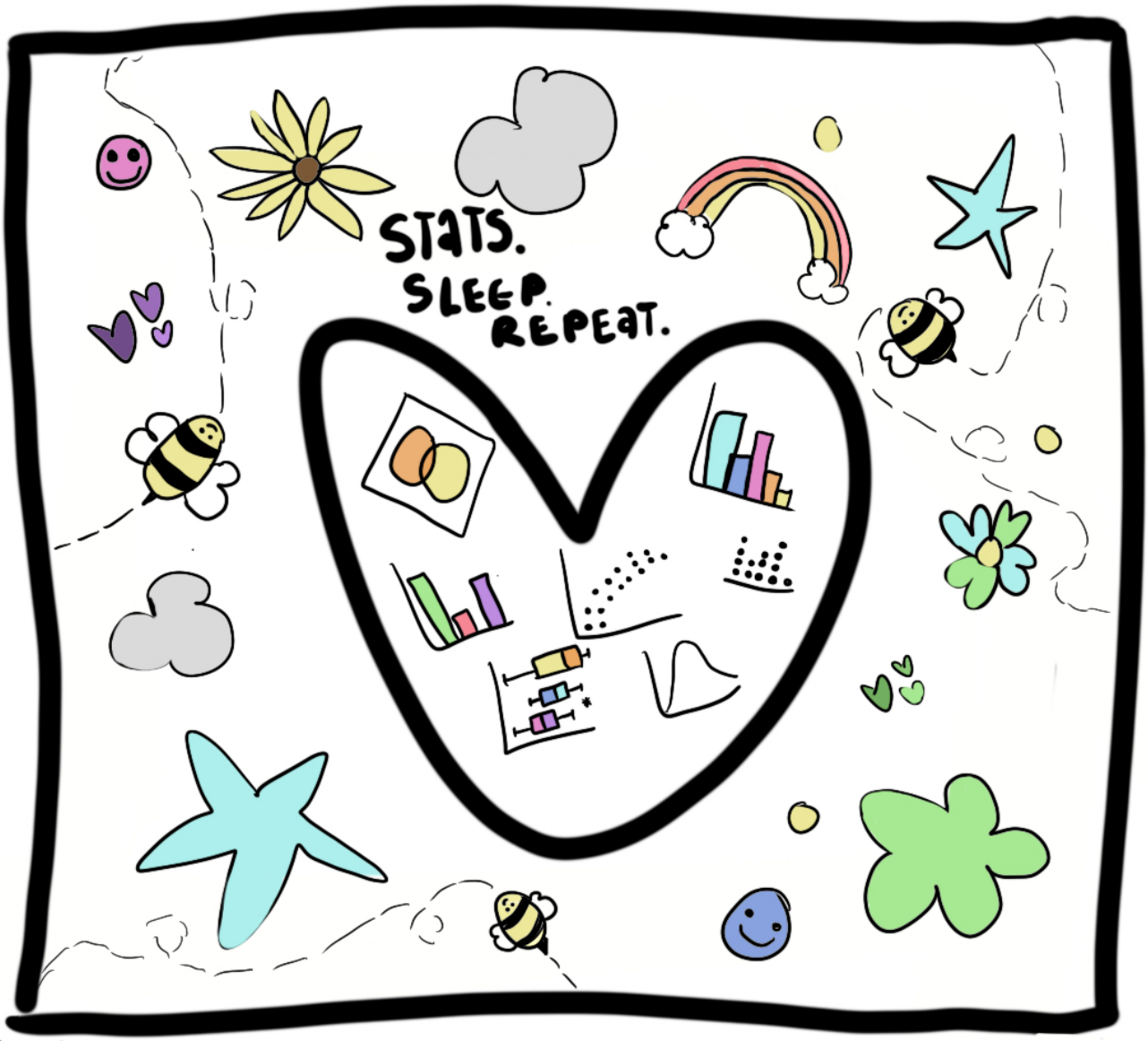
Stat 201 Lecture Notes

Semester: _____ Section: ____

Instructor: _____

These Notes Belong To: _____



@amdesigns

Contents

1 Using Data to Answer Questions

1.1 The Investigation

We begin our exploration into the discipline of statistics with a question. It is our hope that by thinking deeply about answering the problem at hand we will launch an investigative journey together as a class. As we walk this journey together we will explore what it means to employ "statistical thinking" while introducing many of the processes important to any attempt at seeking answers.

The Question

The Academic Common Market (ACM) is a tuition savings program for college students from selected states in the Southern Regional Education Board (SREB) who want to pursue degrees that are not offered by their home state institutions. There are multiple undergraduate programs here at CCU where students can benefit from ACM. For example, a student majoring in marine science who is from a land-locked state can receive in-state tuition. The following table gives a list of undergraduate programs at CCU that are available through the ACM.

Marine Science (BS)	Recreation and Sport Management (BS)
Middle Level Education (BA)	Hospitality, Resort, and Tourism Management (BS)
Digital Culture and Design (BA)	Intelligence and National Security Studies (BA)
Theatre (BS)	Sustainability and Coastal Resilience (BA/BS)
Theatre Arts (BFA)	

Thinking about the ACM, consider the following questions:

- What kind of impact does the ACM have on the diversity of hometowns among majors?
- Is the distance from home higher among those majors available through the ACM than other majors?
- Is there a difference between the percentage of in-state and out-of-state students who choose to major in a program available through ACM?

Think about it

How should we get started in seeking answers to these questions? What kind of actions do we need to take to address these questions? Discuss with a neighbor and as a class. Include key points of the brainstorming session on the next couple of pages.

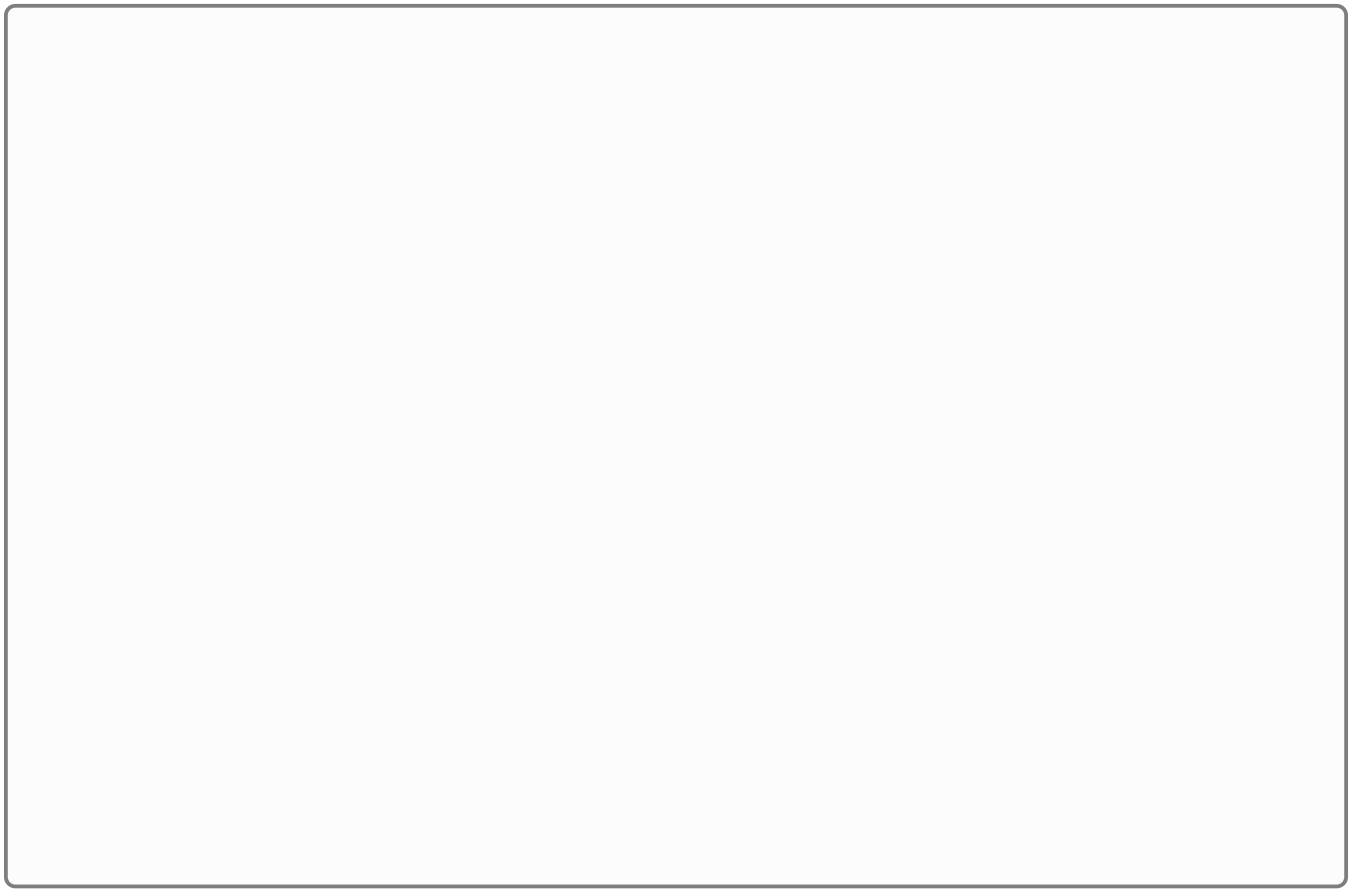
Points from discussion with a neighbor

Points from discussion with the class

1.2 Data Collection

1.2.1 Sample and Population

Our Investigation in a Picture



In each of the following examples, identify the sample, population, statistics, and any parameters.

Example 1.1. A poll commissioned HMD Global consisted of 2,000 smartphone users in the US. Of those polled, 60% said they could not cope without their smartphone for a day. On average, respondents check their phone 20 times a day.

Example 1.2. Tiger sharks are common off the Atlantic coast and in the Myrtle Beach area. They are the fourth largest shark in the world with an average length of 12 feet and an average weight of 1,000 pounds. Some tiger sharks sexed and measured near Garden City had an average length of 9.5 feet and weight of 890 pounds.

1.2.2 Poor Sampling Techniques

Example 1.3. In Shere Hite's widely quoted book *Women and Love: A Cultural Revolution in Progress* (1987) she made a number of claims such as,

- 95% of all women report forms of emotional and psychological harassment from men with whom they are in love relationships.
- 70% of all women married five or more years are having sex outside their marriages.

The data for the study was collected via a survey sent to women's groups, counseling centers, church societies and senior citizen centers. The survey consisted of 127 multiple part essay questions. Out of 100,000 questionnaires, 4.5% were returned. What are your initial thoughts on the conclusions of this study?

1. **Measurement Bias:** Measurements tend to record values larger or smaller than the true value

2. **Sampling Bias:** The sample is not representative of the population

   (a) **Voluntary Sampling:** People with strong feelings one way or the other are the ones tending to respond. (Internet Polls "Disclaimer: this is not a scientific poll")

   (b) **Convenience Sampling:** Only a "convenient" group is surveyed. This may leave out many subgroups of the population.

   (c) **Survivorship Bias:** Focusing on those who made it past some process and not considering those who did not.

   During WWII researchers studied damage done to aircraft returning from missions. They suggested that armor be added to the most frequently damaged areas. Statistician Abraham Wald noted that these were the planes that returned and could do so with damage to those areas. Instead armor should be added to the opposite areas representing planes that did not return.

Question: What kind of biases, if any, are present in our approach to answering the class investigation? What would need to happen to collect data without any sampling bias?

1.2.3 Random Sampling and Variations

The best surveys are based on participants that are randomly selected. The most basic way to make a random selection is the **Simple Random Sample(SRS)**. In such a sample, all possible participants are equally likely to be chosen. It is as if the investigator places all members of the population in a hat, shakes it up, and draws out the desired number for the sample.

- First, all members of the population are identified and given a numerical label. This set of labels is called the **sampling frame**.
- Next, a **random number generator** (ex. calculator or table) is used to select the desired number of labels from the sampling frame.

Example 1.4. Suppose we are doing an audit of the accounts at a particular school that has 60 total accounts. We have the resources (i.e. time and money) to look at 8 of the accounts. Use the table or calculator to obtain an SRS of the accounts to audit. We label the accounts 01 to 60, then:

On the Calculator:

- Set the seed (Optional for reproducible values):
  Enter any number then, `STO>`→`MATH`→`PROB`→`rand`→`ENTER`
- Obtain random numbers: `MATH`→`PROB`→`RandInt(lwr,uppr,number of items)`

Using a Table:

Line	Random Digits Table							
101	27660	75721	02273	06330	26044	51737	25666	27523
102	99656	98374	28768	58389	41526	70670	45652	61148
103	09415	12521	73118	96552	05050	73972	39080	55845
104	28896	43438	38238	73811	86983	53509	42968	07616
105	43670	73457	22563	79758	66677	74798	85780	06266
106	69858	89900	94191	69124	90056	45830	11449	64076
107	53275	23244	19316	50270	50593	63803	79572	94700
108	37477	72467	26595	24618	28088	40413	57703	02567
109	65319	62093	53452	46198	99165	58222	15876	82658
110	62714	34986	10467	12377	77916	96791	64439	57678

Question: How would we take a SRS if there were 120 accounts to choose from?

Extensions of Random Sampling:

Stratified Sampling - divide the population into subgroups (strata) and take an SRS from each strata. Strata are often subgroups of interest such as age groups, different types of habitat, or size of companies. Furthermore, subjects in the same stratum tend to be more similar, so stratification can increase precision.

In our example, suppose the accounts can be divided into strata by department/function such as arts, sciences, grounds and maintenance, and hospitality. We wish to take a stratified sample so that all departments are represented.

Cluster Sampling - members of the populations are aggregated into groups called clusters. First, take an SRS of clusters and then interview all subjects within each selected cluster. This technique is typically selected for matters of convenience and expense reduction. For example when clusters are defined by geographic locations, cluster sampling can save time and money spent travelling to collect the data. Naturally occurring clusters include households or counties. Unlike stratified sampling, cluster sampling rarely increases precision, but it does allow us to take sample without the requirement of listing every member of the population.

In our example, suppose the accounts are aggregated into clusters by department such as arts, sciences, grounds and maintenance, and hospitality and we wish to take a cluster sample (Note: there is no practical reason for a cluster sample in this example)

Systematic Sampling - Every k^{th} subject is sampled

Question: How could we apply these techniques to our data collection in the class investigation?

Revised Investigation: The Office of Institutional Research, Assessment and Analysis (IRAA) took a simple random sample of 100 students enrolled in STAT 201 during the fall 2022 semester. From each student, IRAA observed the following variables:

- **Major** - the active major for each student
- **ACM** - whether or not the active major is part of the Academic Common Market
- **Cumulative Credits** - the cumulative credit hours completed at CCU
- **Class** - classification at time of survey (Freshmen, Sophomore, Junior, Senior)
- **Cumulative GPA** - cumulative GPA for courses completed at CCU
- **Enrolled Credits** - number of credit hours enrolled during fall 2022
- **Residency** - the residency status of each student (in-state/out-of-state/international)
- **Distance** - the number of miles each student's hometown is from CCU

The results of the survey are given on the following two pages. Since the data from this survey was obtained using a proper sampling technique, we will use these data for much of our analysis throughout these notes. Now that we have the data, what do we do with it? Discuss some possibilities of next steps.

Investigation Data Set

Student	Major	ACM	Cumulative Credits	Class	Cumulative GPA	Enrolled Credits	Residency	Distance
1	SOC	NO	54	Sophomore	NA	13	In-State	1.1
2	ESCI	NO	135	Senior	2.674	13	In-State	151.4
3	EXSS	YES	77	Junior	2.839	14	In-State	68.9
4	SOC	NO	78	Junior	2.314	17	In-State	135.0
5	PUBH	NO	112	Senior	3.417	17	In-State	138.3
6	PHYSA	NO	126	Senior	3.461	13	In-State	1.1
7	PUBH	NO	84	Junior	2.851	13	In-State	199.6
8	ESCI	NO	107	Senior	3.294	16	In-State	73.4
9	EXSS	YES	45	Sophomore	3.240	14	In-State	23.3
10	CSCI	NO	149	Senior	2.610	13	In-State	15.2
11	MKTP	NO	38	Sophomore	1.777	14	Out-of-State	462.2
12	EXSS	YES	87	Junior	2.608	16	In-State	165.4
13	PUBH	NO	87	Junior	3.176	13	In-State	179.8
14	MSCI	YES	70	Junior	2.649	13	Out-of-State	727.4
15	INTS	NO	95	Senior	3.584	16	Out-of-State	504.8
16	PUBH	NO	73	Junior	2.785	15	In-State	14.7
17	BIOL	NO	60	Junior	3.125	16	In-State	102.4
18	BIOL	NO	60	Junior	3.133	15	In-State	11.6
19	INTEL	YES	60	Junior	2.803	16	Out-of-State	469.1
20	PUBH	NO	75	Junior	3.357	16	In-State	18.2
21	MSCI	YES	61	Junior	3.234	15	In-State	15.2
22	MSCI	YES	83	Junior	3.967	15	In-State	21.8
23	BIOL	NO	75	Junior	1.500	12	In-State	18.2
24	COMM	NO	65	Junior	3.392	16	Out-of-State	357.4
25	MSCI	YES	56	Sophomore	2.992	15	Out-of-State	454.1
26	MGED	YES	84	Junior	3.627	17	Out-of-State	575.9
27	INTEL	YES	78	Junior	3.534	14	Out-of-State	621.6
28	INTEL	YES	66	Junior	2.812	16	Out-of-State	480.1
29	COMM	NO	62	Junior	4.000	13	In-State	11.6
30	SUST	YES	59	Sophomore	3.136	14	Out-of-State	713.1
31	PUBH	NO	40	Sophomore	2.988	12	In-State	16.0
32	MSCI	YES	34	Sophomore	2.195	13	Out-of-State	357.0
33	EXSS	YES	64	Junior	3.156	15	In-State	107.2
34	PSYC	NO	65	Junior	3.357	13	In-State	90.4
35	MSCI	YES	59	Sophomore	3.475	16	Out-of-State	774.6
36	BIOL	NO	80	Junior	3.968	19	Out-of-State	331.3
37	EXSS	YES	70	Junior	3.221	15	Out-of-State	552.6
38	MSCI	YES	48	Sophomore	2.784	12	Out-of-State	471.8
39	MSCI	YES	69	Junior	2.907	17	Out-of-State	263.4
40	MSCI	YES	58	Sophomore	2.482	15	Out-of-State	353.2
41	INTEL	YES	71	Junior	3.331	16	In-State	88.0
42	BIOL	NO	44	Sophomore	2.357	15	Out-of-State	502.7
43	PUBH	NO	61	Junior	3.361	18	Out-of-State	464.8
44	MSCI	YES	31	Sophomore	3.774	15	Out-of-State	407.7
45	EXSS	YES	30	Sophomore	3.333	19	In-State	121.8
46	MSCI	YES	68	Junior	3.654	12	Out-of-State	341.0
47	EXSS	YES	60	Junior	3.893	17	In-State	21.8
48	BIOL	NO	60	Junior	3.525	13	In-State	124.1
49	INTEL	YES	64	Junior	3.782	13	Out-of-State	415.6
50	MSCI	YES	65	Junior	3.915	17	Out-of-State	399.2

Investigation Data Set Continued ...

Student	Major	ACM	Cumulative Credits	Class	Cumulative GPA	Enrolled Credits	Residency	Distance
51	MGEDP	YES	32	Sophomore	2.531	17	In-State	27.0
52	MSCI	YES	41	Sophomore	3.265	15	In-State	17.7
53	INTEL	YES	53	Sophomore	3.234	16	In-State	116.9
54	EXSS	YES	60	Junior	4.000	14	In-State	24.6
55	PUBH	NO	0	Freshman	NA	17	In-State	92.9
56	BIOL	NO	58	Sophomore	3.862	12	International	NA
57	INFSY	NO	33	Sophomore	3.545	18	Out-of-State	485.8
58	INTEL	YES	41	Sophomore	3.963	20	Out-of-State	478.7
59	MSCI	YES	40	Sophomore	3.214	15	Out-of-State	973.3
60	EXSS	YES	39	Sophomore	3.121	17	Out-of-State	645.3
61	SUST	YES	42	Sophomore	2.419	16	Out-of-State	399.6
62	EXSS	YES	31	Sophomore	3.500	18	In-State	108.3
63	MATHA	NO	34	Sophomore	3.000	18	In-State	135.3
64	CSCI	NO	39	Sophomore	1.841	14	Out-of-State	573.1
65	MKTP	NO	52	Sophomore	2.818	16	Out-of-State	121.7
66	EXSS	YES	48	Sophomore	4.000	15	In-State	132.7
67	EXSS	YES	57	Sophomore	2.719	12	Out-of-State	488.9
68	PUBH	NO	58	Sophomore	3.804	16	Out-of-State	397.1
69	CRMJ	NO	29	Freshman	3.483	16	Out-of-State	499.2
70	SUST	YES	37	Sophomore	2.750	16	Out-of-State	484.8
71	SUST	YES	27	Freshman	2.019	17	Out-of-State	391.3
72	BIOL	NO	34	Sophomore	3.618	18	Out-of-State	447.7
73	EXSS	YES	34	Sophomore	3.603	16	Out-of-State	484.5
74	PUBH	NO	42	Sophomore	3.903	16	In-State	1.1
75	SOC	NO	32	Sophomore	3.750	16	Out-of-State	717.5
76	CSCI	NO	50	Sophomore	3.838	17	In-State	47.7
77	INTEL	YES	30	Sophomore	3.367	13	Out-of-State	386.4
78	EXSS	YES	30	Sophomore	2.926	16	Out-of-State	159.0
79	BIOL	NO	38	Sophomore	3.338	17	Out-of-State	265.0
80	PUBH	NO	32	Sophomore	3.250	16	In-State	108.3
81	MSCI	YES	99	Senior	2.717	16	Out-of-State	1063.5
82	MSCI	YES	49	Sophomore	3.939	16	Out-of-State	371.3
83	EXSS	YES	21	Freshman	2.320	15	In-State	102.7
84	MSCI	YES	91	Senior	3.019	16	In-State	15.4
85	PUBH	NO	58	Sophomore	3.029	14	In-State	92.9
86	CRMJ	NO	30	Sophomore	NA	16	In-State	14.4
87	BIOL	NO	30	Sophomore	3.500	16	Out-of-State	776.7
88	BCHEM	NO	35	Sophomore	4.000	15	In-State	14.7
89	BIOL	NO	33	Sophomore	3.879	17	In-State	20.0
90	STATS	NO	0	Freshman	NA	15	Out-of-State	466.8
91	BIOL	NO	31	Sophomore	3.452	15	Out-of-State	1010.4
92	EXSS	YES	33	Sophomore	2.939	17	Out-of-State	525.6
93	EXSS	YES	53	Sophomore	3.375	15	Out-of-State	288.9
94	EXSS	YES	68	Junior	3.784	14	In-State	21.8
95	BIOL	NO	19	Freshman	NA	17	Out-of-State	283.8
96	MKTP	NO	6	Freshman	NA	17	Out-of-State	620.2
97	EXSS	YES	0	Freshman	NA	17	Out-of-State	403.7
98	FINP	NO	6	Freshman	NA	17	Out-of-State	619.7
99	PSYC	NO	4	Freshman	NA	17	Out-of-State	613.6
100	MSCI	YES	63	Junior	NA	14	Out-of-State	263.3

1.3 Data Visualization for a Single Variable

1.3.1 Types of Data

Consider the following data we have for our study:

- Major
- Distance from home
- Number of credits enrolled

What kind of similarities and differences are there between these three pieces of information?

