INTRO STATS WORKBOOK

Dr Mostafa

2025-03-05

Contents

About The Workbook

This is a *worksheet* written in **Markdown**. It shows the question and answers of the MATH224 workbook.

Table of Contents

Chapter 1

Introduction to Data

1.1 Data Types

1.1.1 Objectives

By the end of this section, students will be able to:

- Understand the importance of statistical methods for answering research questions using data.
- Identify different types of data that can be analyzed using statistical methods.
- Describe basic sampling principles and strategies for the purpose of collecting data for research studies.
- Describe basic principles of designing research experiments.

1.1.2 Overview

In this section, we will delve deeper into the categorization of variables as **numerical and categorical**. This is an important step, as the type of variable helps us determine what summary statistics to calculate, what type of visualizations to make, and what statistical method will be appropriate to answer the research questions we're exploring.

There are two types of variables: numerical and categorical.

- **Numerical**, in other words, quantitative, variables take on numerical values. It is sensible to add, subtract, take averages, and so on, with these values.

- **Categorical**, or qualitative, variables, take on a limited number of distinct categories. These categories can be identified with numbers, for example, it is customary to see likert variables (strongly agree to strongly disagree) coded as 1 through 5, but it wouldn't be sensible to do arithmetic operations with these values. They are merely placeholders for the levels of the categorical variable.

Numerical data

Numerical variables can be further categorized as **continuous or discrete**.

- **Continuous numerical** variables are usually measured, such as height. These variables can take on an infinite number of values within a given range.
- **Discrete numerical** variables are those that take on one of a specific set of numeric values where we are able to count or enumerate all of the possibilities. One example of a discrete variable is number of pets in a household. In general, count data are an example of discrete variables.

When determining whether a numerical variable is continuous or discrete, it is important to think about the nature of the variable and not just the observed value, as rounding of continuous variables can make them appear to be discrete. For example, height is a continuous variable, however we tend to report our height rounded to the nearest unit of measure, like inches or centimeters.

Categorical data

Categorical variables that have ordered levels are called **ordinal**.

Think about a survey question where you're asked how satisfied you are with the customer service you received and the options are very unsatisfied, unsatisfied, neutral, satisfied, and very satisfied. These levels have an inherent ordering, hence the variable would be called ordinal.

If the levels of a categorical variable do not have an inherent ordering to them, then the variable is simply called categorical. For example, do you consume caffeine or not?

Data collection principles

Population versus Sample: In statistics, we almost always want to apply generalizations from a small sample to a large population – you might think of this as a sort of *stereotyping.* The trick here is that for our assertions (generalizations) to be valid, our sample must be *representative of* our population.

The following takeaway is critical: *Results based off of a sample may only be generalized to a population for which that sample is representative.*

Why not take a census?

First, taking a census requires a lot more resources than collecting data from a sample of the population.

Second, certain individuals in your population might be hard to locate or collect data from. If these individuals that are missed in the census are different from those in the rest of the population, the census data will be biased. For example, in the US census, undocumented immigrants are often not recorded properly since they tend to be reluctant to fill out census forms with the concern that this information could be shared with immigration. However, these individuals might have characteristics different than the rest of the population and hence, not getting information from them might result in unreliable data from geographical regions with high concentrations of undocumented immigrants.

Lastly, populations are constantly changing. Even if you do have the required resources and manage to collect data from everyone in the population, tomorrow your population will be different and so the hard work required to collect such data may not pay off.

If you think about it, sampling is actually quite natural.

Sampling is natural

Think about something you are cooking we taste or in other words examine a small part of what we're cooking to get an idea about the dish as a whole. After all, we would never eat a whole pot of soup just to check its taste.

When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, what you're doing is simply exploratory analysis for the sample at hand.

If you then generalize and conclude that your entire soup needs salt, that's making an inference.

For your inference to be valid, the spoonful you tasted, your sample, needs to be representative of the entire pot, your population.

If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not going to be representative of the whole pot.

On the other hand, if you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.

Sampling data is a bit different than sampling soup though.

Steps to Sampling:

1) Identify the research question (then determine the population)

2) Collect data that are reliable and help achieve the research goal (take good samples)

- Population and sample
  - A population is the entire group that you want to draw conclusions about.
  - A sample is the specific group that you will collect data from
- Parameter and Statistic
  - A descriptive measure (for example, average, median, standard deviation and percentages) for an entire population is a ’‘**parameter.**’’
  - A descriptive measure for a sample is referred to as a ’‘**sample statistic**’’
- Observational studies and Experiments
  - Observational studies: research processes where researchers collect data in a way that does not directly interfere with how the data arise (examine something without manipulating it)
  - Experiment: Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables

Four commonly used random sampling techniques: -

```
1.  Simple random sampling
```

-

```
2.  Stratified sample
```

-

```
3.  Cluster sampling
```

-

```
4.  Multistage sampling
```

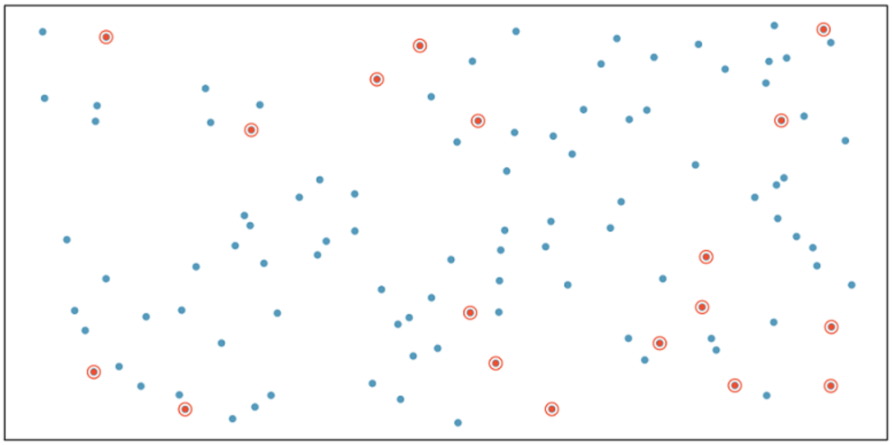
So next, we’ll introduce a few commonly used sampling methods: simple random sampling, stratified sampling, cluster sampling, and multistage sampling.

Sampling Methods

Here we discuss some of the different ways to draw a sample from a population.

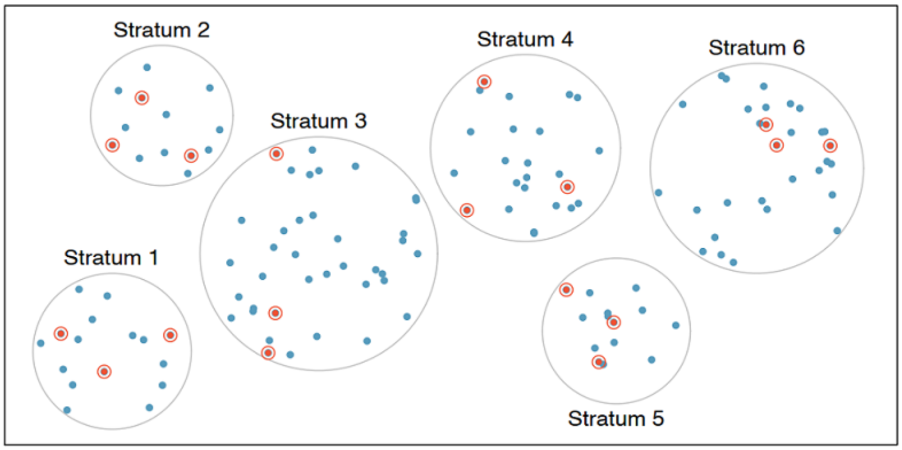
Simple random sample

In *simple random* sampling, we randomly select cases from the population, such that each case is equally likely to be selected. This is similar to randomly drawing names from a hat.



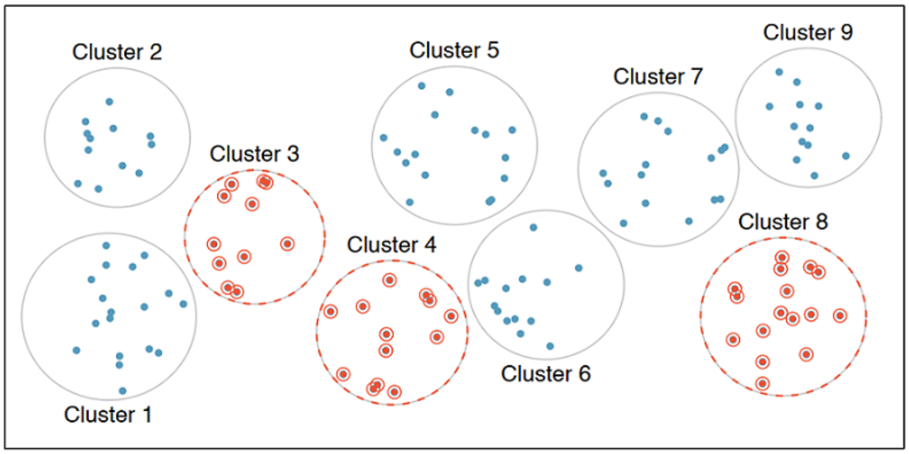
Stratified sample

In *stratified sampling*, we first divide the population into homogeneous groups, called strata, and then we randomly sample from within each stratum. For example, if we wanted to make sure that people from low, medium, and high socioeconomic status are equally represented in a study, we would first divide our population into three groups as such and then sample from within each group.



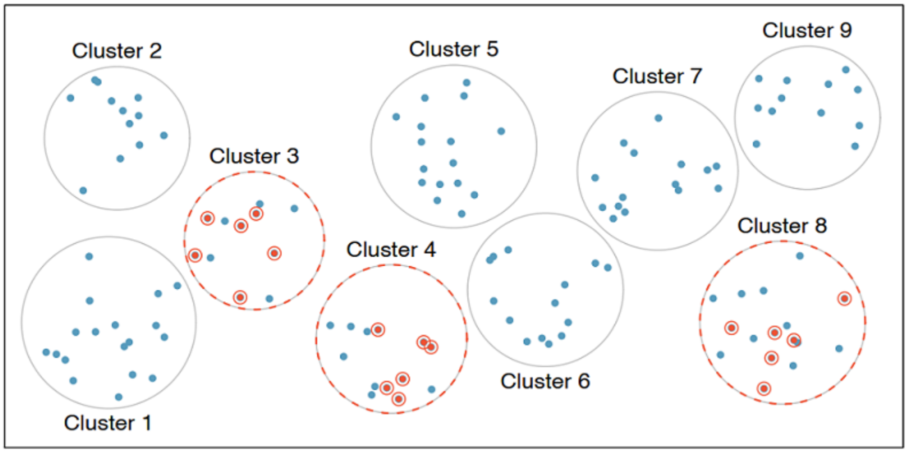
Cluster sample

In cluster sampling, we divide the population into clusters, randomly sample a few clusters, and then sample all observations within these clusters. The clusters, unlike strata in stratified sampling, are heterogeneous within themselves and each cluster is similar to the others, such that we can get away with sampling from just a few of the clusters.



Multistage sample

Multistage sampling adds another step to cluster sampling. Just like in cluster sampling, we divide the population into clusters, randomly sample a few clusters, and then we randomly sample observations from within those clusters.



Note: Cluster and multistage sampling are often used for economical reasons. For example, one might divide a city into geographic regions that are on average similar to each other and then sample randomly from a few randomly picked regions in order to avoid traveling to all regions.

Convenience sample

The *convenience sample* is the most commonly used sampling method. Unfortunately, it is also the worst. When researchers sample from individuals they have "easy access" to, they are conducting a convenience sample. There are always hidden biases in these samples. Do a quick Google search for "FDR versus Alf Landon Sampling Error" to see a very famous example here. In addition, much of the error in predicting the results of the 2016 presidential election may be attributable to convenience sampling.

Sampling strategies, determine which

A consulting company is planning a pilot study on marketing in Boston. They identify the zip codes that make up the greater Boston area, then sample 50 randomly selected addresses from each zip code and mail a coupon to these addresses. They then track whether the coupon was used in the following month.

Sampling strategies, choose worst

A school district has requested a survey be conducted on the socioeconomic status of their students. Their budget only allows them to conduct the survey in some of the schools, hence they need to first sample a few schools.

Students living in this district generally attend a school in their neighborhood. The district is broken into many distinct and unique neighborhoods, some including large single-family homes and others with only low-income housing.

Experimental Design

Experiment versus Observational Study: Beyond just sampling, there are multiple methods for collecting data. We can just *observe* what happens naturally (without manipulating any conditions) or we can run an *experiment.* In experiments we manipulate one or more conditions, utilizing a control and treatment group(s). The advantage to an experiment is that we can infer cause and effect relationships (this is extremely important in medical studies), but in observational studies we can only discuss an association between variables.

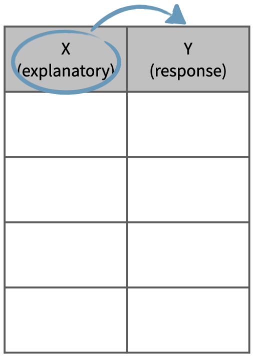
There's lots more to learn about experimental design, but it is beyond the scope of our course. You should read pages 32 through 35 of OpenIntro Statistics, 4Ed as a starting point.

Explanatory and response variables

Often when one mentions "a relationship between variables" we think of a relationship between just two variables, say a so called explanatory variable, x, and response variable, y. However, truly understanding the relationship between two variables might require considering other potentially related variables as well. If we don't, we might find ourselves in a *Simpson's paradox.* So, what is Simpson's paradox?

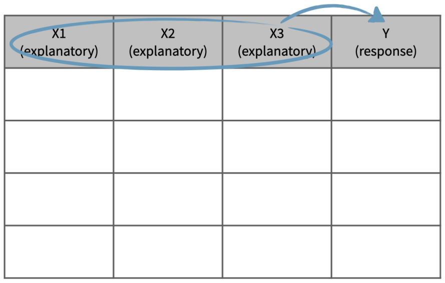
First, let's clarify what we mean when we say explanatory and response variables. Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified. We use these labels only to keep track of which variable we suspect affects the other.

Explanatory and response

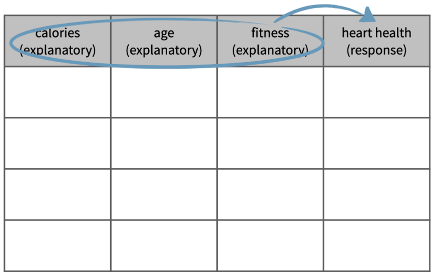


And these definitions can be expanded to more than just two variables. For example, we could study the relationship between three explanatory variables and a single response variable.

Multivariate relationships

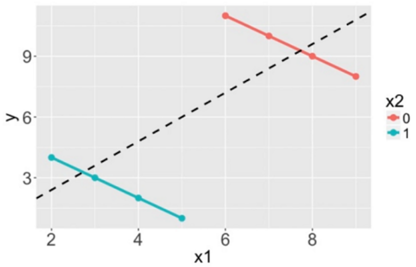


This is often a more realistic scenario since most real world relationships are multivariable. For example, if we're interested in the relationship between calories consumed daily and heart health, we would probably also want to consider information on variables like age and fitness level of the person as well.



Not considering an important variable when studying a relationship can result in what we call a **Simpson's paradox**. This paradox illustrates the effect the omission of an explanatory variable can have on the measure of association between another explanatory variable and the response variable. In other words, the inclusion of a third variable in the analysis can change the apparent relationship between the other two variables.

Consider the eight dots in the scatter plot below (the points happen to fall on the orange and blue lines). The trend describing the points when only considering `x1` and `y`, illustrated by the black dashed line, is reversed when `x2`, the grouping variable, is also considered. If we don't consider `x2`, the relationship between `x1` and `y` is positive. If we do consider `x2`, we see that within each group the relationship between `x1` and `y` is actually negative.



We'll explore Simpson's paradox further with another dataset, which comes from a study carried out by the graduate Division of the University of California, Berkeley in the early 70's to evaluate whether there was a sex bias in graduate admissions. The data come from six departments. For confidentiality we'll call them A through F. The dataset contains information on whether the applicant identified as male or female, recorded as `Gender`, and whether they were admitted or rejected, recorded as `Admit`.

Berkeley admission data

```
| Admitted | Rejected
```

——-| ———|——— Male | 1198 | 1493 Female | 557 | 1278

> Note: At the time of this study, gender and sexual identities were not given distinct names. Instead, it was common for a survey to ask for your "gender" and then provide you with the options of "male" and "female." Today, we better understand how an individual's gender and sexual identities are different pieces of who they are. To learn more about inclusive language surrounding gender and sexual identities see the gender unicorn.

4. Principles of experimental design: 4 principles

- Controlling (assign treatment and control groups, enforce specific treatment in treatment group)

- Randomization (randomly assign treatment group and control group);
- Replication (large sample, or replicate an entire study to verify earlier findings)
- Blocking

1.1.3 Solved Problems

Exercises:

Exercise 1. (page 11 #1.2) Researchers studying the effect of antibiotic treatment for acute sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments such as acetaminophen nasal decongestants, etc. At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. The distribution of responses is summarized below (with some cells missing numbers):

(for b), c), Round answers to within one hundredth of a percent)

	Self-reported improved in symptoms			
		Yes	**No**	**Total**
Treatment		66		85
Control		65		
Total				166

(a). Fill the blank cells in the above table.

(b). What percent of patients in the treatment group experienced improvement in symptoms?

(c). What percent experienced improvement in symptoms in the control group?

(d). In which group did a higher percentage of patients experience improvement in symptoms?

(e). Your findings so far might suggest a real difference in effectiveness of antibiotic and placebo treatments for improving symptoms of sinusitis. However, this is not the only possible conclusion that can be drawn based on your findings so far. What is one other possible explanation for the observed difference between the percentages of patients in the antibiotic and placebo treatment groups that experience improvement in symptoms of sinusitis?

(Answers for reference:

(a).

	Self-reported improved in symptoms		
	Yes	**No**	**Total**
Treatment	66	19	85
Control	65	16	81
Total	131	35	166

(e). Be careful: Do not generalize the results of this study. It is impossible to tell merely by comparing the sample proportions **because the difference could be the result of random error in our sample.**

Exercise 2. The following figure displays data from a lending company.

loan.amount	interest.rate	term	grade	state	total.income	homeownership
7500	7.34	36	A	MD	70000	rent
25000	9.43	60	B	OH	254000	mortgage
14500	6.08	36	A	MO	80000	mortgage
...	...	...	...	...	...	...
3000	7.96	36	A	CA	34000	rent

Variable descriptions

loan amount: Amount of the loan received, in US dollars.

interest rate: Interest rate on the loan, in an annual percentage.

term: The length of the loan, which is always set as a whole number of months.

grade: Loan grade, which takes a values A through G and represents the quality of the loan and its likelihood of being repaid.

state: US state where the borrower resides.

total income: Borrower's total income, including any second income, in US dollars.

homeownership: Indicates whether the person owns, owns but has a mortgage, or rents.

(a). How many cases in the data?

(b). Identify the types of variables.

Exercise 3. (page 19 #1.4) The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence (evidence based only on personal observation) suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were randomly split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.

(a). Identify the main research question of the study.

(b). Who are the subjects in this study and how many are included?

(c). What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous.

(Reference answer:

(a). The effect of Buteyko method on reducing asthma symptoms and improving quality of life.

(b). Asthma patients aged 18-69 who relied on medication for asthma treatment; 600.

(c). The variables and types are: quality of life (categorical), activity (categorical), asthma symptoms (categorical), and medication reduction on a scale from 0 to 10 (numerical discrete).)

Exercise 4. (page 29 #1.13) Exercise 1.3 introduces a study where researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study, air pollution levels were measured by air quality monitoring stations; lengths of gestation data were collected on 143,196 births between the years 1989 and 1993; and air pollution exposure during gestation was calculated for each birth.

(a). Identify the population of interest and the sample in this study.

(b). Comment on whether or not the results of the study can be generalized to the population and if the findings of the study can be used to establish causal relationships.

(Reference answer:

(a) Population: all births in Southern California. Sample: collected length of gestation data of 143,196 births between the years 1989 and 1993.

(b) If the collected lengths of gestation data of births in this time span and geography can be considered representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational, the findings cannot be used to establish causal relationships.)

Exercise 5. A fitness center is interested in the average amount of time a client exercises in the center each week. Match the vocabulary words (a-f) with its corresponding examples (1-6). (Note: 1-1 match)

Examples:

1. All 45 exercise times that were recorded from the participants in the study.
2. The 45 clients from the fitness center who participated in the study.
3. All clients at the fitness center.
4. The average amount of time that all clients from the fitness center exercise.
5. The amount of time that any given client from the fitness center exercises.
6. The average amount of exercise time for the 45 clients from the fitness center who participated in the study.

Vocabulary words:

a. Data

b. Population

c. Variable

d. Sample

e. Parameter

f. Statistic

Exercise 6. (Observational Study or Experiment)

a. You would like to investigate whether listening to music while taking exams affects performance. A group of students are told to listen to music while taking a test and their results are compared to a group not listening to music. Is this an experiment or an observational study?

b. The starting salaries of recent graduates from Ivy League private and public universities are recorded. Is this an experiment or an observational study?

Exercise 7. (page 37 #1.41) In a public health study on the effects of consumption of fruits and vegetables on psychological well-being in young adults, participants were randomly assigned to three groups: (1) diet as usual, (2) an ecological momentary intervention involving text message reminders to increase their fruits and vegetable consumption plus a voucher to purchase them, or (3) a fruit and vegetable intervention in which participants were given two additional daily servings of fresh fruits and vegetables to consume on top of their normal diet. Participants were asked to take a nightly survey on their smartphones. Participants were student volunteers at the University of Otago, New Zealand. At the end of the 14-day study, only participants in the third group showed improvements to their psychological well-being across the 14-days relative to the other groups.

(a). What type of study is this?

(b). Identify the explanatory and response variables.

(c). Comment on whether the results of the study can be generalized to the population.

(d). Comment on whether the results of the study can be used to establish causal relationships.

(e). A newspaper article reporting on the study states, “The results of this study provide proof that giving young adults fresh fruits and vegetables to eat can have psychological benefits even over a brief period of time.” How would you suggest revising this statement so that it can be supported by the study?

