STA2023 Notes

Tuyen Truong

April 9, 2023

Note that these notes do not span the all the topics covered in STA2023, just some of the concepts that I think are most important to have a firm grasp on. I will update these notes. You can find these updates on my website https://maitiennn.github.io/

Contents

§1 Background

In statistics, we can classify variables as categorical or quantitative. Data can be categorical or quantitative. In some research studies one variable is used to predict or explain differences in another variable. In those cases, the explanatory variable is used to predict or explain differences in the response variable. In an experimental study, the explanatory variable is the variable that is manipulated by the researcher.

Definition 1.1 (Quantitative Variables)**.** Numerical values with magnitudes that can be placed in a meaningful order with consistent intervals, also known as numerical.

Definition 1.2 (Categorical Variables)**.** Names or labels (i.e., categories) with no logical order or with a logical order but inconsistent differences between groups (e.g., rankings), also known as qualitative.

Definition 1.3 (Explanatory Variable)**.** The explanatory variable also known as the independent or predictor variable, explains variations in the response variable; in an experimental study, i.e. it is manipulated by the researcher.

Definition 1.4 (Response Variable)**.** The response variable also known as the dependent or outcome variable, its value is predicted or its variation is explained by the explanatory variable; in an experimental study, this is the outcome that is measured following manipulation of the explanatory variable

> **Remark 1.5.** It is important that we distinguish between a parameter and a statistic.

Definition 1.6 (Parameter)**.** We call a measure concerning the population a **parameter**. The population is the entire set of possible cases. The population is who we want to make statistical inferences about.

Definition 1.7 (Statistic)**.** We call a measure concerning the sample a **statistic**. Recall that a sample is a subset of the population.

Example 1.8

Think of what we do in our labs. We are trying to draw inferences about a population (for example, ALL UF students), but the sample is taken from our lab sections which is a subset of the population. However, it may be unrealistic or even impossible to gather data from the entire population. The subset of the population from which data are actually gathered is the sample. A sample should be selected from a population randomly, otherwise it may be prone to bias. Our goal is to obtain a sample that is representative of the population.

Definition 1.9 (Representative Sample)**.** A representative sample is a subset of the population from which data are collected that accurately reflects the population.

Definition 1.10 (Bias)**.** We want to avoid bias when we are collecting data. Bias is the systematic favoring of certain outcomes. There are different types of bias, which we will define below.

Definition 1.11 (Sampling Bias)**.** Sampling bias is the systematic favoring of certain outcomes due to the methods employed to obtain the sample. (See examples below)

Example 1.12 (Weigh Loss Study Volunteers)

A medical research center is testing a new weight loss treatment. They advertise on a social media site that they are looking for volunteers to participate. There is sampling bias because the sample will be limited to people who use the social media site where they advertised. The individuals who choose to participate may be different from the overall population. For example, volunteers may be individuals who are already actively trying to lose weight. This is not a representative sample because the sample may have characteristics that are different from the population of interest.

Definition 1.13 (Non-response Bias)**.** Non-response bias is the systematic favoring of certain outcomes that occurs when the individuals who choose participate in a study differ from the individuals who choose to not participate.

Example 1.14 (Restaurant Experience Survey)

A restaurant invited their recent customers to complete an online survey. Customers who had really strong feelings about their experience, either positive or negative, were very likely to complete the survey while customers who had a neutral experience were much less likely to complete the survey. This is an example of non-response bias because the individuals who chose to participate differed from those who chose to not participate.

Definition 1.15 (Response Bias)**.** Response bias is the systematic favoring of certain outcomes that occurs when participants do not respond truthfully; they may do so to align with social norms or to appease the researcher

Example 1.16 (Academic Cheating)

Using an anonymous online survey, a professor asks his students "Have you cheated on an exam in my class?" Many of the students who have cheated still answered "no." This is an example of response bias because the participants are not responding truthfully; instead their responses are biased toward responses that are less likely to get them in trouble.

Remark 1.17. There are many different ways to select a sample from a population. We Some of these methods are probability-based, such as the simple random sampling method, which you'll read about below and in your textbook. To prevent sampling bias and obtain a representative sample, a sample should be selected using a probability-based sampling design which gives each individual a known equal chance of being selected. The most common probability-based sampling method is the simple random sampling method (SRS). Other probability-based methods include cluster sampling methods and stratified sampling methods. You may learn more about these if you take a research methods course or an advanced statistics course in the future. Other sampling methods are not probability-based, such as convenience sampling methods, which you will read about below.

Definition 1.18 (Simple Random Sample)**.** The Simple Random Sample sampling method is a method of obtaining a sample from a population in which every member of the population has an equal chance of being selected. Using this method, a sample is selected without replacement. This means that once an individual has been selected to be a part of the sample they cannot be selected a second time. If multiple samples are being taken, an individual can appear in more than one sample, but only once in each sample.

Definition 1.19 (Convenience Sample)**.** While probability-based sampling methods are considered better because they can prevent sampling bias, there are times when it is not possible to use one of these methods. For example, a researcher may not have access to the entire population. In cases were probability-based sampling methods are not practical, convenience samples are often used. Convenience sampling is a method of obtaining a sample from a population by ease of accessibility; such a sample is not random and may not be representative of the intended population. See example below.

Example 1.20 (Weight Loss Supplements)

A weight loss company wants to compare how much weight adults lose on their supplement versus a competitor's supplement. To recruit participants, they post an advertisement in a newspaper asking for adults who want to lose weight. This is an example of a volunteer sample which is a convenience sampling method. The researchers are using a sample of individuals who volunteer to participate.

Remark 1.21. Research studies are often classified in terms of their designs. Here, we will make the distinction between experimental and observational research designs. Although researchers aim to make their experiment or observational study as random and free from bias as possible, there can still be issues. A common problem in studies without randomization is that there may be other variables influencing the results. These are known as confounding variables. A confounding variable is related to both the explanatory variable and the response variable.

Definition 1.22 (Experimental Research Design)**.** A study in which the researcher manipulates the treatments (i.e., level of the explanatory variable) received by subjects and collects data; also known as a scientific study.

Definition 1.23 (Observation Research Design)**.** A study in which the researcher collects data without performing any manipulations; also known as a non-experimental study.

Definition 1.24 (Confounding Variables)**.** A confounding variable is a characteristic that varies between cases and is related to both the explanatory and response variables; also known as a lurking variable or a third variable.

Example 1.25 (Ice Cream and Home Invasions)

There is a positive relationship between ice cream sales and home invasions (i.e., as ice cream sales increase throughout the year so do home invasions). It is clear that increases in ice cream sales do not cause home invasions to increase, and home invasions do not cause an increase in ice cream sales. There is a third variable at play here: outdoor temperature. When the weather is warmer both ice cream sales and home invasions increase. In this case, outdoor temperature is a confounding variable because it is related to both ice cream sales and home invasions.

Remark 1.26. In both observational and experimental studies, we often want to compare two or more groups. When comparing two or more groups, cases may be independent or paired.

Definition 1.27 (Independent Groups)**.** Cases in each group are unrelated to one another.

Example 1.28 (Shoes)

A shoe company is studying how many shoes Italian men and women own. In one research study they take a random sample of 500 Italian adults and ask each individual if they identify as a man or women and how many pairs of shoes they own. The men and women in this study are in two independent groups.
In a second study the researchers use a different design. This time they take a random sample of 250 heterosexual married couples in Italy (i.e., 250 husbands and 250 wives). They record the number of shoes owned by each husband and each wife. This is an example of a matched pairs design. Data are paired by couple.

Definition 1.29 (Paired Groups)**.** Cases in each group are meaningfully matched with one another; also known as dependent samples or matched pairs.

Example 1.30 (Exam Scores)

An instructor wants to compare students' scores on the midterm and final exam. This is most often done by obtaining a sample of students and recording each student's midterm exam score and final exam score. In other words, there would be two measurements for each student. This is an example of a matched pairs design because data would be paired by student.

Remark 1.31. We return to the concept of bias. Blinding techniques are used to avoid bias. In a single-blind study the participants do not know what treatment groups they are in, but the researchers interacting with them do know. In a double-blind study, the participants do not know what treatment groups they are in and neither do the researchers who are interacting with them directly. Double-blind studies are used to prevent researcher bias.

Definition 1.32 (Blinding)**.** Blinding is a procedure employed in research to prevent bias in which the participants and/or the researchers interacting with the participants do not know which treatment each case is receiving.

Definition 1.33 (Single Blind Study)**.** Research study in which the participants do not know the treatment group that they have been assigned to.

Definition 1.34 (Double Blind Study)**.** Research study in which neither the participants nor the researchers interacting with them know which cases have been assigned to which treatment groups

Example 1.35 (Caffeine Energy Study)

Researchers want to know if adult males who consume high amounts of caffeine interact more energetically. They obtain a representative sample and randomly assign half of the participants to take a caffeine pill and half to take a placebo pill. The pills are randomly numbered and coded so at the time the researchers do not know which participants have been given caffeine and which have been given the placebo. All participants are told that they may have been given a caffeine pill. After taking the pill, researchers observe the participants interacting with one another and rate the interactions in terms of level of energy.

This is a double-blind study because neither the researchers nor the participants know who is in which group at the time the data are collected. After the data are collected, researchers can look at the pill codes to determine which groups the participants were in to conduct their analyses. A double-blind study is necessary here because the researchers are observing and rating the participants. If the researchers know who is in the caffeine group they may be more likely to rate their levels of energy as very high because that is consistent with their hypothesis.

KNOW THIS TABLE!

	Parameter	**Sample Statistic**
Mean	μ	$\bar{x}$
Difference in Two Means	$\mu_1 - \mu_2$	$\bar{x_1} - \bar{x_2}$
Proportions	p	$\hat{p}$
Difference in Proportions	$p_1 - p_2$	$\hat{p_1} - \hat{p_2}$

Remark 1.36. Without getting too much into probability concepts, I want to talk about distributions (I will make some remarks on this later to clarify what I mean). The one frequently used in this class is a normal distribution and a standard normal distribution. In our labs, we have what is called a sampling distribution. Sample statistics are random variables because they vary from sample to sample. As a result, sample statistics have a distribution called the sampling distribution.

