Section summary

- The **population** is the entire group that the researchers are interested in. Because it is usually too costly to gather the data for the entire population, researchers will collect data from a **sample**, representing a subset of the population.
- A **parameter** is a true quantity for the entire population, while a **statistic** is what is calculated from the sample. A parameter is about a population and a statistic is about a sample. Remember: *p goes with p and s goes with s*.
- Two common summary quantities are **mean** (for numerical variables) and **proportion** (for categorical variables).
- Finding a good estimate for a population parameter requires a random sample; do not generalize from anecdotal evidence.
- There are two primary types of data collection: observational studies and experiments. In an **experiment**, researchers impose a treatment to look for a causal relationship between the treatment and the response. In an **observational study**, researchers simply collect data without imposing any treatment.
- Remember: *Correlation is not causation*! In other words, an association between two variables does not imply that one causes the other. Proving a causal relationship requires a well-designed experiment.

Section summary

- In an **observational study**, one must always consider the existence of **confounding factors**. A confounding factor is a "spoiler variable" that could explain an observed relationship between the explanatory variable and the response. Remember: For a variable to be confounding it must be associated with both the explanatory variable *and* the response variable.
- When taking a sample from a population, avoid **convenience samples** and **volunteer samples**, which likely introduce bias. Instead, use a **random** sampling method.
- Generalizations from a sample can be made to a population only if the sample is random. Furthermore, the generalization can be made only to the population from which the sample was randomly selected, not to a larger or different population.
- Random sampling from the entire population of interest avoids the problem of **undercoverage bias**. However, **response bias** and **non-response** bias can be present in any type of sample, random or not.
- In a **simple random sample**, every *individual* as well as every *group of individuals* has the same probability of being in the sample. A common way to select a simple random sample is to number each individual of the population from 1 to N. Using a random digit table or a random number generator, numbers are randomly selected without replacement and the corresponding individuals become part of the sample.
- A **systematic random sample** involves choosing from of a population using a random starting point, and then selecting members according to a fixed, periodic interval (such as every 10th member).
- A **stratified random sample** involves randomly sampling from *every* **strata**, where the strata should correspond to a variable thought to be associated with the variable of interest. This ensures that the sample will have appropriate representation from each of the different strata and reduces variability in the sample estimates.
- A **cluster random sample** involves randomly selecting a set of **clusters**, or groups, and then collecting data on all individuals in the selected clusters. This can be useful when sampling clusters is more convenient and less expensive than sampling individuals, and it is an effective strategy when each cluster is approximately representative of the population.
- Remember: *Individual strata should be homogeneous (self-similar), while individual clusters should be heterogeneous (diverse).* For example, if smoking is correlated with what is being estimated, let one stratum be all smokers and the other be all non-smokers, then randomly select an appropriate number of *individuals* from *each* strata. Alternately, if age is correlated with the variable being estimated, one could randomly select a *subset* of clusters, where each cluster has mixed age groups.

Section summary

- In an **experiment**, researchers impose a **treatment** to test its effects. In order for observed differences in the response to be attributed to the treatment and not to some other factor, it is important to make the treatment groups and the conditions for the treatment groups as similar as possible.
- Researchers use **direct control**, ensuring that variables that are within their power to modify (such as drug dosage or testing conditions) are made the *same* for each treatment group.
- Researchers **randomly** assign subjects to the treatment groups so that the effects of uncontrolled and potentially confounding variables are *evened out* among the treatment groups.
- **Replication**, or imposing the treatments on many subjects, gives more data and decreases the likelihood that the treatment groups differ on some characteristic due to chance alone (i.e. in spite of the randomization).
- An ideal experiment is **randomized**, **controlled**, and **double-blind**.
- A **completely randomized experiment** involves randomly assigning the subjects to the different treatment groups. To do this, first number the subjects from 1 to N. Then, randomly choose some of those numbers and assign the corresponding subjects to a treatment group. Do this in such a way that the treatment group sizes are balanced, unless there exists a good reason to make one treatment group larger than another.
- In a **blocked experiment**, subjects are first separated by a variable thought to affect the response variable. Then, within *each* block, subjects are randomly assigned to the treatment groups as described above, allowing the researcher to compare like to like within each block.
- When feasible, a **matched-pairs experiment** is ideal, because it allows for the best comparison of like to like. A matched-pairs experiment can be carried out on pairs of subjects that are meaningfully paired, such as twins, or it can involve all subjects receiving both treatments, allowing subjects to be compared to *themselves*.
- A treatment is also called a **factor** or explanatory variable. Each treatment/factor can have multiple **levels**, such as yes/no or low/medium/high. When an experiment includes many factors, multiplying the number of levels of the factors together gives the total number of treatment groups.
- In an experiment, blocking, randomization, and direct control are used to *control for confounding factors*.