Reference answer:

(a). Experiment

(b). Explanatory: treatment group (categorical with 3 levels). Response variable: Psychological well-being.

(c). No, because the participants were volunteers.

(d). Yes, because it was an experiment.

(e). The statement should say “evidence” instead of “proof”.)

1.1.4 Exercises

Answer the following using your knowledge of the dataset and variable types.

1.2 Summarizing Numerical Data

1.2.1 Objectives

By the end of this unit, students will be able to:

- Summarize and describe numerical data using various visual displays including histograms and dotplots.
- Idnetify skewness in the distribution of numerical data.
- Use summary statistics such as mean and median to describe central tendancy of the data.
- Use summary statistics such as variance, standard deviation, and quartiles to describe variability of the data.
- Identify potential outliers in the data using boxplots.

1.2.2 Overview

Summarizing Numerical (Quantitative) Data

Recall: Variables for which computation of measures like the mean (average) or standard deviation are meaningful are numerical variables.

Measuring Central Tendancy

Measures of Central Tendency (Averages): The mean and median both attempt to measure the *center* of a dataset.

- The **mean** of a set of observations is the traditional *average*. We typically denote the mean by $\bar{x}$ (or μ in the case of population-level data) and it is computed as follows: $\bar{x} = \frac{\displaystyle\sum_{i=1}^{n} x_i}{n} = \frac{x_1+x_2+x_3+\ldots+x_n}{n}$
- The **median** is the *middle value* for a set of observations. To compute the median, list the numbers in ascending order and find the number or number(s) in the middle of the list. In the case that there is a single middle number, that is the median. In the case where there are two middle numbers, we take the *mean* of those two.

Aside: Defining my own data

For data which is not already known to `R` (ie. data which is not part of a data frame), we can still use `R` to quickly perform compuations. Consider the distributions of doors knocked on by two political campaign workers last week (Monday - Friday): $\begin{array}{lcl} \text{Worker A} & : & 23,\ 24,\ 25,\ 26,\ 27 \\ \text{Worker B:} & : & 0,\ 15,\ 25,\ 35,\ 50 \end{array}$. We do this below with the help of the `c()` function in `R`, which can be used to create lists of values.

Measuring Spread

Measures of Variability: Clearly, the center of a dataset doesn't tell the entire story. Our two political pollsters obviously have very different door-knocking strategies but both have a mean (and median) of 25 doors per day. We should also measure the *spread* of data.

The *standard deviation* of a set of observations is denoted by s (or σ in the case of population-level data) and is computed as follows:

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

We should also note that if you are certain that you are working with population-level data, then the denominator used to compute the standard deviation should be changed to N (the population size). We can do this because there is no uncertainty in estimating the population standard deviation if we have records from every element of the population.

Explaining the Standard Deviation Formula: The standard deviation seeks to measure an "average deviation" from the mean.

- If we don't look too closely at the formula, we can see the summation symbol ($\sum$) as well as division (by just about the number of values we've added up). That's almost like an average!
- What are we averaging? The quantity $(x - \bar{x})$ denotes an observed value's deviation from the mean. We shouldn't average these values though, since the mean sits in *center* of the data and we would have deviations above the mean (positive) "cancelling out" deviations below the mean (negative).
    - We square the deviations which has two effects: (1) all of the squared deviations are now non-negative, so that no cancellation can occur, and (2) large deviations from the mean carry a larger weight in measuring the standard deviation.
- Since we squared the deviations before computing the "average", the units of measure are no longer comparable to the original units that the variable was measured in – the units are square units now. This is why we see the large square root as the last piece of the formula – taking the square root brings us back to the original units.

The *inter-quartile range* (IQR) of a set of observations measures the spread of the "middle-50-percent" of the observations. The IQR is the distance between $Q1$ (the 25th percentile) and $Q3$ (the 75th percentile).

* The median of a set of observations splits the set into two halves: an upper half and a lower half. The median of the lower half is called the *first quartile* (Q_1) while the median of the upper half is called the *third quartile* (Q_3). The interquartile range is the distance between Q_1 and Q_3. That is,

$$IQR = Q_3 - Q_1$$

A Note on Skew: It is common to refer to data as *skewed* if the presence of outliers cause the mean and median to disagree with one another on the location of the "center" of our data. In this case, we say that the data is *skewed in the direction that those outliers have pulled the mean.* For example, we would say that the `carat` weight data (from above) is *skewed right.*

1. Graphical Presentations

- Scatter graph—present related two numerical data, see if there is any association, outliers
- Dot plot –see overall pattern outliers
- Histogram – see the shape of data distribution: modals, skewness

2. Numerical summaries

- Mean, Median – both measures are for the center of numerical data
- Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

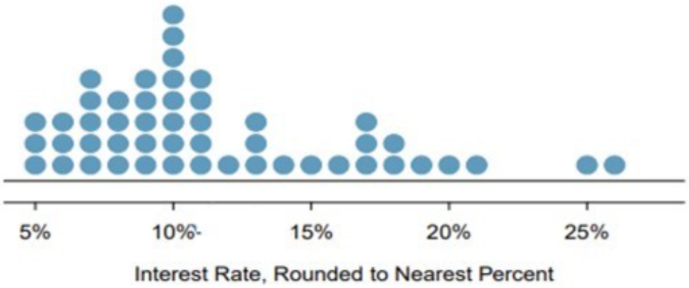
- Median is the middle value of the arranged data
- Standard Deviation —measures the spread or variation
- Quartiles, interquartile range,
- Five number summary: Min, Q1, Q2, Q3, Max
- Use the mean and the standard deviation for symmetrical data or large data
- Use the median and interquartile range for skewed data

3. R functions codes

- `c(?,?,?)` – concatenate function put multiple values in a vector
- `mean(c(?,?,?))` – mean
- `var(c(?,?,?))` – variance
- `sd(c(?,?,?))` – standard deviation
- `median(c(?,?,?))` – median
- `quantile(c(?,?,?), quantile.type=2)` # change type to 6, if n=5
- `IQR(c(?,?,?), quantile.type=2)` # change type to 6, if n=5
- `summary(c(?,?,?), quantile.type=2)` # change type to 6, if n=5

1.2.3 Solved Problem

Exercise 1. Use the following dot plot of an interest rate of some loan data to answer questions.



(a). How many loans?

(b). What is the lowest interest rate? What is the highest interest rate? Find the range.

(c). Find the mean (average).

(d). Find the median.

Help: Using R,

- `x <- c(5, 5, 5, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 11, 12, 13, 13, 13, 14, 15, 16, 17, 17, 17, 18, 118, 19, 20, 21, 25, 26)`
- `length(x)`
- `mean(x)`
- `median(x)`

(Answer: (a) 50 (b) 5; 26; 21 (c) 13.48 (d) 10)

Exercise 2. When we have a distribution where all observations are greater than 0, that is, all $x_i > 0$, the statistic $\frac{\text{mean}}{\text{median}}$ can be used as a measure of skewness. What is the expected shape of the distribution under the following conditions? Sketch the shape to illustrate.

(a) $\frac{\text{mean}}{\text{median}} = 1$

(b) $\frac{\text{mean}}{\text{median}} < 1$

(c) $\frac{\text{mean}}{\text{median}} > 1$

Exercise 3. For given two data sets: Data (1): 0, 2, 4, 6, 8, 10 Data (2) 20, 22, 24, 26, 28, 30

(a). Sketch the dot plots.

(b). Compare their means. What general observation can you draw?

(c). Compare their standard deviations. What general observation can you draw?

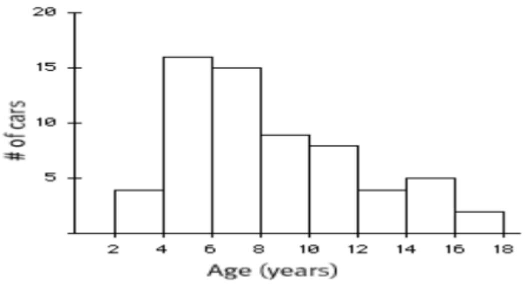
(d). What about their IQRs?

Exercise 4. Find the quartiles and interquartile range (IQR) for each data set.

(a). 2, 5, 10, 12, 16

(b). 2, 5, 10, 11, 12, 16

Exercise 5. The histogram below shows the ages (in years) available for sale in a car dealership on some day.



(a). How many cars are in the first class (between 2 and 4)? In the third class (between 6 and 8)?

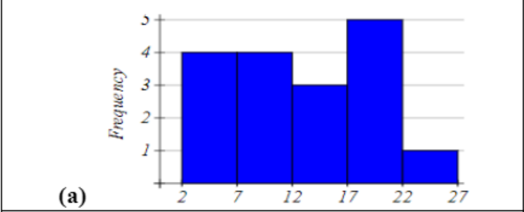
(b). Describe the shape of the histogram – how many modals? Symmetric or skewed (left or right)?

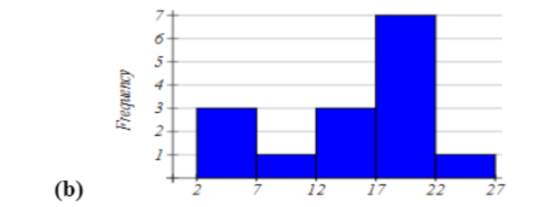
(c). Which measure is more appropriate to use to measure the center? Mean or median?

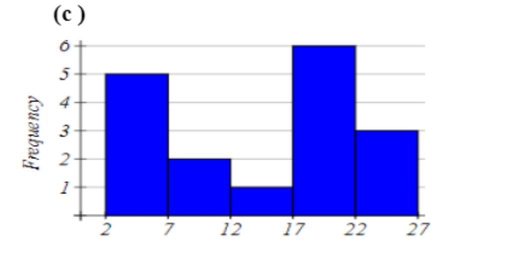
(d). Which measure is more appropriate to use to measure the spread? Standard deviation or IQR?

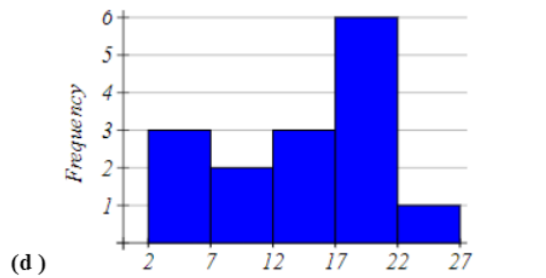
Exercise 6. Identify the histogram for the frequency distribution below.

Bin	Frequency
[2, 7]	3
[7, 12]	2
[12, 17]	3
[17, 22]	6
[22, 27]	1

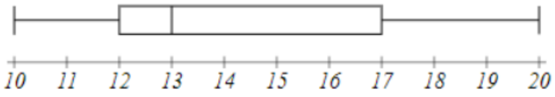

(a)


(b)


(c)


(d)

Exercise 7. Based on the boxplot below:



(a). Write the five number summary.

(b). What percent of data is below 17?

1.2.4 Exercises

1.3 Summarizing Categorical Data

1.3.1 Objectives

By the end of this unit, students will be able to:

- Summarize and describe the distribution of categorical data using contingency tables and various visual displays including bar plots and mosaic plots.
- Explore the association between a numerical variable and a categorical variable using side-by-side box plots.

1.3.2 Overview

Categorical data is best summarized using a frequency table. That is, we use a table that lists how *frequent* each observation occured in the dataset.

1. Tables

- Frequency of a category – the count of that category
- Relative Frequency of a category = $\frac{\text{Frequency in the category}}{\text{Total number of observations}}$
- Table (of one variable) – shows the list of values and the corresponding frequencies (or relative frequencies) of one categorical variable
- Contingency table—presents a summary (of counts or proportions) of two categorical variables bivariate variables
- Computing row and column proportions for contingency table

2. Bar plots

- Bar plot of one variable – visualizes the frequencies (or relative frequencies) of one categorical variable
- Bar lots for two categorical variables – put side by side bar plots
- Stacked bar plot – stack bars over (using different colors)
- Mosaic plot – is a special type of standardized stacked bar plots that represents a contingency table. In detailed words, it shows the percentage of one categorical variable (variable 1) for all categories of another variable (variable 2); can use the width of each bar to represent the ratio of variable 2)

3. Using **side by side Box plots** for exploring categorical-numerical relationships which provide information about how the distribution of the numeric variable changes across categories.

4. R codes (table and bar plot)

1) Create a table — use table()

```
Variable_Name = c ("category 1", "Category 2", ..., "Category n")
(Note: use quotation marks for strings - categories)

table(Variable_Name)
```

2) Create Bar plot – (sample from course notes)

```
# Create a data frame with the Variable_Name

Variable_Name <- data.frame(

Type = c("category 1", "Category 2", ..., "Category n"),

Frequency = c( f1,f2, ..., fn)  # f1,f2, ...,fn are frequencies

# Create a bar plot

barplot(Variable-Name$Frequency, names.arg = Variable-Name$Type,

main = "Frequency of Variable-Name",

xlab = "Type of Variable", ylab = "Frequency",

col = "blue", border = "black",    )
```

1.3.3 Solved Problem

Exercise 1. A survey polled a sample of 350 students for a proposed change of some regulations. The following table summarized the survey response result.

Responses	Frequency	Relative Frequency (Round to 3 decimals)
Support	200	
Neutral	53	
Oppose	97	
Total	350	

(a). How many support the proposed change?
(b). Fill the last column in the table.
(c). What is the percentage of the sampled students who opposed the proposed change?

Exercise 2. The following data is the recorded blood types of 30 volunteers who donated blood at a plasma center.

O O A B A A B O AB O
B A O A AB O B A B B
O O O A A B O B A A

Blood Type	Frequency	Relative Frequency
A		
B		
AB		
O		
Total		

(a). Summarize the data in a frequency table and calculate the relative frequencies.
(b). Draw a histogram for the frequency of the data.

Exercise 3. Four hundred undergraduate students were surveyed about their part time working hours during on semester. The following contingency table summarizes the survey result related to student status and working hours per week.

	Not working	Work 10 hours or less	Work more than 10 hours	Total
Freshman or Sophomore	132	28	20	180
Junior or Senior	120	48	52	220
Total	252	76	72	400

(a). Complete the table for the 2nd row, 3rd row proportions (relative frequencies by class, and overall)
(Divide the 2nd row, 3rd row of the table by 220, by 400)

	Not working	Work 10 hours or less	Work more than 10 hours	Total
Freshman or Sophomore	0.733	0.07	0.05	1
Junior or Senior				1
All UG				1

(b). Find the column proportions. Interoperate the meaning the ratios of 2nd, 3rd, and 4th columns

	Not working	Work 10 hours or less	Work more than 10 hours	Total
Freshman or Sophomore	132/252=	28/76=	20/72=	180/400=
Junior or Senior	120/252=	48/76=	52/72=	220/400
Total	252/252=	76/76=	72/72=	400/400

(c). Find the overall relative frequencies by dividing all by 400 (grant total) Interoperate the meaning of each.

	Not working	Work 10 hours or less	Work more than 10 hours	Total
Freshman or Sophomore	132/400=	28/400=	20/400=	180/400=
Junior or Senior	120/400=	48/400=	52/400=	220/400=
Total	252/400=	76/400=	72/400=	400/400=

1.3.4 Exercise

Chapter 2

Probability and Probability Distribution

2.1 Probability

2.1.1 Objectives

By the end of this unit, students will be able to:

- Explain the concept of randomness and how probability quantifies randomness.
- Recognize the basic concepts of sample space, equally likely outcomes, events, unions and intersections.
- Identify when two events are mutually exclusive, independent or complementary.
- Use probability rules to compute the probability of different types of events.
- Distinguish between marginal probability and conditional probability.

2.1.2 Overview

The starting point in studying probabilities is the concept of an **experiment or random process**, by which we mean some act or observation whose outcome is not known in advance. Simple examples would be

- Rolling a die

- Tossing a coin twice

- Observing the temperature at GSO at 3:00 pm this afternoon (Fo)

Although we cannot predict what we will observe, we can in some cases compile a list of all the outcomes we might observe. This is known as the **sample space** for the experiment, and is a set in mathematical terms, that is to say a collection of distinct items. Generally for a sample space of n possible outcomes we write $S = E1, E2, ..., E_n$. For example in the die rolling experiment we have $S=1,\ 2,\ 3,\ 4,\ 5,\ 6$ and in the coin tossing experiment we could list the outcomes as *S=TT, TH, HT, HH* It is a little harder to list the sample space for the third experiment, given the current temperature a list of the numbers between 40 and 80 would probably suffice.

We use the outcomes in the sample space for computing probabilities for any event according to the following basic rule:

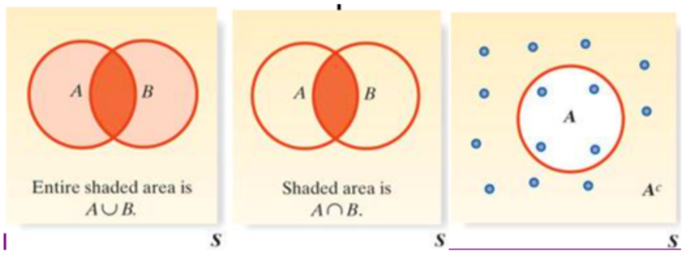
P(A) = the sum of the probabilities of all the outcomes in the event A

Events in probability are just the same as sets in mathematics, so you should know the principle operations on sets:

- UNION: A ∪ B (“A or B”) is the set of all outcomes in A or in B or in both

- INTERSECTION: A ∩ B (“A and B”) is the set of outcomes that are in both A and B

- COMPLEMENT: Ac (“not A”) everything outside of A (but in S)



It is important to realize that A ∪ B, A ∩ B, and A^c are all sets, that is to say, they are **collections of *distinct* items, and no element may be listed twice.**

For example, with reference to the die rolling experiment, define the event B as “a number at least as great as 5 comes up”, so B={5,6}. For A={2,4,6} (an even number comes up), we have

- $A \cup B$=2,4,5,6,

- $A \cap B$=6,

- A^c=1,3,5,

- B^c=1,2,3,4

Suppose the die is assumed to be fair. This, by definition, means that each side is equally likely to come up when the die is rolled. If we assign a total probability of 1 to the entire sample space, then we should assign a probability of 1/6 to each of the 6 outcomes in the sample space, so

$$\mathrm{P}(1) = \mathrm{P}(2) = \mathrm{P}(3) = \mathrm{P}(4) = \mathrm{P}(5) = \mathrm{P}(6) = 1/6.$$

Thus, we should obtain:

$P(A) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

$P(B) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$

$P(A \cup B) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$

$P(A \cap B) = \frac{1}{6}$

$P(A^c) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

$P(A^c) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{2}{3}$

Some points worth noting:

• P(A U B) = 4/6 does NOT equal P(A) + P(B) = 5/6 (this is because the two sets have an element, 6, in common).

• P(Ac) = 1 - P(A) and P(Bc) = 1 – P(B).

• Also note that in cases where all the outcomes in the sample space are equally likely, the rule about adding the probabilities for all of the outcomes in the event simplifies to counting the number of outcomes in the event and dividing by the number out outcomes in the sample space. So, for the events A and B above, P(A) = 3/6 = 1/2, and P(B) = 2/6 = 1/3.

• Generally, if A contains m outcomes, and the sample space has n equally likely outcomes (so the probability for each outcome is 1/n), then P(A) = m/n.

At this stage, let us write down some **general rules or AXIOMS for probability.** The events A, B below are now general events in a sample space, not the specific ones described above.

1) For any sample space, P(S) = 1
2) For any event A, 0 $\leq$ P(A) *le* 1
3) General addition rule: If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

$$\text{P (A or B) = P (A) + P (B) - P (A and B)}$$

where P(A and B) is the probability that both events occur. Here, A or B occurs means A, B, or both A and B occur.

4)Addition rule for disjoint events: For any events A, B which have no common outcomes,

$$\text{P(A U B) = P(A) + P(B)}$$

Events which have no outcomes in common are often referred to as **mutually exclusive** (or *disjoint*). If two events A, B are mutually exclusive, we write P(A $\cap$ B) = $\emptyset$, where $\emptyset$ is referred to as the *empty set.* It is sort of equivalent to the number zero in arithmetic.

Note that using the first and third rules, we have $1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$) , so we have a general rule that for any set A, $P(A^c) = 1˘P(A)$. This is often called the **law of complements**.

2.1.3 Solved Problems

Exercise 1. A set of 11 cards is numbered 1 through 11. A card is picked at random and the following events defined: A - the number on card is odd: B - the number on the card is 5 or higher. Find

a) P(A) =

b) P(B) =

c) P(A and B) =

d) P(A or B) =

Exercise 2. (Sample space where the outcomes are NOT equally likely): Professor Donald Fraser of the University of Toronto constructed a (purposely) uneven die. On inspection it was clear that the sides would not have equal probability. He rolled it 12,800 times, and came up with the following empirical probabilities (based on relative frequency):

Side	1	2	3	4	5	6
Probability	0.186	0.179	0.207	0.137	0.149	0.142

For the events A (even number) and B (at least 5) compute:

a) P(A) =

b) P(B) =

c) $P(A \cup B) =$

d) $P(A \cap B) =$

e) $P(A^c) =$

f) $P(B^c) =$

Exercise 3. A group of 1000 students is classified by gender G_1 (male) or G_2 (female), and by year, $\mathbf{Y}_1$ (freshman), $\mathbf{Y}_2$ (sophomore), $\mathbf{Y}_3$ (junior), or $\mathbf{Y}_4$ (senior). This results in the following table:

	Freshman	Sophomore	Junior	Senior	
	(Y_1)	(Y_2)	(Y_3)	(Y_4)	Total
Male (G_1)	140	120	110	70	440
Female (G_2)	160	130	140	130	560
Total	300	250	250	200	1000

If a student is randomly selected, find the probability that the student:

a) Is a junior, P(Y_3) =

b) Is a female freshman, P(G_2 and Y_1) =

c) Is a male or a junior, P(G_1 or Y_3) =

d)Is not a freshman, P(not Y_1) =

e) Is not a male and is not a junior, P(not G_1 and not Y_3) =

Exercise 4. Here is the "Craps Game" sample space, where a red and a green die are rolled. Each outcome (i,j) represents the red die coming up i and the green die coming up j.

	j=1	j=2	j=3	j=4	j=5	j=6
i=1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
i=2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
i=3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
i=4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
i=5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
i=6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

There are n=36 outcomes, and for fair dice it is reasonable to assume they are equally likely.**Find the probability of each of the following events. Identify which pairs of events are disjoint.**

(a) A "sum is 7":
P(A) =

(b) B "sum is 11":
P(B) =

(c) C "sum is 6":
P(C) =

(d) D "both dice show same number":
P(D) =

(e) E "both dice odd":
P(E) =

(f) F "both dice even":
P(F) =

Exercise 5. True or False

a) If A and B are mutually exclusive (disjoint) events, then P(A and B) = 0.

b) For any event A, P(A) + P(A^c) = 1.

