Section Summaries and Chapter Highlights

from

Advanced High School Statistics

Third Edition

David Diez
Data Scientist
OpenIntro

Mine Çetinkaya-Rundel
Associate Professor of the Practice, Duke University
Professional Educator, RStudio

Leah Dorazio
Statistics and Computer Science Teacher
San Francisco University High School

Christopher D Barr
Investment Analyst
Varadero Capital

May 25, 2022

Updated: May 25, 2022.

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**, which likely introduce bias. Instead, use a **random** sampling method.

- Generalizations from a sample can be made to a population only if the sample is random. Furthermore, the generalization can be made only to the population from which the sample was randomly selected, not to a larger or different population.

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

- In a **simple random sample**, every *individual* as well as every *group of individuals* has the same probability of being in the sample. A common way to select a simple random sample is to number each individual of the population from 1 to N. Using a random digit table or a random number generator, numbers are randomly selected without replacement and the corresponding individuals become part of the sample.

- A **systematic random sample** involves choosing from of a population using a random starting point, and then selecting members according to a fixed, periodic interval (such as every 10th member).

- A **stratified random sample** involves randomly sampling from *every* **strata**, where the strata should correspond to a variable thought to be associated with the variable of interest. This ensures that the sample will have appropriate representation from each of the different strata and reduces variability in the sample estimates.

- A **cluster random sample** involves randomly selecting a set of **clusters**, or groups, and then collecting data on all individuals in the selected clusters. This can be useful when sampling clusters is more convenient and less expensive than sampling individuals, and it is an effective strategy when each cluster is approximately representative of the population.

- Remember: *Individual strata should be homogeneous (self-similar), while individual clusters should be heterogeneous (diverse).* For example, if smoking is correlated with what is being estimated, let one stratum be all smokers and the other be all non-smokers, then randomly select an appropriate number of *individuals* from *each* strata. Alternately, if age is correlated with the variable being estimated, one could randomly select a *subset* of clusters, where each cluster has mixed age groups.

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**.

- **Descriptive statistics** describes or summarizes data, while **inferential statistics** uses samples to generalize or infer something about a larger population.

2.2 Numerical summaries and box plots

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

- Mean and SD are sensitive to outliers. Median and IQR are more robust and less sensitive to outliers.

- 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).

- **Linear transformations of data**. Adding a constant to every value in a data set shifts the mean but does not affect the standard deviation. Multiplying the values in a data set by a constant will multiply the mean and the standard deviation by that constant, except that the standard deviation must always remain positive.

- **Box plots** do not show the *distribution* of a data set in the way that histograms do. Rather, they provide a visual depiction of the **5-number summary**, which consists of: min, Q_1, Q_2, Q_3, max. While a box plot does not indicate modes, it can show skew and outliers.

2.3 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: $Z = \frac{x-\text{mean}}{SD}$.

- The **normal distribution** is the most commonly used distribution in Statistics. Many distribution are approximately normal, but none are exactly normal.

- The empirical rule (68-95-99.7 Rule) comes from the normal distribution. The closer a distribution is to normal, the better this rule will hold.

- It is often useful to use the standard normal distribution, which has mean 0 and SD 1, to approximate a discrete histogram. There are two common types of **normal approximation problems**, and for each a key step is to find a Z-score.

  A: *Find the percent or probability of a value greater/less than a given x-value.*
  1. Verify that the distribution of interest is approximately normal.
  2. Calculate the Z-score. Use the provided population mean and SD to standardize the given x-value.
  3. Use a calculator function (e.g. `normcdf` on a TI) or other technology 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.

2.4 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 chart**. To summarize and compare two categorical variables, use **side-by-side** or **segmented** (stacked) bar charts.
- **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 chart** or a **segmented bar chart**.

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.
- When a distribution is nearly normal we can use the **empirical rule** (68-95-99.7 rule), and we can use a normal model to approximate the histogram.
- **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 and probability distributions

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.

