Module 2: Statistics as a Managerial Tool

Module Two: Statistics as a Managerial Tool

2.01 Learning Objectives

Learning Objectives

After completing this module, you should be able to:

1. Describe how statistics are used in different settings
2. Describe common problems with and misuse of statistics
3. Identify criteria for evaluating statistics
4. Explain the key fundamentals of probability and their real-world application
5. Identify the fundamental concepts of descriptive statistics (populations and samples, measures of central tendency, measures of variability, measures of distribution) and their real-world application
6. Select appropriate graphic methods for displaying descriptive statistics
7. Explain the fundamental concepts of inferential statistics and their real-world application
8. Evaluate a scenario in order to determine the appropriate statistic to use
9. Apply fundamental statistics to a real-world situation
10. Evaluate the appropriateness of statistics used
11. Use statistics to identify the most appropriate decision alternative
12. Translate statistical data into a graphical presentation based on a brief case study

2.03 Video: Statistics for Management

This assignment does not contain any printable content.

2.04 The Misuse of Statistics

Common Misuses of Statistics

Using statistics to analyze data provides a great tool to anyone making decisions. However, it is extremely easy to misuse or misunderstand statistics in application. The misuse of statistics can happen due to a lack of understanding of the appropriate methods to apply or the correct way to interpret results. In some cases, statistics are misused deliberately to try to skew results or misrepresent data. You should always be cautious when reading a report from a statistical study or presenting the results yourself.

The following is a list of common misuses or misunderstandings of statistics that often occur. Please click on each tab to learn more.

Bias in the Data

There are several ways data can be biased before any tests are applied. There are two main types of bias: selection bias and response bias. Each of these types can be further broken down into different categories.

Selection Bias

Selection bias refers to a bias of who is included in the study. In other words, selection bias involves biases present within the sample. When selecting subjects for an experiment or survey, it is important that the sample is representative of the whole population. Otherwise, the study should not be used to make inferences about the entire population. For example, you might not get a good understanding of the spending habits of people in the state of California if you only survey people from Beverly Hills.

Sampling bias occurs when a taken sample is not representative of the whole population.

Example: A survey is sent to customers who shop online via email. The results of this survey would not represent the portion of the population who chose to shop exclusively in person. Therefore, the sample is missing part of the larger population.

Non-response bias typically applies to studies involving surveys when those who respond differ greatly from those who do not respond.

Example: A telephone surveyor asking about the use of over-the-counter antihistamines might only get those who suffer from hay fever to continue to fill out the survey, while people unaffected by allergies might not continue. The respondents (those who suffer from hay fever) and the non-respondents (those who are unaffected by allergies) differ greatly, so the data would be skewed towards those who suffer from allergies.

Convenience sampling occurs when a sample is taken from an easily accessible group.

Example: While testing the quality of items produced, you take a sample of the first 50 items you see. This sample may not be representative of the whole population of items you want to test.

Voluntary bias occurs when members of the population are allowed to decide for themselves whether they want to be in the sample or not.

Example: A movie theater wants to get feedback on customer experience at the movies. They print a survey website on their receipts and make sure the cashiers point out the survey to all guests. Those with strong opinions (both good and bad) are more likely to respond to the survey, so the results may be missing a large portion of the population.

Survival bias is the result of taking a sample of only those who make it past some selection process.

Example: A manager researching successful strategies for their business would likely look at companies where those methods were successful and ignore companies that failed. Companies which no longer exist are likely to be excluded from studies, which can skew results in a more positive way.

Response Bias

Response bias is the result of study participants responding in a way that might not reflect their true opinion. This bias is typically found when conducting surveys. Respondents may say what they believe the "correct" response is based on the survey purpose or any social pressures. For example, a participant may answer a survey on drug use inaccurately to reflect what is socially acceptable. Participants can also be influenced to give misleading answers as a result of the wording of a question or even by how previous questions were worded.

Demand bias occurs when participants change their behavior based on what they think the study is designed for.

Example: After hearing rumors that the study is meant to identify unfavorable working environments for factory workers, participants may report that they have experienced these unfavorable environments because they think that's what the researcher wants to hear. A participant may form an idea of what the experimental hypothesis is and try to confirm or contradict it.

Social response bias occurs when participants report behaviors they think are socially acceptable or desirable.

Example: A clothing company is conducting research on the habits of its customers and asks for information about income, activity level, and education level. Participants of the survey are likely to select options they think make them look good, which can skew the data.

Extreme responding occurs when participants respond with extremely positive or negative responses. On the other hand, neutral bias occurs when participants respond consistently with neutral responses.

Example: A survey containing responses such as "strongly disagree," "disagree," "neither agree nor disagree," "agree," and "strongly agree" may result in respondents exclusively choosing "strongly disagree" or "strongly agree" in extreme responding. Participants will choose something like "neither agree nor disagree" in neutral responding. This may be common in surveys containing Likert scales.

Acquiescence bias occurs when participants respond in agreement with all questions. On the other hand, dissent bias occurs when participants respond in disagreement with all questions.

Example: A survey containing responses such as "strongly disagree," "disagree," "neither agree nor disagree," "agree," and "strongly agree" may result in respondents exclusively choosing to agree to some degree in acquiescence bias. In dissent bias, participants will disagree to some degree for all questions. These respondents will agree (or disagree) to all items, even if some of these responses contradict others. This may be common in surveys containing Likert scales.

Questions in a survey that influence participants to respond in a certain way are called leading questions.

Example: A survey question reads, "Studies have shown that building roundabouts in intersections reduces the number of accidents. Do you support the city spending taxpayer money to fund a project to build a roundabout at the intersection of Main Street and South Street?" Participants are more likely to respond with support for the roundabout because they agree with reducing accidents, even if they do not support the project.

Using Incorrect Methods

There are many steps to conducting an experiment and analyzing the results. If incorrect methods are used at any step of the process, the results may be inaccurate. Incorrect inferences based on these results could mean severe consequences for a company.

Sampling Methods

It is important to consider how a sample is chosen for a study. In addition to various forms of selection or sampling bias to be cautious of, one must consider the method used to obtain a sample from the population. There are several ways to choose a sample, but using a simple random sample is the most straightforward and commonly used method.

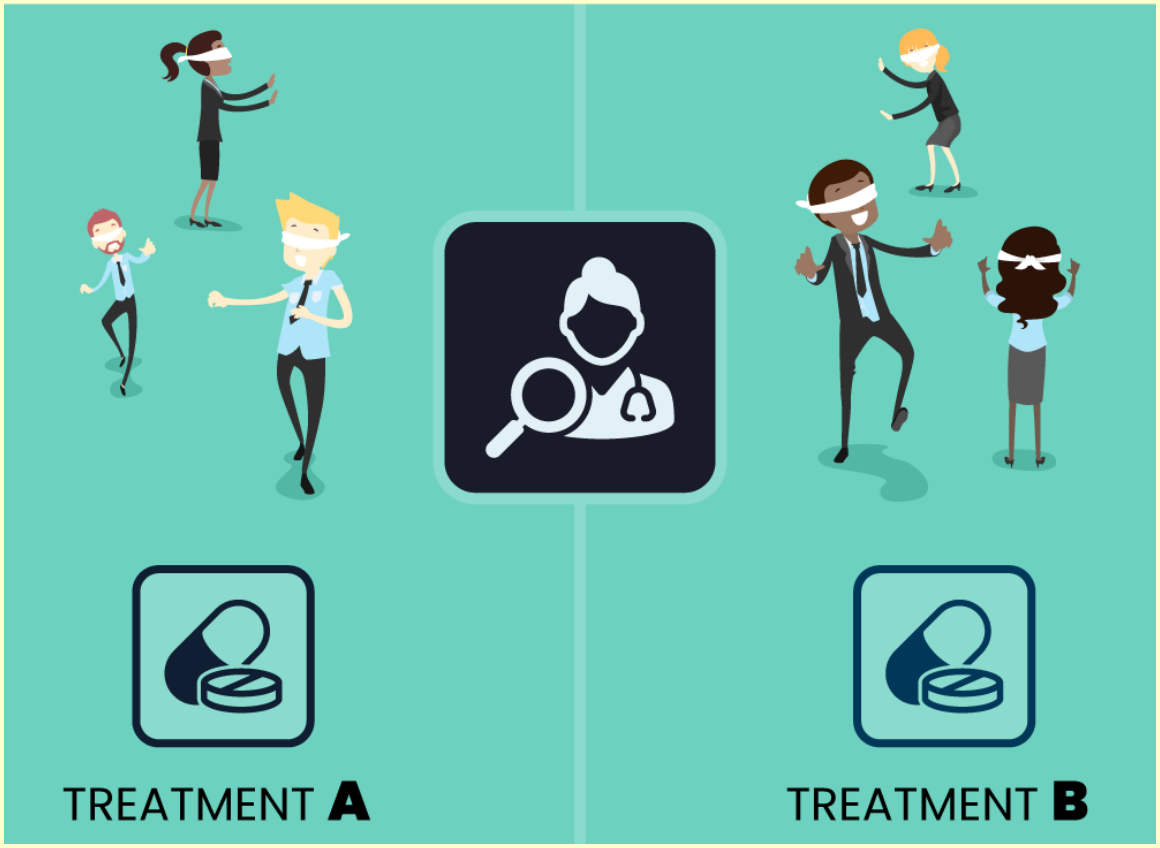
While conducting a study, it is important to consider the size of the sample used. Small samples are not ideal for making inferences about a population. For example, asking 10 New York City residents to report their favorite cuisine is unlikely to give results that are representative of the entire population of New York City residents. A much larger sample is necessary to make inferences about the population in this case.

Blinding

A lack of blinding can cause bias to occur. Blinding is when researchers place barriers between themselves and subjects in order to ensure that the researchers do not influence subjects' behavior during the experiment. Without blinding, subjectivity can be introduced into the results.

Example

Alexa is a doctor trying to determine whether a new drug is effective for patients with a particular type of depression. The experiment group receives a pill bottle that is clearly marked "experiment group." Since the patients know that they are in the experiment group, they are likely to report more of an effect of the drug.



Operationalization

Operationalization refers to the development of specific research procedures that allow for observation and measurement of abstract concepts. For example, if a researcher wants to study how new parents feel about their financial security, they can operationalize this objective by determining a testable hypothesis, developing a mechanism for collecting observations (for example, a survey), identifying a representative sample (200 randomly selected parents of newborns in suburban towns), asking relevant questions, and interpreting the data received. Each of these research decisions (hypothesis, method, sampling, questions, interpretation) is an example of operationalization.

A key aspect of operationalization is defining variables and attributes that adequately represent the concept of the study. For example, to assess attitudes about financial security, researchers might ask respondents to characterize their feelings regarding their financial security as "anxious," "confident," and so forth. These responses can be coded as "1," "2," "3," etc. for analysis. Or they could ask respondents to specify how much money they have in a college savings plan. In both examples, these are measurable dimensions that serve as a proxy for the non-measurable concept the researcher is studying.

If the researcher's reasoning behind any aspect of operationalization is faulty, it can result in misleading or irrelevant findings. For example, data showing high amounts in college savings plans may not necessarily correspond to a strong feeling of financial security. The reasoning making this connection is flawed and leads to inaccurate conclusions.

Example

Datatronics Corporation wants to determine the level of worker satisfaction of its employees. In order to assess a value to this abstract quality of "satisfaction," its research department decides to measure how often employees smile throughout the day. Unfortunately, these employees are customer-facing and know that their supervisors expect them to smile at customers, so smiling is not necessarily an indication of satisfaction for these employees.

Errors in Interpreting Data

It can be easy to misinterpret data based on preconceived ideas about an experiment. Incorrect assumptions or conclusions can be detrimental to a company.

Association and Causality

When two variables are correlated, it is easy to assume that changes in one variable cause changes in the other. However, this is not true in all cases. A common misunderstanding of statistics occurs when a researcher notices a relationship between two variables and assumes that one variable is the cause of the other. In reality, these variables *might* both be caused by a separate variable. In this case, they would merely be correlated, which means they show up together. Or there might be no relationship at all.

Example

A nonprofit sees an upswing in donations after running a media campaign to reach its end-of-the-year goal. The marketing director proclaims the campaign a huge success and promotes the employee who designed the campaign. In reality, the marketing campaign was mediocre, and in fact, the primary motivations for increased donations were the holiday season and the end of the tax year.

Unfounded Assumptions

An unfounded assumption is when an assumption is made that has not been proven.

Example

Middletown has a very large number of churches, synagogues, and mosques. You conclude that residents of the town attend religious services on a consistent basis. This assumption is unfounded—it may be that attendance in the community's houses of worship has actually declined.

2.05 Criteria for Assessing Statistics

Evaluating Statistics

When evaluating statistics, there are a number of questions to ask. The following questions should help determine if the statistic could be useful.

"Was the experiment measuring what the researcher wanted to study?"

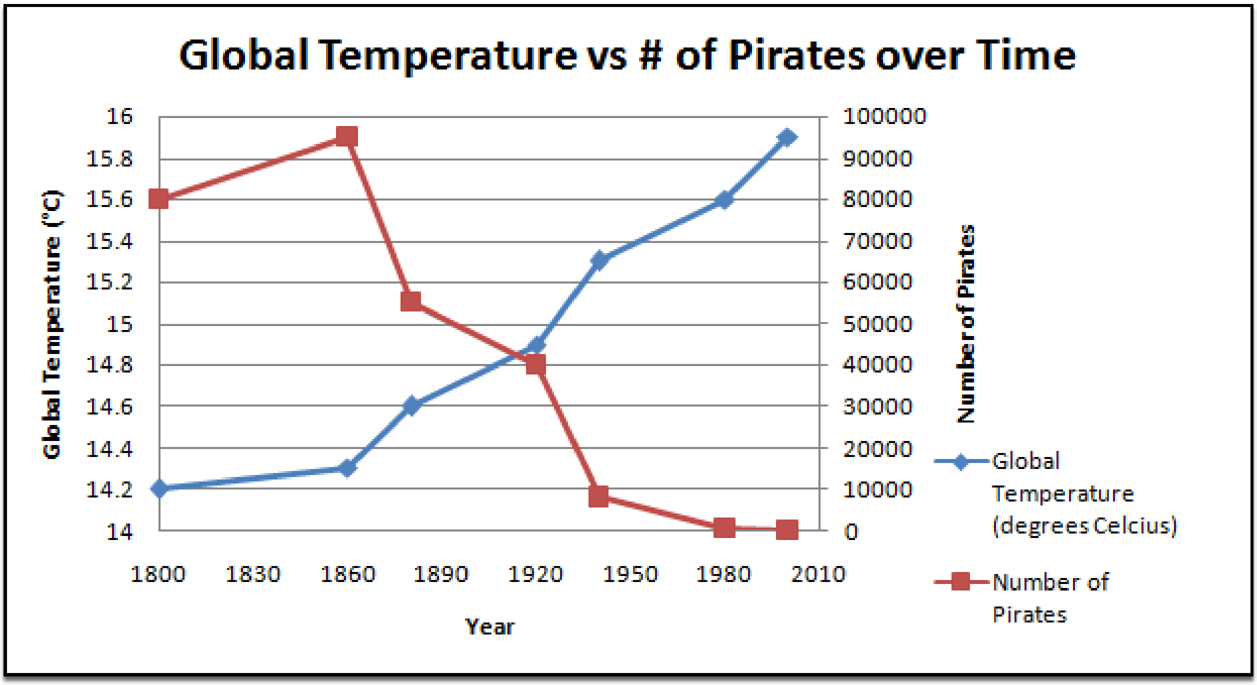
Researchers should be careful to design experiments so that they measure what is intended to be measured. For example, a researcher might use a company's employee turnover rate to represent the overall level of employee satisfaction. If, however, high turnover is due to a large number of employees retiring, it is not necessarily an accurate indicator of employee satisfaction. When designing an experiment, researchers must select variables and attributes that are observable and measurable. At the same time, it is important that they follow a coherent and logical framework to ensure that all aspects of experiment design and method (including research hypothesis, data collection, sample, variables and indicators, data analysis and interpretation) accurately replicate the original research problem.

Was a causal relationship truly existent?

Determining if an action is causing another action is important in evaluating the conclusions made. A high correlation might only signify they are associated and not that one action causes the other.

Here is an example of this.

Let's examine the decrease in pirates over the same time period that there has been an increase in global warming. The following graph shows the relationship.



This is evidently not causation, but there are variables that are related while neither causes the other. As a simpler example, people put on more clothes when leaves fall. Although this is generally the case, neither action causes the other; they are associated.

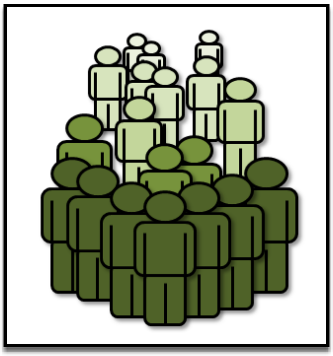
The important thing to remember here is that it's not always clear what is causing a certain outcome. A correlation in data is not enough to conclude that a variable affects another variable. There is a reason for a certain result, and it is important to understand the path that led from the cause to the effect.

Was the sample for the experiment representative of the stated population? Were there parts of the sample or population that were missing?

If you're making an inference about a population, your sample should represent the population by being proportionally distributed through each demographic that might give different responses. If the experiment is applied to a good representative sample, it is still possible that the sample will not respond evenly. This can occur when data is missing, or some respondents choose not to respond. Missing data is a serious error and could cause the conclusions drawn from your experiment to be called into question.

While missing data is common in research studies, it is important to understand the source(s) of missing data and whether it is random or non-random. If the missing data is random across the sample, it will not harm the validity of the data except to reduce the sample size. If, however, missing data is non-random (for example, everyone between the ages of 20 and 40 failed to respond), then it will negatively impact the extent to which the sample responses replicate the population.

While it is difficult to identify if missing data is skewing the results of a study, one possible approach is to obtain some basic statistics for the entire population (geographic mix, demographics, household characteristics) and compare those same statistics for the sample that responded. If the distributions for the population and sample are similar for these statistics, then there is a good chance that the sample replicates the population well.



To ensure a valid sample accurately represents the population, the sample size must be large enough to contain the variation that exists in the target population while allowing for some degree of random missing data. In general, the risk of non-representative samples decreases as the sample size increases. To determine an adequate sample size for a research study, it is necessary to know the confidence interval (or margin of error) that you want to achieve, as well as the confidence level (percent certainty) for your conclusion. A mathematical formula exists to calculate an appropriate sample size for a study, given a specific population size, desired confidence interval and confidence level. For a more expedient approach, sample size calculators are available on the internet to perform the same function.

The link below will open optional content on Confidence Intervals. Confidence Intervals will be covered in greater detail later in the module.

Introduction to Confidence Intervals

> A confidence interval is the range around a sample mean that has a specific probability of containing the true population mean. The "confidence" is the likelihood that a new sample will look like past findings, while the "interval" is the varying range around the existing mean that allows for the different levels of confidence. With the same data set, a greater confidence level needs a greater interval size to allow for more possible outcomes.
>
> **Example: Average Age of Employees**
>
> Hubert, a manager at a large company, is determining the average age of employees at his company. He takes a sample of 100 employees and determines that the mean age of that sample is 33.2 years and the standard deviation is 12.0 years. He wants to determine the age to a 95 percent confidence level. After using confidence intervals, Hubert can say that, with 95 percent confidence, the average age of the company's employees is between 30.85 years and 35.55 years.

Was there a possibility of bias?

An important step in evaluating an experiment is to make sure the experiment is not biased. The goal of an experiment is to collect true data. If bias appears, it will be present in a response or opinion survey or analysis, and the results of your experiment will be compromised.

Bias can occur when the respondent feels a responsibility to answer a certain way, either because they have made (possibly correct) assumptions about the entity surveying them and want to appease them, or they are asked a question in a way that is phrased to elicit a certain response.

Suppose a child is asked by his parents whether he is ready for his big math test. In this situation, the child might feel pressure to respond with a "Yes," because that is what he expects his parents want to hear. A question on a survey designed to elicit a certain response might be: "To decrease crime, should our government decrease the number of guns on the streets or should we just leave people in jail for longer periods of time?" Clearly, this question is biased in favor of gun control laws.

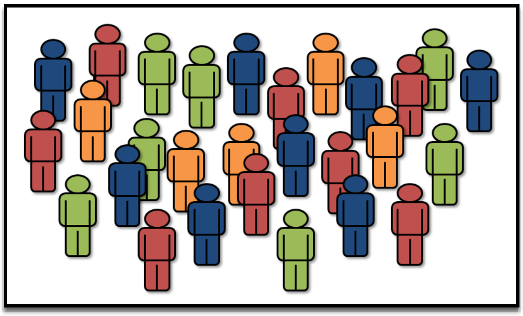
Or consider a researcher interpreting results who is invested in the experiment, and cannot be completely objective. If this is the case, the experiment has been compromised and should instead have blinds. You may recall from Module One the example of the double-blind drug trial, in which neither the treatment allocator nor the subject knew if the subject was in the control group or the treatment group. In this way, neither of the parties could bias the experiment with false expectations or fictional symptoms or treatment effects.

Was the population used in the experiment the basis for the initial hypothesis and therefore proving nothing?

This part of evaluating a statistic is important when looking at a small population. If you were to form a theory about the 10 biggest companies after knowing a lot about them, you could not conduct an experiment to test that theory with those 10 biggest companies. It is a self-fulfilling prophecy.

For a more evident example, suppose you are a farmer. You've noticed that all of your horses have a black dot on their back left leg located in the same spot. To test your theory, you take two of your horses and see if they have a black dot in the same spot. This experiment doesn't make any sense because you already know the answer.

Was the sample size big enough?



It is important when evaluating an experiment to consider the size of a sample. If a sample size is small, people are less likely to believe the theory because the greater the sample size, the greater the precision, and hopefully, accuracy in the results.

For example, Josh is an entry-level financial analyst and has received two positive performance reviews. Would it be wise, using these two data points, to recommend that Josh be hired as the firm's next CFO? With more time and more positive job performance reviews, in other words, a larger sample, the company's executive leadership would be able to make a better-informed decision about the likelihood of Josh's success as a CFO.

For different statistical tests there are minimum sample sizes for the test to be used. To form conclusions, hypothesis testing shows that the likelihood of making statements confidently is directly related to the size of the sample. Therefore, with a greater sample size, one is more likely to come to a conclusion.

Were any assumptions made that were unproven?

For a theory to have the possibility to be true, all of the ideas on which it is based have to be true. If you find the number of minutes in a week by assuming that a week has seven days, 24 hours in each day, and 70 minutes in each hour, then one of your assumptions is incorrect, and so the number of minutes you come up with, 11,760, is also incorrect. If an assumption is made for the basis of the experiment and the assumption is not proven, the entire experiment is compromised. It is important to follow the logic of how the theory was formed to make sure that no untrue assumptions were made.

2.06 Video: Probability

This assignment does not contain any printable content.

2.07 Probability

Introduction to Probability

What is probability? Put simply, it is the chance of an event occurring or happening at some time in the future. Probability is often expressed in quantifiable terms: for example, the chance of a coin landing as heads or tails. It is impossible to correctly predict whether the coin will land heads or tails, which makes the act of flipping the coin a random event. We can, however, calculate the probability of the coin landing heads or tails.