These differences will drive our decisions on the types of plots we will make. More formally,

Types of Variables

- **Quantitative variables** are numerical measurements that record "quantity" in a sense. Examples include concentration of a chemical, time, weight, and number of siblings.
- **Categorical variables** label outcomes into one of several mutually exclusive groups (categories). Some examples include habitat, season, and sport.

Example 1.5. Identify each of the variables in our study (listed above) by type as well as the following:

- T-shirt size
- Number of Black Pines in a tract of land
- Amount of money spent on food by household
- Favorite food

1.3.2 Plots for Categorical Variables

First we introduce the construction of some basic plots that are appropriate for each type of variable. Later we will discuss the information provided by the plots and relevant interpretations. Let's turn our focus on categorical variables first.

A **data distribution** provides the possible values of a variable and how often each value is observed. Creating a data distribution for a categorical variable is the first step in constructing and interpreting related plots. Let's construct a data distribution for some of the data in our course investigation.

As an example, let's consider the classification/year of each student from our investigation.

Classification/Year	Frequency	Relative Frequency
Freshmen	10	
Sophomore	48	
Junior	34	
Senior	8	

Bar charts/bar graphs are simple ways to visualize the the data distribution. The possible values of the variable are given on one axis and the other axis provides the frequency (or relative frequency). Bars are drawn to indicate the frequency with which each value is observed. Using the data distribution above, construct the resulting bar chart from the course investigation.

Comments and questions:

1. What is a general definition of the **mode**? What classification is the mode in our investigation?

2. Are there any other important pieces of information displayed in the bar graph?

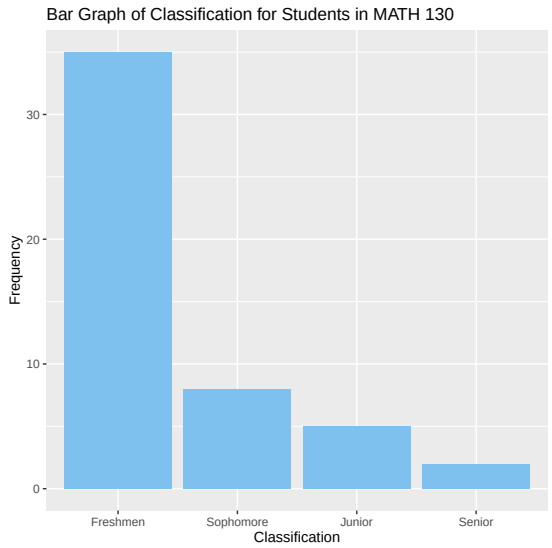
3. Would the overall appearance of the graph change if we used relative frequency rather than frequency on the vertical axis?

4. When would it be important to use relative frequency over frequency?

5. Are we allowed to arrange the categories on the horizontal axis in a different manner?

6. What can we say about the variability/diversity of classes?

7. Consider the following bar graph considering the classification for a random sample of 50 students taking MATH 130 at CCU. Compare the diversity of classifications between MATH 130 and STAT 201.



1.3.3 Plots for Quantitative Variables

Histograms are similar to bar charts, but for quantitative data. They also have a few important differences. First, let's construct the data distribution for the variable distance from home which is one of the quantitative variables in our investigation.

NOTE: The data are presented in numerical order.

1.1	1.1	1.1	11.6	11.6	14.4	14.7
14.7	15.2	15.2	15.4	16.0	17.7	18.2
18.2	20.0	21.8	21.8	21.8	23.3	24.6
27.0	47.7	68.9	73.4	88.0	90.4	92.9
92.9	102.4	102.7	107.2	108.3	108.3	116.9
121.7	121.8	124.1	132.7	135.0	135.3	138.3
151.4	159.0	165.4	179.8	199.6	263.3	263.4
265.0	283.8	288.9	331.3	341.0	353.2	357.0
357.4	371.3	386.4	391.3	397.1	399.2	399.6
403.7	407.7	415.6	447.7	454.1	462.2	464.8
466.8	469.1	471.8	478.7	480.1	484.5	484.8
485.8	488.9	499.2	502.7	504.8	525.6	552.6
573.1	575.9	613.6	619.7	620.2	621.6	645.3
713.1	717.5	727.4	774.6	776.7	973.3	1010.4
1063.5						

Distance from Home	Frequency

Histogram for Distance from Home

Observations:

Now, let's construct the data distribution for the number of cumulative credits completed which is another of the quantitative variables in our investigation. **NOTE:** The data are presented in numerical order.

0	0	0	4	6	6	19	21
27	29	30	30	30	30	30	31
31	31	32	32	32	33	33	33
34	34	34	34	35	37	38	38
39	39	40	40	41	41	42	42
44	45	48	48	49	50	52	53
53	54	56	57	58	58	58	58
59	59	60	60	60	60	60	60
61	61	62	63	64	64	65	65
65	66	68	68	69	70	70	71
73	75	75	77	78	78	80	83
84	84	87	87	91	95	99	107
112	126	135	149				

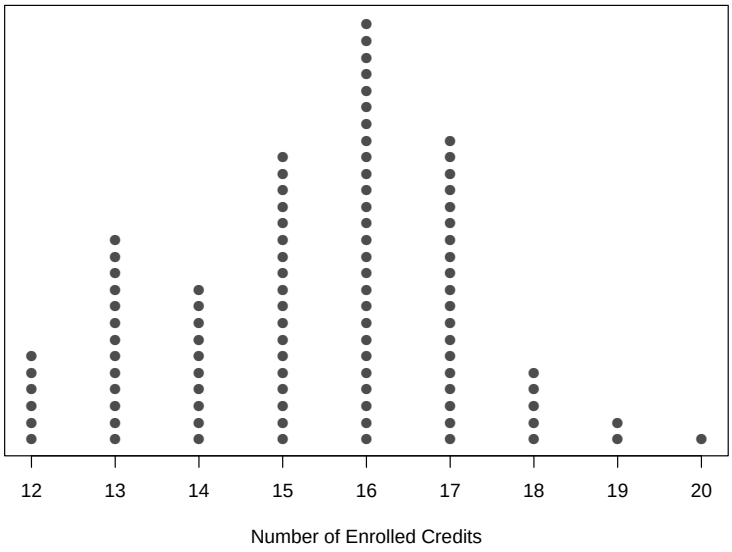
Cumulative Completed Credits	Frequency

Histogram for Number of Cumulative Credits Completed

Observations:

Dot plots are similar to histograms except the original data are represented as dots rather than bars. Any repeated values are represented as stacked dots.

Consider the following dot plot of the number of enrolled credits during fall 2022 for the random sample of students in our investigation.



Questions:

1. What was the most number of credits a student was enrolled within our investigation? How many students were enrolled in this many credits?

2. How many students sampled were enrolled in at least 18 credit hours during fall 2022?

3. What percentage of students sampled were in enrolled in less than 15 credit hours during fall 2022?

1.3.4 Key Features of Plots for Quantitative Variables

We can summarize our observations of the previous histograms under four categories. These are the key features to look for in a plot of quantitative data. We are not doing any formal computations. Rather, we are simply getting an estimate of these features from what we can see in the plot.

1. **Shape** describes the overall layout of the data and typically can be categorized in one of four shapes. Note that real data will rarely follow one of these shapes exactly.

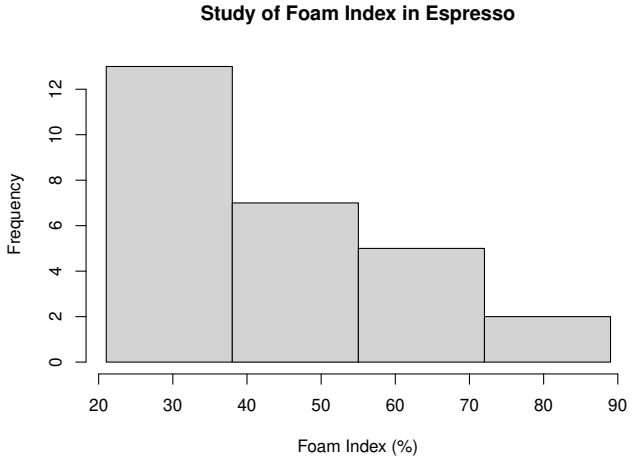
2. **Center** describes the centrality of the data. We can think of it as where the histogram would be "balanced", the approximate location of the middle data point, or the place where we see the majority of the data.

3. **Spread** gives us an idea of where the data lies (minimum value to maximum value) and if the data is mostly concentrated in one area or spread evenly throughout.

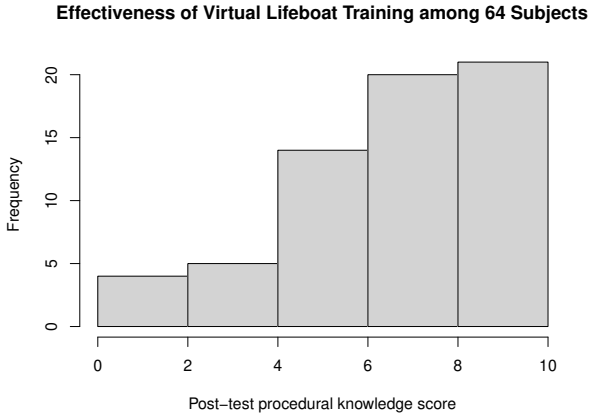
4. **Outliers** are any data points that are extremely large or small compared to the majority of the data. Check to see if there are any points that look as if they do not fit in with the rest.

Check our observations of the previous histograms to ensure we have addressed all the key features. Then identify the key features of the following examples.

Example 1.6. In A. Parenti et. al. (2014) "Comparison of Espresso Coffee Brewing Techniques", researches examined the foam index (%) for brewing espresso. The data consists of 27 brews and is displayed in the following histogram. Assess the key features of the histogram.



Example 1.7. In J.Jung and Y.J. Ahn (2018). "Effects of Interface on Procedural Skill Transfer in Virtual Training: Lifeboat Launching Operation Study," the authors examine the effectiveness of training to launch lifeboats using scores from a follow-up test. First assess the key features of the histogram.



(a). Approximately how many scores were below 4?

(b). Approximately what percent of scores were below 4?

(c). Approximately what percent of scores are at least a 6?

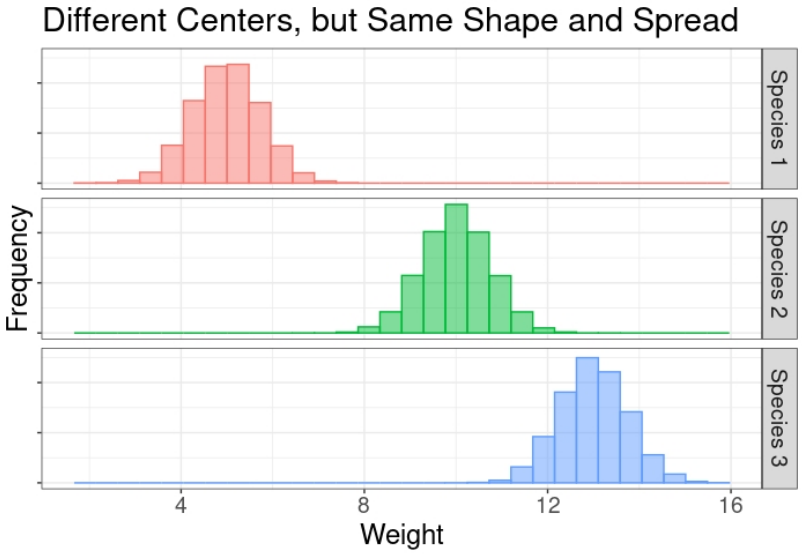
(d). What range of scores occur most frequently among subjects?

1.3.5 Recapping the Meaning of Center and Spread

To tease out the difference between center, spread, and shape, imagine we have data on weights of organisms from three different species we wish to compare.

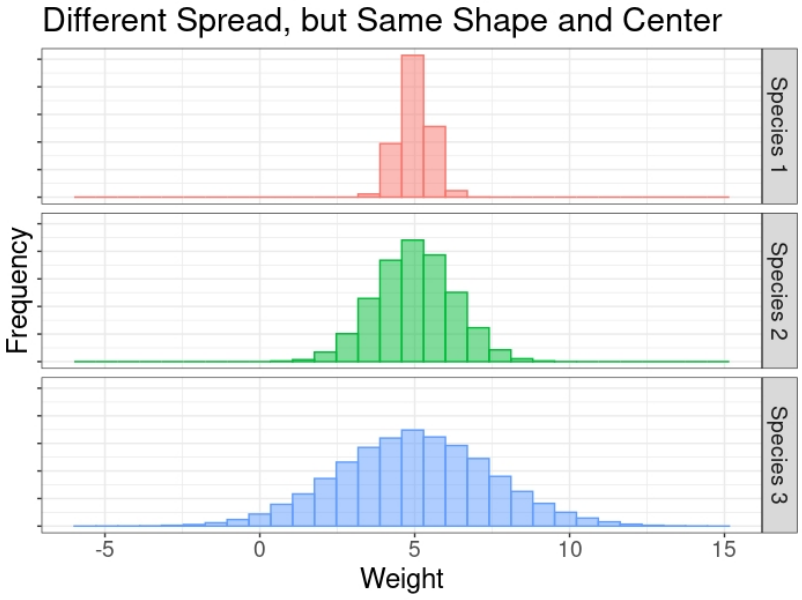
Center: Center measures **location** of the data. That is, center describes **where** the data are grouped. We measured center with mean or median, depending on the shape of the data.

The three species all have weights that are symmetric with the same spread. However, species 3 tends to be heavier with a center around 13 while species 1 tends to be the lightest with a center around 5.



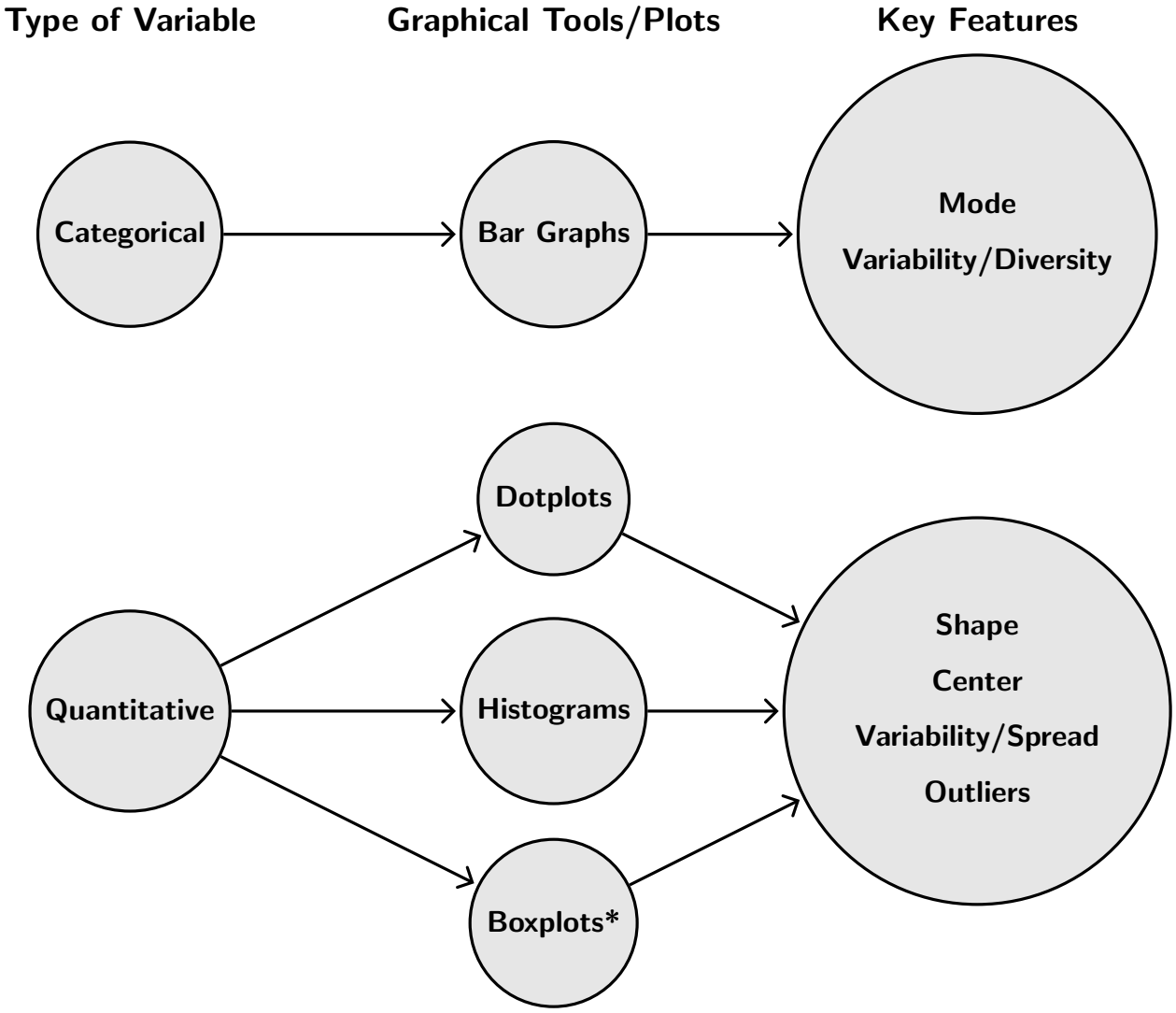
Spread: Spread measures the **variability** of the data. That is, spread describes how tightly or loosely grouped the data are. We measured spread with standard deviation or IQR, depending on the shape of the data.

The three species all have the same center around 5 and all seem to be from the same type of symmetric distribution. However, species 1 tends to have weights that are less varied (lowest spread) while species 3 tends to have weights that are most varied (highest spread).



1.3.6 Overall Summary of Plots

The following map summarizes the graphical tools and key features for both categorical and quantitative variables.



Use the this page to construct some additional plots from our course investigation and comment thoroughly on what you see.

1.4 Data Summarization with Numbers

In the previous section we created graphical summaries of our data and assessed the plots for certain key features. In this section, we revisit those key features in a more formal manner to obtain numerical summaries of center and spread as well as a test for outliers.

1.4.1 Measures of Center

You are likely familiar with methods for summarizing central tendency in the data. List those here and use your calculator to find them for each of the quantitative variables in our study.

TI83/84 Steps for computing measures of center and spread:

- Enter the data: `STAT`→`EDIT`→`enter the data in` L_1
- Compute summary Statistics: `STAT`→`CALC`→`1-VarSTATS`

Now we slow down now to understand exactly how each is measuring central tendency and how these values are influenced by shape of the data.

A silly example to understand the difference between the mean and median:

Suppose we have 5 children sitting on a see-saw at positions 3ft, 4ft, 5ft, 6ft, and 7ft. Draw a picture representing this scenario and find the mean and median by visual inspection first and then performing the calculations.

Now suppose Little John who was sitting at the 7ft marker is feeling rather shy and decides to sit at the 17ft marker away from all the other kids. Draw a picture of the new scenario and discuss what will happen to the mean and median. Perform the calculations to back up your intuition.

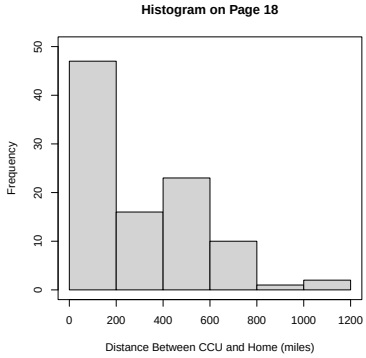
A numerical summary is <u>resistant</u> if it is influenced little by extreme observations. Which measure of center is resistant?

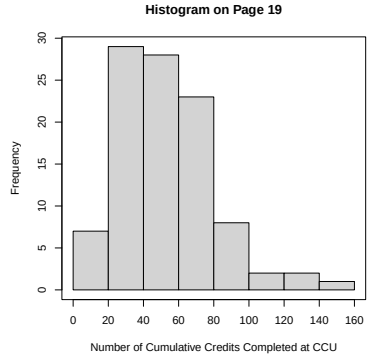
Understanding the Relationship between Shape and Center

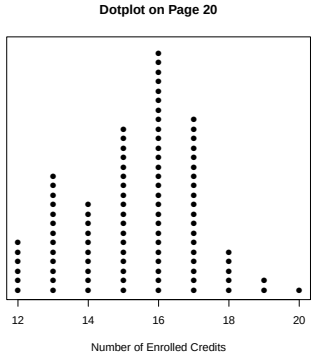
Therefore,

- if the data is skewed (or has outliers) we will use the _____
- if the data is symmetric we will use the _____

Discuss: Consider the plots for several of the quantitative variables in our course investigation. Which measure of center is most appropriate for the some of the quantitative variables in our study and why?







