type of relationship is indicated by the line graph below?



Solution: The line is straight, indicating a linear relationship. It rises from left to right, meaning that the dependent variable increases as the independent variable increases, indicating a positive correlation.

Example B

Which image shows a non-linear graph with a negative correlation?

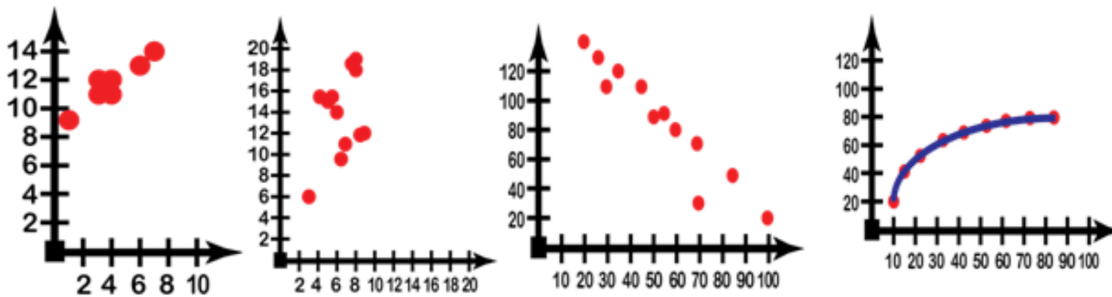


Solution:

- The first image is of a curved line that rises from left to right, this is a non-linear positive correlation
- The second image is a straight line that falls from left to right, this is a linear negative correlation
- The third image is a curved line that falls from left to right, this is a nonlinear negative correlation and is the correct image as described by the question.

Example C

1. Which graph(s) indicate(s) a weak correlation?
2. Which one(s) indicate(s) a strong correlation?
3. Which graph(s) indicate positive correlation(s)?



Solution:

- Only graph 2 indicates a weak correlation, since it is the only one with points that are not clearly arranged in a linear fashion.
- Graphs 1, 3, and 4 all indicate strong correlations, as evidenced by the high percentage of points obviously organized in a line. Graph 4 is obviously a very strong correlation as a clear non-horizontal or vertical line connects all of the points.
- Graphs 1, 2, and 4 are all positive correlations as all three rise from left to right. Another way to put it is that those three graphs have a positive slope (though graph 4 does not have a consistent slope, anywhere on the curve the slope is estimated it would still be positive).

Concept Problem Revisited

Which stock(s) should Steve sell if he needs to make a profit right away?

By looking at the lines on each of the four graphs, we can see that it is important to note that Steve purchased the stocks on the 10th, since only the Walt Disney CO and Amazon are currently valued more highly than they were on the 10th. Both Apple and Microsoft are going up in value now, at the end of the month, but neither has made it back up to where they were on the 10th.

If Steve wants to make a profit right now, he should sell Walt Disney or Amazon or both.

Vocabulary

A **trend** is an estimation of the tendency of data points to move in a certain direction. A **trend line**, also known as a **line of fit**, is a line drawn on a graph to indicate how the data points generally increase or decrease.

A **strong correlation** means that the values of the output variable are strongly affected by the values of the input variable. A strong correlation is indicated on a graph by a large percentage of data points lying in an apparent line, either straight or curved.

A **linear relationship** means that the output values are a simple multiple of the input variable, and appears as a straight line when graphed.

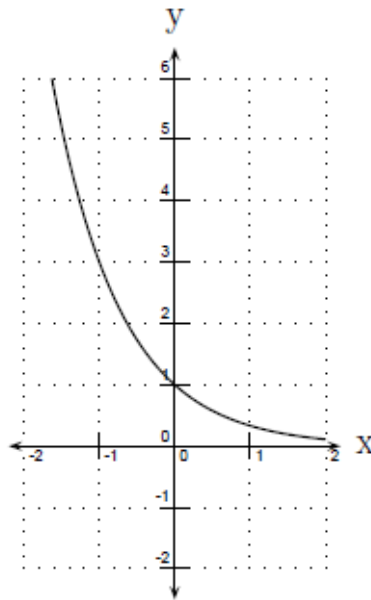
A **non-linear relationship** appears as a curved line on a graph. It indicates an output variable that is a power, a root, or other more complex multiple of the input.

A **direct relationship** means that the variables increase and decrease together, resulting in a **positive correlation** and a line of fit that rises from left to right, whereas an **inverse relationship** is a **negative correlation**, meaning that the output decreases as the input increases, and vice versa.

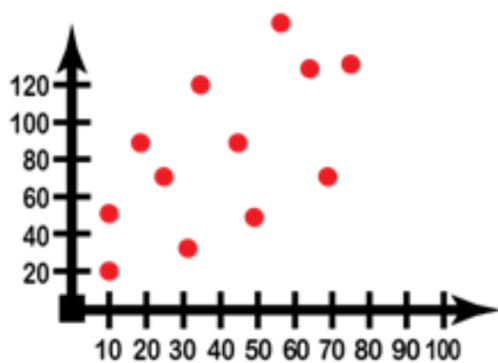
A **slope** is a description of the rate at which the output variable increases or decreases compared to the input variable. This is referred to as the slope because the rate of increase or decrease affects the angle of the line on a graph.

Guided Practice

1. Describe the relationship indicated by the graph:

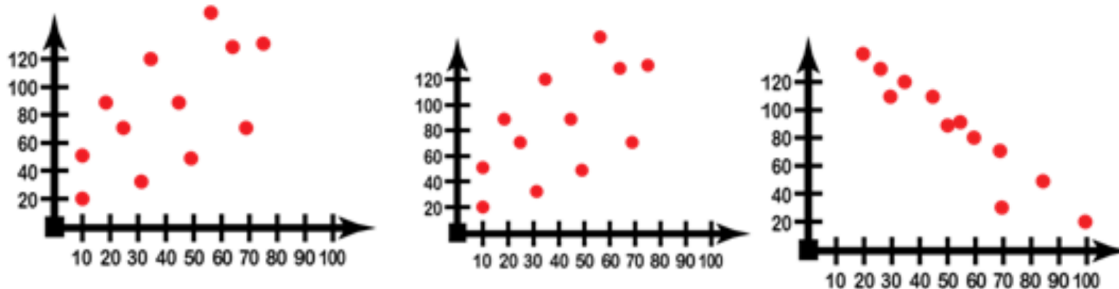


2. Describe the relationship indicated by the graph:

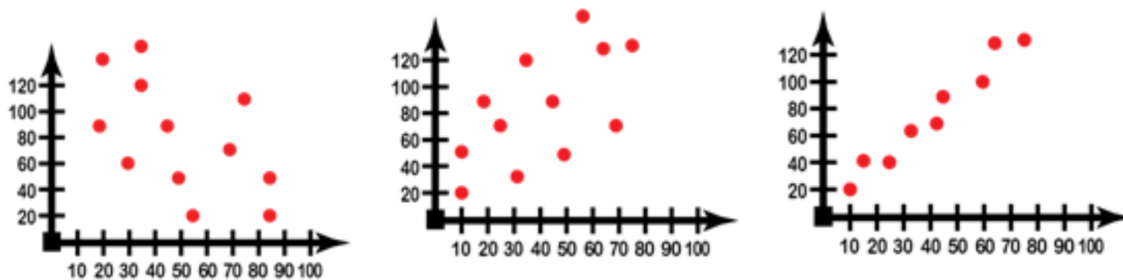


3. Describe the graph that would result from a strongly correlated positive non-linear relationship. Give an example of a function that could result in such a graph.

4. Which scatter plot below indicates the most strongly correlated variables?



5. Which plot below indicates a weakly correlated positive linear relationship?

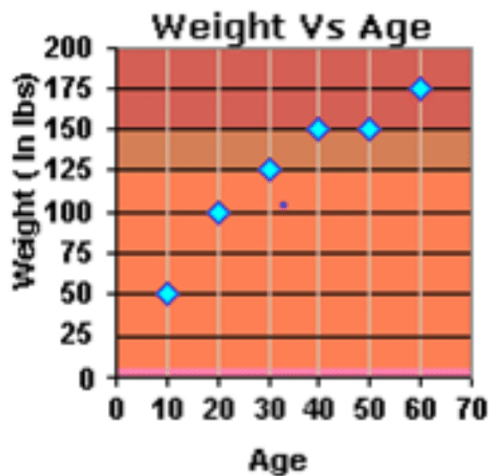


Solutions:

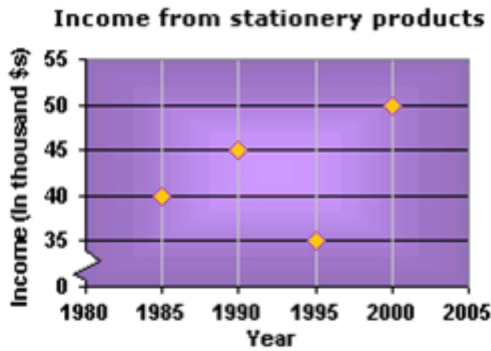
1. This is a very strongly correlated (all the points connected by a line), negative (the line falls from left to right), nonlinear (not a straight line) relationship.
2. This is a weakly correlated (significant scattering of the points), positive (points generally increase in value from left to right), linear (a straight line of fit could be drawn) relationship.
3. A strongly correlated positive non-linear relationship would appear as a well-defined curve of points rising from left to right.
4. The center plot is the most strongly correlated, evidenced by the much cleaner line formed by the data points. Incidentally, this is a negative linear relationship.
5. The left hand plot is weakly correlated, but negative. The right hand plot is positive, but strongly correlated. The center plot is weakly correlated and positive, so it is the one matching the question definition.

Practice

1. What sort of trend is shown in the scatter plot below?



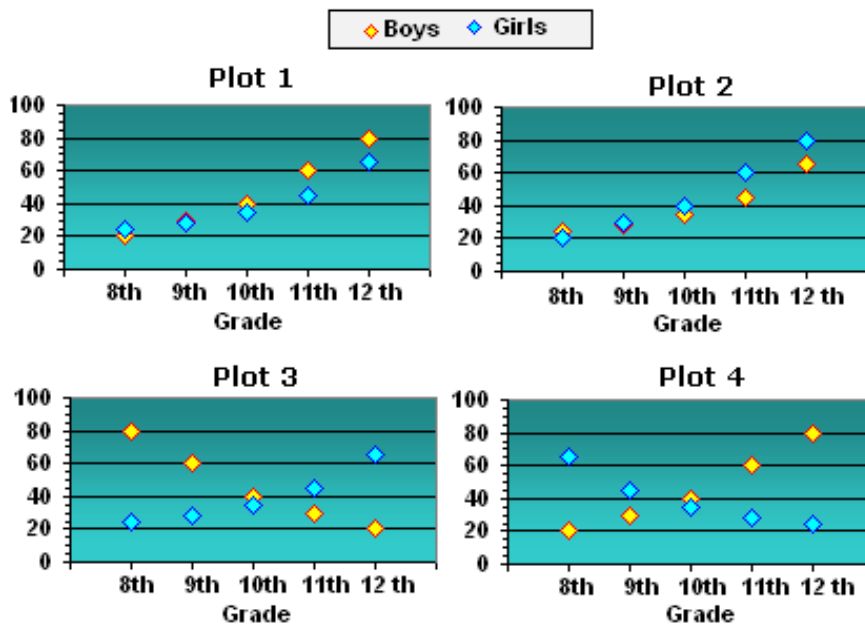
2. A door to door vacuum cleaner sales man plots a scatter diagram of how much he has earned over the years. In which year was his income the highest?



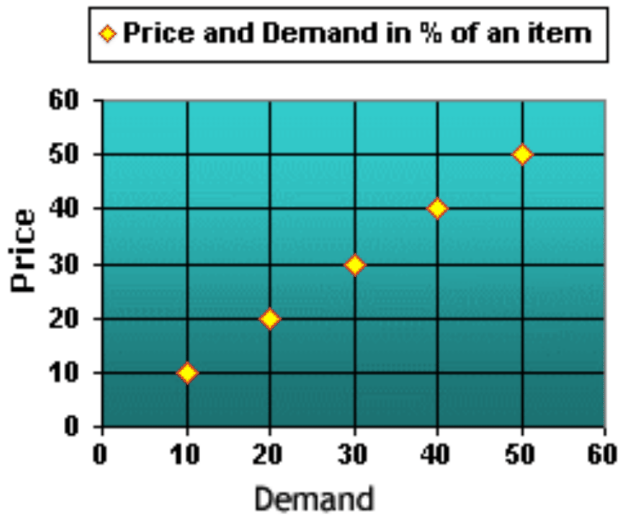
3. The number of children in two different day care centers, and the types of lunch they eat is represented in the table below. Pick the appropriate scatter plot for the data.

TABLE 4.84:

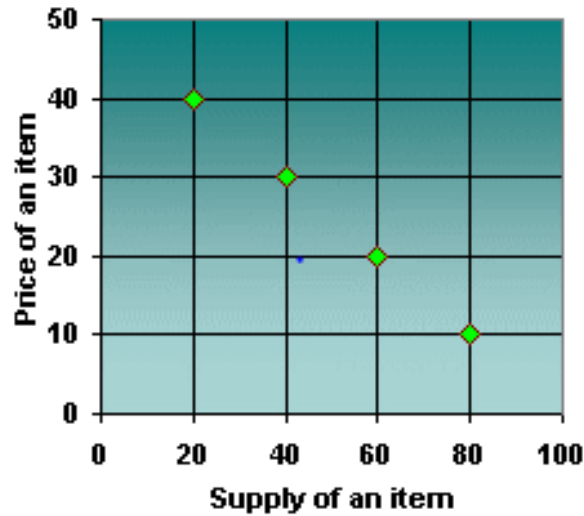
Lunch Served	Center 1 (Yellow)	Center 2 (Blue)
Hamburger	20	25
Mac and Cheese	30	28
Pizza	40	35
Tuna Salad	60	45
Burritos	80	65



4. The plot shown gives the relationship between the demand and price of a trendy consumer good. What trend does the plot follow?



5. The plot represents the relationship between the price and supply of an item. What type of trend does the graph illustrate?



6. Katie recorded the following data relating to how long it took to fill up a horse trough. She measured the depth every two minutes after she began filling it, until it was full. Which scatter plot accurately represents the data?

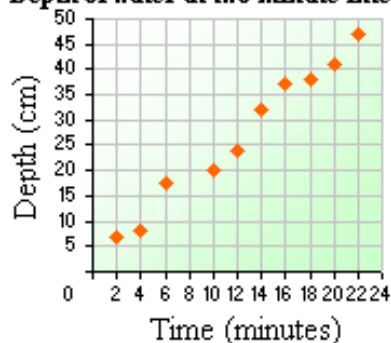
TABLE 4.85:

Time (in minutes)	Dept (in inches)
2	7
4	8
6	13
8	19
10	20
12	24
14	32
16	37
18	38
20	41

TABLE 4.85: (continued)

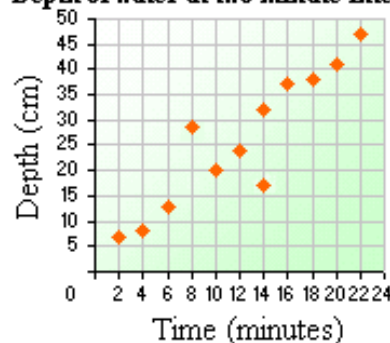
22	47
----	----

Depth of water at two-minute intervals



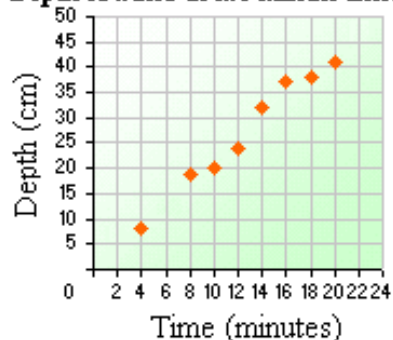
Graph 1

Depth of water at two-minute intervals



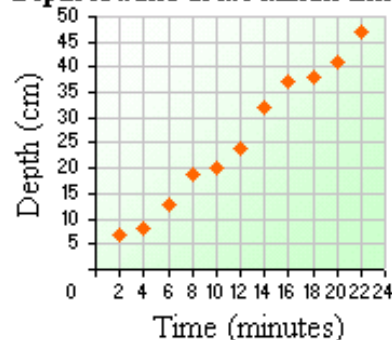
Graph 2

Depth of water at two-minute intervals



Graph 3

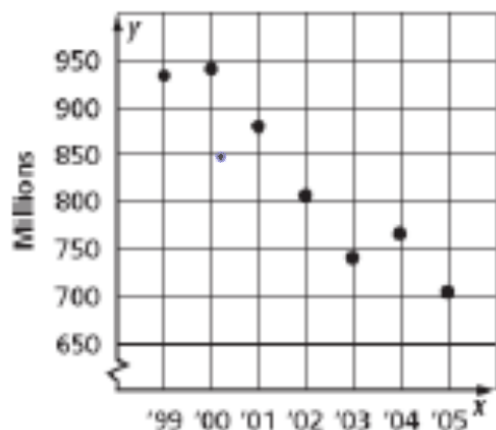
Depth of water at two-minute intervals



Graph 4

7. The Scatter Plot below question 8 shows the number of DVD's sold (in millions) from 2001-2007. Based on the data, about how many DVD's will be sold in 2009?

8. What sort of trend is shown in the scatter plot?



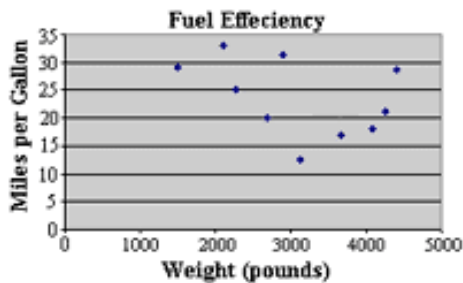
9. The table below shows a relationship between the weight of a car and its average gas mileage. Which plot best represents the data?

TABLE 4.86:

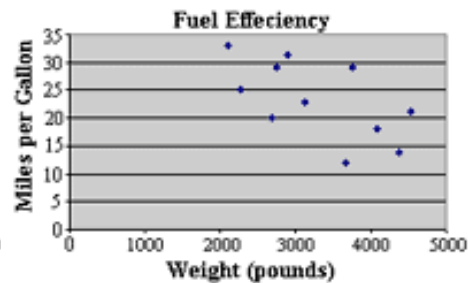
Type of Car	Weight	MPG
-------------	--------	-----

TABLE 4.86: (continued)

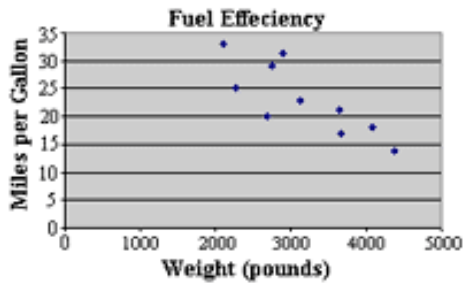
1	3750	29
2	4125	23
3	3100	33
4	5082	18
5	3690	20
6	4640	21
7	5380	14
8	3241	25
9	3895	31
10	4669	17



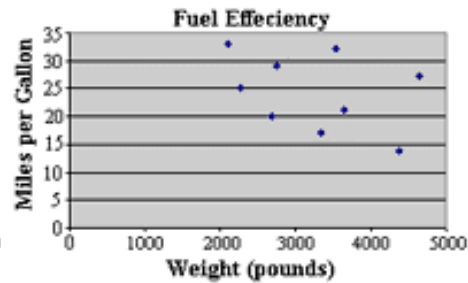
Graph 1



Graph 2

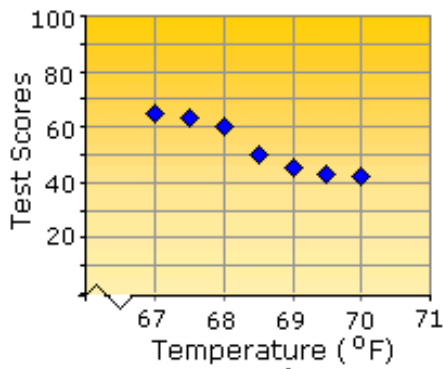


Graph 3

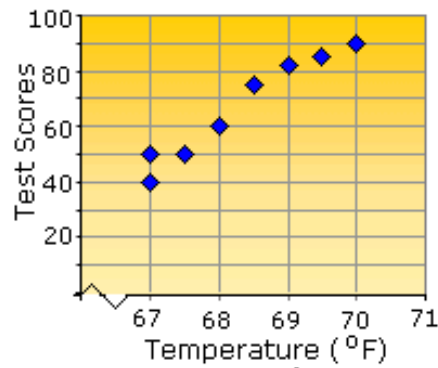


Graph 4

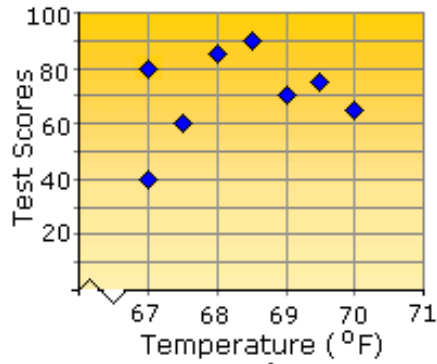
10. Which scatter plot shows no relationship between test scores received by Greg, and the temperature that the classroom was at while taking the test? Why?



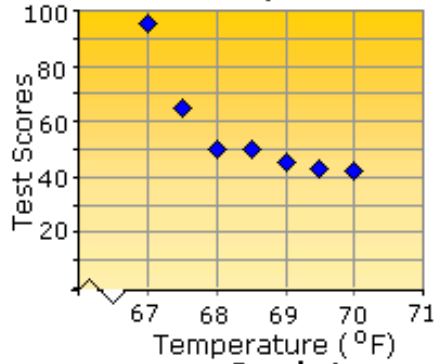
Graph 1



Graph 2

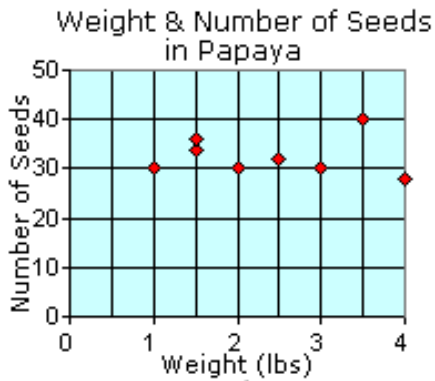


Graph 3

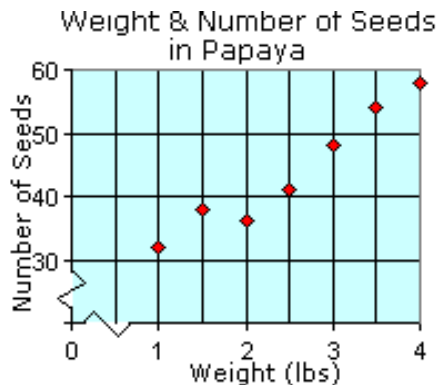


Graph 4

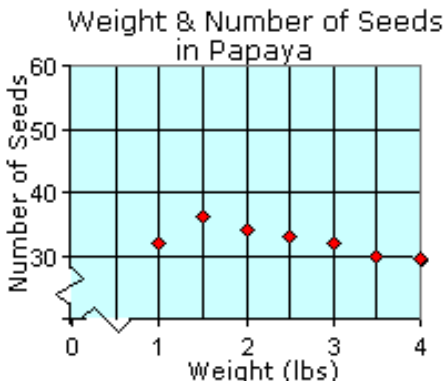
11. Which scatter plot shows a positive relationship between the weight of a mango, and the number of seeds it contains?



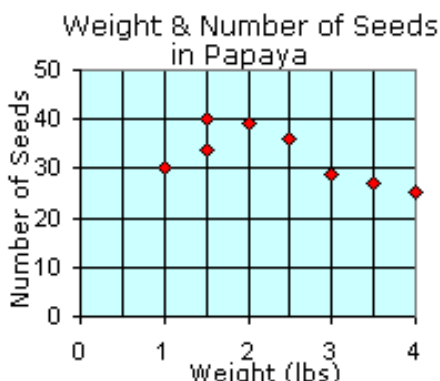
Graph 1



Graph 2

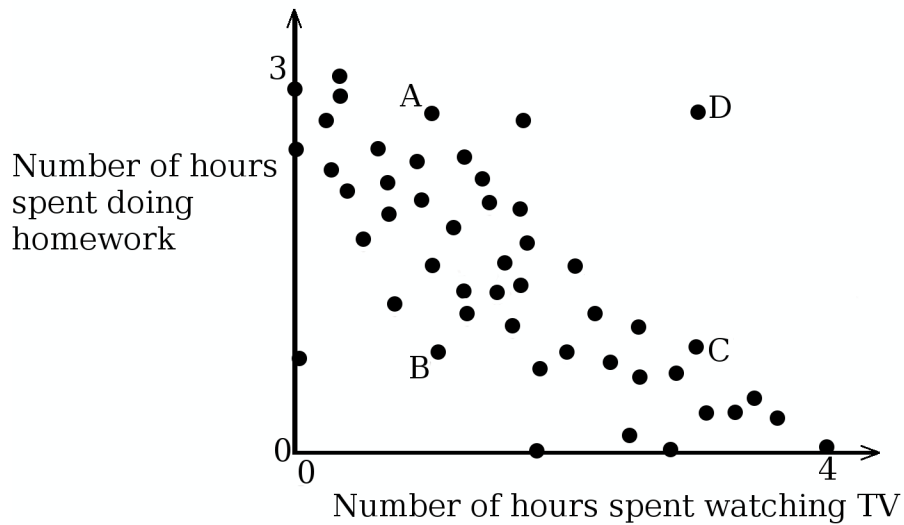


Graph 3



Graph 4

Roy was doing research for a research paper. He questioned students throughout his high school, asking them how much time they spent doing homework and how much time they spent watching TV the previous evening. The following scatter plot shows his results. Based on the information answer the questions that follow.



Choose the best of the 4 points, A, B, C, or D to represent the student's statements below.

12. "I worked on homework almost all night, I only had time to watch my favorite sitcom."
13. "Last night was about half and half for me"
14. "Last night didn't have anything on the screen I wanted to watch, and homework was so light, that I ended up going out."
15. Write a statement that correlates to the 4th point.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.13.

4.14 Creating Pie Charts

Objective

Here you will learn to create an accurate pie chart to display comparative data.

Concept

Kalena's cheerleading squad is raising money for prom by selling candy at school football games. After a month of sales, the squad is running low on candy and decides to review the sales so far to help them decide what to order when they restock.

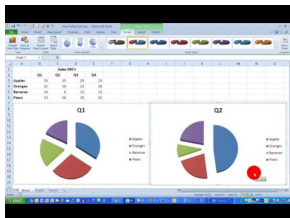
The table below describes the sales that the girls have recorded. How could the girls create a visual representation of the data so they can more easily present their findings to the purchasing committee? Ideally, they would like to order more of the item that is proving the most profitable, and so they want to present the data to the committee in a way that highlights this. We will review this question after the lesson.

TABLE 4.87:

Item	# of Sales
Popsicles \$1 ea	850
Chocolate Bars \$0.85 ea	1300
Bag of Lemon Drops \$1.25 ea	340
Ice Cream Bars \$1 ea	670

Watch This

There are two videos applicable to this lesson, the first is a demonstration of creating a pie chart with the use of spreadsheet software, the second explains creation by hand.



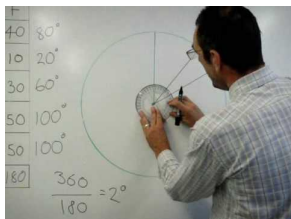
MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63840>

<http://youtu.be/IVIXbH4nczI> eoloughlin - How To... Draw a Simple Pie Chart in Excel 2010

Just for clarity, note that this video shows the same pie chart created *twice*, once incorrectly, and then once correctly, as stated in the first few seconds of the video.

**MEDIA**

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63842>

http://youtu.be/W60F_517PVY krmathsandscience - How to Construct a Pie Chart

Guidance

Pie charts are an excellent way to compare multiple values that make up parts of a whole. Each piece of the pie is called a **sector**, and each sector encompasses an angle that is proportional to the frequency of the data it represents. The formula relating the angle of a sector to frequency is:

$$\text{Sector Angle} = \frac{\text{Frequency of Data in the Sector}}{\text{Total Frequency of Data in Sample}} \times 360^\circ$$

In order to construct an accurate pie chart, you will need to calculate the sector angles for each of the categories or intervals in your sample, verifying that they total 360° .

Once you know the angles for each sector construct a circle and mark the sectors within it with lines running from the center of the circle to the edge (radii). Make sure that the angle between the lines is equal to the calculated sector angle for each category.

Finally you need to either label the sectors directly or create a key similar to the one in the concept question above so your audience can easily identify which sector corresponds to each category in your sample.

Example A

In Karen's school, there were 480 students in 1997, 540 students in 2000, 710 students in 2003, and 900 students in 2006. Construct a pie chart to represent the relative numbers of students each year.

Solution:

First, calculate the total number of students over all four categories (years):

$$\text{Total number of students} = 480 + 540 + 710 + 900 = 2,630$$

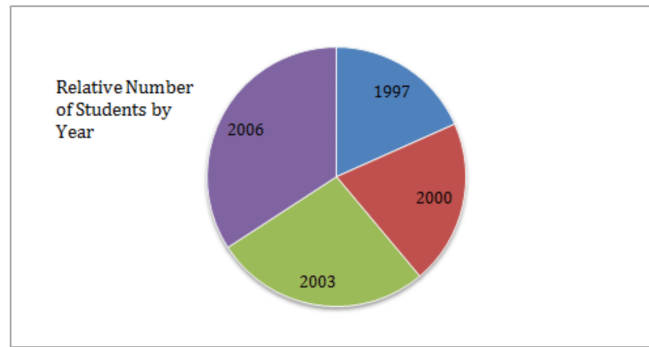
Now calculate the angle of each sector with $\text{Sector Angle} = \frac{\text{Frequency of Data in the Sector}}{\text{Total Frequency of Data in Sample}} \times 360^\circ$

- 1997: $\text{sector angle} = \frac{480}{2630} \times 360^\circ \rightarrow .183 \times 360^\circ \rightarrow \approx 66^\circ$
- 2000: $\text{sector angle} = \frac{540}{2630} \times 360^\circ \rightarrow .205 \times 360^\circ \rightarrow \approx 74^\circ$
- 2003: $\text{sector angle} = \frac{710}{2630} \times 360^\circ \rightarrow .270 \times 360^\circ \rightarrow \approx 97^\circ$
- 2006: $\text{sector angle} = \frac{900}{2630} \times 360^\circ \rightarrow .342 \times 360^\circ \rightarrow \approx 123^\circ$

Verify that your degree measures total 360° :

$$66^\circ + 74^\circ + 97^\circ + 123^\circ = 360^\circ$$

Finally, construct your circle and draw the internal angles equal to the calculated sector angles, and color-code and/or directly label each sector:

**Example B**

Create a pie chart to display the data from the table below:

TABLE 4.88: Makes of Student Cars in Parking Lot

Ford	57
Chevrolet	49
Dodge	36
Toyota	27
Nissan	16
BMW	5
Mercedes	3
All Others	17

Solution: First total the number of cars in the entire population:

$$57 + 49 + 36 + 27 + 16 + 5 + 3 + 17 = 210$$

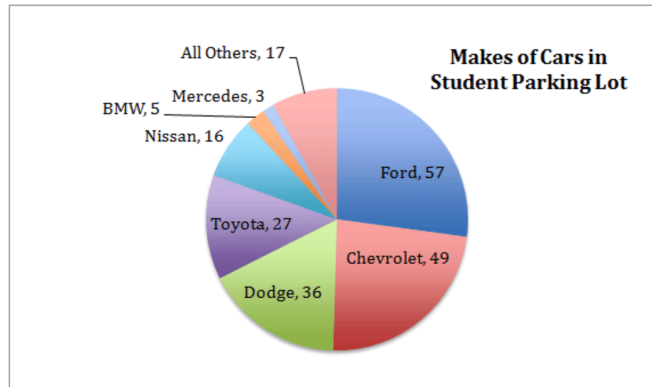
Now find the degree measure of each sector using:

$$\text{Sector Angle} = \frac{\text{Frequency of Data in the Sector}}{\text{Total Frequency of Data in Sample}} \times 360^\circ$$

- Ford: $\text{Angle} = \frac{57}{210} \times 360^\circ \approx 98^\circ$
- Chevrolet: $\text{Angle} = \frac{49}{210} \times 360^\circ = 84^\circ$
- Dodge: $\text{Angle} = \frac{36}{210} \times 360^\circ \approx 62^\circ$
- Toyota: $\text{Angle} = \frac{27}{210} \times 360^\circ \approx 46^\circ$
- Nissan: $\text{Angle} = \frac{16}{210} \times 360^\circ \approx 27^\circ$
- BMW: $\text{Angle} = \frac{5}{210} \times 360^\circ \approx 9^\circ$
- Mercedes: $\text{Angle} = \frac{3}{210} \times 360^\circ \approx 5^\circ$
- All Others: $\text{Angle} = \frac{17}{210} \times 360^\circ \approx 29^\circ$

Verify that the total is 360° : $98^\circ + 84^\circ + 62^\circ + 46^\circ + 27^\circ + 9^\circ + 5^\circ + 29^\circ = 360^\circ$

Construct a circle with sectors representing each degree measure and label directly or create a key:



Example C

Use a spreadsheet or compass and protractor to create two related pie charts of the data in the table below. If using a modern spreadsheet, create a 3-D graph. Highlight the data from the soccer participants in both graphs.

TABLE 4.89:

Sport	Count of Middle School Participants	Count of High School Participants
Football	186	279
Volleyball	28	57
Soccer	66	92
Track	82	124

Solution: By now you are familiar with creating charts using a pencil and paper, so let’s walk through creating a chart in a modern spreadsheet.

If you do not have spreadsheet software on your computer, you can download the *free* ‘Open Office Calc’ spreadsheet software from <http://www.openoffice.org/> . The process described here is essentially the same in any modern spreadsheet software such as Open Office Calc, Microsoft Excel or Numbers.

First, highlight and copy the two columns of data under ‘Sport’ and ‘Count of Middle School Participants’ from the table in the question, include the column headers.

Now open a blank spreadsheet in your software and paste the two columns of data.

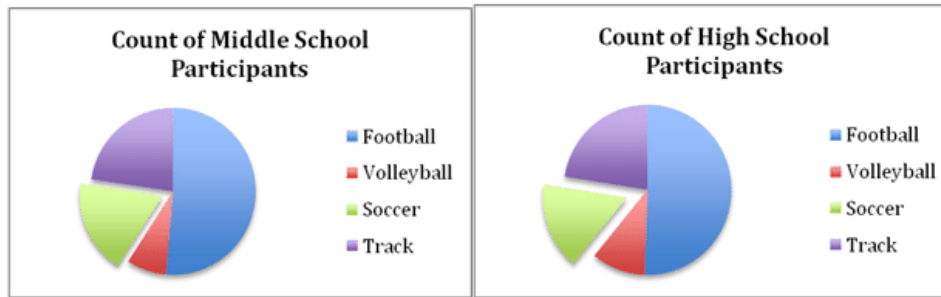
Highlight the data in the spreadsheet (including the headers), and either choose ‘Charts : Pie’ from the toolbar or click the icon that looks like a bar graph and choose ‘Pie.’

The data will be immediately converted to a pie chart for you! Now pick a 3-D style (since that is what was specified in the question), and press ‘enter’ to get a chart like the one below.

To highlight the soccer data, simply select the sector representing the soccer participants and drag it away from the center of the circle just slightly. Finally, select the entire chart in the spreadsheet and copy/paste it into your answer document or print it out to turn it in.

To create the chart for the high school data, just copy the data from the single ‘Count of High School Participants’ category and paste it right overtop of the Middle School column in your spreadsheet and repeat the steps above to convert the revised table into a chart.

The two final products should look something like these:



Concept Problem Revisited

Create a pie chart that compares the income from each product for the cheerleading squad. Use the data in the table of candy sales.

In the example problems, the first step was to total the number of items sold. However, in this problem, we need to compare the *dollar value* of the items rather than just the *number* of items. That means we need to first evaluate the dollar value of each product sale, and then calculate the angle of each slice based on a comparison of the dollar value of each product with the total income from sales. Just to keep things neat, let's add another column to the original table called "dollar value", and another row at the bottom for the total.

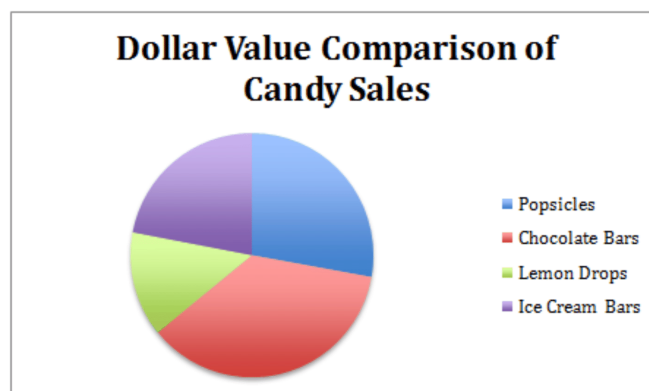
TABLE 4.90:

Item	# of Sales	\$ Value
Popsicles \$1 ea	850	\$850.00
Chocolate Bars \$0.85 ea	1300	\$1,105.00
Bag of Lemon Drops \$1.25 ea	340	\$425.00
Ice Cream Bars \$1 ea	670	\$670.00
TOTAL	3160	\$3,050.00

Now we can use the formula to calculate the angle of each slice:

- Popsicles: $Angle = \frac{\$850}{\$3,050} \times 360^\circ \approx 100^\circ$
- Chocolate Bars: $Angle = \frac{\$1,105}{\$3,050} \times 360^\circ \approx 130^\circ$
- Lemon Drops: $Angle = \frac{\$425}{\$3,050} \times 360^\circ \approx 50^\circ$
- Ice Cream Bars: $Angle = \frac{\$670}{\$3,050} \times 360^\circ \approx 79^\circ$

Finally, we construct our circle and mark the divisions of the sectors based on the angles we have calculated, label the sectors, and label the graph.



Vocabulary

A *sector* is a single 'pie slice' in a circle graph.

The *whole relationship* is represented by the entire circle.

Guided Practice

1. Larry Bird was a well know basketball player. He played for the Boston Celtics. Use the following information to create a pie chart.

TABLE 4.91:

Season:	# of Points Scored
1979-1980	1745
1980-1981	1741
1981-1982	1761
1982-1983	1867
1983-1984	1908
1984-1985	2295
1985-1986	2115
1986-1987	2076
1987-1988	2275
1988-1989	116
1989-1990	1820
1990-1991	1164
1991-1992	908

2. What percent of his career points came from the 1988 season?

3. What percent of his career points came from the 1980 season?

The following table shows the grades achieved by 30 pupils in their end of year exam.

TABLE 4.92:

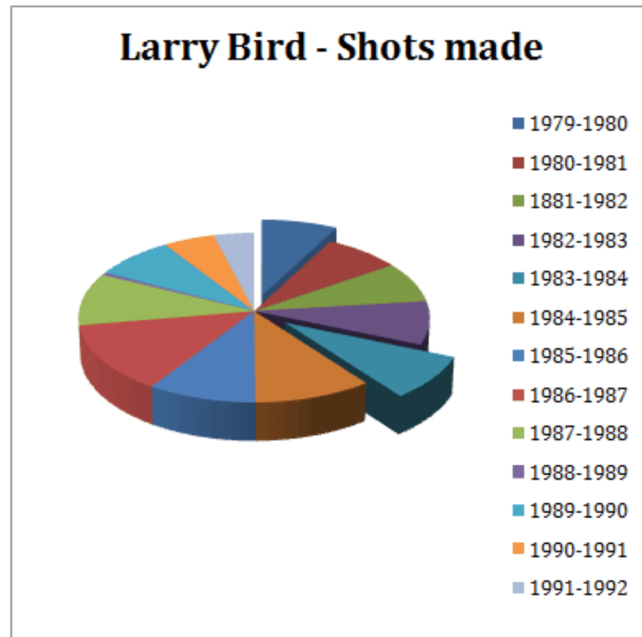
Grade	A	B	C	D	F
Frequency	8	10	7	3	3

4. Based on the number of pupils, how many degrees of a circle graph are allotted per student? What percentage of the pie cart is allotted to "C" grades?

5. Create a pie chart for the information.

Solutions:

1.



2. First we must find the total number of baskets Larry made in the portion of his career represented by the chart. The answer is 21,792. This number represents 100 percent of our chart. To find what percentage of those shots were made in 1988, we write an algebraic word problem that looks this:

What percent of 21,792 is 2275? As an equation: $\boxed{?} \times 21,792 = 2275 \rightarrow \boxed{?} = \frac{2275}{21792} \rightarrow \boxed{?} = 10.4\%$

Percent of career shots made in 1988 is 10.4%

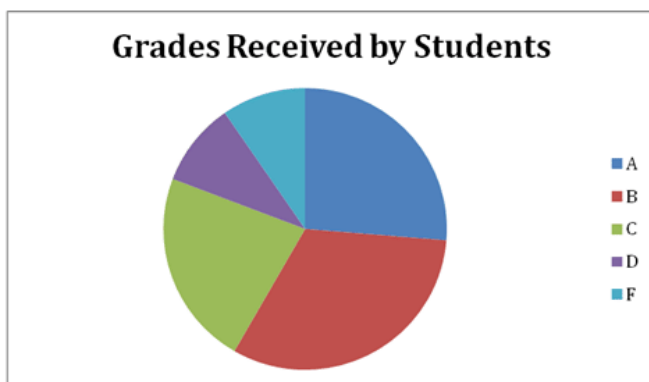
3. Calculate this the same way: $x\% \times 21792 = 1741 \rightarrow \frac{1741}{21792} = .08 \rightarrow x = 8\%$

Percent of shots made in his career in 1980 is 8%

4. There are 31 total grades, 7 of which are “C’s”. To find percentage, just divide: $\frac{7}{31} = .225$, and then multiply by 100 to get percentage. **“C” grades make up 22.5% of the chart.**

5. Your pie chart should look something like the image below, with sectors measuring:

- Grade A: $\frac{8}{31} = .26 = 26\%$ Convert to degrees: $.26 \times 360^\circ = 94^\circ$
- Grade B: $\frac{10}{31} = .32 = 32\%$ Convert to degrees: $.32 \times 360^\circ = 115^\circ$
- Grade C: $\frac{7}{31} = .225 = 23\%$ Convert to degrees: $.225 \times 360^\circ = 81^\circ$
- Grade D: $\frac{3}{31} = .097 = 10\%$ Convert to degrees: $.097 \times 360^\circ = 35^\circ$
- Grade F: $\frac{3}{31} = .097 = 10\%$ Convert to degrees: $.097 \times 360^\circ = 35^\circ$



Practice

1. What is each part of a pie chart called?
2. What part of the circle represents the whole relationship?
3. Based on the data in the table below, which candy sold the most, and which the least?

TABLE 4.93:

Candy Bar	Number sold at the school store	Percentage of Total Sales	Number of Degrees on a Pie Chart
Heath	151		
M	M	191	
Snickers	61		
Skittles	107		
Almond Joy	91		

4. Fill in the chart above with percentages of each candy type sold.
5. Fill in the chart with degrees of a circle required to represent this amount on a pie chart.
6. Create a Pie Chart to represent this data.
7. Based on the data in the table below, which stock has the greatest potential for making money for the investor who owns the stock?

TABLE 4.94:

Company	Shares Owned	Percentage of Total Portfolio	Number of Degrees on a Pie Chart
Hostess	8		
Pepsi	11		
Dell	5		
Conoco	7		
Ford Motor	19		

8. Fill in the chart with relative percentages of each type of stock.
9. Fill in the chart with degrees of a circle required to represent this amount on a pie chart.
10. Create a Pie Chart to represent this data.

Use the data presented in questions 11-15 to create pie charts for each:

11. Julie runs for one hour per day, reads for two and sleeps nine. She spends about two hours eating, at least one hour on the phone with friends. She hangs out with her family on average 4 hours a day, and spends at least five hours a day studying.
12. Mrs. Garcia makes \$1200.00 a month. She puts 10% in savings, spends 20% on her car payment and insurance, and another 20% on groceries. She likes clothes, so 10% of her income goes towards her wardrobe. Of her remaining money she spends 30% on her mortgage, and the remainder on miscellaneous expenses. Determine exactly how much Mrs. Garcia spends in each category. What percent of her income goes to miscellaneous expenses?
13. Katie earned 500.00 doing odd jobs for people in her neighborhood. She spent $\frac{1}{16}$ of her money on the movies, she spent $\frac{3}{8}$ of her money going out with friends. She spent $\frac{1}{2}$ on clothes, and the remainder on books. How much money did Katie spend on movies? What percentage was spent on books?

14. The state of Colorado receives 28 inches of precipitation a year. The winter is when it gets most of it, but it does not see it until it melts and runs off in the spring. However, inches are counted when they accumulate, and so they represent precipitation for Colorado based on a 4 season year. Colorado receives $\frac{3}{5}$ of its precipitation in the winter, $\frac{3}{10}$ in the spring and the rest in the summer and fall months.

15. Students were preparing to go on a field trip, and their teacher let them choose the destination. There were 36 students in the class. 44.4% choose the Nature Preserve, 25% chose an Art Gallery. Half as many students wanted to go to the Symphony as the Nature Preserve, and the rest wanted to go to the Museum of Nature and Science. How many more students were there that wanted to go to the Nature Preserve than wanted to go to the Museum of Nature and Science?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.14.

4.15 Interpreting Pie Charts

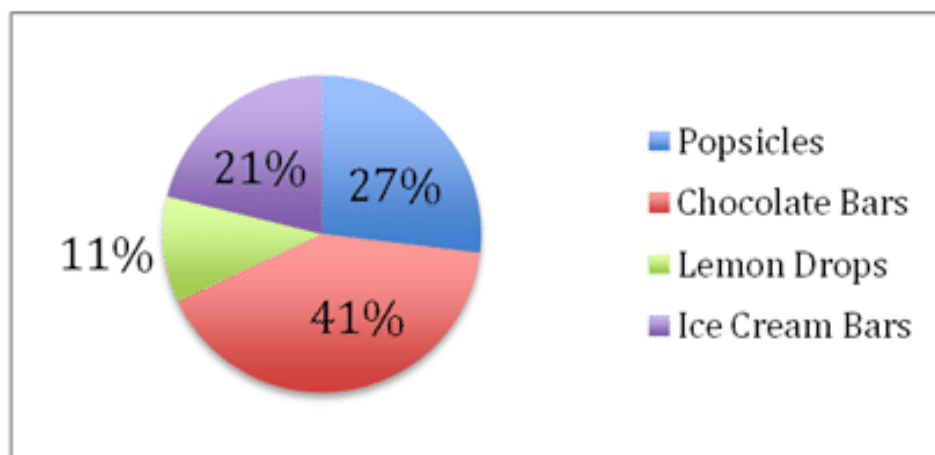
Objective

Here you will learn to read and evaluate the comparative data on a pie chart.

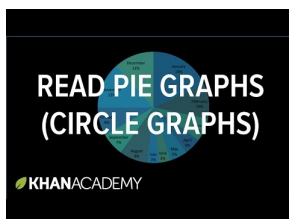
Concept

Kalena's cheerleading squad is raising money for prom by selling candy at school football games. After a month of sales, the squad is running low on candy and decides to review the sales so far to help them decide what to order when they restock.

The pie chart below describes the number of sales that the girls have recorded for each item. If you know that they sold 850 fruit popsicles, how could you calculate the number of chocolate bars or bags of lemon drops they sold? We will review this question after the lesson.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/62548>

<http://youtu.be/4JqH55rLGKY> KhanAcademy - Reading Pie Chart Graphs

Guidance

Reading a pie chart is relatively simple, which is one of the primary values of a pie chart. The only real trick to it is to become very familiar with the *sector angle formula* and to practice using it to deduce the value(s) of unspecified

data.

Recall that each piece of the pie is called a **sector**, and each sector encompasses an angle that is proportional to the frequency of the data it represents. The formula relating the angle of a sector to frequency is:

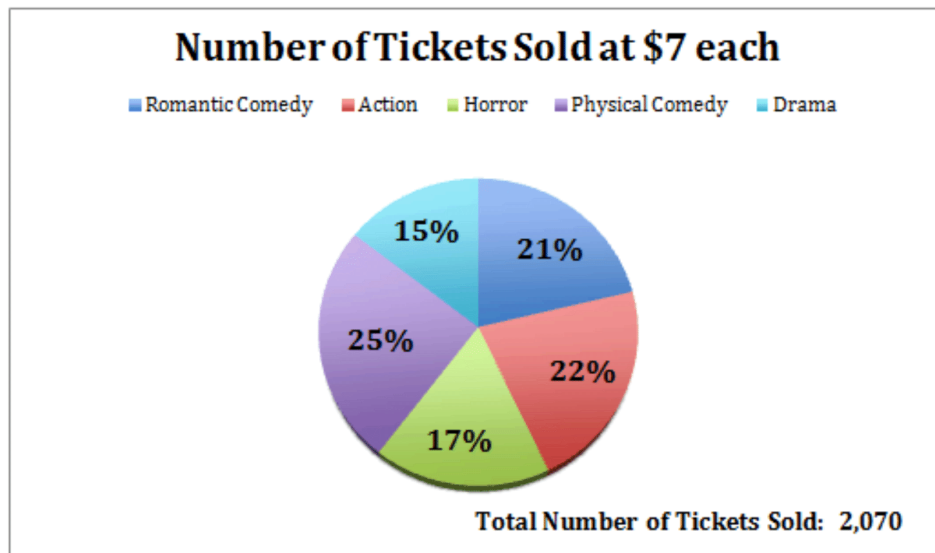
$$\text{Sector Angle} = \frac{\text{Frequency of Data in the Sector}}{\text{Total Frequency of Data in Sample}} \times 360^\circ$$

Because the primary skill here is just familiarity that comes with practice, let's jump right into the examples.

Example A

Given the pie chart below, identify or calculate the following:

- The total number of tickets to comedy-themed movies.
- The dollar value of tickets sold to horror movies.
- The percentage of tickets sold to action and romantic comedies together.



Solution:

a. To calculate the total number of comedy movie tickets, we need to multiply the decimal equivalent of each of the two comedy categories by the total number of tickets sold to learn how many of each were sold, then find the sum.

i. $.21 \times 2070 = 435$ *romantic comedy tickets*

j. $.25 \times 2070 = 518$ *physical comedy tickets*

k. 435 *romantic comedy* + 518 *physical comedy* = 953 *Total Comedy Tickets*

b. To find the dollar value of the horror movies, first multiply the decimal equivalent of the horror movie percentage, and multiply the result by \$7, the price of each ticket.

l. $.17 \times 2070 = 352$ *horror movie tickets*

m. $352 \times \$7 = \$2,464$ *in horror movie ticket sales*

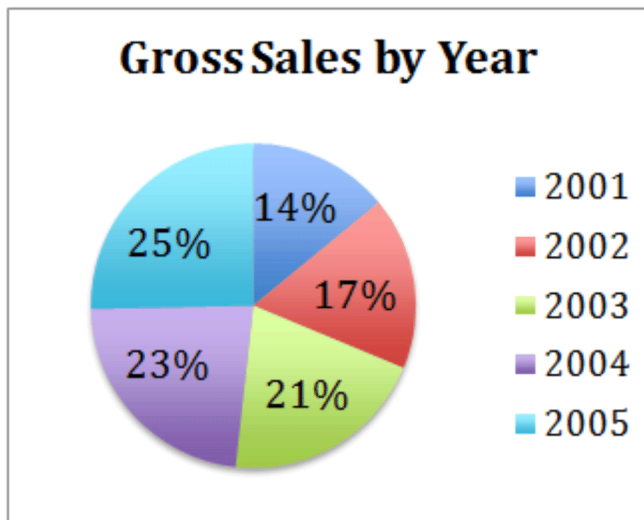
c. To find the percentage of action and romantic comedy tickets together, simply sum the given percentages.

n. $22\% + 21\% = 33\%$ *total action and romantic comedy tickets*

Example B

Given the pie chart below, calculate the actual gross sales in years 2001, 2002, 2003, and 2004.

Gross sales in 2005 were \$94,000



Solution: First we need to calculate the total gross sales for all five years. We know that the \$94,000 in gross sales from 2005 represents 25% of the total, so we can calculate:

$$.25 \times x = \$94,000 \rightarrow \frac{\$94,000}{.25} = \$376,000$$

Now we can simply multiply the total gross sales by the percentage represented by each year to get the estimated dollar value for each category:

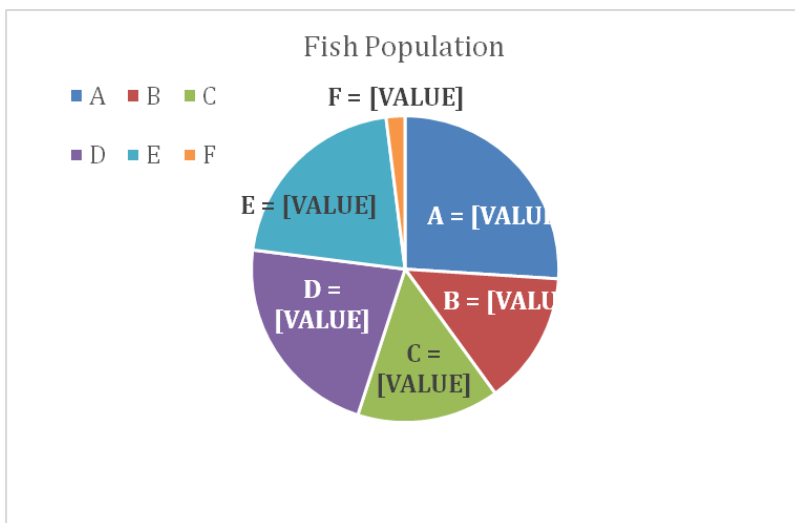
- 2001: 14% of \$376,000 = \$52,640
- 2002: 17% of \$376,000 = \$63,920
- 2003: 21% of \$376,000 = \$78,960
- 2004: 23% of \$376,000 = \$86,480

Verify that the total values add up: $\$52,640 + \$63,920 + \$78,960 + \$86,480 + \$94,000 = \$376,000$

Example C

Evaluate the pie chart below and answer the questions:

- If the population of Lake C is 27,000 fish, what is the total fish population of all lakes on the chart?
- What is the ratio of the fish population in lakes A and C?
- What is the combined population of lakes D and E?



Solution:

a. The population of lake C is 27,000 fish, and represents 15% of the whole, use the data to set up an equation:

$$.15x = 27,000 \rightarrow x = \frac{27000}{.15} \rightarrow x = 180,000$$

The combined population of all lakes on the chart is 180,000 fish.

b. Lake A represents 26% of the total, and Lake C 15%.

$$\text{The ratio is } \frac{26}{15} = 1.73 : 1$$

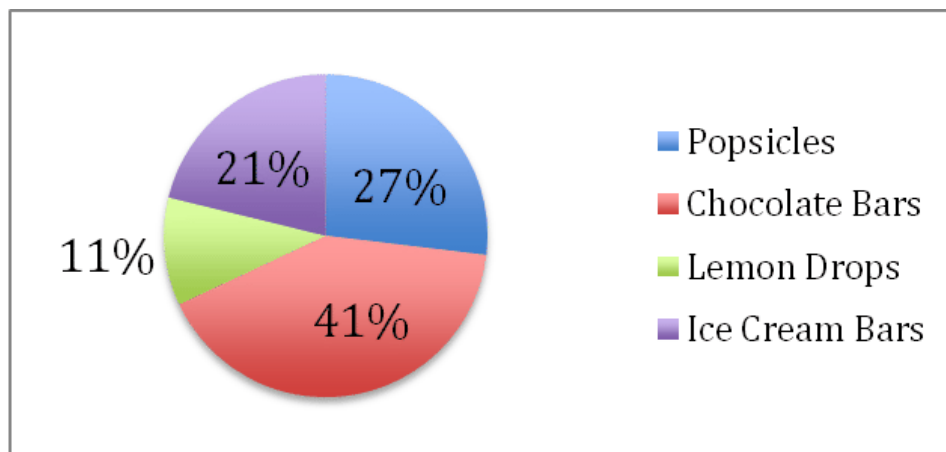
c. Lake D is 22% of 180,000 and Lake E is 21% of 180,000.

Together they represent $22\% + 21\% = 43\%$ of the total.

$$43\% \times 180,000 = 77,400 \text{ fish in lakes D and E together.}$$

Concept Problem Revisited

The pie chart below describes the sales that the girls have recorded as numbers of sales. If you know that they sold 850 fruit popsicles, how could you calculate the number of chocolate bars or bags of lemon drops they sold?



This one should be a piece of... pie... now! 850 fruit popsicles represents 27% of the total. Set up an equation:

$$27\%x = 850 \rightarrow x = \frac{850}{.27} \rightarrow 3,148 \text{ total sales}$$

Chocolate Bars represent 41% of the 3,148 total sales:

$$.41 \times 3148 = 1291 \text{ chocolate bars}$$

Lemon Drops represent 11% of the 3,148 sales:

$$.11 \times 3148 = 346 \text{ lemon drop bags}$$

Vocabulary

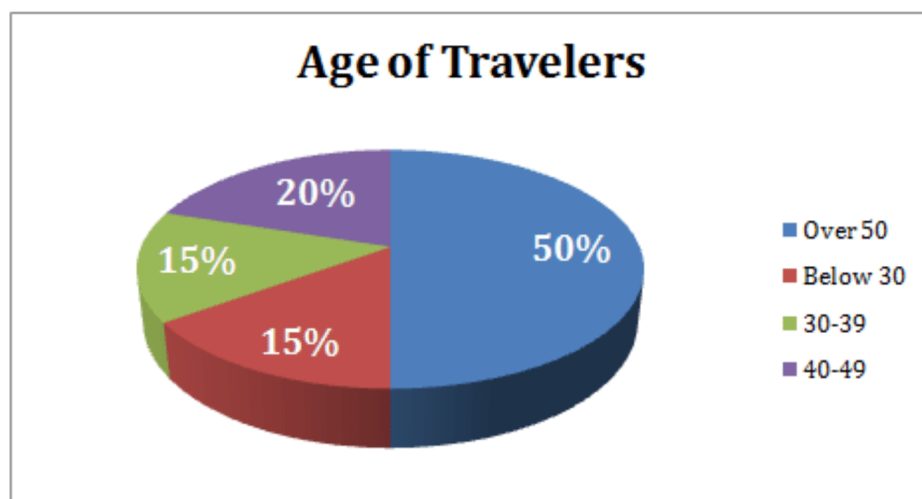
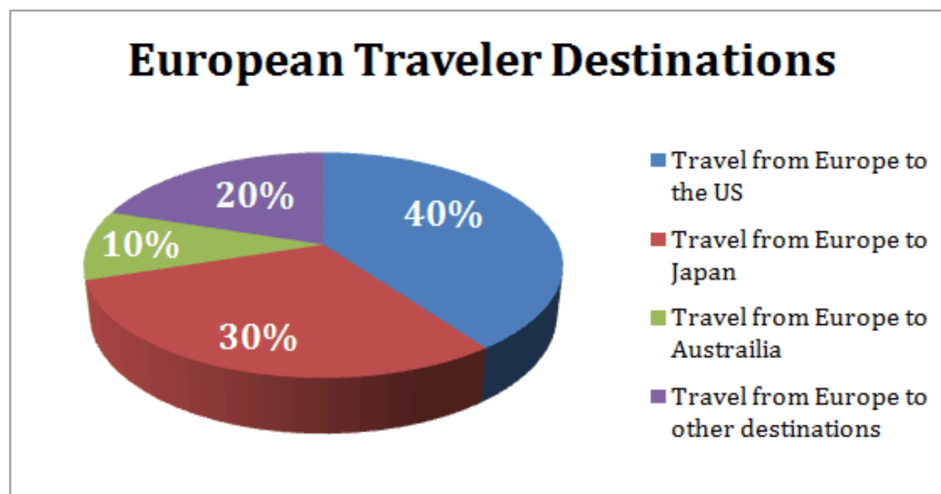
A **sector** is a single portion of a pie or circle graph (one “slice”)

The **sector angle formula** is used to calculate how many degrees of the circle should be allocated to a given value.

The formula is:
$$\text{Sector Angle} = \frac{\text{Frequency of Data in the Sector}}{\text{Total Frequency of Data in Sample}} \times 360^\circ$$

Guided Practice

The two charts below show the distribution of overseas tourist traffic from Europe to other countries. They show the distribution by country and by age, respectively.



1. What percentage of tourists went to either the US or Australia?
2. What is the ratio of European tourists that went to the USA to the number of European tourists who were below the age of 30?
3. If, among the “other destinations”, India accounted for 25% of the travel, and it is known that 25 Europeans went to India, how many 30-39 year old Europeans went abroad in that year?
4. Based on the data, about how many 40-49 year olds traveled to Japan?

Solutions:

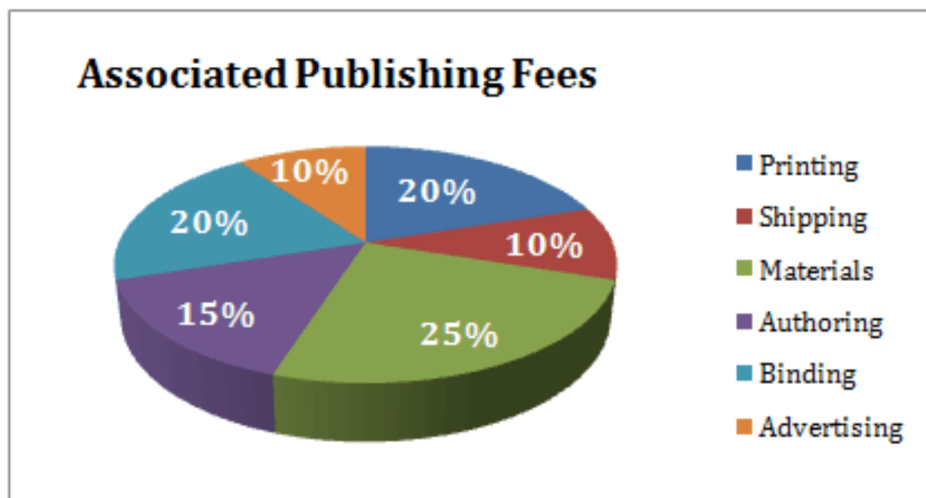
1. We can see from the top chart that U.S. travelers (in blue) accounted for 40% of the travel, and Australian travelers (green) represented 10%. The total percentage who went to either country is $40\% + 10\% = 50\%$ of travelers.
2. 40% of travelers went to the U.S, and 15% were under 30 (bottom graph, in red). **The ratio is 40:15 or 8:3.**
3. If India accounts for 25% of the 20% of travelers who went to “other destinations”, then it accounts for 5% of the total travelers ($25\% \times 20\% = 5\%$). If 25 Europeans went to India, then we know 25 travelers represents

5% of the total. We can calculate the total number of European travelers to be: $5\% \times x = 25 \rightarrow x = \frac{25}{.05} \rightarrow x = 500$ total travelers. We then look at the *Age of Travelers* chart and see that 30-39 year olds accounted for 15% of the travel that year. Multiply 15% by 500 to find our answer, which is **75 travelers**.

4. We know there are a total of 500 travelers. We can look at the number of total travelers who went to Japan, which is 30% of the 500, or **150 people who went to Japan**. Out of that 150, we can guess that 20% (the percentage of 40-49 year old travelers, from the bottom chart) of them were between the age of 40-49. $20\% \times 150 = 30$ **people**.

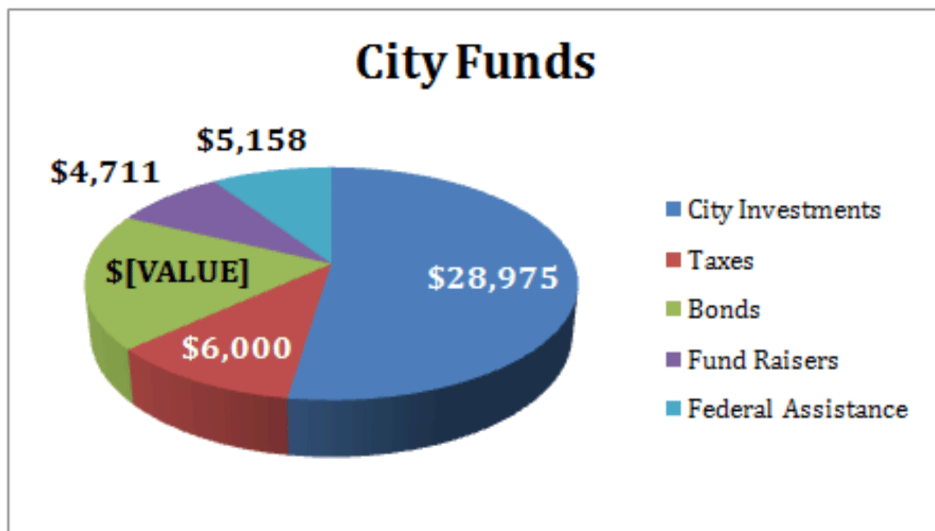
Practice

The following pie-chart shows the percentage distribution of the expenses incurred in publishing a statistics math book. Study the Pie Chart and answer the questions below.



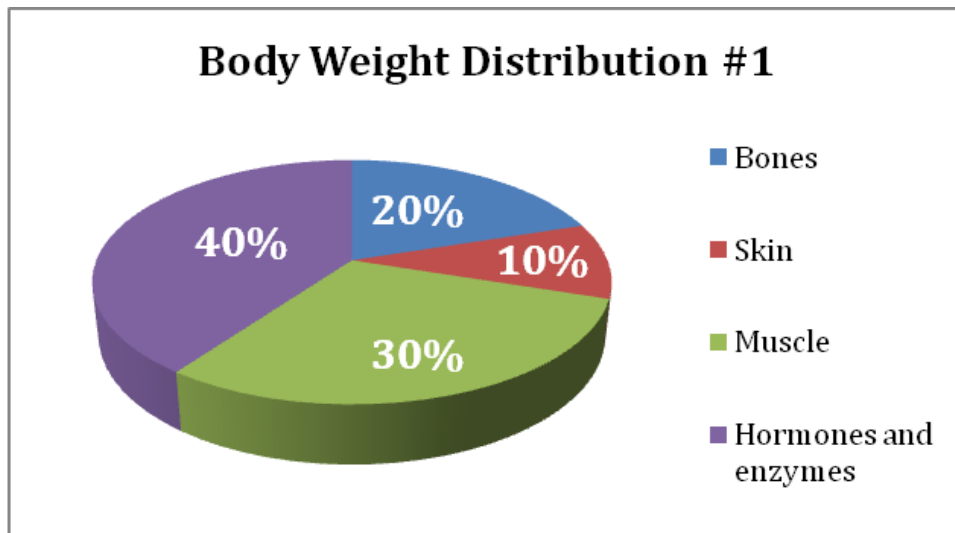
- If the publisher paid \$30,600 for printing, then how much did the publisher pay the author to write the book?
- What is the measurement of the angle of the sector that corresponds to the authoring expense?
- If the book is sold at 20% above the actual cost per book, and the marked price is \$180.00, what is the cost of the paper used in a single copy of the book?
- If 5500 copies of the book are published, and shipping on them amounts to \$82,500, what should the selling price of the book be so that the publisher can earn a profit of 25%?
- Authorship of the book is less than the printing cost by what percentage?
- If the difference between two of the expenditures in the chart are represented by 54 degrees, then which combination(s) of two expenditures could they be?
- If the cost of paper is \$56,250 for printing an edition of the book, then what was the advertising cost?
- Identify two expenditures that together have a central angle of 108 degrees.

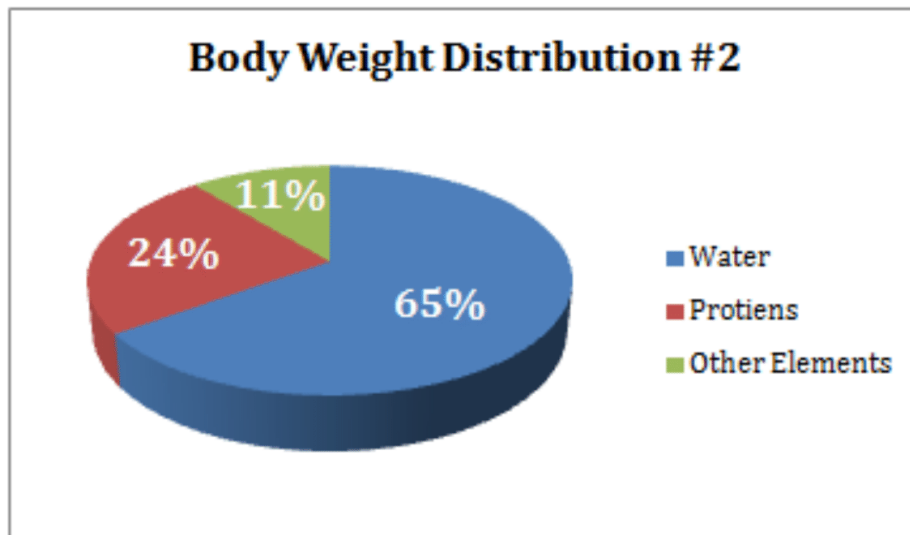
The following pie chart shows the sources of funds that will be collected by Downtown Development to beautify the city. Use the chart to answer the questions that follow.



9. Nearly 20% of the funding will come from which source? 55290
10. If the city could only pull \$9,725 from bonds, by what percent should it increase the amount it borrows against its investments to make up the difference?
11. If the bonds are managed by a third party that collects a 10% commission, how much would the bonds have actually been worth, before being added to the budget represented in the chart above?
12. What is the degree measure corresponding to the Taxes collected?
13. What is the approximate ratio of the funds to be collected through Bonds and through City Investments?

The following charts give information about how weight is distributed throughout the human body, according to different components. Answer the questions below based on the pie charts.





14. What percentage of proteins is equivalent to the weight of skin?
15. How much of the human body is neither made of bones or skin?
16. What is the ratio of the weight of proteins in the muscles to that of the weight of proteins in the bones?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 4.15.

Students were given a quick review of data collection, and were taught about relative and cumulative frequencies before the bulk of the chapter detailed the process of constructing and interpreting various common data visualizations. Histograms, box plots, stem-and-leaf diagrams, scatter plots, line graphs, pie (circle) charts and frequency polygons were each separately taught.

After completing the lessons in this chapter, students should feel confident in their ability to create charts of specific types to display data and to read data from charts to draw conclusions.

4.16 References

1. reynermmedia. <https://www.flickr.com/photos/89228431@N06/11322953266> .
2. ~Pawsitive~Candie_N. <https://www.flickr.com/photos/scjn/4269256333> .
3. Paul Reynolds. <https://www.flickr.com/photos/bigtallguy/139143816> .
4. flash.pro. <https://www.flickr.com/photos/flashpro/4156535452> .
5. Daniel Oines. <https://www.flickr.com/photos/dno1967b/5543841808> .
6. . . CC BY-NC-SA
7. CK-12 Foundation. . CCSA
8. CK-12 Foundation. . CCSA
9. Marc Berry Reid. <https://www.flickr.com/photos/marcberryphotos/2983082444> .
10. CK-12 Foundation. . CCSA
11. CK-12 Foundation. . CCSA
12. CK-12 Foundation. . CCSA
13. CK-12 Foundation. . CCSA
14. . . CC BY-NC-SA
15. . . CC BY-NC-SA
16. . . CC BY-NC-SA
17. . . CC BY-NC-SA
18. . . CC BY-NC-SA
19. CK-12 Foundation. . CCSA
20. CK-12 Foundation. . CCSA
21. CK-12 Foundation. . CCSA
22. CK-12 Foundation. . CCSA
23. CK-12 Foundation. . CCSA
24. CK-12 Foundation. . CCSA
25. CK-12 Foundation. . CCSA
26. CK-12 Foundation. . CCSA
27. CK-12 Foundation. . CCSA
28. CK-12 Foundation. . CCSA
29. CK-12 Foundation. . CCSA
30. . . CC BY-NC-SA
31. CK-12 Foundation. . CCSA
32. CK-12 Foundation. . CCSA
33. CK-12 Foundation. . CCSA
34. CK-12 Foundation. . CCSA
35. CK-12 Foundation. . CCSA
36. CK-12 Foundation. . CCSA
37. CK-12 Foundation. . CCSA
38. CK-12 Foundation. . CCSA
39. CK-12 Foundation. . CCSA
40. . . CC BY-NC-SA
41. CK-12 Foundation. . CCSA
42. . . CC BY-NC-SA
43. CK-12 Foundation. . CCSA
44. CK-12 Foundation. . CCSA
45. CK-12 Foundation. . CCSA

46. CK-12 Foundation. . CCSA
47. CK-12 Foundation. . CCSA
48. CK-12 Foundation. . CCSA
49. CK-12 Foundation. . CCSA
50. CK-12 Foundation. . CCSA
51. CK-12 Foundation. . CCSA
52. CK-12 Foundation. . CCSA
53. CK-12 Foundation. . CCSA
54. CK-12 Foundation. . CCSA
55. CK-12 Foundation. . CCSA
56. CK-12 Foundation. . CCSA
57. CK-12 Foundation. . CCSA
58. CK-12 Foundation. . CCSA
59. CK-12 Foundation. . CCSA
60. CK-12 Foundation. . CCSA
61. CK-12 Foundation. . CCSA
62. CK-12 Foundation. . CCSA
63. CK-12 Foundation. . CCSA
64. CK-12 Foundation. . CCSA
65. CK-12 Foundation. . CCSA
66. CK-12 Foundation. . CCSA
67. CK-12 Foundation. . CCSA
68. CK-12 Foundation. . CCSA
69. CK-12 Foundation. . CCSA
70. CK-12 Foundation. . CCSA
71. CK-12 Foundation. . CCSA
72. CK-12 Foundation. . CCSA
73. CK-12 Foundation. . CCSA
74. CK-12 Foundation. . CCSA
75. CK-12 Foundation. . CCSA
76. CK-12 Foundation. . CCSA
77. . . CC-BY-NC-SA
78. . . CC BY-NC-SA
79. . . CC BY-NC-SA
80. . . CC-BY-NC-SA
81. . . CC-BY-NC-SA
82. . . CC-BY-NC-SA
83. . . CC BY-NC-SA
84. . . CC-BY-NC-SA
85. . . CC-BY-NC-SA
86. . . CC-BY-NC-SA
87. . . CC-BY-NC-SA
88. . . CC-BY-NC-SA
89. . . CC BY-NC-SA
90. . . CC-BY-NC-SA
91. . . CC BY-NC-SA
92. . . CC BY-NC-SA
93. . . CC-BY-NC-SA
94. . . CC-BY-NC-SA
95. . . CC-BY-NC-SA
96. . . CC-BY-NC-SA

97. . . CC-BY-NC-SA
98. . . CC-BY-NC-SA
99. CK-12 Foundation. . CCSA
100. CK-12 Foundation. . CCSA
101. . . CC BY-NC-SA
102. CK-12 Foundation. . CCSA
103. CK-12 Foundation. . CCSA
104. . . CC BY-NC-SA
105. CK-12 Foundation. . CCSA
106. CK-12 Foundation. . CCSA
107. CK-12 Foundation. . CCSA
108. . . CC BY-NC-SA
109. CK-12 Foundation. . CCSA
110. CK-12 Foundation. . CCSA
111. CK-12 Foundation. . CCSA
112. CK-12 Foundation. . CCSA
113. CK-12 Foundation. . CCSA
114. CK-12 Foundation. . CCSA
115. . . CC BY-NC-SA
116. CK-12 Foundation. . CCSA

CHAPTER 5**Central Tendency****Chapter Outline**

- 5.1 ARITHMETIC MEAN**
 - 5.2 GEOMETRIC MEAN**
 - 5.3 HARMONIC MEAN**
 - 5.4 MEDIAN - PROBABILITY AND STATISTICS**
 - 5.5 MODE - PROBABILITY AND STATISTICS**
 - 5.6 CALCULATING VARIANCE**
 - 5.7 VARIANCE PRACTICE**
 - 5.8 CALCULATING STANDARD DEVIATION**
 - 5.9 COEFFICIENT OF VARIATION**
 - 5.10 REFERENCES**
-

In this chapter, you will learn to understand, calculate, and evaluate various measures of central tendency. Central tendencies are measures of a typical or average central value in a set. Later in the chapter, we will also discuss variability, which is the degree to which data varies from the central value.

5.1 Arithmetic Mean

Objective

In this lesson we will discuss the most common single measure of central tendency, the *arithmetic mean*, often called the *average*.

Concept

Suppose you just purchased your first car, and you are wondering what kind of gas mileage it gets. If you collect the data in the table below from your first 4 tanks of gas, how could you calculate your average gas mileage? Is the mean gas mileage different if it is calculated as the average of the mileages from each tank instead of calculated using the total miles and total gallons?

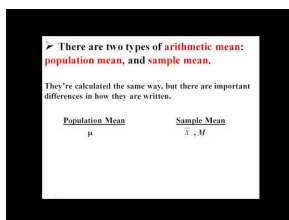


TABLE 5.1:

Gallons of Gas	18.7	17.8	16.3	19.1
Miles Driven	363	347	320	402

After the lesson, we'll return to this question to review the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63711>

<http://youtu.be/GzXIIIfiZUqg> statslectures - Arithmetic Mean for Samples and Populations

Guidance

Although there are actually three different relatively common types of mean, or average, the *arithmetic mean* is far and away the most common method of calculating the “middle” value. The other two, the *geometric mean* and *harmonic mean*, are the topics of other lessons.

The arithmetic mean is a very important calculation to master, because it is the first step in a multitude of more complex calculations. Fortunately, the mean is relatively easy to calculate:

1. Calculate the sum of all of the values in your set
2. Divide the sum by the *number* or *count* of values in the set
3. The quotient of the sum of the values divided by the number of values is the mean

In statistics, you will work with the arithmetic means of different groupings of units. Most importantly, you will need to know the difference between the *population mean* and the *sample mean*. The population mean is the mean of all of the members of an entire population, and the sample mean is the mean only of the members of a *sample* or *subset* of a population.

Example A

Calculate the arithmetic mean of the set.

26	5											
27	5	5										
28	0	0	1	1	1	2	4	5	5	8		

Solution: This set is organized in stem and leaf format, and contains 13 values between 265 and 288.

To find the mean, first calculate the sum of the values in the set:

$$265 + 275 + 275 + 280 + 280 + 281 + 281 + 281 + 282 + 284 + 285 + 285 + 288 = 3,642$$

Since there are 13 values in the set, the mean is:

$$\frac{3,642}{13} = 280.154$$

Rounded to the nearest whole, the arithmetic mean is 280

Example B

Find the arithmetic means of the sets:

- a. 261, 286, 257, 284, 258, 281, 258, 285, 267, 275, 258, 284, 260, 285, 258, 284
- b. 43, 44, 45, 45, 46, 46, 47, 47, 47, 48, 49, 49, 49, 49, 49, 49, 49, 49, 49, 49, 49, 50, 50, 50, 51
- c. 14, 28, 42, 56, 23, 67, 12, 45, 93, 52

Solution: For each set, find the sum of the values in the set, and divide by the number of values:

a) The sum of the values is 4,341. There are 16 values.

$$\frac{4,341}{16} = 271.31$$

b) The sum of the values is 1,198. There are 25 values.

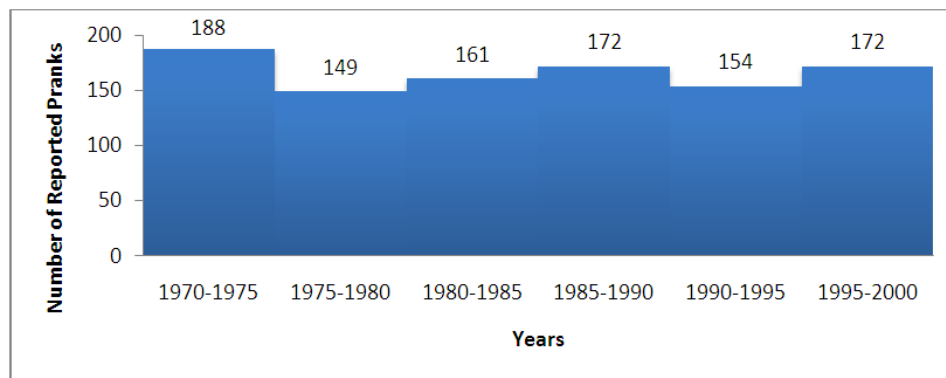
$$\frac{1,198}{25} = 47.92$$

c) The sum of the values is 432. There are 10 values.

$$\frac{432}{10} = 43.2$$

Example C

Using the data from the chart, find: a) the mean number of prank calls per five-year interval, and b) the mean number of reported prank calls per year.



Solution:

a) First, find the sum of the number of reported calls:

$$188 + 149 + 161 + 172 + 154 + 172 = 996 \text{ calls}$$

The chart indicates that there are six five-year intervals with recorded values. So, dividing the sum by the number of values, we get:

$$\frac{996}{6} = 166 \text{ calls per five year interval}$$

Note that the value “166” is not the actual count of any of the recorded intervals. The arithmetic mean obviously does *not* need to be a member of the actual set.

b) The chart spans the years 1970 through 1999 (the end of 1999 is the beginning of 2000), so there are a total of 29 years recorded. Dividing the total number of calls by the number of years yields:

$$\frac{996}{29} \approx 34 \text{ (rounded) calls per year}$$

Concept Problem Revisited

- a) If you collect the data in the table below from your first 4 tanks of gas, how could you calculate your average gas mileage?
- b) Is the mean gas mileage different if calculated as the average of the mileages from each tank instead of the total miles and total gallons?

TABLE 5.2:

Gallons of Gas	18.7	17.8	16.3	19.1
Miles Driven	363	347	320	402

- a) Mileage is measured as miles travelled per gallon, so we can simply total the number of miles and divide by total gallons:

$$\frac{1,432 \text{ miles}}{71.9 \text{ gallons}} = 19.92 \text{ avg miles per gallon}$$

- b) If we calculate the mean mileage for each tank:

$$\frac{363}{18.7} = 19.41 \quad \frac{347}{17.8} = 19.49 \quad \frac{320}{16.3} = 19.63 \quad \frac{402}{19.1} = 21.05$$

and find the mean of the mileages:

$$\frac{19.41 + 19.49 + 19.63 + 21.05}{4} = \frac{79.58}{4} = 19.90 \text{ miles per gallon}$$

we see that the overall mileage appears slightly lower this way.