How is this valuable in a business setting? If we can calculate what *might* happen in a process, or the behavior a group of people (and, in some cases, predict the pattern of events), then it is possible to profit from that knowledge.

For example, market research on consumer preferences relies on probability. It would be impossible to survey every possible consumer to find out what they like—by using principles of probability, marketers can determine how likely it is that the likes and dislikes of a small sample of people reflect the feelings of the total population. With this data, marketers can tailor marketing plans to those preferences.

Trials and Events

When determining the probability of different variables, we need previous data. To gather this data, we perform multiple experiments, or trials, and record the results each time. Events, often represented by a capital letter, are one or a set of these results occurring. When picking a positive single digit number, an event can be picking a six ($\mathrm{E} = \{6\}$) or picking an even number ($\mathrm{E} = \{2, 4, 6, 8\}$).

Expressing Probability

Probability is represented by P(E) which means probability (P) of a certain event (E) occurring. If R= Rain and the weatherman says there is a 40 percent chance of rain, then $\mathrm{P(R)} = 0.40$.

Calculating Probability

Probability is calculated as the number of ways an event can occur, divided by the total number of possible outcomes.

For example, Bob has 5 different suits:

If Bob likes the suits equally, what is the probability he will choose to wear blue?

$$\mathrm{P}\left(\text{Bob wearing a blue suit}\right) = \frac{\text{3 blue suits}}{\text{5 suits total}} = \frac{3}{5}$$

What is the probability of Bob not wearing a blue suit?

The probability of Bob not wearing a blue suit is the same as the chance of Bob wearing a black suit$\frac{1}{5}$), *plus* the chance of Bob wearing a brown suit ($\frac{1}{5}$), or $\frac{2}{5}$

As another way to calculate this, since P(Bob wearing a blue suit) is$\frac{3}{5}$, the probability of Bob not wearing a blue suit can also be calculated as $1 - \frac{3}{5} = \frac{2}{5}$.

The opposite of an event happening (i.e., the event not happening) is called the complement of the event. The sum of the probability of an event and the probability of its complement is always equal to 1.

Independent Events

Events can vary in many ways, from flipping a coin to choosing a suit. Independent events are those that are**not** affected by other trials or events. For example, if you were to flip a coin once, that first result (either heads or tails) would not have any impact when the coin is flipped a second-time — the first event gives no indication of what could result from the second event.

Because each coin flip's result does not influence subsequent coin flips' results, the events are independent of one another.

Complementary Events

Complementary events are those with two outcomes that are the *only* possible outcomes of that event. For example, flipping a coin and it landing on either heads or tails — those are the only two possible outcomes, so the two events are complementary.

If there are more than two possible outcomes, the events are not complementary. For example, picking either a large size shirt or a medium size shirt are not complementary events. There are shirts in many sizes! You could pick a large, medium, small, extra large, etc. So picking a large size shirt or picking a medium size shirt is not complementary.

There must also be two defined events in order to have complementary events. For example, rolling a die and getting an even number would not be a complete instance of complementary events. However, rolling a die and getting either an even number or an odd number *would be* a set of complementary events; there are **only two** possible outcomes and they're both defined in the statement.

Conditional Probability

Just as events can occur independently, there are also those that occur only in the case of another. Conditional probability is the probability of an event occurring, given that another event has already occurred. These events are considered dependent events.

For example, if event A occurs, what is the probability that event B will occur? $P\left(B\middle|A\right) = \frac{P(A \text{ and } B)}{P(A)}$. Event B is *dependent* on event A occurring.

Example

A cookie jar has 4 sugar cookies and 6 chocolate chip cookies. With conditional probability, we're looking to find the probability of choosing certain cookies in different circumstances.

What is the probability that a sugar cookie is chosen first? There's $\frac{4}{10}$ or $\frac{2}{5}$ chance a sugar cookie is chosen first, but what is then the probability of choosing another sugar cookie?

There is one less sugar cookie in a jar that now has only 9 cookies. Picking a second sugar cookie has a probability of $\frac{3}{9}$ or $\frac{1}{3}$ — affected by the event before, the probability of picking another sugar cookie is smaller. The probability of this event is *dependent* on the event prior.

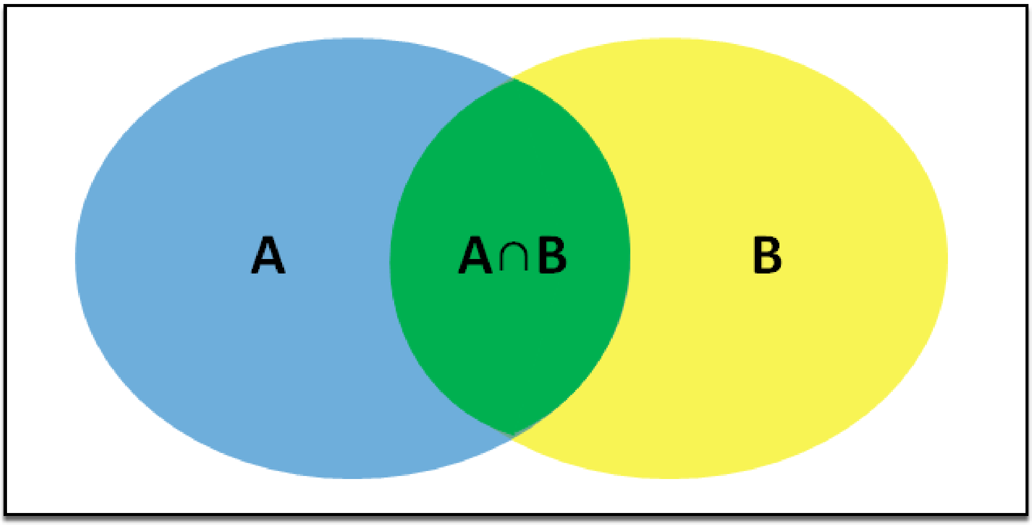
If the first cookie chosen is chocolate chip, the cookie jar would then have all four sugar cookies remaining, among the nine cookies left in the jar. Therefore, probability of choosing a sugar cookie the second time around would be $\frac{4}{9}$.

Notice how the numbers change and will continue to change as the events occur — because the circumstances are changing and the events are dependent upon one another, the probabilities also change.

Probability of an Intersection

The probability of two independent events happening is an intersection, and it can be expressed as ∩ in mathematical notation. The intersection of A and B can be written as P(A∩B).

In the image below, the events A and B are shown in a Venn diagram.



The Venn Diagram shows a visual representation of all of the possible outcomes. The area labeled A represents event A occuring, the area labeled B represents event B happening, while the area P(A∩B) represents both A and B occuring. The white area in the Venn diagram represents neither A nor B occuring. For two *independent* events, the probability of an intersection can be calculated as P(A∩B) $= P(A) \times P(B)$.

We have discussed Bob's suits, now let's look at his shoes. Suppose Bob has 3 pairs of dress shoes: 2 black pairs, and 1 brown pair. If Bob is equally likely to wear any pair of shoes, what is the chance that he will choose to wear a black suit and black shoes?

	black shoes	**black shoes**	**brown shoes**

black suit	black suit, black shoes	black suit, black shoes	black suit, brown shoes
brown suit	brown suit, black shoes	brown suit, black shoes	brown suit, brown shoes
blue suit	blue suit, black shoes	blue suit, black shoes	blue suit, brown shoes
blue suit	blue suit, black shoes	blue suit, black shoes	blue suit, brown shoes
blue suit	blue suit, black shoes	blue suit, black shoes	blue suit, brown shoes

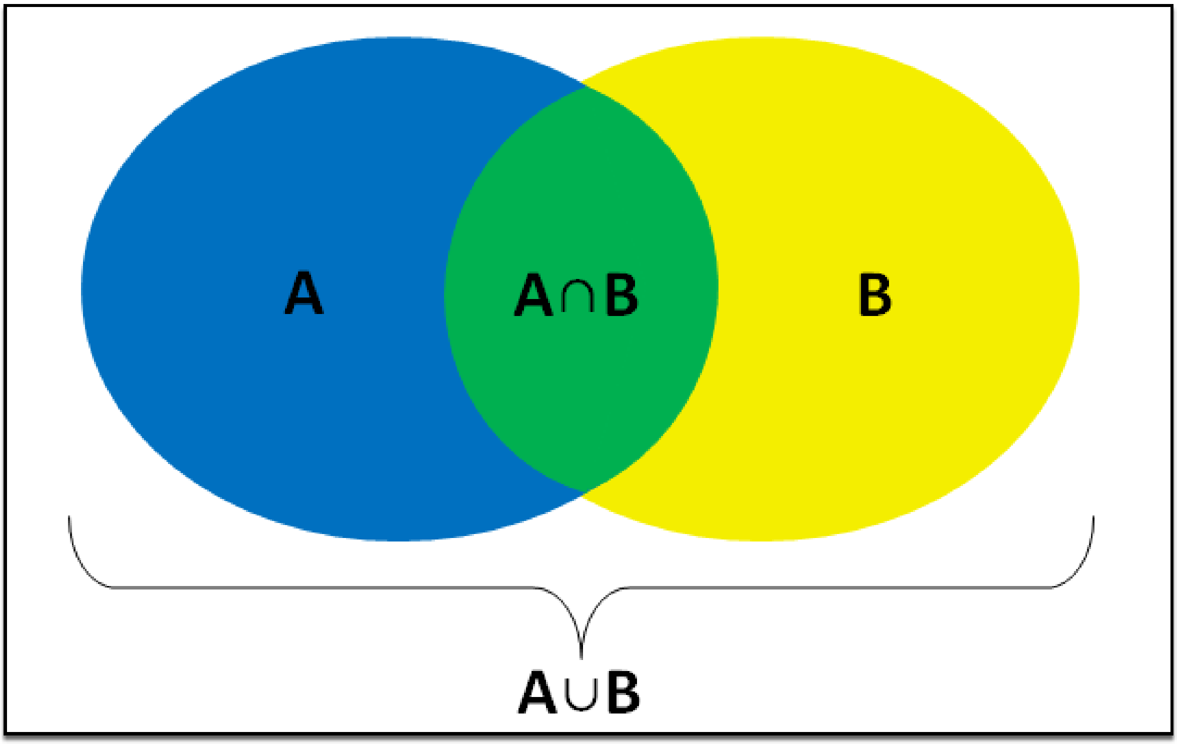
As you can see from the table above, the probability of Bob wearing a *black suit* and *black shoes* is $\frac{2}{15}$ (there are 15 possible combinations of suits and only 2 of them are a black suit with black shoes).

How can you calculate this without creating a table every time? P(A∩B)= $\mathrm{P(A)} \times \mathrm{P(B)}$

P(black suit ∩ black shoes)= $\mathrm{P(black\ suit)} \times \mathrm{P(black\ shoes)} = \frac{1}{5} \times \frac{2}{3} = \frac{2}{15}$

Probability of a Union

A union, written as U, is the chance of, for instance, Bob wearing a black suit OR a black pair of shoes.



Referring again to the table, the probability of Bob wearing a black suit OR a black shoes is$\frac{11}{15}$. A union is the probability of either event happening, including the situation in which both events happen.

	black shoes	**black shoes**	**brown shoes**
black suit	black suit, black shoes	black suit, black shoes	black suit, brown shoes
brown suit	brown suit, black shoes	brown suit, black shoes	brown suit, brown shoes
blue suit	blue suit, black shoes	blue suit, black shoes	blue suit, brown shoes
blue suit	blue suit, black shoes	blue suit, black shoes	blue suit, brown shoes
blue suit	blue suit, black shoes	blue suit, black shoes	blue suit, brown shoes

P (AUB) $= \mathrm{P\ (A)} + \mathrm{P\ (B)} - \mathrm{P\ (A{\cap}B)}$

To calculate this union, sum P (black suit) and P (black shoes), and subtract P (black suit ∩ black shoes). By subtracting the intersection, P (black suit∩black shoes), we avoid counting that area twice in the sum of P (black suit) and P (black shoes).

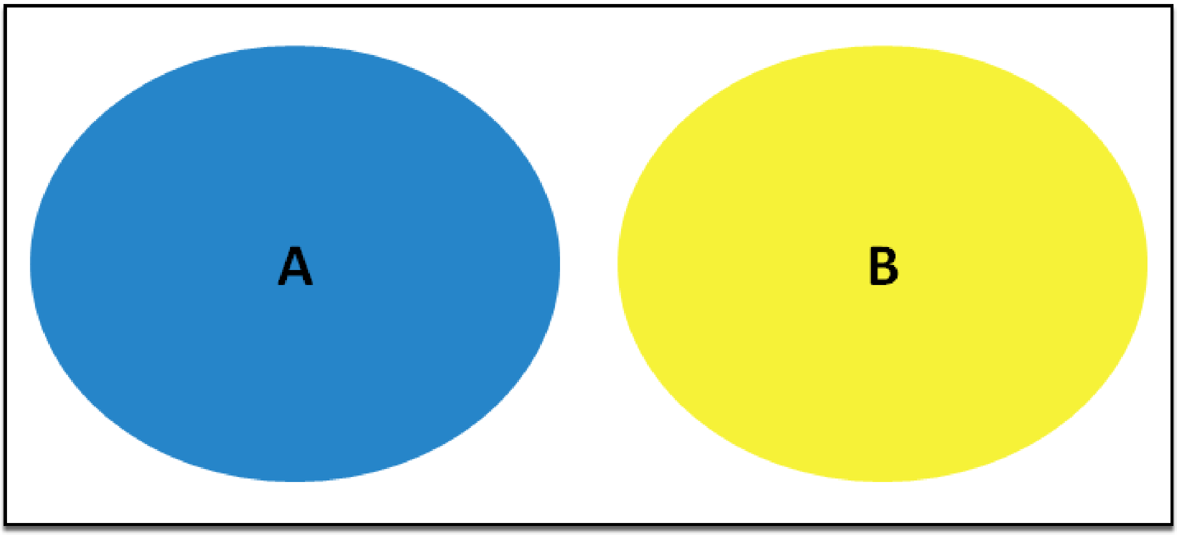
P (black suit U black shoes)$= \frac{1}{5} + \frac{2}{3} - \frac{2}{15} = \frac{11}{15}$

Mutually Exclusive Events

If two events cannot both occur, they are called mutually exclusive events, or disjoint events. This concept applies to any number of multiple events. As long as no two of the events can both occur, the events are mutually exclusive. For example, the three statements below are mutually exclusive:

- I was born in January.
- I was born in March.
- I was born in May.

No two of those statements can both be true. So, these statements are mutually exclusive.



Examine the Venn diagram above. Notice that there is no possibility for both A and B to occur. The only three possibilities are A (the blue area of the diagram), B (the green area of the diagram), or neither (the white area of the diagram). This is because A and B are mutually exclusive events.

If the events A and B are mutually exclusive, then$\mathrm{P}\ (\mathrm{A}\cap\mathrm{B}) = 0$. As you can see, mutually exclusive events are not independent events.

Consider mutually exclusive events, A and B. The probability of the intersection of A and B is 0. P(A∩B)$= 0$. Therefore, the probability of the union of A and B can be calculated as P (AUB) $= \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$.

Key to Probability

In each of the examples above, it is key to understand that when calculating a probability, selections and outcomes are completely random and unbiased. In other words, our findings concerning Bob picking a black suit and black shoes are based solely on numbers — his preference of suits and shoes is not a factor, nor is what he wore yesterday or what the current fashion trends are.

Likewise, if a question of probability involves Lauren baking cookies with nuts versus those without, the question of her coworkers having a nut allergy is not a factor. Though we are concerned about our fictional characters consuming hazardous baked goods, it is not a part of our calculating the probability of doing so.

Random and unbiased selections will not affect all questions of probability, but when calculating, focus solely on the numbers and information provided.

Probability: Assessing Business Risk

In the business environment, many decisions rely on dependent variables whose actual values cannot be predicted with certainty. For example, projections of the cost of raw materials, future interest rates, future employee health insurance claims, and expected product demand are all variables whose values are unknown and will be determined in the future. The uncertainty associated with the value of these variables introduces an element of risk to decision-making problems. The risk is the chance that the decision will not produce the intended results, and it often represents the potential for loss.

In many cases, such as when there is historical data, it is possible to assign probabilities to estimated values for unknown decision variables, representing the likelihood of that value (or range of values) occurring. These probabilities and the likelihood of positive or negative outcomes can be assembled in a matrix or decision tree to analyze expected payoffs and risks.

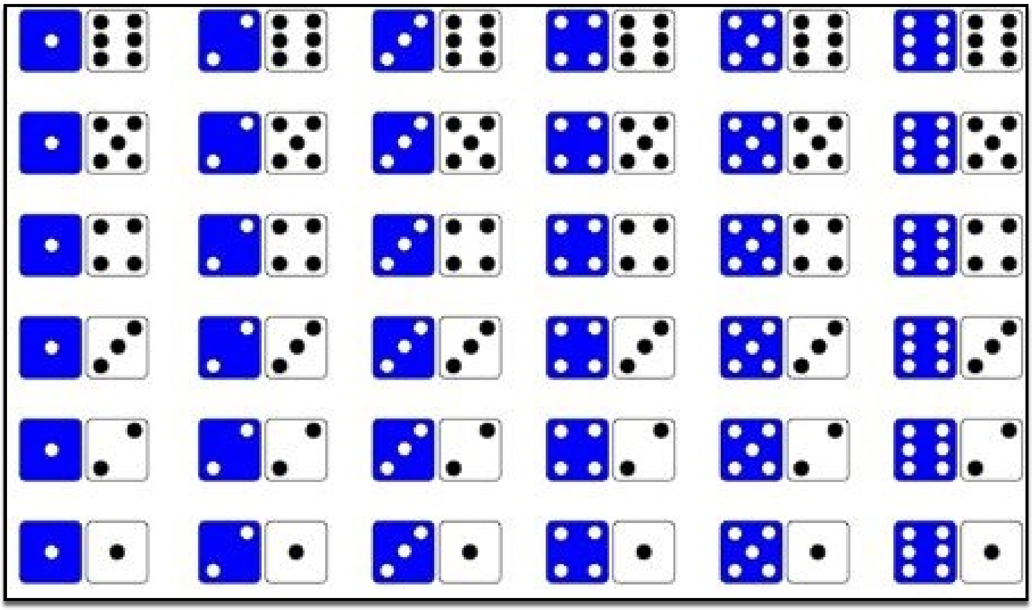
Decision tree will be explained in more detail in 3.17 Decision Analysis.

2.08 Probability: Other Topics

Probability: The Multiplication Principle

In probability, the multiplication principle states that to find the total number of outcomes that several events can have, you multiply each individual event's number of possible outcomes by the number of possible outcomes for all of the others.

For example, rolling a die with six sides has six possible outcomes. When rolling two dice with six sides each there are 36 possible outcomes ($6 \times 6 = 36$). Likewise, when rolling three six-sided dice, there are 216 possible outcomes ($6 \times 6 \times 6 = 216$) and so forth for larger numbers of dice.



Sampling

If you want to test a theory about a large population it might be difficult or impossible to test the theory on every individual in the population. This is when sampling is used. Sampling is the process of testing a number of individuals within a population to make a conclusion about the population as a whole. When you choose a sample there are a number of things to keep in mind: a sample should represent the entire population. It would be a mistake to only test certain individuals from the top ten percent in wealth of the population if one were trying to make a conclusion about that population. It is important to get a representative from each demographic of the population. Making predictions and testing theories about a population from testing a sample is called inferential statistics. An example of this would be a surveyor trying to determine the outcome of an election by surveying some of the people who leave voting areas. This differs from descriptive statistics which test a population and then make conclusions about only that population. An example of descriptive statistics would be a teacher taking a class poll to determine the availability for review session times by surveying the entire class.

Probability: Sampling with and without replacement

Suppose an ice cream store, the IC Delight, offers six different flavors. If IC Delight serves 10 customers, how many possible outcomes are there if each customer can have one scoop (an outcome, in this case, is a unique list of customers with their flavor choices, such as Customer 1 buys chocolate, Customer 2 buys vanilla, Customer 3 buys chocolate, etc.)?

This is an example of sampling with replacement, since each time a customer orders a scoop of ice cream, that ice cream flavor remains available for the next customer. If sampling with replacement and taking a sample of size n from a population of z, there are z^n possible outcomes.

At IC Delight, each customer has six different choices. Here, we examine 10 customers. By using the multiplication principle, we find that:

$$z^n = 6^{10} = 60{,}466{,}176$$

Now suppose IC Delight has only one scoop left of each of its six flavors, and six customers come in to buy one scoop each. How many possible outcomes are there now?

This is an example of sampling without replacement, since each time a customer orders a scoop of ice cream, that ice cream flavor is no longer available for the next customer. If you take a sample size of n without replacement, there are usually n! (n factorial) possible outcomes. The first customer has 6 choices, the 2nd has 5, the 3rd has 4, and so on.

$$n! = 6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$$

Probability: Permutations and Combinations

Permutations

Suppose IC Delight only has one scoop left of each of 8 flavors, and 5 customers come in to buy one scoop each. How many possible outcomes are there now? The first customer has 8 choices, the 2nd has 7, the 3rd has 6, and so on.

$$8 \times 7 \times 6 \times 5 \times 4 = 6{,}720$$

Is there an easier way to write this? Yes, $\frac{8!}{3!}$.

What we're doing is choosing a sample of size 5 without replacement from a population of 8. If you take a sample of size n from a population of size m, then there are $\frac{m!}{(m-n)!}$ possible samples (which is often written as mPn). Each different sample choice is called a permutation.

Combinations

Suppose IC Delight only has one scoop left of each of 4 flavors, and you want to buy 2 scoops. The choices are chocolate, vanilla, coffee, and strawberry. How many combinations of flavors can you make by choosing two at a time?

A combination is a sample chosen from a population where the order of the objects chosen does not matter. This means that in the example at hand, choosing chocolate and vanilla, or vanilla and chocolate, would count as 2 permutations, but only 1 combination. To eliminate these extra permutations and find the combinations of a sample, you need to divide the total number of permutations by the factorial of the sample size.

If mPn $= \frac{m!}{(m-n)!}$, then mCn $= \frac{m!}{((m-n)!n!)}$

To solve for the number of *permutations*, you can make from these flavors we use the permutation formula, expressed as 4P2,

$\frac{4!}{(4-2)!} = 12$ permutations

To solve for the number of *combinations*, you can make from these flavors we use the combination formula, expressed as 4C2.

$= \frac{4!}{[2!\times(4-2)!]}$ combinations

Bayes' Theorem

Your friend, Brian, starts talking to you about his sisters' baby. You are wondering if it is a boy or a girl (we'll assume a 50 percent chance for either). Brian says the baby was wearing red. Also, it is known that baby boys wear red 40 percent of the time while baby girls only wear red 10 percent of the time. From this, you want to calculate the likelihood that Brian has a nephew.

Let's pause and discuss Bayes' Theorem. Bayes' Theorem states that if you know the probability of event A, the probability of event B, and the conditional probability of event A given event B, we can determine the probability of event B given event A. Basically, it's a way to include new information into a calculation to give a better understanding of the situation.