Section summary

- A **scatterplot** is a **bivariate** display illustrating the relationship between two numerical variables. The observations must be **paired**, which is to say that they correspond to the same case or individual. The linear association between two variables can be positive or negative, or there can be no association. **Positive association** means that larger values of the first variable are associated with larger values of the second variable. **Negative association** means that larger values of the first variable are associated with smaller values of the second variable. Additionally, the association can follow a linear trend or a curved (nonlinear) trend.
- When looking at a **univariate** display, researchers want to understand the distribution of the variable. The term **distribution** refers to the values that a variable takes and the frequency of those values. When looking at a distribution, note the presence of clusters, gaps, and **outliers**.
- Distributions may be **symmetric** or they may have a long tail. If a distribution has a long left tail (with greater density over the higher numbers), it is **left skewed**. If a distribution has a long right tail (with greater density over the smaller numbers), it is **right skewed**.
- Distributions may be **unimodal**, **bimodal**, or **multimodal**.
- Two graphs that are useful for showing the distribution of a small number of observations are the **stem-and-leaf plot** and **dot plot**. These graphs are ideal for displaying data from small samples because they show the exact values of the observations and how frequently they occur. However, they are impractical for larger data sets.
- For larger data sets it is common to use a **frequency histogram** or a **relative frequency histogram** to display the distribution of a variable. This requires choosing bins of an appropriate width.
- To see cumulative amounts, use a **cumulative frequency histogram**. A **cumulative relative frequency histogram** is ideal for showing **percentiles**.
- **Descriptive statistics** describes or summarizes data, while **inferential statistics** uses samples to generalize or infer something about a larger population.

Section summary

- In this section we looked at univariate summaries, including two measures of **center** and three measures of **spread**.
- When **summarizing** or **comparing distributions**, always comment on center, spread, and shape. Also, mention outliers or gaps if applicable. Put descriptions in *context*, that is, identify the variable(s) being summarized by name and include relevant units. Remember: *Center, Spread, and Shape! In context!*
- **Mean** and **median** are measures of center. (A common mistake is to report **mode** as a measure of center. However, a mode can appear anywhere in a distribution.)
  - The **mean** is the sum of all the observations divided by the number of observations, n. $\bar{x} = \frac{1}{n}\sum x_i = \frac{\sum x_i}{n} = \frac{x_1+x_2+...+x_n}{n}$
  - In an ordered data set, the **median** is the middle number when n is odd. When n is even, the median is the average of the two middle numbers.
- Because large values exert more "pull" on the mean, large values on the high end tend to increase the mean more than they increase the median. In a **right skewed** distribution, therefore, the mean is greater than the median. Analogously, in a **left skewed** distribution, the mean is less than the median. Remember: *The mean follows the tail! The skew is the tail!*
- **Standard deviation (SD)** and **Interquartile range (IQR)** are measures of spread. SD measures the typical spread from the mean, whereas IQR measures the spread of the middle 50% of the data.
  - To calculate the standard deviation, subtract the average from each value, square all those differences, add them up, divide by $n-1$, then take the square root. Note: The standard deviation is the square root of the variance. $s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$
  - The IQR is the difference between the third quartile Q_3 and the first quartile Q_1. $IQR = Q_3 - Q_1$
- **Range** is also sometimes used as a measure of spread. The range of a data set is defined as the difference between the maximum value and the minimum value, i.e. $max - min$.
- **Outliers** are observations that are extreme relative to the rest of the data. Two rules of thumb for identifying observations as outliers are:
  - more than 2 standard deviations above or below the mean
  - more than $1.5 \times IQR$ below Q_1 or above Q_3