Question 1:

Since, A = {1,3,5,7,9,11}, B= {5,6,7,8,9,10,11}, (A and B) = {5,7,9,11} and (A or B)={1,3,5,6,7,8,9,10,11}

a) P(A) = $\frac{6}{11}$

b) P(B) = $\frac{7}{11}$

c) P(A and B) = $\frac{4}{11}$

d) P(A or B)=P(A)+P(B)-P(A and B)
$= \frac{6}{11} + \frac{7}{11} - \frac{4}{11} = \frac{9}{11}$

Question 2:

A = {2,4,6}, B= {5,6}, (A∪B) = {2,4,5,6} and (A ∩ B)={6}

a) P(A)=P(2)+P(4)+P(6)=$0.179 + 0.137 + 0.142 = 0.458$

b) P(B)=P(5)+P(6) $= 0.149 + 0.142 = 0.291$

c) P(A U B) = P(2)+P(4)+P(5)+P(6) $= 0.179 + 0.137 + 0.149 + 0.142$
$= 0.607$

d) P(A∩B)=P(6)=0.142

e) P(A^c)= 1 - P(A) = 1 - 0.458 = 0.542

f) P(B^c)= 1 - P(B) = 1 - 0.291 = 0.709

Question 3:

The probability that the student :

a) Is a junior, P(Y_3) = $\frac{250}{1000} = 0.25$

b) Is a female freshman, P(G_2 and Y_1)= $\frac{160}{1000} = 0.16$

c) Is a male or a junior, P(G_1 or Y_3)= $\frac{440+140}{1000} = \frac{580}{1000} = 0.580$
or P(G_1)+ P(Y_3) - P(G_1 and Y_3) = $\frac{440}{1000} + \frac{250}{1000} - \frac{110}{1000}$

d) Is not a freshman, P(not Y_1)= 1 - P(Y_1) = 1 - ($\frac{300}{1000}$) =1 - 0.3 = 0.70

e) Is not a male and is not a junior, P (not G_1 and not Y_3) = $\frac{150+130+130}{1000} = \frac{420}{1000} = 0.420$

Question 4:

Note: For this question, also ask them to identify which pair of event are disjoint.

a) A “sum is 7” : {(1,6),(6,1), (2,5), (5,2), (3,4), (4,3)}

P(A) = $\frac{6}{36} = \frac{1}{6}$

b) B “sum is 11”: {(5,6),(6,5)}

P(A) = $\frac{2}{36} = \frac{1}{18}$

c) C “sum is 6”: {(1,5), (5,1), (2,4), (4,2), (3,3)}

P(C) = $\frac{5}{36}$

d) D “both dice show small number”: {(1,1), (2,2), (3,3), (4,4), (5,5),(6,6)}

P(D) = $\frac{6}{36} = \frac{1}{6}$

e) E “both dice odd”: {(1,1), (1,3), (1,5), (3,1), (3,3),(3,5),(5,1),(5,3),(5,5)}

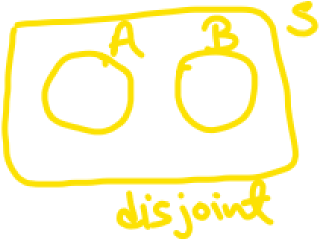
P(E) = $\frac{9}{36} = \frac{1}{4}$

f) F “both dice even”: {(2,2), (2,4), (2,6), (4,2), (4,4),(4,6),(6,2),(6,4),(6,6)}

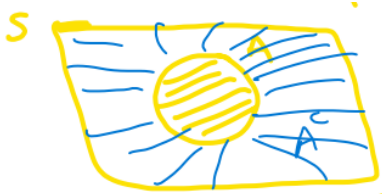
P(F) = $\frac{9}{36} = \frac{1}{4}$

Question 5:

a) If A and B are mutually exclusive (disjoint) events, then P(A and B) = 0.
Ans : TRUE



b) For any event A, P(A) + P(A^c) = 1.
Ans : TRUE



$P(A) + P(A^c) = P(S) = 1$

2.1.4 Exercises

2.2 Discrete Random Variables

2.2.1 Objectives

By the end of this unit, students will be able to:

- Use probability rules to compute the probability of different types of events.
- Distinguish between marginal probability and conditional probability.
- Apply the multiplication rule to compute probabilities when sampling with/without replacement from finite populations.
- Define random variables.
- Compute expectation and variance of discrete random variables.

2.2.2 Overview

2.2.3 Solved Problems

1. What is the probability that exactly five of them support the proposition?

Using the binomial probability formula:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where $n = 10$, $k = 5$, and $p = 0.4$.

2. What is the probability that five or six of them support the proposition?
3. What is the probability that at least three of them support the proposition?

Exercise: the manufacturer of the ColorSmart-5000 television set claims that 95% of its sets last at least five years without requiring a single repair. Suppose that we contact 8 randomly selected ColorSmart-5000 purchasers five years after they purchased their sets and ask each purchaser: Have you needed any repair for your ColorSmart-5000 TV set during the first 5 years after purchasing the set?

1. Find the probability that exactly 7 customers needed at least one repair during the first 5 years.

2. Find the probability that at least 7 purchasers needed at least one repair during the first 5 years.

Solution

Exercise: Profit from crop yield under different weather condi-tions (X). 1. Determine the missing probability in the following distribution.

Weather	**Profit ($)**	**Probability**
Dry	200,000	0.30
Light Rain	300,000	0.50
Storm	150,000	?

2. Find the expected profit from this crop, μ.

$$\begin{aligned}\mu &= \sum x_i P(x_i)\\ &= \$200k \cdot 0.3 + \$300k \cdot 0.5 + \$150k \cdot 0.2\\ &= \$60k + \$150k + \$30k = \$240k\end{aligned}$$

3. Find the variance, σ^2 and the standard deviation, σ.

$$\begin{aligned}\sigma^2 &= \sum (x_i - \mu)^2 P(x_i)\\ &= (\$200k - \$240k)^2\ .0.3 + (\$300k - \$240k)^2\ .0.5 + (\$150k - \$240)^2 .0.2\\ &= \$^2 3,900,000,000\end{aligned}$$

The standard deviation is $\sigma = \sqrt{\$^2 3,900,000,000} =$ $62,450

4. Interpret the value of the expected profit, μ.

The expected profit represents the long-run average, the expected profit on average in the future.

Example/Exercise: 40% of all voters support Proposition A. If a random sample of 10 voters is polled. Find the following probabilities.

1. What is the probability that exactly five of them support the proposition?

$$\begin{aligned}
P(X=5) &= \frac{10!}{(10-5)!5!}(0.40)^5(0.60)^{10-5} \\
&= \frac{10!}{5!5!}(0.40)^5(0.60)^5 \\
&= (252)(0.01024)(0.07776) \\
&= 0.2007
\end{aligned}$$

2. What is the probability that five or six of them support the proposition?

$$\begin{aligned}
P(X=5)+P(X=6) &= 0.2007 + \frac{10!}{(10-6)!6!}(0.40)^6(0.60)^{10-6} \\
&= 0.2007 + \frac{10!}{4!6!}(0.40)^6(0.60)^4 \\
&= 0.2007 + (210)(0.004096)(0.1296) \\
&= 0.2007 + 0.1115 \\
&= 0.312
\end{aligned}$$

3. What is the probability that at least three of them support the proposition?

$$\begin{aligned}
P(x \geq 3) &= P(X=3)+P(X=4)+P(X=5)+P(X=6)+...+P(X=10) \\
&= 1-[P(X=0)+P(X=1)+P(X=2)] \\
&= 1-[\frac{10!}{0!10!}(0.40)^0(0.60)^{10}+\frac{10!}{1!9!}(0.40)^1(0.60)^9+\frac{10!}{2!8!}(0.40)^2(0.60)^8] \\
&= 1-[(1)(1)(0.0060)+(10)(0.40)(0.0101)+(45)(0.16)(0.0168)] \\
&= 1-[0.0060+0.0403+0.1209] \\
&= 1-0.1672 = 0.8328
\end{aligned}$$

Exercise: the manufacturer of the ColorSmart-5000 television set claims that 95% of its sets last at least five years without requiring a single repair. Suppose that we contact 8 randomly selected ColorSmart-5000 purchasers five years after they purchased their sets and ask each purchaser: Have you needed any repair for your ColorSmart-5000 TV set during the first 5 years after purchasing the set?

1. Find the probability that exactly 7 customers needed at least one repair during the first 5years.

$$\begin{aligned}
P(X=7) &= \frac{8!}{(8-7)!7!}(0.05)^7(0.95)^{8-7} \\
&= \frac{8!}{1!7!}(0.05)^7(0.95)^1 \\
&= 0.0000000059375
\end{aligned}$$

2. Find the probability that at least 7 purchasers needed at least one repair during the first 5years.

$$\begin{aligned}
P(X=7) &= P(X=7)+P(X=8) \\
&= 0.0000000059375+\frac{8!}{(8-8)!8!}(0.05)^8(0.95)^{8-8} \\
&= 0.0000000059375+0.0000000000390625 \\
&= 0.0000000059765625
\end{aligned}$$

2.2.4 Exercises

Exercise: Profit from crop yield under different weather conditions (X).

1. Determine the missing probability in the following distribution.

Weather	Profit ($)	Probability
Dry	200,000	0.30
Light Rain	300,000	0.50
Storm	150,000	?

2. Find the expected profit from this crop, μ.

x_i	$P(x_i)$	$x_iP(x_i)$
200,000	0.30	$200,000 \times 0.30$
300,000	0.50	$300,000 \times 0.50$
150,000	?	$150,000 \times ?$

3. Find the variance, σ^2 and the standard deviation, σ.

x_i	$P(x_i)$	μ	$(x_i-\mu)$	$(x_i-\mu)^2$	$(x_i-\mu)^2P(x_i)$
200,000	0.30				
300,000	0.50				
150,000	?				

The standard deviation is $\sigma =$ ______________________

4. Interpret the value of the expected profit, μ.

Example/Exercise: 40% of all voters support Proposition A. If a random sample of 10 voters is polled. Find the following probabilities.

Formula for the probability of exactly x successes from n trials

$$p(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}; \quad \text{where } x = 0, 1, 2, \ldots, n$$

and

$$n! = n(n-1)(n-2)\cdots(3)(2)(1)$$

2.3 Binomial Distribution

2.3.1 Objectives

By the end of this unit, students will be able to:

- Identify when the conditions apply for the Binomial distribution to be used.
- Apply the Binomial distribution to model counts resulting from binary trials.

2.3.2 Overview

Binomial Distribution Condition

Conditions to be satisfied for a Binomial Variable Distribution with a fixed number of trials n:

- The trials are independent
- Each trial has two possible outcomes classified as success or failure

- The probability of a success p is the same for each trial

Probability Mean and Standard Deviation

For a binomial random variable X with n trials and the probability of a single trial being a success p, the probability of observing exactly k successes is

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \quad (k=0,1,\ldots,n)$$

Where: - $n!=1\times 2\times\cdots\times n$ - $0!=1$ - $\binom{n}{k}=\frac{n!}{k!(n-k)!}$ (read as "n choose k", also called the combination coefficient)

The probability of at most k successes is given by

$$\mathbb{P}\left[X\leq k\right]=\sum_{i=0}^{k}\binom{n}{i}\cdot p^i\left(1-p\right)^{n-i}\approx \texttt{pbinom(k,n,p)}$$

In the equations above, $\binom{n}{k}=\frac{n!}{k!(n-k)!}$ counts the number of ways to arrange the k successes amongst the n trials. That being said, the `R` functionality, `dbinom()` and `pbinom()` allow us to bypass the messy formulas – but you'll still need to know what these functions do in order to use them correctly!

Tip: We need to use the binomial distribution to find probabilities associated with numbers of successful (or failing) outcomes in which *we do not know for certain the trials on which the successes (or failures) occur*

Mean: $\mu=np$

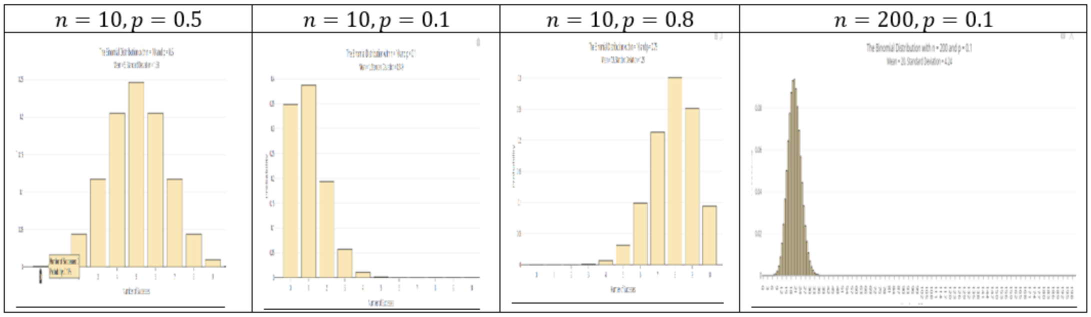
Standard deviation: $\sigma=\sqrt{np(1-p)}$

Observations that are more than 2 standard deviations away from the mean are considered unusual:

Unusual if outside of $\mu-2\sigma$ and $\mu+2\sigma$

Shape of Binomial Distribution

- For $p<0.5$: skew to the left
- For $p>0.5$: skew to the right
- For $p=0.5$: symmetric (centered at μ)
- For large n, if $np\geq 10$ and $n(1-p)\geq 10$, the graph is approximately bell-shaped.



(Generated using online app https://istats.shinyapps.io/BinomialDist/)

Using R

- For $P(X = k)$: `dbinom(k, n, p)`
- For $P(X \leq k) = P(X < k+1) = P(X = 0) + P(X = 1) + \cdots + P(X = k)$: `pbinom(k, n, p, lower.tail = TRUE)` (the `lower.tail = TRUE` can be omitted)
- For $P(X > k) = P(X \geq k+1) = 1 - P(X \leq k) = P(X = k+1) + \cdots + P(X = n)$: `pbinom(k, n, p, lower.tail = FALSE)`
- For $n!$: `factorial(n)`
- For $\binom{n}{k}$: `choose(n, k)`

Using Calculator

1. For $P(X = x)$:
   - `2ND` → `VARS` (`DISTR`) → use arrow to select `binompdf` (enter n, p, x) then `enter`
2. For $P(X \leq x)$:
   - `2ND` → `VARS` (`DISTR`) → use arrow to select `binomcdf` (enter n, p, x) then `enter`
3. For $n!$:
   - Example: 7!
   - Enter 7 then press `Math` key; use (right) arrow key to select `PROB` then use (down) arrow key to select `!` (press enter it then shows `7!`); press the `enter` key again (to get answer 5040)
4. For $\binom{n}{k}$:
   - Example: $\binom{9}{2}$
   - Enter 9 then `MATH` → arrow to `PROB` → choose `nCr` then enter 2 then `enter` to get the result (answer is 36)

2.3.3 Solved Problem

2.3.4 Exercises

Exercise 1. How many ways can we choose 2 students from a group of 6?

Exercise 2 (Combination Formula)

Survey four randomly selected students and record the outcomes as “I” (in state) or “O” (out state). Fill the table below.

# of “I”	Outcomes (list all)	# of outcomes	$\binom{4}{k} = \frac{4!}{k!(4-k)!}$
$k = 0$			
$k = 1$			
$k = 2$			
$k = 3$			
$k = 4$			

Exercise 3 Find the probability of success of the Bernoulli trial with n trials, success probability p, and the success k:

- $n = 3$, $k = 2$, $p = 0.35$
- $n = 5$, $k = 3$, $p = 0.2$

Exercise 4

For a binomial distribution with $n = 4$, $p = 0.7$. (As in exercise 1, assume that 70% are in-state students.)

(a). Write the formula for computing the probability of getting exactly k successes.

(b). Fill the following distribution table. (Round to 4 decimals) (you may use R calculator)

X	$P(X = k)$	$P(X \leq k)$
0		
1		
2		
3		
4		1
Total	1	

(c). What is the expected value?

(d). What is the standard deviation?

Exercise 5

About 75% of dog owners buy holiday presents for their dogs. Suppose twenty dog owners are randomly selected, find the probability of:

(a). Exactly three buy their dog holiday presents

(b). Exactly seventeen do not buy their dog holiday presents

(c). Three or more buy their dog holiday presents

(d). At most four buy their dog holiday presents

(e). Minimum of 11 and maximum of 17 dog owners buy their dog holiday presents

(f). Find the expected number of dog owners in this sample, who buy their dog holiday presents.

(g). Is it unusual if 16 out of 20 randomly selected dog owners buy their dog holiday presents? Why?

(h). Is it unusual if 10 out of 20 randomly selected dog owners buy their dog holiday presents? Why?

Review on Binomial Distribution

Conditionsto be satisfied for a Binomial Variable Distribution:

- The number of trials,is a fixed positive integer
- The trials are independent
- Each trail has two possible outcomes, classified as success or failure
- The probability of a success, p, is the same for each trial

Binomial Distribution

For a binomial random variable with trials and the probability of a single trial being a success the probability of observing exactly successes is

$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$ k= 0,1, ..., n

Accumulative Probability

(at most k success) $P(X \leq k) = P(X < k+1) = \sum_{i=0}^{k} P(X = i)$

(at least k success) $P(X \geq k) = P(X > k-1) = 1 - P(X \leq k-1) = \sum_{i=k}^{n} P(X = i)$

Factorial and Combination Coefficient

n! $= 1 \times 2 \times ... \times n$ 0! $= 1$

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ ($\binom{n}{k}$ is read as “n choose k”)

Solution:

Exercise 1. How many ways can we choose 2 students from a group of 6?

```
choose(6,2)
```

```
## [1] 15
```

Exercise 4. For a binomial distribution with n=4,p=0.2.

a) Write the formula for computing the probability of getting exactly k successes.

$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$ k= 0,1, ..., n

In R you use dbinom(k,n,p)

b) Fill the following distribution table. (Round to 4 decimals) (you may use R calculator)

```
## For P(X=x)
dbinom(c(0,1,2,3,4), 4, 0.2)
```

```
## [1] 0.4096 0.4096 0.1536 0.0256 0.0016
```

```
## P(X <= x)
pbinom(c(0,1,2,3,4), 4, 0.2)
```

```
## [1] 0.4096 0.8192 0.9728 0.9984 1.0000
```

c) What is the expected value?

```
n=4;p=0.2
Expected_mean <-n*p ; Expected_mean
```

```
## [1] 0.8
```

d) What is the standard deviation?

```
n=4
p=0.2
sd <- sqrt(n*p*(1-p)); sd
```

```
## [1] 0.8
```

Exercise 5. About 75% of dog owners buy holiday presents for their dogs. Suppose twenty dog owners are randomly selected, find the probability of a) Exactly three buy their dog holiday presents

$$P(X = 3)$$

```
dbinom(3,20,0.75)
```

```
## [1] 2.799425e-08
```

b) Exactly seventeen do not buy their dog holiday presents

$$P(X = 17)$$

```
dbinom(17,20,0.25)
```

```
## [1] 2.799425e-08
```

c) Three or more buy their dog holiday presents

$$P(X \geq 3) = 1 - P(X \leq 3)$$

```
1 - pbinom(3,20,0.75)
```

```
## [1] 1
```

```
## or
pbinom(3,20,0.75, lower.tail=FALSE)
```

```
## [1] 1
```

d) At most four buy their dog holiday presents

$$P(X \geq 4) = P(X = 0) + ... + P(X = 4)$$

```
pbinom(4,20,0.75)
```

```
## [1] 3.865316e-07
```

```r
## or

sum(dbinom(c(0,1,2,3,4),20,0.75))
```

```
## [1] 3.865316e-07
```

e) Minimum of 11 and maximum of 17 dog owners buy their dog holiday presents

$$P(11 \leq X \leq 17)$$

```r
pbinom(17,20,0.75)-pbinom(10,20,0.75)
```

```
## [1] 0.8948752
```

f) Find the expected number of dog owners in this sample, who buy their dog holiday presents

```r
n=20; p=0.75
E_x = n*p ;
E_x
```

```
## [1] 15
```

g) Is it unusual if 16 out of 20 randomly selected dog owners buy their dog holiday presents? Why?

```r
dbinom(16,20,0.75)
```

```
## [1] 0.1896855
```

Comment

Whether it’s unusual depends on your chosen significance level. If you consider a low probability (e.g., p<0.05) as unusual, then it might be considered unusual.

Is it unusual if 10 out of 20 randomly selected dog owners buy their dog holiday presents? Why?

```r
dbinom(10,20,0.75)
```

```
## [1] 0.009922275
```

Comment

Whether it’s unusual depends on your chosen significance level. If you consider a low probability as unusual, then it might be considered unusual.