1.4.2 Measures of Spread

When we speak about "measures of spread", we are talking about the variation in the data. Are the values very similar or are they very different?

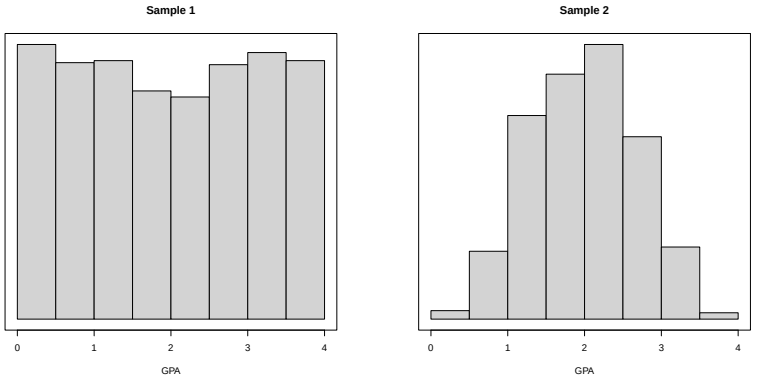
> The simplest measure of spread is the **range**: Largest observation - smallest observation.

→ Find the range for our two data sets. Is the range resistant?

3ft, 4ft, 5ft, 6ft, and 7ft:

3ft, 4ft, 5ft, 6ft, and 17ft:

→ Consider the following two graphs relating to GPA of separate samples of students. For each sample, approximate the value of the following statistics:



- Mean
- Median
- Range

→ Even though the above samples have the same range, would you say that they have the same spread/variability?

> A more informative measure of spread in the data is the **standard deviation**. This measures the typical/average distance of observations from the mean.

→ Find the standard deviation of our data set of 3ft, 4ft, 5ft, 6ft, and 7ft:

→ Do you expect the standard deviation of the data set 3ft, 4ft, 5ft, 6ft, and 17ft to be larger or smaller? Why? Calculate the value by hand or using the calculator to confirm.

→ Is standard deviation resistant to outliers?

→ Determine if the values can or cannot be a standard deviation.

1.2 $\quad 1 \times 10^{-5} \quad -2.1 \quad 100$

A more resistant measure of spread in the data is the **intequartile range (IQR)**. This measures the range of the middle 50% of the data. In order to understand how to compute IQR, we need to cover a few concepts:

- **Percentiles** - the pth percentile is the measurement to which $p\%$ of all measurements fall below it and $(100 - p)\%$ lie above it.
- **Quartiles** - special percentiles that divide the data into quarters.

- **IQR** - $IQR = Q_3 - Q_1$. This measures the spread of the middle 50% of the data.

Example 1.8. The following data gives the HP (Hit Points) for a random sample of generation one Pokemon.

(a). Find the quartiles and the IQR for HP.

10	30	35	39	40	44	45	45	50	55

(b). We will now add Jigglypuff to the data and recompute the quartiles and the IQR.

10	30	35	39	40	44	45	45	50	55	115

NOTE: The standard deviations for the above sets of data are $s_1 = 12.526$ and $s_2 = 25.733$ respectively.
$\rightarrow$ Is the IQR resistant to outliers?

$\rightarrow$ What measures of spread should be used for skewed data? For symmetric data?

1.4.3 Outliers and Boxplots

Test for Outliers

Five Number Summary:

Outliers:

Example 1.9. The following values are a sample of characters from Walking Dead and their corresponding number of zombie kills for the first three seasons. Find the five number summary, test for outliers, and construct boxplots.

Obs.	Character	Kills
1	Gargulio	1
2	Big Tiny	1
3	Cesar	2
4	Morgan	3
5	Beth	3
6	Shupert	4
7	Morgan	4
8	Mexican Father	4
9	Tyrese	5
10	Morales	6
11	Milton	7
12	Woodbury Army	11
13	Merle	11
14	The Govenor	13
15	Carl	14
16	Glenn	19
17	Hershel	21
18	Michonne	31
19	Daryl	61
20	Rick	91

1.5 Data Visualization for Comparing Two Variables

So far we have examined plots of one variable at a time.

- For categorical variables, we can use a bar chart.
- For quantitative variables, we can use a histogram, dotplot, and/or boxplot.

This helps us to better understand the distribution of each individual variable.

Question: Recall our original questions from our course investigations:

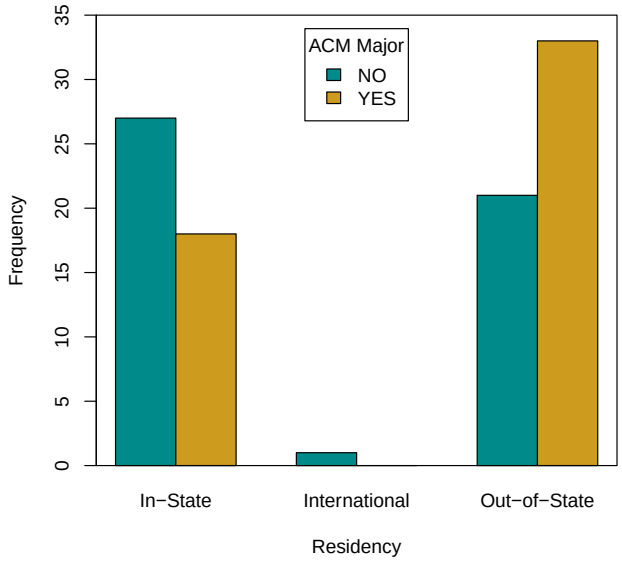
1. What kind of impact does the ACM have on the diversity of hometowns among majors?
2. Is the distance from home higher among those majors available through the ACM than other majors?
3. Is there a difference between the percentage of in-state and out-of-state students who choose to major in a program available through ACM?
4. Let's add an additional question. Is there a relationship between the number of credits enrolled during a semester and the cumulative GPA?

What other visualizations could we look at that might explore the relationship between our variables and get us closer to answering the original questions in our investigation?

1.5.1 Graphs for Comparing Two Categorical Variables

Side-by-side Bar Graphs are bar graphs that compare two categorical variables.

Example 1.10. The following side-by-side bar graph compares the residency of students to choosing a major in a program available through ACM at CCU.



(a). Approximately what proportion of out-of-state students chose a major in a program available through ACM?

(b). Approximately what proportion of in-state students chose a major in a program available through ACM?

(c). Does there seem to be an association between residency and choosing a major in a program available through ACM?

1.5.2 Graphs for Comparing a Quantitative Variable at Different Levels of Categorical Variables

Side-by-Side Boxplots are a powerful tool for examining the distribution of a quantitative variable at different levels of a categorical variable. Essentially, separate boxplots are constructed on the same axis for each level of a categorical variable.

Example 1.11. The following table provides the five number summary for the distance between CCU and their hometown for each student separated by whether or not their major is within the ACM.

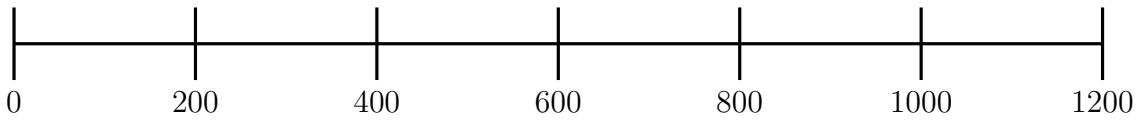
Major in ACM

Min	Q_1	Q_2	Q_3	Max
15.2	107.8	371.3	482.3	1063.5

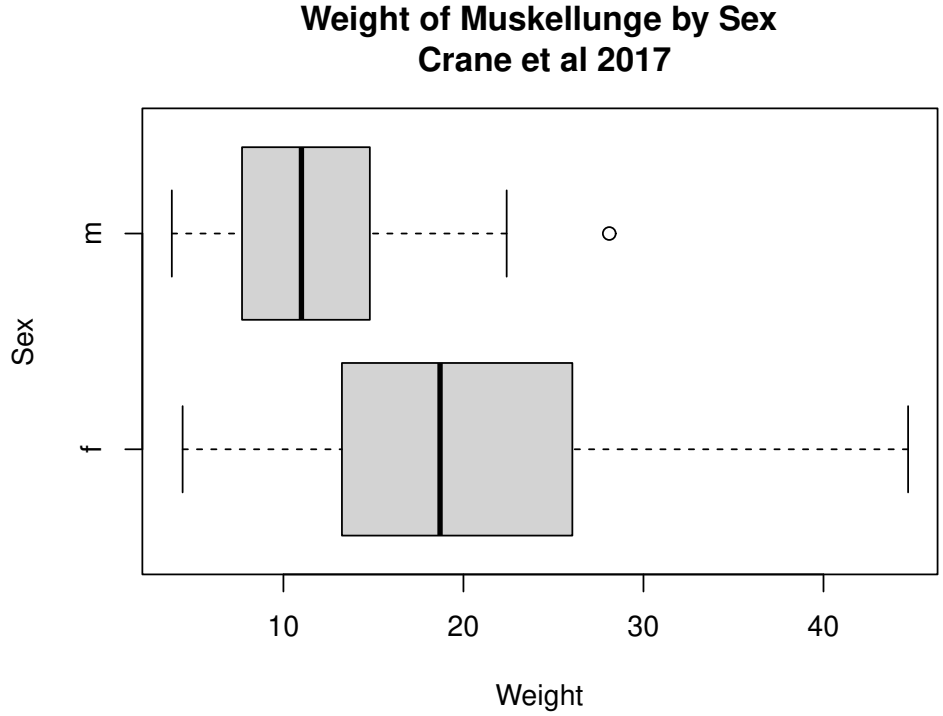
Major Not in ACM

Min	Q_1	Q_2	Q_3	Max
1.10	19.55	136.80	465.30	1010.40

Use the following axes to construct a side-by-side boxplot.



Example 1.12. Crane (2017) and his CCU research group studied the weight, length, girth, sex, and reproductive status of 869 Muskellunge fish. A summary of weight by sex is provided.



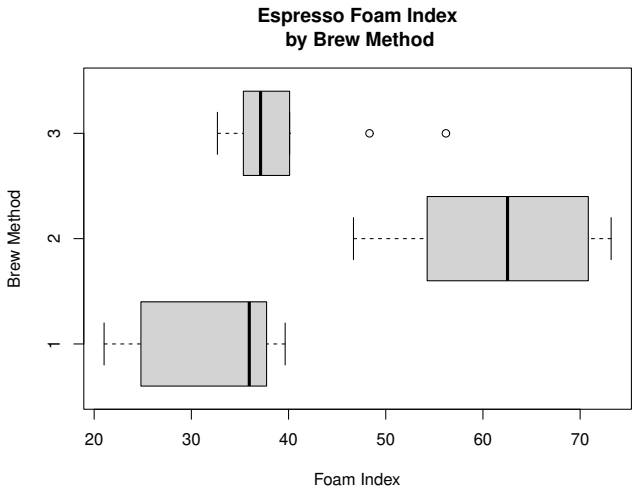
(a). Label the parts of the boxplot with what each represents.

(b). Which sex tends to be heavier? How can you tell?

(c). Which sex tends to have more variability in weights? How can you tell?

(d). Describe the shape of the distribution of weights for each sex.

Example 1.13. Recall, A. Parenti et. al. (2014) "Comparison of Espresso Coffee Brewing Techniques" where researchers examined the foam index (%) for brewing espresso. The data is displayed in the following dot plot. However, it it important to note that the foam index was actually calculated for three different brewing techniques. We view the data by brewing technique using a dot plot. Brewing methods are coded as Method 1=Bar Machine, Method 2 = Hyper-Espresso Method, and Method 3 = I-Espresso System.



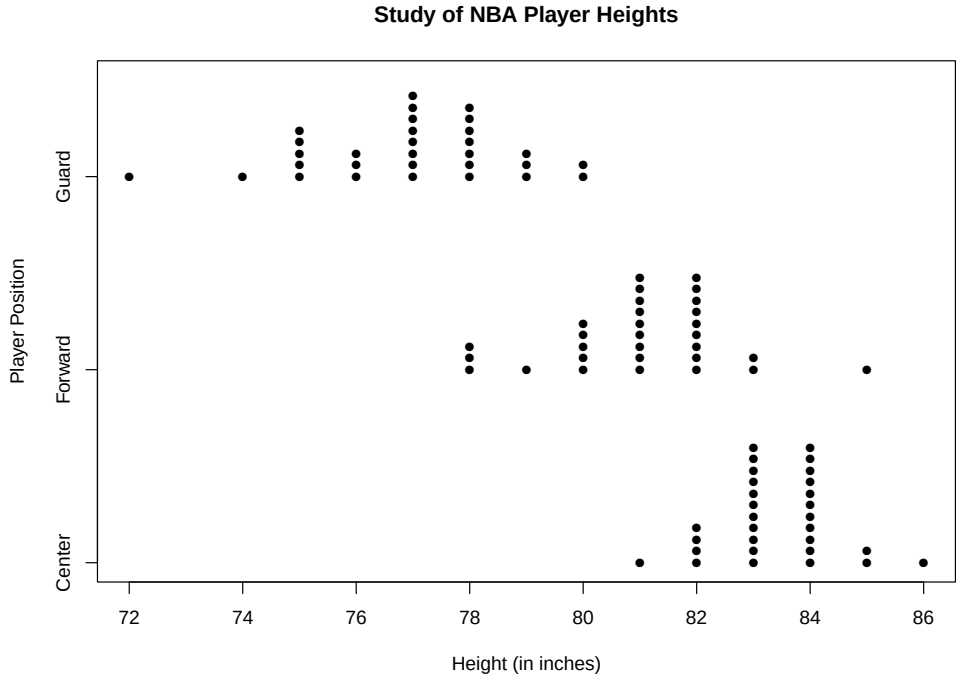
(a). Which method of brewing generally has the highest foam index? Lowest?

(b). Describe the distribution of foam index for method 3 of brewing (I-Espresso System).

(c). Describe the shape of the distributions for brewing methods 1 and 2.

Side-by-Side Dot plots are separate dot plots constructed for each level of a categorical variable. Remember, any repeated values are represented as stacked dots.

Example 1.14. A student was interested in comparing the heights of different player positions in the National Basketball League (NBA). Therefore, she took a stratified random sample of 30 players from each position (Center, Forward and Guard). The results are presented in the following plot.



(a). Out of all the data together, what height was the tallest player observed? What height was the shortest player observed?

(b). Is a particular position typically taller or shorter than the others? Explain.

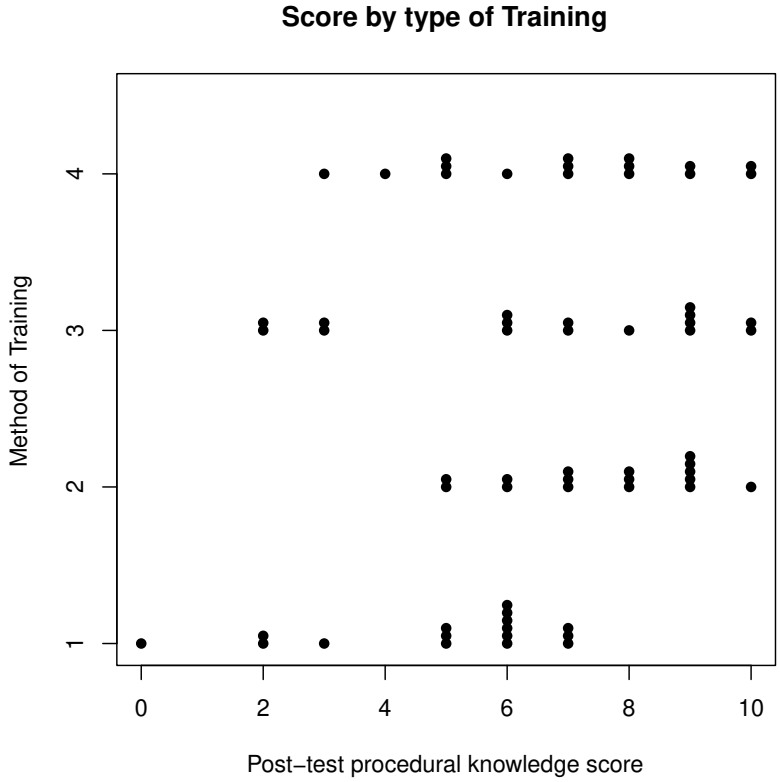
(c). Do there appear to be any outliers within any of the positions?

(d). How would you describe the shape of the data for centers? What about guards?

(e). Which position displays the least amount of variability?

Example 1.15. In J.Jung and Y.J. Ahn (2018). "Effects of Interface on Procedural Skill Transfer in Virtual Training: Lifeboat Launching Operation Study," the authors examine the effectiveness of training to launch lifeboats using scores from a follow-up test. When we also consider the type of training in our plot we see the following. The coding used for the training groups is defined below.

- 1 = control group with traditional lecture and materials
- 2 = monitor and keyboard
- 3 = head monitor display and joypad
- 4 = head monitor display and wearables



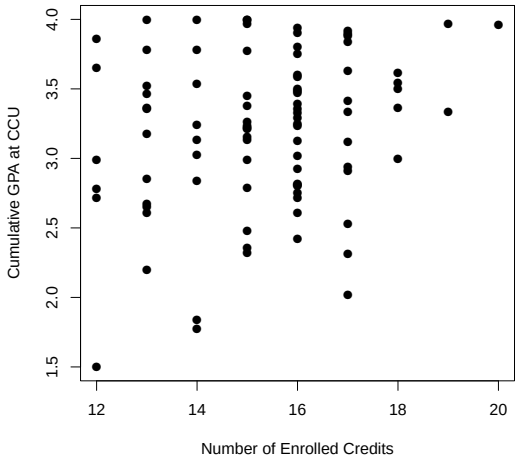
(a). What method of training performed the worst in general? The best? Explain.

(b). What method of training has the most variability? Least?

1.5.3 Graphs for Comparing Two Quantitative Variables

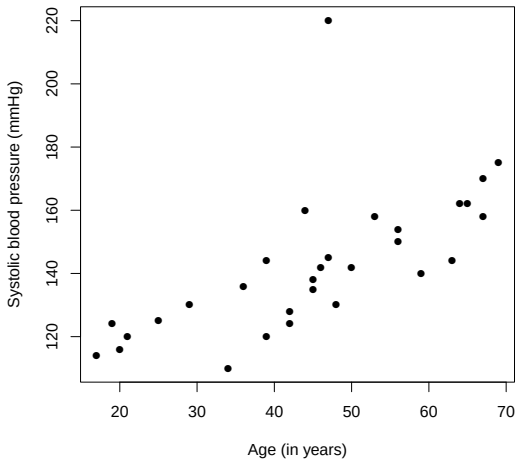
Scatterplots are an easy way to study the relationship between two quantitative variables measured on the same subject or at the same time point. The data is plotted as (x, y) coordinates for each subject/time point. If it is thought that one variable exerts influence on the other, it is plotted on the x-axis and called the **explanatory variable**. The variable on the y-axis is called the **response variable** and may be impacted by or respond to the explanatory variable. We will discuss this in more detail later.

Example 1.16. Consider the following scatterplot between the cumulative GPA and number of enrolled credits for the random sample of students during fall 2022.



What are some basic observations?

Example 1.17. Isolated systolic hypertension, which is an elevation in systolic but not diastolic blood pressure, is the most prevalent type of hypertension (especially in the elderly). A study investigated the relationship between age (in years) and systolic blood pressure (SBP, measured in mmHg) in adult males. There were $n = 30$ individuals in the study. The following scatterplot displays the results of the study.

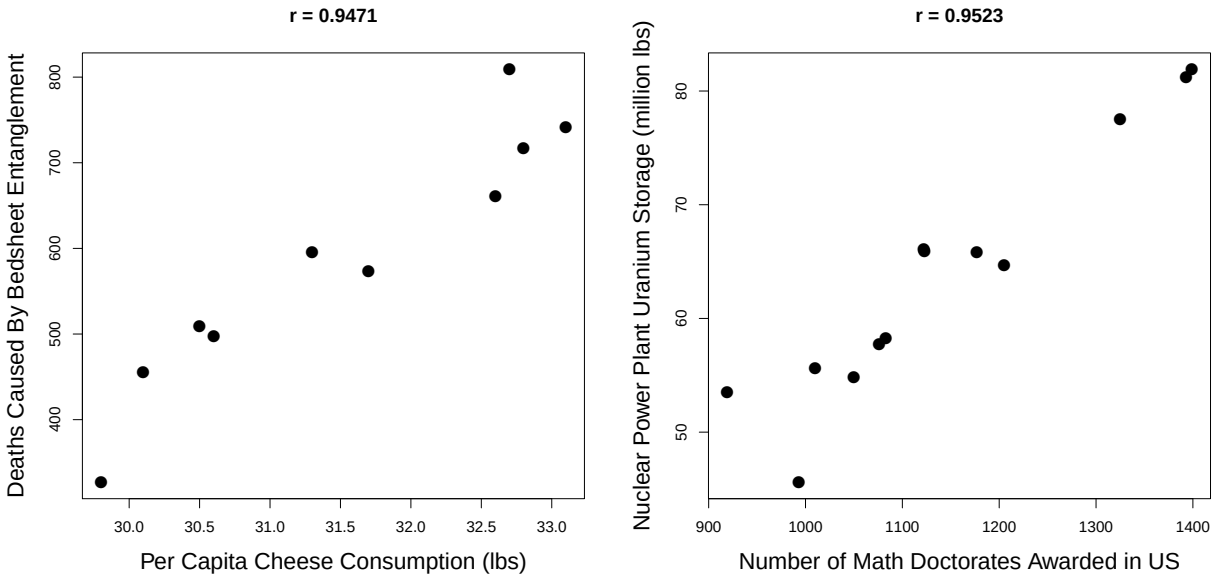


What are some basic observations?

In **Example 1.17**, we examined that systolic blood pressure generally increases as age increases in adult males. Does this mean that aging in mean *causes* SBP to increase?

A Chicago newspaper once reported that "there is a strong correlation (association) between the number of fire trucks at the scene of a fire and the amount of damage that the fire does." Does this mean an increase in the number of fire trucks at the scene of a fire *causes* more damage?