  Note: These rules of thumb generally produce different cutoffs.
- Mean and SD are sensitive to outliers. Median and IQR are more robust and less sensitive to outliers.
- The **empirical rule** states that for normal distributions, about 68% of the data will be within one standard deviation of the mean, about 95% will be within two standard deviations of the mean, and about 99.7% will be within three standard deviations of the mean.
- **Linear transformations of data**. Adding a constant to every value in a data set shifts the mean but does not affect the standard deviation. Multiplying the values in a data set by a constant will multiply the mean and the standard deviation by that constant, except that the standard deviation must always remain positive.
- **Box plots** do not show the *distribution* of a data set in the way that histograms do. Rather, they provide a visual depiction of the **5-number summary**, which consists of: min, Q_1, Q_2, Q_3, max. It is important to be able to identify the median, IQR, and direction of skew from a box plot.

Section summary

- When an outcome depends upon a chance process, we can define the **probability** of the outcome as the proportion of times it would occur if we repeated the process an infinite number of times. Also, even when an outcome is not truly random, modeling it with probability can be useful.
- The **Law of Large Numbers** states that the **relative frequency**, or proportion of times an outcome occurs after n repetitions, stabilizes around the true probability as n gets large.
- The probability of an event is always between 0 and 1, inclusive.
- The probability of an event and the probability of its **complement** add up to 1. Sometime we use $P(A) = 1 - P(\text{not } A)$ when $P(\text{not } A)$ is easier to calculate than $P(A)$.
- A and B are **disjoint**, i.e. **mutually exclusive**, if they cannot happen together. In this case, the events do not overlap and $P(A \text{ and } B) = 0$.
- In the *special case* where A and B are **disjoint** events: $P(A \text{ or } B) = P(A) + P(B)$.
- When A and B are not disjoint, adding $P(A)$ and $P(B)$ will overestimate $P(A \text{ or } B)$ because the overlap of A and B will be added twice. Therefore, when A and B are not disjoint, use the **General Addition Rule**:
  $P(A \text{ or B}) = P(A) + P(B) - P(A \text{ and } B)$.[20]
- To find the probability that *at least one* of several events occurs, use a special case of the rule of **complements**: $P(\text{at least one}) = 1 - P(\text{none})$.
- When only considering two events, the probability that one *or* the other happens is equal to the probability that *at least one* of the two events happens. When dealing with more than two events, the General Addition Rule becomes very complicated. Instead, to find the probability that A or B or C occurs, find the probability that none of them occur and subtract that value from 1.
- Two events are **independent** when the occurrence of one does not change the likelihood of the other.
- In the *special case* where A and B are **independent**: $P(A \text{ and } B) = P(A) \times P(B)$.