2.4 Normal Distribution

2.4.1 Objectives

By the end of this unit, students will be able to:

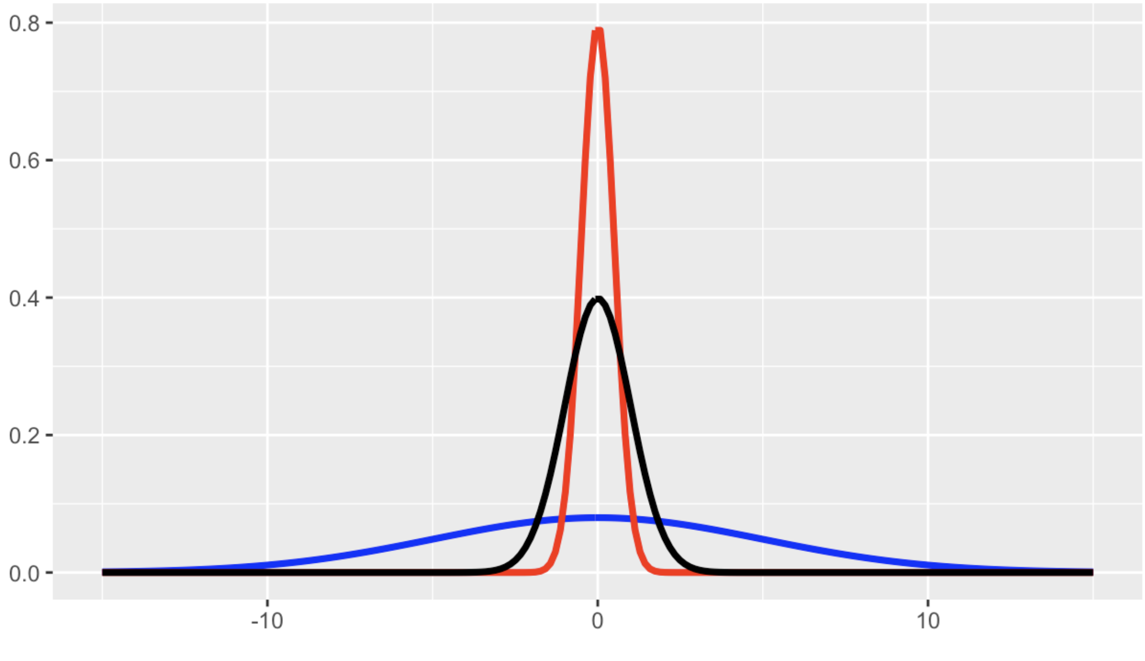
- Understand the notion and characteristics of continuous probability distributions.
- Use the normal distribution to model continuous random variables.

2.4.2 Overview

The Normal Distribution

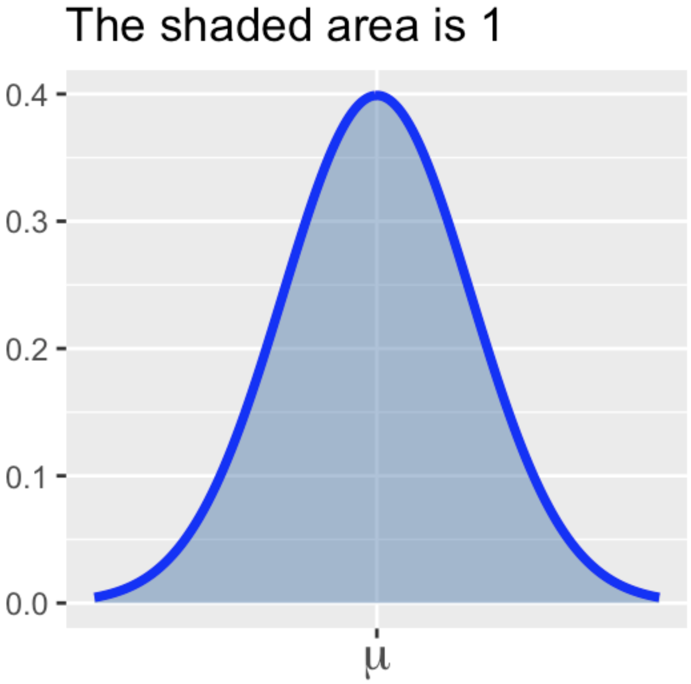
Definition: If a random variable X is normally distributed with mean μ and standard deviation σ, we often write $X \sim N(\mu, \sigma)$. Three different normal distributions appear below.

- In **blue** is a normal distribution with $\mu = 0$ and $\sigma = 5$
- In **red** is a normal distribution with $\mu = 0$ and $\sigma = 0.5$
- In **black** is a normal distribution with $\mu = 0$ and $\sigma = 1$ (the so-called *Standard Normal Distribution*)

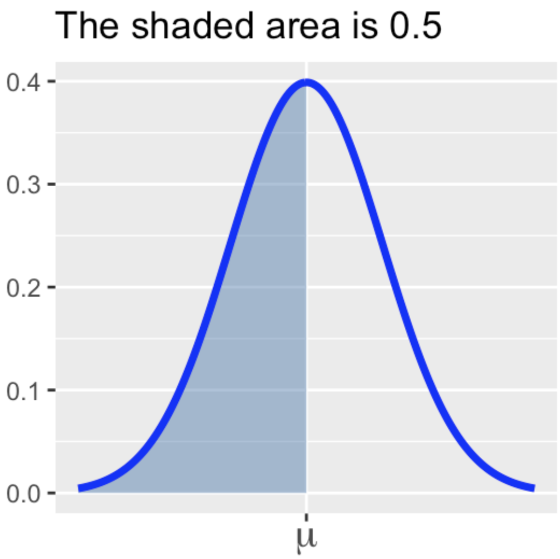


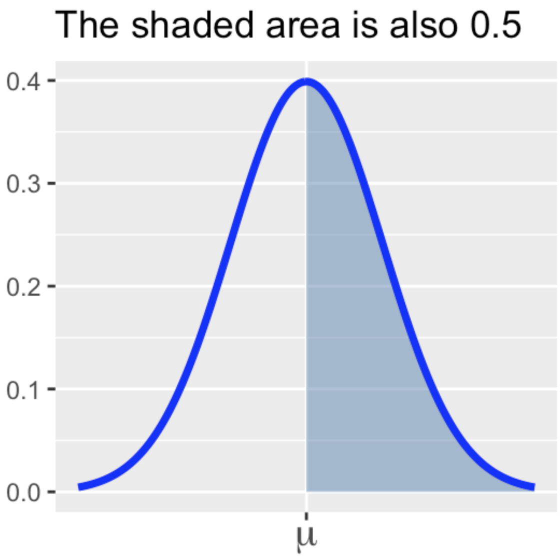
Properties of the Normal Distribution: We have the following properties associated with the normal distribution. Consider $X \sim N(\mu, \sigma)$.

The area beneath the entire distribution is 1 (since this is equivalent to the probability that X takes on any of its possible values).

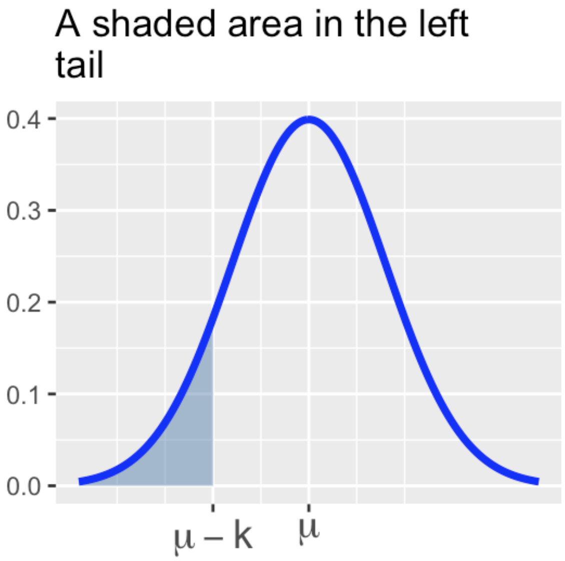


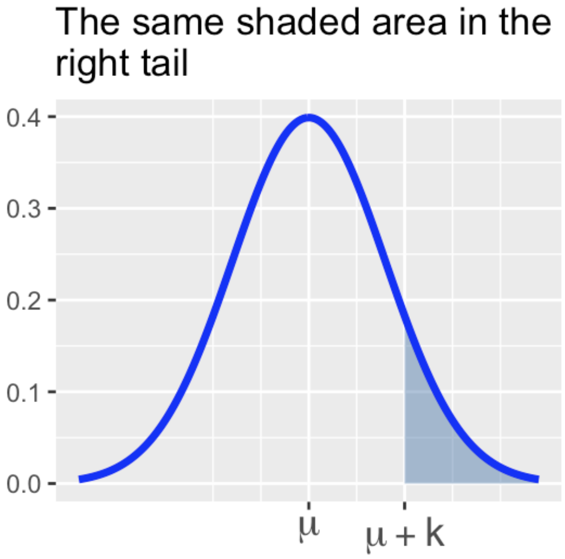
$\mathbb{P}[X \leq \mu] = \mathbb{P}[X \geq \mu] = 0.5$ (the area underneath a full half of the distribution is 0.5)



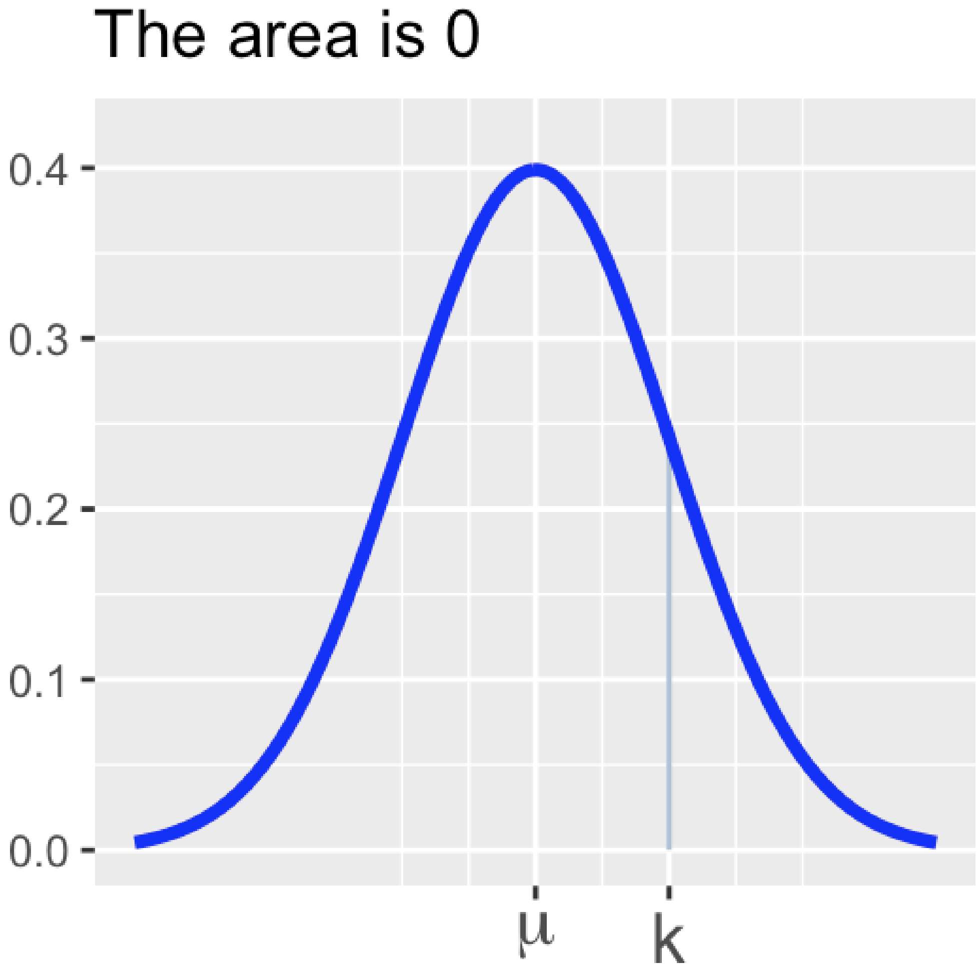


The distribution is symmetric. In symbols, $\mathbb{P}\left[X \leq \mu - k\right] = \mathbb{P}\left[X \geq \mu + k\right]$ for any k.





$\mathbb{P}[X = k] = 0$ (the probability that X takes on any prescribed value exactly is 0)



Sometimes it is useful to be able to estimate probabilities or to estimate the proportion of a population that falls into a range as long as the population is nearly normal. A convenient rule of thumb is the *Empirical Rule*.

The Empirical Rule: If $X \sim N(\mu, \sigma)$, then

$\mathbb{P}[\mu - \sigma \leq X \leq \mu + \sigma] \approx 0.68$ – that is, about 68% of observations lie within one standard deviation of the mean.

$\mathbb{P}[\mu - 2\sigma \leq X \leq \mu + 2\sigma] \approx 0.95$ – that is, about 95% of observations lie within two standard deviations of the mean.

$\mathbb{P}[\mu - 3\sigma \leq X \leq \mu + 3\sigma] \approx 0.997$ – that is, about 99.7% of observations lie within three standard deviations of the mean.

Standardization and z-scores

Scenario: Two students, Bob and Sally, are trying to compare how well they did on a college entrance exam. The difficulty comes in that Bob took the SAT which is known to follow an approximate normal distribution with a mean score of 1068 and a standard deviation of 210 while Sally took the ACT which also follows an approximately normal distribution but with a mean score of 20.8 and a standard deviation of 5.8. If Bob scored a 1400 on the SAT and Sally scored a 31 on the ACT, who scored relatively higher?

How do we answer this question? We'll see two methods.

Method 1: We can standardize the test scores so that they have comparable units.

- **Definition:** If an observation x comes from a nearly normal population with mean μ and standard deviation σ then we compute z-score associated with x as follows:

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

- An observation's z-score is simply the number of standard deviations it falls above or below the mean.

A recap on z-scores: We can use z-scores as a common unit for comparing observations from completely different populations (such as SAT scores and ACT scores). Here's a recap of the most important information so far:

If an observation x comes from a nearly normal population with mean μ and standard deviation σ, we can compute it's z-score using the formula: $z = \dfrac{x - \mu}{\sigma}$.

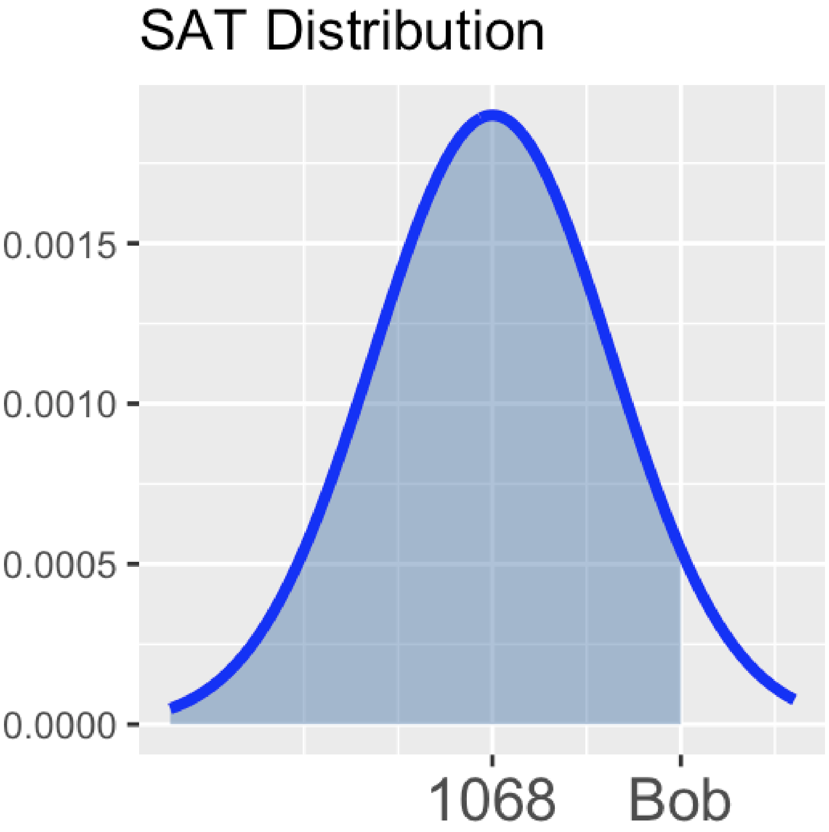
A z-score measures the number of standard deviations which an observation falls above or below the mean.

- A positive z-score means that an observation was above the mean.
- A negative z-score means that an observation was below the mean.
- The larger a z-score is in absolute value, the further the corresponding observation falls from the mean. That is, the larger the magnitude of a z-score, the further into the tail of the distribution the corresponding observation falls.

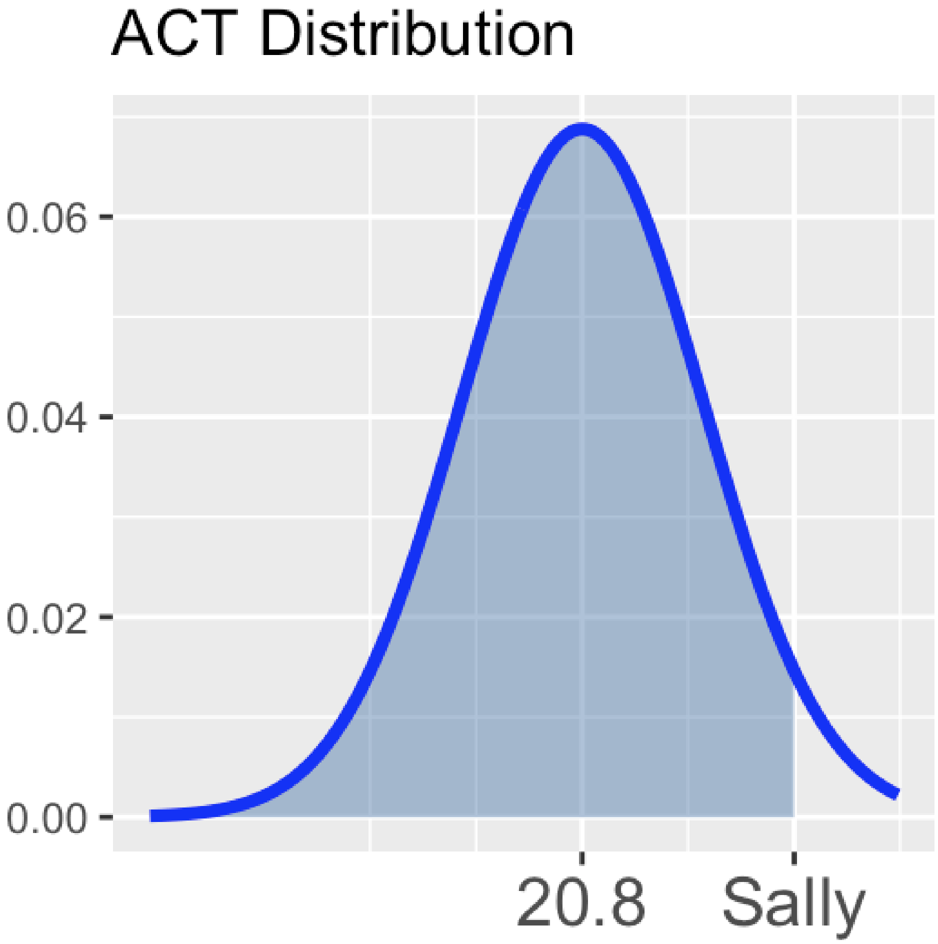
Method 2: We can compute the *percentile* corresponding to Bob's SAT score and the *percentile* corresponding to Sally's ACT score.

- **Definition:** Given an observation x from a population – the *percentile* corresponding to x is the percentage (proportion × 100) of the population which falls below x.

Bob's percentile corresponds to the shaded area in the distribution below.



Sally's percentile corresponds to the shaded area in the distribution below.



There are many ways to compute percentiles. Before the widespread availability of statistical software, people converted observed values to z-scores and then looked up the percentile in a table. Luckily R provides nice functionality for computing percentiles.

Computing Percentiles in R: If $X \sim N(\mu, \sigma)$, then

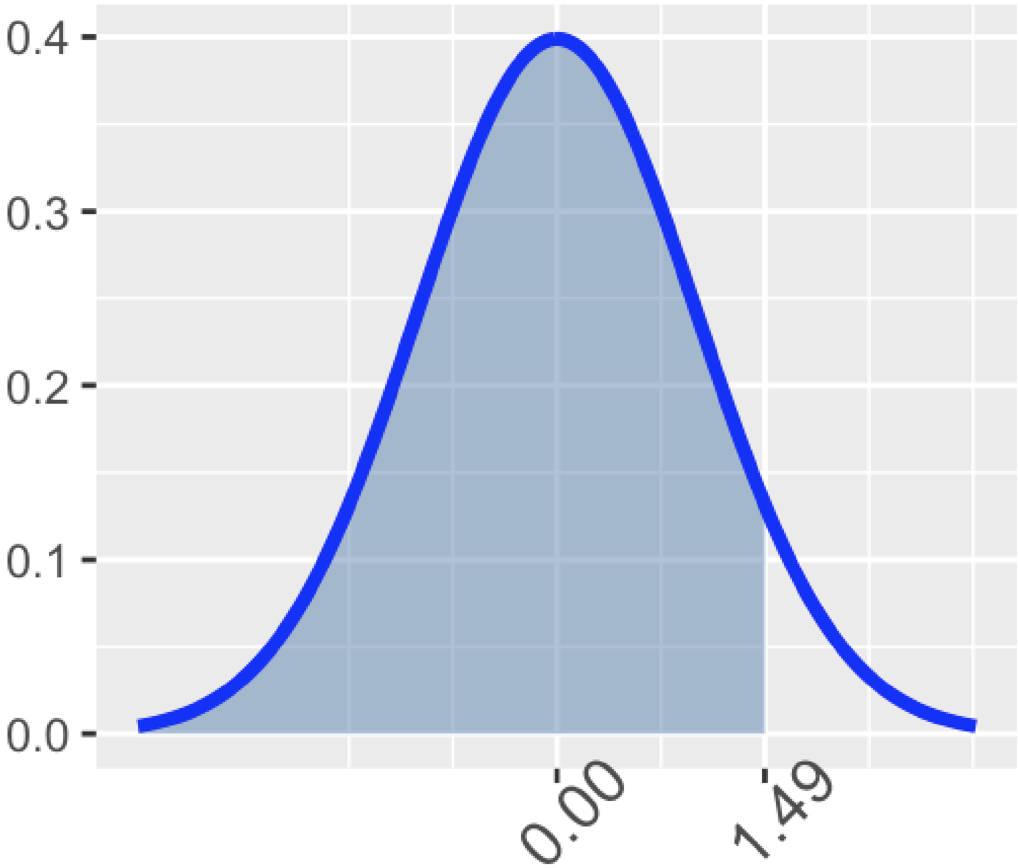
$$\mathbb{P}[X \leq q] \approx \texttt{pnorm}(\texttt{q}, \texttt{mean} = \mu, \texttt{sd} = \sigma)$$

Computing probability from a normal distribution

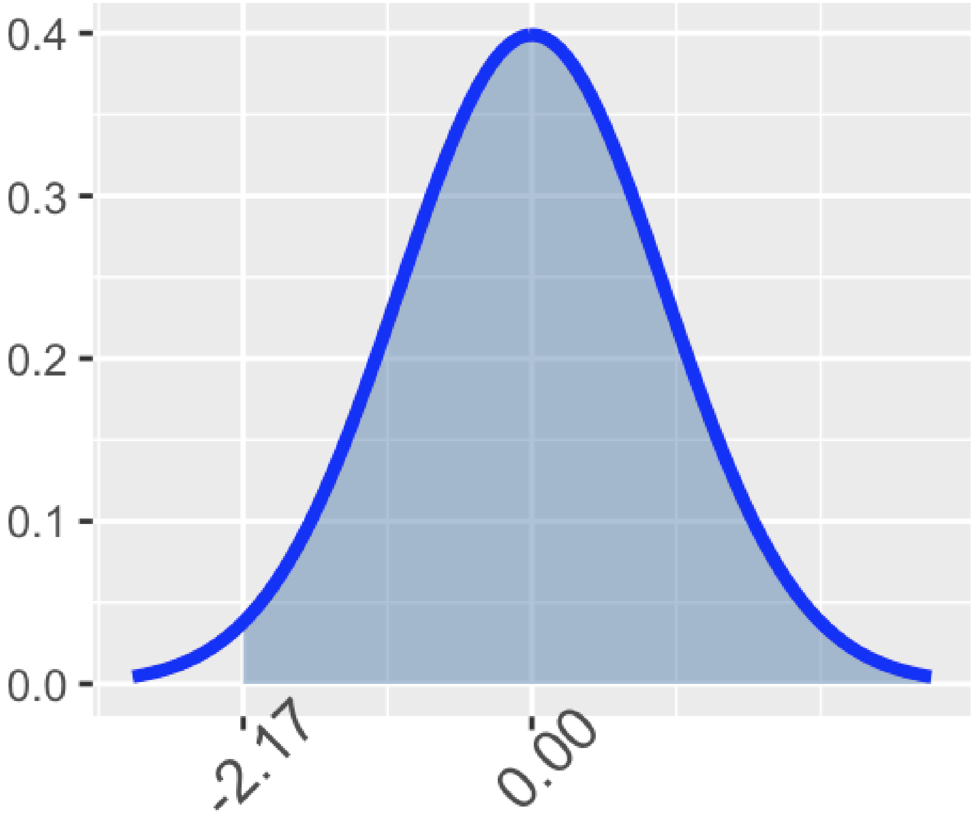
Through this section you'll be getting practice finding probabilities by using R's `pnorm()` function to compute areas. Remember that the `pnorm()` function takes three arguments – the first is a `boundary` value, the second is the `mean` of the distribution, and the third is the `standard deviation`. The value returned by `pnorm()` is the area to the left of the provided boundary value in the distribution with the mean and standard deviation you provided.