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

Instead, to find the probability that A or B or C occurs, find the probability that none of them occur and subtract that value from 1.

- Two events are **independent** when the occurrence of one does not change the likelihood of the other.
- In the *special case* where A and B are **independent**: $P(A \text{ and } B) = P(A) \times P(B)$.

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 $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$.

3.3 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.

3.4 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.

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

  We can think of $P(x_i)$ as the *weight*, and each term is weighted its appropriate amount.
- The **mean** of a probability distribution does not need to be a value in the distribution. It represents the average of many, many repetitions of a random process. The **standard deviation** represents the typical variation of the outcomes from the mean, when the random process is repeated over and over.
- **Linear transformations**. Adding a constant to every value in a probability distribution adds that value to the mean, but it does not affect the standard deviation. When multiplying every value by a constant, this multiplies the mean by the constant and it multiplies the standard deviation by the absolute value of the constant.
- **Combining random variables**. Let X and Y be random variables and let a and b be constants.
  - The expected value of the sum is the sum of the expected values.
    $E(X + Y) = E(X) + E(Y)$
    $E(aX + bY) = a \times E(X) + b \times E(Y)$
  - When X and Y are **independent**: The standard deviation of a sum or a difference is the square root of the sum of each standard deviation squared.
    $SD(X + Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$

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

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

- 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.

3.5 Geometric distribution

- It is useful to model yes/no, success/failure with the values 1 and 0, respectively. 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 x^{th} trial** is given by $(1-p)^{x-1}p$. We can see the reasoning behind this formula as follows: for the first success to happen on the x^{th} trial, it has to *not* happen the first $x-1$ trials (with probability $1-p$), and then happen on the x^{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. If X has a geometric distribution with parameter p, then $P(X=x) = (1-p)^{x-1}p$, where $x = 1, 2, 3 \ldots$.

- 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**: $\mu_X = \frac{1}{p}$ and **standard deviation**: $\sigma_X = \frac{\sqrt{1-p}}{p}$. If the probability of success is $\frac{1}{10}$, then *on average* it takes 10 trials until we see the first success.

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

3.6 Binomial distribution

- $\binom{n}{x}$, the **binomial coefficient**, describes the number of combinations for arranging x successes among n trials. $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, where $n! = 1 \times 2 \times 3 \times ...n$, and 0!=0.

- The **binomial formula** can be used to find the probability that something happens *exactly x times in n trials*. Suppose the probability of a single trial being a success is p. Then the probability of observing exactly x successes in n independent trials is given by

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

- To apply the binomial formula, the events must be **independent** from trial to trial. Additionally, n, the number of trials must be fixed in advance, and p, the probability of the event occurring in a given trial, must be the same for each trial.

- To use the binomial formula, first confirm that the binomial conditions are met. Next, identify the number of trials n, the number of times the event is to be a "success" x, and the probability that a single trial is a success p. Finally, plug these three numbers into the formula to get the probability of exactly x successes in n trials.

- 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.

- The distribution of the *number of successes* in n independent trials gives rise to a **binomial distribution**. If X has a binomial distribution with parameters n and p, then
$P(X = x) = \binom{n}{x} p^x (1-p)^n - x$, where $x = 0, 1, 2, 3 \ldots, n$.

- To write out a binomial probability **distribution table**, list all possible values for x, the number of successes, then use the binomial formula to find the probability of each of those values.

- 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*)

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.
- **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.
- 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 **geometric distribution** provides a model for the number of trials until the first success, when the trials are independent.
- The **binomial distribution** provides a model for the number of successes in n independent trials.
- The geometric distribution is always right skewed. However, when the success-failure rule is met (at least 10 success and 10 failures), the binomial distribution can be modeled using a normal distribution with mean $= np$ and standard deviation $\sqrt{np(1-p)}$.