Definition 1.37 (Sample Distribution)**.** A sample distribution is a distribution of sample statistics with a mean approximately equal to the mean in the original distribution and a standard deviation known as the standard error.

Definition 1.38 (Standard Error)**.** The standard error is the standard deviation of a sampling distribution. (formulas for this are usually provided on the exams). Make sure you use the correct formulas for means/proportions.

When we have certain types of data sets, such as quantitative or categorical data, certain assumptions must be made when using them.

§1.1 Categorical Variables

Definition 1.39. Data concerning one categorical variable can be summarized using a proportion. $p = \frac{\#in\ the\ category}{total\ number}$

The formula for sample proportion ($\hat{p}$) is the same.

Definition 1.40 (Assumptions for Categorical Variables)**.** Let n denote the sample size and p_0 the number of successes in our investigation. Assumptions for categorical variables are as follows:
(1) $np_0 \geq 15$
(2) $n(1-p_0) \geq 15$
(3) Simple Random Sample (SRS)

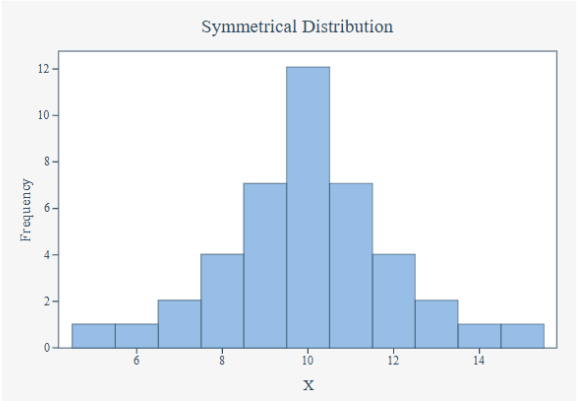
> **Remark 1.41.** We need to check these assumptions before we conduct hypothesis testing, which is covered in the later sections.

§1.2 Quantitative Variables

Note that for quantitative data, we often summarize our data with means, medians, modes, etc. In labs, we typically work with means. Quantitative variables are also often discussed in terms of their shape. Both dot plots and histograms can be used to interpret a distribution's shape. A distribution may be described in terms of symmetry and skewness.

§1.2.1 Shapes

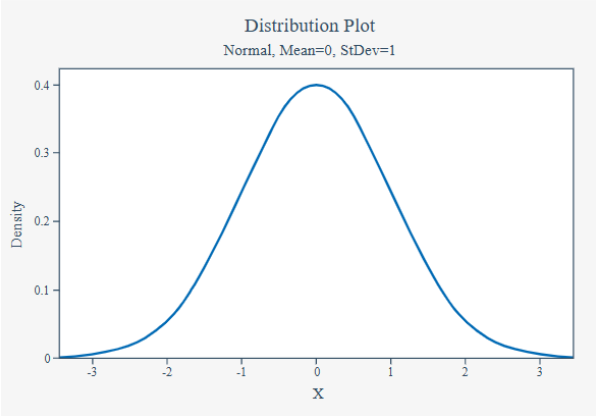
Definition 1.42 (Symmetric)**.** A distribution that is similar on both sides of the center.



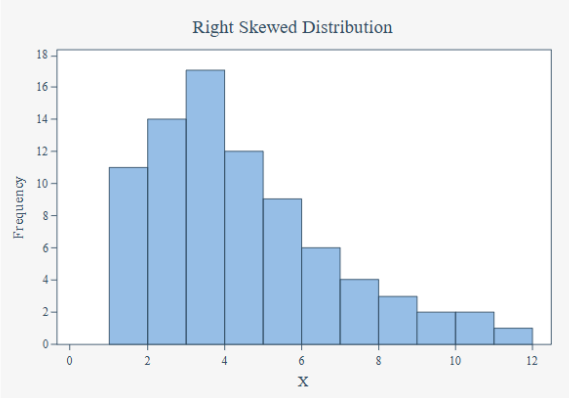
Definition 1.43 (Normal Distribution)**.** One specific type of symmetrical distribution. This is also known as a bell-shaped distribution. The curve of the normal distribution is centered by the mean value, μ, and its spread is measured by the standard deviation σ. These two parameters, μ and σ^2, completely determine the shape and location of the normal density function whose functional form is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty.$$

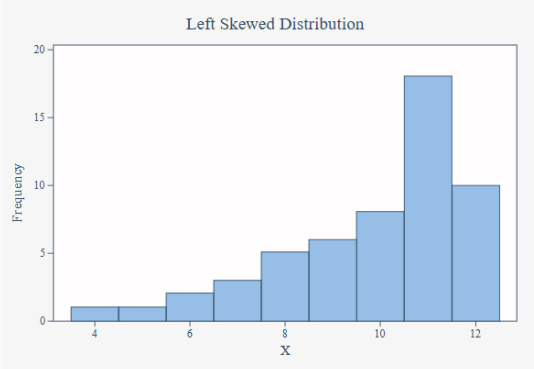
Don't be scared when seeing this formula! Probably won't need it in this course.



Definition 1.44 (Right-Skewed)**.** A distribution in which the higher values (towards the right on a number line) are more spread out than the lower values.



Definition 1.45 (Left-Skewed)**.** A distribution in which the lower values (towards the left on a number line) are more spread out than the higher values.



§1.2.2 Measures of Central Tendency

The mean, median, and mode are three of the most commonly used measures of central tendency.

Definition 1.46 (Mean)**.** The numerical average; calculated as the sum of all of the data values divided by the number of values. Referring to the very important table above, the sample mean is represented as $\bar{x}$ and the population mean is denoted μ.
Formulas:

$$\bar{x} = \frac{\sum x}{n}, \quad \text{where } n \text{ is the sample size.}$$

$$\mu = \frac{\sum x}{N}, \quad \text{where } N \text{ is the population size.}$$

Definition 1.47 (Median)**.** The middle of the distribution that has been ordered from smallest to largest; for distributions with an even number of values, this is the mean of the two middle values

Definition 1.48 (Mode)**.** The most frequently occurring value(s) in the distribution, may be used with quantitative or categorical variables.

Example 1.49 (Household Size)

A group of children are asked how many people live in their household. The following data is collected: $4, 3, 6, 2, 2, 4, 3$.

$$\text{Mean} = \bar{x} = \frac{\sum x}{n} = \frac{4+3+6+2+2+4+3}{7} = \frac{24}{7} = 3.429.$$

Median: First, we need to put all of the values in order from smallest to largest:

2, 2, 3, 3, 4, 4, 6. The value in the middle of this distribution is 3. The median is 3.

Mode:In this distribution, the most common values are 2, 3, and 4.

There are 3 modes: 2, 3, and 4. We say that this distribution is multimodal.

§1.2.3 Measures of Spread

The standard deviation is the most commonly used measure of variability when working with interval variables. In a sample, this is denoted as s. In a population, we denote the standard deviation with the Greek letter "sigma" σ. Before we compute for the standard deviation, we need to find the variance first, which is also known as the standard deviation squared.

Definition 1.50 (Variance)**.** The variance is the average squared distance from the mean. The standard deviation is square root of the variance. We use this for considering how far the data are distributed from the mean. Below are the formulas for the sample and population means.

1. $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$, where n is the sample size.
2. $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$, where N is the population size.
3. Notice that the formulas are slightly different. That is because when we are calculating the sample standard deviation, we are "estimating" the mean with the sample mean $\bar{x}$, so we have to account for that with the minus 1 (think of it as a penalty). This gets into more advanced statistical concepts such as unbiased estimators, which we will not delve into.

§1.2.4 Assumptions for Inference Testing

Definition 1.51 (Assumptions for Quantitative Variables)**.** Let n denote the sample size Assumptions for quantitative variables are as follows:
(1) $n \geq 30$ or original population is normally distributed for Central Limit Theorem to apply.
(2) Simple Random Sample (SRS)

Remark 1.52. We need to check these assumptions before we conduct hypothesis testing, which is covered in the later sections.

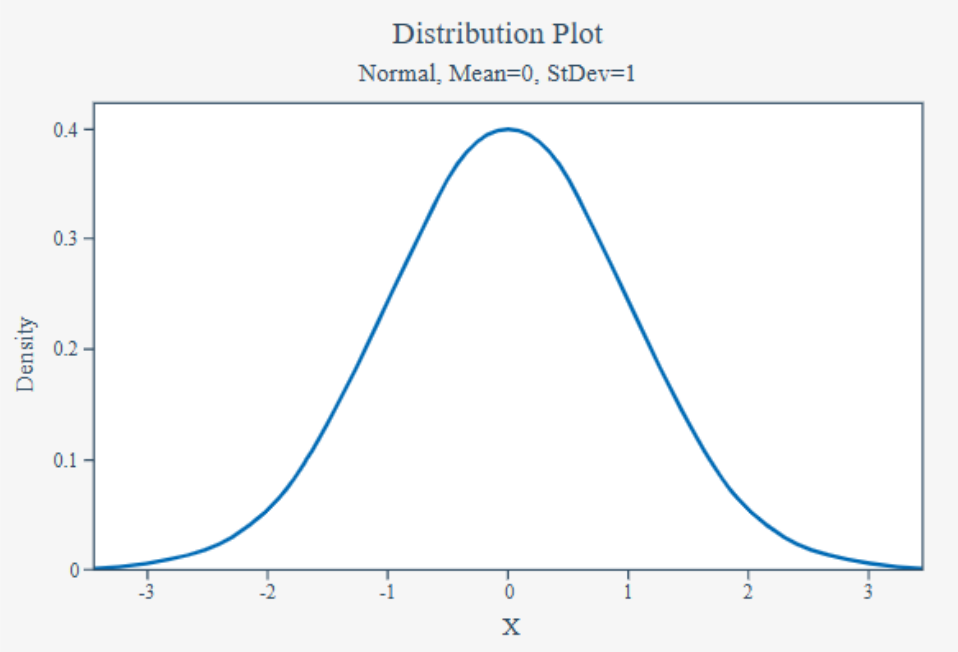
Theorem 1.53 (Central Limit Theorem)
Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then if n is "large enough", $\overline{X}$ has approximately a normal distribution with

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The larger the value of n, the better the approximation.