Example 1.18. Consider the following plots. They demonstrate some level of strong association or correlation between the variables. Do you feel comfortable suggesting the x causes y to happen? That is, can we conclude that just because two variables have a strong correlation that there is a cause and effect relationship? **NOTE:** We will discuss scatterplots and correlation in more detail towards the end of the semester.



NOTE: It is easy to find examples like the ones above. Do not be fooled when someone attempts to convince you that there is a causal relationship between two variables. Chances are good that there isn't one. In the next section, we will discuss the best way to determine a casual relationship.

Use the this page to construct some additional comparative plots from our course investigation and comment thoroughly on what you see.

1.6 Study Designs and Conclusions

1.6.1 Observational vs Experimental Studies

We previously noted that while per capita cheese consumption and deaths by bedsheet entaglement have a strong correlation, we do not feel that eating more cheese causes more such deaths. Similarly, a strong correlation between math doctorates and uranium storage does not necessarily imply that increases in math doctorates cause increased uranium storage. This brings to light an important distinction in vocabulary.

Types of Relationships:

- **Causation (Cause and Effect)** - Changes in one variable (**explanatory variable**/cause), are directly responsible for change in another variable (**response variable**/effect).
- **Association** - The variables are related in some way, but perhaps not directly. As one variable changes, the other variable changes in a predictable way. However, one cannot determine which variable is responsible for the change and perhaps it is neither.

Example 1.19. In the following studies, identify the explanatory and response variables.

(a). Does spending more time outdoors help one to be less afraid of spiders and snakes? (Zsido et. al. 2022)

(b). Is cancer risk increased by increased exposure to radiation from cell phones? (2016 US National Toxicology Program Report)

(c). Is depression more likely with less physical activity? (Lucas et al., 2011)

(d). Does lack of social interaction increase symptoms of Alzheimer's disease? (Hsiao et. al., 2018)

Example 1.20. Suppose we are interested in the long term effects of playing video games as an adolescent. Specifically, does playing more violent video games in adolescence result in an individual being more violent as an adult? To study this suppose we record how many hours of video games are played and of what type in many different households. We then follow-up when the children are grown and record their behavioral tendencies towards violence. Would we be able to conclude that violent video games resulted in violent behavior later on in life? Why or why not? Discuss.

Observational Studies

Experimental Studies

Example 1.21. Determine if each scenario is a controlled experiment or an observational study. Identify if the study can show causation or association. Suggest lurking variables where appropriate.

Study 1(Wilcox, 2012)Researchers at Columbia University have learned that using Facebook may be tied to obesity, due to the negative eating habits that could result from frequent visitation of social network sites. One part of the study surveyed 470 people about their Internet use. Those who used Facebook the most had higher BMI than those who were not as frequently engaged.

Study 2 In a related study by Wilcox, 84 people were randomly assigned to view either Facebook or CNN's website. When presented with both a healthy and an unhealthy snack option after browsing the web, 80% of Facebook browsers chose the unhealthy snack. Only 30% of CNN browsers chose the unhealthy snack.

Study 3 The US National Toxicology Program released partial results of a rodent study on the effects of radio frequency radiation exposure, similar to cell phone exposure in humans. Rats were randomly assigned to GSM or CDMA modulated RFR with specific absorption rates of 0, 1.5, 3 or 6 W/kg. Exposure began in utero and continued for 2 years. There was a statistically significant increase in rates of brain and heart tumors for the exposed mice.

1.6.2 Components of a Good Experiment

Four components of a good experiment:

1. *Control Group Comparison* - The control group does not receive the treatment of interest, but perhaps a placebo or treatment that is currently used. This allows for comparison of treatment results with another group so that the effect of the treatment can actually be measured.

2. *Randomization* - randomizing which participants receive the treatment removes or minimizes the effects of lurking variables. You can also balance the groups with respect to variables that you know will influence the response.

3. *Blinding and Double Blinding* - If possible the subjects should be "blind" to the treatment they are receiving. When those dealing with the subjects and recording information are also "blind" to the subject's treatment group, the study is called double blind. This is ideal since subjects and researchers could intentionally or unintentionally provide support for a particular level of the treatment.

4. *Replication* - Assigning multiple subjects to each treatment in the experiment. If we only have one subject per treatment, we cannot be sure if the results reflect the entire population.

Example 1.22. In "Beyond the beauty of occlusion: medical masks increase facial attractiveness more than other face coverings", Heis and Lewis (2022) studied the effects of facial occlusion on attractiveness. The experiment was set up seven months after wearing masks became mandatory in the UK. Each of the 42 participants was presented with each of the 40 faces four times (medical mask, cloth mask, book, non-occluded) as shown below. Faces were presented in a randomized order. Participants rated the faces using the numbers 1–7.

Faces were rated as significantly more attractive in the medical mask condition compared to in the cloth mask condition ($p = 0.020$), notebook occluder condition ($p < 0.001$), and control condition ($p < 0.001$). In addition, faces in the cloth mask condition were rated as significantly more attractive than in the control condition ($p < 0.001$), but they were only non-significantly more attractive than the notebook condition ($p = 0.123$). Further, faces in the notebook condition were rated as significantly more attractive than in the control condition ($p = 0.005$). Identify any good practices used in this experimental. Are any omitted?

1.6.3 Experimental Design

Suppose we are interested in studying a cause and effect relationship using an experimental study. There are many common experimental designs that a researcher may use to assign subjects to treatments groups. Here we discuss three of those and introduce related vocabulary.

Terminology

- Subjects/experimental units - who or what is receiving the treatment in the study
- Response - the outcome of interest; what is being measured by the researcher
- Factors - variables that may affect the response (explanatory variables)
- Factor levels - the possible categories or levels for each factor
- Treatments - all combinations of the factor levels considered in the study; exactly what is being randomly assigned to the subjects

Design I: Single Factor Design - the effects of only one factor are considered.
Note that the factor levels and treatments are the same for a single factor design!

Example 1.23. A veterinarian wants to study the effect that type of diet has on the weight gain for puppies. For the experiment, the vet divides 144 puppies of approximately the same age and breed into three equal sized treatment groups. Each group is then randomly assigned to one of three types of diet (diet A, diet B, diet C). After three weeks, the weight gain for each puppy is observed.

- Experimental units:
- Response:
- Factor:
- Factor Levels:
- Treatments:
- Outline:

Design II: Multifactor Design (Two Factor) - the effects of two or more factors are considered.

Example 1.24. Now, the veterinarian wants to study the effect that type of diet (diet A, diet B, diet C) and exercise program (none, medium, intense) have on the weight gain for puppies. For the experiment, the vet divides 144 approximately the same age and breed into equal sized treatment groups. Each group is then randomly assigned to one of the combinations of diet and exercise program. After three weeks, the weight gain for each puppy is observed.

- Experimental units:
- Response:
- Factor:
- Factor Levels:
- Treatments:
- Outline:

Design III: Block Design - We are only interested in one factor, but we know a second factor may influence the results. The second factor is taken into consideration and is often something that cannot be randomized. It is sometimes called a "nuisance factor".

Example 1.25. Now, let's go back to the original study where the veterinarian is examining the effect that type of diet has on the weight gain for puppies. The vet has a collection of puppies that are of similar ages but of different breeds. If fact, the vet has 27 Shih Tzu puppies, 27 Goldendoodle puppies, and 27 Beagle puppies. As stated the vet is interested in the effect of diet, but believes that the breed may influence the results. Thus, the vet randomly assigns 9 of each breed to each type of diet (diet A, diet B, diet C). After three weeks, the weight gain for each puppy is observed.

- Experimental units:
- Response:
- Factor:
- Block:
- Outline:

Example 1.26. Let's revisit three studies we looked at previously and determine whether each is single factor, multi(two) factor, or block design.

Study on Web Browsing: In a study by Wilcox, 84 people were randomly assigned to view either Facebook or CNN's website. When presented with both a healthy and an unhealthy snack option after browsing the web, 80% of Facebook browsers chose the unhealthy snack. Only 30% of CNN browsers chose the unhealthy snack.

Study on Masks: In Hies and Lewis (2022), researchers studied the effects of facial occlusion on perceived attractiveness for four different methods of occlusion. Additionally, there was a concern that base attractiveness may influence results so faces were classified as "attractive" or "unattractive". This was done by choosing faces for the study as the 20 most attractive and 20 least attractive faces based on previous ratings included in the Chicago Face Database.

Study on Cell Phone Radiation: The US National Toxicology Program released partial results of a rodent study on the effects of radio frequency radiation exposure, similar to cell phone exposure in humans. Rats were randomly assigned to GSM or CDMA modulated RFR with specific absorption rates of 0, 1.5, 3 or 6 W/kg. Exposure began in utero and continued for 2 years. There was a statistically significant increase in rates of brain and heart tumors for the exposed mice.

Example 1.27. What kinds of conclusions can we make in our course investigation? Explain.

2 Uncertainty in Data

2.1 Probability

Recall our overall goal, which is to estimate parameters using sample statistics. Given a certain statistic from our sample, what values of the parameter are likely? To flip the question around: given an assumed value of the parameter, how likely is an observed statistic? Either way, we will need an understanding of how to judge random events as likely and unlikely. We will need to understand probability.

2.1.1 What is a Probability?

Probability is the study of

The probability of an event is a number between 0 and 1 that measures how likely that event is to occur. While we can never perfectly know whether or not a random event will occur, we can understand which random events are more likely than others.

- The closer to 0, the _____ the event
- The closer to 1, then _____ the event

Events with probabilities of around or higher than 10% (one in ten) should be considered common. Events with probabilities at or less than 1% (one in a hundred) should be considered rare. They are possible, but you'd be surprised to see them!

Example 2.1. Suppose you roll two six-sided dice at the same time.

(a). Consider just one of the dice. Which is more likely: that you roll a 3, or that you roll a 5?

(b). Consider both dice. Which is more likely: that at least one of the dice is a 6, or that both are?

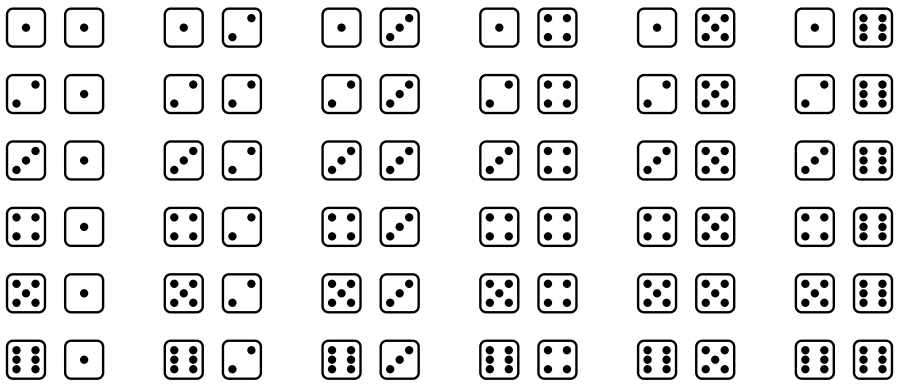
(c). Now, let's conduct a little experiment. Your instructor will pass out a pair of dice to each student. Roll your pair one time. Did at least one of your dice display a 6?

(d). Combining the results for the entire class, compute the relative frequency or the proportion of rolls that resulted in at least one 6.

This type of calculation is known as an **empirical probability**.

Empirical probability:

(e). Prior to rolling the pair of dice, we actually know all outcomes that are possible. The following picture lists all possible outcomes when rolling a pair of dice.



This list of all possible outcomes is known as the **sample space**.

Sample space: The sample space is the set of all possible outcomes of a random experiment. Typically denoted by S.

From this list, how many total outcomes are possible when a pair of dice are rolled? How many times do we observed at least one 6 from this overall list (sample space)?

(f). From a sample space, we often want to compute the probability of an event.

Events: An event is any set of outcomes, in other words, a subset of the sample space. Typically denoted capital letters at the beginning of the alphabet such as A, B, and C.

Let A be the event that your roll at least one 6. Using your results from (e), what is $P(A)$?

This type of probability is known as a **theoretical probability.**

Theoretical probability:

(g). Let B be the event you roll two 6s. What is $P(B)$?

(h). Do the results from (f) and (g) match our intuition in (b)?

(i). Let C be the event that your roll a pair of numbers whose product is 5. What is $P(C)$?

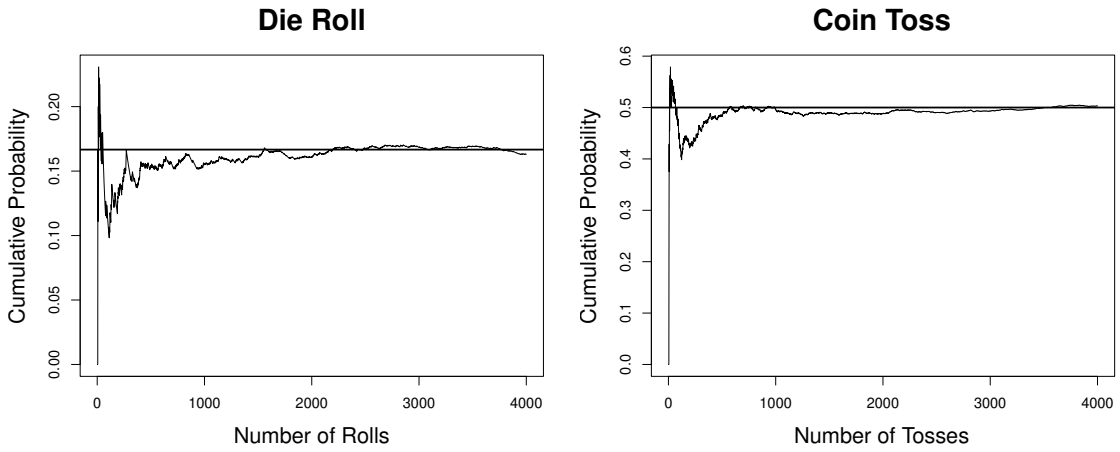
(j). Let D be the event you roll a pair of numbers whose sum is between 2 and 12. What is $P(D)$?

Now that we have a basic understanding of empirical and theoretical probability, consider the example of rolling a single six-sided die.

- Theoretical probability of rolling a 6:
- If you roll a die exactly six times, does that mean you will see exactly one 6? Why or why not?

Now, suppose you flip 10 coins, and 6 of those coins come up heads.

- Theoretical probability of flipping a heads:
- Empirical probability of flipping a heads given this data:
- What do you think would happen if you flipped 100 coins? 1000?



The law of large numbers: The more we repeat an experiment, provided each repetition is identical and independent, the empirical probabilities of the outcomes will approach their theoretical probabilities.

2.1.2 Calculating Probability

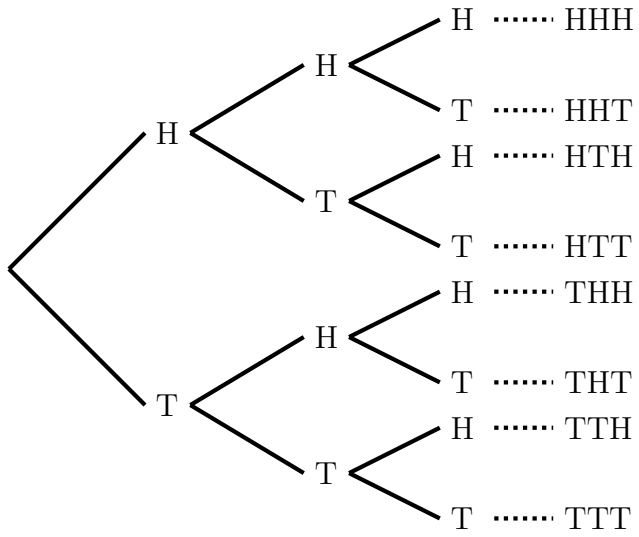
We computed many probabilities within the previous pages. To formalize the process, note that:

- If $P(A) = 0$, then A is impossible, the empty event, $\varnothing$.
- If $P(A) = 1$, then A is certain, the sample space, S.

Probability of an event:

When all outcomes in the sample space S are *equally likely*, then the probability of an event A can be calculated as:

Example 2.2. Suppose we toss a coin three times. The following tree diagram displays the sample space for this experiment.



(a). When a coin is tossed three times, what is the probability of 0 heads? Would this be a theoretical probability or an empirical probability? Explain.

(b). When a coin is tossed three times, what is the probability of 1 or 2 tails?

(c). When a coin is tossed three times, what is the probability of at least 1 tail?

(d). When a coin is tossed three times, what is the probability of 2 and 3 heads?

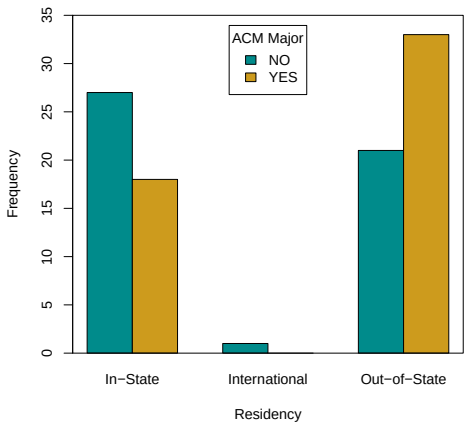
2.1.3 Contingency Tables

We will now introduce a statistical tool called a **contingency table** that is useful for examining the association between two categorical variables along with organizing probabilities of different events. In a contingency table,

- rows represent the categories of one variable
- columns represent the categories of the other variable
- where the rows and columns cross are referred to as cells
- the frequency of observations that fall into each combination of categories go into the cells.

As we will see, organizing probabilities into a contingency table will be a very useful tool.

Example 2.3. In **Example 1.10**, we considered the side-by-side bar graph comparing the residency of students to choosing a major in a program available through ACM at CCU from our investigation. We will now represent the information from this graphical tool in the form of a contingency table.



	ACM Major		
	Yes	No	**Total**
In-State	18	27	45
International	0	1	1
Out-of-State	33	21	54
Total	51	49	100

(a). What proportion of students chose a major in a program available through ACM?

(b). What proportion of students are out-of-state students and chose a major in a program available through ACM?

(c). What percentage of out-of-state students chose a major in a program available through ACM?

(d). Can we consider the above proportions and percentages empirical or theoretical probabilities? Explain.

2.1.4 Special Events Related to Probability

In our previous examples, you have seen different ways of combining multiple events to calculate more probabilities. We will now formalize these ideas.

Unions: The event that *either A or B or both* occur is called the **union** of A and B.

Intersections: The event that *both A and B* occur is called the **intersection** of A and B.

Disjoint events: Two events A and B are **disjoint** or **mutually exclusive** if $P(A \text{ and } B) = 0$.

Complement: The event that A does not happen is called the **complement** of A. **NOTE:** Since A and its complement are disjoint and their union is the entire sample space.

Conditional probability: The probability that A occurs if we already know that B has occurred or will occur is called the **conditional probability** of A **given** B. The idea is that we are "replacing" the sample space with the set of outcomes in B.

Example 2.4. Is there a home team advantage in different sports? The **contingency table** shows results from several games for four professional sports. (Copper, DeNeve, and Mosteller, Chance, Vol. 5)

	Basketball	Baseball	Hockey	Football	Total
Home Team Wins	127	53	50	57	287
Visiting Team Wins	71	47	43	42	203
Total	198	100	93	99	490

(a). Find the probability that a randomly selected game is a basketball game. Would this be a theoretical probability or an empirical probability? Explain.

(b). Find the probability that a randomly selected game is won by the home team.

(c). Find the probability that any randomly selected game is a basketball game **and** is won by the home team.

(d). Are the events "basketball game" and "won by the home team" disjoint events? Explain using probabilities.

(e). Find the probability that any randomly selected game is basketball **or** won by the home team.

(f). Find the probability that a randomly selected game is **not** a basketball game.

(g). Find the probability that, out of the basketball games only, the home team wins.

(h). Find the probability that given a home team win, the game is a baseball game.

(i). Which sport gives the biggest advantage to the home team? Explain.

Example 2.5. Consider the experiment of rolling a die. Let A be the event that you roll an even number. Let B be the event that you roll a number that is greater than or equal to 4. Let C be the event that you roll 1 or a 3.

$$A = \{2, 4, 6\} \qquad B = \{4, 5, 6\} \qquad C = \{1, 3\}$$

Compute the following probabilities:

✱$P(A \text{ or } B) =$ ✱$P(A \text{ and } B) =$ ✱$P(\text{not } A) =$

✱$P(A \text{ or } C) =$ ✱$P(B \text{ and } C) =$ ✱$P[(\text{not } A) \text{ and } C] =$

✱$P(A \text{ given } B) =$ ✱$P(B \text{ given } A) =$ ✱$P(A \text{ given } C) =$

Example 2.6. In 2012, 71% of students graduating from any four-year colleges had student loan debt. 73% of college students attended public universities. 48% of students both graduated with student loan debt and attended a public university. (ticas.org, *Chronicle of Higher Education*)

(a). One way to answer the following questions, though not the only way, is to construct a hypothetical contingency table. Suppose that 100 students were surveyed. Then, 71% of 100—or 71—students in the sample would have graduated with student loan debt. Construct a contingency table from the given information.

	Public University	Other University	Total
Debt			71
No Debt			
Total			100

(b). What is the probability that a student attended a public university or graduated with student loan debt?

(c). What is the probability that a student did not graduate with student loan debt?

(d). Are the events that a student graduated from a public university and that the student graduated with student loan debt disjoint? Explain your answer using probabilities.

Example 2.7. Suppose that Punctual Paige takes Highway 544 to Coastal every day, and is on time 99% of the time. If A is the event that Paige is on time for class, $P(A) = 0.99$. Let B be the event that there is a wreck on 544. We can estimate that this happens roughly 5% of the time; so, $P(B) = 0.05$.

What do you think? If there is a wreck on 544, will Paige still be punctual 99% of the time, or will her probability of being on time change depending on what is happening on 544?

Independent events: Two events A and B are **independent** if knowing that event B has occurred does not affect/change the probability that event A will occur or vice versa.

Dependent events: Two events A and B are **dependent** if knowing that event B has occurred affects/changes the probability that event A will occur or vice versa.