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

Sampling distributions

4.1 Sampling distribution of a sample proportion

- 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: $Z = \frac{x - \text{mean}}{SD}$.
- The standard deviation of $\hat{p}$ describes the typical error or distance of the sample proportion from the population proportion. It also tells us how much the sample proportion is likely to vary from one random sample to another.
- The sampling distribution for the sample proportion $\hat{p}$ for a random sample of size n is identical to the binomial distribution with parameters n and p, but 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 for the sample proportion $\hat{p}$, where the observations can be considered independent:
  - 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** problems involving a sample proportion. *Find the probability of getting more / less than __% yeses in a sample of size n.*
  1. Identify n and p. Verify than observations can be treated as independent and that $np \geq 10$ 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.

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 for the sample average $\bar{x}$ where the observations can be treated as independent:
  - 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 observations can be treated as independent and that either the population is approximately normal or $n \geq 30$.
  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 Sampling distribution for a difference of proportions or means

- When two random variables each follow a nearly normal distribution, the distribution of their difference also follows a nearly normal distribution.

- Both $\hat{p}_1 - \hat{p}_2$ and $\bar{x}_1 - \bar{x}_2$ are statistics that can take on different values from one random sample to the next. As such, they have *sampling distributions* that can be described by their center, spread, and shape.

- Three important facts about the sampling distribution for the difference of sample proportions $\hat{p}_1 - \hat{p}_2$ where the observations can be treated as independent:
  - The mean of the difference of sample proportions, denoted by $\mu_{\hat{p}_1-\hat{p}_2}$, is equal to $p_1 - p_2$. *(center)*
  - The SD of the difference of sample proportions, denoted by $\sigma_{\hat{p}_1-\hat{p}_2}$, is equal to $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. *(spread)*
  - When both groups meet the success-failure condition, the difference of sample proportions can be modeled using a normal distribution. *(shape)*

- Three important facts about the sampling distribution for the difference of sample means $\bar{x}_1 - \bar{x}_2$ where the observations can be treated as independent:
  - The mean of the difference of sample means, denoted by $\mu_{\bar{x}_1-\bar{x}_2}$, is equal to $\mu_1 - \mu_2$. *(center)*
  - The SD of the difference of sample means, denoted by $\sigma_{\bar{x}_1-\bar{x}_2}$, is equal to $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. *(spread)*
  - When both populations are nearly normal or when $n_1 \geq 30$ and $n_2 \geq 30$, the difference of sample means can be modeled using a normal distribution. *(shape)*

- When the difference of sample proportions $\hat{p}_1 - \hat{p}_2$ or the difference of sample means $\bar{x}_1 - \bar{x}_2$ follow a nearly normal distribution, we can find the probability that the difference is greater than or less than a certain amount by finding a Z-score and using the normal approximation.

Chapter Highlights

This chapter began by introducing the idea of a **sampling distribution**. As with any distribution, we can summarize a sampling distribution with regard to its center, spread, and shape. A common thread that ran through this chapter is the application of **normal approximation** (introduced in Section 2.3) to different sampling distributions.
The key steps are included for each of the normal approximation scenarios below. To verify that observations can be considered independent, verify that you have one of the following: a random process, a random sample with replacement, or a random sample without replacement of less than 10% of the population. To satisfy the independence condition when working with two groups, we require 2 independent random samples with replacement, 2 independent samples without replacement of less than 10% of their populations, or an experiment with 2 randomly assigned treatments. For completion and comparison purposes, we include cases introduced in earlier chapters as well in the overview below.

1. Normal approximation for numerical **data**: (introduced in Section 2.3)

   - Verify that observations can be treated as independent and that population is approximately normal.
   - Use a normal model with mean μ and SD σ.