The following is Bayes' Theorem:

$$P\left(A\middle|B\right) = \frac{P(B|A)\times P(A)}{P(B)}$$

$P(A|B) =$ Probability that event A will happen, knowing that event B has happened

With this understanding and equation, let's return to our initial problem with Brian.

$$P(\text{Red}) = P(\text{Boy}) \times P(\text{Red} \mid \text{Boy}) + P(\text{Girl}) \times P(\text{Red} \mid \text{Girl}) = 0.50 \times 0.40 + 0.50 \times 0.10 = 0.25$$

The following is the likelihood that any baby is wearing red:

$$P(\text{Red}) = P(\text{Boy}) \times P(\text{Red} \mid \text{Boy}) + P(\text{Girl}) \times P(\text{Red} \mid \text{Girl}) = 0.50 \times 0.40 + 0.50 \times 0.10 = 0.25$$

From this, using Baye's theorem, we'll determine:

$$P\left(\text{Boy} \middle| \text{Red}\right) = P\left(\text{Red} \middle| \text{Boy}\right) \times \frac{P(\text{Boy})}{P(\text{Red})}$$

$$P\left(\text{Boy} \middle| \text{Red}\right) = \frac{0.40\times 0.50}{0.25} = 0.80$$

Therefore, there is an 80 percent chance Brian's sister had a son.

Bayes' Theorem is very useful in understanding and accounting for new developments as a situation evolves. It determines new probabilities based on the conditions (e.g., the baby is wearing red) that are introduced.

Enhanced Content

Probability example

An example would be in poker. In five-card-draw you get five cards and can exchange any number of cards in hopes to improve your hand. There are 52 cards in a deck, 4 different suits and 13 cards from each suit.

Situation 1: If you have 3 hearts and 2 diamonds in your original hand and decide to pursue a flush (all five cards of the same suit) then you would exchange the diamonds in hopes of a heart AND a heart. There are 10 possible hearts you could get back (you have three) out of 47 cards (you've seen five). This means the probability of the first card being a heart ($\frac{10}{47}$) multiplied by the probability of the second card being a heart, given the first card is a heart $\frac{9}{46}$).

$$P(\text{Flush}) = P(\text{H4}) \times P(\text{H5}) = \frac{10}{47} \times \frac{9}{46} = 0.042$$

Therefore, there is a 4.2% chance that you will get a hearts flush.

Situation 2: You have three aces, and are hoping for four of a kind. You exchange your two non-aces. There is only one ace left in the deck out of 47 cards you haven't seen yet. The probability that the first card is an ace is $\frac{1}{47}$ and the probability that the second is an ace, given the first card is not an ace, is ($\frac{46}{47} \times \frac{1}{46}$).

P(4 of a Kind)$= \mathrm{P(A4)} + \mathrm{P(A5)} = \frac{1}{47} + (\frac{46}{47} \times \frac{1}{46}) = 0.043$

Therefore, there is a 4.3% chance that you will get four of a kind.

Combination example

Example 1: When managing work teams you might look at the number of different combinations with the employees available. If you are assembling a team of 4, choosing from your pool of 24 employees, the order that the 4 team members are chosen is not important. How many unique teams are possible in this scenario? This number can be calculated using the equation above...

$X = 24, Y = 4$

$\frac{24!}{((24-4)! \times 4!)} = \frac{24!}{(20! \times 4!)} = 10{,}626$

This next example calculates probability from all of these combinations we've been learning to calculate. This is rather in-depth and only for a deeper understanding of the statistics themselves and the abilities of combinations.

Example 2: There are one hundred senators. 55 are Democrats, and 45 are Republicans. The Committee for Third-World Support has eight seats. Find the probability that this committee will include 6 Democrats and 2 Republicans.

This is a lot of information. Let's break it down. Let's first start by looking at the big picture. They don't ask for anything to be ordered, so we're working entirely with combinations. We're choosing 6 Democrats and 2 Republicans out of all combinations of 8 Senators. Let's start with the total number of combinations of 8 Senators.

Senators chosen $= A = 8$

Senators possible $= B = 100$

Combinations $= \frac{B!}{((B-A)! \times A!)} = \frac{100!}{(92! \times 8!)} = 186{,}087{,}894{,}000$

So there are about 186 billion combinations of 8 senators. Now let's look at ways to get 6 Democrats and 2 Republicans. Let's start with Democrats.

We need 6 of 55 Democrats. Let's start with finding out how many combinations of 6 Democrats there are from 55 available.

Senators chosen $= 6 = A$

Senators possible $= 55 = B$

Combinations $= \frac{B!}{((B-A)! \times A!)} = \frac{55!}{(49! \times 6!)} = 28{,}989{,}675$

We need 2 or 45 Republican Senators.

Combinations $= \frac{45!}{(43! \times 2!)} = 990$

For the total number of 2 Republican, 6 Democrat combinations we'll multiply these totals together.

$990 \times 28{,}989{,}675 = 28{,}699{,}778{,}250$

This is the total number of combinations for 2 Republicans and 6 Democrats. The total number of combinations for any 8 senators is 186,087,894,000.

This means that there is a ($\frac{28{,}699{,}778{,}250}{186{,}087{,}894{,}000}$) a 15.4 percent chance that an 8-person Senate Committee will have 2 Republicans and 6 Democrats.

In problems like this it's important to break it up into smaller parts. Take some time to figure out what you're looking at and then as you're able to group the terms the picture will gain clarity.

Bayes' Theorem example

To go more in depth on the application of Bayes' Theorem, click on this next example:

Nicole is a professional sprinter and went in for a steroid test. The test correctly identifies users 90 percent of the time and correctly identifies non-users 85 percent of the time. This drug is given to 1000 sprinters, 5 percent of whom are taking steroids. Nicole tests positive. What is the probability that she takes steroids?

To figure this out we need to start with figuring out the likelihood of steroid users getting identified as well as non-steroid users wrongly accused. Basically, we need to determine the probability of a positive test. Then, we will divide the probability of a true positive test (the likelihood someone takes steroids and tested positive) by the probability of a positive test.

P(user) $= 0.05$

P(non-user) $= 0.95$

P(+test|user)$= 0.90$

P(+test|non-user) $= 1 - 0.85 = 0.15$

P(+test) $= P(\text{user}) \times P(+\text{test} \mid \text{user}) + P(\text{non-user}) \times P(+\text{test} \mid \text{non-user})$
$= 0.05 \times 0.90 + 0.95 \times 0.15 = 0.1875$

$$P\left(\text{user} \middle| + \text{test}\right) = \frac{\text{P(user)} \times P(+\text{test} \mid \text{user})}{P(+\text{test})} = \frac{0.05 \times 0.90}{0.1875} = 0.24$$

Therefore, we've determined that there is a 24 percent chance that Nicole is taking steroids. This is not because the test is not reliable but rather because the sample size is so large and therefore that has to be understood and taken under consideration when looking at a positive test.

Remember the steps. We first look at the likelihood of the outcome and then determine the chances that the outcome was reached in a certain way. In this situation, we looked at the probability that people were using and not using and the probability that they could get a positive result. Then we looked at the likelihood that a positive test was really a steroid-user which we were able to relate to the total probability of a positive test.

2.09 Video: Descriptive Statistics

This assignment does not contain any printable content.

2.10 Measures of Central Tendency

In the following table, f stands for frequency, the number of instances something occurs in a given sample, and displays the number of days x books were sold.

Book Sales	f
42	2
43	0
44	1
45	4
46	1
47	1
48	2
49	1
50	2
51	1
52	2
53	1
54	1
55	2

56	3
57	0
58	1
59	1

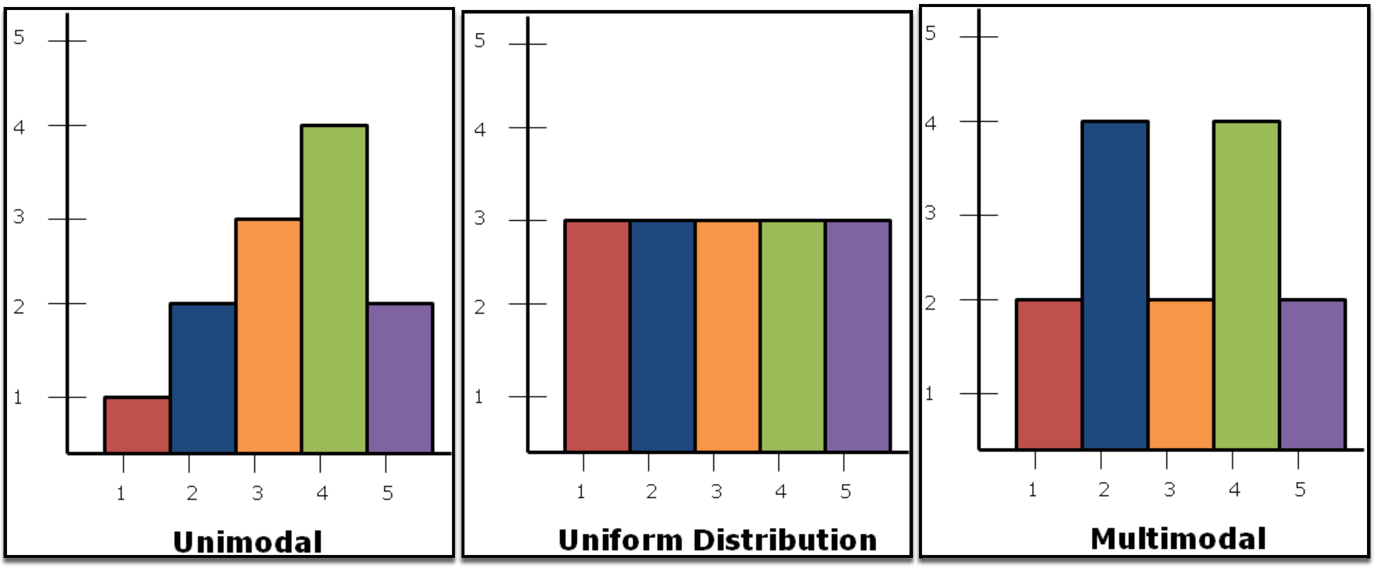
Interval	f
40-44	3
45-49	9
50-54	7
55-59	7

Limitations of Mode

One advantage of the mode is that it can be used with data from any of the four scales of measurement. However, the mode has a couple of important limitations:

1. There can be more than one mode. In fact, if all of the intervals contained the same number of scores, every value would be the mode, rendering that measure useless. For this reason, you should always graph your data. The other two measures of central tendency, median and mean, may not tell you this information about a data set's distribution as well as a graph.

   Below are three graphs, each with a mean of approximately 3.33. Notice the difference in shape.



Unimodal

Uniform Distribution

Multimodal

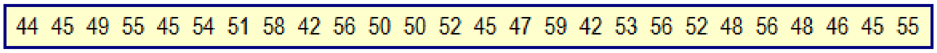
2. The other problem with the mode is that it is not an arithmetic computation. It cannot be used in any of the inferential statistical tools that you will learn later in the course. For that reason, use of the mode as a measure of central tendency is limited.

The Median

A somewhat more sophisticated measure of central tendency is the median. The median of a distribution is the point at which an equal number of scores fall above and below. The median is the "half-way" point of the data. An equal number of values in a distribution are greater than the median and less than the median.

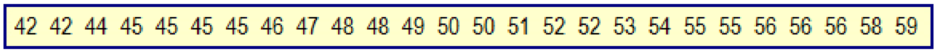
Refer back to the bookstore data from above. To figure out the median of a dataset, follow these steps:

1. List your data.



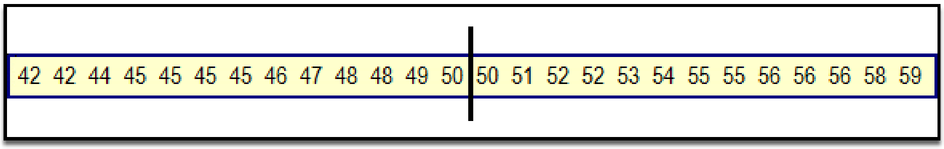
44 45 49 55 45 54 51 58 42 56 50 50 52 45 47 59 42 53 56 52 48 56 48 46 45 55

2. Sort your data from smallest to largest.



42 42 44 45 45 45 45 46 47 48 48 49 50 50 51 52 52 53 54 55 55 56 56 56 58 59

3. Imagine drawing a line down the middle of your data so that half the data points are on the right of the line and half are on the left.



42	42	44	45	45	45	45	46	47	48	48	49	50	50	51	52	52	53	54	55	55	56	56	56	58	59

4. If your line lands on a data point (for example, if you have an odd number of values), that is the median. If your line lands in between 2 data points (for example, if you have an even number of values), your median is the halfway point between the two data points, and the median is calculated by taking the average of those numbers. In this case:

$$\frac{(50+50)}{2} = \mathbf{50}$$

The Mean

The mean is the most sophisticated of the three measures of central tendency. It is generally known as the arithmetic mean or average. Unlike the median and mode, the mean is influenced by the size of the values in a dataset. Because of this, extreme values (values that are very large or very small relative to the other data points) have a greater influence on the mean than they do on the median or mode. Furthermore, you can only calculate a mean for interval and ratio data.

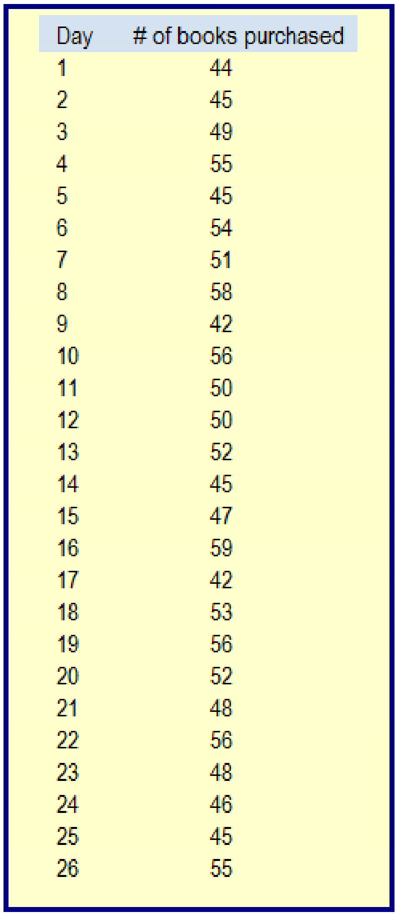
The symbol for the *population mean* is μ. μ is the Greek letter "mu" and pronounced "mew." The symbol for the *sample mean* is $\bar{x}$. $\bar{x}$ is pronounced "x-bar." The formula for the sample mean is:

$$\bar{x} = \frac{\sum x_i}{n}$$

where:
$\bar{x}$ is the sample mean
$\sum x_i$ is the sum of all individual values in a dataset

n is the number of data points in the dataset

Consider the data from the bookstore example. The data to the right represents the number of books purchased each day of our sample. We could find the mean of the data set by summing each of the data points for how many books were sold on a given day, and dividing that sum by the number of days we measured, which is 26. We can see the calculation of the mean as follows:



Day	# of books purchased
1	44
2	45
3	49
4	55
5	45
6	54
7	51
8	58
9	42
10	56
11	50
12	50
13	52
14	45
15	47
16	59
17	42
18	53
19	56
20	52
21	48
22	56
23	48
24	46
25	45
26	55

$$\sum x_i = 44 + 45 + 49 + 55 + \ldots + 55 = 1303$$

Since there are 26 data points, n$= 26$

$$\bar{x} = \frac{1303}{26}$$

$$\bar{x} = \mathbf{50.12}$$

This means that the *average* amount of books sold in a given day is a little more than 50.

Properties of the Mean

A deviation score is calculated by subtracting the mean from an individual score. So, the deviation score for the first data point, 44, is 44 - 50.12 = -6.12. The sum of the deviation scores from the mean is always equal to zero. See the table below and notice that performing this calculation for all data points, then adding up the values equals zero.

Day	# of books purchased	Mean	Deviation Score
1	$44 \quad -$	$50.12 \quad =$	-6.12
2	$45 \quad -$	$50.12 \quad =$	-5.12
3	$49 \quad -$	$50.12 \quad =$	-1.12

4	$55 - 50.12 = 4.88$
5	$45 - 50.12 = -5.12$
6	$54 - 50.12 = 3.88$
7	$51 - 50.12 = 0.88$
8	$58 - 50.12 = 7.88$
9	$42 - 50.12 = -8.12$
10	$56 - 50.12 = 5.88$
11	$50 - 50.12 = -0.12$
12	$50 - 50.12 = -0.12$
13	$52 - 50.12 = 1.88$
14	$45 - 50.12 = -5.12$
15	$47 - 50.12 = -3.12$
16	$59 - 50.12 = 8.88$
17	$42 - 50.12 = -8.12$
18	$53 - 50.12 = 2.88$
19	$56 - 50.12 = 5.88$
20	$52 - 50.12 = 1.88$
21	$48 - 50.12 = -2.12$
22	$56 - 50.12 = 5.88$
23	$48 - 50.12 = -2.12$
24	$46 - 50.12 = -4.12$
25	$45 - 50.12 = -5.12$
26	$55 - 50.12 = 4.88$

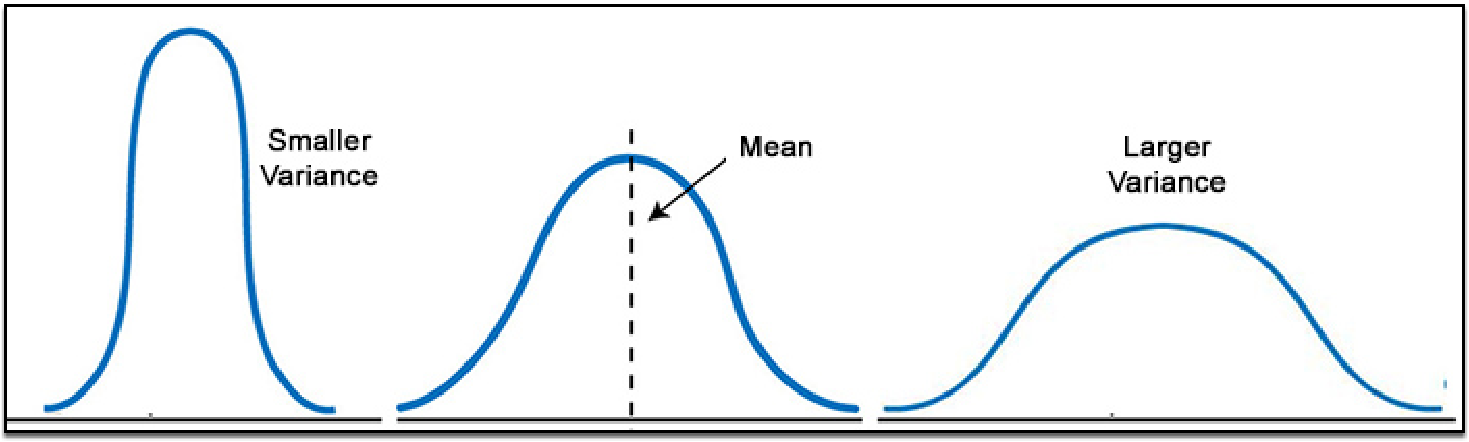
2.11 Measures of Variability: Variance and Standard Deviation

Variance and Standard Deviation

In addition to determining the middle of a distribution by using a measure of central tendency, it is also helpful to know how spread out the data are, or how variable the data are. Measures of variability, the range, and the standard deviation can tell you this. We will focus on the standard deviation and its parent measure, the variance.

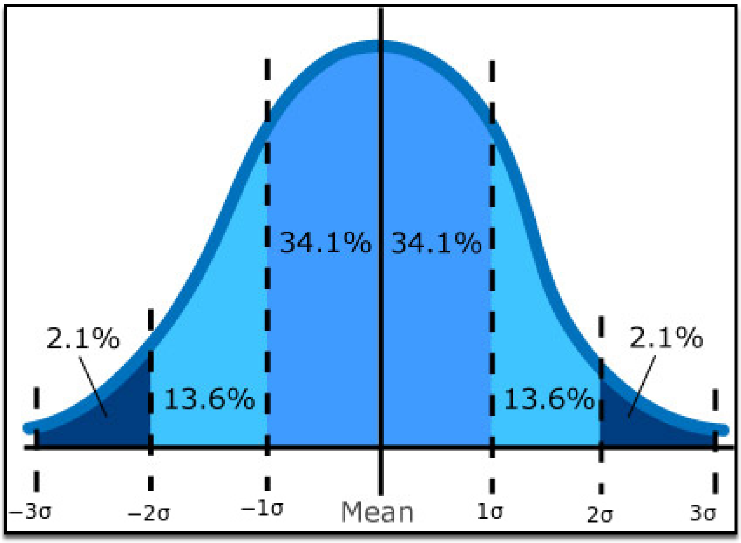
The variance is a measure of *how spread out data are about the mean*. The closer the data are to the mean, the smaller the variance. If most of the data points are spread out relative to the mean, the variance is larger.

The vertical lines through the bell curves below represent the mean. Notice the difference between the graph of the dataset with the larger variance compared to the data with the smaller variance.



Calculating Variance and Standard Deviation

Roughly speaking, the standard deviation tells you how far, on average, the data points are from the mean. As displayed in the graph below, in a normal distribution the 34.1 percent of the data results will fall between the mean and one standard deviation above the mean. 13.6 percent of the data results will fall between one standard deviation above the mean and two standard deviations above the mean. 2.1 percent of the data results will fall between two standard deviations above the mean and three standard deviations above the mean. This is the same for the distribution below the mean as well.



Given two data sets both with means of 100 where one has a standard deviation of 15, and the other has a standard deviation of 30, the data points in the latter will be, on average, roughly twice as far from the mean as the data points in the first data set.

There are two types of variance that each use their own formula: population variance and sample variance. The standard deviation is always equal to the square root of the variance.

Calculating Population Variance and Standard Deviation

We can always calculate the population variance and population standard deviation using spreadsheet programs, such as Microsoft Excel or OpenOffice. Here, we examine the formulas behind population variance and standard deviation.

The variance of a population is denoted by σ^2, and the population standard deviation is σ (σ is the Greek letter sigma). The population variance formula is as follows:

$$\boxed{\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}}$$

where:
$\sigma^2 =$ the population variance
$\mu =$ the population mean
$\sum(x-\mu)^2 =$ the sum of the squared differences between each data point and the mean. That is, you square each data point's difference from the mean, then add these squared numbers together
$=$ the size of the population

Remember, population standard deviation (σ) is the square root of population variance. To calculate a population standard deviation manually, one would take the square root of the population variance formula:

$$\boxed{\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}}$$

Calculating Sample Variance and Standard Deviation

We can always calculate the sample variance and sample standard deviation using spreadsheet programs, such as Microsoft Excel or OpenOffice. Here, we examine the formulas behind sample variance and standard deviation.

The formulas for the variance and standard deviation of a sample (for example, a subset of a population) are slightly different.

The variance of a sample is denoted by s^2, and the sample standard deviation is s. When calculating the variance and standard deviation for a sample, the N term in the formula above is simply replaced by $(n-1)$:

$$\boxed{s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}}$$

where:
$^2 =$ the sample variance
$\bar{x} =$ the sample mean
$(x-\bar{x})^2 =$ the sum of the squared differences between each data point and the mean. That is, you square each data point's difference from the mean, then add these squared numbers together
$n =$ the sample size

The replacement of N with$n-1$ is known as Bassel's correction, and yields a more accurate number for a sample — especially a small sample. When the size of the sample (n) is very large, the difference between the population and sample variances and standard deviations becomes very small.

