Section Summaries and Chapter Highlights
Advanced High School Statistics
First Edition

David M Diez
david@openintro.org

Christopher D Barr
Yale School of Management
chris@openintro.org

Mine Çetinkaya-Rundel
Duke University
mine@openintro.org

Leah Dorazio
San Francisco University High School
leah@openintro.org

August 20, 2018

Preface

This document includes summaries of each section of **Advanced High School Statistics** (AHSS) as well as Chapter Highlights, which draw out and tie together the main concepts of each chapter. Though these summaries follow AHSS, they can be used to complement any introductory statistics text or course.

These summaries do not include any worked examples. They are not intended to be used to introduce or to teach the content. They are intended to be used after previous exposure to the material either in the classroom or via the textbook or its accompanying videos and slides (linked in the guide). These summaries will serve to clarify, consolidate, connect, and reinforce the main terms, concepts, and procedures. It is our hope that these Section Summaries and Chapter Highlights will be helpful to students as they study and review. Please send any and all comments on this document to leah@openintro.org.

Contents

Chapter 1

Data collection

1.1 Case study: using stents to prevent strokes

- To test the effectiveness of a treatment, researchers often carry out an experiment in which they randomly assign patients to a **treatment group** or a **control group**.
- Researchers compare the relevant **summary statistics** to get a sense of whether the treatment group did better, on average, than the control group.
- Ultimately, researchers want to know whether the difference between the two groups is significant, that is, larger than what would be expected by chance alone.

1.2 Data basics

- Researchers often summarize data in a table, where the rows correspond to individuals or **cases** and the columns correspond to the **variables**, the values of which are recorded for each individual.
- Variables can be **numerical** (measured on a numerical scale) or **categorical** (taking on levels, such as low/medium/high). Numerical variables can be **continuous**, where all values within a range are possible, or **discrete**, where only specific values, usually integer values, are possible.
- When there exists a relationship between two variables, the variables are said to be **associated** or **dependent**. If the variables are not associated, they are said to be **independent**.

1.3 Overview of data collection principles

- 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.

1.4 Observational studies and sampling strategies

- 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**. Instead, use a **random** sampling method.

- Random sampling avoids the problem of **selection bias**. However, **response bias** and **non-response** bias can be present in any type of sample, random or not.

- In a **simple random sample**, each individual of the population is numbered from 1 to N. Using a random digit table or a random number generator, numbers are randomly selected and the corresponding individuals become part of the sample.

- In a simple random sample, every *individual* as well as every *group of individuals* has the same probability of being in the sample.

- 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 the different strata and reduces variability in the sample estimates.

- A **cluster random sample** involves 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: *Strata should be self-similar, while clusters should be 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.

1.5 Experiments

- 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*.

Chapter Highlights

Chapter 1 focused on various ways that researchers collect data. The key concepts are the difference between a sample and an experiment and the role that randomization plays in each.

- Researchers take a **random sample** in order to draw an **inference** to the larger population from which they sampled. When examining observational data, even if the individuals were randomly sampled, a correlation does not imply a causal link.

- In an **experiment**, researchers impose a treatment and use **random assignment** in order to draw **causal conclusions** about the effects of the treatment. While often implied, inferences to a larger population may not be valid if the subjects were not also *randomly sampled* from that population.

Related to this are some important distinctions regarding terminology. The terms stratifying and blocking cannot be used interchangeably. Likewise, taking a simple random sample is different than randomly assigning individuals to treatment groups.

- **Stratifying vs Blocking**. Stratifying is used when sampling, where the purpose is to *sample* a subgroup from each stratum in order to arrive at a better *estimate* for the parameter of interest. Blocking is used in an experiment to *separate* subjects into blocks and then *compare* responses within those blocks. All subjects in a block are used in the experiment, not just a sample of them.

- **Random sampling vs Random assignment**. Random sampling refers to sampling a subset of a population for the purpose of inference to that population. Random assignment is used in an experiment to separate subjects into groups for the purpose of comparison between those groups.

When randomization is not employed, as in an **observational study**, neither inferences nor causal conclusions can be drawn. Always be mindful of possible **confounding factors** when interpreting the results of observation studies.

Chapter 2

Summarizing data

2.1 Examining numerical data

- A **scatterplot** is a statistical graph illustrating the relationship between two numerical variables. The variables must be **paired**, which is to say that they correspond to one another. 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 single variable, 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**.

2.2 Numerical summaries and box plots

- In this section we looked at two measures of **center** and two 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.)

  $\bar{x} = \frac{1}{n}\sum x_i = \frac{x_1+x_2+\ldots+x_n}{n}$ The **mean** is the sum of all the observations divided by the number of observations, 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.

  $s = \sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$ 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.

  $IQR = Q_3 - Q_1$ The IQR is the difference between the third quartile Q_3 and the first quartile Q_1.
- **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 approximately symmetric data, about 68% of the data will be within one standard deviation of the mean and about 95% will be within two 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.

- **Range** is defined as the difference between the maximum value and the minimum value, i.e. $max - min$.

- **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.

2.3 Considering categorical data

- **Categorical variables**, unlike numerical variables, are simply summarized by **counts** (how many) and **proportions**. These are referred to as frequency and relative frequency, respectively.

- When summarizing one categorical variable, a **one-way frequency table** is useful. For summarizing two categorical variables and their relationship, use a **two-way frequency table** (also known as a contingency table).

- To graphically summarize a single categorical variable, use a **bar plot**. To summarize and compare two categorical variables, use **side-by-side** or **segmented** (stacked) bar plots.

- **Pie charts** are another option for summarizing categorical data, but they are more difficult to read and bar charts are generally a better option.

Chapter Highlights

A raw data matrix/table may have thousands of rows. The data need to be summarized in order to makes sense of all the information. In this chapter, we looked at ways to summarize data **graphically**, **numerically**, and **verbally**.

Categorical data

- A single **categorical variable** is summarized with **counts** or **proportions** in a **one-way table**. A **bar graph** is used to show the frequency or relative frequency of the categories that the variable takes on.

- Two categorical variables can be summarized in a **two-way table** and with a **side-by-side bar plot** or a **segmented bar plot**.

Numerical data

- When looking at a single **numerical variable**, we try to understand the **distribution** of the variable. The distribution of a variable can be represented with a frequency table and with a graph, such as a **stem-and-leaf plot** or **dot plot** for small data sets, or a **histogram** for larger data sets. If only a summary is desired, a **box plot** may be used.

- The **distribution** of a variable can be described and summarized with **center** (mean or median), **spread** (SD or IQR), and **shape** (right skewed, left skewed, approximately symmetric).

- **Z-scores** and **percentiles** are useful for identifying a data point's relative position within a data set.
- **Outliers** are values that appear extreme relative to the rest of the data. Investigating outliers can provide insight into properties of the data or may reveal data collection/entry errors.
- When **comparing** the distribution of two variables, use two dot plots, two histograms, a back-to-back stem-and-leaf, or parallel box plots.
- To look at the **association** between two numerical variables, use a **scatter plot**.

Graphs and numbers can summarize data, but they alone are insufficient. It is the role of the researcher or statistician to ask questions, to use these tools to identify patterns and departure from patterns, and to make sense of this in the context of the data. Strong writing skills are critical for being able to communicate the results to a wider audience.

Chapter 3

Probability

3.1 Defining probability

- 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 Additional Rule**:
  $P(A \text{ or B}) = P(A) + P(B) - P(A \text{ and } B)$.[1]

- 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.

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

- 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**: $P(A \text{ and } B) = P(A) \times P(B)$.

3.2 Conditional probability

- 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.

3.3 The binomial formula

- $\binom{n}{k}$, the **binomial coefficient**, describes the number of combinations for arranging k successes among n trials. $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, 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 k times in n trials.*
- Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

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

- 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" k, 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 k successes in n trials.