2. Normal approximation for a **sample proportion** (with categorical data):
   - Verify that observations can be treated as independent and that $np \geq 10$ and $n(1-p) \geq 10$.
   - Use a normal model with mean $\mu_{\hat{p}} = p$ and SD $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.
3. Normal approximation for a **sample mean** (with numerical data):
   - Verify that observations can be treated as independent and that population is approximately normal or that $n \geq 30$.
   - Use a normal model with mean $\mu_{\bar{x}} = \mu$ and SD $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
4. Normal approximation for a **difference of sample proportions**:
   - Verify that observations can be treated as independent and that $n_1 p_1 \geq 10$, $n_1(1-p_1) \geq 10$, $n_2(1-p_2) \geq 10$, and $n_2(1-p_2) \geq 10$.
   - Use a normal model with mean $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and SD $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.
5. Normal approximation for a **difference of sample means**:
   - Verify that observations can be treated as independent and that both populations are nearly normal or both n_1 and n_2 are ≥ 30.
   - Use a normal model with mean: $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and SD: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

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

In the case of proportions and counts, we never look to see if the *population* is normal. That would not make sense because a yes/no variable cannot have a normal distribution.

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

- In this section we laid the groundwork for our study of **inference**. Inference involves using known sample values to estimate or better understand unknown population values.

- 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.

- It is helpful to imagine a point estimate as being drawn from a particular sampling distribution.

- The **standard error** (**SE**) of a point estimate tells us the typical error or uncertainty associated with the point estimate. It is also an estimate of the spread of the sampling distribution.

- 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.

- A point estimate has **lower variability** (more precise) when the *standard deviation* of the sampling distribution is smaller.

- 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.

- 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*.

5.2 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 population 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.

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.

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

Margin of error

- Confidence intervals are often reported as: point estimate ± margin of error. The **margin of error** (ME) = critical value × 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 

- A **hypothesis test** is a statistical technique used to evaluate competing claims based on data.
- The competing claims are called **hypotheses** and are often about population parameters (e.g. μ and p); they are 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.
- 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: $\frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$.

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

- 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.

- **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. P(Type I error) = α.
  - 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. P(Type II error) = β.
  - 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.
  - 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 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 $\uparrow$.

- 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 *not practically significant*. That is, the difference detected may be so small as to be unimportant or meaningless.
- The inference procedures in this book all require two broad conditions to be met. The first is that some type of *random sampling* or *random assignment* must be involved. The second condition focuses on sample size and skew to determine whether the point estimate follows the intended distribution.

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,

[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 range 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 probability is about what would happen in a random sample, not about what is true of the population.

- 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 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 explore 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 for $\hat{p}$ is approximately normal when the success-failure condition is met and the observations are independent. When the sampling distribution for $\hat{p}$ is normal, the standardized test statistic also follows a **normal** distribution.
- When verifying the success-failure condition and calculating the SE,
  - use the *sample* proportion $\hat{p}$ for the confidence interval, but
  - use the *null/hypothesized* proportion p_0 for the hypothesis test.
- When there is one sample and the parameter of interest is a single proportion:
  - Estimate p at the C% confidence level using a **1-proportion Z-interval**.
  - Test H_0: $p = p_0$ at the α significance level using a **1-proportion Z-test**.
- The one proportion Z-test and Z-interval require the sampling distribution for $\hat{p}$ to be nearly normal. For this reason we must check that the following conditions are met.
  1. Independence: The data should come from a random sample or random process. When sampling without replacement, check that the sample size is less than 10% of the population size.
  2. Success-failure for Interval: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.
     Success-failure for Test, assuming H_0: $p = p_0$ is true: $np_0 \geq 10$ and $n(1-p_0) \geq 10$.
- When the conditions are met, we calculate the confidence interval and the test statistic as follows.

Confidence interval: point estimate $\pm$ $z^\star \times SE$ of estimate

Test statistic: $Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

Here the point estimate is the sample proportion $\hat{p}$.

The SE of estimate is the SE of the sample proportion.

- For an Interval, use $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- For a Test with H_0: $p = p_0$, use $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$.

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

$$z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < m$$

$z^\star$ depends on the desired confidence level. Unless a particular proportion is given in the problem, use $\hat{p} = 0.5$. We solve for the sample size n. The final answer should be an *integer*, since n refers to a number of people or things.