Can you figure out why this might be?

Vocabulary

The **arithmetic mean** is more commonly known as the average, and is calculated by dividing the sum of a set of values by the count of the number of values.

The **geometric mean** is another type of 'middle value', and is calculated by multiplying the values of each member of a set together and taking the n^{th} root of the product, where n is the number of values in the set.

A **harmonic mean** is calculated by dividing the number of values in the set by the sum of the inverses of the values in the set.

Guided Practice

1. Find the mean of the following data:

237, 258, 232, 232, 241, 238, 233, 245, 242, 243, 237, 232, 241, 242, 233

2. Find the mean of the following data:

22, 30, 24, 42, 24, 42, 18, 32, 24, 48, 24, 45, 28, 46, 22, 42

3. Find the mean of the following data:

19	9
20	0 5
21	6 9
22	0 1 1 2 2 4 7 7 8 9

Solutions:

- To find the mean, add all the numbers and divide by the count of numbers you started with.

$$237 + 258 + 232 + 232 + 241 + 238 + 233 + 245 + 242 + 243 + 237 + 232 + 241 + 242 + 233 = 3586$$

$$= \frac{3586}{15} = 239.07 \text{ (rounded)}$$

- To find the mean, add all the numbers and divide by the quantity of numbers you started with.

$$22 + 30 + 24 + 42 + 24 + 42 + 18 + 32 + 24 + 48 + 24 + 45 + 28 + 46 + 22 + 42 = 513$$

$$= \frac{513}{16} = 32.06 \text{ (rounded)}$$

- To find the mean, add all the numbers and divide by the quantity of numbers you started with.

$$199 + 200 + 205 + 216 + 219 + 220 + 221 + 221 + 222 + 222 + 224 + 227 + 227 + 228 + 229 = 3280$$

$$= \frac{3280}{15} = 218.67$$

Practice

Airon collected a sample of the five cheapest (out of a population of 10) different stores' prices on the new laptop he wanted. His sample was \$285, \$290, \$300, \$305, and \$315. The other five stores' prices were \$345, \$380, \$435, \$480, and \$500.

- What was the arithmetic sample mean of laptop prices among the sample Airon collected?
- What was the arithmetic mean of laptops that Airon did not add to his sample?
- What was the arithmetic population mean of laptop prices?

Questions 4 - 6 refer to the numbers: 4, 7, 2, 5, 7, 8, 4, 9, 5, 6, 2, and 12

- What is the arithmetic mean of the numbers?
- What is the arithmetic mean of the odd numbers?
- What is the arithmetic mean of the even numbers?

Chet's class has 7 boys, ages 15, 16, 14, 15, 17, 14, and 17 years, and 6 girls, ages 15, 16, 14, 14, 15, and 16 years.

- What is the arithmetic population mean of Chet's class?
- What is the arithmetic mean of girls in Chet's class?
- What is the arithmetic mean of boys in Chet's class?

10. What is the arithmetic sample mean of a sample of students consisting of the youngest and oldest boys and youngest and oldest girls (a total of four students)?

Donna works in the customer service department at a large corporation. The employees in her department earn the following incomes: \$24,560, \$32,540, \$29,540, \$39,490, \$42,100, \$27,000, and \$35,750.

11. What is the mean income of the department?

12. What is the mean income above \$30,000?

13. What is the mean income below \$30,000?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 5.1.

5.2 Geometric Mean

Objective

Here you will learn how to calculate the geometric mean of a set of data, and will review some appropriate uses of the geometric mean.

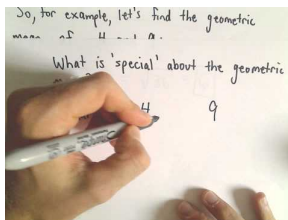
Concept

Suppose you were assigned to help evaluate the scoring from surveys given to students to choose the prom king at your high school prom. The survey asked the students to score the nominees on a five-point scale for each of four categories, on a twenty-five point scale for each of three categories, and on a thirty-five point scale for the last 4 categories.



How could you average the total scores for each nominee without the scores of the higher point categories overshadowing the scores of the lower-point ones?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63709>

http://youtu.be/_UdGUULKN-E partrickJMT - The Geometric Mean

Guidance

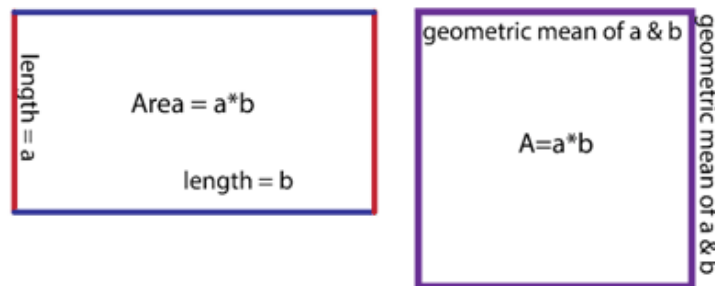
The **geometric mean** (also known as the **mean proportional**) is a method of finding the 'middle' value in a set that contains some values that are intrinsically more influential than others. The geometric mean takes into account the differences in proportion between values in different ranges.

To calculate the geometric mean of a set of data:

- Multiply the value of each member of the set by the next, as in $x_1 \times x_2 \times x_3 \times x_4$, etc.
- Find the n^{th} root of the product of the set values, where n is the number of values in the set
- The n^{th} root of the product of the set values is the **geometric mean** of the set

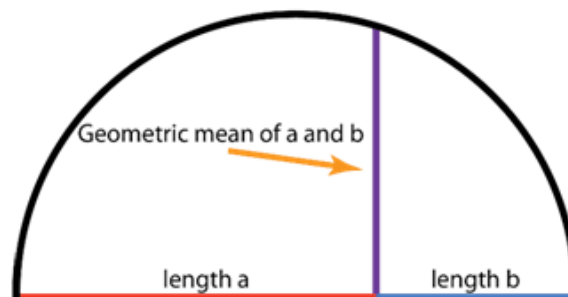
The mean proportional can also be described in a couple of ways using geometry:

- If you have a rectangle with side lengths a and b , then the side length of a square with area equal to the rectangle is the mean proportional of a and b .



Alternatively:

- Construct a semicircle with a diameter length of $a + b$ and a perpendicular line from the diameter to the semicircle located at the intersection of a and b . The length of the perpendicular line is the geometric mean of a and b .



Example A

What is the mean proportional of x ?

$$x = \{8, 12, 14, 6, 9, 15, 22, 7\}$$

Solution: First find the product of the members of the set:

$$8 \times 12 \times 14 \times 6 \times 9 \times 15 \times 22 \times 7 = 167,650,560$$

Now take the 8th root of the product (since there are 8 members in the set):

$$\sqrt[8]{167,650,560} \approx 10.667$$

Example B

Ken has a dog kennel in back of his house. The dimensions of the kennel are 9 feet by 13 feet. What would be the side lengths of a square kennel enclosing the same area? What is another name for this value?

Solution: First find the area of the original kennel:

$$9 \text{ ft} \times 13 \text{ ft} = 117 \text{ ft}^2$$

Now find the lengths of the sides of the square:

$$\sqrt{117 \text{ ft}^2} = 10.82 \text{ ft}$$

10.82 is the geometric mean of 9 and 13.

Example C

Assume that \$5000 was invested at a starting interest rate of 7%, and the rate increased by 2% each year for years 2, 3, 4, and 5 then decreased by 3% for year 6. What would the average rate of return be for the whole period?

Solution: It may seem that you could simply calculate the arithmetic mean of the interest rates to find the average yearly return. However, the arithmetic mean of the 6 yearly rates would only describe the average of the stated percentages each year. The complication is that a 7% interest rate in the first year would only result in 107% of the initial value, whereas a 7% rate in the 6th year would yield 107% of the initial investment *plus* 107% of all of the interest from the prior years! Therefore, a different 'middle number' is needed.

The geometric mean of the interest rates would provide the correct average yearly rate. We need to find the product of all of the yearly rates, and then take the 6th root (since there are 6 rates in this problem) of that product. The correct calculation looks like this:

$$\sqrt[6]{1.07 \times 1.09 \times 1.11 \times 1.13 \times 1.15 \times 1.12} = \sqrt[6]{1.88} = 1.11$$

\therefore The average return is 11% per year

Concept Problem Revisited

Suppose you were assigned to help evaluate the scoring from surveys given to students to choose the prom king at your high school prom. The survey asked the students to score the nominees on a five-point scale for each of four categories, on a twenty-five point scale for each of three categories, and on a thirty-five point scale for the last 4 categories.

How could you average the overall scores for each nominee without the disproportionate weighting that would occur in favor of the higher point categories if you used an arithmetic mean?

By calculating the geometric mean of the scores earned in all 11 categories, you could identify an average score for each contestant that was proportionately *weighted* based on category value.

Vocabulary

A *weighted* value or set of values takes into account varying levels of importance among members of the set.

The *geometric mean* or *mean proportional* is a method of identifying the central value in a set that accounts for weighted values.

Guided Practice

1. Construct a semicircular representation of the mean proportional of the values 5 and 9.
2. Find the average rate of return on an investment that earns 6.04%, 6.89%, 7.22%, 6.92%, and 7.43% over successive years.
3. Construct a visual representation of the geometric mean of the numbers 23 and 38, using quadrilaterals.
4. Find a) the geometric mean of y and b) the arithmetic mean of y .

$$y = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

5. Find the mean proportional of the values: 12.34, 14.52, 16.82, 13.29, and 13.91

Solutions:

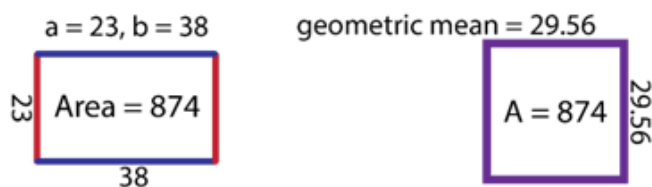
1. Use the geometric method described above: length $a = 9$, length $b = 5$, place them end-to-end and they become the diameter of a semi-circle. Construct a perpendicular from the intersection of a and b to the circumference of the semi-circle, and the length of that perpendicular is the geometric mean. In this case, the geometric mean is 6.8 cm.



2. This problem is just like Example C, just find the geometric mean of the interest rates:

$$\sqrt[5]{6.04 \times 6.89 \times 7.22 \times 6.92 \times 7.43} = \sqrt[5]{15,448.569} = 6.88\%$$

3. This is just like the quadrilateral method described in the “Guidance” section: If you create a rectangle with side lengths equal to the numbers 23 and 38, then the *side lengths* of a square with the same area as the rectangle will be the geometric mean of the two numbers:



Therefore, the geometric mean of 23 and 38 is **29.56**, just like the side length of the square.

4. a) The geometric mean is the 9th root of the product of the values (since there are 9 values):

$$\sqrt[9]{1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8 \times 9} = \sqrt[9]{362,880} = 4.147$$

b) The arithmetic root is the sum of the values divided by the count of the values:

$$\frac{1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9}{9} = 5$$

5. There are five values here, so we need to find the 5th root of the product:

$$\sqrt[5]{12.34 \times 14.52 \times 16.82 \times 13.29 \times 13.91} = \sqrt[5]{557,134.28} = 14.1$$

Practice

For questions 1-10, find the geometric mean of the numbers.

- {5, 8, 9, 7, 6, 8, 3, 4}
- {12, 16, 18, 13, 14, 16}
- {23, 24, 26, 28}
- {33, 37, 38, 36, 35, 35, 36, 34}
- {46, 49, 48, 41, 42, 43, 44}
- {51, 52, 53, 54, 55, 56, 57, 58, 59}
- {156.21, 245.25, 184.64, 222, 32, 218.94, 134.88}
- {554, 564, 585, 525, 534, 500}
- {0.0021, 0.0034, 0.081, 0.009, 0.01, 0.258}
- $\left\{\frac{5}{8}, \frac{13}{21}, \frac{7}{9}, \frac{3}{5}, \frac{11}{23}\right\}$
- Construct a semicircular representation of the mean proportional of the values 12 and 19.
- Construct a visual representation of the geometric mean of the numbers 5 and 8, using quadrilaterals.
- Find the average rate of return on an investment that earns 5.02%, 4.11%, 4.18%, 3.72%, and 3.53% over successive years.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 5.2.

5.3 Harmonic Mean

Objective

Here you will learn how and when to calculate the harmonic mean of a set.

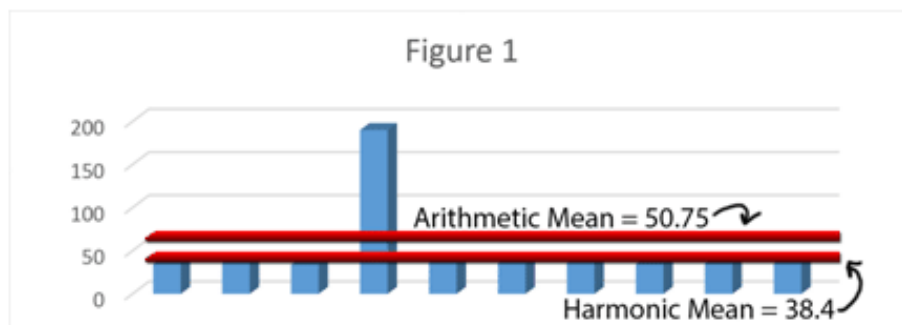
Concept

Can you define the difference between the arithmetic, geometric, and harmonic means? Can you think of a situation when each might be appropriately used?

By the end of this lesson you will!

Guidance

The *harmonic mean* is a 'middle number' you can use when you are worried that the arithmetic mean of your set would be skewed by a few very large values (as in figure 1), and/or when particularly small values are disproportionately important. The harmonic mean is also useful for calculations of average rates, or finding other sorts of *weighted averages*.



The harmonic mean can be a much better average number than the arithmetic mean.

Of the three mean value calculations we discuss in this chapter, the harmonic mean is always the lowest value if calculated using the same data.

Calculating the harmonic mean is a bit more complex than the geometric mean, but it quickly gets easier with practice:

- Count the *number* of values in your data set. This number becomes your numerator.
- Calculate the sum of the inverses of the data values, this sum becomes your denominator.
- Divide the numerator by the denominator, the resulting quotient is the harmonic mean of the values.

$$\frac{\text{number of values}}{\left(\frac{1}{\text{value } 1}\right) + \left(\frac{1}{\text{value } 2}\right) + \left(\frac{1}{\text{value } 3}\right) \cdots \left(\frac{1}{\text{value } n}\right)}$$

For example, to find the harmonic mean of the set {1, 2, 3}:

- First count the values, there are 3. The number 3 becomes the numerator of your final calculation.
- Sum the inverses of the values: $\frac{1}{1} + \frac{1}{2} + \frac{1}{3} = \frac{6}{6} + \frac{3}{6} + \frac{2}{6} = \frac{11}{6}$. $\frac{11}{6}$ becomes the denominator of your final calculation.
- **Final calculation:**
 - Divide the numerator (3) by the denominator ($\frac{11}{6}$):

$$\frac{3}{\frac{11}{6}} = \frac{18}{11} = 1.636$$

1.636 is the harmonic mean of {1, 2, 3}

This type of harmonic mean assumes un-weighted values, such as the *same* distance travelled at different rates. To calculate the harmonic mean of differently weighted values (***the weighted harmonic mean***), such as when calculating the average speed of a trip made of segments with *different* distances *and* different rates, (see Example C), the process is nearly the same:

- First calculate the *sum of the weights* of your values (as opposed to just *counting* the values, as with the basic harmonic mean), this number becomes your numerator.
- Second, find the *sum of the values* (weights divided by amounts - NOT the inverses); this number becomes your denominator.
- Divide the numerator by the denominator, the resulting quotient is the ***weighted harmonic mean*** of the values.

As a formula, this looks like: $WHM = \frac{(\sum w)}{(\sum \frac{w}{a})}$ Where w is the weight (the numerator if your values are fractional) and a is the amount (the denominator if your values are fractional).

Example A

Calculate the harmonic mean of x .

$$x = \{4, 15, 17, 5, 22\}$$

Solution: Follow the steps above for the un-weighted harmonic mean:

- The number of values is 6, this is the numerator
- The sum of the inverses is $\frac{1}{4} + \frac{1}{15} + \frac{1}{17} + \frac{1}{5} + \frac{1}{22} = .621$, this is the denominator
- The harmonic mean is $\frac{6}{.621} = 9.66$

Example B

Kiera decided to take a quick tour around town on her bike, if the information below describes the speeds she travelled over each equal-length segment, what was her average speed for the trip?

TABLE 5.3:

Segment #1	8 mph	3 mi
Segment #2	12 mph	3 mi
Segment #3	14 mph	3 mi
Segment #4	7 mph	3 mi

Solution: There are two ways to solve this:

1. We could find the average speed for the whole trip by dividing the entire distance by the entire time:

- a. Sum the Distances: $3 \text{ mi} + 3 \text{ mi} + 3 \text{ mi} + 3 \text{ mi} = 12 \text{ mi}$ **total distance**
- b. Sum the times:
 - i. Segment 1 $\frac{3 \text{ mi}}{8 \text{ mph}} = .375 \text{ hr}$
 - ii. Segment 2 $\frac{3 \text{ mi}}{12 \text{ mph}} = .25 \text{ hr}$
 - iii. Segment 3 $\frac{3 \text{ mi}}{14 \text{ mph}} = .214 \text{ hr}$
 - iv. Segment 4 $\frac{3 \text{ mi}}{7 \text{ mph}} = .429 \text{ hr}$
 - v. **TOTAL TIME** : $.375 + .25 + .214 + .429 = 1.268 \text{ hrs}$ **total time**
- c. **Average rate** : $\frac{\text{total distance}}{\text{total time}} = \frac{12 \text{ mi}}{1.268 \text{ hrs}} = 9.46 \text{ mph}$

2. The other method is to find the **harmonic mean** of the four rates:

$$\text{a. } HM = \frac{4}{\frac{1}{8} + \frac{1}{12} + \frac{1}{14} + \frac{1}{7}} = \frac{4}{.423} = 9.46 \text{ mph}$$

Obviously, the second method is more efficient!

Example C

Use the weighted harmonic mean to find the average rate of a traveler who records the following itinerary:

TABLE 5.4:

1 st Segment	230 km	66 kph
2 nd Segment	310 km	83 kph
3 rd Segment	199 km	72 kph
4 th Segment	210 km	77 kph
5 th Segment	240 km	91 kph

Solution: Because this problem includes segments with varying speeds and varying distances, we will need to use a weighted harmonic mean to calculate the average rate.

Use the formula: $WHM = \frac{(\sum w)}{(\sum \frac{w}{a})}$ Where w (the weight) is the distance travelled at each rate, and a (the amount) is the rate for each segment.

- First, calculate the sum of the weights (distances):

$$230 + 310 + 199 + 210 + 240 = 1189$$

- Next, calculate the sum of the distances over the rates (remember that $\text{time} = \frac{\text{distance}}{\text{rate}}$):

$$\frac{230 \text{ km}}{66 \text{ kph}} + \frac{310 \text{ km}}{83 \text{ kph}} + \frac{199 \text{ km}}{72 \text{ kph}} + \frac{210 \text{ km}}{77 \text{ kph}} + \frac{240 \text{ km}}{91 \text{ kph}} = 3.48 + 3.735 + 2.764 + 2.727 + 2.637 = 15.343$$

- Finally, divide the sum of the weights by the sum of the times.
- The weighted harmonic mean is $\frac{1189}{15.343} = 77.5 \text{ kph}$

Concept Problem Revisited

Can you define the difference between the arithmetic, geometric, and harmonic means? Can you think of a situation when each might be appropriately used?

The arithmetic mean is the simple average of a set of values, the sum of the values divided by the number of values. It is appropriate for situations such as the average rate for a journey composed of segments of equal time and distance.

The geometric rate is the n^{th} root of the product of the values, and is appropriate for situations such as deriving a single value to represent scores from multiple scales. The single value could then be used to compare an overall ranking of scores.

The harmonic mean is the reciprocal of the arithmetic means of the reciprocals of a set of values. It is useful for calculations such as the average rate for a journey composed of segments of differing times *or* distances. A *weighted* harmonic mean can be used to calculate the average rate of a journey composed of differing times *and* distances.

Vocabulary

A **harmonic mean** is defined as the reciprocal of the mean of the reciprocals of the values in a set.

A **weighted harmonic mean** is a harmonic mean of values with varying frequencies or weights.

A **weighted average** is a common term for a central value meant to account for values of different weights in the same set.

Guided Practice

- Find the harmonic mean of the set: $\{13, 17, 22, 29, 39, 45, 50\}$
- Find the weighted harmonic mean of the set: $\{\frac{1}{3}, \frac{2}{5}, \frac{2}{7}, \frac{1}{2}, \frac{3}{11}\}$
- Use the weighted harmonic mean to **find the average rate** of a traveler who records the following itinerary:

TABLE 5.5:

1 st Segment	110 km	23 kph
2 nd Segment	230 km	56 kph
3 rd Segment	259 km	42 kph
4 th Segment	300 km	102 kph
5 th Segment	330 km	71 kph

- Compare the a) basic and b) weighted, harmonic means of the set: $\{\frac{1}{3}, 5, 3, \frac{3}{5}, 6, \frac{5}{11}\}$

Solutions:

- First, count the values: 7

Second, sum the inverses of the values: $\frac{1}{13} + \frac{1}{17} + \frac{1}{22} + \frac{1}{29} + \frac{1}{39} + \frac{1}{45} + \frac{1}{50} = \frac{4497703}{15862275} = .283$

Finally, divide the count by the sum of the inverses: $\frac{7}{.283} = 24.74$

- First, sum the numerators: $1 + 2 + 2 + 1 + 3 = 9$

Second, sum the values: $\frac{1}{3} + \frac{2}{5} + \frac{2}{7} + \frac{1}{2} + \frac{3}{11} = \frac{4,139}{2,310} = 1.792$

Finally, divide the sum of the numerators by the sum of the values: $\frac{9}{1.792} = 5.022$

- First, find the sum of the distances (the weights of the values):

$$110 \text{ km} + 230 \text{ km} + 259 \text{ km} + 300 \text{ km} + 330 \text{ km} = 1,229$$

Next, find the sum of the times $\left(\frac{\text{distance}}{\text{rate}} = \text{time}\right)$:

$$\frac{110 \text{ km}}{23 \text{ kph}} + \frac{230 \text{ km}}{56 \text{ kph}} + \frac{259 \text{ km}}{42 \text{ kph}} + \frac{300 \text{ km}}{102 \text{ kph}} + \frac{330 \text{ km}}{71 \text{ kph}} = 4.78 + 4.10 + 6.17 + 2.94 + 4.65 = 22.64$$

Finally, divide the distance by the time: $\frac{1,229}{22.65} = 54.26 \text{ kph}$

4. a) Basic harmonic mean:

- First, count the values: there are **6**
- Second, sum the inverses of the values: $\frac{3}{1} + \frac{1}{5} + \frac{1}{3} + \frac{5}{3} + \frac{1}{6} + \frac{11}{5} = \frac{227}{30}$ or 7.57
- Finally, divide the count by the sum of the inverses: $\frac{6}{7.57} = 0.792$

b) Weighted harmonic mean:

- First, find the sum of the numerators: $1 + 5 + 3 + 3 + 6 + 5 = 23$
- Second, find the sum of the values: $\frac{1}{3} + \frac{5}{1} + \frac{3}{1} + \frac{3}{5} + \frac{6}{1} + \frac{5}{11} = 15.39$
- Finally, divide the sum of the numerators by the sum of the values: $\frac{23}{15.39} = 1.494$

Practice

For questions 1 - 10, calculate the harmonic mean.

1. $\{5, 8, 9, 7, 6, 8, 3, 4\}$

2. $\{12, 16, 18, 13, 14, 16\}$

3. $\{23, 24, 26, 28\}$

4. $\{33, 37, 38, 36, 35, 35, 36, 34\}$

5. $\left\{\frac{5}{8}, \frac{13}{21}, \frac{7}{9}, \frac{3}{5}, \frac{11}{23}\right\}$

6. $\left\{\frac{6}{7}, \frac{17}{23}, \frac{17}{19}, \frac{6}{11}, \frac{5}{13}\right\}$

7. $\left\{\frac{2}{5}, \frac{5}{7}, 2\frac{3}{4}, \frac{15}{3}, \frac{9}{11}, \frac{14}{17}\right\}$

8. $\left\{3\frac{3}{4}, 2\frac{2}{3}, 5\frac{7}{8}, 3.25, 1\frac{6}{5}\right\}$

9. $\{5, 8, 9, 7, 6, 8, 3, 4\}$

10. $\{12, 16, 18, 13, 14, 16\}$

11. Brian records the following trip data, use it and the basic harmonic mean to find his average rate for the complete trip:

TABLE 5.6:

Segment #1	18 mph	4.2 mi
Segment #2	21 mph	4.2 mi
Segment #3	24 mph	4.2 mi
Segment #4	17 mph	4.2 mi

12. Use the weighted harmonic mean to **find the average rate** of a traveler who records the following itinerary:

TABLE 5.7:

1 st Segment	21.1 km	23 kph
2 nd Segment	32.0 km	15.6 kph

TABLE 5.7: (continued)

3 rd Segment	29.5 km	14.2 kph
4 th Segment	30.5 km	10.2 kph

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 5.3.

5.4 Median - Probability and Statistics

Objective

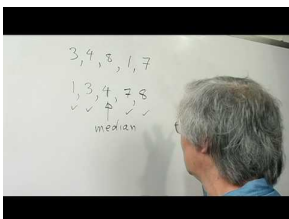
Here you will investigate the calculation and use of the *median* value of a data set.

Concept

What is the median value of a data set? Why would you need to know how to find it if you already know how to find the mean? Is there really any difference? Can you think of a situation where the mean and median are exactly the same number?

In the lesson below we will discuss all of these concepts, and you should have no problem with answering these questions when we review them after example C.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63749>

<http://youtu.be/kZcuGPoqM9c> ehow - Math Definitions: What is a Median Value?

Guidance

The *median* is defined as the value representing the “middle number” of a data set that has been ordered by increasing value, meaning that exactly $\frac{1}{2}$ of the data is greater than the median, and $\frac{1}{2}$ is less. Finding the median value of a set is a very simple process, requiring very little or no actual calculation at all. Regardless, it can be an excellent representation of the *central tendency* of a set. The median is particularly useful when evaluating sets with multiple non-representative outliers, since the median is not sensitive to very large or very small values at the extremes of a data set.

3, 4, 5, 6, 7, 8, 9
M E D I A N

To identify the median:

- First, organized your set in ascending numerical order and count the values:

- Second:
 - If there are an odd number of values, the median is the middle number in the series.
 - If there is an even number of values, the median is the arithmetic mean of the two middle numbers in the series.

Example A

Find the median of z .

$$z = \{1, 10, 3, 8, 5, 6, 7, 4, 9, 2, 13, 12, 15, 19\}$$

Solution: First, organize the numbers in ascending numerical order and count the values:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 15, 19: total of 14 values

Since there is an even number of values (14 of them), the median is the arithmetic mean of the 7th and 8th numbers in the series: $\frac{7+8}{2} = 7.5$

$\therefore 7.5$ is the median of set z

Example B

Find the median of set x .

$$x = \left\{ \frac{1}{8}, \frac{3}{8}, \frac{4}{8}, \frac{19}{21}, \frac{17}{20}, \frac{31}{42}, \frac{2}{3}, \frac{5}{8}, \frac{3}{8} \right\}$$

Solution: First, organize the values in ascending numerical order and count them:

$\frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{2}{3}, \frac{31}{42}, \frac{3}{4}, \frac{4}{5}, \frac{17}{20}, \frac{19}{21}$: Total of 9 values

- Since there are an odd number of values (9 of them), the median is the 5th value: $\frac{31}{42}$.

$\therefore \frac{31}{42}$ is the median of set x

Example C

Brian is looking for an entry-level job at a retail store, and is debating between two options. If he wants to apply for the highest paying job, which of the two employers described below should he choose? Why?



TABLE 5.8:

Employer #1	Employer #2
Company's mean salary: \$48,000	Company's mean salary: \$42,000
Company's median salary \$22,000	Company's median salary: \$28,000

Solution: Since Brian is looking for an entry-level position, he would likely earn more income with Employer #2.

Recall that the mean salary is the average of all the salaries in the company, including any managers or owners on the payroll. Since those positions are undoubtedly higher paid, they would artificially raise the mean income well above that of most employees.

By contrast, the median income is probably closer to the salary of a typical salesman, since there would be many more salesmen than managers or owners.

Concept Problem Revisited

What is the median value of a set? Why would you need to know how to find it if you already know how to find the mean? Is there really any difference? Can you think of a situation where the mean and median are exactly the same number?

The median value is the middle number when the set is organized in ascending numerical order. Depending on the composition of the set, there can actually be quite a large difference between the mean and the median, the mean being much more influenced by outliers. In a *normally distributed* set, the mean, median, and mode are *all* the same.

Vocabulary

The *median* is the middle number in a set that has been organized in ascending numerical order.

Central tendency is a measure of the central or typical value in a set.

A *normally distributed* set has, among other characteristics, an equal mean, median, and mode.

Guided Practice

Identify the median of each set in Q's 1 - 4:

1.

{12, 17, 32, 63, 85, 12, 54, 23, 39}

2.

6	7	7	9									
7	3	4	4	6	8	9						
8	0	1	2	3	3	3	3	4	4	6	7	7
9	0											

3.

{93, 91, 95, 92, 92, 93, 95, 94, 97, 93, 86, 92, 94, 89, 92, 91, 92, 93, 94, 100}

4.

{275, 281, 269, 280, 268, 278, 279, 274, 275, 281, 285, 285, 278, 269, 283, 263, 277, 276, 269, 281, 272, 275, 276}

Solutions:

1. Put the values in numerical order first: 12, 12, 17, 23, 32, 39, 54, 63, 85. There are nine values, an odd count, so we just take the middle number. **32 is the median.**

2. First, put the values in numerical order, remembering that, as a stem plot, this data is listed with the tens place to the left of the vertical, and each separate ones place to the right. That means our values are: 67, 67, 69, 73, 74, 74, 76, 78, 79, 80, 81, 82, 83, 83, 83, 83, 84, 84, 86, 87, 87, 90. Since there are 22 values, an even number, we find the arithmetic mean of the middle two: 81 and 82, to get the median. **The median is 81.5.**

3. Put the values in numerical order: 86, 89, 91, 91, 92, 92, 92, 92, 92, 93, 93, 93, 93, 94, 94, 94, 95, 95, 97, 100. There are 20 numbers, an even count, so we average the middle two, which are both 93. **The median is 93.**

4. Put the values in numerical order: 263, 268, 269, 269, 269, 272, 274, 275, 275, 275, 276, 276, 277, 278, 278, 279, 280, 281, 281, 281, 283, 285, 285. There are 23 numbers, an odd count, so we just take the middle value. **The median is 276.**

Practice

Find the median value for each set in questions 1 - 11.

1. {326, 314, 325, 315, 315, 307, 318, 318, 320, 322, 325, 321, 322, 320, 312, 325, 326, 325}
2. {35, 37, 30, 42, 32, 42, 30, 45, 34, 43, 37, 43, 27, 41, 27, 45, 31, 44, 28, 45}
3. {123, 167, 150, 132, 152, 128, 129, 160, 140, 121}
4. {2120, 3040, 2180, 1892, 923, 9231, 8231}
5. {1, 23, 41, 23, 61, 130, 210, 109, 592, 203, 12}
6. {23.43, 32.52, 23.92, 32.25, 23.43, 29.55, 28.30, 31.54}
7. $\left\{\frac{1}{2}, \frac{4}{9}, \frac{3}{7}, \frac{2}{5}, \frac{21}{23}, \frac{16}{27}, \frac{3}{4}\right\}$
8. $\left\{.57, \frac{23}{100}, .05, \frac{17}{100}, \frac{52}{100}, .42, .44, \frac{45}{100}\right\}$
9. {12, 1.2, .12, 102, 120, .012, .120, 1202}
10. $\left\{123, 12.3, \frac{12}{30}, \frac{123}{120}, \frac{120}{123}, \frac{30}{12}, \frac{1}{123}, 1.23, .123\right\}$
- 11.

2	2	2										
3	3	5	5	6	9	9	9					
4	1	2	2	3	3	4	4	8				
5	2	4	5	5	7	7	7					

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 5.4.

5.5 Mode - Probability and Statistics

Objective

Here you will learn about the *mode*, the measure of central tendency concerned with the value(s) of greatest *frequency* in a data set.

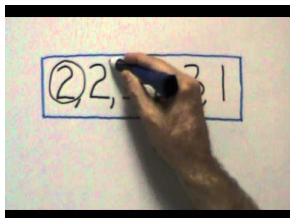
Concept

Noela is working on a homework assignment for her social studies class, and needs to find the 10-year historical period between 1900 and 2000 with the greatest number of recorded hurricanes worldwide. If she uses a data sheet listing all recorded hurricanes, what measure of central tendency would she use to identify the decade with the most hurricanes?



After the lesson below, we'll return to this question to review the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63745>

<http://youtu.be/NZU1omgIZJk> ThirtySecondMath - How to Find the Mode

Guidance

The **mode** is the value(s) in a set that occurs with the greatest frequency. Of the three common measures of central tendency, the mode is the only one that may actually *be* one of the extremes in a set with more than one value. In certain circumstances, the mean of a set with differing values may approach one of the extremes, but only the mode may actually be one of them.

Identification of the mode(s) is simple:

- Organize the set in numerical order (to make it easier to count repeating values) and make note of the frequencies of any repeated values (any values with a frequency greater than 1)
- The value(s) occurring with the greatest frequency are the mode(s)

Because the mode is not directly related to the middle position in the organized series of values, if there are multiple values with the same frequency, do not be concerned if there is a large difference between different modes.

A set with only one mode is called a **unimodal** set. A set with two modes is a **bimodal** set. Technically, there are also **trimodal** sets, but generally any more than two modes are simply referred to as **multimodal**.

Example A

Identify the mode of z .

$$z = \{3, 5, 13, 18, 3, 7, 9, 12, 11, 3, 9, 5, 4, 3, 13\}$$

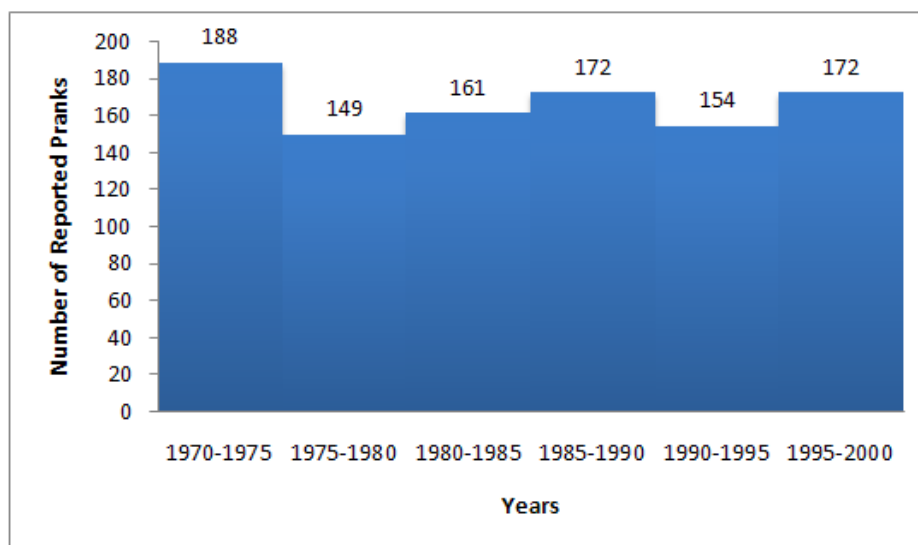
Solution: If we put the values in ascending order, we get:

3, 3, 3, 3, 4, 5, 5, 7, 9, 9, 11, 12, 13, 13, 18

Since 3 is the only value that appears four times, and all other values appear 3 or fewer times, 3 is the mode of z .

Example B

Identify the mode of the set described by the histogram (the mode of the number of reported pranks):



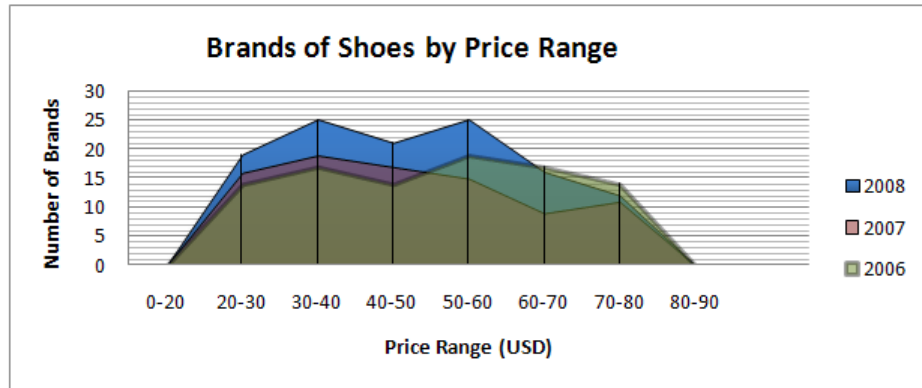
Solution: The intervals 1985-1990 and 1995-2000 are the only two with matching frequencies: 172 reported pranks.

The mode is 172 reported pranks

Example C

Answer the questions using the frequency polygons jointly graphed in the image below.

- Which years demonstrate unimodal distribution?
- Which years demonstrate bimodal distribution?
- Which years demonstrate multimodal distribution OR have no mode?
- Which year(s) have the mode with the greatest frequency?

**Solution:**

- Which years demonstrate unimodal distribution? **2007 has only one mode: 30-40**
- Which years demonstrate bimodal distribution? **2008 has two modes: 30-40 and 50-60**
- Which years demonstrate multimodal distribution OR have no mode? **2006 has 3 modes: 20-30, 40-50, 70-80**
- Which year(s) have the mode with the greatest frequency? **Year 2008 has a mode of 25, the greatest on the chart.**

Concept Problem Revisited

Noela is working on a homework assignment for her social studies class, and needs to find the 10-year historical period between 1900 and 2000 with the greatest number of recorded hurricanes worldwide. If she uses a data sheet listing all recorded hurricanes, what measure of central tendency would she use to identify the decade with the most hurricanes?

Noela needs to organize the data by decade, then identify the *mode*, this will be the decade with the greatest frequency of hurricanes.

Vocabulary

The *mode* is the value occurring with the greatest frequency in a set of data.

A *unimodal* set has only one mode.

A *bimodal* set has two modes.

A *trimodal* set has three modes (may also just be referred to as *multimodal*)

A *multimodal* set has more than two modes.

Guided Practice

Find the mode:

5.6 Calculating Variance

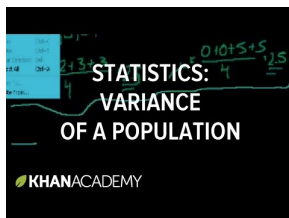
Objective

Here you will learn about *variance*, a measure of the clustering or spread of values around the *mean* of a data set or population.

Concept

If you were told that the mean income at a certain company was \$35,000, you wouldn't really know much about the actual income of the majority of the employees, since there could be a few upper-level managers or owners whose income might *skew* the mean badly. However, if you were also given the *variance* of the incomes, how would that help?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63001>

<http://youtu.be/6JFz11DDyyk> Khan Academy - Variance of a population

Guidance

Variance (commonly denoted σ^2) is a very useful measure of the relative amount of 'scattering' of a given set. In other words, knowing the variance can give you an idea of how closely the values in a set cluster around the mean. The greater the variance, the more the data values in the set are spread out away from the mean.

Variance is an important calculation to become familiar with because, like the arithmetic mean, variance is used in many other more complex statistical evaluations. The calculation of variance is slightly different depending on whether you are working with a population (you do not intend to generalize the results back to a larger group) or a sample (you do intend to use the sample results to predict the results of a larger population). The difference is really only at the end of the process, so let's start with the calculation of the population.

To calculate the variance of a population:

1. First, identify the arithmetic mean of your data by finding the sum of the values and dividing it by the number of values.
2. Next, subtract each value from the mean and record the result. This value is called the *deviation* of each score from the mean.
3. For each value, square the *deviation*.

4. Finally, divide the sum of the squared deviations by the number of values in the set. The resulting quotient is the **variance** (σ^2) of the set.

To calculate the variance of a sample, the only difference is that in step 4, you divide the sum of squared deviations by the number of values in the sample **minus 1**. By dividing the sum of squared deviations by one less than the number of values, you help reduce the effect of outliers in the sample and increase the calculated variance of the sample by a small amount to allow more 'room' for the unknown values in the population.

Example A

Calculate the variance of set x :

$$x = \{12, 7, 6, 3, 10, 5, 18, 15\}$$

Solution: Follow the steps from above to calculate the variance:

- First, calculate the arithmetic mean:

$$\mu = \frac{12 + 7 + 6 + 3 + 10 + 5 + 18 + 15}{8} = 9.5$$

- Subtract each value from the mean to get the deviation of each value, square the deviation of each value:

TABLE 5.9:

Value – Mean = Deviation	Deviation ²
$12 - 9.5 = 2.5$	6.25
$7 - 9.5 = -2.5$	6.25
$6 - 9.5 = -3.5$	12.25
$3 - 9.5 = -6.5$	42.25
$10 - 9.5 = .5$.25
$5 - 9.5 = -4.5$	20.25
$18 - 9.5 = 8.5$	72.25
$15 - 9.5 = 5.5$	30.25
TOTAL (sum of deviation ²):	190.00

- Finally, divide the sum of the squared deviations by the count of values in the data set:

$$\frac{190}{8} = 23.75$$

\therefore The variance of set x is 23.75

Example B

Find the variance of set z :

$$z = \{1, 2, 3, 4, 5, 6, 7, 9\}$$

Solution: Divide the squared deviation of each value from the mean by the total number of values in the set:

$$\begin{aligned}\mu &= \frac{1+2+3+4+5+6+7+9}{8} = 4.625 \\ (1-4.625)^2 + (2-4.625)^2 + (3-4.625)^2 + (4-4.625)^2 \\ &+ (5-4.625)^2 + (6-4.625)^2 + (7-4.625)^2 + (9-4.625)^2 = 49.875 \\ \frac{49.875}{8} &= 6.234 \\ \therefore \text{Variance } (\sigma^2) \text{ of set } z &= 6.234\end{aligned}$$

Example C

Find σ^2 of y :

$$y = \{13, 14, 15, 16, 17, 18, 19, 20, 21\}$$

Solution: Let's do this one differently, using a nifty trick known as the "mean of the squares minus the square of the mean." Start, as before, by finding the arithmetic mean:

$$\mu = \frac{13+14+15+16+17+18+19+20+21}{9} = 17$$

Then, to find the variation, divide the sum of the squares of each value by the number of values (this is the "mean of the squares"), then square the mean we calculated above, 17 (the "square of the mean"), and subtract it from the mean of the squares:

$$\begin{aligned}\sigma^2 &= \frac{13^2+14^2+15^2+16^2+17^2+18^2+19^2+20^2+21^2}{9} - 17^2 = 6.\overline{66} \\ \therefore \sigma^2 \text{ of } y &= 6.\overline{66}\end{aligned}$$

Concept Problem Revisited

If you were told that the mean income at a certain company was \$35,000, you wouldn't really know much about the actual income of the majority of the employees, since there could be a few upper-level managers or owners whose income might **skew** the mean badly. However, if you were also given the variance of the incomes, how would that help?

By learning the variance of the set of incomes, you could get a feel for how representative the \$35,000 figure was of the likely salary of a common employee.

Vocabulary

To **skew** a given set means to cause the trend of data to favor one end or the other.

The **variance** (symbolized by σ^2) of a set is a measure of the average clustering of data points around the mean.

Deviation is a measure of the difference between a given value and the mean.

Guided Practice

1. Find μ and σ^2 of set z .

$$z = \{3.25, 3.5, 2.85, 3.4, 2.95, 3.02, 3.17\}$$

2. If all values of set z , above, were increased by 5, what would the new mean and variance be?
 3. If all values of set z from question #1 were doubled, how would that affect μ and σ^2 ?

Solutions:

1. Let's use the "mean of the squares minus the square of the mean" method:

First find the mean of the set: $\frac{3.25+3.5+2.85+3.4+2.95+3.02+3.17}{7} = 3.16286$

Now divide the sum of each of the values squared by the number of values:

$$\frac{3.25^2+3.5^2+2.85^2+3.4^2+2.95^2+3.02^2+3.17^2}{7} - 10.0036 = 10.0524 - 10.0036 = 0.049 \text{ is the variance.}$$

2. Find the mean of the new set: $\frac{8.25+8.5+7.85+8.4+7.95+8.02+8.17}{7} = 8.16286$

Divide the sum of the values squared by the number of values: $\frac{466.7668}{7} = 66.681$

Subtract the squared mean from the mean of the squares: $66.681 - 66.632 = 0.049$ is the variance.

The variance is the same as before! Does that surprise you? It should, because they actually *aren't* the same, it just appears that way due to rounding. The new set actually has a variance closer to 0.048688, and the original is more accurately 0.04873469. Obviously they are very close, but not exactly the same.

3. The question is what would happen if all of the values were doubled. Do the mean and variance also double? Let's see:

The mean of the new set is $\frac{6.5+7+5.7+6.8+5.9+6.04+6.34}{7} = \frac{44.28}{7} = 6.326$, which is twice the mean of the original set. So far so good.

The "mean of the squares" is $\frac{6.5^2+7^2+5.7^2+6.8^2+5.9^2+6.04^2+6.34^2}{7} = \frac{281.47}{7} = 40.21$, which is *four times* the original mean of the squares, not double after all (which makes sense, given that each doubled value was squared).

Finally, subtract the two values: $40.21 - 6.326^2 = .192$ is the variance. If we compare this to the original: $\frac{.192}{.049} \approx 4$, we can see that doubling the original values quadruples the variance.

Practice

Questions 1-12: find σ^2

- $y = \{4, 50, 63, 2, 82, 99\}$
- Set x is a random sample from a population with 38 members: $x = \{8, 13, 5, 10\}$
- Set z is a random sample from a larger population: $z = \{4, 3, 5, 15, 5\}$
- $y = \{3, 26, 5, 1, 1\}$
- 22, 21, 13, 19, 16, 18
- Sample: 1, 2, 5, 1
- Sample: 10, 6, 3, 4
- 8, 11, 17, 7, 19
- 15, 17, 19, 21, 23, 25, 27, 29
- Sample: 15, 17, 19, 21, 23, 25, 27, 29
- .25, .35, .45, .55, .26, .75
- Find the variance of the data in the table:

TABLE 5.10:

HEIGHTS (rounded to the nearest inch)	FREQUENCY OF STUDENTS
60	35
61	33
62	45
63	4
64	3
65	4
66	7
67	4

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 5.6.

5.7 Variance Practice

Objective

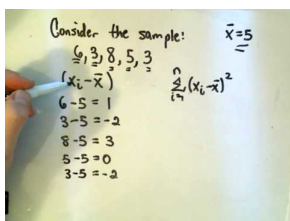
Here you will practice evaluating the variance of data sets.

Concept

Suppose you were given a histogram and asked to find the variance of the data it illustrates? Would you know how? After this lesson, you will understand how to compare visualized data with variance.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63747>

<http://youtu.be/IUixkNvGuWc> patrickJMT - Statistics-Calculating Variance

Guidance

Knowing how to calculate the variance of a set when it is given to you as a list of values is great, but statistical data is often shared and *disseminated* in visual form rather than as raw data. Because of this, it is important to practice evaluating the variance of graphed data as well as *tabular* or raw data so you can actually apply your understanding of variance to real-world statistics.

In general, you will need to:

1. Identify the values of the dependent variable, as these are the values you will be finding the variance of.
2. Sum the values and calculate the arithmetic mean.
3. Subtract the mean from each value to find the deviation and square the deviation
4. Sum the squared deviations and divide the total by the count of values in the data set, the result is the variance.

Example A

Find the μ and σ^2 of the number of students in each classroom at Toni's school:

TABLE 5.11:

Classroom	Number of Students
A	6
B	5
C	9
D	13
E	12
F	16
G	14

Solution: Follow the steps from above to find mean and variance of the students:

1. The frequency of students in each classroom is the dependent variable.
2. There are 7 values, listed in ascending order they are: 5, 6, 9, 12, 13, 14, and 16.
3. The sum of the values is: $5 + 6 + 9 + 12 + 13 + 14 + 16 = 75$, the **mean** is $\frac{75}{7} = 10.714$.
4. The deviances and squared deviances are:

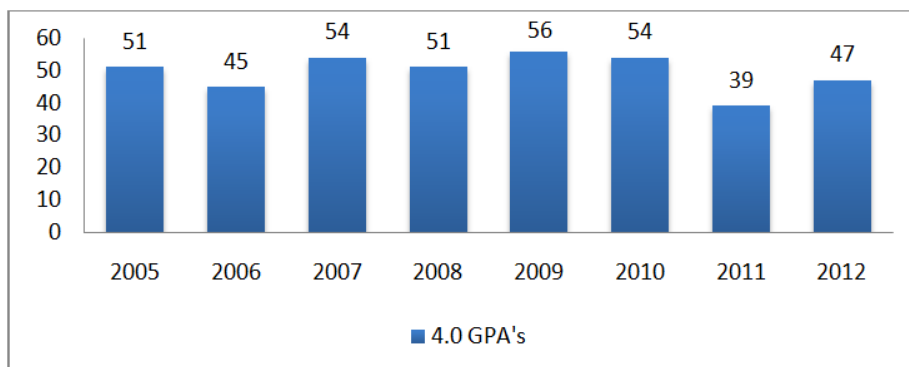
TABLE 5.12:

Value – Mean = Deviance	Deviance ²
$5 - 10.714 = -5.714$	32.65
$6 - 10.714 = -4.714$	22.22
$9 - 10.714 = -1.714$	2.94
$12 - 10.714 = 1.286$	1.654
$13 - 10.714 = 2.286$	5.226
$14 - 10.714 = 3.286$	10.798
$16 - 10.714 = 5.286$	27.942

5. The sum of the squared deviances is 103.43. The **variance** is $\frac{103.43}{7} = 14.776$

Example B

Find the μ and σ^2 of the graphed data.



Solution: Follow the steps outlined above:

1. Most often, the dependent variable is represented by the vertical axis, and this histogram is no exception. The number of 4.0's each year is the dependent variable, while the year is the independent variable.

2. In ascending order, the dependent variable values are:

$$39, 45, 47, 51, 51, 54, 54, 56$$

3. The sum of the values is: $39 + 45 + 47 + 51 + 51 + 54 + 54 + 56 = 397$.

The mean (μ) is: $\frac{397}{8} = 49.625$ which suggests that a year with 50 or more 4.0 GPA's would be considered an above average year.

4. The deviation and squared deviation of each value is:

TABLE 5.13:

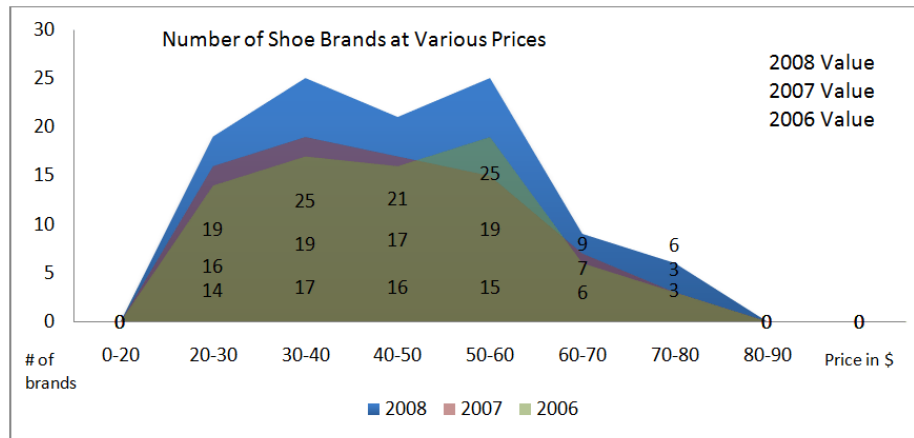
Deviance	Deviance ²
$39 - 49.625 = -10.625$	$(-10.625)^2 = 112.89$
$45 - 49.625 = -4.625$	$(-4.625)^2 = 21.39$
$47 - 49.625 = -2.625$	$(-2.625)^2 = 6.89$
$51 - 49.625 = 1.375$	$(1.375)^2 = 1.89$
$51 - 49.625 = 1.375$	$(-10.625)^2 = 112.89$
$54 - 49.625 = 4.375$	$(4.375)^2 = 19.14$
$54 - 49.625 = 4.375$	$(4.375)^2 = 19.14$
$56 - 49.625 = 6.375$	$(6.375)^2 = 40.64$

5. The sum of the squared deviances is 334.87, making the variance $\frac{334.87}{8} = 41.86$.

$$\therefore \sigma^2 = 41.86$$

Example C

Based on the data in the frequency polygon, which year had the greatest variance in number of shoe brands at various prices, and which had the least variance?



Solution: Each of the three data sets contains 6 values, and the mean of each set is:

- 2008: $Sum: 19 + 25 + 21 + 25 + 9 + 6 = 105$ $Mean: \frac{105}{6} = 17.5$
- 2007: $Sum: 16 + 19 + 17 + 19 + 7 + 3 = 81$ $Mean: \frac{81}{6} = 13.5$
- 2006: $Sum: 14 + 17 + 16 + 15 + 6 + 3 = 71$ $Mean: \frac{71}{6} = 11.83$

The sum of the squared deviances for each year is:

- 2008: $(19 - 17.5)^2 + (25 - 17.5)^2 + (21 - 17.5)^2 + (25 - 17.5)^2 + (9 - 17.5)^2 + (6 - 17.5)^2 = 331.5$
- 2007: $(16 - 13.5)^2 + (19 - 13.5)^2 + (17 - 13.5)^2 + (19 - 13.5)^2 + (7 - 13.5)^2 + (3 - 13.5)^2 = 231.5$
- 2006: $(14 - 11.83)^2 + (17 - 11.83)^2 + (16 - 11.83)^2 + (15 - 11.83)^2 + (6 - 11.83)^2 + (3 - 11.83)^2 = 170.833$

The variance of each set is:

- 2008: $\frac{331.5}{6} = 55.25$
- 2007: $\frac{231.5}{6} = 38.583$
- 2006: $\frac{170.833}{6} = 28.472$

\therefore 2008 has the greatest variance and 2006 has the least variance

Concept Problem Revisited

Could you find the variance of a data set presented as a histogram?

After your practice above, this should no longer be a problem!

Vocabulary

Disseminated data is data that has been given out to others.

Tabular data is data presented in the form of a table, or, depending on the use, it may refer to data points separated by tabs.

Guided Practice

The number of cars of various colors in a parking lot with 5 levels is summarized by the table below, use the data to answer questions 1-4.

TABLE 5.14:

	Red	Yellow	Blue	White
Level 1	11	4	9	14
Level 2	9	3	8	11
Level 3	13	5	10	12
Level 4	14	4	7	9
Level 5	12	6	13	7

1. What is the variance of red cars among the 5 levels?
2. What is the color variance of cars above level 3?
3. What is the variance of blue cars across the 5 levels?
4. If we take a sample of levels by rolling a die and end up with levels 1, 3, and 5, what is the variance of white cars in the sample?

Solutions:

1. The population of red cars across the 5 levels is: 11, 9, 13, 14, and 12.
 - Add the values and divide by five to get the mean of 11.8.
 - Square each of the values and sum the squares: $11^2 + 9^2 + 13^2 + 14^2 + 12^2 = 711$
 - Divide the sum of the squares by the number of values in the set (since this is the whole population of red cars), getting $\frac{711}{5} = 142.2$, and subtract the mean squared ($11.8^2 = 139.24$)
 - The variance of the population of red cars is $142.2 - 139.24 = 2.96$
2. The levels above level 3 include only levels 4 and 5. The total number of red, yellow, blue, and white cars is 26, 10, 20, and 16, respectively.
 - The mean number of cars of each color is $\frac{26+10+20+16}{4} = 18$
 - Square the values and find the sum: $26^2 + 10^2 + 20^2 + 16^2 = 1432$
 - Divide the sum of the squares by the number of values: $\frac{1432}{4} = 358$. Subtract the squared mean ($18^2 = 324$) to get the variance: $358 - 324 = 34$
3. The blue car counts are: 9, 8, 10, 7, and 13
 - The mean number of blue cars is $\frac{47}{5} = 9.4$
 - The sum of the squared values is $9^2 + 8^2 + 10^2 + 7^2 + 13^2 = 463$, divided by the number of levels (5), gives us 92.6
 - Subtract the squared mean ($9.4^2 = 88.36$) to get the variance
 - **The variance is** $92.6 - 88.36 = 4.24$
4. The number of white cars on levels 1, 3, and 5 is 14, 12, and 7.
 - The mean number of white cars in this sample is $\frac{33}{3} = 11$
 - **Since this is a sample, we need to use the individual deviations:** subtract the mean from each value, and square the result of each subtraction, then find the sum: $(14 - 11)^2 + (12 - 11)^2 + (7 - 11)^2 = 26$
 - Divide the sum of the deviations by the number of values *minus 1* (remember, this is a sample!): $\frac{26}{2} = 13$
 - **The sample variance is 13.**

Practice

Find the variance:

1. 365, 400.7, 303, 479, 514.2, 500, 489
2. 7200, 7020, 7165.9, 7000, 7796, 7012, 7016.1
3. 17, 10.3, 30.7, 70, 66, 76, 40, 53
4. 3607, 3600, 3600, 3631, 3600.6
5. 700, 700, 712, 756, 741, 716, 782
6. 3370, 3300.5, 3366, 3306.6, 3310, 3336, 3301.3

Calculate the sample variance:

7. 34.4, 34, 34.7, 34.6, 34, 34.1, 31, 31.3
8. 989.22, 990.6, 992, 996.9, 981.1, 986, 975
9. 10, 16, 10.33, 10.63, 18, 17, 16.36, 10.46
10. 3240, 3260, 3250, 3280, 3280, 3300, 3310, 3270

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 5.7.

5.8 Calculating Standard Deviation

Objective

Here you will learn to calculate the *standard deviation* of sample sets and populations.

Concept

What is standard deviation? How is the standard deviation of a set related to variance? Is the standard deviation of a sample different from that of a population, the way it is with variation?

This lesson details the process of calculating standard deviation, and introduces a few examples of its use. After the lesson we'll review the questions above, using the knowledge we have gained.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63755>

<http://youtu.be/HvDqbzu0i0E> Khan Academy - Statistics-Standard Deviation

Guidance

Standard deviation (σ) is a very common term in statistics and it is not particularly difficult to calculate, particularly if you have already identified the variance of a set. The standard deviation is sort of a “reference difference from the mean” that you can use to evaluate the spread of the data in a set.

For instance, assume the mean of a particular set is 6 and the standard deviation is 4. If you are considering a value of 21, it is probably a very rare occurrence in that set since 21 is nearly 4 standard deviations away from the mean ($4 \times 4 = 16$ and 16 more than the mean would be 22). However, a value of 8 is much more likely, given that it is only $\frac{1}{2}$ of a standard deviation (SD) away from the mean.

Recall from the lesson *Calculating Variance* that calculating the variance of a set involves finding the arithmetic mean, subtracting each data point value from the mean and squaring the result, then finding the sum of the squared results and dividing by either the number of members of the set (population) or the number of members - 1 (sample). See the first part of Example B for a review of finding the variance.

- Once you have the variance of a population, you are practically done finding the SD.
- **To find the SD, simply take the square root of the variance.** That's it!
- One important difference between the variance and the standard deviation is that the units associated with variance are the **square** of the units of the original values, but **the units associated with the standard deviation are the same as the units in the original set.**

We will return to SD in our chapter on “Normal Distribution”, when we will further discuss the uses of the SD of both samples and populations.

Example A

What is the standard deviation of a set with σ^2 of 14.6?

Solution: The standard deviation (σ) is simply the square root of the variance (σ^2). As a formula: $\sqrt{\sigma^2}$.

In this case we have: $\sqrt{14.6} = 3.821 \therefore \sigma = 3.821$

Example B

What is the σ of set x ?

$$x = \{3, 4, 5, 6, 7, 8, 9\}$$

Solution: First find the variation of the set:

- $\mu(\text{mean}) = \frac{3+4+5+6+7+8+9}{7} = 6$
- Deviations and squared deviations:

$$\begin{aligned} - 3 - 6 &= -3 \rightarrow (-3)^2 = 9 \\ - 4 - 6 &= -2 \rightarrow (-2)^2 = 4 \\ - 5 - 6 &= -1 \rightarrow (-1)^2 = 1 \\ - 6 - 6 &= 0 \rightarrow (0)^2 = 0 \\ - 7 - 6 &= +1 \rightarrow (+1)^2 = 1 \\ - 8 - 6 &= +2 \rightarrow (+2)^2 = 4 \\ - 9 - 6 &= -3 \rightarrow (-3)^2 = 9 \end{aligned}$$

- Sum of squared deviations = $9 + 4 + 1 + 0 + 1 + 4 + 9 = 28$
- Variation = $\frac{28}{7} = 4$

$$\therefore \text{Standard deviation of set } x = \sqrt{4} = 2$$

Example C

Katrina wants to use the average scores of the top long jumpers at the 5 schools in her district to predict the average long jumps for top competitors at all schools in her state. Data for her district is below. Find the appropriate variance and standard deviation of the jumps.



School #1	24'10.5"
School #2	24'8.5"
School #3	24'4.25"
School #4	24'1.75"
School #5	23'10.5"

Solution: Since Katrina intends to generalize from her sample data back to the population of jumpers in her state; we need to find the *sample* variance and corresponding *sample* standard deviation.

- Start by finding the mean distance: $\mu = \frac{24'10.5'' + 24'8.5'' + 24'4.25'' + 24'1.75'' + 23'10.5''}{5} = 24'4.7''$
 - As a decimal: 24.39'
- Deviations and squared deviations of each value:
 - $24'10.5'' = 24.875'$: $24.875 - 24.39 = .485 \rightarrow (.485)^2 = .235$
 - $24'8.5'' = 24.71'$: $24.71' - 24.39' = .32 \rightarrow (.32)^2 = .102$
 - $24'4.25'' = 24.35'$: $24.35' - 24.39' = -.04 \rightarrow (-.04)^2 = .001$
 - $24'1.75'' = 24.15'$: $24.15' - 24.39' = -.24 \rightarrow (-.24)^2 = .058$
 - $23'10.5'' = 23.875'$: $23.875 - 24.39' = -.52 \rightarrow (-.52)^2 = .270$
- Sum of squared deviations = $.235 + .102 + .001 + .058 + .270 = .666$
- Sample variance = $\frac{.666}{4} = .167'$ (Remember to divide by $n - 1$, since this is a sample)
- Standard deviation = $\sqrt{.167} = .409' = 4.9''$

Concept Problem Revisited

What is standard deviation? How is the standard deviation of a set related to variance? Is the standard deviation of a sample different from that of a population, the way it is with variation?

By now you should know that standard deviation is a measure of the spread of data, and is calculated as the square root of the variance. Since variance is calculated slightly differently for a sample than for a population, the deviation will differ similarly.

Vocabulary

Standard deviation (σ) is calculated by finding the square root of the variance. The standard deviation acts as a reference unit of difference from the mean in a set of data.

The **variance** (σ^2) is calculated as the sum of the squared differences from the mean, divided by either the number of values (for populations) or the number of values minus one (for samples).

Guided Practice

1. Find the mean (μ), variance (σ^2), and standard deviation (σ) of set z .

$$z = \{12.3, 12.5, 12.2, 11.9, 12.6, 12.35\}$$

2. Find the mean (μ), variance (σ^2), and standard deviation (σ) of set y .

$$y = \{9.1, 10.1, 8.27, 7.9, 8.6, 10.0\}$$

3. Which set has the greater standard deviation, x or y ?

$$x = \{2, 4, 6, 8, 10\} \quad y = \{3, 5, 7, 9, 11, 13\}$$

4. Kevin takes a random sample of ages of students in his class, and gets the following values, what is the sample variance and standard deviation of the set?

$$a = \{15, 16, 16, 15, 17, 17, 18, 16, 17, 16, 18, 18, 15\}$$

Solutions:

1. Let's start by finding the mean, since we will need it to calculate the others:

$$\frac{12.3 + 12.5 + 12.2 + 11.9 + 12.6 + 12.35}{6} = 12.30833$$

The mean (μ) = 12.30833

Now we calculate the deviation of each value from the mean and square it:

$$12.3 - 12.30833 = -0.00833 \rightarrow -0.00833^2 = 0.00007$$

$$12.5 - 12.30833 = 0.19167 \rightarrow 0.19167^2 = 0.03674$$

$$12.2 - 12.30833 = -0.10833 \rightarrow -0.10833^2 = 0.01174$$

$$11.9 - 12.30833 = -0.40833 \rightarrow -0.40833^2 = 0.16673$$

$$12.6 - 12.30833 = 0.29167 \rightarrow 0.29167^2 = 0.08507$$

$$12.35 - 12.30833 = 0.04167 \rightarrow 0.04167^2 = 0.00173$$

Now we sum the squared deviations: $0.00007 + 0.03674 + 0.01174 + 0.16673 + 0.08507 + 0.00173 = 0.30208$, and divide the total by the number of values: $\frac{0.30208}{6} = 0.050347$ to get the variance.

The variance (σ^2) = 0.05347

Finally, to get the standard deviation (σ), just take the square root of σ^2 .

The standard deviation (σ) is $\sqrt{0.05347} = 0.23124$

2. Start by finding μ : $\frac{9.1+10.1+8.27+7.9+8.6+10.0}{6} = 8.995$

Next, find the squared variation from the mean for each value:

- $9.1 - 8.995 = 0.105 \rightarrow 0.105^2 = 0.011025$
- $10.1 - 8.995 = 1.105 \rightarrow 1.105^2 = 1.221025$
- $8.27 - 8.995 = -0.725 \rightarrow 0.525625^2 = 0.276281$
- $7.9 - 8.995 = -1.095 \rightarrow -1.095^2 = 1.199025$
- $8.6 - 8.995 = -0.395 \rightarrow 0.105^2 = 0.156025$
- $10.0 - 8.995 = 1.005 \rightarrow 0.105^2 = 1.010025$

Sum the squared deviations and divide by the number of values to get the variance:

$$\sigma^2 = \frac{3.873406}{6} = 0.6455677$$

Finally, take the square root of the variance to get the standard deviation:

$$\sigma = \sqrt{0.6455677} = .8034723$$

3. Follow the same series of steps to find the standard deviation of each set.

- $x = \{2, 4, 6, 8, 10\} : \mu = 6, \sigma^2 = 10, \sigma = 3.16228$
- $y = \{3, 5, 7, 9, 11, 13\} : \mu = 8, \sigma^2 = 14, \sigma = 3.74166$

Set y has the greater standard deviation

4. There are 13 values, with $\mu = 16.46154$

- The sum of the squared deviations is: 15.2308, divide by 12 (since this is a sample!), to get the *sample variance*: $\frac{15.2308}{12} = 1.26923$
- The square root of the sample variance is the sample standard deviation: $\sqrt{1.26923} = 1.0824$

Practice

Find μ , σ^2 and σ :

1. 265, 280.7, 293, 279, 314.2, 300, 289
2. 7200, 7020, 7165.9, 7100, 7196, 7112, 7116.1
3. 27, 20.3, 30.7, 40, 46, 36, 40, 33

4. 3607, 3600, 3600, 3631, 3600.6

5. 700, 700, 712, 736, 741, 716, 782

6. 3370, 3300.5, 3366, 3306.6, 3310, 3336, 3301.3

Calculate the sample standard deviation:

7. 34.4, 34, 34.7, 34.6, 34, 34.1, 31, 31.3

8. 989.22, 990.6, 992, 996.9, 981.1, 986, 975

9. 10, 16, 10.33, 10.63, 18, 17, 16.36, 10.46

10. 3240, 3260, 3250, 3280, 3280, 3300, 3310, 3270

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 5.8.

5.9 Coefficient of Variation

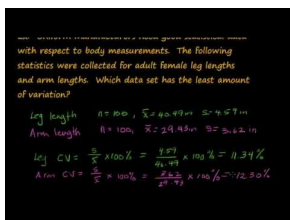
Objective

Here you will learn how to evaluate variance as a percentage of mean. This process allows you to consider how a particular variance relates to the data it describes.

Concept

Suppose you were given three different sets of data, one with a variance of 3.2 and mean of 9.2, another with a variance of 16 and mean of 45, and the third with a variance of 155 and mean of 2100. If you were asked which set was the least centrally clustered, how could you find out?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63753>

<http://youtu.be/iLI81ULCwkk> Kathy Arcangeli - coefficient of variation

Guidance

In a prior lesson, we touched on the idea that variance is calculated as a single value, but that the level of clustering that it represents depends on the mean of the data. One measure that accounts for the differences between means when comparing variance is called the *coefficient of variation*, which is defined as:

$$\frac{\sigma}{\mu} \times 100 = CV\%$$

Where σ = standard deviation, μ = arithmetic mean, and $CV\%$ = coefficient of variation

Recall that σ , the standard deviation, is simply the square root of σ^2 , the variance.

There are many ways to compare the relative spread of different data sets, and we will review some of them in more detail in later lessons, particularly in the chapter on ANOVA.

Example A

What is the $CV\%$ of a data set with a variance of 23.91 and mean of 283?

Solution: Recall that $CV\%$ (coefficient of variance percentage) is equal to 100 times the ratio of the standard deviation to the mean. This means that we should start by finding the standard deviation.

- $\sigma = \sqrt{\sigma^2}$ So the standard deviation would be $\sqrt{23.91}$, or 4.89
- $CV\% = \frac{4.89}{283} \times 100 = 1.728\%$

Example B

What is the $CV\%$ of the data in the table below?

TABLE 5.15:

Spinner	Frequency
1	4
2	9
3	5
4	8
5	9
6	10
7	7

Solution: First find the population variance and standard deviation.

- $\mu = \frac{4+9+5+8+9+10+7}{52} = 7.43$
- Sum of squared deviances = 32.615
- Variance = $\frac{32.615}{7} = 4.659$
- Standard deviation = $\sqrt{4.659} = 2.16$

$$CV\% = \frac{2.16}{7.43} \times 100 = 29.07\%$$

Example C

Which population data set has the highest and which the lowest coefficient of variation?

$$x = \{14, 16, 17, 19, 16, 19\} \quad y = \{22, 24, 27, 24, 29, 35, 31\} \quad z = \{41, 44, 47, 44, 40, 49, 52\}$$

Solution: First find the mean and standard deviation of each set:

- Set x :
 - Mean : $\frac{101}{6} = 16.83$
 - Variance : 3.139
 - Standard Deviation : $\sqrt{3.139} = 1.772$
- Set y :
 - Mean : $\frac{192}{7} = 27.43$
 - Variance : 17.96
 - Standard Deviation : $\sqrt{17.96} = 4.24$
- Set z :
 - Mean : $\frac{317}{7} = 45.29$
 - Variance : 15.92
 - Standard Deviation : $\sqrt{15.92} = 3.99$

Divide the standard deviation of each set by its mean, multiply by 100, and compare the percent coefficients of variation:

- Coefficient of variation set x : $\frac{1.772}{16.83} = 10.53\%$
- Coefficient of variation set y : $\frac{4.24}{27.43} = 15.46\%$
- Coefficient of variation set z : $\frac{3.99}{45.29} = 8.8\%$

Set z has the lowest coefficient of variation and set y has the highest.

Concept Problem Revisited

Suppose you were given three different sets of data, one with a variance of 3.2 and mean of 9.2, another with a variance of 16 and mean of 45, and the third with a variance of 155 and mean of 2100. If you were asked which set was the least centrally clustered, how could you find out?

By finding the square root of the variance (the standard deviation), and dividing the standard deviation by the mean, you can find the coefficient of variation. Comparing the coefficients of variation allows you to directly compare the data clustering of each set, since a higher $CV\%$ means the data is more spread out.

Vocabulary

The **coefficient of variation** is a measure of data clustering calculated by dividing the standard deviation by the mean, and may be used to compare the spreads of different data sets.

Guided Practice

Find the coefficient of variation:

1. 3240, 3260, 3250, 3280, 3280, 3300, 3310, 3270
2. 34.4, 34, 34.7, 34.6, 34, 34.1, 31, 31.3
3. 989.22, 990.6, 992, 996.9, 981.1, 986, 975

Solutions:

1. Start by finding the mean and the standard deviation:

- Arithmetic **mean**: $\frac{26,190}{8} = 3273.75$
- Find the variance (here I am using “mean of squares minus the square of mean”) :
 - $\frac{3240^2+3260^2+3250^2+3280^2+3280^2+3300^2+3310^2+3270^2}{8} = 10,717,937.5$
 - Subtract the squared mean ($3273.75^2 = 10,715,802.25$) to get the variance: $10,717,937.5 - 10,715,802.25 = 2135.25$
- The square root of the variance is the **standard deviation**: $\sqrt{2135.25} = 46.209$

Divide the standard deviation by the mean, and multiply by 100 to get $CV\%$

- $\frac{46.209}{3273.75} \times 100 = 1.4115\%$

2. Find the mean and standard deviation:

- **Arithmetic mean**: $\frac{34.4+34+34.7+34.6+34+34.1+31+31.3}{8} = 33.5125$
- **Standard deviation** (square root of the “mean of squares minus square of mean”):

$$\sqrt{\frac{(34.4^2 + 34^2 + 34.7^2 + 34.6^2 + 34^2 + 34.1^2 + 31^2 + 31.3^2)}{8} - 33.5125^2} = 1.388$$

Divide the standard deviation by the mean, and multiply by 100 to get $CV\%$

- $\frac{1.388}{33.5125} \times 100 = 4.0414\%$

3. Find the mean and standard deviation:

- **Arithmetic mean:** $\frac{989.22+990.6+992+996.9+981.1+986+975}{7} = 987.26$

- **Standard deviation:**

$$\sqrt{\frac{989.22^2 + 990.6^2 + 992^2 + 996.9^2 + 981.1^2 + 986^2 + 975^2}{7} - 987.26^2} = 6.764$$

Divide the standard deviation by the mean and multiply by 100 to get CV%:

- $\frac{6.764}{987.26} \times 100 = .685\%$

Practice

Find the coefficient of variation %:

- 10, 11.1, 10.33, 10.63, 11, 11.2, 11.36, 10.46
- 275, 280.7, 283, 279, 284.2, 280, 282
- 7100.5, 7080, 7065.9, 7100, 7096, 7112, 7116.1
- 37, 35.3, 32.7, 34, 36, 36.2, 33.3, 33.8
- 3607, 3600, 3604, 3631, 3606
- 702, 704, 712, 716, 721, 716, 722
- 3370, 3300.5, 3366, 3306.6, 3310, 3336, 3301.3
- 34.4, 34, 34.7, 34.6, 34, 34.1, 31, 31.3
- 989.22, 990.6, 992, 996.9, 981.1, 986, 985
- 10.2, 16.34, 10.33, 10.63, 10.2, 10.44, 16.36, 10.46
- 3240, 3260, 3250, 3280, 3280, 3300, 3310, 3270

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 5.9.

Various real-world uses of probability and statistics were discussed and explored. Students were given examples of multiple ways in which an understanding of probability and statistics could contribute to their own lives and activities.

5.10 References

1. John Liu. <https://www.flickr.com/photos/8047705@N02/5560539738> .
2. CK-12 Foundation. . CCSA
3. JD Hancock. <https://www.flickr.com/photos/jdhancock/7238856518> .
4. CK-12 Foundation. . CCSA
5. CK-12 Foundation. . CCSA
6. CK-12 Foundation. . CCSA
7. CK-12 Foundation. . CCSA
8. CK-12 Foundation. . CCSA
9. CK-12 Foundation. . CCSA
10. Nicholas Eckhart. <https://www.flickr.com/photos/fanofretail/13957466674> .
11. Chuck Simmins. <https://www.flickr.com/photos/chucksimmins/2859050569> .
12. CK-12 Foundation. . CCSA
13. CK-12 Foundation. . CCSA
14. Anders Sandberg. <https://www.flickr.com/photos/arenamontanus/2243540719> .
15. CK-12 Foundation. . CCSA
16. CK-12 Foundation. . CCSA
17. Oscar Rethwell. <https://www.flickr.com/photos/rethwill/8635466165> .

CHAPTER 6**Probability****Chapter Outline**

- 6.1 BASIC PROBABILITY - PROBABILITY AND STATISTICS**
 - 6.2 UNION OF COMPOUND EVENTS**
 - 6.3 INTERSECTION OF COMPOUND EVENTS**
 - 6.4 MULTIPLICATION RULE**
 - 6.5 MUTUALLY INCLUSIVE EVENTS - PROBABILITY AND STATISTICS**
 - 6.6 CALCULATING CONDITIONAL PROBABILITIES**
 - 6.7 IDENTIFYING THE COMPLEMENT**
 - 6.8 FINDING PROBABILITY BY FINDING THE COMPLEMENT**
 - 6.9 REFERENCES**
-

Probability is the study of chance, calculating how likely a particular outcome may be as compared to all others. This chapter introduces some of the main concepts involved with the study of probability.

6.1 Basic Probability - Probability and Statistics

Objective

In the study of probability, singular events are the simplest events to learn about. However, by building your understanding of this concept, you will more easily understand the more complex probabilities of compound events.

Concept

Most people have heard, I think, of the old adage that buttered bread always lands buttered side down. However, from a scientific standpoint, what is the real statistical and experimental probability of buttered bread landing butter side up? For that matter, what is the difference between a statistical and an experimental probability?



Watch the video below and read through the lesson and we'll return to this question afterward.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63764>

<http://youtu.be/zzwuiuqsLCE0?t=1m47s> Myth Busters - Is Yawning Contagious?

Guidance

Probability is the study of chance. When studying probability, there are two very general classifications: *theoretical probability* and *experimental probability*.

- **Theoretical probability** is the calculated probability that a given outcome will occur if the same experiment were completed an infinite number of times.
- **Experimental probability** is the observed result of an experiment conducted a limited number of times.

For example, ignoring the very slight differences between the figures stamped onto each side of a coin, the *statistical probability* of a coin landing heads-up is 50%. However, if you flip a coin 10 times, you may very well find that the observed *experimental probability* results in 60% or 70% or even greater probability of one side landing up. This discrepancy is perfectly natural and expected when conducting experiments, and it is important to recognize it.

In this lesson we will confine our study to the probability of a simple event. The probability of a simple event is the calculated chance of a specific direct outcome of a single experiment where in all possible outcomes are equally likely. To calculate the probability of such an outcome, we use a very simple and intuitive formula:

$$P(x) = \frac{\text{number of events where } x \text{ is true}}{\text{total number of possible events}}$$

Where “ $P(x)$ ” is the probability that x will occur

In other words, just as you might expect, the probability of randomly picking one of the three blue marbles out of a bag with ten marbles total would be $\frac{3}{10}$.

Example A

You are given a bag containing 15 equally sized marbles. You know there are 10 yellow marbles and 5 green marbles in the bag. What is the probability that you would pull a yellow marble out, if you reach in the bag and grab a marble at random?

Solution: Use the formula for the probability of a simple event:

$$P(x) = \frac{\text{number of outcomes where } x \text{ is true}}{\text{total number of possible outcomes}}$$

In this case, we have:

$$P(\text{yellow}) = \frac{10 \text{ yellow marbles}}{15 \text{ total marbles}}$$

Which would reduce to:

$$P(\text{yellow}) = \frac{2}{3} \text{ or } 66.\bar{6}\%$$

Example B

What is the probability of rolling an odd number on a standard six-sided die?

Solution: A standard die has three odd numbers (1, 3, 5) and three even numbers (2, 4, 6). Therefore, the probability of rolling an odd number is:

$$P(\text{odd}) = \frac{3 \text{ odd}}{6 \text{ total}}$$

Reducing to:

$$P(\text{odd}) = \frac{1}{2} \text{ or } 50\%$$

Example C

If Lawrence is playing with a standard 52-card deck, then the statistical probability of him pulling a single Queen at random is: $\frac{4 \text{ queens}}{52 \text{ cards}} = \frac{1}{13} = 7.7\%$. If he decides to test it out and ends up pulling a Queen at random 6 times in 52 trials of “pull a card, record it, put it back”, what is the experimental probability of pulling a Queen?

Solution: Recall that experimental probability is the observed probability of a number of identical experiments. Experimental probability is not *affected* by statistical probability (it may be *predicted* by it, but not *affected*), therefore the experimental probability is:

$$P(y) = \frac{6 \text{ Queens}}{52 \text{ trials}}$$

Reducing to:

$$P(y) = \frac{3}{26} = 11.5\%$$

Concept Problem Revisited

From a scientific standpoint, what is the real statistical and experimental probability of buttered bread landing butter side up? For that matter, what is the difference between a statistical and an experimental probability?

Remember that the difference is that *statistical probability* is the calculated probability of a specific outcome, and *experimental probability* is the observed probability.

The statistical probability of the bread landing butter side up can be assumed to be $\frac{1}{2}$, based on bread having two sides.

According to the “MythBusters” experiment in the video, the observed probability was $\frac{29}{45}$. However, you should know that your results might be different!

Vocabulary

Theoretical probability is the calculated probability that a given outcome will occur if the same experiment were completed an infinite number of times.

Experimental probability is the observed result of an experiment conducted a limited number of times.

A **trial** is one “run” of a particular experiment.

An **event** is any collection of the outcomes of an experiment.

An **outcome** is the result of a single trial.

Guided Practice

1. What is the probability of pulling the 1 red marble out of a bag with 12 marbles in it?
2. What is the probability of a spinner landing on “6” if there are 6 equally spaced points on the spinner?
3. What is the probability of pulling a red card at random from a standard deck?

- What is the experimental probability of heads in an experiment where Scott flipped a coin 50 times and got heads 21 times?
- What is the probability of shaking the hand of a female student if you randomly shake the hand of one person in a room with 23 female students and 34 male students?

