Edited: 1:21pm, October 3, 2024

DRAFT —— DRAFT —— DRAFT —— DRAFT

The Story of Statistics: Inquiry based notes for introductory statistics

Lindsey Bell, Nicholas Pritchard, Douglas Weathers

Coastal Carolina University

AMS Open Math Notes: Works in Progress; Reference # OMN:202412.111448; Last Revised: 2024-12-15 09:20:44

Contents

Acknowledgments **viii**

1 Using Data to Answer Questions **1**
1.1 The Investigation 1
1.2 Data Collection 4
1.3 Data Visualization for a Single Variable 14
1.4 Data Summarization with Numbers 28
1.5 Data Visualization for Comparing Two Variables . . . 37
1.6 Study Designs and Conclusions 49

2 Uncertainty in Data **59**
2.1 Probability 59
2.2 Discrete Distributions 82
2.3 Normal Distribution 91

3 Statistical Inference for Proportions **104**
3.1 Sampling Distribution of the Proportion 104
3.2 Confidence Intervals for One Proportion 111
3.3 Hypothesis Test for One Proportion 119
3.4 Inference for Two Proportions 132

4 Statistical Inference for Means **137**
4.1 Sampling Distribution of the Mean 137
4.2 Inference for One Mean 145
4.3 Inference for Two Means 153
4.4 ANOVA 158

5 Associations Between Quantitative Variables 165
5.1 Scatterplots 165
5.2 Correlation 168
5.3 Simple Linear Regression 172

To the Instructor

The introductory statistics course at the author's medium-sized, four-year institution satisfies the core requirement for quantitative literacy and is required by many majors, predominately in the sciences. Approximately 14 sections of the course are taught each semester in a highly coordinated fashion. For many years, all faculty have had some version of their own course notes. However, after the COVID-19 pandemic we were inspired to think critically about how we teach the course. In spring 2022, all instructors of the course met weekly during our lunch hour. This was an energizing and productive time as we discussed new ideas for approaching the course broadly and also in the minute details. The inquiry-based course notes were created as a result of these discussions. The discussions changed the way we teach from a set of topics to a story about the process of statistics. The notes have been used in the classroom for all sections of the course from summer 2022 - fall 2023. During this time we have continued to discuss the material as a group and make revisions as indicated by experience in the classroom. The current version is the result of writing from three instructors, but several semesters worth of input and testing from nine regular instructors of the course.

The topic of probability was a large part of our discussions. Traditionally the students would memorize and apply the standard formulas including those for unions, intersections, and conditional probability. This portion of the material was often the most challenging for the students. Those that did well were good at applying formulas, but lacked a conceptual understanding. One of Douglas Weathers' students, asked, "Can't we view all of these problems as a contingency table?" The answer was a resounding "Yes!" Students tended to understand probabilities and the idea of independence far better without reliance on formulas. We decided to revamp the entire section to use this approach since we value understanding over memorization. The result has been a success!

The notes include hands-on activities to introduce important topics. The activities aim to use previous knowledge for building intuition surrounding new topics. Some activities require additional materials such as dice or playing cards.

The notes have been used in classes of approximately 30 students. The class meets in-person three days a week at 50 minutes per meeting. There are occasionally more examples in the notes than there is time to cover in the class. In these cases, instructors have some flexibility to focus on examples more fitting to their class. Uncovered examples are suggested as extra practice. The solutions from the teacher's edition can be posted on the school's learning management system for students to check their work.

The class meets a fourth day for hands-on laboratory activities. These activities involve further data collection and analysis to reinforce concepts from class. The activities encourage group learning, building relationships between students, and observing the real challenges of data collection and analysis in practice. The laboratory manual is intended to be a low cost supplement to the course providing laboratory activities, homework sets that align with the notes, and additional practice problems with fully worked out solutions. The following is a citation of the supplemental workbook.

Bell, L. Jagannathan, K. Pritchard, N. (2023). *Understanding Statistics: Activities and Exercises for a First Statistics Course* (3rd ed.). Kendall Hunt Publishing. ISBN 9798765784525.

The laboratory manual is required at the author's institution. Additionally, the students are charged a small printing fee ($5) so that print copies of the interactive notes are provided to each student on the first day of class. A digital copy is provided on the learning management system for students who wish to annotate a digital copy. A calculator is also required (TI-84 recommended). Most students have such a calculator from their time in high school or are able to check one out from the library. This means the total cost of materials for the student is under $50. Classes do not meet in a computer lab, but it is helpful to have an instructor computer in which students can enter data to be projected for the class. Desks that can be rearranged are also a nice way to promote group work.

An instructor who wishes to use additional programs, e.g. Excel and R, can contact the authors for the LaTeX files to edit. Alternatively, an instructor might develop supplementary pages for using additional statistical programs if those are favored over the calculator.

The course assessment used at the author's institution has been broken down in the following manner:

- Homework (11%): Homework is turned in every Wednesday and assigned from the supplemental text. Upper level students who have successfully completed the course and are recommended by their instructors are responsible for grading the homework.
- Quizzes (11%): Quizzes are given at the end of class most Fridays. They cover material that was on the homework turned-in Wednesday. Quizzes are about 10-20 minutes in length (one page) and graded by the instructor. The short quizzes allow instructors to gauge understanding in an efficient manner throughout the course and encourage students to stay engaged with the material.
- Labs (8%): Students meet once a week for lab activities. These come from the required supplement previously mentioned. The labs reinforce the notes, homework, and quizzes with hands-on group work. Some lab activities require extra materials such as candy, scissors, dice, and cards. Student graders who are responsible for the homework also grade the lab activities.
- Unit Exams (51%): Three unit exams are given throughout the course. Exams are a mixture of multiple choice and free response questions that test student's ability in both computations and interpretations. The exams are typically broken down in the following manner. Slight variations may occur due to the alignment of breaks (e.g. Thanksgiving, spring break) during the semester.
  - Exam 1: Chapter 1 material
    - Data collection
    - Data visualization
    - Numerical summaries of data
    - Study designs
  - Exam 2: Chapter 2 and 3 material
    - Probability
    - Discrete distributions
    - Normal distribution
    - Inference for one proportion
  - Exam 3: Chapter 3 and 4 material
    - Inference for two proportions

    - Inference for one mean
    - Inference for two means
    - ANOVA

- Cumulative final exam (19%): Includes all previous material with the addition of chapter 5 material
    - Scatter plots
    - Correlation
    - Simple linear regression

The first 3-5 minutes of each class can be spent reviewing concepts from the previous class. The blank pages at the end of the notes are intended to be a place where students can write these summaries. They end up with a nice overview of all the important topics from the course. Students regularly provide positive feedback regarding this practice. Finally, some instructors enjoy the flexibility of these pages to include examples or discussions as they arise organically during class discussions.

Example schedules with weekly coverage from the course notes, homework problems, and lab activities are available upon request. The materials have been successfully used in both four and five week long summer classes. Pacing guides for summer courses are also available upon request.

One of the authors (Nick Pritchard) has created a website to supplement the course. The first portion of the page contains digital practice problems for each section of the notes. The second portion contains video explanations of the notes. The website may be found at `https://ww2.coastal.edu/npritcha/stat201resourcepage.html`.

Acknowledgments

The authors have enjoyed the creative process of sharing ideas and engaging with our colleagues to create and refine the introductory statistics experience for our students. The diversity of ideas shared has resulted in what we believe is a much stronger curriculum that we are excited to share with you. We would like to specifically acknowledge the other instructors of the course who tested the material and shared ideas: Dr. Keshav Jagannathan, Dr. Paul Hill, Dr. Joseph Njuki, Mrs. Patricia Whitaker, Mr. Craig Cook, Mrs. Jennifer Green, and Ms. Joyce Keenan.

Douglas Weathers would like to additionally thank his student, Lauralyn Clifford, who prompted the discussion which led to the reimagining of the probability section.

Chapter 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 who want to pursue degrees that are not offered by their home state institutions. There are multiple undergraduate programs at a particular university 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. Below is a list of undergraduate programs at this particular university 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

This is a great activity to encourage interaction and meeting classmates on the first day of class. Have students work with one or two others to introduce themselves and discuss the problem at hand. An alternative idea is to assign groups that will work together for the semester. This will be the first opportunity for the groups to get acquainted.

Points from class discussion

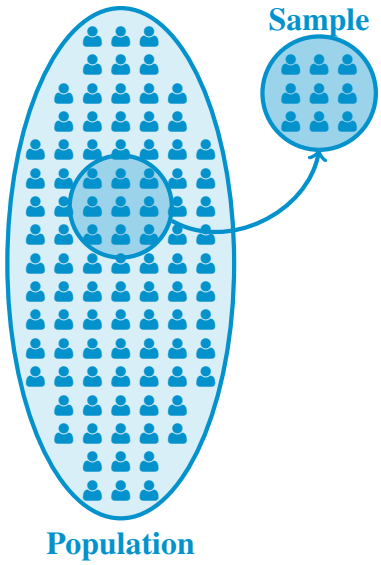
Summarize student's ideas on the board. Ideas should address two main points:

1. What information should we gather to answer the primary questions? That is, what variables should we observe? Some possibilities include:
   - Major
   - Distance from home (How would we measure this?)
   - Residency
   - Participation in ACM (Y/N)
2. How do we go about gathering this information? This allows for insightful discussion on the benefits and drawbacks to each method, which are discussed in more detail in the following pages. Possible discussion includes:
   - Asking students in your class
   - Creating a survey to be emailed to students
   - Asking students walking between classes
   - Contacting Institutional Research for the information

1.2 Data Collection

Sample and Population

Our Investigation in a Picture



- **Population:** the entire group of interest.

  ex. All students at the university

  The population is described or summarized by *population parameters* which are often denoted by Greek letters.

  ex. μ = average GPA for all students at the university

- **Sample:** A subset of the population

  Example: A sample of 100 students in an introductory statistics course

  The sample is described or summarized by *sample statistics* which are often denoted by Roman letters.

  ex. $\bar{x}$ = average GPA for 100 students

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

Example 1. *A poll commissioned by 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.*

- **Sample:** The 2,000 smartphone users in the US
- **Population:** All smartphone users in the US
- **Statistics:** From the sample, we have 60% that could not cope without their smartphone for a day along with an average of 20 times per day checking their phone.
- **Parameters:** No parameters are given

Example 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.*

- **Sample:** Tiger sharks sampled near Garden City
- **Population:** All tiger sharks
- **Statistics:** $\bar{x}_L = 9.5$ ft; $\bar{x}_W = 890$ lbs
- **Parameters:** $\mu_L = 12$ ft; $\mu_W = 1,000$ lbs

Poor Sampling Techniques

Example 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?

The percentages seem high. Based on the sampling methods we are suspicious if those sampled truly represent ALL women.

Measurement bias: Measurements tend to record values larger or smaller than the true value. This type of bias is typically introduced by incorrectly calibrated instruments such as a tape measure affixed too high on the wall or a survey with poorly worded questions.

Sampling bias: The sample is not representative of the population

1. **Voluntary sampling:** People with strong feelings one way or the other are the ones tending to respond.

   Consider product reviews, Rate My Professor and online polls. Responses tend towards the extremes. Offering incentives may help get more responses from the middle.

2. **Convenience sampling:** Only a "convenient" group is surveyed. This may leave out many subgroups of the population.

   Consider taking a survey at the mall at 10am on Thursday while enjoying Cinnabon. This would miss many segments of the population such as those who work standard 9-5 jobs.

AMS Open Math Notes: Works in Progress; Reference # OMN:202412.111448; Last Revised: 2024-12-15 09:20:44

3. **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?

This can be a brief discussion. For example, one suggestion may have been to survey only students in introductory statistics courses. Students in the course are most likely science majors and we would miss many other groups of students who are not in science programs.

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 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 calculator to obtain an SRS of the accounts to audit.*

We label the accounts 01 to 60, then on the TI-83 or 84 calculator:

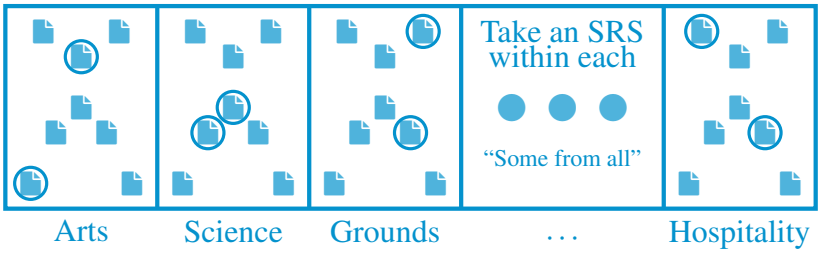
- Set the seed (Optional for reproducible values):
  Enter any number then, `STO>`→`MATH`→`PROB`→`rand`→`ENTER`
- Obtain random numbers:
  `MATH`→`PROB`→`RandInt(lwr,uppr,number of items)`

Walk through the steps on the calculator with the students. For our example, we could use `RandInt(1,60,8)`. Discuss with students what to do if values are repeated on the output.

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.

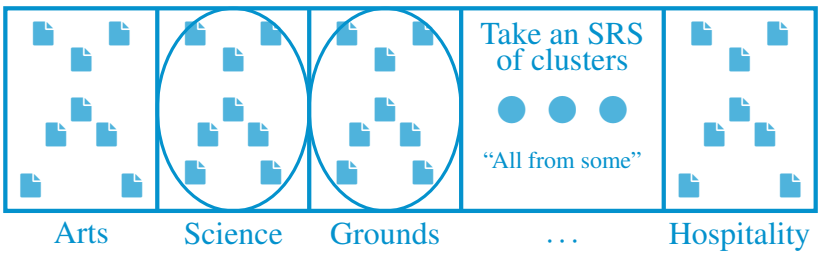
In our example, suppose the accounts can be divided into strata by department/function such as arts, sciences, grounds and maintenance, and hospitality. We wish to take a stratified sample so that all departments are represented.

The memory tool of alliteration with stratified sampling and "Some from all" helps students distinguish this sampling approach.



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.

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. There is no practical reason for a cluster sample in this example.



Systematic sampling: Every k^{th} subject is sampled.

As an example, consider sampling every 10^{th} paper towel from a roll to test for absorbency.

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

Feel free to have an open discussion with your students about possible options using random sampling. Here are a few examples:

- Using a database, we could take a simple random sample of students from the school population.
- Using stratified random sampling, we could take a simple random sample of students from each major or classification.
- Using cluster sampling, we could take a simple random sample of classes and include all students within those selected classes in our sample.
- Using systematic random sampling, we could sample every 50th student listed withing the student database.

Revised Investigation: The Office of Institutional Research, Assessment and Analysis (IRAA) took a simple random sample of 100 students enrolled in introductory statistics 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 so far
- **Enrolled Credits** - number of credit hours enrolled during the fall 2022 semester
- **Residency** - the residency status of each student (in-state/out-of-state/international)
- **Distance** - the number of miles each student's hometown is from campus

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.

It is difficult to ascertain any useful information from the two-page table of raw data. Constructing graphical and numerical summaries would help us gain a better understanding of the "layout" (distribution) of the data and any relationships that might exist.

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

Investigation Data Set Continued ...

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

1.3 Data Visualization for a Single Variable

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?

Some data are numerical and could be summarized by averages (ex. distance from home, number of credits). Some data are categories and could not be summarized by an average (ex. major). Proportions and counts might be better for this kind of variable.

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

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

Plots for Categorical Variables

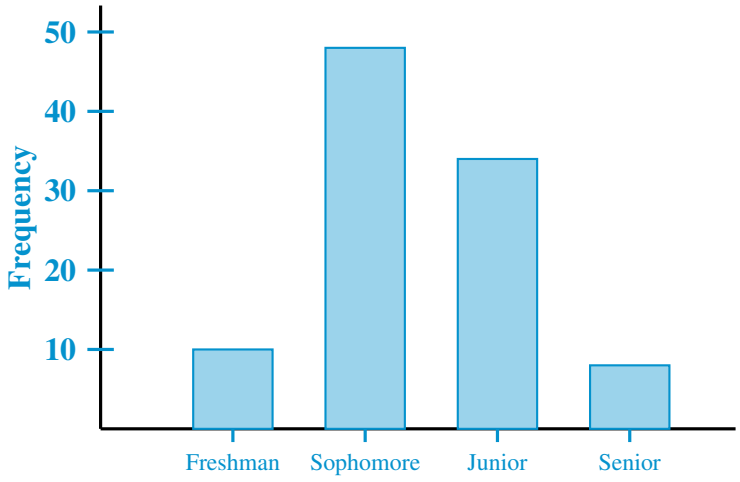
First, we introduce the construction of some basic plots that are appropriate for each type of variable. Later, we will discuss the information provided by the plots and relevant interpretations. Let's turn our focus on categorical variables first.

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.

Example 6. *Consider the classification/year of each student.*

Classification/Year	Frequency	Relative Frequency
Freshmen	10	10/100 = 0.10
Sophomore	48	48/100 = 0.48
Junior	34	34/100 = 0.34
Senior	8	8/100 = 0.08

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:

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

The mode is the most frequently observed category. Sophomores are the mode of classification in our data.

(b). Are there any other important pieces of information displayed in the bar graph?

Juniors are also quite frequent, while freshmen and seniors are least common in introductory statistics courses according to the data. (Note that this is data from the fall semester, and the spring typically has a higher proportion of freshmen.)

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

No! The proportions/relationship between categories would remain the same. The scale on the vertical axis would be the only thing that changes.

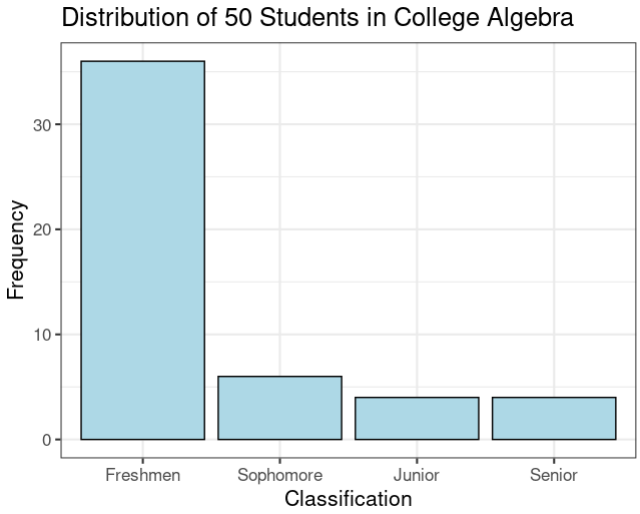
(d). When would it be important to use relative frequency over frequency?

It would be especially important to use relative frequency, or percentages, if we wish to compare our results to those in another data set that may differ in sample size.

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

Absolutely! In this example, there is a natural ordering to the categories (ordinal). In other instances there may not be a natural ordering (nominal) and we might order alphabetically or by frequency for example.

(f). Consider the following bar graph with classification for a random sample of 50 students taking College Algebra at the same university. Compare the diversity of classifications between students in College Algebra and Introductory Statistics.



Students in College Algebra have less diversity in classification because they are overwhelmingly freshmen. Students in Introductory Statistics tend to have more diversity in classification because there are many sophomores and juniors rather than primarily one classification.

Plots for Quantitative Variables

Histograms are similar to bar charts, but for quantitative data. The data are grouped into bins of equal width. The frequency of observations within each bin is indicated by the height of the bar.

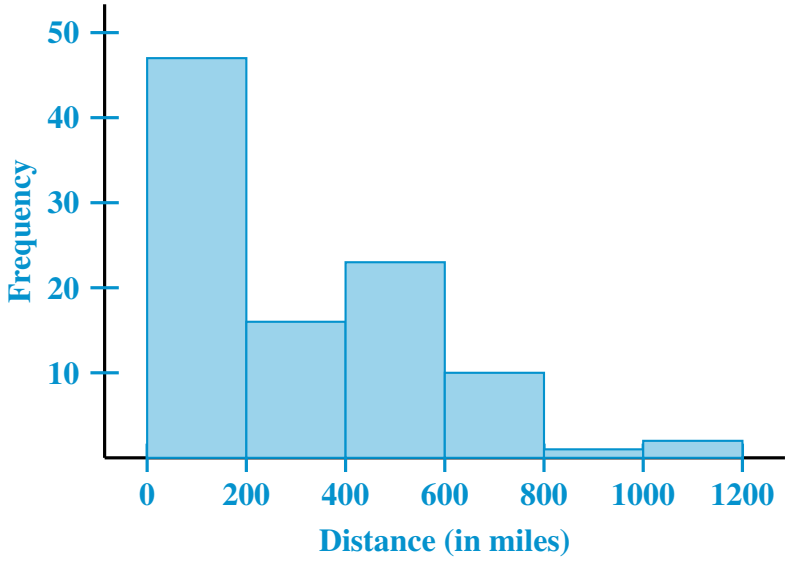
Let's construct histograms for two of our quantitative variables: distance from home and cumulative credits.

Example 7. *The data for "distance from home (miles)" are presented in numerical order below. Use this to first construct a data distribution. Then create a histogram and make observations on what you see.*

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
$[0, 200)$	47
$[200, 400)$	16
$[400, 600)$	23
$[600, 800)$	10
$[800, 1000)$	1
$[1000, 1200)$	2

Histogram for Distance from Home



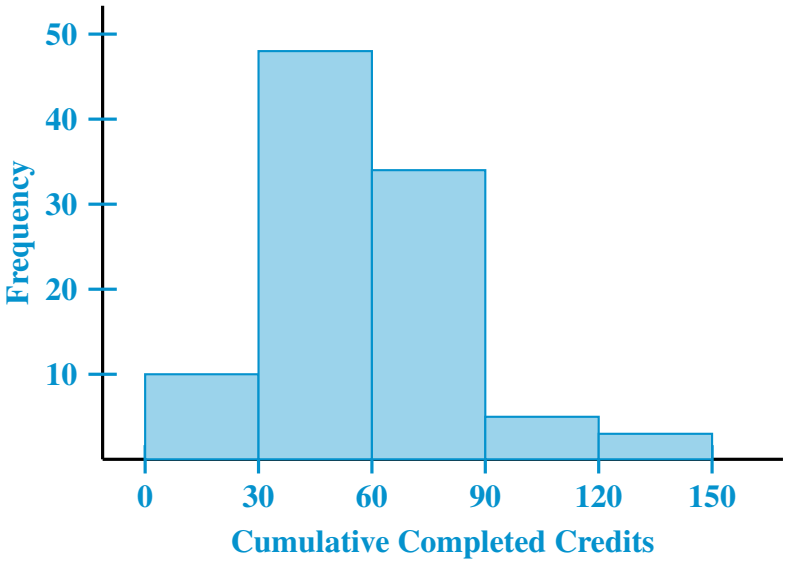
Observations: Most students are from nearby while only some come from far away (right skewed). The distances are centered around 400 miles and cover a range of 0 to 1200 miles. There are a couple students from over 1000 miles away that might be considered outliers.

Example 8. *The data for "cumulative credits" are presented in numerical order below. Use this to first construct a data distribution. Then create a histogram and make observations on what you see.*

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 Credits	Freq.
$[0,30)$	10
$[30,60)$	48
$[60,90)$	34
$[90,120)$	5
$[120,150)$	3

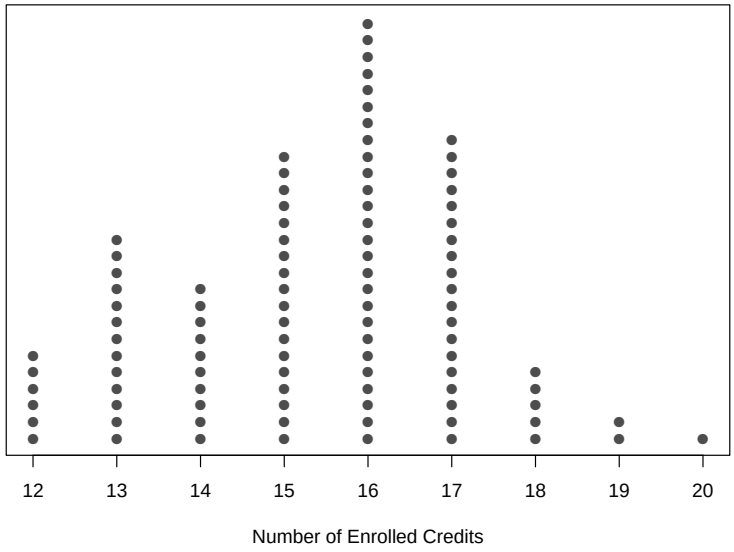
Histogram for Number of Cumulative Credits Completed



Observations: The number of cumulative credits are fairly symmetric around 60 credits and range from 0 to 150 credits.