6.2 Difference of two proportions

In the previous section, we looked at inference for a single proportion. In this section, we *compared* 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. The difference, $p_1 - p_2$ tells us whether p_1 is greater than, less than, or equal to p_2.
- 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 for $\hat{p}_1 - \hat{p}_2$ is nearly normal when the success-failure condition is met for *both* groups and when the observations are independent between and within groups. When the sampling distribution for $\hat{p}_1 - \hat{p}_2$ is nearly 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** $\hat{p}_c = \frac{x_1+x_2}{n_1+n_2}$, i.e. the combined number of yeses over the combined sample sizes, when verifying the success-failure condition and when finding the SE. For the confidence interval, do not use the pooled sample proportion; use the separate values of $\hat{p}_1$ and $\hat{p}_2$.
- 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):
  - Estimate $p_1 - p_2$ at the C% confidence level using a **2-proportion Z-interval**.

  - Test H_0: $p_1 - p_2 = 0$ at the α significance level using a **2-proportion Z-test**.

- The two proportion Z-interval and Z-test require the sampling distribution for $\hat{p}_1 - \hat{p}_2$ to be nearly normal. For this reason we must check that the following conditions are met.

  1. Independence: Data come from 2 independent random samples or from a randomized experiment with 2 treatments. When sampling without replacement, check that the sample size is less than 10% of the population size for both samples.
  2. Success-failure for CI: $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$.

     Success-failure for Test: $n_1\hat{p}_c \geq 10$, $n_1(1-\hat{p}_c) \geq 10$, $n_2\hat{p}_c \geq 10$, and $n_2(1-\hat{p}_c) \geq 10$.

- When the conditions are met, we calculate the confidence interval and the test statistic using the same structure as in the previous section.

  Confidence interval: point estimate $\pm$ $z^\star \times SE$ of estimate

  Test statistic: $Z = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

  Here the point estimate is the difference of sample proportions $\hat{p}_1 - \hat{p}_2$.

  The SE of estimate is the SE of a difference of sample proportions.

  - For a CI, use: $SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.
  - For a Test, use: $SE = \sqrt{\hat{p}_c(1-\hat{p}_c)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

6.3 Testing for goodness of fit using chi-square

The inferential procedures we saw in the first two sections of this chapter are based on the test statistic following a *normal distribution*. In this section, we introduced 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 increases: the center increases, the spread increases, and 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 chi-square goodness of fit test, we calculate a **χ^2-statistic**, which is a measure of how far the observed counts in the sample are from the expected counts, relative to the expected counts, under the null hypothesis. $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$.

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

  - 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.

- A larger χ^2 represents greater deviation between the observed values and the expected values, relative to the expected values. For a fixed degrees of freedom, a larger χ^2 value leads to a smaller p-value, providing greater evidence against H_0.

- **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&M's with the specified distribution of color, the hypotheses can often be written 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.

  We test these hypotheses at the α significance level using a **χ^2 goodness of fit test**.

- For the χ^2 goodness of fit test, we check the following conditions to verify that the test statistic follows a chi-square distribution.

  1. Independence: Data come from a random sample or random process. When sampling

     without replacement, check that sample size is less than 10% of the population size.

  2. Expected counts: All expected counts are ≥ 5.

- We calculate the test statistic as follows:

  test statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$; $\qquad df = \#$ of categories $- 1$

- The p-value is the area to the *right* of the χ^2-statistic under the chi-square curve with the appropriate df.

- For a χ^2 test, the p-value corresponds to the probability of getting a test statistic as large as we got or larger, assuming the null hypothesis is true and assuming the chi-square model holds.

6.4 Homogeneity and independence in two-way tables

- When there are two categorical variables, rather than one, the data can be arranged in a **two-way table**.

- When working with a two-way table, the **expected count** for each row,column combination is calculated as: expected count $= \frac{(\text{row total}) \times (\text{column total})}{\text{table total}}$.