[20] Often written: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Section summary

- A **conditional probability** can be written as $P(A|B)$ and is read, "Probability of A given B". $P(A|B)$ is the probability of A, given that B has occurred. In a conditional probability, we are given some information. In an **unconditional probability**, such as $P(A)$, we are not given any information.
- Sometimes $P(A|B)$ can be deduced. For example, when drawing without replacement from a deck of cards, $P(\text{2nd draw is an Ace} \mid \text{1st draw was an Ace}) = \frac{3}{51}$. When this is not the case, as when working with a table or a Venn diagram, one must use the conditional probability rule $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$.
- In the last section, we saw that two events are **independent** when the outcome of one has no effect on the outcome of the other. When A and B are independent, $P(A|B) = P(A)$.
- When A and B are **dependent**, find the probability of A *and* B using the **General Multiplication Rule**: $P(A \text{ and } B) = P(A|B) \times P(B)$.
- In the *special case* where A and B are **independent**, $P(A \text{ and } B) = P(A) \times P(B)$.
- If A and B are **mutually exclusive**, they must be **dependent**, since the occurrence of one of them changes the probability that the other occurs to 0.
- When sampling **without replacement**, such as drawing cards from a deck, make sure to use **conditional probabilities** when solving *and* problems.
- Sometimes, the conditional probability $P(B|A)$ may be known, but we are interested in the "inverted" probability $P(A|B)$. **Bayes' Theorem** helps us solve such conditional probabilities that cannot be easily answered. However, rather than memorize Bayes' Theorem, one can generally draw a tree diagram and apply the conditional probability rule $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$. The resulting answer often has the form $\frac{w \times x \,+\, y \times z}{w \times x}$, where w, x, y, z are numbers from a tree diagram.

Section summary

- $\binom{n}{x}$, the **binomial coefficient**, describes the number of combinations for arranging x successes among n trials. $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, where $n! = 1 \times 2 \times 3 \times ...n$, and 0!=0.
- The **binomial formula** can be used to find the probability that something happens *exactly x times in n trials*. Suppose the probability of a single trial being a success is p. Then the probability of observing exactly x successes in n independent trials is given by

$$\binom{n}{x} p^x (1-p)^{n-x} \quad = \quad \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

- To apply the binomial formula, the events must be **independent** from trial to trial. Additionally, n, the number of trials must be fixed in advance, and p, the probability of the event occurring in a given trial, must be the same for each trial.
- To use the binomial formula, first confirm that the binomial conditions are met. Next, identify the number of trials n, the number of times the event is to be a "success" x, and the probability that a single trial is a success p. Finally, plug these three numbers into the formula to get the probability of exactly x successes in n trials.
- The $p^x(1-p)^{n-x}$ part of the binomial formula is the probability of just one combination. Since there are $\binom{n}{x}$ combinations, we add $p^x(1-p)^{n-x}$ up $\binom{n}{x}$ times. We can think of the binomial formula as: [# of combinations] $\times$ P(a single combination).
- To find a probability involving *at least* or *at most*, first determine if the scenario is binomial. If so, apply the binomial formula as many times as needed and add up the results. e.g. P(at least 3 Heads in 5 tosses of a fair coin) $=$ P(exactly 3 Heads) + P(exactly 4 Heads) + P(exactly 5 Heads), where each probability can be found using the binomial formula.

Section summary

- When a probability is difficult to determine via a formula, one can set up a **simulation** to estimate the probability.
- The **relative frequency** theory of probability and the **Law of Large Numbers** are the mathematical underpinning of simulations. A larger number of trials should tend to produce better estimates.
- The first step to setting up a simulation is to assign digits to represent outcomes. This should be done in such a way as to give the event of interest the correct probability. Then, using a random number table, calculator, or computer, generate random digits (outcomes). Repeat this a specified number of trials or until a given stopping rule. When this is finished, count up how many times the event happened and divide that by the number of trials to get the estimate of the probability.