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.

Example 9. *Consider the following dot plot of the number of enrolled credits during fall 2022 for the random sample of students in our investigation.*



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

The most number of credits a student was enrolled in is 20 credits. Only one student is enrolled in 20 credits.

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

Eight students were enrolled in 18 or more credits.

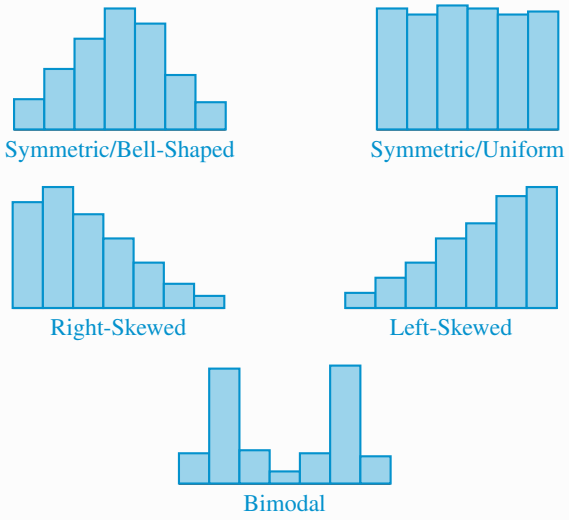
(c). What percentage of students sampled were in enrolled in less than 15 credit hours during fall 2022?

29 out of 100 students were enrolled in less than 15 credit hours (29%).

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.

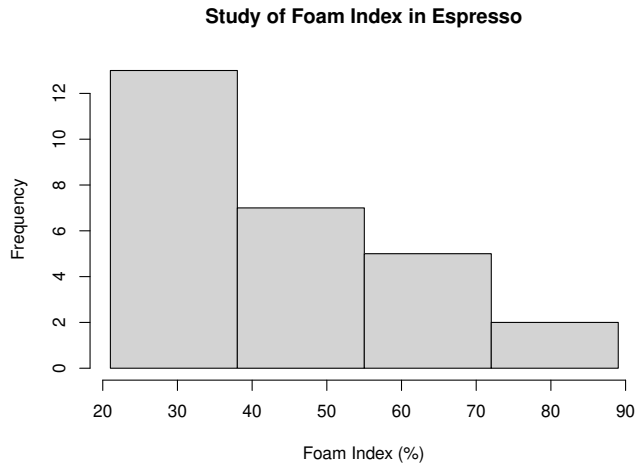
1. **Shape** describes the overall layout of the data and typically can be categorized in one of three broad categories. Some categories can appear a few different ways. Note that real data will rarely follow one of these shapes exactly.



2. **Center** describes the centrality or location 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 variation in the data (minimum value to maximum value). It describes 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.

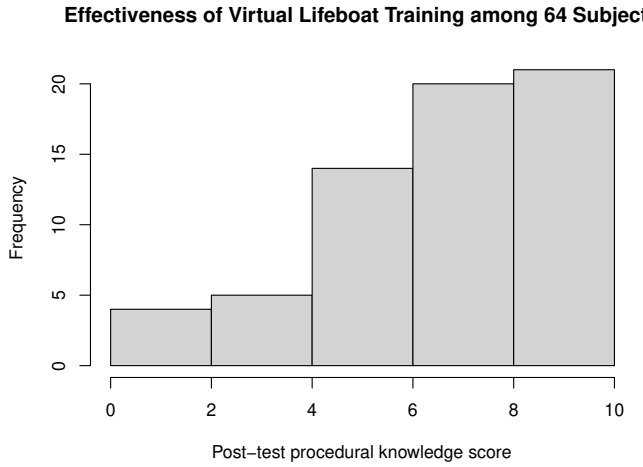
Check our observations of the previous histograms to ensure we have addressed all the key features. Then identify the key features of the following examples.

Example 10. *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.*



The shape is right skewed with a center about 40-45% foam index. Values are spread between 20% and 90% foam index with no apparent outliers.

Example 11. *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.*



(a). Assess the key features of the plot.

The plot is left skewed with center around 7 points. Scores ranged from 0 to 10 points and there are no apparent outliers.

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

About $4+5=9$ *scores were below 4 points.*

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

Approximately $9/64\times 100\% = 14.06\%$ *of scores are below 4 points.*

(d). Approximately what proportion of scores are at least a 6?

The proportion of scores that are at least a 6 is $41/64 = 0.6406$.

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

Scores between 8 and 10 points are most common.

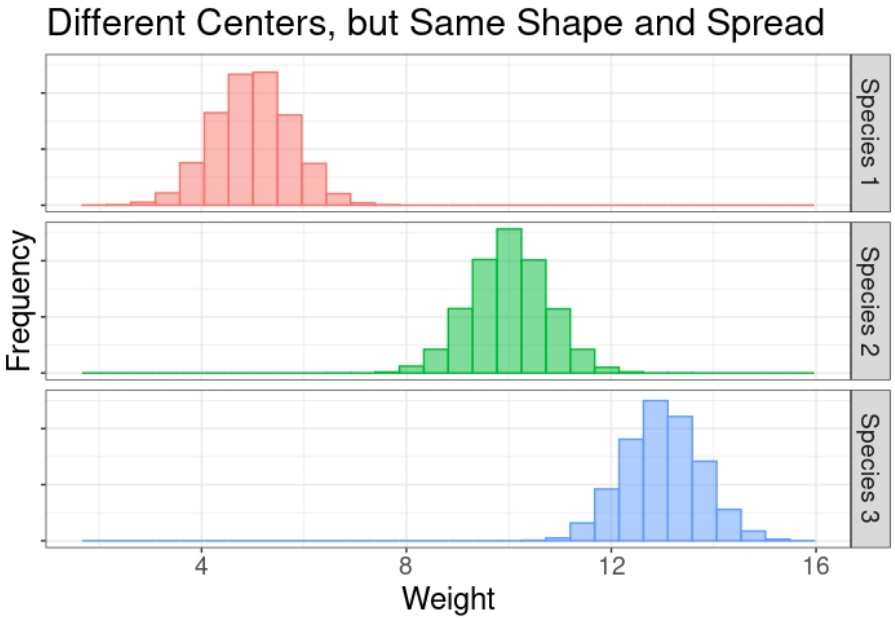
Recapping the Meaning of Center and Spread

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 will measure center with mean or median, depending on the shape of the data.

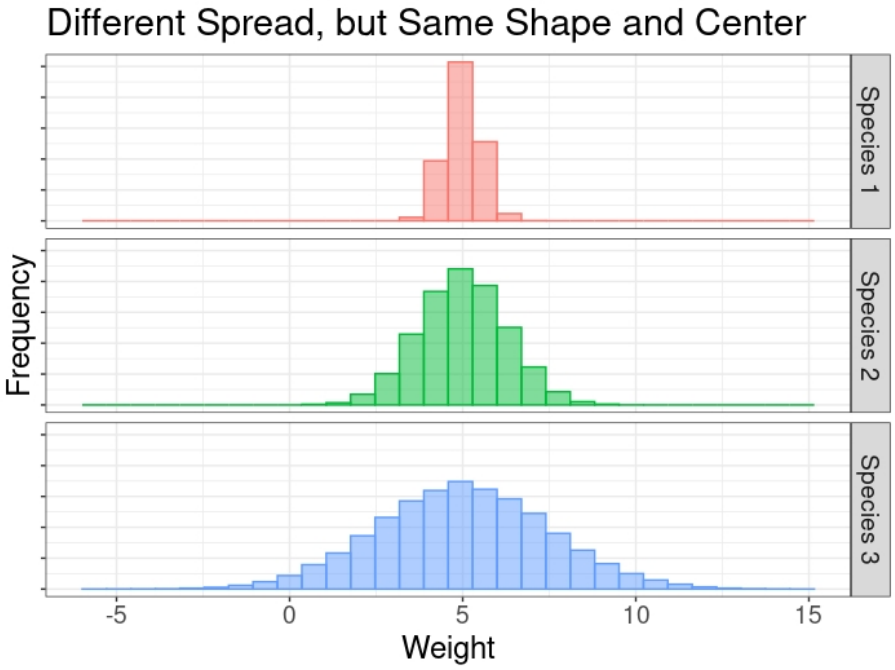
The three species all have weights that are symmetric with the same spread. However, Species 3 tends to be heavier with a center around 13 while Species 1 tends to be the lightest with a center around 5.

While the distinction of center and spread may seem straightforward, we have found that many students struggle with differentiating these ideas. This is especially clear when we interpret comparative plots in the following sections. Reiterate the meaning of center and spread as much as possible.



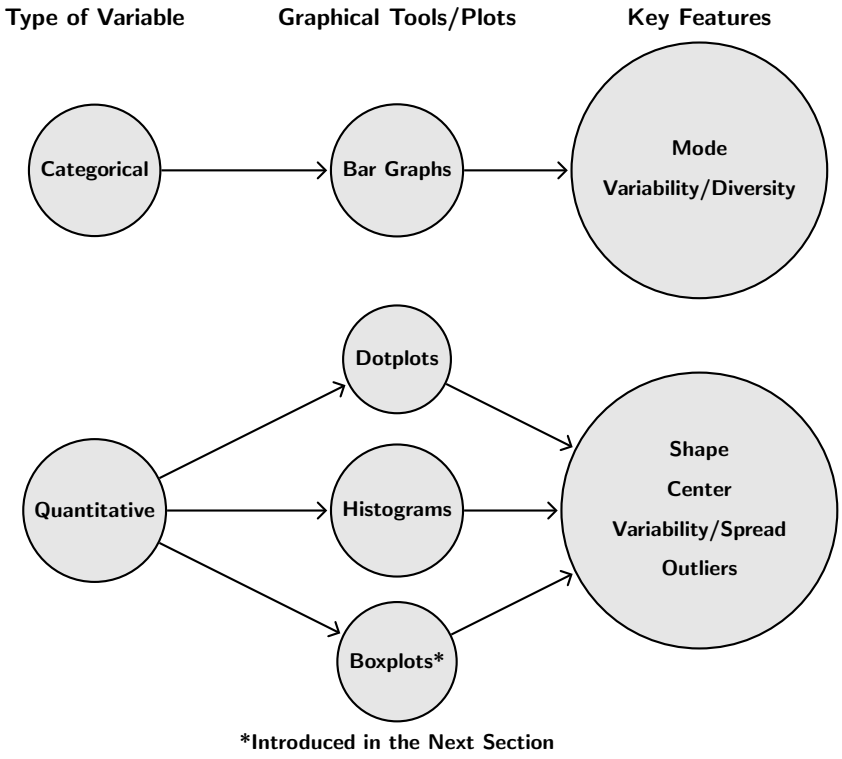
Spread: Spread measures the **variability** of the data. That is, spread describes how tightly or loosely grouped the data are. We will measure spread with standard deviation or IQR, depending on the shape of the data.

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).



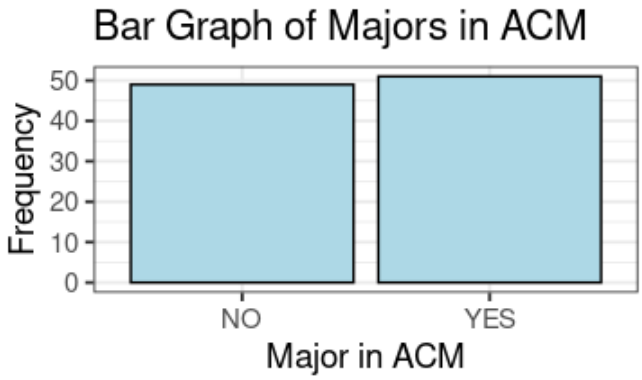
Overall Summary of Plots

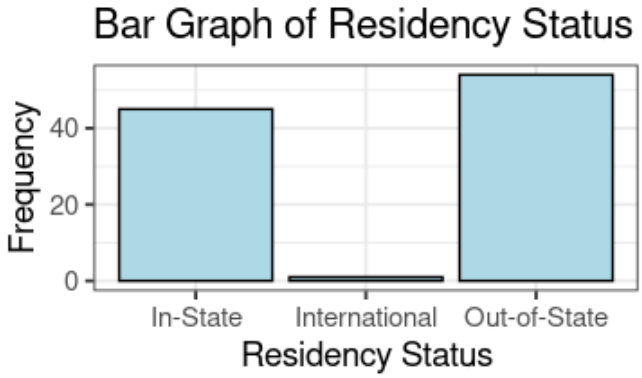
The following map summarizes the graphical tools and key features for both categorical and quantitative variables.

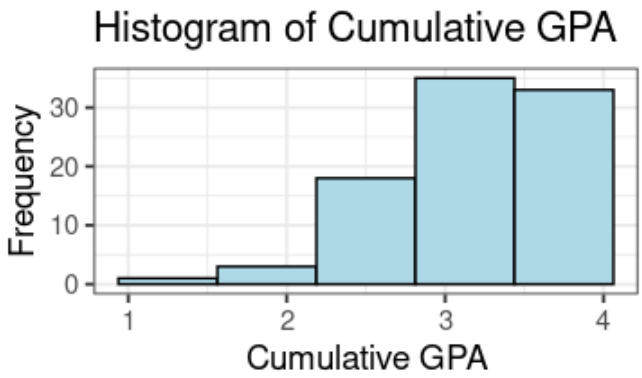


If time allows, use the this page to construct some additional plots from our course investigation and comment thoroughly on what you see.

The following plots were constructed using R. There may or may not be time to do this based on pacing of the course up to this point.







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.

Measures of Center

You are likely familiar with methods for summarizing central tendency in the data. List those here. Identify a quantitative variable from our course investigation and if time allows, compute measures of center.

- **Mean:** the average; can be viewed as the "balance point" of the data
  - Population Mean: μ
  - Sample Mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\text{add up the data points}}{\text{total number of data points}}$$

- **Median:** the middle data point (of the ordered data set)
  - When n is odd, the median is the middle observation
  - When n is even, the median is the average of the two middle observations.

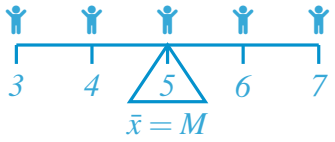
Caution: Do not forget to order the data before finding the median by hand!

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 and presence of outliers.

Example 12. *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.*



- $\bar{x} = \frac{3+4+5+6+7}{5} = 5$
- $M = 5$
- *The data are symmetric and we notice that the "balance point" and the middle point are the same.*

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.

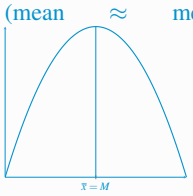
We expect the mean to shift right (that is, increase) to "balance" Little John's move.

A numerical summary is **resistant** if it is influenced little by extreme observations. Which measure of center is resistant?

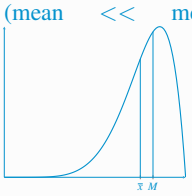
The median is resistant while the mean is not resistant.

Understanding the Relationship between Shape and Center

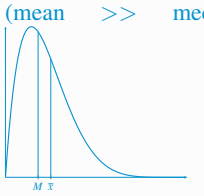
Symmetric
(mean ≈ median)



Left-skewed
(mean << median)



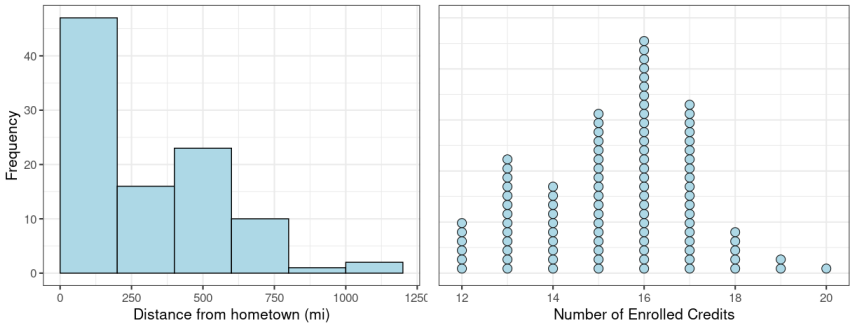
Right-skewed
(mean >> median)



Therefore,

- if the data is skewed or has outliers we will use the median
- if the data is symmetric we will use the mean

Discuss: Consider the plots for additional 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?



Shape is right-skewed
Use the median

Shape is roughly symmetric
Use the mean

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?

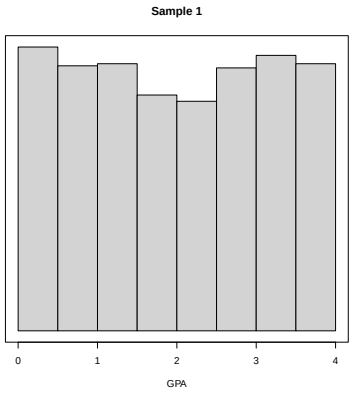
The simplest measure of spread is the **range**: The largest observation minus the smallest observation.

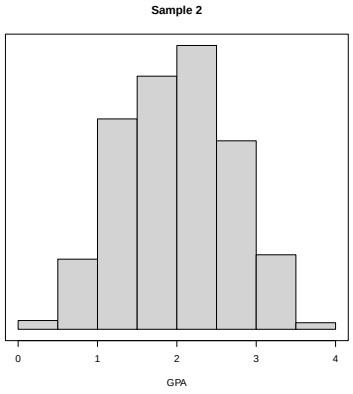
Example 13. *Find the range for our two data sets. Is the range resistant? No!*

3 ft, 4 ft, 5 ft, 6 ft, and 7 ft: $Range = 7 - 3 = 4\ ft$

3 ft, 4 ft, 5 ft, 6 ft, and 17 ft: $Range = 17 - 3 = 14\ ft$

Example 14. *Consider the following two graphs relating to GPA of separate samples of students. For each sample, approximate the value of the following statistics:*





- *Mean*

  $\bar{x} \approx 2$ *for both samples*

- *Median*

  $M \approx 2$ *for both samples*

- *Range*

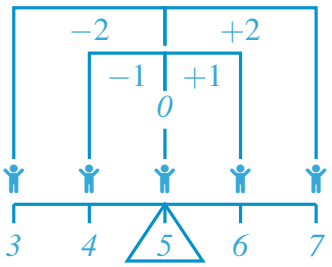
  $Range \approx 4$ *for both samples*

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

  No. Sample 2 has less variability with most of the observations very close to 2. Sample 1 has much more variability with samples equally distributed between 0 and 4.

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

Example 15. *Find the standard deviation of our data set of 3ft, 4ft, 5ft, 6ft, and 7ft:*



(1) $x-\bar{x}$	*(2)* $(x-\bar{x})^2$
$3-5=-2$	*4*
$4-5=-1$	*1*
$5-5=0$	*0*
$6-5=1$	*1*
$7-5=2$	*4*
$\sum(x-\bar{x})=0$	$\sum(x-\bar{x})^2=10$

(3) ***Variance:***

$$s^2=\frac{\sum(x-\bar{x})^2}{n-1}=\frac{10}{5-1}=2.5ft^2$$

This represents the "typical" squared distance from the mean.

(4) ***Standard deviation:***

$$s=\sqrt{s^2}=\sqrt{2.5}=1.5811ft$$

This represents the "typical" distance of points from the mean.

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.

The standard deviation for the this set should be larger since distances from the mean are larger.

$s=5.7009$

Is standard deviation resistant to outliers?

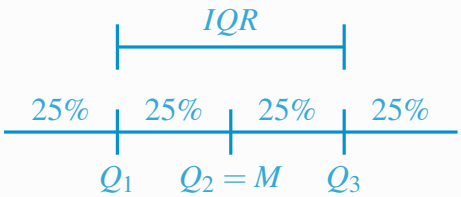
No! The standard deviation (sd) changed greatly when we moved one observation.

Example 16. *Determine whether each value can or cannot be a standard deviation.*

1.2 1×10^{-5} -2.1 100

A more resistant measure of spread in the data is the **interquartile 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:

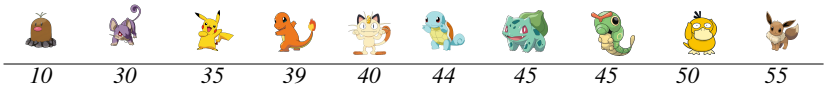
- **Percentiles** - the pth percentile is the measurement to which $p\%$ of all measurements fall below it and $(100-p)\%$ lie above it.
- **Quartiles** - special percentiles that divide the data into quarters.
  - Q_1 - First Quartile - 25th Percentile
  - Q_2 - Second Quartile - 50th Percentile (Median)
  - Q_3 - Third Quartile - 75th Percentile



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

Example 17. *The following data gives the HP (Hit Points) for a random sample of first generation Pokemon.*

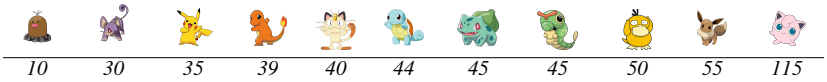
(a). Find the quartiles and the IQR for HP.



10	30	35	39	40	44	45	45	50	55

$$IQR = Q_3 - Q_1 = 45 - 35 = 10$$

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



$IQR = Q_3 - Q_1 = 50 - 35 = 15$

The standard deviations for the data sets are $s_1 = 12.5$ and $s_2 = 25.7$ respectively. Notice the standard deviation changed greatly with this addition. Is the IQR resistant to outliers?

Yes! Jigglypuff could have 115,000 HP and the IQR would be the same.

Recap: What measures of center and spread should be used for skewed data? For symmetric data?

Shape	Center	Spread
Skewed/Outliers	Median	IQR
Symmetric	Mean	Standard Deviation

Outliers and Boxplots

Test for Outliers

Five Number Summary:

Minimum, Q_1, M, Q_3, Maximum

Outliers:

- $IQR = Q_3 - Q_1$
- $Step = 1.5(IQR)$
- Inner Fences: $LF = Q_1 - Step$ $\quad UF = Q_3 + Step$
- Any observation beyond the fences is considered an outlier.

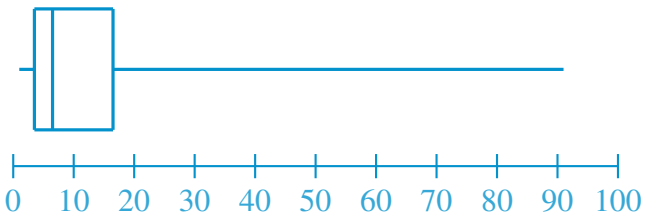
Example 18. *The following values are a sample of characters from Walking Dead and their corresponding number of zombie kills for the first three seasons. Find the five number summary, test for outliers, and construct boxplots.*

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

Five Number Summary:

Min $= 1$
$Q_1 = 3.5$
$Q_2 = 6.5$
$Q_3 = 16.5$
Max $= 91$

Boxplot:

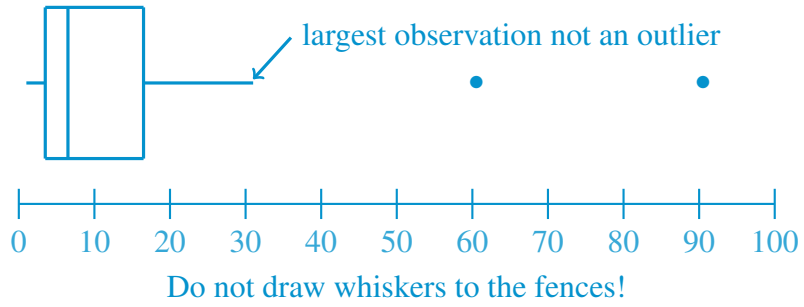


Test for Outliers:

- $IQR = 16.5 - 3.5 = 13$
- $Step = 1.5(13) = 19.5$
- $LF = 3.5 - 19.5 = -16$

  $UF = 16.5 + 19.5 = 36$
- Daryl and Rick are considered outliers (61 and 91)

Modified Boxplot:



Do not draw whiskers to the fences!