- When categorical data are arranged in a two way table, use the χ^2 test for homogeneity or the χ^2 test for independence. These tests are almost identical; the differences lie in the data collection method and in the hypotheses.
- When there are **multiple random 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, the hypotheses can be written as follows:

  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.

  We test these hypotheses at the α significance level using a **χ^2 test for homogeneity**.
- When there is **one random sample** and we are looking for association or dependence between two categorical variables, e.g. testing for an association between gender and political party, the hypotheses can be written as:

  H_0: [variable 1] and [variable 2] are independent.

  H_A: [variable 1] and [variable 2] are dependent.

  We test these hypotheses at the α significance level using a **χ^2 test for independence**.
- In addition to the independence/random condition, all expected counts must be at least 5 for the test statistic to follow a chi-square distribution.
- The chi-square statistic and associated *df* are found as follows:

  test statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

  $df = (\#\text{ of rows} - 1)(\#\text{ of cols} - 1)$
- The p-value is the area to the *right* of χ^2-statistic under the chi-square curve with the appropriate *df*.

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&M's.
- **χ^2 test for 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 for 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.

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.

Chapter 7

Inference for numerical data

7.1 Inference for a single mean with the t-distribution

- The t-distribution.
  - When calculating a test statistic for a mean, using the sample standard deviation in place of the population standard deviation gives rise to a new distribution called the t-distribution.
  - 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 can use the t-table at $df = \infty$ to find the value of $z^\star$.
- When carrying out inference for a single mean, we use the t-distribution with $n - 1$ degrees of freedom.
- When there is one sample and the parameter of interest is a single mean:
  - Estimate μ at the C% confidence level using a **1-sample *t*-interval**.
  - Test H_0: $\mu = \mu_0$ at the α significance level using a **1-sample *t*-test**.
- The one-sample t-interval and t-test require that the sampling distribution for $\bar{x}$ be nearly normal. For this reason we must check that the following conditions are met.
  1. Independence: The data come from a random sample or random process. When sampling without replacement, check that the sample size is less than 10% of the population size.
  2. Large sample or normal population: $n \geq 30$ or population distribution is nearly normal. - If the sample size is less than 30 and the population distribution is unknown, check for strong skew or outliers in the data. If neither is found, then the condition that the population distribution is nearly normal is considered reasonable.

- When the conditions are met, we calculate the confidence interval and the test statistic as we did in the previous chapter, except that we use $t^\star$ for the critical value and we use T for the test statistic.

  Confidence interval: point estimate $\pm$ $t^\star \times SE$ of estimate

  Test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

  Here the point estimate is the sample mean: $\bar{x}$.

  The SE of estimate is the SE of the sample mean: $\frac{s}{\sqrt{n}}$.

  The degrees of freedom is given by $df = n - 1$.

- To calculate the minimum sample size required to estimate a mean with C% confidence and a margin of error no greater than m, we set up an inequality as follows:

$$z^\star \frac{\sigma}{\sqrt{n}} \leq m$$

  $z^\star$ depends on the desired confidence level and σ is the standard deviation associated with the population. We solve for the sample size, n. Always round the answer up to the next *integer*, since n refers to a number of people or things.

7.2 Inference with paired data

- Paired data can come from a random sample or a matched pairs experiment. With paired data, we are often interested in whether the *difference* is positive, negative, or zero. For example, the difference of paired data from a matched pairs 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.

- To carry out inference with paired data, we first find all of the sample differences. Then, we perform a one-sample procedure using the *differences*. For this reason, the confidence interval and hypothesis test with paired data use the one-sample t-procedures, where the degrees of freedom is given by $n_{diff} - 1$.

- When there is paired data and the parameter of interest is a mean of differences:

  – Estimate μ_{diff} at the C% confidence level using a **1-sample *t*-interval** with paired data.

  – Test H_0: $\mu_{diff} = 0$ at the α significance level using a **1-sample *t*-test** with paired data.