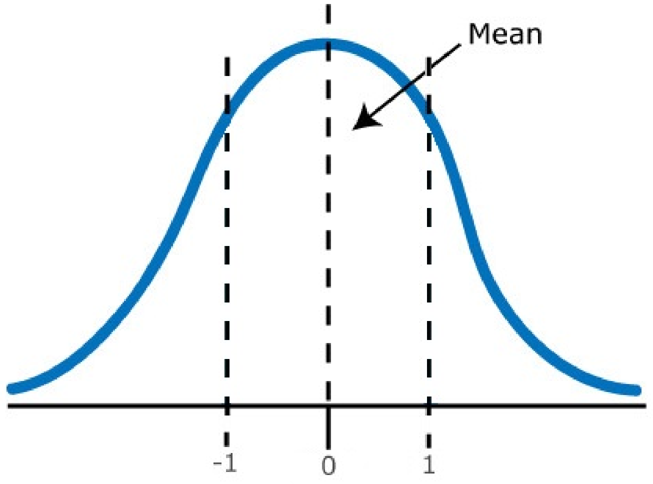
Remember, sample standard deviation (*s*) is the square root of sample variance. To calculate a sample standard deviation manually, one would take the square root of the sample variance formula:

$$\boxed{s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}}$$

The Normal Distribution

When data tends to occur around a central value with no bias right or left, it gets close to a normal distribution. All normal distributions look like a symmetric, bell-shaped curve.

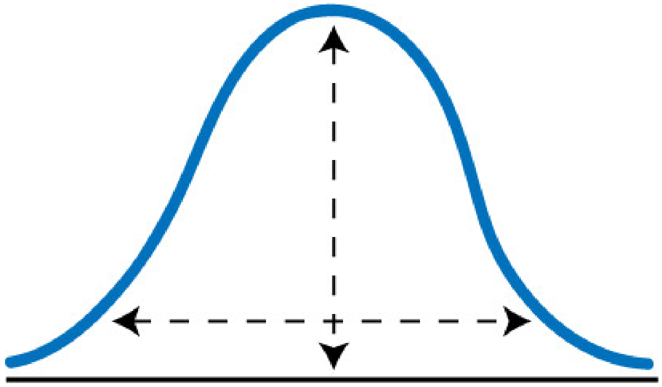
Slide 1



The graph of the normal distribution depends on two factors—the mean and the standard deviation.

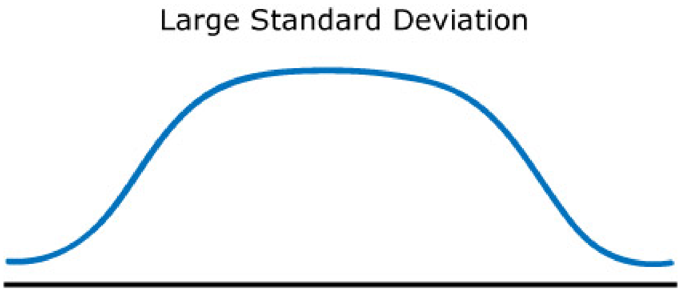
The mean of the distribution determines the location of the center of the graph.

Slide 2



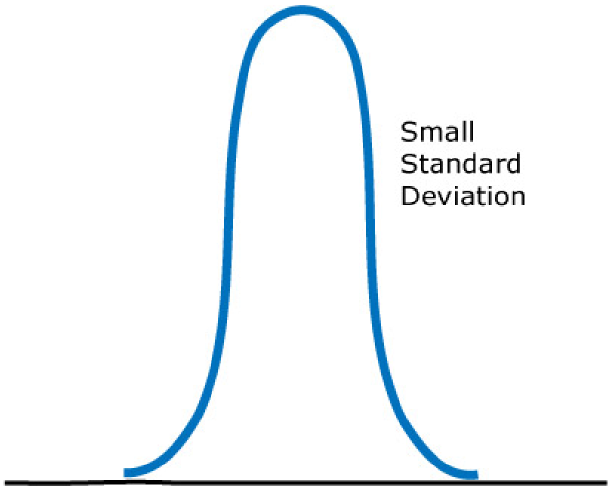
Standard deviation is a measure of variance, or how spread out the data is. The standard deviation determines the height and width of the graph.

Slide 3



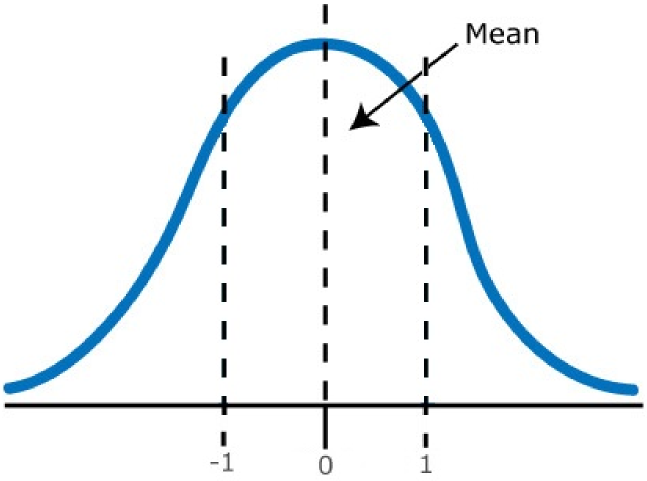
When the standard deviation is large, the curve is short and wide.

Slide 4



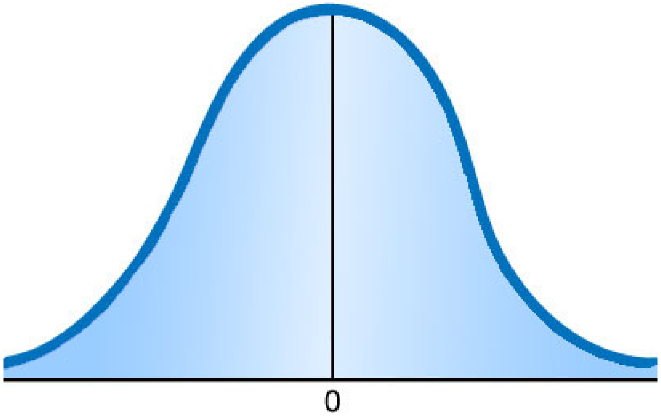
When the standard deviation is small, the curve is tall and narrow.

Slide 5



The standard normal distribution is the normal distribution that has a mean of zero and a standard deviation of one.

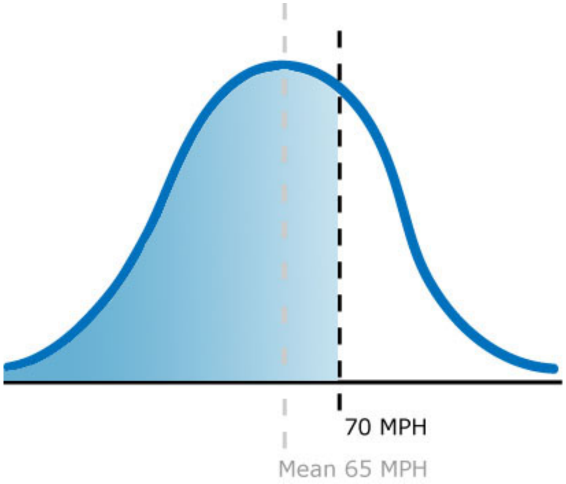
Slide 6



The area under the curve represents the probability of events happening. The total area under the normal curve, as with any distribution, is 1.

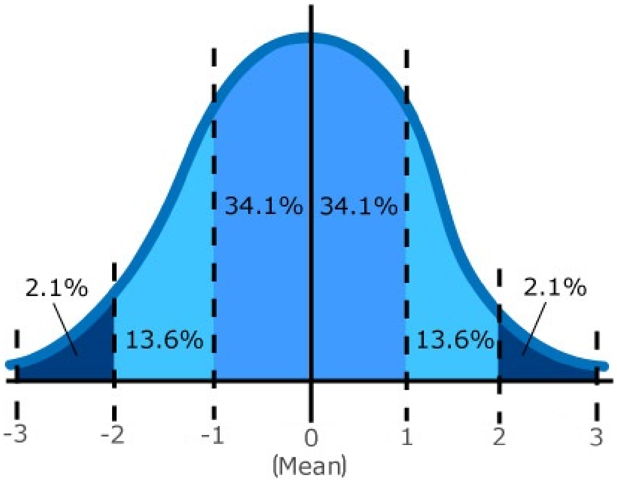
Cumulative probability is a sum of probabilities. In connection with the normal distribution, a cumulative probability refers to the probability that a randomly selected score will be less than or equal to a specified value.

Slide 7



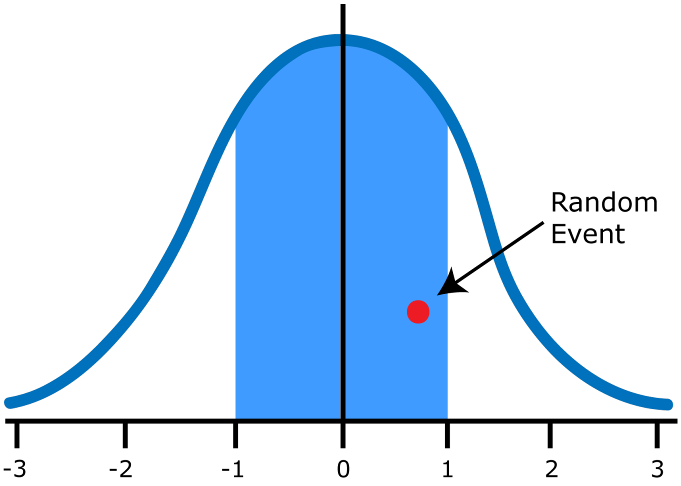
For example, the probability that a random car on a particular stretch of road is traveling at less than or equal to 70 mph, is the sum of probabilities (i.e. the probability that the car is traveling at exactly 70 mph plus the probability that it is traveling at 69 mph plus the probability that it is traveling at 68 mph, etc.). This is cumulative probability.

Slide 8



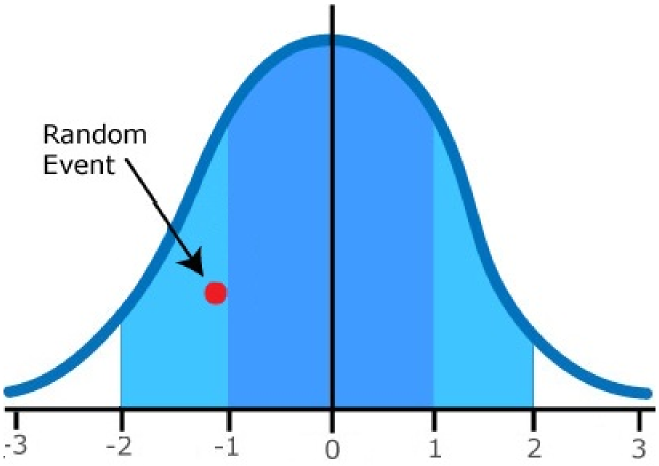
According to The Empirical Rule, approximately 68.3% of the data points in a dataset will be within 1 standard deviation of the mean. Approximately 95.4% of the data points in a dataset will be within 2 standard deviations of the mean. And almost all (99.7%) of the data points in a dataset will be within 3 standard deviations of the mean.

Slide 9



If we were to randomly choose one event from a normally distributed dataset, we would pick an event with a value that is within 1 standard deviation of the mean 68.3% of the time. This percentage is represented by shading the area under the curve within 1 standard deviation of the mean.

Slide 10



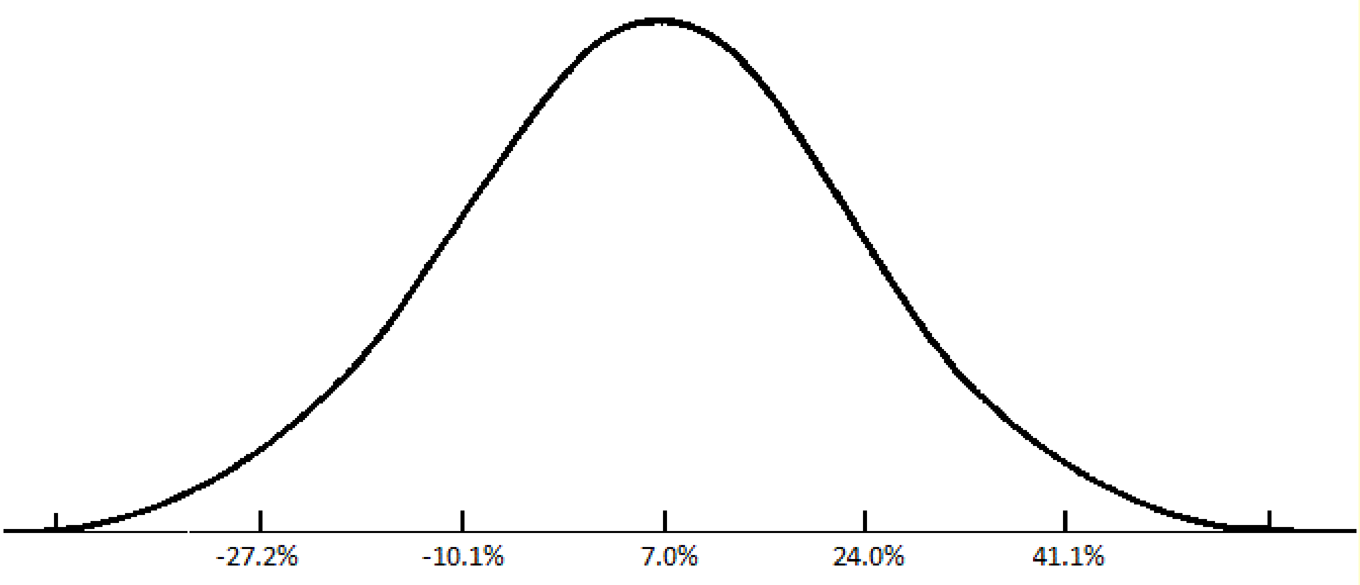
Similarly, examine the proportion of probabilities falling +/- 2 standard deviations from the mean. Randomly picking a data point from a normally distributed dataset, you would have a 95% chance of getting a value that falls within this part of the distribution. Overall, the probability of getting a value under the curve is 100%, and thus the area under the curve is 1.

Applications

The normal distribution exists in many different fields. Click on each field below to see an example.

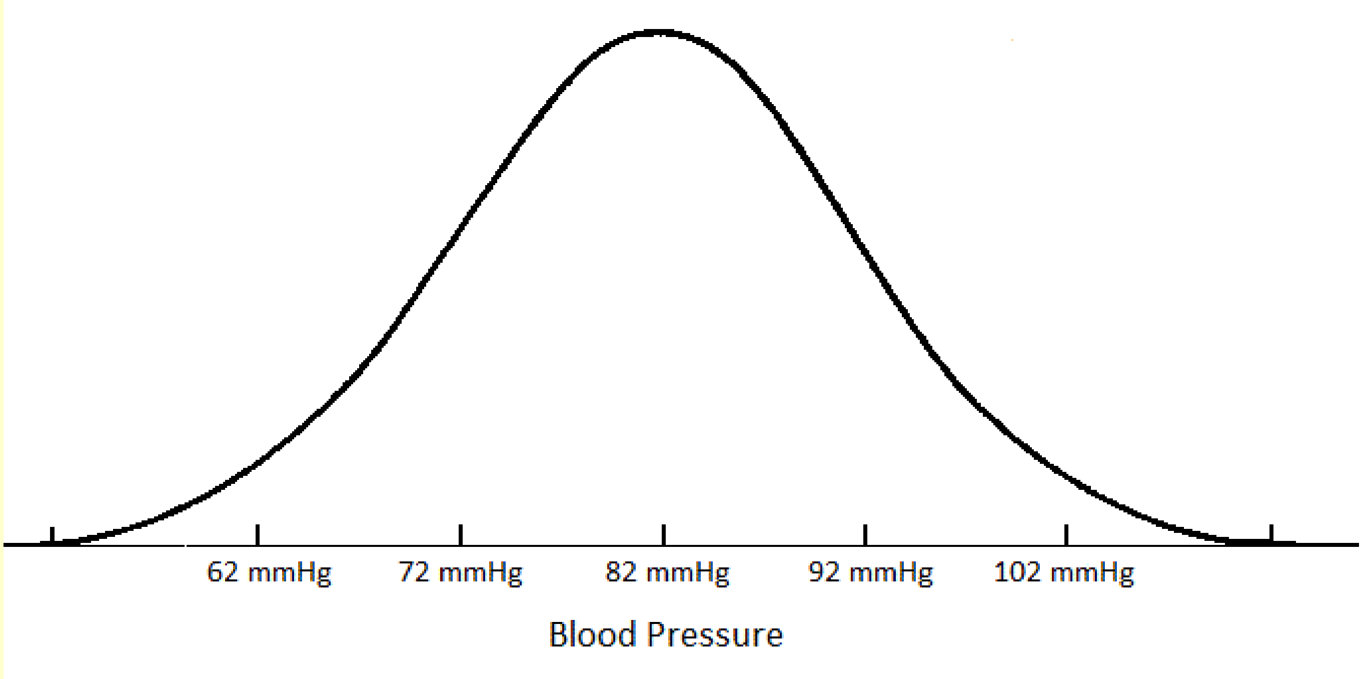
Business

The following is a table of inflation-adjusted annual returns on investment in the S&mp;P 500, with dividends reinvested, over the 50-year period from 1963 through 2012:



The normal distribution can help a savvy investor make a well-informed plan. An investor who has money invested in the S&P 500 could look at this graph and determine what plan he needs to make. Knowing that the mean, median and mode, which is approximately 7%, is the most likely outcome, an investor may decide that he should expect to make about 7% annually over the long term if he keeps his money invested and reinvests his dividends. Another investor may decide that he needs a contingency plan if the investment does anything between losing 27.2% of its value and gaining 41.1% of its value in a given year. Referring to the empirical rule, 95.4% of years will experience an outcome within these two data points, 2 standard deviations from the mean. Both of these investors would be using the normal distribution to their advantage.

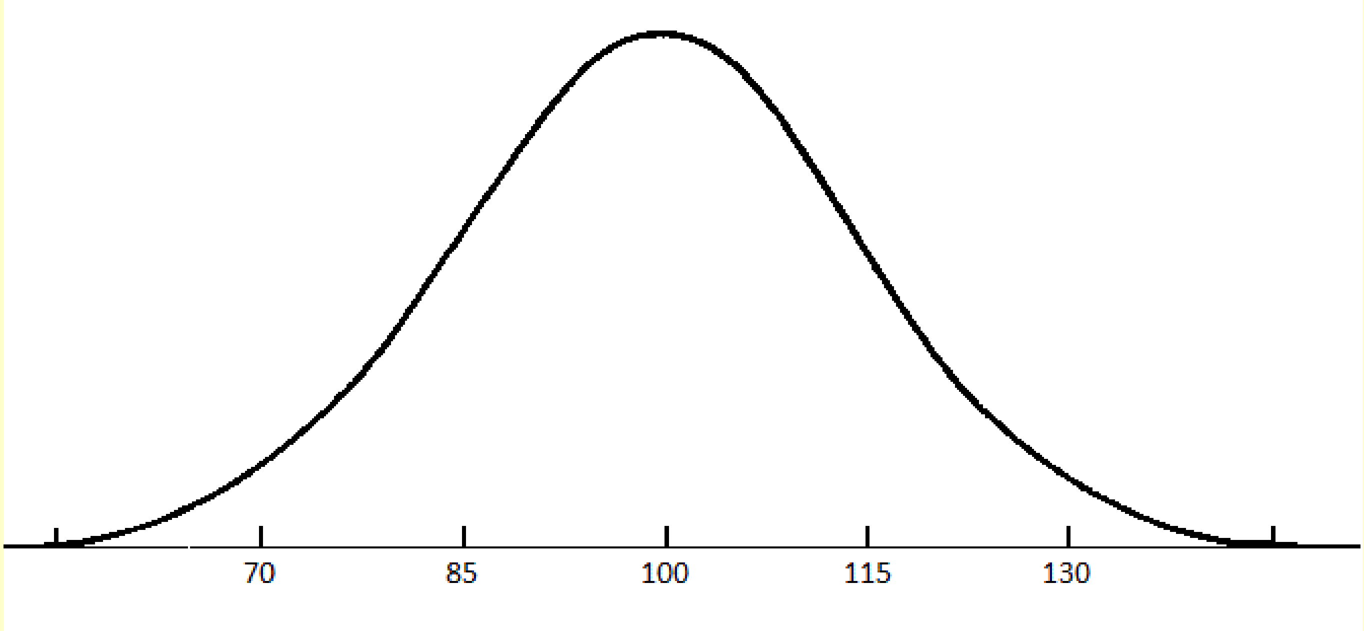
Healthcare



The graph above shows a distribution of adult diastolic blood pressure in a population. We see that it is normally distributed. This graph can help a doctor or patient determine if a blood pressure reading is abnormal. To measure the level of risk a patient in a population has of being diagnosed with high blood pressure, a doctor would examine the population's mean and standard deviation of blood pressure. Clinical guidelines define high blood pressure as diastolic blood pressure greater than 90mmHg. In the healthcare field, the normal distribution is also seen in BMI. There are many ways somebody working in healthcare could use the normal distribution.

Education

IQ tests are normally distributed based on a mean of 100 with a standard deviation of 15. Here is a graph of IQ test scores:



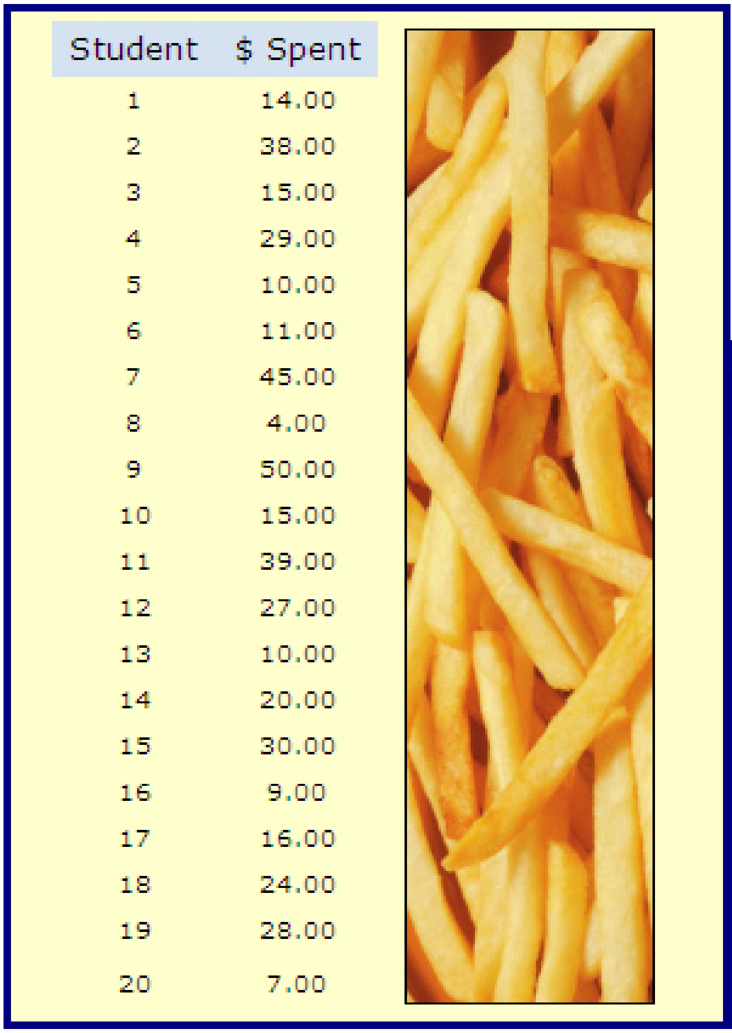
2.12 Video: Graphic Display of Statistics

This assignment does not contain any printable content.