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, dot plot, and/or boxplot.

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

Question: Recall our original questions from the course investigation:

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?

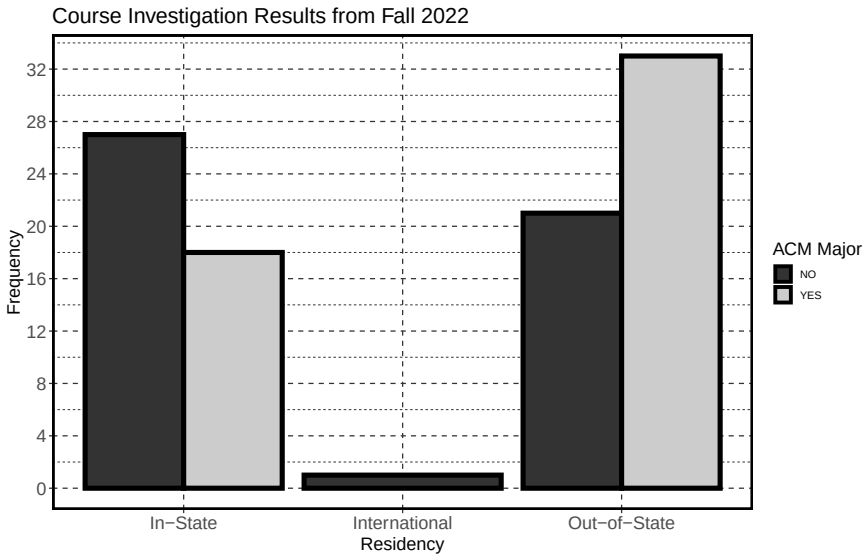
Students may offer a variety of ideas. Some possibilities include:

- Bar chart of residency shaded by proportion participating in ACM
- Two different histograms or boxplots for ACM status
- A scatter plot of number of credits and GPA

Graphs for Comparing Two Categorical Variables

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

Example 19. *The following side-by-side bar graph compares the residency of students to choosing a major in a program available through ACM at the university.*



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

There are roughly 54 out-of-state students, about 33 of whom chose a major available through ACM. Therefore, $33/54 = 0.6111$

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

There are about 45 in-state students, 18 of whom chose a major available through ACM. Therefore, $18/45 \times 100\% = 40\%$

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

There does seem to be an association since a much higher proportion of out-of-state students chose a major available through ACM than in-state students.

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 20. *The following table provides the five number summary for the distance between the university and hometown for each student, grouped by ACM status.*

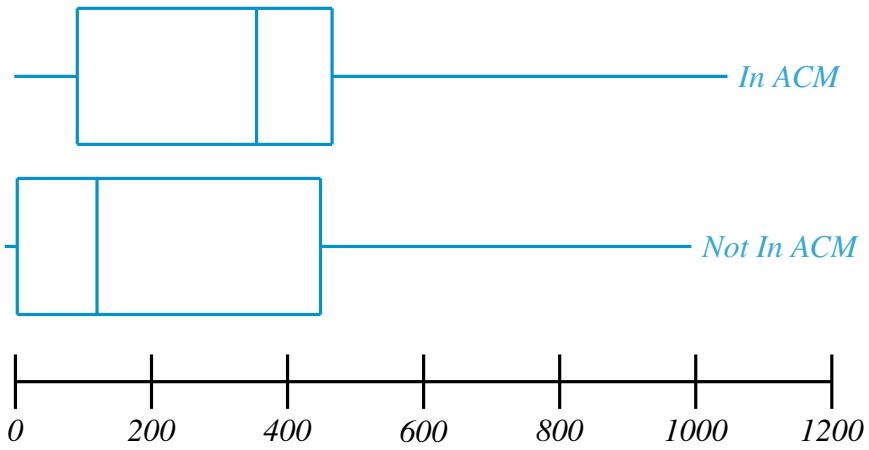
Major in ACM

Min	Q_1	Q_2	Q_3	*Max*
15.2	*107.8*	*371.3*	*482.3*	*1063.5*

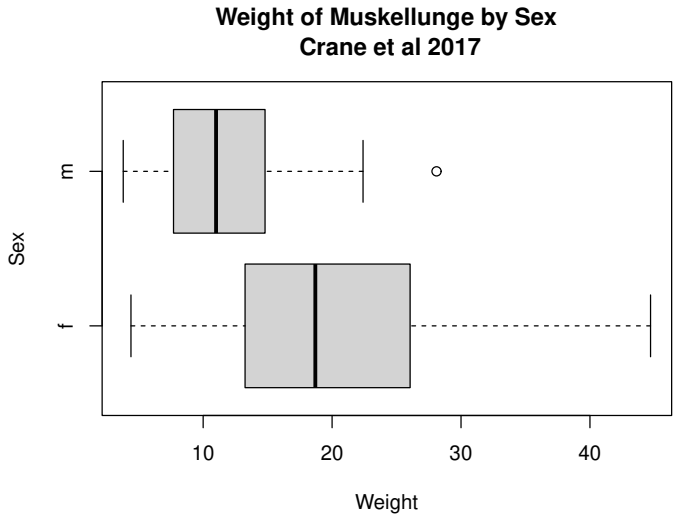
Major Not in ACM

Min	Q_1	Q_2	Q_3	*Max*
1.10	*19.55*	*136.80*	*465.30*	*1010.40*

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



Example 21. *Crane (2017) 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?

Females tend to be heavier. Their median weight is closer to 20 units while males have a median weight closer to 10 units.

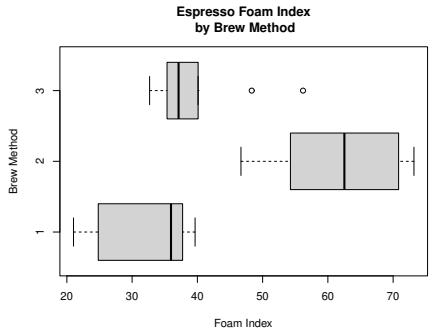
(c). Which sex tends to have more variability in weights? How can you tell?

Females have more variability in weights than males. The boxplot shows a larger IQR and range for females.

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

Males have a symmetric distribution with an outlier on the right. Females have a fairly symmetric distribution of weights with a slight right skew.

Example 22. *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 boxplot. However, it is important to note that the foam index was actually calculated for three different brewing techniques. We view the data by brewing technique using a boxplot. 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?

Method 2 results in the highest foam index while Method 1 results in the lowest. We can see this by comparing medians.

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

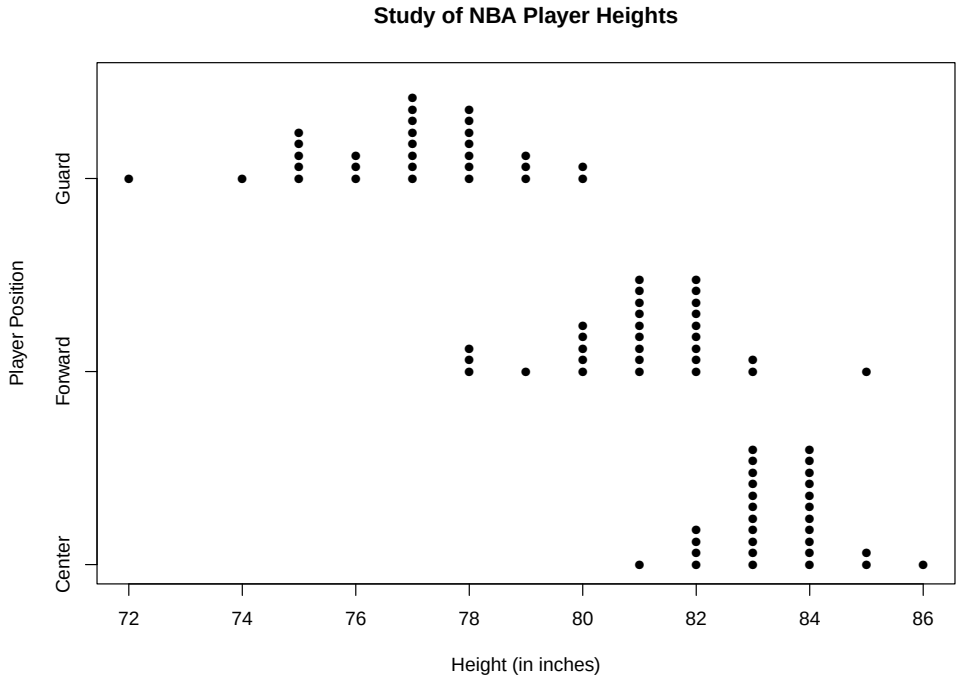
Method 3 has an interesting distribution because there is no discernible line to the right. This means the largest observation that is not an outlier is equal to or very close to Q_3. There are also two outliers in the distribution and Method 3 appears to have much lower variability in responses than the other two methods.

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

Method 1 has a left skewed distribution while Method 2 is fairly symmetric.

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 23. *A student was interested in comparing the heights of different player positions in the National Basketball League (NBA). She took a stratified 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?

The tallest player is the 86 inch center and the shortest player is the 72 inch guard.

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

Centers tend to be tallest with a central value around 83 or 84 inches. Guards tend to be shortest with a central value around 77 or 78 inches.

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

The 72 inch guard appears to be unusually short compared to the other guards. The 85 inch forward may be unusually tall compared to the other forwards.

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

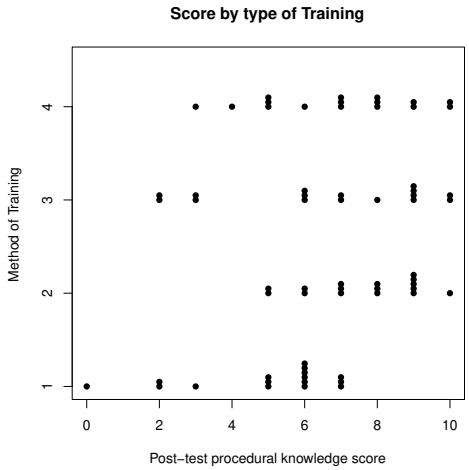
The distribution for centers is fairly symmetric. The same can be said for guards.

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

Centers have the least variability. The range of values is the shortest and the vast majority of observations are either 83 or 84 inches.

Example 24. *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.

Method 1 (traditional lecture and materials) performed the worst because the scores were centered around 6 points out of 10. Method 2 (monitor and keyboard) performed the best because the scores were centered around 8 points out of 10.

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

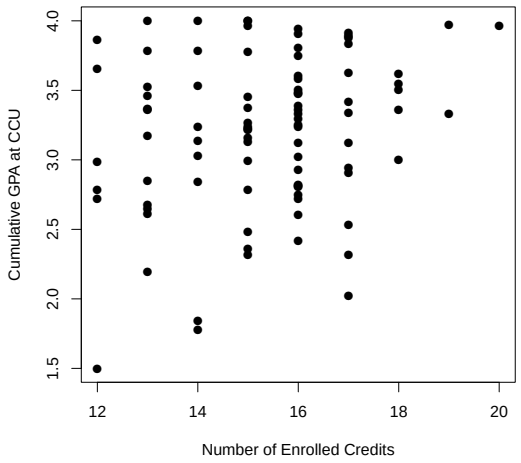
Method 4 (head monitor and display wearables) had the most variability because the range is largest (7 points) and the scores were evenly distributed in that range. Method 1 (control group) had the least variability. While the range is the same as method 4, most observations were either 5, 6, or 7 points rather than evenly distributed.

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.

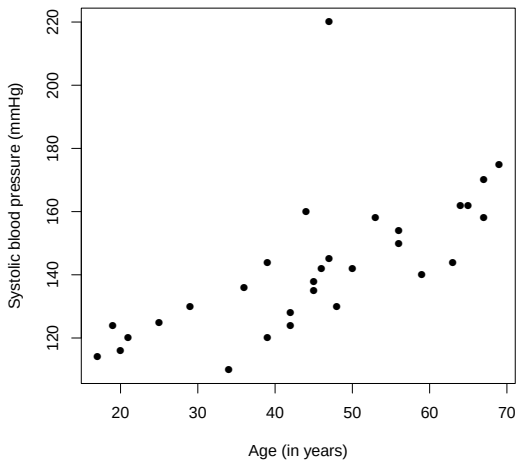
If it is thought that one variable exerts influence on the other, it is plotted on the x-axis and called the **explanatory variable**. The variable on the y-axis is called the **response variable** and may be impacted by or respond to the explanatory variable. We will discuss this in more detail in the following section.

Example 25. *Consider the scatterplot between the cumulative GPA and number of enrolled credits for the random sample of students during fall 2022. What are some basic observations?*



There appears to be little relationship between number of enrolled credits and cumulative GPA. That is, as number of enrolled credits increases, there is no clear impact on GPA.

Example 26. *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?*



As age increases, systolic blood pressure also tends to increase in a linear fashion. This is a fairly strong/consistent trend. There is one outlier. The individual who is about 50 years old has much higher blood pressure than the general trend would suggest.

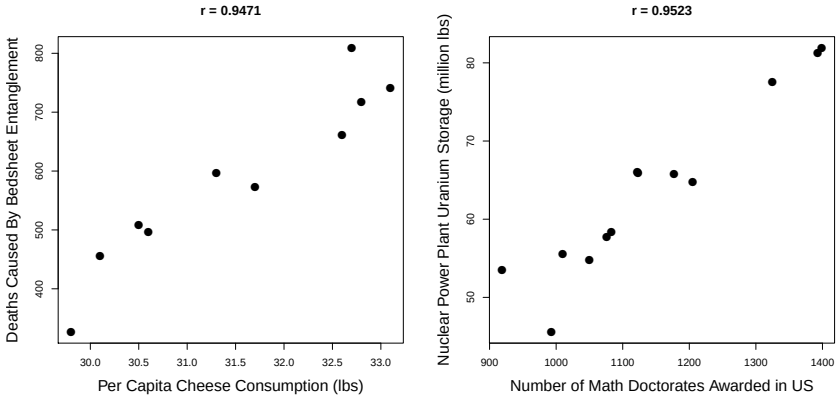
Example 27. *In* ***Example 26****, we examined that systolic blood pressure generally increases as age increases in adult males. Does this mean that aging in men causes SBP to increase?*

No! There are many other factors related to aging that could in fact be the reason for the increase (ex. more sedentary lifestyles).

Example 28. *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?*

No! For example, more fire trucks may be called to larger fires.

Example 29. *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?*



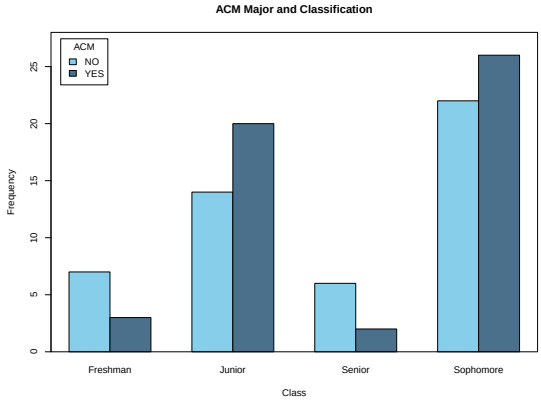
Of course not!

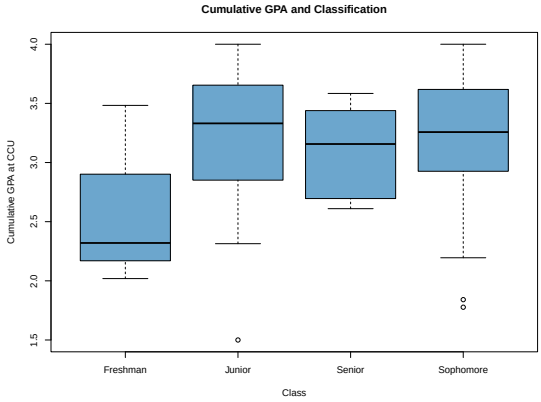
It is easy to find examples like the ones above. Do not be fooled when someone attempts to convince you that there is a causal relationship between two variables. Chances are good that there isn't one. In the next section, we will discuss the best way to determine a casual relationship.

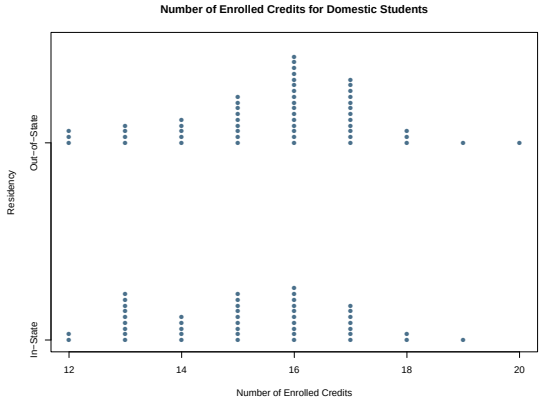
We will discuss scatterplots and correlation in more detail towards the end of the semester.

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

The following plots were constructed using R. This page can be done as time allows.







AMS Open Math Notes: Works in Progress; Reference # OMN:202412.111448; Last Revised: 2024-12-15 09:20:44

1.6 Study Designs and Conclusions

Observational vs Experimental Studies

We previously noted that while per capita cheese consumption and deaths by bed sheet entanglement 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 (i.e. **explanatory variable**), are directly responsible for change in another variable (i.e. **response variable**).
- **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 30. *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)

Response: fear of spider; Explanatory: time outdoors

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

Response: cancer risk; Explanatory: radiation exposure

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

Response: depression; Explanatory: physical activity levels

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

Response: Alzheimer's symptoms; Explanatory: level of social interaction

Example 31. *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.*

We are not able to conclude that video games played while children caused violent tendencies as adults. There may be other factors, such as poor home life, that tend to occur in those groups more commonly as well.

Observational studies

Subjects are not randomly assigned to treatment groups. Groups may exist, but they are not randomly assigned (ex. classification). Without randomization to groups, we may only conclude **association** between variables.

Experimental studies

Subjects are randomly assigned to groups. In this case we may conclude **causal relationships (cause and effect)**.

By randomly assigning subjects to groups, the effects of any **lurking variables** should be equally dispersed among all groups and not have an out-sized effect on any one group.

Example 32. *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.*

(a). (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.

This is an observational study because subjects are simply surveyed rather than randomly assigned to groups. As an observational study, we may only conclude association between Facebook and obesity due to several potential lurking variables (ex. physical activity).

(b). 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.

This is an experimental study because subjects were randomly assigned to view a particular website. As an experimental study, we may only conclude causation because lurking variables should be dispersed equally in both groups.

(c). 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.

This is an experimental study because subjects were randomly assigned to one of several radiation types and levels. As an experimental study, we may only conclude causation because lurking variables should be dispersed equally in both groups.

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 33. *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 experiment. Are any omitted?

- **Control group** - the face with no occlusion serves as the control
- **Randomization** - the images were shown in a randomized order to those rating the appearance
- **Blinding and double blinding** - blinding is not possible in this study
- **Replication** - replication is achieved by the use of 40 different faces and 42 participants rating the faces

Experimental Design

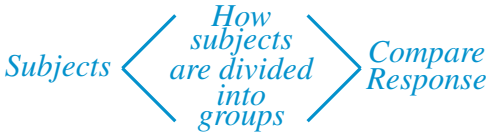
Suppose we are interested in studying a cause and effect relationship using an experimental study. There are many common experimental designs that a researcher may use to assign subjects to treatments groups. Here we discuss three of those and introduce related vocabulary.

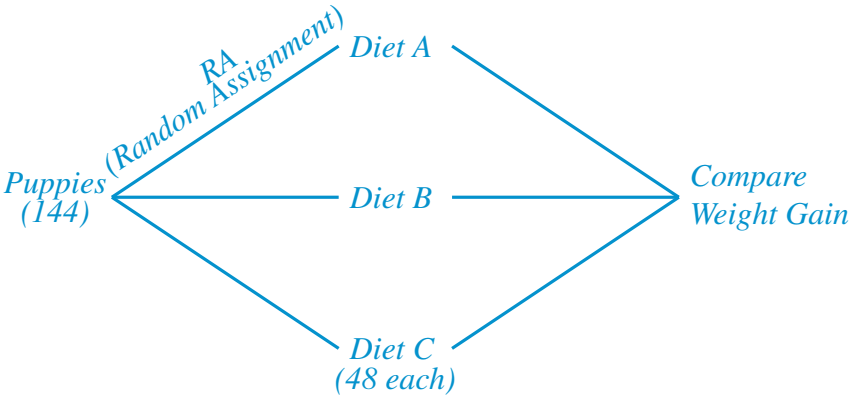
- Subjects/**experimental units** - who or what is receiving the treatment in the study
- **Response** - the outcome of interest; what is being measured by the researcher
- **Factors** - variables that may affect the response (explanatory variables)
- **Factor levels** - the possible categories or levels 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 34. *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 (RA) 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: 144 puppies*
- *Response: Weight gain after three weeks*
- *Factor: Diet*
- *Factor Levels: A, B, C*
- *Treatments: A, B, C*
- *Outline:*

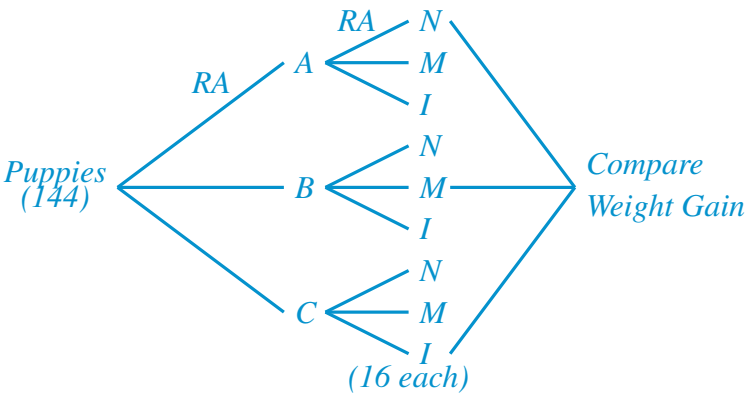




Design II: Multifactor Design (Two Factor) - the effects of two or more factors are considered.

Example 35. *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 puppies of 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.*

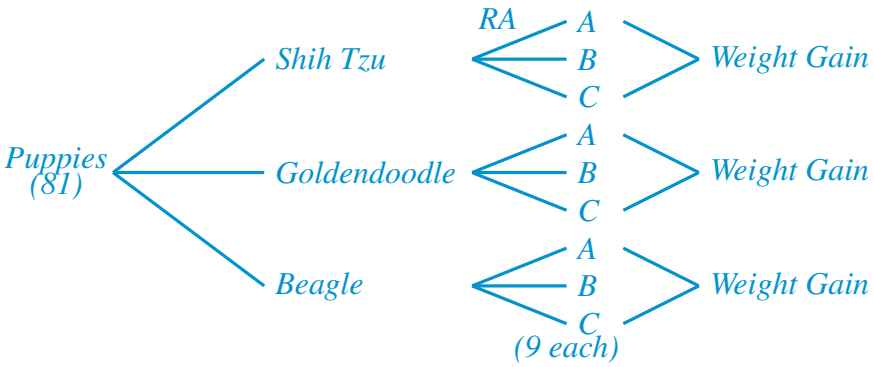
- *Experimental units:* *144 puppies*
- *Response:* *Weight gain after three weeks*
- *Factor:* *Diet and exercise*
- *Factor Levels:* *A, B, C and none (N), medium (M), intense (I)*
- *Treatments:*
  $T_1 : AN$ $T_2 : AM$ $T_3 : AI$
  $T_4 : BN$ $T_5 : BM$ $T_6 : BI$
  $T_7 : CN$ $T_8 : CM$ $T_9 : CI$
- *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 36. *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:* *81 puppies*
- *Response:* *Weight gain after three weeks*
- *Factor:* *Diet*
- *Block:* *Breed of dog*
- *Outline:*



Example 37. *Let's revisit three studies we looked at previously and determine whether each is single factor, multi(two) factor, or block design.*

(a). ***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.*

This is a single factor study. The factor is website with levels CNN and Facebook. The response is snack choice.