- The $p^k(1-p)^{n-k}$ part of the binomial formula is the probability of just one combination. Since there are $\binom{n}{k}$ combinations, we add $p^k(1-p)^{n-k}$ up $\binom{n}{k}$ 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.

3.4 Simulations

- 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.

3.5 Random variables

- 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.

$$\begin{aligned}
E(X) = \mu_x &= \sum (x_i \times p_i) \\
&= x_1 \times p_1 + x_2 \times p_2 + \cdots + x_n \times p_n \\
Var(X) = \sigma_x^2 &= \sum (x_i - \mu_x)^2 \times p_i \\
&= (x_1 - \mu_x)^2 \times p_1 + (x_2 - \mu_x)^2 \times p_2 + \cdots + (x_n - \mu_x)^2 \times p_n \\
SD(X) = \sigma_x &= \sqrt{Var(X)}
\end{aligned}$$

We can think of p_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.

  $$E(X + Y) = E(X) + E(Y)$$
  $$E(aX + bY) = a \times E(X) + b \times E(Y)$$

  The expected value of the sum is the sum of the expected values.

  When X and Y are **independent**:

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

  The standard deviation of the sum is the square root of the sum of each standard deviation squared.

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

  The standard deviation of the difference is the square root of the sum of each standard deviation squared.[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.

3.6 Continuous distributions

- Histograms use bins with a specific width to display the distribution of a variable. When there is enough data and the data does not have gaps, as the bin width gets smaller and smaller, the histogram begins to resemble a smooth curve, or a **continuous distribution**.

- Continuous distributions are often used to approximate relative frequencies and probabilities. In a continuous distribution, the *area under the curve* corresponds to relative frequency or probability. The total area under a continuous probability distribution must equal 1.

- Because the area under the curve for a single point is zero, the probability of any specific value is zero. This implies that, for example, $P(X < 5) = P(X \leq 5)$ for a continuous probability distribution.

- Finding areas under curves is challenging; it is common to use distribution tables, calculators, or other technology to find such areas.

Chapter Highlights

This chapter focused on understanding likelihood and chance variation, first by solving individual probability questions and then by investigating probability distributions.

The main probability techniques covered in this chapter are as follows:

- The **General Multiplication Rule** for **and** probabilities (intersection), along with the special case when events are **independent**.

- The **General Addition Rule** for **or** probabilities (union), along with the special case when events are **mutually exclusive**.

- The **Conditional Probability Rule**.

- Tree diagrams and **Bayes' Theorem** to solve more complex conditional problems.

- The **Binomial Formula** for finding the probability of exactly k successes in n independent trials.

- **Simulations** and the use of random digits to estimate probabilities.

Fundamental to all of these problems is understanding when events are independent and when they are mutually exclusive. Two events are **independent** when the outcome of one does not affect the outcome of the other, i.e. $P(A|B) = P(A)$. Two events are **mutually exclusive** when they cannot both happen together, i.e. $P(A \text{ and } B) = 0$.

Moving from solving individual probability questions to studying probability distributions helps us better understand chance processes and quantify expected chance variation.

- For a **discrete probability distribution**, the **sum** of the probabilities must equal 1. For a **continuous probability distribution**, the **area under the curve** represents a probability and the total area under the curve must equal 1.

- As with any distribution, one can calculate the mean and standard deviation of a probability distribution. In the context of a probability distribution, the **mean** and **standard deviation** describe the average and the typical deviation from the average, respectively, after many, many repetitions of the chance process.

- A probability distribution can be summarized by its **center** (mean, median), **spread** (SD, IQR), and **shape** (right skewed, left skewed, approximately symmetric).

- 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.

- The mean of the sum of two random variables equals the sum of the means. However, this is not true for standard deviations. Instead, when finding the standard deviation of a sum or difference of random variables, take the square root of the sum of each of the standard deviations squared.

The study of probability is useful for measuring uncertainty and assessing risk. In addition, probability serves as the foundation for inference, providing a framework for evaluating when an outcome falls outside of the range of what would be expected by chance alone.

Chapter 4

Distribution of random variables

4.1 Normal distribution

- 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 **68-95-99.7 Rule**, otherwise known as the empirical 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.

4.2 Sampling distribution of a sample mean

- 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.

4.3 Geometric distribution

Previously in this chapter we looked at the normal distribution as a model for *numerical* data. In the next sections of this chapter we turn our attention to **categorical variables** with two levels.

- It is useful to model yes/no, success/failure with the values 1 and 0, respectively.[1] We call the **probability of success** p and the **probability of failure** $1 - p$.

- When the trials are **independent** and the value of p is constant, the probability of finding **the first success on the n^{th} trial** is given by $(1-p)^{n-1}p$. We can see the reasoning behind this formula as follows: for the first success to happen on the n^{th} trial, it has to *not* happen the first $n-1$ trials (with probability $1-p$), and then happen on the n^{th} trial (with probability p).

- When we consider the *entire distribution* of possible values for the how long *until* the first success, we get a discrete probability distribution known as the geometric distribution. The **geometric distribution** describes the waiting time *until* the first success, when the trials are independent and the probability of success, p, is constant.

- The geometric distribution is always *right skewed* and, in fact, has no maximum value. The probabilities, though, decrease exponentially fast.

- Even though the geometric distribution has an infinite number of values, it has a well-defined **mean**, $\mu = \frac{1}{p}$. If the probability of success is $\frac{1}{10}$, then *on average*, it takes 10 trials until we see the first success.[2]

- Note that when the trials are not independent, we can simply modify the geometric formula to find the probability that the first success happens on the n^{th} trial. Instead of simply raising $(1-p)$ to the $n-1$, multiply the appropriate *conditional* probabilities.

4.4 Binomial distribution

In the previous chapter, we introduced the binomial formula to find the probability of exactly k 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 or in a random sample of size n gives rise to a **binomial distribution**.

[1] Formally, we call such a variable a Bernoulli random variable.

[2] The geometric distribution also has a well defined standard deviation given by $\sigma = \sqrt{\frac{(1-p)}{p^2}}$.

- To write out a binomial probability **distribution table**, list all possible values for k, 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.

4.5 Sampling distribution of a sample proportion

The binomial distribution shows the distribution of the number of successes in n trials. Often, we are interested in the *proportion* of successes rather than the *number* of successes.

- To convert from "number of yeses" to "proportion of yeses" we simply divide the number by n. The sampling distribution of the sample proportion $\hat{p}$ is identical to the binomial distribution with a change of scale, i.e. different mean and different SD, but same shape.

- The same **success-failure condition** for the binomial distribution holds for a sample proportion $\hat{p}$.

- Three important facts about the sampling distribution of the sample proportion $\hat{p}$:
  - The mean of a sample proportion is denoted by $\mu_{\hat{p}}$, and it is equal to p. (*center*)
  - The SD of a sample proportion is denoted by $\sigma_{\hat{p}}$, and it is equal to $\sqrt{\frac{p(1-p)}{n}}$. (*spread*)

  - When $np \geq 10$ and $n(1-p) \geq 10$, the distribution of the sample proportion will be approximately normal. (*shape*)

- We use these properties when solving the following type of **normal approximation** problem involving a sample proportion. *Find the probability of getting more / less than x% yeses 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 $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ to standardize the sample proportion.
  3. Find the appropriate area under the normal curve.

Chapter Highlights

This chapter began by introducing the normal distribution. A common thread that ran through this chapter is the use of the **normal approximation** in various contexts. The key steps are included for each of the normal approximation scenarios below.

1. Normal approximation for **data**:
   - Verify that population is approximately normal.
   - Use the given mean μ and SD σ to find the Z-score for the given x value.

2. Normal approximation for a **sample mean/sum**:
   Verify that population is approximately normal or that $n \geq 30$.
   Use $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to find the Z-score for the given/calculated sample mean.