Section summary

- A **discrete probability distribution** can be summarized in a table that consists of all possible outcomes of a random variable and the probabilities of those outcomes. The outcomes must be disjoint, and the sum of the probabilities must equal 1.
- A probability distribution can be represented with a histogram and, like the distributions of data that we saw in Chapter 2, can be summarized by its **center**, **spread**, and **shape**.
- When given a probability distribution table, we can calculate the **mean** (expected value) and **standard deviation** of a random variable using the following formulas.

$$\begin{aligned} E(X) = \mu_X &= \sum x_i \cdot P(x_i) \\ &= x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \cdots + x_n \cdot P(x_n) \\ Var(X) = \sigma_x^2 &= \sum (x_i - \mu_X)^2 \cdot P(x_i) \\ SD(X) = \sigma_X &= \sqrt{\sum (x_i - \mu_X)^2 \cdot P(x_i)} \\ &= \sqrt{(x_1 - \mu_X)^2 \cdot P(x_1) + (x_2 - \mu_X)^2 \cdot P(x_2) + \cdots + (x_n - \mu_X)^2 \cdot P(x_n)} \end{aligned}$$

  We can think of $P(x_i)$ as the *weight*, and each term is weighted its appropriate amount.
- The **mean** of a probability distribution does not need to be a value in the distribution. It represents the average of many, many repetitions of a random process. The **standard deviation** represents the typical variation of the outcomes from the mean, when the random process is repeated over and over.
- **Linear transformations**. Adding a constant to every value in a probability distribution adds that value to the mean, but it does not affect the standard deviation. When multiplying every value by a constant, this multiplies the mean by the constant and it multiplies the standard deviation by the absolute value of the constant.
- **Combining random variables**. Let X and Y be random variables and let a and b be constants.
  - The expected value of the sum is the sum of the expected values.

    $E(X + Y) = E(X) + E(Y)$

    $E(aX + bY) = a \times E(X) + b \times E(Y)$
  - When X and Y are **independent**: The standard deviation of a sum or a difference is the square root of the sum of each standard deviation squared.

    $SD(X + Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$

    $SD(X - Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$

    $SD(aX + bY) = \sqrt{(a \times SD(X))^2 + (b \times SD(Y))^2}$

  The SD properties require that X and Y be independent. The expected value properties hold true whether or not X and Y are independent.

Section summary

- A **Z-score** represents the number of standard deviations a value in a data set is above or below the mean. To calculate a Z-score use: $Z = \frac{x - \text{mean}}{SD}$.
- *Z-scores do not depend on units.* When looking at distributions with different units or different standard deviations, Z-scores are useful for comparing how far values are away from the mean (relative to the distribution of the data).
- The **normal distribution** is the most commonly used distribution in Statistics. Many distribution are approximately normal, but none are exactly normal.
- The empirical rule (68-95-99.7 Rule) comes from the normal distribution. The closer a distribution is to normal, the better this rule will hold.
- It is often useful to use the standard normal distribution, which has mean 0 and SD 1, to approximate a discrete histogram. There are two common types of **normal approximation problems**, and for each a key step is to find a Z-score.

  A: *Find the percent or probability of a value greater/less than a given x-value.*
  1. Verify that the distribution of interest is approximately normal.
  2. Calculate the Z-score. Use the provided population mean and SD to standardize the given x-value.
  3. Use a calculator function (e.g. `normcdf` on a TI) or a normal table to find the area under the normal curve to the right/left of this Z-score; this is the *estimate* for the percent/probability.

  B: *Find the x-value that corresponds to a given percentile.*
  1. Verify that the distribution of interest is approximately normal.
  2. Find the Z-score that corresponds to the given percentile (using, for example, `invNorm` on a TI).
  3. Use the Z-score along with the given mean and SD to solve for the x-value.
- Because the sum or difference of two normally distributed variables is itself a normally distributed variable, the normal approximation is also used in the following type of problem.

  Find the probability that a sum $X + Y$ or a difference $X - Y$ is greater/less than some value.
  1. Verify that the distribution of X and the distribution of Y are approximately normal.
  2. Find the mean of the sum or difference. Recall: the mean of a sum is the sum of the means. The mean of a difference is the difference of the means.
     Find the SD of the sum or difference using:
     $SD(X + Y) = SD(X - Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$.
  3. Calculate the Z-score. Use the calculated mean and SD to standardize the given sum or difference.
  4. Find the appropriate area under the normal curve.