(b). ***Study on masks:*** *In Hies and Lewis (2022), researchers studied the effects of facial occlusion on perceived attractiveness for four different methods of occlusion. Suppose 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.*

This is a block design. The block is baseline attractiveness. The factor is facial occlusion with levels none, cloth mask, surgical mask, and notebook. The response is perceived attractiveness.

(c). ***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.*

This is a two factor design with factors being the type and level of radiation. Type of radiation has levels GSM and CDMA while level of radiation has levels of 0, 1.5, 3 or 6 W/kg. There would be 8 treatments. The response is rates of brain and heart tumors.

Example 38. *What kinds of conclusions can we make in our course investigation? Explain.*

Because our study is observational, we may only conclude association.

Chapter 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.

What is a Probability?

> **Probability** is the study of random processes. Probability is used to describe the chances of possible outcomes, but the actual outcome is not known until the random process is complete (ex. who wins coin toss a football game, weather for the day, etc.)

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 less likely the event
- The closer to 1, then more likely the event

Events with probabilities 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 39. *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?

Rolling a three and five are equally likely (1/6).

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

Rolling at least one 6 is more likely because it includes both being 6, but also many other options.

(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?

Answers will vary. Yes or no.

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

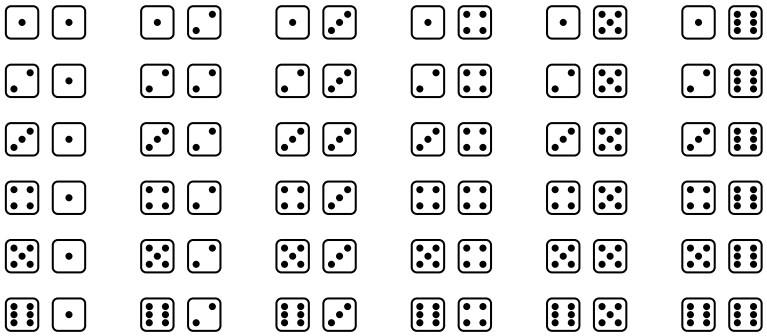
Answers will vary. Compute,

$$\textit{Relative Frequency} = \frac{\textit{Part}}{\textit{Whole}} = \frac{\textit{Number that have at least one 6}}{\textit{Number in the class}}$$

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

Empirical probability: is a probability estimated from data.

(e). Prior to rolling the pair of dice, we actually know all outcomes that are possible. The following picture displays 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 observe at least one 6 from this overall list (sample space)?

There are $6 \times 6 = 36$ possible outcomes in the sample space and 11 of those contain at least one 6 in the pair.

(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 with capital letters from 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 $P(A)$?

$P(A) = \frac{part}{whole} = \frac{11}{36} = 0.3056$

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

Theoretical probability: *Theoretical probabilities are long run probabilities assigned to events for some mathematical reason.*

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

$P(B) = \frac{1}{36}$

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

Yes, in (b) we reasoned that rolling at least one 6 is more likely than rolling two sixes. Then we computed $P(A) > P(B)$.

(i). Let C be the event that your roll a pair of numbers whose product is 13. What is $P(C)$?

From the sample space it is not possible to have a pair of numbers whose product is 13. Therefore, C is an impossible event and $P(C) = 0$.

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

The sample space allows us to see that the sum of the two numbers will ALWAYS be between 2 and 12. Therefore, D is a certain event and $P(D) = 1$.

Example 40. *Now that we have a basic understanding of empirical and theoretical probability, consider the example of rolling a single six-sided die.*

(a). Theoretical probability of rolling a 6:

$P(\textit{roll a six}) = \frac{1}{6} \approx 0.1667$

(b). If you roll a die exactly six times, does that mean you will see exactly one 6? Why or why not?

No. The roll of the die is completely random.

Example 41. *Now, suppose you flip 10 coins, and 6 of those coins come up heads.*

(a). Theoretical probability of flipping a heads:

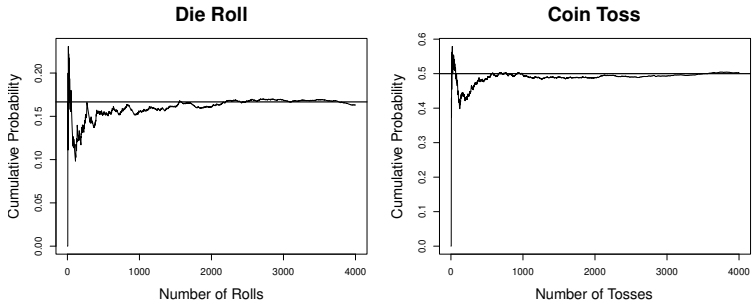
$P(\textit{Heads}) = \frac{1}{2} = 0.5$

(b). Empirical probability of flipping a heads given this data:

$P(\textit{Heads}) = \frac{6}{10} = 0.6$

(c). What do you think would happen if you flipped 100 coins? 1000?

As the number of tosses increases, the proportion of heads (empirical probability) would tend towards the theoretical probability of 1/2.



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.

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

Probability of an event: is a number between 0 and 1 that measures how likely an event is.

Use the remaining space to recap ideas presented so far as needed:
$P(A)$ is the notation for "probability event A will occur."

The empirical probability of an event can be found by simple proportions in the form of

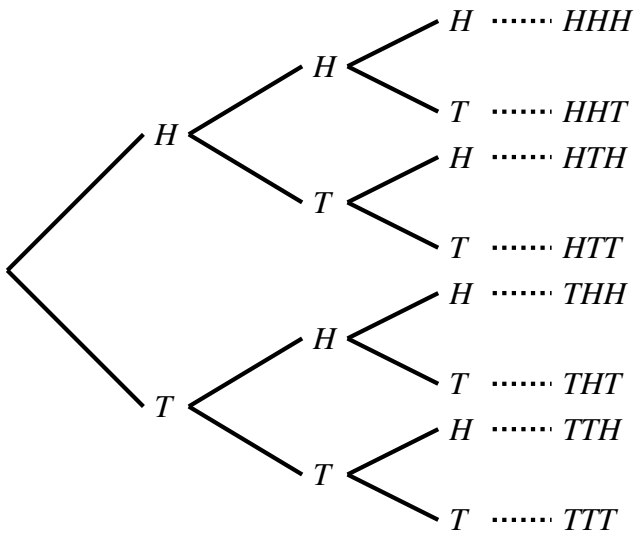
$$\frac{\text{part}}{\text{whole}}$$

A theoretical probability can be computed by adding the probabilities of the individual outcomes in the event. If the probabilities of the individual outcomes are not known, we often need an additional assumption.

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

$$P(A) = \frac{\text{Number of outcomes in A}}{\text{Number of outcomes in sample space}}$$

Example 42. *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.

$P(0 \text{ heads}) = 1/8 = 0.125$. *This would be a theoretical probability since probabilities are assigned for the mathematical reason of each outcome being equally likely rather than collected from data.*

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

$P(1 \text{ or } 2 \text{ tails}) = 6/8 = 0.75$

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

$P(\text{at least one tail}) = 7/8 = 0.875$

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

$P(2 \text{ and } 3 \text{ heads}) = 0$ *since this is impossible.*

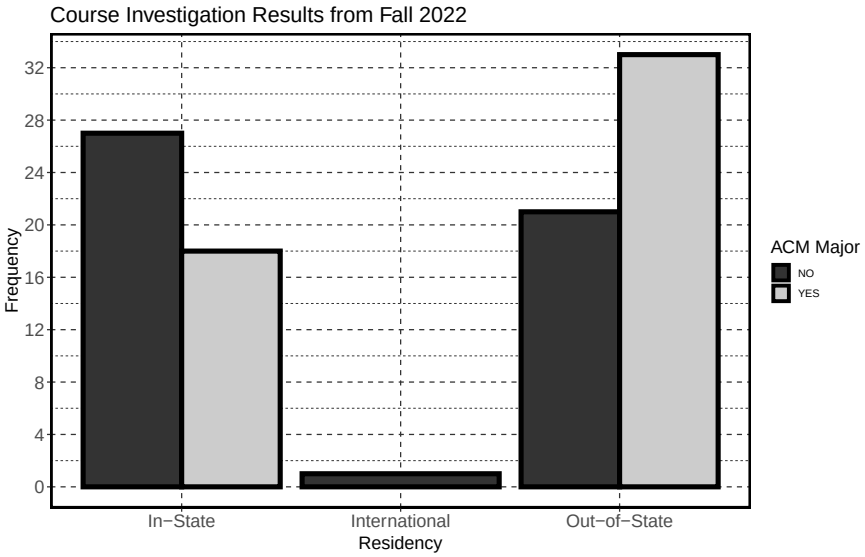
Contingency Tables

Contingency tables are a useful tool 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 intersect are referred to as cells
- the frequency of observations are provided in the corresponding cells

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

Example 43. *Previously, we considered the side-by-side bar graph comparing the residency of students to choosing a major in a program available through ACM 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?

$P(ACM) = 51/100 = 0.51$

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

$P(\textit{Out-of-state and ACM}) = 33/100 = 0.33$

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

$P(\textit{ACM given out-of-state}) = 33/54 \to 61.11\%$
You do not need to introduce conditional probabilities formally at this point. Focus on the idea that the "whole" is now different in this question.

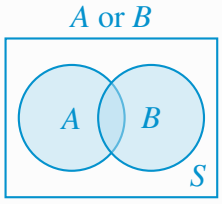
(d). Can we consider the above proportions and percentages empirical or theoretical probabilities? Explain.

These values are empirical probabilities because they are computed from data.

Special Events Related to Probability

Let's slow down and formalize the ideas we have seen thus far.

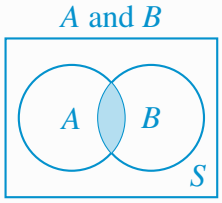
Unions: The event that *either A or B or both* occur is called the **union** of A and B.



Practical Notation: A or B

Mathematical Notation: $A \cup B$

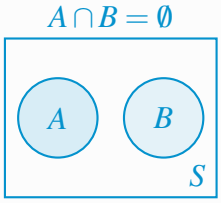
Intersections: The event that *both A and B* occur is called the **intersection** of A and B.



Practical Notation: A and B

Mathematical Notation: $A \cap B$

Disjoint events: Two events A and B are **disjoint** or **mutually exclusive** if $P(A \text{ and } B) = 0$. That is, events A and B share no common outcomes.

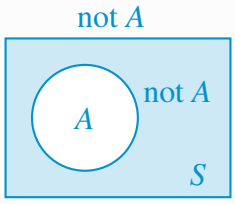


Practical Notation: $P(A \text{ and } B) = 0$

Mathematical Notation: $A \cap B = \emptyset$

Complement: The event that A does not happen is called the **complement** of A.

Note: Since A and its complement are disjoint, their union is the entire sample space.

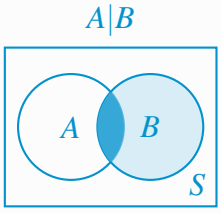


Practical Notation: not A

Mathematical Notation: A^c, A'

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.



Practical Notation: $P(A \text{ given } B)$

Mathematical Notation: $P(A|B)$

Example 44. *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.

$P(BK) = \frac{198}{490} = 0.4041$. *This is an empirical probability since it was computed from data.*

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

$P(H) = \frac{287}{490} = 0.5857$. *The fact that the value is over half suggests there is an advantage to the home team.*

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

$P(BK \text{ and } HT) = \frac{127}{490} = 0.2592$

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

Based on the previous question, around 26% of games are both basketball and won by the home team. Because there is overlap, $P(BK \text{ and } HT) \neq 0$, *the events are not disjoint.*

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

$P(BK \text{ or } HT) = \frac{198+287-127}{490} = \frac{358}{490} = 0.7306.$

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

$P(\text{not } BK) = \frac{100+93+99}{490} = 0.5959, \text{ or } 1 - P(BK) = 1 - 0.4041$

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

$P(HT \text{ given } BK) = \frac{127}{198} = 0.6414$

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

$P(BS \text{ given } HT) = \frac{53}{287} = 0.1847$

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

- $P(HT \text{ given } BK) = \frac{127}{198} = 0.6441$
- $P(HT \text{ given } HK) = \frac{50}{93} = 0.5376$
- $P(HT \text{ given } BS) = \frac{53}{100} = 0.53$
- $P(HT \text{ given } FB) = \frac{57}{99} = 0.5758$

The largest home team advantage is with basketball because they have the highest percent of home team wins.

Example 45. *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:

- $P(A \text{ or } B) =$ *4/6*
- $P(A \text{ and } B) =$ *2/6*
- $P(\text{not } A) =$ *3/6*
- $P(A \text{ or } C) =$ *5/6*
- $P(B \text{ and } C) =$ *0/6*
- $P[(\text{not } A) \text{ and } C] =$ *2/6*
- $P(A \text{ given } B) =$ *2/3*
- $P(B \text{ given } A) =$ *2/3*
- $P(A \text{ given } C) =$ *0/2*

Example 46. *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. When we say "71 percent" graduated with debt, that literally means "71 per 100". We could represent this in our table as follows. Construct a contingency table from the remaining information in a similar manner.

	Public University	Other University	Total
Debt	48	23	71
No Debt	25	4	29
Total	73	27	100

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

$P(P \text{ or } D) = \frac{73+71-48}{100} = 0.96$

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

$P(\text{not } D) = \frac{29}{100} = 0.29$

(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.

These events are not disjoint because 48% of students attended a public university AND had student loan debt. In other words, $P(P \text{ and } D) \neq 0$.

Example 47. *Suppose that Punctual Paige takes Highway 544 to school 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 highway 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 highway 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 highway 544?

Knowing there is a wreck on highway 544 will most likely decrease her chances of being on time for class. That is, the probability of Paige being on time DEPENDS on whether or not there is a wreck.

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. In terms of probabilities:

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

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.

It is worth noting that two events A and B depending on one another does not mean that A forces B to be true, or vice-versa. It is still possible for one to happen without the other. What it does mean, however, is that knowledge of A will change the probability that B occurs, or vice-versa.

Example 48. *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. For the total number of days, consider using a power of 10 like 100, 1000, or 10,000.

	Wreck on 544	No wreck on 544	Total
Paige on time	41	949	990
Paige not on time	9	1	10
Total	50	950	1,000

The students seem to like keeping the total as 100 ("per cent") and do not mind decimals in their table.

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

Use this to help students fill in the table in (a).
$P(\text{on time AND wreck}) = \frac{82\% \text{ of } 50}{1000} = \frac{0.82 \times 50}{1000} = 0.041$

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

$P(\text{on time}) = 0.99$. *However, if there is a wreck, the chances of Paige being on time changes,* $P(\text{on time GIVEN wreck}) = 0.82$. *Therefore, the two events are dependent.*

(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?

$$P(\textit{wreck GIVEN not on time}) = \frac{9}{10} = 0.9$$

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

$$P(\textit{not a wreck GIVEN not on time}) = \frac{1}{10} = 0.1$$

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

From part (a) we see that Paige is on time and there is a wreck on highway 544 around 4.1% of the time. Therefore, the events are not disjoint since $P(\textit{on time AND wreck}) \neq 0$.

Additional Examples

Depending on time, select some of the problems in this section to give students more practice in areas it is needed. Solutions to the remaining can be posted as extra practice.

Example 49. *From our course investigation, 51% of students in introductory statistics 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 introductory statistics 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.

	$\geq$ **300 miles**	$<$ **300 miles**	***Total***
ACM	*29*	*22*	*51*
No ACM	*18*	*31*	*49*
Total	*47*	*53*	*100*

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

$P(no\ ACM) = 49/100 = 0.49$

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

$P(ACM\ and < 300) = 22/100 = 0.22$

(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?

$P(ACM \text{ given} \geq 300) = 29/47 = 0.617$

(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.

The events are not disjoint because 29% of students chose a major in ACM and live at least 300 miles from campus. That is, $P(ACM \text{ and} \geq 300) \neq 0$.

(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.

$P(ACM) = 0.51$ *while* $P(ACM \text{ given} \geq 300) = 0.617$. *Therefore, knowing a student comes from further away increases the probability that they will be participating in ACM. The events are then dependent because the probability of participating in ACM depends on the distance from home.*

Example 50. *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.*

	B	***Not B***	***Total***
A	*5 (50% of 10)*	*25*	*30*
Not A	*5*	*65*	*70*
Total	*10*	*90*	***100***

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

$P(A \text{ and } B) = 5/100 = 0.05$

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

$P(A \text{ or } B) = \frac{30+10-5}{100} = 0.35$

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

$P(\text{not } B) = 90/100 = 0.90$

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

$P(\text{not } A \text{ and not } B) = 65/100 = 0.65$

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

$P(A) = 0.30$ *while* $P(A \text{ given } B) = 0.50$. *Therefore, the events are dependent because the probability of symptom A depends on (changes based on) the presence of symptom B.*

Example 51. *For independent events A and B, suppose that* $P(A) = 0.5$ *and* $P(B) = 0.4$. *Using this information, compute the following probabilities.*

	B	***Not B***	***Total***
	(40% of 50) ↓		
A	*20*	*30*	*50*
	(50% of 40) ↑		
Not A	*20*	*30*	*50*
Total	*40*	*60*	*100*

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

$P(A \text{ given } B) = 20/40 = 0.5 = P(A)$, *since A and B are independent*

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

$P(A \text{ and } B) = 20/100 = 0.20$

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

$P(A \text{ or } B) = \frac{50+40-20}{100} = 0.70$

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

$P(B \text{ given } A) = 20/50 = 0.40 = P(B)$, *see part (a)*

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

$P(A \text{ and } B) = 20/100 = 0.20 \neq 0$, *therefore the events are not disjoint.*

Example 52. *For disjoint events A and B, suppose that* $P(A) = 0.3$ *and* $P(B) = 0.2$*. Using this information, compute the following probabilities.*

	B	***Not B***	***Total***
A	*0*	*30*	*30*
Not A	*20*	*50*	*70*
Total	*20*	*80*	*100*

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

$P(A \text{ and } B) = 0/100 = 0$, *since A and B are disjoint*

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

$P(A \text{ given } B) = 0/20 = 0$

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

$$P(A \text{ or } B) = \frac{20 + 30 - 0}{100} = 0.50$$

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

$P(A) = 30/100 = 0.30$ *while* $P(A \text{ given } B) = 0/20 = 0$, *therefore the events are dependent. The probability of A occurring changes if we know B occurred.*

2.2 Discrete Distributions

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:

- Discrete - values take distinct points, often "counting"
  Examples: # of siblings, # of credits taken
- Continuous - values are anywhere within an interval, often "measuring"
  Examples: time, distance, chemical concentration

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 one

Example 53. *Give the probability distribution that describes rolling a fair die once.*

Let the random variable $X =$ # of dots

Probability distribution of X:

x	1	2	3	4	5	6
$P(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

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

$$\mu = \sum xP(x)$$

NOTE:

- Each outcome is weighted by its probability
- The value μ is a theoretical value not a sample value like $\bar{x}$.

Example 54. *Find the expected value when rolling a fair die once.*

$$\mu = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$$

If we roll a fair die many times, we expect the average face value to be 3.5. Note that we do not round averages.

Example 55. *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 2.1: 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 2.2: 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?

$$P(X=1) = 1 - P(X \neq 1) = 1 - (0.275 + 0.242 + 0.064 + 0.013) = 0.406$$

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

$$P(X_G \geq 3) = 0.242 + 0.064 + 0.013 = 0.319$$

$$P(X_A \geq 3) = 0.197 + 0.031 + 0.004 = 0.232$$

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

It is nice to have students think about a reasonable answer before computing. For example, they may observe that 1, 2, and 3 have the highest probabilities in this case so the average will be close to 2.

$$\begin{aligned}\mu_G &= 1(0.406)+2(0.275)+3(0.242)+4(0.064)+5(0.013)\\ &= 2.003\end{aligned}$$

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

$$\begin{aligned}\mu_A &= 1(0.513)+2(0.255)+3(0.197)+4(0.031)+5(0.004)\\ &= 1.758\end{aligned}$$

(e). Which region typically experiences the worse hurricanes? How do you know?

Parts (b), (c), and (d) indicate that hurricanes are typically worse in the gulf. Explanations may vary: students may discuss which region has a higher average/expected value or talk about the probabilities.

Binomial Distribution

The **binomial distribution** is a special kind of **discrete** distribution. It must meet all four of the following requirements (BINS):
To motivate the setting, consider tossing a coin 10 times and let X denote the number of heads.

1. **B**inary Outcomes (success and failure)
   ex. heads/tails
2. **I**ndependent Trials (each trial is not influenced by the others)
   ex. coin tosses are independent, the use of random sampling
3. **N**umber of trials is fixed ahead of time (n)
   ex. $n = 10$
4. **S**ame probability of success each trial (p)
   ex. $p = 0.5$

Shorthand notation for Binomial random variables:

$$X \sim \text{Binomial}(n, p) \rightarrow X \sim \text{Binomial}(10, 0.5)$$

Example 56. *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

Binary Outcomes - Roll snake eyes or not
Indepedent trials - Rolls of the dice are independent
Number of trials fixed - $n = 10$
Same probability of success - $p = 1/36$
Therefore, $X \sim$ *Binomial*$(10, 1/36)$

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

This is not binomial because the number of trials is not defined.

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

This is not binomial because the trials are not independent.

Mean and standard deviation of a Binomial variable:

$$\mu = np \text{ and } \sigma = \sqrt{np(1-p)}$$

Example 57. *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.

$$X \sim Binomial(n = 10, p = 0.90)$$

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

You can ask students, "If a person makes 90% of their shots and takes 10 shots, how many do you expect them to make?" They immediately say, "9 out of 10" and have gained an intuitive understanding of expected value in this setting rather than simply relying on the formula.

$$\mu = 10(0.9) = 9 \text{ and } \sigma = \sqrt{10(0.9)(0.1)} = 0.9487$$

Example 58. *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.

Binary Outcomes - tattoo or no tattoo
Independent trials - random sample
Number of trials fixed - $n = 20$
Same probability of success - $p = 0.26$

Therefore, $X \sim Binomial(20, 0.26)$

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

$$\mu = 20(0.26) = 5.2 \text{ and } \sigma = \sqrt{20(0.26)(1-0.26)} = 1.9616$$

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.

$$P(X = 10) = binompdf(20, 0.26, 10) = 0.0128$$

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

$$P(X = 5) = binompdf(20, 0.26, 5) = 0.2013$$

To compute the probability that X is less than or equal to k, $P(X \leq k)$, 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.

$$P(X \leq 10) = \texttt{binomcdf(20,0.26,10)} = 0.9945$$

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

$$P(X < 10) = P(X \leq 9) = \texttt{binomcdf(20,0.26,9)} = 0.9817$$

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

$$\begin{aligned} P(X > 5) &= 1 - P(X \leq 5) \\ &= 1 - \texttt{binomcdf(20,0.26,5)} \\ &= 0.4235 \end{aligned}$$

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 59. *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?