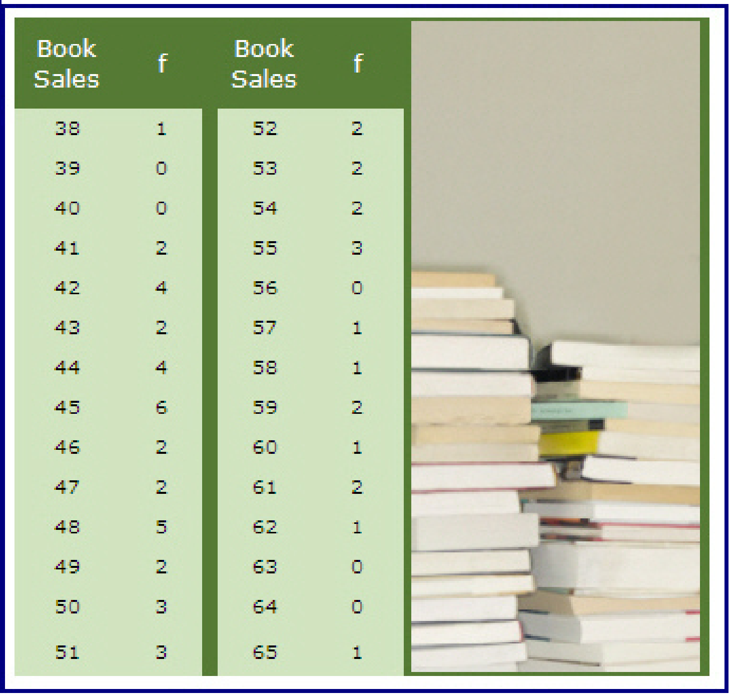
2.13 Standard Scores

Standard Scores (z-scores)

Often, if we are comparing two or more distributions to one another, the scales will be different. Consider the book sales data and the restaurant spending data:



Student	$ Spent
1	14.00
2	38.00
3	15.00
4	29.00
5	10.00
6	11.00
7	45.00
8	4.00
9	50.00
10	15.00
11	39.00
12	27.00
13	10.00
14	20.00
15	30.00
16	9.00
17	16.00
18	24.00
19	28.00
20	7.00



Book Sales	f	Book Sales	f
38	1	52	2
39	0	53	2
40	0	54	2
41	2	55	3
42	4	56	0
43	2	57	1
44	4	58	1
45	6	59	2
46	2	60	1
47	2	61	2
48	5	62	1
49	2	63	0
50	3	64	0
51	3	65	1

It is very difficult to see how these distributions are similar or different from one another because of the differences in scale. Z-scores will transform different datasets to the same scale. In order to do this, we figure out how far away each individual data point is from its respective mean. In other words, a z-score tells us the number of standard deviations a data point is from its mean. If a data point from the book sale data has the same z-score as a data point from the restaurant spending data, the two data points have the same relative location in their respective datasets.

The formula for z-score is:

$$z = \frac{x - \bar{x}}{s}$$

where:
$z =$ the number of standard deviation units a raw score is from its mean
$x =$ the raw score
$\bar{x} =$ the sample mean
$=$ the sample standard deviation (also denoted by $\boldsymbol{\sigma}$ for a population)

Be sure to replace s with σ when dealing with population data.

Because the mean is at the center of the distribution if the score falls below the mean, the z-score is negative. If the score falls above the mean, the z-score is positive. Once transformed to z-scores, all distributions have a mean equal to zero and a standard deviation of 1.

Example

Consider the book sales data. The mean of this data sample is 49.481. The sample standard deviation is 6.345. How many standard deviations away from the 49.481 is the data point of 44?

Here we are using sample data:

$$z = \frac{x - \bar{x}}{s}$$

$z = \frac{(44 - 49.481)}{6.345}$

$z = -0.864$

With a z-score of -0.864, the data point of 44 books is 0.864 standard deviations*below* the mean for this sample of book sales. We know this data point is below the mean because the z-score is a negative number.

Further, if you wanted to find a comparable data point among the restaurant spending data, you can use the same formula, solving for a different variable. The sample standard deviation of the restaurant spending data is 13.31, and the mean is 22.05:

$$z = \frac{x - \bar{x}}{s}$$

$-0.864 = \frac{(x - 22.05)}{13.31}$

$10.55 = x$

Since both data points, 44 from the book sale and 10.55 from the restaurant spending, have the same z-score, we know that they fall at the same relative point in their respective datasets even though one set of data is listed in quantity of books and the other in dollars.

Business Example

A manager at a specialty spice store is running low on real, organic wasabi, cultivated from the Japanese plant. The store only has 21 ounces of the product in stock. She decides to reorder some wasabi, which will arrive in two weeks. The manager has determined that demand of wasabi during this replenishment lead-time is normally distributed with a mean of 15 ounces and a standard deviation of 6 ounces. The manager would like to know the probability of a stockout during replenishment lead-time. In other words, what is the probability that demand over the next two weeks will exceed 21 ounces?

Here we are using population data:

$$z = \frac{x - \bar{x}}{\sigma}$$

$z = \frac{(21 - 15)}{6}$
$z = 1$

Because 21 ounces of demand during the lead-time has a z-score of 1.0, we know that the probability of exceeding that amount is related to a greater than one standard deviation above the mean. Recall, from our introduction to the normal distribution, that 50.0 percent of the data points will be below the mean and 34.1 percent of data points will be between the mean and one standard deviation above it. From subtraction of the total distribution ($1 - 0.5 - 0.341$), we see that 15.9 percent of the distribution falls greater than one standard deviation above the mean. Therefore, there is a 15.9 percent probability that demand will exceed 21 ounces.

2.14 Graphic Displays: Ranges and Box Plots

Range

The range represents the array of possibilities in which a value can exist, from the minimum value to the maximum value. The size of a range is the difference between the minimum and maximum values.

The following is a set of test grades:

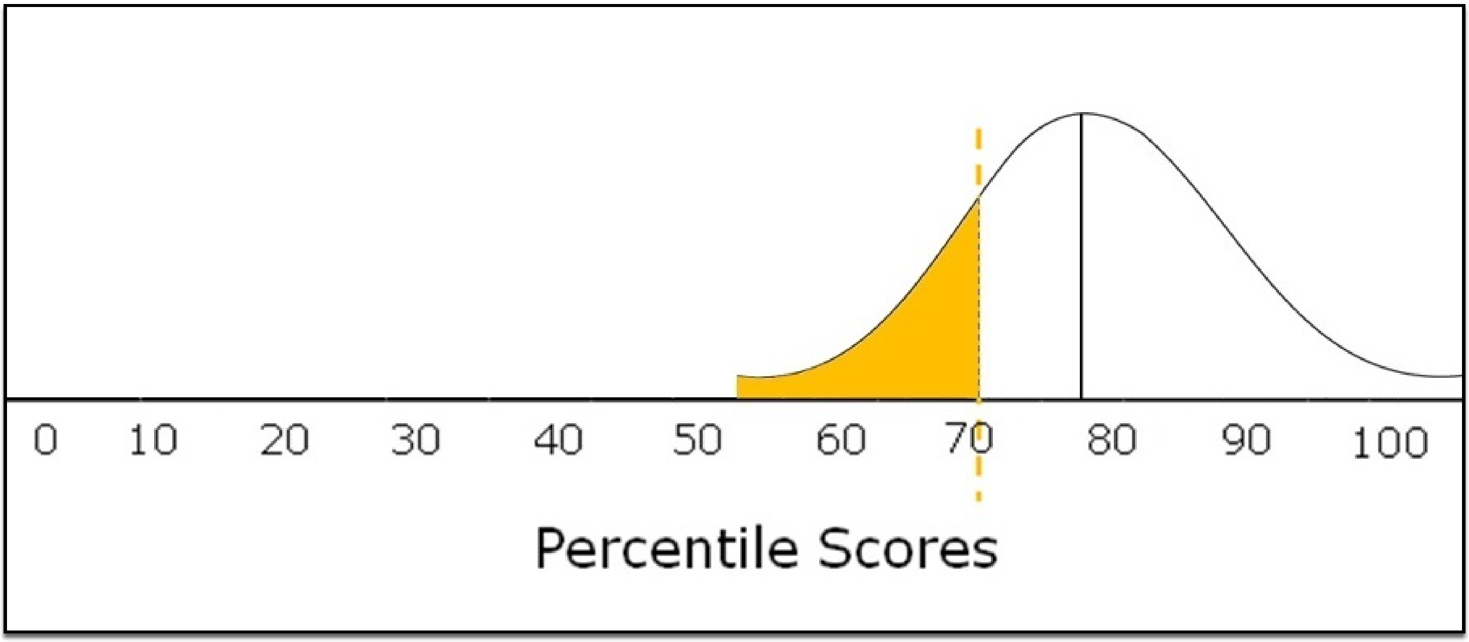
58, 95, 83, 72, 86, 85, 91, 80, 74, 78, 69

The range of these values is from 58 (the lowest value) to 95 (the highest value). The size of this range is 37.

Percentiles

A percentile is a unit of measurement that gives a value of which a percentage of the population falls below.

If there are 100 people taking a test and the twentieth lowest score is a 71, then the twentieth percentile is at 71.



It is very important to remember that percentile refers to a percentage of the population and not a percentage of the sum of the values. Percentiles are on the same scale as medians, not averages (means).

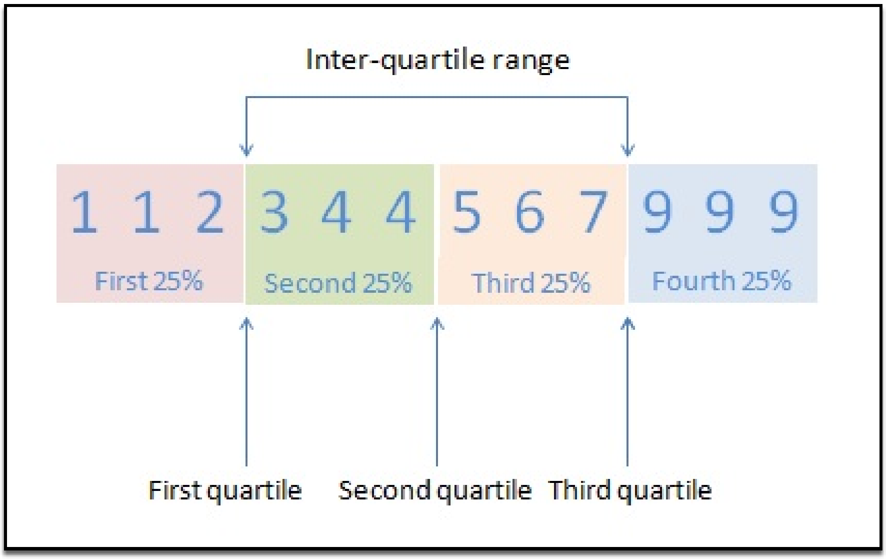
For example, polling 10 people about how many cigarettes they smoke per day might yield the following data:

10, 0, 0, 0, 30, 0, 0, 0, 0, 0

The average number of cigarettes per day of this 10-person sample is four cigarettes. This does NOT mean that four cigarettes is the 50th percentile. In fact, a majority of people do not smoke. In this case, the 50th percentile is zero cigarettes a day.

Inter-quartile Range

A quartile is each of four equal groups into which a population can be divided. The inter-quartile range measures the difference between the third quartile and the first quartile. To determine the first and third quartiles, order the data from lowest value to highest value. Then separate the data into four equal groups.



As you can see in the chart, specific numbers that have multiple occurrences are included the number of times they occur.

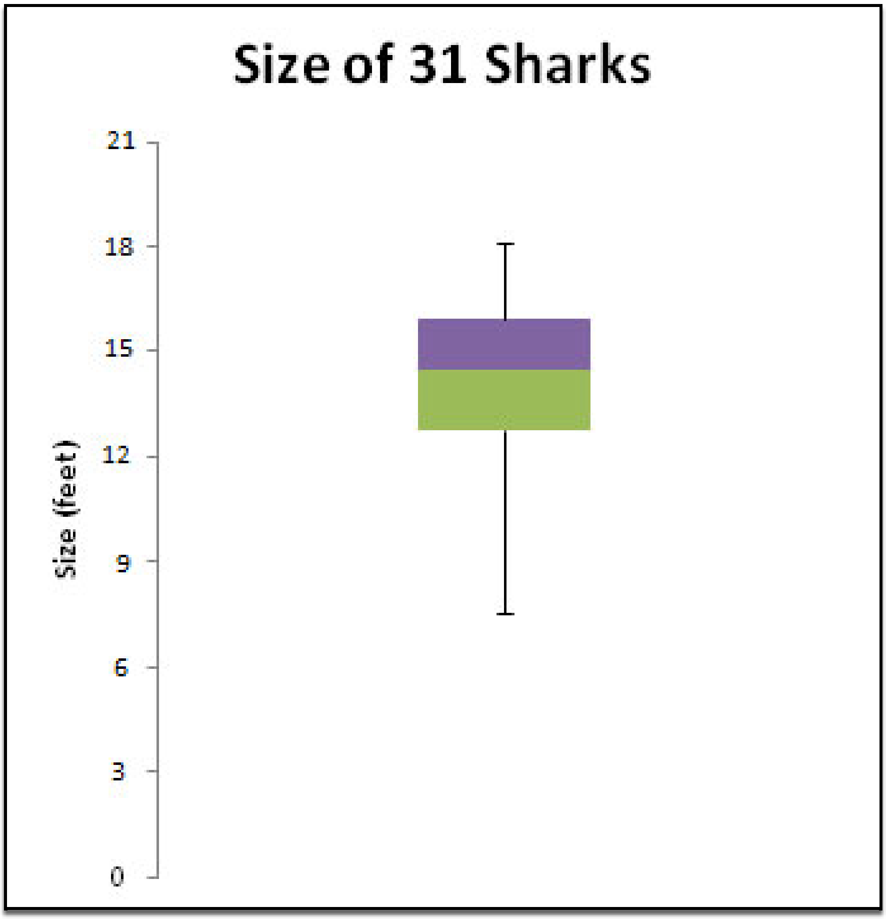
The inter-quartile range is an indicator of the distribution of a sample and can also help single out an outlier. Outliers are observation points (numbers) that are distant from other observations. It is helpful to identify any outliers and determine whether they should be used.

Follow these steps to find the interquartile ranges of a data set.

1. Put the data set in order.
   2 4 6 8 10 12 14 16 18
2. Find the median, or mid-point, of the data set.
   2 4 6 8 10 12 14 16 18
3. Identify the median of the lower half of the data set, label as Q1.
   2 4 | 6 8 10 12 14 16 18
   In this case, the median of the lower half of the data set is between 4 and 6, which averages to 5.
4. Identify the median of the upper half of the data set, label as Q3.
   2 4 | 6 8 10 12 14 | 16 18
   In this case, the median of the upper half of the data set is between 14 and 16, which averages to 15.
5. Subtract Q1 from Q3 to determine the interquartile range.
   15 - 5 = 10

Box Plots

A box plot can also be known as a box-and-whiskers or hinge plot. Box plots are used while studying the composition of a data set to examine the distribution. The top of the line is the maximum value in the set, the upper limit of the box represents the third quartile (75th percentile), The line in the middle is the median, the lower limit of the box represents the first quartile (25th percentile), and the bottom of the line represents the minimum value of the data set. Box plots are useful when showing non-parametric data as it takes into account median and percentiles rather than averages.



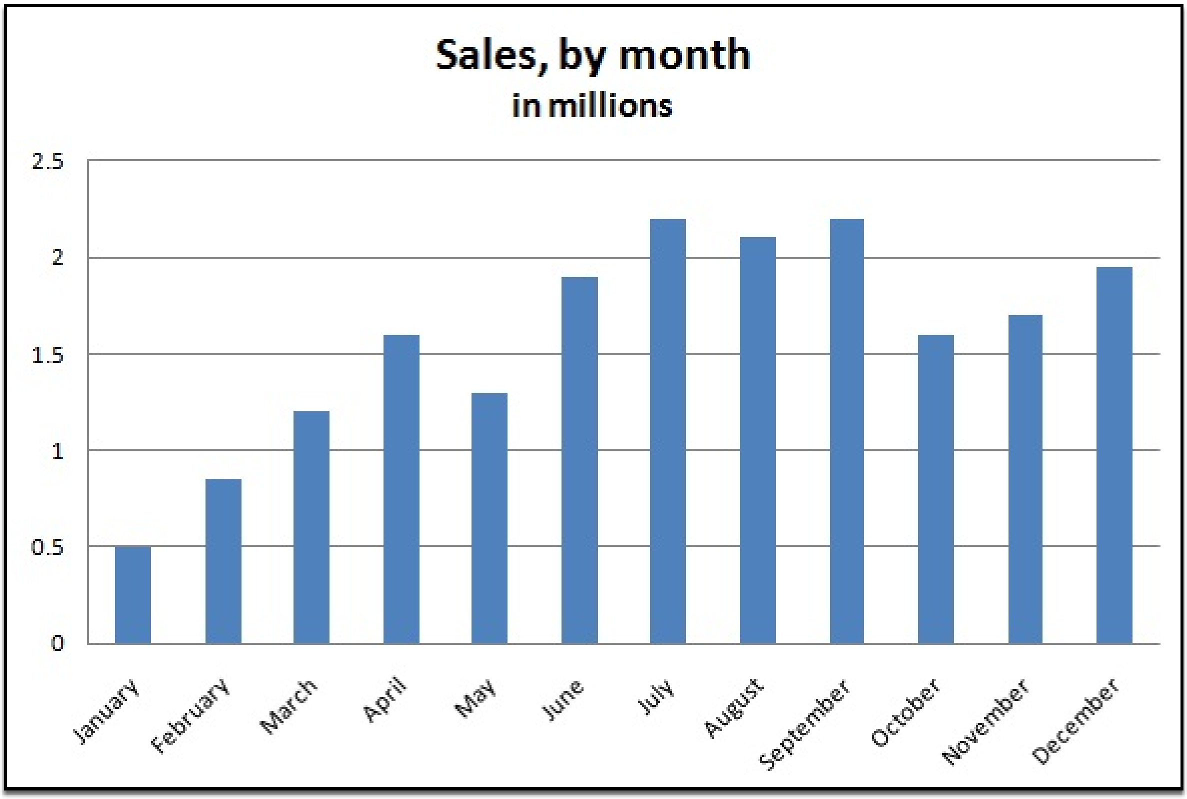
From this box plot one can see that most of the sharks are between 13 and 16 feet. One can also see that sharks don't grow that much past the standard shark and one might be able to make the assumption that the smaller sharks measured might have still been growing.

2.15 Graphic Displays: The Histogram and Bar Charts

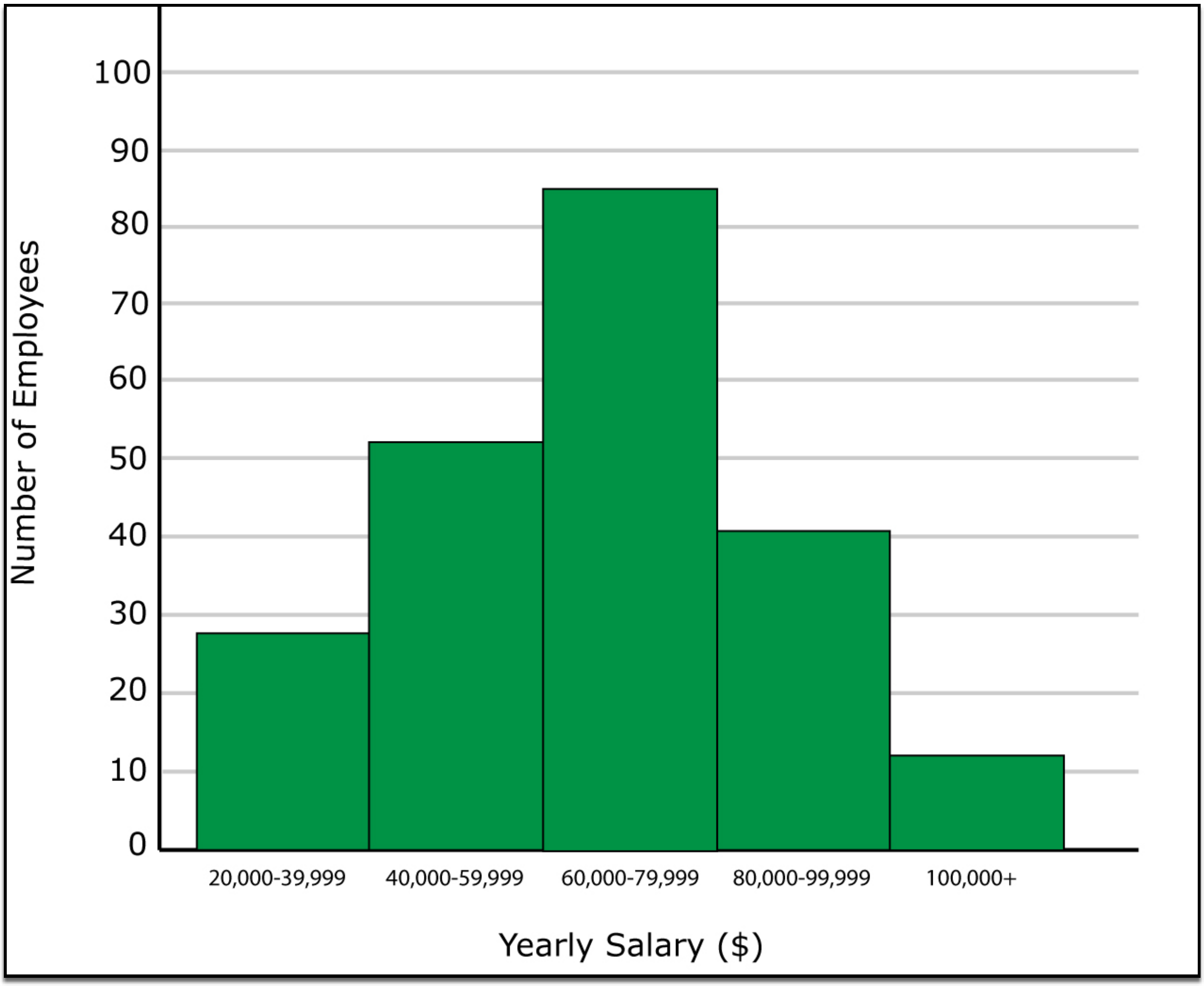
Histograms and Bar Charts

A histogram is a graph that displays continuous data. The vertical bars in a histogram show the counts or numbers in each range. A comparison of the ranges, or a review of the graph as a whole, helps the audience understand the information presented.