For these first few questions I'll draw pictures for you, but you should be prepared to draw your own shortly.

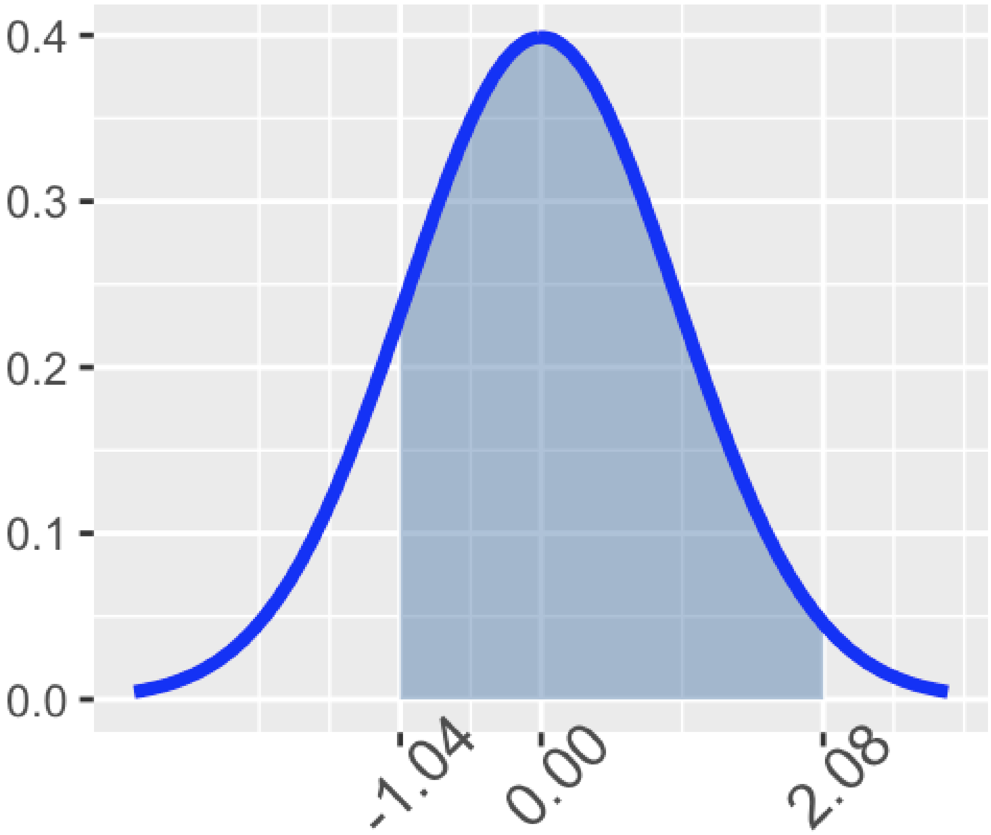
Question 1: Remember that $Z \sim N\left(\mu = 0, \sigma = 1\right)$.



Question 2: Find $\mathbb{P}\left[Z > \ \right]$.



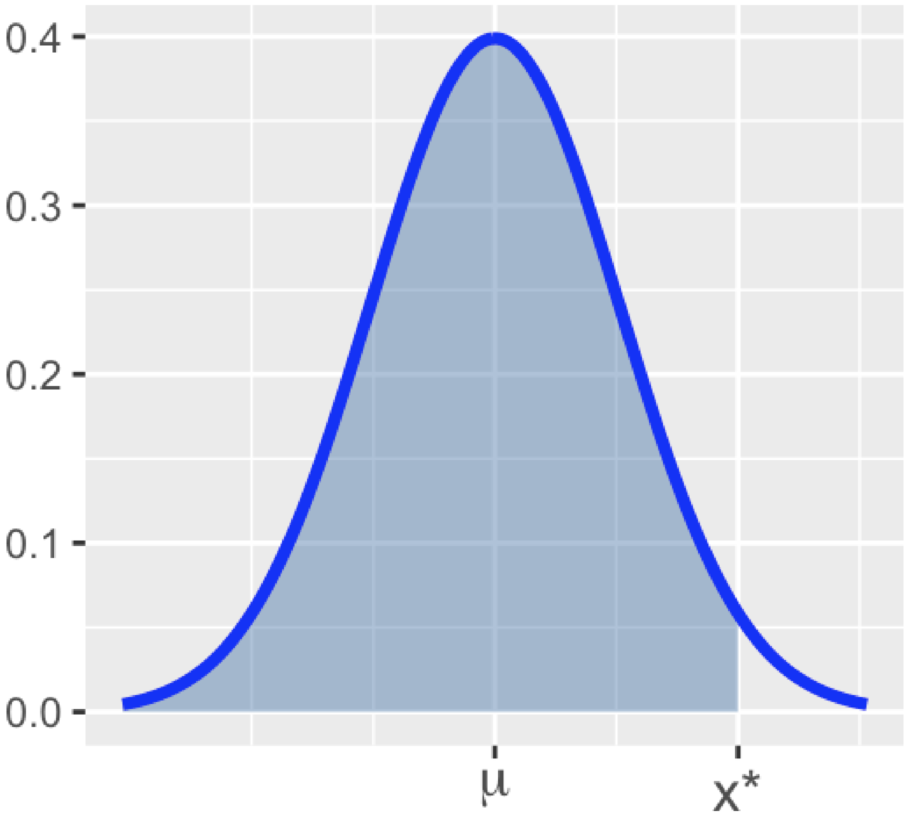
Question 3: Find $\mathbb{P}[\ < Z < \ ]$.



Through the last three problems you only worked with the standard normal distribution – that's the Z-distribution, which is $N(\mu = 0, \sigma = 1)$. We can find probabilities from arbitrary normal distributions (normal distributions with any mean and any standard deviation) using R's 'pnorm()' functionality – just supply the appropriate 'mean' and 'sd' arguments to 'pnorm()' instead of the 0 and 1 that we passed earlier.

Finding percentile cutoffs on a normal distribution

Recall from earlier that the p^{th} percentile of a random variable X is the value x^* such that $\mathbb{P}[X < x^*] = p$.



If $X \sim N(\mu, \sigma)$, then to find the cutoff x^* for which $\mathbb{P}[X < x^*] = p$, we can use R's 'qnorm()' function. Similar to `pnorm()`, this function takes three arguments. The first is the `area to the LEFT` of the desired cutoff, the second is the `mean` of the distribution, and the third is the `standard deviation` of the distribution.

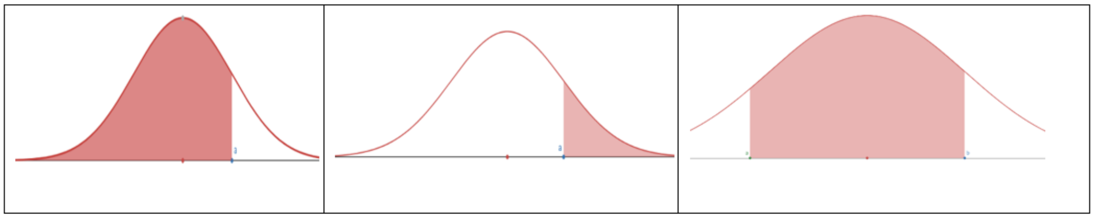
Recall from earlier that SAT scores followed $N(\mu = 1068, \sigma = 210)$ and ACT scores followed $N(\mu = 20.8, \sigma = 5.8)$. The code block below is set up to find the minimum required SAT score to fall in the 95th percentile (to do better than 95% of other test-takers). Execute the code and note the required score. Adapt the code to find the minimum ACT score required to fall into the top 10% of all ACT test takers. Does your answer seem right? How can you judge?

Using R to compute cumulative probability for $X \sim N(\mu, \sigma)$

- For $P(X < b) = P(X \leq b)$: `pnorm(b, \mu, \sigma)`
- For $P(X > a) = P(X \geq a)$: `pnorm(a, \mu, \sigma, lower.tail = FALSE)` or `1 - pnorm(a, \mu, \sigma)`
- For $P(a < X < b)$: `pnorm(b, \mu, \sigma) - pnorm(a, \mu,`

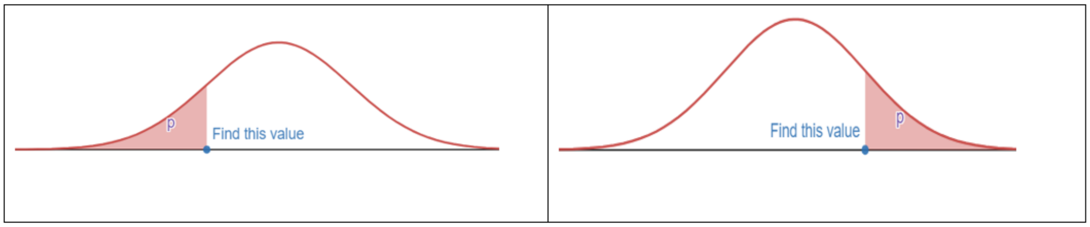
`\sigma)` or `1 - (pnorm(a, \mu, \sigma) + pnorm(b, \mu, \sigma, lower.tail = FALSE))`

- For $Z \sim N(0,1)$: the mean and SD can be omitted in 1)-3):
    - $P(Z < b)$: `pnorm(b)`



To Compute Inverse Cumulative Probability (Finding x for Given Cumulative Probability)

- Find x for $P(X < x) = p$: `qnorm(p, \mu, \sigma)`
- Find x for $P(X > x) = p$: `qnorm(1 - p, \mu, \sigma)` or `qnorm(p, \mu, \sigma, lower.tail = FALSE)`



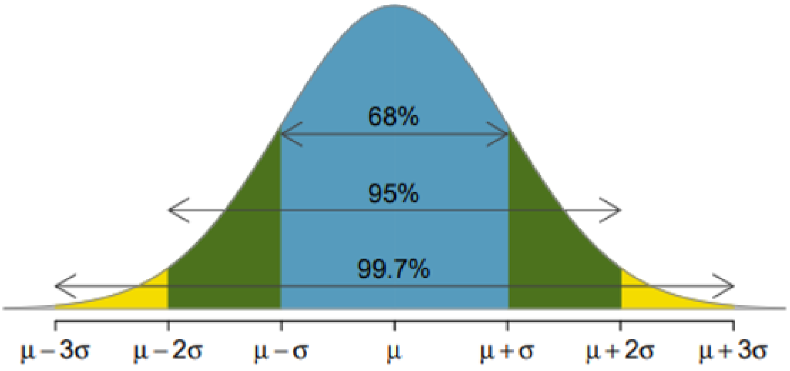
Z-score

- If $X \sim N(\mu, \sigma)$, the z-score of x is computed by $z = \frac{x-\mu}{\sigma}$.
- The z-score measures how many standard deviations of x from the mean.
- $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- $X = \mu + Z \cdot \sigma$
- $x = \mu$ if $z = 0$; $x > \mu$ if $z > 0$; $x < \mu$ if $z < 0$

Empirical Rule (68-95-99.7 Rule)

For a nearly normally distributed data, the empirical rule predicts that:

- 68% of observations fall within the first standard deviation ($\mu \pm \sigma$).
- 95% within the first two standard deviations ($\mu \pm 2\sigma$).
- 99.7% within the first three standard deviations ($\mu \pm 3\sigma$) of the mean.



2.4.3 Solved Problem

2.4.4 Exercises

Exercise 1 For $Z \sim N(0,1)$ (the standard normal distribution, the mean = 0, the standard deviation = 1), use R to find the probability and sketch the region that represents the probability.

(a). $P(Z < -1.5)$ (b). $P(Z > 1.75)$ (c). $P(-1.5 < Z < 1.75)$ (d). $P(|Z| < 2.5)$ (e). $P(Z > 1)$

Exercise 2 For $X \sim N(-3,2)$ (the normal distribution, the mean = -3, the standard deviation = 2), use R to find the probability and sketch the region that represents the probability.

1. $P(X < -3.25)$
2. $P(X > 1.75)$
3. $P(-3.25 < X < -1.25)$

Exercise 3 For $X \sim N(-3,2)$, compute the z-score of the given x:

1. $x = -3.25$
2. $x = -3$
3. $x = -1.25$

Exercise 4

(a). State the Empirical Rule.

(b). Use R to verify the Empirical Rule: find $P(|Z| < 1)$, $P(|Z| < 2)$, $P(|Z| < 3)$.

Exercise 5

The scores on a college entrance exam follow a normal distribution with a mean of 50 and standard deviation of 10. Find the probability that a student will score:

(a). Over 65

(b). Less than 25

(c). Between 33 and 68

Exercise 6

The scores on a college entrance exam follow a normal distribution with a mean of 50 and standard deviation of 10.

(a). What is the cut off score of the lowest 20%? (Round to 1 decimal)

(b). What is the cut off score of the highest 10%? (Round to 1 decimal)

Exercise 7

The hours of sleep of college students fits a normal distribution with mean of 7.2 hours and standard deviation of 1.3 hours. Find the (standardized) z-score corresponding to 6.5 hours.

Exercise 8

John scored a 92 on a test with a mean of 88 and a standard deviation of 2.7. Jessica scored an 86 on a test with a mean of 82 and a standard deviation of 1.8. Find the Z-scores for John's and Jessica's test scores and use them to determine who did better on their test relativ*e to their class.

Exercise 9

The score data of the verbal portion of the Graduate Record Examination (GRE) is approximately normally distributed with a mean of 462 points and a standard deviation of 119 points. Fill in the following blanks: approximately

(a) 68% of students who took the verbal portion of the GRE scored between ___________ and _____________

(b) 95% of students who took the verbal portion of the GRE scored between _________ and ____________

(c) 99.7% of students who took the verbal portion of the GRE scored between _________ and ____________

Chapter 3

Inference for Proportion

3.1 Point Estimates and Sampling Variability

3.1.1 Objectives

By the end of this unit, students will be able to:

- Understand the meaning of sampling distributions.
- Apply the central limit theorem to define the sampling distribution of a sample proportion.
- Identify the conditions needed for the central limit theorem to apply for sample proportions.

3.1.2 Overview

Basic Terms

- Sample proportion $\hat{p} = \frac{x}{n}$

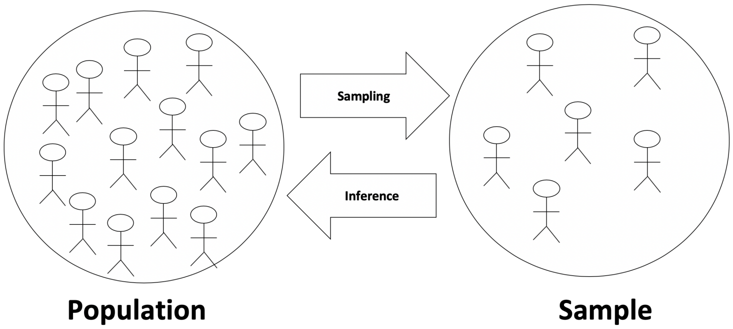
- Sample proportion $\hat{p}$ is the unbiased point estimator for the population proportion p
- A value of $\hat{p}$ is a point estimate
- Error $= \hat{p} - p$

What is statistical inference?

Statistical inference is *the process of making claims about a population based on information from a sample* of data.

Typically, the data represent only a small portion of the larger group which you'd like to summarize. For example, you might be interested in how a drug treats diabetes. Your interest is in how the drug treats all people with diabetes, not just the few dozen people in your study.

At first glance, the logic of statistical inference seems to be backwards, but as you become more familiar with the steps in the process, the logic will make much more sense.



In this section we'll begin investigating the true power of statistics – using sample data to make accurate claims about a population (even when we don't have access to the entire population). We start by exploring the connection between a *Population Distribution* and the distribution of sample means, often called the *Sampling Distribution*. We'll do this through a series of simple, interactive code blocks which you will run and use to answer questions.

Exploring the connection between population and sampling distributions

Start by viewing the following following video from the New York Times.

So the video claimed that the sampling distribution can help us answer questions about the population. This is really important because, as we mentioned in our first tutorial, Census is almost always impossible. Use the code blocks below to explore the connection between the population and the sampling distribution for various different populations. Note that you do not need to understand all of the code contained in the code blocks – you should focus, instead, on the pictures resulting each time you run the code. In general, you are invited to change the first few lines of code in each block, and you are not expected to look at the remaining code.

Sampling Distribution

Point Estimate error

Suppose a poll suggested the US President's approval rating is 45%. We would consider 45% to be a **point estimate** of the approval rating we might see if we collected responses from the entire population. This entire-population response proportion is generally referred to as the **parameter** of interest. When the parameter is a proportion, it is often denoted by p, and we often refer to the sample proportion as $\hat{p}$ (pronounced p - hat). Unless we collect responses from every individual in the population, p remains unknown, and we use $\hat{p}$ as our estimate of p. The difference we observe from the poll versus the parameter is called the **error** in the estimate. Generally, the error consists of two aspects: sampling error and bias.

- **Sampling error**, sometimes called sampling uncertainty, describes how much an estimate will tend to vary from one sample to the next. For instance, the estimate from one sample might be 1% too low while in another it may be 3% too high. Much of statistics, including much of this book, is focused on understanding and quantifying sampling error, and we will find it useful to consider a sample's size to help us quantify this error; the **sample size** is often represented by the letter n.

- **Bias** describes a systematic tendency to over- or under-estimate the true population value. For example, if we were taking a student poll asking about support for a new college stadium, we'd probably get a biased estimate of the stadium's level of student support by wording the question as, *Do you support your school by supporting funding for the new stadium?* We try to minimize bias through thoughtful data collection procedures, which were discussed in Chapter 1 and are the topic of many other books.

Point Estimate for Proportion

A point estimate is the value of a statistic (based on a sample) that estimates the value of a population parameter.

Suppose we want to estimate the proportion of adult Americans who believe that immigration is a good thing for the U.S. It is unreasonable to expect that we could survey every adult American. Instead, we use a sample of adult Americans to arrive at an estimate of the proportion. We call this estimate a point estimate.

The sample proportion $\hat{p}$

We now study categorical data and draw inference on the proportion, or percentage, of the population with a specific characteristic.
If we call a given categorical characteristic in the population "success" then the sample proportion of successes, p, is:

$$\hat{p} = \frac{x}{n}$$

Where x is the number of individuals in the sample with a specified characteristic and n is the sample size.

Example 1: The Gallup Organization conducted a poll in which a simple random sample of 1,520 adults, living in all 50 U.S. states and the District of

Columbia were asked the following question. “Thinking now about immigrants – that is, people who come from other countries to live here in the United States, in your view, do you think legal immigration is a good thing or a bad thing for this country today?” If 1135 responded “Yes”.

Obtain a point estimate for the proportion of Americans 18 and older who believe that immigration is good for the US.

Solution:

$$\hat{p} = \frac{1135}{1520} = 0.747$$

We estimate for the proportion of Americans 18 and older who believe that immigration is good for the US is 74.7%
Note: We agree to round proportions to three decimal places.

Sampling distribution of sample proportions $\hat{p}$

Central Limit Theorem (for Sampling Distribution of Sample Proportions)

When observations are independent (take random samples of fixed size n without replacement); the sample size n is large enough, i.e. $np \geq 10$ and $n(1-p) \geq 10$; and sample size $n < 10\%$ of the population size then the sample proportion $\hat{p}$ is approximately normal with mean $= p$ and standard deviation $= \sqrt{\frac{p(1-p)}{n}}$: $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$

That is, $z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$.

Notes:

- The requirement that the sample size $n < 10\%$ of population size is to have the standard deviation $\sqrt{\frac{p(1-p)}{n}}$.
- If the sample size $n < 10\%$ of population size, then the $\sqrt{\frac{p(1-p)}{n}}$ is overly estimated SD, and the SD will typically adjust by a factor $\sqrt{\frac{N-n}{N-1}}$,i.e. $\sqrt{\frac{N-n}{N-1}} \times \sqrt{\frac{p(1-p)}{n}}$.
- When using $\hat{p}$ to estimate p, the Standard Error of $\hat{p}$ is the standard deviation of its sampling distribution: $S.E._{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- When p is unknown, $\hat{p}$ is used to replace p then check $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ (success and failure condition) estimated $S.E. \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

The Central Limit Theorem for proportions

The Central Limit Theorem for proportions states that if n is large enough, then:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Therefore, a different random sample of adult Americans might result in a different point estimate of the population proportion, such as $\hat{p} = 0.71, \hat{p} = 0.78, \ldots$.