3. Normal approximation for the **number of successes** (binomial):
   - Verify that $np \geq 10$ and $n(1-p) \geq 10$.
   - Use $\mu_x = np$ and $\sigma_x = \sqrt{np(1-p)}$ to find the Z-score for the given number of successes.

4. Normal approximation for a **sample proportion**:
   - Verify that $np \geq 10$ and $n(1-p) \geq 10$.
   - Use $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ to find the Z-score for the given sample proportion.

5. Normal approximation for the **sum of two independent random variables**:
   - Verify that each random variable is approximately normal.
   - Use $E(X+Y) = E(X)+E(Y)$ and $SD(X+Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$ to find the Z-score for the given sum.

Cases 1 and 2 apply to **numerical** variables, while cases 3 and 4 are for **categorical** yes/no variables. Case 5 applies to both numerical and categorical variables.

Note that in the binomial case and in the case of proportions, we never look to see if the *population* is normal. That would not make sense because the "population" is simply a bunch of no/yes, 0/1 values and could not possibly be normal.

The **Central Limit Theorem** is the mathematical rule that ensures that when the sample size is sufficiently large, the sample mean/sum and sample proportion/count will be approximately normal.

Chapter 5

Foundations for inference

5.1 Estimating unknown parameters

Probability vs inference

- **Probability** involves using a known population value (parameter) to make a prediction about the likelihood of a particular sample value (statistic).
- **Inference** involves using a calculated sample value (statistic) to estimate or better understand an unknown population value (parameter).

Point estimates

- A sample statistic can serve as a **point estimate** for an unknown parameter. For example, the sample mean is a point estimate for an unknown population mean, and the sample proportion is a point estimate for an unknown population proportion.
- A point estimate is **unbiased** (accurate) if the sampling distribution (i.e., the distribution of all possible outcomes of the point estimate from repeated samples from the same population) is *centered* on the true population parameter. The sample mean and sample proportion are unbiased estimators for the population mean and population proportion, respectively.
- A point estimate has **lower variability** (more precise) when the *standard deviation* of the sampling distribution is smaller.
- In general, we want a point estimate to be unbiased and to have low variability. Remember: the terms unbiased (accurate) and low variability (precise) are properties of generic point estimates, which are variables that have a *sampling distribution*. These terms do not apply to individual values of a point estimate, which are *numbers*.

The standard error

- The **standard error** (**SE**) of a point estimate tells us the typical error or uncertainty associated with the point estimate. The standard error is the best estimate of the standard deviation when population parameters are unknown.
- In a random sample, increasing the sample size n will make the standard error smaller. This is consistent with the intuition that larger samples tend to be more reliable, all other things being equal.

5.2 Confidence intervals

Constructing **confidence intervals**

- A point estimate is not perfect; there is almost always some error in the estimate. It is often useful to supply a plausible *range of values* for the parameter, which we call a **confidence interval**.
- A confidence interval is centered on the point estimate and extends a certain number of standard errors on either side of the estimate, depending upon how *confident* one wants to be. For a fixed sample size, to be more confident of capturing the true value requires a wider interval.
- When the sampling distribution of a point estimate can reasonably be modeled as *normal*, such as with a **sample proportion**, then the following are true:
  - A 68% confidence interval is given by: point estimate $\pm$ SE of estimate.
    We can be 68% confident this interval captures the true value.
  - A 95% confidence interval is given by: point estimate $\pm$ $1.96 \times SE$ of estimate.
    We can be 95% confident this interval captures the true value.
  - A C% confidence interval is given by: point estimate $\pm$ $z^\star \times SE$ of estimate.
    We can be C% confident this interval captures the true value.
- For a C% confidence interval described above, we select $z^\star$ such that the area between -$z^\star$ and $z^\star$ under the standard normal curve is C%. Use the t-table at row ∞ to find the critical value $z^\star$.[1]
- After interpreting the interval, we can usually draw a conclusion, with C% confidence, about whether a given value X is a reasonable value for the true parameter. When drawing a conclusion based on a confidence interval, there are three possibilities.
  - We *have evidence* that the true [parameter]:
    ...is greater than X, because the entire interval is *above* X.
    ...is less than X, because the entire interval is *below* X.
  - We *do not have evidence* that the true [parameter] is not X, because X is *in* the interval.
- AP exam tip: A full confidence interval procedure includes the following steps.
  1. Choose and name the type of confidence interval.
  2. Check conditions to ensure that the point estimate is unbiased and follows the intended distribution and that the standard error estimate is reasonable.
  3. Calculate the interval: point estimate $\pm$ critical value $\times$ SE of estimate.
  4. Evaluate the CI and write the result in the form (____, ____).
  5. Interpret the interval: "We are C% confident that the true [parameter] is between ____ and ____." Describe the parameter in context.
  6. If applicable, draw a conclusion about whether a given value is a reasonable value for the parameter, based on whether that value falls inside or outside the interval.

[1] We explain the relationship between z and t in the next chapter.

Interpreting **confidence intervals** and **confidence levels**

- 68% and 95% are examples of **confidence levels**. A correct interpretation of a 95% confidence level is that if many samples of the same size were taken from the population, about 95% of the resulting confidence intervals would contain the true population parameter. Note that this is a *relative frequency interpretation*.

- We cannot use the language of probability to interpret an *individual* confidence interval, once it has been calculated. The confidence level tells us what percent of the intervals will contain the population parameter, not the probability that a calculated interval contains the population parameter. Each calculated interval either does or does not contain the population parameter.

Margin of error

- Confidence intervals are often reported as: point estimate $\pm$ margin of error. The **margin of error** (ME) = critical value $\times$ SE of estimate, and it tells us, with a particular confidence, how much we expect our point estimate to deviate from the true population value due to chance.

- The margin of error depends on the *confidence level*; the standard error does not. Other things being constant, a higher confidence level leads to a larger margin of error.

- For a fixed confidence level, increasing the sample size decreases the margin of error. This assumes a random sample.

- The margin of error formula only applies if a sample is random. Moreover, the margin of error measures only *sampling error*; it does not account for additional error introduced by response bias and non-response bias. Even with a perfectly random sample, the actual error in a poll is likely higher than the reported margin of error.[2]

5.3 Introducing hypothesis testing

Setting up, carrying out, and interpreting the results of a hypothesis test.

- A **hypothesis test** is a statistical technique used to evaluate competing claims based on data.

- The competing claims are called **hypotheses** and are always about population parameters (e.g. μ and p), never about sample statistics.

  - The **null hypothesis** is abbreviated H_0. It represents a skeptical perspective or a perspective of no difference or *no change*.
  - The **alternative hypothesis** is abbreviated H_A. It represents a new perspective or a perspective of a real difference or change. Because the alternative hypothesis is the stronger claim, it bears the burden of proof.

[2] nytimes.com/2016/10/06/upshot/when-you-hear-the-margin-of-error-is-plus-or-minus-3-percent-think-7-instead.html

- The **logic of a hypothesis test**: In a hypothesis test, we begin by *assuming that the null hypothesis is true.* Then, we calculate how unlikely it would be to get a sample value as extreme as we actually got in our sample, assuming that the null value is correct. If this likelihood is too small, it casts doubt on the null hypothesis and provides evidence for the alternative hypothesis.

- We set a **significance level**, denoted α, which represents the threshold below which we will reject the null hypothesis. The most common significance level is $\alpha = 0.05$. If we require more evidence to reject the null hypothesis, we use a smaller α.

- After verifying that the relevant **conditions are met**, we can calculate the test statistic. The **test statistic** tells us *how many* standard errors the point estimate (sample value) is from the null value (i.e. the value hypothesized for the parameter in the null hypothesis). When investigating a single mean or proportion or a difference of means or proportions, the test statistic is calculated as: $\frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$.