§2 The Normal Distribution

Among all the distributions we see in practice, one is overwhelmingly the most common. The symmetric, unimodal, bell curve is ubiquitous throughout statistics. Indeed it is so common, that people often know it as the normal distribution. Read more about probability distributions in the probability section!



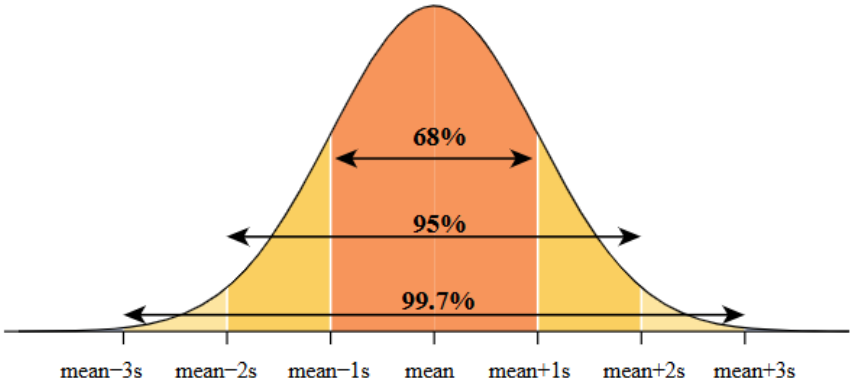
If a normal distribution has mean μ and standard deviation σ, we denote the distribution in shorthand as $\sim N(\mu, \sigma)$. Because the mean and standard deviation describe a normal distribution exactly, they are called the distribution's parameters.

§2.1 Standard Normal Distribution

Definition 2.1 (Standard Normal Distribution)**.** The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the standard normal distribution.

§2.2 The Empirical Rule

Definition 2.2 (The Empirical Rule)**.** The Empirical Rule is a statement about normal distributions. It is also known as the 95% Rule, because 95% is the most commonly used interval. The 95% Rule states that approximately 95% of observations fall within two standard deviations of the mean on a normal distribution. On a normal distribution about 68% of data will be within one standard deviation of the mean, about 95% will be within two standard deviations of the mean, and about 99.7% will be within three standard deviations of the mean



§2.3 z-Score

We often want to put data onto a standardized scale, which can make comparisons more reasonable.

Definition 2.3 (z-Score)**.** The z-score of an observation is defined as the number of standard deviations it falls above or below the mean. If the observation is one standard deviation above the mean, its z-score is 1. If it is 1.5 standard deviations below the mean, then its z-score is -1.5. If x is an observation from a distribution $N(\mu, \sigma)$, we define the z-score mathematically as:

$$z = \frac{x - \mu}{\sigma}.$$

The formula for the z-score of the sample is:

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

Example 2.4

A study of 66,831 chocolate dairy cows found that the mean chocolate milk yield was 12.5 kg per milking with a standard deviation of 4.3 kg per milking. A chocolate cow produces 18.1 kg per milking. What is this cow's z-score?

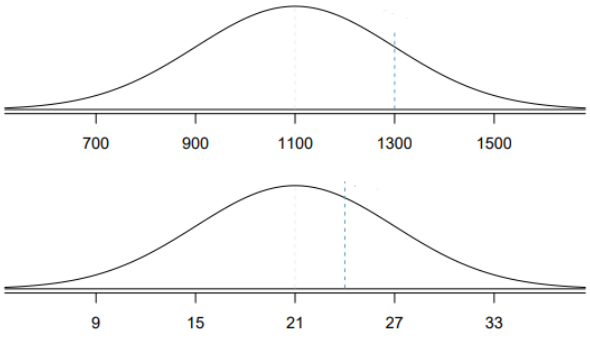
Proof. $z = \frac{x-\bar{x}}{s} = \frac{18.1-12.5}{4.3} = 1.302$.
Therefore, the cow's milk production was 1.302 standard deviations above the mean. □

Example 2.5 (SAT and ACT Scores)
SAT scores are normally distributed with a mean of 1100 and standard deviation of 200. ACT scores are also normally distributed with a mean score 21 and standard deviation 6. Suppose Patty scored 1300 on her SAT and Bill scored 24 in his ACT. Who performed better?

Proof. We will begin by finding Patty's z-score:

$$z_{patty} = \frac{x_{patty} - \mu_{SAT}}{\sigma_{SAT}} = \frac{1300 - 1100}{200} = 1$$
$$z_{bill} = \frac{x_{bill} - \mu_{ACT}}{\sigma_{ACT}} = \frac{24 - 21}{6} = 0.5$$

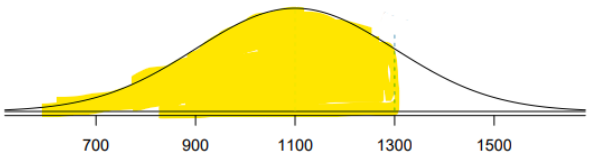
Therefore, Patty performed better.



□

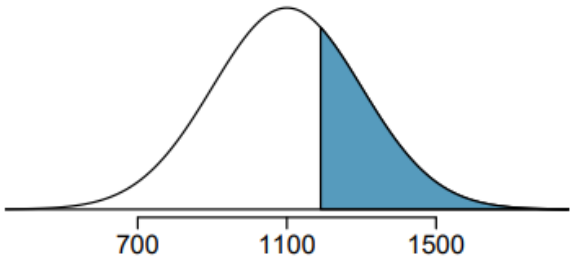
§2.4 Finding Tail Areas/Percentiles

It is very useful in statistics to be able to identify tail areas of distributions. For instance, what fraction of people have an SAT score below Patty's score of 1300? This is the same as the percentile Ann is at, which is the percentage of cases that have lower scores than Ann. We can visualize such a tail area like the curve and shading shown below.



Example 2.6
Carmen is a randomly selected SAT test taker and no information is known about Carmen's SAT aptitude. What is the probability Carmen scores **at least** 1190 on her SAT?

Proof. We will always begin by drawing a picture, more importantly a diagram of the normal distribution. We are interested in a score of 1190 and above, so we shade the area to the right as follows:



The simplest way to find the shaded area under the curve makes use of the z-score of the cutoff value. So, we calculate the z-score as we did before,

$$z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = 0.45.$$

Now, we will look up on the z-table what lower tail value or probability value corresponds to our z-score of 0.45, which is 0.6736 (you can verify this for yourself). But remember, this value gives us the area to the left of the normal distribution; we are interested in the right! So, to find the area above a z-score of 0.45, we compute one minus the area of the lower tail: $1 - 0.6736 = 0.3264$. □

§3 Introduction to Probability

§3.1 Risks and Odds

You may have heard the terms risk and odds before. They are both ways to communicate the likelihood of an event. Risk and odds are often confused with one another. The formulas for computing risk and odds are different and their interpretations are different.

Definition 3.1 (Risk)**.** Risk is the probability that an event will occur. It can be expressed as a decimal, a fraction, or a percent.

$$\text{risk} = \frac{\text{number with the outcome}}{\text{total number of outcomes}}.$$

Example 3.2 (Flu Risk)
45 out of 100 children get the flu each year. risk $= \frac{45}{100} = 0.45$.

Definition 3.3 (Odds)**.** Odds expresses risk by comparing the likelihood of an event happening to the likelihood it does not happen.

$$\text{odds} = \frac{\text{number with the outcome}}{\text{number without the outcome}}$$

OR

$$\text{odds} = \frac{\text{risk}}{1 - \text{risk}}$$

We often interpret odds in relation to the value of 1. For example, if the odds of a game are in favor of the house 2 to 1, that means for every 2 games the house wins it will lose 1.

Example 3.4 (Flu Odds)
The risk of a child getting the flu is which can also be written as 0.45. Because we have the risk, we can use the second odds formula:

$$\text{odds} = \frac{0.45}{1 - 0.45} = 0.818$$

So, the odds of a child getting the flu is 0.818 to 1.

§3.2 Loosely Defining Probability

Definition 3.5 (Sample Space)**.** A sample space S is a set that includes all possible outcomes for a random experiment listed in a mutually exclusive and exhaustive way. The phrase mutually exclusive means that the outcomes of the set do not overlap.

Example 3.6 (Identifying Sample Space in Dice Roll)
Consider the set $S = \{1, 2, 3, 4, 5, 6, \text{even}, \text{odd}\}$. This set is not an appropriate sample space according to our definition for a dice roll experiment because the outcomes are not mutually exclusive. Instead, we write the sample space $S = \{1, 2, 3, 4, 5, 6\}$.

Definition 3.7 (Event)**.** An event is a subset of a sample space. We often denote events as capital letters, for example $A, B, \ldots$.

Remark 3.8. Know some basic set theory (it help to draw the Venn diagrams):

1. $A \cup B$
2. $A \cap B$
3. $\overline{A \cup B} = \bar{A} \cap \bar{B}$ ...
4. De Morgan Laws (communicative, associative, distributive laws) (not really necessary)

Note that I denote the complement of an event say A with a bar, so the complement of A is $\bar{A}$.