Section summary

- The symbol $\bar{x}$ denotes the sample average. $\bar{x}$ for any particular sample is a number. However, $\bar{x}$ can vary from sample to sample. The distribution of all possible values of $\bar{x}$ for repeated samples of a fixed size from a certain population is called the **sampling distribution** of $\bar{x}$.
- The standard deviation of $\bar{x}$ describes the typical error or distance of the sample mean from the population mean. It also tells us how much the sample mean is likely to vary from one random sample to another.
- The standard deviation of $\bar{x}$ will be *smaller* than the standard deviation of the population by a factor of $\sqrt{n}$. The larger the sample, the better the estimate tends to be.
- Consider taking a simple random sample from a population with a fixed mean and standard deviation. The **Central Limit Theorem** ensures that regardless of the shape of the original population, as the sample size increases, the distribution of the sample average $\bar{x}$ becomes more normal.
- Three important facts about the sampling distribution of the sample average $\bar{x}$:
  - The mean of a sample mean is denoted by $\mu_{\bar{x}}$, and it is equal to μ. (*center*)
  - The SD of a sample mean is denoted by $\sigma_{\bar{x}}$, and it is equal to $\frac{\sigma}{\sqrt{n}}$. (*spread*)
  - When the population is normal or when $n \geq 30$, the sample mean closely follows a normal distribution. (*shape*)
- These facts are used when solving the following two types of **normal approximation** problems involving a *sample mean* or a *sample sum*.

  A: *Find the probability that a sample average will be greater/less than a certain value.*
  1. Verify that the population is approximately normal or that $n \geq 30$.
  2. Calculate the Z-score. Use $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to standardize the sample average.
  3. Find the appropriate area under the normal curve.

  B: *Find the probability that a sample sum/total will be greater/less than a certain value.*
  1. Convert the sample sum into a sample average, using $\bar{x} = \frac{sum}{n}$.
  2. Do steps 1-3 from Part A above.

Section summary

In the previous chapter, we introduced the binomial formula to find the probability of exactly x successes in n trials for an event that has probability p of success. Instead of looking at this scenario piecewise, we can describe the entire *distribution* of the number of successes and their corresponding probabilities.

- The distribution of the *number of successes* in n independent trials gives rise to a **binomial distribution**. If X has a binomial distribution with parameters n and p, then $P(X = x) = \binom{n}{x}p^x(1-p)^n - x$, where $x = 0, 1, 2, 3 \ldots, n$.
- To write out a binomial probability **distribution table**, list all possible values for x, the number of successes, then use the binomial formula to find the probability of each of those values.
- Because a binomial distribution can be thought of as the *sum* of a bunch of 0s and 1s, the **Central Limit Theorem** applies. As n gets larger, the shape of the binomial distribution becomes more normal.
- We call the rule of thumb for when the binomial distribution can be well modeled with a normal distribution the **success-failure** condition. The success-failure condition is met when there are at least 10 successes and 10 failures, or when $np \geq 10$ and $n(1-p) \geq 10$.
- If X follows a binomial distribution with parameters n and p, then:
  - The mean is given by $\mu_x = np$. $(center)$
  - The standard deviation is given by $\sigma_x = \sqrt{np(1-p)}$. $(spread)$
  - When $np \geq 10$ and $n(1-p) \geq 10$, the binomial distribution is approximately normal. $(shape)$
- It is often easier to use **normal approximation to the binomial distribution** rather than evaluate the binomial formula many times. These three properties of the binomial distribution are used when solving the following type of problem.

  Find the probability of getting more than / fewer than x yeses in n trials or in a sample of size n.

  1. Identify n and p. Verify than $np \geq 10$ and $n(1-p) \geq 10$, which implies that normal approximation is reasonable.
  2. Calculate the Z-score. Use $\mu_x = np$ and $\sigma_x = \sqrt{np(1-p)}$ to standardize the x value.
  3. Find the appropriate area under the normal curve.