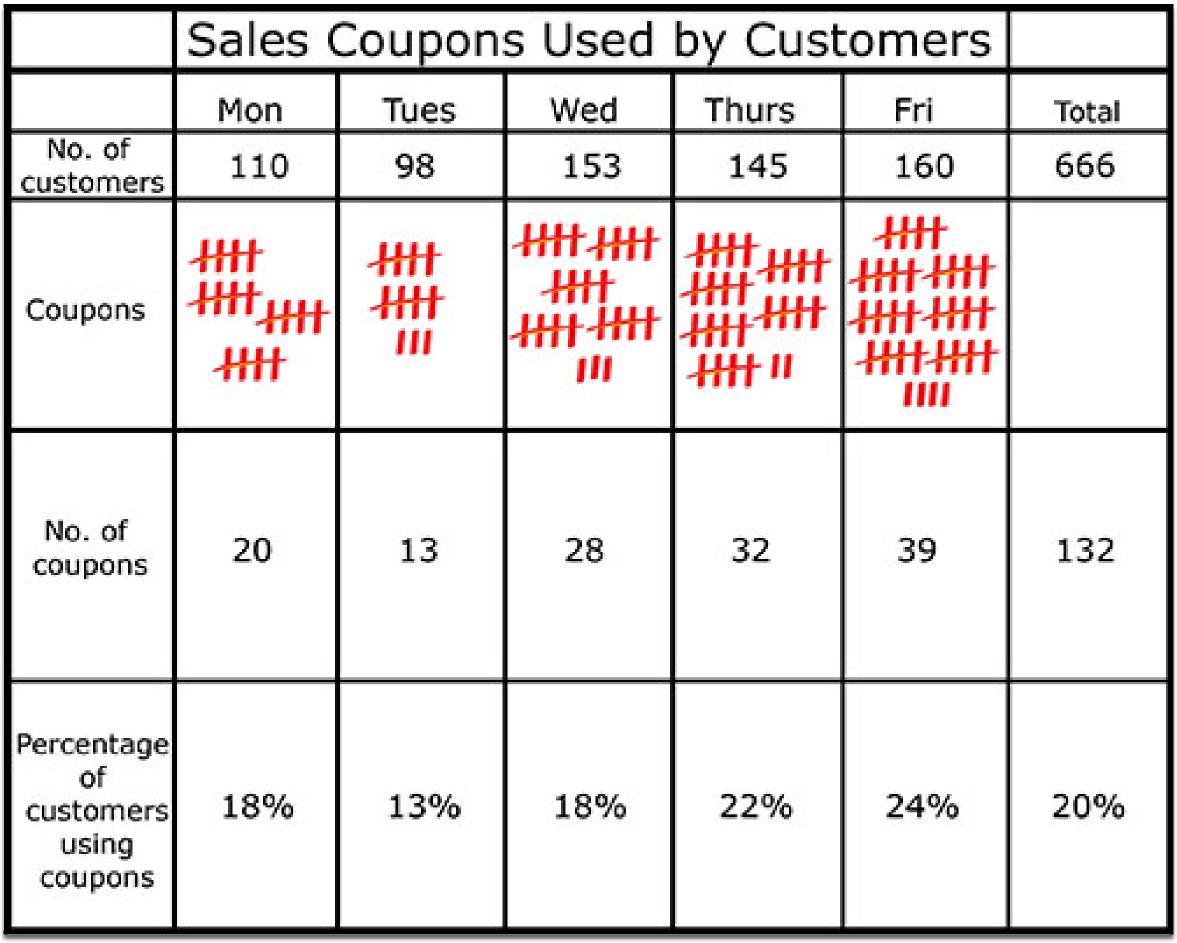
The distinction between a histogram and a bar chart is an important distinction to make. While a histogram measures how continuous data is distributed over various ranges, a bar chart measures data that is distributed over groups or categories. For example, a histogram would be appropriate to display how many people fall in various ranges of heights, as height is an example of continuous data. Bar charts, though, would be more appropriate to measure how many people are from each state, as states are an example of discrete categories. Here is an example of a bar chart:



Histograms allow team members and stakeholders to view a significant amount of data at one time, and to see how data is distributed across values and ranges. The histogram's bars represent the values or ranges in the study. The height of each bar shows how many observations or events fall into each range. The shape of the graph illustrates how the data is distributed. Here is an example of a histogram:



To create a histogram or bar chart, first collect the data using a check sheet. For a histogram, decide what ranges, or values, you are going to use and note how the data is distributed. For a bar chart, select the categories for your chart and separate your data into these categories. A sample check sheet that can be used for a bar chart is pictured below. This displays the number of coupons a cashier accepts from customers, by day, throughout the week.



	Sales Coupons Used by Customers					
	Mon	Tues	Wed	Thurs	Fri	Total
No. of customers	110	98	153	145	160	666
Coupons	𝍸 𝍸 𝍸 𝍸	𝍸 𝍸 III	𝍸 𝍸 𝍸 𝍸 𝍸 III	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 II	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 IIII	
No. of coupons	20	13	28	32	39	132
Percentage of customers using coupons	18%	13%	18%	22%	24%	20%

The categories in a bar chart, or ranges in a histogram, need to encompass all of the data collected. For a histogram, make sure the minimum and maximum values are accounted for, as well as every value in between. For a bar chart, use an appropriate number of categories; if you have difficulty determining how many categories to use, you can refer to the rough guidelines laid out in the chart below:

If you have:	Divide the data into:
Fewer than 50 measurements	5 to 7 categories
50 to 100 measurements	6 to 10 categories
100 to 250 measurements	7 to 12 categories
Greater than 250 measurements	10 to 20 categories

Make sure that the categories or ranges are comparable and exhaustive. For a bar chart, the categories should be similar, such as each day of the week being a category or each baseball team being a category. This will yield results that can be easily interpreted through visualization. For a histogram, the ranges should run consecutively, so that all data is accounted for and the visual representation of the graph is accurate.

2.16 Graphic Displays: The Scatter Diagram

Bivariate Charts: The Scatter Diagram

One of the most common types of charts are bivariate charts. Bivariate charts have a vertical y-axis and a horizontal x-axis that measure two different variables. A common mistake when drawing bivariate charts is to put the independent variable on the y-axis. This is important because when we look at a graph, we naturally look from left to right to see how the dependent vertical variable has reacted over time. Time is the most common independent variable. Because of this, we assume the independent variable is on the x-axis. There are two common bivariate charts: Scatter Diagrams and Line Graphs.

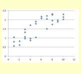
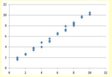
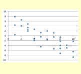
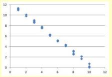
A scatter diagram, also known as a scatterplot, helps to show potential relationships or correlations between two variables. Data points are plotted as dots along an XY axis, and the concentration or dispersion of these dots shows the strength of the interaction between the variables.

To construct a scatter diagram, collect data as "pairs" based on the axes you need for the display. Place the data points as dots on the chart.

Correlations between variables are very easy to spot on a scatter diagram. If a correlation does exist, the data points on the diagram will line up along a curve or straight line across the chart. If the data does not fall along a curve or line, it is likely that no correlation exists between the variables.

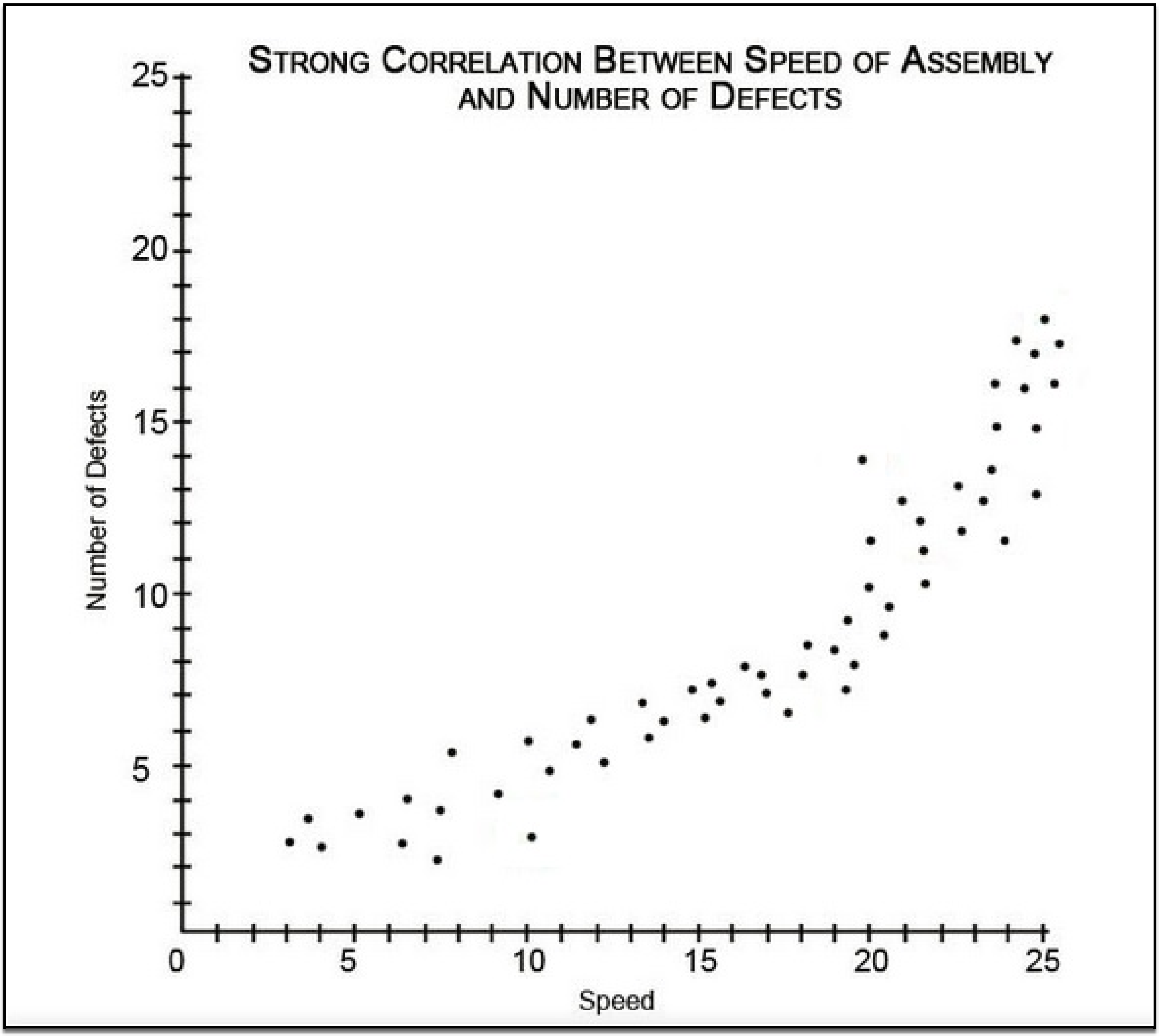
The correlations suggested by scatter diagrams can be positive (as one variable increases, the other also increases) or negative (as one variable increases, the other variable decreases). They can also be classified as weak (the data points are loosely collected) or strong (the data points are clumped together in a tight band).

Correlation in Scatter Diagrams

	WEAK	**STRONG**
Positive		
Negative		

Regardless of the correlation suggested, practitioners should be careful not to assume that a correlation among data proves that one variable causes another—it is possible that both of the variables in question are affected by some other factor, or that one variable is a subset of the other. Care must also be taken to ensure that the diagram is not taken at face value since the shape of the diagram and the implied correlation can be easily misunderstood or manipulated by adjusting the length or the scales of the diagram's axes.

Example



In this scatter diagram, there is a strong positive correlation between the speed of assembly and the number defects that occur. It is strong because all of the points are grouped rather close to each other. It is positive because as the speed of assembly increases, the number of defects increases, and vice versa. From this scatter diagram, a company might realize that they have to determine if they favor a slower speed of assembly and fewer defects, or a faster speed of assembly but more defects.

2.17 Exercise: The Scatter Diagram

This assignment does not contain any printable content.

2.18 Graphic Displays: Line Graphs

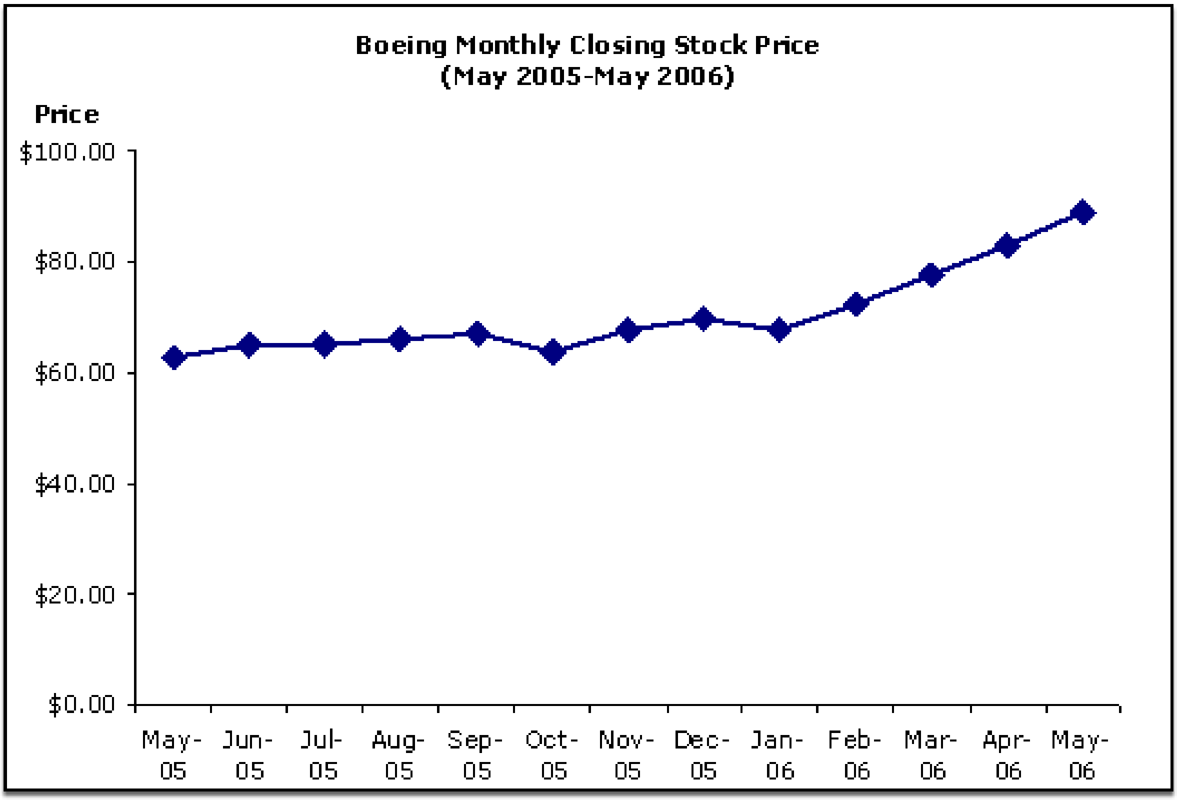
Bivariate Charts: Line Graphs

The other common bivariate chart is a line graph. A line graph plots the relationship between two or more variables by using connected data points. Line graphs are very useful where there is time series data to be summarized. They are appropriate where the data are continuous.

Consider the following data table for the stock price of Boeing in 2005 and 2006:

Date	Closing Stock Price
2-May-2005	$62.94
1-Jun-2005	$65.01
1-Jul-2005	$65.02
1-Aug-2005	$66.26
1-Sept-2005	$67.18
3-Oct-2005	$63.91
1-Nov-2005	$67.67
1-Dec-2005	$69.71
3-Jan-2006	$67.79
1-Feb-2006	$72.44
1-Mar-2006	$77.66
3-Apr-2006	$83.17
1-May-2006	$88.94

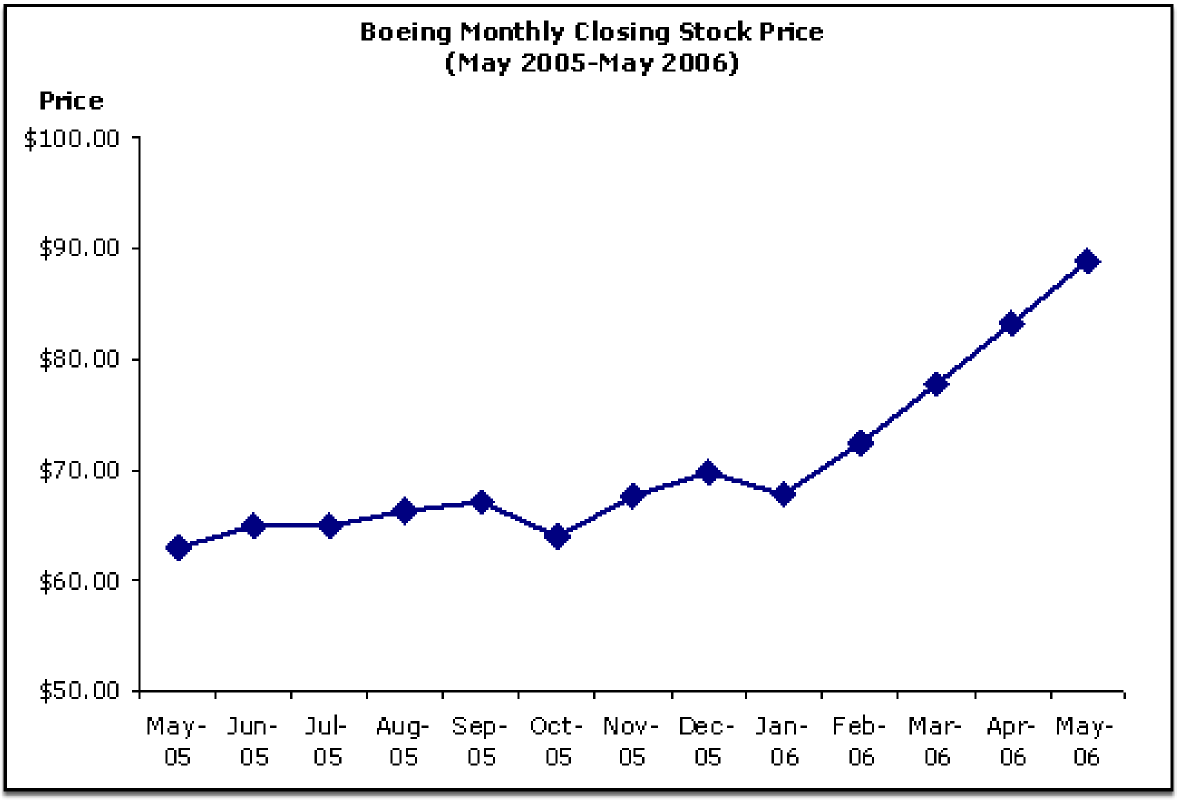
Now consider that same data presented as a line graph:



As you can see, the graph provides the better summary of the information.

Be Wary of Exaggerating Scales!

By adjusting the x- or y-axis scales you can exaggerate the effect a line chart displays. In the Boeing example, with a y-axis scale (closing price) of $0 to $100, we see a gradually increasing price trend over 12 months. However, consider the same data depicted in a line chart where the scale is adjusted to $50 to $100 instead of starting at $0.

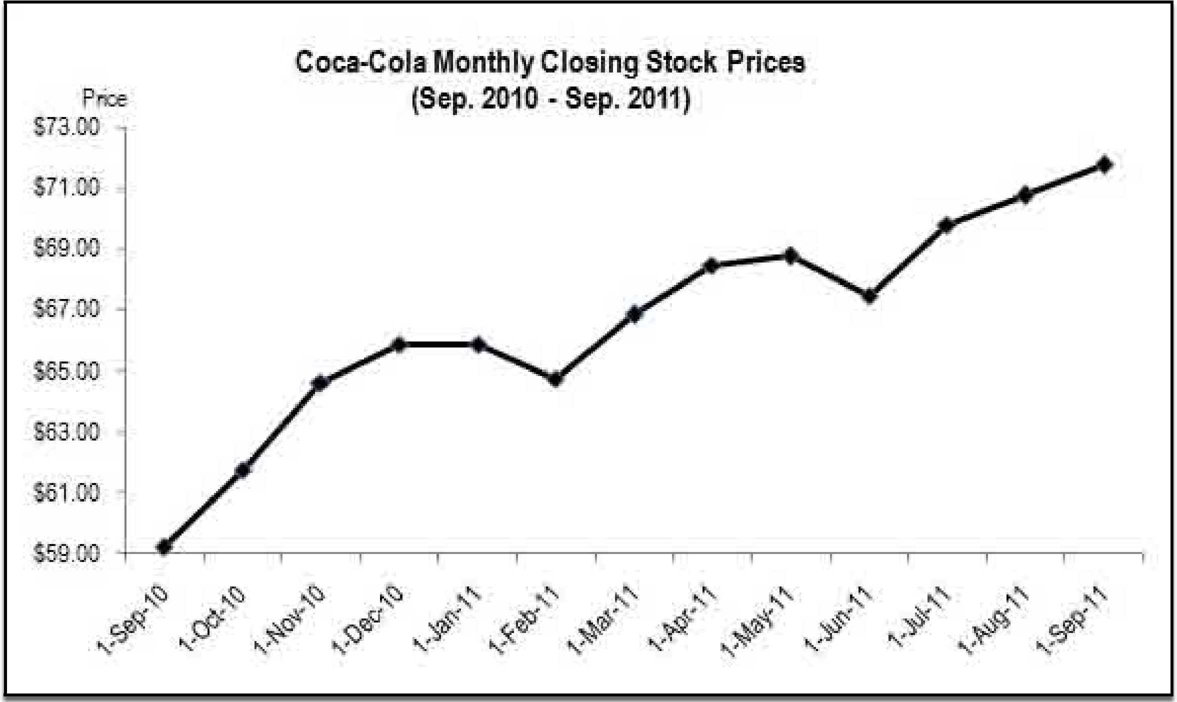


If the y-axis starts at $50 instead of $0, Boeing's stock increase looks far more dramatic than is actually the case.