Example 2.8. Continuing from the prior example, recall that $P(A) = 0.99$ is the probability of Paige being on-time on any given day and $P(B) = 0.05$ is the probability of there being a wreck on 544. If there is a wreck on 544, we can expect Paige to have a higher chance of being late. Suppose that $P(A \text{ given } B) = 0.82$.

(a). Construct a contingency table for a hypothetical sample of days that Paige travels to CCU via HWY 544. **NOTE:** For the total number of days, use a power of 10 like 100, 1000, or 10,000.

	Wreck on 544	No wreck on 544	Total
Paige on time			
Paige not on time			
Total			

(b). Are the events "Paige being on-time" and "wreck on 544" independent? Explain using probabilities.

(c). What is the probability on a randomly selected day, Paige is on time and there is a wreck on 544?

(d). Suppose that Paige's first professor of the day notices she is late. What is the probability that there was a wreck on 544?

(e). If Paige is late, what is the probability that there is not a wreck on 544?

(f). Are the events "Paige being on-time" and "wreck on 544" disjoint? Explain using probabilities.

2.1.5 Additional Examples

Example 2.9. From our course investigation, 51% of students in STAT 201 chose a major in a program available through ACM. Around 47% of hometowns are at least 300 miles from campus. Furthermore, about 29% of students in STAT 201 chose a major in a program available through ACM and have a hometown that is at least 300 miles from campus.

If you find it helpful, you may summarize the results in the following contingency table for a hypothetical sample of students.

		Total
Total		

(a). What is the probability that a randomly selected student in STAT 201 chose a major in a program not available through ACM?

(b). What is the probability that a randomly selected student in STAT 201 chose a major in a program available through ACM and whose hometown is less than 300 miles from campus?

(c). For a student whose hometown is at least 300 miles from campus, what is the probability that they chose a major in a program available through ACM?

(d). Are the events "hometown is at least 300 miles from campus" and "chose a major in a program available through ACM" disjoint events? Explain.

(e). Are the events "hometown is at least 300 miles from campus" and "chose a major in a program available through ACM" independent events? Explain.

Example 2.10. A large group of people is to be checked for two common symptoms of new virus. It is thought that 30 percent of people possess symptom A while 10 percent of people possess symptom B. For those who possess symptom B, about 50 percent will possess symptom A.

If you find it helpful, you may summarize the results in the following contingency table for a hypothetical sample of people.

		Total
Total		

(a). What is the probability that a randomly selected person possesses both symptom A and symptom B?

(b). What is the probability that a randomly selected person has at least one symptom?

(c). What is the probability that a randomly selected person does not possess symptom B?

(d). What is the probability that a randomly selected person does not possess either symptom?

(e). Are the events "possessing symptom A" and "possessing symptom B" independent or dependent events? Explain using probabilities.

Example 2.11. For independent events A and B, suppose that $P(A) = 0.5$ and $P(B) = 0.4$. Using this information, compute the following probabilities.

If you find it helpful, you may summarize the results in the following contingency table for a hypothetical sample of items.

		Total
Total		

(a). Compute $P(A \text{ given } B)$.

(b). Compute $P(A \text{ and } B)$.

(c). Compute $P(A \text{ or } B)$.

(d). Compute $P(B \text{ given } A)$.

(e). Are events A and B disjoint? Explain using probabilities.

Example 2.12. For disjoint events A and B, suppose that $P(A) = 0.3$ and $P(B) = 0.2$. Using this information, compute the following probabilities.

If you find it helpful, you may summarize the results in the following contingency table for a hypothetical sample of items.

	B	**Not B**	**Total**
A			
Not A			
Total			

(a). Compute $P(A \text{ and } B)$.

(b). Compute $P(A \text{ given } B)$.

(c). Compute $P(A \text{ or } B)$.

(d). Are events A and B independent? Explain using probabilities.

2.2 Discrete Distributions

2.2.1 Overview of Discrete Distributions

Random Variable: A function whose input is the outcome of a random experiment and whose output is a number. There are two types of numerical outputs possible.

Types of Quantitative Variables:

A **discrete probability distribution** gives

1. the possible outcomes of an experiment and
2. the probability of observing each outcome.

*To be a valid distribution, the sum of all the probabilities must be _____

Example 2.13. Give the probability distribution that describes rolling a fair die once.

The **expected value (mean/average)** of a discrete probability distribution can be found as the weighted average of outcomes:

Example 2.14. Find the expected value when rolling a fair die once.

Example 2.15. Hurricanes are a fact of life when living on the coast. However, do they look different depending on where you live? Consider the following distributions of hurricane strength for direct hits to the US mainland at two locations. (Probabilities are from all hurricanes making landfall between 1851 and 2004.)

Table 1: Distribution of Direct Hits in the Gulf Coast

Saffir-Simpson	1	2	3	4	5
Probability	???	0.275	0.242	0.064	0.013

Table 2: Distribution of Direct Hits in the Atlantic

Saffir-Simpson	1	2	3	4	5
Probability	0.513	0.255	0.197	0.031	0.004

(a). What is the probability a hurricane making landfall on the Gulf Coast will be category 1?

(b). What is the probability that a hurricane making landfall on the Gulf Coast is at least a category 3? Atlantic?

(c). What is the expected strength of hurricanes making landfall on the Gulf Coast?

(d). What is the mean strength of hurricanes making landfall on the Atlantic coast?

2.2.2 Binomial Distribution

The **binomial distribution** is a special kind of **discrete** distribution. It must meet all four of the following requirements (BINS):

1. **B**inary Outcomes (success and failure)
2. **I**ndependent Trials (each trial is not influenced by the others)
3. **N**umber of trials is fixed ahead of time (n)
4. **S**ame probability of success each trial (p)

Shorthand notation for Binomial random variables:

Example 2.16. Determine if the following are Binomial random variables.

(a). X = the number of times "snake eyes" are rolled in a game of dice with 10 rolls

(b). X = the number of shots it takes for a player to make 10 free throws

(c). X = the number of children in a family of five children that get the flu during a given year

We can calculate the mean and standard deviation for Binomial distribution by:

Example 2.17. In the NBA, the top free-throw shooters usually have a probability of about 0.9 of making any given free throw. Suppose a player shoots 10 free throws and let X = the number of free throws made.

(a). Find n and p for the binomial distribution. State the distribution of X using proper notation.

(b). Find the mean and standard deviation for the binomial distribution.

Example 2.18. An Ipsos poll done in 2019 looked at tattoos as a cultural phenomenon for different generations. They found that 26% of Generation Z have a tattoo. Define the binomial random variable X = the number of young adults with a tattoo in a random sample of 20.

(a). State and verify the four conditions for this to be a Binomial random variable. State the random variable using proper notation.

(b). Find the expected number in the sample with tattoos. How much does this vary by?

To compute the probability that X takes on a single value k, P(X = k), we can use our calculators: `2nd -> DISTR -> binompdf`. That is, $P(X = k) =$ `binompdf(n,p,k)`.

(c). Find the probability that exactly half of the sample will have a tattoo.

(d). Find the probability that five people will have a tattoo.

To compute the probability that X is less than or equal to k, P(X ≤ k), we can use our calculators: `2nd -> DISTR -> binomcdf`. That is, $P(X \leq k) =$ `binomcdf(n,p,k)`

(e). Find the probability that 10 or fewer will have a tattoo.

(f). Find the probability that fewer than 10 will have a tattoo.

(g). Find the probability that more than 5 will have a tattoo.

✱ In case you are interested the percentage of people with tattoos in earlier generations: 38% of Millennials, 32% of Gen X's, 15% of Baby Boomers, and 6% of the Silent Generation have tattoos according to the poll.

Example 2.19. A large study on gaming by *Earnest* found that the least likely profession to be into gaming is that of dentistry. In fact, only 5.8% of dentists spend any money on gaming. This is compared to the 60% of Americans who play video games daily! Suppose we take a random sample of 50 dentists and let X = the number of dentists in a random sample of 50 that spend money on gaming.

(a). How many are expected to spend money on video games?

(b). What is the probability that at most 4 spend money on video games?

(c). What is the probability that more than 4 spend money on video games?

(d). What is the probability that 4 spend money on video games?

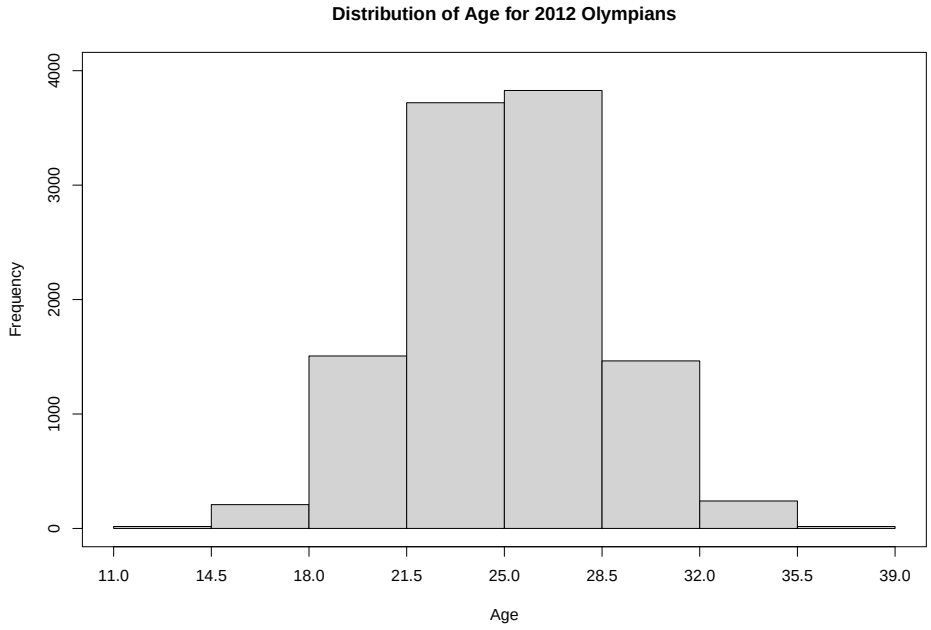
(e). What is the probability that at least 12 spend money on video games?

2.3 Normal Distribution

2.3.1 Introduction

In the last section, we examined discrete random variables and a specific case in the binomial random variable. It is important to recall that discrete random variables take distinct values only. In addition, we defined continuous random variables which can take on values anywhere within an interval. For example, when a student arrives at a trolley/bus stop on campus, the time (in minutes) they will have to wait until the next trolley/bus arrives could be any value between 0 and 15 minutes.

Example 2.20. The following histogram displays the distribution for the age of nearly 11,000 athletes that participated in the 2012 London Olympics. The average age for the athletes was 25 years old with standard deviation of 3.5 years.



(a). What is the shape of this distribution?

(b). Approximately what proportion of athletes are younger than 18?

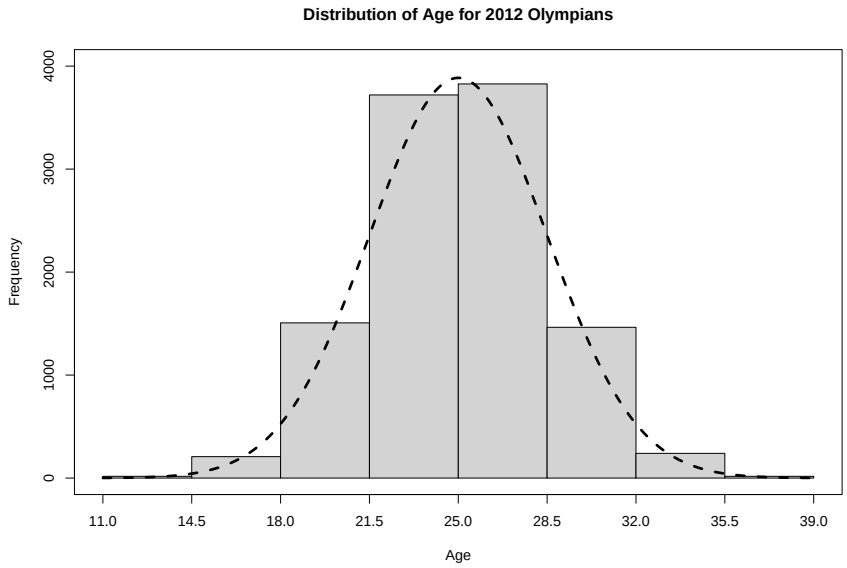
(c). Approximately what percent of athletes are between 21.5 and 28.5 years old?

(d). Using the above graph, can we find the approximate percent of athletes are exactly 25 years old? Explain.

(e). During the 2012 London Olympics, Gabby Douglas won a gold medal for the all around in women's gymnastics. Gabby was around 16.5 years old at the time. Using the above graph, could we find the percent of athletes that were younger than Gabby? Explain

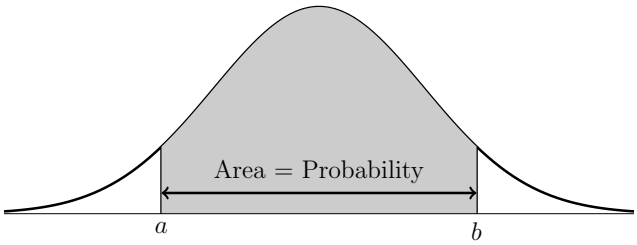
In the previous example, would could only approximate percentages (probabilities) for ranges of values referring to specific ages outlined by the histogram. As in considered in part (e), it is possible to find probabilities for any range of ages.

For continuous random variables, the probability distribution can only be represented by a mathematical function since it is impossible to list all the possible values. This mathematical function for a continuous random variable is often an curve and describes the distribution of the random variable. The following picture displays a continuous distribution that models the age of athletes at the 2012 London Olympics.

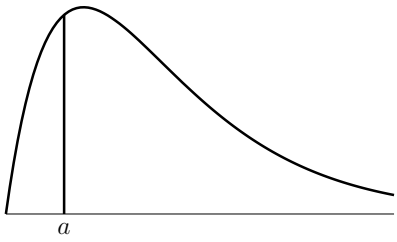


For a continuous distribution:

✱Probabilities are the same as area under the curve.



✱$P(X = a) = 0$ for any value a.



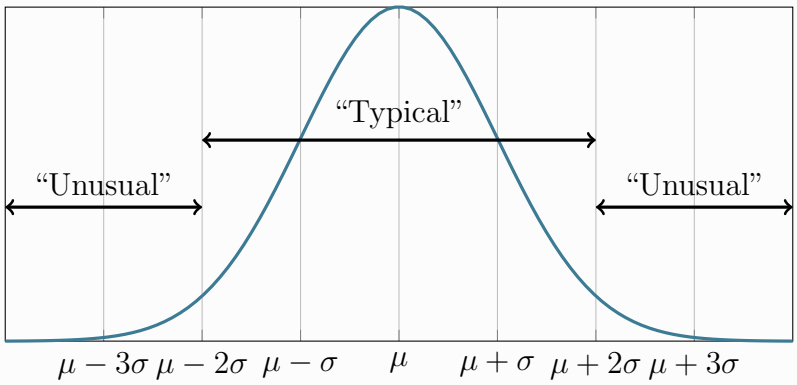
✱The total area under the curve is 1.

Continuous distributions can be used to model any continuous random variable such as time, salary, and length. We will see later that many discrete distributions may also be approximated by a continuous curve.

The symmetric bell-curve continuous distribution we saw in the ages of Olympians example is common enough and important for theoretical reasons that we give its own name.

The Normal Distribution: As viewed in real world applications, many continuous variables have a *bell shaped distribution*. When this is the case, a Normal distribution is commonly used to model this type of shape. The Normal distribution can be characterized by its two parameters:

- $\mu =$ mean
- $\sigma =$ standard deviation



Notation:

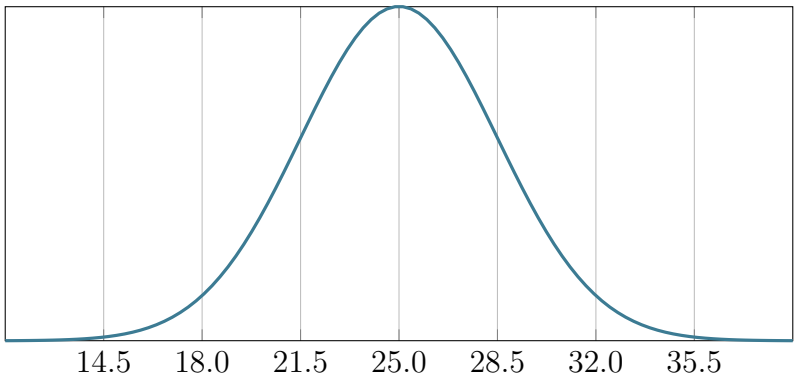
2.3.2 Probability and Percentile Computations with Z-Scores

There are an infinite number of possible Normal distributions depending on the mean and standard deviation of the random variable we are describing. However, we can relate every Normal distribution to the **Standard Normal Distribution** using the **Z-score** formula.

Z-scores measure the distance of an observation from the mean, in terms of standard deviations. Positive z-scores indicate the observation is above average and negative z-scores indicate an observation is below average.

The **Standard Normal Distribution** is a normal distribution with mean zero and standard deviation 1, denoted by $Z \sim N(0, 1)$. This standardized distribution of z-scores is useful for comparing measurements on a common, unit-less scale.

Example 2.21. For the example considering the ages of 2012 Olympic athletes, the ages of athletics is said to follow a Normal distribution with mean 25 and standard deviation 3.5. In notation, that is $X \sim N(25, 3.5)$. Consider the following picture of this distribution,



(a). Add the values of the z-scores for the values in the above plot.

(b). What ages are considered "typical" for athletes during the 2012 London Olympics?

(c). What ages are considered "unusual" for athletes during the 2012 London Olympics?

Example 2.22. In the 2012 London Olympics, the average age of male gymnasts was about 23 with standard deviation of about 3 years. The average age of males participating in equestrian events was about 40 with a standard deviation of 6.5 years. Hiroshi Hoketsu was a 71 year old equestrian from Japan and Iordan Iovtchec was a 39 year old gymnast from Bulgaria. Who is oldest relative to their sport and are either of the athletes potential outliers with respect to age?

The Standard Normal Distribution on the Calculator:

I. To find the area(probability) between two z-scores use `normalcdf(lwr,uppr)`

(a). Find the area below z = 1.52.

(b). Find the area above z = 1.52.

(c). Area below z = -1.

(d). Area between z = -2 and z = 2.

II. To find the z-score with a particular area below (percentile) use `invNorm(left area)`

(a). Find the 10% percentile.

(b). Find the z-score in the top 15%

2.3.3 Probability and Percentile Computations with General Settings

Example 2.23. Returning to our example where the ages of 2012 Olympic athletes are $X \sim N(25, 3.5)$, we can now find probabilities for any observation of interest.

(a). Find the probability that a randomly selected athlete will be younger than 30.

(b). Find the probability that a randomly selected athlete will be between 20 and 25 years old.

(c). 97.5% of athletes are younger than what age?

(d). 30% of athletes are older than what age?

Example 2.24. In March 2022, YouTube users spent an average of 18 minutes on the site at a time. Suppose times follow a Normal distribution with standard deviation 5 minutes.

(a). What percent of YouTube users spend more than 15 minutes at a time on the site?

(b). What time is the 60^{th} percentile?

(c). What proportion of users spend between 15 and 30 minutes on the YouTube?

(d). Would 10 minutes be an unusual amount of time? Explain.

3 Statistical Inference for Proportions

3.1 Sampling Distribution of the Proportion

3.1.1 Introductory Activity

Consider standard playing cards and suppose we are interested in p = the proportion of all standard playing cards that are hearts (♥).

- We know the parameter $p = 0.25$ because of our prior knowledge of playing cards.
- Consider a random sample of 5 cards. Let the random variable Y = the number of hearts in that sample.

1. Verify that Y is a binomial random variable and give its distribution in the correct shorthand notation.

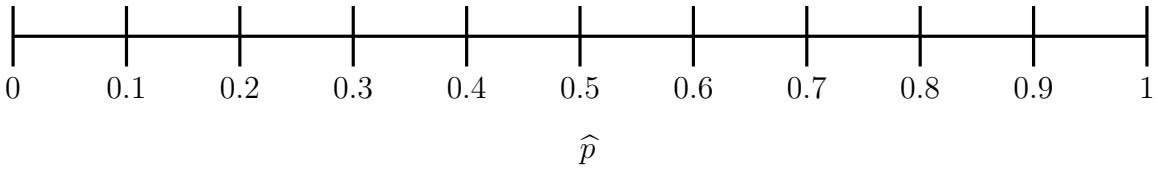
2. How many hearts do you expect to have in a random sample of 5 cards? By how much do you expect the values to vary from this?

3. Your instructor will pass out a random sample of five playing cards to each student. How many hearts cards are in your hand? What is the probability of observing that number of hearts cards?

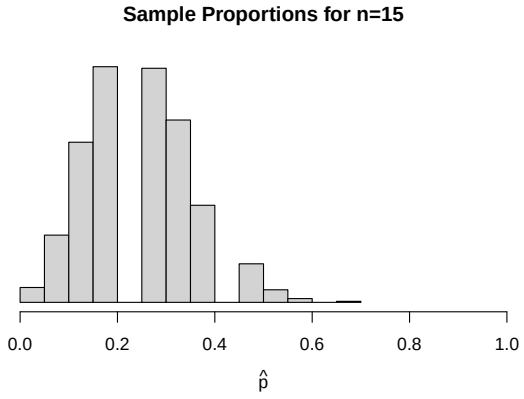
4. What is the probability of observing at most that many cards?

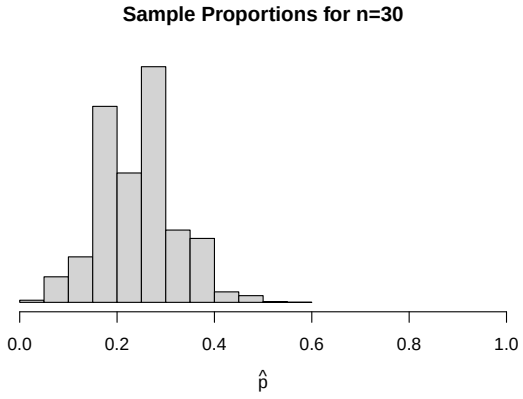
5. The sample proportion $\hat{p}$ is the fraction of the sample that are hearts; that is, $\hat{p} = y/5$. Calculate *your* value of $\hat{p}$.