Solutions:

- $P(\text{red}) = \frac{1 \text{ red marble}}{12 \text{ total marbles}} = \frac{1}{12}$ or 8.3%
- $P(6) = \frac{1 \text{ number 6}}{6 \text{ total numbers}} = \frac{1}{6}$ or 16.7%
- $P(\text{red}) = \frac{26 \text{ red cards}}{52 \text{ total cards}} = \frac{26}{52} = \frac{1}{2}$ or 50%
- $P(\text{heads}) = \frac{21 \text{ heads}}{50 \text{ flips}} = \frac{21}{50}$ or 42%
- $P(\text{female}) = \frac{23 \text{ females}}{57 \text{ students}} = \frac{23}{57}$ or 40.4%

Practice

Questions 1-10, find the probability:

- Rolling a 4 on a standard die
- Pulling a King from a standard deck
- Pulling a green candy from an opaque bag with 5 red, 3 yellow, 3 blue, and 6 green candies.
- Getting a 5 from one spin on a spinner numbered 1-8 (equally spaced)
- Rolling an even number on a 20-sided die
- Rolling an odd number on a standard die
- Pulling a red card from a standard deck
- Pulling a face card from a standard deck
- Spinning red on a spinner with Red, Orange, Yellow, Green, Blue and Purple (equally spaced)
- Pulling a club from a standard deck
- Pulling a brown candy from a box of 25 candies, containing equal numbers of brown, red, green, blue, and yellow candies
- Getting a prime number with a random number generator that has an equal chance of generating any number between 1 and 50
- Getting a composite number with the same generator

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 6.1.

6.2 Union of Compound Events

Objective

In this lesson, you will learn about calculating the probability that any one of multiple *mutually exclusive independent events* will occur in a single experiment.

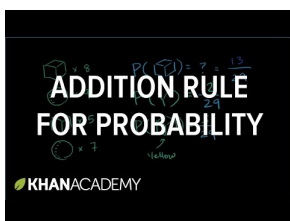
Concept

If you think it through, it should make sense that the probability of pulling one Queen at random from a standard deck is $\frac{4}{52}$ or $\frac{1}{13}$, since there are 4 Queens in a standard 52 card deck. How then would you calculate the probability of pulling a Queen OR a King from the same deck?



After this lesson on the *union of compound events*, we'll return to this question and work out the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63801>

<http://youtu.be/QE2uR6Z-NcU> Khan Academy - Addition Rule for Probability

Guidance

When multiple independent events may occur during a particular experiment, there are a couple of different types of outcomes you may need to consider:

- **Intersection:** the probability of *both* or *all* of the events you are calculating happening at the same time (less likely).

- **Union:** the probability of *any one* of multiple events happening at a given time (more likely).

In this lesson, we will focus on **union**. Calculating the union is relatively easy, you just add up the individual probabilities of the events:

$$P(x \text{ or } y) = P(x) + P(y)$$

This can also be thought of as:

$$P(x \text{ or } y) = \frac{(\text{number of outcomes where } x \text{ is true}) + (\text{number of outcomes where } y \text{ is true})}{\text{total number of possible outcomes}}$$

It is really just that simple! It is intuitive also, assuming there is no overlap (which we will consider later), it just makes sense to think that if you have a 20% probability of one thing happening, and a 30% probability of another, then you have a 50% probability of one of the two of them happening during a given experiment.

Example A

You are given a bag containing 15 equally sized marbles. You know there are 5 yellow marbles, 5 blue marbles, and 5 green marbles in the bag. What is the statistical probability that you would pull a yellow *or* green marble out, if you reach in the bag and grab a marble at random?

Solution: Recall the formula for the union of simple probabilities:

$$P(x \text{ or } y) = \frac{(\text{number of outcomes where } x \text{ is true}) + (\text{number of outcomes where } y \text{ is true})}{\text{total number of possible outcomes}}$$

In this case, we have:

$$P(\text{yellow or green}) = \frac{5 \text{ yellow marbles} + 5 \text{ green marbles}}{15 \text{ total marbles}}$$

Which would reduce to:

$$P(\text{yellow or green}) = \frac{2}{3} \text{ or } 66.\bar{6}\%$$

Example B

What is the probability of rolling an odd or even number on a standard six-sided die?

Solution: A standard die has three odd numbers (1, 3, 5) and three even numbers (2, 4, 6). Therefore, the probability of rolling an odd or even number is:

$$P(\text{odd or even}) = \frac{3 \text{ odd} + 3 \text{ even}}{6 \text{ total}} = \frac{6}{6}$$

Reducing to:

$$P(\text{odd or even}) = 1 \text{ or } 100\%$$

Example C

If Lawrence is playing with a standard 52-card deck, what is the probability of pulling a 2, a 4, or a 6 out of the deck at random?

Solution: Let's solve this one as the total of the individual probabilities. Lawrence's probability of pulling a 2, 4, or 6 is the same as the union of the probability of each possible outcome:

$$P(2, 4, \text{ or } 6) = P(2) + P(4) + P(6) = \frac{1}{13} + \frac{1}{13} + \frac{1}{13} = \frac{3}{13} \text{ or } 23.1\%$$

Concept Problem Revisited

It should make sense now that the probability of pulling one Queen at random from a standard deck is $\frac{4}{52}$ or $\frac{1}{13}$, since there are 4 Queens in a standard 52 card deck. How then would you calculate the probability of pulling a Queen OR a King from the same deck?

Remember that the *union* of multiple probabilities is simple the total sum of all of the individual probabilities:

$$P(\text{Queen or King}) = \frac{4 \text{ Queens} + 4 \text{ Kings}}{52 \text{ Cards}} = \frac{8}{52} \text{ or } \frac{4}{26} \text{ or } \frac{2}{13} = 15.4\%$$

Vocabulary

A **statistical probability** is the calculated probability that a given outcome will occur if the same experiment were completed an infinite number of times.

The probability of the **union** of multiple, mutually exclusive, events is the sum of the probabilities of each of the individual outcomes occurring during a given trial.

An **event** is any collection of the outcomes of an experiment.

Mutually exclusive events cannot occur at the same time (they have no overlap). For instance, a single coin flip cannot be *both* heads and tails.

An **independent event** is an event that is unaffected by any other event occurring before or after it.

An **outcome** is the result of a single trial.

Guided Practice

1. What is the statistical probability of pulling either the only red or the only blue marble out of a bag with 12 marbles in it?
2. What is the probability of a spinner landing on "2", "3", or "6" if there are 6 equally spaced points on the spinner?
3. What is the probability of pulling a red or black card at random from a standard deck?
4. What probability of picking a red or green marble from a bag with 5 red, 7 green, 6 blue, and 14 yellow marbles in it?
5. What is the of shaking the hand of a student wearing red if you randomly shake the hand of one person in a room containing the following mix of students?

- 13 female students wearing blue
- 7 male students wearing blue
- 6 female students wearing red
- 9 males students wearing red
- 18 female students wearing green
- 21 male students wearing green

Solutions:

1. $P(\text{red or blue}) = \frac{1 \text{ red marble} + 1 \text{ blue marble}}{12 \text{ total marbles}} = \frac{2}{12}$ or $\frac{1}{6}$ or 16.6%
2. $P(2 \text{ or } 3 \text{ or } 6) = \frac{1 \text{ number } 6 + 1 \text{ number } 2 + 1 \text{ number } 3}{6 \text{ total numbers}} = \frac{3}{6}$ or $\frac{1}{2}$ or 50%
3. $P(\text{red or black}) = \frac{26 \text{ red cards} + 26 \text{ black cards}}{52 \text{ total cards}} = \frac{52}{52} = \frac{1}{1}$ or 100%
4. $P(\text{red or green}) = \frac{5 \text{ red marbles} + 7 \text{ green marbles}}{32 \text{ total marbles}} = \frac{12}{32}$ or $\frac{6}{16}$ or $\frac{3}{8}$ or 37.5%
5. $P(\text{red}) = \frac{6 \text{ females wearing red} + 9 \text{ males wearing red}}{74 \text{ total students}} = \frac{15}{74}$ or 20.3%

Practice

1. What is the probability of rolling a standard die and getting between a 1 and 6 (inclusive)?
2. What is the probability of pulling one card from a standard deck and it being an 8, a 3, or a queen?
3. What is the probability of rolling a 5 or a 2 on an 8-sided die?
4. What is the probability of pulling one card from a standard deck and it being a spade, a diamond, or a club?
5. What is the probability of rolling a 1, 3, or 5 on a 7-sided die?
6. What is the probability of pulling one card from a standard deck and it being a king, a 4, or a 8?
7. What is the probability of pulling a yellow or blue candy from a bag containing 35 candies equally distributed among yellow, blue, green, red, and brown candies?
8. What is the probability of spinning 2, 4, or 7 on a 10-space spinner (equally spaced)?
9. What is the probability of rolling a 1, 3, 5, or 6 on a 20-sided die?
10. A car factory creates cars in the following ratio: 3 green, 2 blue, 7 white, 2 black and 1 brown. What is the probability that a randomly selected car will be either blue or brown?
11. There are 4 flavors of donuts on the shelf: glazed, sprinkles, plain, and powdered sugar. If there are equal numbers of each of the non-plain donuts, and half as many plain as any one of the others, what is the probability of randomly choosing a plain donut out of all donuts on the shelf?
12. What is the probability of randomly choosing a red Ace or a black King from a standard deck?
13. What is the probability of rolling a prime number or an even number on a standard die?
14. Mr. Spence's class has 13 students. 4 students are wearing coats, 3 are wearing vests, 3 are wearing hoodies, and the rest are in t-shirts. What is the probability that Mr. Spence will randomly call the name of a student wearing a coat or a vest?
15. In the same class, what is the probability that Mr. Spence will randomly call the name of a student in a hoodie or t-shirt?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 6.2.

6.3 Intersection of Compound Events

Objective

In this lesson, you will learn about calculating the probability that *all* of a series of *independent events* will occur in a single experiment.

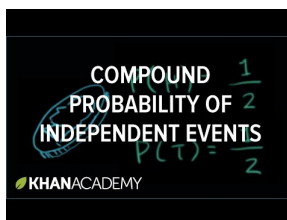
Concept

There is a classic example of probability studies involving a coin flip. Everyone knows that the probability of getting heads on a single flip is 50%, which means that every time you flip it, there is also a 50% probability of getting tails. The question is, if you have flipped a coin 99 times and got heads every time, what is the probability of getting heads the next time?



After this lesson on the intersection of compound events, we'll return to this question and see how it does (and doesn't!) fit with the concept.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63796>

<http://youtu.be/xSc4oLA9e8o> Khan Academy - Compound Probability of Independent Events

Guidance

It should make sense intuitively that the more specific or restricted you make the details of an event, the less probable it becomes for that event to occur. The concept of calculating the total probability of multiple events strung together is the same idea.

If I flip a coin once there are only two possible outcomes:

$$H \text{ (heads) or } T \text{ (tails)}$$

If I flip the coin twice, there are four possibilities:

$$H + T \text{ or } H + H \text{ or } T + H \text{ or } T + T$$

We know there are a total of four possible outcomes from two coin flips: HT , HH , TH , and TT , and only one of them: HH , results in the outcome we want to calculate. Using the simple probability formula, we get:

$$P(HH) = \frac{1 \text{ outcome}}{4 \text{ possible outcomes}} = \frac{1}{4} \text{ or } 25\%$$

Example A

What is the probability of flipping a coin four times and getting tails all four times?

Solution: Create a table listing all of the possible outcomes:



Now we can look at the bottom row and see that there are a total of 16 possibilities, only one of which is four tails in a row. The probability, therefore, is:

$$P(4 \text{ tails}) = \frac{1 \text{ outcome}}{16 \text{ outcomes}} = \frac{1}{16} = 6.25\%$$

Example B

What is the probability of rolling two even numbers in a row on a standard six-sided die?

Solution: Create a table listing all possible outcomes:



$$P(\text{two evens}) = \frac{9 \text{ favorable outcomes}}{36 \text{ total outcomes}} = \frac{9}{36} = \frac{1}{4}$$

Reducing to:

$$P(\text{two evens}) = \frac{1}{4} \text{ or } 25\%$$

Example C

What is the probability of spinning two 2's in a row OR two 4's in a row on a spinner with the numbers 1-4?

Solution: Create a table listing all possible outcomes, and highlight the favorable ones:

Spin 1	1				2				3				4			
Spin 2	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4

Out of a total of 16 possible outcomes, only 2 fit our description, which gives us:

$$P(2's \text{ or } 4's) = \frac{2 \text{ favorable outcomes}}{16 \text{ possible outcomes}} = \frac{2}{16} = \frac{1}{8} \text{ or } 12.5\%$$

Concept Problem Revisited

The question is, if you have flipped a coin 99 times and got heads every time, what is the probability of getting heads the next time?

This is a very common example of something called the gambler's fallacy. It is *not* a good example of calculating the intersection of compound events *because of the way it is worded*. The question as written is essentially asking about a single flip of the coin, which is always $\frac{50}{50}$, because a coin has no memory.

From the standpoint of an example of what we have been studying in this chapter, the more useful, and dramatically more difficult question would be:

What is the probability of flipping a coin 100 times and getting heads every time?

See the difference? The first question assumes that 99 flips have already occurred and asks about the last flip, the second question asks about all 100 flips.

If you want to know the probability of flipping 100 heads in a row, you could either draw a *really* long chart of all of the possibilities (like the one in Example A, but much longer), or you could use the **multiplication rule** that we will be learning in the next lesson. Check it out!

Vocabulary

An **independent event** is an event whose outcome is not directly affected by another event. (a coin flip, for example)

A **favorable outcome** is an outcome of an event that meets a set of initial specifications.

Two **mutually exclusive** events cannot both occur at the same time, (e.g. *both* heads *and* tails on the same coin flip).

Guided Practice

1. What is the probability of pulling 1 red marble, replacing it, then pulling another red marble out of a bag containing 4 red and 2 white marbles?
2. What is the probability of a spinner landing on “2” and then a “3”, or “6” if there are 6 equally spaced points on the spinner?
3. What is the probability of pulling a red and then a black card at random from a standard deck (replacing the first card after drawing)?
4. What probability of picking a red and then a green marble from a bag with 5 red and 1 green marbles in it (replacing the first marble after the draw)?
5. What is the probability of shaking the hand of a student wearing red and then a student wearing blue if you randomly shake the hands of two people in a row in a room containing 3 students in blue and 2 in red?

Solutions:

1. Make a chart:

first pull: $\frac{r \quad r \quad r \quad r \quad w \quad w}{\text{second pull: } rrrrww \quad rrrrww \quad rrrrww \quad rrrrww \quad rrrrww \quad rrrrww}$

The four sets of four red “r”s represent the favorable outcomes out of the total of 36, therefore $P(2 \text{ red}) = \frac{16}{36} = \frac{4}{9}$ or 44.44%

2. Make a chart:

first spin: $\frac{1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6}{\text{second spin: } 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6}$

The red numbers 3 and 6 represent the two favorable outcomes out of 36 total, therefore $P(2 \text{ and } (3 \text{ or } 6)) = \frac{2}{36}$ or $\frac{1}{18}$ or 5.55%

3. There are 26 black and 26 red cards in the deck, so the probability on the *first* pull is $P(\text{red}) = \frac{26 \text{ red cards}}{52 \text{ total cards}} = \frac{26}{52} = \frac{1}{2}$ or 50%. On the *second* pull, we again have a 50% chance of favorable outcome, but that 50% only applies to the half of the first pulls that were favorable. Therefore: $P(\text{red then black}) = 50\% \text{ of } 50\% = 25\%$

4. Make a chart:

first pull: $\frac{r \quad r \quad r \quad r \quad r \quad g}{\text{second pull: } rrrrrg \quad rrrrrg \quad rrrrrg \quad rrrrrg \quad rrrrrg \quad rrrrrg}$

Of the 36 possible outcomes, only 5 fit the description of red the first time, and green the second time (noted by the red “g”s). **Therefore** $P(\text{red then green}) = \frac{5}{36}$

5. Make a chart:

B					B					B					R					R									
B	B	B	R	R	B	B	B	R	R	B	B	B	R	R	B	B	B	R	R	B	B	B	R	R	B	B	B	R	R

So, out of the 25 possible handshake possibilities, 6 of them fit the requirements of red first, then blue:

$$P(\text{red then blue}) = \frac{6}{25}$$

Practice

Questions 1-6: Suppose you have an opaque bag filled with 4 red and 3 green balls. Assume that each time a ball is pulled from the bag, it is random, and the ball is replaced before another pull.

1. Create a chart of all possible outcomes of an experiment consisting of pulling one ball from the bag at random, noting the color and replacing it, then pulling another.
2. How many possible outcomes are there?
3. What is the probability of randomly pulling a red ball from the bag, returning it, and pulling a green ball on your second pull?
4. What is the probability of randomly pulling a red ball both times?
5. What is the probability of pulling a green ball both times?
6. Is the probability of pulling a red followed by a green different than pulling a green followed by a red?

Questions 7 - 12: Suppose you have two standard dice, one red and one blue.

7. Construct a probability distribution table or diagram for an experiment consisting of one roll of the red die followed by one roll of the blue one.
8. How many possible outcomes are there?
9. Is there an apparent mathematical relationship between the number of sides on the dice and the number of possible outcomes?
10. What is the probability of rolling a 2 on the red die and a 1, 3, or 5 on the blue one?
11. What is the probability of rolling an even number on the red die and an odd on the blue one?
12. Do the probabilities of a particular outcome change based on which die is rolled first? Why or why not?

Questions 13 - 16: Suppose you have a spinner with 5 equally-spaced color sections: red, blue, green, yellow, and orange.

13. Construct a probability distribution detailing the possible outcomes of three consecutive spins. You may wish to use only the first letter, or a single color-coded hash mark, to represent each possibility, as there will be many of them.
14. How many possible outcomes are there?
15. Is there an apparent mathematical relationship between the number of sections on the spinner, the number of spins, and the number of possible outcomes? If so, what is the relationship?
16. What is the probability of spinning red, then green, and then orange?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 6.3.

6.4 Multiplication Rule

Objective

Here you will learn how to quickly calculate the probability of the intersection of multiple independent events without building a frequency table.

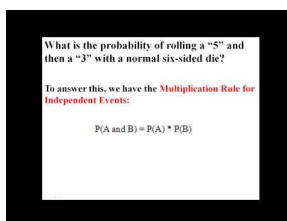
Concept

Finding the probability of getting two or three heads in a row when flipping a fair coin is straightforward enough by building a frequency table. However, the process becomes somewhat unwieldy when the experiment is more complex, such as calculating the probability of pulling 3 queens in a row from a standard deck of cards. Building a frequency table for all 52 cards would be time consuming at best.



There must be an easier way, right?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63805>

http://youtu.be/Q_7PR9kRXWs StatsLectures - Multiplication Rule (Probability “and”)

Guidance

If I bet someone \$1 that I can roll a standard die and get a 6, and then *do* roll the die and get a 6, I would probably get my \$1, along with a clap on the back and a “congratulations!” from my friends. However, if I bet someone \$20

that I can roll 20 times and get 6 *every* time, and then I do just that, I would probably be dealing with a very angry group of people who want to know how I cheated! That's because it would be, at best, *very* improbable that I could get a series of 20 6's in a row under normal circumstances. Let's see if we can find out just *how* improbable.

Let's start with the probability of just one roll of a 6:

Since there are 6 sides, the probability is:

$$P(6) = \frac{1 \text{ outcome}}{6 \text{ possible outcomes}} = \frac{1}{6} \text{ or } 16.7\%$$

For two 6's in a row:

If we create a table of the possible rolls where the only total outcome yielding two 6's is highlighted in blue, we get:

KEY

1 st roll possibilities					
2 nd	roll	roll	roll	roll	roll
1					
2					
3					
4					
5					
6					

With this already pretty unwieldy table, we can see that there are 36 possible outcomes of rolling a standard die twice. Therefore, the probability of rolling two 6's in a row is:

$$P(\text{two } 6's) = \frac{1 \text{ outcome}}{36 \text{ possible outcomes}} = \frac{1}{36} \text{ or } 2.8\%$$

It should be apparent that things only get crazier from here, since calculating 3 6's this way would require another row of 6 possibilities for each of the 36 outcomes of the 2nd roll! However, look at the difference between the two probabilities:

$$P(\text{one } 6) = \frac{1}{6} \text{ and } P(\text{two } 6's) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

The fact that the probability of getting two 6's is $\frac{1}{6^{\text{th}}}$ of $\frac{1}{6}$ is no coincidence, of course. In fact, to understand how to calculate more complex intersections of independent compound probabilities, it may help to remember something you likely learned when practicing word problems:

To translate English to math, the word *of* and the multiplication sign \cdot or \times mean the same thing.

Since the probability of getting two 6's in a row is $\frac{1}{6^{\text{th}}}$ of $\frac{1}{6^{\text{th}}}$ we can say:

$$P(\text{two } 6's) = \frac{1}{6} \text{ of } \frac{1}{6} = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

This is an example of the *multiplication rule* of compound probability:

$$P(\text{total}) = P(\text{1st outcome}) \times P(\text{2nd outcome}) \dots \times P(\text{last outcome})$$

Now we can actually calculate the probability of 20 6's in a row:

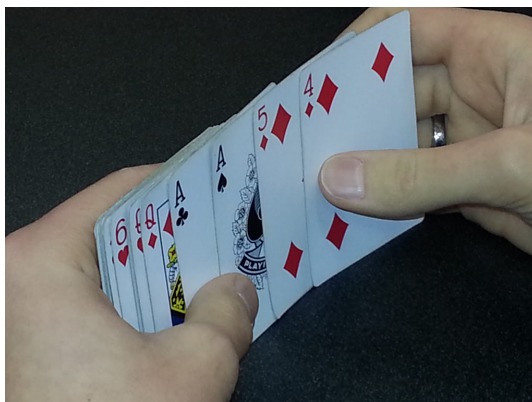
$$P(20 \text{ 6's}) = \frac{1}{6} \times \frac{1}{6} \dots \frac{1}{6} = \left(\frac{1}{6}\right)^{20} = \frac{1}{3,656,158,440,062,976}$$

or approximately one in three and one-half quadrillion, which I would consider *not* good odds!

Which also illustrates the practical impossibility of solving such a question with a frequency table, since it would take approximately 116,000,000 *years* just to write out the 6th row at one number per second! (not to mention the 1 *trillion* sheets of paper...)

Example A

What would be the theoretical probability of randomly pulling a queen from a deck of 52 cards, putting it back, randomly pulling a queen again, and so on until you have pulled 5 queens in a row?



Solution: The theoretical probability of pulling a single queen from a standard deck is:

$$P(\text{queen}) = \frac{4 \text{ queens}}{52 \text{ cards}} = \frac{1}{13} \text{ or } 7.7\%$$

If we use the multiplication rule for five pulls, we get:

$$\frac{1}{13} \times \frac{1}{13} \times \frac{1}{13} \times \frac{1}{13} \times \frac{1}{13} = \frac{1}{13^5} = \frac{1}{371293}$$

Example B

What is the theoretical probability of rolling a 1, 2, 3, 4, 5, and then 6, in order, on six successive rolls of a standard die?

Solution: The probability of rolling any single number on a standard die is $\frac{1}{6}$. Use the multiplication rule:

$$P(1-6) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{6^6} = \frac{1}{46656}$$

Example C

What is the theoretical probability that you might deal the King of Hearts, Jack of Diamonds, and then any Ace, in order, from a standard deck of cards, assuming you replace each card after drawing?

Solution: Let's start by evaluating the individual probabilities.

- $P(\text{first card}) = \frac{1 \text{ King of Hearts}}{52 \text{ total cards}} = \frac{1}{52}$
- $P(\text{second card}) = \frac{1 \text{ Jack of Diamonds}}{52 \text{ total cards}} = \frac{1}{52}$
- $P(\text{third card}) = \frac{4 \text{ Aces}}{52 \text{ total cards}} = \frac{4}{52} = \frac{1}{13}$

Now we can use the multiplication rule:

$$P(\text{King of Hearts, Jack of Diamonds, Ace}) = \frac{1}{52} \times \frac{1}{52} \times \frac{1}{13} = \frac{1}{35152}$$

Concept Problem Revisited

Finding the probability of getting two or three heads in a row when flipping a fair coin is straightforward enough by building a frequency table. However, the process becomes somewhat unwieldy when the experiment is more complex, such as calculating the probability of pulling 3 queens in a row from a standard deck of cards. Building a frequency table for all 52 cards would be time consuming at best.

There must be an easier way, right?

Of course, now you know this isn't even really a question anymore, the multiplication rule makes this question pretty easy:

$$P(3 \text{ queens}) = \frac{1}{52} \times \frac{1}{51} \times \frac{1}{50} = \frac{1}{132600}$$

Vocabulary

The **multiplication rule of probability** states that, for independent events: $P(\text{total}) = P(\text{case 1}) \times P(\text{case 2}) \dots \times P(\text{case } n)$

Guided Practice

1. What is the probability of rolling an odd number, followed by an even number, followed by a prime number, on three successive rolls of a 20-sided die?
2. What is the probability of pulling a heart, replacing it, pulling a club, replacing it, pulling a diamond, replacing it, then pulling a spade, all from a standard deck?
3. What is the probability of flipping a coin ten times in a row, and getting heads every time?
4. What is the probability of spinning red 5 times in a row on a spinner with 6 equally spaced color segments, only one of which is red?

Solutions:

1. Apply the multiplication rule: $P(\text{total}) = P(\text{case 1}) \times P(\text{case 2}) \dots \times P(\text{case } n)$

$$P(\text{odd, even, prime}) = P(\text{odd}) \times P(\text{even}) \times P(\text{prime})$$

$$P(\text{odd, even, prime}) = \frac{10}{20} \times \frac{10}{20} \times \frac{8}{20} = \frac{800}{8000} = \frac{1}{10} \text{ or } 0.1 \text{ or } 10\%$$

$$P(\text{odd, even, prime}) = 10\%$$

2. We can apply the multiplication rule here also:

$$P(\text{heart, club, diamond, spade}) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{256} \text{ or } 0.00391 \text{ or } .4\%$$

$$P(\text{heart, club, diamond, spade}) = 0.4\%$$

3. Since we are looking for the same outcome from the same experiment repeated ten times, we can 'shortcut' the multiplication rule by using an exponent:

$$P(\text{heads ten times}) = \left(\frac{1}{2}\right)^{10} = \frac{1^{10}}{2^{10}} = \frac{1}{1024} \text{ or } 0.001 \text{ or } 0.1\%$$

$$P(\text{heads ten times}) = 0.1\%$$

4. This one is similar to the last, in that we are looking for the same outcome of the same experiment, multiple times (5 times, in this case).

$$P(\text{red five times}) = \left(\frac{1}{6}\right)^5 = \frac{1^5}{6^5} = \frac{1}{7776} \text{ or } 0.0001 \text{ or } 0.01\%$$

$$P(\text{red five times}) = 0.01\%$$

Practice

1. What are independent events?
2. What is the multiplication rule?

Questions 3-7: Suppose you have an opaque bag filled with 6 red, 4 green, 7 blue and 5 purple balls.

3. What is the probability of randomly pulling a purple ball from the bag, returning it, and pulling a purple ball again on your second pull?
4. What is the probability of randomly pulling a red ball from the bag, returning it, and pulling a blue ball on your second pull?
5. What is the probability of randomly pulling a green ball from the bag, returning it, and pulling a green ball again on your second pull?
6. What is the probability of randomly pulling a blue ball from the bag, returning it, and pulling a red ball on your second pull?
7. What is the probability of randomly pulling a purple ball from the bag, returning it, and pulling a blue ball on your second pull?

Questions 8 - 12: Suppose you have two standard dice, one red and one blue.

8. What is the probability of rolling a 3 on the red die and a 5 on the blue one?
9. What is the probability of rolling a 3 or 4 on the red die and a 5 on the blue one?

10. What is the probability of rolling an even number on the red die and an odd on the blue one?
11. What is the probability of rolling a 6 on the red die and an odd number on the blue one?
12. What is the probability of rolling a 1 on the red die and prime number on the blue one?

Questions 13 - 16: Suppose you are dealing with a standard deck of cards, calculate the probability of each outcome as described, assuming you replace each card after drawing it.

13. Pulling a queen, then any club, then any red card.
14. Pulling the Ace of Spades, then a red 6, then any king.
15. Pulling any face card three times in a row.
16. Pulling a face card, then an ace, then the 5 of clubs.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 6.4.

6.5 Mutually Inclusive Events - Probability and Statistics

Objective

Here you will learn how to calculate the probability that any one of multiple events will occur, even though two or more of them could happen at the same time.

Concept

How does the Addition Rule for the union of probabilities: $P(A \text{ or } B) = P(A) + P(B)$ work when at least some of the events overlap? For example, in a room with 20 people, there are 5 women wearing red and 5 wearing yellow, and there are 5 men wearing red and 5 wearing green. What is the probability of randomly picking one person who is either wearing red or is male?

Since $P(\text{wearing red})$ actually overlaps with $P(\text{male})$, we can't just use the addition rule, so how do we find the answer?



We will return to this question at the end of the lesson.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/64524>

<http://vimeo.com/46752439> CK-12 - MutuallyInclusiveEventsA

Guidance

Mutually inclusive events are events that have at least some amount of “overlap”, in other words, at least one of the favorable outcomes of one event is the same as a favorable outcome of another event. Because one or more outcomes may satisfy multiple cases, we cannot simply add up the probabilities as we did with **mutually exclusive** events or we would end up with some events being counted twice. To account for the duplication, we just need to subtract the duplicated probabilities from the sum.

The modified formula looks like this:

$$P(A \text{ or } B)(\text{mutually inclusive}) = P(A) + P(B) - P(A \text{ and } B)$$

Prob that either A or B will occur = Prob of A occurring + Prob of B occurring – Prob of both at once

Example A

Consider a bag with five marbles in it. If there are three large marbles, one green, one blue, and one red, and also one each small red and small blue marbles, what is the probability that a random choice would be small or red?

Solution:

If we were to try to solve this with the simple addition rule, we would **wrongly** get $\frac{2}{5} + \frac{2}{5} = \frac{4}{5}$ or 80%.

In fact, if we check that answer with a table of possible outcomes, we can *see* that it is incorrect since there are only three marbles that could qualify as either small or red: 1) The large red marble, 2) The small red marble, and 3) The small blue marble.

The correct solution is:

$$\frac{1 \text{ large red} + 2 \text{ small}}{5 \text{ total marbles}} = \frac{3}{5} \text{ or } 60\%$$

What went wrong when we used the simple addition rule? The problem is that we ended up counting the small red marble *twice*.

Incorrectly calculated: $\frac{(1 \text{ large red} + 1 \text{ small red}) + (1 \text{ small blue} + 1 \text{ small red})}{5 \text{ total marbles}} = \frac{4 \text{ marbles}}{5 \text{ marbles}}$

Properly calculated: $\frac{((1 \text{ large red} + 1 \text{ small red}) + (1 \text{ small blue}))}{5 \text{ total marbles}} = \frac{3 \text{ marbles}}{5 \text{ marbles}}$

This leads us back to the modified addition rule for mutually inclusive events:

$$P(A \text{ or } B) = (P(A) + P(B) - P(A \text{ and } B))$$

Example B

Suppose you are playing with the spinner in the image below. What is the theoretical probability that the spinner would randomly land on either a top quadrant or a red quadrant?



Solution: There are two favorable events here, red and top. To apply the modified addition rule, we need to know the probability of each case, and the probability of the intersection of the two cases:

The probability of the spinner landing on red is:

$$P(\text{red}) = \frac{2 \text{ red spaces}}{4 \text{ total spaces}} = \frac{1}{2} \text{ or } 50\%$$

The probability of the spinner landing on a top space is:

$$P(\text{top}) = \frac{2 \text{ top spaces}}{4 \text{ total spaces}} = \frac{1}{2} \text{ or } 50\%$$

The probability of the spinner landing on a top red space is:

$$P(\text{top AND red}) = \frac{1 \text{ top red space}}{4 \text{ total spaces}} = \frac{1}{4} \text{ or } 25\%$$

Inserting those values into the formula, we get:

$$P(\text{top OR red}) = \left(\frac{1}{2} + \frac{1}{2} \right) - \frac{1}{4} = \frac{3}{4} \text{ or } 75\%$$

Example C

If $P(A) = 40\%$, $P(B) = 30\%$, and $P(A \text{ and } B) = 2\%$, are $P(A)$ and $P(B)$ mutually inclusive or mutually exclusive?

Solution: This one is easier than it looks. If $P(A \text{ and } B)$ is greater than 0% , then they are inclusive, since it is possible for there to be outcomes that are both A and B .

Concept Problem Revisited

In a room with 20 people, there are 5 women wearing red and 5 wearing yellow, and there are 5 men wearing red and 5 wearing green. What is the probability of randomly picking one person who is either wearing red or is male?

Now we know to use the modified addition rule for inclusive events to solve this sort of problem:

$$P(\text{red}) = \frac{5 \text{ women in red} + 5 \text{ men in red}}{20 \text{ total people}} = \frac{1}{2} \text{ or } 50\%$$

$$P(\text{male}) = \frac{5 \text{ men in green} + 5 \text{ men in red}}{20 \text{ total people}} = \frac{1}{2} \text{ or } 50\%$$

$$P(\text{red, male}) = \frac{5 \text{ men in red}}{20 \text{ total people}} = \frac{1}{4} \text{ or } 25\%$$

Therefore:

$$P(\text{red} \cup \text{male}) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4} \text{ or } 75\%$$

Vocabulary

Mutually inclusive events are events that have at least some amount of “overlap”, in other words, at least one of the favorable outcomes of one event is the same as a favorable outcome of another event.

Mutually exclusive events are events that cannot both be correct at the same time, such as a single coin flip, which cannot be both heads and tails at once.

Guided Practice

For questions 1 - 4, suppose you have a bag containing 5 quarters, 3 dimes, 4 nickels, 4 pennies, and 5 gold \$1 coins.

1. What is the probability that a random coin will be either silver or worth less than 10 cents?
2. What is the probability that a random coin will be either gold or worth more than 10 cents?
3. What is the probability that you pick two coins in a row that are each either silver or worth more than 10 cents?
4. What is the probability that a random coin is either worth more than 9 cents or silver?

Solutions:

1. Use the modified addition rule for inclusive events:

$$P(A \text{ or } B) = (P(A) + P(B) - P(A \text{ and } B))$$

$$P(\text{silver or } < 10 \text{ cents}) = (P(\text{silver}) + P(< 10 \text{ cents}) - P(\text{silver and } < 10 \text{ cents}))$$

$$= \left(\frac{12}{21} + \frac{8}{21} - \frac{4}{21} \right)$$

$$P(\text{silver or } < 10 \text{ cents}) = \frac{16}{21} \text{ or } 76.2\%$$

- 2.

$$P(\text{gold or } > 10 \text{ cents}) = (P(\text{gold}) + P(> 10 \text{ cents}) - P(\text{gold and } > 10 \text{ cents}))$$

$$= \left(\frac{5}{21} + \frac{10}{21} - \frac{5}{21} \right)$$

$$P(\text{gold or } > 10 \text{ cents}) = \frac{10}{21} \text{ or } 47.6\%$$

3. This is a two-step problem, first we need to calculate the probability of a single choice being silver or worth more than 10 cents, and then we can apply the multiplication rule to calculate the total probability.

$$\begin{aligned} P(\text{silver or } > 10 \text{ cents}) &= (P(\text{silver}) + P(> 10 \text{ cents}) - P(\text{silver and } > 10 \text{ cents})) \\ &= \left(\frac{12}{21} + \frac{10}{21} - \frac{5}{21} \right) \\ P(\text{silver or } > 10 \text{ cents}) &= \frac{17}{21} \text{ or } 81\% \end{aligned}$$

Now we use the multiplication rule: $P(A \text{ then } B) = P(A) \times P(B)$

$$\begin{aligned} P(\text{silver or } > 10 \text{ cents then silver or } > 10 \text{ cents}) &= 0.81 \times 0.81 = 0.656 \\ P(\text{silver or } > 10 \text{ cents then silver or } > 10 \text{ cents}) &= 66\% \end{aligned}$$

4.

$$\begin{aligned} P(> 9 \text{ cents or silver}) &= (P(> 9 \text{ cents}) + P(\text{silver}) - P(> 9 \text{ cents and silver})) \\ P(> 9 \text{ cents or silver}) &= \left(\frac{13}{21} + \frac{12}{21} - \frac{8}{21} \right) \\ P(> 9 \text{ cents or silver}) &= \left(\frac{17}{21} \right) \text{ or } 81\% \end{aligned}$$

Practice

1. What is the probability that the outcome of 1 roll of a 10-sided die will be either even or greater than 5?
2. What is the probability that the outcome of one roll of a 12 sided die will be either prime or odd?
3. What is the probability of randomly pulling either a king or a heart from a standard deck?
4. What is the probability of randomly pulling either an even numbered card or a black card from a standard deck?
5. What is the probability that one roll of two standard dice will either result in a number either even or less than 7?
6. What is the probability that a randomly chosen month will either start with a "J" or have 30 days?

For problems 7 - 11, suppose you have a bag containing 4 blue, 3 green, 5 yellow, and 2 red marbles. All of the green and 2 of the yellow marbles are larger than normal, and 3 of the blue and 1 of the red marbles are smaller than normal.

7. What is the probability of randomly pulling a marble that is either large or yellow?
8. What is the probability that a randomly chosen marble is not large or is yellow?
9. What is the probability that a randomly chosen marble is either small or red?
10. What is the probability that a randomly chosen marble is either normally sized or blue?
11. What is the probability that a randomly chosen marble is small or blue?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 6.5.

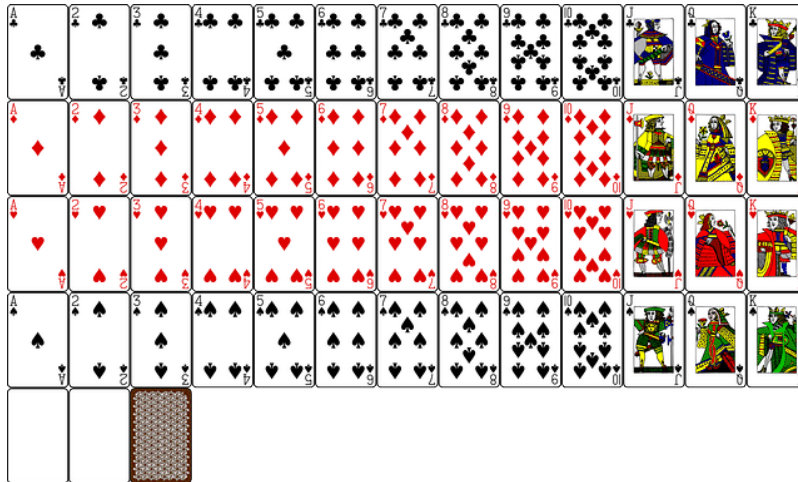
6.6 Calculating Conditional Probabilities

Objective

Here you will learn to calculate a probability that depends on another, different, probability in order to occur.

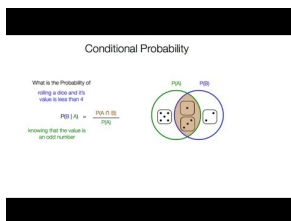
Concept

Suppose you wanted to calculate the probability of pulling the King of Hearts, then the Jack of Diamonds, and then any of the four Aces, from a standard deck of 52 cards, in that order, and without replacing any cards between pulls. Would the probability be significantly different than if you put the cards back after drawing each time?



In this lesson we will discuss conditional probabilities that are different for each trial. We'll return to this question after the lesson.

Watch This



MEDIA

Click image to the left or use the URL below.
 URL: <https://www.ck12.org/flx/render/embeddedobject/63809>

<http://youtu.be/H02B3aMnKzE> statisticsfun - How to Calculate Conditional Probability

Guidance

In a previous lesson, we discussed compound probabilities and reviewed some situations involving the probability of multiple occurrences of the same event in a row. A standard example would be the probability of throwing a fair

coin three times and getting three heads. In this lesson, we will be introducing a slightly more complex situation, where the coin may or may *not* be fair.

A new concept we will be introducing in this lesson is the “given that” concept. The idea is that we sometimes need to calculate a probability with a specific condition, for example:

The probability of rolling a “2” on a standard die is $\frac{1}{6}$. What is the probability of rolling a “2”, *given that* I know already that I have rolled an even number? As described in the video above, this is a *conditional probability*, and we notate it this way: $P(2|even)$, which is read as “The probability of rolling a “2” *given that* we roll an even number”.

The difference in calculations is:

$$P(2) = \frac{1}{6} \text{ (on number on the 6-sided die is a 2)}$$

$$P(2|even) = \frac{1}{3} \text{ (one number out of the three even numbers is a 2)}$$

To calculate a “given that” type of problem, we use the ***conditional probability formula***:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This is read as: “The probability that *A* will occur, given that *B* will occur (or has occurred), is equal to the intersection of probabilities *A* and *B* divided by the probability of *B* alone.

We have practiced the use of the addition rule and the multiplication rule for calculating probabilities, here we will also be using those again, but this time we will need to combine them for some of the problems.

For review:

Multiplication Rule for independent events: $P(A \text{ then } B) = P(A) \times P(B)$

Addition Rule for mutually exclusive events: $P(A \text{ or } B) = P(A) + P(B)$

Addition Rule for mutually inclusive events: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Example A

What is the probability that you have pulled the Jack of Hearts from a standard deck, given that you know you have pulled a face card?

Solution: Let’s solve this using the conditional probability formula first (A), then check by looking at the question another way (B):

A. The problem asks us to calculate the probability of a card being the Jack of Hearts, given that the card is a face card: $P(\text{Jack of Hearts}|\text{face card})$. Apply the conditional probability formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Putting in the information from the problem gives us:

$$P(\text{Jack of Hearts}|\text{face card}) = \frac{P(\text{Jack of Hearts} \cap \text{face card})}{P(\text{face card})}$$

$$P(\text{Jack of Hearts}|\text{face card}) = \frac{\left(\frac{1}{52}\right)}{\frac{12}{52}}$$

$$P(\text{Jack of Hearts}|\text{face card}) = \frac{1}{52} \times \frac{52}{12} = \frac{1}{12}$$

$$P(\text{Jack of Hearts}|\text{face card}) = \frac{1}{12} \text{ or } 8.33\%$$

B. The other way to view this is that we are looking for the probability of pulling the Jack of Hearts from the sample space including only face cards, which means we are looking for one specific card from a set including only 12 cards:

$$P(\text{Jack of Hearts}) = \frac{1 \text{ Jack of Hearts}}{12 \text{ face cards}} = \frac{1}{12}$$

$$P(\text{Jack of Hearts}) = \frac{1}{12} \text{ or } 8.33\%$$

We calculate 8.33% both ways, looks like we got it!

Example B

What is the probability that you could roll a standard die and get a 6, then grab a deck of cards and pull the King of Clubs, keep it, and then pull the Jack of Hearts?



Solution: This one looks rather complex, but it can be seen as just three individual probabilities:

1. $P(\text{roll } 6) = \frac{1 \text{ side with a } 6}{6 \text{ sides}} = \frac{1}{6}$
2. $P(\text{King}) = \frac{1 \text{ King of Clubs}}{52 \text{ cards}} = \frac{1}{52}$
3. $P(\text{Jack}) = \frac{1 \text{ Jack of Hearts}}{51 \text{ cards left after first pull}} = \frac{1}{51}$

The overall probability can, and should, be calculated with the multiplication rule, since the 2nd and 3rd are dependent:

$$P(\text{roll } 6 | \text{King} | \text{Jack}) = P(\text{roll } 6) \times P(\text{King}) \times P(\text{Jack})$$

$$P(\text{roll } 6 \text{ then pull King then pull Jack}) = \frac{1}{6} \times \frac{1}{52} \times \frac{1}{51} = \frac{1}{15912} \text{ OR } .167 \times .019 \times .020 = .000063$$

Example C

You reach into a bag containing 6 coins, 4 are 'fair' coins (they have an equal chance of heads or tails), and 2 are 'unfair' coins (they have only a 35% chance of tails). If you randomly grab a coin from the bag and flip it 3 times, what is the probability of getting 3 heads?

Solution: We actually have two different situations here:

1. We flip a 'fair' coin 3 times and get 3 heads

2. We flip an 'unfair' coin and get 3 heads

Since there are 4 fair coins, and 2 unfair coins, we can say the probability of: $P(\text{choose fair}) = \frac{4}{6}$ or $\frac{2}{3}$ and $P(\text{choose unfair}) = \frac{2}{6} = \frac{1}{3}$.

Note that $P(\text{choose unfair})$ would be the same thing as $P(\text{choose fair})'$, (see the apostrophe?) which is the **complement** of $P(\text{choose fair})$. In other words: the probability of choosing an unfair coin is 100% minus the probability of choosing a fair coin.

Let's calculate the probabilities of flipping each 3 times using the multiplication rule:

- The fair coin has a .5 chance of heads each flip: $P(\text{fair 3 heads}) = .5 \times .5 \times .5 = .125$
- The unfair coin has a .65 chance: $P(\text{unfair 3 heads}) = .65 \times .65 \times .65 = .275$

So now we can put them together to find the overall probability (the union) by applying the addition rule:

$$\begin{aligned} P(3 \text{ heads either coin}) &= .6\bar{6} \times P(\text{fair 3 heads}) + .3\bar{3} \times P(\text{unfair 3 heads}) \\ P(3 \text{ heads either coin}) &= .6\bar{6} \times .125 + .3\bar{3} \times .275 \\ P(3 \text{ heads either coin}) &= .083 + .092 = .175 \end{aligned}$$

The probability that we can randomly grab a coin from the bag and flip three heads in a row with it is 17.5%

Concept Problem Revisited

Suppose you wanted to calculate the probability of pulling the King of Hearts, then the Jack of Diamonds, and then any of the four Aces, from a standard deck of 52 cards, in that order and without putting any back. Would the probability be significantly different than if you put the cards back after drawing each time?

The probability would be different, but perhaps less different than you might think, at least as a percentage. Let's look at the two cases, with $P(A)$ representing the probability with each choice coming out of a full deck of 52 cards, and $P(B)$ representing the probability when the deck gets smaller each pull:

$$\begin{aligned} P(A) &= \frac{1 \text{ King of Hearts}}{52 \text{ cards}} \times \frac{1 \text{ Jack of Diamonds}}{52 \text{ cards}} \times \frac{4 \text{ aces}}{52 \text{ cards}} = \frac{1}{52} \times \frac{1}{52} \times \frac{1}{13} = \frac{1}{35152} \\ P(B) &= \frac{1 \text{ King of Hearts}}{52 \text{ cards}} \times \frac{1 \text{ Jack of Diamonds}}{51 \text{ Cards}} \times \frac{4 \text{ aces}}{50 \text{ cards}} = \frac{1}{52} \times \frac{1}{51} \times \frac{4}{50} = \frac{4}{132600} = \frac{1}{33150} \end{aligned}$$

The difference in probability is approximately $\frac{1}{2000}$ or $\frac{5}{100}$ of 1%, pretty small difference!

Vocabulary

A **conditional probability** is a probability that depends on the outcome of another event.

The **conditional probability formula** is $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Guided Practice

For questions 1 - 3: Suppose you have two coins, one is a normal, fair coin, and the other is an unfair coin that has a 75% chance of landing on heads. For each question, assume you reach into the bag, grab one of the two coins at random, and perform the experiment using that coin.

1. What would be the probability of the coin landing heads on your first flip?

2. What would be the probability of flipping tails four times in a row?
3. What would be the probability of flipping heads five times in a row?
4. Assume you are using a limited portion of a deck of cards that only includes face cards (no number cards). Assume also that each time you pull a card, you keep it until the end of the experiment. What would be the probability of pulling three kings in a row?
5. What would be the probability of rolling a 5, given that you know you rolled an odd number?

Solutions:

1. To calculate the probability of flipping heads, we need to calculate the union of 50% of the probability of flipping heads on each coin. (Why 50% of each probability? There are two coins, so the chance that you will pull either one is 50%)

$$P(\text{heads}|\text{either coin}) = 50\% \times P(\text{heads}|\text{fair coin}) + 50\% \times P(\text{heads}|\text{unfair coin})$$

$$P(\text{heads}|\text{either coin}) = 50\%(50\%) + 50\%(75\%)$$

$$P(\text{heads}|\text{either coin}) = 25\% + 37.5\%$$

$$P(\text{heads}|\text{either coin}) = 62.5\%$$

2. To calculate the probability of flipping four tails in a row, we calculate the union of 50% of the probability of flipping four tails in a row with each coin, much like in question 1.

$$P(4 \text{ tails}|\text{either coin}) = 50\% \times P(4 \text{ tails}|\text{unfair coin}) + 50\% \times P(4 \text{ tails}|\text{fair coin})$$

$$P(4 \text{ tails}|\text{either coin}) = 50\%(25\% \times 25\% \times 25\% \times 25\%) + 50\%(50\% \times 50\% \times 50\% \times 50\%)$$

$$P(4 \text{ tails}|\text{either coin}) = 0.1953\% + 3.125\%$$

$$P(4 \text{ tails}|\text{either coin}) = 3.32\%$$

3. Just like question 2, only this time the probability will end up greater, since the unfair coin has a large chance of heads:

$$P(4 \text{ heads}|\text{either coin}) = 50\% \times P(4 \text{ heads}|\text{unfair coin}) + 50\% \times P(4 \text{ heads}|\text{fair coin})$$

$$P(4 \text{ heads}|\text{either coin}) = .50 \times (.75 \times .75 \times .75 \times .75)^4 + .50 \times (.25 \times .25 \times .25 \times .25)^4$$

$$P(4 \text{ heads}|\text{either coin}) = .1582 + .0020$$

$$P(4 \text{ heads}|\text{either coin}) = 16.02\%$$

4. The key here is to note that you do *not* replace the card between pulls. That means that the probability changes with each trial. Let's look at the situation for each trial individually:

- (T1): Since we are only dealing with face cards, our first trial will have 12 possible outcomes, four for each of three face cards. Four of the outcomes are favorable, since there are four kings.
- (T2) Our second trial will have only 11 outcomes, since we are keeping the first card. There are only three favorable outcomes this time, since we "used up" a king if (T1) was favorable.
- The third pull (T3) only has 10 outcomes, since we will already have the other two cards. Two of the outcomes are favorable, since there would be only two kings left.

$$P(\text{three kings}|\text{face card}) = P(T1) \times P(T2) \times P(T3)$$

$$P(\text{three kings}|\text{face card}) = \frac{4 \text{ kings}}{12 \text{ face cards}} \times \frac{3 \text{ kings}}{11 \text{ face cards}} \times \frac{2 \text{ kings}}{10 \text{ face cards}}$$

$$P(\text{three kings}|\text{face card}) = .3\bar{3} \times .27\bar{2}\bar{7} \times .2$$

$$P(\text{three kings}|\text{face card}) = 0.0182 \text{ or } 1.82\%$$

5. This is a 'given that' problem, so we can use the conditional probability formula:

$$P(\text{roll } 5|\text{roll odd}) = \frac{P(\text{roll } 5) \cap P(\text{roll odd})}{P(\text{roll odd})}$$

$$P(\text{roll } 5|\text{roll odd}) = \frac{\frac{1}{6}}{\frac{1}{2}}$$

$$P(\text{roll } 5|\text{roll odd}) = \frac{2}{6} \text{ or } \frac{1}{3} \text{ or } 33.33\%$$

Practice

1. What is the probability that you roll two standard dice, and get 4's on both, given that you know that you have already rolled a 4 on one of them?
2. Assuming you are using a standard deck, what is the probability of drawing two cards in a row, without replacement, that are the same suit?
3. What is the probability that a single roll of two standard dice will result in a sum greater than 8, given that one of the dice is a 6?
4. Assuming a standard deck, what is the probability of drawing 3 queens in a row, given that the first card is a queen?
5. There are 130 students in your class, 50 have laptops, and 80 have tablets. 20 of those students have both a laptop and a tablet. What is the probability that a randomly chosen student has a tablet, given that she has a laptop?
6. Thirty percent of your friends like both Twilight and The Hobbit, and half of your friends like The Hobbit. What percentage of your friends who like the Hobbit also like Twilight?
7. Assume you pull and keep two candies from a jar containing sweet candies and sour candies. If the probability of selecting one sour candy and one sweet candy is 39%, and the probability of selecting a sweet candy first is 52%, what is the probability that you will pull a sour candy on your second pull, given that you pulled a sweet candy on your first pull?
8. The probability that a student has called in sick and that it is Monday is 12%. The probability that it is Monday and not another day of the school week is 20% (there are only five days in the school week). What is the probability that a student has called in sick, given that it is Monday?
9. A neighborhood wanted to improve its parks so it surveyed kids to find out whether or not they rode bikes or skateboards. Out of 2300 children in the neighborhood that ride something, 1800 rode bikes, and 500 rode skateboards, while 200 of those ride both a bike and skateboard. What is the probability that a student rides a skateboard, given that he or she rides a bike?
10. A movie theatre is curious about how many of its patrons buy food, how many buy a drink, and how many buy both. They track 300 people through the concessions stand one evening, out of the 300, 78 buy food only, 113 buy a drink only and the remainder buy both. What is the probability that a patron buys a drink if they have already bought food?
11. A sporting goods store want to know if it would be wise to place sports socks right next to the athletic shoes. First they keep the socks and shoes in separate areas of the store. They track purchases for one day, Saturday,

their busiest day. There were a total of 147 people who bought socks, shoes, or both in one given day. Of those 45 bought only socks, 72 bought only shoes and the remainder bought both. What is the probability that a person bought shoes, if they purchased socks?

12. The following week they put socks right next to the shoes to see how it would affect Saturday sales. The results were as follows; a total of 163 people bought socks, shoes, or both. Of those 52 bought only socks, 76 bought only shoes and the remainder bought both. What is the probability that a person bought socks, if they purchased shoes?
13. A florist wanted to know how many roses and daisies to order for the upcoming valentines rush. She used last year's statistics to determine how many to buy. Last year she sold 52 arrangements with roses only, 15 arrangements with daisies only, and 36 arrangements with a mixture of roses and daisies. What is the probability that an arrangement has at least one daisy, given that it has at least one rose?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 6.6.

6.7 Identifying the Complement

Objective

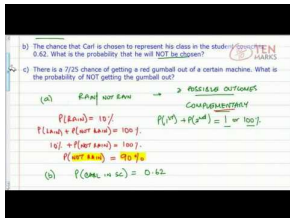
Here you will learn how to identify the complement of a given theoretical or experimental probability.

Concept

What does it mean to find the **complement** of an event? Why would you want to do so?

This lesson is all about what this lesson is not about, so stick about and we'll figure it out!

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63803>

<http://youtu.be/SOHzm-dSJRI0> TenMarksInstructor - Complement of an Event - Probability

Guidance

The **complement** of an event is the sample space of all outcomes that are *not* the event in question. The complement of the event “a flipped coin lands on heads” is “a flipped coin lands on tails”. The complement of “A six-sided die lands on 1 or 2”. Is “A six-sided die lands on 3, 4, 5, or 6”. Complements are notated using the prime symbol ' as in: $P(A')$ is the complement of $P(A)$.

The probability of the complement of an event is always whatever probability it would take to reach 100%. If the probability of pulling a green marble out of a bag is 26%, then the probability of the complement (pulling a *not* green marble) is 74%.

By convention, we most often see probabilities described as either a percentage or a fraction. Despite this, every calculated or experimental probability can be expressed as a value between 0 and 1 since percentages and fractions can all be converted to decimals and a probability must be between 0% and 100%. For example, a probability of $\frac{3}{4}$ or 75% could also be expressed as the decimal .75, and a probability of 20% or $\frac{1}{5}$ could be expressed as the decimal .20.

One benefit of viewing probabilities as decimals is that it is easy to calculate the **complementary** probability of a given event by subtracting the event probability (expressed as a decimal) from 1.

Example A

What is the complement to the event “Brian chooses one of the 2 red shirts from his drawer containing 10 shirts”?



Solution: The complement would be the other possibility: “Brian chooses one of the *not* red shirts from his drawer.”

Example B

If the probability of randomly choosing a Queen from a standard deck of 52 cards is .077, what is the probability of the complementary event?

Solution: The complement would be choosing a card that is *not* a Queen, and the complement probability would be the difference between .077 and 1:

$$P(\text{Queen}') = 1 - .077 = .923$$

Therefore, if the probability of choosing a Queen is 7.7%, then the probability of choosing a card *not* a Queen is 92.3%

Example C

What is the probability of the complement of the event: “Roll a standard die and get an even number”?

Solution: There are three even numbers on a standard die: 2, 4, and 6. That means that the probability that you *do* roll and get an even number is:

$$P(\text{even}) = \frac{3}{6} \text{ or } 50\%$$

Therefore, the complement is:

$$P(\text{even}') = 1 - .50 = .5 \text{ or } 50\%$$

Concept Problem Revisited

The complement of an event is the set of all outcomes that are *not* the event.

Vocabulary

The **complement** of an event is notated using the prime symbol ' such as: “The complement of $P(A)$ is $P(A')$ ”. $P(A')$ is the sample space of all outcomes not a part of $P(A)$, and can be calculated as $1 - P(A)$.

Guided Practice

1. If $P(X) = \frac{1}{6}$, what is $P(X')$?
2. What is the probability of the complement to a probability of 74%?
3. What is the complement to the event: “flipped coin lands on heads”?
4. What is the probability of the complement of randomly choosing one of the 3 quarters from a set of 10 coins?
5. What is the percent probability of Y' if $P(Y) = \frac{1}{8}$?

Solutions:

1. $P(X') = 1 - P(X) = 1 - \frac{1}{6} = \frac{5}{6}$
2. The complement probability is $100\% - 74\% = 26\%$
3. “flipped coin lands on tails”
4. The event probability is $\frac{3 \text{ quarters}}{10 \text{ coins}} = \frac{3}{10}$. The complement is $1 - \frac{3}{10} = \frac{7}{10}$
5. $P(Y') = 1 - P(Y) = 1 - 12.5\% = 87.5\%$

Practice

For problems 1 - 10, identify the **percent probability** of the **complement** of the described event.

1. Roll a standard die once and get an even number.
2. Pull a red card from a standard deck.
3. Pull a face card from a standard deck.
4. Roll two standard dice and get a sum greater than 9.
5. Pull two cards from a deck, without replacement, get at least one face card.
6. Roll a 10-sided die twice, get a 6 both times.
7. The probability that a student in your class likes chocolate is 34%.
8. Of the 76 students in your math class, 26 earned an A.
9. 23% of million-mile cars are Toyotas.
10. A candy machine has 24 green, 32 red, and 14 yellow candies in it. You choose a yellow candy.
11. There are 150 students in your class, 40 have laptops, and 110 have tablets. 26 of those students have both a laptop and a tablet. What is the probability that a randomly chosen student has a tablet, given that she has a laptop?
12. Roll two standard dice, and get 4's on both, given that you know that you have already rolled a 4 on one of them.
13. Draw two cards in a row, without replacement, that are the same suit from a standard deck.
14. Roll of two standard dice once, getting a sum greater than 8, given that one of the dice is a 6.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 6.7.

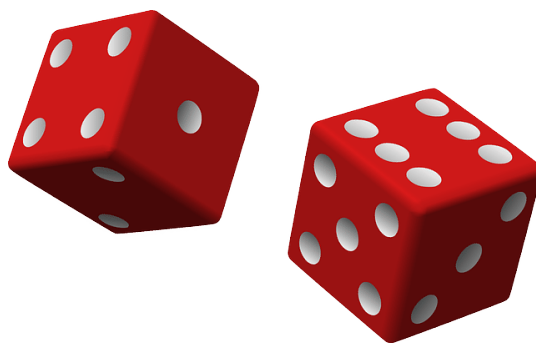
6.8 Finding Probability by Finding the Complement

Objective

Here you will learn how to quickly calculate the probability of an event by finding the probability of the complement.

Concept

Suppose you were asked to calculate the probability of rolling two dice and getting a different number on each. How could you find the answer without needing to enumerate all of the possibilities?



This lesson is about a shortcut to the calculation of some probabilities. We'll return to this question after the lesson.

Watch This

Theoretical Probability
Example 28 - Finding The Complement of an Event
For a standard deck of cards, what is the probability of not drawing a queen?
 $P(\text{not a Q}) = 1 - P(Q)$
 $= 1 - \frac{4}{52}$
 $= \frac{52}{52} - \frac{4}{52} = \frac{48}{52} = \frac{12}{13}$

Mr. Pi

MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/63807>

<http://youtu.be/uofEzYbTric> MrPilarski - How to find the complement of an event and odds

Guidance

Sometimes the probability of an event is difficult or impossible to calculate directly. Particularly when calculating the probability of “at least one. . .” types of problems, or when the sample space of the complement is smaller than that of the event, it may be worth looking first for the probability of the complement of the event you are trying to measure.

Recall that the **complement** of an event is the sample space containing all the outcomes that are not a part of the event itself. That means that the probability of an event + the probability of the complement = 100% or 1.00, or, to

say the same thing as a formula: $P(A) + P(A') = 1$. Once you know the probability of the complement, you can just subtract it from 1 to find the probability of the event.

Example A

What is the probability of randomly drawing a king other than the King of Hearts from a bag containing one king of each standard suit: hearts, clubs, diamonds, and spades.

Solution: Let's say that $P(KH)$ is the probability that you *do* draw the King of Hearts. In that case, $P(KH')$ is the probability of *not* drawing the King of Hearts, the complement of $P(KH)$. Since there are 4 kings and only 1 is the King of Hearts, we can say:

$$P(KH) = \frac{1}{4}$$

therefore:

$$P(KH') = 1 - \frac{1}{4} \text{ or } \frac{3}{4} = 75\%$$

So the probability of *not* drawing the King of Hearts is 75%

Example B

What is the probability of getting a cherry sour from a bag that starts with 9 cherry sours, 5 lemon sours, and 8 lime sours, given that you keep anything you choose and you choose up to 3 times?



Solution: This problem combines conditional probabilities with complements. A key point to this problem is the “keep anything you choose” part. Because we are not putting back the candy after each pull, the probability changes each time, and the chance to get a cherry improves each time we don't choose one. Let's deal with each pull separately at first:

1. First pull: $P(C1) = \frac{9 \text{ cherry}}{22 \text{ candies}} = \frac{9}{22} \text{ or } 41\%$
2. Second pull: $P(C2) = \frac{9 \text{ cherry}}{21 \text{ candies}} = \frac{9}{21} \text{ or } 43\%$
3. Third Pull: $P(C3) = \frac{9 \text{ cherry}}{20 \text{ candies}} = \frac{9}{20} \text{ or } 45\%$

It is tempting to think that the probability of getting at least 1 cherry is just the sum of the three probabilities, but obviously there can't be a $41\% + 43\% + 45\% = 129\%$ chance. The twist is that we only choose a 2nd or 3rd time if

we *don't* get a cherry the time before. That means we need to calculate the chance of choosing a 2nd or 3rd time and multiply the probability of a cherry on that pull by the chance of pulling that time at all.

Fortunately we can use the complement rule to save time. The chance of needing a 2nd pull is the same as the chance that we *don't* pull a cherry on the 1st pull, in other words:

$$P(2nd\ pull) = P(C1') = 1 - 41\% = 59\%$$

The chance of needing a 3rd pull is the same as *not* getting a cherry the 2nd time:

$$P(3rd\ pull) = P(C1') \times P(C2') = 59\% \times 57\% = 34\%$$

Now we can find the overall probability of getting a cherry in three pulls or less:

$$\begin{aligned} P(cherry) &= (\text{chance of cherry of pull 1}) \\ &+ (\text{chance of needing pull 2} \times \text{chance of cherry on pull 2}) \\ &+ (\text{chance of needing pull 3} \times \text{chance of cherry on pull 3}) \\ P(cherry) &= 41\% + (59\% \times 43\%) + (34\% \times 45\%) = 82\% \end{aligned}$$

Example C

What is the probability of rolling two dice and at least one die showing a factor of 6?

Solution: This is an “at least one. . .” problem. To satisfy the requirements, either one of the two dice or both need to land on 1, 2, 3, or 6. That is quite a few possibilities to solve for! However, there are many fewer possible outcomes where *neither die* shows 1, 2, 3, or 6 - in other words, where *both* dice show 4 or 5! There are only 4 ways that could happen:

1. 1st die rolls 5 and 2nd rolls 4
2. 1st die rolls 4 and 2nd rolls 5
3. Both dice roll 4
4. Both dice roll 5

There are $6 \times 6 = 36$ total possible outcomes. If we say that the event “at least one die shows a factor of 6” is A , then A' would be the complement, so we can say:

$$P(A') = \frac{4\ \text{favorable outcomes}}{36\ \text{total outcomes}} = \frac{1}{9}$$

If the *complement* of the event we want to calculate is $\frac{1}{9}$, then the event itself can be calculated as:

$$P(A) = 1 - P(A') = \frac{8}{9}\ \text{or}\ 88.9\%$$

Therefore, we can say that the probability of rolling two dice and getting at least one factor of six is $\frac{8}{9}$.

Concept Problem Revisited

Suppose you were asked to calculate the probability of rolling two dice and getting a different number on each. How could you find the answer without needing to enumerate all of the possibilities?

The probability of rolling two *not* matching numbers is the complement of rolling a matching pair. Since there are only 6 numbers to make matches from, there are only 6 matching pairs. The total number of possible outcomes of two dice is $6 \times 6 = 36$.

$$\therefore P(\text{matching}) = \frac{6 \text{ matches}}{36 \text{ possibilities}} = \frac{1}{6} \text{ or } .167$$

That means that the complement, rolling *not* matching numbers is:

$$1 - .167 = .833 \text{ or } 83.3\%$$

That was quite a bit faster than trying to enumerate all of the possible non-matching rolls!

Vocabulary

The **complement** of an event has a probability equal to 100% minus the probability of the event. As a formula, this looks like: $P(A') = 1 - P(A)$. The probability of the complement of event “A” equals one minus the probability of event “A”.

To **enumerate** a list of things is to mention each member of the list independently.

Guided Practice

1. What is the probability of rolling a number other than 5 on a number cube?
2. What is the probability of choosing a card that isn't a club from a standard deck?
3. What is the probability of not rolling a factor of 10 on a single roll of a 20-sided die?
4. You have a bag of crazy-flavor jellybeans. You know that the flavors are distributed as: 12 earwax, 14 belly-button lint, 26 dog food, 38 gym sock, and 10 of your favorite fruit in the bag. What is the probability that you will be unhappy with the taste of your choice if you choose one jellybean at random?
5. What is the probability of rolling a number that is not a factor of ten or twelve on a single standard die?

Solutions:

1. There are six sides on a number cube, so the probability of rolling any single number is $\frac{1}{6}$. Therefore, we can say: $P(5) = \frac{1}{6}$ so $P(5') = \frac{5}{6}$. Therefore the probability of *not* rolling a 5 is $\frac{5}{6}$.
2. There are four suits, so the probability of choosing a club is $\frac{1}{4}$. The complement is the probability of *not* choosing a club, and it is $1 - \frac{1}{4} = \frac{3}{4}$ or 75%
3. There are 4 factors of 10: 1, 2, 5, and 10. Therefore the probability of rolling a factor of 10 on a 20-sided die is $\frac{4}{20}$. The complement is the probability of *not* rolling one of the four factors: $1 - \frac{4}{20} = \frac{16}{20}$
4. Let $P(F)$ be the probability of pulling a fruit-flavor, then $P(F')$ is the probability of *not* pulling a fruit flavor (and getting something disgusting instead!). There are 10 fruit-flavored beans in the bag of 100 beans, so:

$$P(F) = \frac{10}{100}$$

$$P(F') = 1 - P(F) = \frac{90}{100}$$

Therefore, you have a 90% probability of being unhappy with your choice.

5. The factors of ten or twelve are: 1, 2, 3, 4, 5, 6, 10, and 12. Since all six numbers from a standard die are in that set, the probability of rolling one of them is 100% or 1.0. Therefore, the probability of *not* rolling one of them is $1 - 1 = 0$.

Practice

1. If you randomly pull a single card from a standard deck, what is the probability that the card is anything other than a king?
2. What is the probability of not pulling a face card from a standard deck?
3. What is the probability that a single roll of a 10-sided die will not land on a 7?
4. What is the probability that a single roll of a standard die will be 1, 2, 3, 4, or 5?
5. A candy jar contains 6 red, 7 blue, 8, green, and 9 yellow candies. What is the probability that choosing a single candy at random will result in a piece that is either red, blue, or green?
6. What is the probability that a single choice from the jar in Q 5 will result in a piece that is either red or yellow?
7. What is the probability that a single choice from the same jar will not be blue?
8. You have \$39 in cash, composed of the largest bills possible. What is the probability that a randomly chosen bill from the \$39 will not be a \$1 bill?
9. There are 450 songs on the .mp3 player you share with your father and sister. If your dad has 125 80's songs, and your sister has twice that many country music songs, what is the percent probability that a randomly chosen song will not be one of yours?
10. What is the percent probability that a random roll of a fair die will not result in an even or prime number?
11. The train station has 47 active trains. 5 are late by less than 10 minutes, 4 are between 11 and 15 minutes late, and 10 are more than 15 minutes late. What is the probability that a randomly chosen train will be on time?
12. The probability that a student has called in sick and that it is Monday is 12%. The probability that it is Monday and not another day of the school week is 20% (there are only five days in the school week). What is the probability that a student has not called in sick, given that it is Monday?
13. A neighborhood wanted to improve its parks so it surveyed kids to find out whether or not they rode bikes or skateboards. Out of 2300 children in the neighborhood that ride something, 1800 rode bikes, and 500 rode skateboards, while 200 of those ride both a bike and skateboard. What is the probability that a student does not ride a skateboard, given that he or she rides a bike?
14. A movie theatre is curious about how many of its patrons buy food, how many buy a drink, and how many buy both. They track 300 people through the concessions stand one evening, out of the 300, 78 buy food only, 113 buy a drink only and the remainder buy both. What is the probability that a patron does not buy a drink if they have already bought food?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 6.8.

Students were introduced to many of the major concepts involved with probability studies. Basic instruction and examples include for the concepts of: singular and compound events, conditional probabilities, compliments and inclusive and exclusive events.

6.9 References

1. Jarkko Laine. <https://www.flickr.com/photos/jarkkolaine/8184789759> .
2. Jamie. <https://www.flickr.com/photos/jamiesrabbits/5791400156/> .
3. Expert Infantry. <https://www.flickr.com/photos/expertinfantry/5458168907/> .
4. CK-12 Foundation. . CCSA
5. CK-12 Foundation. . CCSA
6. . . CC BY-NC-SA
7. . . CC BY-NC-SA
8. PhotoAtelier. <https://www.flickr.com/photos/glenbledsoe/12126341973> .
9. . . CC BY-NC-SA
10. . . CC BY-NC-SA
11. Vivid Image. https://www.flickr.com/photos/minnesota_social_marketing/8511416728 .
12. OpenClips. <http://pixabay.com/en/people-dance-dancing-silhouette-160789/?oq=people> .
13. CK-12 Foundation. . CCSA
14. OpenClips. <http://pixabay.com/en/card-deck-deck-cards-playing-cards-161536/?oq=card%20deck> .
15. Images Money. <https://www.flickr.com/photos/59937401@N07/5857256935> .
16. . . CC BY-NC-SA
17. OpenClips. <http://pixabay.com/en/t-shirt-shirt-red-clothing-156081/?oq=red%20t> .
18. Nemo. <http://pixabay.com/en/red-icon-two-recreation-cartoon-25637/?oq=dice> .
19. Mariolh. <http://pixabay.com/en/jellies-candies-sweets-colors-204206/?oq=cAndy> .

CHAPTER

7

Probability Distribution

Chapter Outline

- 7.1 UNDERSTANDING DISCRETE RANDOM VARIABLES**
 - 7.2 UNDERSTANDING CONTINUOUS RANDOM VARIABLES**
 - 7.3 PROBABILITY DISTRIBUTION**
 - 7.4 VISUALIZING PROBABILITY DISTRIBUTION**
 - 7.5 PROBABILITY DENSITY FUNCTION**
 - 7.6 BINOMIAL EXPERIMENTS**
 - 7.7 EXPECTED VALUE**
 - 7.8 RANDOM VARIABLE VARIANCE**
 - 7.9 TRANSFORMING RANDOM VARIABLES I**
 - 7.10 TRANSFORMING RANDOM VARIABLES II**
 - 7.11 REFERENCES**
-

Probability distributions are explanations of the probabilities that a random process will result in each of the possible specific outcomes. For instance, the probability distribution of a single roll of a standard die might look like:

- 1: 16.7%
- 2: 16.7%
- 3: 16.7%
- 4: 16.7%
- 5: 16.7%
- 6: 16.7%

Different types of probability measures require different probability distributions. Continuous random variables, for instance, have an infinite number of possible outcomes in any given interval, making it impossible to create a list of individual probabilities like the ones for the die above. In this chapter, you will learn how to create and interpret all kinds of probability distributions.

7.1 Understanding Discrete Random Variables

Objective

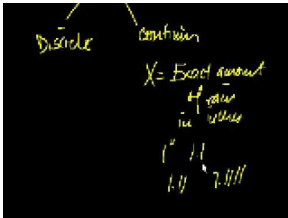
In this lesson, you will learn what discrete random variables are, and how they apply to the study of probability.

Concept

What is the purpose of a random variable? How do random variables differ from algebraic variables?

Look to the end of the lesson for the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67352>

<http://youtu.be/IYdiKeQ9xEI> Khan Academy - Introduction to Random Variables

Guidance

You are familiar with variables through your extensive use of them in Algebra. In this lesson we are going to introduce the concept of **random variables**. A random variable assigns a unique *numerical* value to the outcome of a *random* experiment. Note how I pointed out the importance that the values are numerical and that they are chosen randomly. It is also important to note that the total probability of all of the possible values of the random variable should be 100%.

Random variables are often used to represent the number of times you get a specific result of a random process. For instance, the random variable C might represent the number of times you get a total of nine by rolling two six-sided dice three times.



A **discrete random variable** is a random variable with a specific and countable set of possible values. If a set is defined as a series of specific values, whether there are 5, or 10, or 100,000,000 values, that set could be described by a discrete random variable.

Example A

If T is the outcome of flipping a fair coin once, is T a random variable?

Solution:

No, T is not a random variable, because it is not a *numeric* result. The outcome of flipping a coin would be either “heads” or “tails”. To make T a random variable, you would have to set a number for each outcome, such as: $T = 1$ if heads and 2 if tails.

Example B

If Trina designates Y to be the number of yellow marbles she gets during nine trials of randomly pulling 1 marble from a bag filled with marbles of various colors and returning it, is Y a random variable?

Solution:

Yes, Y is a random variable, since it is the random numerical result of a limited number of independent trials of an experiment.

Example C

If N is the number of nines you get when rolling two standard dice three times:

- Is N a binomial random variable?
- What are the possible values of N ?
- What is the probability distribution of N ?

Solution:

- N is a binomial random variable, because it is the result of a specific and limited number of independent trials of a random process and each outcome is either nine or not nine.
- Since you could only roll a total of 9 once each trial, N could be 0, 1, 2, or 3.
- Since there are only four ways that a 9 could be rolled: $6 + 3$, $5 + 4$, $4 + 5$, and $3 + 6$, out of the $6 \times 6 = 36$ possible combinations, your chances of getting a 9 on any given roll would be $\frac{4}{36} = \frac{1}{9}$. The probabilities of each of the possible values of N would be:

- $N = 0$ (You do not roll 9 at all): $\frac{8}{9} \times \frac{8}{9} \times \frac{8}{9} = \frac{512}{729}$
- $N = 1$ (You roll a single 9, no more or less): $\frac{1}{9} \times \frac{8}{9} \times \frac{8}{9} = \frac{64}{729} \times 3 = \frac{192}{729}$
 - Multiply by 3, because there are 3 possible positions for the single “9”: 1st, 2nd, or 3rd
- $N = 2$ (You roll 9 exactly twice): $\frac{1}{9} \times \frac{1}{9} \times \frac{8}{9} = \frac{8}{729} \times 3 = \frac{24}{729}$
 - Multiply by 3, because there are 3 possible positions for the single “not 9”
- $N = 3$ (Three 9’s in a row): $\frac{1}{9} \times \frac{1}{9} \times \frac{1}{9} = \frac{1}{729}$

Concept Problem Revisited

What is the purpose of a random variable? How do random variables differ from algebraic variables?

A random variable is used to simplify the expression of the possible outcomes of a random process. Random variables differ from algebraic variables most prominently in that algebraic variables commonly only represent a single value, where as random variables represent a range of possible numeric outcomes *and* the associated probability of each.

Vocabulary

A **random variable** is the numeric result of a specific and limited number of independent trials of a random process.

A **discrete random variable** has a specific and countable number of possible values.

A **continuous random variable** is a random variable that can take on all values in an interval. For instance, if a continuous random variable can be any value in the interval between 0 and 1, then it could be .1, .11, .111, .1111, etc. There are an infinite number of possible values in any given interval.

Guided Practice

1. If random variable A represents the age of a single randomly chosen student from your classroom, is A a discrete random variable?
2. If B is the age of a randomly chosen student in your classroom, is B a discrete random variable?
3. If C is defined by the function $C = \frac{102}{3}$, is C a discrete random variable?

Solutions:

1. No, A is not discrete because age could represent any of an infinite number of values based on how accurately you measure.
2. No, B is not discrete because although there must be a limited number of students in the classroom, the number of possible ages is not countable (15 years, 5 months, 3 days, 2 hours, 10 minutes and 1 second... or 2 seconds... or 2.33 seconds, etc.)
3. No, based on the given function, C is not random since it will always represent 34.

Practice

For questions 1-10, state why the example does or does not describe a random variable:

1. K is the number of Kings you get over 10 trials of randomly drawing 1 card at a time from a standard deck.
2. S is the number of cards you draw before drawing the Seven of Hearts.
3. One person is chosen at random from a classroom, H is her height in inches.
4. A fair coin is flipped 10 times, T is the number of tails.

5. B is the number of barks you hear in the city park during one-sixth of an hour (10 minutes), the sixth of an hour you choose to listen is selected by roll of a fair die.
6. S is the sound you hear first during a ten minute period at the park, the 10 minute period is chosen by the roll of a fair die.
7. N is the number that makes the statement $N + 4 = 9$ true.
8. Choose a number between 1 and 10, multiply it by the age, in years, of the person next to you and add your own age in years. Do this for 5 people. A is the average of the results from the six trials.
9. H is the average maximum vertical jump of the girls on the girls' basketball team.
10. H is the number of jumps greater than 22 inches in a basketball game.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.1.

7.2 Understanding Continuous Random Variables

Objective

In this lesson, you will learn the difference between a discrete and a continuous random variable, and will learn of a few examples of real-world uses of continuous random variables.

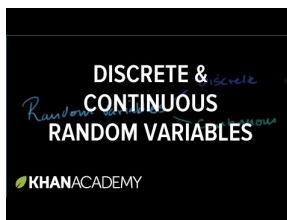
Concept

If H is the height of 5 boys chosen at random from your school. Is H a continuous random variable or is it discrete? What is it about the experiment that makes H continuous or discrete? What could you do to change it?

Look for the answers at the end of the lesson.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67358>

<http://youtu.be/dOr0NKyD31Q> Khan Academy - Discrete and Continuous Random Variables

Guidance

Recall that *random variables* assign numeric values to the outcomes of independent random events. A *discrete random variable* is used to represent a specific and countable number of values, such as the number of people in a room, or the number of times you roll a four during ten rolls of a fair die.

A **continuous random variable** is used to represent all of the possible values in a particular interval, such as the distance around a randomly chosen lake, or the weight of a randomly chosen rock from a pile. In either of these cases, the result could literally be an infinite number of possibilities, since the exact distance around the lake (around each molecule of water), and the exact weight of the rock (again, to the smallest of measures) is impossible to pinpoint.

In short: If the possible values of a random variable are countable, it is a discrete random variable. If the values are uncountable, it is a continuous random variable.

Example A

If Brian uses the variable Y to represent the top bicycling speed of a randomly chosen student in class, is Y a continuous random variable?

Solution: Yes, speed is a common continuous variable, and the value is chosen by a random process. We know it is continuous because there is always another possible value between any two speed values. It would not be possible to count all of the possible speeds that Y could be.

Example B

Would the arm length of a randomly chosen preschooler be represented by a continuous random variable?

Solution: Yes, length is another common continuous variable. There is always another possible value for length between any two values.

Example C

Would a continuous variable be used to represent the number of people in your city?

Solution: No, even though there may be a huge number, the number of people in your city is finite, and there are no fractions or decimals of people.

Concept Problem Revisited

If H is the total height of 5 boys chosen at random from your school. Is H a continuous random variable or is it discrete? What is it about the experiment that makes H continuous or discrete? What could you do to change it?

H is a continuous random variable. We know it is continuous because the possible height values aren't countable. One way to make H a discrete random variable would be to say that $H = 1$ if the total heights were less than 25 feet, $H = 2$ if between 25 and 30 feet, and $H = 3$ if more than 30 feet. Then H would be discrete with the possible values 1, 2, and 3, which are quite countable.

Vocabulary

A **random variable** is the numeric result of a specific and limited number of independent trials of a random process.

A **discrete random variable** has a specific and countable number of possible values.

A **continuous random variable** is a random variable that can take on all values in an interval. For instance, if a continuous random variable can be any value in the interval between 0 and 1, then it could be .1, .11, .111, .1111, etc. There are an infinite number of possible values in any given interval.

Guided Practice

1. If B represents the mean time spent baking a batch of cookies by 5 students chosen at random from Home Ec class, is B a continuous random variable?
2. Would you use a continuous or discrete random variable to represent the number of atoms in a baseball?

- Roll a fair die three times, the total is the number of minutes you run water from a hose into a small pool. Would the volume of water in the pool be represented by a continuous or discrete random variable? What about the number of minutes you run the hose?
- Roll two dice, the total is the number of people you count before asking the next how many games they purchased. If C is the number of dollars he or she spent, is C discrete or continuous?

**Solutions:**

- Yes, time is continuous, and the result of the experiment is a random value.
- Even though there is a seemingly uncountable number of atoms, the number is finite. Therefore, this would be a discrete variable.
- Volume is continuous, so the amount of water would be represented by a continuous random variable. The number of minutes is countable, so it would be a discrete variable.
- C is discrete, money is measured in specific denominations.

Practice

For questions 1-10, assume the “student” in the situation is chosen at random, and state whether the variable is discrete or continuous.

- The body temperature of the student.
- The number of brothers or sisters the student has.
- The exact age of the student.
- The number of classes the student signed up for this semester.
- The number of books the student has in his or her bag.
- The student’s favorite teacher.
- The weight of the student.
- The percentage of students in the school taller than the student, to the nearest 1%.
- The student’s G.P.A.
- The mean number of people in the student’s classes.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.2.