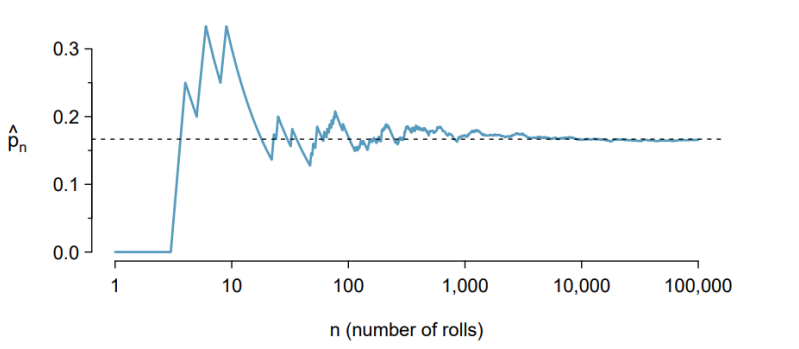
Definition 3.9 (Probability and Properties)**.** Suppose that a random experiment has associated with it a sample space, which we will call S. A probability is a numerically valued function that assigns a number $P(A)$ to every event A so that the following axioms hold:

1. $P(A) \geq 0$;
2. $P(S) = 1$;
3. if $A_1, A_2, \ldots$ is a sequence of mutually exclusive events (that is, a sequence in which $A_i A_j = \varnothing$ (we use that symbol to denote the empty set) for any $i \neq j$), then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

Theorem 3.10 (The Law of Large Numbers)

We now know that probability is defined as a proportion, and it always takes values between 0 and 1 (inclusively). Probability is often illustrated by rolling a die many times. Let $\hat{p_n}$ be the proportion of outcomes that are 1 after the first n rolls. As the number of rolls increases $\hat{p_n}$ will converge to the probability of rolling a 1, $p = \frac{1}{6}$. The figure below shows the convergence of 100,000 dice rolls. The tendency of $\hat{p_n}$ to stabilize around p is described by the Law of Large Numbers. More generally, we can state the Law of Large Numbers in the following way: As more observations are collected, the proportion $\hat{p_n}$ of occurrences with a particular outcome converges to the probability p of that outcome.



Remark 3.11. Don't get too hang up on notation here. Just know that $P(A)$ is a value between 1 and 0 such that $0 \leq P(A) \leq 1$, and know the formulas that follows from this definition below.

Corollary 3.12

1. if A and B are mutually exclusive events, $P(A \cup B) = P(A) + P(B)$;
2. if A and B are NOT DISJOINT, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
3. $P(\bar{A}) = 1 - P(A)$;

Example 3.13 (Probability of Getting Spades)

What is the probability that a randomly selected card from a standard 52-card deck will be a spade? There are 13 spades in the deck of 52.

Proof. Let S be the event that we get a spades. So, $P(S) = \frac{13}{52} = 0.25$. In words, the probability of pulling a spades is 0.25. □

§3.3 Conditional Probability

We will begin by building some intuition for conditional probability. Given the probability that an event B occurs, it is the case that A occurs if and only if $A \cap B$ occurs. Thus the conditional probability of A given B should be proportional to $P(A \cap B)$, which is to say that $c \cdot P(A \cap B)$ for some constant $c = c(B)$. The conditional probability of S (recall that this is our sample space) given B must equal 1, and thus $c \cdot P(S \cap B) = 1$, giving us $c = \frac{1}{P(B)}$.

Definition 3.14. If $P(B) > 0$ then the conditional probability that A occurs given that B occurs is defined to be $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

§3.4 Independence

In general, the occurrence of some event B changes the probability that another event A occurs, the original probability $P(A)$ being replaced by $P(A|B)$. If the probability remains unchanged, that is to say $P(A|B) = P(A)$, then we call A and B independent

Definition 3.15. Events A and B are called independent if $P(A \cap B) = P(A)P(B)$. We can generalize this for n events as well. If there n events $A_1, \ldots, A_n$ from n independent processes, then the probability they all occur is: $P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)$.

Example 3.16

Choose a card at random from a pack of 52 2 playing cards, each being picked with equal probability $\frac{1}{52}$. We claim that the suit of the chosen card is independent of its rank. For example, $P(\text{king}) = 4/52$, $P(\text{king}|\text{spade}) = 1/13$. Or we can say, $P(\text{spade king}) = 1/52 = 1/13 \cdot 1/4 = P(\text{king})P(\text{spade})$.

Definition 3.17 (Conditional Independence)**.** Let C be an event with $P(C) > 0$. Two events A and B are called conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$.

§3.5 False Positives Example

Example 3.18 (To be or not to be ill?)

A rare disease affects one person in 10^5. A test for the disease shows positive with probability $\frac{99}{100}$ when applied to an ill person, and with probability $\frac{1}{100}$ when applied to a healthy person. What is the probability that you have the disease given that the test shows positive? Note that $+$ is the event that the test is a positive result.

Proof.

$$\begin{aligned} P(\text{ill}|+) &= \frac{P(+|\text{ill}) \cdot P(\text{ill})}{P(+|\text{ill})P(\text{ill}) + P(+ \text{ — healthy})P(\text{healthy})} \\ &= \frac{\frac{99}{100} \cdot 10^{-5}}{\frac{99}{100} \cdot 10^{-5} + \frac{1}{100}(1 - 10^{-5})} = \frac{99}{99 + 10^5 - 1}. \end{aligned}$$

The chance of being ill is rather small. Indeed it is more likely that the test was incorrect. □

§4 Introduction to Probability Distributions

Sometimes, we are not always interested in an experiment itself, but rather in some consequence of its random outcome. Most of the experiments we encounter generate outcomes that can be interpreted in terms of real numbers, such as heights of people, numbers of voters favoring various candidates, and numbers of accidents at specified intersections. These numerical outcomes, whose values can change from experiment to experiment, are called **random variables**.

Definition 4.1 (Random Variable)**.** A random variable is a real-valued function whose domain is a sample space.

Example 4.2

Let's say we have a random variable called X. Suppose X is normally distributed, then we say $X \sim N(\mu, \sigma)$ i.e. X is a random variable that follows the normal distribution with mean μ and standard deviation σ.

§4.1 Discrete Probability Distributions

Here are a few types of discrete probability distributions we will discuss:

- Bernoulli Distribution
- Geometric Distribution
- Binomial Distribution (we will also discuss the normal approximation to the binomial distribution)

§4.1.1 Bernoulli Distribution

Definition 4.3 (Bernoulli Random Variable)**.** When an individual trial in an experiment only has two possible outcomes, often labeled as success or failure, it is called a Bernoulli random variable. The mathematical framework we will build does not depend on which outcome is labeled a success and which a failure, as long as we are consistent. Bernoulli random variables are often denoted as 1 for a success and 0 for a failure. In addition to being convenient in entering data, it is also mathematically handy (see example below).

Example 4.4

Suppose we observe 10 trials: $1, 1, 1, 0, 1, 0, 0, 1, 1, 0$. Then the sample proportion, $\hat{p}$ is the sample mean of these observations:

$$\hat{p} = \frac{\#\text{of successes}}{\#\text{of trials}} = \frac{1+1+1+0+1+0+0+1+1+0}{10} = 0.6.$$

Definition 4.5 (Mean and Std. Deviation of a Bernoulli)**.** If X is a random variable that takes value 1 with probability of success p and 0 with probability $1-p$, then X is a Bernoulli random variable with mean and standard deviation:

$$\mu = p \quad \sigma = \sqrt{p(1-p)}$$

§4.1.2 Geometric Distribution

The geometric distribution is used to describe how many trials it takes to observe a success.

Definition 4.6 (Geometric Distribution)**.** If the probability of a success in one trial is p and the probability of a failure is $1-p$, then the probability of finding the first success in the n-th trial is given by $(1-p)^{n-1} \cdot p$. The mean and standard deviation, as well as variance of this "wait" times are as follows:

$$\mu = \frac{1}{p} \quad \sigma^2 = \frac{1-p}{p^2} \quad \sigma = \sqrt{\frac{1-p}{p^2}}$$

Example 4.7 (Til Heads)

Suppose we flip a fair coin repeatedly until the first head appears. Let X be the number of flips required to get the first head. Then X follows a geometric distribution with parameter $p = \frac{1}{2}$.

$$P(X = k) = (1-p)^{k-1}p = \left(\tfrac{1}{2}\right)^{k-1}\left(\tfrac{1}{2}\right), \quad k = 1, 2, 3, \ldots$$

For example:

- $P(X = 1) = \frac{1}{2}$ (a head on the first flip),
- $P(X = 2) = \frac{1}{4}$ (a tail then a head),
- $P(X = 3) = \frac{1}{8}$ (two tails then a head).

§4.1.3 Binomial Distribution

The binomial distribution is used to describe the number of successes in a fixed number of trials. This is different from the geometric distribution, which described the number of trials we must wait before we observe a success.

Definition 4.8. Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

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

Note:

- p^k comes from the probability of getting k successes.
- $(1-p)^{n-k}$ comes from the probability of getting the remaining $n-k$ failures.
- $\binom{n}{k}$ counts how many different orders of successes/failures give exactly k successes.

The mean, variance, and standard deviation of the number of observed successes are

$$\mu = np, \quad \sigma^2 = np(1-p), \quad \sigma = \sqrt{np(1-p)}.$$

Theorem 4.9 (Assumptions: Is it Binomial?)
1. The trials are independent.
2. The number of trials, n, is fixed.
3. Each trial outcome can be classified as a *success* or *failure*.
4. The probability of a success, p, is the same for each trial.

Example 4.10

Suppose we flip a fair coin $n = 5$ times. Let X be the number of heads observed. Then X follows a binomial distribution with parameters $n = 5$ and $p = \frac{1}{2}$.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, 3, 4, 5.$$

In this case:

$$P(X = k) = \binom{5}{k} \left(\tfrac{1}{2}\right)^k \left(\tfrac{1}{2}\right)^{5-k} = \binom{5}{k} \left(\tfrac{1}{2}\right)^5.$$

For example:

- $P(X = 0) = \frac{1}{32}$ (all tails),
- $P(X = 2) = \frac{10}{32} = \frac{5}{16}$ (exactly two heads),
- $P(X = 5) = \frac{1}{32}$ (all heads).

Example 4.11 (Binomial Bulbs)

A factory produces lightbulbs, and each has a 2% chance of being defective. If we test $n = 20$ bulbs, the probability that exactly $k = 1$ bulb is defective is

$$P(X = 1) = \binom{20}{1} (0.02)(0.98)^{19}.$$

§4.1.4 Normal Approximation to the Binomial Distribution

Using the binomial formal can get tedious and lengthy real fast as n (sample size) increases. In some cases we may use the normal distribution as an easier and faster way to estimate binomial probabilities.