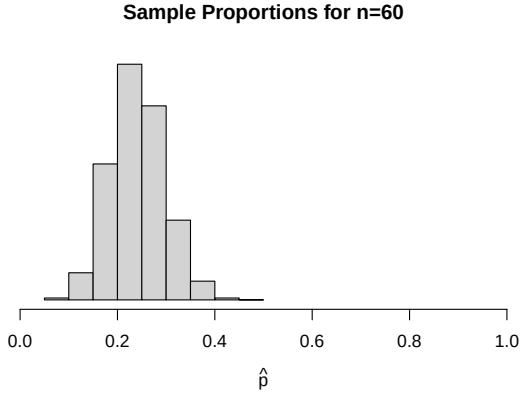
6. Add your value of $\widehat{p}$ to the class plot on the board. Record your answer along with your classmate's on the axis below.

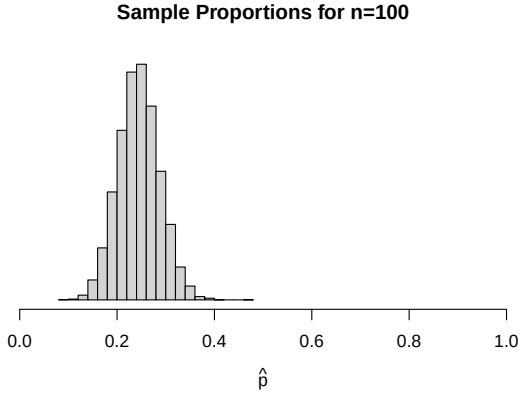


The remaining four plots were created by taking many samples of $n = 15$, $n = 30$, $n = 60$, and $n = 100$ cards and computing $\widehat{p}$. The different values of $\widehat{p}$ were then displayed as a histogram.









7. What do you observe about the central tendency of the previous plots?

8. What do you observe about the variability of the previous plots?

9. What do you observe about the shape of the previous plots?

3.1.2 Formal Result and Examples

The Sampling Distribution of the Sample Proportion

Notation check:

- $p =$
- $\widehat{p} =$

We have seen that different random samples give us different statistics ($\widehat{p}$'s in our case). The distribution of the statistic computed from all possible samples of a fixed size is the **sampling distribution**. We have observed that for proportions,

10. Does the result apply to our sample of size 5?

11. Does the result apply if we have a random sample of 100? If so, state the sampling distribution for the proportion of hearts in a sample of 100 playing cards.

12. Use the previous result to find the probability that 40 or fewer are hearts in a sample of 100. Does this appear reasonable from the generated sampling distribution?

13. Recall that y = number of hearts in our sample is also a binomial random variable. Use this approach to compute that at most 40 are hearts in our sample of 100. Compare and discuss.

Example 3.1. According to the National Safety Council, 28% of all traffic crashes (1.6 million per year) are due to drivers using cell phones.

(a). Find the sampling distribution for the proportion of accidents caused by cell phone usage in random sample of 200 accidents.

(b). What is the probability that a random sample of 200 contains at least 35% of accidents that were caused by cell phone usage?

(c). What is the probability that a random sample of 200 contains between 30% and 40% of the accidents that were caused by cell phone use?

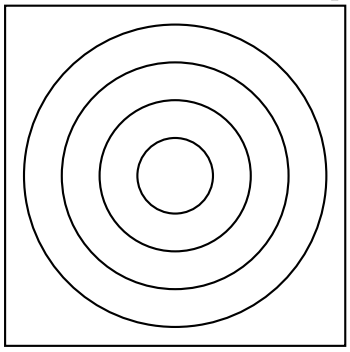
(d). What is the sampling distribution for the sample proportion of accidents caused by cell phones for a random sample of 300 accidents? Describe in symbols and words.

Statistics are called **unbiased** if the center of the possible values is indeed the parameter of interest. For example, we have seen that the sample proportions ($\widehat{p}$'s) are centered around the population proportion (p) and are therefore unbiased.

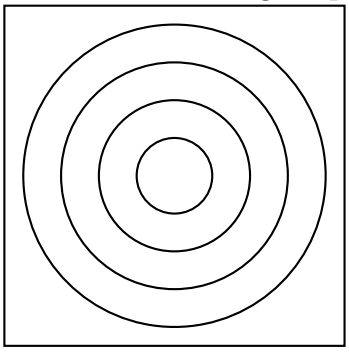
How much a statistic varies around the parameter can be described by **precision** or **accuracy**. The less a statistic varies around the parameter the more precise, or accurate, the estimator. We have seen that the precision of the $\widehat{p}$'s increases as the sample size increases.

Example 3.2. Consider the analogy of playing darts. The bulls-eye serves are the parameter (e.g. p) and the dart throws serve as possible values of the statistic. Sketch an example of each of the following settings:

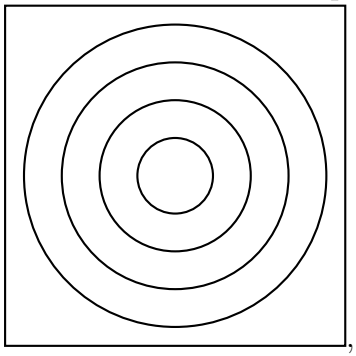
1. Biased with low precision/accuracy



3. Biased with high precision/accuracy

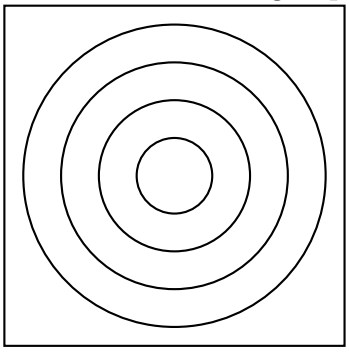


2. Unbiased with low precision/accuracy



,

4. Unbiased with high precision/accuracy



Which setting is ideal?

3.2 Confidence Intervals for One Proportion

3.2.1 Introduction

Our discussion of statistics will be moving to the area of **statistical inference**. Statistical inference uses sample statistics to estimate population parameters.

Q: Suppose we want to estimate the population proportion. What value from our sample could we use? Why is this number alone insufficient (even though media often reports only this single number)?

Rather than using a single point to estimate a population value, we will use a range of numbers called a **confidence interval**. The basic framework for any confidence interval is given by:

Confidence intervals for a single proportion

3.2.2 Examples

Example 3.3. A OnePoll online survey was conducted in March 2022 with 1,000 Americans. The survey examined respondents views on April Fools' day pranks. Despite the potential for hurt feelings, April Fools' Day still remains a popular holiday with 640 of the respondents saying they enjoy it. Filling a room full of helium balloons and putting googly eyes on unexpected household objects ranked as two of the most acceptable pranks.

(a). Suppose we wish to estimate the proportion of all American adults who enjoy April Fools day. Based on the sample of 1,000 respondents, what is the point estimate for this value?

(b). What is the estimated standard error of the point estimate?

(c). Find the multiplier for a 95% confidence interval and calculate the margin of error.

(d). Construct a 95% confidence interval for the population proportion.

(e). Interpret the interval.

(f). One of your friends claims that 50% of people enjoy April Fools day. Based on the interval, is there any validity to this claim? Explain.

Example 3.4. The U.S. is one of six countries that doesn't offer paid family leave for parents on a national level. In a survey of 2,000 parents of children 0 to 18 conducted on January 14th 2022, researchers found that 77% feel "outraged that the US has no federal paid family leave laws for new moms and dads." Construct and interpret a 99% confidence interval for the proportion of all American adults who are outraged by lack of federal family leave laws.

Example 3.5. Consider our course investigation. Out of a random sample of 100 students, 45% were classified as in-state students.

(a). Construct and interpret a 90% confidence interval for the proportion of all students at CCU who are in-state students.

(b). According to the fall 2021 demographic report, 49% are actually in-state. Does our interval contain this value? If not, discuss reasons why not.

3.2.3 Wrap it up

Summary of most common z*'s:

Assumptions for the confidence interval to be valid:

1.

2.

→ Are the intervals from our three examples all valid?

Q: What affects the width of the confidence interval?

Interpretation of the confidence level

3.3 Hypothesis Test for One Proportion

3.3.1 Introductory Activity

In this section we introduce the idea of hypothesis testing with an activity. A hypothesis test makes a claim about a population, then gathers evidence (data) to determine the plausibility of that claim. Suppose our sample is a bag of stones that your teacher has been gifted. The giver claimed the stones in the bag came from a very large pile (the population) where half of the stones were of a particular color.

1. State the default assumption or **null hypothesis** about the parameter p, the proportion of stones in the bag that are this particular color, in both words and symbols.

2. Suppose we have an **alternative hypothesis**. State it in words and symbols.

3. Your class will consider the stones from the bag as a random sample of stones. Write down your sample proportion $\hat{p}$.

4. Do you still believe the null hypothesis?

5. Suppose that the null hypothesis is true. Does the sampling distribution of the sample proportion apply to the random sample we took? If so, state the distribution of $\hat{p}$.

6. Based on your response above, calculate the z-score for the class's $\hat{p}$ assuming that the null hypothesis is true. This particular z-score is called the **test statistic**.

7. Do you still believe the null hypothesis?

8. Since we know the $\hat{p}$'s follow a normal distribution and we know the z-score of our particular $\hat{p}$, we can calculate the probability, called the **p-value** of getting our $\hat{p}$ or one "more extreme." Do that.

9. Interpret the p-value. Remember, this probability was calculated under the assumption of the null hypothesis being true.

10. Do you still believe the null hypothesis?

11. State the conclusion to the hypothesis test in terms of the alternative hypothesis.

Congratulations! You have just completed your first hypothesis test. We will formalize and summarize these ideas in the next pages and practice additional examples.

3.3.2 Formalizing the Hypothesis Test for One Proportion

1. **State the hypotheses:**

   H_0 : is the **null hypothesis** and is what we initially assume is true.

   H_A : is the **alternative hypothesis** and is where we state our research question.

   Things to remember:

   - Always use a population parameter in the hypotheses (ex. μ, p) because we are testing a claim about the entire population. Never use statistics ($\bar{x}$, $\widehat{p}$).
   - The same number should be used in both hypotheses. This number comes from our research question, not the data. We ask our question before going out to collect the evidence(data).
   - H_0 has a statement of equality ($=$). H_A has a statement of inequality ($<, >, \neq$).

2. **Compute the test statistic:** This computes how far our evidence(data) is from the initial assumption (H_0).

   In general: $TS = \frac{PE - H_0}{SE}$

   For proportions: $z_c = \frac{\widehat{p} - p_0}{SE}; SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

3. **Compute the p-value:** The p-value quantifies the evidence from the test statistic in the form of a probability. It tells us the probability of obtaining our test statistic (data) or something more extreme (depends on H_A), if the null hypothesis is assumed to be true.

4. **Interpret the p-value:** The p-value is a probability that tells us the likelihood of observing our data if the null hypothesis is true. Example interpretation: "If H_0 *is true*, we would see data like ours, or more extreme, *p-value*$\times$*100%* of the time."

5. **State your conclusion:** The less likely our evidence/data under H_0 (smaller p-value), the stronger the evidence for our research claim (H_A). Note that the conclusion is written in terms of the amount of evidence for our research question, H_A.

p-value > 0.1	$0.05 <$ p-value < 0.10	$0.01 <$ p-value < 0.05	p-value < 0.01
(more than 10%)	(between 5% and 10%)	(between 1% and 5%)	(less than 1%)
"Little to no evidence in favor of H_a"	"Some evidence in favor of H_a"	"Strong evidence in favor of H_a"	"Very strong evidence in favor of H_a"

3.3.3 Examples

Example 3.6. Consider our course investigation. According to fall 2021 enrollment data, 49% of all undergraduate students at CCU are in-state students (South Carolina residents). From our random sample of 100 students taking STAT 201 in fall 2022, 45% were in-state students. Conduct a test to determine if the current proportion of in-state students at CCU has changed.

Further Questions:

1. When we state a conclusion in hypothesis testing, are we 100% certain in those conclusions? If we did not find a significant difference, does that mean one does not exist? If we did find a significant difference, does that mean it is for sure different? Discuss.

2. Based only on the results of the hypothesis test, do you think a 95% confidence interval for p in this setting would contain 0.49? Explain your reasoning.

Example 3.7. A 2009 Sports Illustrated study said that 78 percent of NFL retirees have "gone bankrupt or are under financial stress because of joblessness or divorce" within two years of their careers ending. (espn.go.com/nfl/) Suppose a more recent study wants to determine if that proportion has decreased. They found that 140 out of 200 randomly sampled players were faced with financial ruin. Perform a hypothesis test to determine if the proportion has decreased.

3.3.4 Wrap it up

Common names of the alternative hypothesis

Alternative	Common Name
$H_a : p < p_0$	left-sided/left-tailed test
$H_a : p > p_0$	right-sided/right-tailed test
$H_a : p \neq p_0$	two-sided/two-tailed test

General conceptual ideas

1. If the null is true in reality, we would expect the test statistic to be _____.

2. The greater the magnitude (absolute value) of the test statistic, the _____ the p-value of a two-sided test.

3. As sample size increases and all else stays the same, the value of the test statistic _____.

The duality between hypothesis tests and confidence intervals

- By examining the results of a two-sided hypothesis test, we can get a general idea on whether or not the hypothesized value would be in a corresponding confidence interval.

Strong evidence in favor of $H_a : p \neq p_0$	$\Longrightarrow$	p_0 is most likely not the value of p	$\Longrightarrow$	p_0 will most likely not be in the CI
Little to no evidence in favor of $H_a : p \neq p_0$	$\Longrightarrow$	p_0 could be the value of p	$\Longrightarrow$	p_0 will most likely be in the CI

- By examining the confidence interval, we can get a general idea on the type of evidence in favor of the H_a for a two-sided test.

p_0 is not in the CI	$\Longrightarrow$	p_0 is most likely not the value of p	$\Longrightarrow$	Strong evidence in favor of $H_a : p \neq p_0$
p_0 is in the CI	$\Longrightarrow$	p_0 could be the value of p	$\Longrightarrow$	Little to no evidence in favor of $H_a : p \neq p_0$

3.3.5 Additional Examples

Example 3.8. Suppose we test the hypothesis of $H_0 : p = 0.2$ vs $H_A : p \neq 0.2$ (concerning the proportion of orange M&Ms). Based on the data collected, we obtain a p-value of 0.201. Based on this information, would a 95% confidence interval for the population proportion of orange M&Ms contain 0.2? Explain without calculating the interval?

Example 3.9. Suppose you are interested in estimating the proportion of first down plays in the *National Football League* (NFL) that are run plays. For a random sample of first down plays, you obtain the following 95% confidence interval for the true proportion of first down plays in the NFL that are run plays: $(0.543, 0.677)$. Suppose an analyst claims that two-thirds of first down plays in the NFL are run plays. Based only on the confidence interval, is there enough evidence to refute this claim?

Example 3.10. Consider our course investigation. Suppose an individual claimed that less than half of students chose an ACM major. From our random sample of 100 students taking STAT 201 in fall 2022, 51% chose an ACM major. Conduct a test to determine if the sample evidence supports this individual's claim.

Example 3.11. Are ads on streaming services like YouTube becoming too repetitive? Out of a random sample of 1,500 American adults, 1,035 said they think ads on streaming services are repetitive. Is there enough evidence to conclude that more than two-thirds (66.7%) of Americans think that ads on streaming services are repetitive?

3.4 Inference for Two Proportions

3.4.1 Introduction

Confidence Intervals for Two Proportions

Hypothesis Tests for Two Proportions

3.4.2 Examples

Example 3.12. In a national poll conducted Oct. 6-8, 2021, researchers found that 117 of GenZer's in a sample of 212 subscribed to Amazon Prime Video. Of the 673 Millennials sampled, 424 subscribed to Amazon Prime Video.

(a). Estimate the difference in the proportion of Amazon Prime Video subscribers between the two age groups with a 95% confidence interval.

(b). Based on the confidence interval, is there evidence that a difference exists in the proportion of subscribers in each age group? Discuss.

Example 3.13. In the same national poll conducted Oct. 6-8, 2021, researchers found that out of the 793 participants who consider themselves "avid fans of film", 330 prefer watching foreign films with dubbing. Out of the 1,251 who consider themselves "casual fans of film", 437 prefer watching foreign films with dubbing.

(a). Conduct the appropriate test to determine if avid fans of film are more likely to prefer dubbing than casual fans.

(b). Based on the results of the hypothesis test, would a 95% confidence interval for $p_1 - p_2$ contain 0? Discuss.

Example 3.14. From our course investigation, it is of interest to estimate the difference in the proportion of ACM and non-ACM students who are out-of-state. The following contingency table compares the residency and type of major (ACM or Non-ACM) for each student surveyed.

	ACM Major		
	Yes	No	**Total**
In-State	18	27	45
International	0	1	1
Out-of-State	33	21	54
Total	51	49	100

Use these results to construct and interpret a 90% confidence interval for the difference in the proportion of ACM and non-ACM students who are out-of-state.

Example 3.15. From our course investigation, there were a total of 51 students whose major was in a program available through ACM and 49 students whose major was not available through ACM. Of those majors in the ACM, 56.9% of students' hometowns were at least 300 miles from campus while only 36.7% of students' hometowns were at least 300 miles from campus for majors not in the ACM. Is there enough evidence to conclude that a difference exists in the proportion of students' hometowns that are at least 300 miles from campus between majors in the ACM and majors not in the ACM? Conduct the appropriate test.

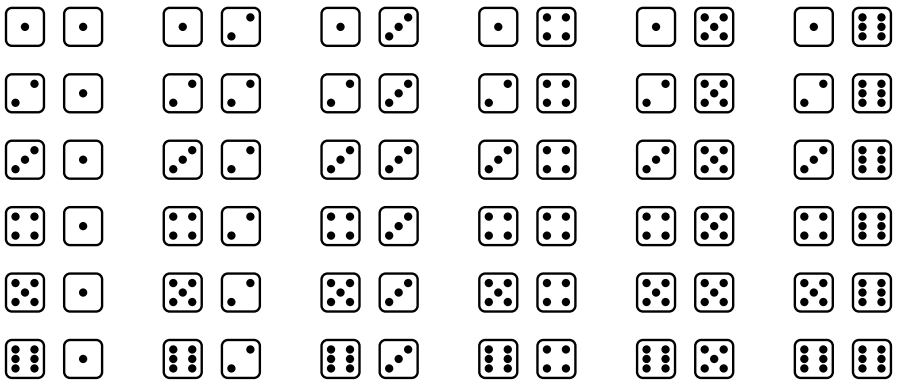
4 Statistical Inference for Means

4.1 Sampling Distribution of the Mean

4.1.1 Introductory Activity

We have seen with sample proportions that the value we get for the statistic will vary based on our sample. The same is true for any statistic computed from data. In this section we explore the distribution of the sample mean using an activity before formalizing and applying the results.

Suppose we roll a pair of six-sided dice (or roll one six-sided die twice). The sample space for this experiment is given below.

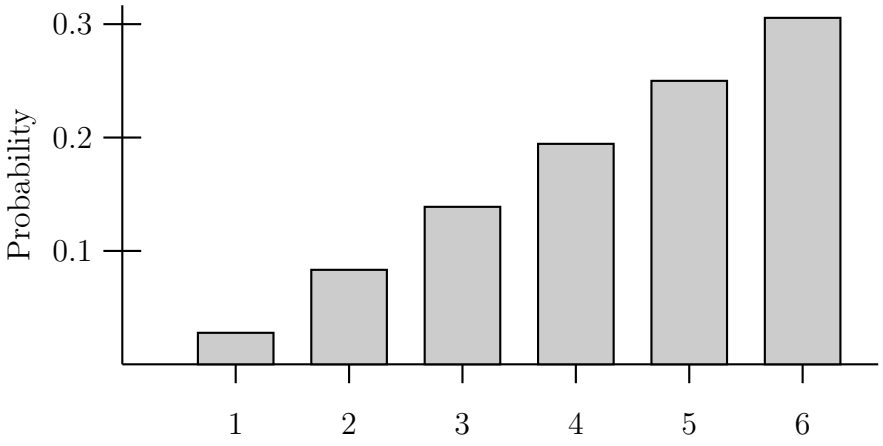


Let X be the the maximum value rolled.

1. Is X a discrete or continuous random variable? What are the possible values for X?

2. What is the probability distribution for the random variable X?

The following is a chart for the probability distribution of X.



3. What is the mean of X?

4. Let $\overline{X}$ be the mean of $n = 10$ rolls of a pair of dice (or rolling a single die twice). In other words, $\overline{X}$ is the average of a random sample from the population X. In the problems that follow, we will seek to understand the distribution of the random variable $\overline{X}$.

   Is $\overline{X}$ a discrete or continuous random variable?

5. Roll a pair of six-sided dice (or one six-sided die twice) ten times. Record the results here, then calculate the mean $\overline{x}$ of your rolls using the `1-Var Stats` command.

$\overline{x} =$ _____

6. Record the class's values for $\overline{x}$ here.

7. Reproduce a histogram or dot plot for the class's empirical distribution of $\overline{X}$ here.

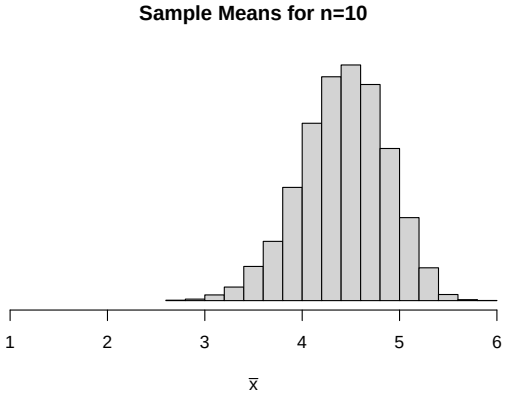
8. Using `1-Var Stats`, find the sample mean of the $\overline{X}$ from your class. What did you expect the mean to be close to? Is it close?