- After the test statistic, we calculate the p-value. We find and interpret the **p-value** according to the nature of the alternative hypothesis. The three possibilities are:

  H_A: $p > p_0$. The p-value corresponds to the area in the *upper tail* and is the probability of observing a sample value *as large as* our sample value, if H_0 were true.

  H_A: $p < p_0$. The p-value corresponds to the area in the *lower tail* and is the probability of observing a sample value *as small as* our sample value, if H_0 were true.

  H_A: $p \neq p_0$. The p-value corresponds to the area in *both tails* and is the probability of observing a sample value *as far from* the null value as our sample value, if H_0 were true.

- The conclusion or decision of a hypothesis test is based on whether the p-value is smaller or larger than the preset significance level α.
  - When the p-value $< \alpha$, we say the results are **statistically significant** at the α level and we have evidence of a real difference or change. The observed difference is beyond what would have been expected from chance variation alone. This leads us to reject H_0 and gives us evidence for H_A.
  - When the p-value $> \alpha$, we say the results are not statistically significant at the α level and we do not have evidence of a real difference or change. The observed difference was within the realm of expected chance variation. This leads us to not reject H_0 and does not give us evidence for H_A.

- AP exam tip: A full hypothesis test includes the following steps.
  1. Choose and name the test.
  2. Check conditions to ensure that the point estimate is unbiased and follows the intended distribution and that the standard error estimate is reasonable.
  3. Set up hypotheses.
  4. Identify significance level α.
  5. Calculate the test statistic, which usually has the form: $\frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$.
  6. Find the p-value and compare it to α to determine whether to reject or not reject H_0.
  7. Write the conclusion in the context of H_A.

Decision errors and **power**

- **Decision errors**. In a hypothesis test, there are two types of decision errors that could be made. These are called Type I and Type II errors.
  - A **Type I error** is rejecting H_0, when H_0 is actually true. We commit a Type I error if we call a result significant when there is *no* real difference or effect.
  - A **Type II error** is not rejecting H_0, when H_A is actually true. We commit a Type II error if we call a result not significant when there *is* a real difference or effect.

  Once a decision is made, only one of the two types of errors is possible. If the test rejects H_0, for example, only a Type I error is possible.

- The probability of a Type I and a Type II error.
  - The significance level, α, is the probability of a Type I error. Setting α defines the Type I error rate.
  - To calculate the probability of a Type II error, we must hypothesize a specific value for the parameter in H_A.
  - The probability of a Type I error (α) and a Type II error (β) are *inversely related*. Decreasing α makes β larger; increasing α makes β smaller.

- The power of a test.
  - When a particular H_A is true, the probability of not making a Type II error is called **power**. Power $= 1 - \beta$.
  - The power of a test is the probability of detecting an effect of a particular size when it is present.
  - Increasing the significance level decreases the probability of a Type II error and increases power. $\alpha \uparrow, \beta \downarrow$, power $\uparrow$.
  - For a fixed α, increasing the sample size n makes it easier to detect an effect and therefore decreases the probability of a Type II error and increases power. $n \uparrow, \beta \downarrow$, power $\uparrow$.

5.4 Does it make sense?

The inference procedures in this book require *two conditions* to be met.

- The first is that some type of **random sampling** or **random assignment** must be involved. If this is not the case, the point statistic may be biased and may not follow the intended distribution. Moreover, without a random sample or random assignment, there is no way to accurately measure the standard error. (When sampling without replacement, the sample size should be less than 10% of the population size in order for the standard error formula to apply. In sample surveys, this condition is generally met.)

- The second condition focuses on **sample size** and **skew** to determine whether the point estimate follows the intended distribution.

Understanding what the term **statistically significant** does and does not mean.

- *A small percent of the time (α), a significant result will not be a real result.* If many tests are run, a small percent of them will produce significant results due to chance alone.[3]

- *With a very large sample, a significant result may point to a result that is real but unimportant.* With a larger sample, the power of a test increases and it becomes easier to detect a small difference. If an extremely large sample is used, the result may be statistically significant, but not be *practically significant*. That is, the difference detected may be so small as to be unimportant or meaningless.

Chapter Highlights

Statistical inference is the practice of making decisions from data in the context of uncertainty. In this chapter, we introduced two frameworks for inference: **confidence intervals** and **hypothesis tests**.

- Confidence intervals are used for *estimating* unknown population parameters by providing an *interval of reasonable values* for the unknown parameter with a certain level of confidence.

- Hypothesis tests are used to assess how reasonable a *particular* value is for an unknown population parameter by providing *degrees of evidence* against that value.

- The results of confidence intervals and hypothesis tests are, generally speaking, *consistent*.[4] That is:
  - Values that fall *inside* a 95% confidence interval (implying they are reasonable) will *not be rejected* by a test at the 5% significance level (implying they are reasonable), and vice-versa.
  - Values that fall *outside* a 95% confidence interval (implying they are not reasonable) will *be rejected* by a test at the 5% significance level (implying they are not reasonable), and vice-versa.
  - When the confidence level and the significance level add up to 100%, the conclusions of the two procedures are consistent.

- Many values fall inside of a confidence interval and will not be rejected by a hypothesis test. "Not rejecting H_0" is NOT equivalent to *accepting* H_0. When we "do not reject H_0", we are asserting that the null value is *reasonable*, not that the parameter is exactly *equal to* the null value.

- For a 95% confidence interval, 95% is not the probability that the true value lies inside the confidence interval (it either does or it doesn't). Likewise, for a hypothesis test, α is not the probability that H_0 is true (it either is or it isn't). In both frameworks, the probability is about what would happen in a random sample, not about what is true of the population.

[3]Similarly, if many confidence intervals are constructed, a small percent (100 - C%) of them will fail to capture a true value due to chance alone. A value outside the confidence interval is not an *impossible* value.

[4]In the context of proportions there will be a small number of cases where this is not true. This is because when working with proportions, the SE used for confidence intervals and the SE used for tests are slightly different, as we will see in the next chapter.

- The confidence interval procedures and hypothesis tests described in this book should not be applied unless particular conditions (described in more detail in the following chapters) are met. If these procedures are applied when the conditions are not met, the results may be unreliable and misleading.

While a given data set may not always lead us to a correct conclusion, statistical inference gives us tools to *control and evaluate how often these errors occur.*

Chapter 6

Inference for categorical data

6.1 Inference for a single proportion

Most of the confidence interval procedures and hypothesis tests of this book involve: a **point estimate**, the **standard error** of the point estimate, and an assumption about the **shape of the sampling distribution** of the point estimate. In this section, we look at inference when the parameter of interest is a *proportion*.

- We use the sample proportion $\hat{p}$ as the *point estimate* for the unknown population proportion p. The sampling distribution of $\hat{p}$ is approximately normal when the success-failure condition is met and the observations are independent. The observations can generally be considered independent when the data is collected from a random sample or come from a stable, random process analogous to flipping a coin. When the sampling distribution of $\hat{p}$ is normal, the standardized test statistic also follows a **normal** distribution.

- When verifying the success-failure condition and calculating the SE,
  - use the *null/hypothesized* proportion p_0 for the hypothesis test, but
  - use the *sample* proportion $\hat{p}$ for the confidence interval.

- **Confidence intervals** and **hypothesis tests** for a single **proportion** p. When there is one sample and the parameter of interest is a single proportion, e.g. the true proportion of U.S. residents that are over age 65:

  Choose:

  - **1-proportion z-interval** to estimate p with C% confidence, or
  - **1-proportion z-test** to test H_0: $p = p_0$ at a specified α.

  Check:

  1. Data come from a random sample.
  2. For a CI: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ (Make sure to plug in numbers
     For a Test: $np_0 \geq 10$ and $n(1-p_0) \geq 10$. for n and $\hat{p}$, or for n and p_0!)