Theorem 4.12 (Normal Approximation of the Binomial)

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1-p)$ are both at least 15. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:
$\mu = np \quad \sigma = \sqrt{np(1-p)}$

§4.2 Continuous Probability Distributions

Like before, here is a list of the few continuous probability distributions we will discuss:

- Normal Distribution
- Standard Normal Distribution

Remark 4.13. Note that we have already covered the Normal and Standard Normal Distribution in the previous sections. Please refer back if needed.

§5 Introduction to Simple Linear Regression (SLR)

Example 5.1 (Real-Life Example Using SLR: Weight with Height)
Suppose that we are interested in the average weight of undergraduate students at the University of Florida. We input each student's name (**the population**) in a random name generator and randomly select 100 names (**sample**). Then their weight is measured in kilograms. We denote the selected measurement $Y_1, Y_2, \ldots, Y_{100}$. In addition, suppose that we also measure the selected student's height in meters and denote these by $X_1, X_2, \ldots, X_{100}$.
Here are some questions for you:

1. How would you use this data to estimate the average weight of a male undergrad?
2. How would you visualize this data?

Let us build some intuition for simple linear regression with some definitions.

Definition 5.2 (Response/Dependent Variable)**.** We will denote the response variable Y, which is our variable of interest (i.e. typically what we are measuring).

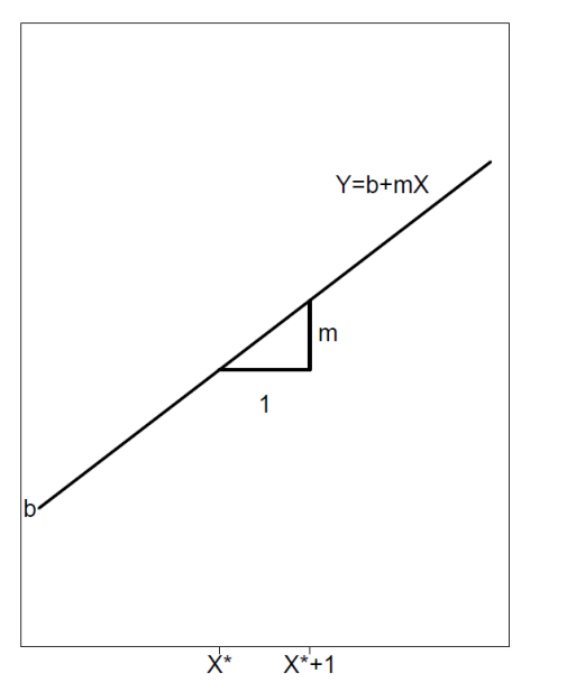
Definition 5.3 (Explanatory/Predictor Variable)**.** Let us denote the explanatory variable X, and it is also often called the independent variable (i.e. the variable we are changing).

Definition 5.4 (Discrete Variable)**.** A discrete variable has a countable sample space. These include gender, year in school, etc.

Definition 5.5 (Continuous Variable)**.** A continuous variable has an uncountable sample space. These include height, weight, length, time, etc.

Example 5.6 (Scatterplot)

A scatter plot of 100 points with the ordered pairs (*height*, *weight*) shows that there is a linear trend.



The equation of the line looks like something from the good old years of your algebra class except with different notation: $Y = \beta_0 + \beta_1 X$, where β_0 is the intercept and β_1 is the slope.

Definition 5.7 (Simple Linear Regression in a SAMPLE)**.** Simple linear regression formula for a SAMPLE: $\hat{y} = b_0 + b_1 x$

- $\hat{y}$ =predicted value of y for a given value of x.
- $b_0 = y$-intercept: The point on the y-axis where a line crosses (i.e., value of y when $x = 0$); in regression, also known as the constant.
- b_1 =slope: A measure of the direction (positive or negative) and steepness of a line; for every one unit increase in x, the change in y. For every one unit increase in x the predicted value y of increases by the value of the slope.

In a population, the y-intercept is denoted β_0 and the slope is denoted β_1.

Theorem 5.8 (Formal Statement of Simple Linear Regression)

Suppose we have observations/data $(X_1, Y_1), \ldots, (X_n, Y_n)$, where each $X_i \in \mathbb{R}$ and $Y_i \in \mathbb{R}$. The simple linear regression model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where:

- Y_i is the value of the response variable in the i-th trial,
- $\beta_0, \beta_1 \in \mathbb{R}$ are unknown regression coefficients or **unknown parameters**,
- X_i is a known **fixed** constant, namely the value of the predictor variable in the i-th trial
- ε_i are random error terms satisfying $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$,
- $\{\varepsilon_i\}$ are uncorrelated.

The method of least squares estimates $\hat{\beta}_0, \hat{\beta}_1$ by minimizing

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

which is the sum of the squares of the vertical distances from the points on the line. We would minimize this by taking partial derivatives with respect to β_0 and β_1.

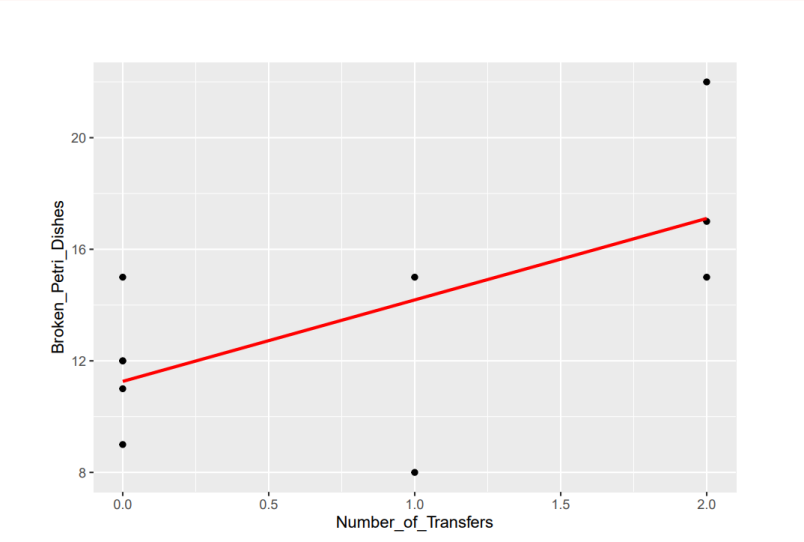
Remark 5.9. Some textbooks use slightly different notation. For example, you might in the future see the following notation used:

- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\hat{y} = a + bx \leftarrow$ I believe this is the notation used in STA2023

Definition 5.10 (Regression)**.** A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion. Here are some properties:

- There is a probability distribution of Y for each level of X.
- The means of these probability distributions vary in some systematic fashion with X.

Example 5.11 (Interpreting the Regression Coefficients for Height and Weight)
Let X be the number of times a carton of 1000 petri dishes was s transferred from one aircraft to another over the shipment route. Let Y be the the number of petri dishes found to be broken upon arrival.



The data above, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route. The estimated regression function is $\hat{y} = 11.263 + 2.921x$.
Interpretation of the Coefficients

- b_0: When the carton is not transferred at all (i.e. when $x = 0$) the model predicts about 11.3 broken dishes per 1000. It makes since to interpret the intercept since the case $x = 0$ is plausible.
- b_1: For each additional transfer, the expected number of broken dishes increases by about 2.92 per 1000, on average.

Example 5.12 (Interpreting Coefficients With Customers and Costs)
A restaurant owner wishes to model the association between mean daily costs y and the number of customers x served using a simple linear regression model. The owner collected a week of data and the values for 7 days are shown in the following table:

Day	1	2	3	4	5	6	7
Costs	1000	2180	2240	2410	2590	2820	3060
Customers	0	60	120	133	143	175	175

Our regression equation is $\hat{y} = 1191.930 + 9.872x$.

Interpretation of the Coefficients

- b_0: When there are no customers, the average daily cost is 1191.93. We can interpret the intercept since at least one observed X is equal to 0.
- b_1:For every additional customer the daily cost increases by 9.87 on average.

Remark 5.13. As we have seen so far, Simple linear regression uses data from a sample to construct the line of best fit. But what makes a line "best fit"? The most common method of constructing a regression line is the least squares method. The least squares method computes the values of the intercept and slope that make the sum of the squared residuals as small as possible.

Definition 5.14 (Residuals)**.** Residuals are often symbolized by ε or e. We will use e to not confuse it with our random error term. As with most predictions, you expect there to be some error. For example, if we are using height to predict weight, we wouldn't expect to be able to perfectly predict every individuals weight using their height. There are many variables that impact a person's weight, and height is just one of those many variables. These errors in regression predictions are called prediction error or residuals. **A residual is calculated by taking an individual's observed y value minus their corresponding predicted y value** such that

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

where y_i is the actual value for the i-th observation and $\hat{y}_i$ is the predicted value of y for the i-th observation.

Definition 5.15 (Sum of Squared Residues (SSE))**.** Given by $SSE = \sum_{i=1}^{n} (y_i - \hat{y_i}^2)$.

Formulas for Intercept and Slope

- Intercept: $b_0 = \bar{y} - b_1 \cdot \bar{x}$, where $\bar{y}$ and $\bar{x}$ are the means of x and y, respectively. **it is often easier to find the slope first and then substituting that in to find the y-intercept**
- Slope: $b_1 = r \cdot \frac{s_y}{s_x}$, where r is the Pearson's correlation coefficient between x and y, and s_x, s_y is the standard deviation of x and y (of the sample) respectively.
- Slope (Don't need to know this formula): $b_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

§5.0.1 Assumptions of Simple Linear Regression

Theorem 5.16 (Assumptions of SLR)

In order to use simple linear regression for data analysis, the following assumptions must be met.