- The one-sample t-interval and t-test with paired data require the sampling distribution for $\bar{x}_{diff}$ to be nearly normal. For this reason, we must check that the following conditions are met.

1. Independence: Data should come from one random sample (with paired observations) or from a randomized matched pairs experiment. If sampling without replacement, check that the sample size is less than 10% of the population size.
2. Large sample or normal population: $n_{diff} \geq 30$ or population of differences nearly normal. - If the number of differences is less than 30 and it is not known that the population of differences is nearly normal, we argue that the population of differences could be nearly normal if there is no strong skew or outliers in the sample differences.

- When the conditions are met, we calculate the confidence interval and the test statistic as we did in the previous section. Here, our data is a list of differences.

  Confidence interval: point estimate $\pm$ $t^\star \times SE$ of estimate

  Test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

  Here the point estimate is the mean of sample differences: $\bar{x}_{diff}$.

  The SE of estimate is the SE of a mean of sample differences: $\frac{s_{diff}}{\sqrt{n_{diff}}}$.

  The degrees of freedom is given by $df = n_{diff} - 1$.

7.3 Inference for the difference of two means

- This section introduced inference for a difference of means, which is distinct from inference for a mean of 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. To calculate a mean of difference, $\bar{x}_{diff}$, we first calculate all of the differences, then we find the mean of those differences.

- Inference for a difference of means is based on the t-distribution. The degrees of freedom is complicated to calculate and we rely on a calculator or other software to calculate this.[1]

- When there are two samples or treatments and the parameter of interest is a difference of means:
  - Estimate $\mu_1 - \mu_2$ at the C% confidence level using a **2-sample *t*-interval**.
  - Test H_0: $\mu_1 - \mu_2 = 0$ at the α significance level using a **2-sample *t*-test**.

- The 2-sample t-test and t-interval require the sampling distribution for $\bar{x}_1 - \bar{x}_2$ to be nearly normal. For this reason we must check that the following conditions are met.

  1. Independence: The data should come from 2 independent random samples or from a randomized experiment with 2 treatments. When sampling without replacement, check that the sample size is less than 10% of the population size for each sample.

[1] If this is not available, one can use $df = min(n_1 - 1, n_2 - 1)$.

2. Large samples or normal populations: $n_1 \geq 30$ and $n_2 \geq 30$ or both population distributions are nearly normal.
- If the sample sizes are less than 30 and it is not known that both population distributions are nearly normal, check for excessive skew or outliers in the data. If neither exists, the condition that both population distributions could be nearly normal is considered reasonable.

- When the conditions are met, we calculate the confidence interval and the test statistic as follows.

  Confidence interval: point estimate $\pm$ $t^\star \times SE$ of estimate

  Test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$

  Here the point estimate is the difference of sample means: $\bar{x}_1 - \bar{x}_2$.

  The SE of estimate is the SE of a difference of sample means: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

  Find and record the *df* using a calculator or other software.

Chapter Highlights

We've reviewed a wide set of inference procedures over the last 3 chapters. Let's revisit each and discuss the similarities and differences among them. The following confidence intervals and tests are structurally the same – 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
- 1-sample t-test/interval with paired data
- 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.

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 for homogeneity - compares a categorical variable across multiple groups.
- χ^2 test for 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 the *distribution* of one or more categorical variables, 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 parameter to be estimated and the confidence level. For a hypothesis test, identify the hypotheses to be tested and the significance level.
- 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.

For a summary of these hypothesis test and confidence interval procedures (including one more that we will encounter in the next chapter, see the Inference Guide.

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 call x the **explanatory variable** 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.
- When we write a linear model, we use $\hat{y}$ to indicate that it is the model or the prediction. The value $\hat{y}$ 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 modeled value, computed as $y - \hat{y}$. The order of the difference matters, and the sign of the residual will tell us if the model overpredicted or underpredicted 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.
- **Correlation**, denoted with the letter r, measures the strength and direction of a linear relationship. The following are some important facts about correlation.
  - 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.
- Just like Z-scores, the correlation has no units. Changing the units in which x or y are measured does not affect the correlation.
- Correlation is sensitive to outliers. Adding or removing a single point can have a big effect on the correlation.
- As we learned previously, correlation is not causation. Even a very strong correlation cannot prove causation; only a well-designed, controlled, randomized experiment can prove causation.