Calculate:

– CI: point estimate $\pm$ $z^{\star} \times SE$ of estimate, or

– test statistic: $Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$, then find p-value

point estimate: sample proportion $\hat{p}$

SE of estimate: for a CI, use $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$; for a Test, use $\sqrt{\frac{p_0(1-p_0)}{n}}$

(`1-PropZInt` or `1-PropZTest` on a TI)

Conclude:

– For a CI: We are C% confident that the true *proportion* of [...] is between ___ and ___. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value of interest.

– For a hypothesis test:

- p-value $< \alpha$, reject H_0; we have evidence that [H_A in context], or
- p-value $> \alpha$, do not reject H_0; we do not have evidence that [H_A in context].

- The **margin of error** (ME) for one-sample confidence interval for a proportion is $z^{\star}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

- To find the **minimum sample size** needed to estimate a proportion with a given confidence level and a given margin of error, m, set up an inequality of the form: $z^{\star}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < m$. Unless a particular proportion is given in the problem, use $\hat{p} = 0.5$. Make sure to round the final answer up to an *integer*, since this answer refers to a number of people or things.

6.2 Difference of two proportions

In the previous section, we estimated and hypothesized about a single proportion. In this section, we *compare* two groups to each other with respect to a proportion or a percent.

- We are interested in whether the true proportion of yeses is the same or different between two distinct groups. Call these proportions p_1 and p_2. One of three things must be true.

  – The proportions are the *same*, i.e. $p_1 = p_2$, or equivalently, $p_1 - p_2 = 0$.

  – The proportion in group 1 is *greater than* the proportion in group 2, i.e. $p_1 > p_2$, or equivalently, $p_1 - p_2 > 0$.

  – The proportion in group 1 is *less than* the proportion in group 2, i.e. $p_1 < p_2$, or equivalently, $p_1 - p_2 < 0$.

- The sign of the difference of the proportions tells us whether the proportion of yeses in group 1 is equal to, greater than, or less than the proportion of yeses in group 2. Therefore, when *comparing* two proportions to each other, the parameter of interest is the *difference of proportions*, $p_1 - p_2$, and we use the difference of sample proportions, $\hat{p}_1 - \hat{p}_2$, as the *point estimate*.

- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal when the success-failure condition is met for *both* groups and when the data is collected using 2 independent random samples or 2 randomly assigned treatments. When the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is normal, the standardized test statistic also follows a **normal** distribution.

- When the null hypothesis is that the two populations proportions are *equal* to each other, use the **pooled sample proportion** $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$, i.e. the combined number of yeses over the combined sample sizes, when finding the SE. For the confidence interval, do not use the pooled sample proportion; use the separate values of $\hat{p}_1$ and $\hat{p}_2$.

- **Confidence intervals** and **hypothesis tests** for a **difference of proportions** $p_1 - p_2$. When there are two samples or treatments and the parameter of interest is a difference of proportions, e.g. the true difference in proportion of 17 and 18 year olds with a summer job (proportion of 18 year olds − proportion of 17 year olds):

  Choose:

  - **2-proportion z-interval** to estimate $p_1 - p_2$ with C% confidence, or
  - **2-proportion z-test** to test H_0: $p_1 - p_2 = 0$ (i.e. $p_1 = p_2$) with a specified α.

  Check:

  1. Data come from 2 independent random samples or 2 randomly assigned treatments.
  2. $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$ (Make sure to plug in numbers!)

  Calculate:

  - CI: point estimate $\pm$ $z^\star \times SE$ of estimate, or
  - test statistic: $Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$, then find p-value

    point estimate: difference of sample proportions $\hat{p}_1 - \hat{p}_2$
    SE of estimate:
    for a CI, use $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$; for a Test, use $\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

  (`2-PropZInt` or `2-PropZTest` on a TI)

  Conclude:

  - For a CI: We are C% confident that the true *difference in the proportion* of [...] is between ___ and ___. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value 0.
  - For a hypothesis test:
    - p-value $< \alpha$, reject H_0; we have evidence that [H_A in context], or
    - p-value $> \alpha$, do not reject H_0; we do not have evidence that [H_A in context].

6.3 Testing for goodness of fit using chi-square

The inferential procedures we have seen thus far are based on the test statistic following a *normal distribution*. In this section, we introduce a new distribution called the chi-square distribution.

- While a normal distribution is defined by its mean and standard deviation, the chi-square distribution is defined by just one parameter called **degrees of freedom**.
- For a chi-square distribution, as the degrees of freedom increase:
  - the center increases.
  - the spread increases.
  - the shape becomes more symmetric and more normal.[1]
- When we want to see if a model is a good fit for observed data or if data is representative of a particular population, we can use a χ^2 **goodness of fit test.** This test is used when there is one variable with multiple categories (bins) that can be arranged in a **one-way table**.
- In a χ^2 goodness of fit test, we calculate a χ^2**-statistic**, which is a measure of how far the observed values in the sample are from the expected values under the null hypothesis. $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$
  - Always use whole numbers (counts) for the observed values, not proportions or percents.
  - For each category, the expected counts can be found by multiplying the sample size by the expected proportion under the null hypothesis. Expected counts do *not* need to be integers.
  - For each category, find $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$, then add them all together to get the χ^2-statistic.
- When there is a random sample and all of the expected counts are at least 5, the χ^2-statistic follows a **chi-square distribution** with degrees of freedom equal to number of categories $-$ 1.
- For a χ^2 test, the p-value corresponds to the probability that observed sample values would differ from the expected values by *more than* what we observed in this sample. The p-value, therefore, corresponds to the area *to the right* of the calculated χ^2-statistic (the area in the upper tail).
- A larger χ^2 represents greater deviation between the observed values and the expected values under the null hypothesis. For a fixed degrees of freedom, a larger χ^2 value leads to a smaller p-value, providing greater evidence against H_0.

[1] Technically, however, it is always right skewed.

- χ^2 **tests for a one-way table**. When there is one sample and we are comparing the distribution of a categorical variable to a specified or population distribution, e.g. using sample values to determine if a machine is producing M&Ms with the specified distribution of color:

  Test the hypotheses at a specified α. The hypotheses can often be set up as:

  H_0: The distribution of [...] matches the specified or population distribution.

  H_A: The distribution of [...] doesn't match the specified or population distribution.

  Choose: χ^2 **goodness of fit test**

  Check:

  1. Data come from a random sample.
  2. All expected counts ≥ 5. (Make sure to calculate expected counts!)

  Calculate:

  test statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ (χ^2`GOF-Test` on TI-84)
  df = # of categories $-$ 1
  p-value = (area to the *right* of test statistic)

  Conclude:

  – p-value $< \alpha$, reject H_0; we have evidence that [H_A in context], or
  – p-value $> \alpha$, do not reject H_0; we do not have evidence that [H_A in context].

6.4 Homogeneity and independence in two-way tables

- When there are two categorical variables, rather than one, the data must be arranged in a **two-way table** and a χ^2 test of homogeneity or a χ^2 test of independence is appropriate.
- These tests use the same χ^2-statistic as the χ^2 goodness of fit test, but instead of number of categories $-$ 1, the **degrees of freedom** is (# of rows$-$1)$\times$(# of columns$-$1). All expected counts must be at least 5.
- When working with a two-way table, the **expected count** for each row,column combination is calculated as: expected count $= \frac{(\text{row total})\times(\text{column total})}{\text{table total}}$.
- The χ^2 test of homogeneity and the χ^2 test of independence are almost identical. The differences lie in the data collection method and in the hypotheses.

- χ^2 **tests for two-way tables**:

- When there are multiple samples or treatments and we are comparing the distribution of a categorical variable across several groups, e.g. comparing the distribution of rural/urban/suburban dwellers among 4 states:

  Test the hypotheses at a specified α. The hypotheses can often be set up as:

  H_0: The distribution of [...] is the same for each population/treatment.