$\mu = 50(0.058) = 2.9$ *dentists*

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

$P(X \leq 4) = \texttt{binomcdf(50,0.058,4)} = 0.8372$

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

$$\begin{aligned} P(X > 4) &= 1 - P(X \leq 4) \\ &= 1 - \texttt{binomcdf(50,0.058,4)} \\ &= 0.1628 \end{aligned}$$

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

$P(X = 4) = \texttt{binompdf(50,0.058,4)} = 0.1669$

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

$$\begin{aligned} P(X \geq 12) &= 1 - P(X \leq 11) \\ &= 1 - \texttt{binomcdf(50,0.058,11)} \\ &= 2.206 \times 10^{-5} \end{aligned}$$

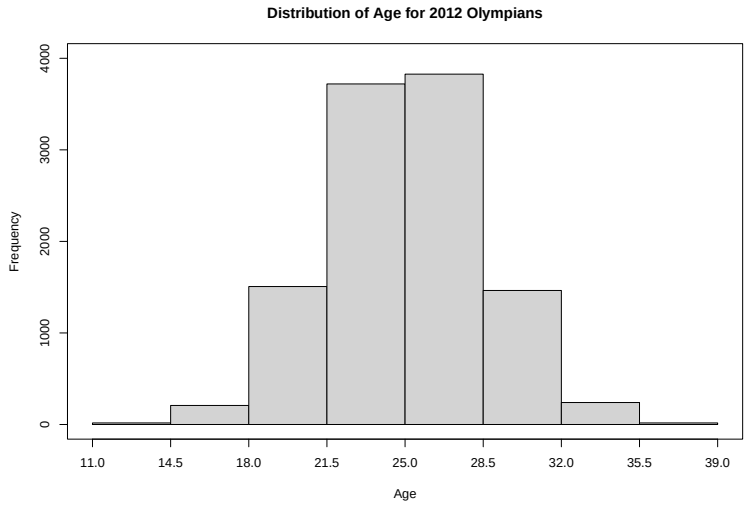
Remind your students of scientific notation since the calculator will report `2.206e-5` *which is also 0.00002206.*

2.3 Normal Distribution

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 60. *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?

The shape is symmetric.

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

Based on the graph, around 200 athletes are less than 18 years old. Thus, the proportion is $200/11000 = 0.018$.

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

Based on the graph, around 3700 athletes are between 21.5 and 25 years while around 3800 athletes are between 25 and 28.5 years. So, the percent is approximately $(3700 + 3800)/11000 \times 100 = 68.18\%$.

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

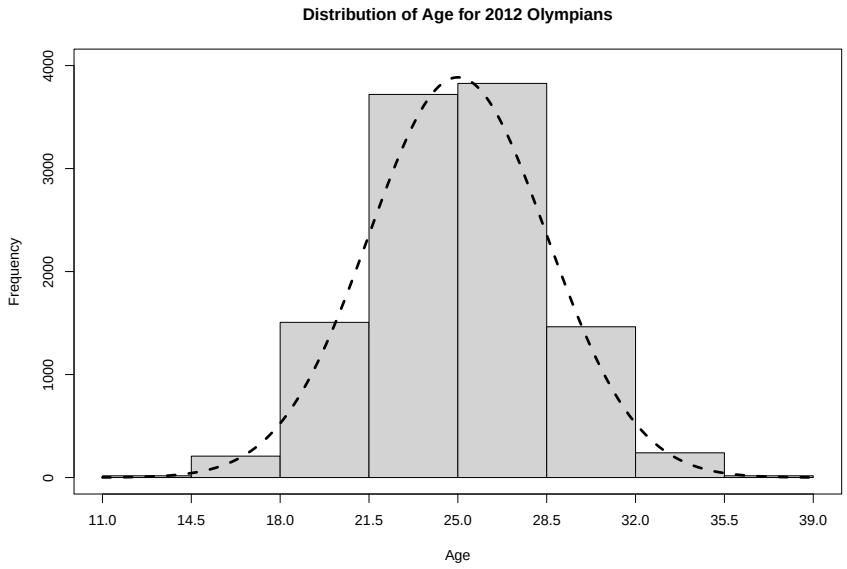
We cannot find probabilities for exact ages. The histogram only provides frequencies for different ranges of age.

(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.

Using this graph, we cannot find an approximate percentage for ages less than 16.5 since 16.5 is not one of our "bounds" or "tick marks". However, this is possible if we have a mathematical model for the distribution.

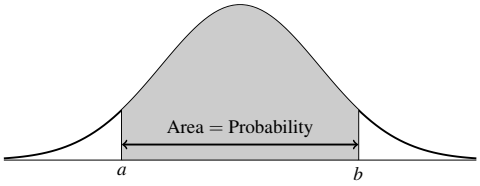
In the previous example, would could only approximate percentages (probabilities) for ranges of values referring to specific ages outlined by the histogram. As considered in part (e), it is possible to find probabilities for any range of ages.

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 a 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.

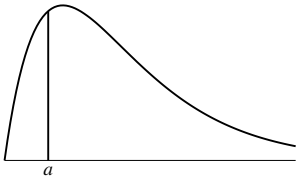


For a continuous distribution:

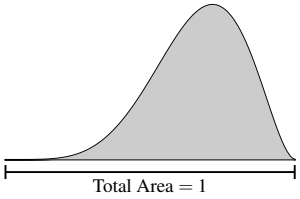
- Probabilities are the same as area under the curve.



- $P(X = a) = 0$ for any value a.



- The total area under the curve is 1.

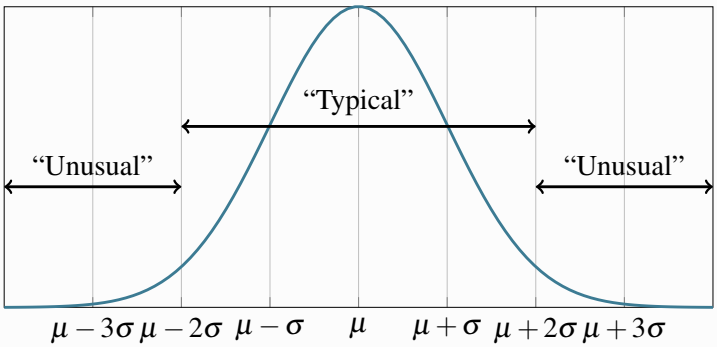


Continuous distributions can be used to model any continuous random variable such as time, salary, and length. We will see later that many discrete distributions may also be approximated by a continuous curve.

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

μ = mean and σ = standard deviation



Notation: $X \sim N(\mu, \sigma)$

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.

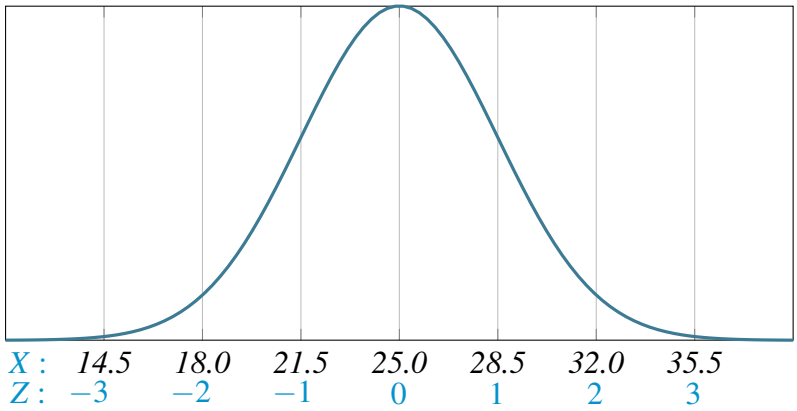
z-scores measure the distance of an observation from the mean, in terms of standard deviations. Positive z-scores indicate the observation is above average and negative z-scores indicate an observation is below average.

$$Z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}}$$

z-scores have no units. We consider z-scores beyond ± 2 to be unusual, and z-scores beyond ± 3 to be highly unusual.

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 61. *For the example considering the ages of 2012 Olympic athletes, the ages of athletes 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?

Athletes within 2 standard deviations from the mean are considered "typical". Thus, athletes between 18 and 32 years old are considered "typical" for the 2012 Olympics.

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

Althletes beyond 2 standard deviations from the mean are considered "unusual". Thus, olympic athletes younger than 18 and older than 32 are considered "unusual".

Example 62. *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?*

Gymnasts: $\mu = 23$; $\sigma = 3$

$$Z_I = \frac{39-23}{3} = 5.3333$$

Equestrian: $\mu = 40$; $\sigma = 6.5$

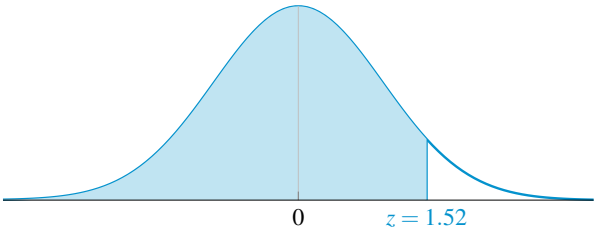
$$Z_H = \frac{71-40}{6.5} = 4.7649$$

The ages for both athletes were highly unusual relative to their sport. However, the gymnast is older relative to their sport due to a higher z-score.

The Standard Normal Distribution on the Calculator:

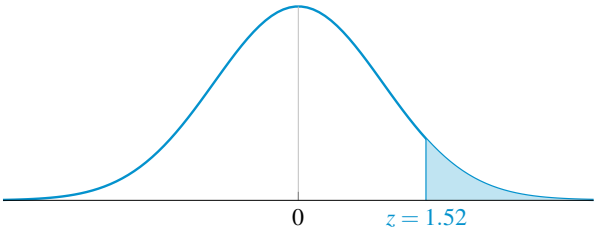
To find the area (probability) between two z-scores use `normalcdf(lwr,uppr)`. There is no infinity (∞) button on the calculator. In this application you may use some large number such as 9999.

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



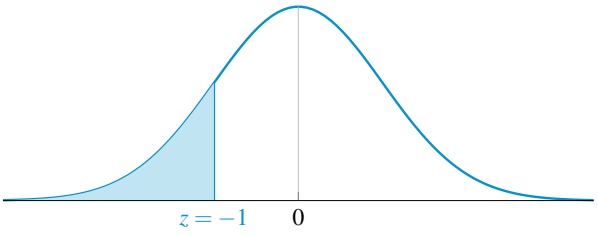
$P(Z < 1.52) =$ `normalcdf(-9999,1.52)` $= 0.9357$

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



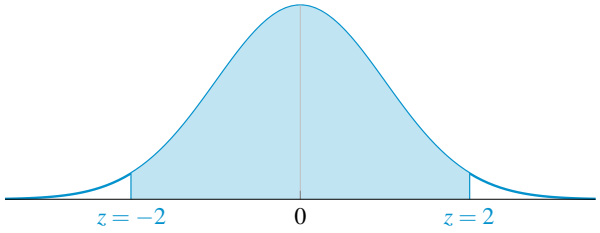
$P(Z > 1.52) =$ `normalcdf(1.52,9999)` $= 0.0643$

(c). Find the area below z = -1.



$P(Z < -1) =$ `normalcdf(-9999,-1)` $= 0.1587$

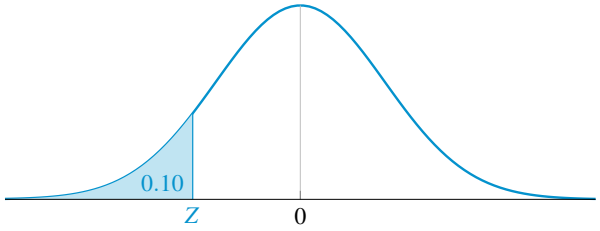
(d). Find the area between z = -2 and z = 2.



$P(-2 < Z < 2) = \texttt{normalcdf(-2,2)} = 0.9545$

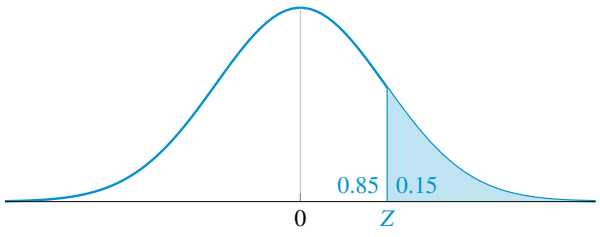
To find the z-score with a particular area below (percentile) use `invNorm(left area)`. Some calculators have the option to change the location of the area, but we assume the area is always on the left.

(a). Find the 10th percentile.



$Z = \texttt{invNorm(0.10)} = -1.282$

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



$Z = \texttt{invNorm(0.85)} = 1.036$

Probability and Percentile Computations with General Settings

Example 63. *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.

$$\begin{aligned} P(X < 30) &= P\left(Z < \frac{30-25}{3.5}\right) \\ &= P(Z < 1.4286) \\ &= \texttt{normalcdf(-9999,1.4286)} \\ &= 0.9234 \end{aligned}$$

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

$$\begin{aligned} P(20 < X < 25) &= P\left(\frac{20-25}{3.5} < Z < \frac{25-25}{3.5}\right) \\ &= P(-1.4286 < Z < 0) \\ &= \texttt{normalcdf(-1.4286,0)} \\ &= 0.4234 \end{aligned}$$

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

First, we need to compute the 97.5th percentile for Z:

$$Z = \texttt{invNorm(0.975)} = 1.960$$

Second, we need to solve the z-score equation for age. Note that

$$Z = \frac{X-\mu}{\sigma} \implies X = Z\sigma + \mu$$

Thus, we have

$$X = 1.960(3.5) + 25 = 31.86 \text{ years}$$

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

First, we need to compute the 70th percentile for Z (since we are given that 30% are above):

$$Z = \texttt{invNorm(0.70)} = 0.5244$$

Second, we need to compute the 70th percentile for age:

$$X = Z\sigma + \mu = 0.5244(3.5) + 25 = 26.8354 \text{ years}$$

Example 64. *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?

$$\begin{aligned} P(X > 15) &= P\left(Z > \frac{15-18}{5}\right) \\ &= P(Z > -0.6) \\ &= \texttt{normalcdf(-0.6,9999)} \\ &= 0.7257 \end{aligned}$$

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

First, we need to compute the 60th percentile for Z:

$$Z = \texttt{invNorm(0.60)} = 0.2533$$

Second, we need to compute the 60th percentile for age:

$$X = Z\sigma + \mu = 0.2533(5) + 18 = 19.2665 \text{ minutes}$$

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

$$\begin{aligned} P(15 < X < 30) &= P\left(\frac{15-18}{5} < Z < \frac{30-18}{5}\right) \\ &= P(-0.6 < Z < 2.4) \\ &= \texttt{normalcdf(-0.6,2.4)} \\ &= 0.7175 \end{aligned}$$

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

Note that

$$Z = \frac{10-18}{5} = -1.6$$

Thus, 10 minutes is 1.6 standard deviations below the mean. Since it is not beyond 2 standard deviations from the mean, this value would not be unusual.

Chapter 3

Statistical Inference for Proportions

3.1 Sampling Distribution of the Proportion

Introductory Activity Consider several standard decks of cars all shuffled together to simulate sampling with replacement.

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.
   - B - binary outcomes - each card is a heart or not
   - I - independent trials - reasonable due to random sampling (with replacement)
   - N - n fixed trials - $n = 5$ and is fixed prior
   - S - same probability - $p = 0.25$ and is the same for each card

   So, $Y \sim \text{Binomial}(5, 0.25)$.

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?

   Using the binomial distribution, we have

   $$\mu = 5(0.25) = 1.25 \qquad \sigma = \sqrt{5(0.25)(1-0.25)} = 1.25$$

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?

   The instructor will likely need to shuffle several decks of cards together to ensure every student is given 5 cards at that $p \approx 0.25$. Each student should observe either 0, 1, 2, 3, 4, or 5 hearts within their sample. Suppose a student observes 2 hearts. Then they would need to perform the following calculation:

   $$P(Y = 2) = \texttt{binompdf(5,0.25,2)} = 0.2637$$

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

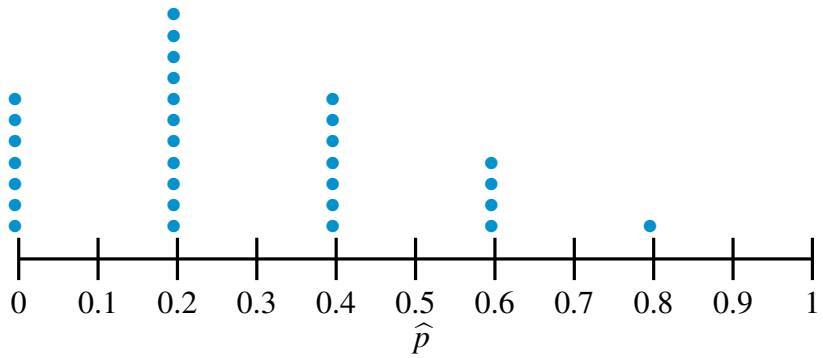
   Again, each student should observe either 0, 1, 2, 3, 4, or 5 hearts within their sample. Suppose a student observes 2 hearts. Then they would need to perform the following calculation:

   $$P(Y \leq 2) = \texttt{binomcdf(5,0.25,2)} = 0.8965$$

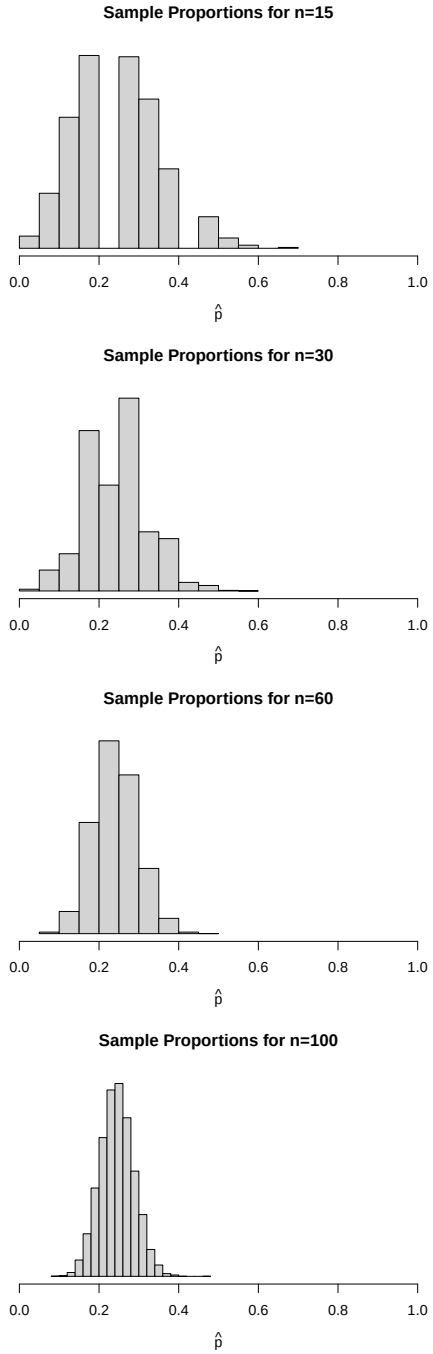
5. The sample proportion $\hat{p}$ is the fraction of the sample that are hearts; that is, $\hat{p} = y/5$. Calculate *your* value of $\hat{p}$.

   If a student observed $Y = 2$, $\widehat{p} = 2/5 = 0.4$

6. Add your value of $\widehat{p}$ to the class plot on the board. Record your answer along with your classmates' on the axis below. The following plot was created from a simulated class of 30 students.



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?

   The center of the previous plots is roughly the same at around $p = 0.25$.

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

   As the sample size increases, the variability of each distribution decreases. We measure the variability of the statistic with the **Standard Error** (SE).

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

   As the sample size increases, the distribution becomes more continuous. Also, the shape shape of the distribution becomes more symmetric (normal) as the sample size increases.

Formal Result and Examples

The Sampling Distribution of the Sample Proportion

Notation check:

- $p =$ the proportion of successes from the **population**
- $\widehat{p} =$ the proportion of successes from the **sample**

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,

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

The sample size n is considered "large" enough if both $np \geq 10$ and $n(1-p) \geq 10$. That is, we can expect at least 10 success and 10 failures in our sample.

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

    No since $np = 5(0.25) = 1.25$ and $n(1-p) = 5(1-0.25) = 3.75$ are both less than 10.

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.

    Since $100(0.25) = 25$ and $100(1-0.25) = 75$ are both greater than 10, the sampling distribution result for $\widehat{p}$ does apply. Thus,

$$\widehat{p} \sim N\left(0.25, \sqrt{\frac{0.25(1-0.25)}{100}}\right) \implies \widehat{p} \sim N(0.25, 0.0433)$$

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?

$$\begin{aligned} P(\widehat{p} \leq 0.4) &= P\left(Z \leq \frac{0.4-0.25}{0.0433}\right) \\ &= P(Z < 3.464) \\ &= \texttt{normalcdf(-9999,3.464)} = 0.9997 \end{aligned}$$

    By looking at the bottom right plot on Page 105, we can clearly see that nearly all of the samples of 100 yielded a $\widehat{p}$ less than 0.4. So, this probability seems reasonable.

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.

    Now, $X \sim \text{Binomial}(100, 0.25)$. Thus,
    $P(X \leq 40) = \texttt{binomcdf(100,0.25,40)} = 0.9997$

Example 65. *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.

$$\widehat{p} \sim N\left(0.28, \sqrt{\frac{0.28(1-0.28)}{200}}\right) \implies \widehat{p} \sim N(0.28, 0.0317)$$

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

$$\begin{aligned} P(\widehat{p} \geq 0.35) &= P\left(Z \geq \frac{0.35-0.28}{0.0317}\right) = P(Z \geq 2.208) \\ &= \texttt{normalcdf(2.208,9999)} = 0.0136 \end{aligned}$$

(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?

$$\begin{aligned} P(0.3 < \widehat{p} < 0.4) &= P\left(\frac{0.3-0.28}{0.0317} < Z < \frac{0.4-0.28}{0.0317}\right) \\ &= P(0.6309 < Z < 3.7855) \\ &= \texttt{normalcdf(0.6309,3.7855)} = 0.2640 \end{aligned}$$

(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.

$$\widehat{p} \sim N\left(0.28, \sqrt{\frac{0.28(1-0.28)}{300}}\right) \implies \widehat{p} \sim N(0.28, 0.0259)$$

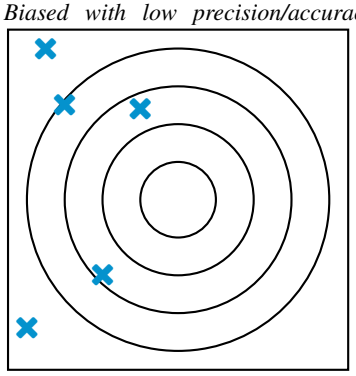
We take many samples of $n = 300$ accidents and compute $\widehat{p} =$ sample proportion of accidents caused by phones. Considering all the different values of $\widehat{p}$, we will see that they follow a Normal distribution centered at 0.28 with a standard deviation of 0.0259.

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.

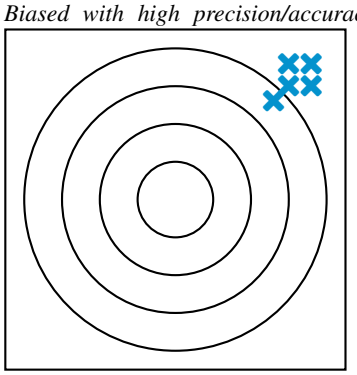
How much a statistic varies around the parameter can be described by **precision** or **accuracy**. The less a statistic varies around the parameter the more precise, or accurate, the estimator. We have seen that the precision of the $\widehat{p}$'s increases as the sample size increases.