If the method used to select the sample of Americans was done appropriately, both point estimates would be good guesses of the population proportion. Due to variability in the sample proportion, we need to report a range (or *interval*) of values, including a measure of the likelihood that the interval includes the unknown population proportion.

3.1.3 Solved Problem

Suppose the proportion of American adults who support the expansion of solar energy is p = 0.88, which is our parameter of interest. Is a randomly selected American adult more or less likely to support the expansion of solar energy?

Answer: More likely.

Let's see an example of sampling

Suppose that you don't have access to the population of all American adults, which is a quite likely scenario. In order to estimate the proportion of American adults who support solar power expansion, you might sample from the population and use your sample proportion as the best guess for the unknown population proportion.

We will simulate a data to play the role of the population. As discusses above, we will assume that 88% of the population support the expansion and the remaining 12% don't.

```r
pop_size <- 250000000
possible_entries_solar <- c(rep("support", 0.88 * pop_size),
                            rep("not", 0.12 * pop_size))
```

First we will sample, without replacement, 1000 American adults from the population, and record whether they support or not solar power expansion.

```r
sampled_entries <- sample(possible_entries_solar,
                          size = 1000, replace = F)
```

Second we will find the sample proportion.

```
sum(sampled_entries == "support")/1000
```

```
## [1] 0.869
```

Interesting thing about sampling from a population is that its always random. So the first sample might give a completely different sample proportion compared to the second sample, third sample and so on.

For example, we will perform the same sampling using the same code and we will obtain a different sample proportion.

```
sampled_entries <- sample(possible_entries_solar,
                          size = 1000, replace = F)
sum(sampled_entries == "support")/1000
```

```
## [1] 0.897
```

```
sampled_entries <- sample(possible_entries_solar,
                          size = 1000, replace = F)
sum(sampled_entries == "support")/1000
```

```
## [1] 0.888
```

> Run the code to see that you obtain a different result everytime you perform sampling.

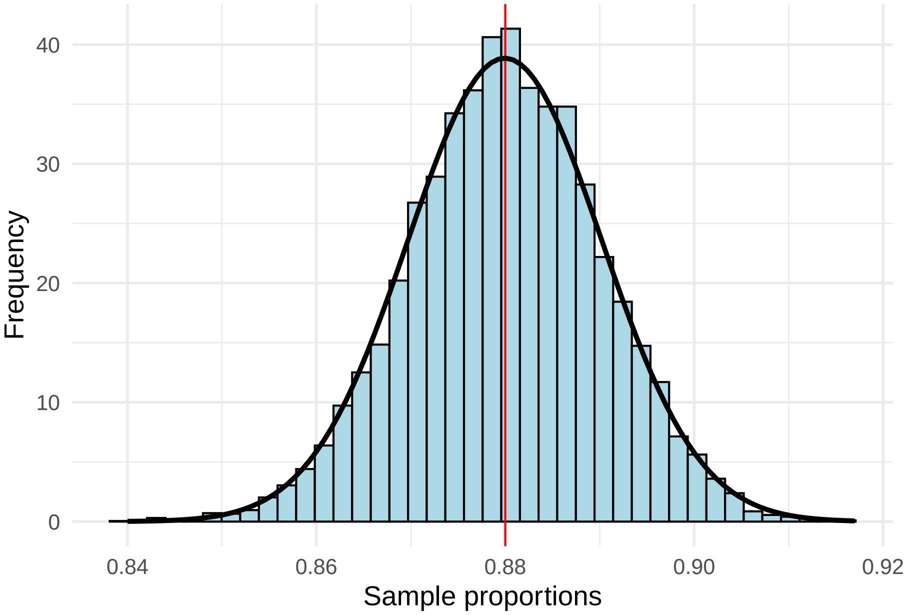
Third we will use the fact that the sample proportion changes with every sample and collect multiple samples and find the sample proportion for all of those samples. This will help us to create a distribution for the sample proportion to understand the spread, center and shape of the sample proportions.

```
library(tidyverse)
set.seed(123)
# Creating 10000 different sample proportions
phat <- rep(NA, 10000)
for(i in 1:10000){
  sampled_entries <- sample(possible_entries_solar, size = 1000, replace = F)
  phat[i] <- sum(sampled_entries == "support") / 1000
}

sampling <- tibble(phat = phat)

# Plot
```

```
ggplot(sampling, aes(x = phat)) +
  geom_histogram(aes(y=..density..),bins = 40,col = "black", fill = "lightblue") +
  geom_vline(xintercept = 0.88, col = "red")+
  theme_minimal(base_size = 14) +
  labs(x = "Sample proportions", y = "Frequency")+
  stat_function(fun = dnorm, args = list(mean = mean(phat), sd = sd(phat)), size = 1.2)
```



This distribution of sample proportions is called a sampling distribution. We can characterize this sampling distribution as follows:

Center. The center of the distribution is $\bar{xp} = 0.880$, which is the same as the parameter. Notice that the simulation mimicked a simple random sample of the population, which is a straightforward sampling strategy that helps avoid sampling bias.

Spread. The standard deviation of the distribution is $s_{\hat{p}} = 0.010$. When we're talking about a sampling distribution or the variability of a point estimate, we typically use the term **standard error** rather than *standard deviation*, and the notation $SE_{\hat{p}}$ is used for the standard error associated with the sample proportion.

Shape. The distribution is symmetric and bell-shaped, and it resembles a normal distribution.

These findings are encouraging! When the population proportion is p = 0.88 and the sample size is n = 1000, the sample proportion ˆp tends to give a pretty good estimate of the population proportion. We also have the interesting observation that the histogram resembles a normal distribution.

3.1.4 Exercises

Exercise 1

In a random sample with size $n = 9000$, the count of "yes" is $x = 250$.

(a) Compute the sample proportion $\hat{p} = \frac{x}{n}$.

(b) Compute the estimated standard error of the sample proportion $S.E. \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Exercise 2

In a random sample of 765 adults in the U.S., 322 say they could not cover $600 unexpected expense without borrowing money or going to debt.

(a) What population is under consideration in the data set?

(b) What parameter is being estimated?

(c) Compute a point estimate for the parameter using the given information above?

(d) What is the estimated standard error?

Exercise 3

Of all freshmen at a large college, 19% made the dean's list.

(a) What is the value of the interested parameter? State the sampling distribution of sample proportion for sample size 90.

(b) If a random sample of 90 freshmen selected 14 made the dean's list. Compute the sample proportion and the Z-score.

(c) If a random sample of 90 freshmen selected 20 made the dean's list. Compute the sample proportion and the Z-score.

(d) What is the probability that at most 14 of selected 90 freshmen made the dean's list?

(e) What is the probability that between 14 to 20 students of selected 90 freshmen made the dean's list?

3.2 Confidence Interval

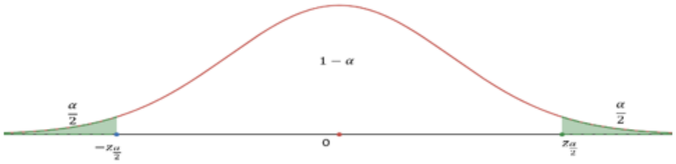
3.2.1 Objectives

- Understand the meaning of sampling distributions.

- Apply the central limit theorem to define the sampling distribution of a sample proportion.
- Identify the conditions needed for the central limit theorem to apply for sample proportions.
- Construct and interpret confidence intervals for the population proportion.

3.2.2 Overview

About critical value z^* or $z_{\alpha/2}$: For $N(0,1)$, $z_{\alpha/2}$ is the cut-off point with upper tail of probability $\alpha/2$



How to find $z_{\alpha/2}$:

(1) For $100(1-\alpha)\%$ confidence level, find α then $\alpha/2$

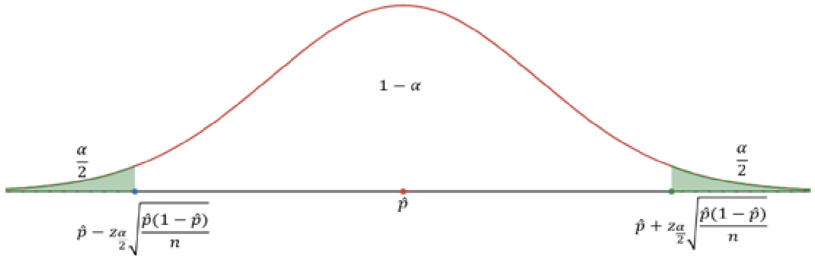
(2) Use R: $qnorm\left(\frac{\alpha}{2}, \text{lower.tail} = \text{FALSE}\right)$ or $qnorm\left(1-\frac{\alpha}{2}\right)$

Common $z_{\alpha/2}$ values:

Confidence level	α	$z_{\alpha/2}$
90%	0.10	$z_{0.05} = 1.644854 \approx 1.645$
95%	0.05	$z_{0.025} = 1.959964 \approx 1.96$
98%	0.02	$z_{0.01} = 2.326348 \approx 2.326$
99%	0.01	$z_{0.005} = 2.575829 \approx 2.576$

Construct $100(1-\alpha)\%$ confidence interval: Use $\hat{p}$, n, and $z_{\alpha/2}$

$\hat{p} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (or $(\hat{p} - z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$)


$1-\alpha$
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$\hat{p}$
$\hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Margin of Error (M.E.)

$M.E. = z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

So, C.I.: point estimate $\pm M.E.$

Note: the point estimate is the middle point; the $M.E.$ = half of the length of C.I.)

Interpretation of C.I.:

With the confidence level of $100(1-\alpha)\%$ and a sample proportion $\hat{p}$ with sample size n, we are $100(1-\alpha)\%$ confident that the population proportion p is in the confidence interval $(\hat{p} - z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$

Minimum sample size to guarantee the specified $M.E. \leq a$

For known p: $n = \text{ceiling}\left[\frac{p(1-p) \times z^2_{\alpha/2}}{a^2}\right]$

For unknown p: $n = \text{ceiling}\left[\frac{1}{4} \times \frac{z^2_{\alpha/2}}{a^2}\right]$

3.2.3 Solved Problem

3.2.4 Exercises

Exercise 1

(a) Construct a 95% confidence interval using a sample proportion $\hat{p} = 0.3$ and sample size $n = 1000$.

(b) Construct a 90% confidence interval using a sample proportion $\hat{p} = 0.3$ and sample size $n = 1000$.

(c) Construct a 95% confidence interval using a sample proportion $\hat{p} = 0.3$ and sample size $n = 100$.

Exercise 2. Circle the proper choices.

(a) The confidence interval is ________________(wider/narrower) if the sample size is increasing.

(b) The confidence interval is ________________(wider/narrower) if the confidence level is increasing.

Exercise 3

(a) Construct a 98% confidence interval using a sample proportion 45% and standard error 1.2%. (Assume that the CLT can be applied)

(b) Compute the margin of error using the same information of (a).

Exercise 4

A website is trying to increase registration of first-time visitors using a new site design. Of 752 randomly sampled visitors over a month who saw the new design, 64 registered.

(a) Compute the sample proportion.

(b) Compute the standard error.

(c) Construct and interpret a 90% confidence interval for the fraction of first-time visitors of the site who would register under the new design.

Exercise 5. For a confidence interval of proportion $(0.291, 0.309)$ find the following:

(a) The sample proportion that was used to create this C.I.

(b) The M.E. (Margin of Error)

Exercise 6

A public health survey is going to estimate the proportion of a population p having defective vision. How many persons should be examined if the public health commissioner wishes to be 95% certain that the margin of error is below 0.04 when:

(a) p is known to be about 0.45.

(b) There is no knowledge about the value of p?

Note. Similar result for mean:

- CLT: In random sampling from a population with mean μ and standard deviation σ, when the sample size n is large ($n \geq 30$), the distribution of sample mean $\bar{X}$ is approximately normal: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.
- Use $\bar{x}$ as point estimate for μ.
- $S.E. = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$
- $100(1-\alpha)\%$ confidence interval for mean: $\bar{x} \pm z_{\alpha/2} \times \sqrt{\frac{s}{n}}$ or $(\bar{x} - z_{\alpha/2} \times \sqrt{\frac{s}{n}}, \bar{x} + z_{\alpha/2} \times \sqrt{\frac{s}{n}})$
- $M.E. = z_{\alpha/2} \times S.E. = z_{\alpha/2} \times \sqrt{\frac{s}{n}}$

Exercise 7

The GSS (General Social Survey) asked the question: "For how many days during the past 30 days was your mental health not good (stress, depression, with emotions)?" Based on responses from 1151 US residents, a 95% confidence interval (3.40, 4.24) (days) was reported in 2014.

(a) Determine the sample mean (days).

(b) Determine the margin of error (M.E.).

(c) What is the value of $z_{\alpha/2}$ for 95% confidence level?

(d) Write a sentence to interpret this confidence interval.

Confidence Intervals for Proportion

A confidence interval for an unknown parameter consists of an interval of numbers based on a point estimate.

The level of confidence represents the expected proportion of intervals that will contain the parameter if a large number of different samples is obtained.

The level of confidence is denoted $(1 - \alpha) \times 100\%$. For example, a 95% level of confidence ($\alpha = 0.05$) implies that if 100 different confidence intervals are constructed, each based on a different sample from the same population, we will expect 95 of the intervals to contain the parameter and 5 not to include the parameter.

Confidence interval estimates for the population proportion are of the form:

Point estimate ± margin of error.

The margin of error of a confidence interval estimate of a parameter is a measure of how accurate the point estimate is. The margin of error depends on three factors:

- **Level of confidence:** As the level of confidence increases, the margin of error also increases.
- **Sample size:** As the size of the random sample increases, the margin of error decreases.
- **Standard deviation of the population:** The more spread there is in the population, the wider our interval will be for a given level of confidence.
- **Point estimate**
  - statistics vs. parameter
  - point estimate for population proportion is $\hat{p} = \frac{x}{n}$
  - may also represent a probability of a binomial distribution, such as p=0.5 for a fair coin.
- **Confidence interval** for an unknown parameter
- **Margin of error**
- Interpretation of what is meant by being “95%” confident (think simulations)

- Formula for confidence interval for population proportion, p.
- Know z - multipliers for 90%, 95%, and 99% confidence intervals.
- Requirements:
  - Random sample (independent)
  - AND large sample size (at least 10 successes & 10 Failures) np>10 and n(1-p) > 10
  - Calculating sample size to obtained desired margin of error E.
    - ∗ Using educated guess for population proportion.
    - ∗ Using p=0.5 for conservative (large) sample size.
  - Round up to nearest integer.

Suppose that a simple random sample of size n is taken from a population.
A $(1-\alpha)\times 100\%$ confidence interval for p is given by the following quantities:

$\hat{p} \pm z^{*} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Where p = Theoretical or “True” population proportion

Confidence			
Level	$1-\alpha\times 100\%$	α	$Z^{*}=Z_{\alpha/2}$
0.90	0.10	1.645	
0.95	0.05	1.96	
0.98	0.02	2.33	
0.99	0.01	2.58	

The Z-table is used to find the critical values, Z^{*}, for confidence intervals on the true proportion p. This abbreviated table gives the most common z-scores for the centered values of confidence for the normal curve.

Example 2:

In July of 2008, a University Poll asked 1783 registered voters nationwide whether they favored or opposed the death penalty for persons convicted of murder 1123 were in favor.

Obtain a 90% confidence interval for the proportion of registered voters nationwide who are in favor of the death penalty for persons convicted of murder.

Solution:
$\hat{p} = \frac{1123}{1783} = 0.63$

Where:

n = 1783

$$np \;\&\; n(1-p) > 10$$

Lower bound: $0.63 - 1.645 \times \sqrt{\frac{0.63(1-0.63)}{1783}} \approx 0.61$
Upper bound: $0.63 + 1.645 \times \sqrt{\frac{0.63(1-0.63)}{1783}} \approx 0.65$

We are 90% confident that the proportion of registered voters who are in favor of the death penalty for those convicted of murder is between 0.61 and 0.65.

Estimating the margin of error on p for a given confidence level

Consider the scenarios for the product of proportion and complements for the margin of error. What happens as the proportion changes?

$$(1-\alpha)\%CI \text{ on } p = \hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where:

- $\hat{p}$ = Point estimate
- $Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ = Margin of Error = $Z_{critical} \times StandardError$

Determine the Sample Size Necessary for Estimating a Population Proportion within a Specified Margin of Error

Two possible solutions:

1. Use an estimate of p based on a pilot study
2. Use the value of p which gives the largest possible value of n for a given confidence level & margin of error.

Sample Size Needed for Estimating the Population Proportion

The sample size required to obtain a $(1-\alpha)$ - 100% confidence interval for p with a margin of error E is given by:

$$n = \frac{p(1-p)\left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2}{\textit{rounded} \text{ up to the next integer}}, \text{ where } p \text{ is a prior estimate of } p.$$

n	Approximate Margin of Error
500	0.045 (or 4.5%)
800	0.035 (or 3.5%)
1000	0.032 (or 3.2%)

n	Approximate Margin of Error
1500	0.026 (or 2.6%)

If a prior estimate on the proportion is unavailable, using $\hat{p} = 0.5$ will give the following estimate for the sample size:

$$n = 0.25 \times \left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2$$

Gallup and other polling agencies report at the 95% confidence level and assume p = 0.5 in calculating the margin of error:

Example 3

The statistic presented below appeared in the weekly magazine TIME, August 23, 1993, under the article *Danger in the Safety Zone*. Consider the tiny print "From telephone poll of 500 adult Americans taken for TIME/CNN. Margin of error is ±0.45".

Do you favor the death penalty?

YES	NO
77%	17%

From a telephone poll of 500 adult Americans taken for TIME/CNN on Aug. 12 by Yankelovich Partners, Inc. Margin of error is ±0.45

Explain how the margin of error can be calculated.

Solution: Margin of Error in the article (under 95% confidence) is about:

$$1.96\sqrt{\frac{0.5 \times 0.5}{500}} \approx \frac{1}{\sqrt{500}} = 0.04472 \approx 4.5\%$$

Example 2 Illustrating the Meaning of Level of Confidence Using Simulation :

Let's illustrate what "95% confidence" means in a 95% confidence interval in another way. We will simulate obtaining 200 different random samples of size n=50, m=50 i.e., n equals 50 from a population with p=0.7, p equals 0.7. Figure 4 shows the confidence intervals in groups of 100. A green interval is a 95% confidence interval that includes the population proportion, 0.7. A red interval is a confidence interval that does not include the population proportion. (For now, ignore the blue intervals.) Notice that the red intervals that do not capture the population proportion 0.7 have centers that are far away (more than

1.96 standard errors) from 0.7. Of the 200 confidence intervals obtained, 10 (the red intervals) do not include the population proportion. For example, the first interval to miss has a sample proportion that is too small to result in an interval that captures 0.7.

(Note: The actual image illustrating the confidence intervals would be embedded here using ![ ]() if the image file was provided.)

Appeal: Simulating Confidence Intervals (rossmanchance.com)

A 95% level of confidence means that 95% of all possible samples result in confidence intervals that include the parameter (and 5% of all possible samples result in confidence intervals that do not include the parameter).

Caution!

A 95% confidence interval does *not* mean that there is a 95% probability that the interval contains the parameter (such as p). Remember, probability describes the likelihood of *undetermined* events. Therefore, it does not make sense to talk about the probability that the interval contains the parameter since the parameter is a *fixed* value. Thinking of this way: If a coin and obtain a head. If I ask you to determine the probability that the flip resulted in a head, it would not be 0.5, because the outcome has already been determined. Instead, the probability is 0 or 1. Confidence intervals work the same way. Because p or μ is already determined, we do not say that there is a 95% probability that the interval contains μ.

Exercises

1. As a potential worldwide pandemic, avian influenza H5N1 (commonly called the bird flu) poses a serious health risk. As of January 24, 2012, there have been 583 human cases of this virus in the world. Of these cases, 344 have resulted in death. Consider the outcomes of these cases as a random sample of all possible outcomes.

    a. Find a point estimate for the proportion of people who would die if infected with the bird flu.
    b. Construct a 90% confidence interval for the proportion of cases that would be expected to result in death if a pandemic occurred.
    c. Interpret the confidence interval.

2. A sociologist wanted to determine the percentage of residents of America that only speak English at home. What size sample should be obtained if she wishes her estimate to be within 3 percentage points with 90% confidence assuming she uses the estimate obtained from the Census 2000 Supplementary Survey of 82.4%?

3. Nitrates are groundwater contaminants derived from fertilizer, septic tank seepage, and other sewage. Nitrate poisoning is particularly hazardous to

infants under the age of 6 months. The Maximum Contaminant Level (MCL) is the highest level of a contaminant that government allows in drinking water. For nitrates, the MCL is 10 mg/L. The health department wants to know the proportion of wells in Madison County that have nitrate levels above the MCL. A worker has been assigned to take a simple random sample of wells in the county, measure the nitrate levels, and assess compliance. What size sample should the health department obtain if the estimate is desired to be within 2 percent with 95% confidence if:

a. there is no prior information available?
b. a study conducted two years ago showed that approximately 7% of the wells in Madison County had nitrate levels exceeding the MCL.

Confidence interval for two-sided test

For $H_0 : p = p_0$, $H_a : p \neq p_0$ (two-sided) and significance level α, construct a $100(1-\alpha)\%$ confidence interval: $\hat{p} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Make decision:

- If p_0 is not in the C.I. then we reject H_0 in favor of H_a;
- If p_0 is in the C.I. then we fail to reject H_0.

Type I and Type II errors

Truth	Test conclusion	do not reject H_0	reject H_0 in favor of H_A
H_0 true		okay	Type 1 Error
H_A true		Type 2 Error	okay

Example

Claim: $p < 0.32$, $n = 120$, $p = 0.233$, $\alpha = 0.05$

Solution

(a) Formulate the hypotheses: $H_0 : p = 0.32$, $H_a : p < 0.32$

(b) Compute the z-test statistic: $z = \frac{0.233-0.32}{\sqrt{\frac{0.32\times0.68}{120}}} = -2.043057$ (R-code: `(0.233-0.32)/sqrt(0.32*0.68/120)`)

(c) Compute the p-value: p-value $= P(Z < -2.043057) = 0.020534$ (R-code: `pnorm(-2.043057)`)

(d) Draw conclusion: since the p-value is less than α, $0.0205 < 0.05$, we reject the null hypothesis and support the alternative hypothesis: We have strong evidence to support the claim that the proportion is less than 0.32 ($p < 0.32$).