- **Linearity**: The relationship between x and y must be linear. Check this assumption by examining the scatter plot og x and y.
- **Independence of Errors**: There is not a relationship between the residuals and the predicted values. Check this assumption by examining a scatterplot of"residuals versus fits." The correlation should be approximately 0.
- **Normality of Errors**: The residuals must be approximately normally distributed. Check this assumption by examining a normal probability plot; the observations should be near the line. You can also examine a histogram of the residuals; it should be approximately normally distributed. The distribution will not be perfectly normal because we're working with sample data and there may be some sampling error, but the distribution should not be clearly skewed.
- **Equal Variances**: The variance of the residuals should be consistent across all predicted values. Check this assumption by examining the scatterplot of "residuals versus fits." The variance of the residuals should be consistent across the x-axis. If the plot shows a pattern (e.g., bowtie or cone-like shape), then variances are not consistent and this assumption has not been met.

§5.0.2 Coefficient of Determination

Definition 5.17 (Coefficient of Determination i.e. R-squared)**.** The amount of variation in the response variable that can be explained by (i.e. accounted for) the explanatory variable is denoted by R^2.

Example 5.18 (Kilometers of a Car and Car Prices)

Suppose we have some data of the kilo-mileage of cars and the price of each car. I did some data analysis on R and found that the R-squared value is 37.04%. This means that 37.04% of the variation in cars prices can be explained by the kilo-mileage of the cards.

S	R-sq	R-sq(adj)	R-sq(pred)
9.71152	37.04%	35.73%	29.82%

Remark 5.19. Observe that the coefficient of determination is equal to the correlation coefficient squared. In other words, $R^2 = r^2$. If you want r, take the square root of R^2. See the following example for more clarity.

Example 5.20 (R-squared and r)
The correlation between quiz averages and final exam scores was $r = 0.608630$. So, the coefficient of determination (R^2) is: $R^2 = (0.608630)^2 = 0.3704$.

§5.0.3 Cautions of Simple Linear Regression

I will just list a few here.

- Influence of Outliers
- Extrapolation: a regression equation should not be used to make predictions for values that are far from those that were used to construct the model or for those that come from a different population.
- Interpretation of Causation (**very bad!**)

§6 Confidence Intervals

Definition 6.1 (Confidence Interval)**.** A confidence interval is a range (with a lower and upper limit) computed using sample statistics to estimate an unknown population parameter with a stated level of confidence (for example, 90%, 95%, ...)

§6.1 Confidence Intervals for Proportions

Now, let's build some intuition for constructing a confidence interval (CI). At the center of a confidence interval is the sample statistic, such as a sample mean or sample proportion (the thing we want). This is known as the point estimate or estimator (if you would like to call it as such). The width of the confidence interval is determined by the margin of error. The margin of error is the amount that is subtracted from and added to the point estimate to construct the confidence interval. Here's a general formula to remember when you are asked to construct a CI:

$$\textbf{point estimator} \pm \textbf{margin of error}$$

The margin of error can be further broken down into a multiplier (the level of confidence that you are given determine for the multiplier. For example, at 95% confidence, I have a multiplier of 1.96) times the standard error as follows:

Let k be the multiplier and SE be shorthand for the standard error. Then, the
$$\text{margin of error} = \text{k} \cdot \text{SE}$$

Remark 6.2. Now that we've established what confidence intervals are and their general purpose in statistical inference, we need to examine the foundational assumptions that make them valid and reliable. Before we can construct a confidence interval for a population proportion, we must verify that our data and sampling method satisfy certain key conditions. The most critical of these are the assumptions underlying the Central Limit Theorem and the success-failure condition, which together ensure that our sampling distribution will be approximately normal and that our confidence interval will provide accurate coverage of the true population parameter.

Theorem 6.3 (The Central Limit Theorem Again)
When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with the following mean and standard error: $\mu_{\hat{p}} = p \quad \text{standard error}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 15$ and $n(1-p) \geq 15$, which is called the success-failure condition.

Theorem 6.4 (How to verify sample observations are independent?)

- Subjects in an experiment are considered independent if they undergo random assignment to the treatment groups.
- If the observations are from a simple random sample, then they are independent.
- If a sample is from a seemingly random process, e.g. an occasional error on an assembly line, checking independence is more difficult. In this case, use your best judgement.

[**Constructing a 95% Interval**]
Our sample proportion $\hat{p}$ the most plausible value of the population proportion, so it makes sense to build a confidence interval around this point estimate. The standard error provides a guide for how large we should make the confidence interval. The standard error represents the standard deviation of the point estimate, and when the Central Limit Theorem conditions are satisfied, the point estimate closely follows a normal distribution.In a normal distribution, 95% of the data is within 1.96 standard deviations of the mean. Using this principle, we can construct a confidence interval:

$$\text{point estimate} \pm 1.96 \cdot \text{standard error}$$

$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}$$

[**Summary**]
Once you've determined a one-proportion confidence interval would be helpful for an application, there are steps to constructing the interval:

1. Prepare: Identify $\hat{p}$ and n, and determine what confidence level you wish to use or what is asked of you.
2. Verify the conditions to ensure $\hat{p}$ is nearly normal.(see above for more information).
3. If the conditions hold, compute the standard error using the formula above, and find the z-score that corresponds to the given confidence level.
4. Interpret the confidence interval in the context of the problem.

§6.2 Confidence Intervals for Means

Now instead of proportions, we will focus on means ($\bar{x}$). In this case, the sample mean ($\bar{x}$) tends to follow a normal distribution centered at the population mean, μ.

Theorem 6.5 (Central Limit Theorem for Sample Mean)
When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ will be nearly normal with:

$$\bar{x} = \mu \quad \text{standard error}_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Theorem 6.6 (Conditions Required for Modeling $\bar{x}$)
There are two conditions.

1. **Independence**: The sample observations must be independent, The most common way to satisfy this condition is when the sample is a simple random sample from the population
2. **Normality**: When a sample is small, we also require that the sample observations come from a normally distributed population. We can relax this condition more and more for larger and larger sample sizes. There is no perfect check for normality, but a good rule of thumb is:
   - If $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.
   - If $n \geq 30$: If the sample size n is at least 30 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

§6.2.1 Introduction to the t-distribution

In practice, we cannot directly calculate the standard error for $\bar{x}$ since we do not know the population standard deviation, σ. We encountered a similar issue when computing the standard error for a sample proportion, which relied on the population proportion, p. Our solution in the proportion context was to use sample value in place of the population value when computing the standard error. We'll employ a similar strategy for computing the standard error of $\bar{x}$, using the sample standard deviation s in place of σ:

$$SE = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

This strategy tends to work well when we have a lot of data and can estimate σ using s accurately. However, the estimate is less precise with smaller samples, and this leads to problems when using the normal distribution to model $\bar{x}$.

Definition 6.7 (The t-distribution)**.** The t-distribution, also called Student's t-distribution, is a family of distributions that are symmetric and bell-shaped, similar to the normal distribution, but with heavier tails. The exact shape of a t-distribution depends on the degrees of freedom, and as the degrees of freedom increase, the t-distribution approaches the standard normal distribution.

Definition 6.8 (Degrees of Freedom)**.** Degrees of freedom (df) represent the number of independent pieces of information available to estimate a parameter. For a single sample mean, the degrees of freedom are $df = n - 1$, where n is the sample size. We lose one degree of freedom because we use the sample mean $\bar{x}$ to calculate the sample standard deviation s.

When the sample size is small and we don't know the population standard deviation σ, the sampling distribution of the standardized sample mean follows a t-distribution rather than a normal distribution:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

where T follows a t-distribution with $n - 1$ degrees of freedom.

Theorem 6.9 (Properties of the t-distribution)

- Symmetric and bell-shaped, centered at 0
- Heavier tails than the normal distribution (more spread out)
- As degrees of freedom increase, the t-distribution approaches the standard normal distribution
- When $df \geq 30$, the t-distribution is nearly identical to the standard normal distribution
- The variance of a t-distribution is $\frac{df}{df-2}$ for $df > 2$

Example 6.10 (Comparing t-distributions with Different Degrees of Freedom)
Consider three t-distributions with different degrees of freedom:

- t_1: 1 degree of freedom (very heavy tails)
- t_9: 9 degrees of freedom (moderate tails)
- t_{30}: 30 degrees of freedom (nearly normal)

As the degrees of freedom increase, the distribution becomes more concentrated around the mean and approaches the standard normal distribution. The critical values (like $t_{0.025}$) decrease as degrees of freedom increase, approaching the corresponding z-scores.

§6.3 Finding a t-Confidence Interval for the Mean

[**Constructing a t-Confidence Interval for μ**] When we don't know the population standard deviation σ, we use the t-distribution to construct confidence intervals for the population mean. The general form is:

$$\bar{x} \pm t_{\alpha/2,df} \cdot \frac{s}{\sqrt{n}}$$

where:

- $\bar{x}$ is the sample mean (point estimate)
- $t_{\alpha/2,df}$ is the critical t-value with $df = n - 1$ degrees of freedom
- $\frac{s}{\sqrt{n}}$ is the estimated standard error
- $\alpha = 1 -$ confidence level

[**Steps for Constructing a One-Mean t-Confidence Interval**]

1. **Prepare**: Identify $\bar{x}$, s, n, and the desired confidence level.
2. **Check Conditions**:
   - *Independence*: Sample observations must be independent (random sampling)
   - *Normality*:
     - If $n < 30$: Data should come from approximately normal population with no clear outliers
     - If $n \geq 30$: Central Limit Theorem applies; normality condition is relaxed unless there are extreme outliers
3. **Calculate**:
   - Determine degrees of freedom: $df = n - 1$
   - Find the critical t-value: $t_{\alpha/2,df}$ using a t-table or calculator
   - Calculate the standard error: $SE = \frac{s}{\sqrt{n}}$
   - Compute the margin of error: $ME = t_{\alpha/2,df} \cdot SE$
   - Construct the interval: $\bar{x} \pm ME$
4. **Interpret**: State the confidence interval in context of the problem.