Example 66. *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 (e.g. $\widehat{p}$). Sketch an example of each of the following settings:*

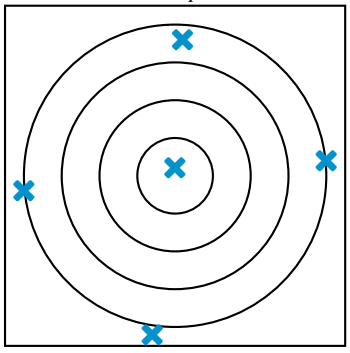
1. *Biased with low precision/accuracy*



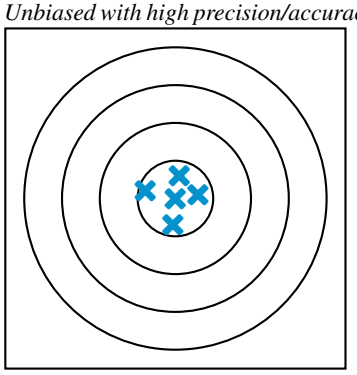
3. *Biased with high precision/accuracy*



2. *Unbiased with low precision/accuracy*



4. *Unbiased with high precision/accuracy*



Which setting is ideal?

Unbiased with high accuracy

3.2 Confidence Intervals for One Proportion

Introduction

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

Suppose we want to estimate the population proportion. What value from our sample could we use? Why is this number alone insufficient?

- We could use $\widehat{p}$, the sample proportion, to estimate p, the population proportion. (Remember, statistics estimate parameters.)
- We call statistics **point estimates** (PE) because we estimate a parameter with a single value.
- Using a point estimate (like $\widehat{p}$) alone is not good enough because different samples will give different estimates. We measured this expected variability with SE.

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:

Basic Form of a CI:

PE $\pm$ multiplier $\times$ SE

- PE = point estimate (our statistic)
- SE = standard error (the standard deviation of the point estimate/statistic)
- The **multiplier** is the # of SEs to allow for and is chosen to achieve a desired level of confidence
- The **margin of error** (MOE) is the product of the multiplier and the SE

Confidence intervals for a single proportion (p)

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

Example 67. *Consider our course investigation. Out of a random sample of 100 students, 45% were classified as out-of-state students.*

(a). What is the point estimate (PE) for the proportion of all students at the university who are out-of-state? What do we use this value to estimate?

From the data, the point estimate is $PE = \widehat{p} = 0.45$. *This statistic from the sample can be used to estimate the parameter,* p *from the population.*

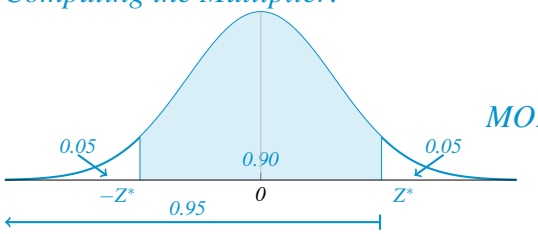
(b). If we take a different sample of $n = 100$ *students, will we get the same point estimate? If not, by how much do we expect it to vary? What do we call this measurement of expected variability in a statistic?*

No, a different sample is expected to have a different estimate due to the nature of random sampling. We expect it to vary by the estimated standard error which is:

$$SE = \sqrt{\frac{0.45(1-0.45)}{100}} = 0.0497$$

(c). Find the multiplier for a 90% confidence interval and calculate the margin of error.

Computing the Multiplier:



$$\begin{aligned} MOE &= mult. \times SE \\ &= 1.645(0.0497) \\ &= 0.0818 \end{aligned}$$

$$\begin{aligned} Z^* &= |\texttt{invNorm(0.05)}| \\ &= \texttt{invNorm(0.95)} \\ &= 1.645 \end{aligned}$$

(d). Construct a 90% confidence interval for the proportion of all students at the university who are out-of-state students.

CI: PE $\pm$ *MOE* $\rightarrow (0.45 \pm 0.0818 \longrightarrow (0.3682, 0.5318)$

(e). What is this interval estimating?

This interval is estimating the population proportion, i.e. the parameter p. Note that the statistic from the sample will always be in the middle of the interval.

(f). Interpret the interval in context.

We are 90% confident the ***proportion of all students*** *at the university who are out-of-state is between 0.3682 and 0.5318.*

(g). According to the spring 2021 demographic report, 47.3% were actually out-of-state at the university ($p = 0.473$*). Does our interval contain this value? If not, discuss reasons why not.*

The value 47.3% is within the 90% confidence interval limits. If the value had not been in our interval, it could be because a sample of introductory statistics students is not representative of the entire university.

Example 68. *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?

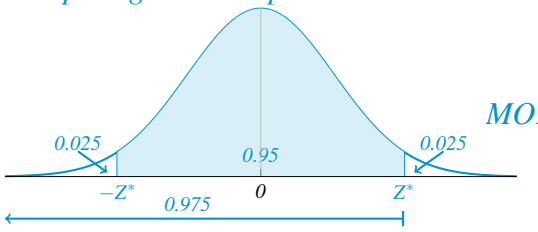
$$PE: \widehat{p} = \frac{640}{1000} = 0.64$$

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

$$SE = \sqrt{\frac{0.64(1-0.64)}{1000}} = 0.0152$$

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

Computing the Multiplier:



$$\begin{aligned} MOE &= mult. \times SE \\ &= 1.960(0.0152) \\ &= 0.0298 \end{aligned}$$

$$\begin{aligned} Z^* &= |\texttt{invNorm(0.025)}| \\ &= \texttt{invNorm(0.975)} \\ &= 1.960 \end{aligned}$$

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

$$CI \pm MOE \longrightarrow 0.64 \pm 0.028 \longrightarrow (0.6102, 0.6698)$$

(e). Interpret the interval.

We are 95% confident that between 61.02% and 66.98% of Americans enjoy April Fool's Day.

(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.

No, since 50% is not within the limits of our confidence interval.

Example 69. *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.*

- *PE:* $\widehat{p} = 0.77$
- *SE:* $\sqrt{\dfrac{0.77(1-0.77)}{2000}} = 0.0094$
- *Mult.:* $Z^* = \texttt{invNorm}(0.975) = 2.576$

$MOE = 2.576(0.0094) = 0.0242$

CI: $0.77 \pm 0.0242 \longrightarrow (0.7458, 0.7942)$

We are 99% confident that between 74.58% and 79.42% of all American adults are outraged by lack of federal family leave laws.

Wrap it up

Summary of most common z*'s:

- 90% Confidence $\implies Z^* = \texttt{invNorm}(0.95) = 1.645$
- 95% Confidence $\implies Z^* = \texttt{invNorm}(0.975) = 1.96$
- 99% Confidence $\implies Z^* = \texttt{invNorm}(0.995) = 2.576$

Assumptions for the confidence interval to be valid:

1. The data come from a random sample.
2. There are at least 10 successes and 10 failures observed.

→ Are the intervals from our three examples all valid? Yes!

Q: What affects the width of the confidence interval?
MOE = mult × SE

1. **Level of Confidence:** As the confidence level increases, the width of the CI increase.

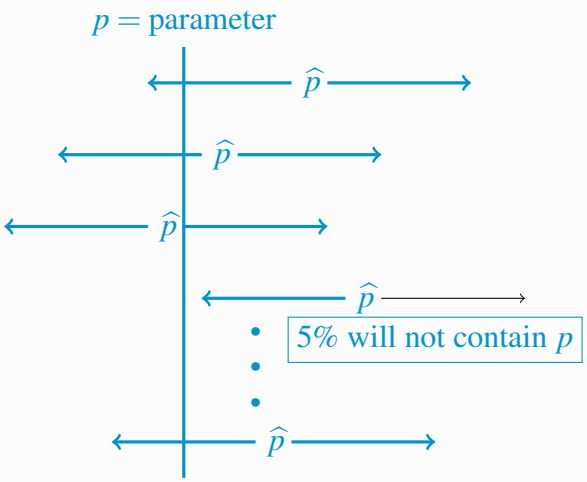
   NOTE: This is because Z^* increases to allow more area. As a visual, you could draw a big circle on the board and say you are 90% confidence that you will throw a ball from 10 feet and hit somewhere within the circle. Then draw a bigger circle and ask the students if you are more or less confident to throw a ball and hit somewhere within.

2. **Sample Size:** As n increases, the width of the CI decreases.

   NOTE: This is because n is in the denominator of the SE, Thus as n increases, SE decreases. Thus, the MOE decreases.

Interpretation of the confidence level

Suppose our level of confidence is 95%. This means that 95% of all samples will yield confidence intervals that contain the true parameter and 5% will not.



3.3 Hypothesis Test for One Proportion

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. As an example, we will use green as our 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.

   H_0 : the proportion of green stones from the population is 0.5 (50%)

   $H_0 : p = 0.5$

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

   As an introductory example, let's consider a one-sided alternative.

   H_a : the proportion of green stones from the population is less than half

   $H_a : p < 0.5$

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

   Results will vary from bag to bag. Let the class pass around the bag with each student drawing out a stone.

   Suppose that the bag of gifted stones contains $n = 40$ with 6 green. Then, $\widehat{p} = 6/40 = 0.15$

4. Do you still believe the null hypothesis?

   Our statistic ($\widehat{p} = 0.15$) does seem far below the hypothesized value of $p = 0.5$. Remind our students that we must be careful about looking at simple differences since the statistics will almost never equal the hypothesized value. There is always random variation that is expected. Is this within what is expected or not?

5. Suppose that the null hypothesis is true. Does the sampling distribution of the sample proportion apply to the random sample we took? State the distribution of $\widehat{p}$.

   If $H_0 : p = 0.5$ is true, we have that $np = 40(0.5) = 20$ and $n(1-p) = 40(0.5) = 20$. Thus, normal distribution does apply. So,

$$\widehat{p} \sim N\left(0.5, \sqrt{\frac{0.5(1-0.5)}{40}}\right) \text{ or } \widehat{p} \sim N(0.5, 0.0791)$$

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**.

$$Z_c = \frac{0.15 - 0.5}{0.0791} = -4.4248$$

   Thus, our statistic is 4.4248 standard deviations (standard errors) below the hypothesized value of 0.5. Notice, that this is in the highly unusual range for z-scores.

7. Do you still believe the null hypothesis?

   Since our z-score is in the highly unusual range, either our data (statistic) is just rare or unusual, or, our hypothesized value is incorrect. So, there is a strong possibility that our null hypothesis is not correct. Let's see if there is a way we can find the possibility or probability of this.

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.

   Assuming that $H_0 : p = 0.5$ is true, we can compute the probability of getting our value of $\widehat{p}$ or one "more extreme" as follows:

$$\begin{aligned} P(\widehat{p} \leq 0.15) &= P(Z \leq -4.4248) \\ &= \texttt{normalcdf}(-9999, -4.4248) \\ &= 4.8266 \times 10^{-6} \end{aligned}$$

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

   If 50% of the stones are truly green, there is a 0.00048% chance of observing data/results like the one we observed or more extreme.

10. Do you still believe the null hypothesis?

    Since the p-value is really small, it is most likely that our null hypotheses is not true. Thus, our evidence from our data is more in favor of the alternative.

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

    There is strong evidence that the true proportion of green stones (from the population) is less than 50%.

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

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.

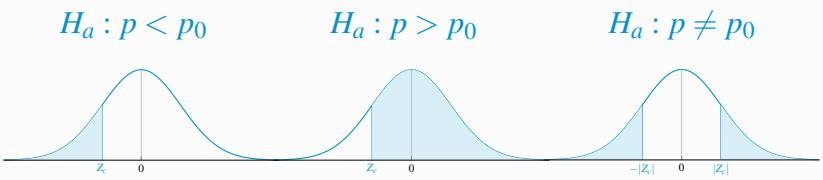
Things to remember:

- 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:	For proportions:
$TS = \dfrac{PE - H_0}{SE}$	$z_c = \dfrac{\widehat{p} - p_0}{SE}$; $SE = \sqrt{\dfrac{p_0(1-p_0)}{n}}$

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.

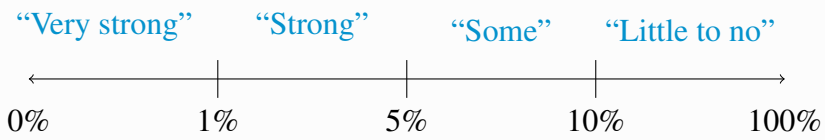


normalcdf$(-9999, Z_c)$ normalcdf$(Z_c, 9999)$ $2 *$ normalcdf$(|Z_c|, 9999)$

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 < 0.01 (less than 1%) $\rightarrow$ "There is **very strong** evidence in favor of the alternative hypothesis"
- p-value > 0.1 (more than 10%) $\rightarrow$ "There is **little to no** evidence in favor of the alternative hypothesis"
- $0.05 <$ p-value < 0.1 (between 5% and 10%) $\rightarrow$ "There is **some evidence** in favor of the alternative hypothesis"
- $0.01 <$ p-value < 0.05 (between 1% and 5%) $\rightarrow$ "There is **strong evidence** in favor of the alternative hypothesis"



Example 70. *Consider our course investigation. According to fall 2021 enrollment data, 47% of all undergraduate students at the university are in-state students. We wish to determine if the the current proportion of in-state students at the university has changed. State the hypotheses of the test.*

$H_0 : p = 0.47$ *versus* $H_a : p \neq 0.47$

From our random sample of 100 students taking introductory statistics in fall 2022, 45% were in-state students. We know a different sample would give a different statistic. If H_O is true, by how much does the statistic ($\widehat{p}$) vary?

$SE = \sqrt{\frac{0.47(1-0.47)}{100}} = 0.0499$

Based on the data and expected variation in the statistic, does it seem like the proportion of out-of-state students has changed from 47%? Explain.

Our sample had 45% in-state students. Under our initial assumption $H_O : p = 0.47$, we expect 47% in-state with estimates that vary by about 5%. Our observed data is within one standard error, so there is likely no difference.

Conduct the test to confirm.

- ***Test Statistic:*** $Z_c = \frac{0.45-0.47}{0.0499} = -0.4008$
- ***Compute the p-value:*** *Since* $H_a : p \neq 0.47$*, we have*

$$\begin{aligned} p\text{-value} &= 2P(Z \geq |Z_c|) \\ &= 2 * \texttt{normalcdf}(0.4008, 9999) \\ &= 0.6886 \end{aligned}$$

- ***Interpret the p-value:*** *If 47% of all current undergraduate students at the university are in-state, we would expect to see results like this or more extreme 68.86% of the time.*
- ***Conclusion:*** *There is little to no evidence to conclude that the current proportion of undergraduate students who are in-state has changed from 0.47 (p-value > 0.10).*

Based only on the results of the hypothesis test, do you think a 95% confidence interval for p in this setting would contain 0.47? Explain your reasoning.

The hypothesis test above concluded that there was little to evidence that p differs from 0.47. Thus, 0.47 could be the value of p. As a result, 0.47 will be in the confidence interval for p since it gives a range of plausible values.

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.

We are making conclusions about a population parameter based off one sample. Thus, it is possible that an error could happen. For example, it could just be that our data is "unusual" by chance as opposed to null hypothesis being incorrect.

Example 71. *A 2009 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. 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.*

- ***Hypotheses:*** $H_0 : p = 0.78$ *versus* $H_a : p < 0.78$
- ***Test Statistic:*** *First, note that* $\widehat{p} = 140/200 = 0.70$ *and* $n = 200$. *Thus,*

$$SE = \sqrt{\frac{0.78(1-0.78)}{200}} = 0.0293$$
$$Z_c = \frac{0.70 - 0.78}{0.0293} = -2.7304$$

- ***Compute the p-value:*** *Since* $H_a : p < 0.78$, *we have*

$$\begin{aligned} p\text{-value} &= P(Z \leq Z_c) \\ &= \texttt{normalcdf}(-9999, -2.7304) \\ &= 0.0032 \end{aligned}$$

- ***Interpret the p-value:*** *If 78% of all NFL retirees find themselves in financial ruin within three years, we would expect to see results like this or more extreme 0.32% of the time.*
- ***Conclusion:*** *There is very strong evidence to conclude that the true percentage of NFL retirees that find themselves in financial ruin within three years of retirement is less than 78% (p-value* < 0.01*).*

**** After introducing hypothesis testing and confidence intervals in the context of one proportion, the students generally have a good idea of the process. When introducing inference in other settings (two proportions, means) time typically allows for one example of each. Choose which example is interesting to the class and post solutions for the others so that students may practice as desired.****

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 _____small_____.

   Use this space to explain as needed.

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

   Use this space to explain as needed.

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

   Use this space to explain as needed.

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 $H_a : p \neq p_0$

Additional Examples

Example 72. *Suppose we test the hypothesis of* $H_0 : p = 0.2$ *vs* $H_A : p \neq 0.2$. *Using the data collected we obtain a p-value of 0.201. Based on this information, would a 95% confidence interval for p contain 0.2? Explain without calculating the interval.*

With a p-value of 0.201, there is little to no evidence to conclude that the proportion is different from 0.2. Thus, the confidence interval for p will contain 0.2.

Example 73. *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 that the proportion is something other than two-thirds?*

Since the value $p = 2/3 = 0.667$ *is within the limits of the confidence interval, there is little to no evidence to conclude that the true proportion of first down plays in the NFL differs from two-thirds.*

Example 74. *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 introductory statistics in fall 2022, 51% chose an ACM major. Conduct a test to determine if the sample evidence supports this individual's claim.*

- **Hypotheses:** $H_0 : p = 0.5$ versus $H_a : p < 0.5$
- **Test Statistic:** First, note that $\widehat{p} = 0.51$ and $n = 100$. Thus,

$$SE = \sqrt{\frac{0.5(1-0.5)}{100}} = 0.05$$
$$Z_c = \frac{0.51 - 0.5}{0.05} = 0.2$$

- **Compute the p-value:** Since $H_a : p < 0.5$, we have

$$\begin{aligned} \text{p-value} &= P(Z \leq Z_c) \\ &= \texttt{normalcdf}(-9999, 0.2) \\ &= 0.5793 \end{aligned}$$

- **Interpret the p-value:** If 50% of all undergrads chose an ACM major, we would expect to see results like this or more extreme 57.93% of the time.
- **Conclusion:** There is little to no evidence to conclude that less than half of students chose an ACM major (p-value > 0.10).

Example 75. *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?*

- **Hypotheses:** $H_0 : p = 0.667$ versus $H_a : p > 0.667$
- **Test Statistic:** First, note that $\widehat{p} = 1035/1500 = 0.69$ and $n = 1500$. Thus,

$$SE = \sqrt{\frac{0.667(1-0.667)}{1500}} = 0.0122$$
$$Z_c = \frac{0.69 - 0.667}{0.0122} = 1.8852$$

- **Compute the p-value:** Since $H_a : p > 0.667$, we have

$$\begin{aligned} \text{p-value} &= P(Z \geq Z_c) \\ &= \texttt{normalcdf}(1.8852, 9999) \\ &= 0.0297 \end{aligned}$$

- **Interpret the p-value:** If two-thirds of Americans think that ads on streaming services are repetitive, we would expect to see results like this or more extreme 2.97% of the time.
- **Conclusion:** There is strong evidence to conclude that more than two-thirds of Americans think that ads on streaming services are repetitive (p-value $= 0.0297$).

3.4 Inference for Two Proportions

Introduction

Confidence Intervals for Two Proportions

- **Target Parameter:** $p_1 - p_2$

  The sign ($\pm$) of the target parameter tells us which group has a larger proportion:

  - If $p_1 - p_2 < 0 \Rightarrow p_2 > p_1$
  - If $p_1 - p_2 > 0 \Rightarrow p_1 > p_2$
  - If $p_1 - p_2 = 0 \Rightarrow p_1 = p_2$

- **Point Estimate:** $\widehat{p}_1 - \widehat{p}_2$
- **Standard Error:**

$$SE = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

- **Multiplier:** Z^*

Hypothesis Tests for Two Proportions

- **Hypotheses:**

$$\begin{aligned} H_0: &\quad p_1 = p_2 \ (p_1 - p_2 = 0) \\ H_a: &\quad p_1 < p_2 \ (p_1 - p_2 < 0) \\ &\quad p_1 > p_2 \ (p_1 - p_2 > 0) \\ &\quad p_1 \neq p_2 \ (p_1 - p_2 \neq 0) \end{aligned}$$

- **Test Statistic:**

$$SE = \sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad Z_c = \frac{\widehat{p}_1 - \widehat{p}_2}{SE}$$

  where

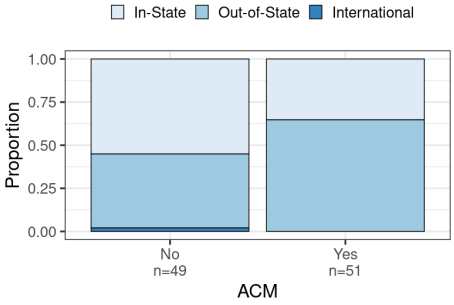
$$\widehat{p} = \text{pooled sample proportion} = \frac{x_1 + x_2}{n_1 + n_2}$$

- **Compute the p-value:** Same expressions as the one sample proportion setting.

AMS Open Math Notes: Works in Progress; Reference # OMN:202412.111448; Last Revised: 2024-12-15 09:20:44

Example 76. *In Section 1.5 Example 19 and Section 2.1 Example 43 we explored the relationship between residency status and ACM with comparative bar charts and contingency tables. Let's take a look again.*

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



What are your initial impressions for a comparison of out-of-state students in the two types of majors? What does this mean for a confidence interval for $p_1 - p_2$?

There appears to be a greater proportion of out-of-state students in ACM majors. This means a confidence interval will not include zero.

Inference can help determine if this represents a true difference in the populations or is within expected sampling variation. 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.

Group 1: ACM Major	**Group 2:** Non-ACM Major
$x_1 = 33, n_1 = 51$	$x_2 = 21, n_2 = 49$
$\widehat{p}_1 = 33/51 = 0.6471$	$\widehat{p}_2 = 21/49 = 0.4286$

- **Point Estimate:** $0.6471 - 0.4286 = 0.2185$
- **Standard Error:**

$$SE = \sqrt{\frac{0.6471(1-0.6471)}{51} + \frac{0.4286(1-0.4286)}{49}} = 0.0973$$

- **Multiplier:** $Z^* = \texttt{invNorm}(0.95) = 1.645$
- **MOE:** $\text{MOE} = \text{mult} \times \text{SE} = 1.645(0.0973) = 0.1601$
- **CI:** $\text{PE} \pm \text{MOE} \longrightarrow 0.2185 \pm 0.1601 \longrightarrow (0.0584, 0.3786)$
- **Interpretation:** We are 90% confident that the true difference in the proportion of out-of-state students between ACM majors and non-ACM majors is between 0.0584 and 0.3786.

  We are 90% confident that the percentage of ACM majors who are out-of-state students is between 5.48% and 37.86% more than the true percentage of non-ACM majors who are out-of-state students.

Example 77. *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.*

(a). Is the proportion of students whose hometowns were at least 300 miles away greater for ACM than non-ACM majors according to our sample?
Yes!

(b). Does this necessarily mean same is true for the entire population?
No!

(c). Conduct the appropriate test using the data to infer results for the population using statistics from the sample of students.

Group 1: ACM Majors	**Group 2:** Non-ACM Majors
$x_1 = 29$	$x_2 = 18$
$n_1 = 51$	$n_2 = 49$
$\widehat{p}_1 = 0.569$	$\widehat{p}_2 = 0.367$

$$\widehat{p} = 47/100 = 0.47$$

- **Hypotheses:** $H_0 : p_1 = p_2$ versus $H_a : p_1 > p_2$
- **Test Statistic:**

$$SE = \sqrt{0.47(1-0.47)\left(\frac{1}{51}+\frac{1}{49}\right)} = 0.0998$$

$$Z_c = \frac{0.569-0.367}{0.0998} = 2.0240$$

- **Compute the p-value:** Since $H_a : p_1 > p_2$, we have

$$\text{p-value} = P(Z \geq Z_c) = \texttt{normalcdf}(2.024, 9999) = 0.0215$$

- **Interpret the p-value:** If proportion of students whose hometowns were at least 300 miles away is the same for ACM and non-ACM majors, we would expect to see results like this or more extreme 2.15% of the time.
- **Conclusion:** There is strong evidence to conclude that the proportion of students with hometowns at least 300 miles from campus is greater for ACM majors than non-ACM majors ($p = 0.0215$).