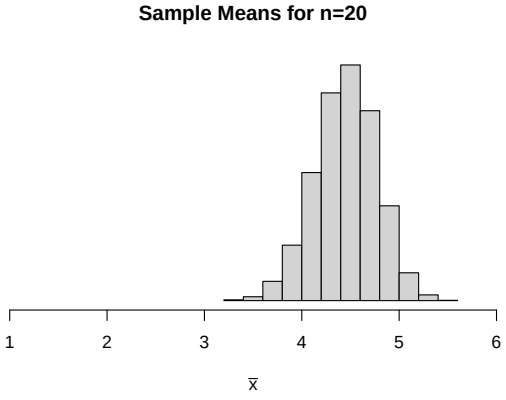
9. Using `1-Var Stats`, find the sample standard deviation of the $\overline{X}$ from your class. What did you expect the standard deviation to be close to? Is it close?

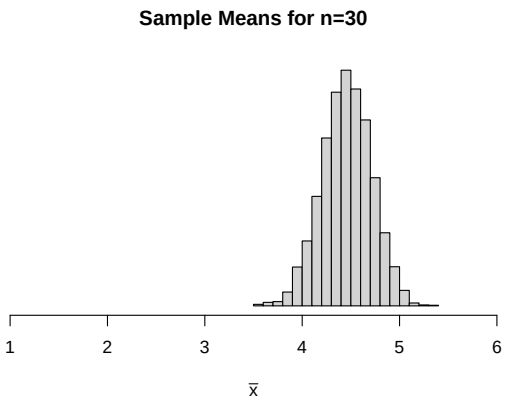
10. Suppose you did not know $\mu = 4.47222$, and you tried to use one of the $\overline{X}$ to estimate μ. How wrong would you expect to be? In other words, what is the "standard" amount of "error"?

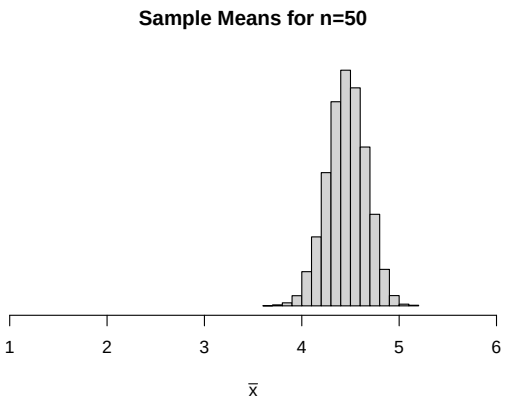
11. What is the shape of the distribution of $\overline{X}$ when $n = 10$?

12. The following four plots were created by taking many samples of $n = 10$, $n = 20$, $n = 30$, and $n = 50$ rolls and computing $\overline{X}$. The different values of $\overline{X}$ were then displayed as a histogram. What happens to the shape of the distribution of $\overline{X}$ as n increases?









4.1.2 Formal Results and Examples

The Sampling Distribution of $\overline{x}$ - Central Limit Theorem:

1. Center: The mean of the sampling distribution of $\overline{x}$ is equal to the mean of the population. In other words, all possible sample means are centered at the true mean. Specifically, $\mu_{\overline{x}} = \mu$.
2. Spread: The standard deviation of all possible sample means decreases as the sample size increases. Specifically, $SE_{\overline{x}} = \sigma/\sqrt{n}$.
3. Shape:
   - If the population of X is Normal, then the sampling distribution of $\overline{x}$ is Normal regardless of the sample size, n.
   - If the population of X is not Normal, then the sampling distribution of $\overline{x}$ is approximately Normal for large n (at least 30).

The Central Limit Theorem says for large enough n, $\overline{x} \sim N(\mu, \sigma/\sqrt{n})$.

Example 4.1. According to the AAA Foundation for Traffic Safety and the Urban Institute, motorists age 16 years and older drive, on average, 29.2 miles per day (10,658 miles per year). The distribution of distances is unknown. Let's assume the standard deviation of distances is 10 miles per day.

(a). Find the probability that a randomly selected driver travels less than 25 miles per day.

(b). What is the sampling distribution of the mean distance for a sample of 40 drivers?

(c). Find the probability that a random sample of 40 drivers will travel less than 25 miles per day, on average.

Example 4.2. Yearly chocolate consumption by American adults is Normally distributed. Americans consume 12 pounds of chocolate on average per year with a standard deviation of 2.7 pounds. Suppose we take a random sample of 10 American adults.

(a). What is the sampling distribution of the mean for samples of size 10?

(b). What is the probability that a randomly selected adult consumes more than 15 pounds of chocolate per year?

(c). What is the probability that a random sample of 10 adults will consume more than 15 pounds per year, on average?

(d). Suppose we take several different samples of 50 American adults and find the average chocolate consumption for each sample. What will be the mean of the sample averages?

(e). Suppose we take several different samples of 50 American adults and find the average chocolate consumption for each sample. What will be the standard deviation of the sample averages?

4.2 Inference for One Mean

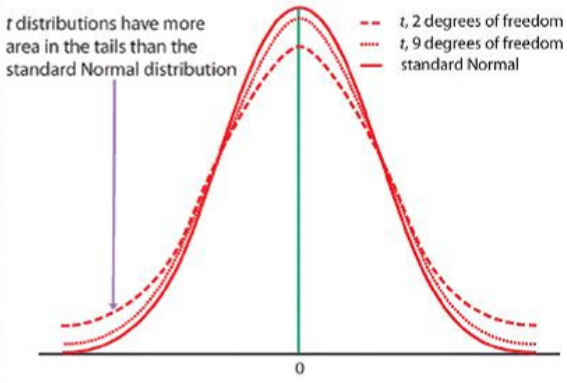
4.2.1 Introduction and the T-Distribution

Confidence Intervals for One Mean

Influences on the width of the CI:

Hypothesis Testing for One Mean

Student-t Distribution We use the t-distribution to obtain multipliers corresponding to the desired level of confidence. The t-distribution was invented by William Gosset under the pseudonym, Student, while working at the Guinness brewery in Dublin. Later in life he moved back home to London and took a job as Head Brewer at a new Guinness brewery. He is well known for developing statistical methods to deal with small sample sizes. (Pictures courtesy of Wikipedia)



Practice using the t-table and calculator

When df is between two rows on the table, always go with the:

Practice finding the multiplier for confidence intervals using `invT(%,df)` or t-table

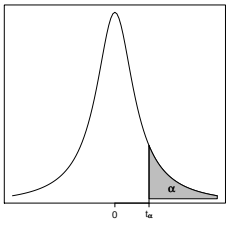
- 95% Confidence with n = 15
- 99% Confidence with n = 50
- 90% Confidence with n = 35

Practice finding the p-value using `tcdf(min,max,df)`

- $H_A : \mu < 100$, $t_c = -2.4$, $n = 20$
- $H_A : \mu > 100$, $t_c = -1.2$, $n = 23$
- $H_A : \mu \neq 100$, $t_c = 2.9$, $n = 30$

Student's t Table

The entries in this table are the critical values for Student's t-distribution for which the area under the curve in the right-hand tail is α. Critical values for the left-hand tail are found by symmetry.



	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$	$t_{0.001}$
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.091

4.2.2 Examples

Example 4.3. It is difficult to imagine the size of the blue whale, the largest animal inhabiting the earth. There are records of individuals over 100 feet (30.5 m) long, but 80 feet is probably average. A good way to visualize their length is to remember that they are about as long as three school buses. An average weight for an adult is 200,000 to 300,000 pounds (100-150 tons). Its heart alone is as large as a small car. (www.marinemammalcenter.org) Suppose a researcher wonders if environmental factors such as climate change, pollution of the oceans, and whalers have have caused any changes in the length of the blue whale. A random sample of 15 whales from the coast of California yielded a mean length of 75 feet and a standard deviation of 13 feet. Answer the researcher's question using a 95% confidence interval. Note the duality.

Example 4.4. From our course investigation of a random sample of 100 students, the number of credits enrolled during the fall 2022 semester was observed. The sample mean number of credits was 15.38 with a sample standard deviation of 1.757. Compute a 90% confidence interval for the true mean number of credits enrolled for CCU students. Interpret your interval and note any limitations.

Example 4.5. From our course investigation of a random sample of 100 students, the cumulative GPA of courses taken at CCU was available for only 90 of those students. The sample average cumulative GPA is 3.1998 with a standard deviation of 0.5584. Conduct a test to determine if the true mean GPA of students at CCU exceeds 3.0.

Example 4.6. According to a Limelight survey of online gamers worldwide, the average time spent playing video games was 8.5 hours per week in 2021. Suppose a 2022 study of 45 gamers found that the average time spent was 8.9 hours per week with standard deviation 1.2 hours.

(a). Conduct a test to determine if there is a difference in time average spent gaming in 2022?

(b). Based on the results of the hypothesis test, would a 95% confidence interval for μ contain the value 8.5? Discuss.

4.3 Inference for Two Means

4.3.1 Introduction

Confidence Intervals for Two Means

Hypothesis Tests for Two Means

4.3.2 Examples

Example 4.7. Sixgill sharks sampled in the Puget Sound were measured and sexed. A summary of the results is given below. Is there significant evidence that the adult males are smaller on average than the females? Answer the researcher's question using a 95% confidence interval for the difference in mean size of sixgill sharks. Note the duality.

Group	n	Mean	SD
Females	51	3.7	0.40
Males	26	3.2	0.35

Example for two mean inference continued...

Example 4.8. In "Influence of alcohol and marijuana use on academic performance in college students" (Meda et. al., 2017) researchers studied the effects of alcohol and marijuana use on college GPA. A group of 463 participants were classified as medium to high alcohol use with little to no marijuana use. Their mean GPA was 3.03 with standard deviation of 0.64. A second group of 188 students was identified to have high alcohol and high marijuana use. The mean GPA of this group was 2.66 with standard deviation of 0.83.

(a). Conduct a test to determine if there a difference in the GPA of college students based on substance use.

(b). Based on the results of the hypothesis test, would a 95% confidence interval for the difference in means contain 0? Discuss.

Example 4.9. Using data from our course investigation, the following table displays summary statistics for the distance between a student's hometown and the CCU campus separated by major being either available or not available through the ACM.

ACM Major	n	Mean	SD
Yes	51	340.902	257.8415
No	48	257.3104	255.1598

NOTE: The overall sample size is 99 since the distance was unavailable for one of the students sampled.

Is there enough evidence to conclude that the distance between a student's hometown and the CCU campus is typically greater for students whose major is available through the ACM than for students whose major is not available through the ACM? Test this claim.

Example 4.10. Using data from our course investigation, the following table gives summary statistics for the cumulative GPA of students separated by residency.

Residency	n	Mean	SD
In-State	42	3.2314	0.5446
Out-of-State	47	3.1575	0.5716

Compute and interpret a 95% confidence interval for the difference in mean cumulative GPA between in-state and out-of-state students.

4.4 ANOVA

4.4.1 Introduction

Purpose: To examine the differences between two or more means. We will expand the procedures for two means to any number of means.

ANOVA stands for

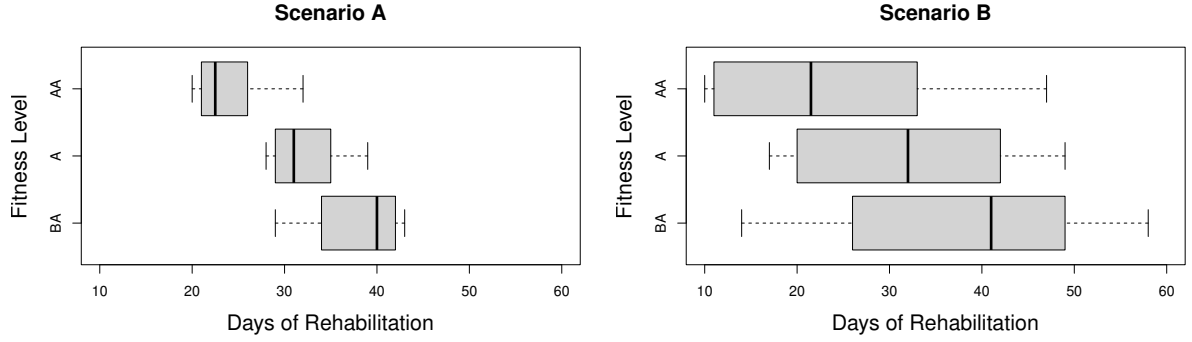
Question: Why does a test about means involve variances? Consider the following.

A rehabilitation center researcher was interested in examining the relationship between physical fitness prior to surgery of persons undergoing corrective knee surgery and time required in physical therapy until successful rehabilitation. Twenty-four male subjects ranging in age from 18 to 30 years who had undergone similar corrective knee surgery during the past year were selected for the study. Patients were randomly drawn from each type of prior fitness level (below average, average, and above average). The number of days required for successful completion of the physical therapy was recorded for each patient.

- Response (outcome variable) -
- Factor Levels -
- Factor (explanatory variable) -
- Experimental units (subjects) -

The average recovery times for the three groups (BA,A, AA) were 38, 32, and 24 days, respectively. Does this suggest that the average recovery time is different in the three groups?

Now, consider the following two scenarios. Both have the same group means (38,32, and 24 days). Which one provides more evidence for a difference in population means and why?



Therefore, we must consider two types of variability:

- The variability *between* each of the groups

- The variability *within* each treatment group

We summarize the results in an **ANOVA Table**:

Source	DF	SS	MS	F	P-value
Factor	No.Groups $-$ 1	SSFactor	MSFactor $=$ $\frac{\text{SSFactor}}{\text{No.Groups}-1}$	$F_c = \frac{\text{MSFactor}}{\text{MSError}}$	Provided or use calculator
Error	$n -$ No.Groups	SSError	MSError $=$ $\frac{\text{SSError}}{n-\text{No.Groups}}$	-	-
Total	$n - 1$	SSTotal	-	-	-

4.4.2 Examples

Obtain the ANOVA table using the calculator for our example. The data is provided below.

	1	2	3	4	5	6	7	8	9	10
Below Average	29	42	38	40	43	40	30	42		
Average	30	35	39	28	31	31	29	35	29	33
Above Average	26	32	21	20	23	22				

ANOVA on the TI 83/84 Calculator:

1. Enter the data into three columns or lists (Stat $\rightarrow$ Edit $\rightarrow \ldots$)
2. Run ANOVA (Stat $\rightarrow$ Tests $\rightarrow$ ANOVA(L_1, L_2, L_3)

Record the results in the ANOVA table and perform a hypothesis test to determine if there are any differences in the mean recovery time based on prior fitness level.

Source	DF	SS	MS	F	P-value
Factor					
Error				-	-
Total			-	-	-

Example 4.11. Let's explore the computation of sums of squares using our knee surgery data.

	1	2	3	4	5	6	7	8	9	10
Below Average	29	42	38	40	43	40	30	42		
Average	30	35	39	28	31	31	29	35	29	33
Above Average	26	32	21	20	23	22				

Summary Statistics

Overall Mean: 32 Mean of BA: 38 Mean of A: 32 Mean of AA: 24

Deviations from overall mean:

Group	1	2	3	4	5	6	7	8	9	10
Below Average										
Average										
Above Average										

SSTotal = dfTotal =

Deviations between group mean and overall mean:

Group	1	2	3	4	5	6	7	8	9	10
Below Average										
Average										
Above Average										

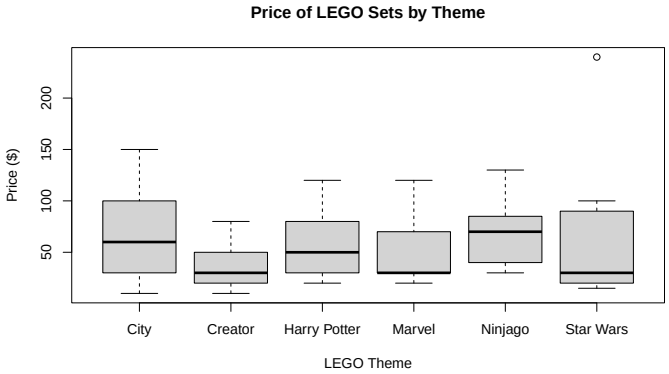
SSFactor = dfFactor = MSFactor =

Deviations from group means:

Group	1	2	3	4	5	6	7	8	9	10
Below Average										
Average										
Above Average										

SSError = dfError = MSError =

Example 4.12. An honors student at CCU was interested in comparing the price of LEGO sets across a variety of themes. He chose to focus on the following themes: *City*, *Creator*, *Harry Potter*, *Marvel*, *Ninjago*, and *Star Wars*. From each theme, he randomly sampled 9 sets and observed the price in dollars for each set. He was interested to see if the theme created a difference in the typical price of the set. The following side-by-side boxplot displays the distribution of the data across the difference themes.



(a). Based on the boxplot, does their appear to be more variability *within* or *between* the groups? Explain your answer and what it means in context of the student's research.

(b). Complete the following ANOVA table.

Source	DF	SS	MS	F	P-value
Factor		6773			0.618
Error				-	-
Total		98179	-	-	-

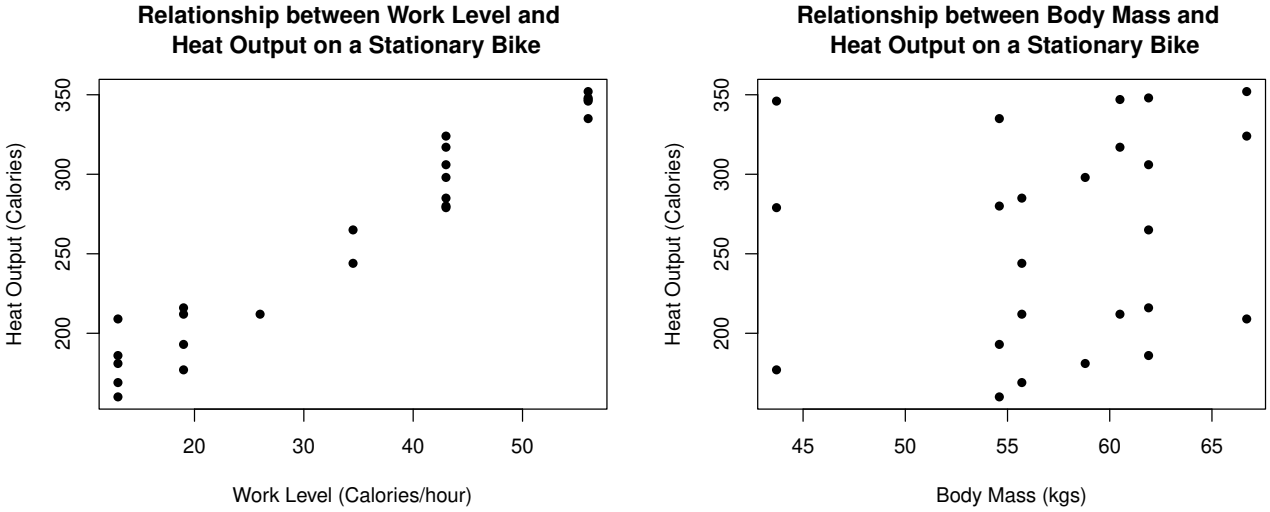
(c). State the hypotheses, p-value interpretation and conclusion of the test.

5 Associations Between Quantitative Variables

5.1 Scatterplots

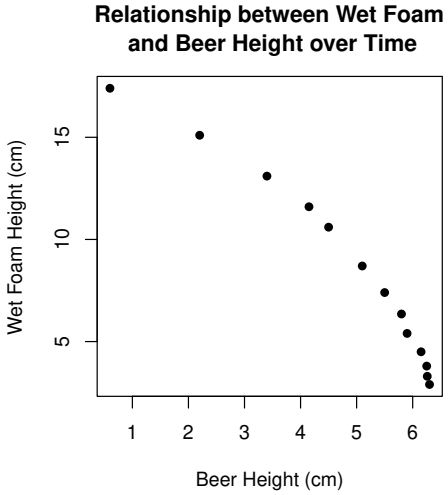
As discussed earlier in Section 1.5.3, **Scatterplots** are an easy way to study the relationship between two quantitative variables measured on the same subject or at the same time point. The data is plotted as (x, y) coordinates for each subject/time point. If it is thought that one variable exerts influence on the other, it is plotted on the x-axis and called the **explanatory variable**. The variable on the y-axis is called the **response variable** and may be impacted by or respond to the explanatory variable.

Example 5.1. The paper by M. Greenwood (1918) "On the Efficiency of Muscular Work," examined the relationship between body mass (kg) and work level (calories/hour) to the amount of heat production (calories) when riding a stationary bike.



Observations:

Example 5.2. In Hackbarth (2006). "Multivariate Analyses of Beer Foam Stand," researchers recorded measurements of wet foam height and beer height at various time points for Shiner Bock.