Example 6.11 (Constructing a 95% t-Confidence Interval)
A random sample of 16 students took a statistics exam, and their scores had a mean of $\bar{x} = 78.5$ and standard deviation of $s = 12.3$. Construct a 95% confidence interval for the population mean exam score.
Solution:

1. **Given**: $n = 16$, $\bar{x} = 78.5$, $s = 12.3$, confidence level = 95%
2. **Check Conditions**:
   - Independence: Assume random sample
   - Normality: $n = 16 < 30$, so we assume exam scores are approximately normal
3. **Calculate**:
   - $df = n - 1 = 16 - 1 = 15$
   - For 95% confidence: $\alpha = 0.05$, so $\alpha/2 = 0.025$
   - $t_{0.025,15} = 2.131$ (from t-table)
   - $SE = \frac{s}{\sqrt{n}} = \frac{12.3}{\sqrt{16}} = \frac{12.3}{4} = 3.075$
   - $ME = 2.131 \times 3.075 = 6.55$
   - Confidence interval: $78.5 \pm 6.55 = (71.95, 85.05)$
4. **Interpret**: We are 95% confident that the interval (71.95, 85.05) captures the true mean exam score for all students in the population.

Remark 6.12. Notice that the t-critical value (2.131) is larger than the corresponding z-critical value (1.96) for 95% confidence. This is because the t-distribution has heavier tails, accounting for the additional uncertainty introduced by estimating σ with s. As the sample size increases, the t-critical values approach the z-critical values.

§6.4 Interpreting Confidence Intervals

Template: We are 95% confident that the [input interval here] captures the true mean/proportion of [insert parameter here in words] (based on what the problem deals with) for [insert population here] [add some context].
It is very important you include all the necessary components in a confidence interval: parameter, population, confidence level, interval, and a statement whether your data deals with means or proportions.
There will be further interpretations that can be done with confidence intervals later when we talk about hypothesis testing.

§6.5 Effect of Sample Size on Confidence Intervals

Definition 6.13 (Adjusting Sample Size)**.** As n (sample size) increases, the width of the confidence interval gets narrower because the standard error decreases. Recall that standard error is proportional to $1/\sqrt{n}$.

Definition 6.14 (Adjusting Confidence Levels)**.** As the confidence level increases (i.e 90, 95, 99), the width of the confidence interval gets wider because a higher confidence level requires a larger critical value (from the z- or t-distribution). While you're more confident that the interval contains the true parameter, you're also allowing for a broader range of values to ensure that higher level of certainty.

§6.6 Bootstrapping/Bootstrap Confidence Intervals

Definition 6.15 (Bootstrapping Sampling)**.** Bootstrapping is a resampling procedure that uses data from one sample (with replacement) to generate a sampling distribution by repeatedly taking random samples from the known sample.

[**Bootstrap Procedure**]

1. Take a random sample of size n with replacement from the original sample
2. Calculate the statistic of interest (mean, proportion, etc.) for this bootstrap sample
3. Repeat steps 1-2 many times (typically 1,000 or more)
4. Use the distribution of bootstrap statistics to construct confidence intervals

Once we have a bootstrap sampling distribution there are two methods for constructing a confidence interval:

1. The standard deviation of the bootstrap distribution is the standard error which we can use to construct a bootstrap confidence interval. Recall that for a 95% confidence interval, given that the sampling distribution is approximately normal, the 95% confidence interval will be:
sample statistic $\pm$ 2 $\cdot$ standard error
2. For a 95% confidence interval we can find the middle 95% bootstrap statistics. This is known as the percentile method. This is the preferred method because it works regardless of the shape of the sampling distribution.

Theorem 6.16 (Bootstrap Confidence Interval Methods)
Standard Error Method: Use the standard deviation of bootstrap statistics as the standard error: CI = original statistic $\pm\, t_{\alpha/2} \cdot SE_{bootstrap}$
Percentile Method (Preferred): Use the appropriate percentiles from the bootstrap distribution: 95% CI = (2.5th percentile, 97.5th percentile)

Example 6.17 (Bootstrap Confidence Interval)
Original sample of exam scores: 78, 82, 76, 85, 79, 81, 77, 83 Sample mean: $\bar{x} = 80.125$ After 1,000 bootstrap samples, suppose the bootstrap means have:

- 2.5th percentile: 77.8
- 97.5th percentile: 82.4

95% Bootstrap CI: (77.8, 82.4)
Interpretation: We are 95% confident that the true population mean exam score is between 77.8 and 82.4.

§6.7 Paired Samples

Definition 6.18 (Paired Data)**.** Paired data occurs when each observation in one group is naturally matched or paired with an observation in another group. Common examples include:

- Before-and-after measurements on the same subjects
- Measurements on twins or matched pairs
- Left vs. right measurements on the same individuals

Theorem 6.19 (Paired t-Test Procedure)

For paired data, we analyze the differences $d = x_1 - x_2$ and test:

- $H_0 : \mu_d = 0$ (no difference)
- $H_A : \mu_d \neq 0$ (there is a difference)

The test statistic is: $t = \frac{\bar{d}-0}{s_d/\sqrt{n}}$ with $df = n - 1$
The confidence interval for μ_d is: $\bar{d} \pm t_{\alpha/2,n-1} \cdot \frac{s_d}{\sqrt{n}}$

Example 6.20 (Paired Sample Analysis)

A study measures blood pressure before and after a new exercise program for 12 participants:

Participant	Before	After	Difference
1	145	142	3
2	150	144	6
⋮	⋮	⋮	⋮

If $\bar{d} = 4.2$, $s_d = 3.1$, and $n = 12$:
95% CI for mean difference: $4.2 \pm 2.201 \cdot \frac{3.1}{\sqrt{12}} = 4.2 \pm 1.97 = (2.23, 6.17)$
Interpretation: We are 95% confident that the exercise program reduces blood pressure by between 2.23 and 6.17 mmHg on average.

Remark 6.21. Paired analysis is generally more powerful than comparing two independent groups because it controls for individual variation between subjects. The pairing removes much of the variability that would otherwise obscure the treatment effect.

§7 Hypothesis Testing

§7.1 Introduction to Hypothesis Testing

Previously we used confidence intervals to estimate unknown population parameters. We compared confidence intervals to specified parameter values and when the specific value was contained in the interval, we concluded that there was not sufficient evidence of a difference between the population parameter and the specified value. In other words, any values within the confidence intervals were reasonable estimates of the population parameter and any values outside of the confidence intervals were not reasonable estimates. Here, we are going to look at a more formal method for testing whether a given value is a reasonable value of a population parameter. To do this we need to have a hypothesized value of the population parameter.

Definition 7.1 (Null Hypothesis)**.** The null hypothesis (H_0) is the statement that there is no difference in the population(s). It represents the status quo or the claim we are testing. The null hypothesis always includes an equality ($=$, $\leq$, or $\geq$).

Definition 7.2 (Alternative Hypothesis)**.** The alternative hypothesis (H_A or H_1) is the statement that there is some difference in the population(s). It represents what we are trying to find evidence for and is the complement of the null hypothesis.

§7.2 Writing Hypotheses

[**Steps for Writing Hypotheses**] When writing hypotheses, there are three key components to identify:

1. **Parameter**: What population parameter are we testing? (μ, p, $\mu_1 - \mu_2$, etc.)
2. **Direction**: Is the test one-tailed or two-tailed?
   - Two-tailed: "different from," "not equal to" $\Rightarrow H_A : \theta \neq \theta_0$
   - Right-tailed: "greater than," "more than" $\Rightarrow H_A : \theta > \theta_0$
   - Left-tailed: "less than," "fewer than" $\Rightarrow H_A : \theta < \theta_0$
3. **Hypothesized Value**: What specific value are we testing against?

Theorem 7.3 (Common Hypothesis Formats)

Test Type	Two-tailed	Right-tailed	Left-tailed
Single Mean	$H_0 : \mu = \mu_0$ $H_A : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_A : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_A : \mu < \mu_0$
Single Proportion	$H_0 : p = p_0$ $H_A : p \neq p_0$	$H_0 : p \leq p_0$ $H_A : p > p_0$	$H_0 : p \geq p_0$ $H_A : p < p_0$
Difference in Means	$H_0 : \mu_1 - \mu_2 = 0$ $H_A : \mu_1 - \mu_2 \neq 0$	$H_0 : \mu_1 - \mu_2 \leq 0$ $H_A : \mu_1 - \mu_2 > 0$	$H_0 : \mu_1 - \mu_2 \geq 0$ $H_A : \mu_1 - \mu_2 < 0$
Paired Differences	$H_0 : \mu_d = 0$ $H_A : \mu_d \neq 0$	$H_0 : \mu_d \leq 0$ $H_A : \mu_d > 0$	$H_0 : \mu_d \geq 0$ $H_A : \mu_d < 0$

Example 7.4 (Writing Hypotheses)

Scenario: A manufacturer claims that their light bulbs last an average of 1000 hours. A consumer group wants to test if the bulbs last significantly less than advertised.
Solution:

- Parameter: μ (population mean lifetime)
- Direction: Left-tailed (testing if mean is less than 1000)
- Hypothesized value: 1000 hours

Hypotheses:

$$H_0 : \mu \geq 1000 \text{ hours} \quad (1)$$
$$H_A : \mu < 1000 \text{ hours} \quad (2)$$

§7.3 Randomization Procedures

Definition 7.5 (Randomization Test)**.** A randomization test (also called a permutation test) is a statistical hypothesis testing method that uses random reassignment of observed data to generate a null distribution. This approach does not rely on theoretical distributions but instead creates the sampling distribution through simulation.

Theorem 7.6 (Logic of Randomization Tests)

The fundamental principle behind randomization tests:

1. If the null hypothesis is true, then group labels are essentially meaningless
2. We can randomly shuffle/reassign the group labels many times
3. Each randomization gives us one possible outcome under the null hypothesis
4. The collection of these outcomes forms our null distribution
5. We compare our observed test statistic to this null distribution

[**Randomization Test Procedure**]

1. **Calculate the observed test statistic** from the original data
2. **Assume H_0 is true** and randomly reassign group labels (or shuffle data)
3. **Calculate the test statistic** for this randomized dataset
4. **Repeat steps 2-3** many times (typically 1,000 to 10,000 times)
5. **Create a histogram** of the randomized test statistics (null distribution)
6. **Find the p-value** by determining what proportion of randomized statistics are as extreme or more extreme than the observed statistic

Example 7.7 (Two-Sample Randomization Test)
Scenario: Testing if a new drug reduces blood pressure more than a placebo.
Data: Drug group (n=8): 12, 15, 9, 6, 11, 8, 13, 10; Placebo group (n=7): 18, 20, 16, 22, 19, 17, 21
Observed difference: $\bar{x}_{drug} - \bar{x}_{placebo} = 10.5 - 19.0 = -8.5$
Randomization: Randomly reassign the 15 values to two groups (8 and 7), calculate the difference, repeat 10,000 times.
If only 23 out of 10,000 randomizations produce differences $\leq$ -8.5, then p-value = 0.0023.