Example 78. *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. Interpret the interval as well.

Group 1: *GenZer's*	***Group 2:*** *Millennials*
$x_1 = 117$	$x_2 = 424$
$n_1 = 212$	$n_2 = 673$
$\widehat{p}_1 = 117/212 = 0.5519$	$\widehat{p}_2 = 424/673 = 0.6300$

- ***Point Estimate:*** $0.5519 - 0.6300 = -0.0781$
- ***Standard Error:***

$$SE = \sqrt{\frac{0.5519(1-0.5519)}{212} + \frac{0.6300(1-0.6300)}{673}} = 0.0389$$

- ***Multiplier:*** $Z^* = \texttt{invNorm}(0.975) = 1.96$
- ***MOE:*** $MOE = mult \times SE = 1.96(0.0389) = 0.0762$
- ***CI:***

$$PE \pm MOE \longrightarrow -0.0781 \pm 0.0762 \longrightarrow (-0.1543, -0.0019)$$

- ***Interpretation:*** *We are 95% confident that the true difference in the proportion of Amazon Prime Video subscribers between the two age groups is between* -0.1543 *and* -0.0019*. We are 95% confident that the percentage of Millennials who are Amazon Prime Video subscribers is between 0.19% and 15.43% more than the true percentage of GenZer's who are Amazon Prime Video subscribers.*

(b). Based on the confidence interval, is there evidence that a difference exists in the proportion of subscribers in each age group? Discuss.
Since the value 0 is not in the confidence interval, there is strong evidence that a difference exists in the proportion of Amazon Prime Video subscribers between the two age groups.

Example 79. *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.

Group 1: *"avid film fans"*	***Group 2:*** *"casual film fans"*
$x_1 = 330$	$x_2 = 437$
$n_1 = 793$	$n_2 = 1251$
$\widehat{p}_1 = 330/793 = 0.4161$	$\widehat{p}_2 = 437/1251 = 0.3493$

$$\widehat{p} = (330+437)/(793+1251) = 767/2044 = 0.3752$$

- ***Hypotheses:*** $H_0 : p_1 = p_2$ *versus* $H_a : p_1 > p_2$
- ***Test Statistic:***

$$SE = \sqrt{0.3752(1-0.3752)\left(\frac{1}{793}+\frac{1}{1251}\right)} = 0.0220$$

$$Z_c = \frac{0.4161-0.3493}{0.0220} = 3.0364$$

- ***Compute the p-value:*** *Since* $H_a : p_1 > p_2$*, we have*

$$p\text{-value} = P(Z \geq Z_c) = \texttt{normalcdf}(3.0364, 9999) = 0.0012$$

- ***Interpret the p-value:*** *If proportion of "avid film fans" who prefer dubbing is the same as the proportion of "casual film fans" who prefer dubbing, we would expect to see results like this or more extreme 0.12% of the time.*
- ***Conclusion:*** *There is very strong evidence to conclude that avid film fans are more likely to prefer dubbing than casual fans (p-value* < 0.01*).*

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

The results of the above hypothesis test indicated that there was strong evidence of a positive difference since we concluded that avid film fans are more likely to prefer dubbing than casual fans. Thus, the value of 0 will NOT be in the confidence interval.

Chapter 4

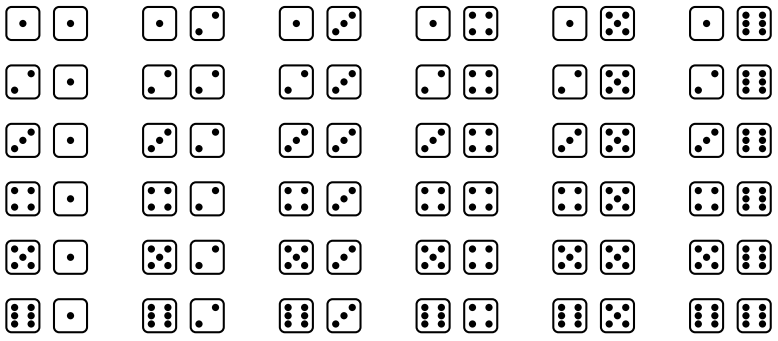
Statistical Inference for Means

4.1 Sampling Distribution of the Mean

Introductory Activity (If time does not allow for this activity, you may brifely discuss and proceed to #12 on pg 141.)

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.



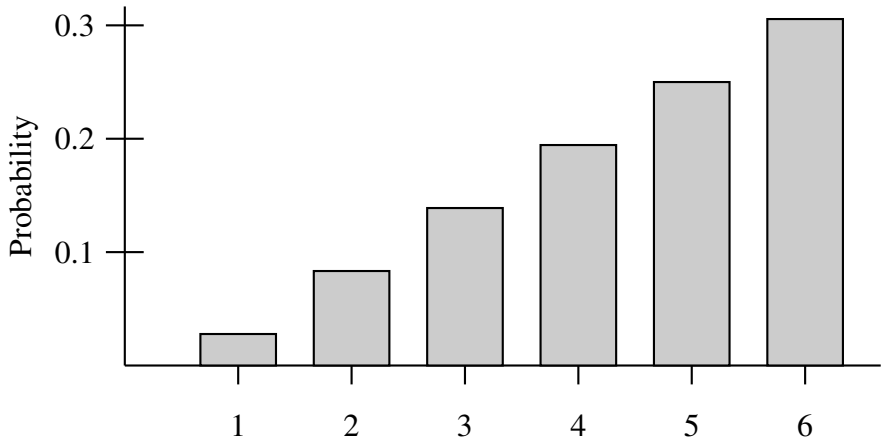
Let X be the larger of the two numbers showing.

1. Is X a discrete or continuous random variable? What are the possible values for X?
   Since the maximum value rolled has possible values 1,2,3,4,5, or 6, the random variable X is discrete.

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

x	1	2	3	4	5	6
$P(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

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



3. What is the mean of X?

$$\mu = 1\left(\frac{1}{36}\right) + 2\left(\frac{3}{36}\right) + \cdots + 6\left(\frac{11}{36}\right) = 4.47222$$

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?

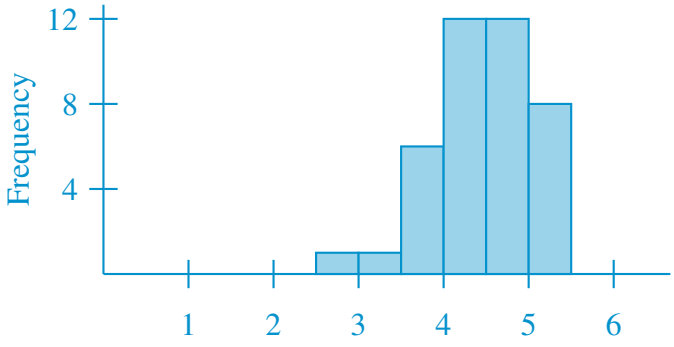
   Since $\overline{X}$ is the average of 10 numbers, it would be considered continuous.

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

$\bar{x} =$ _____

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

7. Reproduce a histogram or dot plot for the class's empirical distribution of $\overline{X}$ here. Answers will vary on this page. The class distribution should be similar to the following.



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?

   The mean of the $\bar{x}$'s should be close to $\mu = 4.47222$. From our histogram/dot plot in the previous question it appears fairly close.

9. Using `1-Var Stats`, find the sample standard deviation of the $\overline{X}$ from your class.

   Answers will vary per class. For comparison, you may wish to note that the theoretical standard deviation of X is $\sigma = 1.404$. Thus, $SE = 1.404/\sqrt{10} = 0.444$.

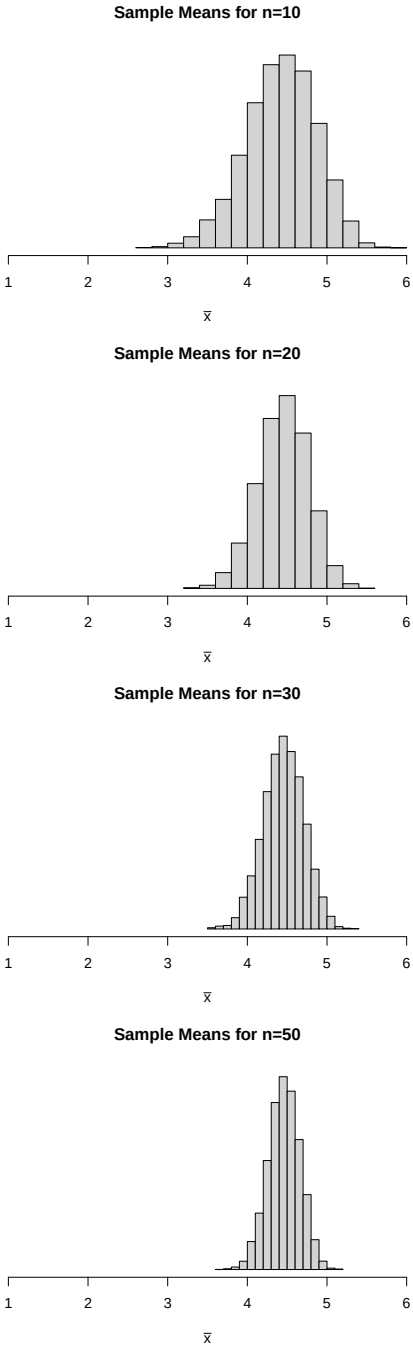
10. Suppose you did not know $\mu = 4.47222$, and you tried to use one of the $\overline{X}$ to estimate μ. How wrong would you expect to be? In other words, what is the "standard" amount of "error"?

    We would expect our results to differ, on average, by the value of the standard deviation of the $\overline{X}$.

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

    The shape of the distribution is slightly left skewed.

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?



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 80. *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.

We do not know the shape of the population distribution. So, it not appropriate to find the z-score and use `normalcdf()`

$$P(X < 25) \quad = \quad ???$$

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

Since $n > 30$, we can appeal to the CLT:

$$\overline{X} \sim N\left(29.2, \frac{10}{\sqrt{40}}\right) \quad or \quad \overline{X} \sim N(29.2, 1.5811)$$

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

$$\begin{aligned} P(\overline{X} < 25) &= P\left(Z < \frac{25 - 29.2}{1.5811}\right) \\ &= P(Z < -2.6564) \\ &= \texttt{normalcdf}(-9999, -2.6564) \\ &= 0.0039 \end{aligned}$$

Example 81. *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). State the sampling distribution of the mean for samples of $n = 10$?

$$\overline{X} \sim N\left(12, \frac{2.7}{\sqrt{10}}\right) \quad or \quad \overline{X} \sim N(12, 0.8538)$$

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

$$\begin{aligned} P(X > 15) &= P\left(Z > \frac{15-12}{2.7}\right) \\ &= P(Z > 1.1111) \\ &= \texttt{normalcdf}(1.1111, 9999) \\ &= 0.1333 \end{aligned}$$

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

$$\begin{aligned} P(\overline{X} > 15) &= P\left(Z > \frac{15-12}{0.8538}\right) \\ &= P(Z > 3.5137) \\ &= \texttt{normalcdf}(3.5137, 9999) \\ &= 2.21 \times 10^{-4} \end{aligned}$$

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

$\mu = 12$

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

$SE_{\bar{x}} = \dfrac{2.7}{\sqrt{50}} = 0.3818$

4.2 Inference for One Mean

Introduction and the t-Distribution

Confidence Intervals for One Mean

- **Target Parameter:** μ = population mean
- **Point Estimate:** $\bar{x}$ = sample mean
- **Standard Error:** Since μ is now unknown, the population standard deviation, σ, will also be unknown. So, we will estimate σ with s, the sample standard deviation.

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

- **Multiplier:** t^*; $df = n-1$

Influences on the width of the CI:

1. Sample size, n: as n ↑, the width of the CI ↓
2. Confidence level (t^*): Confidence ↑, t^* ↑, Width ↑

Hypothesis Testing for One Mean

- **Hypotheses:**

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_a : \mu < \mu_0; \mu > \mu_0; \mu \neq \mu_0$$

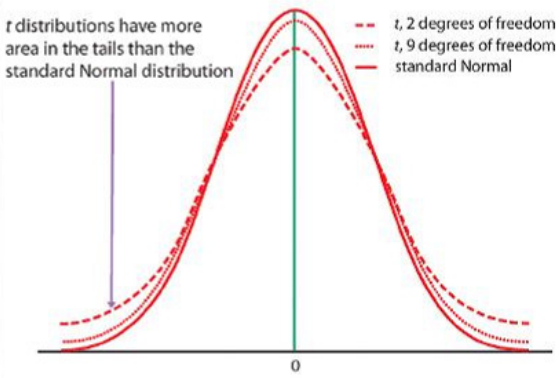
- **Test Statistic:**

$$SE = \frac{s}{\sqrt{n}} \qquad t_c = \frac{\bar{x} - \mu_0}{SE}$$

- **Compute the p-value:** If H_0 is true, $t_c \sim t_{df=n-1}$.

  For $H_a : \mu < \mu_0$,
  p-value $= P(t \leq t_c)$ = `tcdf(-9999,tc,df)`
  For $H_a : \mu > \mu_0$,
  p-value $= P(t \geq t_c)$ = `tcdf(tc,9999,df)`
  For $H_a : \mu \neq \mu_0$,
  p-value $= 2P(t \geq |t_c|)$ = `2tcdf(abs(tc),9999,df)`

Student's 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)



The entries in the following table are the critical values for Student's t-distribution for a selection of confidence levels. You may need to use this table, depending on the functions available on your calculator.

	Confidence Level					
df	80%	90%	95%	98%	99%	99.8%
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

Practice using the t-table and calculator

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

smaller df (round down; conservative)

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

- 95% Confidence with n = 15

  $t^* = \texttt{invT(0.975,14)} = |\texttt{invT(0.025,14)}| = 2.145$

- 99% Confidence with n = 50

  $t^* = \texttt{invT(0.995,49)} = |\texttt{invT(0.005,49)}| = 2.6800$
  (or $t^* = 2.704$ with $df = 40$ from table)

- 90% Confidence with n = 35

  $t^* = \texttt{invT(0.95,34)} = |\texttt{invT(0.05,34)}| = 1.6909$
  (or $t^* = 1.697$ with $df = 30$ from table)

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

- $H_A : \mu < 100, t_c = -2.4, n = 20$

  p-value $= \texttt{tcdf(-9999,-2.4,19)} = 0.0109$

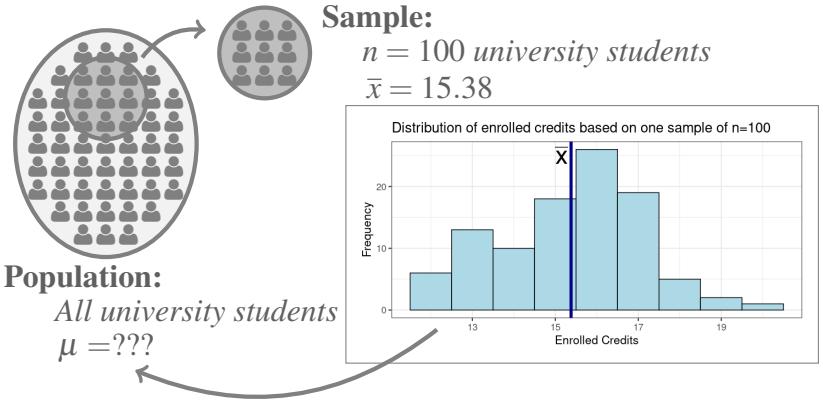
- $H_A : \mu > 100, t_c = -1.2, n = 23$

  p-value $= \texttt{tcdf(-1.2,9999,22)} = 0.8785$

- $H_A : \mu \neq 100, t_c = 2.9, n = 30$

  p-value $= \texttt{2tcdf(2.9,999,29)} = 0.0070$

Example 82. *From our course investigation of a random sample of 100 students enrolled in Introductory Statistics, 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. The scenario is depicted below.*



Use the sample of 100 university students to estimate the average number of credits enrolled for all students at the university (μ) with 90% confidence.

- **Point Estimate:** $\bar{x} = 15.38$
- **Standard Error:** $SE = s/\sqrt{n} = 1.757/\sqrt{100} = 0.1757$
- **Multiplier:** $df = 99$; $t^* = \texttt{invT}(0.95, 99) = 1.6604$ (or $t^* = 1.664$ from the table with $df = 80$)
- **MOE:** $\text{MOE} = \text{mult} \times \text{SE} = 1.6604(0.1757) = 0.2917$
- **CI:**

$$\text{PE} \pm \text{MOE} \longrightarrow 15.38 \pm 0.2917 \longrightarrow (15.0883, 15.6717)$$

- **Interpretation:** We are 90% confident that the true mean number of credits enrolled for all students is between 15.0883 and 15.6717 during fall 2022.

  Since this sample was taken from students taking introductory statistics, this might not be representative of the entire undergraduate population of the university.

Based on the CI, is there evidence that the mean number of credits taken by all students is greater than 15? Explain.

Example 83. *Administration wants to know if the average cumulative GPA for all students at the university is above 3.0. State the hypotheses for the question in statistical notation.*

$H_0 : \mu = 3$ *versus* $H_a : \mu > 3$

Let's use data from our course investigation to help answer the question. Based on a summary of cumulative GPA from our sample, does there appear to be evidence that the average is above 3.0 for the population of all students?

n	$\bar{x}$	s	M	IQR
90	3.1998	0.5584	3.257	0.775

The mean and median from the sample are both above 3.0. In fact, 75% of the data are greater than 2.82. Initial findings from the sample suggest that the mean cumulative GPA for the population could also be above 3.0.

Conduct a test to determine if the true mean GPA of students at the university exceeds 3.0.

- ***Test Statistic:***

$$SE = \frac{0.5584}{\sqrt{90}} = 0.0589 \qquad t_c = \frac{3.1998 - 3}{0.0589} = 3.3922$$

- ***Compute the p-value:*** *Since* $H_a : \mu > 3$*, we have*

$$p\text{-value} = P(t \geq t_c) = \texttt{tcdf}(3.3922, 9999, 89) = 0.0005$$

- ***Interpret the p-value:*** *If the true mean cumulative GPA for students at the university is 3.0, we would expect to see results like this or more extreme 0.05% of the time.*
- ***Conclusion:*** *There is very strong evidence to conclude that the true mean GPA of students at the university exceeds 3.0 (p-value* < 0.01*).*

Based on the hypothesis test, would we expect to see 3.0 in a confidence interval for μ*? Explain. No, since we concluded the mean GPA is likely larger than 3.0.*

Example 84. *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 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.*

- ***Point Estimate:*** $\bar{x} = 75$
- ***Standard Error:***

$$SE = \frac{s}{\sqrt{n}} = \frac{13}{\sqrt{15}} = 3.3566$$

- ***Multiplier:*** $df = 14$; $t^* = \texttt{invT}(0.975, 14) = 2.145$
- ***MOE:*** $MOE = mult \times SE = 2.145(3.3566) = 7.1999$
- ***CI:***

$$PE \pm MOE \longrightarrow 75 \pm 7.1999 \longrightarrow (67.8001, 82.1999)$$

- ***Interpretation:*** *We are 95% confident that the true mean length of all blue whales is between 67.8001 and 82.1999 feet.*

Note the duality: What hypothesis test would correspond to your confidence interval, and what results would you expect? Records state that 80 feet is most likely the average length for a blue whale. Since 80 is within the limits of our confidence interval, there is little to no evidence to refute this value. Thus, we have little to no evidence that the mean length has changed from this value.

Example 85. *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 average time spent gaming in 2022?

- ***Hypotheses:*** $H_0 : \mu = 8.5$ *versus* $H_a : \mu \neq 8.5$
- ***Test Statistic:***

$$SE = \frac{1.2}{\sqrt{45}} = 0.1789 \qquad t_c = \frac{8.9 - 8.5}{0.1789} = 2.2359$$

- ***Compute the p-value:*** *Since* $H_a : \mu \neq 8.5$*, we have*

$$p\text{-value} = 2P(t \geq |t_c|) = 2 * \texttt{tcdf}(2.2359, 9999, 44) = 0.0305$$

- ***Interpret the p-value:*** *If the true mean amount of time gamers spend playing video games is 8.5 hours per week, we would expect to see results like this or more extreme 3.05% of the time.*
- ***Conclusion:*** *There is strong evidence to conclude that the true amount of time gamers spend playing video games is no longer 8.5 hours per week (p-value* $= 0.0305$*).*

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

Since there was strong evidence to conclude that the mean was no longer 8.5, the value 8.5 will not be in the confidence interval.

4.3 Inference for Two Means

Confidence Intervals for Two Means

Target Parameter: $\mu_1 - \mu_2$

Like the difference in proportions, the sign ($\pm$) for the difference in means indicates which group has a larger mean.

- **Point Estimate:** $\bar{x}_1 - \bar{x}_2$
- **Standard Error:** $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- **Multiplier:** t^*; $df = \min(n_1 - 1, n_2 - 1)$

Hypothesis Tests for Two Means

- **Hypotheses:**

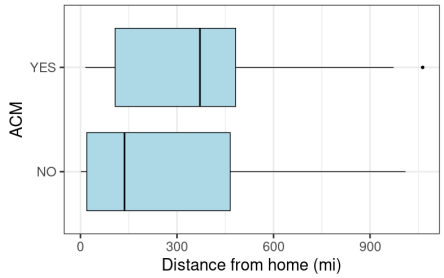
$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 \ (\mu_1 - \mu_2 = 0) \\ H_a: &\quad \mu_1 < \mu_2 \ (\mu_1 - \mu_2 < 0) \\ &\quad \mu_1 > \mu_2 \ (\mu_1 - \mu_2 > 0) \\ &\quad \mu_1 \neq \mu_2 \ (\mu_1 - \mu_2 \neq 0) \end{aligned}$$

- **Test Statistic:**

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad t_c = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

- **Compute the p-value:** Same as one sample t test with $df = \min(n_1 - 1, n_2 - 1)$

Example 86. *Consider the following summaries from our course investigation.*



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

Recall this comparison with visual and numerical summaries in Section 1.5, Example 20. What were our initial impressions? Does it appear that average distance from home varies based on ACM status?

Comparing the sample means and boxplots, it appears that the mean and median distances from home for ACM majors are greater than non ACM majors.

Let's take those initial impressions from the data and see how they can inform us about the population through statistical inference. Is there enough evidence to conclude that the distance from a student's hometown is typically greater for students whose major is available through the ACM? Test this claim.

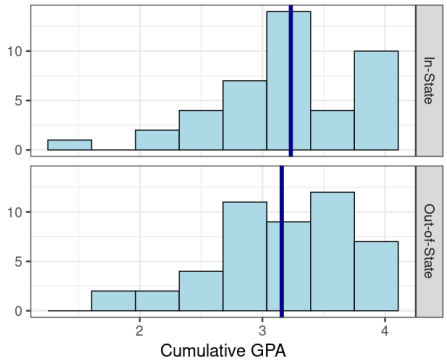
- ***Hypotheses:*** $H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 > \mu_2$
- ***Test Statistic:***

$$SE = \sqrt{\frac{257.8415^2}{51} + \frac{255.1598^2}{48}} = 51.5748$$

$$t_c = \frac{340.902 - 257.3104}{51.5748} = 1.6208$$

- ***Compute the p-value:*** *Since* $H_a : \mu_1 > \mu_2$, *we have p-value* $= P(t \geq t_c) = \texttt{tcdf}(1.6208, 9999, 47) = 0.0559$
- ***Interpret the p-value:*** *If there is no difference in the mean distance from home between the ACM majors and non-ACM majors, we would expect to see results like this or more extreme 5.59% of the time.*
- ***Conclusion:*** *There is some evidence to conclude that the mean distance from home is greater for ACM majors than for non-ACM majors (p-value* $= 0.0559$*).*

Example 87. *Consider the following summaries from our course investigation.*



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

Does there appear to be a difference in average cumulative GPA for the two groups? We can take this initial impression from the data and formally use it to make an inference to compare the populations. Based on our initial impression, what do you expect the confidence interval for $\mu_1 - \mu_2$ to look like?