Do the following hypotheses testing for given claims, sample size, sample proportion, and significance level:

1. Claim: $p < 0.56$, $n = 86$, $\hat{p} = 0.387$, $\alpha = 0.10$
2. Claim: $p > 0.75$, $n = 228$, $\hat{p} = 0.818$, $\alpha = 0.02$
3. Claim: $p \neq 0.60$, $n = 77$, $\hat{p} = 0.709$, $\alpha = 0.02$
4. Claim: $p \neq 0.60$, $n = 77$, $\hat{p} = 0.565$, $\alpha = 0.02$

3.3 Hypothesis Testing for p

3.3.1 Objectives

By the end of this unit, students will be able to:

- Formulate claims about a population proportion in the form of a null hypothesis and alternative hypothesis.
- Identify the types of errors associated with statistical hypothesis testing.
- Conduct a large-sample z-test about population proportions.

3.3.2 Overview

Hypothesis Testing Steps

1. Formulate the null hypothesis and the alternative hypothesis
2. Use sample to compute p-value (or compute the test statistic and use the rejection region based on given significance level)
3. Make decision for specified significance level

Formulation of Hypothesis

The null hypothesis: $H_0 : p = p_0$

The alternative hypothesis is one of the three:

- $H_a : p > p_0$ (right-sided) (or right-tailed)
- $H_a : p < p_0$ (left-sided) (or left-tailed)
- $H_a : p \neq p_0$ (two-sided) (or two-tailed)

What is p-value

A p-value is the calculated probability of observing data at least as favorable to the alternative hypothesis, assuming that the null hypothesis is true (in this section, we use the observed $\hat{p}$, and under assumption the sample proportion $\sim N(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}))$

How to compute p-value

Use a sample of sample size n with sample proportion $\hat{p} = p_1$, under the assumption that:

$\hat{p} \sim N(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}))$.

Let $z_1 = \frac{p_1 - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ (this is called the z-test statistic)

- For left-sided test, p-value is $P(\hat{p} < p_1)$, or $P(Z < z_1)$,

use pnorm $(p_1, p_0, \sqrt{\frac{p_0(1-p_0)}{n}})$ or pnorm (z_1)

- For right-sided test, p-value is $P(\hat{p} > p_1)$, or $P(Z > z_1)$, use pnorm$(p_1, p_0, \sqrt{\frac{p_0(1-p_0)}{n}}, lower.tail = FALSE)$ or

pnorm $(z_1, lower.tail = FALSE)$

- For two-sided test, p-value is $P(|Z| > |z_1|)$, use 2*pnorm $(-|z_1|)$ or

2*pnorm $(|z_1|, lower.tail = FALSE)$

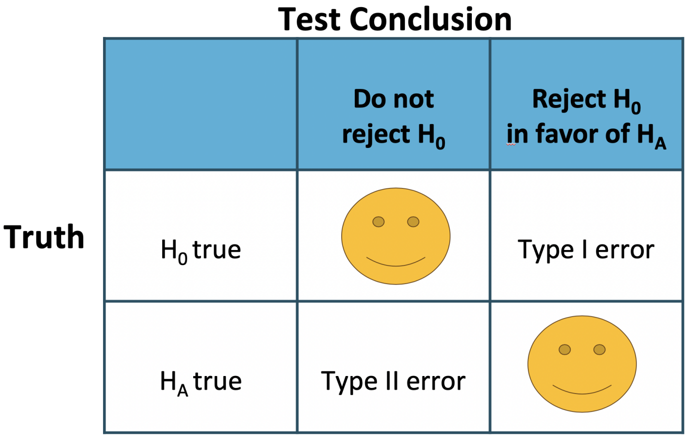
Make decision by comparing p-value with significance level α

- If p-value $\leq \alpha$, then we have enough evidence to reject H_0 and substantiate H_a;
- If p-value $> \alpha$, then we do not have enough evidence to reject H_0
- The default value of significance level is $\alpha = 0.05$

Decision errors

Hypothesis tests are not flawless: we can make an incorrect decision in a statistical hypothesis test based on the data. For example, in the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free. One key distinction with statistical hypothesis tests is that we have the tools necessary to probabilistically quantify how often we make errors in our conclusions.

Recall that there are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios, which are summarized in the picture below.

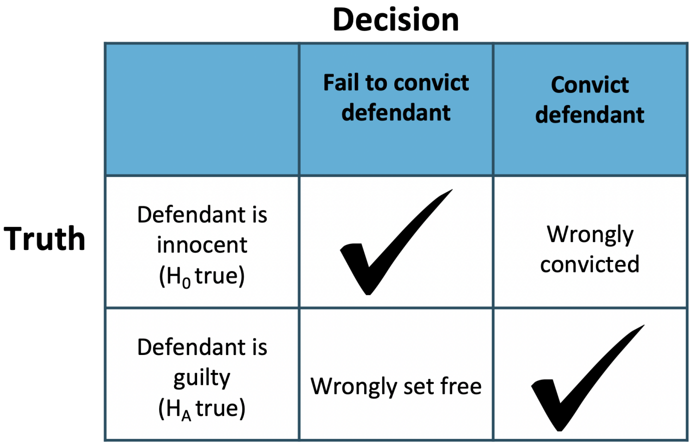


Test Conclusion

Truth		Do not reject H_0	Reject H_0 in favor of H_A
	H_0 true	☺	Type I error
	H_A true	Type II error	☺

- **Type 1 Error** is rejecting the null hypothesis when H_0 is actually true.
- **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.

In a US court, the defendant is either innocent (H_0) or guilty (H_A). What does a Type 1 Error represent in this context? What does a Type 2 Error represent?



Decision

Truth		Fail to convict defendant	Convict defendant
	Defendant is innocent (H_0 true)	✓	Wrongly convicted
	Defendant is guilty (H_A true)	Wrongly set free	✓

The logic of hypothesis testing shares many key elements with the US judicial system. The jury does not know whether the defendant committed the crime, but they must decide whether or not to convict the individual. The jury is presented with evidence, akin to data, and asked whether the evidence is consistent with innocence. If the evidence is outside of what would be expected, the defendant is charged with a crime.

One-sided hypothesis tests

So far we have only considered what are called **two-sided hypothesis tests**, where we care about detecting whether p is either above or below some null value p_0 i.e. $H_A : p \neq p_0$. There is a second type of hypothesis test called a **one-sided hypothesis test**. For a one sided hypothesis test, the hypotheses take one of the following forms:

- There's only value in detecting if the population parameter is *less than* some value p_0. In this case, the alternative hypothesis is written as $p < p_0$ i.e. $H_A : p < p_0$ for some null value p_0.
- There's only value in detecting if the population parameter is *more than* some value p_0. In this case, the alternative hypothesis is written as $p > p_0$ i.e. $H_A : p > p_0$ for some null value p_0.

While we adjust the form of the alternative hypothesis, we continue to write the null hypothesis using an equals-sign i.e. $H_0 : p = p_0$ in the one-sided hypothesis test case.

In the entire hypothesis testing procedure, there is only one difference in evaluating a one-sided hypothesis test vs a two-sided hypothesis test: how to compute the p-value. In a one-sided hypothesis test, we compute the p-value as the tail area in the *direction of the alternative hypothesis only*, meaning it is represented by a single tail area.

Herein lies the reason why one-sided tests are sometimes interesting: if we don't have to double the tail area to get the p-value, then the p-value is smaller and the level of evidence required to identify an interesting finding in the direction of the alternative hypothesis goes down.

However, one-sided tests aren't all sunshine and rainbows: the heavy price paid is that any interesting findings in the opposite direction much be **disregarded**.

Why can't we simply run a one-sided test that goes in the direction of the data?

Answer:

We have been building a careful framework that controls for the Type 1 Error, which is the significance level α in a hypothesis test. We will use the $\alpha = 0.05$ below to keep things simple.

Imagine we could pick the one-sided test after we saw the data. **What will go wrong?**

- If $\hat{p}$ is *smaller* than the null value, then a one-sided test where $p < p_0$ would mean that any observation in the *lower* 5% tail of the null distribution would lead to us **rejecting** H_0.
- If $\hat{p}$ is *larger* than the null value, then a one-sided test where $p > p_0$ would mean that any observation in the *upper* 5% tail of the null distribution would lead to us **rejecting** H_0.

Now, if H_0 were **true**, there is a 10% chance of being in one of the two tails, so our testing error is actually $\alpha = 0.10$, not 0.05. That is, not being careful about when to use one-sided tests effectively undermines the methods we are working so hard to develop and utilize.

3.3.3 Solved Problem

Before we start, the following question comes from a book written by Hans Rosling, Anna Rosling Ronnlund, and Ola Rosling called **Factfulness**:

How many of the world's 1 year old children today have been vaccinated against some disease:

A) 20%
B) 50%
C) 80%

In this tutorial, we will be exploring how people with a 4-year college degree perform on this and other world health questions as we learn about hypothesis tests, which are a framework used to rigorously evaluate competing ideas and claims

Hypothesis testing framework

We're interested in understanding how much people know about world health and development. If we take a multiple choice world health question, then we might like to understand if $\mathbf{H_0}$: People never learn these particular topics and their responses are simply equivalent to random guesses.

$\mathbf{H_A}$: People have knowledge that helps them do better than random guessing, or perhaps, they have false knowledge that leads them to actually do worse than random guessing.

These competing ideas are called **hypotheses**. We call H_0 the null hupothesis and H_A the alternative hypothesis. When there is a subscript 0 like in H_0, data scientists pronounce it as "nought" (e.g. H_0 is pronounced "H-nought").

Null and Alternative Hypothesis

The **null hypothesis** (H_0) often represents a skeptical perspective or a claim to be tested.

The **alternative hypothesis** (H_A) represents an alternative claim under consideration and is often represented by a range of possible parameter values.

The null hypothesis often represents a skeptical position or a perspective of "no difference". In our first example, we'll consider whether the typical person does any different than random guessing on Rosling's question about infant vaccinations.

The alternative hypothesis generally represents a new or stronger perspective. In the case of the question about infant vaccinations, it would certainly be interesting to learn whether people do better than random guessing, since that would mean that the typical person knows something about world health statistics. It would also be very interesting if we learned that people do worse than random guessing, which would suggest people believe incorrect information about world health.

The hypothesis testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative. The hallmarks of hypothesis testing are also found in the US court system.

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Even if the jurors leave unconvinced of guilt beyond a reasonable doubt, this does not mean they believe the defendant is innocent. This is also the case with hypothesis testing: *even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as true.* Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis.

When considering Roslings' question about infant vaccination, the null hypothesis represents the notion that the people we will be considering - college-educated adults - are as accurate as random guessing. That is, the proportion p of respondent who pick the correct answer, that 80% of 1 year olds have been vaccinated against some disease, is about 33.3% (or 1-in-3 if wanting to be perfectly precise). The alternative hypothesis is that this proportion is something other than 33.3%.

While its helpful to write these hypotheses in words, it can be useful to wrtie them using mathematical notation:

$H_0 : p = 0.333$

$H_A : p \neq 0.333$

In this hypothesis setup, we want to make a conclusion about the population parameter p. The value we are comparing the parameter to is called the **null value**, which in this case is 0.333. It's common to label the null value with the same symbol as the parameter but with a subscript "0". That is, in this case, the null value is $p_0 = 0.333$ (pronounces "p-nought equals 0.333")

3.3.4 Exercises

Exercise 1

Formulate the null and alternative hypotheses in the following situations:

(a) A company claims that the proportion of its customers that have complaints against the company is now less than 0.13.

(b) An inspector wants to establish that 2×4 lumber at a mill does not meet a specification that requires at most 5% break under a standard load.

(c) A university official believes that the proportion of students who currently hold part-time jobs has changed from the value 0.26 that prevailed four years ago.

Exercise 2

A census recorded five years ago that 20% of the families in a large community lived below the poverty level. To determine if this percentage has changed, a random sample of 400 families is studied and 70 are found to be living below the poverty level. Does this finding indicate that the current percentage of families earning income below the poverty level has changed from what it was five years ago? (Use significance level $\alpha = 0.05$)

Follow the steps to conduct the hypotheses testing:

(a) Formulate the hypotheses

(b) Compute the sample proportion

(c) Compute the test statistic

(d) Compute the p-value

(e) Draw conclusion using the significance level $\alpha = 0.05$.

Exercise 3

Conduct hypotheses testing.

Follow the example (Please redo the example by yourself)

Chapter 4

Inference for Mean

4.1 Quick Review on Inference for Mean

4.1.1 Objectives

By the end of this unit, students will be able to:

- Distinguish between normal distribution and t distribution.
- Compute point estimates and confidence intervals for estimating one population mean based on one sample and paired samples.

4.1.2 Overview

One-sample means with the t-distribution

Similar to how we can model the behavior of the sample proportion $\hat{p}$ using a normal distribution, the sample mean $\bar{x}$ can also be modeled using a normal distribution when certain conditions are met. However, we will soon learn that a new distribution, called the t-distribution, then we will use it to construct confidence intervals and conduct hypothesis tests for the mean.

The sampling distribution of $\bar{x}$

The sample mean tends to follow a normal distribution centered at the population mean, μ, when certain conditions are met. Additionally, we can compute a standard error for the sample mean using the population standard deviation σ and the sample size n.

Central limit theorem for the 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

$$\text{Mean} = \mu \qquad \text{Standard Error}(SE) = \frac{\sigma}{\sqrt{n}}$$

Before diving into confidence intervals and hypothesis tests using $\bar{x}$, we first need to conver two topics:

- When we modeled $\hat{p}$ using the normal distribution, certain conditions had to be satisfied. The conditions for working with $\bar{x}$ are a little more complex, and we will spend the next section discussing how to check conditions for inference.
- The standard error is dependent on the population standard deviation, σ. However, we rarely know σ, and instead we must estimate it. Because this estimation is itself imperfect, we use a new distribution called the t-distribution to fix this problem.

Evaluating the two conditions required for modeling $\bar{x}$

Two conditions are required to apply the Central limit theorem for a sample mean $\bar{x}$:

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. If the data comes from a random process, analogus to rolling a die, this would also satisfy the independence condition.

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. This condition is obviously vague, making it difficult to evaluate, so next we introduce a couple rules of thumb to make checking this condition easier.

Rules of thumb: How to perform the normality check

There is no perfect way to check the normality condition, so instead we use two rules of thumb:

- $\mathbf{n < 30}$ **:** 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.
- $\mathbf{n \geq 30}$: 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.

In this section, you aren't expected to develop perfect judgement on the normality condition. However, you are expected to be able to handle clear cut cases based on the rules of thumb.

In practice, it's typical to also do a mental check to evaluate whether we have reason to believe the underlying population would have moderate skew (if $n < 30$) or have aprticularly extreme outliers ($n \geq 30$) beyond what we observe in the data. For example, consider the number of followers for each individual account on Twitter, and then imagine the distribution. The large majority of accounts have built up a couple thousand followers or fewer, while relatively tiny fraction have amassed tens of millions of followers, meaning the distribution is extremely skewed. When we know the data come from such an extremely skewed distribution, it takes some effor to understand what sample size is large enough for the normality condition to be satisfied.

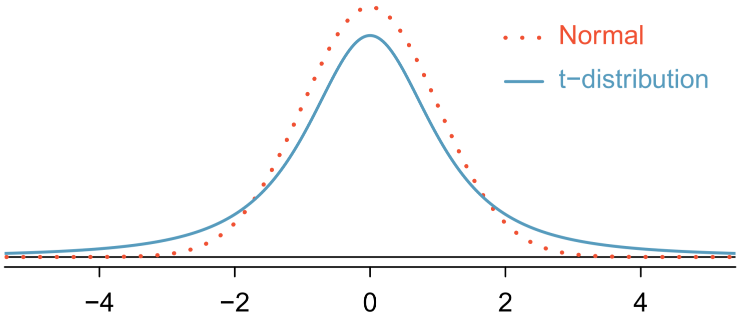
Introducing the t-distribution

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 will employ a similar strategy for computing the standard error of $\bar{x}$, using the sample standard deviation s in place of σ:

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

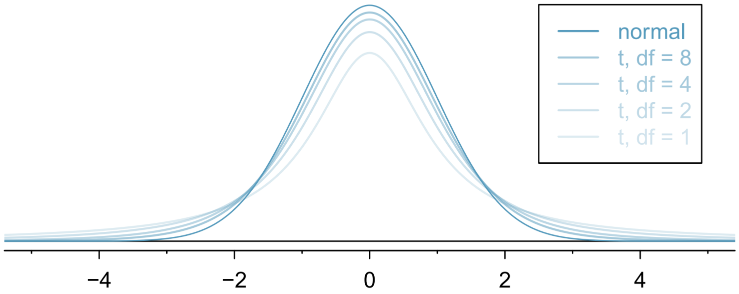
This strategy tends to work well when we have a lot of data and can estimate σ using s accurately. However, the estimate is less precise with smaller samples, and this leads to problems when using the normal distribution to model $\bar{x}$.

We will find it useful to use a new distribution for inference called the **t-distribution**. A t- distribution, shown as a solid line in the figure below, has a bell shape. However, its tails are thicker than the normal distribution's, meaning observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution. The extra thick tails of the t-distribution are exactly the correction needed to resolve the problem of using s in place of σ in the SE calculation.



The t-distribution is always centered at zero and has a single parameter: degrees of freedom. The **degrees of freedom** (**df**) describes the precise form of the bell-shaped t-distribution. Several t-distributions are shown in the figure below in comparison to the normal distribution.

In general, we will use a t-distribution with $df = n - 1$ to model the sample mean when the sample size is n. That is, when we have more observations, the degrees of freedom will be larger and the t-distribution will look more like the standard normal distribution: when the degrees of freedom is about 30 or more, the t-distribution is nearly indistinguishable from the normal distribution.



Degrees of Freedom (df)

The degrees of freedom describes the shape of the t-distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

When modeling $\bar{x}$ using the t-distribution, use $df = n - 1$.

The t-distribution allows us greater flexibility than the normal distribution when analyzing numerical data. In practice, its common to use a statistical software, such as R, Python or SAS for these analyses. Alternatively, a graphing calculator or a **t-table** may be used; the t-table is similar to the normal distribution table.

In the normal model, we used z^* and the standard error to determine the width of a confidence interval. We revise the confidence interval formula slightly when using the t-distribution:

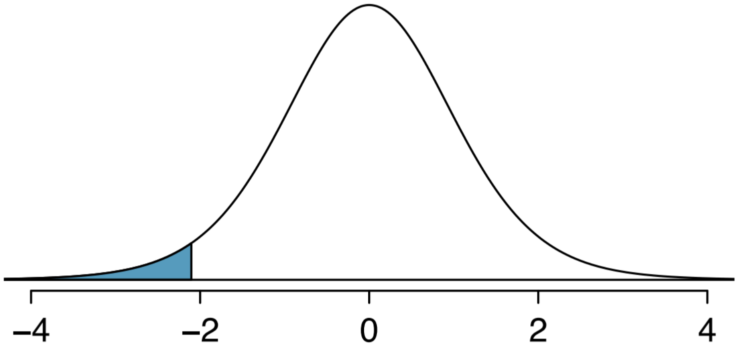


Figure 4.1: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded

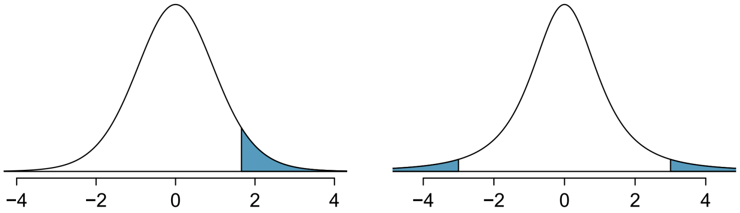


Figure 4.2: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

$$point\ estimate \pm t^*_{df} \times SE \qquad \rightarrow \qquad \bar{x} \pm t^*_{df} \times \frac{s}{\sqrt{n}}$$

Confidence interval for a single mean

Once you've determined a one-mean confidence interval would be helpful for an application, there are four steps to constructing the interval:

Prepare. Identify $\bar{x}, s, n$, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\bar{x}$ is nearly normal.

Calculate. If the conditions hold, compute, SE, find t^*_{df} and construct the interval.

Conclude. Interpret the confidence interval in the context of the problem.

One sample *t*-tests

Is the typical US runner getting faster or slower over time? We consider this question in the context of the Cherry Blossom Race, which is a 10-mile race in Washington, DC each spring.

The average time for all runner who finished the Cherry Blossom Race in 2006 was 93.29 minutes (93 minutes and about 17 seconds). We want to determine using data from 100 participants in the 2017 Cherry Blossom Race whether runners in this race are getting faster or slower, versus the other possibility that there has been no change.

When completing a hypothesis test, for the one-sample mean, the process is nearly identical to completing a hypothesis test for a single proportion. First, we find the Z-score using the observed value, null value, and standard error; however, we call it a **T-score** since we use a t-distribution for calculating the tail area. Then we find the p-value using the same ideas we used previously: find the one-tail area under the sampling distribution, and double it.

With both the independence and normality conditions satisfied, we can proceed with a hypothesis test using the t-distribution.

4.1.3 Solved Problem

Just like the normal distribution, we can use R to find the area to the below a certain standard deviation.

We can find the area to the left of 1.75 standard deviations with 12 degrees of freedom by using the following:

```r
pt(1.75, df = 12)
```

```
## [1] 0.9471902
```

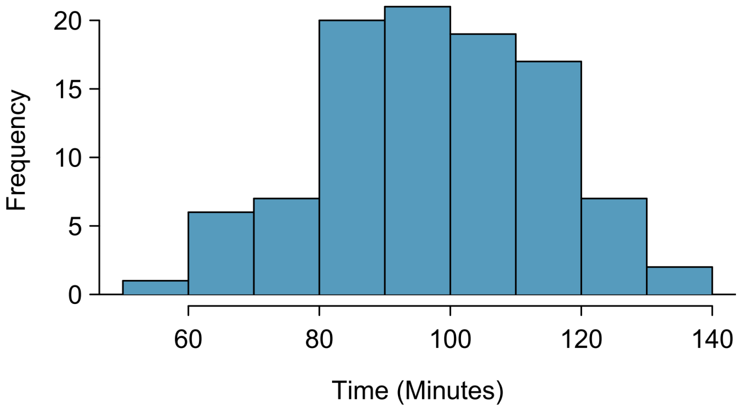


Figure 4.3: A histogram of time for the sample Cherry Blossom Race data.