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: $\hat{y} = a + bx$, and we can calculate a and b based on the summary statistics as follows:

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

- *Interpreting* the **slope** and **y-intercept** of a linear model
  - The slope, b, describes the *average* increase or decrease in the y variable if the explanatory variable x is one unit larger.
  - The y-intercept, a, 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 know 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. 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.

  - The R^2 for a linear model describes the *proportion of variation* in the y variable that is *explained by* the regression line.
  - 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 $r = -\sqrt{R^2}$ or $r = \sqrt{R^2}$. 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 or curvature, such as a $\cup$-shape, a linear model is not appropriate.

- **Outliers** in regression are observations that fall far from the "cloud" of points.

- An **influential 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.

8.3 Transformations for skewed data

- A **transformation** is a rescaling of the data using a function. When data are very skewed, a log transformation often results in more symmetric 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 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.

8.4 Inference for the slope of a regression line 

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

- We use the slope b as a *point estimate* for the slope β of the population regression line. The slope of the population regression line is the true increase/decrease in y for each unit increase in x. If the slope of the population regression line is 0, there is no linear relationship between the two variables.

- Under certain assumptions, the sampling distribution for b is *normal* and the distribution of the standardized test statistic using the standard error of the slope follows a ***t*-distribution** with $n - 2$ degrees of freedom.

- When there is (x, y) data and the parameter of interest is the slope of the population regression line, e.g. the slope of the population regression line relating air quality index to average rainfall per year for each city in the United States:
  - Estimate β at the C% confidence level using a ***t*-interval for the slope**.
  - Test H_0: $\beta = 0$ at the α significance level using a ***t*-test for the slope**.
- The conditions for the t-interval and t-test for the slope of a regression line are the same.
  1. Independence: Data come from a random sample or randomized experiment. If sampling without replacement, check that the sample size is less than 10% of the population size.
  2. Linearity: Check that the scatterplot does not show a curved trend and that the residual plot shows no $\cup$-shape pattern.
  3. Constant variability: Use the residual plot to check that the standard deviation of the residuals is constant across all x-values.
  3. Normality: The population of residuals is nearly normal or the sample size is ≥ 30. If the sample size is less than 30 check for strong skew or outliers in the sample residuals. If neither is found, then the condition that the population of residuals is nearly normal is considered reasonable.
- The confidence interval and test statistic are calculated as follows:

  Confidence interval: point estimate $\pm$ $t^\star \times SE$ of estimate, or

  Test statistic: $T = \frac{\text{point estimate} - \text{null value}}{SE \text{ of estimate}}$ and p-value

  point estimate: the slope b of the sample regression line

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

  $df = n - 2$
- The confidence interval for the slope of the population regression line estimates the true average increase in the y-variable for each unit increase in the x-variable.
- The t-test for the slope and the 1-sample t-test for a mean of differences both involve *paired*, numerical data. However, the t-test for the slope asks if the two variables have a linear *relationship*, specifically if the *slope* of the population regression line is different from 0. The 1-sample t-test for a mean of differences, on the other hand, asks if the two variables are, on average, *different*, specifically if the *mean* of the population differences is not equal to 0.

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.

- 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: $\hat{y} = a + bx$, where a is the y-intercept and b is the slope. It is important to be able to *calculate* a and b 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.
- $\hat{y} = a + bx$ provides the best fit line for the *observed data.* To estimate or hypothesize about the slope of the population regression line, first confirm that the residual plot has no pattern and that a linear model is reasonable, then use a $\boldsymbol{t}$-interval for the slope or a $\boldsymbol{t}$-test for the slope 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 distributions.
- 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.