7.3 Probability Distribution

Objective

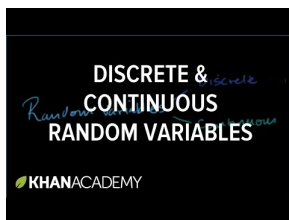
In this lesson, you will learn about probability distributions and how they describe the probabilities associated with different possible values of a random variable.

Concept

Assume you have an unfair coin that is weighted to land on heads 65% of the time. If you flip that coin 3 times and let T represent the number of tails you get, what is the probability distribution for T ?

Look to the end of the lesson for the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67358>

<http://youtu.be/s2S1oD3ovps> statslectures - Discrete Probability Distributions

Guidance

A **probability distribution** is a list of each value a random variable can attain, along with the probability of attaining each value. In other words, the probability distribution of an event is sort of a map of how each possible outcome relates to the chance it will happen.



For instance, the probability distribution of flipping a coin twice is:

heads, heads = 25%, heads, tails = 25%, tails, heads = 25%, and tails, tails = 25%.

If we define the random variable X to be the number of heads you get when you flip a coin twice, we could create the following probability distribution table for X :

TABLE 7.1:

X	0	1	2
$P(X)$	$\frac{1}{8}$	$\frac{3}{4}$	$\frac{1}{8}$

There are various ways of visualizing a probability distribution, and we will review that concept in another lesson. For now, we focus on identifying what a probability distribution is, and how to calculate it for a particular event.

Example A

In Chi's class, 4 students have one parent, 7 have two parents, and 1 student lives with his uncle. Let P be the number of parents of a randomly selected student from the class. Create a probability distribution for P .

Solution:

Set random variable P to be the number of parents:

$$P(P) = \% \text{ probability that a student has } P \text{ parents}$$

Now find the probability of each P , noting that there are 12 students total:

$$\begin{aligned} \text{1 student has 0 parents: } P(0) &= \frac{1}{12} \text{ or } 8.3\bar{3}\% \\ \text{4 students have 1 parent: } P(1) &= \frac{4}{12} \text{ or } 33.3\bar{3}\% \\ \text{7 students have 2 parents: } P(2) &= \frac{7}{12} \text{ or } 58.3\bar{3}\% \end{aligned}$$

Example B

Roll two fair six-sided dice. Let D equal the sum of the dice. Create a probability distribution for D .

Solution: Make a list of the individual probabilities of each of the 36 possible outcomes:

1 possibility with a sum of 2:	$P(D = 2) = \frac{1}{36} = 0.0278$
2 possibilities with a sum of 3:	$P(D = 3) = \frac{2}{36} = 0.0556$
3 possibilities with a sum of 4:	$P(D = 4) = \frac{3}{36} = 0.0833$
4 possibilities with a sum of 5:	$P(D = 5) = \frac{4}{36} = 0.1111$
5 possibilities with a sum of 6:	$P(D = 6) = \frac{5}{36} = 0.1389$
6 possibilities with a sum of 7:	$P(D = 7) = \frac{6}{36} = 0.1667$
5 possibilities with a sum of 8:	$P(D = 8) = \frac{5}{36} = 0.1389$
4 possibilities with a sum of 9:	$P(D = 9) = \frac{4}{36} = 0.1111$
3 possibilities with a sum of 10:	$P(D = 10) = \frac{3}{36} = 0.0833$
2 possibilities with a sum of 11:	$P(D = 11) = \frac{2}{36} = 0.0556$
1 possibility with a sum of 12:	$P(D = 12) = \frac{1}{36} = 0.0278$

Example C

Janie wants to evaluate the probabilities of pulling various cards from a deck. She sets the discrete random variable C to be the number of diamonds she gets over the course of three trials, if each trial consists of pulling, recording, and replacing one random card from a standard deck. What is the probability distribution of C ?



Solution: To evaluate the probability distribution of C , Janie needs to identify the probability of each of the possible values of C . Note that the chance she will pull a diamond is $\frac{13}{52}$ or .25, meaning that the chance she will *not* pull a diamond is $1 - .25 = .75$:

- **For $C = (1)$** , the total probability is: $.14 + .14 + .14 = .42$ or 42% (see the three possible outcomes resulting in $C = 1$ below)

- Diamond, other, other : $.25 \times .75 \times .75 = .14$
- Other, Diamond, other : $.75 \times .25 \times .75 = .14$
- Other, other, Diamond : $.75 \times .75 \times .25 = .14$
- For $C = (2)$, the total probability is: $.047 + .047 + .047 = .141$ or 14.1%
 - Diamond, Diamond, other : $.25 \times .25 \times .75 = .047$
 - Diamond, other, Diamond : $.25 \times .75 \times .25 = .047$
 - Other, Diamond, Diamond : $.75 \times .25 \times .25 = .047$
- For $C = (3)$, the probability is : $.25 \times .25 \times .25 = .016$ or 1.6%
 - Diamond, Diamond, Diamond: $.25 \times .25 \times .25 = .016$

Concept Problem Revisited

Assume you have an unfair coin that is weighted to land on heads 65% of the time. If you flip that coin 3 times and let T represent the number of tails you get, what is the probability distribution for T ?

If each throw has a 65% chance of heads, then it has a 35% chance of tails:

- For $T = 1$, we could have THH, HTH, or HHT. Each of those has a $.35 \times .65 \times .65 = .15$ chance of occurring, so $P(T = 1) = .15 \times 3 = .45$ or 45%
- For $T = 2$, we could have TTH, THT, or HTT. Each has a $.35 \times .35 \times .65 = .08$ chance, so $P(T = 2) = .08 \times 3 = .24$ or 24%
- For $T = 3$, we could have only TTT, with a chance of $.35 \times .35 \times .35 = .043$ or 4.3%

Vocabulary

A **probability distribution** is a list of each value a random variable can attain, along with the probability of attaining each value.

Guided Practice

- Create a probability distribution for number of heads when you flip a coin 3 times.
- Let C be the number of chocolate chip cookies you get if you randomly pull and replace two cookies from a jar containing 6 chocolate chip, 4 peanut butter, 8 snickerdoodle, and 12 sugar cookies. Create a probability distribution for C .
- Let S be the score of a single student chosen at random from Mr. Spence's class. Create a probability distribution for S , given the following:

TABLE 7.2:

Number of Students	Test
11	87
7	89
13	92
9	94
6	96

Solutions:

- Write out all the possibilities:

TTT has 0 heads.

TTH has 1 heads.

THT has 1 heads.

TTH has 2 heads.

HTT has 1 heads.

HTH has 2 heads.

HHT has 2 heads.

HHH has 3 heads.

So we have 1 possibility with 0 heads:

$$P(0) = \frac{1}{8} = 0.125$$

3 possibilities with 1 heads:

$$P(1) = \frac{3}{8} = 0.375$$

3 possibilities with 2 heads:

$$P(2) = \frac{3}{8} = 0.375$$

1 possibility with 3 heads:

$$P(3) = \frac{1}{8} = 0.125$$

2. There are a total of 30 cookies, the probability of pulling a chocolate chip cookie is $\frac{6}{30} = .20$, so the probability of not pulling a chocolate chip is $\frac{24}{30} = .80$

- For $C = 0$ we have to pull a non-chocolate chip both times: $.8 \times .8 = .64$ or 64%
- For $C = 1$ we could either pull the chocolate chip cookie first or second, so we get $(.2 \times .8) + (.8 \times .2) = .32$ or 32%
- For $C = 2$ we have to pull chocolate chip both times, so we have $.2 \times .2 = .04$ or 4%

3. There are a total of 46 students in Mr. Spence's class, so there are 46 scores. The probability of a random student having score S is the same as that score's portion of the total number of scores:

- $P(S = 87) = \frac{11}{46}$
- $P(S = 89) = \frac{7}{46}$
- $P(S = 94) = \frac{9}{46}$
- $P(S = 96) = \frac{6}{46}$

Practice

1. What is a probability distribution?
2. What is a random variable?
3. What is the difference between a discrete and a continuous random variable?

For problems 4-7, refer to the following table:

TABLE 7.3:

S	2	3	4	5	6	7	8	9	10
$P(S)$.04		.12	.16		.16	.12		.04

4. Assuming the table is a probability distribution for discrete random variable S , which is the sum of two dice rolled once, how many sides does each die have?
5. What is $P(3)$?
6. What is $P(6)$?

7. What is $P(9)$?
8. Roll two seven-sided dice once. Let S be the sum of the two dice. Create a probability distribution for S .
9. Flip a fair coin 3 times, let H be the number of heads. Create a probability distribution for H .
10. Let S be the sum of two standard fair dice. Create a probability distribution for S , if the experiment consists of a single roll of both dice.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.3.

7.4 Visualizing Probability Distribution

Objective

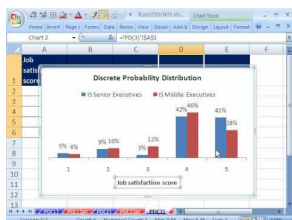
Here you will learn how to create a visual representation of a probability distribution.

Concept

Writing down all of the various probabilities of outcomes of an event is fine, but it can get a little tedious both to create and to read a long list of different probabilities. How else can we display the information from a probability distribution?

Watch This

This video is a quick lesson on how to create a discrete probability distribution in “Excel”, and the process is similar in any standard spreadsheet software. Statistics and spreadsheets go very much hand-in-hand, so I certainly recommend you begin practicing with one, if you have not already. If you do not have access to Excel, there is a very similar and free software called “OpenOffice” available online.



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67354>

<http://youtu.be/qSu-Rk-6apw> ExcellsFun - Discrete Probability Chart

(This is video 46 in the series from “ExcellsFun”, if you would like a more detailed explanation on how to begin setting up the distribution, you may wish to watch the (much longer) video #45.)

Guidance

Probability distributions can convey a fantastic amount of useful information, but there may be so much information to view that the important points get lost in the data. Because of this, it is very common to create a graphical representation of the data to highlight important or interesting values.

Tables, histograms and bar charts in particular are excellent means of visualizing the data from discrete probability distributions. If you use a histogram or bar chart, by enumerating the various outcomes along the x -axis and the expected probability of occurrence on the y -axis, you create a very concise and easily read summary of the distribution of outcome probabilities.

Example A

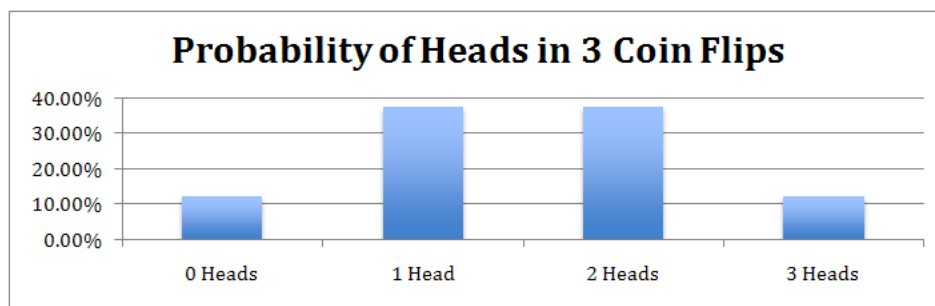
Let C be a discrete random variable representing the number of heads that might result from flipping a coin three times. Create a bar chart to illustrate the probability distribution of C .

Solution: Start by identifying the possible outcomes of flipping a coin three times:

TTT has 0 heads.	TTH has 1 heads.
THT has 1 heads.	TTH has 2 heads.
HTT has 1 heads.	HTH has 2 heads.
HHT has 2 heads.	HHH has 3 heads.

So we have 1 possibility with 0 heads: $P(0) = \frac{1}{8} = 0.125$

3 possibilities with 1 heads :	$P(1) = \frac{3}{8} = 0.375$
3 possibilities with 2 heads :	$P(2) = \frac{3}{8} = 0.375$
1 possibility with 3 heads :	$P(3) = \frac{1}{8} = 0.125$

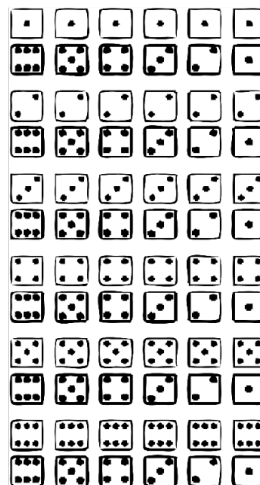


Example B

Create a table showing the probability distribution of the possible outcomes of rolling two standard dice.

Solution: Let random variable S represent the sum of the pips showing on the roll of both dice. We know then than $2 \leq S \leq 12$.

Find all of the possible outcomes of rolling two dice, as shown in the image on the below.



Create a table showing the probabilities of each possible outcome of S :

TABLE 7.4:

S	$P(S)$	S	$P(S)$
2	$\frac{1}{36}$	8	$\frac{5}{36}$
3	$\frac{2}{36}$	9	$\frac{4}{36}$
4	$\frac{3}{36}$	10	$\frac{3}{36}$
5	$\frac{4}{36}$	11	$\frac{2}{36}$
6	$\frac{5}{36}$	12	$\frac{1}{36}$
7	$\frac{6}{36}$		

Example C

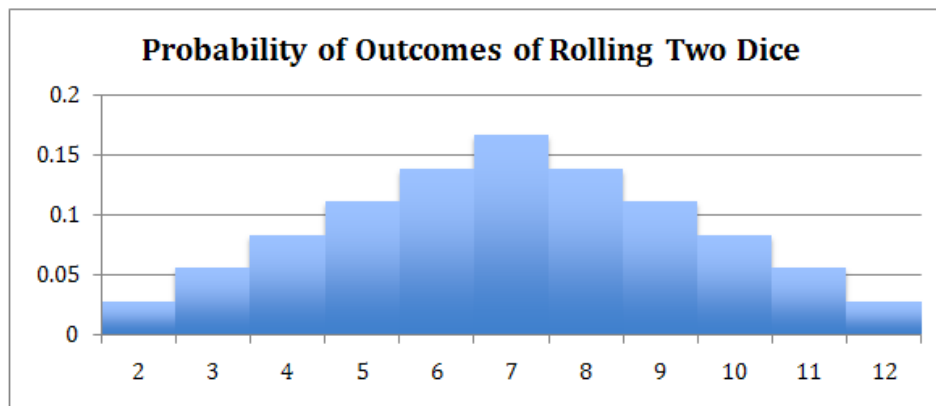
Create a probability histogram of the possible outcomes of rolling two dice. You may use your data from Example B.

Solution: In Example B, we created a table of the probabilities of each outcome of rolling two dice, designated as discrete random variable S . Let's add one more column for each value so we can convert the fractional probability to decimal:

TABLE 7.5:

S	$P(S)$	$P(S)$ <i>decimal</i>	S	$P(S)$	$P(S)$ <i>decimal</i>
2	$\frac{1}{36}$.028	8	$\frac{5}{36}$.139
3	$\frac{2}{36}$.056	9	$\frac{4}{36}$.111
4	$\frac{3}{36}$.083	10	$\frac{3}{36}$.083
5	$\frac{4}{36}$.111	11	$\frac{2}{36}$.056
6	$\frac{5}{36}$.139	12	$\frac{1}{36}$.028
7	$\frac{6}{36}$.167			

We can use this data to create a histogram, setting the y -axis to the probability and the x -axis to the values of S :

**Concept Problem Revisited**

Writing down all of the various probabilities of outcomes of an event is fine, but it can get a little tedious both to create and to read a long list of different probabilities. How else can we display the information from a probability distribution?

Tables, histograms, bar graphs and pie charts are the most common visual representations of probability distributions.

Vocabulary

A *histrogram* is a specific form of a bar chart where the bars are proportional in area to the frequency and proportional in width to the interval represented by the bar.

Guided Practice

1. Let R be a discrete random variable representing the number of red marbles pulled over three trials of pulling and replacing one marble out of a bag containing 4 red, 4 yellow, and 4 green marbles. Create a probability distribution table for R .
2. Let S be a discrete random variable representing the number of 2's you spin over 5 spins on a spinner with 4 equally-spaced points. Create a histogram showing the probability distribution of S .
3. Create a probability distribution table for the outcomes of the sum of two 5-sided dice.

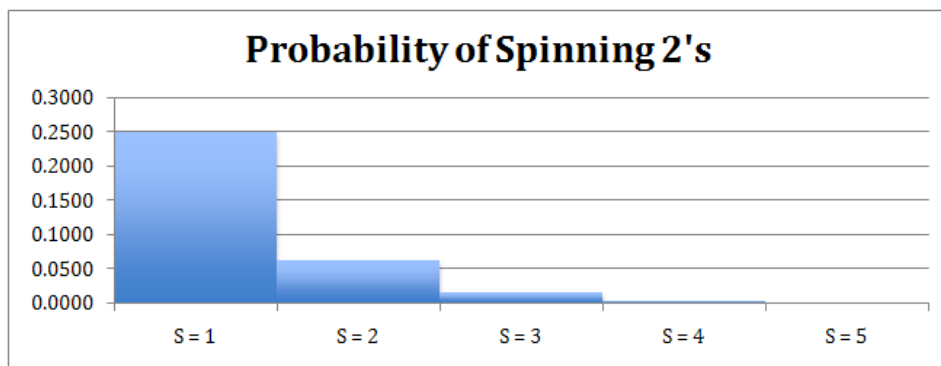
Solutions:

1. The possible values of R are 1, 2, and 3. There is a $\frac{1}{3}$ chance of red on each pull. The probability distribution for R would thus be:

TABLE 7.6:

$P(1)$.333
$P(2)$	$(.333)^2 = .111$
$P(3)$	$(.333)^2 = .037$

2. The possible values of S are 1, 2, 3, 4, and 5. There is a $\frac{1}{4}$ chance of a 2 on each spin:



3. Let's start by creating a grid to show all of the possible combinations:

TABLE 7.7:

	1	2	3	4	5
1	2	3	4	5	6
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10

Now we can create a distribution based on the probability of each possible outcome 2-10, let R be a discrete random variable representing the sum of the dice:

TABLE 7.8:

R	2	3	4	5	6	7	8	9	10
$P(R)$	$\frac{1}{25} = .04$	$\frac{2}{25} = .08$	$\frac{3}{25} = .12$	$\frac{4}{25} = .16$	$\frac{5}{25} = .20$	$\frac{4}{25} = .16$	$\frac{3}{25} = .12$	$\frac{2}{25} = .08$	$\frac{1}{25} = .04$

Practice Problems

1. Create a probability distribution table for a single roll of two 7-sided dice.
2. Create a histogram to visualize the data from problem 1.
3. Create a pie chart showing the same data.
4. There are 12 green, 9 blue, and 4 red candies in an opaque bag. Let R be a discrete random variable representing the number of red candies you get *in a row* by pulling and replacing one candy four times. Create a probability distribution table illustrating the possible outcomes of R .
5. Create a histogram illustrating the information from problem 4.
6. Create a pie chart showing the same data.
7. Let discrete random variable S represent the number of 7's you get when rolling two 5-sided dice three times. Create a probability distribution table for S .
8. Create a histogram illustrating the information from problem 7.
9. Create a pie chart with the same information.
10. Let T be the number of tails you get when you flip a fair coin 4 times, create a probability distribution table for T .
11. Create a histogram or bar chart for T , from problem 10.
12. Create a pie chart for T from problem 10.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.4.

7.5 Probability Density Function

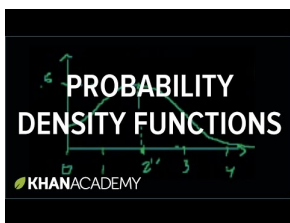
Objective

Here you will learn about visualizing *probability density functions*, which are probability distributions of continuous random variables.

Concept

Bar graphs and histograms are great for visualizing the probability distributions of *discrete random variables*, but neither of those are appropriate for *continuous random variables*, since there would need to be an infinite number of bins or columns to represent the possible outcomes. How then do we visualize the probability of a continuous random variable?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67360>

http://youtu.be/Fvi9A_tEmXQ Khan Academy - Probability Density Functions

Guidance

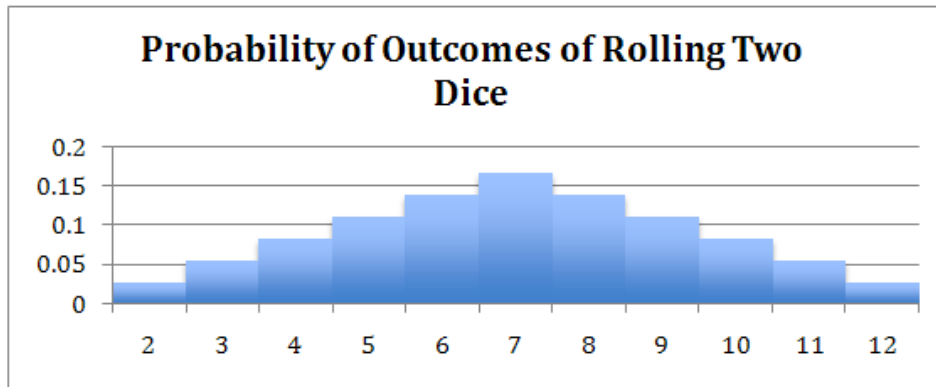
When it comes to describing the probability of a particular value of a random variable, there is one particularly *huge* difference between discrete and continuous variables:

- The probability that a *continuous* random variable will take on a particular value is, for all intents and purposes, **always zero**. Therefore, we calculate the probability that the variable will take on a particular value within a *range* of values.
- The probability that a discrete random variable will take on a particular value can be calculated for that value.

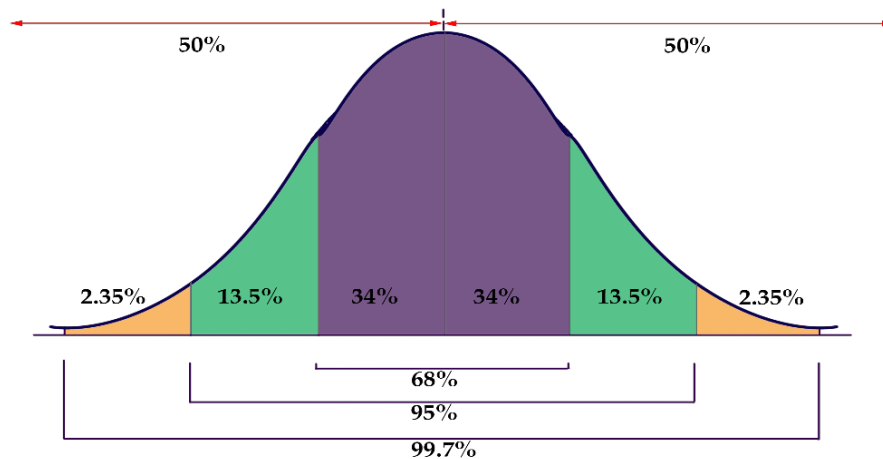
Perhaps the most immediately obvious result of this difference in the two types of variables appears in the way we visualize the probability distribution of each:

- **Discrete Random Variable:** A bar chart or histogram illustrating the probability that a discrete random variable will take on a particular value is straightforward: the height of the bar represents the probability that the variable will take on the value represented by the bar. In the example below, we have a bar graph from a previous lesson showing the probability distribution of rolling two dice. We can see by looking at the graph that the probability that the two dice will total seven is between 15% and 20% (the actual value is apx 16.7%).

The number of possible outcomes is specific and limited, and it is clear that if we were to roll the two dice, we would expect to see exactly seven pips relatively often.



- Continuous Random Variable:** A continuous random variable does not represent a specific or countable number of outcomes. Between any two values of a continuous variable, there are always more values. Because of this, we cannot visualize the distribution of probabilities with bars, since there is no way to create a bar for each possible value of the variable. The solution is to use a graph to identify the probability that a particular value we are interested in will occur within a *range* of values. This graph, called a **probability density** graph, illustrates that probability as the area under a portion of a curve like the **normal curve** below.

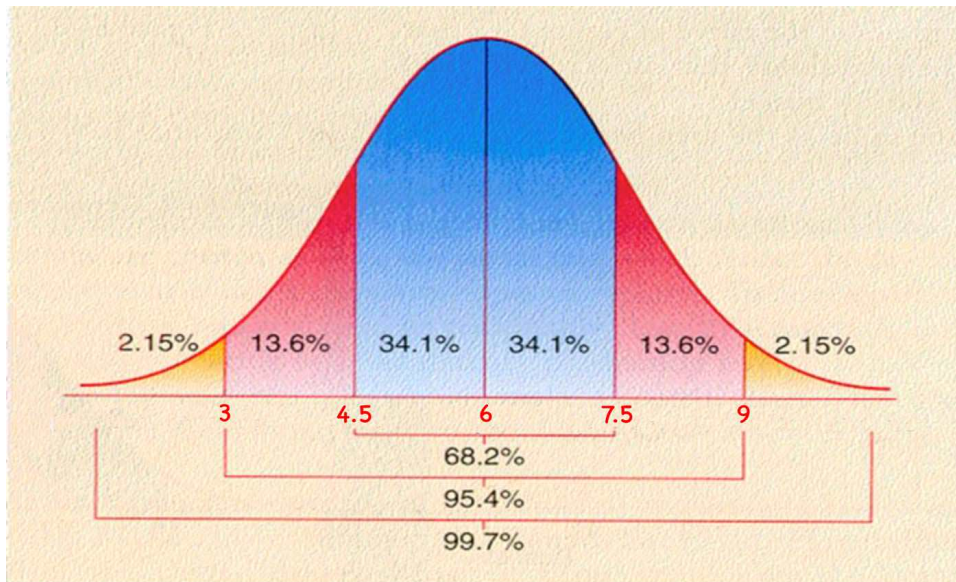


This probability density graph is a representation of a **normal curve**, which will be spending more time on later. For now we use it to learn how a probability density graph works.

The center of the graph is the tallest point, and represents the **mean** of the data. Each of the colors in this image represent one **standard deviation** (sort of a standard 'step size' - we will discuss this more later) from the mean. The percentages marked on the image describe the probability that a random selection from the data represented on the graph will fall within that range.

Example A

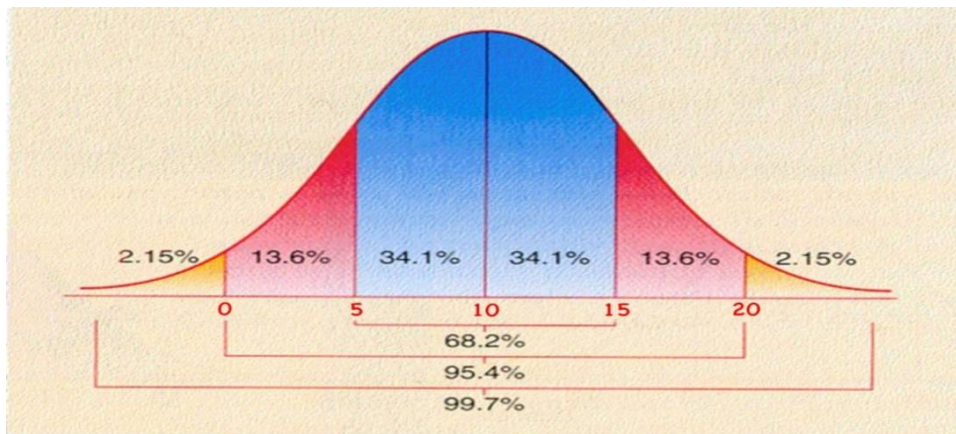
What is the probability that a random value selected from the normal graph below will be between 3 and 9?



Solution: According to the graph, the mean of the data is 6, and the standard deviation is 1.5. We can see from the color-coding that the area between two standard deviations below and 2 standard deviations above the mean represents a probability of 95.4%.

Example B

Look at the image below, where we have a normal curve with a mean of 10, and a standard deviation of 5. What is the probability that a randomly chosen value from the function illustrated will be between 5 and 15?



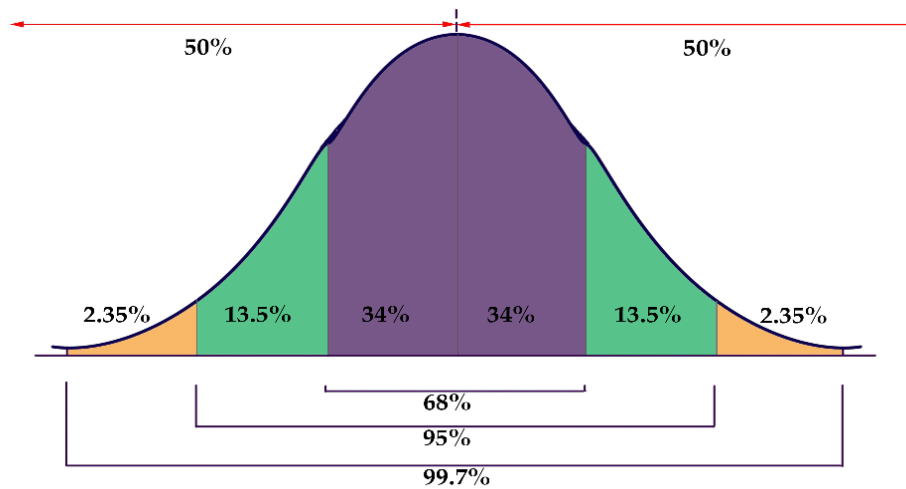
Solution:

As you can see by the area colored blue, there is a 68.2% probability that a value within 5 units of 10 will occur. So there is approximately a two-thirds probability that a random value selected from the data will be between 5 and 15.

Example C

Use the image of the normal curve below to answer questions a-c:

- How much of the area under the curve below is located to the left of the mean?
- How much of the total area is located to the right of +1 standard deviation?
- How much area is located to the left of +2 standard deviations?

**Solution:**

- On a normal curve, the median is the same as the mode, which is the center of the graph. Since the center is the median, one-half of the data is to the left, and one-half to the right.
- The entire graph is 100% of the data, and half of it is to the left of the median. One standard deviation incorporates another 34.1%, so the total area to the *left* of the +1 standard deviation mark would be 50% + 34.1% = 84.1%. **The area to the right of the mark is 100% – 84.1% = 15.9%.**
- The area below two SD's above the mean would incorporate the 50% below the median, plus the 34.1% between the median and +1 SD, plus the 13.6% between +1 SD and +2 SD's = 97.7%.

Concept Question Revisited

Bar graphs and histograms are great for visualizing the probability distributions of discrete random variables, but neither of those are appropriate for continuous random variables, since there would need to be an infinite number of bins or columns to represent the possible outcomes. How then do we visualize the probability of a continuous random variable?

Since the probability of randomly choosing an *exact* value in a given interval is essentially zero, we calculate the probability of randomly choosing a value in a given interval. We graph the trends of the probability as a smooth curve, and calculate the probability that a value will lie within a particular interval as the area between the smooth curve and the x -axis on a Cartesian graph.

Vocabulary

A **probability density function** is a function that defines a probability distribution for a continuous random variable.

A **probability density graph** is a visual depiction of a probability density function, commonly a curved line drawn on a Cartesian coordinate graph.

The **normal curve** is the curve that defines the probability density graph for a normally distributed variable.

The **standard deviation** of a random variable or data set is the square root of the variance.

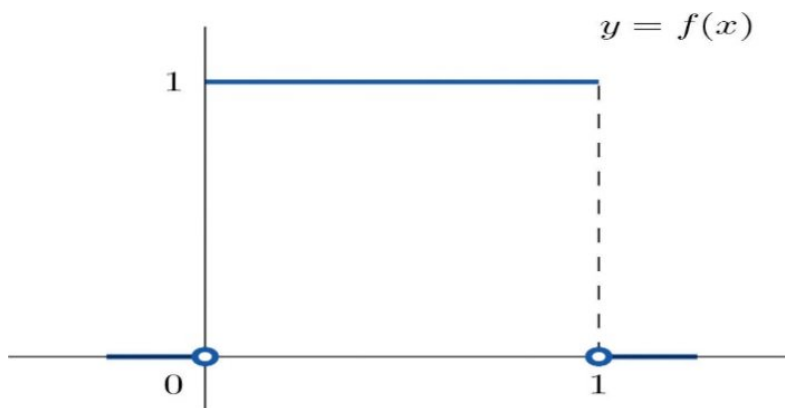
The **variance** of a data set is the average squared distance from the mean value of the set.

A **discrete random variable** is a variable with a countable number of outputs determined by a random process.

A **continuous random variable** is a variable defined by a random process that can take on any value in an interval.

Guided Practice

Use the graph of the uniform distribution (all values in the range have the same probability of occurring) below to answer questions 1-3:

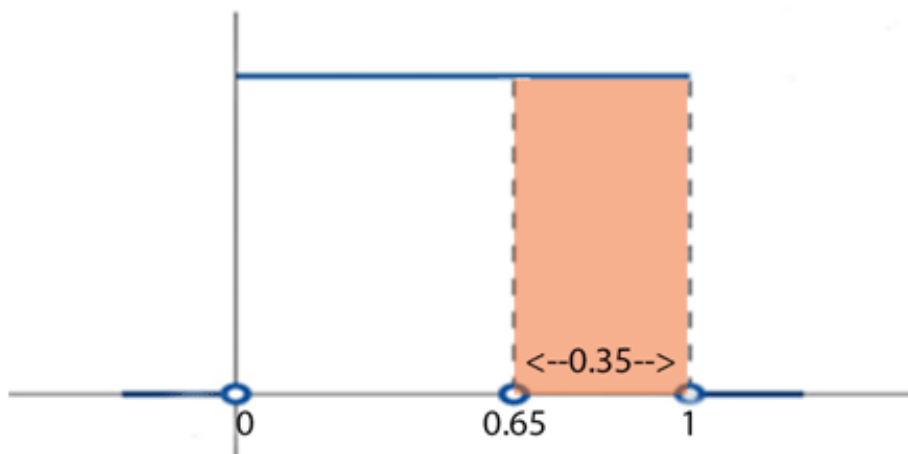


1. Find $P(0 < X < 1)$, the probability that X will be between 0 and 1.
2. Find $P(X > 0.65)$, the probability that X will be greater than 0.65.
3. Find $P(X < .40)$, the probability that X will be less than 0.40.

Solutions:

1. Recall that the probability that a value will occur within a given range is represented by the area between the defining line and the X -axis. Since the interval is 0 to 1, and the defining line is parallel to the X -axis and 1 unit above the X -axis, the area is a rectangle and can be calculated as $length \times width = 1 \times 1 = 1$ sq unit. Therefore $P(0 < X < 1) = 1$ or 100%.

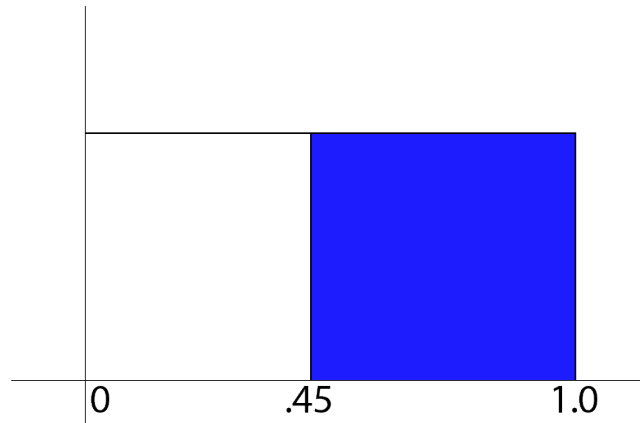
2. (See the image) If $X > 0.65$, then the width of the area is $1 - 0.65 = 0.35$. The height is 1 unit, so the area is $1 \times 0.35 = 0.35$ sq units. Therefore $P(X > 0.65) = .35$ or 35%.



3. If $X < .40$, then the width of the area representing the probability is .40. Since the height is still 1, the area is $0.40 \times 1 = 0.40$ sq units. Therefore $P(X < 0.40) = .40$ or 40%.

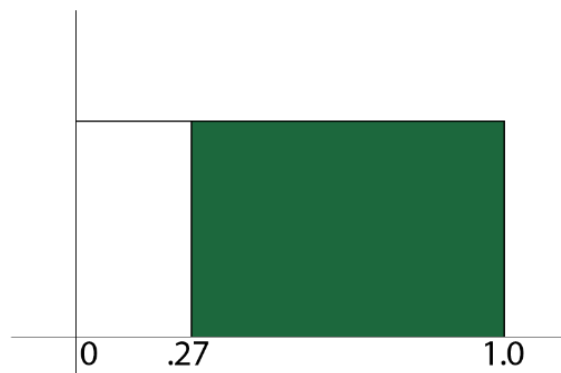
Practice

The graph below illustrates the probability distribution of random variable Y , use it to calculate the answers to problems 1-4:



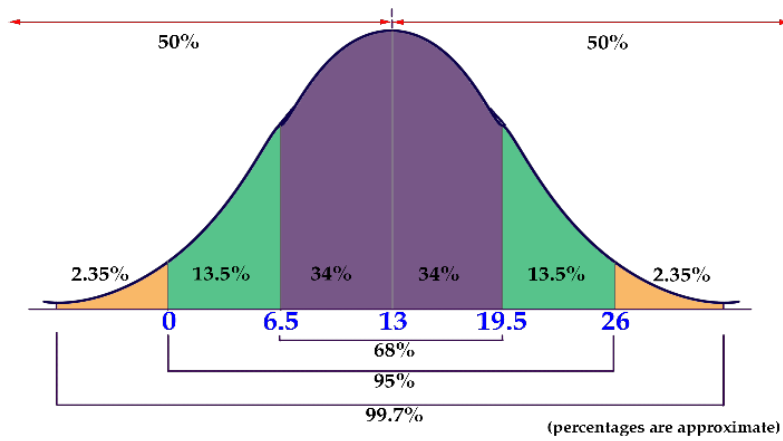
1. What is $P(Y > .45)$?
2. What is $P(Y > 0)$?
3. What is $P(Y < .45)$?
4. What is $P(Y > 1)$?

The graph below illustrates the uniform probability distribution of random variable Z , use it to calculate the answers to problems 5-8:



5. What is $P(Z < .27)$?
6. What is $P(Z < 0)$?
7. What is $P(Z > .27)$?
8. What is $P(Z < 1)$?

The graph below illustrates the normal probability distribution of random variable A , use it to calculate the answers to problems 9-15:



9. What is $P(6.5 < A < 19.5)$?
10. What is $P(0 < A < 19.5)$?
11. What is $P(0 < A)$?
12. What is $P(A < 26)$?
13. What is $P(-6.5 < A < 0)$?
14. What is $P(32.5 < A)$?
15. What is $P(A = 13)$?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.5.

7.6 Binomial Experiments

Objective

In this lesson, you will learn about a specific sort of an experiment called a *binomial experiment*.

Concept

Binomial experiments are very popular for studies because the probability of one possibility or the other can be calculated quickly and accurately. How do you identify a binomial experiment? Can an experiment that is not binomial be easily converted into a binomial experiment?

Look to the end of the lesson for the answer.

Watch This

Binomial
 2 outcomes
 Success Failure
 repeat n times - n trials
 X = no. of Successes in n trials
 $P(X=k)$
 Test case H T
 heads tails
 flip coin
 1. 2 outcomes
 2. each trial is independent
 3. Prob success same in each trial

MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67364>

http://youtu.be/6u8Kgc_dNL8 westofvideo - Binomial Experiments

Guidance

Binomial experiments give rise to *binomial random variables*, which will be the topic of our next couple of lessons. A *binomial experiment* is a very specific type of experiment. In order to be a binomial experiment, there are four qualifications that the experiment must meet:

1. **There must be a fixed number of trials.** The experiment cannot just be to roll a die until you get a 2, because the number of rolls (trials) is not fixed.
2. **Each trial must be independent of the others.** You cannot have a situation like “If you flip a coin and get heads, flip twice more, and if you get tails, flip three more times”.
3. **Each trial must have a “success” and a “failure”.** Depending on the trial, these may be identified as “yes” and “no” or “0” and “1” or “black” and “white”, etc. However, from a statistics standpoint, the outcome you are studying is generally called the “success” and the other is called “failure”.
4. **The probability of success must be the same for all trials.** The experiment cannot be 10 trials of pulling and *keeping* a card from a deck to see how many are hearts, because the probability of getting a heart would change each trial. To make this a binomial experiment, you need to *replace* the card each time.

Example A

If a fair coin is flipped 10 times, and T is the number of tails, is T a binomial random variable?

Solution:

Yes, T is a binomial random variable, and this is a binomial experiment. It meets all four qualifications:

1. There is a specific number of trials: 10 flips
2. Trials are independent: the outcome of one coin flip does not affect the next flip
3. There are only two possible outcomes: a “success” and “failure”. Since we are counting tails, every tails is a “success” and every heads is a “failure”
4. The probability is the same for all trials: The probability of getting tails is always 50% if flipping a fair coin

Example B

If Trina designates Y to be the number of yellow marbles she gets during nine trials of randomly pulling 1 marble from a bag filled with marbles of various colors and returning it, is Y a random variable? Is it binomial?

Solution:

Yes, Y is a random variable, since it is the random numerical result of a limited number of independent trials of an experiment. It is also binomial, since each of the limited trials is independent, has a success/failure (yellow/not yellow), and has the same probability of success.

Example C

If N is the number of nines you get when rolling two standard dice three times:

- a. Is N a binomial random variable?
- b. What are the possible values of N ?
- c. Create a histogram or pie chart showing the probability distribution of N .

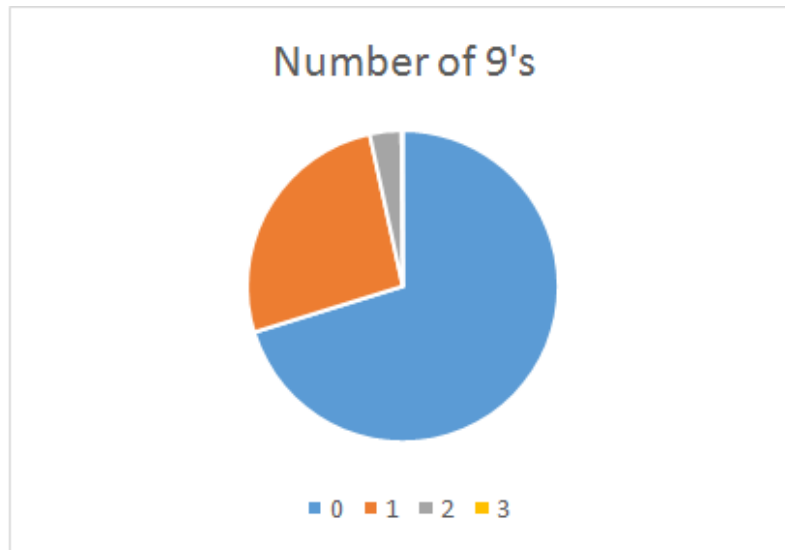
Solution:

- a. N is a binomial random variable, because it is the result of a specific and limited number of independent trials of a random process and each outcome is either nine or not nine.
- b. Since you could only roll a total of 9 once each trial, N could be 0, 1, 2, or 3.
- c. The probabilities of each of the possible values of N would be:

(see the lesson: Understanding Discrete Random Variables, Example C, for the calculations)

- $N = 0 : \frac{512}{729}$
- $N = 1 : \frac{192}{729}$
- $N = 2 : \frac{24}{729}$
- $N = 3 : \frac{1}{729}$

A pie chart would look like this: (note that total probability = $\frac{512}{729} + \frac{192}{729} + \frac{24}{729} + \frac{1}{729} = \frac{729}{729}$).



Concept Problem Revisited

Binomial experiments are very popular for studies because the probability of one possibility or the other can be calculated quickly and accurately. How do you identify a binomial experiment? Can an experiment that is not binomial be easily converted into a binomial experiment?

A binomial experiment must consist of a limited number of independent trials, where each trial outcome is either a success or a failure, and each trial has the same probability of success as all other trials.

A non-binomial experiment can often be viewed as binomial by carefully stating the outcome of each trial in a binomial format. For example, a non-binomial experiment might be “Count the number of heads and tails resulting from 8 flips of a fair coin”. Viewed as a binomial experiment, the same results could be collected from “How many tails do you get by flipping a fair coin 8 times”? You could then subtract the result from 8 to get the number of “not tails”, e.g. “heads”.

Vocabulary

A **random variable** is the numeric result of a specific and limited number of independent trials of a random process.

A **binomial experiment** must consist of a limited number of independent trials, where each trial outcome is either a success or a failure, and each trial has the same probability of success as all other trials.

A **discrete random variable** has a specific and countable number of possible values.

A **continuous random variable** is a random variable that can take on all values in an interval. For instance, if a continuous random variable can be any value in the interval between 0 and 1, then it could be .1, .11, .111, .1111, etc. There are an infinite number of possible values in any given interval.

Guided Practice

1. Mariska spins a spinner 40 times, recording the number of 4's she gets. Is this a binomial experiment?
2. Heidi has a bag containing 4 blue, 3 green, 5 red, and 7 yellow marbles. She defines a trial as pulling a marble, recording the color, and replacing it. She records the number of trials it takes to pull a green marble. Is this a binomial experiment?
3. Evan notes that 24% of online game players he polled are between 30 and 39 years old. Evan decides to create a team of players from that age range by randomly choosing names from among those he polled, keeping each

one he chooses that is in his/her 30's. If he chooses a name only 10 times, no matter the number of players he gets, is this a binomial experiment?

Solutions:

1. Yes, this is a binomial experiment because Mariska is conducting a limited number of independent random “4” or “not 4” trials, and the probability of spinning a “4” does not change,
2. No, Heidi is not conducting a binomial experiment because the number of trials is not specified, she just keeps pulling until she gets a green.
3. No, Evan is not conducting a binomial experiment because the probability that a random player will be between 30 and 39 changes each time he keeps one for his team.

Practice

For questions 1-12, state that a particular experiment is or why it is not binomial:

1. A spinner has a 35% probability of landing on blue. Let B be the number of blues spun in 5 spins.
2. A bag contains 6 blue, 4 green, and 3 red candies. Let G be the number of green candies you pull out and eat in 5 trials.
3. One trial of an experiment consists of pulling a random card from a standard deck, noting it, and replacing it, you conduct 12 trials.
4. One trial consists of pulling two cards from a standard deck, noting them, and replacing them. Let T be the number of trials until you pull two face cards at the same time.
5. A 20-sided die is rolled ten times, and S is the number of sevens rolled.
6. Assume that 15% of word game players create at least 12 words out of 50 that have more than 5 letters, and you let W be the number of letters in words from 20 trials of 1 game each.
7. A die is rolled 20 times. What is the probability of rolling a 1 exactly 5 times?
8. You plan on choosing students (with replacement) from a population of 28, 17 of which are Juniors. You want to know how many will have to be picked before getting a Junior.
9. A new reality show is so popular that an estimated 47% of households watch it every week. You choose 20 households at random. Let X be the number of households watching the show.
10. H is the number of heads tallied over ten flips of a fair coin.
11. F is the number of 5's you roll before rolling a 6, on a standard die.
12. O is the number of 1's you roll in fifteen rolls of a standard die.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.6.

7.7 Expected Value

Objective

Here you will learn how to calculate the mean and variance or standard deviation of the probability distributions of discrete random variables.

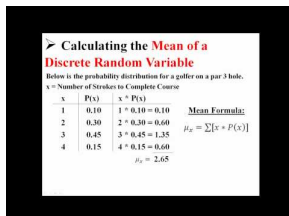
Concept

Suppose you were given a six-sided die that was weighted to land on one value more often than normal. You roll the die one hundred times, record the results, and display them on the frequency table below. How could you use this information to determine the probability of a particular value appearing on any given roll, and the value you would expect to be the average, if you were to continue rolling?

TABLE 7.9:

Roll Value	1	2	3	4	5	6
Frequency	11	11	34	11	11	11

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67356>

<http://youtu.be/EDNhzQMLTqE> statslectures - Mean and Expected Value of Discrete Random Variables

Guidance

A random variable yields outputs that are random by definition, however that does not necessarily mean that all possible values have the same chance of appearing. In the video above, the instructor uses a golf player's past performance to calculate the *expected value* of future performance in a similar situation. We can see from the video that the player more often completes similar holes in three strokes than in two or four. To find the expected value, the instructor calculates the mean of the random variable.

Since the *expected value* of a random variable is the *mean* output of the observed trials, it is value we would expect the average of *lots* (perhaps thousands) of trials to approach over time. Interestingly, the expected value is *not* necessarily the value we would expect to see on any given roll. In fact, as we can see from the video where the expected number of strokes to complete the hole is 2.65, it is quite possible to have an expected value that never has, and never will, be an outcome of the experiment (kind of difficult to swing a golf club 2.65 times to sink the ball!).

The formula we use to calculate the mean is not the arithmetic mean formula you have used in the past, since it does not require you to divide by the count of values, but rather to multiply each value by the probability of it appearing as an outcome. The formula looks like this:

$$\mu_x = \sum [x \times P(x)]$$

The mean of random variable x is the sum of possible outcomes of x , each multiplied by its percent probability of occurrence.

The application of the formula is more straight forward than the explanation, so let's look at a few examples.

Example A

What is the mean of discrete random variable Y , which has a probability distribution given by the table below?

TABLE 7.10:

x	1	2	3	4
$P(Y = x)$.26	.24	.35	.15

Solution:

To calculate the mean, we simply add up each of the values multiplied by its probability of occurrence:

$$(1 \times .26) + (2 \times .24) + (3 \times .35) + (4 \times .15) = 2.39$$

$$\mu_Y = 2.39$$

The mean of random variable Y is 2.39.

Example B

What is the expected value of a weighted six-sided die that has a 50% probability of landing on 5, and an equal probability of landing on each other possibility?

Solution:

Start by creating a probability distribution for random variable X :

TABLE 7.11:

x	1	2	3	4	5	6
$P(X = x)$.10	.10	.10	.10	.50	.10

Now we can apply the formula for calculating the mean:

$$\mu_X = (1 \times .1) + (2 \times .1) + (3 \times .1) + (4 \times .1) + (5 \times .5) + (6 \times .10) = 4.1$$

$$\mu_X = 4.1$$

Example C

Let random variable C be one-half of the sum of two standard dice. What is the expected value of C ?

Solution:

To calculate μ_C , the mean or expected value of C , start by creating a probability distribution:

The distribution of the sum of two standard dice is:

TABLE 7.12:

x	1	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	0	.028	.056	.083	.111	.139	.167	.139	.111	.083	.056	.028

Which makes the distribution of C :

TABLE 7.13:

x	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6
$P(X = x)$.028	.056	.083	.111	.139	.167	.139	.111	.083	.056	.028

Now we can apply the formula for calculating the mean of a discrete random variable:

$$\begin{aligned}
 &(1 \times 0.028) + (1.5 \times 0.056) + (2 \times 0.083) + (2.5 \times 0.111) + (3 \times 0.139) + (3.5 \times 0.167) \\
 &\quad + (4 \times 0.139) + (4.5 \times 0.111) + (5 \times 0.083) + (5.5 \times 0.056) + (6 \times 0.028) \\
 &(0.03) + (.083) + (0.17) + (.28) + (.42) + (.59) + (.56) + (.5) + (.42) + (.31) + (.17) = 3.5 \\
 &\mu_X = 3.5
 \end{aligned}$$

So, if you were to conduct many, many trials, and find the mean, the *expected value* of that mean would be apx. 3.5.

Concept Problem Revisited

Suppose you were given a six-sided die that was weighted to land on one value more often than normal. You roll the die one hundred times, record the results, and display them on the probability distribution below. How could you use this information to determine the probability of a particular value appearing on any given roll, and the value you would expect to be the average, if you were to continue rolling?

TABLE 7.14:

Roll value	1	2	3	4	5	6
Frequency	11	11	34	11	11	11

Because there were 100 rolls, we can convert the frequency table to a probability distribution just by considering each frequency as a percentage, which gives us the probability of rolling each value. We can then set a random variable, say, D to equal the outcome of a roll of the die. The ***expected value*** is the mean of the random variable D , found by applying the formula:

$$\begin{aligned}
 1 \times .11 + 2 \times .11 + 3 \times .34 + 4 \times .11 + 5 \times .11 + 6 \times .11 &= 3 \\
 \text{So } \mu_D &= 3
 \end{aligned}$$

Vocabulary

The **expected value** of a random variable is the value you expect the average of the outcomes to approach as the number of trials increases.

The **mean** of a random variable is similar to the arithmetic mean of a set, but is calculated by finding the sum of each outcome multiplied by its probability of occurrence.

Guided Practice

- Suppose you take all of the number cards two through five from a standard deck, and set random variable C to be the sum of two cards drawn at random, without replacement. What is the expected value of C ?
- Over the past year, Sally has compiled a probability distribution of the number of kids she baby sits on each day of the week. Based on her data from the table below, what is the expected number of kids she will baby sit on any random day?

TABLE 7.15:

x	1	2	3	4	5
$P(X = x)$.35	.40	.15	.05	.05

- Tuscany works for a hot dog vendor at the Colorado Rockies baseball stadium. Over the last couple of years, she has created the probability distribution below of the number of drinks a spectator will consume during a baseball game. If the drinks cost \$5 each, how much is a spectator expected to spend on drinks?

TABLE 7.16:

# of drinks	1	2	3	4	5
Probability	28%	42%	20%	7%	3%

Solutions:

- Create a probability distribution for C :

TABLE 7.17:

x	4	5	6	7	8	9	10
# possible combinations	6	16	22	32	22	16	6
$P(C = x)$.05	.133	.183	.267	.183	.133	.05

Apply the expected value formula:

$$(4 \times .05) + (5 \times .133) + (6 \times .183) + (7 \times .267) + (8 \times .183) + (9 \times .133) + (10 \times .05) = .2 + .665 + 1.098 + 1.869 + 1.464 + 1.197 + .5 \approx 7$$

The expected value of the random variable C is 7.

- We already have the probability distribution, just apply the formula:

$$(1 \times .35) + (2 \times .4) + (3 \times .15) + (4 \times .05) + (5 \times .05)$$

$$.35 + .8 + .45 + .2 + .25 = 2.05$$

Sally should expect to average two kids per day for her baby sitting business.

3. Start by applying the probability distribution values to the expected value formula:

$$1 \times .28 + 2 \times .42 + 3 \times .2 + 4 \times .07 + 5 \times .03$$

$$.28 + .84 + .6 + .28 + .15 = 2.15$$

So the average customer is expected to buy 2.15 drinks per game. Since the drinks cost \$5 each, that means that **Tuscany's hot dog cart can expect to average \$10.75 per customer in drink sales.**

Practice

For questions 1 - 10, calculate the expected value of the random variable with the given probability distribution:

1.

TABLE 7.18:

x	4.1	4.4	4.7	4.9	5.1
$P(X = x)$.30	.45	.10	.05	.10

2.

TABLE 7.19:

x	4	8	12	16	20
$P(X = x)$.50	.25	.15	.05	.05

3.

TABLE 7.20:

x	15	30	45	60	75
$P(X = x)$.20	.25	.15	.27	.13

4.

TABLE 7.21:

x	30	60	90	120	150	170
$P(X = x)$.18	.16	.24	.22	.20	.00

5.

TABLE 7.22:

x	3	11	19	27
-----	---	----	----	----

TABLE 7.22: (continued)

$P(X = x)$.07	.08	.65	.20
------------	-----	-----	-----	-----

6.

TABLE 7.23:

x	13	17	21	25	29	33	37
$P(X = x)$.15	.17	.23	.30	.10	.03	.02

7.

TABLE 7.24:

x	26	39	52	65	78
$P(X = x)$	6%	14%	30%	28%	22%

8.

TABLE 7.25:

x	22	43	64	85	106
$P(X = x)$	10.5%	22.5%	31.5%	22.8%	12.7%

9.

TABLE 7.26:

x	.65	.84	1.03	1.22	1.41
$P(X = x)$.16	.29	.14	.28	.13

10.

TABLE 7.27:

x	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{5}{8}$	$\frac{3}{4}$	$\frac{7}{8}$
$P(X = x)$	$\frac{2}{5}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{20}$

11. Carrie shines shoes for money on weekday mornings, and she has compiled the following probability distribution of the number of clients she is likely to get each day. If she earns \$3.50 per shine, how much should she expect to earn each day, on average?

TABLE 7.28:

# clients	20	25	30	35	40
probability	.15	.35	.30	.15	.05

12. Vincente works for a fast food restaurant, where he earns \$8.50 per hour. The number of hours he works each week varies between 20 and 40, based on how busy the restaurant is during the week. Over the past year, he has compiled the probability distribution below describing the percent probability of his getting 20, 25, 30, 35, or 40 hours in any random week. If Vincente wants to move out, and knows that he shouldn't spend more than $\frac{1}{3}$ of his average income on housing, how much can he afford for rent?

TABLE 7.29:

# hours	20	25	30	35	40
probability	.15	.28	.32	.15	.10

13. How much more could Vincente afford for rent if he were given a raise of \$2 per hour?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.7.

7.8 Random Variable Variance

Objective

Here you will learn to calculate the variance and standard deviation of discrete random variables.

Concept

Recently, we discussed the process of finding the mean of a discrete random variable. The process resembled that of finding the arithmetic mean of a set of basic numbers, yet had some significant differences as well. Suppose you needed to know the variance or standard deviation of a random variable. Would these values be calculated differently for random variables than for standard numerical data sets, or not?

Watch This

x	P(x)	x^2	$x^2 \cdot P(x)$
1	0.10	$1^2 = 1$	$1 \cdot 0.10 = 0.10$
2	0.30	$2^2 = 4$	$4 \cdot 0.30 = 1.20$
3	0.45	$3^2 = 9$	$9 \cdot 0.45 = 4.05$
4	0.15	$4^2 = 16$	$16 \cdot 0.15 = 2.40$

Mean Formula: $\mu_x = \sum(x \cdot P(x))$ $\mu_x = 2.65$	Variance Formula: $\sigma^2_x = \sum(x^2 \cdot P(x)) - \mu^2_x$
---	--

MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67366>

<http://youtu.be/7h0TfaYVCv0> statslectures - Variance and Standard Deviation of Discrete Random Variables

Guidance

As we discussed some time ago, sometimes it is not enough to know the average, or mean, value of a data set when trying to get a feel for the trend(s) of the set. It is the same with a random variable, sometimes you need to know about the spread of a variable to get a better idea of the overall behavior.

One of the other additional pieces of information we learned to calculate in order to evaluate sets before was the **variance**, which is the square of the **standard deviation**. Both of these measures help to create an understanding of the tendency of values to cluster around the mean. By evaluating the variance and standard deviation of a random variable, we can get a better idea of the spread of the values than with the mean alone.

Just as with the mean, or **expected value**, we have a formula to apply in order to calculate the **variance**:

Then, to find the **standard deviation**, just take the square root of the variance:

$$\sigma_X = \sqrt{\sigma^2_X}$$

Example A

In another lesson, we calculated the *expected value* of the number of the number of kids that Sally baby sits on any given day from the data in the table below. Using the table and the mean, calculate the variance and standard deviation of the number of kids she baby sits.

TABLE 7.30:

x	1	2	3	4	5
$P(X = x)$.35	.40	.15	.05	.05

$$\mu_X = 2$$

Solution:

Use the data given in the question to fill in the formula and find the variance:

$$\text{Formula: } \sigma^2_X = \sum (x_i - \mu_X)^2 p_i$$

$$\begin{aligned} \sigma^2_X &= (1 - 2)^2 \times .35 + (2 - 2)^2 \times .4 + (3 - 2)^2 \times .15 + (4 - 2)^2 \times .05 + (5 - 2)^2 \times .05 \\ &= (1 \times .35) + (0 \times .4) + (1 \times .15) + (4 \times .05) + (9 \times .05) \\ &= .35 + 0 + .15 + .2 + .45 \\ \sigma^2_X &= 1.15 \end{aligned}$$

Since the variance is 1.15, the standard deviation is $\sqrt{1.15} = 1.07$

$$\sigma_X = 1.07$$

Example B

Random variable X has mean 18.84, and the probability distribution show below. Calculate the variance and standard deviation.

TABLE 7.31:

x	3	11	19	27
$P(X = x)$.07	.08	.65	.20

Solution:

Using the variance formula: $\sigma^2_X = \sum (x_i - \mu_X)^2 p_i$

$$\begin{aligned} \sigma^2_X &= (3 - 18.84)^2 \times .07 + (11 - 18.84)^2 \times .08 + (19 - 18.84)^2 \times .65 + (27 - 18.84)^2 \times .2 \\ \sigma^2_X &= (-15.84)^2 \times .07 + (-7.84)^2 \times .08 + (.16)^2 \times .65 + (8.16)^2 \times .2 \\ \sigma^2_X &= 251 \times .07 + 61.5 \times .08 + .03 \times .65 + 66.59 \times .2 \\ \sigma^2_X &= 17.6 + 4.9 + .02 + 13.3 \\ \sigma^2_X &= 35.8 \end{aligned}$$

Since the variance is 36.3, the standard deviation is $\sqrt{35.8} = 5.98$

$$\sigma_X = 5.98$$

Example C

The random variable Z has a probability distribution shown below, find μ_Z , σ^2_Z , and σ_Z .

TABLE 7.32:

x	.65	.84	1.03	1.22	1.41
$P(X = x)$.16	.29	.14	.28	.13

Solution: Start by finding the mean of Z :

$$\mu_Z = (.65 \times .16) + (.84 \times .29) + (1.03 \times .14) + (1.22 \times .28) + (1.41 \times .13) = 1.02$$

Now that we have the mean, we can use it to find the variance:

$$\sigma^2_Z = (.65 - 1.02)^2 \times .16 + (.84 - 1.02)^2 \times .29 + (1.03 - 1.02)^2 \times .14 + (1.22 - 1.02)^2 \times .28 + (1.41 - 1.02)^2 \times .13$$

$$\sigma^2_Z = (-.37)^2 \times .16 + (-.18)^2 \times .29 + (.01)^2 \times .14 + (.2)^2 \times .28 + (.39)^2 \times .13$$

$$\sigma^2_Z = .14 \times .16 + .03 \times .29 + 0 \times .14 + .04 \times .28 + .16 \times .13$$

$$\sigma^2_Z = .02 + .01 + .01 + .02$$

$$\sigma^2_Z = .06$$

Finally, the standard deviation is just the square root of the variance:

$$\sigma_Z = \sqrt{.06} = .25$$

Concept Problem Revisited

Suppose you needed to know the variance or standard deviation of a random variable. Would these values be calculated differently for random variables than for standard numerical data sets or not?

The variance and standard deviation are the same concept when dealing with random variable as with numerical data sets. However, the process of calculating the values is slightly different. Instead of dividing the squared difference of each number and the mean by the count of values: $\frac{(x-\mu)^2}{n}$ you multiply the square of the difference of each number and the mean by the probability of that value: $(x - \mu)^2 \times P(x)$.

In either case, the standard deviation is the square root of the variance.

Vocabulary

The **variance** of a random variable is a measure of how closely the values of the random variable tend to cluster around the mean, or expected value of the variable.

The **mean** or **expected value** of a random variable is the value that is expected to be the average of the outputs of the variable, over many, many trials.

Guided Practice

1. Calculate the variance and standard deviation of random variable Y , given: $\mu_Y = 43.2$ and:

TABLE 7.33:

x	15	30	45	60	75
$P(Y = x)$.20	.25	.15	.27	.13

Find μ_X , σ_X , and σ^2_X given:

TABLE 7.34:

x	4	8	12	16	20
$P(Y = x)$.50	.25	.15	.05	.05

3. Marie has a part-time job walking dogs to earn money on weekends. The following probability distribution represents the probability of having a particular number of clients on any given day. If she earns \$2.75 per client, how much could she expect to earn each day, on average, and what is the standard deviation of her expected earnings?

TABLE 7.35:

# clients	20	25	30	35	40
probability	.15	.35	.30	.15	.05

Solutions:

1. All of the values we need for this one are given, it is really just a “plug-n-chug” using the variance formula:

$$\sigma^2_X = \sum (x_i - \mu_x)^2 p_i$$

$$\sigma^2_Y = (15 - 43.2)^2 \times .2 + (30 - 43.2)^2 \times .25 + (45 - 43.2)^2 \times .15 + (60 - 43.2)^2 \times .27 + (75 - 43.2)^2 \times .13$$

$$\sigma^2_Y = (-28.2)^2 \times .2 + (-13.2)^2 \times .25 + (-1.8)^2 \times .15 + (16.8)^2 \times .27 + (31.8)^2 \times .13$$

$$\sigma^2_Y = 159 + 43.6 + .5 + 76.2 + 131.5$$

$$\sigma^2_Y = 410.8$$

$$\sigma_Y = \sqrt{410.8} = 20.27$$

2. First, calculate the mean

$$\mu_X = (4 \times .5) + (8 \times .25) + (12 \times .15) + (16 \times .05) + (20 \times .05) = 7.6$$

Then, use the mean to calculate the variance:

$$\sigma^2_X = (4 - 7.6)^2 \times .5 + (8 - 7.6)^2 \times .25 + (12 - 7.6)^2 \times .15 + (16 - 7.6)^2 \times .05 + (20 - 7.6)^2 \times .05$$

$$\sigma^2_X = 20.64$$

$$\sigma_X = \sqrt{20.64} = 4.5$$

3. Start by finding the mean:

$$\mu_X = 20 \times .15 + 25 \times .35 + 30 \times .3 + 35 \times .15 + 40 \times .05 = 28$$

Use the mean to find the variance:

$$\sigma_X^2 = (20 - 28)^2 \times .15 + (25 - 28)^2 \times .35 + (30 - 28)^2 \times .30 + (35 - 28)^2 \times .15 + (40 - 28)^2 \times .05 = 28.5$$

Use the variance to find the standard deviation: $\sigma_X = \sqrt{28.5} = 5.3$

Now we can find her average income by multiplying the mean, 28 by Marie's rate, \$2.75, to get her **average daily income of \$77**.

Finally, we can multiply the calculated standard deviation, 5.3, by the rate, \$2.75, to get the standard deviation of her income: $5.3 \times \$2.75 = \14.58

What all this means is that Marie can expect to average \$77 per day, on average, give or take about \$14.50.

Practice

For questions 1 - 9, find the variance and standard deviation of the random variable, given the mean and probability distribution.

1. $\mu_x = 4.435$

TABLE 7.36:

x	4.1	4.4	4.7	4.9	5.1
$P(X = x)$.30	.45	.10	.05	.10

2. $\mu_x = 7.6$

TABLE 7.37:

x	4	8	12	16	20
$P(X = x)$.50	.25	.15	.05	.05

3. $\mu_x = 43.2$

TABLE 7.38:

x	15	30	45	60	75
$P(X = x)$.20	.25	.15	.27	.13

4. $\mu_X = 93$

TABLE 7.39:

x	30	60	90	120	150	170
$P(X = x)$.18	.16	.24	.22	.20	.00

5. $\mu_X = 12.92$

TABLE 7.40:

x	5	9	13	17
$P(X = x)$.07	.08	.65	.20

6. $\mu_X = 21.80$

TABLE 7.41:

x	13	17	21	25	29	33	37
$P(X = x)$.15	.17	.23	.30	.10	.03	.02

7. $\mu_X = 57.98$

TABLE 7.42:

x	26	39	52	65	78
$P(X = x)$	6%	14%	30%	28%	22%

8. $\mu_X = 64.99$

TABLE 7.43:

x	22	43	64	85	106
$P(X = x)$	10.5%	22.5%	31.5%	22.8%	12.7%

9. $\mu_X = 7.46$

TABLE 7.44:

x	3.65	5.84	7.03	9.22	11.41
$P(X = x)$.16	.25	.18	.24	.17

10. Dorian works for a construction company, where he earns \$11.50 per hour. The number of hours he works each week varies between 25 and 40. Based on prior experience, Dorian has compiled the probability distribution below describing the probability that he will work a given number of hours. Can Dorian afford to buy a new truck that has a payment of \$525/month, if he wants to be sure not to put more than 25% of his average monthly income into car payments? What is the standard deviation of his monthly income?

TABLE 7.45:

# hours	25	28	31	34	37	40
probability	.15	.14	.26	.18	.14	.13

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.8.

7.9 Transforming Random Variables I

Objective

Here you will learn how adding or subtracting a constant from a random variable affects the mean and standard deviation.

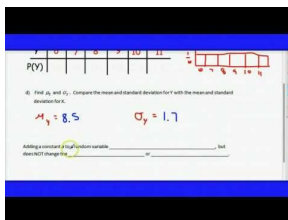
Concept



Suppose you are running a lemonade stand, and you know that you have a mean of 15 customers per hour, with a standard deviation of 5 customers. If you know that it costs you 30 cents per glass in materials, and that you could hire your little sister to work the stand for you for \$4.50 per hour, how could you calculate the mean and standard deviation of the cost per hour of having your sister run the stand?

At the end of this lesson we will know one part of the answer. We will learn the other part in the next lesson, so stay tuned!

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/82777>

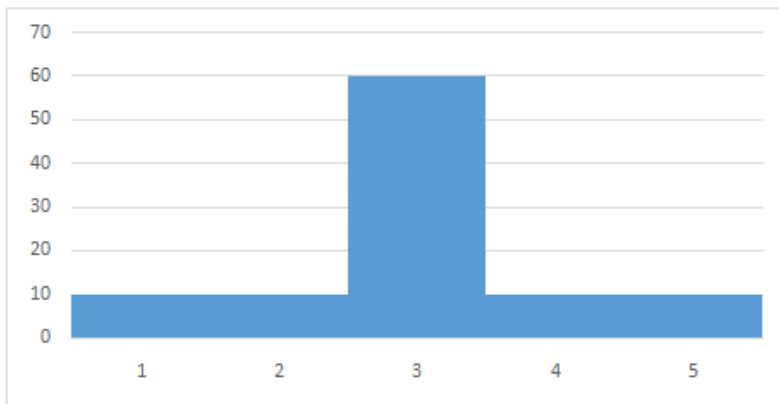
<http://youtu.be/CtcuN4wBDMQ> Leigh Nataro — Transforming Random Variables

Guidance

Sometimes it is useful to be able to calculate how the mean and standard deviation of a random variable is affected by adding or subtracting a constant to each outcome. For example, suppose you were able to calculate the mean

labor and materials cost per hour at your pizza store, based on the probabilities of there being different numbers of orders per hour. If you wanted to learn how it would affect your bottom line to hire an employee, you might want to compare the increased income and expense of adding 8 orders per hour. Since you could use a random variable to represent number of customers per hour, you would need to add 8 to every possible outcome of the random variable, which could take quite a while. Rather than having to then recalculate the mean, variance, and standard deviation using all new data, it would be great to know how adding a constant (like 8, for instance) affects the mean and variance of a random variable directly.

Perhaps the most important realization here is that adding or subtracting a constant from the outcomes of a random variable *changes all of the outcomes by the same amount*. Look at the distribution graph here, it represents the probability distribution of random variable X :

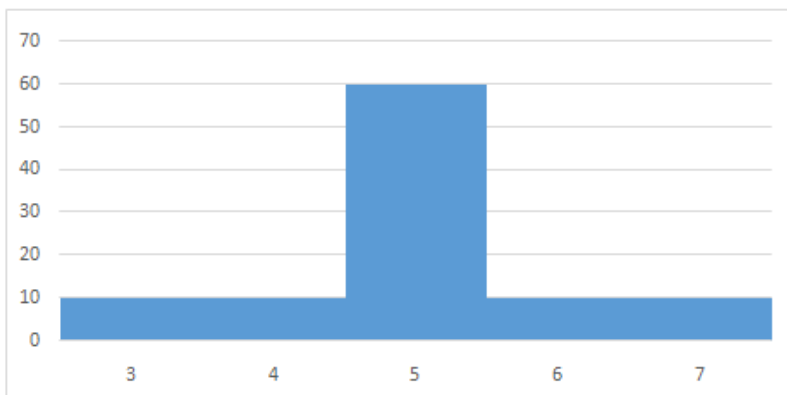


This distribution has a mean of 3 and a standard deviation of 1. In other words:

$$\begin{aligned}\mu_X &= 3 \\ \sigma_X &= 1\end{aligned}$$

If all of the values on the chart were increased or decreased by the same amount, then the position of the mean would move left or right across the graph, but the *relative position* of each bar would not change.

To see the effect of adding a constant to the outcomes, let's look at the distribution of $X + 2$:

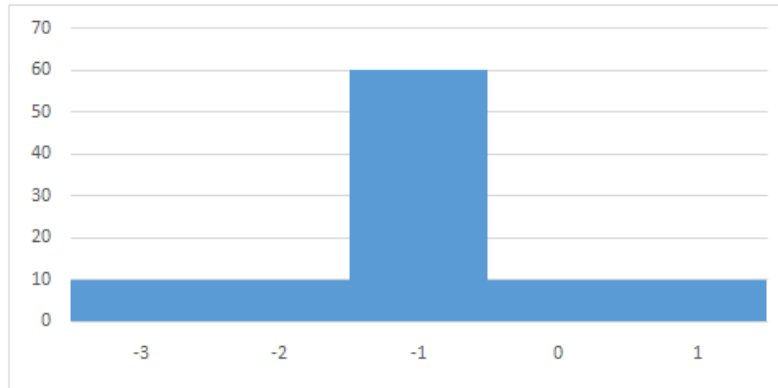


Now the distribution has a mean of 5, but the standard deviation is still 1:

$$\mu_{X+2} = 5$$

$$\sigma_{X+2} = 1$$

What happens if we subtract a value? Let's take a look at $X - 4$:



Now the mean has gone down by four, to -1, but the standard deviation is still unchanged:

$$\mu_{X-4} = -1$$

$$\sigma_{X-4} = 1$$

Example A

If $\mu_X = 5$, and $\sigma_X = 2.4$, what is $\mu_{X+3.5}$ and $\sigma_{X+3.5}$?

Solution:

When adding or subtracting a constant to/from a random variable, the mean is changed directly by the constant, but the standard deviation remains unchanged.

In this example, $\mu_X = 5$, so $\mu_{X+3.5} = 5 + 3.5 = 8.5$

Since we are simply adding a constant, 3.5, to X , the standard deviation remains unchanged, so $\sigma_{X+3.5} = 2.4$.

Example B

If $\mu_Y = 13.7$, and $\sigma_Y = 6.7$, what is $\mu_{Y-1.4}$ and $\sigma_{Y-1.4}$?

Solution:

Similar to the last example, here we are subtracting a constant from a random variable, so we expect the mean to be reduced, and the standard deviation to remain unchanged.

$\mu_Y = 13.7$, so $\mu_{Y-1.4} = 13.7 - 1.4 = 12.3$

$\sigma_Y = 6.7$, and standard deviation does not change when adding or subtracting a constant from a random variable, so $\sigma_{Y-1.4} = 6.7$ also.

Example C

Sayber walks dogs for money on weekends, and he has compiled the following probability distribution of the number of dogs he is likely to walk each day.

- If he earns \$7.50 per dog, how much should he expect to earn each day, on average?
- How would it affect his mean income if he were to get 5 more dogs each day?

TABLE 7.46:

# clients	13	17	22	27	32
probability	.10	.25	.30	.25	.10

Solution:

a. Start by setting a random variable, we'll use D , to represent the number of dogs he walks. Then we can calculate the mean of D , using the formula: $\mu_x = \sum[x \times P(x)]$.

$$\mu_D = (13 \times .1) + (17 \times .25) + (22 \times .3) + (27 \times .25) + (32 \times .10) = 22.1$$

$$\mu_D = 22.1$$

That means **Sayber could expect to earn $22.1 \times \$7.50 = \165.75 each day, on average.**

b. If the mean of the number of dogs was 22.1 to start with, and Sayber were to add 5 dogs to each of the values he started with on his probability distribution, that would be the same as calculating μ_{D+5} , which we now know would result in an average of $22.1 + 5 = 27.1$. His income then would be $27.1 \times \$7.5 = \203.25 .

Concept problem Revisited

Suppose you are running a lemonade stand, and you know that you have a mean of 15 customers per hour, with a standard deviation of 5 customers. If you know that it costs you 30 cents per glass in materials, and that you could hire your little sister to work the stand for you for \$4.50 per hour, how could you calculate the mean and standard deviation of the cost per hour of having your sister run the stand?

By using the skills we learned in this lesson, you could convert the cost of running the stand yourself to hiring your sister, by adding \$4.50 to the average hourly cost. Unfortunately, we can't actually solve the problem yet, because we haven't learned how to deal with the \$0.30 per glass in materials cost, that is the next lesson!

Vocabulary

The **mean** of a random variable is the value you would expect the average of all outcomes of the random variable to approach as the number of trials increases.

The **variance** and **standard deviation** of a random variable are measures of the spread of the values in a distribution around the mean.

Guided Practice

- If $\mu_C = 12$, and $\sigma_C = .7$, what is $\mu_{C+6.4}$ and $\sigma_{C+6.4}$?
- If $\mu_Y = 113$, and $\sigma_Y = 12.22$, what is $\mu_{Y-17.4}$ and $\sigma_{Y-17.4}$?
- If a production line has an error tolerance of 4 parts per ten thousand within 1 standard deviation of the mean, and currently has a mean error rate of .0002, with a standard deviation of .00005, would it be acceptable to increase the speed of production by 12% if it resulted in an increased error rate of .00013?

Solutions:

1. Since $\mu_C = 12$, and $\sigma_C = .7$, and adding a constant to a random variable affects the mean directly, but does not change the variance or standard deviation, $\mu_{C+6.4} = 12 + 6.4 = 18.4$ and $\sigma_{C+6.4} = .7$.
2. Since $\mu_Y = 113$, and $\sigma_Y = 12.22$, that means $\mu_{Y-17.4} = 113 - 17.4 = 95.6$, and $\sigma_{Y-17.4} = 12.22$ (unchanged).
3. If the line must maintain an error rate of $< .0004$ within one standard deviation of the mean, and currently ranges from $-1SD = .0002 - .00005 = .00015$ to $+1SD = .0002 + .00005 = .00025$. There are two different ways to look at this:
 1. If the current error rate has a maximum of $.00025$, that means there is a “safety net” of $.0004 - .00025 = .00015$. Since $.00013 < .00015$, the increased error is acceptable, just barely.
 2. Alternately, we can recall that adding a constant to a random variable affects the mean directly, but does not affect the standard deviation. So adding 0.00013 to the error rate would increase the mean to $.00025 + .00013 = .00038$, which, again, is just barely within tolerance.

Practice

For questions 1-10, add or subtract the given constant from the random variable:

1. If $\mu_X = .07$, and $\sigma_X = .002$, what is $\mu_{X+.02}$ and $\sigma_{X+.02}$?
2. If $\mu_C = 144$, and $\sigma_C = 17$, what is $\mu_{C+13.6}$ and $\sigma_{C+13.6}$?
3. If $\mu_Y = 22$, and $\sigma_Y = 1.8$, what is $\mu_{Y-3.49}$ and $\sigma_{Y-3.49}$?
4. If $\mu_Z = 68.33$, and $\sigma_Z = 4.87$, what is the μ and σ of $Z + 4.25$?
5. If $\mu_X = 2.071$, and $\sigma_X = .807$, what is the μ and σ of $X - 1.035$?
6. If $\mu_A = \frac{3}{5}$, and $\sigma_A = \frac{1}{5}$, what is the μ and σ of $A + \frac{3}{10}$?
7. If $\mu_B = 17.031$, and $\sigma_B = 2.101$, what is the μ and σ of $B - .035$?
8. If $\mu_A = \frac{2}{7}$, and $\sigma_A = \frac{1}{7}$, what is the μ and σ of $A - \frac{3}{14}$?
9. If $\mu_Y = 13.1$, and $\sigma_Y = 3.01$, what is the μ and σ of $Y + .27$?
10. If $+1SD$ of $\mu_X = 4.25$ and $-1SD$ of $\mu_X = 3.75$, what is μ_{X+2} and σ_{X+2} ?
11. If $+1SD$ of $\mu_X = 11.25$ and $-1SD$ of $\mu_X = 9.75$, what is $\mu_{X-1.25}$ and $\sigma_{X-1.25}$?
12. If $+2SD$ of $\mu_X = 25$ and $-2SD$ of $\mu_X = 5$, what is μ_{X-4} and σ_{X-4} ?
13. If $-2SD$ of $\mu_Z = 9$ and $+2SD$ of $\mu_Z = 15$, what is $\mu_{Z-.75}$ and $\sigma_{Z-.75}$?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.9.

7.10 Transforming Random Variables II

Objective

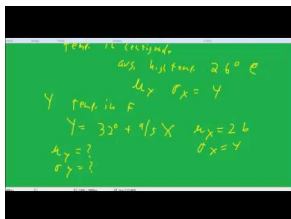
Here you will learn about multiplying random variables by a constant.

Concept

In the previous lesson, Transforming Random Variables I, we considered the following problem, and partially solved it, but did not yet know how to handle the \$0.30 per customer materials cost. In this lesson we will learn how to multiply random variables by a constant (like 0.30), so we can finish solving the problem.