§7.4 p-Values

Definition 7.8 (P-Value)**.** Given that the null hypothesis is true, the p-value is the probability of obtaining a sample statistic as extreme or more extreme than the one observed in the sample, in the direction of the alternative hypothesis.

$$\text{p-value} = P(\text{test statistic as extreme or more extreme} | H_0 \text{ is true})$$

Theorem 7.9 (Interpreting P-Values by Size)
- $p > 0.10$: Little or no evidence against H_0
- $0.05 < p \leq 0.10$: Weak evidence against H_0
- $0.01 < p \leq 0.05$: Moderate evidence against H_0
- $0.001 < p \leq 0.01$: Strong evidence against H_0
- $p \leq 0.001$: Very strong evidence against H_0

Remark 7.10 (Common P-Value Misconceptions)**. What p-values ARE**:
- The probability of our data (or more extreme) given H_0 is true
- A measure of evidence against H_0

What p-values are NOT:
- The probability that H_0 is true
- The probability that the results are due to chance
- The probability of making an error

Example 7.11 (P-Value Interpretation)
In a clinical trial, we test whether a new medication is more effective than current treatment. We obtain p = 0.03.
Correct interpretation: "If the new medication were no more effective than current treatment, there would be only a 3% chance of observing a difference as large as (or larger than) what we observed."
Incorrect interpretation: "There's a 3% chance the null hypothesis is true" or "There's a 97% chance the new medication is better."

§7.5 Randomization Test Examples in StatKey

Definition 7.12 (StatKey)**.** StatKey is an online statistical software tool that allows users to perform randomization tests, bootstrap confidence intervals, and theoretical probability calculations through an intuitive web interface.

[**Using StatKey for Randomization Tests**] **Steps for a Two-Sample Mean Test**:

1. Go to StatKey → Randomization Tests → Difference in Means
2. Upload your data or use built-in datasets
3. Click "Show Original Sample" to see observed difference
4. Click "Generate 1000 Samples" to create null distribution
5. Observe the histogram and p-value calculation
6. Use "Left Tail," "Right Tail," or "Two Tail" based on your H_A

Example 7.13 (StatKey Randomization Example)

Research Question: Do male and female students differ in average hours of sleep per night?
Data: Males: 7.2, 6.8, 7.5, 6.9, 7.1; Females: 8.1, 7.8, 8.3, 7.9, 8.0
StatKey Steps:

1. Enter data in StatKey's Difference in Means tool
2. Observed difference: $\bar{x}_M - \bar{x}_F = 7.1 - 8.02 = -0.92$
3. Generate 5,000 randomizations
4. Two-tailed p-value ≈ 0.008

Conclusion: Strong evidence that male and female students differ in average sleep hours ($p = 0.008$).

§7.6 Summary

[**Key Concepts Summary**] **Hypothesis Testing Framework**:

- State H_0 (status quo) and H_A (what we're testing for)
- Choose significance level α before collecting data
- Calculate test statistic from sample data
- Find p-value: probability of observing our result (or more extreme) if H_0 is true
- Make decision: Reject H_0 if p-value $\leq \alpha$

Two Main Approaches:

- **Theoretical**: Use known distributions (normal, t, etc.)
- **Simulation**: Use randomization/permutation tests

Decision Making:

- p-value $\leq \alpha$: Reject H_0 (statistically significant)
- p-value $> \alpha$: Fail to reject H_0 (not statistically significant)

§8 Hypothesis Testing, Part 2

§8.1 Type I and Type II Errors

Definition 8.1 (Type I Error)**.** A Type I error occurs when we reject a true null hypothesis. The probability of making a Type I error is denoted by α (the significance level):

$$P(\text{Type I Error}) = P(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha$$

Definition 8.2 (Type II Error)**.** A Type II error occurs when we fail to reject a false null hypothesis. The probability of making a Type II error is denoted by β:

$$P(\text{Type II Error}) = P(\text{Fail to reject } H_0 | H_0 \text{ is false}) = \beta$$

Theorem 8.3 (Error Types Summary Table)

Decision	H_0 **is True**	H_0 **is False**
Reject H_0	Type I Error (α)	Correct Decision
Fail to Reject H_0	Correct Decision	Type II Error (β)

Example 8.4 (Medical Testing Context)
Consider a medical test for a disease where:

- H_0: Patient does not have the disease
- H_A: Patient has the disease

Type I Error (False Positive): Test indicates the patient has the disease when they don't. This could lead to unnecessary stress, treatment, and medical costs.
Type II Error (False Negative): Test indicates the patient doesn't have the disease when they do. This could delay necessary treatment and worsen health outcomes.

§8.2 Significance Levels

Definition 8.5 (Significance Level)**.** The significance level α is the probability threshold below which we reject the null hypothesis. It represents the maximum probability of Type I error we are willing to accept.
Common significance levels:

- $\alpha = 0.05$ (5%): Most common in social sciences and general research
- $\alpha = 0.01$ (1%): More stringent, used when Type I errors are costly
- $\alpha = 0.10$ (10%): Less stringent, used in exploratory research

Remark 8.6 (Trade-offs in Choosing α)**.** Choosing a smaller significance level (like 0.01 instead of 0.05) makes it harder to reject the null hypothesis:

- **Benefit**: Reduces the chance of Type I errors
- **Cost**: Potentially increases Type II errors (reduces power)

The choice of α should reflect the relative costs of each type of error in the specific context.

§8.3 Practical Significance

Definition 8.7 (Practical Significance)**.** Practical significance refers to whether an observed effect is large enough to be meaningful in real-world applications, regardless of statistical significance. An effect can be statistically significant but not practically significant, especially with large sample sizes.

Remark 8.8 (Statistical vs. Practical Significance)**.** Here is a summary of their differences:

- **Statistical significance**: Tells us the effect is likely real (not due to chance)
- **Practical significance**: Tells us the effect is large enough to matter

With very large sample sizes, even tiny, practically meaningless differences can become statistically significant. Always consider both types of significance when interpreting results.

§8.4 Power

Definition 8.9 (Statistical Power)**.** Power is the probability of correctly rejecting a false null hypothesis. It represents the test's ability to detect a real effect when one exists:

$$\text{Power} = 1 - \beta = P(\text{Reject } H_0 | H_0 \text{ is false})$$

Theorem 8.10 (Factors Affecting Power)

Statistical power is influenced by four main factors:

1. **Effect Size**: Larger true differences from H_0 increase power
2. **Sample Size**: Larger samples increase power
3. **Significance Level**: Higher α increases power (but increases Type I error risk)
4. **Population Variability**: Lower σ increases power

Example 8.11 (Power Calculation)

A researcher wants to detect a 10-point increase in test scores (current mean = 75).
Scenario Analysis:

- With n = 20, $\sigma = 15$, $\alpha = 0.05$: Power $\approx 60\%$
- With n = 50, $\sigma = 15$, $\alpha = 0.05$: Power $\approx 90\%$
- With n = 20, $\sigma = 10$, $\alpha = 0.05$: Power $\approx 80\%$

Interpretation: Increasing sample size or reducing variability substantially improves our ability to detect the effect.

[**Improving Statistical Power**]

1. **Increase sample size**: Most common and reliable method
2. **Reduce measurement error**: Use more precise instruments
3. **Control for confounding variables**: Use blocking or matching
4. **Use repeated measures**: Reduce individual variation
5. **Increase effect size**: Stronger treatments or interventions
6. **Use one-tailed tests**: When direction is clearly predicted

§8.5 Confidence Intervals & Hypothesis Testing

Theorem 8.12 (Duality Between Hypothesis Tests and Confidence Intervals)

For two-tailed hypothesis tests, there is a direct relationship with confidence intervals:

- If the hypothesized value falls **within** the confidence interval $\Rightarrow$ fail to reject H_0
- If the hypothesized value falls **outside** the confidence interval $\Rightarrow$ reject H_0

The confidence level corresponds to $1 - \alpha$:

- $\alpha = 0.05 \Leftrightarrow 95\%$ confidence interval
- $\alpha = 0.01 \Leftrightarrow 99\%$ confidence interval
- $\alpha = 0.10 \Leftrightarrow 90\%$ confidence interval

Example 8.13 (CI and Hypothesis Test Connection)

Testing $H_0 : \mu = 50$ vs. $H_A : \mu \neq 50$ with $\alpha = 0.05$.
95% CI for μ: (48.2, 53.7)
Decision: Since 50 falls within (48.2, 53.7), we fail to reject H_0.
Alternative scenario: If 95% CI was (52.1, 57.8), then since 50 falls outside this interval, we would reject H_0.

Remark 8.14 (Advantages of Confidence Intervals)**.** Confidence intervals provide more information than hypothesis tests alone:

- Show the range of plausible parameter values
- Indicate the precision of the estimate
- Allow assessment of practical significance
- Work for any hypothesized value, not just the one being tested

Example 8.15 (CI Providing Additional Insight)

Research Question: Is the mean customer satisfaction score different from 7.0?
Sample Results: $\bar{x} = 7.2$, 95% CI: (6.8, 7.6)
Hypothesis Test: Since 7.0 falls within (6.8, 7.6), we fail to reject $H_0 : \mu = 7.0$.
Additional Insights from CI:

- The true mean could plausibly be anywhere from 6.8 to 7.6
- Values below 6.8 or above 7.6 are unlikely
- The estimate is relatively precise (narrow interval)
- Even if significantly different from 7.0, the practical difference would be small