Observations:

Key Features of Scatterplots

We can assess the information presented in a scatterplot by looking for the following:

1. **Form**

3. **Strength**

2. **Association/Direction**

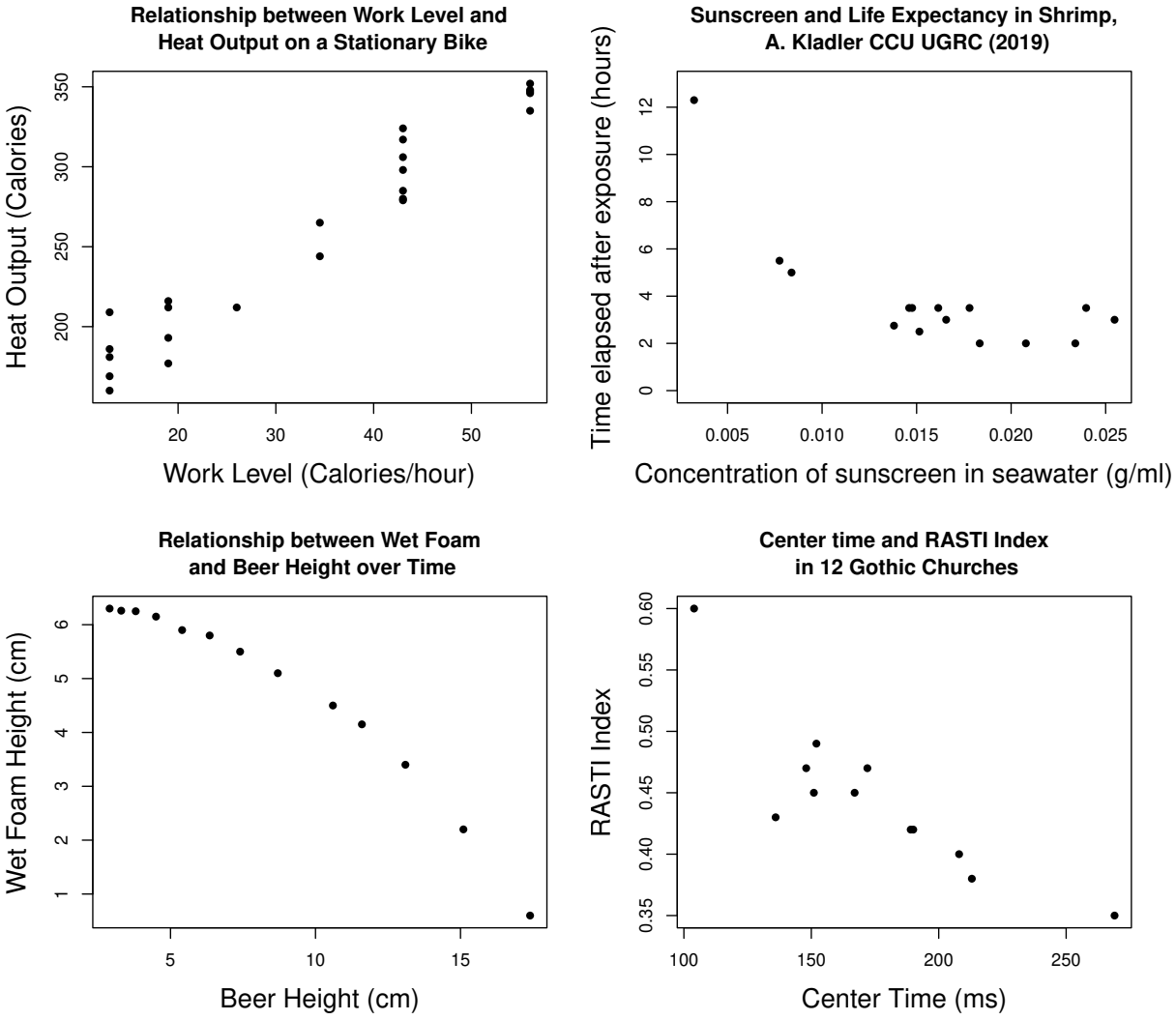
4. **Outliers**

Reevaluate our observations of the previous scatterplots to make sure we addressed all of the key features in each plot.

5.2 Correlation

We have seen that scatterplots can be used to visualize the relationship between two quantitative variables. Scatterplots gave us an idea of the form, direction, and strength of the relationship along with any potential outliers. Here we look at formalizing our understanding of strength with a numerical value.

For linear relationships between two numerical variables, we can more formally measure the strength using **correlation (r)**.

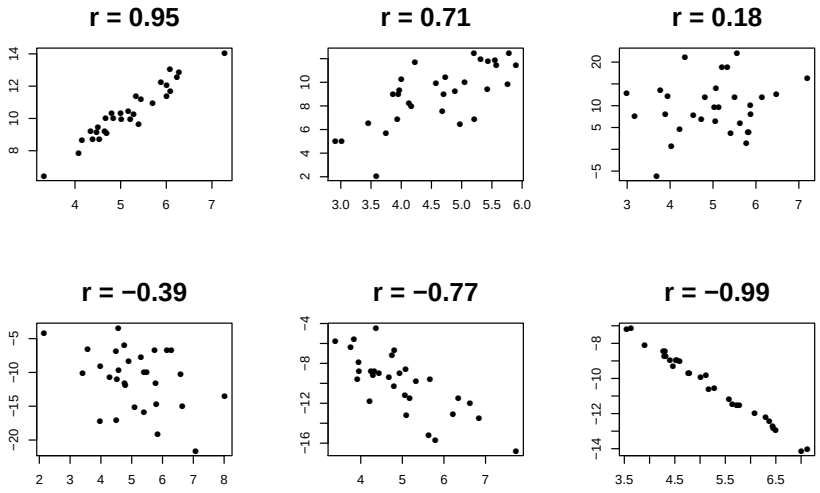


According to this, in which of the above scenarios could we compute correlation (r)?

Some Properties of Correlation:

1. Correlation is always between -1 and 1.

2. The sign of the correlation indicates the association/direction.



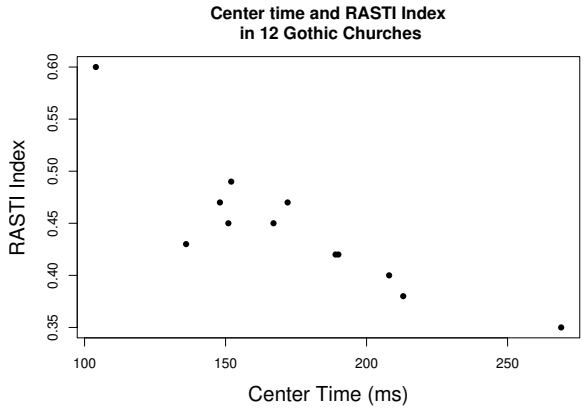
3. Correlation has no units! Its value does not depend on the units of the two variables.

4. Correlation is the same regardless of how you assign the x and y variables.

5. Correlation does not imply causation!!!

Finding correlation on the calculator

Example 5.3. Consider the scatterplot from the previous page (and repeated below) for the acoustical properties of 12 Gothic Churches.



Guess the correlation for the Gothic Church data based on the scatterplot. Then use your calculator and the data provided to find the actual correlation. (Continued on the next page)

Guessed: $r =$ _____ Actual: $r =$_____

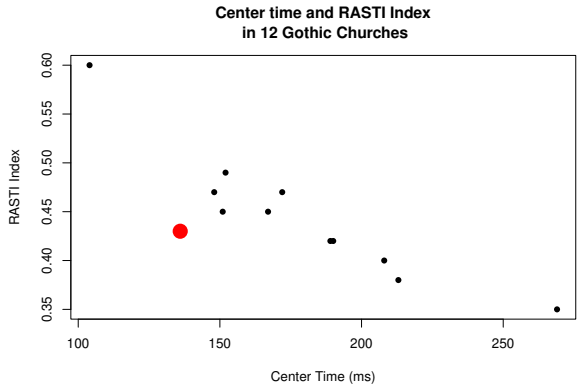
Table 3: Gothic Church Acoustic Data

Church	1	2	3	4	5	6	7	8	9	10	11	12
Center Time	213	208	172	190	152	167	148	151	136	269	104	189
RASTI	0.38	0.40	0.47	0.42	0.49	0.45	0.47	0.45	0.43	0.35	0.60	0.42

Calculator Steps for Correlation (TI 83/84):

0. By default the calculator will not display correlation. Once the default settings are changed, you do not have to repeat this step each time. Go to the catalog ($2^{nd} \rightarrow 0$). Scroll down to DiagnOn. Press enter twice.
1. Enter the X variable into L_1 and the Y variable into L_2. To do this, go to STAT→Edit.
2. To obtain correlation, go to STAT→CALC→LinReg(a+bx).

Question: What do you think would happen to correlation if we removed observation 9, $(136, 0.43)$ (enlarged below)? Discuss, then compute to confirm.



Based on the results, would you consider correlation a resistant calculation?

5.3 Simple Linear Regression

We have seen that **scatterplots** can be used to visualize the relationship between two quantitative variables. When it is clear, the **explanatory** variable is represented on the x-axis and the **response** variable is represented on the y-axis. Scatterplots gave us an idea of the **form, direction, and strength of the relationship along with any potential outliers**. We have also seen that the strength and association of linear relationship can be formally measured by **correlation**. How do we move on to study such relationships in more depth? To exam this, recall the study on heat output and work on stationary bicycles.

Key Features and Correlation:

Question: We have seen that the form of the relationship is linear. Draw your best estimate of a line to describe the trend in the plot above. Your line is probably similar to your classmate's line, but not exactly the same. How can we determine which line is the best? Discuss.

5.3.1 Computing and Interpreting the LSR Line

Residuals are the error between the predicted values ($\widehat{y}$) according to our line and the observed values of the response (y). That is,

$$\text{Residual} = \text{Observed } y - \text{Predicted } y, \text{ i.e. } e = y - \widehat{y}$$

The **Least Squares Regression (LSR) Line** is the line of best fit that produces the "least squared" error as defined by the residuals. Using calculus, one can find that the slope and intercept of such a line are computed with the following formulas:

Slope: $b = r\frac{s_Y}{s_X}$

Y-Intercept: $a = \overline{y} - b\overline{x}$

Using the slope and intercept as computed above, we can obtain the **final equation of the LSR line** as shown below. Note that x and y are often written out in words according to the context of the problem.

$$\widehat{y} = a + bx$$

Computing the LSR Line on the Calculator (TI 83/84)

1. Enter the X variable into L_1 and the Y variable into L_2. To do this, go to STAT→Edit.
2. To obtain the slope and intercept for the LSR line, go to STAT→CALC→LinReg(a+bx).

Example 5.4. Using the data for the study on energy output on stationary bikes, compute the regression line and write it in context.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Work Level	19	43	56	13	19	43	56	13	26	34.5	43	13
Heat Output	177	279	346	160	193	280	335	169	212	244	285	181
Subject	13	14	15	16	17	18	19	20	21	22	23	24
Work Level	43	19	43	56	13	19	34.5	43	56	13	43	56
Heat Output	298	212	317	347	186	216	265	306	348	209	324	352

Example 5.5. Using the data and estimated line of best fit, answer the following questions.

(a). What heat output is predicted for a work level of 19 calories? 20 calories? Use the table below to find the following predicted values.

x	$\widehat{y}$
19	
20	
42	
43	

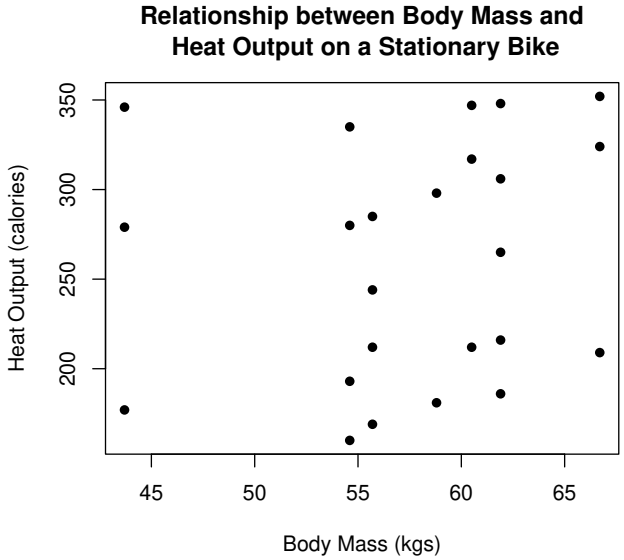
(b). Subject 1 in the study had an observed work level of 19 calories/hour and 177 calories of heat output. The observed data for subject 13 is $(43, 298)$. Compute the residuals for these two observations.

(c). Using your work in (a), how does the heat output change when work level changes from 19 calories/hour to 20 calories/hour? How does the heat output change when work level changes from 42 calories/hour to 43 calories/hour? What do you notice?

Interpretation of the slope: As x increases by one unit, the predicted/average y increases/decreases by slope units.

Example 5.6. Interpret the slope of our bicycle example in context.

Question: In the same study, researchers also recorded body mass (kg) of the 24 subjects. The scatterplot of relationship is given below. How would you summarize the relationship? Give a rough estimate of the slope for the least squares regression line without performing any computations.



Question: Clearly we can always plug values into our formulas to obtain a least squares regression line. However, that does not always indicate that the regression line is useful or meaningful. How can we determine the usefulness of an estimated LSR line?

5.3.2 Determining the Usefulness of a Regression Line

In this class we will explore two approaches for testing the usefulness of a regression line:

1. Computing the **coefficient of determination**, r^2.

   Recall that correlation, r, tells us the strength and direction of a linear relationship. Also recall that,

   $$\leq r \leq$$

   This implies that,

   $$\leq r^2 \leq$$

   It is interesting to note that we may also compute the same value for r^2 using the ANOVA setting.

   $$r^2 = \frac{\text{SSRegression}}{\text{SSTotal}} = \frac{\text{Variability in Y explained by the line}}{\text{Total Variability in Y}}$$

   Therefore, r^2 is giving us a proportion. Specifically, r^2 tells us **the proportion of variation in y that is explained by the line with x.**

2. Conducting **inference about the slope**. Is is something other than zero?

   Hypotheses:

   Test Statistic:

   P-value:

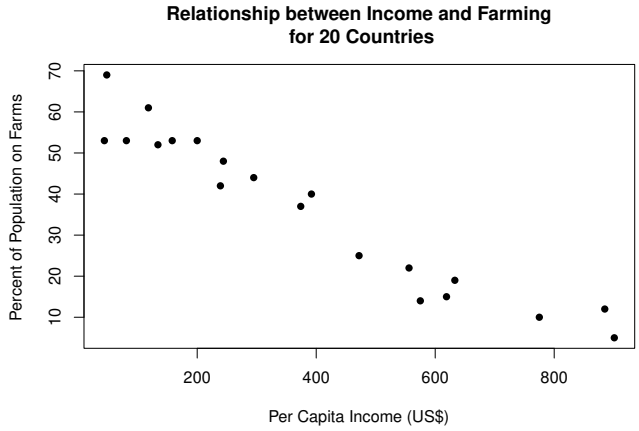
Example 5.7. Using your calculator, compute the coefficient of determination for the regression model predicting heat output based on work level. Interpret this value. Does this indicate that work level is a useful predictor of heat output? Why or why not?

Example 5.8. The correlation between heat output and body mass was found to be $r = 0.1434$. Compute and interpret the coefficient of determination. Does this indicate that body mass is a useful predictor of heat output? Why or why not?

Example 5.9. Previously we have found the estimated slope between heat output and work level. In addition to this, we can compute $SE_b = 0.1965$. Test if there is a significant linear relationship between heat output and work level.

Example 5.10. It can be found that the estimated slope between heat output and body mass is $b = 1.434$ and $SE_b = 2.110$. Test if there is a significant linear relationship between heat output and body mass.

Example 5.11. Consider the following data on income and the percent of the population living on farms for 20 different countries in the year 1953. Note that $r = -0.963$.



Country	1	2	3	4	5	6	7	8	9	10
Farm Percent	40	61	53	53	53	37	12	53	14	69
Income	392	118	44	158	81	374	885	200	575	48
Country	11	12	13	14	15	16	17	18	19	20
Farm Percent	42	48	25	52	19	44	10	15	5	22
Income	239	244	472	134	633	295	775	619	901	556

(a). Compute the LSR line and report in context.

(b). Interpret the slope in context.

(c). Greece (country 14) had \$134 per capita income and 52% of the population living on farms that year. What is the predicted proportion living on farms based on Greece's per capita income? What is the residual for Greece?

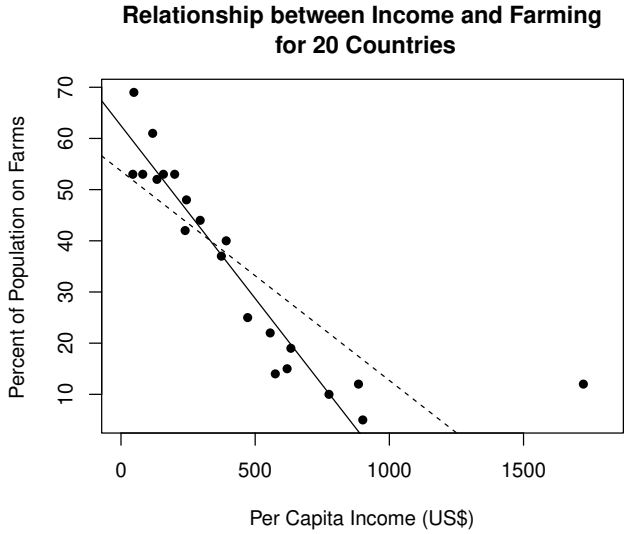
(d). What proportion of variation in the percent of individuals living on farms in explained by a linear regression on per capita income? In view of this value, is the regression equation useful for predictions? Explain your answer.

(e). We can compute that $SE_b = 0.0044$. Test if there is a significant linear relationship between the per capita income of a country and the percent of the population living on farms.

5.3.3 Influential Observations and Extrapolation

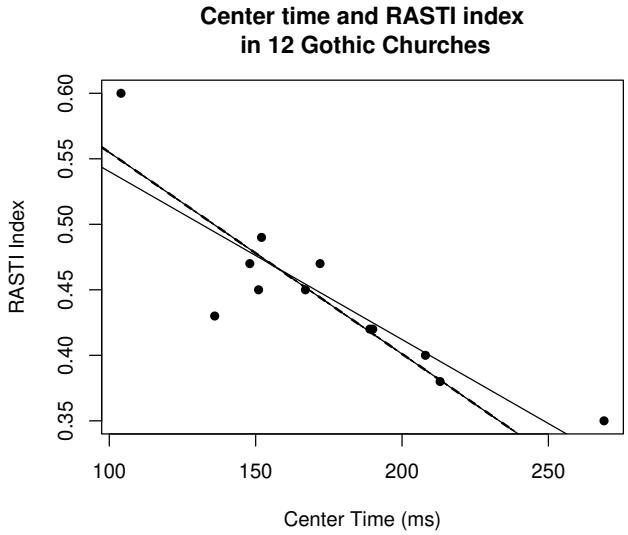
Influential points in regression are those who have a very large or small x-value compared to the majority of the data. In addition, influential points do not follow the overall pattern of the data. It is important to note and carefully examine any influential points in the data because they will influence the computations.

Example 5.12. The data presented on income and the population on farms was only a subset of the original data. The United States was not previously included. We can see that the United States is an influential observation in this data set.



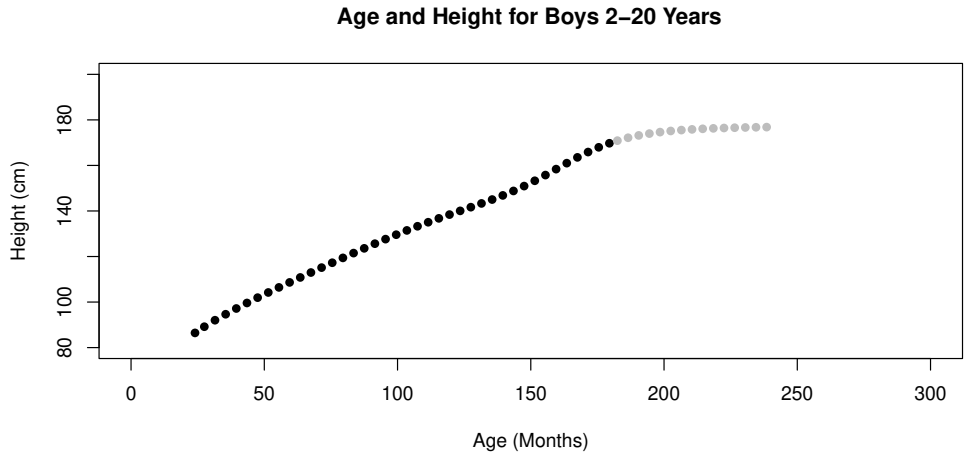
Data	Slope	Intercept	r	r^2
Original	-0.068	62.494	-0.963	0.927
With US	-0.041	53.635	-0.828	0.686

Example 5.13. Recall our example of acoustics in Gothic churches. An influential observation existed there as well.

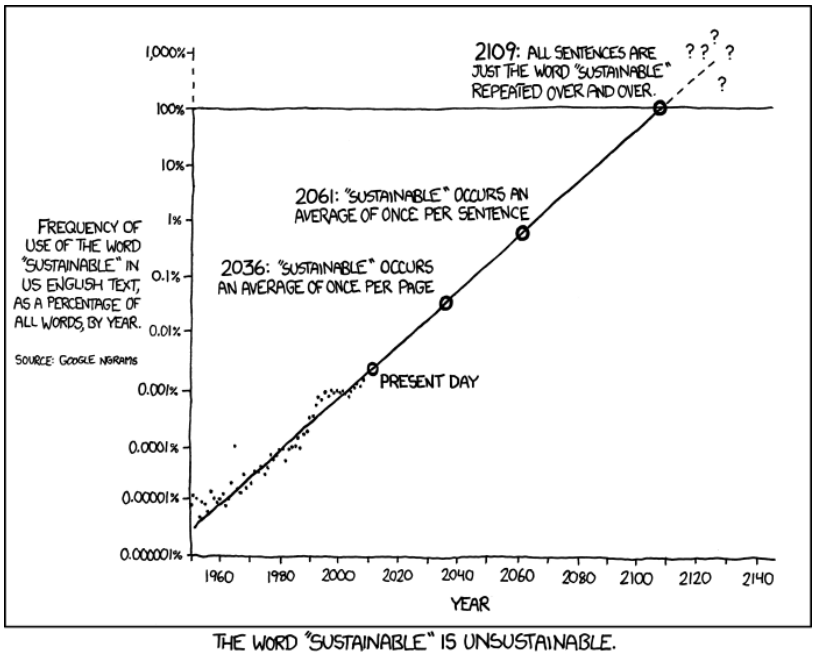


Extrapolation is the practice of using your line of best fit to make predictions for values of x that were never modeled. It is bad practice to use the line to make predictions for values of x that are smaller or larger than the observed data because the trend may change beyond the observed data.

Example 5.14. Age and height provide a classic example of the dangers of extrapolation. We tend to grow at rapid rates when we are younger according to a linear trend. However, once we have reached a certain age, growth starts to slow down. As seen in the height and age data from the CDC, we can model height using one trend up to about 180 months of age (15 years). If we used this same trend to predict the height of someone older than 15, say 20 (240 months) or 50 (600 months), we would predict an unusually large height because the trend changes!



Example 5.15. To end on a comic note (https://xkcd.com/1007/). The word "sustainable" is unsustainable."



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

6 Summary Pages

This section is provided as extra space for daily reviews, extra examples, or any other need.

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page

Summary Page