Suppose you are running a lemonade stand, and you know that you have a mean of 15 customers per hour, with a standard deviation of 5 customers. If you know that it costs you 30 cents per glass in materials, and that you could hire your little sister to work the stand for you for \$4.50 per hour, how could you calculate the mean and standard deviation of the cost per hour of having your sister run the stand?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/82780>

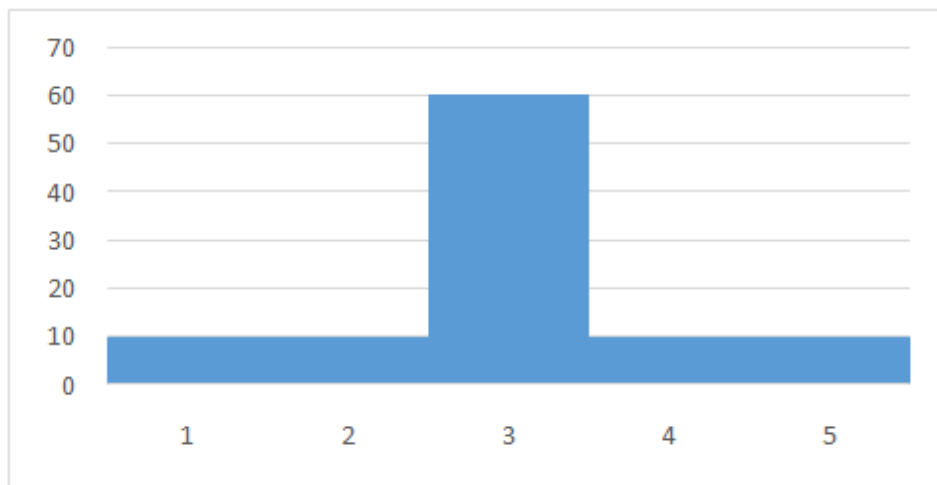
<http://youtu.be/hpQfhHQS4Xk> Arnold Kling — transformations of random variables

Guidance

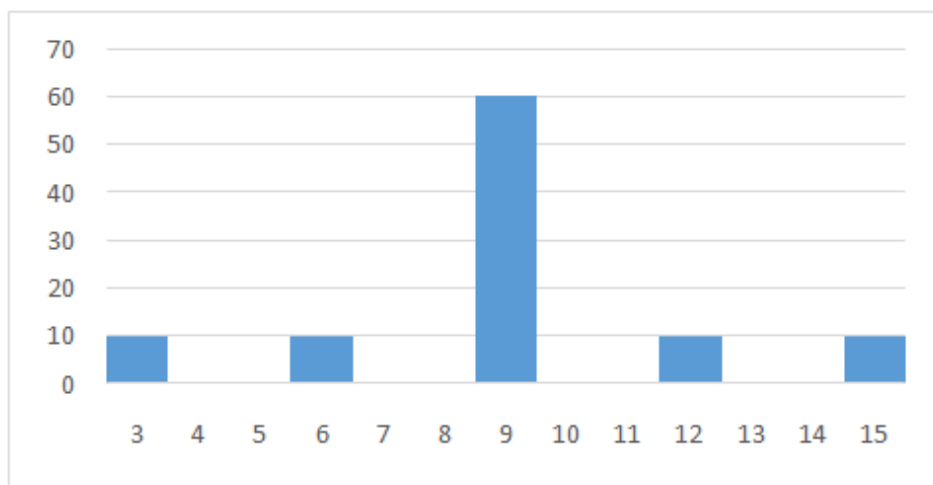
We discussed previously the process of adding or subtracting a constant to a random variable, but often, you may need to multiply the random variable by a constant as well, or instead. For instance, suppose you had a random variable that represented the temperature in your classroom in degrees Celsius, and wanted to convert the values to degrees Fahrenheit. You could either start over with your calculations, having first converted each random variable outcome to the Fahrenheit equivalent, or you could just convert the mean and variance you already calculated to Fahrenheit by multiplying by $\frac{9}{5}$ and adding 32.

Since we discussed the process of adding a constant to a random variable in the previous lesson, let's look at the multiplication process.

The graph below represents the probability distribution of random variable Y , with $\mu_Y = 3$ and $\sigma_Y = 1$:



Let's see what happens if we multiply each outcome by 3, and find μ_{3Y} and σ_{3Y} :



By multiplying every outcome by three, not only was the mean multiplied by three: $\mu_{3Y} = 3 \times 3 = 9$, but the standard deviation was also increased by a multiple of three: $\sigma_{3Y} = 1 \times 3 = 3$. The variance, since it is the square of the standard deviation, would be multiplied by 3^2 or 9.

Example A

Given that $\mu_X = 3$, $\sigma_X = .5$, and $\sigma^2_X = .25$, what is μ_{3X} , σ_{3X} and σ^2_{3X} ?

Solution:

When a random variable is multiplied by a constant, the mean and standard deviation are multiplied by the same constant, and the variance is multiplied by the square of the constant:

- $\mu_{3X} = \mu_X \times 3 = 3 \times 3 = 9$
- $\sigma_{3X} = \sigma_X \times 3 = .5 \times 3 = 1.5$
- $\sigma^2_{3X} = \sigma^2_X \times 3^2 = .25 \times 9 = 2.25$

Example B

Given random variable Y with $\mu_Y = .75$, $\sigma_Y = 0.125$, and $\sigma^2_Y = 0.0156$, what is $\mu_{-2.2Y}$, $\sigma_{-2.2Y}$ and $\sigma^2_{-2.2Y}$?

Solution:

As above, when a random variable is multiplied by a constant, the mean and standard deviation are multiplied by the same constant, and the variance is multiplied by the square of the constant:

- $\mu_{-2.2Y} = \mu_Y \times -2.2 = 0.75 \times -2.2 = -1.65$
- $\sigma_{-2.2Y} = \sigma_Y \times -2.2 = 0.125 \times -2.2 = -0.275$
- $\sigma^2_{-2.2Y} = \sigma^2_Y \times (-2.2)^2 = 0.0156 \times 4.84 = 0.075$

Example C

Tony the troll charges a fee of 3 gold pieces per person + 2 gold pieces per sword in order to cross his bridge. If random variable S represents the number of swords carried by a traveler wanting to cross his bridge, then $\mu_S = 2.8$, and $\sigma_S = 1.095$. What is the mean and standard deviation of Tony's income per traveler?

Solution:

Since we know the mean and variance of the traffic across Tony's bridge, we are really just looking to multiply that data by 2 and then add 3 in order to convert the number of swords to the number of gold pieces and add the per traveler charge.

We can describe the converted values mathematically, using the random variable S from the question, as: μ_{2S+3} , and σ_{2S+3} .

We know that adding a constant affects the mean, but not the variance or standard deviation, and that multiplying by a constant affects both, so the standard deviation should be multiplied by two, and the mean should be multiplied by 2, and then increased by 3:

- $\mu_{2S+3} = 2(\mu_S) + 3 = 2(2.8) + 3 = 5.6 + 3 = 8.6$
- $\sigma_{2S+3} = 2(\sigma_S) = 2(1.095) = 2.19$ (remember that adding a constant to a random variable doesn't affect the standard deviation.)

Concept Problem Revisited

Suppose you are running a lemonade stand, and you know that you have a mean of 15 customers per hour, with a standard deviation of 5 customers. If you know that it costs you 30 cents per glass in materials, and that you could hire your little sister to work the stand for you for \$4.50 per hour, how could you calculate the mean and standard deviation of the cost per hour of having your sister run the stand?

Now we should have all of the pieces to solve this problem. If we let random variable D equal the number of customers per hour, then $\mu_D = 15$ and $\sigma_D = 5$. We need to multiply the number of customers, D , by \$0.30, and add that to the base cost of \$4.50 that we have to pay our sister regardless of the number of customers, to get the hourly cost in dollars. Mathematically, we need to find the mean and standard deviation of $.3D + 4.5$. In other words, we are looking for $\mu_{.3D+4.5}$ and $\sigma_{.3D+4.5}$.

- $\mu_{.3D+4.5} = .3(\mu_D) + 4.5 = .3(15) + 4.5 = 4.5 + 4.5 = \9.00 per hour
- $\sigma_{.3D+4.5} = .3(\sigma_D) = .3(5) = \1.50 per hour

So, based on our mean number of customers, and paying our sister \$4.50 per hour, we can expect to have a total of \$9.00 per hour in materials and labor, with a standard deviation of \$1.50.

Vocabulary

A **random variable** takes on the numerical outcomes of a random process.

The **mean** of a random variable is also known as the **expected value**, and is the value you would expect to be the average outcome of many, many trials.

The **transformation** of a random variable describes the effect of performing operations (+, -, *, /) on the random variable.

Guided Practice

1. If $\mu_X = 5.5$, and $\sigma_X = .8$, what is $\mu_{2.5X}$ and $\sigma_{2.5X}$?
2. Given $\mu_Y = 3.1$, $\sigma_X = .35$, and $\sigma^2_X = .1225$, what would $\mu_{-1.3X}$, $\sigma_{-1.3X}$ and $\sigma^2_{-1.3X}$ be?
3. If random variable Z has $\mu = 14.9$, and $\sigma^2 = 16$, what is the mean and the standard deviation of $4Z + 2$?

Solutions:

1. Multiplying a random variable by a constant affects both mean and standard deviation:

- $\mu_{2.5X} = (2.5)(\mu_X) = (2.5)(5.5) = 13.75$
- $\sigma_{2.5X} = (2.5)(\sigma_X) = (2.5)(0.8) = .2$

2. As above, in Q1:

- $\mu_{-1.3X} = (-1.3)(\mu_X) = (-1.3)(3.1) = -4.03$
- $\sigma_{-1.3X} = (-1.3)(\sigma_X) = (-1.3)(.35) = -0.455$
- $\sigma^2_{-1.3X} = (-1.3)^2(\sigma^2_X) = (1.69)(0.1225) = 0.207$

3. Multiplying by a constant affects both mean and variance, adding a constant affects only the mean, so here we get:

- $\mu_{4Z+2} = 4(\mu_Z) + 2 = 4(14.9) + 2 = 59.6 + 2 = 61.6$
- $\sigma^2_{4Z+2} = 4^2(\sigma^2_Z) = (16)(16) = 256$

Practice

1. If $\mu_X = .23$, and $\sigma_X = .03$, what is $\mu_{3.1X}$ and $\sigma_{3.1X}$?
2. If $\mu_C = 124$, and $\sigma_C = 12$, what is μ_{4C} and σ_{4C} ?
3. If $\mu_Y = 19.2$, and $\sigma_Y = 2.3$, what is $\mu_{-1.9Y}$ and $\sigma_{-1.9Y}$?
4. If $\mu_Z = 48.38$, and $\sigma_Z = 2.27$, what is the μ and σ of $2.3Z$?
5. If $\mu_X = 7.84$, and $\sigma_X = .72$, what is the μ and σ of $-1.98X$?
6. If $\mu_A = \frac{5}{7}$, and $\sigma_A = \frac{1}{7}$, what is the μ and σ of $\frac{3}{14}A$?
7. If $\mu_B = 14.11$, and $\sigma_B = 1.15$, what is the μ and σ of $-0.35B + 2$?
8. If $\mu_A = \frac{3}{7}$, and $\sigma_A = \frac{1}{7}$, what is the μ and σ of $2A - \frac{3}{14}$?
9. If $\mu_Y = 21.3$, and $\sigma_Y = 2.94$, what is the μ and σ of $-3.8Y + 2.13$?
10. If $+1SD$ of $\mu_X = 4.75$ and $-1SD$ of $\mu_X = 3.25$, what is $\mu_{3X+2.25}$ and $\sigma_{3X+2.25}$?
11. If $+1SD$ of $\mu_X = 11.25$ and $-1SD$ of $\mu_X = 9.75$, what is $\mu_{2X-2.25}$ and $\sigma_{2X-2.25}$?
12. If $+2SD$ of $\mu_X = 25$ and $-2SD$ of $\mu_X = 5$, what is μ_{3X-4} and σ_{3X-4} ?
13. If $-2SD$ of $\mu_Z = 9$ and $+2SD$ of $\mu_Z = 15$, what is $\mu_{2Z-.75}$ and $\sigma_{2Z-.75}$?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 7.10.

Students were introduced to the concepts of random variables and probability distributions. Instruction and exercises were provided to familiarize students with the interpretation and construction of various probability distribution tables and graphs. Binomial experiments were introduced, as were transformations of discrete random variables.

7.11 References

1. akshayapatra. <http://pixabay.com/en/school-class-room-children-boys-298680/?oq=boy%20school> .
2. James. <https://www.flickr.com/photos/71217725@N00/126070445> .
3. WikiImages. <http://pixabay.com/en/dollar-coin-currency-money-67726/?oq=coin> .
4. Steven Depolo. <https://www.flickr.com/photos/stevendepolo/4028160990> .
5. . . CC BY-NC-SA
6. . . CC BY-NC-SA
7. . . CC BY-NC-SA
8. . . CC BY-NC-SA
9. . . CC BY-NC-SA
10. . . CC BY-NC-SA
11. . . CC BY-NC-SA
12. . . CC BY-NC-SA
13. . . CC BY-NC-SA
14. . . CC BY-NC-SA
15. amy gizenski. <https://www.flickr.com/photos/agizenski/3778965891> .
16. . . CC BY-NC-SA
17. . . CC BY-NC-SA
18. . . CC BY-NC-SA
19. . . CC BY-NC-SA
20. . . CC BY-NC-SA

CHAPTER 8**Combinations and Permutations****Chapter Outline**

- 8.1 COMBINATIONS AND PERMUTATIONS**
 - 8.2 CALCULATING PERMUTATIONS**
 - 8.3 PERMUTATIONS WITH REPEATS**
 - 8.4 PERMUTATIONS WITH INDISTINGUISHABLE MEMBERS**
 - 8.5 CALCULATING COMBINATIONS**
 - 8.6 CALCULATING COMBINATIONS II**
 - 8.7 USING TECHNOLOGY**
 - 8.8 REFERENCES**
-

The mathematics of combinations, “combinatorics”, is a study in it’s own right, but it is also an important part of the mathematics of probability. There are two basic divisions of combinatorics: *permutations*, which are groupings of items where the *order* of the items is important, and *combinations*, where only the *identity* of the items is important, regardless of the order in which they appear.



In this chapter, you will study both kinds of groupings, and you will learn to calculate the number of possible unique ways to combine or arrange items of all kinds.

8.1 Combinations and Permutations

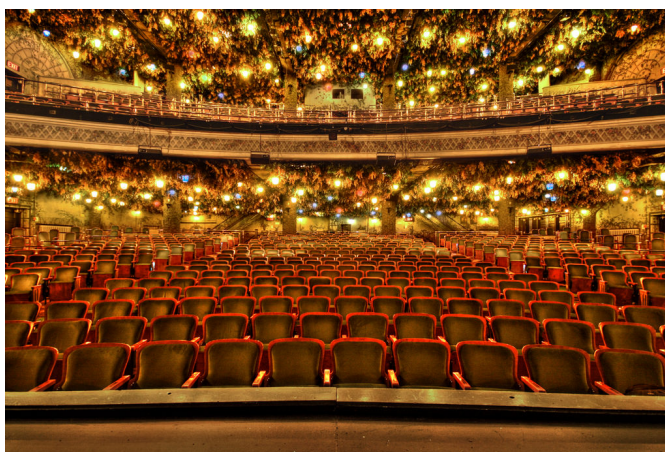
Objective

Here you will learn about combinations and permutations and how to recognize the difference between them.

Concept

Guidance

Combinations and *Permutations* are each different methods of counting the possible number of ways that the members of a set may be selected or arranged. The difference between the two is that *permutations* consider the order in which objects are placed to be important, and *combinations* only consider *which* objects are chosen, ignoring the order.

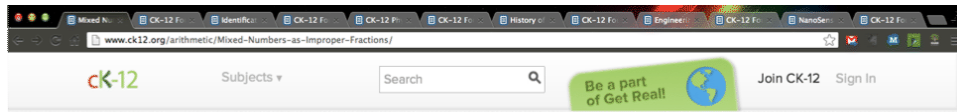


One way to remember the difference between the two is to consider the meanings of the root words: combine and permute. You may not be familiar with the word 'permute', which basically means to mix up the order of something, but you are likely quite comfortable with 'combine', which means group things together.

- **Permutations:** If you are trying to find the number of ways that different items can be put in order, meaning that 1, 2, 3 is one permutation, and 3, 2, 1 is another permutation, each counting as separate entries, you are identifying the count of the different ways you could *permute* or *permute* (mix up) the items. It is quite possible to have multiple **permutations** using exactly the same items.
- **Combinations:** If you are identifying the number of different groups of items, meaning that 1, 2, 3 and 3, 2, 1 are the same group, regardless of order, then you are counting the number of ways you could *combine* (group up) the items. **Combinations** can also be thought of as putting items in a bucket, the order in which you drop the items in does not change *which* items are in the bucket when you finish. In order to have different combinations, you need *different items*, you can't just drop them into the bucket in a different order.

Example A

Suppose you have 12 tabs open in your web browser, and you are curious how many ways they could be reorganized. Is this a combination problem, or a permutation problem, and why?

**Solution:**

Since you are counting the number of different ways the same 12 tabs could be organized or put in order, this is a permutation problem. Each different permutation will have the same twelve tabs, they will just be in a different order.

Example B

Your sports team is planning a trip to a nearby town, about 2 hours away by bus. The driver will allow each rider to bring 2 electronic devices for the trip. You have access to a smartphone, a tablet, an mp3 player, and a laptop computer, and you are curious how many options that allows you to choose. Is this a combination problem, or a permutation problem, and why?

**Solution:**

Regardless of which item you choose first, and which you choose second, the same two items will result in the same devices available to you for the trip, so order does not matter. This is a combinations problem.

Example C

In the United States, telephone numbers are numerical series composed of a 3-digit area code, a 3-digit exchange, and a 4-digit subscriber line identifier. For each of the over 250 different area codes, there are 7,920,000 useable phone numbers (some numbers, particularly those starting with 0 or 1, are unavailable for public use). Is the count of useable phone numbers a product of a permutation calculation, or a combination calculation, and why?

Solution:

Phone numbers are a permutation function, since 555-1234 would ring a different person than 555-4321.

Concept Problem Revisited

Evan, his girlfriend, and 5 of his other friends are going to the movies. Evan wants to be sure that he sits next to his girlfriend, and is curious how many different ways he and his friends can be arranged to sit in the same row. Is this a combination problem or a permutation problem? How can you tell?

This is a permutation problem, since Evan is interested in learning how many different ways he could order the same 7 friends. If he is curious about how many different groups of 3 friends could be chosen to go get drinks and popcorn after they have seats, then he would be dealing with a combination problem.

Vocabulary

Permutations consider the order of items in a group, and count each different order as a different option. A permutation is a unique arrangement of items.

Combinations consider only *which* items are in a group, regardless of order. A combination is a unique group of items.

Guided Practice

1. The back of the box of jellybeans that I have says that there are hundreds of different flavor combinations possible by eating multiple beans at the same time. If I wanted to calculate how many possibilities there were if eating only two beans at a time, would that be a permutation calculation, or a combination calculation?
2. Tuscany has 500 songs in her .mp3 collection, and she wants to make a playlist for her boyfriend. The trick is that she wants to hide a secret message in the playlist by choosing songs so that the first word of each song forms a message. If she wanted to know how many different messages were possible from a playlist 15 songs long, including messages that are gibberish, is this a permutation or combination problem?
3. Scott has 5 dogs, and he feeds them one at a time so they don't fight over the food. If he wants to discover how many different ways he could choose to order them for feeding, is he considering permutations or combinations?
4. Vicki is making gift bags for a party. She has 8 kinds of gifts and the bags hold 3 gifts each. If she wants to know how many different gift bags she can make, is she dealing with permutations or combinations?

Solutions:

1. Since two jellybeans of given flavors will result in the same combined flavor regardless of which I grab from the box first, this is a combination problem.
2. Since the message would be different if the same words were read in a different order, this is a permutation problem.
3. The order of feeding is all that Scott is considering, so this is a permutation problem. The combination of 5 dogs is the same regardless.
4. If Vicki chooses a particular group of 3 gifts for a bag, that bag will end up the same regardless of which of the 3 she chooses first, second, or third. This is a combination problem.

Practice

For problems 1 - 14, identify each situation as either a permutation or a combination consideration:

1. An ice cream store has fifteen different flavors of ice cream, you wonder how many different three-scoop bowls can be made.
2. There is a red, a blue, a green, and a yellow chair around the table. You wonder how many ways can four friends sit around the table.
3. Two friends decide to stay home on a Saturday and watch a movie marathon. Of the twenty movies the friends have to choose from, each friend chooses a first, second, and third choice. How many different six-movie marathons are possible?
4. There are seven rides at the carnival, but you can only afford to ride four of them, how many different groups of four rides are there?
5. How many five-letter groups can you make from the word "grandmother"?
6. Fourteen friends decide to visit each other's houses to trick-or-treat, but they only have time to choose seven of the houses. How many ways could the friends put the houses in order for visiting?
7. How many different pizzas can be made, assuming no double toppings, from seven toppings?

8. How many 5-card hands are possible with 20 different cards?
9. You are responsible for selecting a lead and an understudy for a school play.
10. You are team captain for a tug-of-war, how many different teams could you create from a pool of 30 players?
11. Scott is a safety-conscious rider, and knows he should wear a leather jacket and helmet when he rides a motorcycle. There are 10 different helmets and 7 different jackets to choose from, he wonders many options are possible.
12. John is practicing a card trick to show his granddaughter, he asks her to pick three cards from the deck and put them back into the pile. John wants to know many groups of three he needs to guess from to get it right.
13. Robin is kind of a health food fanatic, and she is making whole wheat pizza with organic sauce, for dinner. She has spinach, kale, fresh tomato, organic anchovy, free-range chicken, and grass-fed beef. Her kids want to know how many different pizzas she could make.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 8.1.

8.2 Calculating Permutations

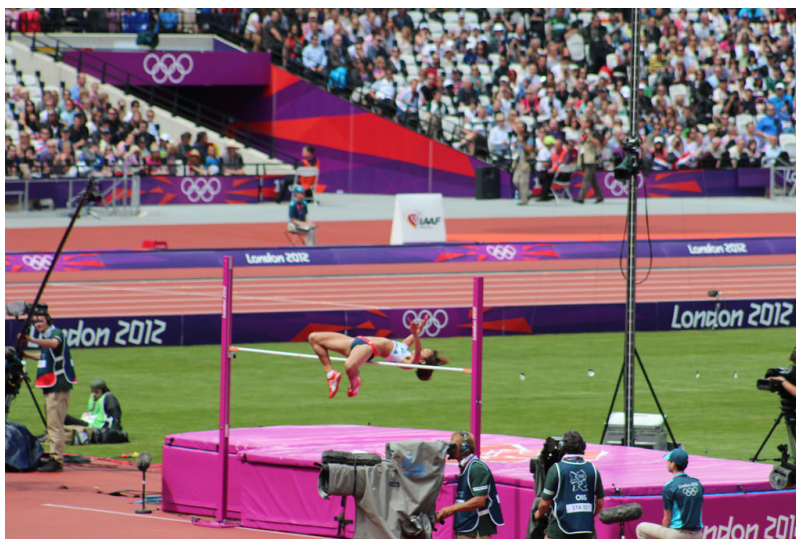
Objective

Here you will learn to calculate basic permutations without repeats or indistinguishable members.

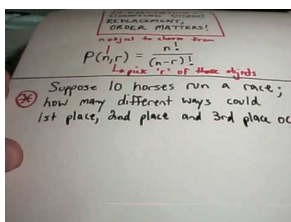
Concept

If the Olympic High Jump has 24 semifinalists, how many possible ways can the competitors be arranged into Gold, Silver, and Bronze winners?

This is a permutation calculation, by the end of the lesson you will have no problem calculating the answer.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67368>

http://youtu.be/-mC_QK6dBIY PatrickJMT - Permutations - Counting Using Permutations

Guidance

Calculating the number of *permutations* possible from a group of objects is a rather simple calculation, but understanding the concept can be a little tricky.



One way to understand the concept is to think of a group of friends seated around a table. If there are three friends and three seats, then the number of ways the friends can sit around the table is a question of permutations.

Suppose the friends are Amber, Brian, and Kelli, and the chairs are one each green, blue, and red.

Let's consider the options, starting with Amber in the green chair:

- | | | |
|------------------|---------------|--------------|
| 1. Green: Amber, | blue: Brian, | red: Kelli |
| 2. Green: Amber, | red: Brian, | blue: Kelli |
| 3. Blue: Amber, | green: Brian, | red: Kelli |
| 4. Blue: Amber, | red: Brian, | green: Kelli |
| 5. Red: Amber, | blue: Brian, | green: Kelli |
| 6. Red: Amber, | green: Brian, | blue: Kelli |

There are six permutations. We can see that each time the first color is chosen, there are two possibilities left for the next person to choose from, and the last person does not get to choose. That's why there are two of each color in each column. If Amber chooses green, then Brian can either choose blue, leaving red for Kelli, or Brian can choose red, leaving blue for Kelli. The same goes for Amber choosing blue or red. Each time, Brian has two colors left to choose from, and Kelli gets the remaining color. The chart would look the same regardless of who chooses first, because no matter who the first person is, there are only two ways the others could sit, meaning there should be (and are) two entries for each color under each person's name.

Fortunately, we don't need to draw out a diagram every time we want to find the number of permutations possible in a situation like this. As long as there are no duplicates or items so similar they can't be told apart (two green chairs, for instance), all we need to know is how to use *factorials*.

Factorials are notated with an exclamation point “!”, and they indicate that you should start with the number before the exclamation point, and count down to 1, multiplying each number by the next. For instance, the first few factorials are:

$$0! = 1 \text{ (by definition)}$$

$$1! = 1$$

$$2! = 2 \times 1 = 2$$

$$3! = 3 \times 2 \times 1 = 6$$

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

To calculate the number of permutations possible given n non-repeating values, calculate $n!$.

To calculate the number of permutations of r values from set n , calculate the first r numbers of $n!$

Alternatively, the formula for counting permutations is:

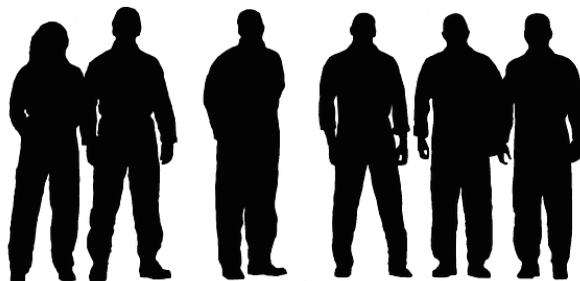
$$\frac{n!}{(n-r)!}$$

Where n is the number of available values, and r is the number of values to be selected.

You may also see this information in the form: ${}_n P_r$

Example A

How many ways can six people line up as they wait in line to buy tickets?



Solution:

Since this is a basic permutation question, with no duplicates and no indistinguishable members, all we need to do is find six **factorial**:

$$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$$

There are 720 ways that six people could line up.

Example B

How many ways can the letters in the word “*factory*” be arranged?

Solution:

“*factory*” has seven letters, all unique, and we want the possible number of arrangements using all seven, so the number of permutations is seven factorial:

$$7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5040$$

There are 5,040 ways to arrange the letter in the word “factory”.

Example C

How many different three-letter permutations are possible using the letters in the word “bread”?

Solution:

This is a slightly different problem than the first two, since we aren’t looking for permutations using all five letters. Since we are only selecting three of the five letters for each arrangement, there will be fewer possible arrangements. As I mentioned at the end of the “Guidance” section, there are a couple of ways to view this type of problem:

- Conceptually, we need to calculate only the first three numbers of five factorial:

$$5 \times 4 \times 3 = 120$$

- Using the formula: $\frac{n!}{(n-r)!}$

$$\frac{5!}{(5-3)!}$$

$$\frac{5!}{2!}$$

$$\frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1}$$

$$5 \times 4 \times 3 = 120$$

There are 120 three-letter permutations of the letters in the word “bread”.

Concept Problem Revisited

If the Olympic High Jump has 24 semifinalists, how many possible ways can the competitors be arranged into Gold, Silver, and Bronze winners?

For this question, we need to calculate the number of permutations of three competitors out of the set of 24 semifinalists. Since we only want arrangements of three, we need to calculate the first three numbers in 24!

$$24 \times 23 \times 22 = 12,144$$

There are 12,144 possible gold, silver, bronze rankings of the 24 semifinalists.

Vocabulary

Permutations are unique *arrangements* of items.

Combinations are unique *groups* of items.

The *factorial* of n is: $n \times (n-1) \times (n-2) \dots 1$. In other words, to calculate five factorial, which would be written “5!”, you would multiply $5 \times 4 \times 3 \times 2 \times 1 = 120$.

Guided Practice

1. How many ways can the letters in “education” be arranged?
2. How many different four-letter arrangements can be made from the word “document”?
3. How many permutations are represented by ${}_8P_5$?
4. How many different ways can the basic colors of a rainbow: red, orange, yellow, green, blue, indigo, and violet, be arranged?

Solutions:

1. There are nine letters in “education”, so we need to calculate $9!$:

$$9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 362,880$$

2. There are eight letters in “document”, but we only want arrangements of four, so we calculate the first four numbers of $8!$:

$$8 \times 7 \times 6 \times 5 = 1,680$$

3. ${}_8P_5$ is read as “pick five items from the eight available”. The formula for permutations is $\frac{n!}{(n-r)!}$:

$$\frac{8!}{(8-5)!}$$

$$\frac{8!}{3!}$$

$$\frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1}$$

$$8 \times 7 \times 6 \times 5 \times 4 = 6720$$

4. There are seven colors in a basic rainbow, and we are looking for the number of unique permutations of all seven.

$$7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5,040$$

Practice

For all questions 1-13, assume no duplicated units are allowed.

1. All the answers for a 5 question multiple - choice test are A, B, C, or D. Find the number of possible answer keys a teacher could have.
2. Calculate $6!$
3. Calculate $4!$
4. You are distributing from a group of 5 items one to each of 2 people. How many different ways can you do this?
5. Calculate ${}_6P_3$
6. Find the number of permutations of 7 distinct items.
7. To unlock a school locker, you need a locker combination consisting of 3 unique numbers from 1 to 9. How many possible locker combinations are there?

8. Calculate ${}_6P_2$
9. Calculate ${}_8P_8$
10. If a bank account number consists of seven unique digits 0-9, how many possible accounts are there?
11. For a dinner party, you need to make a seating arrangement of 9 people. Find the number of different ways of arranging the party.
12. Why can't you calculate $4.5!$?
13. How many five-letter arrangements can be made from the word "number"?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 8.2.

8.3 Permutations with Repeats

Objective

Here you will learn how to calculate the number of permutations possible when repeated uses of set items are allowed.

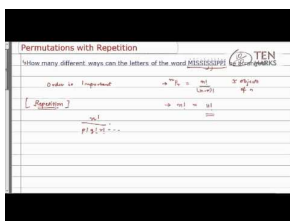
Concept

If a bank account consists of eight digits between zero and nine, and repeated digits are allowed, how many possible account numbers are there?

This is not the same question as it would be if repeats were not allowed. At the end of this lesson, we'll return to this question and review the difference.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67370>

http://youtu.be/L_v4e4Yaypw Ten Marks - Permutations with Repetition

Guidance

When we consider permutations, we often specify that repeated values are not allowed, simply because many real-world situations would not support the concept. For instance, if you are calculating the number of possible seating

arrangements for six people, it would be pretty silly to include any possibilities with the same person in multiple seats (unless you have a cloning machine in the closet!).

That said, there are times when you do need to include duplicates, such as the bank account question in the concept section above. The number of possible permutations in a set can be much greater when repeated values are allowed, but the calculation is actually simpler in concept.

To calculate the number of possible permutations of r items from n available items, simply raise n to the power of r :

$$n^r : (\text{number available})^{\text{number chosen}}$$

Example A

The ice cream shop on the corner carries 27 flavors of ice cream, how many different 4-scoop cones can be created there?

Solution:

There are 27 flavors, so $n = 27$. We are creating 4-scoop cones, so $r = 4$.

$$\text{Number of unique cones} = 27^4 = 531,441$$

Example B

Lockers in your school are each three digits 0-9. How many different combinations are possible?

Solution:

There are ten digits, so $n = 10$. We are looking for arrangements of three digits each, so $r = 3$.

$$\text{Number of locker combinations} = 10^3 = 1000$$

Example C

Keith's trivia challenge team competes in trivia competitions all over the U.S. During a competition, the members are numbered 1-6, and a die is cast before each of the twenty questions in the challenge to decide which team member must answer the question. How many possible ways are there for Keith's team to step up to the podium and answer trivia questions in a single meet?

Solution:

Keith's team has six members, and there are twenty questions. There are six possible choices for the first question, and six possible 2^{nd} choices for each of them, resulting in $6^2 = 36$ possibilities for the first two questions. Since there are twenty questions, the total number of possible lineups is $6^{20} \approx 3.656 \times 10^{15}$.

Concept Problem Revisited

If a bank account consists of eight digits between zero and nine, and repeated digits are allowed, how many possible account numbers are there?

Each digit has ten possibilities, and there are eight digits: $10^8 = 100,000,000$ possible bank accounts.

Vocabulary

Permutations are unique arrangements of items.

Combinations are unique groups of items.

Guided Practice

1. How many different ice cream cones can be made with four scoops of ice cream, if there are 12 flavors to choose from?
2. How many permutations are possible with seven units composed of the digits 0-9, duplication allowed?
3. How many four-letter permutations are possible using the letters of the alphabet?

Solutions:

1. Since duplication is allowed, the number of possible permutations can be calculated with the formula n^r :

$$12^4 = 20,736$$

2. There are ten digits to choose from, and we are making permutations of seven digits each:

$$10^7 = 10,000,000$$

3. There are twenty-six letters, and we are building permutations of four letters each:

$$26^4 = 456,976$$

Practice

For questions 1-12, calculate the number of possible permutations, duplicate values are allowed.

1. Using five decks of cards, permutations of five cards each.
2. Using the letters A-G and numbers 1-5, arrangements of eight units each.
3. Using the letters in the word “combine”.
4. Seven digit arrangements of the numbers 0-9.
5. Ice cream cones with three scoops chosen from 19 flavors.
6. 1 cent, 5 cent, 10 cent, and 25 cent coins, in arrangements of five coins at a time.
7. Letters A-F, in arrangements of six letters each.
8. Roll a 10-sided die seven times.
9. Roll a standard die five times.
10. How many locker combinations are possible using three digits on a ten-digit dial?
11. How many unique passwords can be made from the letters of the word “remix”?
12. How many unique passwords can be made from the letters in “portable”?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 8.3.

8.4 Permutations with Indistinguishable Members

Objective

Here you will learn to calculate the number of permutations possible using strings that have some members that cannot be told apart.

Concept

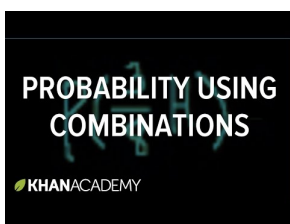
Password

How many unique passwords can be created using the letters in the word “password”?

Since the two “s” letters appear exactly the same, we cannot simply calculate the number of permutations of 8 letters. We also cannot simply calculate as if there were only seven letters, since there are eight locations in which to place letters.

By the time we return to this question after the lesson, you will know how to handle these types of permutations, and the answer should be clear.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67422>

<http://youtu.be/Xqfcy1rqMbl> Khan Academy - Probability using Combinations

Guidance

When a collection of units includes some members that appear the same, the number of permutations is reduced by the number of arrangements that result only from exchanging the identical members.

In order to exclude the number of permutations that are effectively the same due to identical members, we need to divide the number of possible permutations of all the items by the product of the factorials of the number of indistinguishable members. As a formula, this looks like:

$$\frac{n!}{n_1! \times n_2! \times \dots \times n_k!}$$

The explanation of *why* the division works is, unfortunately, beyond the scope of this book. At this point, I recommend learning the formula and applying it as needed, that is, whenever you need to calculate the permutations possible in a collection of items including some that are indistinguishable.

Example A

How many ways can the letters ABCCCB be arranged so that the permutations are distinguishable?

Solution:

Use the formula for calculating permutations with indistinguishable members: $\frac{n!}{n_1! \times n_2! \times \dots \times n_k!}$

- There are six letters, so $n = 6$
- There is one “A”, so $n_1 = 1$
- There are two “B’s”, so $n_2 = 2$
- There are three “C’s”, so $n_3 = 3$

$$\frac{6!}{1! \times 2! \times 3!}$$

$$\frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(1) \times (2 \times 1) \times (3 \times 2 \times 1)}$$

$$\frac{720}{(1) \times (2) \times (6)}$$

$$\frac{720}{12}$$

$$60$$

There are 60 ways to arrange the letters ABCCCB that appear unique.

Example B

A classroom has eight students in it. Three students are wearing red shirts, two are wearing white shirts, two are wearing blue shirts, and one is wearing a green shirt. If the students line up at the door for lunch break, how many ways could the shirt colors be arranged?

Solution:

- There are eight students, so $n = 8$
- There are three red shirts, so $n_1 = 3$
- There are two white shirts, so $n_2 = 2$
- There are two blue shirts, so $n_3 = 2$
- The one green shirt is not a multiple, so we don’t need to worry about it, but it does not hurt to set $n_4 = 1$, since multiplying by one doesn’t change anything anyway.

$$\frac{8!}{(3!)(2!)(2!)(1!)}$$

$$\frac{40,320}{(6)(2)(2)(1)}$$

$$\frac{40,320}{24}$$

$$3,360$$

Example C

Suppose there are twenty-five students in the auditorium and you are responsible for forming them into teams A, B, and C, with six students each, and team D of seven students. How many ways could the students be arranged?

**Solution:**

This problem looks different, but the concept is the same, and you can use the same formula for calculation.

- There are twenty-five students to choose from: $n = 25$
- There are six students in team A: $n_1 = 6$
- There are six students in team B: $n_2 = 6$
- There are six students in team C: $n_3 = 6$
- There are seven students in team D: $n_4 = 7$

$$\frac{25!}{(6!)(6!)(6!)(7!)}$$

$$\frac{1.55 \times 10^{25}}{(720)(720)(720)(5040)}$$

8, 245, 512, 475, 200 - Wow, that is a lot of possible teams!

Concept Problem Revisited

How many unique passwords can be created using the letters in the word “password”?

This one should be easy now, we’ll use the formula $\frac{n!}{n_1! \times n_2! \times \dots \times n_k!}$:

- There are eight letters total, so $n = 8$
- The only doubled letter is “s”, and it appears twice, so $n_1 = 2$
- We don’t need to worry about the other letters, since they would each be 1

$$\frac{8!}{2!}$$

$$\frac{40,320}{2}$$

$$20,160$$

Vocabulary

Permutations are unique arrangements of items.

Indistinguishable items are items that appear the same and cannot be told apart.

Guided Practice

1. How many ways can the letters in the word “banana” be arranged?
2. How many passwords can be made from the letters in “summers”?
3. How many ways can twelve students be organized evenly into classrooms A, B, C, and D?

Solutions:

1. There are six letters, one has three multiples, one has two multiples, and one is singular:

$$\frac{6!}{(3!)(2!)} = 60 \text{ arrangements}$$

2. There are seven letters, two of which each repeat twice:

$$\frac{7!}{(2!)(2!)} = 1260 \text{ passwords}$$

3. This one is just like Example C, there are twelve students, and four “teams” of three students each:

$$\frac{12!}{(3!)(3!)(3!)(3!)} = 369600 \text{ arrangements}$$

Practice

1. How many anagrams are possible using the letters in “possible”?
2. How many anagrams are possible using the letters in “anagram”?
3. Bobby knows he used the letters in his name, and the numbers in his age, 22, to make his e-mail password, mixing them up to make it hard to guess. Unfortunately, he made it *too* hard to guess, and forgot it. If the “reset password” option is unavailable, how many permutations might he need to guess?
4. How many ways can fourteen students be seated, if one desk holds five students, one holds four students, one holds three, and one holds two?
5. How many anagrams are possible using the letters in the word “letters”?
6. There are thirteen pairs of shoes on the rack at a shoe store. There are four pairs of tennis shoes, three pairs of dress shoes, four pairs of high heels, and two pairs of sandals. How many ways can the pairs be arranged on the display shelf?
7. There are five chocolate bars, three vanilla cookies, four chocolate chip cookies, and six snickerdoodle cookies on the counter. How many ways can the sweets be arranged on the counter?

8. There are five apples, three bananas, four oranges, and three pears on a shelf. How many ways can they be arranged?
9. How many ways can thirteen friends be sorted into two groups of four and one group of five?
10. How many ways can six nickels, five pennies, and four quarters be arranged?
11. How many arrangements are possible using the numbers 1, 1, 3, 3, 3, 4, 5 and 5?
12. How many ways can three trucks, four motorcycles, four convertibles, and three SUV's be parked all in a row?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 8.4.

8.5 Calculating Combinations

Objective

Here you will learn how to calculate basic *combinations*, which are arrangements of items that ignore the order of the items.

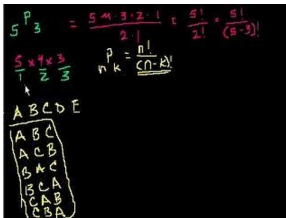
Concept

Suppose there are fifteen students in a classroom, and you are looking to build a team of five students for a tug-of-war. How many different groups of five students could you create?

After the following lesson, you should have no problem solving combination question of this type.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67425>

<http://youtu.be/bCxMhncR7PU> Khan Academy - Combinations

Guidance

In other lessons, we have spent quite a bit of time learning how to count *permutations*, which are arrangements of items that take order into account. Now we will be considering *combinations*, which are groups of items only concerned with which items are in the arrangement, regardless of the order.

Naturally there are fewer unique combinations of a particular group of items than there are permutations, since there may be many permutations of the same single combination of items. In fact, the formula for counting combinations without repetition is actually just a modification of the formula for permutations where the denominator is increased. Take a look at the two formulas side-by-side:

$$\text{Permutations: } \frac{n!}{(n-r)!} \quad \text{Combinations: } \frac{n!}{(n-r)! \times r!}$$

Where n = number of objects to choose from, and r = the number chosen

Notice how the combination formula is the same, with only the addition of the $r!$ in the denominator of the fraction. This makes sense conceptually, given that we need to reduce the number of arrangements that are the same items in a different order.

In summary, to calculate the number of **combinations** possible from n items chosen r at a time, use the combination formula: $\frac{n!}{(n-r)! \times r!}$

Example A

The local ice cream shop carries twelve flavors. How many different three-scoop bowls are possible?

Solution:

There are twelve flavors, so $n = 12$. We are creating bowls with three scoops each, so $r = 3$:

$$\begin{aligned} & \frac{12!}{(12-3)! \times 3!} \\ & \frac{12!}{9! \times 3!} \\ & \frac{12 \times 11 \times 10 \times \cancel{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}}{(\cancel{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1})(3 \times 2 \times 1)} \quad (\text{cancel terms}) \\ & \frac{1320}{6} \\ & 220 \text{ different flavors} \end{aligned}$$

Example B

If there are thirteen students in a room, how many ways can a group of five be formed from them?

Solution:

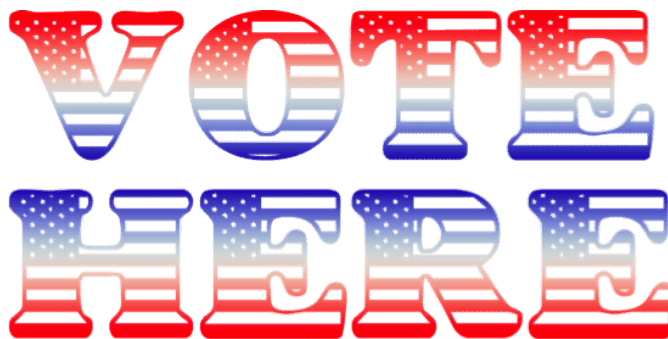
There are thirteen students to choose from, so $n = 13$

We are choosing five at a time, so $r = 5$

$$\begin{aligned} & \frac{13!}{(13-5)! \times 5!} \\ & \frac{13!}{(8!)(5!)} \\ & \frac{13 \times 12 \times 11 \times 10 \times 9}{5!} \\ & \frac{154440}{120} \\ & 1287 \text{ possible groups of five} \end{aligned}$$

Example C

If your student council has seven members, how many ways are there to reach a majority vote?

**Solution:**

There are seven members to choose from, so $n = 7$. A majority vote is either four, five, six, or seven votes, so we need to calculate the number of combinations for each $r = 4, 5, 6$, and 7 , and add them together.

a. Four votes:

$$\frac{7!}{(7-4)! \times 4!}$$

$$\frac{7 \times 6 \times 5 \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}}{(\cancel{3} \times \cancel{2} \times \cancel{1})(4 \times 3 \times 2 \times 1)}$$

$$\frac{840}{24} = 35$$

b. Five votes:

$$\frac{7!}{(7-5)! \times (5!)}$$

$$\frac{7 \times 6 \times 5 \times 4 \times 3}{5 \times 4 \times 3 \times 2 \times 1}$$

$$\frac{2520}{120} = 21$$

c. Six votes:

$$\frac{7!}{(7-6)! \times 6!}$$

$$\frac{7 \times 6 \times 5 \times 4 \times 3 \times 2}{6 \times 5 \times 4 \times 3 \times 2 \times 1}$$

$$\frac{5040}{720} = 7$$

d. Seven votes:

$$\frac{7!}{(7-7)! \times 7!}$$

$$\frac{7!}{7!} = 1$$

Total = $35 + 21 + 7 + 1 = 64$ ways to reach a majority vote

Concept Problem Revisited

Suppose there are fifteen students in a classroom, and you are looking to build a team of five students for a tug-of-war. How many different groups of five students could you create?

This is a combinations problem requiring us to choose items five at a time from a pool of 15, which can be notated as $15C5$ (read as "fifteen, choose five"). Use the combinations formula with $n=15$ and $r=5$:

$$\begin{aligned} \text{Combinations} &= \frac{n!}{(n-r)! \times r!} \\ &= \frac{15!}{((15-5)! \times 5!)} \\ &= \frac{15!}{10! \times 5!} \\ &= \frac{15 \times 14 \times 13 \times 12 \times 11}{5!} \\ &= \frac{360,360}{120} \end{aligned}$$

3,003 different teams

Vocabulary

A **combination** is a unique *collection* of items.

A **permutation** is a unique *arrangement* of items.

Guided Practice

1. How many ways can six cars be chosen from a lot containing twenty-one cars?
2. How many unique groups of four can be made from nine pieces of candy?
3. How many unique five-card hands are possible using a standard deck of cards?

Solutions:

1. There are twenty-one cars to choose from, $n = 21$

There are six cars in each group, $r = 6$

$$\begin{aligned} &\frac{21!}{((21-6)! \times 6!)} \\ &\frac{21 \times 20 \times 19 \times 18 \times 17 \times 16}{6!} \\ &54264 \text{ combinations} \end{aligned}$$

2. Nine pieces to choose from, $n = 9$, and four in each group, $r = 4$:

$$\frac{9!}{(9-4)! \times 4!}$$

$$\frac{9 \times 8 \times 7 \times 6}{4!}$$

756 groups

3. There are fifty-two cards in a standard deck, $n = 52$, and five in each hand, $r = 5$:

$$\frac{52!}{(52-5)! \times 5!}$$

$$\frac{52 \times 51 \times 50 \times 49 \times 48}{5!}$$

62,375,040 possible hands

Practice

- In the ball closet, there are 6 basketballs, footballs, and baseballs. You go in and pick 2 items at random. Assuming all the balls are marked differently, how many different groups of balls can you end up with?
- You go to a store and buy 4 various items from 6 distinct choices. How many different possibilities were there?
- Find the number of combinations of choosing 5 items from 9 distinct items.
- Find the number of combinations of choosing 6 items from 9 distinct items.
- You have a pile of 7 distinctly different coins. How many ways can you take 2 coins from the pile?
- You are dealt a hand of 4 cards from a deck of 6 cards. Find the number of possible hands you can get.
- You hire 2 out of 9 people to build a garden. Find the number of possible ways of choosing your workers.
- Find the number of combinations of choosing 4 items from 7 distinct items.
- Find the number of combinations of choosing 5 items from 5 distinct items.
- How many five-card hands are possible from a deck of twenty-six cards?
- How many unique 7-person combinations are possible using 32 people?
- If Pandi has nineteen chickens in her coop, how many unique groups of six chickens are possible from them?
- If one of her chickens lays colored eggs, every egg unique, how many four-color groups could she make from a dozen eggs?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 8.5.

8.6 Calculating Combinations II

Objective

Here you will learn about calculating combination in special cases, such as repeating or indistinguishable members, and you will practice using generalized combination notation.

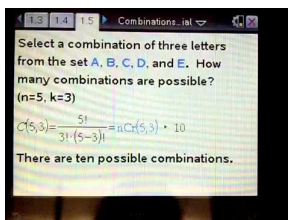
Concept

How many different ways could you select three coins from a pile of coins containing quarters, dimes, nickels, and pennies, if you can take as many of each kind as you like to make up the total of three?

At the end of the lesson, we'll return to this problem and see if we can figure it out.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67427>

<http://youtu.be/sgicxaIux8Y> ews773 - Combinations (with and without repetition)

Guidance

Counting combinations when some members of the source set either cannot be told apart or may be used multiple times is a bit different from calculating them otherwise. In order to practice these types of combination calculations, we are going to use **combination notation**, which is similar to the **permutation notation** that we discussed in a prior lesson.

There are two common types of **combination notation**: ${}_nC_r$ and $C(n, r)$

- n = number of different units to choose from
- r = number of units in each group.

Once you know n and r , you can use the formula we reviewed in *Calculating Combinations I* if your problem involves combinations **without repeats**:

$$\frac{n!}{(n-r)! \times r!}$$

However, to calculate combinations **with repetition** allowed, you need a different formula:

$$\frac{(n+r-1)!}{(n-1)! \times r!}$$

Example A

How many unique three-scoop bowls of ice cream are possible with twelve flavors to choose from?

Solution:

This is a combination question, with repeated choices allowed (it is OK to have a bowl with three scoops of chocolate). Since there are 12 choices, and we choose 3, the notation is: ${}_{12}C_3$, and we need to use the formula for repeat-allowed combinations: $\frac{(n+r-1)!}{(n-1)! \times r!}$. Here, $n = 12$ and $r = 3$.

$$\begin{aligned} & \frac{(12 + (3 - 1))!}{(12 - 1)! \times 3!} \\ & \frac{14!}{11! \times 3!} \\ & \frac{14 \times 13 \times 12}{3 \times 2 \times 1} \\ & \frac{2184}{6} \end{aligned}$$

364 unique bowls of ice cream

Example B

How many unique teams of four can be made from thirteen people?

Solution:

This is a combination without repetition question (the same person cannot be two different members of the team). Since there are thirteen different people, and we are choosing four, the notation is ${}_{13}C_4$, and we need to use the combination without repetition formula: $\frac{n!}{(n-r)! \times r!}$, where $n = 13$ and $r = 4$.

$$\frac{13!}{(13-4)! \times 4!}$$

$$\frac{13!}{9! \times 4!}$$

$$\frac{13 \times 12 \times 11 \times 10}{4!}$$

$$\frac{17160}{24}$$

$$715 \text{ teams}$$

Example C

How many unique handfuls of candy, containing less than five pieces each, are possible with nine flavors to choose from?

**Solution:**

Since the problem specifies “less than five pieces”, we need to complete multiple steps because we need to know how many combinations (repetition allowed) are possible from four, three, two, and one piece of candy. For each number, we need to calculate the number of possible combinations, using the “repetition allowed” formula: $\frac{(n+r-1)!}{(n-1)! \times r!}$

1. One piece: This is easy, one possible option from each of nine flavors: **9 combinations**
2. Two pieces: ${}_9C_2 = \frac{(9+2-1)!}{(9-1)! \times 2!} = \frac{10!}{8! \times 2!} = \frac{10 \times 9}{2 \times 1} = \frac{90}{2} = 45 \text{ combinations}$
3. Three pieces: ${}_9C_3 = \frac{(9+3-1)!}{(9-1)! \times 3!} = \frac{11!}{8! \times 3!} = \frac{11 \times 10 \times 9}{3 \times 2 \times 1} = \frac{990}{6} = 165 \text{ combinations}$
4. Four pieces: ${}_9C_4 = \frac{(9+4-1)!}{(9-1)! \times 4!} = \frac{12!}{8! \times 4!} = \frac{12 \times 11 \times 10 \times 9}{4 \times 3 \times 2 \times 1} = \frac{11880}{24} = 495 \text{ combinations}$

TOTAL = $9 + 45 + 165 + 495 = 714$ possible handfuls

Concept Problem Revisited

How many different ways could you select three coins from a pile of coins containing quarters, dimes, nickels, and pennies, if you can take as many of each kind as you like to make up the total of three?

Now that we have completed a few similar examples, you should recognize that this could be written as a ${}_4C_3$ question, with repeats allowed, meaning that we should use the formula: $\frac{(n+r-1)!}{(n-1)! \times r!}$

$$n = 4 \text{ and } r = 3$$

$$\frac{6!}{3! \times 3!}$$

$$\frac{6 \times 5 \times 4}{6}$$

$$\frac{120}{6}$$

$$20 \text{ ways}$$

Vocabulary

Combination notation can be seen in the form of ${}_n C_r$ or $C(n, r)$, and indicates the number of possible ways to *combine* n objects in groups of r objects each.

Permutation notation is the form ${}_n P_r$ or $P(n, r)$, and indicates the number of ways that n objects can be *ordered* into groups of r items each.

Guided Practice

1. How many ways can a five-person team be chosen from twenty-three people?
2. How many three-color combinations can be selected from the seven-color basic rainbow?
3. How should ${}_8 C_2$ be read? How many combinations does it indicate?

Solutions:

1. There are 23 people to choose from, so $n = 23$. We are choosing members in groups of 5, so $r = 5$. Use the formula for non-repeating combinations: $\frac{n!}{(n-r)! \times r!}$

$$\frac{23!}{(23-5)! \times 5!}$$

$$\frac{23 \times 22 \times 21 \times 20 \times 19}{5!}$$

$$\frac{4037880}{120}$$

$$33,649 \text{ possible teams}$$

2. This is a ${}_7 C_3$ calculation, repeats allowed. $n = 7, r = 3$, use the formula for repeating combinations: $\frac{(n+r-1)!}{(n-1)! \times r!}$

$$\frac{(7+3-1)!}{(7-1)! \times 3!}$$

$$\frac{9!}{6! \times 3!}$$

$$\frac{9 \times 8 \times 7}{3 \times 2 \times 1}$$

$$\frac{504}{6}$$

$$84 \text{ color combinations}$$

3. ${}_8C_2$ is read as “Choose 2 items at a time from 8 possibilities”

To calculate the number of combinations indicated, note that $n = 8$ and $r = 2$. There are two possible solutions, depending on whether the situation allows for repeats.

If repeats are allowed, we use $\frac{(n+r-1)!}{(n-1)! \times r!}$, and we get: $\frac{(8+2-1)!}{(8-1)! \times 2!} = \frac{9!}{7! \times 2!} = \frac{9 \times 8}{2} = 36$

If repeats are not allowed, use $\frac{n!}{(n-r)! \times r!}$, to get: $\frac{8!}{(8-2)! \times 2!} = \frac{8!}{6! \times 2!} = \frac{8 \times 7}{2} = 28$

Practice

For questions 1-10, calculate the number of combinations indicated by the combination notation, on **odd-numbered questions, assume repeats are allowed**, otherwise assume no repeats.

1. ${}_5C_3$
2. ${}_7C_2$
3. ${}_8C_3$
4. ${}_{13}C_7$
5. ${}_{12}C_4$
6. ${}_6C_6$
7. ${}_6C_5$
8. ${}_8C_5$
9. ${}_{11}C_3$
10. ${}_{11}C_3$
11. How many unique 6-player teams can be picked from a pool of 19 players?
12. You are part of an 8-person trivia team. When you compete in trivia tournaments, players are chosen to answer questions based on random numbers generated between 1 and 8. If a tournament consists of 20 questions, how many ways could the team members be chosen to answer trivia questions?
13. You work in an ice cream shop. The most popular cone in the store is the “Monster Bellyache”, a massive four-scoop ice-cream bowl. If the shop carries seventeen flavors, how many unique “Monster Bellyaches” are there?
14. The same ice cream shop carries three other options for ice-cream bowls: “Big Bellyache” (3 scoops), “Brain Freeze” (2-scoops), and “Diet Denter” (1-scoop). How many unique bowls of ice cream are possible, counting all four types of bowls?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 8.6.

8.7 Using Technology

Objective

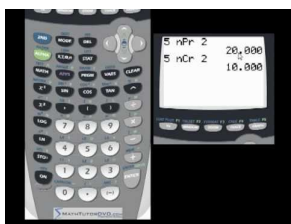
Here you will learn of a few ways that combinations and permutations can be quickly calculated using technology.

Concept

Suppose you wanted to compare the number of possible permutations and combinations possible by choosing six, seven, or eight cards from either one or two standard card decks. Manually calculating all of those different cases could take quite a while. How could you use technology to simplify the process?

Watch This

This video describes how to use a *TI-84 calculator* to calculate factorials, combinations, and permutations. The video will start partway through, as the first couple of minutes discussed random numbers, which are not part of our current studies.



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67429>

<http://youtu.be/TuowYf5LULw?t=2m28s> mathtutordvd - Permutations, Combinations, Factorials with the TI-84 Calculator

Guidance

Although calculating combinations and permutations 'by hand' is an excellent skill worthy of practice, since it is the best way to come to understand combinations and permutations, there are much more efficient ways to find the number of possibilities.

As you saw in the video above, the TI-84 calculator, which is very common in upper mathematics courses of all kinds, can be used to calculate combinatorics questions. Even more efficient, however, are some of the free and freely available online calculators. Here are a few examples:

<http://www.mathsisfun.com/combinatorics/combinations-permutations-calculator.html>

- This all-in-one combinatorics calculator is very simple to use, and can handle most basic combinatorics problems. The interface is self-explanatory, simply enter the number of items available to choose from (the n count), and the number to be chosen (the r count), specify if order is important (permutation or combination), and if repeats are allowed.

Example A

Use a calculator to evaluate ${}_7C_4$ twice, once with repetitions allowed, and once without.

Solution:

Let's use the *Calculator soup* calculator:

- First, to calculate the number of combinations without repeats, choose the calculator labeled “**Combinations Calculator (nCr)**”. Enter “7” for n and “4” for r , and click “*calculate*”. Your screen should look like the image here.

Find the Combinations:
 $C(n,r)$ where:
 $n = 7$
 $r = 4$
 Calculate
 $C(7,4) = 7! / (4! (7 - 4)!) =$
35

There are 35 possible combinations without repeats.

- Second, to calculate the number of combinations possible when repeats are allowed, choose the calculator labeled “**Combinations Replacement Calculator CR(n,r)**”. Again, enter “7” for “ n ”, and “4” for “ r ”, and click “*calculate*”. Your screen should look like the image here.

Find the Combinations:
 $C^R(n,r)$ where:
 $n = 7$
 $r = 4$
 Calculate
 $CR(7,4) = (7+4-1)! / 4! (7 - 1)! =$
210

There are 210 possible combinations with repeats.

Example B

Use technology to calculate the number of possible permutations of the numbers 1-6 and letters A-F, both with and without repeats.

Solution:

Let's use the “Math is Fun” website this time:

- First, to calculate without repeats, enter “12” in both the “*Types to choose from*”, and “*Number chosen*” fields (since we want to know the number of permutations using all six numbers and all six letters), and set “*Is Order Important?*” to “yes”. Your screen should look like the image here.

Combinations and Permutations

Types to choose from? (n)
 Number Chosen? (r)
 Is Order important? ▼
 Is Repetition allowed? ▼

Permutations:

There are apx 479,000,000 permutations possible without repeats.

- Second, to calculate with repeats, all we need to do is change the dropdown menu labeled “*Is Repetition Allowed*” to “yes”. Your screen should show something similar to the image here.

Combinations and Permutations

Types to choose from? (n)
 Number Chosen? (r)
 Is Order important? ▼
 Is Repetition allowed? ▼

Permutations:

There are apx 8,916,100,000,000 permutations possible with repeats.

Example C

Use technology to evaluate ${}_{13}P_{11}$, both with and without repeats.

Solution:

Let’s use the “Calculator soup” calculator for this last one.

- To calculate without repeats, choose the “*Permutations Calculator (nPr)*” and enter “13” for n and “11” for r .

Click “*calculate*”, and you should see the output:

$$P(13, 11) = \frac{13!}{(13 - 11)!} = 3,113,510,400$$

- To calculate with repeats, return to the list of statistics calculators, choose “*Permutations Replacement Calculator PR(n,r)*”, and enter “13” for n and “11” for r .

Click “*calculate*”, and you should get:

$$PR(13, 11) = nr = 1311 = 1,792,160,394,037$$

Concept Problem Revisited

Suppose you wanted to compare the number of possible permutations and combinations possible by choosing six, seven, or eight cards from either one or two standard card decks. Manually calculating all of those different cases could take quite a while. How could you use technology to simplify the process?

By now, you should have no problems finding a technology resource to simplify this question. If you use the “**Math is Fun**” site, you just need to run through the various inputs for $n = 52$ (one deck) and $n = 104$ (two decks), $r = 6, 7$, and 8 (for the number of cards chosen), and “*order matters*” = yes/no (to evaluate combinations and permutations).

Vocabulary

The *TI-84 calculator* is sort of an “industry standard” for more advanced calculations. Many upper-level mathematics classes assume the availability of a TI-84 or the equivalent.

Combinatorics is the study of permutations and combinations.

Guided Practice

Evaluate the combinations and permutations, you may use technology.

- ${}_9C_5$: No repeats
- ${}_7P_6$: Repeats allowed
- ${}_{10}C_2$: Repeats allowed
- ${}_8P_4$: No repeats

Solutions: Using “**Math is Fun**”

- $n = 9, r = 5$, “Is Order Important” = no (since these are combinations), “Is Repetition Allowed” = no

There are 126 possible combinations

- $n = 7, r = 6$, “Is Order Important” = yes (since these are permutations), “Is Repetition Allowed” = yes

There are 117,649 possible permutations

- $n = 10, r = 2$, “Is Order Important” = no (combinations), “Is Repetition Allowed” = yes

There are 55 unique combinations

- $n = 8, r = 4$, “Is Order Important” = yes (permutations), “Is Repetition Allowed” = no

There are 1680 unique permutations

Practice

Evaluate the combinations and permutations, you may use technology.

- ${}_5C_3$: No repeats

2. ${}_{12}P_6$: Repeats allowed
3. ${}_8C_3$: Repeats allowed
4. ${}_{18}P_{17}$: No repeats
5. ${}_9C_9$: No repeats
6. ${}_7P_7$: Repeats allowed
7. ${}_{10}C_{10}$: Repeats allowed
8. ${}_3P_{12}$: Repeats allowed
9. ${}_5C_9$: Repeats allowed
10. ${}_4P_6$: Repeats allowed
11. ${}_3C_{21}$: Repeats allowed
12. ${}_6P_3$: No repeats
13. ${}_6C_{15}$: Repeats allowed
14. ${}_7P_{61}$: Repeats allowed
15. ${}_{10}C_2$: No repeats

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 8.7.

Students were introduced to the concepts of combinatorics. The differences between permutations and combinations were discussed. Students practiced the use of the formulas for calculating the number of possible combinations or permutations in various situations allowing or disallowing repeats and/or indistinguishable members.

8.8 References

1. Dirk Haun. <https://www.flickr.com/photos/dhaun/6855772085> .
2. Canon in 2D. <https://www.flickr.com/photos/16462767@N00/3286021531> .
3. . . CC BY-NC-SA
4. Yutaka Tsutano. <https://www.flickr.com/photos/ivyfield/4486938721> .
5. Steve Fair. <https://www.flickr.com/photos/buster1976/7724059864> .
6. . . CC BY-NC-SA
7. savid. <http://pixabay.com/en/silhouette-group-people-279758/?oq=people> .
8. Ben Salter. https://www.flickr.com/photos/ben_salter/5068541282/ .
9. . . CC BY-NC-SA
10. Rebecca MacKinnon. <https://www.flickr.com/photos/rebeccamack/2541386134> .
11. Holidayextras. <https://www.flickr.com/photos/holiday-extras/9270793598> .
12. . . CC BY-NC-SA
13. Jason Eppink. <https://www.flickr.com/photos/jasoneppink/4063640521> .
14. Steven Depolo. <https://www.flickr.com/photos/stevendepolo/4605640584> .
15. . . CC BY-NC-SA
16. . . CC BY-NC-SA
17. . . CC BY-NC-SA
18. . . CC BY-NC-SA
19. . . CC BY-NC-SA
20. . . CC BY-NC-SA
21. . . CC BY-NC-SA

CHAPTER

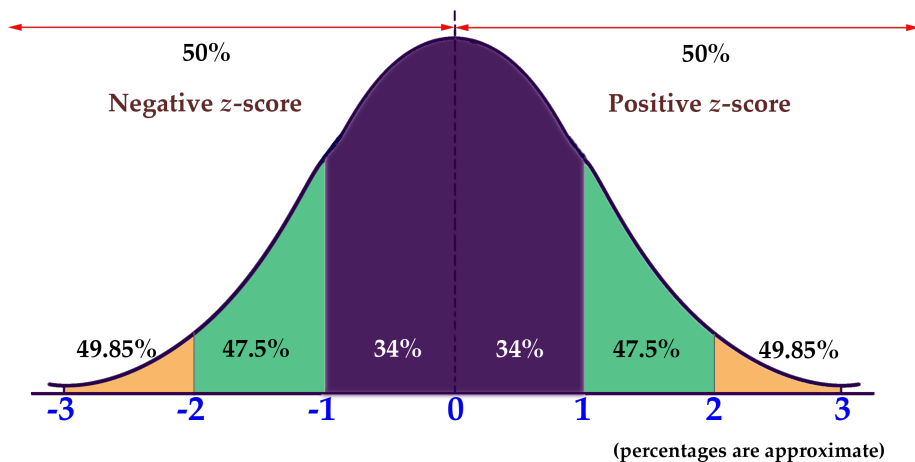
9

The Normal Distribution

Chapter Outline

- 9.1 UNDERSTANDING NORMAL DISTRIBUTION
- 9.2 THE EMPIRICAL RULE
- 9.3 Z-SCORES
- 9.4 Z SCORES II
- 9.5 Z-SCORES III
- 9.6 THE MEAN OF MEANS
- 9.7 CENTRAL LIMIT THEOREM
- 9.8 APPROXIMATING THE BINOMIAL DISTRIBUTION
- 9.9 REFERENCES

The normal distribution can be found practically everywhere, from the distribution of human heights to IQ scores. Understanding the application of the normal distribution can simplify the calculation of binomial probabilities, such as the probability of flipping heads 37 or more times out of 60. By using the Central Limit Theorem, you can evaluate the probabilities associated with many different random variables.



9.1 Understanding Normal Distribution

Objective

Here you will learn about the Normal Distribution. You will learn what it is and why it is important, and you will begin to develop an intuition for the rarity of a value in a set by comparing it to the mean and standard deviation of the data.

Concept



If you knew that the prices of t-shirts sold in an online shopping site were *normally distributed*, and had a mean cost of \$10, with a standard deviation of \$1.50, how could that information benefit you as you are looking at various t-shirt prices on the site? How could you use what you know if you were looking to make a profit by purchasing unusually inexpensive shirts to resell at prices that are more common?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67462>

<http://youtu.be/47099905> CK-12.org - Chapter6SpreadofaNormalDistributionA

Guidance

A *distribution* is an evaluation of the way that points in a data set are clustered or spread across their *range* of values. A *normal distribution* is a very specific symmetrical distribution that indicates, among other things, that exactly $\frac{1}{2}$ of the data is below the mean, and $\frac{1}{2}$ is above, that approximately 68% of the data is within 1, approximately 96% of the data is within 2, and approximately 99.7% is within 3 *standard deviations* of the mean.

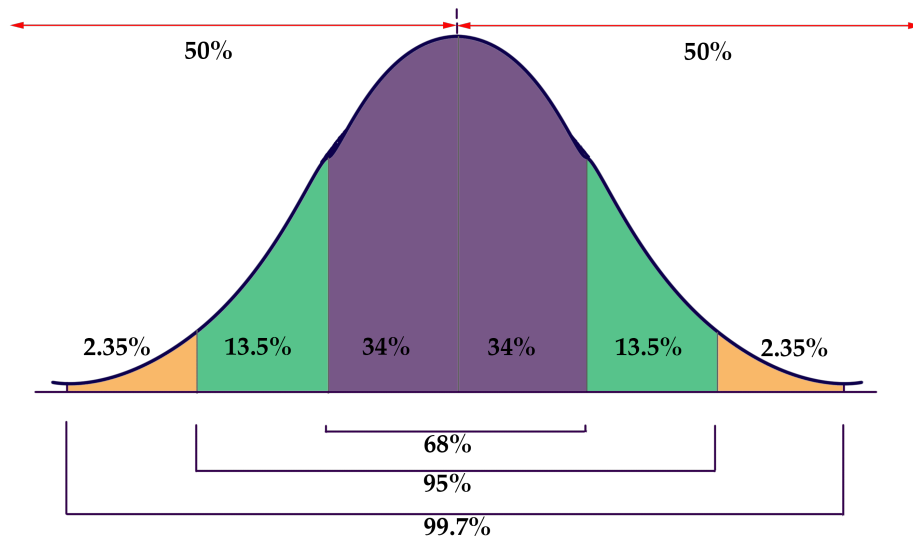
There are a number of reasons that it is important to become familiar with the normal distribution, as you will discover throughout this chapter. Examples of values associated with normal distribution:

- Physical characteristics such as height, weight, arm or leg length, etc.
- The percentile rankings of standardized testing such as the ACT and SAT
- The volume of water produced by a river on a monthly or yearly basis
- The velocity of molecules in an ideal gas

Knowing that the values in a set are exactly or approximately normally distributed allows you to get a feel for how common a particular value might be in that set. Because the values of a normal distribution are predictably clustered around the mean, you can estimate in short order the rarity of a given value in the set. In our upcoming lesson on the Empirical Rule, you will see that it is worth memorizing that normally distributed data has the characteristics mentioned above:

- 50% of all data points are above the mean and 50% are below
- Apx 68% of all data points are within 1 standard deviation of the mean
- Apx 95% of all data points are within 2 standard deviations of the mean
- Apx 99.7% of all data points are within 3 standard deviations of the mean

In this lesson, we will be practicing a 'rough estimate' of the probability that a value within a given range will occur in a particular set of data, just to develop an intuition of the use of a normal distribution. In subsequent lessons, we will become more specific with our estimates. The image below will be used in greater detail in the lesson on the Empirical Rule, but you may use it as a reference for this lesson also.



Example A

Human height is commonly considered an approximately normally distributed measure. If the mean height of a male adult in the U.S.A. is 5'10", with a standard deviation of 1.5", how common are men with heights greater than 6'2"?

Solution:

Since each standard deviation of this normally distributed data is 1.5", and 6'2" is 4" above the mean for the population, 6'2" is nearly 3 standard deviations above the mean. That tells us that men taller than 6'2" are quite rare in this population.

Example B

If the fuel mileage of a particular model of car is normally distributed, with a mean of 26 mpg and a standard deviation of 2 mpg, how common are cars with a fuel efficiency of 24 to 25 mpg?



Solution:

We know that apx 68% of the cars in the population have an efficiency of between 24 and 28 mpg, since that would be 1 SD below and 1 SD above the mean. That suggests that apx 34% have an efficiency of 24 to 26 mpg, so we can say that it is uncommon to see a car with an efficiency between 24 and 25 mpg, but not extremely so.

Example C

If the maximum jumping height of U.S. high school high jumpers is normally distributed with a mean of 5'11.5" and a SD of 2.2", how unusual is it to see a high school jumper clear 6'3"?

Solution:

If the mean is 5'11.5", then 1 SD above is 6'1.7" and 2 SD's is 6'3.9". That means that less than 2.5% of jumpers 6'3.9", so it would be pretty uncommon to see a high-school competitor exceed 6'3".

Concept Problem Revisited

*If you knew that the prices of t-shirts sold in an online shopping site were **normally distributed**, and had a mean cost of \$10, with a standard deviation of \$1.50, how could that information benefit you as you are looking at various t-shirt styles and designs on the site? How could you use what you know if you were looking to make a profit by purchasing unusually inexpensive shirts to resell at prices that are more common?*

By knowing the mean and SD of the shirt prices, and knowing that they are normally distributed, you can estimate right away if a shirt is priced at a point significantly below the norm. For instance, with this data, we can estimate that a shirt priced at \$7.00 is less expensive than apx 97.5% of all shirts on the site, and could likely be resold at a profit (assuming there is not something wrong the shirt that is not obvious from the listing).

Vocabulary

A **distribution** is an arrangement of values of a variable showing their observed or theoretical frequency of occurrence.

The **range** of values of a distribution is the difference between the least and greatest values.

The **normal distribution** is a very specific distribution that is symmetric about its mean. Half the values of the random variable are below the mean and half are above the mean. Approximately 68% of the data is within 1 **standard deviation** of the mean, apx 96% is within 2 SD's, and 99.7% within 3 SD's.

A **standard deviation** is measure of how spread out the data is from the mean. To determine if a data value is far from the mean, determine how many standard deviations it is from the mean. The SD is calculated as the square root of the variance.

Guided Practice

For questions 1-4, assume the data to be normally distributed, and describe the rarity of an event using the following scale: 0% – < 1% probability = very rare, 1% – < 5% = rare, 5% – < 34% = uncommon, 34% – < 50% = common, 50% – 100% = likely.

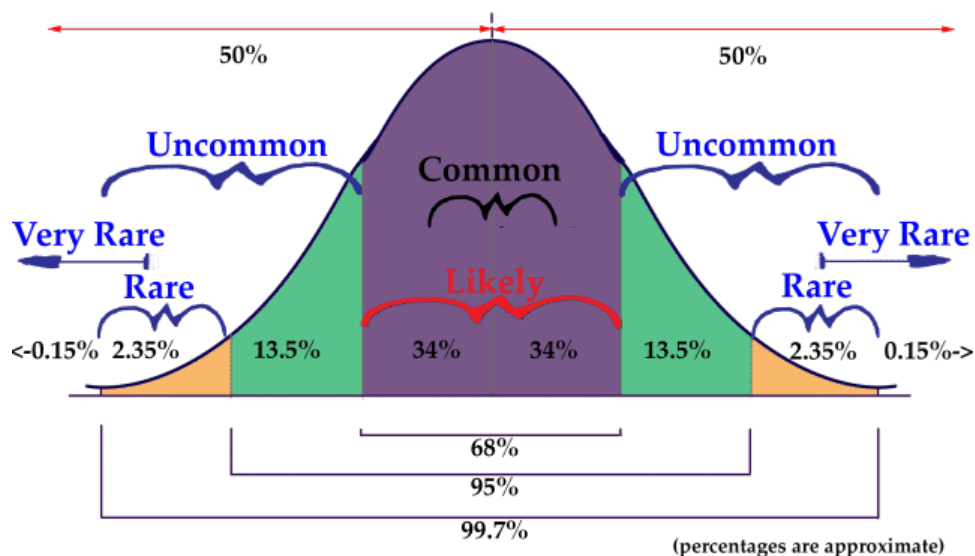
1. If the mean (μ) of the data is 75, and the standard deviation (σ) is 5, how common is a value between 70 and 75?
2. If the μ is .02 and the σ is .005, how common is a value between .005 and .01?
3. If the μ is 1280 and the σ is 70, how common is a value between 1210 and 1350?
4. If the mean defect rate at a cellphone production plant is .1%, with a standard deviation of .03%, would it seem reasonable for a quality assurance manager to be concerned about 3 defective phones in a single 1000 unit run?

Solutions:

1. A value of 70 is only 1 standard deviation below the mean, so a value between 70 and 75 would be expected approximately 34% of the time, so it would be **common**.
2. A value of .01 is 2 SD's below the mean, and .005 is 3 SD's below, so we would expect there to be about a 2.5% probability of a value occurring in that range. A value between 0.005 and 0.01 would be **rare**.
3. 1210 is 1 SD *below* the mean, and 1350 is 1 SD *above* the mean, so we would expect approximately 68% of the data to be in that range, meaning that it is **likely** that a value in that range would occur.
4. .1% translates into 1 per thousand, with a standard deviation of 3 per ten thousand. That means that 3 defects in the same thousand is nearly 7 SD's above the mean, well into the **very rare** category. While it is not impossible for random chance to result in such a value, it would certainly be prudent for the manager to investigate.

Practice

Assume all sets/populations to be approximately normally distributed, and describe the rarity of an event using the following scale: 0% – < 1% probability = very rare, 1% – < 5% = rare, 5% – < 34% = uncommon, 34% – < 50% = common, 50% – 100% = likely. You may reference the image below:



1. Scores on a certain standardized test have a mean of 500, and a standard deviation of 100. How common is a score between 600 and 700?

2. Considering a full-grown show-quality male Siberian Husky has a mean weight of 52.5 lbs, with SD of 7.5 lbs, how common are male huskies in the 37.5 - 45 lbs range?
3. A population $\mu = 125$, and $\sigma = 25$, how common are values in the 100 - 150 range?
4. Population $\mu = 0.0025$ and $\sigma = 0.0005$, how common are values between 0.0025 and 0.0030?
5. A 12 oz can of soda has a mean volume of 12 oz, with a standard deviation of .25 oz. How common are cans with between 11 and 11.5 oz of soda?
6. $\mu = 0.0025$ and $\sigma = 0.0005$, how common are values between 0.0045 and 0.005?
7. If a population $\mu = 1130$ and $\sigma = 5$, how common are values between 0 and 1100?
8. Assuming population $\mu = 1130$ and $\sigma = 5$, how common are values between 1125 and 1135?
9. The American Robin Redbreast has a mean weight of 77 g, with a standard deviation of 6 g. How common are Robins in the 59 g – 71 g range?
10. Population $\mu = \frac{3}{5}$ and $\sigma = \frac{1}{10}$, how common are values between $\frac{2}{5}$ and 1?
11. Population $\mu = 0.25\%$ and $\sigma = 0.05\%$, how common are values between 0.35% and 0.45%?
12. Population $\mu = 156.5$ and $\sigma = 0.25$, how common are values between 155 and 156?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 9.1.

9.2 The Empirical Rule

Objective

Here you will learn how to use the Empirical Rule to estimate the probability of an event.

Concept

If the price per pound of USDA Choice Beef is normally distributed with a mean of \$4.85/lb and a standard deviation of \$0.35/lb, what is the estimated probability that a randomly chosen sample (from a randomly chosen market) will be between \$5.20 and \$5.55 per pound?

Watch This



MEDIA

Click image to the left or use the URL below.

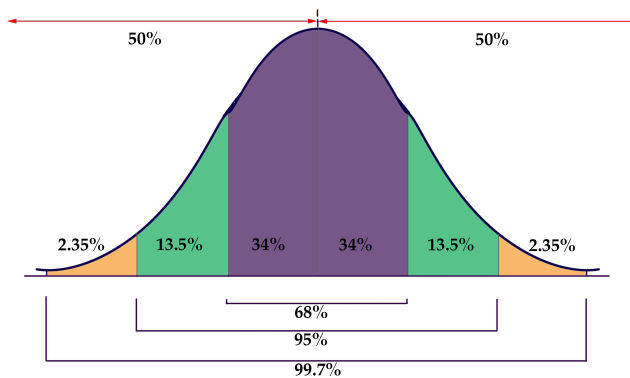
URL: <https://www.ck12.org/flx/render/embeddedobject/67465>

<http://vimeo.com/47111393> CK-12 Foundation: Chapter6EmpiricalRuleA

Guidance

This lesson on the *Empirical Rule* is an extension of the previous lesson “Understanding the Normal Distribution”. In the prior lesson, the goal was to develop an intuition of the interaction between decreased probability and increased distance from the mean. In this lesson, we will practice applying the Empirical Rule to estimate the specific probability of occurrence of a sample based on the range of the sample, measured in standard deviations.

The graphic below is a representation of the Empirical Rule:



The graphic is a rather concise summary of the 'vital statistics' of a Normal Distribution. Note how the graph resembles a bell? Now you know why the normal distribution is also called a “*bell curve*”.

- 50% of the data is above, and 50% below, the mean of the data
- Approximately 68% of the data occurs within 1 SD of the mean
- Approximately 95% occurs within 2 SD's of the mean
- Approximately 99.7% of the data occurs within 3 SD's of the mean

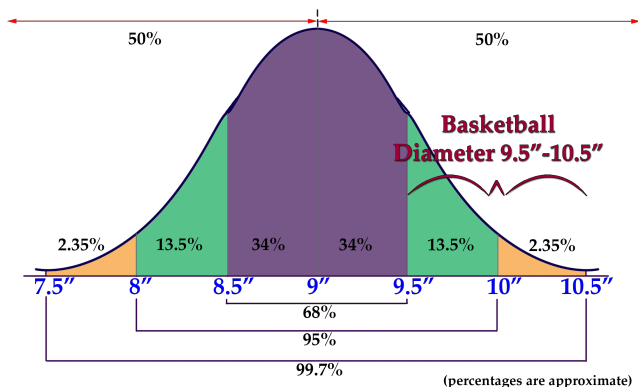
It is due to the probabilities associated with 1, 2, and 3 SD's that the Empirical Rule is also known as the “68 – 95 – 99.7 rule”.

Example A

If the diameter of a basketball is normally distributed, with a mean (μ) of 9", and a standard deviation (σ) of 0.5", what is the probability that a randomly chosen basketball will have a diameter between 9.5" and 10.5"?

Solution:

Since the $\sigma = 0.5''$ and the $\mu = 9''$, we are evaluating the probability that a randomly chosen ball will have a diameter between 1 and 3 standard deviations above the mean. The graphic below shows the portion of the normal distribution included between 1 and 3 SD's:



The percentage of the data spanning the 2nd and 3rd SD's is 13.5% + 2.35% = 15.85%

The probability that a randomly chosen basketball will have a diameter between 9.5 and 10.5 inches is 15.85%.

Example B

If the depth of the snow in my yard is normally distributed, with $\mu = 2.5''$ and $\sigma = .25''$, what is the probability that a randomly chosen location will have a snow depth between 2.25 and 2.75 inches?



Solution:

2.25 inches is $\mu - 1\sigma$, and 2.75 inches is $\mu + 1\sigma$, so the area encompassed approximately represents $34\% + 34\% = 68\%$.

The probability that a randomly chosen location will have a depth between 2.25 and 2.75 inches is 68%.

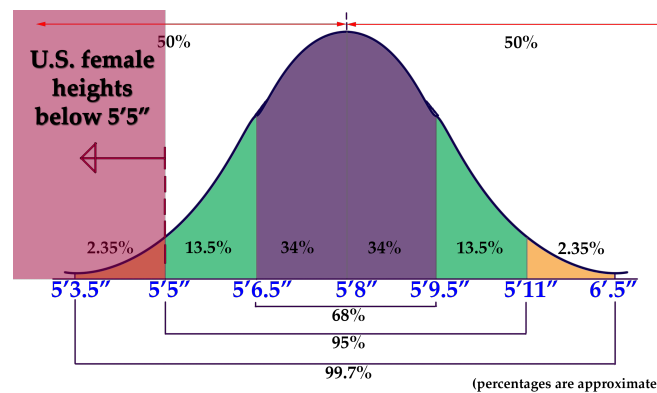
Example C

If the height of women in the U.S. is normally distributed with $\mu = 5'8''$ and $\sigma = 1.5''$, what is the probability that a randomly chosen woman in the U.S. is shorter than 5'5''?

Solution:

This one is slightly different, since we aren't looking for the probability of a limited range of values. We want to evaluate the probability of a value occurring anywhere below 5'5''. Since the domain of a normal distribution is infinite, we can't actually state the probability of the portion of the distribution on 'that end' because it has no 'end'! What we need to do is add up the probabilities that we do know and subtract them from 100% to get the remainder.