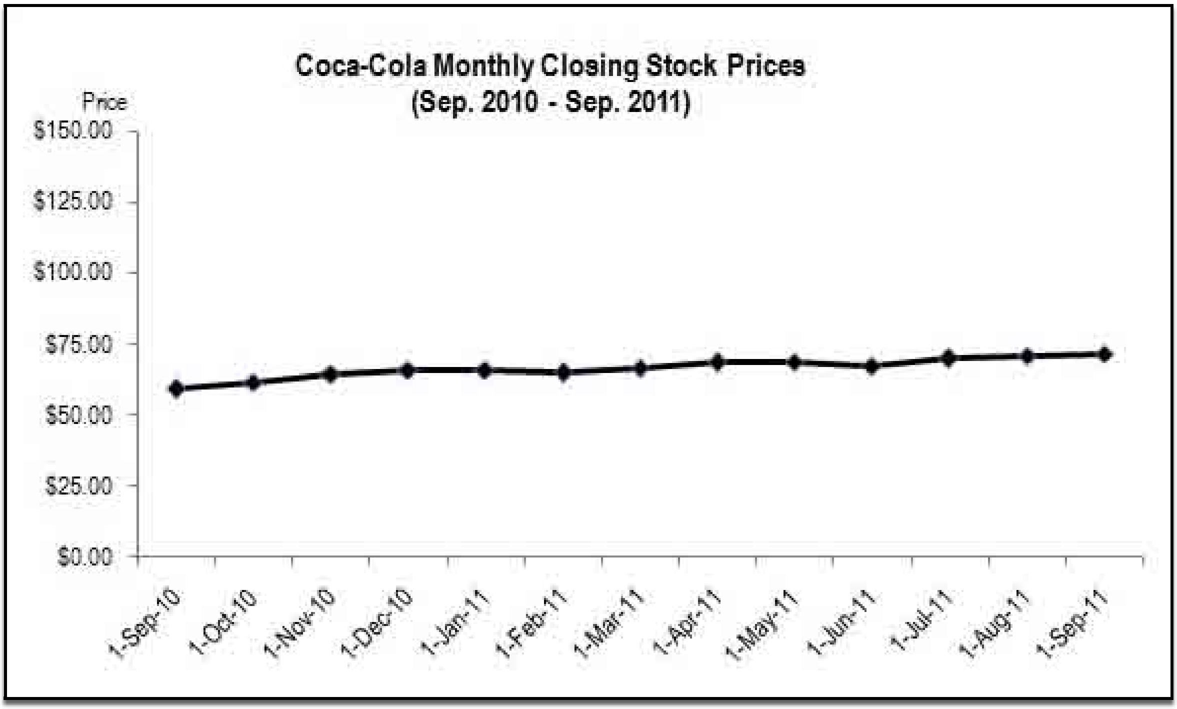
Coca-Cola's Stock Price

Even more manipulated graphs within the stock market can include the following graphs of Coca-Cola's stock price from September 2010 to September 2011:

Greatly Increasing Stock



Very Stable Stock

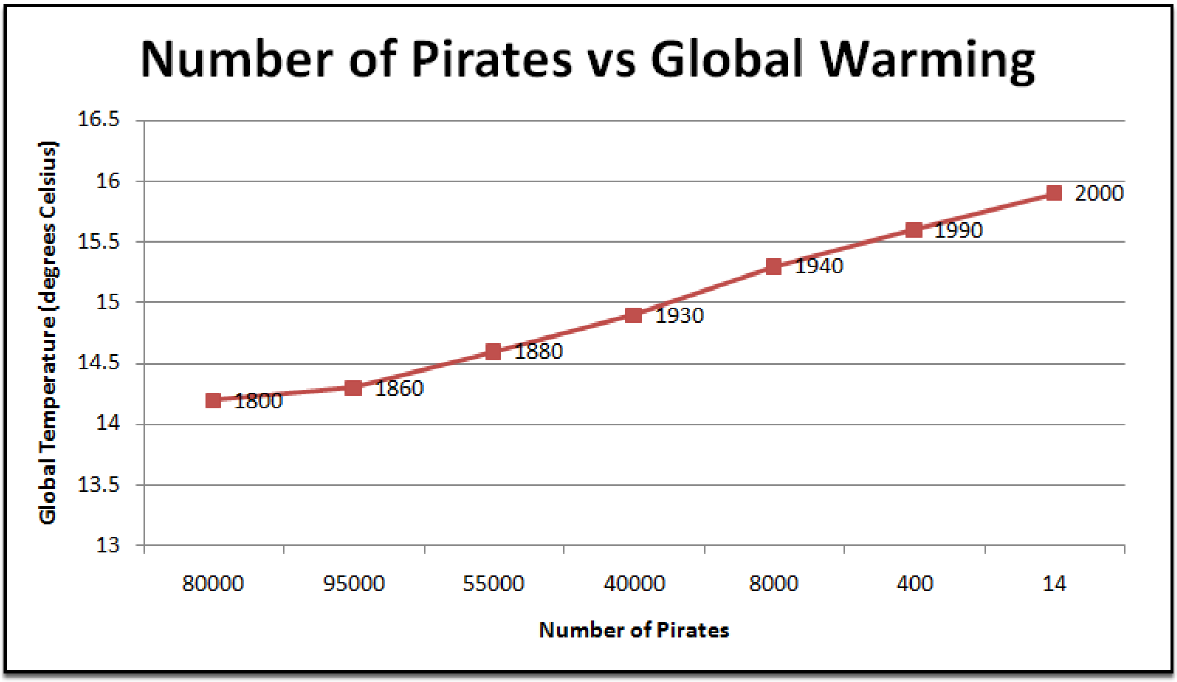


Both of these graphs are being manipulated. The first shows a drastic change in the stock price, while the other shows almost no change at all. There is some change occuring within Coca-Cola's stock throughout the year, but not as great a change as the first graph is communicating and a lot more of a change than the second graph is communicating. Within both of these graphs, the scale is being changed to support the agenda of the graph's creator.

Bottom Line: do not manipulate a graphic to tell an incorrect story and be sure to understand the scales shown on a graph's axes.

Common Mistakes in Graphs

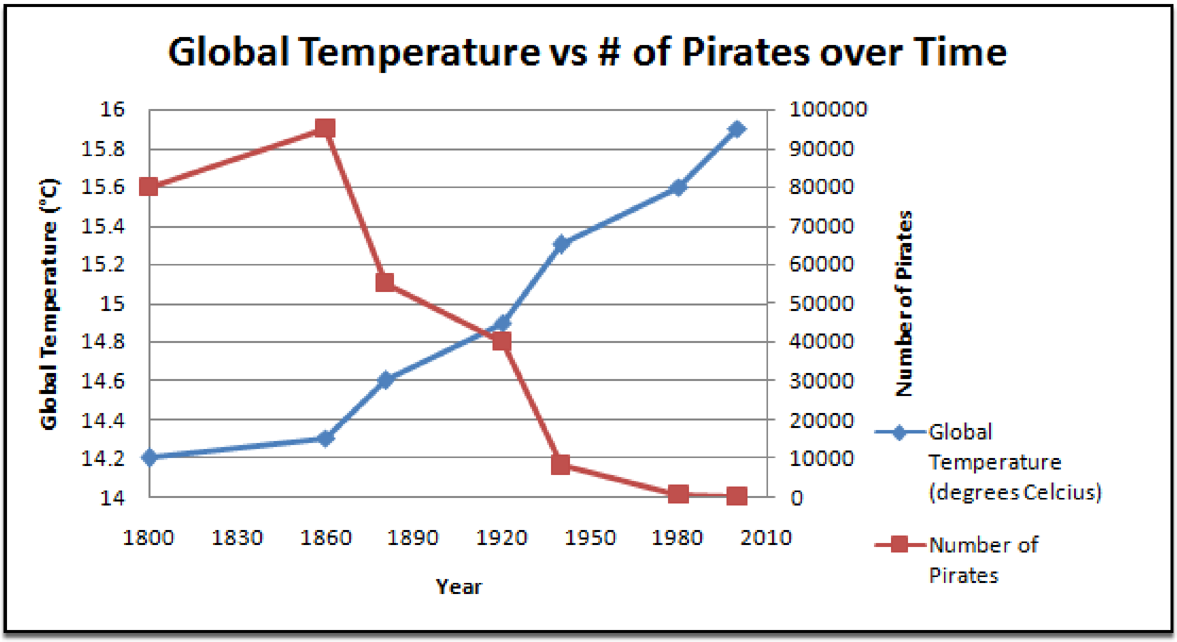
Let's take a look at the Pirates vs. Global Warming graph from Causation vs. Association. How many flaws can you find in the graph below?



The following flaws are in this graph.

- The title says "Number of Pirates vs. Global Warming". This is not what is shown in the graph. The graph shows the relationship of global average temperature over time. The line in the graph itself does not show any information about the number of pirates.

- The x-axis information is not to scale. The scale's increments vary both in the number of pirates per interval and by both increasing and decreasing. The scale of a graph is important as it gives an idea of how the variables are changing. The scale of the graph does not always have to have equal increments, but it does have to have consistently changing increments. This graph does not.
- Even if the label of the x-axis (Pirates) actually represented time (as it is supposed to), it should go up in number from left to right. There are rare occasions to go down from left to right. This graph does neither. It jumps up from 80,000 pirates to 95,000 and then falls down to 55,000. Changing from an increase to a decrease along an axis, and vice-versa only creates confusion.
- The main problem with this graph is that it is trying to show three variables (global temperature, time, number of pirates) while only showing one line, which, in a two-dimensional plane, can only show two variables. This problem can be solved by either having two lines on one graph or two separate graphs with the independent variable as time in both. The following is what the graph would look like if one were to have all three variables on the same graph, although one could also present the data with two different graphs.



2.19 Video: Inferential Statistics

This assignment does not contain any printable content.

2.20 Vocabulary Game

This assignment does not contain any printable content.

2.21 Flashcards

Flashcards

Term	Definition

2.22 Case Study: 5th Grade Weight and BMI Statistics

Case Study: 5th Grade Weight and BMI Statistics

The Oak Meadow Elementary School is currently conducting a study of 5th graders at the school to measure their health and fitness levels through multiple key indicators. In addition to a series of physical tests and questions about lifestyle habits, school officials have collected the weight and body mass index (BMI) for all boys and girls in the grade which is included in the study. BMI is a measure, based on a person's weight and height, that is used to classify people as underweight or overweight. The following table shows the data:

Student ID	Gender	Gender Variable	Age	Weight	BMI (%)
1	Male	1	10.30	60	11.9
2	Male	1	10.40	60	12.6
3	Female	2	10.30	62	13.5
4	Male	1	10.70	64	12.6
5	Male	1	11.1	65	17.1
6	Male	1	10.80	65	14.2
7	Female	2	10.70	65	12.2
8	Female	2	11.00	67	15.6
9	Female	2	10.3	68	12.9
10	Female	2	10.3	68	14.6
11	Female	2	10.30	68	14.4
12	Male	1	11.1	69	13.4
13	Male	1	10.5	69	14.4
14	Male	1	10.7	69	13.3
15	Male	1	10.4	70	13.7
16	Female	2	10.50	70	12.4
17	Female	2	11.00	71	15.1
18	Male	1	11.20	72	13.7
19	Female	2	11.2	72	19.6
20	Female	2	10.50	72	16.2
21	Male	1	10.90	73	16.8
22	Female	2	11.10	73	15.6
23	Female	2	10.3	74	18.8
24	Male	1	10.3	75	17.9
25	Male	1	10.40	75	15.5
26	Female	2	10.9	75	16.4
27	Female	2	10.3	75	17.5
28	Female	2	10.80	75	16.1
29	Female	2	10.70	75	16.1
30	Female	2	10.90	75	16.2
31	Female	2	10.90	76	16.7
32	Female	2	10.50	76	15.4
33	Female	2	10.40	77	14.5
34	Female	2	11.00	77	17
35	Female	2	10.30	77	14.1
36	Female	2	10.50	78	18.3
37	Male	1	10.40	79	15.2
38	Female	2	10.8	79	19.7
39	Male	1	10.80	80	17.3
40	Female	2	10.40	80	15.3
41	Female	2	10.50	80	16.5
42	Female	2	10.40	80	16.7
43	Male	1	10.8	81	16.3
44	Male	1	10.30	82	15.3
45	Female	2	11.10	82	16.2
46	Female	2	10.80	83	18.2
47	Female	2	10.70	83	17.3
48	Male	1	11	84	18.1

49	Male	1	10.70	84	16.9
50	Female	2	11.10	84	16.1
51	Male	1	10.60	85	15.8
52	Female	2	10.30	85	19.2
53	Male	1	10.80	86	16.9
54	Male	1	10.60	86	15.8
55	Female	2	10.80	86	14.6
56	Female	2	11	87	17.3
57	Male	1	11.2	88	19.7
58	Male	1	11.10	89	15.6
59	Male	1	10.3	90	18.2
60	Female	2	10.60	90	18.3
61	Male	1	12.20	91	14.8
62	Female	2	10.6	91	17
63	Female	2	10.40	91	17.6
64	Female	2	11.00	92	17.9
65	Male	1	10.60	93	18.9
66	Female	2	11.00	94	19.1
67	Male	1	10.70	95	18.4
68	Female	2	10.60	97	20.8
69	Male	1	11.20	98	14.9
70	Male	1	10.40	100	18.8
71	Male	1	10.50	103	19.4
72	Male	1	10.8	104	20.1
73	Male	1	10.8	104	22.5
74	Male	1	11	107	18.2
75	Male	1	11.1	113	19.2
76	Female	2	10.60	116	15.2
77	Male	1	10.40	117	11.8
78	Male	1	10.6	118	20.4
79	Male	1	10.5	120	21.9
80	Male	1	10.80	123	21.4

Review Checkpoint

Question 1. What variables are included in this data set?

Question 2. What type of variables are these?

Question 3. Is this sample or population data?

Question 4. What is the sample size (N)?

Question 5. Approximately what is the split between males and females?

Review the chart below showing the descriptive stats for Weight and BMI, then answer questions 6, 7, and 8 below.

	Weight	**BMI**	**Male Weight**	**Female Weight**	**Male BMI**	**Female BMI**
Mean	82.9	16.51	86.821	79.171	16.638	16.395
Median	80	16.35	85	77	16.8	16.2
Mode	75	18.2 OR 17.3 OR 16.2	69	75	18.2 OR 12.6	16.2
Population Standard Deviation	14.572	2.399	17.319	10.024	2.770	1.975
Population Variance	212.34	5.754	299.942	100.483	7.672	3.901
Skewness	0.927	0.166	0.429	1.258	0.172	-0.039
Range	63	10.7	63	54	10.7	8.6
Minimum	60	11.8	60	62	11.8	12.2
Maximum	123	22.5	123	116	22.5	20.8
Sum	6632	1321.1	3386	3246	648.9	672.2
Count	80	80	39	41	39	41

Question 6. How do the results in the table differ for males and females?

Question 7. What does the difference between the mean, median and mode for the above categories tell you? Based on the relationships between mean, median and mode, which categories (e.g. male BMI, female weight, overall BMI, etc.) will have the most skewed (asymmetrical) curves?

Question 8. What does the range of the four sets of descriptive statistics tell you? If you look at the ranges for the three different BMI data sets (male, female, overall), which data set do you think would be the most reliable.

Suggested/Sample Responses

Question 1:

Gender, age, weight and BMI

Question 2:

Gender is a discrete (or categorical) variable, further categorized as a nominal variable. There is no intrinsic ordering to the categories. Weight and BMI are continuous (or quantitative) variables. For these variables, there can be ordering of the categories into levels (such as high, medium, low — or in the case of BMI, underweight, average, overweight or obese).

Question 3:

Population data, as it includes all fifth grade students in the Oak Meadow Elementary School (the group that is being studied). Sample data would include a randomly selected (representative) subset of the 80 fifth graders.

Question 4:

80

Question 5:

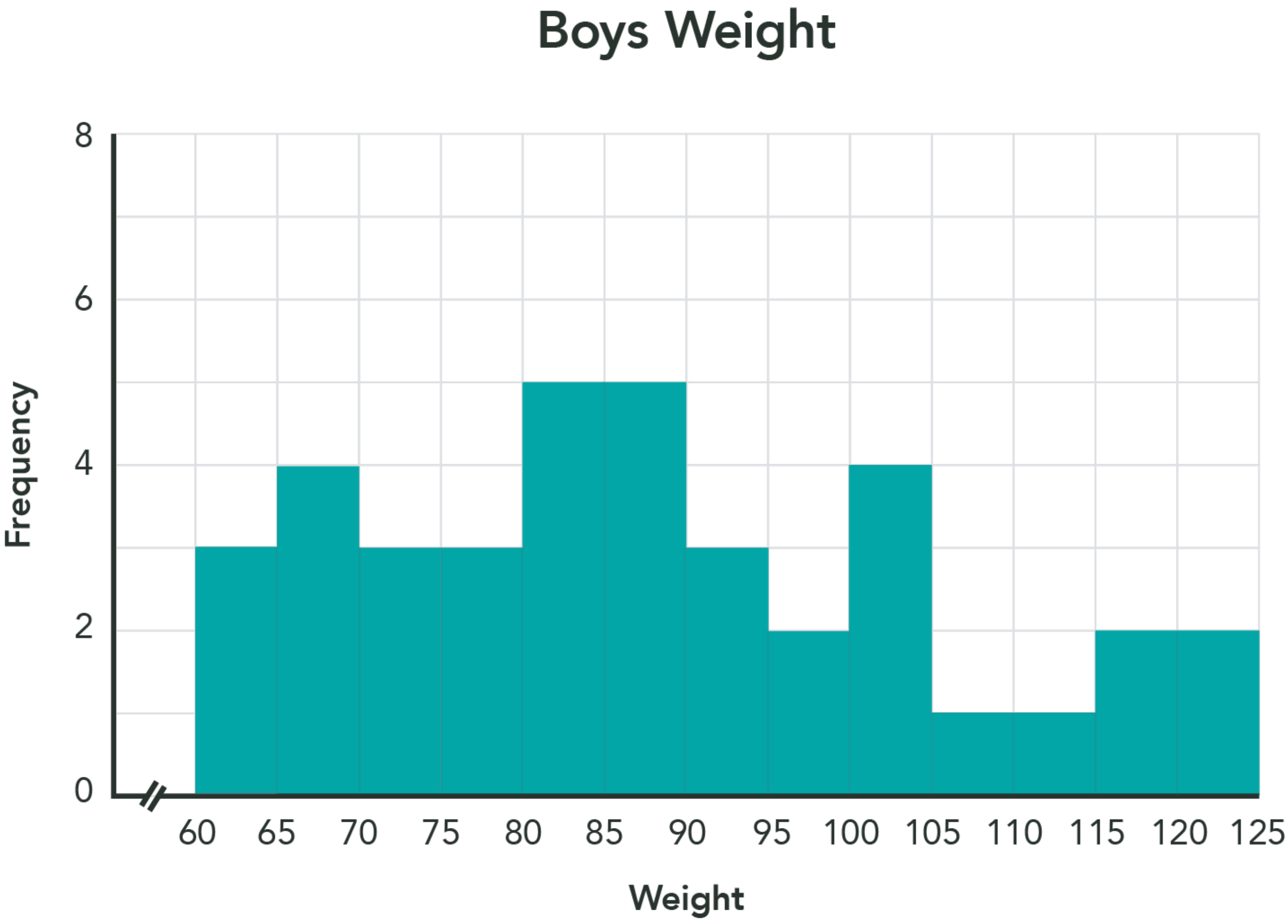
About 50/50 — using the "dummy variables" for male and female ("1" and "2"), it is possible to calculate a mean/average of 1.5, which is halfway between 1 and 2.

Question 6:

In general, males weigh more than females and have slightly higher BMIs. There is also more variation in the results for boys than for girls, with a broader range of values for males in both categories.

Question 7:

For most of the categories, the mean, median and mode are all quite close, which means they will have relatively symmetrical distributions. However, the male weight category has more variation between these measures, so there will be a greater degree of skew in its distribution.



Question 8:

The range is the difference between the highest and lowest data points in a set, and it measures the spread of the data values (or rather, how well the mean represents the data). The female BMI descriptive statistics have the smallest range (8.6, compared to 10.7 for overall and male) which means that the mean is a better representation of the overall data for females. In other words, there is less variability.

2.23 Case Study: Statistics and the Effectiveness of a Mayor's Crime Program

Case Study: Statistics and the Effectiveness of a Mayor's Crime Program

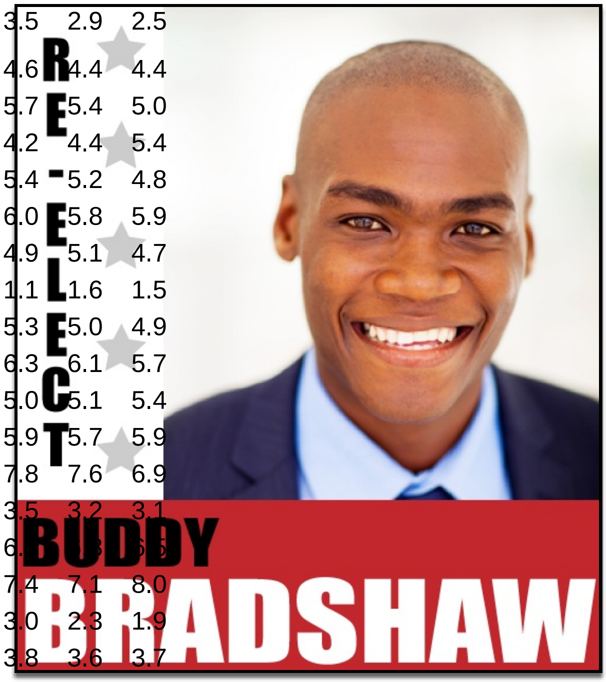
Meg Tarullo is the campaign manager for Buddy Bradshaw, an incumbent who is running for reelection as the mayor of the city of Brightville (population 560,000). At a meeting with an advocacy group called the "Concerned Citizens for a Safe Brightville," Tarullo will discuss the group's concerns about violent crime rates and gang activity in the city's neighborhoods.

Tarullo feels she has a very positive story to tell. Since Bradshaw entered office in 2006, he has worked with the Police Department to fund and promote a "CrimeStoppers" program that, according to his staff's analysis, has dramatically reduced violent crime.

Tarullo has spent the last few days working with the Brightville public safety record-keeping department to gather crime data and build charts that will show Bradshaw's success in battling crime during his two terms.

Tarullo began by compiling an analysis of all reported violent crimes committed in Brightville, broken down by neighborhood, during an eight-year period. The following table shows the data Tarullo gathered for the rate of violent crime per 100,000 residents.