  H_A: The distribution of [...] is not the same for each population/treatment.

  Choose: $\boldsymbol{\chi^2}$ **test of homogeneity**.

- When there is one sample and we are looking for association or dependence between two categorical variables, e.g. seeing if there is an association between gender and political party:

  Test the hypotheses at a specified α. The hypotheses can often be set up as:
  H_0: [variable 1] and [variable 2] are independent.
  H_A: [variable 1] and [variable 2] are dependent.

  Choose: χ^2 **test of independence**.

- For both χ^2 tests:

  Check:
  1. Data come from a random sample.
  2. All expected counts ≥ 5. (Calculate using `2ND MATRIX, EDIT, 2:[B]` *after* doing `X²-Test` on a TI)

  Calculate:
  test statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ (`X²-Test` on TI)
  df = (# of rows − 1)(# of cols − 1)
  p-value = (area to the *right* of test statistic)

  Conclude:
  - p-value $< \alpha$, reject H_0; we have evidence that [H_A in context], or
  - p-value $> \alpha$, do not reject H_0; we do not have evidence that [H_A in context].

Chapter Highlights

Calculating a confidence interval or a test statistic and p-value are generally done with statistical software. It is important, then, to focus not on the calculations, but rather on

1. choosing the correct procedure
2. understanding when the procedures do or do not apply, and
3. interpreting the results.

Choosing the correct procedure requires understanding the *type of data* and the *method of data collection*. All of the inference procedures in Chapter 6 are for categorical variables. Here we list the five tests encountered in this chapter and when to use them.

- **1-proportion z-test**
  - *1 random sample*, a yes/no variable
  - Compare the sample proportion to a fixed / hypothesized proportion.
- **2-proportion z-test**
  - *2 independent random samples or randomly allocated treatments*
  - Compare two populations or treatments to each other with respect to one yes/no variable; e.g. comparing the proportion over age 65 in two distinct populations.
- χ^2 **goodness of fit test**
  - *1 random sample*, a categorical variable (generally at least three categories)
  - Compare the distribution of a categorical variable to a fixed or known population distribution; e.g. looking at distribution of color among M&Ms.
- χ^2 **test of homogeneity**:
  - *2 or more independent random samples or randomly allocated treatments*
  - Compare the distribution of a categorical variable across several populations or treatments; e.g. party affiliation over various years, or patient improvement compared over 3 treatments.
- χ^2 **test of independence**
  - *1 random sample*, 2 categorical variables
  - Determine if, in a single population, there is an association between two categorical variables; e.g. grade level and favorite class.

A careful reader may have noticed that some of the above descriptions seem to overlap. In fact, there are times when both a z-test and a χ^2 test could be used. In these cases, the results are equivalent, meaning that the two tests produce the same p-value.

- A two-sided 1-proportion z-test is equivalent to a χ^2 goodness of fit test with $df = 1$ (2 categories).
- A two-sided 2-proportion z-test is equivalent to a χ^2 test of homogeneity with $df = 1$ (a 2×2 table).[2]

Even when the data and data collection method correspond to a particular test, we must *verify that conditions are met* to see if the assumptions of the test are reasonable. All of the inferential procedures of this chapter require some type of random sample or process. In addition, the 1-proportion z-test/interval and the 2-proportion z-test/interval require that the success-failure condition is met and the three χ^2 tests require that all expected counts are at least 5.

Finally, understanding and communicating the logic of a test and being able to accurately *interpret* a confidence interval or p-value are essential. For a refresher on this, review Chapter 5: Foundations for inference.

[2]Sometimes the success-failure condition is weakened to require the number of successes and failures to be at least 5, making it consistent with the chi-square condition of all all expected counts at least 5.

Chapter 7

Inference for numerical data

7.1 Inference for a single mean with the t-distribution

In Chapter 7 we turn our attention to inference for **numerical variables**. We begin by investigating **confidence intervals** and **hypothesis tests** for a *single mean*, and in doing so, we encounter a new distribution, the t-distribution.

- The sample mean $\bar{x}$ is a *point estimate* for the population mean μ.
- Introducing the **t–distribution**
  - Previously, we saw that $SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. However, when doing inference on an unknown population mean, the population standard deviation is usually unknown as well. In this case we use the sample standard deviation s to estimate σ and we standardize $\bar{x}$ using $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$.
  - Because s introduces extra variability, the standardized test statistic using s in place of σ no longer has a normal distribution. Instead, it has a t-distribution with $n - 1$ degrees of freedom.
  - As the sample size and degrees of freedom increase, s becomes a more stable estimate of σ, and the corresponding t-distribution has smaller spread.
  - As the degrees of freedom go to ∞, the t-distribution approaches the normal distribution. This is why we use the t-table at $df = \infty$ to find the value of $z^\star$.
- If the population is normal or if the sample size is sufficiently large, the sampling distribution of $\bar{x}$ will be *normal* and the standardized test statistic using $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$ will follow a **t-distribution** with $n - 1$ degrees of freedom.
- If $n < 30$ and we are unsure whether the population is approximately normal, we look at the distribution of sample data for evidence that the population is not approximately normal.
  - If the sample data has strong skew or outliers, this gives us evidence to believe that the population is not approximately normal.
  - Just as data cannot prove H_0 to be true, the sample data cannot prove that that population is approximately normal; however, if the sample data does not have

strong skew or outliers, we take the assumption that the population distribution is approximately normal to be *reasonable*.

- To find the critical value $t^\star$ for a confidence interval, use the t-table with row corresponding to the degrees of freedom. Here $df = n - 1$.

- **Confidence intervals** and **hypothesis tests** for a single **mean** μ. When there is one sample and the parameter of interest is a single mean, e.g. the true mean (or average) miles per gallon of a 2018 Honda Civic:

  Choose:

  - **1-sample t-interval** to estimate μ with C% confidence, or
  - **1-sample t-test** to test H_0: $\mu = \mu_0$ at a specified α.

  Check:

  1. Data come from a random sample.
  2. $n \geq 30$, OR population is known to be approximately normal, OR population could be approximately normal because data shows no strong skew or outliers. (If arguing the third case, make sure to also sketch a graph of the data!)

  Calculate:

  - CI: point estimate $\pm$ $t^\star \times SE$ of estimate, or
  - test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$, then find p-value

    point estimate: sample mean $\bar{x}$

    SE of estimate: $\frac{s}{\sqrt{n}}$

    $df = n - 1$

  (`TInterval` or `T-Test` on a TI)

  Conclude:

  - For a CI: We are C% confident that the true *mean* of [...] is between ____ and ____. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value of interest.
  - For a hypothesis test:
    - p-value $< \alpha$, reject H_0; we have evidence that [H_A in context], or
    - p-value $> \alpha$, do not reject H_0; we do not have evidence that [H_A in context].

7.2 Inference for paired data

In this section we investigate **confidence intervals** and **hypothesis tests** for **paired, numerical data**.

- Paired data can come from a **random sample** or a **matched pairs experiment**. With paired data, the *difference* tells us something important. For example, if we take a random sample of U.S. counties and record population in 2017 and population in 2018 for those counties, the difference tells us whether population increased, decreased, or stayed the same in each county. Similarly, the difference of paired data from an experiment tells us whether one treatment did better, worse, or the same as the other treatment for each subject.

- We use the notation $\bar{x}_{diff}$ to represent the mean of the sample differences. Likewise, s_{diff} is the standard deviation of the sample differences, and n_{diff} is the number of sample differences.

- We use $\bar{x}_{diff}$ as the *point estimate* for μ_{diff}. When the population of differences is normal or when the number of sample differences is sufficiently large, the sampling distribution of $\bar{x}_{diff}$ will be *normal* and the standardized test statistic using $SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}}$ will follow a **t–distribution** with $n_{diff} - 1$ degrees of freedom.