What proportion of the *t*-distribution with 18 degrees of freedom falls below -2.10?

To find the two tailed proportion with 2 degrees of freedom more than 3 units from the mean, we will do the following

```
2*pt(-3, df = 2, lower.tail = TRUE)
```

```
## [1] 0.09546597
```

OR

```
pt(-3, df = 2, lower.tail = TRUE) + pt(3, df = 2, lower.tail = FALSE)
```

```
## [1] 0.09546597
```

OR

```
pt(-3, df = 2, lower.tail = TRUE) + 1 -  pt(3, df = 2, lower.tail = TRUE)
```

```
## [1] 0.09546597
```

We can see that there are multiple ways to acquire the same results by modifying the arguments with the use of the complement rule.

4.1.4 Exercises

- Sample mean $\bar{x}$ is the unbiased **point estimator** for the population mean μ
- A value of $\bar{x}$ is a point estimate
- Error $= \mu - \bar{x}$

Central Limit Theorem (Sampling distribution of sample mean)

When taking samples of fixed size n from a population with mean μ and standard deviation σ, when the observations are independent (take random samples of fixed size n, without replacement); the sample size $n \geq 30$, then the sample proportion $\bar{x}$ is approximately normal: $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

When we know the population is normal, no matter what sample size, $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

Notes:

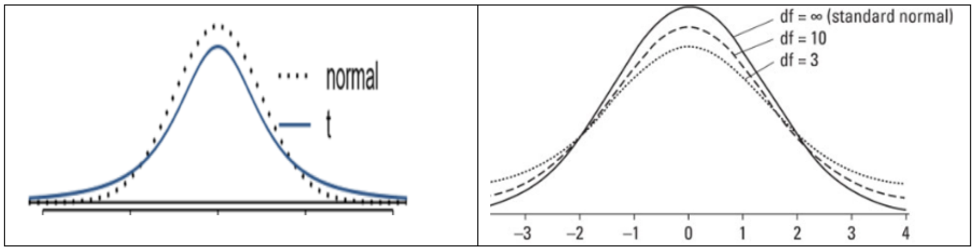
- When using $\bar{x}$ to estimate μ the Standard Error of $\bar{x}$ is the standard deviation of its sampling distribution: $S.E. = \frac{\sigma}{\sqrt{n}}$
- Usually σ is unknown use s to replace σ: $S.E. \approx \frac{s}{\sqrt{n}}$

When can the CLT be applied

- If $n \geq 30$ and σ is known
- If the population is normal and σ is known
- Otherwise we use t-distribution: $T = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{df}$, where $df = n - 1$ is the degree of freedom.

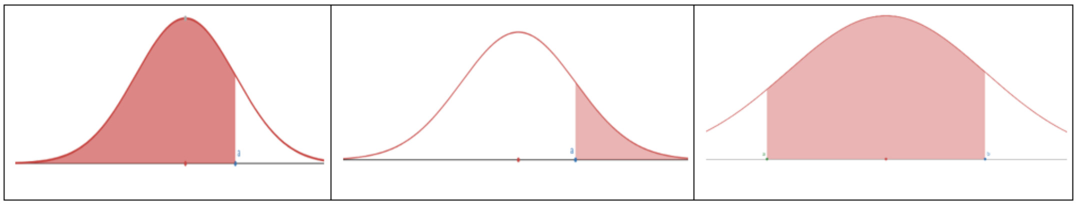
t-distribution

Similar to the standard normal distribution: the probability density curve of a t-distribution is centered at 0, and it is bell-shaped. But tails of a t-distribution are thicker than that of the standard normal distribution; moreover, the lower df, the thicker the tails.



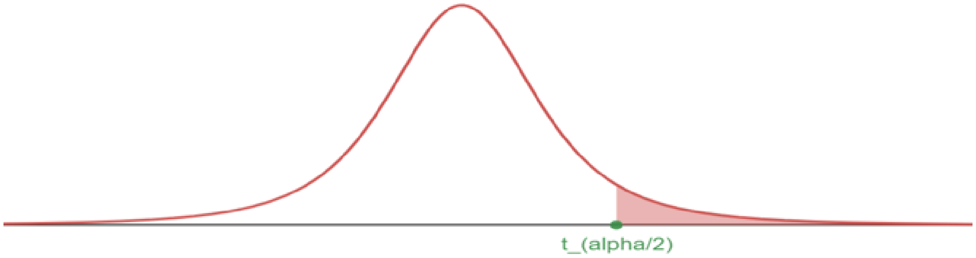
Using R to find probability under t-distribution with $df = n - 1$:

- For $P(T < b)$: `pt(b, df)`
- For $P(T > a)$: `pt(a, df, lower.tail = FALSE)` or `1 - pt(a, df)`
- For $P(a < T < b)$: `pt(b, df) - pt(a, df)`



To find the cut-off point t (critical value t^* or $t_{\alpha/2}$) for a given cumulative probability with $df = n - 1$:

- Find t for $P(T < t) = p$: `qt(p, df)`
- Find t for $P(T > t) = p$: `qt(1 - p, df)` or `qt(p, df, lower.tail = FALSE)`
- $t_{\alpha/2}$: $P(T > t_{\alpha/2}) = \alpha/2$: $qt(\alpha/2, df, lower.tail = FALSE)$



100$(1 - \alpha)$% Confidence interval for mean μ

Using sample with size n, sample mean $\bar{x}$, sample standard deviation s, the critical value $t_{\alpha/2}$: $\bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$

Margin of Error (M.E.)

$M.E. = t_{\alpha/2} \times S.E. = t_{\alpha/2} \times \frac{s}{\sqrt{n}}$

4.2 Hypothesis Testing for mean μ

4.2.1 Objectives

By the end of this unit, students will be able to:

- Formulate claims about a population mean in the form of a null hypothesis and alternative hypothesis.
- Conduct t-tests for testing claims about a single population mean based on one sample and paired samples.

4.2.2 Overview

One sample t-test (Same framework as the Hypothesis Testing for proportion)

Steps:

1. Set up the hypotheses
2. Compute the t test statistic

   Using sample with size n, sample mean $\bar{x}$, sample standard deviation s, null value μ_0,

   $T = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
3. Compute the p-value

   Let t-test statistic $t_1 = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ (from step 2)

   - For left-sided test, p-value is $P(T < t_1)$ use `pt(t, df)`
   - For right-sided test, p-value is $P(T > t_1)$ use `pt(t, df, lower.tail = FALSE)`
   - For two-sided test, p-value is $P(|T| > |t_1|)$ use $2 * pt(-|t_1|, df)$ or $2 * pt(|t_1|, df, lower.tail = FALSE)$
4. Compare the p-value with the significance level α and make decision

   - If p-value $\leq \alpha$, then we have enough evidence to reject H_0 and substantiate H_a;
   - If p-value $> \alpha$, then we do not have enough evidence to reject H_0
   - The default value of significance level is $\alpha = 0.05$

Hypothesis testing for a single mean

Once you have determined a one-mean hypothesis test is the correct procedure, there are four steps to completing the test:

Prepare. Identify the parameter of interest, list out hypotheses, identify the significance level, and identify $\bar{x}, s$ and n.

Check. Verify conditions to ensure $\bar{x}$ is nearly normal.

Calculate. If the conditions hold, compute SE, compute the T-score, and identify the p-value.

Conclude. Evaluate the hypothesis test by comparing the p-value to α, and provide a conclusion in the context of the problem/.

Paired Data

In an earlier edition of this text book, we found that Amazon prices were, on average, lower than those of the UCLA Bookstore for UCLA courses in 2010. It's been several years, and many stores have adapted to the online market, so we wondered, how is the UCLA Bookstore doing today?

We sampled 201 UCLA courses. Of those, 68 required books could be found on Amazon. A portion of the data set from these courses is shown in the table below, where the prices are in US dollars.

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
68	Jewish studies	M10	35.96	32.40	3.56

Paired observations:

Each textbook has two corresponding prices in the data set: one for the UCLA Bookstore and one for Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.

Paired data

Two sets of observations are *paired* if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In the textbook data, we look at the differences in prices, which is represented as the `price_difference` variable in the data set. Here the differences are taken as

$$\text{UCLA Bookstore price} - \text{Amazon price}$$

It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. The first difference shown in the table above is computed as $47.97-47.45 = 0.52$. Similarly, the second difference is computed as $14.26 - 13.55 = 0.71$, and the third is $13.50 - 12.53 = 0.97$. A histogram of the differences is shown in the figure below. Using differences between paired observations is a common and useful way to analyze paired data.

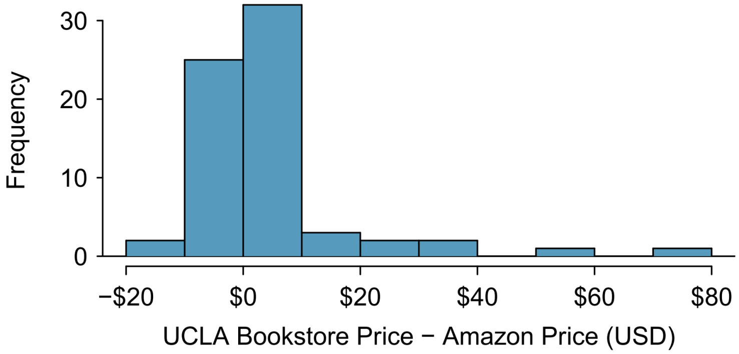


Figure 4.4: A histogram of difference in prices for each book sampled.

Inference for paired data

To analyze a paired data set, we simply analyze the differences.

$$n_{diff} = 68 \qquad \bar{x}_{diff} = 3.58 \qquad s_{diff} = 13.42$$

Lets set up a hypothesis test to determine whether, on average, there is a difference between Amazon's price for a book and the UCLA bookstore's price. Also, check the conditions for whether we can move forward with the test using the t-distribution.

We are considering two scenarios: there is no difference or there is some difference in average prices.

$$H_0 : \mu_{diff} = 0. \text{ There is no difference in the average textbook price.}$$

$$H_A : \mu_{diff} \neq 0. \text{ There is a difference in average prices.}$$

Next, we will check the independence and normality conditions.

The observations are based on a simple random sample, so independence is reasonable. While there are some outliers, $n = 68$ and none of the outliers are particularly extreme, so the normality of $\bar{x}$ is satisfied. With these conditions satisfied, we can move forward with the t-distribution.

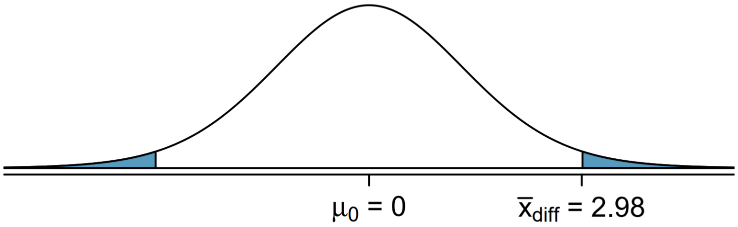
To compute the test, we need to compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences ($s_{diff} = 13.42$) and the number of differences ($n_{diff} = 68$):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = 1.63$$

The test statistics is the T-score of $\bar{x}_{diff}$ under the null condition that the actual mean difference is 0:

$$T = \frac{\bar{x}_{diff} - 0}{SE_{\bar{x}_{diff}}} = \frac{3.58 - 0}{1.63} = 2.20$$

To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true, and the p-value is represented by the two shaded tails:



The degrees of freedom is $df = 68 - 1 = 67$. Using R, we find the one-tail area of 0.0156. Doubling this area gives the p-value: 0.0312.

```
2*pt(2.20, df = 67, lower.tail = FALSE)
```

```
## [1] 0.03125996
```

Because the p-value is less than 0.05, we reject the null hypothesis. Amazon prices are, on average, lower than the UCLA Bookstore prices for UCLA courses.

4.2.3 Solved Problem

4.2.4 Exercises

Exercise 1

Without finding the values, arrange the numbers from small to large:

a) $P(Z < -1.25)$

b) $P(T < -1.25)$ with $df = 10$

c) $P(T < -1.25)$ with $df = 15$

d) $P(Z > 1.35)$

e) $P(T > 1.35)$ with $df = 10$

f) $P(T > 1.25)$ with $df = 15$

________ < ________ < ________ < ________ < ________ < ________

Exercise 2

Use R calculator to find the values of the probability of t-distribution. Sketch the t-curve and shaded region.

- $P(T < -1.25)$ with $df = 10$
- $P(T < -1.25)$ with $df = 15$
- $P(T > 1.35)$ with $df = 10$
- $P(T > 1.25)$ with $df = 15$

Exercise 3

Use R calculator to find the critical t-value ($t_{\alpha/2}$), rounded the result to 4 decimal places.

- CL = 90%, $n = 7$
- CL = 98%, $n = 20$
- CL = 99%, $n = 28$
- CL = 95%, $n = 9$

Exercise 4

Find confidence interval with the sample information:

(a) $n = 5, \bar{x} = 4.1, s = 1.2$, 90% confidence level

(b) $n = 15, \bar{x} = 4.1, s = 1.2$, 90% confidence level

(c) $n = 5, \bar{x} = 4.1, s = 1.2$, 98% confidence level

(d) $n = 15, \bar{x} = 4.1, s = 1.2$, 98% confidence level

Exercise 5

What affects the width of the confidence interval? (You may use your observations from Exercise 4 for reference)

Exercise 6

(Working backwards) A 95% confidence interval for a population mean μ is given as (18.98, 20.02). This confidence interval is based on a simple random sample of 36 observations. Calculate the following:

(a) The sample mean

(b) The margin of error

(c) The critical t-value (use t-distribution)

(d) The standard error (use the result of c)

(e) The sample standard deviation (use the result of d)

Exercise 7

Find the P-value for the given sample sizes and test statistic:

(a) $n = 26$, $T = 2.485$, for right-sided test

(b) $n = 18$, $T = -1.45$, for left-sided test

(c) $n = 26$, $T = 2.485$, for two-sided test

(d) $n = 18$, $T = -1.45$, for two-sided test

Exercise 8

A random sample of 25 New Yorkers were asked how much sleep they get per night. The result shows:

$n = 25, \bar{x} = 7.73, s = 0.77$

The point estimate suggests that New Yorkers sleep less than 8 hours per night on average. Is the result statistically significant?

Follow the steps to conduct the hypothesis test.

(a) Write the hypotheses in symbols: H_0: __________________ H_a: __________________

(b) Calculate the test statistic

(c) Compute the P-value and draw a picture

(d) What is the conclusion of the hypothesis test, using the significance level $\alpha = 0.05$

(e) If you were to construct a 90% confidence interval that corresponds to this hypothesis test, would you expect 8 hours a night on average to be in the interval?

Exercise 9

Georgianna claims that in a small city, the average child takes less than 5 years of piano lessons. We have a random sample of 20 children from the city, with a mean of 4.6 years of piano lessons and a standard deviation of 2.2 years. Evaluate Georgianna's claim using a hypothesis test.

(a) Write the hypotheses in symbols: H_0: __________________ H_a: __________________

(b) Calculate the test statistic

(c) Compute the P-value and draw a picture

(d) What is the conclusion of the hypothesis test, using the significance level $\alpha = 0.05$

Chapter 5

Linear Regression

5.1 Introduction to Linear Regression

5.1.1 Objectives

By the end of this unit, students will be able to:

- Describe linear associations between numerical variables using correlations.
- Use linear regression to model linear relationships between two numerical variables.
- Evaluate the statistical significance of linear relationships between numerical variables.

5.1.2 Overview

Regression analysis concerns the study of relationships between quantitative variables: identifying, estimating, and validating the relationship.

- (Simple) linear regression is to study if the relationship between two numerical variables is linear, and the strength of the linear association.
- We begin with the scatter plot of two numerical variables, to observe if there is a linear association.
- If there seems to be a linear relationship, we use the linear model $y = \beta_0 + \beta_1 x$ to best fit the data.
- Using a sample data set (x_i, y_i) for $i = 1, \ldots, n$ and least squares error, we derive an estimated model $\hat{y} = b_0 + b_1 x$.

Let's check in with a few short videos from our friends at OpenIntro.org to help develop the notion of linear regression for us.

Simple linear regression uses a single numerical feature (predictor variable) to predict a numerical response. Simple linear regression uses the form of a straight line $y = mx + b$, where m denotes slope of the relationship and b denotes the intercept (the value of y if x is 0). With regression, we are fitting a straight line to data, where noise is present – that is, the line we fit is not expected to pass through all of the data points. The form for a simple regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Notice that β_0 is the intercept, β_1 is the slope, and ε denotes the unexplained error (noise). We typically do not write ε as part of the model, since we assume that it is random noise with a mean of 0 and a constant standard deviation (σ) – in our earlier notation, we assume $\varepsilon \sim N\left(\mu = 0, \sigma\right)$. Instead, we often write the regression model as

$$\mathbb{E}\left[y\right] = \beta_0 + \beta_1 x$$

Let's see regression in action as we consider an application to understanding biases in course evaluations.

The Data

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. The article titled, "Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity" (Hamermesh and Parker, 2005) found that instructors who are viewed to be better looking receive higher instructional ratings. (Daniel S. Hamermesh, Amy Parker, Beauty in the classroom: instructors pulchritude and putative pedagogical productivity, *Economics of Education Review*, Volume 24, Issue 4, August 2005, Pages 369-376, ISSN 0272-7757, 10.1016/j.econedurev.2004.07.013. http://www.sciencedirect.com/science/article/pii/S0272775704001165.)

In this workbook we will analyze the data from this study in order to learn what goes into a positive professor evaluation.

The data were gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin. In addition, six students rated the professors' physical appearance. (This is a slightly modified version of the original data set that was released as part of the replication data for *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Gelman and Hill, 2007).) The result is a data frame where each row contains a different course and columns represent variables about the courses and professors.

variable	description
`score`	average professor evaluation score: (1) very unsatisfactory - (5) excellent.
`rank`	rank of professor: teaching, tenure track, tenured.
`ethnicity`	ethnicity of professor: not minority, minority.
`gender`	gender of professor: female, male.
`language`	language of school where professor received education: english or non-english.
`age`	age of professor.
`cls_perc_eval`	percent of students in class who completed evaluation.
`cls_did_eval`	number of students in class who completed evaluation.
`cls_students`	total number of students in class.
`cls_level`	class level: lower, upper.
`cls_profs`	number of professors teaching sections in course in sample: single, multiple.
`cls_credits`	number of credits of class: one credit (lab, PE, etc.), multi credit.
`bty_f1lower`	beauty rating of professor from lower level female: (1) lowest - (10) highest.
`bty_f1upper`	beauty rating of professor from upper level female: (1) lowest - (10) highest.
`bty_f2upper`	beauty rating of professor from second upper level female: (1) lowest - (10) highest.
`bty_m1lower`	beauty rating of professor from lower level male: (1) lowest - (10) highest.
`bty_m1upper`	beauty rating of professor from upper level male: (1) lowest - (10) highest.
`bty_m2upper`	beauty rating of professor from second upper level male: (1) lowest - (10) highest.
`bty_avg`	average beauty rating of professor.
`pic_outfit`	outfit of professor in picture: not formal, formal.
`pic_color`	color of professor's picture: color, black & white.

Prediction (Predicted value)

If the least square regression model is given by $\hat{y} = b_0 + b_1 x$, then for a given x, the predicted value is $\hat{y} = b_0 + b_1 x$ – plug in the value of x.

Interpreting the slope and the y-intercept of a regression line

- The slope b_1 is the amount by which the predicted value y changes when x is increased by one unit.
- The y-intercept b_0 is the predicted value of y when $x = 0$.

Residual

For a data set (x_i, y_i) for $i = 1, \ldots, n$, the error of using the model is: $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

The Correlation Coefficient

- The mathematical formula is:

$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$

- The value of R: $-1 \leq R \leq 1$
- The closer $|R|$ is to 1, the stronger the linear association.

The Coefficient of Determination: R^2

- The coefficient of determination R^2 is a measure used in statistical analysis to assess how well a model explains and predicts future outcomes.
- R^2 is the proportion (fraction) of the variation in the response variable that is predictable (can be explained) from the explanatory variable.

Conditions to have the least squares regression

Visually inspect the scatter plot:

- The relationship between the explanatory and the response variable should be linear.
- The histogram of residuals distribution should be normal (symmetric, bell-shaped).
- The variability of points should be roughly constant.
- No extreme outliers.

Computing the coefficients in $\hat{y} = b_0 + b_1 x$

$b_1 = \frac{s_y}{s_x} R$

$b_0 = \bar{y} - b_1 \bar{x}$

Notes

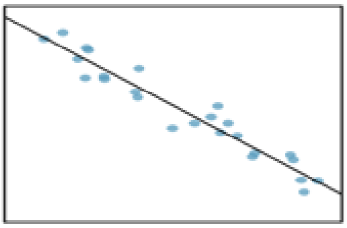
1. R and b_1 have the same sign.
2. Equivalently, $R = \frac{s_x}{s_y} b_1$

5.1.3 Solved Problem

5.1.4 Exercises

Exercise 1

Describe the linear relationship from the scatter plot.



Select the correct choice.

(a) Strong positive relationship

(b) Strong negative relationship

(c) Weak positive relationship

(d) Weak negative relationship

Exercise 2

The mean travel time from one stop to the next on the Coast Starlight is 129 minutes, with a standard deviation of 113 minutes. The mean distance from one stop to the next is 108 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636.

(a) Write the equation of the regression line for predicting travel time (based on the distance).

(b) Interpret the slope and the intercept in this context.

(c) Calculate and interpret R^2.

(d) The distance between Santa Barbara and LA is 103 miles. Use this model to estimate the time it takes to travel between these two cities.

(e) It actually takes the Coast Starlight about 168 minutes to travel between Santa Barbara and LA. Calculate the residual. Is the model over or underestimating the time?