Here is that normal distribution graphic again, with the height data inserted:



Recall that a normal distribution always has 50% of the data on each side of the mean. That indicates that 50% of U.S. females are taller than 5'8'', and gives us a solid starting point to calculate from. There is another 34% between 5'6.5'' and 5'8'' and a final 13.5% between 5'5'' and 5'6.5''. Ultimately that totals: $50\% + 34\% + 13.5\% = 87.5\%$.

Since 87.5% of U.S. females are 5'5'' or taller, that leaves 12.5% that are less than 5'5'' tall.

Concept Problem Revisited

If the price per pound of USDA Choice Beef is normally distributed with a mean of \$4.85/lb and a standard deviation of \$0.35/lb, what is the estimated probability that a randomly chosen sample (from a randomly chosen market) will be between \$5.20 and \$5.55 per pound?

\$5.20 is $\mu + 1\sigma$, and \$5.55 is $\mu + 2\sigma$, so the probability of a value occurring in that range is approximately 13.5%.

Vocabulary

A **normal distribution** is a common, but specific, distribution of data with a set of characteristics detailed in the lesson above.

The **Empirical Rule** is a name for the way in which the normal distribution divides data by standard deviations: 68% within 1 SD, 95% within 2 SD's and 99.7 within 3 SD's of the mean.

The "**68-95-99.7 rule**" is another name for the Empirical Rule.

A **bell curve** is the shape of a normal distribution.

Guided Practice

1. A normally distributed data set has $\mu = 10$ and $\sigma = 2.5$, what is the probability of randomly selecting a value greater than 17.5 from the set?
2. A normally distributed data set has $\mu = .05$ and $\sigma = .01$, what is the probability of randomly choosing a value between .05 and .07 from the set?
3. A normally distributed data set has $\mu = 514$ and an unknown standard deviation, what is the probability that a randomly selected value will be less than 514?

Solutions:

1. If $\mu = 10$ and $\sigma = 2.5$, then $17.5 = \mu + 3\sigma$. Since we are looking for all data above that point, we need to subtract the probability that a value will occur *below* that value from 100%:

The probability that a value will be less than 10 is 50%, since 10 is the mean. There is another 34% between 10 and 12.5, another 13.5% between 12.5 and 15, and a final 2.35% between 15 and 17.5.

$100\% - 50\% - 34\% - 13.5\% - 2.35\% = 0.15\%$ **probability of a value greater than 17.5**

2. 0.05 is the mean, and 0.07 is 2 standard deviations above the mean, so the probability of a value in that range is $34\% + 13.5\% = 47.5\%$

3. 514 is the mean, so the probability of a value less than that is 50%.

Practice

For questions 1-15, assume all distributions to be normal or approximately normal, and calculate percentages using the 68 – 95 – 99.7rule.

1. Given mean 63 and standard deviation of 168, find the approximate percentage of the distribution that lies between -105 and 567.
2. Approximately what percent of a normal distribution is between 2 standard deviations and 3 standard deviations from the mean?
3. Given standard deviation of 74 and mean of 124, approximately what percentage of the values are greater than 198?
4. Given $\sigma = 39$ and $\mu = 101$, approximately what percentage of the values are less than 23?
5. Given mean 92 and standard deviation 189, find the approximate percentage of the distribution that lies between -286 and 470.
6. Approximately what percent of a normal distribution lies between $\mu + 1\sigma$ and $\mu + 2\sigma$?
7. Given standard deviation of 113 and mean 81, approximately what percentage of the values are less than -145?
8. Given mean 23 and standard deviation 157, find the approximate percentage of the distribution that lies between 23 and 337.
9. Given $\sigma = 3$ and $\mu = 84$, approximately what percentage of the values are greater than 90?
10. Approximately what percent of a normal distribution is between μ and $\mu + 1\sigma$?
11. Given mean 118 and standard deviation 145, find the approximate percentage of the distribution that lies between -27 and 118.
12. Given standard deviation of 81 and mean 67, approximately what percentage of values are greater than 310?
13. Approximately what percent of a normal distribution is less than 2 standard deviations from the mean?
14. Given $\mu + 1\sigma = 247$ and $\mu + 2\sigma = 428$, find the approximate percentage of the distribution that lies between 66 and 428.
15. Given $\mu - 1\sigma = -131$ and $\mu + 1\sigma = 233$, approximately what percentage of the values are greater than -495?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 9.2.

9.3 Z-Scores

Objective

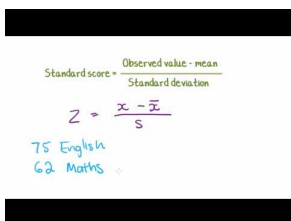
Here you will learn how z-scores can be used to evaluate how extreme a given value is in a particular set or population.

Concept

Using the Empirical Rule can give you a good idea of the probability of occurrence of a value that happens to be exactly one, two or three to either side of the mean, but how do you compare the probabilities of values that are in between standard deviations?

Watch This

The British video below is very clear and easy to follow. It is worth noting, particularly for U.S. students, that the instructor uses the notation \bar{x} rather than μ for mean, and pronounces z as “zed”.



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67459>

<http://youtu.be/2JjaWQZChqsvce> furthermaths - Maths Tutorial: Z scores

Guidance

Z-scores are related to the Empirical Rule from the standpoint of being a method of evaluating how extreme a particular value is in a given set. **You can think of a z-score as the number of standard deviations there are between a given value and the mean of the set.** While the Empirical Rule allows you to associate the first three standard deviations with the percentage of data that each SD includes, the z-score allows you to state (as accurately as you like), just how many SD's a given value is above or below the mean.

Conceptually, the z-score calculation is just what you might expect, given that you are calculating the number of SD's between a value and the mean. You calculate the z-score by first calculating the difference between your value and the mean, and then dividing that amount by the standard deviation of the set. The formula looks like this:

$$z\text{-score} = \frac{(\text{value} - \text{mean})}{\text{standard deviation}} = \frac{(x - \mu)}{\sigma}$$

In this lesson, we will practice calculating the z-score for various values. In the next lesson, we will learn how to associate the z-score of a value with the probability that the value will occur.

Example A

What is the z-score of a value of 27, given a set mean of 24, and a standard deviation of 2?

Solution:

To find the z-score we need to divide the difference between the value, 27, and the mean, 24, by the standard deviation of the set, 2.

$$z \text{ score} = \frac{27 - \mu}{\sigma}$$

$$\frac{27 - 24}{2}$$

$$\frac{3}{2}$$

$$z \text{ score of } 27 = +1.5$$

This indicates that 27 is 1.5 standard deviations above the mean.

Example B

What is the z-score of a value of 104.5, in a set with $\mu = 125$ and $\sigma = 6.2$?

Solution:

Find the difference between the given value and the mean, then divide it by the standard deviation.

$$z \text{ score} = \frac{104.5 - \mu}{\sigma}$$

$$\frac{104.5 - 125}{6.2}$$

$$\frac{-20.5}{6.2}$$

$$z \text{ score of } 104.5 = -3.306$$

Note that the z-score is negative, since the measured value, 104.5, is less than (below) the mean, 125.

Example C

Find the value represented by a z-score of 2.403, given $\mu = 63$ and $\sigma = 4.25$.

**Solution:**

This one requires that we solve for a missing value rather than for a missing z-score, so we just need to fill in our formula with what we know and solve for the missing value:

$$z \text{ score} = \frac{x - \mu}{\sigma}$$

$$2.403 = \frac{x - 63}{4.25}$$

$$10.213 = x - 63$$

$$73.213 = x$$

73.213 has a z-score of 2.403

Concept Problem Revisited

Using the Empirical Rule can give you a good idea of the probability of occurrence of a value that happens to be right on one of the first three standard deviations to either side of the mean, but how do you compare the probabilities of values that are in between standard deviations?

The z-score of a value is the count of the number of standard deviations between the value and the mean of the set. You can find it by subtracting the value from the mean, and dividing the result by the standard deviation.

Vocabulary

The **z-score** of a value is the number of standard deviations between the value and the mean of the set.

Guided Practice



1. What is the z-score of the price of a pair of skis that cost \$247, if the mean ski price is \$279, with a standard deviation of \$16?
2. What is the z-score of a 5-scoop ice cream cone if the mean number of scoops is 3, with a standard deviation of 1 scoop?
3. What is the z-score of the weight of a cow that tips the scales at 825 lbs, if the mean weight for cows of her type is 1150 lbs, with a standard deviation of 77 lbs?
4. What is the z-score of a measured value of 0.0034, given $\mu = 0.0041$ and $\sigma = 0.0008$?

Solutions:

1. First find the difference between the measured value and the mean, then divide that difference by the standard deviation:

$$\frac{247 - 279}{16} = \frac{-32}{16}$$

$$z\text{-score} = -2$$

2. This one is easy: The difference between 5 scoops and 3 scoops is +2, and we divide that by the standard deviation of 1, so the **z-score is +2**.

3. First find the difference between the measured value and the mean, then divide that difference by the standard deviation:

$$\frac{825 \text{ lbs} - 1150 \text{ lbs}}{771 \text{ lbs}} = \frac{-325}{77}$$

$$z\text{-score} = -4.2207$$

4. First find the difference between the measured value and the mean, then divide that difference by the standard deviation:

$$\frac{0.0034 - 0.0041}{0.0008} = \frac{-0.0007}{0.0008}$$

$$z\text{-score} = -0.875$$

Practice

- Given a distribution with a mean of 70 and standard deviation of 62, find a value with a z-score of -1.82.
- What does a z-score of 3.4 mean?
- Given a distribution with a mean of 60 and standard deviation of 98, find the z-score of 120.76.
- Given a distribution with a mean of 60 and standard deviation of 21, find a value with a z-score of 2.19.
- Find the z-score of 187.37, given a distribution with a mean of 185 and standard deviation of 1.
- What does a z-score of -3.8 mean?
- Find the z-score of 125.18, given a distribution with a mean of 101 and standard deviation of 62.
- Given a distribution with a mean of 117 and standard deviation of 42, find a value with a z-score of -0.94.
- Given a distribution with a mean of 126 and standard deviation of 100, find a value with a z-score of -0.75.
- Find the z-score of 264.16, given $\mu = 188$ and $\sigma = 64$.
- Find a value with a z-score of -0.2, given $\mu = 145$ and $\sigma = 56$.
- Find the z-score of 89.79 given $\mu = 10$ and $\sigma = 79$.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 9.3.

9.4 Z Scores II

Objective

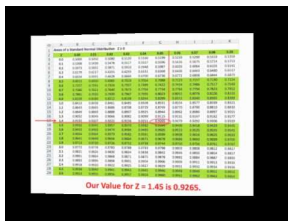
Here you will learn to evaluate z -scores as they relate to probability.

Concept

Knowing the z -score of a given value is great, but what can you do with it? How does a z -score relate to probability? What is the probability of occurrence of a z -score less than $+2.47$?

Watch This

The video below provides a demonstration of how to use a z -score probability reference table, as we do in this lesson. The table he uses in the video is slightly different, but the concept is the same.



MEDIA

Click image to the left or use the URL below.

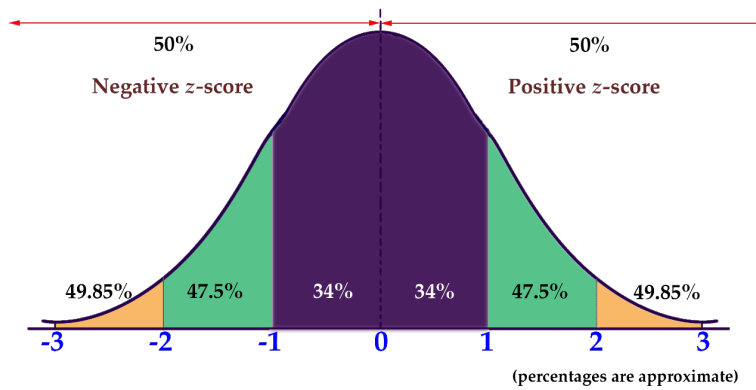
URL: <https://www.ck12.org/flx/render/embeddedobject/67464>

<http://youtu.be/rEmNUkKSpbU> TAMUC - Dr. Dawg

Guidance

Since z -scores are a measure of the number of SD's between a value and the mean, they can be used to calculate probability by comparing the location of the z -score to the area under a normal curve either to the left or right. The area can be calculated using calculus, but we will just use a table to look up the area.

I believe that the concept of comparing z -scores to probability is most easily understood with a graphic like the one we used in the lesson on the Empirical Rule, so I included one below. Be sure to review the Examples to see how the scores work.



Like the graphic we viewed in the Empirical Rule lesson, this one only provides probability percentages for integer values of z -scores (standard deviations). In order to find the values for z -scores that aren't integers, you can use a table like the one below. To find the value associated with a given z -score, you find the first decimal of your z -score on the left or right side and then the 2nd decimal of your z -score across the top or bottom of the table. Where they intersect you will find the decimal expression of the percentage of values that are less than your sample (see Ex. A).

TABLE 9.1:

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
0.0	0.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359	0.0
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.1
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.2
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517	0.3
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879	0.4
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224	0.5
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.6
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	0.7
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.8
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389	0.9
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	1.0
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883	1.1
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015	1.2
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	1.3
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	1.4
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	1.5
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	1.6
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	1.7
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	1.8
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767	1.9
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	2.0
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857	2.1
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989	2.2
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	2.3
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	2.4
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	2.5
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964	2.6

TABLE 9.1: (continued)

2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974	2.7
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981	2.8
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	2.9
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999	3.0
3.1	0.999	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993	3.1
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	3.2
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	3.3
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	3.4
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	3.5
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.6
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.7
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.8
3.9	1	1	1	1	1	1	1	1	1	1	3.9
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z

Z-score tables like the one above describe the probability that a given value, or any value less than it, will occur in a given set. This particular table assumes you are looking to find the probability associated with a *positive* z-score. You may have additional work to do if the z-score is negative.

- To find the percentage of values **greater** than a **negative** Z score, just look up the matching positive Z score value.
- To find the percentage of values **less** than a **negative** z-score, subtract the chart value from 1.
- To find the percentage of values **greater** than a **positive** z-score, subtract the chart value from 1.

Example A

What is the probability that a value with a z-score less than 2.47 will occur in a normal distribution?

Solution:

Scroll up to the table above and find “2.4” on the left or right side. Now move across the table to “0.07” on the top or bottom, and record the value in the cell: **0.9932**. That tells us that **99.32% of values in the set are at or below a z-score of 2.47**.

Example B

What is the probability that a value with a z-score greater than 1.53 will occur in a normal distribution?

Solution:

Scroll up to the table of z-score probabilities again and find the intersection between 1.5 on the left or right and 3 on the top or bottom, record the value in the cell: **0.937**.

That decimal lets us know that 93.7% of values in the set are *below* the z-score of 1.53. To find the percentage that is *above* that value, we subtract 0.937 from 1.0 (or 93.7% from 100%), to get **0.063 or 6.3%**.

Example C

What is the probability of a random selection being less than 3.65, given a normal distribution with $\mu = 5$ and $\sigma = 2.2$?

Solution:

This question requires us to first find the z-score for the value 3.65, then calculate the percentage of values below that z-score from a reference.

1. Find the z-score for 3.65, using the z-score formula: $\frac{(x-\mu)}{\sigma}$

$$\frac{3.65 - 5}{2.2} = \frac{-1.35}{2.2} \approx -0.61$$

2. Now we can scroll up to our z -score reference above and find the intersection of 0.6 and 0.01, which should be **.7291**.

3. Since this is a negative z -score, and we want the percentage of values *below* it, we subtract that decimal from 1.0 (reference the three steps highlighted by bullet points below the chart if you didn't recall this), to get $1 - .7291 = .2709$

There is approximately a 27.09% probability that a value less than 3.65 would occur from a random selection of a normal distribution with mean 5 and standard deviation 2.2.

Concept Problem Revisited

Knowing the z -score of a given value is great, but what can you do with it? How does a z -score relate to probability? What is the probability of occurrence of a z -score less than 2.47?

A z -score lets you calculate the probability that a randomly selected value will be greater or less than a particular value in a set.

To find the probability of a z -score below +2.47, using a reference such as the table in the lesson above:

1. Find 2.4 on the left or right side
2. Move across to 0.07 on the top or bottom.
3. The cell you arrive at says: 0.9932, which means that apx 99.32% of the values in a normal distribution will occur below a z -score of 2.47.

Vocabulary

A **z -score table** associates the various common z -scores between 0 and 3.99 with the decimal probability of being less than or equal to that z -score.

Guided Practice

1. What is the probability of occurrence of a value with z -score greater than 1.24?
2. What is the probability of $z < -.23$?
3. What is $P(Z < 2.13)$?

Solutions:

1. Since this is a positive z -score, we can use the value for $z = 1.24$ directly from the table, and just express it as a percentage: **0.8925 or 89.25%**
2. This is a negative z -score, and we want the percentage of values *greater* than it, so we need to subtract the value for $z = +0.23$ from 1: $1 - 0.591 = .409$ **or 40.9%**
3. This is a positive z -score, and we need the percentage of values below it, so we can use the percentage associated with $z = +2.13$ directly from the table: **0.9834 or 98.34%**

Practice

Find the probabilities, use the table from the lesson above.

1. What is the probability of a z -score less than $+2.02$?
2. What is the probability of a z -score greater than $+2.02$?
3. What is the probability of a z -score less than -1.97 ?
4. What is the probability of a z -score greater than -1.97 ?
5. What is the probability of a z -score less than $+0.09$?
6. What is the probability of a z -score less than -0.02 ?
7. What is $P(Z < 1.71)$?
8. What is $P(Z > 2.22)$?
9. What is $P(Z < -1.19)$?
10. What is $P(Z > -2.71)$?
11. What is $P(Z < 3.71)$?
12. What is the probability of the random occurrence of a value greater than 56 from a normally distributed population with mean 62 and standard deviation 4.5?
13. What is the probability of a value of 329 or greater, assuming a normally distributed set with mean 290 and standard deviation 32?
14. What is the probability of getting a value below 1.2 from the random output of a normally distributed set with $\mu = 2.6$ and $\sigma = .9$?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 9.4.

9.5 Z-scores III

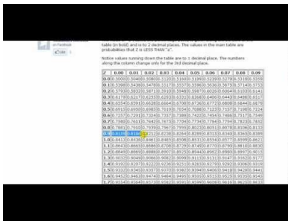
Objective

Here you will learn to calculate the probability of a z -score between two others.

Concept

Do z -score probabilities always need to be calculated as the chance of a value either above or below a given score? How would you calculate the probability of a z -score between -0.08 and $+1.92$?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67469>

<http://youtu.be/tnVr1Qbk4YU?t=9m23sAH> Metrics - Reading probabilities from the Z table

Guidance

To calculate the probability of getting a value with a z -score between two other z -scores, you can either use a reference table to look up the value for both scores and subtract them to find the difference, or you can use technology. In this lesson, which is an extension of Z-scores and Z-scores II, we will practice both methods.

Historically, it has been very common to use a z -score probability table like the one below to look up the probability associated with a given z -score:

TABLE 9.2:

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
0.0	.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359	0.0
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.1
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.2
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517	0.3
0.4	.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879	0.4
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224	0.5
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.6
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	0.7
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.8
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389	0.9

TABLE 9.2: (continued)

1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	1.0
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883	1.1
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015	1.2
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	1.3
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	1.4
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	1.5
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	1.6
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	1.7
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	1.8
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767	1.9
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	2.0
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857	2.1
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989	2.2
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	2.3
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	2.4
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	2.5
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964	2.6
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974	2.7
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981	2.8
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	2.9
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999	3.0
3.1	0.999	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993	3.1
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	3.2
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	3.3
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	3.4
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	3.5
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.6
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.7
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.8
3.9	1	1	1	1	1	1	1	1	1	1	3.9
Z	0.000	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z

Since the proliferation of the Internet, however, you can also use a free online calculator such as one of these three:

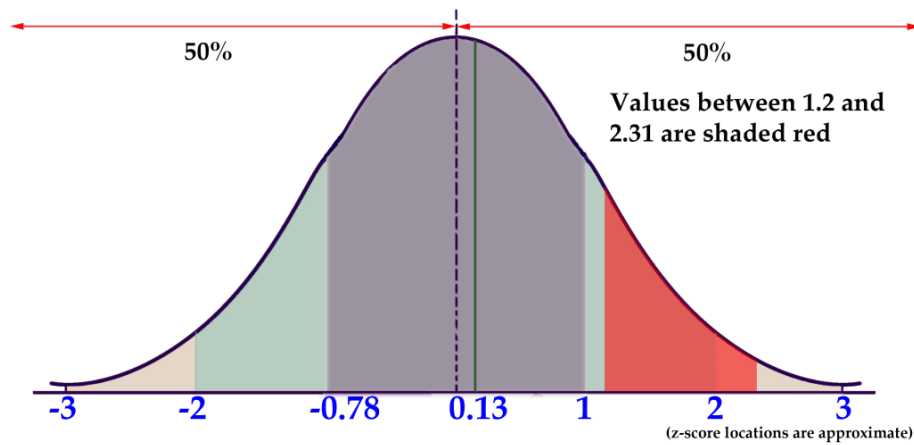
- <http://www.fourmilab.ch/rpkp/experiments/analysis/zCalc.html>
- <http://easycalculation.com/statistics/p-value-for-z-score.php>
- <http://www.mathportal.org/calculators/statistics-calculator/z-score-calculator.php>

Example A

What is the probability associated with a z -score between 1.2 and 2.31?

Solution:

To evaluate the probability of a value occurring within a given range, you need to find the probability of both the upper and lower values in the range, and subtract to find the difference.



- First find $z = 1.2$ on the z -score probability reference above: **.8849** Remember that value represents the percentage of values *below* 1.2.
- Next, find and record the value associated with $z = 2.31$: **.9896**
- Since approximately 88.49% of all values are below $z = 1.2$ and approximately 98.96% of all values are below $z = 2.31$, there are $98.96\% - 88.49\% = 10.47\%$ of values **between**.

Example B

What is the probability that a value with a z -score between -1.32 and $+1.49$ will occur in a normal distribution?

Solution:

Let's use the online calculator at mathportal.org for this one.

When you open the page, you should see a window like this:

MATHPORTAL.ORG

Z - Score Calculator

Enter cutoff points in order to find the area under normal curve. Important: The form will NOT let you enter wrong characters (like y, p, ;, ...)

Use the standard normal distribution to find one of the following probabilities:

$P(\text{ } < Z < \text{ })$

$P(\text{ } < Z)$

$P(\text{ } > Z)$

All you need to do is select the radio button to the left of the first type of probability, input “ -1.32 ” into the first box, and 1.49 into the second. When you click “Compute”, you should get the result

$$P(-1.32 < Z < 1.49) = 0.8385$$

Which tells us that there is approximately and 83.85% probability that a value with a z -score between 1.32 and 1.49 will occur in a normal distribution.

Notice that the calculator also details the steps involved with finding the answer:

1. Estimate the probability using a graph, so you have an idea of what your answer should be.
2. Find the probability of $z < 1.49$, using a reference. (0.9319)
3. Find the probability of $z < -1.32$, again, using a reference. (0.0934)
4. Subtract the values: $0.9319 - 0.0934 = 0.8385$ or 83.85%

Example C

What is the probability that a random selection will be between 8.45 and 10.25, if it is from a normal distribution with $\mu = 10$ and $\sigma = 2$?

Solution:

This question requires us to first find the z -scores for the value 8.45 and 10.25, then calculate the percentage of value between them by using values from a z -score reference and finding the difference.

1. Find the z -score for 8.45, using the z -score formula: $\frac{(x-\mu)}{\sigma}$

$$\frac{8.45 - 10}{2} = \frac{-1.55}{2} \approx -0.78$$

2. Find the z -score for 10.25 the same way:

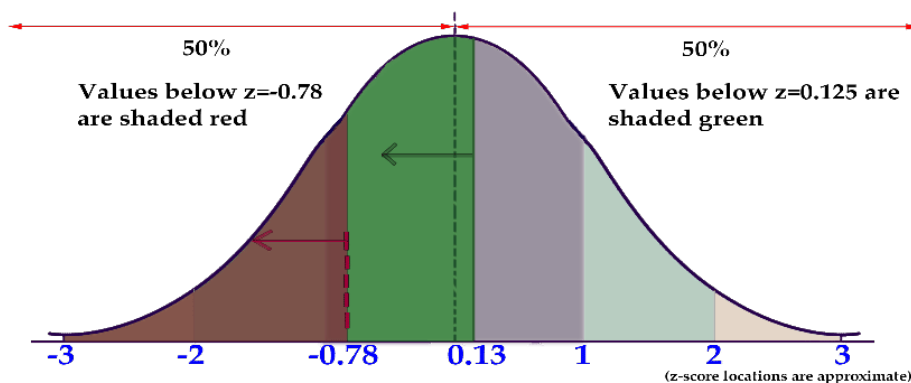
$$\frac{10.25 - 10}{2} = \frac{0.25}{2} \approx .13$$

3. Now find the percentages for each, using a reference (don't forget we want the probability of values *less* than our negative score and *less* than our positive score, so we can find the values between):

$$P(Z < -0.78) = .2177 \text{ or } 21.77\%$$

$$P(Z < .13) = .5517 \text{ or } 55.17\%$$

4. At this point, let's sketch the graph to get an idea what we are looking for:



Because the values we are interested in are between the two z -scores, we subtract the red area from the green area to get only the green "stripe" between -0.78 and 0.13.

5. Finally, subtract the values to find the difference:

$$.5517 - .2177 = .3340 \text{ or about } 33.4\%$$

There is approximately a 33.4% probability that a value between 8.45 and 10.25 would result from a random selection of a normal distribution with mean 10 and standard deviation 2.

Concept Problem Revisited

Do z-score probabilities always need to be calculated as the chance of a value either above or below a given score? How would you calculate the probability of a z-score between -0.08 and +1.92?

After this lesson, you should know without question that z-score probabilities do not need to assume only probabilities above or below a given value, the probability between values can also be calculated.

The probability of a z-score below -0.08 is 46.81%, and the probability of a z-score below 1.92 is 97.26%, so the probability between them is $97.26\% - 46.81\% = 50.45\%$.

Vocabulary

A **z-score** is a measure of how many standard deviations there are between a data value and the mean.

A **z-score probability table** is a table that associates z-scores to area under the normal curve. The table may be used to associate a Z-score with a percent probability.

Guided Practice

1. What is the probability of a z-score between -0.93 and 2.11?
2. What is $P(1.39 < Z < 2.03)$?
3. What is $P(-2.11 < Z < 2.11)$?

Solutions:

1. Using the z-score probability table above, we can see that the probability of a value below -0.93 is .1762, and the probability of a value below 2.11 is .9826. Therefore, the probability of a value between them is $.9826 - .1762 = .8064$ or 80.64%
2. Using the z-score probability table, we see that the probability of a value below $z = 1.39$ is .9177, and a value below $z = 2.03$ is .9788. That means that the probability of a value between them is $.9788 - .9177 = .0611$ or 6.11%
3. Using the online calculator at mathportal.org, we select the top calculation with the associated radio button to the left of it, enter “-2.11” in the first box, and “2.11” in the second box. Click “Compute” to get “.9652”, and convert to a percentage. The probability of a z-score between -2.11 and +2.11 is about **96.52%**.

Practice

Find the probabilities, use the table from the lesson or an online resource.

1. What is the probability of a z-score between +1.99 and +2.02?
2. What is the probability of a z-score between -1.99 and +2.02?

3. What is the probability of a z -score between -1.20 and -1.97 ?
4. What is the probability of a z -score between $+2.33$ and -0.97 ?
5. What is the probability of a z -score greater than $+0.09$?
6. What is the probability of a z -score greater than -0.02 ?
7. What is $P(1.42 < Z < 2.01)$?
8. What is $P(1.77 < Z < 2.22)$?
9. What is $P(-2.33 < Z < -1.19)$?
10. What is $P(-3.01 < Z < -0.71)$?
11. What is $P(2.66 < Z < 3.71)$?
12. What is the probability of the random occurrence of a value between 56 and 61 from a normally distributed population with mean 62 and standard deviation 4.5?
13. What is the probability of a value between 301 and 329, assuming a normally distributed set with mean 290 and standard deviation 32?
14. What is the probability of getting a value between 1.2 and 2.3 from the random output of a normally distributed set with $\mu = 2.6$ and $\sigma = .9$?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 9.5.

9.6 The Mean of Means

Objective

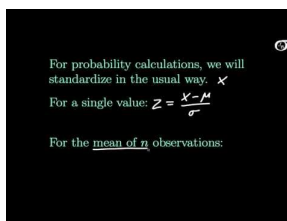
Here you will learn how to calculate the mean of means, which is the mean value of several sample means.

Concept

Suppose you have taken several samples of 10 units each from a population of 500 students, and calculated the mean of each sample. How might you use the data you now have to estimate a mean for the entire population?



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67531>

<http://youtu.be/0zqNGDVNKgAJB> Statistics - The Sampling Distribution of the Sample Mean

Guidance

In statistics, you often need to take data from a small number of samples and use it to extrapolate an estimate of the parameters of the population the samples were pulled from. Since one of the more common parameters of interest is the mean, it is common to see a distribution of the means of a number of samples (I realize this may be confusing, “sample” here actually refers to the results of several individual samples) from the same population. This

distribution is called, appropriately, the “*sampling distribution of the sample mean*”. We will be investigating the sampling distribution of the sample mean in more detail in the next lesson “*The Central Limit Theorem*”, but in essence it is simply a representation of the spread of the means of several samples.

Here we will be focusing on a single value in that sampling distribution, the “*mean of means*”. The mean of means is simply the mean of all of the means of several samples. By calculating the mean of the sample means, you have a single value that can help summarize a lot of data.

The mean of means, notated here as $\mu_{\bar{x}}$, is actually a pretty straightforward calculation. Simply sum the means of all your samples and divide by the number of means.

As a formula, this looks like:

$$\mu_{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 \dots + \bar{x}_n}{n}$$

The second common parameter used to define sampling distribution of the sample means is the “*standard deviation of the distribution of the sample means*”. The only significant difference between the standard deviation of a population and the standard deviation of sample means is that you need to divide the population standard deviation by the square root of the sample size.

As a formula, this looks like:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

I recognize that the terminology in this lesson may be getting a bit scary, but the actual concept and the required calculations are actually not particularly difficult. **Work your way through the examples below**, and I think you will find that the hardest part of this lesson is getting past the wording!

Example A

Given the following sample means, what is the mean of means?

$$\bar{x}_1 = 4.35, \bar{x}_2 = 4.62, \bar{x}_3 = 4.29, \bar{x}_4 = 4.39, \bar{x}_5 = 4.55$$

Solution:

To calculate the mean of means, sum the sample means and divide by the number of samples:

$$\begin{aligned} \mu_{\bar{x}} &= \frac{4.35 + 4.62 + 4.29 + 4.39 + 4.55}{5} \\ &= \frac{22.20}{5} \\ \mu_{\bar{x}} &= 4.44 \end{aligned}$$

Example B

Brian works at a pizza restaurant, and has been carefully monitoring the weight of cheese he puts on each pizza for the past week. Each day, Brian tracks the weight of the cheese on each pizza he makes, and calculates the mean weight of cheese on each pizza for that day. If the weights below represent the mean weights for each day, what is the mean of means weight of cheese over the past week? If Brian makes 25 pizzas per day and knows the standard deviation of cheese weight per pizza is 0.5 oz, what is the standard deviation of the sample distribution of the sample means?



TABLE 9.3:

DAY	WEIGHT (OZ)
Monday	7.84
Tuesday	7.93
Wednesday	7.79
Thursday	8.03
Friday	8.14
Saturday	8.09
Sunday	7.88

Solution:

First calculate the mean of means by summing the mean from each day and dividing by the number of days:

$$\begin{aligned}\mu_{\bar{x}} &= \frac{7.84 + 7.93 + 7.79 + 8.03 + 8.14 + 8.09 + 7.88}{7} \\ &= \frac{55.7}{7} \\ \mu_{\bar{x}} &= 7.96\end{aligned}$$

Then use the formula to find the standard deviation of the sampling distribution of the sample means:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where σ is the standard deviation of the population, and n is the number of data points in each sampling.

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{.05 \text{ oz}}{\sqrt{25}} \\ &= \frac{.05}{5} \\ \sigma_{\bar{x}} &= .01\end{aligned}$$

Brian's research indicates that the cheese he uses per pizza has a mean weight of 7.96 oz, with a standard deviation of .01 oz.

Example C

Calculate $\mu_{\bar{x}}$, given the following:

$$\bar{x}_1 = 352.7$$

$$\bar{x}_2 = 351.9$$

$$\bar{x}_3 = 349.97$$

$$\bar{x}_4 = 352.33$$

$$\bar{x}_5 = 353.1$$

$$\bar{x}_6 = 349.63$$

Solution:

The $\mu_{\bar{x}}$ (mean of means) of the given data is:

$$\begin{aligned}\mu_{\bar{x}} &= \frac{352.7 + 351.9 + 349.97 + 352.33 + 353.1 + 349.63}{6} \\ &= \frac{2109.63}{6} \\ \mu_{\bar{x}} &= 351.61\end{aligned}$$

Concept Problem Revisited

Suppose you have taken several samples of 10 units each from a population of 500 students, and calculated the mean of each sample. How might you use the data you now have to estimate a mean for the entire population?

You could sum the means of each 10-unit sampling, and divide by the number of samples to get the mean of the means. You could further divide the standard deviation of the entire 500 students (if known) by $\sqrt{10}$ (since each sampling contained 10 data points), to find the standard deviation of the distribution of the sample mean.

Vocabulary

The **sampling distribution of the sample mean** is the distribution that describes the spread of the means of multiple samples from the same population.

The **mean of means** is the overall mean value of the means of several samples from the same population.

The **standard deviation of the distribution of the sample means** is a measure of the spread of the random variable \bar{x} .

Guided Practice

1. Calculate $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$, given the following data:

$$\bar{x}_1 = 251.6$$

$$\bar{x}_2 = 242.8$$

$$\bar{x}_3 = 248.79$$

$$\bar{x}_4 = 245.33$$

$$\bar{x}_5 = 253.21$$

$$\bar{x}_6 = 256.31$$

Sample size = 9, $\sigma = 5.8$

2. If the σ of a population is 2.94, and 25 samples of 12 samples each are taken, what is $\sigma_{\bar{x}}$?

3. Given the population: {1, 2, 3, 4, 5}, create a sampling distribution by finding the mean of all possible samples that include four units. How does $\mu_{\bar{x}}$ compare to μ ?

Solutions:

1. First calculate $\mu_{\bar{x}}$, using $\mu_{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 \dots + \bar{x}_n}{n}$.

$$\begin{aligned} & \frac{251.6 + 242.8 + 248.79 + 245.33 + 253.21 + 256.31}{6} \\ & \frac{1498.04}{6} \\ & \mu_{\bar{x}} = 249.67 \end{aligned}$$

Next calculate $\sigma_{\bar{x}}$, using $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$:

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{5.8}{\sqrt{9}} \\ &= \frac{5.8}{3} \\ \sigma_{\bar{x}} &= 1.93\bar{3} \end{aligned}$$

2. To calculate $\sigma_{\bar{x}}$, use $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$:

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{2.94}{\sqrt{12}} \\ &= \frac{2.94}{4} \\ \sigma_{\bar{x}} &= .735 \end{aligned}$$

3. This one requires a few steps, first we need to find the mean of each possible sample of four units:

\bar{x}_1 (1234) has a mean of 2.5

\bar{x}_2 (1235) has a mean of 2.75

\bar{x}_3 (1245) has a mean of 3

\bar{x}_4 (1345) has a mean of 3.25

\bar{x}_5 (2345) has a mean of 3.5

Next we calculate the mean of means, $\mu_{\bar{x}}$, using $\mu_{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 \dots + \bar{x}_n}{n}$:

$$\begin{aligned} \mu_{\bar{x}} &= \frac{2.5 + 2.75 + 3 + 3.25 + 3.5}{5} \\ &= \frac{15}{5} \\ \mu_{\bar{x}} &= 3 \end{aligned}$$

Now we need to calculate μ , using all the population data:

$$\begin{aligned}\mu &= \frac{1+2+3+4+5}{5} \\ &= \frac{15}{5} \\ \mu &= 3\end{aligned}$$

With the given data, $\mu_{\bar{x}} = \mu$

Practice

- Find $\mu_{\bar{x}}$, given $\bar{x}_1 = 21.0$, $\bar{x}_2 = 24.3$, $\bar{x}_3 = 25.0$, $\bar{x}_4 = 20.6$, $\bar{x}_5 = 22.3$, and $\bar{x}_6 = 22.3$
- Find $\mu_{\bar{x}}$, given $\bar{x}_1 = 15.1$, $\bar{x}_2 = 15.77$, $\bar{x}_3 = 15.55$, $\bar{x}_4 = 15.99$, $\bar{x}_5 = 15.42$, and $\bar{x}_6 = 15.37$
- Find $\mu_{\bar{x}}$, given $\bar{x}_1 = 341.52$, $\bar{x}_2 = 345.16$, $\bar{x}_3 = 343.66$, $\bar{x}_4 = 345.86$, and $\bar{x}_5 = 336.10$
- Find $\mu_{\bar{x}}$, given $\bar{x}_1 = 1.41$, $\bar{x}_2 = 0.59$, $\bar{x}_3 = 1.44$, $\bar{x}_4 = 0.93$, $\bar{x}_5 = 1.44$, $\bar{x}_6 = 1.01$, and $\bar{x}_7 = 0.74$
- Find $\mu_{\bar{x}}$, given $\bar{x}_1 = 218.19$, $\bar{x}_2 = 279.70$, $\bar{x}_3 = 262.86$, and $\bar{x}_4 = 243.88$
- If the σ of a population is 292.66, and samples of 17 units each are taken, what is $\sigma_{\bar{x}}$?
- If the σ of a population is 41.39, and 23 samples of 30 samples each are taken, what is $\sigma_{\bar{x}}$?
- If the σ of a population is 193.61, and samples of 19 units each are taken, what is $\sigma_{\bar{x}}$?
- If the σ of a population is 91.85, and 129 samples of 11 samples each are taken, what is $\sigma_{\bar{x}}$?
- If the σ of a population is 255.19, and 43 samples of 31 samples each are taken, what is $\sigma_{\bar{x}}$?
- Given the population: {1, 2, 3, 4, 5, 6}, create a sampling distribution by finding the mean of all possible samples that include two units. How does $\mu_{\bar{x}}$ compare to μ ?
- Given the population: {1, 2, 3, 4}, create a sampling distribution by finding the mean of all possible samples that include two units. What is $\mu_{\bar{x}}$?
- Given the population: {1, 2, 3, 4, 5}, create a sampling distribution by finding the mean of all possible samples that include two units. How does $\mu_{\bar{x}}$ compare to μ ?
- Given the population: {1, 2, 3, 4, 5}, create a sampling distribution by finding the mean of all possible samples that include three units. What is $\mu_{\bar{x}}$?
- Given the population: {1, 2, 3, 4}, create a sampling distribution by finding the mean of all possible samples that include three units. What is $\mu_{\bar{x}}$?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 9.6.

9.7 Central Limit Theorem

Objective

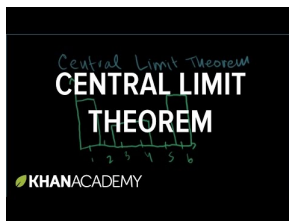
Here you will learn about one of the more remarkable theorems in all of mathematics, the Central Limit Theorem.

Concept

What is the Central Limit Theorem? How does the Central Limit Theorem relate other distributions to the normal distribution?

This lesson describes the relationship between the normal distribution and the Central Limit Theorem.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67467>

<http://youtu.be/JNm3M9cqWyc> Khan Academy - Central Limit Theorem

Guidance

The *Central Limit Theorem* is a very powerful statement in statistics, saying that as you take more and more samples from a random variable, the distribution of the means of the samples (If you completed the lesson titled “The Mean of Means”, you will recognize this as “the sampling distribution of the sample means”) will approximate a normal distribution. This is true *regardless of the original distribution of the random variable* (if the number of data points in each sample is 30 or more)! In fact, as demonstrated in the video above, even a discrete random variable with a pretty odd distribution will output an approximately normal distribution from the means of enough samples.

Formally, the CLT says:

If samples of size n are drawn at random from any population with a finite mean and standard deviation, then the sampling distribution of the sample means, \bar{x} , approximates a normal distribution as n increases.

In “normal English”:

If you collect many samples from an ordinary random variable, and calculate the mean of each sample, then the means will be distributed in an approximate bell-curve, and the “mean of means” will be the same as the mean of the population. The larger the size of the samples you collect, the more closely the distribution of their means will approximate a normal distribution.

Notes to remember:

- As long as your sample size is **30 or greater**, you may assume the distribution of the sample means to be approximately normal, meaning that you can calculate the probability that the mean of a single sample of size 30 or greater will occur by using the z -score of the mean.
- The mean of the distribution created from many sample means approaches the mean of the population. Formally: $\mu_{\bar{x}} = \mu$
- The standard deviation of the distribution of the means is estimated by dividing the standard deviation of the population by the square root of the sample size. Formally: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Use the notation \bar{x} (x -bar) rather than the random variable x to indicate that the random variable you are describing is a sample mean.

You may use the z -score percentage reference table below as needed:

TABLE 9.4:

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
0.0	.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359	0.0
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.1
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.2
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517	0.3
0.4	.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879	0.4
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224	0.5
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.6
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	0.7
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.8
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389	0.9
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	1.0
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883	1.1
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015	1.2
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	1.3
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	1.4
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	1.5
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	1.6
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	1.7
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	1.8
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767	1.9
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	2.0
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857	2.1
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989	2.2
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	2.3
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	2.4
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	2.5
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964	2.6
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974	2.7
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981	2.8
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	2.9
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999	3.0
3.1	0.999	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993	3.1
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	3.2
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	3.3
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	3.4

TABLE 9.4: (continued)

3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	3.5
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.6
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.7
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.8
3.9	1	1	1	1	1	1	1	1	1	1	3.9
Z	0.000	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z

Example A

Mack asked 42 fellow high-school students how much they spent for lunch, on average. According to his research online, the amount spent for lunch by high school students nation wide has $\mu = \$15$, with $\sigma = \$9$. What is the probability that Mack's random sample will result within \$0.01 of the national average?

**Solution:**

There are a few important facts to note here:

- Mack's sample is 42 students, since $42 \geq 30$, he can safely assume that the distribution of his *sample* is approximately normal, according to the Central Limit Theorem.
- The range we are considering is \$14.99 to \$15.01, since that represents \$0.01 above and below the mean.
- The mean of the sample should approximate the mean of the population, in other words $\mu_{\bar{x}} = \mu$
- The standard deviation of Mack's sample, $\sigma_{\bar{x}}$, can be calculated as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where $n = 42$

Let's start by finding the standard deviation of the sample, $\sigma_{\bar{x}}$:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{9}{\sqrt{42}} \\ &= \frac{9}{6.48} \\ \sigma_{\bar{x}} &= 1.389\end{aligned}$$

Since Mack's sample of 42 samples can be assumed to be normally distributed, and since we now know the standard deviation of the sample, 1.39, we can calculate the z-scores of the range using $Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$:

$$Z_1 = \frac{15.01 - 15.00}{1.389} = +0.01$$

$$Z_2 = \frac{14.99 - 15.00}{1.389} = -0.01$$

Finally, we look up Z_1 and Z_2 on the Z-score probability table to get a range of 50.4% to 49.6% = **0.80%**

The probability that Mack's sample will have a mean within \$0.01 of the population mean of \$15.00 is a little less than 1%.

Example B

The time it takes a student to complete the mid-term for Algebra II is a bi-modal distribution with $\mu = 1$ hr and $\sigma = 1$ hr. During the month of June, Professor Spence administers the test 64 times. What is the probability that the average mid-term completion time for students during the month of June exceeds 48 minutes?

Solution:

Important facts:

- There are more than 30 samples, so the Central Limit Theorem applies
- The mean of the sample should approximate the mean of the population, in other words $\mu_{\bar{x}} = \mu$
- The standard deviation of Professor Spence's sample, $\sigma_{\bar{x}}$, can be calculated as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where $n = 64$ (the number of tests/samples)
- 48 minutes is the same as $\frac{48}{60} = 0.8$ hrs, so the range we are interested in is $x > 0.8$ hrs

First calculate the standard deviation of the sample, using $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$:

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{64}}$$

$$\sigma_{\bar{x}} = 0.125$$

Since the sample is normally distributed, according to the CLT, we can use the standard deviation of the sample to calculate the z-score of the minimum value in the relevant range, 0.80 hrs:

$$Z = \frac{0.80 - 1}{0.125} = -1.60$$

Finally, we use the z-score probability reference above to correlate the z-score of -1.60 to the probability of a value greater than that

$$P(Z \geq -1.6) = .9452 \text{ or } 94.52\%$$

Example C

Evan price-checked 123 online auction sellers to record their average asking price for his favorite game. According to a major nation price-checking site, the national average online auction cost for the game is \$35.00 with a standard deviation of \$3.00. Evan found the prices less than \$34.86 on average. How likely is this result?

Solution:

Since there are more than 30 samples ($123 > 30$), we can apply the CLT theorem and treat the sample as a normal distribution.

The standard deviation of the sample is: $\sigma_{\bar{x}} = \frac{3}{\sqrt{123}} = \frac{3}{11.09} = .27$

The z -score for Evan's price point of \$34.86 is:

$$Z = \frac{34.86 - 35}{.27} = \frac{-.14}{.27} = -0.518$$

Consulting the z -score probability table, we learn that the area under the normal curve less than 0.52 is .3015 or 30.15%

The likelihood of 123 samples having a mean of \$34.86 is approximately 30.15%

Concept Problem Revisited

What is the Central Limit Theorem? How does the Central Limit Theorem relate other distributions to the normal distribution?

The Central Limit Theorem says that the larger the sample size, the more the mean of multiple samples will represent a normal distribution. Since that is true regardless of the original distribution, the CLT can be used to effect a bridge between other types of distributions and a normal distribution.

Vocabulary

The **Central Limit Theorem** states that if samples are drawn at random from any population with a finite mean and standard deviation, then the sampling distribution of the sample means approximates a normal distribution as the sample size increases beyond 30.

The **sampling distribution of the sample means** is a distribution of the means of multiple samples. It is commonly assumed to be a normal distribution, though technically it is normal only if the sample size is greater than 30.

Guided Practice



1. The time it takes to drive from Cheyenne WY to Denver CO has a μ of 1 *hr* and σ of 15 *mins*. Over the course of a month, a highway patrolman makes the trip 55 times. What is the probability that his average travel time exceeds 60 minutes?
2. Abbi polls 95 high school students for their G.P.A.. According to the school, the average G.P.A. of high school students has a mean of 3.0, and standard deviation of .5. What is the probability that Abbi's random sample will have a mean within 0.01 of the population?

3. A recipe website has calculated that the time it takes to cook Sunday dinner has a μ of 1 hr with σ of 25 mins. Over the course of a month, 172 users report their time spent cooking Saturday dinner, what is the probability that the average user reports spending less than 45 mins cooking dinner?

Solutions:

1. The sample mean, $\mu_{\bar{x}}$ is the same as the population mean: 1 hr = 60 mins.

The sample standard deviation is $\frac{15 \text{ mins}}{\sqrt{55}} = \frac{15}{7.42} = 2.02 \text{ min}$

The 55 trips made by the patrolman exceed the minimum sample size of 30 required to apply the CLT, so we may assume the sample means to be normally distributed.

The z-score of the patrolman's average time is: $\frac{60-60}{2.02} = \frac{0}{2.02} = 0$

According to the z-score percentage reference, a z-score of 0 corresponds to .50 or 50%

There is a 50% probability that the patrolman's mean travel time is greater than 60 mins.

2. The sample mean of the 95 polled G.P.A. scores is the same as the population mean: **3.0**

The sample standard deviation is $\frac{.5}{\sqrt{95}} = \frac{.5}{9.75} = .05$

The 95 sampled G.P.A.'s exceed the minimum sample size of 30, so we may apply the CLT.

The z-scores of the minimum and maximum values in the range of interest, 2.99 to 3.01 is:

$$Z_1 = \frac{2.99 - 3.00}{.05} = \frac{-.01}{.05} = -0.2$$

$$Z_2 = \frac{3.01 - 3.00}{.05} = \frac{.01}{.05} = +0.2$$

Referring to the z-score reference table, **the z-scores -0.2 and 0.2 cover a range of apx. 15.86%**

3. The sample mean, $\mu_{\bar{x}}$ is the same as the population mean: 1 hr = 60 mins.

The sample standard deviation is $\frac{25 \text{ mins}}{\sqrt{172}} = \frac{25}{13.11} = 1.91 \text{ min}$

The 172 users reporting cooking times exceed the minimum sample size of 30 required to apply the CLT, so we may assume the sample means to be normally distributed.

The z-score of the average reported cooking time is: $\frac{45-60}{1.91} = \frac{-15}{1.91} = -7.85$

According to the z-score percentage reference, a z-score of -7.85 corresponds to 0%.

There is essentially zero probability that 172 users would average only 45 mins.

Practice

- 128 randomly-sampled students reported how much they spent on a movie at the theater. If the national average amount spent at the movies has a mean of \$15 and standard distribution of \$8, what is the probability that the random sample will give a result within \$0.01 of the true value?
- The time an American family spends doing dishes in the evening has $\mu = 60 \text{ mins}$ and $\sigma = 60 \text{ mins}$. 58 Americans were polled to find the time they spend doing dishes. What is the probability that their average time exceeds 60 minutes?
- Rachel asked 65 second year college students how many credits they have taken. According to the colleges, the average number of credits taken by 2nd year students is 15, with a standard deviation of 7. How likely is it that Rachel got less than 17.17 on average?
- What do you need in order to apply the Central Limit Theorem to sample means?

5. 117 business women were asked how much they spend for lunch, on average. If the national average has a mean of \$30, and standard distribution of \$9, what is the probability that the random sample will return a result within \$0.01 of the true value?
6. According to the phone company, the daily average number of calls made by Americans is 30, with a standard deviation of 10. What is the probability that 117 Americans reported less than 30.92 calls per day, on average?
7. The time spent by the average technician repairing a laptop is governed by an exponential distribution where μ and σ are each 60 minutes. In the month of June, a technician repairs 76 laptops. How likely is it that the average repair time is greater than 77 minutes?
8. 46 teenagers were asked how many .mp3's they purchase each month. According .mp3 sales data, the average has a mean of 15, with a standard distribution of 2. How likely is it that the 46 polled teens averaged within 0.02 of the national average?
9. 44 classrooms were investigated to see how many students they contained. According to school data, the average number of students per classroom is 35, with a standard deviation of 10. How likely is it that the 44 classrooms averaged fewer than 33.49 students?
10. 100 bags of candy were counted to see how many pieces they contained. According to the company that fills the bags, the average number of candies per bag has a mean of 50, and standard distribution of 10. What is the probability that the 100 bags will have an average number within 0.02 of the production average?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 9.7.

9.8 Approximating the Binomial Distribution

Objective

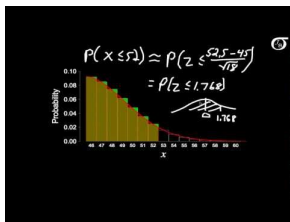
Here you will learn when it is reasonable to approximate a *binomial distribution* with a *normal distribution*, making quick work of probability calculations.

Concept

Suppose you were completing a multiple-choice test, and you are worried that you don't know the information well enough. If there are 75 questions, each with 4 answers, what is the probability that you would get at least 60 correct just by guessing randomly?

You could probably answer this question if you have completed prior lessons on binomial probability, but it would be quite a calculation, requiring you to individually calculate the probability of getting 60 correct, adding it to the probability of getting 61 correct, and so on, all the way up to 75! At the end of the lesson, we will review this question in light of the normal distribution, and see how much more efficient it can be.

Watch This



MEDIA

Click image to the left or use the URL below.

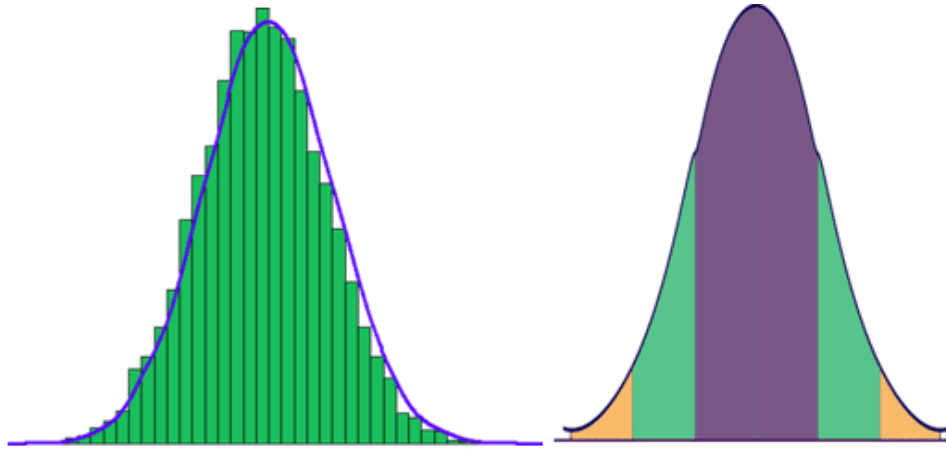
URL: <https://www.ck12.org/flx/render/embeddedobject/67471>

http://youtu.be/CCqWkJ_pqNUjb statistics - The Normal Approximation to the Binomial Distribution

Guidance

Many real life situations involve binomial probabilities, as we saw in prior lessons on binomial experiments. In fact, even many questions that don't appear binomial at first can be formatted so that they are, allowing the probability of success or failure of a given study to be calculated as a binomial probability. Unfortunately, if the probability of success spans a wide range of possible values, the calculation can become very burdensome.

The good news is that there is another way to approximate the probability of success, and you can see what it is by comparing the following graphs. The first graph displays the probability of getting various numbers of heads over 100 flips of a fair coin, in other words, the distribution of a binomial random variable with $P(\text{success}) = .50$. The second graph is a normal distribution. Notice any similarities?



They are extremely similar in shape, in fact, if you follow a “rule of thumb”, you can use a normal distribution to estimate the results of a binomial distribution with quite acceptable accuracy. The rule of thumb for knowing when the normal distribution will provide a good approximation of a binomial distribution with the same mean and standard deviation is:

$$n \times P > 10 \text{ and } n(1 - p) > 10$$

Where n is the number of trials, and p is the probability of success.

If you have determined that a given binomial distribution is a candidate for approximation using a normal distribution, you can calculate the μ and σ of the normal distribution using:

$$\begin{aligned}\mu &= np \\ \sigma &= \sqrt{np(1 - p)}\end{aligned}$$

If you are interested in the comparison between the binomial probability and normal approximation for a particular n or p value, there is an excellent Java applet at http://onlinestatbook.com/stat_sim/normal_approx/index.html that will show the actual values and a graph for any combination of n and p values.

Example A

Can the results of a binomial experiment consisting of 40 trials with a 72% probability of success of be acceptably approximated by a normal distribution?

Solution:

Here, $n = 40$, and $p = 0.72$

- First, is $n \times p > 10$?

$$\begin{aligned}n \times p &= 40 \times .72 = 28.8 \\ 21.6 &> 10 \text{ YES}\end{aligned}$$

- Second, is $n(1 - p) > 10$?

$$n(1 - p) = 40(1 - 0.72) = 40(.28) = 11.2$$

$$11.2 > 10 \text{ YES}$$

Yes, based on our rule of thumb, you could use a normal distribution to approximate the results of this binomial experiment.

Example B

If Kaile wants to estimate the probability of correctly guessing at least 9 answers out of 50 on a true/false exam, can she estimate using a normal distribution?

Solution:

Here, $n = 50$ and $p = 0.50$ (true/false):

$$n \times p = 50 \times .50 = 25$$

$$25 > 10 \text{ Yes}$$

$$n(1 - p) = 50(1 - 0.50) = 50 \times .50 = 25$$

$$25 > 10 \text{ Yes}$$

Example C

Ciere works in a production plant. Due to the balance of speed and accuracy in production, each part off the line has a 98.8% probability of defect free production.



- Can a binomial experiment based on 98.8% probability be approximated if Ciere produces 1000 parts?
- What would be the mean and standard deviation of the appropriate normal distribution?
- What is the probability that Ciere will produce at least 990 parts without a defect in a 1000 part run?

Solution:

a. Here, $n = 1000$ and $p = 0.988$, does this satisfy our 'rule of thumb'?

$$n \times p = 1000 \times 0.988 = 988 \text{ Yes}$$

$$n(1 - p) = 1000(1 - 0.988) = 1000 \times 0.012 = 12 \text{ Yes}$$

If Ciere produces 1000+ parts, this experiment may be approximated using a normal distribution.

b. The mean, μ , and standard deviation, σ , can be evaluated as follows:

$$\mu = np = 1000 \times .988 = 988$$

$$\sigma = \sqrt{np(1 - p)} = \sqrt{988(0.012)} = 3.44$$

c. To calculate this, we need the z -score of 990, which we can calculate using: $Z = \frac{(x - \mu)}{\sigma}$. Once we have the z -score, we can reference a z -score probability table to find the probability of a value above it:

$$z = \frac{(x - \mu)}{\sigma} = \frac{990 - 988}{3.44} = \frac{2}{3.44} = .52$$

Using a reference table (which you can find online or in the previous lesson), we see that the probability of a z -score greater than .52 is 34.60%

The probability of Ciere producing at least 990 parts in a row without a defect is about 34.60%.

Concept Problem Revisited

Suppose you were completing a multiple-choice test, and you are worried that you don't know the information well enough. If there are 75 questions, each with 4 answers, what is the probability that you would get at least 60 correct just by guessing randomly?

Here, since there are 75 questions, $n = 75$, and since each has 4 possible answers, $p = .25$. First we check to see if we can use the normal approximation:

$$n \times p = 75 \times .25 = 18.75$$

$$18.75 > 10 \text{ Yes}$$

Since we can use the normal distribution, we need to calculate the mean and SD of the distribution:

$$\mu = np = 75 \times .25 = 18.75$$

$$\sigma = \sqrt{np(1 - p)} = \sqrt{18.75(.75)} = 3.75$$

Now we need the z -score of our minimum number of correct guesses, 60:

$$z = \frac{(x - \mu)}{\sigma} = \frac{60 - 18.75}{3.75} = \frac{41.25}{3.75} = 11$$

Ha! We don't even need a z -score reference for this one, your chances of randomly guessing 60 or more correctly are virtually nil, since most tables only go up to $z = 3$ or $z = 3.99$. Better study more next time!

Vocabulary

A **binomial distribution** is a frequency distribution of the possible number of successful outcomes in a given number of success/failure trials, each of which has the same probability of success.

A **normal distribution** is a continuous bell-shaped distribution that is balanced about the mean, median, and mode. Sixty-eight percent of the observations fall within plus or minus one standard deviation of the mean, approximately 95% fall within plus or minus two standard deviations, and approximately 99.5% fall within plus or minus three standard deviations.

THE CONTINUITY CORRECTION FACTOR

WE WILL NOT BE DEALING WITH THIS IN OUR EXAMPLES, BUT YOU MAY FIND IN THE FUTURE THAT SINCE A BINOMIAL DISTRIBUTION IS DISCRETE, AND THE NORMAL DISTRIBUTION IS CONTINUOUS, YOU MAY NEED TO TREAT AN INTEGER VALUE OF THE BINOMIAL DISTRIBUTION AS THE INTERVAL BETWEEN .5 BELOW AND .5 ABOVE THE INTEGER, PARTICULARLY IF THE VALUE OF N IS RATHER SMALL.

Guided Practice

1. Is a binomial experiment consisting of 35 trials, each with a 55% probability of success, a good candidate for normal curve approximation?
2. A binomial random variable has a 0.821 probability of success. If data is collected from 48 trials, can the results be viably approximated with a normal distribution?
3. What is the approximate probability of correctly guessing at least 20 questions out of 50, on a true/false exam?

Solution:

1. $n = 35$ and $p = 0.55$, use the rule of thumb from the lesson to evaluate:

$$n \times p = 35 \times .55 = 19.25$$

$$19.25 > 10 \text{ Yes}$$

$$n \times (1 - p) = 35 \times .45 = 15.75$$

$$15.75 > 10 \text{ Yes}$$

Since the experiment meets both qualifications, a normal approximation should be variable.

2. $n = 48$ and $p = 0.721$, check with our rule of thumb:

$$n \times p = 48 \times 0.721 = 39.41$$

$$39.41 > 10 \text{ Yes}$$

$$n \times (1 - p) = 48 \times 0.179 = 7.05$$

$$7.05 < 10 \text{ No!}$$

3. Verify that we can use a normal approximation, using $n = 50$ and $p = 0.50$:

$$n \times p = 50 \times .5 = 25$$

$$25 > 10 \text{ Yes}$$

$$n \times (1 - p) = 50 \times .5 = 25$$

$$25 > 10 \text{ Yes}$$

Since we can use the normal approximation, we need to calculate the mean and SD, so we can get the z -score for 20 (the minimum number of correct answers we want to get).

$$\begin{aligned}\mu &= np = 50 \times .5 = 25 \\ \sigma &= \sqrt{np(1-p)} = \sqrt{25(.50)} = 3.54 \\ z &= \frac{(x-\mu)}{\sigma} = \frac{20-25}{3.54} = 1.41\end{aligned}$$

Consulting our reference, we learn that the probability that a z -score greater than 1.41 will occur is approximately 94%.

You have approximately a 94% probability of correctly guessing at least 20 questions correctly on a 50 question exam.

Practice

- Karen is playing a game of chance with a probability of success of 33%. If she plays the game 43 times, what is the probability that she will win more than 19 times?
- When approximating a binomial distribution, how do you calculate the standard deviation?
- Gregory has created a card game where you either draw a black card or a red card. If you draw a red card, you get a dollar. But if you draw a black card you owe him a dollar. The chance of drawing a red card is 61%. You decide to play against Gregory 26 times. Can you approximate this situation with a normal curve? Why or why not?
- Sue has organized her closet into summer clothing and winter clothing. She closes her eyes and reaches in her closet to pick an outfit. If she selects summer clothing she will be cold (it is winter). The chance of Sue selecting summer clothing is 41%. She decides to select 27 outfits this way. If you were to approximate this with a normal curve, what would the standard deviation be?
- Sharon can't decide between two guys that she likes. She picks a daisy from the garden and decides to play "I like Greg more, I like Stan more" with the petals. The chance of the last petal being "I like Greg more" is 67%. She decides to go through this process with 48 daisies. What is the probability that she will select Greg more than 36 times?
- Vern has to choose between two summer jobs. He painted a wheel red and blue. If the spinner lands in the red area, he works for a landscaping company. If it lands in the blue, he works for a fast food restaurant. The chance of the spinner landing in the red area is 52%. He decides to spin 33 times. What standard deviation would you use to approximate this situation with a normal curve?
- When approximating a binomial distribution, what is the mean?
- George has devised a scheme where he flips a coin to earn money. If it lands on heads you get a quarter. If it lands on tails you give him a quarter. The chance of the coin landing on heads, is 68%. You play against George 22 times. Can you approximate this situation with a normal curve? Why or why not?
- Jade has been practicing shooting a bow and arrow. Based on her target practice she has a 30% chance of hitting the bull's-eye with the bow and arrow. She shoots the bow and arrow 27 times. Can you approximate this situation with a normal curve? Why or why not?
- Steve has created a "grab the marble" game. If you grab a green marble you get a dollar, if you grab a yellow marble you get nothing. There are 31 green marbles, and 69 yellow ones. You decide to reach into the bag and grab a marble 39 times, replacing the marble you grab each time. What is the probability that you will win more than 9 dollars?
- Steve asks another friend to play "grab the marble". If you grab a green marble you get a dollar, if you grab a yellow marble you get nothing. Now, however, there are 49 green and 51 yellow marbles. You decide to reach into the bag and grab a marble 32 times, replacing the marble you grab each time. By approximating this situation with a normal curve, try to predict the expected outcome.

12. Kyle is reading a “Choose Your Own Adventure Book”. He has decided to leave his decisions to chance, so before making a choice between decision 'a' and decision 'b', he flips a coin. If it lands on heads, he selects choice 'a', if it lands on tails, he selects choice 'b'. The trick is that he uses an unfair coin with a probability of heads of 74%. He has to flip the coin 42 times to get to the end of the story. Can you approximate this situation with a normal curve? Why or why not?
13. 19 Students at a local high school are all applying to two different colleges. The chance of each of them getting into their first college of choice is 50%. Can you approximate this situation with a normal curve? Why or why not?
14. Tammy is at the circus, playing a game where she has to throw darts at pink and green balloons on a spinning dart board. If she hits a pink balloon, she earns a ticket. If she hits a green balloon, she receives nothing. If she misses entirely, the throw is not counted. There are 4 pink and 6 green balloons. She decided to play the game 29 times. Can you approximate this situation with a normal curve? Why or why not?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 9.8.

Students practiced the application of the normal distribution, and were introduced to the Empirical rule and the concept and calculation of z -scores. Z -score probabilities were introduced, and students practiced associating z -scores with probabilities via reference tables and technology. The concept of the sampling distribution of the sample mean was introduced, along with the Central Limit Theorem. Finally, students practiced using the normal distribution to approximate the output of binomial random variables.

9.9 References

1. . . CC BY-NC-SA
2. OpenClips. <http://pixabay.com/en/t-shirts-shirts-clothing-apparel-154312/?oq=shirt> .
3. . . CC BY-NC-SA
4. . . CC BY-NC-SA
5. stux. <http://pixabay.com/en/gasoline-prices-petrol-fuel-206098/?oq=gas%20price> .
6. . . CC BY-NC-SA
7. . . CC BY-NC-SA
8. . . CC BY-NC-SA
9. PublicDomainPictures. <http://pixabay.com/en/snow-day-georgia-back-yard-13962/?oq=snow%20yard> .
10. . . CC BY-NC-SA
11. Simon . <http://pixabay.com/en/touring-skis-ski-touring-binding-262028/?oq=skis> .
12. . . CC BY-NC-SA
13. . . CC BY-NC-SA
14. . . CC BY-NC-SA
15. . . CC BY-NC-SA
16. geralt. <http://pixabay.com/en/personal-group-silhouettes-man-112391/?oq=group> .
17. PublicDomainPictures. <http://pixabay.com/en/spinach-pizza-cheese-yellow-white-72123/?oq=cheese%20pizza> .
18. U.S. Department of Agriculture. <https://www.flickr.com/photos/usdagov/6277240734/> .
19. Highway Patrol Images. <https://www.flickr.com/photos/special-fx/6966663529/> .
20. . . CC BY-NC-SA
21. . . CC BY-NC-SA
22. ideowl. <https://www.flickr.com/photos/special-fx/6966663529/> .

CHAPTER

10**Predicting and Testing****Chapter Outline**

- 10.1 THE NULL HYPOTHESIS**
 - 10.2 CRITICAL VALUES**
 - 10.3 TAILS**
 - 10.4 CONFIDENCE INTERVALS**
 - 10.5 THE T-TEST**
 - 10.6 PUTTING IT TOGETHER**
 - 10.7 REFERENCES**
-

In this chapter you will learn about testing claims and hypotheses. The process of testing a claim using statistics is actually rather simple. You identify the value you are testing and the value you are comparing it to and determine the likelihood that they could both come from the same process by simple chance. If the probability of both occurring by chance is extremely small, beyond a specified threshold, you determine that there must be something causing the discrepancy.

You will also learn about predicting values for which you have no observed data. By calculating the location of a line representing the best trend of observed data, you can make educated predictions regarding the values of data points beyond the scope of your observations.

10.1 The Null Hypothesis

Objective

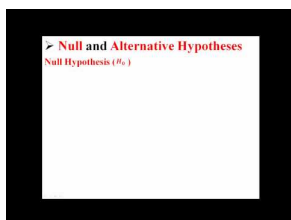
Here you will learn about the first step in testing a hypothesis, identifying the Null Hypothesis.

Concept

Suppose you wanted to show that a coin was fair, what would be involved with setting up the experiment(s) to validate or deny the claim?



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67533>

<http://youtu.be/5N7L1cGCL-w> statslectures - Null and Alternative Hypothesis

Guidance

One of the main uses for statistics is the testing of hypothesis. Commonly, a statement is made, such as “I think that the probability of a student in my class preferring red candy is 70%,” and experiments are conducted to determine the validity of the claim. However, to strengthen the results, specific steps should be followed as the experiment is

set up, carried out, and reviewed or reported. The first step is the establishment of the *null hypothesis* notated as H_0 and *alternative hypothesis*, notated as H_1 .

- The *null hypothesis*, H_0 , is the opposite of what you are hoping to claim. In the case of the claim above, the null hypothesis would be “The probability of a student in class preferring red candy is less than 70%” The null hypothesis may also be referred to as the “no-change” value, as it is the default conclusion.
- The *alternative hypothesis*, H_1 , is the clear and concise statement of the initial claim. In the case of the candy above, the hypothesis could be: “An average of 70% or greater of the students in school preferred candy to green candy,” or simply $H_1 : p \geq 70\%$, where p stands for “population percent”.
- It is extremely important that the alternative hypothesis and null hypothesis be mutually exclusive, meaning that if one is true, the other must be false.

In this lesson, you will practice identifying the null hypothesis for a number of claims.

Example A

The claim is made that a certain medication relieves headaches for more than 75% of patients who take it. What null hypothesis could be assumed during the investigation of this claim?

Solution:

The null hypothesis for the claim “More than 75% of patients who take medication Z experience headache relief,” would be the mutually exclusive statement “75% or fewer of the patients who take medication Z experience headache relief”. This could be written as $H_0 : p \leq 75\%$.

Example B

A researcher wants to demonstrate that an average of more than 8 out of 10 dogs prefer the taste of a new dog food. What alternative hypothesis and null hypothesis might he work with? Use notation.



Solution:

The researcher’s hypothesis could be stated as: “On average, greater than 80% of dogs prefer the taste of dog food X.” In notation that would be:

$$H_1 : p > 80\%$$

The null hypothesis would be: “On average, 80% or fewer of dogs prefer the taste of dog food X.” In notation: $H_0 : p \leq 80\%$.

Example C

What alternative hypothesis would correlate to the null hypothesis $H_0 : p < 60\%$?

Solution:

If the null hypothesis is $H_0 : \mu < 60\%$, then the alternative hypothesis is $H_1 : p \geq 60\%$.

Concept Problem Revisited

Suppose you wanted to show that a coin was fair, what would be the first steps involved with setting up the experiment(s) to validate or deny the claim?

The first step would be to identify the hypothesis and null hypothesis:

$$H_1 : p = 50\%$$

$$H_0 : p \neq 50\%$$

Vocabulary

A **hypothesis** is a claim or supposition. In statistics, an **alternative hypothesis** is commonly defined with the notation H_1 .

A **null hypothesis** is the mutually exclusive corollary to a hypothesis, commonly denoted H_0 .

Guided Practice

1. What is the null hypothesis to the claim that more people like cereal X than cereal Y?
2. What alternative hypothesis would correlate to the null hypothesis $H_0 : 0 \leq \sigma \leq 5$?
3. What alternative hypothesis and null hypothesis could be stated to define the claim that more people own blue cars than red cars? Use notation.

Solutions:

1. The null hypothesis is the opposite of what you are hoping to claim, so H_0 : More people like cereal Y than cereal X.
2. $H_1 : 0 \geq \sigma$ or $\sigma \geq 5$
3. $H_0 : R \geq B$ and $H_1 : R < B$

Practice

For questions 1-8, state the null hypothesis for the given alternative hypothesis.

1. The number of sales of mp3 players would not go down if the price were raised by \$10.
2. The average dog owner owns 2 or more cats.
3. The average cat owner does not own a dog.
4. The average subcompact car gets more than 30 mpg.
5. The average computer gamer owns 4 or more games.
6. The average temperature in Northern CO during the month of March is less than 70 degrees.
7. At least 28% of astronomers can name more than 50 stars.

8. Less than 35% of astronomers can name at least 75 stars.

For questions 9-16, use notation to state the alternative hypothesis and null hypothesis.

9. Less than 1% of U.S. citizens participate in a militia.

10. Between 20% and 35% of students watch cartoons on Saturday morning.

11. More than 71% of high school students can name a favorite fantasy book.

12. Less than 1 in 5 U.S. college students ride a bike to school.

13. Less than 2% of U.S. adults have milked a cow.

14. Between 2% and 10% of Americans are vegetarians.

15. Less than 4% of high school students take Statistics.

16. More than 80% of students say they learn better through video than with a textbook.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 10.1.

10.2 Critical Values

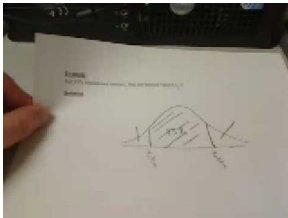
Objective

Here you will learn about *critical values*, which are related to z -scores, and are used to determine *confidence intervals* (among other things).

Concept

What are critical values? Why is it important to know if a test is left, right, or two-tailed, when calculating critical values?

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67543>

<http://youtu.be/ZkCAHlzkMS0> hknuth123 - Finding Critical Values

Guidance

When you calculate the probability that a range of values will occur given a random variable with a particular distribution, you often use a z -score reference or calculator.

Critical values are the values that indicate the edge of the *critical region*. *Critical regions* (also known as *rejection regions*) describe the entire area of values that indicate you reject the null hypothesis. In other words, the *critical region* is the area encompassed by the values *not* included in the initial claim - the area of the 'tails' of the distribution.

In this lesson, we will use a table to find critical values. Although critical values may refer to other types of distributions, for the moment we will be dealing only with the Z -score critical values of a normal distribution. There is a table of Z -score values that you may refer to here:

TABLE 10.1:

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z
0.0	0.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359	0.0
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.1
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.2
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517	0.3
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879	0.4

TABLE 10.1: (continued)

0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224	0.5
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.6
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	0.7
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.8
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389	0.9
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	1.0
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883	1.1
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015	1.2
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	1.3
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	1.4
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	1.5
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	1.6
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	1.7
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	1.8
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767	1.9
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	2.0
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857	2.1
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989	2.2
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	2.3
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	2.4
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	2.5
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964	2.6
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974	2.7
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981	2.8
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	2.9
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999	3.0
3.1	0.999	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993	3.1
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	3.2
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	3.3
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	3.4
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	3.5
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.6
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.7
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	3.8
3.9	1	1	1	1	1	1	1	1	1	1	3.9
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	Z

In the next lesson, we will discuss how to identify a test as Left, Right, or Two-tailed. For now, the problems will specify which type you are working with since it affects the location of the critical value(s), and therefore the area of the critical region.

If you have completed the prior lesson(s) on Z-scores, you should recognize Z-score critical values as the Z-scores associated with a given percentage. The reason there is another term, *critical values*, instead of just **Z-scores**, is that the concept of critical values is also applicable to other types of distributions, such as the student's *t*-score distribution discussed in the lesson **Degrees of Freedom**.

Example A

What is the critical value $\left(Z_{\frac{\alpha}{2}}\right)$ for a 95% confidence level, assuming a two-tailed test?

Solution:

A 95% confidence level means that a total of 5% of the area under the curve is considered the critical region.

Since this is a two-tailed test, $\frac{1}{2}$ of 5% = 2.5% of the values would be in the left tail, and the other 2.5% would be in the right tail. Looking up the Z -score associated with 0.025 on a reference table, we find 1.96. Therefore, +1.96 is the critical value of the right tail and -1.96 is the critical value of the left tail.

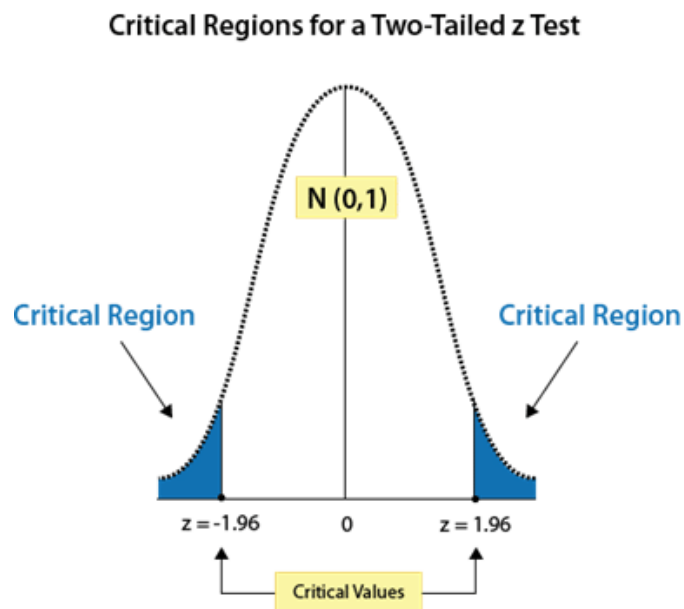
The critical value for a 95% confidence level is $Z = +/-1.96$.

Example B

Sketch the Z -score critical region for Ex A.

Solution:

Sketch the graph of the normal distribution with the given values and mark the critical values from Ex. A, then shade the area from the critical values away from the center. The shaded areas are the critical regions.

**Example C**

What would be the critical value for a right-tailed test with $\alpha = 0.01$?

Solution:

If $\alpha = 0.01$, then the area under the curve representing H_1 , the alternative hypothesis, would be 99%, since α (alpha) is the same as the area of the rejection region. Using the Z -score reference table above, we find that the Z -score associated with 0.9900 is approximately 2.33.

It appears that the critical value is $Z = 2.33$.

Let's see if that answer makes sense. Since this is a right-tailed test, α is on the right end of the graph, and $1 - \alpha$ is on the left. A Z -score of +2.33 is well to the right of the center of the graph, and describes the area under the curve from that point to the left, nearly the entire graph. Since the initial question specified $\alpha = 0.01$, indicating that only 1% of the area is in the critical region, $Z = +2.33$ is quite reasonable.



Concept Problem Revisited

What are critical values? Why is it important to know if a test is left, right, or two-tailed, when calculating critical values?

Critical values are the values which separate the area indicating the null hypothesis should be rejected from the area suggesting it not be rejected. It is important to know what kind of test you are dealing with, because the critical value is often calculated from a probability range, and the Z-score of that range changes depending on what portion of the distribution you are evaluating.

Vocabulary

Critical values are values separating the values that support or reject the null hypothesis.

Critical regions are the areas under the distribution curve representing values that support the null hypothesis.

Guided Practice

1. What would be the critical value for a left-tailed test with $\alpha = 0.01$?
2. What would be the critical *region* for a two-tailed test with $\alpha = 0.08$?
3. What would be the α for a right-tailed test with a critical value of $Z = 1.76$?

Solutions:

1. A left-tailed test with $\alpha = 0.01$ would have 99% of the area under the curve outside of the critical region. If we use a reference to find the Z-score for 0.99, we get approximately 2.33. However, a Z-score of 2.33 is significantly to the right of the center of the distribution, including all the area to the left and only leaving a very small alpha value on the right. While we are indeed looking for a critical value with only a very small alpha, this is a *left-tailed* test, so the critical value we need is *negative*.

$$Z = -2.33$$

2. We are looking for the critical region here, but let's start by finding the critical values. This is a two-tailed test, so half of the alpha will be in the left tail, and half in the right. That means that we are looking for a positive/negative critical value associated with an alpha of 0.04, which indicates that we need to find the Z-score for $(1 - 0.04) = 0.96$. Referring to the Z-score table, we see that 0.96 corresponds to approximately 1.75. The critical values, then are ± 1.75 , and the critical region would be $Z < -1.75 \cup Z > 1.75$.

3. The area under the curve associated with a Z-score of 1.76, according to the reference table above, is 0.9608. Since 96.08% of the area is to the left of $Z = 1.76$, that leaves approximately $1 - 0.9608 = 0.0392$ as the area in the critical region.

$$\alpha = 0.0392$$

Practice

For questions 1-9, identify the critical value(s) for each α :

1. $\alpha = 0.05$, left-tailed
2. $\alpha = 0.02$, right-tailed
3. $\alpha = 0.05$, two-tailed
4. $\alpha = 0.02$, left-tailed
5. $\alpha = 0.01$, right-tailed
6. $\alpha = 0.01$, two-tailed
7. $\alpha = 0.1$, left-tailed
8. $\alpha = 0.1$, right-tailed
9. $\alpha = 0.1$, two-tailed

For questions 10-15, find α for the given critical value(s):

10. $Z = 1.28$, right-tailed
11. $Z = 1.65, -1.65$
12. $Z = -3.10$, left-tailed
13. $Z = 2.58, -2.58$
14. $Z = -1.65$, left-tailed
15. $Z = 2.33$, right-tailed

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 10.2.

10.3 Tails

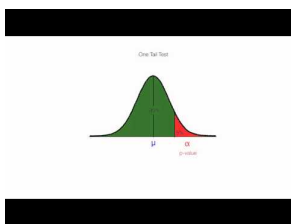
Objective

Here you will learn to recognize left-tailed, right-tailed, and two-tailed tests.

Concept

How many tails are there in the test for the claim: “Less than 5% of high school students are familiar with the tails of a distribution?”

Watch This



MEDIA

Click image to the left or use the URL below.

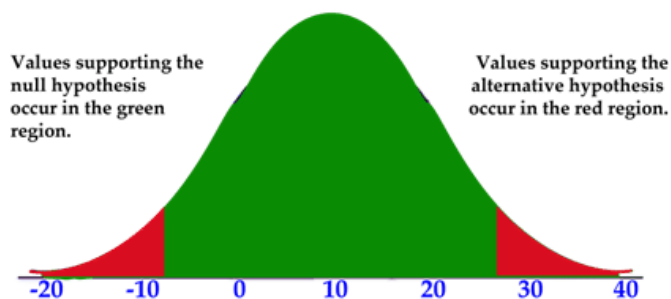
URL: <https://www.ck12.org/flx/render/embeddedobject/68324>

http://youtu.be/B9u_grPccUs Statisticsfun - How to calculate One Tail and Two Tail Tests For Hypothesis Testing.

Guidance

The “tails” of a test are the values outside of the *critical values*. In other words, the tails are the ends of the distribution, and they begin at the greatest or least value included in the alternative hypothesis (the critical values).

In the graph below, the tails are in red and the rest of the distribution is in green. The critical values of the test in the image are -8 and 28, as these are the dividers between values supporting the alternative and null hypothesis. The area in red can also be seen as the *rejection region*, since an observed value in this region indicates that the null hypothesis be rejected.