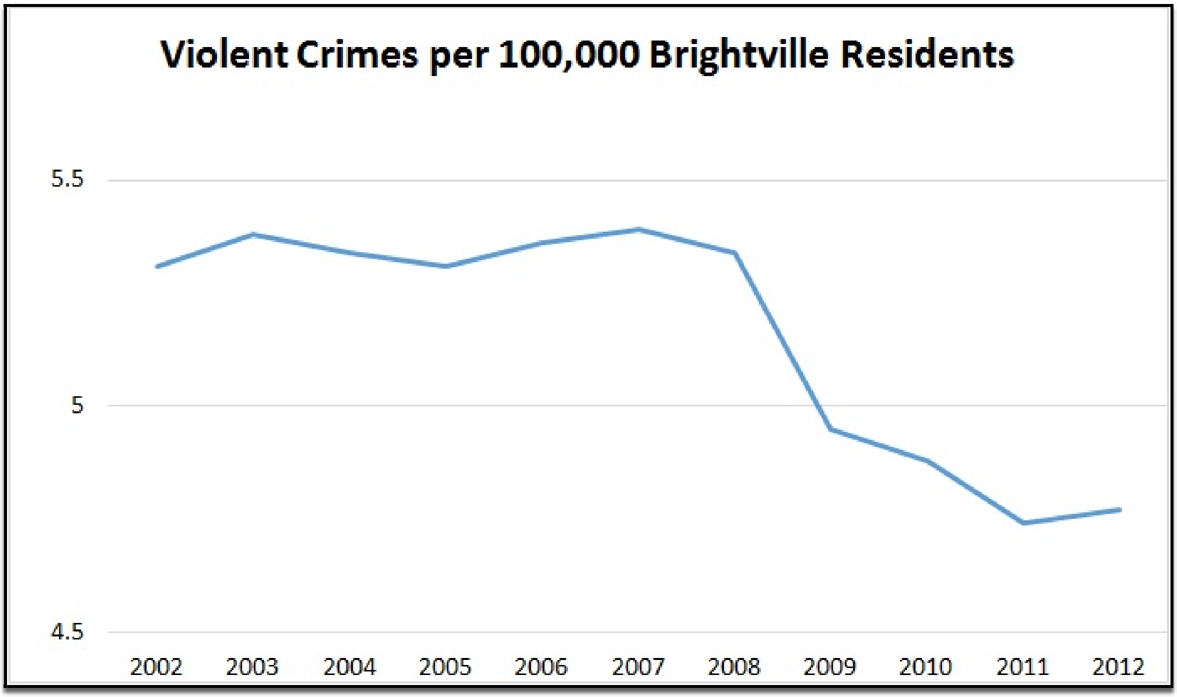
Neighborhood/District	2005	2006	2007	2008	2009	2010	2011	2012



Anson	**	**	**	**	3.7	3.5	2.9	2.5
Bingham	4.4	4.5	4.7	4.7	4.8	4.6	4.4	4.4
Covington	5.5	5.7	6.2	5.1	5.6	5.7	5.4	5.0
Drew	4.0	4.3	4.1	4.3	4.5	4.2	4.4	5.4
Eltonville	5.2	5.1	5.4	5.0	5.7	5.4	5.2	4.8
Franklin	5.8	5.6	5.7	6.0	6.6	6.0	5.8	5.9
Glover	4.7	4.4	3.4	3.4	4.1	4.9	5.1	4.7
Harrison	**	**	**	**	1.5	1.1	1.6	1.5
Ingersoll	5.1	5.5	6.3	6.5	5.1	5.3	5.0	4.9
Jasper	6.1	5.7	5.0	5.6	5.3	6.3	6.1	5.7
Kingsley	4.8	4.9	5.0	4.3	5.8	5.0	5.1	5.4
Lovell	5.7	5.9	5.8	6.9	6.2	5.9	5.7	5.9
Maxwell	7.6	7.9	7.4	8.1	8.1	7.8	7.6	6.9
Northside	**	**	**	**	3.6	3.5	3.2	3.1
Olson	6.4	6.3	6.3	5.9	6.6	[illegible]	6.3	6.5
Parklane	7.2	7.1	7.6	7.4	7.1	7.4	7.1	8.0
Quinlan	**	**	**	**	2.7	3.0	2.3	1.9
Ragsdale	3.6	3.3	3.1	3.3	4.1	3.8	3.6	3.7
Shelby	4.6	5.0	4.8	4.3	4.9	4.8	5.1	5.5
Tate	5.2	4.8	4.9	4.7	5.8	5.4	5.2	5.4
Usher	5.2	5.3	5.6	5.7	5.3	5.4	5.1	5.6
Vassar	4.8	**	**	**	1.6	1.7	1.2	1.1
Whitman	5.0	5.2	5.8	5.0	5.2	5.0	5.2	6.0
City Average Incidents per 100K Residents	**5.31**	**5.26**	**5.39**	**5.34**	**4.95**	**4.88**	**4.74**	**4.77**

*** Five surrounding neighborhoods were officially incorporated into the City of Brightville in 2009.*

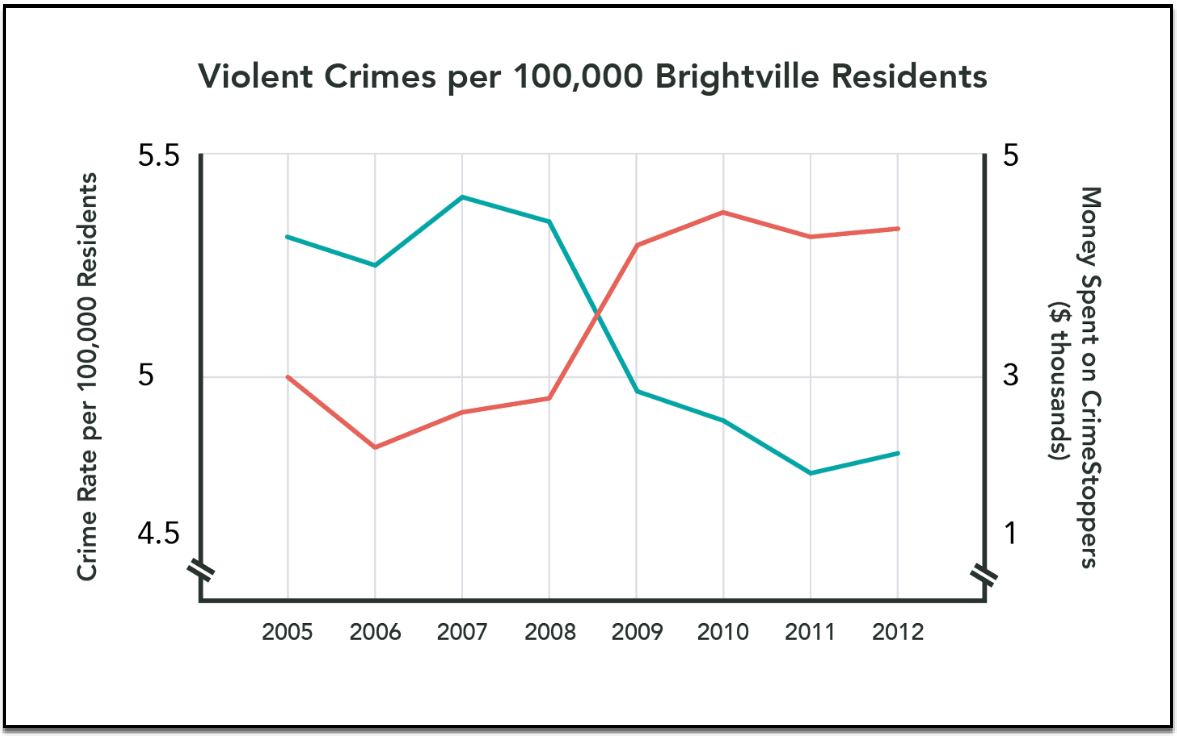
Based on this data, Tarullo created the following chart for the Advocacy Group presentation. She believes it graphically illustrates the success of Mayor Bradshaw's crime reduction program and his productive relationship with Brightville's Chief of Police. Her talking points emphasize the dramatic decline in the number of violent crimes (armed robberies, assaults, and murders) since Bradshaw took office.



Question 1. What are two major problems with the chart presented above? In what ways do these problems distort the

statistics to support Tarullo's narrative?

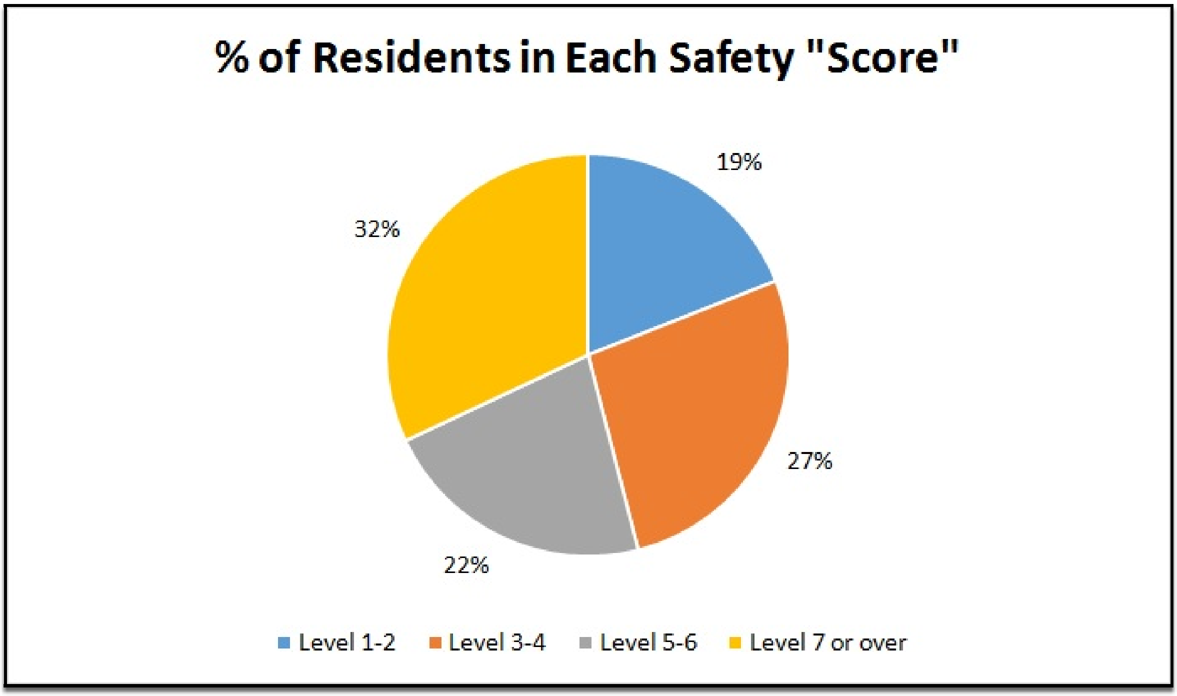
Next, Tarullo pulled together data to show the budget for CrimeStoppers throughout Bradshaw's term. She then charted the year-over-year decline in the city's murder rate and plotted each year's decrease against the amount of money allocated to the program during that same year. Noting that the city's budget tended to fluctuate based on the economy (tax revenue typically takes a hit during tough fiscal times), she was pleased to find a clear correlation between the amount of CrimeStoppers program spending and the size of corresponding murder rate declines. Further demonstrating the Mayor's effectiveness of program spending to decrease the murder rate, Tarullo added the following chart to illustrate this correlation:



Question 2. Do you agree that the chart above shows a clear causation between program spending and murder rate declines? Can you identify any problems with Tarullo's conclusion?

Question 3. How would hypothesis testing be useful for Tarullo here?

Next, Tarullo included the findings from a recent survey of households in Brightville. In this survey, which received a 15% response rate, city residents answered questions about their experiences (such as the number of times they have witnessed or been involved in a violent incident). The survey also asked them to rate their degree of personal security and safety, on a scale of 1-10. The following pie chart shows a high level of confidence on this measure, judging from the percent of residents giving high scores:



Rank	Percent of responses
1	6
2	13
3	12
4	15
5	12
6	10
7	10
8	9
9	8
10	5

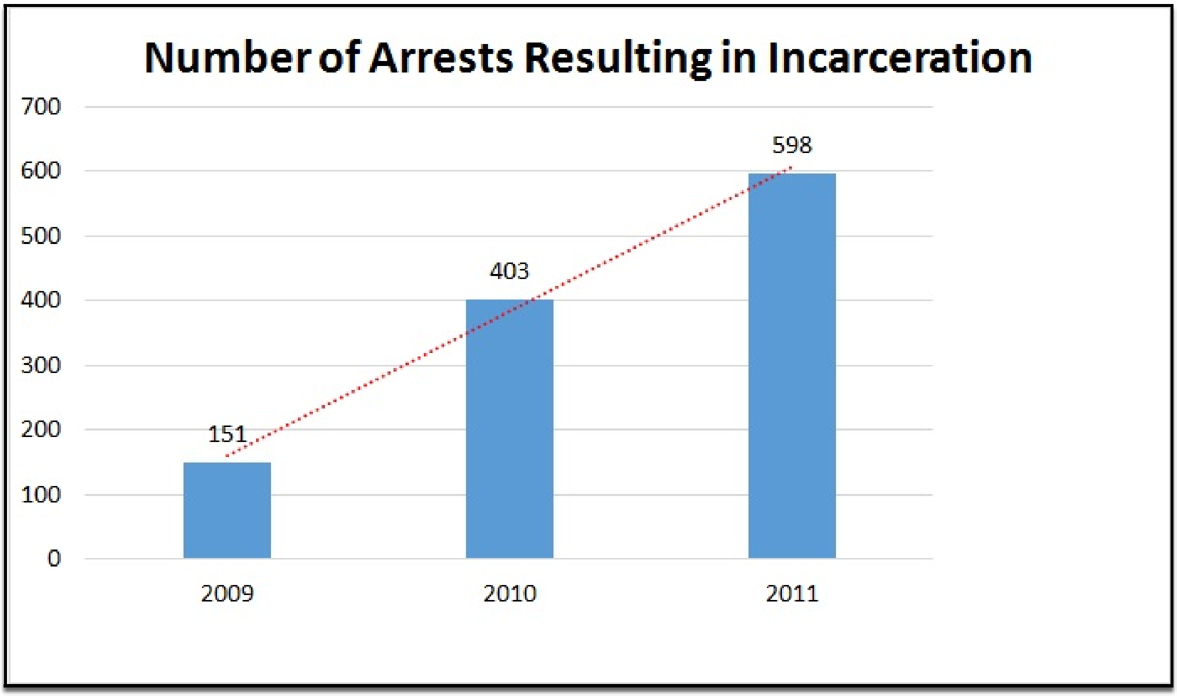
Question 4. The pie chart above and the data table above represent the same information. What's wrong with the pie chart? Why is it misleading? How would you adjust the pie chart to show an accurate representation of the data?

To summarize her presentation, Tarullo asserts that, since the Mayor's CrimeStoppers program has been enacted, Brightville has been a safer place. She believes that a key indicator of this is the number of arrests leading to incarceration and she highlights this data in a table:

Year	# of Arrests resulting in Incarceration
2002	317
2003	165
2004	344
2005	322
2006	386
2007	357
2008	232

2009	151
2010	403
2011	598
2012	396

She uses these numbers to prepare a chart to depict what her presentation characterizes as a "dramatic increase in the number of arrests resulting in incarceration."



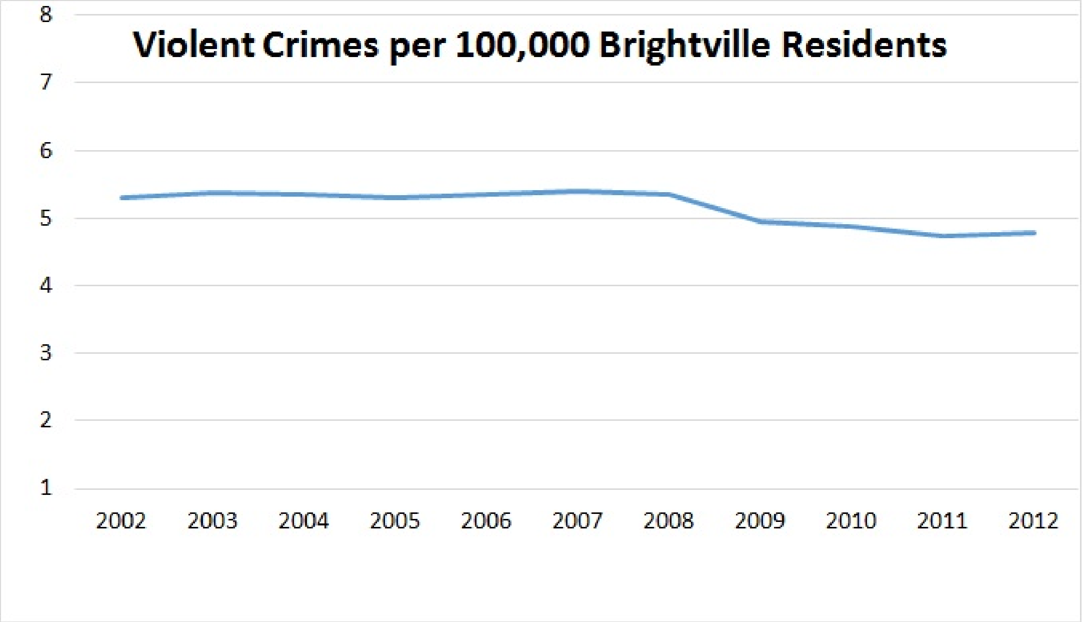
Question 5. How might the above chart be misleading and/or erroneous? Prepare a chart that shows a more accurate depiction of the data above.

Suggested/Sample Responses

Question 1:

1) Distortion of the sample set. In 2009, Brightville "grew" to encompass new, safer neighborhoods in its official population. This resulted in an inconsistency in measurement from the early years to the latter, as the safer neighborhoods skewed the results from 2009 onward. To get a meaningful trend, it would be necessary to compare apples to apples, either by excluding the 5 new neighborhoods for all years studied, or to include the data for those neighborhoods prior to their incorporation in Brightville.

2) The second problem is the manipulation of the axis scale on the graph. To make a set of relatively flat statistics (even with the inclusion of the safer neighborhoods from 2009-2012) look more dramatic, Tarullo compressed the y-axis scale to make the distinction between relatively small data points look much more significant. Compare her chart to the following chart based on the same data:



Question 2:

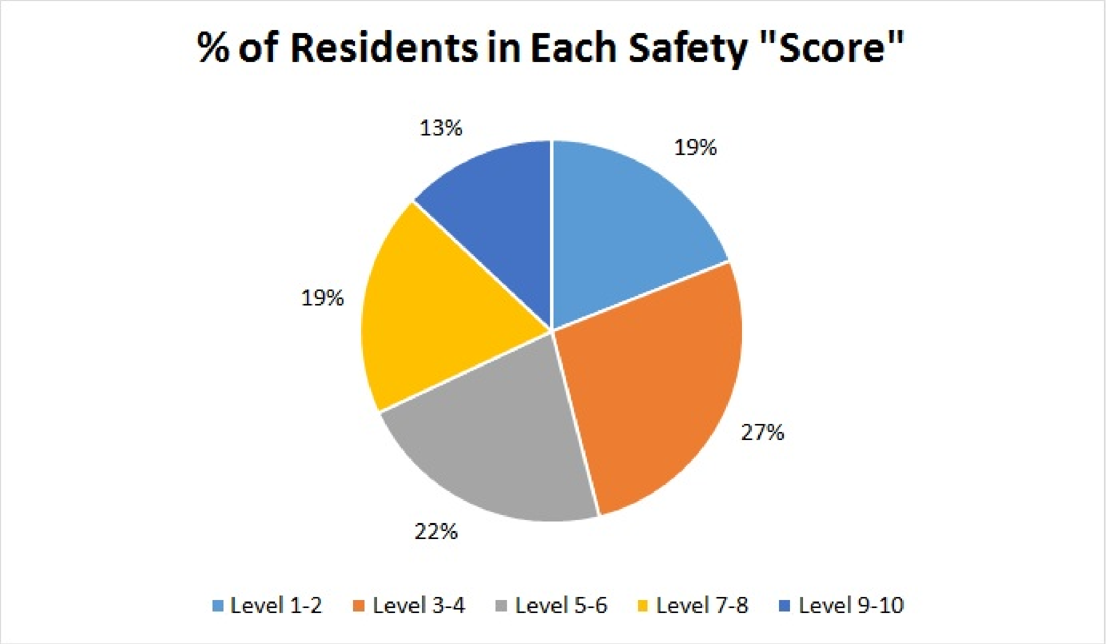
In a vacuum, the chart above tells a compelling story that correlates the budgetary spend with the effectiveness of CrimeStoppers. However, there are likely to be many variables (in addition to spend) that account for how much the murder rate declines or increases in a given year. For example, improvements in economic conditions (the conditions that create more opportunity for public spending) are the same conditions that often make it easier for people to find jobs, achieve better prosperity, allow for other economic development in communities, etc. In other words, the data shows correlation (two things that tend to happen at the same time) but not necessarily "causation." There could be many reasons for a decline in murder rates, aside from the fact that the Mayor spends more money on a program. In fact, many of these variables themselves are highly correlated with the economy, which could be the underlying cause of both the higher program spending and the lower murder rates. Therefore, it is impossible to separate out each one's individual effect on the decline in murder rates and to determine clear cause and effect. In short, Tarullo is right to conclude that there is a correlation between spending and the murder rate — that they tend to go together — but jumping to the conclusion that the spending is the "cause" is problematic.

Question 3:

Hypothesis testing would compel Tarullo to continue to test other possible causes and rule them out before concluding that this one dynamic is the true cause. By proving that other causes are incorrect, Tarullo would increase the certainty that her hypothesis (that program spending caused murder rate declines) cannot be proven wrong.

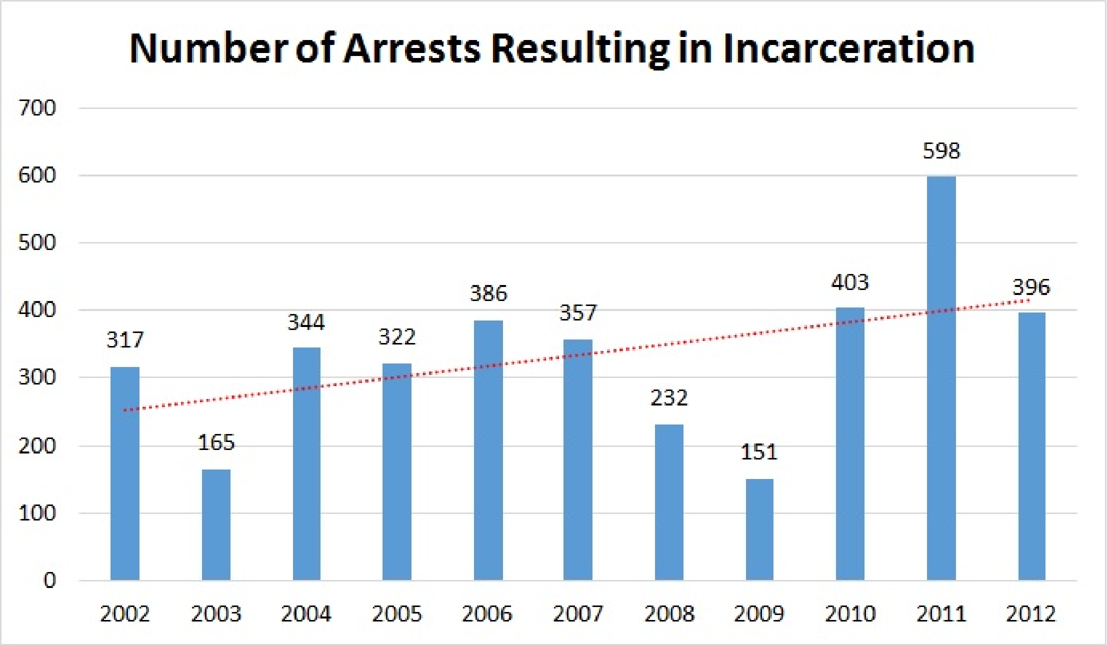
Question 4:

Inconsistent "intervals" or "Groupings" of results that skew the number of respondents in the "highest" category. If respondents were asked to rank their feelings with a number ranging from 1-10, the "buckets" of results must be equal — in other words, the pie chart is distorted by including 4 rankings (7, 8, 9, and 10) in one segment while only including 2 rankings in each of the other segments. This makes residents' ranking of personal safety and security seem higher, on the whole, than it actually is. An accurate pie chart would look more like this:



Question 5:

The chart above only displays portions of the data. In fact, the data for all years since the Mayor was elected (and several years prior) looks more like this. While 2011 saw a large number of arrests and 2009 saw a particularly low level, those appear to be outliers compared to the overall trend.



2.24 Module 2 Printable PDF

This assignment does not contain any printable content.

2.26 Module Feedback

This assignment does not contain any printable content.