- To carry out inference on paired data, we first find all of the sample differences. Then, we perform a one-sample procedure on the *differences*. For this reason, the confidence interval and hypothesis test for paired data use the same one-sample procedures from the previous section.

- **Confidence intervals** and **hypothesis tests** for a **mean of differences** μ_{diff}. When there is paired data and the parameter of interest is a mean of differences, e.g. the true mean of the differences in county population (year 2018 − year 2017):

  Choose:

  - **matched pairs t-interval** to estimate μ_{diff} with C% confidence, or
  - **matched pairs t-test** to test H_0: $\mu_{diff} = 0$ at a specified α.

  Check:

  1. There is paired data from a random sample or matched pairs experiment.
  2. $n_{diff} \geq 30$, OR population of *differences* is known to be approximately normal, OR population of *differences* could be approximately normal because sample differences show no strong skew or outliers. (If arguing the third case, make sure to also graph the differences!)

  Calculate:

  - CI: point estimate $\pm$ $t^{\star} \times SE$ of estimate, or
  - test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$, then find p-value

    point estimate: sample proportion $\bar{x}_{diff}$
    SE of estimate: $\frac{s_{diff}}{\sqrt{n_{diff}}}$
    $df = n_{diff} - 1$

  (`TInterval` or `T-Test` on a TI)

  Conclude:

  - For a CI: We are C% confident that the true *mean of the differences in* [...] is between ___ and ___. If applicable, draw a conclusion based on whether the interval is entirely positive, is entirely negative, or contains the value 0.
  - For a hypothesis test:
    - p-value $< \alpha$, reject H_0; we have evidence that [H_A in context], or
    - p-value $> \alpha$, do not reject H_0; we do not have evidence that [H_A in context].

7.3 Difference of two means using the t-distribution

In the last section we carried out a confidence interval and hypothesis test for the *mean of differences*. In this section we consider a confidence interval and hypothesis test for the *difference of means*.

- **Mean of differences** vs **difference of means**. To calculate a mean of difference, $\bar{x}_{diff}$, we first calculate all of the differences, then we find the mean of those differences. To calculate a difference of means, $\bar{x}_1 - \bar{x}_2$, we first calculate the mean of each group, then we take the difference between those two numbers.

- The difference of sample means, $\bar{x}_1 - \bar{x}_2$, is a *point estimate* for the difference of population means, $\mu_1 - \mu_2$.

- When both populations are normal or both sample sizes are sufficiently large, $\bar{x}_1 - \bar{x}_2$ follows a *normal* distribution, and the standardized test statistic using the sample standard deviations follows a **t-distribution**, where the degrees of freedom is given by a complicated formula involving the sample sizes and standard deviations. In practice, one finds the *df* for this test using computer software or a calculator.[1]

- **Confidence intervals** and **hypothesis tests** for a **difference of means** $\mu_1 - \mu_2$. When there are two samples or treatments and the parameter of interest is a difference of means, e.g. the true difference in mean cholesterol reduction (mean treatment A − mean treatment B):

  Choose:

  - **2-sample t-interval** to estimate $\mu_1 - \mu_2$ with C% confidence, or
  - **2-sample t-test** to test H_0: $\mu_1 - \mu_2 = 0$ (i.e. $\mu_1 = \mu_2$) at a specified α.

  Check:

  1. Data come from 2 independent random samples or 2 randomly assigned treatments.
  2. $n_1 \geq 30$ and $n_2 \geq 30$, OR both populations are known to be approximately normal, OR both populations could be normal because both data sets show no strong skew or outliers. (If arguing the third case, make sure to also graph *both* data sets!)[2]

  Calculate:

  - CI: point estimate $\pm$ $t^\star \times SE$ of estimate, or
  - test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$, then find p-value

    point estimate: difference of sample means $\bar{x}_1 - \bar{x}_2$
    SE of estimate: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
    Find and record the *df* using a calculator or other software.

  (`2-SampTInt` or `2-SampTTest` on a TI)

[1] If this is not possible, one can use $df = min(n_1 - 1, n_2 - 1)$.
[2] Yes, a mix-and-match of conditions works!

Conclude:

- For a CI: We are C% confident that the true *difference in mean* [...] is between ___ and ___. If applicable, draw a conclusion based on whether the interval is entirely above, is entirely below, or contains the value 0.
- For a hypothesis test:
  - p-value $< \alpha$, reject H_0; we have evidence that [H_A in context], or
  - p-value $> \alpha$, do not reject H_0; we do not have evidence that [H_A in context].

Chapter Highlights

The following confidence intervals and tests are structurally the same in that they all involve inference on a population parameter, where that parameter is a proportion, a difference of proportions, a mean, a mean of differences, or a difference of means.

- 1-proportion z-test/interval
- 2-proportion z-test/interval
- 1-sample t-test/interval
- matched pairs t-test/interval
- 2-sample t-test/interval

The above inferential procedures all involve a **point estimate**, a **standard error** of the estimate, and an assumption about the **shape of the sampling distribution** of the point estimate and of the standardized test statistic.

From Chapter 6, the χ^2 tests and their uses are as follows:

- χ^2 goodness of fit - compares a categorical variable to a known/fixed distribution.
- χ^2 test of homogeneity - compares a categorical variable across multiple groups.
- χ^2 test of independence - looks for association between two categorical variables.

χ^2 is a measure of *overall* deviation between observed values and expected values. These tests stand apart from the others because when using χ^2 there is not a parameter of interest. For this reason there are no confidence intervals using χ^2. Also, for χ^2 tests, the hypotheses are usually written in words, because they are about *distributions* of a categorical variable, not about a single parameter.

While formulas and conditions vary, all of these procedures follow the same basic logic and process.

- For a confidence interval, identify the confidence level and the parameter to be estimated. For a hypothesis test, identify the significance level and set up the hypotheses using population parameters, where applicable.
- Choose the correct procedure.
- Check that both conditions for its use are met.
- Calculate the confidence interval or the test statistic and p-value, as well as the *df* if applicable.
- Interpret the results and draw a conclusion based on the data.

Chapter 8

Introduction to linear regression

8.1 Line fitting, residuals, and correlation

- In Chapter 2 we introduced **scatterplots**, which show the relationship between two numerical variables. When we use x to predict y, we usually call x the **explanatory** or **predictor variable**, and we call y the **response variable**.

- A linear model can be useful for prediction when the variables have a constant, linear trend. Linear models should not be used if the trend between the variables is curved.

- y refers to an actual data value. When we write a linear model, we use $\hat{y}$ to indicate that it is the model or the prediction. $\hat{y}$ can be understood as a **prediction** for y based on a given x, or as an **average** of the y values for a given x.

- The **residual** is the **error** between the true value and the model, $y - \hat{y}$, for a given x-value. The order of the difference matters, and the sign of the residual will tell us if the model over-predicted or under-predicted a particular data point.

- The symbol s in a linear model is used to denote the standard deviation of the residuals, and it measures the typical prediction error by the model.

- A **residual plot** is a scatterplot with the residuals on the vertical axis. The residuals are often plotted against x on the horizontal axis, but they can also be plotted against y, $\hat{y}$, or other variables. Two important uses of a residual plot are the following.
  - Residual plots help us see patterns in the data that may not have been apparent in the scatterplot.
  - The standard deviation of the residuals is easier to estimate from a residual plot than from the original scatterplot.

- The **correlation coefficient**, r, measures the strength and direction of a linear relationship. The following are some important properties of r.
  - The value of r is always between -1 and 1, inclusive, with an $r = -1$ indicating a perfect negative relationship (points fall exactly along a line that has negative

slope) and an $r = 1$ indicating a perfect positive relationship (points fall exactly along a line that has positive slope).