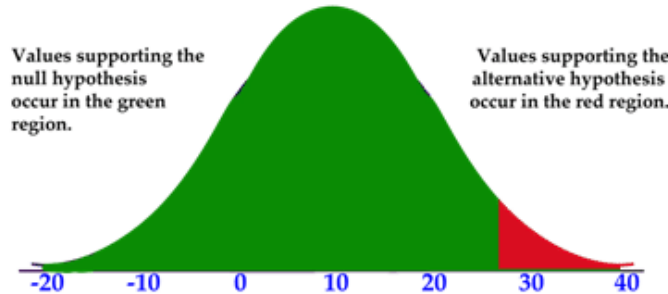


There are three types of test from a “tails” standpoint:

- A *left-tailed test* only has a tail on the left side of the graph:



- A **right-tailed test** only has a tail on the right side of the graph:



- A **two-tailed test** has tails on both ends of the graph. This is a test where the null hypothesis is a claim of a specific value. For example: $H_0 : X = 5$



Example A

A researcher claims that black horses are, on average, more than 30 lbs heavier than white horses, which average 1100 lbs. What is the null hypothesis, and what kind of test is this?

Solution:

The null hypothesis would be notated $H_0 : \mu \leq 1130 \text{ lbs}$

This is a right-tailed test, since the tail of the graph would be on the right. Recognize that values above 1130 would indicate that the null hypothesis be *rejected*, and the red area represents the *rejection region*.

Example B

A package of gum claims that the flavor lasts more than 39 minutes. What would be the null hypothesis of a test to determine the validity of the claim? What sort of test is this?

Solution:

The null hypothesis would be notated as $H_0 : \mu \leq 39$.

This is a right-tailed test, since the rejection region would consist of values greater than 39.

Example C

An ice pack claims to stay cold between 35 and 65 minutes. What would be the null hypothesis of a test to determine the validity of the claim? What sort of test would it be?

Solution:

The null hypothesis would be $H_0 : 35 > \mu \cup \mu > 65$.

This is a two-tailed test, since the null hypothesis contains the values above and below a given range.

Concept Problem Revisited

How many tails are there in the test for the claim: "Less than 5% of high school students are familiar with hypothesis tails?"

This is a left-tailed test, since the null hypothesis is $H_0 : X \geq 5\%$ meaning that the *rejection* region contains only values below a given value.

Vocabulary

A **left-tailed test** is identified by a rejection region (tail) only on the left side of a graph.

A **right-tailed test** is identified by a rejection region (tail) only on the right side of a graph.

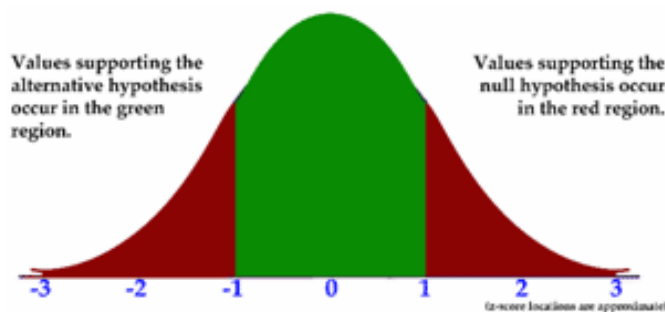
A **two-tailed test** is identified by a rejection region (tails) on both ends of the graph.

Critical values are values which separate values supporting the alternative and null hypothesis.

Guided Practice

For each question: a) State the null and alternative hypothesis for a test to determine the validity of the claim, and b) Identify left, right, or two-tailed test type.

1. The weatherman claims an average of 2" - 2.5" of snow throughout the county.
2. The average Siberian Husky weighs more than 35 pounds.
3. The United Parcel Service claims less than 5-day shipping times for ground packages.

**Solutions:**

1. a) $H_1 : 2'' < \mu < 2.5''$ and $H_0 : 2'' > \mu \cup \mu > 2.5''$

b) This is a two-tailed test, since the rejection region is at the ends of the graph.

2. a) $H_1 : \mu > 35 \text{ lbs}$ and $H_0 : 35 \text{ lbs} \geq \mu$

b) This is a right-tailed test, since the rejection region is located on the right end of the graph.

3. a) $H_1 : \mu < 5 \text{ days}$ and $H_0 : \mu \geq 5 \text{ days}$

b) This is a left-tailed test, since the 'reject' values are on the left end of the graph.

4. a) $H_1 : -1 < Z < 1$ and $H_0 : -1 > Z \cup Z > 1$

b) This is a two-tailed test, since the tails are located at the ends of the graph.

Practice

For each question: a) State the null and alternative hypothesis for a test to determine the validity of the claim, and b) Identify left, right, or two-tailed test type.

- The average miniature horse is less than 34 inches at the shoulder. Tiny Equines claims that their miniature horses are an average of 3 inches less than the breed average.
- Speedy Solutions is a delivery company that claims all deliveries average less than 72 hours.
- Terrific Textiles claims that 45% to 55% of t-shirts sold in CO are red.
- Gas Hater Car Sales claims that 70% or more of the cars on the lot average more than 40 mpg.
- The 2011 average service time for Fast Fatz Burgers was 56 seconds. The store manager claims that the 2012 average is more than 5 seconds faster.
- Colorful Cavity Causers sells multi-colored candies by the bag. Each bag is claimed to average between 5.5 and 6.5 oz.
- Aaron works for an electronics company, quality-testing batteries. The company claims a minimum of 5 hrs battery life.
- The lessons in this Probability and Statistics text average 13 practice problems each.
- The outside temperature in Maui averages between 72 and 83 degrees year-round.
- 72% of elementary school teachers are female.



Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 10.3.

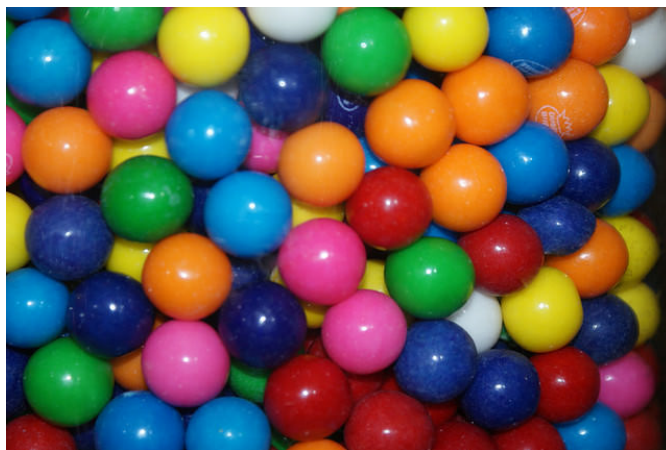
10.4 Confidence Intervals

Objective

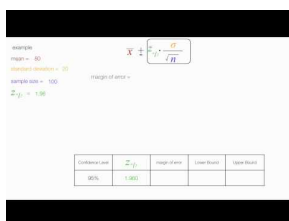
Here you will learn about confidence intervals, which are ranges within which you expect the population parameter to occur with a particular probability

Concept

Suppose you were at a county fair and saw a large jar full of gumballs, maybe 1000 of them, with a sign that said “Guess the Number, Win a Prize!” If the rules of the game are that you could win a \$10 prize by guessing within 200 gumballs either way, or a \$50 prize by guessing within five gumballs either way, but you have to specify which prize you are trying for before submitting your guess, which would you choose?



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/68326>

<http://youtu.be/dNfpsVLaaEE> statisticsfun - How to calculate Confidence Intervals and Margin of Error

Guidance

The general concept of confidence intervals is pretty intuitive: It is easier to predict that an unknown value will lie somewhere within a wide range, than to predict it will occur within a narrow range. In other words, if you are

making an educated guess about an unknown number, you are more likely to be correct if you predict it will occur within a wider range. This idea is reflected in the concept question above, where the reward is greater if you guess within a smaller range, because the contest creator knows that your chance of guessing correctly is much less if you have to guess within a smaller range.

A **confidence interval**, centered on the mean of your sample, is the range of values that is expected to capture the population mean with a given level of confidence. A wider confidence interval is a greater range of values, resulting in a greater **confidence level** that the range will include the population mean. By convention, you will mostly be concerned with identifying the intervals associated with 90%, 95%, and 99% confidence levels.

Calculate the confidence interval by combining the sample mean with the **margin of error**, found by multiplying the standard error of the mean by the z -score of the percent confidence level:

$$\text{confidence interval} = \bar{x} \pm \text{margin of error}$$

$$\text{margin of error} = Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

It is common, but incorrect, to assume that a confidence level indicates the probability that the mean of the population will occur within a given range of the mean of your sample. A 95% confidence interval means that if you took 100 samples, all of the same size, and formed 100 confidence intervals, 95 of these intervals would capture the population mean.

The confidence level indicates the number of times out of 100 that the mean of the population will be within the given interval of the sample mean.

Example A

Suppose you took 100 unbiased random samples of the heights of U.S. women (recall that height is normally distributed), each sample containing 30 women. What can you say about the means of the samples ($\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{100}$) compared to the population mean?

Solution:

Since height is normally distributed, we know that approximately 95% of women will have a height within two standard deviations of the mean (remember the Empirical Rule?). That means that out of 100 samples, we can assume that 95 of them will have a mean within 2 standard deviations of the population mean.

Example B

Suppose the mean of the means of our 100 samples from Example A is 5'5", in other words, $\bar{X} = 5'5''$. Within what range of heights can we expect the population mean to be, with 95% confidence? Assume a standard deviation of 1.5".



Solution:

Remember that since height is normally distributed, 95% of the values lie within 2 standard deviations of the mean, we need to identify that range of values.

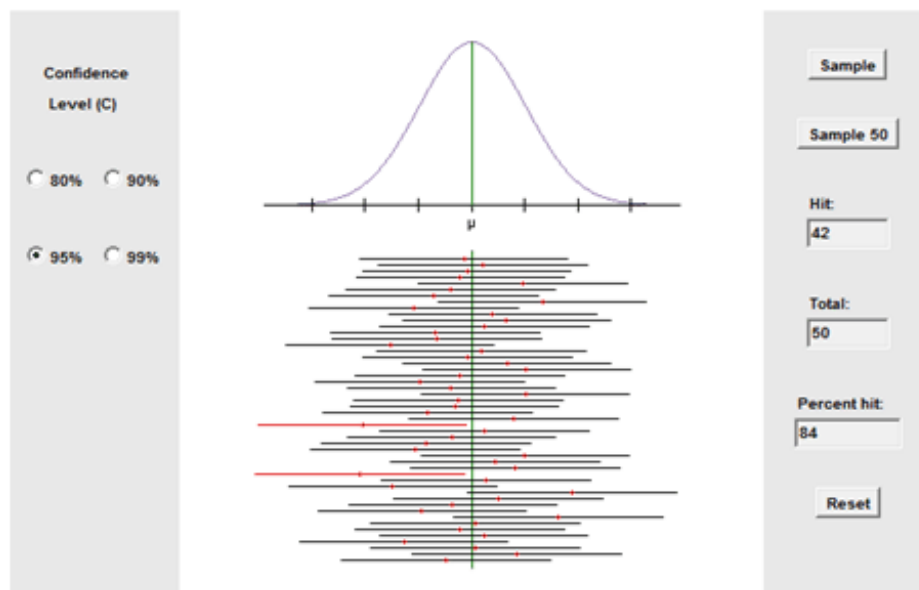
- First we need to use $Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ to identify the margin of error (since we are looking for a 95% confidence level, this is the range of values within 2 standard deviations of the sample mean). Since $\sigma = 1.5''$, in this case we get $2 \times \frac{1.5}{\sqrt{100}} = 2 \times \frac{1.5}{10} = 2 \times 0.15 = 0.3''$ above and below \bar{X} .
- The interval then is $5'2''$ to $5'8''$, or three inches above and below the mean of $5'5''$.

We can say that there is a 95% probability that the mean of our 100 samples would be within 0.3 inches either way of the population mean. Since the mean of our sample is $5'5''$, we can say that the population mean is between $5'4.7''$ and $5'5.3''$ with 95% confidence.

Mathematically: $5'4.7'' < \mu < 5'5.3''$

Example C

Suppose you plot the mean of each of your height samples on a graph, and drawing a line each way of the mean of each sample to represent 2 standard deviations. If you were to do this for 50 of the samples, you might end up with an image like the one below.



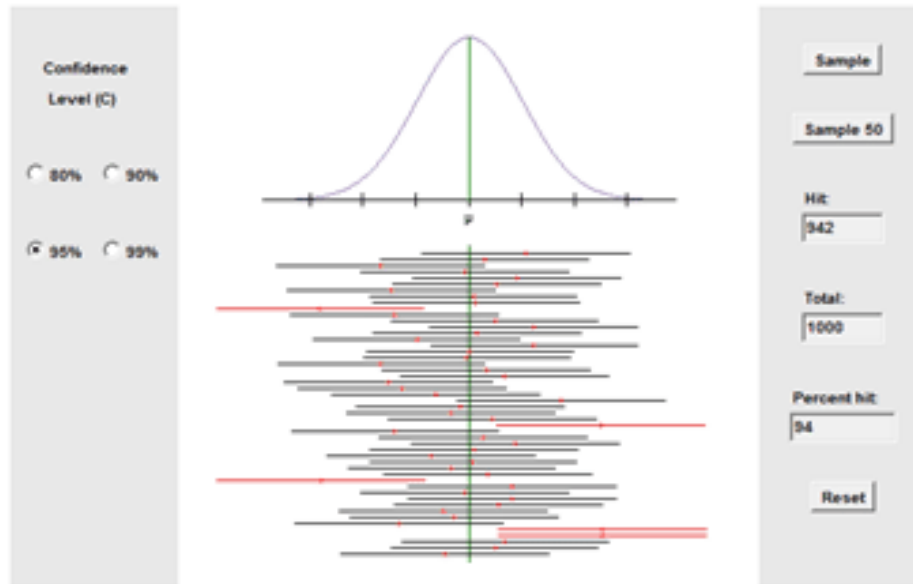
(The image is a screen capture from the interactive applet at: http://bcs.whfreeman.com/ips5e/content/cat_010/applets/confidenceinterval.html .)

At the top of the image is a normal curve. Each of the lines below the curve has a length that represents a 95% confidence interval, centered on the mean (in red) of the sample.

- What is indicated by the lines that are all red in color?
- What value is indicated by the vertical red center line on each interval?
- What does the “percent hit” number mean? How would it change if you were to continue taking more and more samples of 60 each?

Solution:

- The lines that are colored entirely red have a mean that is greater than 2 standard deviations away from the population mean. In other words, the mean of those two samples was not within the stated *confidence interval* (95%).
- The vertical red center line represents the mean of each sample.
- The “percent hit” number indicates the percentage of times that the population mean was included in the confidence interval of sample means. If you were to continue plotting sample means and confidence intervals, the percent hit would approach 95%. In fact, here is the same graph after 1000 sample runs:



Concept Problem Revisited

Suppose you were at a county fair and saw a large jar full of gumballs, maybe 1000 of them, with a sign that said “Guess the Number, Win a Prize!” If the rules of the game are that you could win a \$10 prize by guessing within 200 gumballs either way, or a \$50 prize by guessing within five gumballs either way, but you have to specify which prize you are trying for before submitting your guess, which would you choose?

This problem/question is meant to give you an intuitive feeling for the concept of a confidence interval or confidence level. It should be clear that you would have a greater *level of confidence* in trying for a \$10 prize that you would win simply by guessing within $\pm 20\%$ of the number, than in trying for \$50 by guessing within $\pm 0.5\%$ of the number!

Vocabulary

A **confidence interval** is the interval within which you expect to capture a specific value. The confidence interval width is dependent on the confidence level.

A **confidence level** is the probability value associated with a confidence interval.

Guided Practice

- Suppose you took 40 unbiased random samples of the number of candies in a \$0.75 bag of candy from a particular factory. The factory states that the number of candies per bag is normally distributed. What can you say about the mean number of candies in your sample?

- Suppose the factory states that the number of candies per bag has $\sigma = 2$. If each sample includes data from 40 bags of candies ($n = 40$), what is the standard error of the mean $\left(\frac{\sigma}{\sqrt{n}}\right)$?
- If the sample mean is 38 candies, within what interval could we expect 99 out of each 100 samples to contain the population mean? What is that interval known as?
- What is the more common way to describe the fact that “expect 99 out of each 100 samples contain the population mean”?

Solutions:

- Since the population is normally distributed, we can state that the mean of the sample follows the Empirical Rule.
- The standard error of the mean is calculated as $\frac{\sigma}{\sqrt{n}}$, so $SEM = \frac{2}{\sqrt{40}} = \frac{2}{6.32} = .31$
- The interval is called the **confidence interval**, and it is calculated as $\bar{x} \pm z_{\frac{\alpha}{2}} \times \bar{\sigma}$:

$$38 \pm z_{0.005} \times .316$$

$$38 \pm 2.58 \times .316$$

$$38 \pm 0.81528$$

Therefore, the confidence interval is approximately 37.18 to 38.82.

- Saying that you “expect 99 out of each 100 samples contain the population mean”, is the same as saying that the interval has a 99% confidence level.

Practice

- What is a confidence interval?
- What is the formula for calculating the confidence interval?
- What is the difference between a confidence interval and a confidence level?
- What is a margin of error?
- How is the margin of error calculated?
- What common misconception about confidence level is corrected by stating that a 99% confidence level means that 99 out of 100 samples are expected to contain the population mean?
- If a population is known to have an approximately normal distribution, but the standard deviation is unknown, how can the population standard deviation be approximated?
- If the sample mean is unknown, is it safe to use the population mean as the sample mean?
- What Z-score corresponds to a 98% confidence interval?
- What confidence interval is associated with a Z-score of 2.576, assuming a two-tailed test?
- Which confidence level would describe a wider confidence interval, 80% or 85%?
- A factory produces bags of marbles for a toy store. The factory has previously calculated that the $\sigma = 1$ marble per bag. If you were to sample 35 bags and calculate $\bar{\mu} = 40$, within what range could you predict μ , with 98% confidence?
- Interpret your results from question 12, in context.
- The manager of a clothing store is attempting to estimate the mean number of customers that pass through her store each day. If the data from past estimates and other franchises suggests that $\sigma = 78$, and the manager

has collected the customer counts in the table below from a SRS (Simple Random Sample), what can the manager predict the range of customers to be, with 50% confidence?

TABLE 10.2:

148	298	210	213	315	129	145	148	131	281	317
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

15. Interpret your answer from problem 14, in context.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 10.4.

10.5 The T-Test

Objective

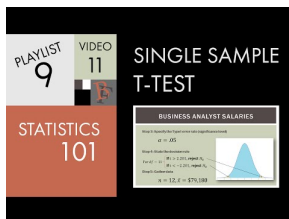
Here you will learn about the *Student's t-distribution* and *degrees of freedom*.

Concept

In previous lessons, we calculated confidence intervals for the mean of a population based on data from a large sample ($n \geq 30$). Does that mean that sample sizes of less than 30 are useless? If not, how do you calculate a confidence interval based on data from a smaller sample?

Watch This

This video is quite a bit longer than the average, as it is quite detailed in following the steps involved with testing a hypothesis using a t -test. Should you opt to watch the entire video, do not be concerned if the information past 22:30 doesn't make sense, as we have not discussed p -values in this course!



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/68439>

<http://youtu.be/dDsKP7wVpzM> Brandon Foltz - Statistics 101: Single Sample Hypothesis T-test

Guidance

When you are attempting to estimate the mean of a population, it is generally best to collect as many data points as possible from the population. Unfortunately, in the 'real world', samples are not always easy to collect. Sometimes you simply do not have access to the 30 or more data points required to use the Central Limit Theorem. If you happen to have good data giving you the standard deviation of the population (σ), then it is generally permissible to use the z -score to calculate a confidence interval regardless. However, when you do not know σ , and you do not have enough data to estimate it with your sample ($n < 30$), a z -score confidence interval is not reliable.



Fortunately, there is a solution. A confidence interval can be calculated from a small sample when we do not know the population standard deviation, if we do two things differently:

- Since we do not know σ , we instead use the best approximation for it that we have: s (the sample standard deviation).
- We modify the confidence interval formula to use a *Student's t -distribution* reference, rather than the z -score percentage reference.

$$\text{Confidence Interval for } n < 30 : \bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

The Student's t -distribution is similar to the normal distribution, except it is more spread out and wider in appearance, and has thicker tails. The differences are more exaggerated when there are fewer data points, and therefore fewer *degrees of freedom*. Degrees of freedom are essentially the number of samples that have the 'freedom' to change without necessarily affecting the sample mean. A clear description of degrees of freedom is beyond the scope of this lesson, but you can find many online lessons describing them if you are interested. For our purposes, all you really need to know about degrees of freedom is that there is always one less degree of freedom than the number of data points:

$$df = n - 1$$

The reason you need to know how to find the number of degrees of freedom is quite simple: the t -distribution has a different value for each number of degrees of freedom, as you can see in the reference below:

df	$t_{0.200}$	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	$t_{0.0025}$	$t_{0.001}$	$t_{0.0005}$
1	1.376	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.061	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.978	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.941	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.920	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.906	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.896	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.889	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.883	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.879	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.876	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.873	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.870	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.868	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.866	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.865	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.863	1.333	1.740	2.110	2.576	2.898	3.222	3.646	3.965
18	0.862	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.861	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.860	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.859	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.858	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.858	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.857	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.856	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.856	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.855	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.855	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.854	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.854	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646

To use the reference, find the point of intersection of the degrees of freedom on the left with the $\frac{\alpha}{2}$ (for 2-tailed tests) or α (for one-tailed tests) across the top. If you need a refresher on the calculation of α , you can reference the lesson **Critical Values** (essentially the alpha value is the area in the tail(s) of the distribution, that is $1.0 - \text{confidence level}$).

Example A

How many degrees of freedom are there in a sample of size $n = 11$?

Solution:

Recall that $df = n - 1$:

$$df = 11 - 1$$

$$df = 10$$

Example B

What is the t -score multiplier for a two-tailed test with a 98% confidence level, given $n = 16$?

Solution:

Begin by finding α :

$$\alpha = 1 - .98$$

$$\alpha = .02$$

Then, identify the number of degrees of freedom, $n - 1$:

$$df = 16 - 1$$

$$df = 15$$

Now we reference the t -score table.

Since this is a two-tailed test, we need to look up $t_{\frac{\alpha}{2}} = t_{.01}$, cross-referenced with $df = 15$:

df	$t_{0.200}$	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$
1	1.376	3.078	6.314	12.706	31.821
2	1.061	1.886	2.920	4.303	6.965
3	0.978	1.638	2.353	3.182	4.541
4	0.941	1.533	2.132	2.776	3.747
5	0.920	1.476	2.015	2.571	3.365
6	0.906	1.440	1.943	2.447	3.143
7	0.896	1.415	1.895	2.365	2.998
8	0.889	1.397	1.860	2.306	2.896
9	0.883	1.383	1.833	2.262	2.821
10	0.879	1.372	1.812	2.228	2.764
11	0.876	1.363	1.796	2.201	2.718
12	0.873	1.356	1.782	2.179	2.681
13	0.870	1.350	1.771	2.160	2.650
14	0.868	1.345	1.761	2.145	2.624
15	0.866	1.341	1.753	2.131	2.602

The t -score multiplier is 2.602.

Example C

You are given a sample of $n = 19$, where $s = 4.3$, and $\bar{x} = 26$. What confidence interval would you use to bracket μ with a confidence level of 95%?

Solution:

The first thing to note is that we do *not* have a large enough sample to use a z -test, since $n < 30$, so we will instead use a t -test.

Recall the t -test formula for calculating confidence interval:

$$\bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

We are given the values of \bar{x} , s and n in the question text. In order to use the t -test, we will need to know the t -score multiplier $t_{\frac{\alpha}{2}}$, so we will need the values of α and the degrees of freedom (df).

The α value is a direct result of the confidence level, which is 95% in this case. With a 95% confidence level, $\alpha = 0.05$ or 5% of the total area under the curve.

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$

Remember that the degrees of freedom (df) are 1 less than the number of data points.

$$df = 19 - 1$$

$$df = 18$$

Now we can use the t -score table to reference $df = 19$ and $t_{0.025}$:

$$t_{\frac{\alpha}{2}} = 2.101$$

Putting everything together:

$$\begin{aligned} \bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) \\ 26 \pm 2.101 \left(\frac{4.3}{\sqrt{19}} \right) \\ 26 \pm 2.101 \left(\frac{4.3}{4.359} \right) \\ 26 \pm 2.101(.986) \\ 26 \pm 2.072 \end{aligned}$$

The confidence interval is 23.928 to 28.072, with a 95% confidence level.

Concept Problem Revisited

In previous lessons, we calculated confidence intervals for the mean of a population based on data from a large sample ($n \geq 30$). Does that mean that sample sizes of less than 30 are useless? If not, how do you calculate a confidence interval based on data from a smaller sample?

Sample sizes less than 30 are certainly not useless. A confidence interval from a sample of size $n < 30$, can certainly be calculated, even if σ is unknown. Rather than using a z -test, however, you use a t -test, referencing values from a Student's t -distribution, and estimate the value of σ with the sample standard deviation, denoted $\bar{\sigma}$ or s .

Vocabulary

A **Student's t -distribution** is a distribution similar to the normal distribution, with slightly greater spread and thicker tails. It is commonly used in the calculation of confidence intervals when $n < 30$.

A **t -test** is a hypothesis test with a rejection area calculated from a t -distribution.

Guided Practice

1. If you are conducting a hypothesis test with $n = 27$ and $\sigma = 3$, is it permissible to conduct a z -test?
2. What is α for a two-tail t -test with a 90% confidence level?
3. What is the 95% confidence interval for a one-tailed test with $n = 17$, $\bar{x} = 142.23$, and $\bar{\sigma} = 13$?

Solutions:

1. Since you know the value of σ , a z -test is permissible.

- Any test with a 90% confidence level would have $\alpha = 0.1$. Since it is a two-tail test, we would look up $t_{0.05}$ on the reference chart, because $\frac{1}{2}$ of the alpha would be on each end of the curve.
- The first thing to decide is what type of test to use. Since $n < 30$, and we do not know σ , it would not be appropriate to use a z -test, so we will use a t -test instead. To calculate the confidence interval for a t -test, use the formula: $\bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$.

- \bar{x} is given: **142.23**
- $t_{\frac{\alpha}{2}} = \frac{t_{1-0.9}}{2} = \frac{t_{0.1}}{2} = t_{0.05}$
- s is given: 13
- n is given: 17, meaning that $df = 16$

If we reference $t_{0.05}$ with $df = 16$ on the table from the lesson above, we find: **2.921**

Now we can put it all together:

$$\begin{aligned} \text{Confidence Interval} &= \bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) \\ &= 142.23 \pm 2.921 \left(\frac{13}{\sqrt{17}} \right) \\ &= 142.23 \pm 2.921 \left(\frac{13}{4.123} \right) \\ &= 142.23 \pm 2.921(3.153) \\ \text{Confidence Interval} &= 142.23 \pm 9.210 \end{aligned}$$

Practice

- What is a t -test?
- What conditions indicate the use of a t -test, rather than a z -test?
- How does the shape of a Student's t -distribution differ from a normal distribution?

For questions 4-8, identify the t -score multiplier for the given confidence level:

- 99% CL, one tail, $n = 17$
- 99% CL, two-tail, $df = 9$
- 95% CL, one tail, $n = 22$
- 95% CL, two-tail, $df = 17$
- $\alpha = 0.1, n = 13$, one-tail

Questions 9-12 refer to the following:

The school director at Desiderata School wants to determine if the mean GPA for the entire student body for the current year is above 3.0, with a 95% confidence level. He collects the following sample GPA's, using a SRS: 2.97, 3.21, 3.10, 2.81, 3.35, 4.0, 2.51, 2.38, 3.85, 3.24, 3.81, 3.01, 2.85, 3.4, 2.94.

- What kind of test should he use?
- What are the null and alternative hypotheses?
- What is s ?
- What is \bar{x} ?

13. How many degrees of freedom are there?
14. What is the confidence interval?
15. Should he reject or fail to reject, the null hypothesis?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 10.5.

10.6 Putting it Together

Objective

Here you will practice using the skills you have built to test hypotheses.

Concept

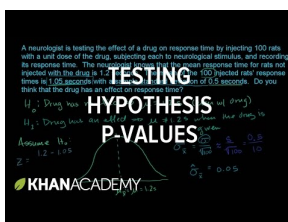
Professor Reitz believes that she has more than 8" of snow on her deck, if she takes the following measurements using a SRS, can she be 95% sure that the mean depth is greater than 8"?

7.92", 8.3", 7.98", 8.12", 8.31", 8.05", 8.27"

Look to the end of the lesson for the answer.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/68441>

<http://youtu.be/-FtIH4svqx4> Khan Academy - Hypothesis Testing and P-values

Guidance

By now, you have learned all you need to know to test some real-life hypotheses. You know how to calculate the mean of means, \bar{x} , the sample standard deviation, $\bar{\sigma}$, and how to reference z and t scores. You are familiar with confidence intervals and confidence levels. You know when to use a z -test, and when to use a t -test, and you know the difference between a null and alternative hypothesis. If are unsure of your grasp of any of these concepts, return to the appropriate lesson and review before moving on, otherwise, let's put it all together!

Just for reference, here are the formulas for calculating confidence intervals of t and z tests:

- T -test confidence interval = $\bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$
- Z -test confidence interval = $\bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$

The steps to test a single-mean hypothesis:

1. Choose an appropriate test:
 1. Z -test if σ is known and if the population is known to be normal or $n \geq 30$.
 2. T -test if σ is unknown and/or $n < 30$.
2. Identify your null and alternative hypotheses
3. Identify or calculate any needed values, probably \bar{x} , s or σ , and n .
4. Find the Z or T score of the sample mean.
5. Draw a curve, mark your critical value(s), and shade α (the rejection region), based on the appropriate number of tails in your test.
6. Decide to reject or fail to reject the null hypothesis.
7. Interpret your results.

Example A

You have been asked to determine if the mean weight of cheese use on the large pizzas at Speedy G's is indeed 5 oz, as it should be. After sampling 87 large pizzas with a SRS, you have calculated $\bar{x} = 4.86$ oz and $s = .27$ oz. Is the mean cheese weight correct? Show your work.



Solution:

Let's use the steps from the lesson as a reference:

1. Since $n \geq 30$, we can use a z -test.
2. The null hypothesis is that the cheese weight is 5 oz. The alternative hypothesis is that the cheese weight is *not* 5oz.

In other words, null $\mu = 5$ oz, alternative $\mu \neq 5$ oz.

3. We are given the values for \bar{x} : 4.86 oz, s : 0.27 oz and n : 87.

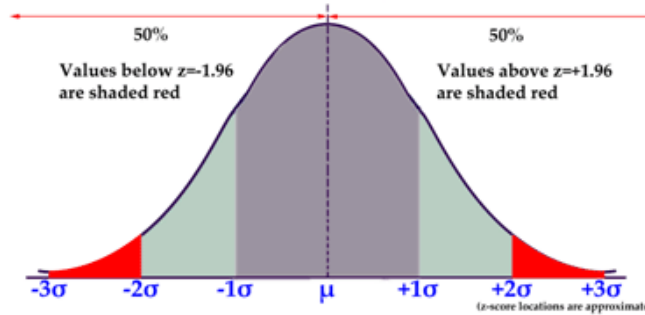
4. The z -score for \bar{x} is $\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ but we do not know σ , the standard deviation for cheese weight company-wide, so we will use our "best guess" of σ , the standard deviation of our sample: $s = .27$ oz:

$$\frac{4.86 - 5}{\frac{0.27}{\sqrt{87}}} = -4.84$$

5. Draw a curve, mark your critical value(s), and shade α (the rejection region), based on the appropriate number of tails in your test.

Since there is no confidence level given, we assume 95%. Since this is a two-tail test, that gives us critical values of $Z = \pm 1.96$. If our sample mean has $Z < -1.96$ or $Z > 1.96$, we may reject the null hypothesis.

The Z-score of -4.84 of our sample mean is clearly well below the lower critical Z-score critical value of -1.96, in fact it would be well off to the left of our entire chart!



6. Decide to reject or fail to reject the null hypothesis.

Since the sample mean has a Z-score of +4.84, clearly not between the critical values of ± 1.96 that we calculated from the 95% confidence level, we **should reject** the null hypothesis.

7. Interpret your results.

Since the data clearly suggests that we reject the null hypothesis, we reject the claim that the mean weight of cheese is 5 oz. If the mean weight of cheese really is 5 oz, our conclusions would have been extremely likely (about 1 in 1.5 million!).

Example B

Rachel collected a SRS of 70 people, asking how many texts they send per day on average, and calculated $\bar{x} = 42.72$. Phone company data suggests that $\mu = 45$, and $\sigma = 8$. Rachel thinks the actual mean is less than the phone company claims. Does Rachel's sample data support her hypothesis?

Solution:

Following the steps as outlined in the lesson:

1. Choose an appropriate test: Since $n \geq 30$, we can use a z-test (one-tail).

2. State the null and alternate hypotheses:

Null: $\mu \geq 45$

Alternative: $\mu < 45$

3. Calculate any needed values:

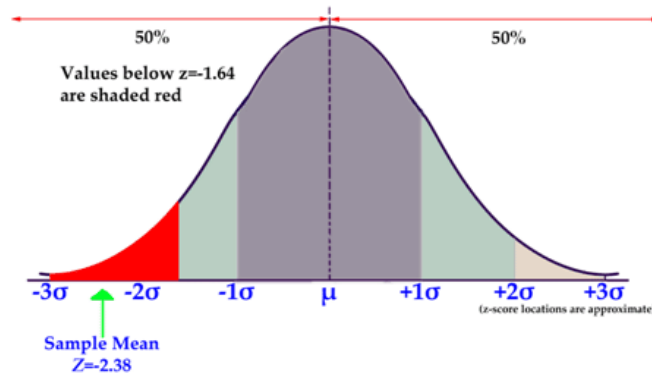
We have what we need

4. Find the Z or T score of the sample mean:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \frac{42.72 - 45}{\frac{8}{\sqrt{70}}} = -\frac{2.28}{8.366} = -2.38$$

5. Draw a curve, mark your critical value(s), and shade α (the rejection region), based on the appropriate number of tails in your test.

6. Since there is no confidence level given, we assume 95%. Since this is a one-tail test, that gives us a critical value of $Z = -1.645$. If our sample mean has $Z < -1.645$, we may reject the null hypothesis.



7. Decide to reject or fail to reject the null hypothesis.

- Since the sample mean has a Z -score of -2.38 , well below the critical value of -1.645 that we calculated from the 95% confidence level, we can **reject the null hypothesis**.

8. Interpret your results.

Since the data clearly suggests that we reject the null hypothesis, we can reject the claim that the mean number of texts per day is ≥ 45 . If the mean number of texts per day were actually 45 or greater, Rachel's conclusions would have a likelihood of 0.86%, well less than 5%.

Example C

A SRS of 51 of cheerleaders, asking how much they spend on a pair of jeans, yields $\bar{x} = \$38.77$. Department store data suggests that the average selling price of a pair of teen girl jeans is \$40, with $\sigma = \$4$. Is it reasonable to reject the claim based on the research?

Solution:

Following the steps as outlined in the lesson:

1. Choose an appropriate test: Since $n \geq 30$, and we also know μ and σ , we can use a z -test (two-tail).
2. State the null and alternate hypotheses:

Null: $\mu = 40$

Alternative: $\mu \neq 40$

3. Calculate any needed values:

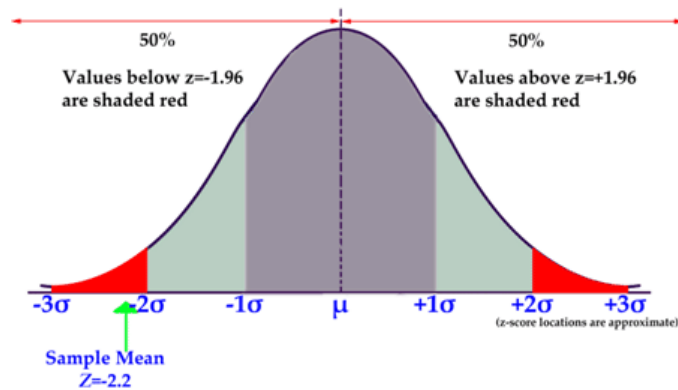
We have everything we need.

4. Find the Z or T score of the sample mean:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \frac{38.77 - 40}{\frac{4}{\sqrt{51}}} = -\frac{1.23}{0.560} = -2.2$$

5. Draw a curve, mark your critical value(s), and shade α (the rejection region), based on the appropriate number of tails in your test.

Since there is no confidence level given, we assume 95%. Since this is a two-tail test, that gives us critical values of $Z = \pm 1.96$. If our sample mean has $Z < -1.96$ or $Z > 1.96$, we may reject the null hypothesis.



6. Decide to reject or fail to reject the null hypothesis.

Since the sample mean has a Z -score of -2.2 , below the critical value of -1.96 that we calculated from the 95% confidence level, we can **reject the null hypothesis**.

7. Interpret your results.

Since the data suggests that we reject the null hypothesis, we can reject the claim that the mean cost of the jeans is actually \$40. If the mean cost of the jeans was actually \$40, our conclusions would have a likelihood of 1.4%, which is less than 5%.

Concept Problem Revisited

Professor Reitz believes that she has more than 8" of snow on her deck, if she takes the following measurements using a SRS, can she be 95% sure that the mean depth is greater than 8"?

7.92", 8.3", 7.98", 8.12", 8.31", 8.05", 8.27"

Technically, no, and the measurements actually do not matter. All she could state is that if she took 100 similar samples (7 data points each), 95 out of 100 of the samples would have a mean within two standard deviations of the actual mean.

Vocabulary

A *hypothesis* is a claim or supposition. In statistics, an *alternative hypothesis* is the initial claim under investigation, commonly defined with the notation H_1 .

A *null hypothesis* is the mutually exclusive corollary to the alternative hypothesis, commonly denoted H_0 .

A *p-value* is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that H_0 is true.

Guided Practice

A simple random sample of 32 homes were polled to see how many birds visited their back yard each summer day, yielding $\bar{x} = 60.72$ birds. The city claims that the nickname, "City of Birds" is appropriate, since homes in the city have an average of 60 or more birds in the back yard each summer day, with a standard deviation of 2.

1. What kind of test is appropriate in this instance?
2. What are the null and alternative hypotheses?
3. What is the value of s ?
4. What is the z -score of \bar{x} ?

5. Assuming 95% CL, does the sampled data support the claim?

Solutions:

1. Since $n \geq 30$, and we know σ , a z -test is appropriate.
2. $H_1 : \mu < 60, H_0 : \mu \geq 60$
3. $s = \frac{\sigma}{\sqrt{n}} \rightarrow \frac{2}{\sqrt{32}} = 0.3536$
4. $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \frac{60.72 - 60}{0.3536} = 2.04$
5. Since the z -score for the 95% of a one-tail test is -1.65, and the z -score of \bar{x} is 2.04, we **cannot reject** the null hypothesis, because we would need a z -score *below* -1.65!

Practice

Questions 1-4 refer to the following:

A simple random sample of 59 third year college students were asked, on average, how much they spent for lunch, resulting in $\bar{x} = \$48.29$. The college claims that the students spend an average of \$50 for lunch with a standard deviation of \$6.

1. Can the claim be viably tested with a Z -test? Why/Why not?
2. What are H_0 and H_1 ?
3. What is the Z -score of \bar{x} ?
4. Can you reject this claim using a significance level of 0.001?

Questions 5-10 refer to the following:

A simple random sample of 77 models were polled to learn how much they spent for a T-shirt, and got \$43.33 on average. The models' manager says that models in his company spend an average \$45 with a standard deviation of \$6. The polling company thinks the actual value is less than the manager claims.

5. Can the claim be viably tested with a Z -test? Why/Why not?
6. What are H_0 and H_1 ?
7. What is the Z -score of \bar{x} ?
8. Can you reject this claim using a significance level of 0.05?
9. Can you say that you are 95% sure that the claim is correct?
10. What can you say, with 95% confidence?

Questions 11-15 refer to the following:

A group of 26 high school seniors chosen via SRS were asked how many texts they make per day on average, yielding $\bar{x} = 42.49$, with $s = 8$. The phone company states that the average high school senior sends an average of 45.

11. Can the claim be viably tested with a Z -test? Why/Why not?
12. What are H_0 and H_1 ?
13. What is the SEM ?
14. How many df are there?
15. What is the T -score multiplier?

16. What is the confidence interval $\bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$?

17. Can you reject this claim using a significance level of 0.05?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 10.6.

Students were introduced to the concepts of hypothesis testing and predicting values. The concept of the null and alternative hypothesis were introduced and practiced. The tails of a curve were described as were critical values and critical regions and their association with alpha values. Students learned about confidence intervals and learned the meaning of confidence level. The similarities between T and Z testing were reviewed and students practiced using both to test hypotheses in word problems.

10.7 References

1. Kevin Dooley. <https://www.flickr.com/photos/pagedooley/3302597972> .
2. bykst. <http://pixabay.com/en/food-bowl-fressnapf-dog-food-281980/?oq=dog%20food> .
3. . . CC BY-NC-SA
4. . . CC BY-NC-SA
5. . . CC BY-NC-SA
6. . . CC BY-NC-SA
7. . . CC BY-NC-SA
8. . . CC BY-NC-SA
9. . . CC BY-NC-SA
10. . . CC BY-NC-SA
11. . . CC BY-NC-SA
12. . . CC BY-NC-SA
13. J.B. Hill. <https://www.flickr.com/photos/jbhill/2747439395> .
14. geralt. <http://pixabay.com/en/women-wall-stones-shadow-225399/?oq=girl%20group> .
15. . . CC BY-NC-SA
16. . . CC BY-NC-SA
17. John Liu. <https://www.flickr.com/photos/8047705@N02/5634141468/> .
18. . . CC BY-NC-SA
19. . . CC BY-NC-SA
20. . . CC BY-NC-SA
21. Jenn Durfey. <https://www.flickr.com/photos/dottiemae/5379273654> .
22. . . CC BY-NC-SA
23. . . CC BY-NC-SA
24. . . CC BY-NC-SA

CHAPTER **11**

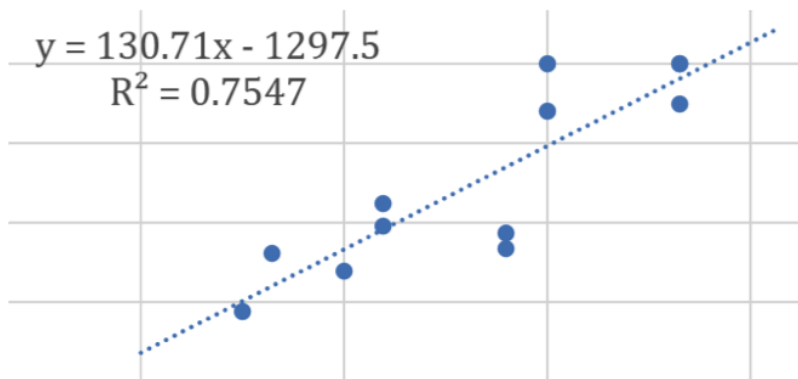
Linear Regression and Chi-Squared

Chapter Outline

- 11.1 LINEAR RELATIONSHIPS
- 11.2 LINEAR CORRELATION COEFFICIENT
- 11.3 LEAST SQUARES
- 11.4 CONTINGENCY TABLES
- 11.5 CHI SQUARED STATISTIC
- 11.6 CHI-SQUARED II - TESTING FOR INDEPENDENCE
- 11.7 REFERENCES

In this chapter, you will learn how to use observed data to predict the value of an observation that you do not have. One way to do this is via **linear regression**, the process of calculating a line that represents the best average rate of change of points in a scatter plot.

You will learn to evaluate a set of data to see if it supports a hypothesized distribution, and you will also learn to create contingency tables to organize information to compare variables and see if they are related. The chi-squared (χ^2) statistic is a value you will learn to use for this purpose.



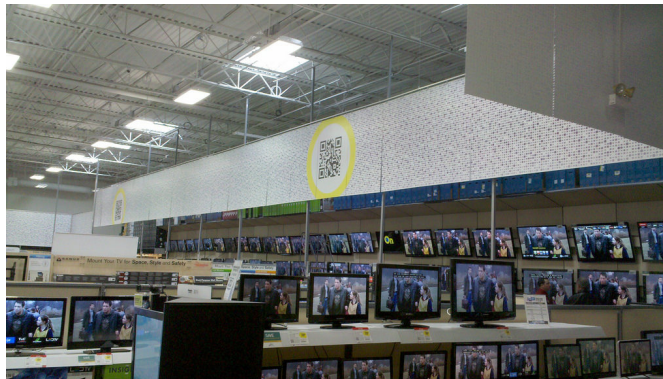
11.1 Linear Relationships

Objective

Here you will learn about pairs of variables that are related in a linear fashion, including those with values occurring in a slightly random manner.

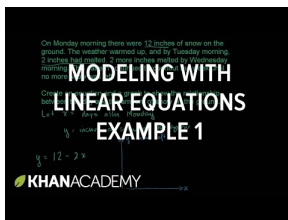
Concept

Imagine walking through the electronics section of your local department store. On the wall are examples of dozens of television sets, from little 19" units made to sit on a kitchen counter to 72"+ monsters meant to be the centerpiece of a home theatre. Looking at the prices, you note without surprise that the 72" model is more expensive than the 19", and a 42" model is priced in between. It seems rather clear that as the TV gets larger, the price goes up. Does that mean increased screen size causes increased price?



Look to the end of the lesson for the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67548>

<http://youtu.be/qPx7i1jwXXX> Khan Academy - Exploring Linear Relationships

Guidance

When two quantities are compared, it is not uncommon to note a relationship between them that indicates both quantities increase and decrease at the same time, or that one increases as the other decreases. If both quantities are plotted on coordinate axes, the data points show a general or definite linear trend.

If the points actually form a clearly defined line, the variables may be an example of a *deterministic* relationship. A deterministic relationship indicates that the value of one variable can be reliably and accurately determined by the manipulation of the other variable. An example might be inches and centimeters: one inch is the same as 2.54 centimeters. If you know how many inches long something is, you can reliably and accurately calculate the number of centimeters long the same item is.

As you likely recall from Algebra, the *slope* describes the angle of the line created by plotting points from a linear relationship, and the point where the *explanatory variable* has a value of zero is called the y-intercept (commonly denoted b).

Often, particularly in research situations when one or both variables are measured, the plotted values are generally linear, but do not line up precisely. When two variables seem to show a linear relationship, but the values display some amount of randomness, we commonly visually describe the relationship with a *scatter plot*. As you will see throughout this chapter, the strength of the linear relationship of the variables can be described through mathematics.

Example A

Given the equation $y = 2.3x + 5$:

- Create an $X - Y$ table to describe the values of at least four points
- What is the slope of the line?
- What is the y-intercept?

Solution:

- Pick a value for x , substitute the chosen value for x in the equation, and calculate y :

TABLE 11.1:

X	calculation	Y
1	$y = 2.3(1) + 5$	7.3
2	$y = 2.3(2) + 5$	9.6
0	$y = 2.3(0) + 5$	5
-1	$y = 2.3(-1) + 5$	2.7

$$y = mx + b$$

The equation in the problem is in $y = mx + b$ form (also known as *slope-intercept form*), where b is the y-value when $x = 0$, and m is the slope of the line. Therefore:

- $m = 2.3$
- $b = 5$

Example B

Given the equation $y = -3x + 3.9$:

- Create an $X - Y$ table to describe the values of at least four points

- What is the slope of the line?
- What is the y-intercept?

Solution:

- Pick a value for x , substitute the chosen value for x in the equation, and calculate y :

TABLE 11.2:

X	calculation	Y
1	$y = -3(1) + 3.9$	6.9
2	$y = -3(2) + 3.9$	9.9
0	$y = -3(0) + 3.9$	3.9
-1	$y = -3(-1) + 3.9$	0.9

The equation in the problem is in $y = mx + b$ form (also known as *slope-intercept form*), where b is the y -value when $x = 0$, and m is the slope of the line. Therefore:

- $m = -3$
- $b = 3.9$

Example C

Given the equation $y = -2.8x - 9.1$:

- Create an $X - Y$ table to describe the values of at least four points
- What is the slope of the line?
- What is the y-intercept?

Solution:

- Pick a value for x , substitute the chosen value for x in the equation, and calculate y :

TABLE 11.3:

X	calculation	Y
1	$y = -2.8(1) - 9.1$	-11.9
2	$y = -2.8(2) - 9.1$	-14.7
0	$y = -2.8(0) - 9.1$	-9.1
-1	$y = -2.8(-1) - 9.1$	6.3

The equation in the problem is in $y = mx + b$ form (also known as *slope-intercept form*), where b is the y -value when $x = 0$, and m is the slope of the line. Therefore:

- $m = -2.8$
- $b = -9.1$

Concept Problem Revisited

Imagine walking through the electronics section of your local department store. On the wall are examples of dozens of television sets, from little 19" units made to sit on a kitchen counter to 72"+ monsters meant to be the centerpiece of a home theatre. Looking at the prices, you note without surprise that the 72" model is more expensive than the 19", and a 42" model is priced in between. It seems rather clear that as the TV gets larger, the price goes up. Does

that mean increased screen size causes increased price?

No, it does not. This is an example of the difficulty associated with examining linear relationships. **Correlation does not imply causation.** Just because a pair of variables exhibit a relationship, linear or otherwise, does not mean that one variable *causes* changes in the other variable.

Vocabulary

A **Cartesian Graph** is a “plus-shaped” graph, with the **explanatory variable** (the input value) plotted horizontally on the x -axis, and the **response variable** (the output value) plotted vertically on the y -axis.

A **deterministic** linear relationship is a relationship that plots a reliably straight and accurate single line.

The **slope** of a line (commonly denoted m) describes the angle of a plotted line on a graph.

A **scatter plot** is a graph of individual points on an $X - Y$ graph.

Guided Practice

1. If a linear graph exhibits a positive slope, what can you predict will happen to the response variable as the explanatory variable increases?
2. If a linear graph has no slope, what does that mean?
3. Given the linear equation $2y = 5.2x + 7$:
 - a. What is the slope?
 - b. What is the y -intercept?
 - c. What happens to y as x increases?
4. Given the equation $y = 2x^2 + 4$:
 - a. Is this a linear equation? Why or why not?
 - b. Does this equation represent a relationship?

Solutions:

1. A positive slope indicates that the variables increase and decrease together.
2. A line with no slope is a horizontal line, since the only defined variable is the output. No matter what value is given for the explanatory variable, the response is the same.
3. a. The slope, m , is 5.2.
b. The y -intercept, b , is 7.
c. Since this line has a positive slope, y increases as x increases.
4. a. No, the explanatory variable is squared, this graph would form a parabola.
b. Yes! It is just not a *linear* relationship.

Practice

For questions 1-8, find the x and y intercepts of the given equations.

1. $-x + 4y = 8$
2. $3x + 5y = 15$

3. $-3x + 4y = 36$

4. $-8x + 5y = 40$

5. $5x - 6y = -30$

6. $-9x - 3y = -54$

7. $-x + 5y = -10$

8. $-3x + 8y = -72$

For questions 9-15, graph the line.

9. $x + 3y = 2$

10. $m = -4, b = \frac{4}{3}$

11. x -intercept = -1 , y -intercept = 2

12. $y = -4x + 2$

13. $m = -1, b = \frac{1}{2}$

14. $x + 2y = 5$

15. $-3x + 2y = -3$

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 11.1.

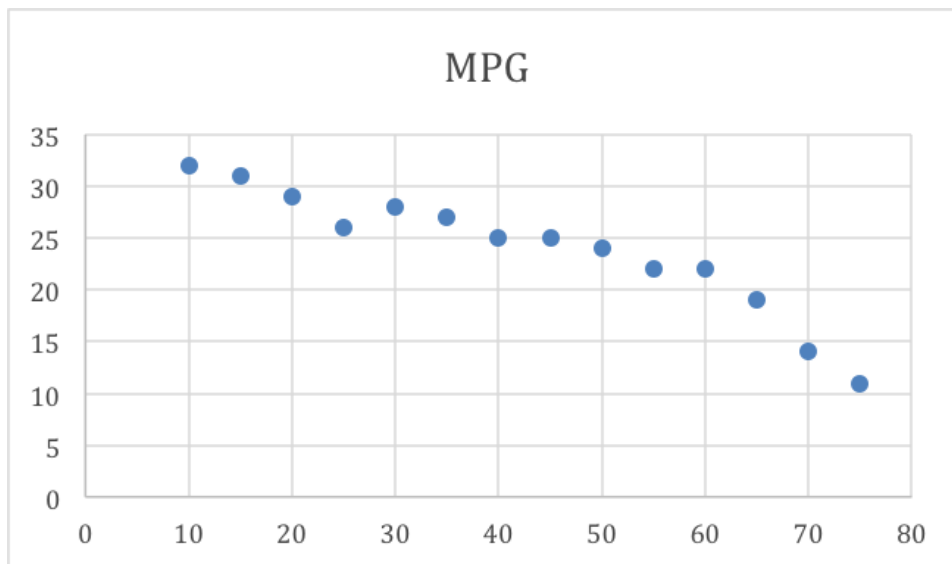
11.2 Linear Correlation Coefficient

Objective

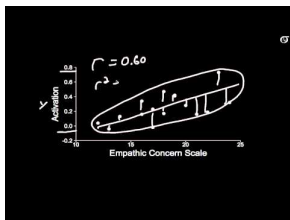
Here you will learn to calculate the *linear correlation coefficient*, and how to use it to describe the relationship between an explanatory and response variable.

Concept

Suppose you have noted that your car seems to use more gas when you drive fast than when you drive more slowly. You decide to see how strong the relationship is, so you do some research, collect the data, and plot the data on the graph below, where the explanatory variable x is mph, and the response variable y is mpg. How can you describe how strong the correlation is without the graph?



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67535>

<http://youtu.be/dCgIavyFWIo> jbststatistics - The Correlation Coefficient and Coefficient of Determination

Guidance

The linear correlation coefficient (sometimes called Pearson's Correlation Coefficient), commonly denoted r , is a measure of the strength of the *linear* relationship between two variables. The value of r has the following properties:

- r is always a value between -1 and +1
- The further an r value is from zero, the stronger the relationship between the two variables.
- The sign of r indicates the nature of the relationship: A positive r indicates a positive relationship, and a negative r indicates a negative relationship.



Generally speaking, you may think of the values of r in the following manner:

- If $|r|$ is between 0.85 and 1, there is a strong correlation.
- If $|r|$ is between 0.5 and 0.85, there is a moderate correlation.
- If $|r|$ is between 0.1 and 0.5, there is a weak correlation.
- If $|r|$ is less than 0.1, there is no apparent correlation.

Naturally, r -value can be calculated, but the formula is a bit beyond the scope of this course. Fortunately, there are many excellent and free online calculators for determining the r -value of a set of data. In this lesson, I will be using the one at <http://easycalculation.com/statistics/correlation.php>, but a search for “correlation calculator online” will yield the most current options.

At the risk of overloading you with new terms, there is one more that I think it is worth learning in this lesson, the **coefficient of determination**. The coefficient of determination is very simple to calculate if you know the correlation coefficient, since it is just r^2 . The reason I mention it is that the coefficient of determination can be interpreted as the percentage of variation of the y variable that can be attributed to the relationship. In other words, a value of $r^2 = .63$ can be interpreted as “63% of the changes between one y value and another can be attributed to y 's relationship with x ”.

Example A

Elaina is curious about the relationship between the weight of a dog and the amount of food it eats. Specifically, she wonders if heavier dogs eat more food, or if age and size factor in. She works at the Humane Society, and does some research. After some calculation, she determines that dog weight and food weight exhibit an r -value of 0.73.



What can Elaina say about the relationship, based on her research? What percentage of the increases in food intake can she attribute to weight, according to her research?

Solution:

The calculated r -value of 0.73 tells us that Elaina's data demonstrates a moderate to strong correlation between the variables.

Since the coefficient of determination tells us the percentage of changes in the output variable that can be attributed to the input variable, we need to calculate r^2 :

$$r^2 = (0.73)^2 = .5329$$

Approximately 53% of increases in food intake can be attributed to the linear relationship between food intake and the weight of the dog, suggesting that other factors, perhaps age and size, are also involved.

Example B

Tuscany wonders if barrel racing times are related to the age of the horse. Specifically, she wonders if older horses take longer to complete a barrel racing run. As a member of the Pony Club, she does some research, and determines that horse age to barrel run time exhibits an r -value of 0.52.

What can Tuscany say about horse age vs barrel race time, according to her research?

Solution:

Tuscany's research suggests that there is a moderate to weak correlation between horse age and barrel run time. In other words, the research suggests that $(0.52)^2 = .27 = 27\%$ of the differences between barrel run times could be attributable to the linear relationship between barrel run time and the age of the horse.

Example C

Sayber has collected the following data regarding player score vs age in his favorite online game. He suspects that increased age is not a good indicator of gaming ability. What are the linear correlation coefficient and coefficient of determination values of his data, and how do they support or not support Sayber's hypothesis?

TABLE 11.4:

Age	Avg. Player Score
12	5,120
14	6,328
18	7,892

TABLE 11.4: (continued)

22	7,340
28	6,987
34	7,750
42	5,421

Solution:

Let's use the online calculator at easycalculation.com for this one.

I entered the explanatory (Age) and response (Player Score) values into the calculator:

To Calculate Correlation Co-efficient:

X Value	Y Value
12	5120
14	6328
18	7892
22	7340
28	6987
34	7750
42	5421

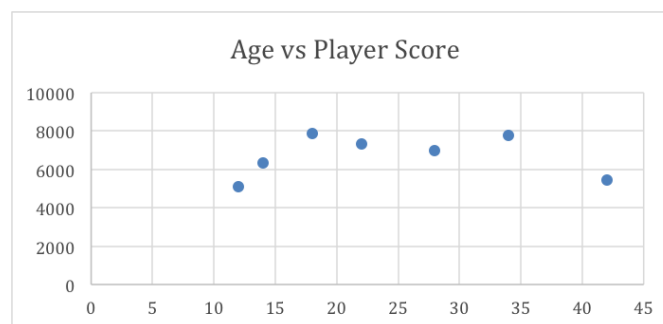
[Add More..](#) [Fewer..](#)

Results:

Total Numbers :	7
Correlation :	0.03983186004487671

The linear correlation coefficient of approximately 0.04 suggests that there is no appreciable linear correlation. The coefficient of determination of 0.0016 suggests that perhaps 0.16% (practically none) of the variability of the player score is dependent on age.

Looking at the scores, however, something seems a miss with our findings. The scores suggest that age has no bearing on player score, but look at the graph of the same data:



The graph suggests that the youngest and oldest polled players score less than players in late teens to mid-thirties, which seems reasonable.

This is an important example of the weakness of using just one indicator of the relationship between two variables. As I noted early in the lesson, the r -value is only an indicator of *linear* correlation, it says nothing at all about other kinds of variable relationships. It is always a good idea to review your data in different ways to evaluate your initial conclusions.

Concept Problem Revisited

Suppose you have noted that your car seems to use more gas when you drive fast than when you drive more slowly. You decide to see how strong the relationship is, so you do some research, collect the data, and plot the data on the graph below, where the explanatory variable x is mph, and the response variable y is mpg. How can you describe how strong the correlation is without the graph?

Calculate Correlation Co-efficient:

X Value	Y Value
10	32
15	31
20	29
25	26
30	28
35	27
40	25
45	25
50	24
55	22
60	22
65	19
70	14
75	11

[Add More..](#) [Fewer..](#)

Results:

Total Numbers :	14
Correlation :	-0.9425121291679348

After the lesson above, we know that the r -value or r^2 -value of the relationship between MPG and MPH would describe the strength of the linear relationship in a single value.

By taking the data points detailed on the graph (in practice, of course, I would have had them in table format already, since I would have needed them to *build* the graph in the first place), and entering them into a free [linear coefficient calculator](#) online, I get an **r -value of -0.943** , indicating a strong negative relationship. This also translates into an **r^2 -value of $(-0.943)^2 = 0.89$** , indicating that the research suggests that approximately 89% of the *decrease* in MPG from left to right across the graph can be attributed to the *increase* in MPH.

Vocabulary

A **linear correlation coefficient** or ***r*-value** of a relationship between two variables describes the strength of the linear relationship.

The **coefficient of determination** or ***r*²value** of a relationship indicates the approximate percentage of variation in the response variable that can be attributed to the linear relationship between the response and explanatory variables, according to the data presented.

Guided Practice

1. What can you say about the strength of a linear relationship with an *r*-value of -0.87?
2. What can you say about the level of negative correlation of a relationship if you know the coefficient of determination is 0.82?
3. How much of the variability of *y* is attributable to *x* in a relationship with an *r*-value of 0.76?

Solutions:

1. An $|r|$ of >0.85 indicates a strong linear relationship. The fact that *r* is negative indicates that as *x* increases, *y* decreases.
2. Nothing! The coefficient of determination is r^2 , and therefore always positive. We know that $|r| = \sqrt{.82} \approx .91$, so this is a strong linear correlation, but we have no idea if it is positive or negative.
3. The coefficient of determination describes the variation in *y* attributable to *x*, so we need to find r^2 : $(0.76)^2 = .5776$. Approximately 57.76% of the change in *y*-values can be attributed to the change in *x*.

Practice

For questions 1-5, describe the relationship based on the *r*-value.

1. $r = 0$
2. $r = 0.91$
3. $r = -0.49$
4. $r = 0.05$
5. $r = 1$

For questions 6-10, describe the relationship based on the coefficient of determination:

6. $r^2 = 0.82$
7. $r^2 = 0.15$
8. $r^2 = 0.47$
9. $r^2 = 1$
10. $r^2 = 0$

Questions 11-15 refer to the data in the following table:

TABLE 11.5:

<i>X</i>	<i>Y</i>
5	70
7	69

TABLE 11.5: (continued)

13	58
22	47
36	36
38	25
45	14

11. What is the linear correlation coefficient of the data?
12. What does r tell you about the relationship?
13. What is the r^2 value of the data?
14. What does the coefficient of determination tell you about this relationship?
15. What would a graph of the data look like?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 11.2.

11.3 Least Squares

Objective

Here you will learn about lines of best fit, and how to calculate such a line using the least squares method.

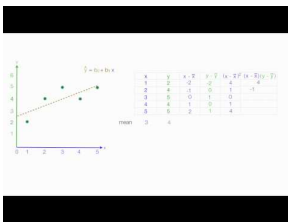
Concept

You have calculated the linear correlation coefficient value $r = -0.9542$ to determine that there is a strong linear relationship between the speed you drive and the gas you use. Suppose you want to predict the average MPG at a speed that you did not measure, like 85 mph. How could you use your research data (below) to provide an estimated value?

TABLE 11.6:

MPG	25	24	22	22	19	14	11
MPH	45	50	55	60	65	70	75

Watch This



MEDIA

Click image to the left or use the URL below.

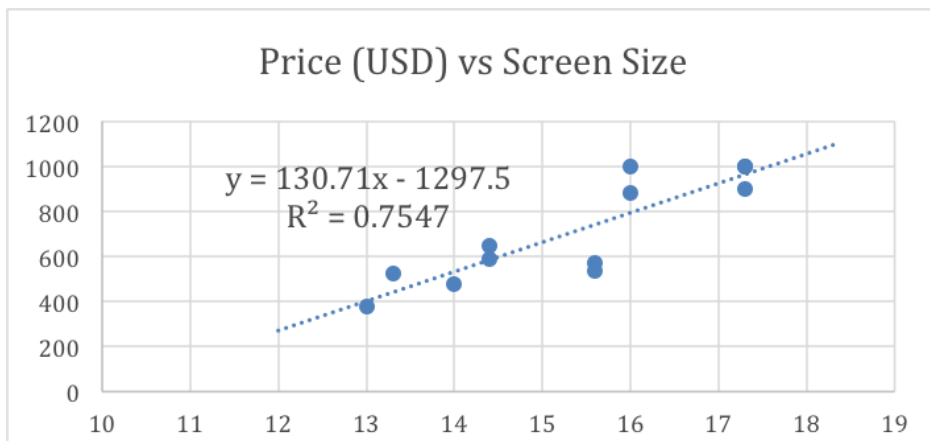
URL: <https://www.ck12.org/flx/render/embeddedobject/67541>

<http://youtu.be/JvS2triCgOY> Statisticsfun - How to calculate linear regression using least square method

Guidance

The linear correlation coefficient r and coefficient of determination r^2 can assist in determining the strength of a linear relationship between two variables, but are not helpful if you need to predict a value that has not been observed. In order to *predict* a value of a relationship, we need to find a line that represents the best average change in y based on change in x (you should recognize this as the slope of a line).

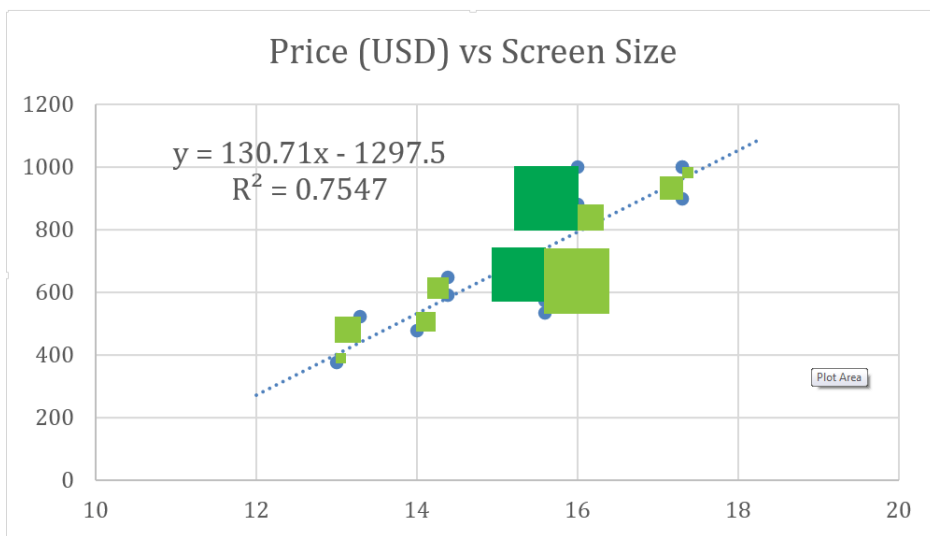
Consider the graph below:



This scatterplot represents the price of laptops online related to screen size. The data points exhibit a clear positive trend, but are certainly not in a straight line. The r^2 value of 0.755 suggests that approximately 75% of the increase in price as we move left to right may be attributed to the relationship between price and increased screen size.

The dotted line $y = 130.71x - 1297.5$, drawn on the graph, is the **line of best fit**, calculated using the **least squares method**. The line of fit may be used to *predict* the likely values of price (y), based on screen size (x), for sizes not plotted on the graph. We can see just by looking that the estimated average price for a laptop with an 18" screen size is about \$1,075, and the estimated average price for a laptop with a 12" screen is less than \$300.

To understand what is meant by the least squares method, imagine a square formed from the distance of each point to the line. The square has four equal sides each equal to the shortest *vertical* distance between each point and the line, as illustrated below:



The line of best fit is the line resulting in the least total area of these squares. The line will always go through the point at the intersection of the mean x and mean y values: (\bar{x}, \bar{y}) .

To identify the equation of the least squares line, you will need the following values calculated from your data:

- \bar{x} : The mean value of x .
- σ_x or s_x : The standard deviation of the x values
- \bar{y} : The mean value of y .
- σ_y or s_y : The standard deviation of the y values
- r : The linear correlation coefficient

Once you know this data, you can find the equation of the line of best fit in slope-intercept form $Y = bX + a$, with two easy formulas:

$$b = r \left(\frac{s_y}{s_x} \right)$$

$$a = \bar{Y} - b\bar{X}$$

The least squares line is a powerful tool for predicting values, but the reliability of the predictions reduces as you get further from the observed data. Take care to recognize that *any* data point that you predict beyond the observed data carries an element of uncertainty that increases the further “out” you predict.

Example A

Find the equation of the line of best fit given $\bar{X} = 13$, $\bar{Y} = 6$, $s_x = 4$, $s_y = 1.5$, and $r = 0.65$.

Solution:

To find the equation of the line, we need to calculate b , and a to substitute into the equation $Y = bX + a$:

- $b = r \left(\frac{s_y}{s_x} \right) \rightarrow b = 0.65 \left(\frac{1.5}{4} \right) = 0.65 \times 0.375 \approx 0.244$
- $a = \bar{Y} - b\bar{X} \rightarrow a = 6 - 0.244(13) = 6 - 3.172 \approx 2.83$

Now we have a and b , we just substitute them into the equation:

$$Y = 0.244X + 2.83$$

Example B

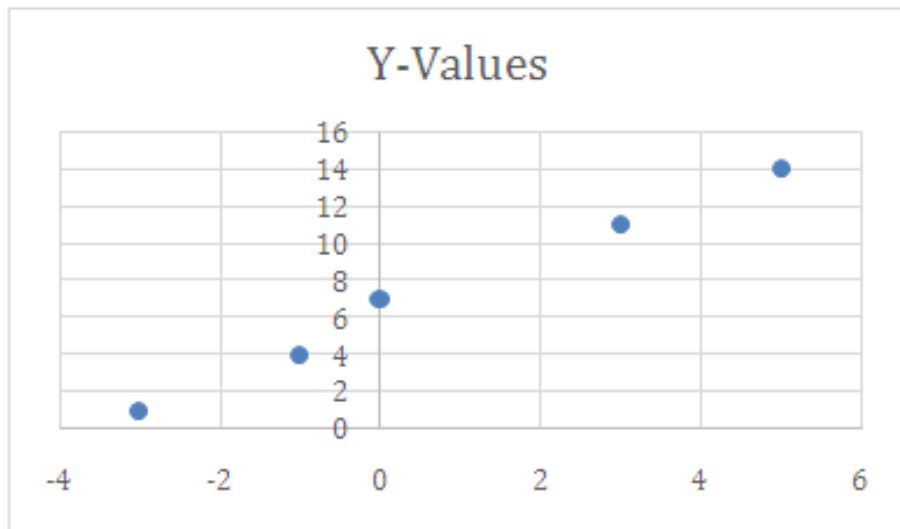
Graph and find the equation of the line of best fit given the data from the following table:

TABLE 11.7:

X	Y
-3	1
-1	4
0	7
3	11
5	14

Solution:

Let's graph the data first, using a scatter plot:



Now we need to calculate our required values, as listed in the lesson:

- $\bar{X} : \frac{-3-1+0+3+5}{5} = 0.80$
- $\sigma_x \text{ or } s_x := \sqrt{\left(\frac{(-3-0.8)^2 + (-1-0.8)^2 + (0-0.8)^2 + (3-0.8)^2 + (5-0.8)^2}{5-1} \right)} = 3.19$
- $\bar{Y} : -7.4$
- $\sigma_y \text{ or } s_y := 5.225$
- $r := 0.9948$ (calculated using the online tool at nccalculators.com)

Now we can use our two new formulas to find a and b :

$$b = r \left(\frac{s_y}{s_x} \right) \rightarrow b = 0.9948 \left(\frac{5.225}{3.19} \right) \approx 0.9948(1.638) \approx 1.63$$

$$a = \bar{Y} - b\bar{X} \rightarrow -7.4 - 1.63(0.80) = -7.4 - 1.304 = -8.704$$

The equation of the line of best fit is: $Y = 1.63X - 8.704$

Example C

Pandi is shopping for rice at her local supermarket, and notes that the rice comes in different size packages, 6 oz for \$1.75, 12 oz for \$3.50, 18 oz for \$4.68, 28 oz for \$7.90, 44 oz for \$13.09 and the “Family Size” package of 64 oz that has no price listed. If Pandi wants the Family Size package, what price would you predict it sells for?

Solution:

First we need to find the equation of best fit, just as we did in Example B.

Necessary values:

- $\bar{X} : 21.6$
- $\sigma_x \text{ or } s_x := 14.9265$
- $\bar{Y} : 6.184$
- $\sigma_y \text{ or } s_y := 4.4655$
- $r := 0.9978$ (Calculated online at nccalculators.com)

Finding a and b :

$$b = r \left(\frac{s_y}{s_x} \right) \rightarrow b = 0.9978 \left(\frac{4.4655}{14.9265} \right) \approx 0.299$$

$$a = \bar{Y} - b\bar{X} \rightarrow 6.184 - 0.299(21.6) = -0.2744$$

The equation of the line of best fit is: $Y = 0.299x - 0.2744$

Now that we have an equation for the line of fit, we can just substitute in the x -value of 64 oz to calculate a predicted price:

$$Y = 0.299(64) - 0.2744$$

$$Y = 19.135 - 0.2744$$

$$Y \approx \$18.86$$

Concept Problem Revisited

You have calculated the linear correlation coefficient value $r = -0.9542$ to determine that there is a strong linear relationship between the speed you drive and the gas you use. Suppose you want to predict the average MPG at a speed that you did not measure, like 85 mph. How could you use your research data (below) to provide an estimated value?

TABLE 11.8:

MPG	25	24	22	22	19	14	11
MPH	45	50	55	60	65	70	75

This is very much like Example C, we need to find the equation of best fit, and then substitute 85 mph in for X to find the predicted MPG.

Let's use the online tool at ncalculators.com to find our necessary data:

- \bar{X} : 60
- σ_x or s_x := 10.8012
- \bar{Y} : 19.57
- σ_y or s_y := 5.2554
- r := -0.9542
- $b = r \left(\frac{s_y}{s_x} \right) \rightarrow b = -0.9542 \left(\frac{5.2554}{10.8012} \right) \approx -0.464$
- $a = \bar{Y} - b\bar{X} \rightarrow 19.57 + 0.464(60) = 47.41$

The equation of the line of best fit is: $Y = -0.464X + 47.41$

Substituting 85 in for X yields:

$$Y = -0.464(85) + 47.41$$

$$Y = 7.97$$

At 85 mph, we predict the fuel efficiency to be 7.97 mpg.

Vocabulary

The **line of best fit** on a scatterplot is a draw line that best represents the trend of the data. If calculated using the **least squares method**, the line represents the least possible average squared vertical difference from the mean for all values.

Guided Practice

Questions 1-5 use the following data:

$$x_1 = 14.26, x_2 = 12.82, x_3 = 11.29, x_4 = 10.02, \text{ and } x_5 = 9.71$$

$$y_1 = 29.43, y_2 = 34.92, y_3 = 40.29, y_4 = 46, \text{ and } y_5 = 49.78$$

1. What are the μ and σ values for x and y ?
2. What is the linear correlation coefficient?
3. What is the equation of the line of best fit?
4. What would you predict y_6 to be, if $x_6 = 7.42$?
5. What would you predict y_0 to be, if $x_0 = 16.28$?

Solutions:

$$1. \mu_x = 11.62, \mu_y = 40.084, \sigma_x = 1.918, \text{ and } \sigma_y = 8.204$$

$$2. r = -0.9906$$

$$3. b = -0.9906 \left(\frac{8.204}{1.918} \right) = -4.237$$

$$a = 40.084 + 4.237(11.62) = 89.318$$

$$Y = -4.237X + 89.318$$

$$4. Y_6 = -4.237(7.42) + 89.318$$

$$Y_6 = 57.88$$

$$5. Y_0 = -4.237(16.28) + 89.318$$

$$Y_0 = 20.34$$

Practice

1. What does the symbol μ_x represent in the context of the lesson?
2. What does the symbol σ_y represent in the context of the lesson?
3. What does the symbol r represent in the context of the lesson?
4. As compared to the standard slope-intercept form of an equation that you likely first learned about in Algebra I, what do a and b represent?
5. What is a line of best fit?
6. What is meant by referring to the 'least squares'?

Questions 7-11 refer to the following data:

$$x_1 = 2, x_2 = 4, x_3 = 5, x_4 = 8, \text{ and } x_5 = 11$$

$$y_1 = -4, y_2 = -7, y_3 = -9, y_4 = -12, \text{ and } y_5 = -16$$

7. What are the μ and σ values for x and y ?
8. What is the linear correlation coefficient?
9. What is the equation of the line of best fit?

10. What would you predict y_6 to be, if $x_6 = 14$?

11. What would you predict y_0 to be, if $x_0 = 0$?

Questions 12-16 refer to the following:

Brian wonders if more expensive skis slide better, and collects the following data:

TABLE 11.9:

Ski Cost in USD	Sliding Coefficient
237.43	0.06
283.92	0.056
343.50	0.05
373.89	0.049
422.99	0.051
487.50	0.05
505.24	0.046

12. What are the μ and σ values for x and y ?

13. What is the linear correlation coefficient?

14. What is the equation of the line of best fit?

15. What would you predict the sliding coefficient of a \$650 pair of skis to be?

16. How would you describe the relationship between price and sliding coefficient, based on Brian's data?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 11.3.

11.4 Contingency Tables

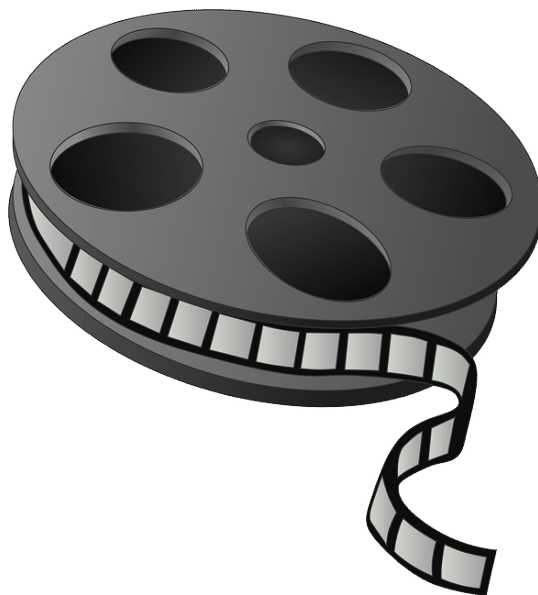
Objective

Here you will learn how to create and use *contingency tables*.

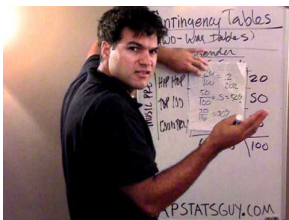
Concept

Suppose you wanted to evaluate how gender affects the type of movie chosen by movie-goers, how might you organize data on Male and Female watchers, and Action, Romance, Comedy, and Horror movie types, so it would be easy to compare different combinations?

See the end of the lesson where this question is reviewed.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67536>

<http://youtu.be/sYkt4rXif78> MrNystrom - Contingency Table Intro

Guidance

Contingency tables are used to evaluate the interaction of statistics from two different categorical variables. They are often used to organize data from different random variables in preparation for a contingency test (which we will be discussing further in the next lesson).

Contingency tables are sometimes called **two-way tables** because they are organized with the outputs of one variable across the top, and another down the side. Consider the table below:

TABLE 11.10:

	Male	Female
Chocolate Candy	42	77
Fruit Candy	58	23

This is a contingency table comparing the variable 'Gender' with the variable 'Candy Preference'. You can see that, across the top of the table are the two gender options for this particular study: 'male students' and 'female students'. Down the left side are the two candy preference options: 'chocolate' and 'fruit'. The data in the center of the table indicates the reported candy preferences of the 100 students polled during the study.

Commonly, there will be one additional row and column for totals, like this:

TABLE 11.11:

	Male	Female	TOTAL
Chocolate Candy	42	77	119
Fruit Candy	58	23	81
TOTAL	100	100	200

Notice that you can run a quick check on the calculation of totals, since the "total of totals" should be the same from either direction: $119 + 81 = 200 = 100 + 100$.

The benefits of a contingency square will be apparent the more you use it. As you begin to evaluate different bits of information, each combination of variable outputs is easily noted.

Example A

Construct a contingency table to display the following data: "250 mall shoppers were asked if they intended to eat at the in-mall food court or go elsewhere for lunch. Of the 117 male shoppers, 68 intended to stay, compared to only 62 of the 133 female shoppers".

Solution:

First, let's identify our variables and set up the table with the appropriate row and column headers.

The variables are gender and lunch location choice:

TABLE 11.12:

	Male	Female	TOTAL
<i>Food Court</i>			
<i>Out of Mall</i>			
TOTAL			

Now we can fill in the values we have directly from the text:

TABLE 11.13:

	<i>Male</i>	<i>Female</i>	<i>TOTAL</i>
<i>Food Court</i>	68	62	
<i>Out of Mall</i>			
<i>TOTAL</i>	117	133	250

Now we can fill in the missing data with simple addition/subtraction:

TABLE 11.14:

	<i>Male</i>	<i>Female</i>	<i>TOTAL</i>
<i>Food Court</i>	68	62	130
<i>Out of Mall</i>	49	71	120
<i>TOTAL</i>	117	133	250

Example B

Referencing data from Example A, answer the following:

- What percentage of food-court eaters are female?
- What is the distribution of male lunch-eaters?
- What is the *marginal distribution* of the variable “lunch location preference”?
- What is the marginal distribution of the variable “Gender”?
- What percentage of females prefer to eat out?

Solutions:

- If we read across the row “Food Court”, we see that there were a total of 130 shoppers eating “in”, and that 62 of them were female. To calculate percentage, we simply divide: $\frac{62}{130} \approx .477$ or **47.7%**.
- The male shoppers were distributed as **68 food court and 49 out of mall**.
- The *marginal distribution* is the distribution of data “in the margin”, or in the TOTAL column. In this case, we are interested in the data on lunch location preference, which is found in the far right column: **130 food court and 120 out of mall**.
- The marginal distribution of gender can be found in the bottom row: **117 males and 133 females**.
- Here we are interested in data from the females, so we will be dealing with the ‘female’ column. From the data in the column, we see that 71 of the 133 females preferred to eat out. This is a percentage of: $\frac{71}{133} \approx .534$ or **53.4%**.

Example C

Using the given data:

- Construct a contingency table
- Identify the marginal distributions
- Identify 3 different percentage-based observations

“Out of 213 polled amateur drag racers, 47 drove cars with turbo-chargers, 59 had superchargers, and the rest were normally aspirated. The racers themselves were split between 102 rookies and 111 veterans. The rookies evidently preferred turbos, since 29 of them had turbo-charged vehicles, and avoided superchargers, since there were only 12 of them”.

**Solutions:**

a. Set up the table with the appropriate headers, and fill in the data we know. Note that this time we will need a 3×2 table instead of a 2×2 (it is still a two- way table though, as there are only two variables: engine aspiration and driver experience):

TABLE 11.15:

	Turbocharger	Supercharger	Normal Aspiration	TOTAL
Rookie	29	12		102
Veteran				111
TOTAL	37	59	117	213

Now we can update the table with the missing data, calculated using addition or subtraction:

TABLE 11.16:

	Turbocharger	Supercharger	Normal Aspiration	TOTAL
Rookie	29	12	61	102
Veteran	8	47	56	111
TOTAL	37	59	117	213

b. The marginal data refers to the overall data for each of the two variables:

- Aspiration type is distributed as follows: **37 Turbos, 59 Superchargers, and 117 normally aspirated.**
- Driver experience distribution: **102 Rookies and 111 Veterans.**

c. Three percentage-based observations:

- $\frac{61}{102} = 0.598$ or 59.8% of Rookies drive normally aspirated cars.
- $\frac{47}{59} = 0.7966$ or 79.66% of the Superchargers were in cars driven by Veterans.
- $\frac{47}{111} = 0.4234$ or 42.34% of Veterans use Superchargers.

Concept Problem Revisited

Suppose you wanted to evaluate how gender affects the type of movie chosen by movie-goers, how might you organize data on Male and Female watchers, and Action, Romance, Comedy, and Horror movie types, so it would be easy to compare different combinations?

A contingency table would be excellent for this purpose. By listing gender categories in one direction and movie type in the other, it would be a simple matter to evaluate different combinations of variables.

Vocabulary

A *contingency table* or *two-way* table is used to organize data from multiple categories of two variables so that various assessments may be made.

A *marginal distribution* is the distribution of data “in the margin” of a table. It may also be described as the distribution of the data for a single variable.

Guided Practice

1. Complete the data in the contingency table:

TABLE 11.17:

	A	B	TOTAL
X	47		
Y		32	100
TOTAL	100		200

2. What is the marginal distribution of the variable consisting of categories A and B?

3. What percentage of B's are Y's?

4. What portion of A's are X's? Express your answer as a decimal.

Solutions:

1.

TABLE 11.18:

	A	B	TOTAL
X	47	$100 - 47 = 53$	$200 - 100 = 100$
Y	$100 - 47 = 53$	32	100
TOTAL	100	$200 - 100 = 100$	200

2. The variable consisting of categories A and B is distributed as A: 100 and B: 100.

3. There are 32 B's that are also Y's, out of the total of 100 B's: $\frac{32}{100} = 32\%$

4. 47 of the 100 A's are X's, $\frac{47}{100} = 0.47$

Practice

Questions 1-9 refer to the following table:

TABLE 11.19:

	Sports Cars	Pickup Trucks	Luxury Cars	TOTAL
Male Drivers	72	67	36	175
Female Drivers	36	71	68	175

TABLE 11.19: (continued)

TOTAL	108	138	104	350
--------------	-----	-----	-----	-----

1. What is the marginal distribution of vehicle types?
2. What is the marginal distribution of driver gender?
3. What decimal portion of male drivers have luxury cars?
4. What percentage of female drivers have pickups?
5. How many drivers were polled?
6. What is the overall most popular vehicle type, by percentage?
7. Which vehicle type has the single largest cell value, and what percentage does it represent of that gender category?
8. What percentage of pickup trucks are driven by females?
9. What percentage of females drive pickup trucks?

Questions 10-18 refer to the following data:

“One hundred eighty dogs were studied to determine if breed affected food preference. Of the 70 Huskies, 30 preferred beef flavor and 40 preferred chicken. Of the 50 Poodles, 27 preferred beef, the rest chicken. The rest of the dogs, English Mastiffs, were obviously beef-lovers, as only 19 preferred chicken over beef”.

10. Create a contingency table to display the data.
11. What is the marginal distribution of dog breeds?
12. What is the marginal distribution of food types?
13. What percentage of Mastiffs preferred beef?
14. What percentage of beef-lovers were Mastiffs?
15. What flavor/dog combination indicated the strongest preference? What percentage of the breed did it represent?
16. What is the distribution of chicken preference?
17. What is the distribution of beef preference?
18. Which breed shows the least defined preference, as a percentage?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 11.4.

11.5 Chi Squared Statistic

Objective

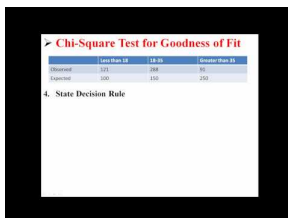
Here you will learn how to use a Chi-Squared statistic to evaluate the fit of a hypothesized distribution. This is known as a Goodness of Fit test.

Concept

Suppose you wanted to evaluate a recent statistic stating that iOS represents 32% and Android 51% of active smart phones. You would like to know if the statistic actually reflects the distribution of phones among your friends. How could you evaluate the data you collect to see if it supports this hypothesis?

Look to the end of the lesson for the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67547>

Guidance



The Greek letter “chi”, written as χ , is the symbol used to identify a *chi-square statistic*, which we will use here to evaluate how well a set of observed data fits a corresponding expected set.

Conducting a *Chi-Square test* is much like conducting a *Z-test* or *T-test* as we did in Chapter 10. We will follow the same basic series of steps and compare a calculated value to a chart to evaluate the probability of getting the results we have if the null hypothesis is true, just as we did with the *Z* and *F* tests. Additionally, as was the case with the *F*-testing, we will be evaluating the number of *degrees of freedom*, and choosing values from a chart based on the number.

The primary difference between a Chi-Square test and the tests we have work with before is that previous tests have all been primarily dedicated to comparing single parameters, whereas Chi-Square tests are used to determine if two random variables are independent or related and so deal with multiple values for each variable. Additionally, the Chi-Square statistic is useful for looking at categorical data rather than quantitative data.

The Chi-Square statistic is actually pretty straightforward to calculate:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Example A

The *American Pet Products Association* conducted a survey in 2011 and determined that 60% of dog owners have only one dog, 28% have two dogs, and 12% have three or more. Supposing that you have decided to conduct your own survey and have collected the data below, determine whether your data supports the results of the *APPA* study. Use a significance level of 0.05.

Data: Out of 129 dog owners, 73 had one dog and 38 had two dogs.

Solution:

- **Step 1: Clearly state the null and alternative hypotheses**

H_0 : The survey agrees with the sample.

H_1 : The survey does not agree with the sample.

- **Step 2: Identify an appropriate test and significance level**

Since we are comparing two sets of data, and not just a single value, a Chi-Square test is appropriate. In the absence of a stated significance level in the problem, we assume the default 0.05.

- **Step 3: Analyze sample data**

Create a table to organize data and compare the observed data to the expected data:

TABLE 11.20:

	One Dog	Two Dogs	3+ Dogs	TOTAL
Observed	73	38	18	129
Expected				

To identify the expected values, multiply the expected % by the total number observed:

TABLE 11.21:

	One Dog	Two Dogs	3+ Dogs	TOTAL
Observed	73	38	18	129
Expected	$0.60 \times 129 = 77.4$	$0.28 \times 129 = 36.1$	$0.12 \times 129 = 15.5$	129

To calculate our chi-square statistic, we need to sum the squared difference between each observed and expected value divided by the expected value:

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ \chi^2 &= \frac{(73 - 77.4)^2}{77.4} + \frac{(38 - 36.1)^2}{36.1} + \frac{(18 - 15.5)^2}{15.5} \\ \chi^2 &= \frac{(-4.4)^2}{77.4} + \frac{(1.9)^2}{36.1} + \frac{(2.5)^2}{15.5} \\ \chi^2 &= \frac{19.36}{77.4} + \frac{3.61}{36.1} + \frac{6.25}{15.5} \\ \chi^2 &= 0.2501 + 0.1000 + 0.4032 \\ \chi^2 &= 0.7533\end{aligned}$$

Now that we have our chi-square statistic, we need to compare it to the chi-square value for the significance level 0.05. We can use a reference table such as the one below, or a [chi-square value calculator](#). Just as with the T -tests in Chapter 10, we will need to know the **degrees of freedom**, which equal the number of observed category values minus one. In this case, there are three category values: one dog, two dogs, and three or more dogs. The degrees for freedom, therefore, are $3 - 1 = 2$.

Using the calculator or the table, we find that the critical value for a 0.05 significance level with $df = 2$ is 5.9915. That means that 95 times out of 100, a survey that agrees with a sample will have a χ^2 critical value of 5.9915 or less. If our chi-square value is *greater* than 5.9915, then the measurements we took only occur 5 or fewer times out of 100, or the null hypothesis is incorrect. Our chi-square statistic is only **0.7533**, so we will **not reject** the null hypothesis.

- **Step 4: Interpret the results**

Since our chi-square statistic was less than the critical value, we do not reject the null hypothesis, and we can say that our survey data does support the data from the APPA.

Example B

Rachel told Eric that the reason her car insurance is less expensive is that female drivers get in fewer accidents than male drivers. Specifically, she says that male drivers are held responsible in 65% of accidents involving drivers under 23.



If Eric does some research of his own and discovers that 46 out of the 85 accidents he investigates involve male drivers, does his data support Rachel's hypothesis?

Solution:

- **Step 1: Clearly state the null and alternative hypotheses**

H_0 : The survey agrees with the sample.

H_1 : The survey does not agree with the sample.

- **Step 2: Identify an appropriate test and significance level**

Since we are comparing two sets of data, and not just a single value, a Chi-Square test is appropriate. In the absence of a stated significance level in the problem, we assume the default 0.05.

- **Step 3: Analyze sample data**

Create a table to organize data and compare the observed data to the expected data:

TABLE 11.22:

	Male Drivers	Female Drivers	TOTAL
Observed	46	39	85
Expected			

To identify the expected values, multiply the expected % by the total number observed:

TABLE 11.23:

	Male Drivers	Female Drivers	TOTAL
Observed	46	39	85
Expected	$0.65 \times 85 = 55.25$	$0.35 \times 85 = 29.75$	85

To calculate our chi-square statistic, we need to sum the squared differences between each observed and expected value divided by the expected value:

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ \chi^2 &= \frac{(46 - 55.25)^2}{55.25} + \frac{(39 - 29.75)^2}{29.75} \\ \chi^2 &= \frac{(-9.25)^2}{55.25} + \frac{(9.25)^2}{29.75} \\ \chi^2 &= \frac{85.5625}{55.25} + \frac{85.5625}{29.75} \\ \chi^2 &= 1.5486 + 2.8760 \\ \chi^2 &= 4.4246\end{aligned}$$

Now that we have our chi-square statistic, we need to compare it to the chi-square critical value for 0.05 with *one degree of freedom*, since we have two categories. Using the [chi-square value calculator](#), we find the critical value to be 3.8414. The critical value indicates that only 0.05, or 5%, of values would be as high as 3.8414. If the χ^2 of our data is greater than 3.8414, then fewer than 5 times out of 100 would we expect to get that result if the null hypothesis is true.

- **Step 4: Interpret your results**

Our calculated data value of $\chi^2 = 4.4246$ is greater than the 0.05 significance level critical value of 3.8414, so we reject the null hypothesis. The data that Eric observed does not support the distribution that Rachel claimed.

Example C





The online car magazine “*Camaro5.com*” claims that 51% of Ford Mustang or Chevy Camaro owners own Camaros. Ellen is a Mustang lover and decides to do some research. If Ellen collects the data below, does her data support the magazine’s claim?

Data: Mustang owners: 28, Camaro owners: 34

Solution:

- **Step 1: Clearly state the null and alternative hypotheses**

H_0 : The survey agrees with the sample.

H_1 : The survey does not agree with the sample.

- **Step 2: Identify an appropriate test and significance level**

Since we are comparing two sets of data, and not just a single value, a Chi-Square test is appropriate. In the absence of a stated significance level in the problem, we assume the default 0.05.

- **Step 3: Analyze sample data**

We will start by creating a table to organize our data:

TABLE 11.24:

	Mustang	Camaro	TOTAL
Observed	28	34	62
Expected	$0.49 \times 62 = 30.4$	$0.51 \times 62 = 31.6$	62

Now we can calculate our chi statistic:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\chi^2 = \frac{(28 - 30.4)^2}{30.4} + \frac{(34 - 31.6)^2}{31.6}$$

$$\chi^2 = \frac{(-2.4)^2}{30.4} + \frac{(2.4)^2}{31.6}$$

$$\chi^2 = .3718$$

The chi-square critical value for $df = 1$ and a significance level of 0.05 is 3.8414 (the same as in Example B).

• **Step 4: Interpret your results**

Our calculated data value of $\chi^2 = 0.3718$ is significantly less than the 0.05 significance level critical value of 3.8414, so we fail to reject the null hypothesis. This means that, unfortunately for Ellen, her research did not allow her to deny the claim that Camaros are more popular.

Concept Problem Revisited

Suppose you wanted to evaluate a recent statistic stating that iOS represents 32% and Android 51% of active smart phones. You would like to know if the statistic actually reflects the distribution of phones among your friends. How could you evaluate the data you collect to see if it supports this hypothesis?

You could evaluate the hypothesis by collecting data from a SRS of cell phone owners and using a chi-square test to see if your data supports the hypothesis.

Vocabulary

A **chi-square** statistic is a derived value used in a **chi-square test** to calculate the probability that a given distribution is a good fit for observed data.

The **degrees of freedom** of a variable are the number of values in the final calculation of a statistic that are free to vary. The degrees of freedom are calculated as $n - 1$, where n is the number of samples or categories in the variable.

Guided Practice

Questions 1-5 refer to the following data:

Tuscany claims that 70% of dog or cat owners own a dog, and 30% own a cat. Sayber decides to test her claim and learns that 23 of the 40 people he asks own dogs, and 17 own cats.

1. What kind of test could you use to see if Sayber's data supports Tuscany's claim?
2. What would be the null and alternative hypotheses?
3. What would be the expected values of dog and cat owners?
4. What is the chi-square statistic of the observed data?
5. Assuming a 0.1 significance level, does Sayber's data support Tuscany's claim?

Solutions:

1. A chi-square test would be appropriate.
2. The null hypothesis, H_0 , would be that the research *does* support the hypothesis, the alternative hypothesis would be that it does not.
3. The expected number of dog owners, according to Tuscany's claim, would be 70% of the 40 people that Sayber polled, or 28 dog owners. The expected number of cat owners would be 30% of the 40 people polled, or 12.
4. The χ^2 statistic is the sum of the squared differences between the observed and expected values, divided by the expected values:

$$\begin{aligned}\chi^2 &= \frac{(23 - 28)^2}{28} + \frac{(17 - 12)^2}{12} \\ &= \frac{25}{28} + \frac{25}{12} \\ \chi^2 &= 2.9762\end{aligned}$$

5. The critical value of chi-squared for 1 degree of freedom at a significance level of 0.1 is 2.705. Since the chi-square statistic we calculated is 2.9762, and is therefore more extreme than the critical value, we may *reject the hypothesis*, and say that Sayber's data does not support Tuscany's claim.

Practice

Questions 1-5 refer to the following:

Evan claims that 15% of computer gamers have played "Team Fortress 2", and 35% have played "World of Warcraft". Evan's brother is skeptical of those figures and decides to do some research. He discovers that 60 of the 200 computer gamers he polls have played "Team Fortress 2", and 90 have played "World of Warcraft".

1. Create a table to organize the data and prepare for hypothesis testing.
2. What sort of test would be appropriate to determine if the observed data supports Evan's claim?
3. What would be H_0 and H_1 ?
4. What would be the χ^2 statistic for the observed data?
5. How many degrees of freedom are there in the variable "played game"?
6. Assuming a significance level of 0.05, what is the χ^2 critical value?
7. Does the observed data support Evan's claim? Explain your findings.

Questions 8-15 refer to the following:

Mack claims that 84% of street racers drive import cars, and 16% drive domestic muscle cars. Abbi likes domestic cars and thinks Mack is overstating the percentage of imports, so she does some research of her own and finds that 57 of the street racers she interviewed drive imports, and 31 drive American muscle.

8. Create a table to organize the data and prepare for hypothesis testing.
9. What sort of test would be appropriate to determine if the observed data supports Mack's claim?
10. What would be H_0 and H_1 ?
11. What would be the χ^2 statistic for the observed data?
12. How many degrees of freedom are there in the variable "played game"?
13. Assuming a significance level of 0.10, what is the χ^2 critical value?
14. Does the data indicate that Abbi should reject, or fail to reject H_0 ?
15. Interpret your results.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 11.5.

11.6 Chi-Squared II - Testing for Independence

Objective

Here you will learn how to use a chi-square test to determine whether two variables are dependent or independent.

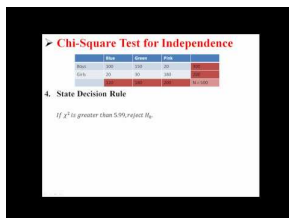
Concept

It is a common belief that age influences food preference. Suppose you wanted to test that hypothesis. How could you test observed data to see if the two variables (age and food preference) influence each other?

Look to the end of the lesson to see the answer.



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67551>

http://youtu.be/LE3AIyY_cn8 statslectures - Chi-Square Test for Independence

Guidance

A chi-square (χ^2) test can be used to determine if observed data indicates that two variables are dependent in much the same way that the test can be used to determine goodness of fit.

Just as with a goodness of fit test, we will calculate expected values, calculate a chi-square statistic, and compare it to the appropriate chi-square value from a reference to see if we should reject H_0 , which is that the *variables are not related*.

In fact, the only major difference in process between a goodness of fit test and a test of independence is how we calculate the expected values, as you will see in Example A.

Just for reference:

- The formula to calculate chi-square is:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

- A good resource for chi-square critical values is:

<http://www.danielsoper.com/statcalc3/calc.aspx?id=12>

or

<http://easycalculation.com/statistics/chi-squared-distribution-table.php>

- The formula for calculating expected values in a test of independence is:

$$\text{expected cell value} = \frac{C \times R}{n}$$

Where C is the observed column total for the cell, R is the observed row total for the cell, and n is the total number of samples. (see Example A for clarification of the use of the formula)

- The degrees of freedom in a test of independence are calculated as:

$$df = (\text{rows} - 1)(\text{columns} - 1)$$

Example A

Given the following contingency table, what is the expected value for each of the four cells in the body of the table?

TABLE 11.25:

	A	B	TOTAL
X	23	37	60
Y	19	41	60
TOTAL	42	78	120

Solution:

To calculate the expected values, use the formula $\text{expected cell value} = \frac{C \times R}{n}$ for each cell:

- Cell $\frac{X}{A}$: $\frac{42 \times 60}{120} = 21$
- Cell $\frac{Y}{A}$: $\frac{42 \times 60}{120} = 21$
- Cell $\frac{X}{B}$: $\frac{78 \times 60}{120} = 39$

- Cell $\frac{Y}{B}$: $\frac{78 \times 60}{120} = 39$

Example B

Using the contingency table and data from Example A, calculate χ^2 .

Solution:

Start by adding the expected values you calculated in Ex A to the table. Use parentheses to set off the expected values:

TABLE 11.26:

	A	B	TOTAL
X	23 (21)	37 (39)	60
Y	19 (21)	41 (39)	60
TOTAL	42	78	120

Now use the chi-square formula $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ to calculate the statistic:

$$\begin{aligned}\chi^2 &= \frac{(23 - 21)^2}{21} + \frac{(19 - 21)^2}{21} + \frac{(37 - 39)^2}{39} + \frac{(41 - 39)^2}{39} \\ \chi^2 &= \frac{(2)^2}{21} + \frac{(-2)^2}{21} + \frac{(2)^2}{39} + \frac{(-2)^2}{39} \\ \chi^2 &= \frac{4}{21} + \frac{4}{21} + \frac{4}{39} + \frac{4}{39} \\ \chi^2 &= 0.5861\end{aligned}$$

Example C

Rachel claims that girls take more black and white and color photographs than boys, but Jack (who is a photographer) is skeptical. If Jack collects the following data, would it be correct to say that he should reject Rachel's claim that gender affects tendency to take photographs?

TABLE 11.27:

	Black/White	Color	TOTAL
<i>Female</i>	72	489	561
<i>Male</i>	48	530	578
TOTAL	120	1019	1139

Solution:

The question here is whether gender affects tendency to take more photographs, or, in other words, are gender and photograph-taking tendency dependent?



To run a chi-squared test, we need to know the expected value for each of the four cells containing observations. In a test for independence, this is calculated with the formula: $expected\ cell\ value = \frac{C \times R}{n}$.

1. The upper-left cell, female X black/white:

$$\begin{aligned}
 expected\ cell\ value &= \frac{C \times R}{n} \\
 &= \frac{(column\ total) \times (row\ total)}{total\ number\ of\ observations} \\
 &= \frac{120 \times 561}{1139} \\
 expected\ cell\ value &= 59.1
 \end{aligned}$$

2. The cell below that, male X black/white:

$$\begin{aligned}
 expected\ cell\ value &= \frac{C \times R}{n} \\
 &= \frac{(column\ total) \times (row\ total)}{total\ number\ of\ observations} \\
 &= \frac{120 \times 578}{1139} \\
 expected\ cell\ value &= 60.9
 \end{aligned}$$

3. Top-right cell, female X color:

$$\begin{aligned}
 expected\ cell\ value &= \frac{C \times R}{n} \\
 &= \frac{(column\ total) \times (row\ total)}{total\ number\ of\ observations} \\
 &= \frac{1019 \times 561}{1139} \\
 expected\ cell\ value &= 501.9
 \end{aligned}$$

4. Bottom-right cell, male X color:

$$\begin{aligned}
 \text{expected cell value} &= \frac{C \times R}{n} \\
 &= \frac{(\text{column total}) \times (\text{row total})}{\text{total number of observations}} \\
 &= \frac{1019 \times 578}{1139} \\
 \text{expected cell value} &= 517.1
 \end{aligned}$$

Now we can add the expected values to our initial table, placing the expected value for each cell in parentheses:

TABLE 11.28:

	Black/White	Color	TOTAL
Female	72 (59.1)	489 (501.9)	561
Male	48 (60.9)	530 (517.1)	578
TOTAL	120	1019	1139

Now we can calculate our χ^2 statistic as before, using: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ and each of the four values in the body of the table:

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\
 \chi^2 &= \frac{(72 - 59.1)^2}{59.10} + \frac{(48 - 60.1)^2}{60.89} + \frac{(489 - 501.9)^2}{501.89} + \frac{(530 - 517.1)^2}{517.10} \\
 \chi^2 &= \frac{(12.9)^2}{59.10} + \frac{(-12.9)^2}{60.89} + \frac{(-12.9)^2}{501.89} + \frac{(12.9)^2}{517.10} \\
 \chi^2 &= \frac{166.41}{59.10} + \frac{166.41}{60.89} + \frac{166.41}{501.89} + \frac{166.41}{517.10} \\
 \chi^2 &= 2.82 + 2.73 + .33 + .32 \\
 \chi^2 &= 6.20
 \end{aligned}$$

To see if our χ^2 statistic is greater or less than the critical value at the default significance level of 0.05, we need the number of *degrees of freedom*: $df = (\text{rows} - 1)(\text{columns} - 1)$

$$\begin{aligned}
 df &= (2 - 1)(2 - 1) \\
 df &= 1
 \end{aligned}$$

Using our chi-squared critical value reference, we find that the critical value for 0.05 with $df = 1$ is 3.8414.

Finally, we compare our calculated chi-squared value of 6.2 to the critical value of 3.8414 and determine that since $6.2 > 3.8414$, we can *reject* H_0 , in other words, we reject the independence of the variables. **The observed data indicates that there is a gender bias on picture-taking tendency.**

Concept Problem Revisited

It is a common belief that gender influences movie genre preference. Suppose you wanted to test that hypothesis. How could you test observed data to see if the two variables (gender and movie genre preference) influence each

other?

A chi-square test of independence could be used in this situation. Create a contingency table to organize observed data on movie preference and gender, calculate the χ^2 value of the data, and compare it to the χ^2 critical value with the appropriate number of degrees of freedom. If the calculated value is greater than the critical value, then the variables are not related.

Vocabulary

A **chi-squared** (χ^2) statistic is calculated using $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$, and may be used to evaluate variable independence.

The **expected value**, as used in a chi-square test, is the value you would expect to get if the null hypothesis is correct.

Guided Practice

Questions 1-4 refer to the following:

Kato claims that single people prefer different pizzas than married people do. Kato's brother doesn't think that is true, so he conducts some research of his own, and collects the data below.



TABLE 11.29:

	Pepperoni	Sausage	Cheese	TOTAL
Single	29	12	61	102
Married	8	47	56	111
TOTAL	37	59	117	213

1. Fill in the expected values of the 6 cells in the body of the table using parenthesis.
2. What is the value of χ^2 ?
3. How many degrees of freedom are there?
4. If we plan to test the claim, what are H_0 and H_1 ?
5. Assuming a significance level of 0.05, does the observed data support Kato's claim?

Solutions:

1. Completed table, using *expected cell value* = $\frac{C \times R}{n}$:

TABLE 11.30:

	Pepperoni	Sausage	Cheese	TOTAL

TABLE 11.30: (continued)

Single	29 (17.71)	12 (28.25)	61 (56.02)	102
Married	8 (19.28)	47 (30.74)	56 (60.97)	111
TOTAL	37	59	117	213

2. Using $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$:

$$\chi^2 = \frac{(29 - 17.71)^2}{17.71} + \frac{(12 - 28.25)^2}{28.25} + \frac{(61 - 56.02)^2}{56.02} + \frac{(8 - 19.28)^2}{19.28} + \frac{(47 - 30.74)^2}{30.74} + \frac{(56 - 60.97)^2}{60.97}$$

$$\chi^2 = \frac{(11.29)^2}{17.71} + \frac{(-16.25)^2}{28.25} + \frac{(4.98)^2}{56.02} + \frac{(-11.28)^2}{19.28} + \frac{(16.26)^2}{30.74} + \frac{(-4.97)^2}{60.97}$$

$$\chi^2 = 7.2 + 9.35 + 0.44 + 6.6 + 8.6 + 0.41$$

$$\chi^2 = 32.6$$

3. $(3 - 1)(2 - 1) = (2)(1) = 2 \text{ df}$

4. H_0 : The observed data supports the hypothesis, H_1 : The data does not support the hypothesis.

5. The critical value for χ^2 with $2df$ at 0.05 significance is 5.991. Since our calculated value of $\chi^2 = 32.6$, and $32.6 > 5.991$ we can **reject the null hypothesis that the data supports the claim**.

Practice

1. What use of the χ^2 statistic is used in this lesson?
2. How is an expected value calculated, in the context of this lesson?
3. How are degrees of freedom calculated in a chi-square test of independence?
4. What type of table is commonly used to organize information for a chi-square test of independence?
5. What is the default level of significance for a chi-squared test of independence?

Questions 6 - 10 refer to the following breakdown of favorite flavor by gender:

TABLE 11.31:

	Cherry	Lemon	Strawberry	Other	TOTAL
Male	13	11	7	13	44
Female	15	18	11	5	49
TOTAL	28	29	18	18	93

6. Fill in the expected values of the 6 cells in the body of the table using parenthesis.
7. What is the value of χ^2 ?
8. How many degrees of freedom are there?
9. If we plan to test the claim that gender affects favorite flavor, what are H_0 and H_1 ?
10. Assuming a significance level of 0.05, does the observed data indicate that we reject or fail to reject H_0 ?

Questions 10 - 15 refer to the following:

Are hamburger cooking preferences dependent on gender? 1087 people were asked their preference among three

ways of cooking burgers, grilling, frying, and broiling. The men stated their preferences as: 137: grilling, 193: frying, and 212: broiling. The women were distributed as: 110: grilling, 215: frying, and 220: broiling.

11. Create a contingency table to organize the information.
12. What is the value of χ^2 ?
13. How many degrees of freedom are there?
14. If we plan to test the claim that gender affects cooking preference, what are H_0 and H_1 ?
15. Assuming a significance level of 0.05, does the observed data indicate that we reject or fail to reject H_0 ?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 11.6.

Students were introduced to the concept of linear regression and Pearson's correlation coefficient. Students learned to calculate a line of best fit using the least squares method.

After a review of the creation of contingency tables, students practiced extracting data from them to calculate the Chi-Squared statistic. Finally the students learned to use chi-square statistic to run chi-square tests of goodness of fit and variable independence.

11.7 References

1. . . CC BY-NC-SA
2. Eliot Phillips. <https://www.flickr.com/photos/hackaday/5334437242> .
3. . . CC BY-NC-SA
4. . . CC BY-NC-SA
5. . . CC BY-NC-SA
6. Christian Haugen. <https://www.flickr.com/photos/christianhaugen/3721602063> .
7. . . CC BY-NC-SA
8. . . CC BY-NC-SA
9. . . CC BY-NC-SA
10. . . CC BY-NC-SA
11. . . CC BY-NC-SA
12. . . CC BY-NC-SA
13. OpenClips. <http://pixabay.com/en/film-reel-cinema-film-movie-reel-147631/?oq=film> .
14. StooMathiesen. <https://www.flickr.com/photos/stoo57/5773404346> .
15. . . CC BY-NC-SA
16. Derek Hatfield. <https://www.flickr.com/photos/loimere/6772341183> .
17. PGHAuto2010. <https://www.flickr.com/photos/47591094@N08/4354783689> .
18. George Rigato. <https://www.flickr.com/photos/georgerigato/2880099644> .
19. Robert Lopez. [CK-12 Foundation](#) .
20. PublicDomainPictures. <http://pixabay.com/en/camera-digital-equipment-female-15673/?oq=camera> .
21. jeffreyw. <https://www.flickr.com/photos/jeffreyww/6057076521> .

CHAPTER 12**Reasoning****Chapter Outline**

- 12.1** **INDUCTIVE AND DEDUCTING REASONING**
 - 12.2** **ARGUMENTS**
 - 12.3** **EULER DIAGRAMS**
 - 12.4** **VALID FORMS**
 - 12.5** **HIDDEN PREMISES**
 - 12.6** **STRUCTURAL FALLACIES**
 - 12.7** **CONTENT FALLACIES**
 - 12.8** **REFERENCES**
-

Logic and Reasoning go hand in hand with Probability and Statistics. Understanding how to conduct a survey or poll and make sense out of the data you collect can help you gain some great insights into how things work. However, to share what you discover and, in some cases, debate its value, you need to understand how to think logically and argue your case with reason. In this lesson you will learn the technical definition of an argument, how to structure an argument viably, and how to identify logical fallacies or faults.



12.1 Inductive and Deducting Reasoning

Objective

Here you will learn the difference between inductive reasoning and deductive reasoning.

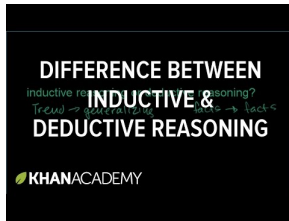
Concept



Suppose you were given the task of collecting data from each class in your school on the ratio between male and female students. After reviewing the M:F ratios of each classroom, would you use inductive reasoning or deductive reasoning to come up with a hypothesis regarding an average M:F ratio for the school? What kind of reasoning would be involved if your friend asked you to review your data to see if her theory about ratios being different in different grades was supported by your observations?



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67539>

<http://youtu.be/GEId0GonOZM> KhanAcademy - Difference between inductive and deductive reasoning

Guidance

One of the primary uses of probability and statistics is to learn about parameters of a population, and to do that, one must be able to reason from a sample to a population. Either a person observes something and tries to explain it by collecting and distilling data into a conclusion, or else he/she begins with a hypothesis and seeks data to support or renounce it. In this lesson, we will discuss these two types of reasoning, Inductive and Deductive.

- **Deductive Reasoning** - Begins with the question or theory and works toward specific examples or evidences to support or renounce it.
 - Every morning, I eat eggs for breakfast. Every day, I am not hungry again until lunchtime. This morning if I eat eggs for breakfast, I will not be hungry until lunchtime.
- **Inductive Reasoning** - Begins with specific observations or data and works toward a general statement to explain it.
 - This morning I ate eggs for breakfast and was not hungry until lunchtime. As long as I eat eggs for breakfast, I'll never be hungry until lunchtime.

In scientific study, both sorts of reasoning are used, often in conjunction and to support each other. However, as you will see over the next few lessons, there are a lot of ways to make errors in reasoning (called *fallacies*), and knowing what type of reasoning you are using will help you to learn which fallacies to watch out for!

Example A

What sort of reasoning is applicable to finding the solution to a five-step linear equation such as the one below?

$$\begin{aligned}
 2(x + 3) - 7 &= x + 4 \\
 2(x + 3) &= x + 11 \\
 2x + 6 &= 11 \\
 2x - x &= 11 - 6 \\
 x &= 5
 \end{aligned}$$

Solution:

This is deductive reasoning, since we started with a statement or theory: $2(x + 3) - 7 = x + 4$, and used a step-by-step process to find a specific example supporting it, namely that if $x = 5$, then $2(5 + 3) - 7 = 5 + 4$, so the original statement is supported by a specific example.

Since we progressed from general to specific, this was deductive reasoning.

Example B

Assuming the sequence below, what type of reasoning would you use to conjecture the next number in the sequence?

1, 4, 10, 19, 31, 46, 64, ...

Solution:

This is an example of inductive reasoning, since we started with a number of specific observations, namely the 1st, 2nd, 3rd, 4th, and so on numbers in a sequence, and use the observations to make the statement that the pattern is to add $3n$, where n is the count, to each number to get the next: $1 + 3(1) = 4$, $4 + 3(2) = 10$, $10 + 3(3) = 19$, $19 + 3(4) = 31$, and so on. That tells us that the next number in the series should be: $64 + 3(7) = 85$.

Since we progressed from specific to general, this was inductive reasoning.

Example C

What sort of reasoning is expressed in the following statements?

Chloe took her umbrella to work today, and it rained.

Every time Chloe takes her umbrella, it will rain.

**Solution:**

This is inductive reasoning, beginning with the specific statement about a specific day and action, and progressing to a general statement about *all* days with the same action.

Concept Problem Revisited

Suppose you were given the task of collecting data from each class in your school on the ratio between male and female students. After reviewing the M:F ratios of each classroom, would you use inductive reasoning or deductive reasoning to come up with a hypothesis regarding an average M:F ratio for the school? What kind of reasoning would be involved if your friend asked you to review your data to see if her theory about ratios being different in different grades was supported by your observations?

First, you begin with specific examples of the ratios of males and females and use them to create a general statement about the ratio of the entire school. That was inductive reasoning: specific to general. Second, you started with the general statement that the ratios are different in different grades, and considered the specific data to support or not support the statement. That was deductive reasoning: general to specific.

Vocabulary

Inductive reasoning describes reasoning from specific examples to general statements.

Deductive reasoning describes reasoning from general statements to specific examples.

Guided Practice

For questions 1-4, describe the type of reasoning demonstrated in each passage.

1. Scott leaves for school at 8:15 in the morning every day, it takes him 15 minutes to get to school, and he arrives on time. If Scott leaves at 8:15 this morning, he will arrive at school on time.
2. On Monday, Sophie went to lunch at the local fast-food joint on her lunch break and arrived back at school in time for class. On Tuesday, she did the same thing and was on time again. If Sophie goes to the same fast-food place for lunch on every day, she will be back in time for class.
3. $3(x - 4) - 7 = 6x$, therefore, $x = -6.\overline{33}$.
4. If $y = 7$, and $x = 4$, therefore $x \times \frac{7}{4} = y$.

Solutions:

1. This is deductive reasoning, starting with a general statement about Scott's actions everyday and progressing to the specific occurrence of today.
2. This is inductive reasoning, starting with specific examples of actions and progressing to a general statement about every similar action.
3. Deductive reasoning, from a general statement to a specific example of the statement being true.
4. Inductive reasoning, from specific stated values of x and y to a general statement about them both.

Practice

For each question, state whether the reasoning is an example of inductive or deductive logic.

1. All housecats are felines. All felines have claws. Therefore all housecats have claws.
2. My dog has fleas. My neighbor's dog has fleas. Therefore all dogs must have fleas.
3. All cows like hay. My cow will like hay.
4. My Mac laptop is fast. All Mac laptops are fast.
5. My tennis shoes are comfortable. My friend's tennis shoes are comfortable. All tennis shoes are comfortable.
6. The scalloped potatoes I took from the oven were cheesy. The enchiladas I took from the oven were cheesy. If I take cookies from the oven, they will be cheesy.
7. Everything cooked on the stove gets hot. If I cook macaroni on the stove, it will get hot.
8. iPads are popular. iPhones are popular. Every phone or tablet is popular.
9. Roses are red. Tomatoes are red. All red things come from plants.
10. Rock music is loud. Sayber listens to rock music. Sayber's music is loud.
11. Milk is good with cookies. Snicker doodles are cookies. Milk is good with snicker doodles.
12. Hummers use a lot of gas. Suburbans use a lot of gas. Large SUV's use a lot of gas.
13. My garden has pumpkins. My dad's garden has pumpkins. All gardens have pumpkins.
14. Prob and Stats students are smart. You are a Prob and Stats student. You are smart.
15. Students who study hard get good grades. You are a student who studies hard. You will get good grades.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 12.1.

12.2 Arguments

Objective

Here you will learn about the formal terminology involved with logical reasoning and arguments.

Concept

What does it mean to state a sound concrete argument with premises A and B and conclusion C? If one or more of the premises is untrue, does that make the argument unsound (quiet, maybe?), or not concrete (muddy, perhaps?).

Guidance

Formal logical reasoning can seem somewhat... illogical to someone not familiar with the terminology involved. Jargon such as 'affirm the disjunct' or 'denying the consequent', can certainly sound impressive, but what does it *mean*? Hearing terms such as these may make you think that logical reasoning is really only for lawyers or politicians. The truth is, understanding the basics of logical reasoning is an excellent skill for "the rest of us" who just want to be able to tell fact from fiction.



By now you should know that statistics can be a pretty complex study, and that forming conclusions from questionable or faulty data is chancy at best. Logical reasoning is very similar. It is pretty easy for someone who really understands reasoning to make an argument that *seems* sound even though it may be based on faulty information, or to make true information seem to support an incorrect conclusion.

The goal of these lessons on logical thinking and argument is to help you recognize invalid and unsound reasoning so that you can make decisions in your life that are based on true data, rather than just someone else's interpretation of data.

Let's start with some definitions:

- An **argument** is a series of statements, progressing (usually in order, but not necessarily) from the **premises**, which are the assumptions (true or untrue), to the **conclusion**.
- The purpose of an **argument** is to present the **premises** in such a way as to support the truth of the **conclusion**.

- A **concrete** statement is one that provides a specific example of a concept rather than just a generalization. For instance:
 - **Generalization**: If **A**, then **B**. **B**, therefore **A**.
 - **Concrete**: If it rains, I carry an umbrella. It is raining, therefore I am carrying an umbrella.
- An **argument** is **valid** if the truth of its **premises** assures the truth of its **conclusion**.
- An **argument** is **invalid** or **fallacious** if it is not **valid**.
- A **sound** argument has both true **premises** and **valid** reasoning.

Example A

Is the following a valid argument?

Stars are holiday lights in the curtain of night. Holiday lights are only lit from November to February. Therefore, stars are only lit from November to February.

Solution:

Yes! This is indeed a **valid** argument. Remember that an argument is **valid** if the truth of the premises assures the truth of the conclusion. If indeed the stars were holiday lights, and if holiday lights were only lit for those few months, then the stars would go out at the end of January, not to be seen again until after Halloween.

This is a good example of the fact that an argument certainly need not be **sound**, or true, in order to be **valid**.

Example B

Is the following argument sound?

If it is snowing, it is cold. It is snowing, therefore it is cold.

Solution:

Yes, this is a **sound** argument. It is **valid**, since the truth of the premises “If it is snowing, it is cold”, and “It is snowing”, assures the truth of the **conclusion** “It is cold”. Since the statements are both true (at least they are right now, since it is snowing outside as I write this question!), the conclusion is also true.

Example C

Is the following argument **sound**?

If the grass is green, it is not winter. It is late fall, therefore the grass is green.



Solution:

No, this is not a **sound** argument, in fact, it is not **valid**. The problem is that even if both premises are true, they do *not* assure the truth of the conclusion. This is actually an example of a common **fallacy** that we will explore further in another lesson.

Concept Problem Revisited

What does it mean to state a sound argument with premises *A* and *B* and conclusion *C*? If one or more of the premises is untrue, does that make the argument unsound (silent, maybe?).

A **sound argument** is one with both valid reasoning and true premises. In this case, that means that both premise *A* and premise *B* are true, and they ensure that the conclusion, *C*, is also true. If either *A* or *B* prove to be untrue, then the argument will still be valid, but will no longer be considered sound.

Vocabulary

An **argument** is a series of statements, progressing (usually in order, but not necessarily) from the **premises**, which are the assumptions (true or untrue), to the **conclusion**.

An **argument** is **valid** if the truth of its **premises** assures the truth of its **conclusion**, and **invalid** or **fallacious** if it is not **valid**.

A **sound** argument has both true **premises** and **valid** reasoning.

Guided Practice

Questions 1-4 refer to the following argument:

All people who drive red cars get speeding tickets. I drive a red car. I get speeding tickets.

1. What are the premises of this argument?
2. What is the conclusion?
3. Is the argument valid?
4. Is the argument sound?

Solutions:

1. The premises are: a) "All people who drive red cars get speeding tickets". b) "I drive a red car".
2. The conclusion is: "I get speeding tickets".
3. The argument is valid, since the conclusion must be true if both premises are true.
4. The argument is *not* sound, since the first premise, "All people who drive red cars get speeding tickets", is not true.

Practice

Questions 1-4 refer to the following:

All students who listen to comedy shows while studying get distracted. Evan listens to comedy shows while studying. Therefore, Evan gets distracted.

1. What are the premises?
2. What is the conclusion?
3. Is the argument valid?
4. Is the argument sound?

Questions 5-8 refer to the following:

Basketball is great exercise. Sam plays basketball. Sam is in great shape.

5. What are the premises?

6. What is the conclusion?

7. Is the argument valid?

8. Is the argument sound?

Questions 9-12 refer to the following:

All students that fall asleep in class are male. Trisha falls asleep in class. Therefore Trisha is male.

9. What are the premises?

10. What is the conclusion?

11. Is the argument valid?

12. Is the argument sound?

Questions 13-16 refer to the following:

All dogs chase cats. Mack chases cats. Therefore Mack is a dog.

13. What are the premises?

14. What is the conclusion?

15. Is the argument valid?

16. Is the argument sound?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 12.2.

12.3 Euler Diagrams

Objective

Here you will learn to use Euler Diagrams to help test the validity of arguments.

Concept

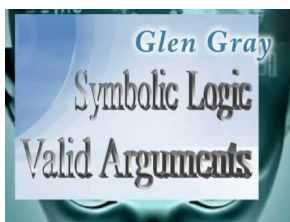
Consider the following argument:

Dogs learn fast. Frank does not learn fast. Frank is a cat.

How could you use a diagram to help you evaluate the validity of this argument?



Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67545>

<http://youtu.be/DXtLux0XYms> Glen Gray - Logic 3 - Valid Arguments

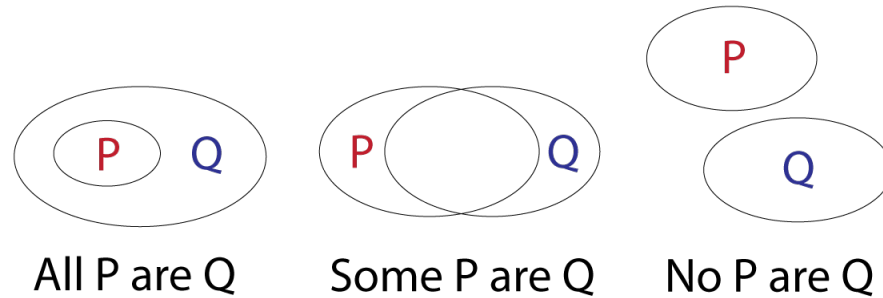
Guidance

You may not immediately associate art with Reason and Logic, but sometimes drawing diagrams can greatly simplify the process of evaluating a statement for validity or soundness.

Euler (sounds like “oiler”) diagrams look very much like Venn Diagrams, but Euler was using them to describe mathematical concepts a very long time before the classic Venn diagram was recognized. The purpose of Euler

diagrams is to create a visual representation of each of the aspects in a logical argument so that the conclusion may be clearly evaluated.

Generally, one oval is constructed to represent each set described in the argument, and an “X” is used to represent solitary units. Possible relationships can be expressed by the location of the ovals and “X’s”.



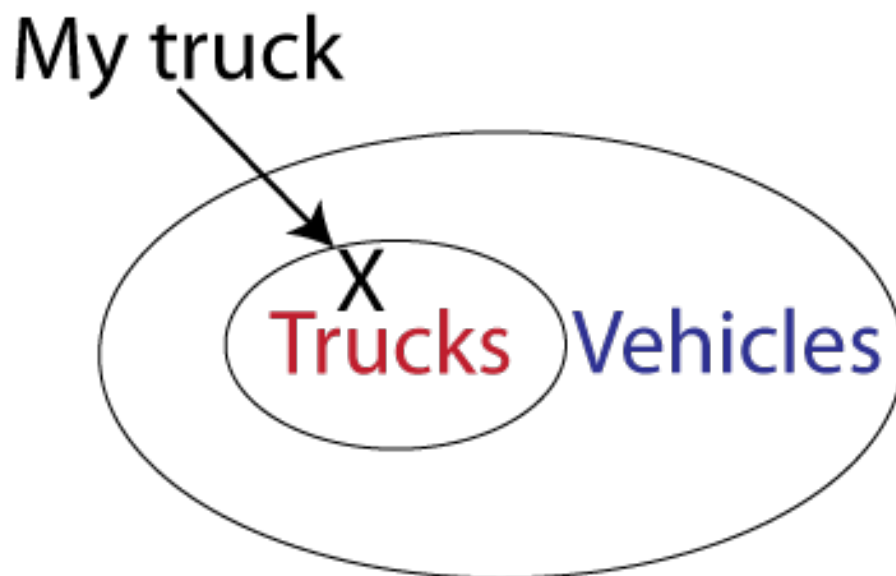
Example A

Express the argument using an Euler diagram.

All trucks are vehicles. I drive a truck. Therefore I drive a vehicle.

Solution:

The set “all trucks” is a subset of the set “vehicles”. My truck is a member of the set “all trucks”, so my truck is also within the set “vehicles”.



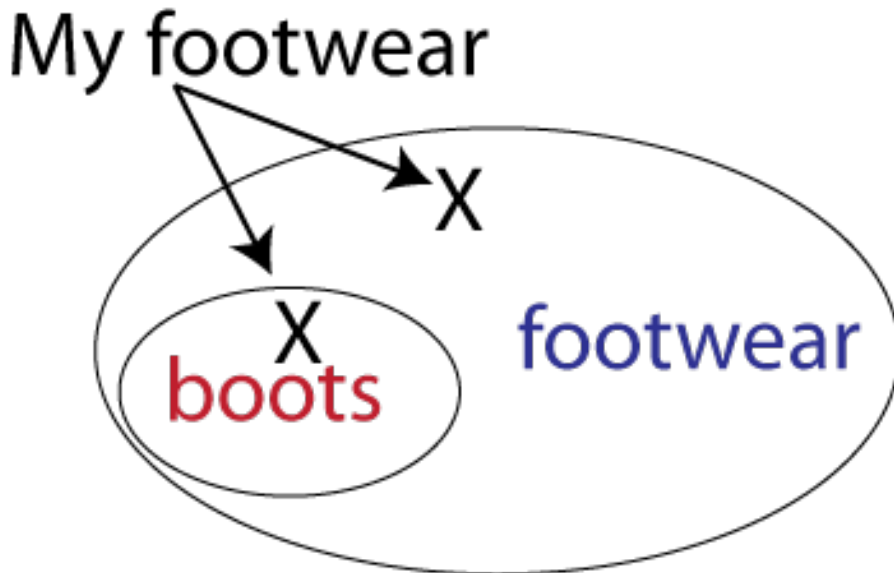
Example B

Evaluate the validity of the argument using an Euler diagram.

Boots are a type of footwear. I wear footwear. Therefore, I wear boots.

Solution:

The set “boots” is a subset of the set “footwear”. My footwear is in the set “footwear”, but we do not know if it is in the set “boots” or not.



We can see from the diagram that this argument cannot be valid, since both statements may be true but I might be wearing sandals, making the conclusion false.

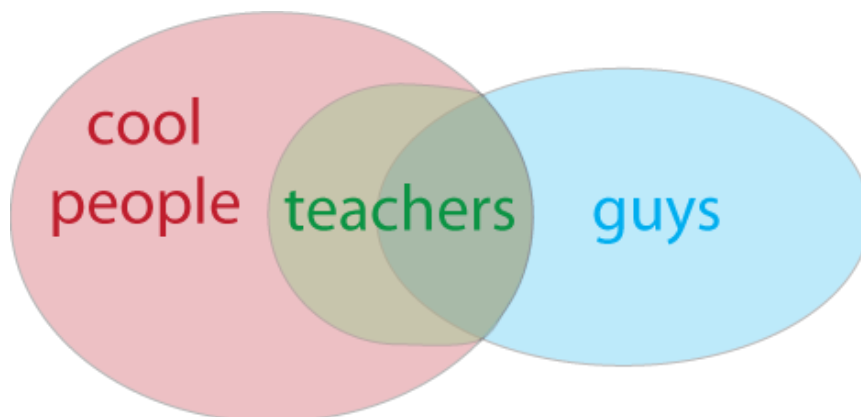
Example C

Evaluate the validity of the argument using an Euler diagram.

All teachers are cool. Some guys are teachers. Some guys are cool.

Solution:

The set “teachers” is a subset of the set “cool people”. The set “guys” intersects with the set “teachers”. The set “cool people” intersects with the set “guys” where “guys” includes “teachers”.



If the premises are true, the conclusion must be. The only way for there to be no cool guys is for one of the premises to be false. The argument is valid.

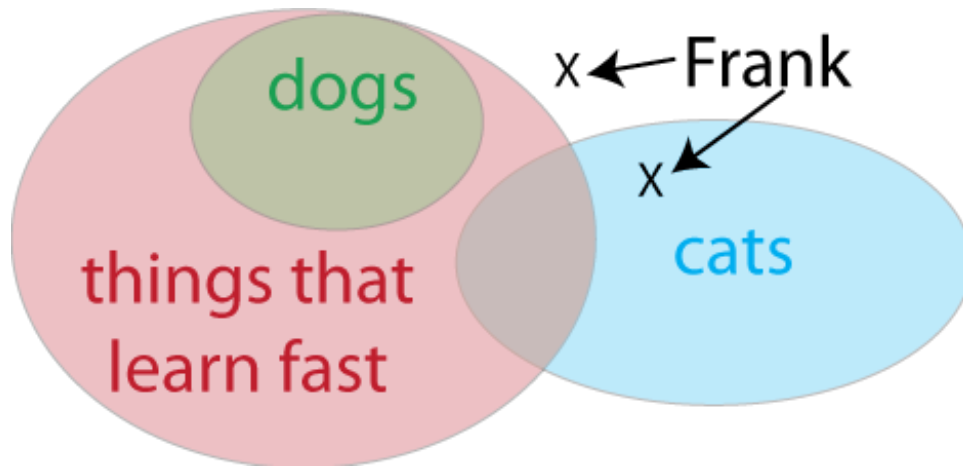
Concept Problem Revisited

Consider the following argument:

All dogs learn fast. Frank does not learn fast. Frank is a cat.

How could you use a diagram to help you evaluate the validity of this argument?

Create an Euler diagram to illustrate the argument:



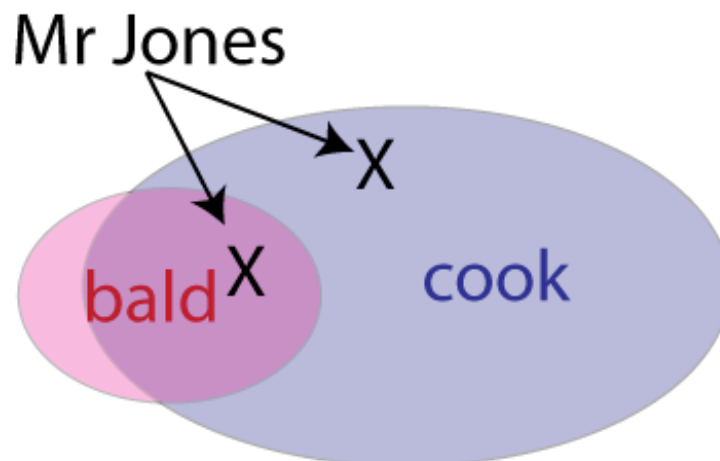
“Dogs” are a subset of “things that learn fast”. Frank is not a member of the set “things that learn fast”, so he is not a member of “dogs” either. However, that does not mean he must be a member of “cats”. It is possible for both premises to be true, while the conclusion is false. The argument is invalid.

Vocabulary

An *Euler diagram* is similar to a Venn diagram. It is a visual representation of the relationship between sets, subsets, and members. Euler diagrams do not necessarily need to be composed of circles or ovals, polygons are also acceptable.

Guided Practice

Use the Euler diagram below to answer questions 1-3:



1. Which pair of premises could be illustrated by the Euler diagram?
 1. All bald people are cooks. Mr Jones is bald.

2. Some cooks are bald. Mr Jones is a cook.
 3. Mr Jones is a cook. Mr Jones is bald.
2. Given the premises from question 1, what conclusion is illustrated as being incorrect by the diagram?
 3. Is the argument valid? Why or why not?
 4. What valid conclusion *could* be drawn from the given premises?

Solutions:

1. The correct answer is “b: Some cooks are bald. Mr Jones is a cook”. The set of bald people intersects with the set of cooks, but neither is a subset of the other. Mr. Jones is somewhere within the set of cooks, but may or may not be within the set of bald people.
2. The conclusion that Mr Jones is/is not bald is show to be incorrect by the diagram, since he could be in either set, based on the given premises.
3. The argument is not valid, since both premises could be true while the conclusion remains false.
4. The only possible conclusion is that Mr Jones *may* be bald, which is the case regardless, so the premises are weak.

Practice

Create Euler diagrams to represent each of the situations in questions 1-7:

1. All P are Q .
2. Some P are Q .
3. All Q are *not* P .
4. Some P are *not* Q .
5. All P are Q . R is a member of P . R is a member of Q .
6. All P are Q . R is not Q . R is not P .
7. All P are Q . All Q are R . All P are R .
8. Create a Euler diagram to represent the following argument:

If a bull has been gelded, it is a steer. Ferdinand is not a steer. Therefore, Ferdinand is not a gelded bull.

9. Is the argument in question 8 valid or invalid? Why?
 10. Create an Euler diagram to represent the following argument:
- Ignoring problems makes them go away. I ignore my problems. My problems go away.
11. Is the argument in question 10 valid or invalid? Why?
 12. Create an Euler diagram to represent the following argument:

Arguments must be valid or invalid. This argument is invalid. This argument is valid.

13. Is the argument is question 12 valid or invalid? Why?
14. Create an Euler diagram to represent the following argument:

If you study Reason, you will better understand Logic. If you better understand Logic, you will make better use of Statistics. Therefore if you study Reason, you will be better at math.

15. Is the argument in question 14 valid or invalid? Why?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 12.3.

12.4 Valid Forms

Objective

Here you will learn to use some of the valid forms of argument.

Concept

If an argument is valid and the premises are true, then the conclusion must be true. How then can you be sure that your argument is valid to start with? Are there some standard forms of valid arguments to refer to?

Watch This

The link below is a playlist including a number of short videos specifically detailing the valid forms discussed in this lesson. There are also some of the *invalid* forms that we will be discussing in other lessons, which you may choose to review now or when you got to them later.



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67642>

<http://www.youtube.com/playlist?list=PL6B84FAC7296D01CC> Kevin deLaplante - Common Valid and Invalid Argument Forms

Guidance

An argument may be *valid* without being *sound*, but it cannot be sound without being valid. In addition, if a *valid* argument has true premises, then it must be sound. That means that one way to make sure that your arguments will be sound is to start by stating them in a particular form that you know to be valid. That way, you need only convince your audience that your premises are true in order to make your argument *persuasive*.

In this lesson, we will practice some valid forms of argument. Don't worry if the names of the forms seem odd, logical thinking and reasoning rules are sometimes literally *thousands* of years old, and so may have names based on ancient languages (primarily Latin).

If Q, then P

It is common when describing forms of argument to replace sentences or phrases with single letters, such as P and Q . By using letters to **generalize** an argument form, we can more easily evaluate a **concrete** argument for validity. It is a common, and useful, practice to replace P and Q with statements of your own in order to clarify the use of a particular form.

- **Modus ponens** (affirm by affirming): **If P , then Q . P , therefore Q .**
 - If water is frozen, then it is below 32 degrees Fahrenheit. This water is frozen, therefore it is below 32 degrees Fahrenheit.
- **Modus tollens** (denying the consequent): **If P , then Q . Not Q , therefore not P .**
 - If water is frozen, then it is below 32 degrees Fahrenheit. This water is not below 32 degrees Fahrenheit, therefore it is not frozen.
- **Hypothetical syllogism** (the chain argument): **If P , then Q . If Q , then R . Therefore, if P then R .**
 - If you wear sunscreen, you won't get sunburn. If you don't get sunburn, you will not get skin cancer. Therefore, if you wear sunscreen, you won't get skin cancer.
- **Disjunctive syllogism: P or Q . Not P . Therefore Q .** (also works in reverse)
 - You are either dead or alive. You are not dead. Therefore you are alive.
 - You are either dead or alive. You are not alive. Therefore you are dead.

Example A

Is the following a valid form of argument? If so which form is it?

If you overeat, you will get a bellyache. You do not have a bellyache. Therefore you did not overeat.

Solution:

This is a valid form. It is an example of *modus tollens*, denying the consequent. Because the initial premise is that every time you overeat, you get a bellyache, not having a bellyache must mean that you did not overeat.

Note that this does *not* necessarily mean that this is a **sound** argument. Since it is entirely possible to overeat without getting a bellyache, you might indeed have overeaten and felt fine. The important thing is that the *form* of the argument is valid, so that the only question is the truth of the premises.

Example B

Is the following argument stated in a valid form? If so, which form is it?

If you are a teenager with a smartphone, you send text messages. You are a teenager with a smartphone. Therefore you send text messages.

**Solution:**

This is a valid form, an example of *modus ponens*, affirm by affirming. Since the initial premise is that every teenager who owns a smartphone sends texts, if you are a teenager with a smartphone, you must send texts.

Example C

Is the following a valid form of argument? If so, which form?

If you wear a helmet, you won't hurt your head in a crash. If you don't hurt your head in a crash, you won't get a headache. Therefore, if you wear a helmet, you won't get a headache.

Solution:

This is an example of the valid form known as *hypothetical syllogism*, the chain argument. The premise about not getting a headache if you don't hurt your head is *chained* to the premise that you won't hurt your head if you wear a helmet.

As in prior examples, the *validity* of the argument does not necessarily lead to the *soundness* of it. Obviously you might get a headache for some reason other than hitting your head, and wearing a helmet won't prevent that.

Concept Problem Revisited

If an argument is valid and the premises are true, then the conclusion must be true. How then can you be sure that your argument is valid to start with? Are there some standard forms of valid arguments to refer to?

By making sure your own premises follow valid forms of reasoning, you will know that your conclusions are true as long as your premises are. There are many standard valid forms of argumentation, including *modus ponens*, *disjunctive syllogism*, *hypothetical syllogism*, *modus tollens*, and others.

Vocabulary

A *valid* argument is one which is phrased such that true premises ensure a true conclusion.

A *sound* argument is a *valid* argument with true premises.

A *persuasive* argument is a *valid* argument with obviously true, or previously accepted, premises.

Guided Practice

Describe the form of the logical arguments in questions 1-5.

1. If a bull has been gelded, it is a steer. Ferdinand is not a steer. Therefore, Ferdinand is not a gelded bull.
2. Ignoring problems makes them go away. I ignore my problems. My problems go away.

- Arguments must be valid or invalid. This argument is not valid. This argument is invalid.
- If you study Reason, you will better understand Logic. If you better understand Logic, you will make better use of Statistics. Therefore if you study Reason, you will make better use of Statistics.

Solutions:

- This argument is in the form: If P , then Q . Not Q , therefore not P . It is an example of the form *Modus tollens* or *denying the consequent*.
- This argument is in the form: If P , then Q . P , therefore Q . It is an example of *Modus ponens* or *affirm by affirming*.
- This argument is in the form: P or Q , not P , therefore Q . It is an example of *Disjunctive syllogism*.
- This argument is in the form: If P , then Q . If Q , then R . Therefore, if P , then R . It is an example of: *Hypothetical syllogism*, also known as *The Chain Argument*.

Practice

Describe the form of the logical arguments in questions 1-13.

- If P , then Q . P , therefore Q .
- If P , then Q . Not Q , therefore not P .
- If P , then Q . If Q , then R . Therefore, if P then R .
- P or Q . Not P . Therefore Q .
- If it is snowing, then it is below freezing. It is snowing, therefore it is below freezing.
- If your homework is not done, you cannot go out. You are going out, therefore your homework is done.
- If you wear drink too many energy drinks, you will not be able to sleep. If you aren't able to sleep, you will be tired tomorrow. If you are tired tomorrow, you won't do well on your exam. Therefore, if you drink too many energy drinks, you won't do well on your exam.
- You are productive or lazy. You are not lazy. Therefore you are productive.
- The sky is either cloudy or clear. The sky is not clear. Therefore the sky is cloudy.
- Minivans get good gas mileage. Bob drives a minivan. Bob gets good gas mileage.
- Girls drive pink cars. Sam does not drive a pink car. Therefore Sam is not a girl.
- If you eat too much candy, you will get cavities. If you get cavities, you will have to spend money on the dentist. If you spend money on the dentist, you cannot go to the movies. Therefore, if you eat too much candy, you cannot go to the movies.
- If you tell students to come in from recess, you are a teacher. You are not a teacher. You do not tell students to come in from recess.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 12.4.

12.5 Hidden Premises

Objective

Here you will learn about hidden premises in logical argument.

Concept

Consider the argument below:

“You should eat Veggie-O’s for breakfast because they contain more frammisspilts than other cereals”.

This argument is not valid as it stands, but why not? Is something missing?

See the end of the lesson for the answer.

Guidance

One of the most important skills to learn in order to become skilled with logic and reason is to understand the concept of the *hidden premise*. Many, many arguments contain a hidden premise, and, although it can be used as a sort of sneaky way to avoid an obvious flaw in a line of reasoning, a hidden premise does not necessarily make an argument invalid. The trick to handling a hidden premise is to recognize it right away for what it is, and then state it clearly so that it may be correctly included in your evaluation of the argument.

A *hidden premise* is a premise that is required in order to reach the stated conclusion, but is not itself stated clearly in the argument. Consider the following:

“My bag of candy is better than yours, because mine has more red pieces”.



This is not a valid argument as written, what is wrong with it?

Let’s break it down and see:

Premise 1: My bag of candy has more red pieces

Hidden premise: Red candy pieces are better than other-colored pieces.

Conclusion: My bag of candy is better than yours.

Without the assumption of the *hidden premise*, the conclusion makes no sense, and the argument is *invalid*. In order to make a decision about the soundness of the argument, you will need to decide if you accept the premise “red candies are best”. If you agree that “red candies are best” is a viable premise, the argument is sound, and the conclusion is reasonable. If you believe that yellow candies are better than red ones, then you will obviously reject the premise, and the conclusion will no longer seem reasonable. Regardless of your feelings about red candy, however, the important point here is that you *must* take the hidden premise into account as you evaluate the argument.

Example A

Consider the following:

“We should reduce the penalty for drunk driving, it would result in more convictions”

What hidden premise(s) are in this argument?

Solution:

Let’s break down what we have:

Premise 1: Reducing the penalty for driving drunk would result in more convictions

Conclusion: We should reduce the penalty.

Something is missing, isn’t it? The logic seems like it might be ok, but there is an important premise that is assumed to be true, but unstated:

Hidden premise: More convictions for drunk driving is better.

Now it makes sense. Assuming the hidden premise is solid, the argument may be considered.

Example B

Consider the following:

“It should not be illegal to smoke pot, I know it does not harm anyone”.

What hidden premise is this argument hinged upon?

Solution:

Let’s break it down:

Premise: I know smoking pot is harmless

Conclusion: Smoking pot should not be illegal

What is missing? The fact that harmless and legal are not the same thing.

Hidden premise: Anything I consider harmless should be legal

Now the weakness in the argument is much more apparent. While it may be a challenge for my opponent to prove that smoking pot is harmful, he or she should easily be able to demonstrate that my personal beliefs should not be consulted before the passage of every single law!

Example C

Consider the following:

Everyone should drink raw cow’s milk, because it is natural and not processed.

What is the hidden premise?

**Solution:**

Break it down:

Premise: Raw milk is natural

Premise: Raw milk is not processed

Conclusion: Everyone should drink raw milk

What is missing? The assumption that natural and unprocessed are preferable for everyone.

Hidden premise: It is better for everyone to drink things that are natural and unprocessed.

Concept Problem Revisited

“You should eat Veggie-O’s for breakfast because they contain more frammisspilts than other cereals”.

This argument is not valid as it stands, but why not? Is something missing?

The hidden premise here is the assumption that “frammisspilts” are important or desirable.

Vocabulary

A **hidden premise** is a premise that is required in order to reach the stated conclusion, but is not itself stated clearly in the argument.

Guided Practice

Questions 1-3 refer to the following:

“No one wants to kiss a person with bad breath, therefore you shouldn’t smoke”.

1. Is the argument valid as written?
2. Is the argument sound as written?
3. Is there a hidden premise? If so what is it?
4. Is the argument sound if the hidden premise is accepted?

Solutions:

1. No, even if the premise “No one wants to kiss a person with bad breath” is false, the conclusion may be true, and vice versa.

2. No, an invalid argument cannot be sound.

3. Yes, there are actually *two* hidden premises:

Smoking causes bad breath

Having people want to kiss you is desirable.

4. Yes, the argument is sound if the hidden premises are accepted:

No one wants to kiss a person with bad breath

Smoking causes bad breath

Being kissable is desirable

You should not smoke

Practice

Questions 1-4 refer to the following:

“Abortion is morally wrong because it is murder”.

1. Is the argument valid as written?
2. Is there a hidden premise? If so what is it?
3. Is the argument sound if the hidden premise is accepted?
4. Rewrite the argument with the hidden premise stated.

Questions 5-8 refer to the following:

“Great actors make great movies. Will Smith is a great actor. Therefore *Legend* must be a great movie”

5. Is the argument valid as written?
6. Is there a hidden premise? If so what is it?
7. Is the argument sound if the hidden premise is accepted?
8. Rewrite the argument with the hidden premise stated.

Questions 9-12 refer to the following:

“You should get your hamburger from Christie’s Corner Market, because they sell grass-fed beef”

9. Is the argument valid as written?
10. Is there a hidden premise? If so what is it?
11. Is the argument sound if the hidden premise is accepted?
12. Rewrite the argument with the hidden premise stated.

Questions 13-16 refer to the following:

“Diet Cola is bad for you because it contains as part a me”

13. Is the argument valid as written?
14. Is there a hidden premise? If so what is it?
15. Is the argument sound if the hidden premise is accepted?

16. Rewrite the argument with the hidden premise stated.

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 12.5.

12.6 Structural Fallacies

Objective

Here you will learn how to recognize some of the *formal fallacies* in logic, both to strengthen your own arguments and help you identify weaknesses in others.

Concept

Consider the following statement:

“A black cat ran across the road on my way to school last Thursday and I had a horrible day, therefore black cats are bad luck”.



What is wrong with this argument? A black cat *did* cross my path, and I *did* have a bad day afterward, so both premises are true, but the conclusion is suspect. What went wrong with the argument?

See the end of the lesson for the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67553>

<http://youtu.be/WmIb2Jb-KC8> Michael Austin - Logical Fallacies

Guidance

Logical arguments are practically everywhere you look. Humans, almost by definition, are self-aware creatures with the ability to reason and the desire to share their reasoning with others. Because of this tendency, it is very valuable

to be more than a little bit familiar with the rules of valid argument, and the types of logical fallacies that make arguments invalid.

In this lesson, we will practice identifying some common *formal fallacies*. It is important to note that identifying an argument as invalid because it follows the form of a common fallacy may require that you first reconstruct the argument in a standard form, since arguments often rely on unstated *hidden premises* (see Example A).

Common Fallacies:

- **Affirming the Consequent:** If A then B. B, therefore A.
 - If it is snowing, I wear my boots. I am wearing my boots, therefore it is snowing.
 - Just because I wear my boots when it is snowing does not mean I don't also wear my boots for some other reason.
- **Appeal to Ignorance:** Use the absence of proof for a premise as evidence in favor of the opposing premise.
 - There are no fossilized remains of a winged snake, so snakes must not have evolved into birds.
 - The lack of proof of winged snakes is not, in and of itself, proof either for or against the evolution of snakes to birds.
- **Diversion:** Trying to support one premise by arguing for other premise.
 - ABC Dog Food is flavored with beef-like flavoring. According to studies, dogs choose hamburgers 3:2 over chicken tenders, so ABC Dog Food is the best.
 - Showing that dogs prefer hamburger to chicken tenders is not evidence that ABC Dog Food tastes better than any other dog food.
- **Equivocation:** Using one meaning of a word in the premise, and another in the conclusion.
 - Criminal actions are illegal. All murder trials are criminal actions. Therefore all murder trials must be illegal.
- **Coincidental Correlation** (also known as “post hoc ergo propter hoc”, which means “after this, therefore because of this” or just “post hoc”): Falsely assuming that just because one thing occurs after another, it must have been *caused* by the other.
 - Public school attendance has skyrocketed in the past 10 years, and so has the number of kids in juvenile hall, so school must be corrupting children.

Example A

Identify the logical fallacy in the argument below:

The once-blind man could obviously see, since he picked up his hammer and saw.



Solution:

This is an example of **equivocation**, since the premise uses the word “see” to describe the ability to perceive an object with one’s eyes, while the conclusion uses the word “saw”, meaning the cutting implement rather than the past tense of the word in the premise.

Example B

Identify the logical fallacy:

My dad refused to pay me the allowance I earn by doing my chores, even after I proved that gas prices have gone up by \$0.25 per gallon. He just kept pointing out that my chores weren’t done. I think he should pay me extra so I can afford gas.

Solution:

This is an example of a **diversion**. The allowance is based on the completion of chores, so any evidence of an unrelated premise such as gas pricing is not going to strengthen the argument that allowance should be paid.

The fallacy is clear if the argument is stated in a standard form:

Premise 1: Dad refuses to pay allowance

Premise 2: Allowance is based on chores

Premise 3: Gas prices have increased

Conclusion: Allowance should be paid, and increased

Stated this way, it seems pretty clear that the conclusion is based only on the unrelated premise that gas prices have increased, rather than on the valid premises that allowance is linked to chores and that chores weren’t done.

Example C

Identify the logical fallacy:

Scientists have been trying for years to prove that ghosts do not exist. Since there is no proof yet that they don’t exist, they must be real.

Solution:

This is an **Appeal to Ignorance**. The premise is that the lack of proof *against* ghosts may be taken as proof *for* the existence of ghosts.

Concept Problem Revisited

“A black cat ran across the road on my way to school last Thursday and I had a horrible day, therefore black cats are bad luck”.

What is wrong with this argument? A black cat did cross my path, and I did have a bad day afterward, so both premises are true, but the conclusion is suspect. What went wrong with the argument?

This is an example of “post hoc”, meaning that the conclusion is based on the false assumption that the bad day that occurred after the black cat crossed the path was *caused* by the black cat.

Vocabulary

Formal fallacies are fallacies based on the *form* of the argument. In the case of a formal fallacy, the conclusion may or may not be true, but it does not follow from the premises.

A **hidden premise** is a premise that is not explicitly stated, but must be assumed to exist based on the wording of the argument.

Guided Practice



Consider the following statements:

Damien said he would ask Carrie to the dance if he won the lottery. Damien is at the dance with Carrie, so he must have won the lottery.

1. What are the premises to the argument?
2. What is the conclusion?
3. Is the argument valid?
4. What fallacy, if any, is demonstrated?

Solutions:

1. Premise 1: Damien said he would ask Carrie to the dance if he won the lottery Premise 2: Damien is at the dance with Carrie.
2. Conclusion: Damien won the lottery.
3. The argument is invalid. We can tell because it fits the form of a common formal fallacy.
4. The conclusion is based on the logic: If A, then B. B, so A. This is the fallacy known as **affirming the consequent**.

Practice

Questions 1-4 refer to the following argument:

If it is sunny outside, I wear sandals. I am wearing sandals, so it must be sunny outside.

1. What are the premises in the argument?
2. What is the conclusion?
3. Is the argument valid?
4. What fallacy, if any, is demonstrated?

Questions 5-8 refer to the following:

You can't prove I threw the water balloon, so my sister must have done it.

5. What are the premises in the argument?
6. What is the conclusion?

7. Is the argument valid?
8. What fallacy, if any, is demonstrated?

Questions 9-12 refer to the following:

I teach math and science classes. Physics is a science class and everyone thinks it is the coolest subject in science. Therefore I am the best teacher in school.

9. What are the premises in the argument?
10. What is the conclusion?
11. Is the argument valid?
12. What fallacy, if any, is demonstrated?

Questions 13-16 refer to the following:

My mom always told me not to talk to strangers. You are as strange as anyone I know, so mom wouldn't want me to speak with you.

13. What are the premises in the argument?
14. What is the conclusion?
15. Is the argument valid?
16. What fallacy, if any, is demonstrated?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 12.6.

12.7 Content Fallacies

Objective

Here you will learn about logical fallacies involving the *content* of an argument, as opposed to the *structure*.

Concept

Consider the following statements:

“Hitler was a bad person, and he had a mustache, so mustaches are bad”.

“She thinks that movie was great, but she is stupid, so the movie must be bad”.

“He thinks that skiing is fun, but he believes in UFO’s, so skiing must be boring”.



All three of these arguments exemplify the same form of logical content fallacy. What fallacy is at play here, and how can it be avoided?

See the end of the lesson for the answer.

Watch This



MEDIA

Click image to the left or use the URL below.

URL: <https://www.ck12.org/flx/render/embeddedobject/67556>

<http://youtu.be/AiUrSFAIktY> Michael Austin - Logical Fallacies, Part Two

Guidance

There are two broad classifications of logical fallacy: fallacies of structure and fallacies of content. In this lesson, we will consider **content fallacies**, also known as **informal fallacies**.

A content fallacy is a logical fallacy that is not due to the way the argument is stated, but rather due to what the argument actually says. Although there are effectively infinite ways to devalue an argument by using faulty content of one sort or another, there are some types of content fallacies that are common enough to warrant particular consideration. Learning to recognize the more common types of content fallacies can greatly simplify the process of identifying faulty arguments.

Common Content Fallacies:

- **Ad Hominem:** This fallacy is committed when an argument is based on the perceived failings of an adversary.
 - My sister likes that book, and she is annoying. The book must be bad.
- **Bandwagon:** This is an argument based on the concept that the majority is always right.
 - That video has 100,000 hits, it must be really good!
- **Begging the Question (circular argument):** An argument that assumes the truth of its conclusion.
 - Executions are moral because we must have a death penalty to discourage violent crime.
- **False Dilemma:** An argument which over simplifies a complex situation into only two possible alternatives.
 - Bad people make bad decisions, good people good ones. I lied once, so I must be a bad person.
- **Non-Sequitur:** An argument where the conclusion is not based on the premises.
 - I am a math teacher, so I know all about fashion.
- **Straw Man:** An argument based on misrepresenting the opponent's argument so it may be easily defeated.
 - “*Straw man* has always been a stock-in-trade of advertisers.... A Post Office commercial once pictured competitors trying to deliver packages with rickety old planes that fell apart on camera.” (H. Kahane and N. Cavender, *Logic and Contemporary Rhetoric*. Wordsworth, 1998)

Example A

Consider the following argument, what content fallacy does it represent?

“He thinks Ferraris are the best cars, but he likes VW Bugs, so what does he know”?



Solution:

Begin by breaking down the argument and rewriting it in a standard form:

- Premise 1: He thinks Ferraris are the best sports cars
- Premise 2: He likes VW Bugs
- (assumed premise): VW Bugs are obviously bad cars, anyone who likes them must not know anything about nice sports cars.
- Conclusion: Ferraris must not be good cars.

The conclusion that Ferraris are not good is based on the premise that “He” is a bad judge of cars so any car he likes can’t possibly be any good. This is a clear example of **Ad Hominem**, since the premise is a character attack and the conclusion has no basis in any evidence about the product in question.

Example B

Consider the following argument, what content fallacy does it exemplify?

“The last three days I walked to school and it rained, so we deserve a longer lunch break”.

Solution:

Break the argument down into a standard form:

- Premise 1: I walked to school the last three days
- Premise 2: It rained the last three days
- Conclusion: We deserve a longer lunch break

The conclusion is based on the false assumption that there is some connection between walking to school in the rain and the length of a lunch break. Since there is no apparent connection, this is an example of a **non-sequitur**.

Example C

Consider the following argument, what content fallacy does it exemplify?

“Mom, why can’t I have a slice of my birthday cake”?

“You can’t eat nothing but sugar all the time, it is unhealthy”.

Solution:

This is an example of a **Straw Man**. Mom cleverly avoided answering the initial question by setting up the argument that an all sugar diet is unhealthy - she knows that is an easy argument to win.

Concept Problem Revisited

“Hitler was a bad person, and he had a mustache, so mustaches are bad”.

“She thinks that movie was great, but she is stupid, so the movie must be bad”.

“He thinks that skiing is fun, but he believes in UFO’s, so skiing must be boring”.

All three of these arguments exemplify the same form of logical content fallacy. What fallacy is at play here, and how can it be avoided?

All three arguments above are examples of **Ad Hominem**, which means they are based in a personal attack on a person. This fallacy is easily avoided by not basing an argument on a perceived flaw in an opponent.

Vocabulary

A **content fallacy** or **informal fallacy** is a logical fallacy based on what is stated in the premises, rather than the form in which they are presented.

A **formal fallacy** is a logical fallacy that is based on the form in which the argument is presented.

Guided Practice

“You are either a winner or a loser. Winners eat Yummy-O’s cereal! Are you a winner?”

1. What are the premises of this argument?
2. What is the conclusion?
3. What fallacy is represented?

Solutions:

1. Premise 1: You are a winner or a loser

Premise 2: Winners eat Yummy-O’s cereal

Hidden premise: It is good to be a winner

2. You are a loser if you don’t eat Yummy-o’s

3. This is a *false dilemma*, by making it look as if there are only two types of people, winners that eat Yummy-O’s and losers that don’t, you are set up to believe the conclusion that you must eat the cereal to be a winner. The other possibilities, that you might be a winner that does not eat Yummy-O’s, or a loser that does, are not represented.

Practice

Questions 1-4 refer to the following:

“Bob says that we should not fund the proposed laser defense program. I disagree entirely. I can’t understand why he wants to leave us defenseless like that”.

1. What are the premises to the argument?
2. What is the conclusion?
3. Is the argument valid?
4. What fallacy, if any, is demonstrated?

Questions 5-8 refer to the following:

“I am the best player in school because no one is better than I am”.

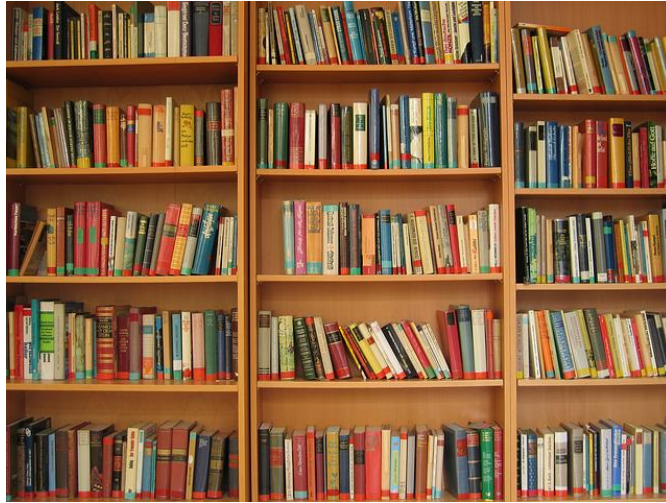
5. What are the premises to the argument?
6. What is the conclusion?
7. Is the argument valid?
8. What fallacy, if any, is demonstrated?

Questions 9-12 refer to the following:

“Karen says that being a vegetarian is great, but she is crazy anyway”.

9. What are the premises to the argument?
10. What is the conclusion?
11. Is the argument valid?
12. What fallacy, if any, is demonstrated?

Questions 13-16 refer to the following:



“Reading encourages you to use your imagination more than TV, so there should be more comic-book stores in town”.

13. What are the premises to the argument?
14. What is the conclusion?
15. Is the argument valid?
16. What fallacy, if any, is demonstrated?

Answers for Explore More Problems

To view the Explore More answers, open this [PDF file](#) and look for section 12.7.

Students were introduced to the concepts of argument and inductive and deductive reasoning. Methods of evaluating argument validity such as Euler diagrams and rewriting premises and conclusions were also introduced and practiced. Students learned about structural and content fallacies and how to identify hidden premises. They practiced recognizing valid and invalid forms of argument throughout the lesson.

12.8 References

1. bngnaranjo. <http://pixabay.com/en/girl-sitting-sunset-rest-relax-325399/?oq=sitting> .
2. . . CC BY-NC-SA
3. . . CC BY-NC-SA
4. Roberto Trm. https://www.flickr.com/photos/massimo_riserbo/5268982083 .
5. Ramunas Gečiauskas. <https://www.flickr.com/photos/qisur/4351188236> .
6. [U+9673] [U+30DD] [U+30FC] [U+30CF] [U+30F3]. <https://www.flickr.com/photos/pohan-camera/3732165202> .
7. Nikita. <https://www.flickr.com/photos/malfet/1413379559> .
8. Robert Scoble. <https://www.flickr.com/photos/scobleizer/4695901494> .
9. Graniers. <https://www.flickr.com/photos/graniers/6824787235> .
10. . . CC BY-NC-SA
11. Vladimir Agafonkin. <https://www.flickr.com/photos/mourner/3273125130> .
12. James Bowe. <https://www.flickr.com/photos/jamesrbowe/6371964415/> .
13. Morgan. <https://www.flickr.com/photos/meddygarnet/3087083501> .
14. OpenClips. <http://pixabay.com/en/ufo-flying-saucer-flying-disc-alien-146541/?oq=ufo> .
15. picturemaker123. <http://pixabay.com/en/vw-beetle-beetle-auto-old-vw-247925/?oq=red%20vw> .
16. geralt. <http://pixabay.com/en/book-books-bookshelf-read-67049/?oq=books> .