- An $r = 0$ indicates no *linear* association between the variables, though there may well exist a quadratic or other type of association.
- The calculation of r involves converting the x and y values into Z-scores. Because of this, changing the units of x and y does not affect the value of r. Another way to say this is that adding a constant, subtracting a constant, or multiplying a positive constant to all values of x or y does not affect the correlation.
- Multiplying all values of x or y by a *negative* constant will flip the graph over the x or y axis and therefore change the *sign* of r.

8.2 Fitting a line by least squares regression

- We define the *best fit line* as the line that minimizes the sum of the squared residuals (errors) about the line. That is, we find the line that minimizes $(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 = \sum (y_i - \hat{y}_i)^2$. We call this line the **least squares regression line**.
- We write the least squares regression line in the form: $\hat{y} = b_0 + b_1 x$, and we can calculate b_0 and b_1 based on the summary statistics as follows:

$$b_1 = r\frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1\bar{x}.$$

- b_0 and b_1 are point estimates for the parameters β_0 and β_1 in the true model, given by: $y = \beta_0 + \beta_1 x + \mathit{error}$.
- *Interpreting* the **y-intercept** and **slope** of a linear model
  - The slope, b_1, describes the *average* increase or decrease in the y variable if the explanatory variable x is one unit larger.
  - The y-intercept, b_0, describes the average or predicted outcome of y if $x = 0$. The linear model must be valid all the way to $x = 0$ for this to make sense, which in many applications is not the case.
- Two important considerations about the regression line
  - The regression line provides *estimates* or *predictions*, not actual values. It is important to how large s, the standard deviation of the residuals, is, in order to know about how much error to expect in these predictions.
  - The regression line estimates are only reasonable within the domain of the data. Predicting y for x values that are outside the domain, known as **extrapolation**, is unreliable and may produce ridiculous results.
- Using R^2 to assess the fit of the model
  - R^2, called **R-squared** or the **explained variance**, is a measure of how well the model explains or fits the data.
  - The R^2 of a linear model describes the *proportion of variation* in the y variable that is *explained by* the regression line.

  - R^2 is always between 0 and 1, inclusive, or between 0% and 100%, inclusive. The higher the value of R^2, the better the model "fits" the data.
  - R^2 applies to any type of model, not just a linear model, and can be used to compare the fit among various models.
  - The correlation coefficient $r = \pm\sqrt{R^2}$. The value of R^2 is always positive and cannot tell us the *direction* of the association. If finding r based on R^2, make sure to use either the scatterplot or the slope of the regression line to determine the *sign* of r.

- When a residual plot of the data appears as a random cloud of points, a linear model is generally appropriate. If a residual plot of the data has any type of pattern, such as a ∪-shape, a linear model is not appropriate.

8.3 Types of outliers in linear regression

- **Outliers** in regression are observations that fall far from the "cloud" of points.
- An **infuential point** is a point that has a big effect or pull on the slope of the regression line.
- Points that are outliers in the x direction will have more pull on the slope of the regression line and are more likely to be influential points.
- Outliers in the data should not be removed or ignored without a good reason; often these points communicate important aspects about the association.

8.4 Inference for the slope of a regression line

In Chapter 6, we used a χ^2 test of independence to test for association between two categorical variables. In this section, we test for association/correlation between two numerical variables.

- We use the sample slope b_1 as a *point estimate* for the population slope β_1. The population slope is the true increase/decrease in y for each unit increase in x. If the population slope is 0, there is no linear relationship between the two variables.
- Under certain assumptions, the sampling distribution of b_1 is *normal* and the distribution of the standardized test statistic using the sample standard deviation of the residuals follows a **t-distribution** with $n - 2$ degrees of freedom.
- **Confidence intervals** and **hypothesis tests** for the **slope of a regression line** β_1. When there is (x, y) data and the parameter of interest is the true slope of a regression line, e.g. the true slope of the regression line relating air quality index to average rainfall per year for each city in the United States:

  Choose:

  - **Linear regression t-interval** to estimate β_1 with C% confidence, or
  - **Linear regression t-test** to test H_0: $\beta_1 = 0$ at a specified α.

  Check:

1. Data come from a random sample or randomized experiment.
2. The residual plot shows no pattern.[1]

Calculate:

- CI: point estimate $\pm$ $t^{\star} \times SE$ of estimate, or
- test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$, then find p-value

 point estimate: the sample slope b_1

 SE of estimate: SE of slope (find using computer output)

 $df = n - 2$

 (`LinRegTInt` or `LinRegTTest` on a TI)

Conclude:

- For a CI: We are C% confident that the true *slope* of the regression line relating [y] to [x] is between ___ and ___. If applicable, draw a conclusion based on whether the interval is entirely positive, is entirely negative, or contains the value 0.
- For a hypothesis test:
  - p-value $< \alpha$, reject H_0; we have evidence that [H_A in context], or
  - p-value $> \alpha$, do not reject H_0; we do not have evidence that [H_A in context].

- The linear regression t-test and the matched pairs t-test both involve *paired*, numerical data. However, the linear regression t-test asks if the two variables have a linear *relationship*, specifically if the *slope* of the population regression line is 0. The matched pairs t-test, on the other hand, asks if the two variables are in some way the *same*, specifically if the *mean* of the population differences is 0.

8.5 Transformations for nonlinear data

- Regression analysis is easier to perform on linear data. When data are nonlinear, we sometimes **transform** the data in a way that results in a linear relationship. The most common transformation is *log* (or *ln*) of the y values. Sometimes we also apply a transformation to the x values.

- To assess the model, we look at the **residual plot** of the *transformed* data. If the residual plot of the original data has a pattern, but the residual plot of the transformed data has no pattern, a linear model for the transformed data is reasonable, and the transformed model provides a better fit than the simple linear model.

Chapter Highlights

This chapter focused on describing the linear association between two numerical variables and fitting a linear model.

- The **correlation coefficient**, r, measures the strength and direction of the linear association between two variables. However, r alone cannot tell us whether data follow a linear trend or whether a linear model is appropriate.

[1] Technically, the residuals should be independent, nearly normal, and have constant standard deviation.

- The **explained variance**, R^2, measures the proportion of variation in the y values explained by a given model. Like r, R^2 alone cannot tell us whether data follow a linear trend or whether a linear model is appropriate.

- Every analysis should begin with *graphing* the data using a **scatterplot** in order to see the association and any deviations from the trend (outliers or influential values). A **residual plot** helps us better see patterns in the data.

- When the data show a linear trend, we fit a **least squares regression line** of the form: $\hat{y} = b_0 + b_1 x$, where b_0 is the y-intercept and b_1 is the slope. It is important to be able to *calculate* b_0 and b_1 using the summary statistics and to *interpret* them in the context of the data.

- A **residual**, $y - \hat{y}$, measures the error for an *individual point*. The **standard deviation of the residuals**, s, measures the typical size of the residuals.

- $\hat{y} = b_0 + b_1 x$ provides the best fit line for the *observed data*. To estimate or hypothesize about the true slope, first confirm that the residual plot has no pattern and that a linear model is reasonable, then use a **linear regression t-interval** or a **linear regression t-test** with $n - 2$ degrees of freedom.

In this chapter we focused on simple linear models with one explanatory variable. More complex methods of prediction, such as multiple regression (more than one explanatory variable) and nonlinear regression can be studied in a future course.

Final words

The main topics we have covered in this introduction to statistics are:

- Methods of data collection, with an emphasis on understanding and minimizing bias.
- Summarizing univariate and bivariate data graphically, numerically, and verbally.
- Probability and probability models.
- Sampling distributions and inferential procedures to better understand randomness and make conclusions based on data.

We have only scratched the surface of each of these topics; however, we hope that this introduction has generated curiosity and excitement for future study of statistics.