There appears to be little difference in average cumulative GPA for the two groups. Inferring for the entire population we might conclude the same. Therefore, we expect zero to be in the CI.

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

- ***Point Estimate:*** $\bar{x}_1 - \bar{x}_2 = 3.2314 - 3.1575 = 0.0739$
- ***Standard Error:***

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.5446^2}{42} + \frac{0.5716^2}{47}} = 0.1184$$

- ***Multiplier:*** $df = 41$; $t^* = \texttt{invT}(0.975, 41) = 2.0195$
- ***MOE:*** $MOE = mult \times SE = 2.0195(0.1184) = 0.2391$
- ***CI:***

$$PE \pm MOE \longrightarrow 0.0739 \pm 0.2391 \longrightarrow (-0.1652, 0.3130)$$

- ***Interpretation:*** *We are 95% confident that the true difference in mean cumulative GPA for in-state and out-of-state students is between -0.1652 and 0.3130.*

Example 88. *Sixgill sharks sampled in the Puget Sound were measured and sexed. A summary of the results is given below. Is there significant evidence that a difference exists in the average size of adult male and adult female sixgill sharks? 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

Let Group 1 be female sharks and Group 2 be male sharks.

- ***Point Estimate:*** $\bar{x}_1 - \bar{x}_2 = 3.7 - 3.2 = 0.5$
- ***Standard Error:***

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{0.4^2}{51} + \frac{0.35^2}{26}} = 0.0886$$

- ***Multiplier:*** $df = 25$; $t^* = \texttt{invT}(0.975, 25) = 2.0595$
- ***MOE:*** $MOE = mult \times SE = 2.0595(0.0886) = 0.1825$
- ***CI:***

$$PE \pm MOE \longrightarrow 0.5 \pm 0.1825 \longrightarrow (0.3175, 0.6825)$$

- ***Interpretation:*** *We are 95% confident that the true difference in mean size of female and male sixgill sharks is between 0.3175 and 0.6825 meters.*

  Since 0 is not within the limits of our confidence interval, there is strong evidence that a difference exists in mean size of female and male sixgill sharks.

Example 89. *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.

Let Group 1 represent medium to high alcohol use with little to no marijuana use and Group 2 represent high alcohol and high marijuana use.

- ***Hypotheses:*** $H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 \neq \mu_2$
- ***Test Statistic:***

$$SE = \sqrt{\frac{0.64^2}{463} + \frac{0.83^2}{188}} = 0.0674 \qquad t_c = \frac{3.03 - 2.66}{0.0674} = 5.4896$$

- ***Compute the p-value:*** *Since* $H_a : \mu_1 \neq \mu_2$*, we have*

$$\begin{aligned} p\text{-value} &= 2P(t \geq |t_c|) \\ &= 2 * \texttt{tcdf}(5.4896, 9999, 187) \\ &= 1.2997 \times 10^{-7} \end{aligned}$$

- ***Interpret the p-value:*** *If there is no difference in the mean GPA between the two substance use groups, we would expect to see results like this or more extreme* $1.2997 \times 10^{-5}\%$ *of the time.*
- ***Conclusion:*** *There is very strong evidence to conclude that a difference exists in the mean GPA of college students between the two substance use groups (p-value* < 0.01*).*

(b). Based on the results of the hypothesis test, would a 95% confidence interval for the difference in means contain 0? Discuss. *Since there is strong evidence of a difference, the value 0 will not be in the confidence interval.*

4.4 ANOVA

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

Analysis **o**f **Va**riance

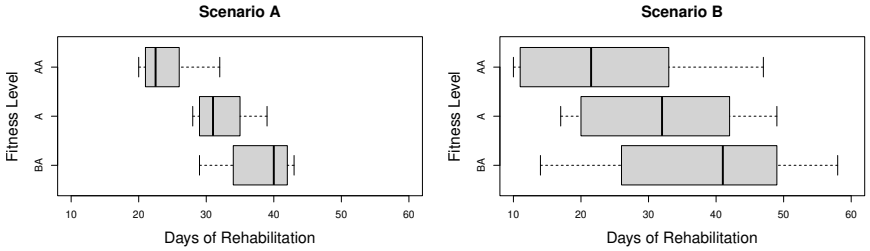
Why does a test about means involve variances? Consider the following.

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):
  number of days to complete PT
- Factor (explanatory variable):
  prior fitness level
- Factor levels:
  below average (BA), average (A), above average (AA)
- Experimental units (subjects):
  24 male 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?
Better fitness levels appear to be associated with shorter recovery times. However, we know that a difference in sample means is not always indicative of a difference in population means. This is why we conduct hypothesis tests!

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?
Scenario A gives stronger evidence because there is less variability in recovery time within each fitness level.



Therefore, we must consider two types of variability:

- The variability *between* each of the groups is measured by **SSFactor**. SSFactor computes the variability in the response (recovery time) that is *explained* by the factor (prior fitness level). In our case, it does seem that better fitness has lower average recovery times (38 vs. 32 vs. 24 days).

- The variability *within* each treatment group is measured by **SSError**. SSError computes the variability in the response (recovery time) that is *not explained* by the factor (prior fitness level). In scenario B, even though we might consider those with below average fitness, there is still a great deal of variability in recovery times (high SSError). The opposite is true for scenario A.

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	-	-	-

Example 90. *Obtain the ANOVA table using the calculator for the example of the physical rehabilitation time. 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 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	2	672	336	16.96	4.1×10^{-5}
Error	21	416	19.8095	-	-
Total	23	1088	-	-	-

- **Hypotheses:** $H_0 : \mu_1 = \mu_2 = \mu_3$ versus H_a : not all μ equal
- **Test Statistic:** $F_c = 16.96$
- **Compute the p-value:** 4.1×10^{-5}
- **Interpret the p-value:** If the mean recovery time is the same across all fitness levels, we would see results like this or more extreme 0.0041% of the time.
- **Conclusion:** There is very strong evidence to conclude that at least two fitness levels differ with respect to mean recovery time (p-value < 0.01).

Example 91. *Let's explore the computation of sums of squares using our knee surgery data.*
Optional activity/example as time and interest allows.

	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	
Below Average	
Average	
Above Average	

SSTotal = *dfTotal =*

Deviations between group mean and overall mean:

Group	
Below Average	
Average	
Above Average	

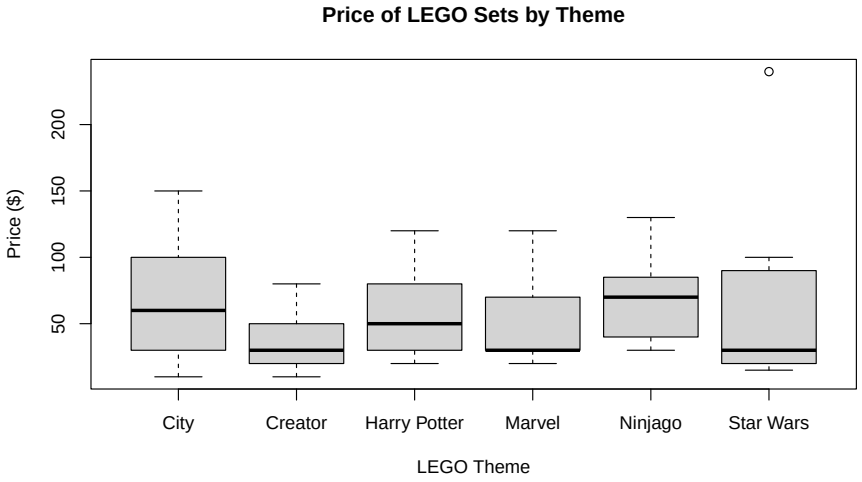
SSFactor = *dfFactor =* *MSFactor =*

Deviations from group means:

Group	
Below Average	
Average	
Above Average	

SSError = *dfError =* *MSError =*

Example 92. *An honors student 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. He randomly sampled 9 sets for each theme and recorded 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.

There appears to be more variability within the different themes as compared to the variability between the different themes/groups. Thus, theme most likely will not have an impact on the mean price of a LEGO set.

(b). Complete the following ANOVA table.

Source	DF	SS	MS	F	P-value
Factor	5	6773	1354.6	0.7113	0.618
Error	48	91406	1904.292	-	-
Total	53	98179	-	-	-

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

- ***Hypotheses:*** $H_0 : \mu_1 = \mu_2 = \cdots = \mu_6$ *versus* H_a *: not all* μ *equal*
- ***Interpret the p-value:*** *If the LEGO themes have the same mean price, we would expect to see results like this or more extreme 61.8% of the time.*
- ***Conclusion:*** *There is little to no evidence that at least two LEGO themes differ with respect to mean price (p-value* > 0.10*).*

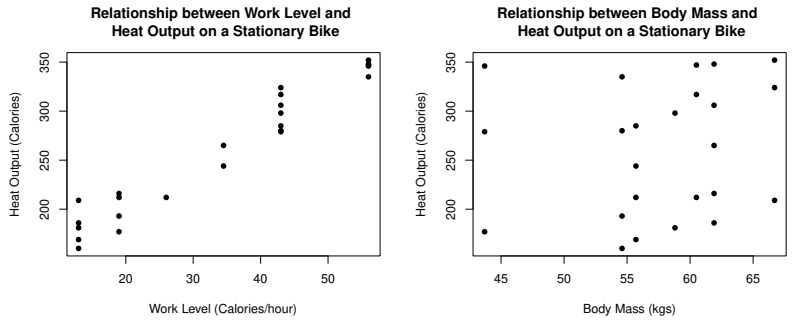
Chapter 5

Associations Between Quantitative Variables

5.1 Scatterplots

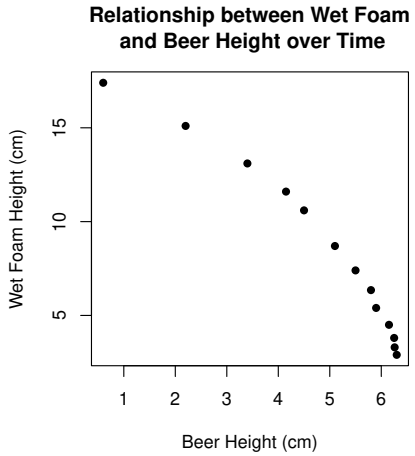
Scatterplots provide a way to study the relationship between two quantitative variables measured on the same subject or at the same time point. The data is plotted as (x, y) coordinates. If it is thought that one variable exerts influence on the other, it is plotted on the x-axis and called the **explanatory variable**. The variable on the y-axis is called the **response variable** and may be impacted by the explanatory variable.

Example 93. *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: *In the first plot we can see a fairly strong positive trend. As the work level increases, the heat output also increases in a linear fashion. In the plot on the right, there appears to be little relationship between body mass and heat output.*

Example 94. *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:

As beer height increases, there is a consistent decrease in wet foam height. Also, the trend looks more like a curve than a line.

Key Features:

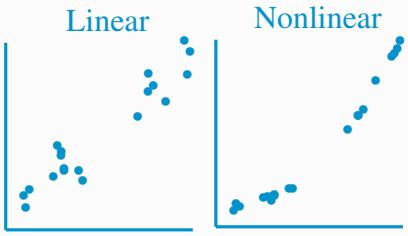
(Please discuss after introducing key features)

- ***Form:*** *Nonlinear*
- ***Association:*** *Negative*
- ***Strength:*** *Very Strong*
- ***Outliers:*** *None*

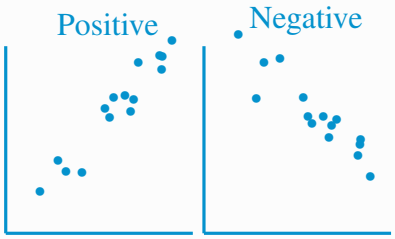
Key Features of Scatterplots

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

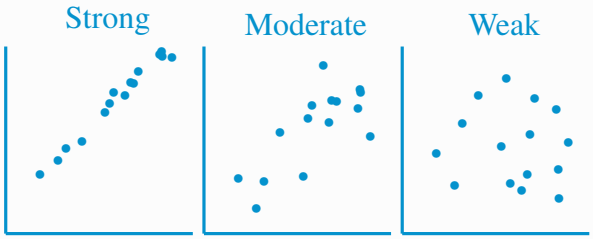
1. **Form**



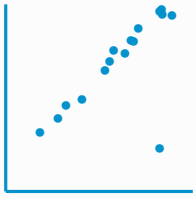
2. **Association/Direction**



3. **Strength**



4. **Outliers**

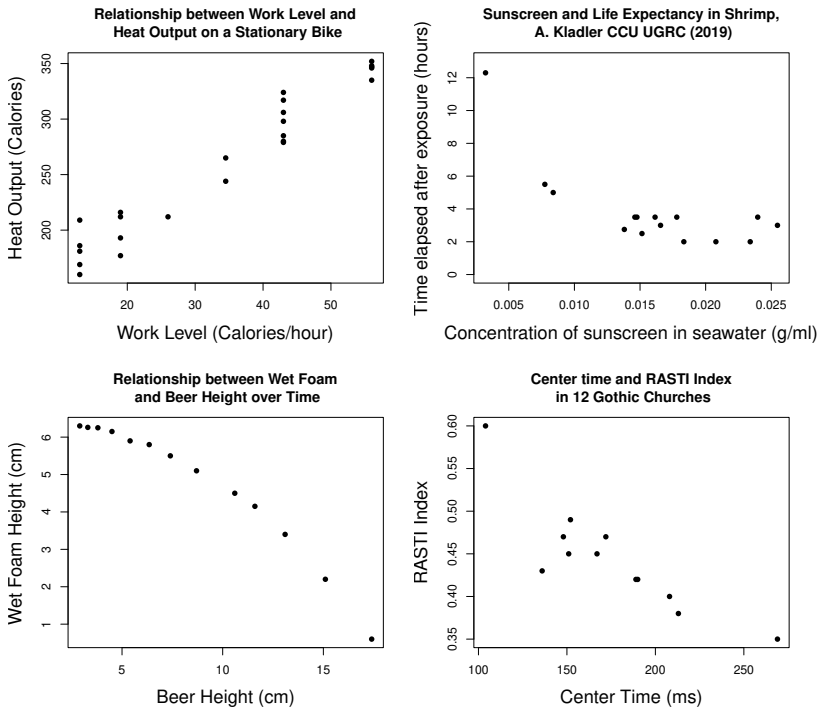


Reevaluate our observations of the previous scatterplots to make sure we addressed all of the key features in each plot.

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)**.

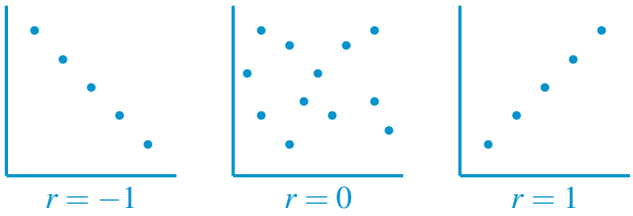


According to this, in which of the above scenarios could we compute correlation (r)?

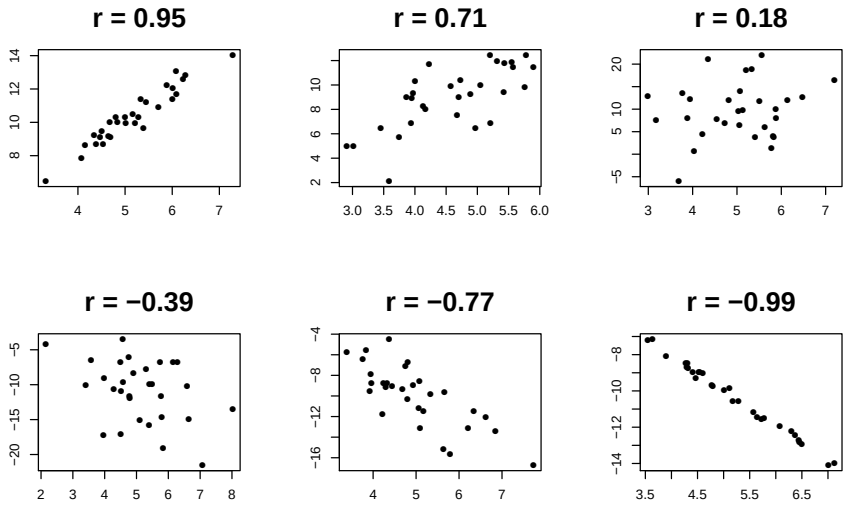
Relationship is linear so *r* is appropriate	Relationship is nonlinear so *r* is NOT appropriate
Relationship is nonlinear so *r* is NOT appropriate	Relationship is linear so *r* is appropriate

Some Properties of Correlation:

1. Correlation is always between -1 and 1.



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

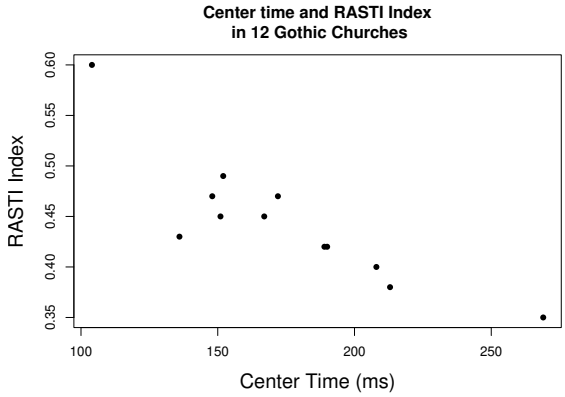


3. Correlation has no units! Its value does not depend on the units of the two variables.
4. Correlation is the same regardless of how you assign the x and y variables.
5. Correlation does not imply causation!!!

If time allows, you may use this space to explain how correlation is computed with z-scores to help with understanding in general and of points 3 and 4 above.

Finding correlation on the calculator

Example 95. *Consider the scatterplot 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.

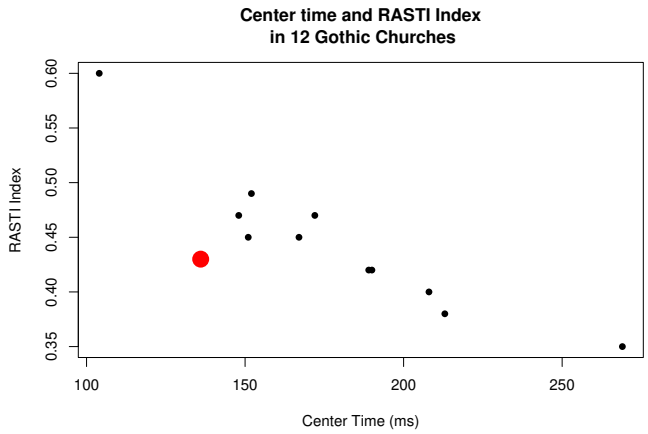
Guessed: $r =$ *value close to* -1 *Actual:* $r =$*-0.8691*

Church	1	2	3	4	5	6
Center Time	213	208	172	190	152	167
RASTI	0.38	0.40	0.47	0.42	0.49	0.45
Church	7	8	9	10	11	12
Center Time	148	151	136	269	104	189
RASTI	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 ($2^{\text{nd}} \to$ `0`). Scroll down to `DiagnosticOn`. Press enter twice.
1. Enter the X variable into $\texttt{L}_1$ and the Y variable into $\texttt{L}_2$. To do this, go to `STAT` $\to$ `Edit`.
2. To obtain correlation, go to `STAT` $\to$ `CALC` $\to$ `LinReg(a+bx)`.

Example 96. *What do you think would happen to correlation if we removed observation 9,* $(136, 0.43)$ *(enlarged below)? Discuss, then compute to confirm.*



The trend will be stronger without the observation so the value of r should get closer to -1. In fact, $r = -0.9302$ when the observation is removed.

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

One observation has the power to change the value of correlation greatly so correlation is not a resistant calculation.

By the way, STI stands for "speech transmission index," and RASTI stands for "room acoustics STI." It has been the standard since 1980 to measure the intelligibility of human speech over some channel (such as a cell phone connection), and for RASTI specifically, that channel is a room.

5.3 Simple Linear Regression

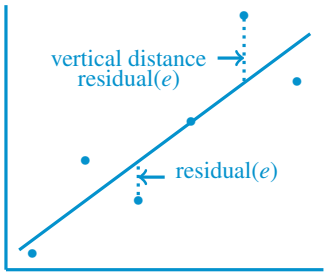
We have seen that **scatterplots** can be used to visualize the relationship between two quantitative variables. When it is clear, the **explanatory** variable is represented on the x-axis and the **response** variable is represented on the y-axis. Scatterplots gave us an idea of the **form, direction, 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:

- **Form:** Linear
- **Association:** Positive
- **Strength:** Strong
- **Outliers:** None

Since the relationship is positively strong and linear, r will be close to 1.

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.



The vertical distances between the observed values of Y and the predicted value of Y ($\widehat{Y}$, the regression line) are called residuals, e.

Method of Least Squares: Using calculus, we find the line that minimizes the sum of the squared residuals (errors). The resulting line is known as the least squares regression line.

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,

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

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 $\texttt{L}_1$ and the Y variable into $\texttt{L}_2$. To do this, go to $\texttt{STAT} \to \texttt{Edit}$.
2. To obtain the slope and intercept for the LSR line, go to $\texttt{STAT} \to \texttt{CALC} \to \texttt{LinReg(a+bx)}$.

Example 97. *Consider the data below for the study on energy output on stationary bikes.*

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

(a). Compute the regression line and write it in context.

Using the calculator, we get the following regression equation:

$$\widehat{y} = 126.5588 + 3.9212x$$

In the terms of the problem, we have:

$$\widehat{Heat} = 126.5588 + 3.9212(Work)$$

(b). 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	*201.0616*
20	*204.9828*
42	*291.2492*
43	*295.1704*

- *For $x = 19$:* $\widehat{y} = 126.5588 + 3.9212(19) = 201.0616$
- *For $x = 20$:* $\widehat{y} = 126.5588 + 3.9212(20) = 204.9828$
- *For $x = 42$:* $\widehat{y} = 126.5588 + 3.9212(42) = 291.2492$
- *For $x = 43$:* $\widehat{y} = 126.5588 + 3.9212(43) = 295.1704$

(c). 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.*

- *Subject 1:* $\widehat{y} = 126.5588 + 3.9212(19) = 201.0616$*. Thus,*

$$e = y - \widehat{y} = 177 - 201.0616 = -24.0616$$

The observed value for subject 1 is 24.0616 units below the regression line.

- *Subject 2:* $\widehat{y} = 126.5588 + 3.9212(43) = 295.1704$*. Thus,*

$$e = y - \widehat{y} = 298 - 295.1704 = 2.8296$$

The observed value for subject 13 is 2.8296 units above the regression line.

(d). 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?

- *When the work level changes from 19 to 20, the heat output changes from 201.0616 to 204.9828 which is a difference of*

$$204.9828 - 201.0616 = 3.9212 \leftarrow \textit{slope}$$

- *When work level changes from 42 to 43, the heat output changes from 291.2492 to 295.1704 which is a difference of*

$$295.1704 - 291.2492 = 3.9212 \leftarrow \textit{slope}$$

- *Recall from a previous math course that*

$$\textit{slope} = \frac{\textit{rise}}{\textit{run}}$$

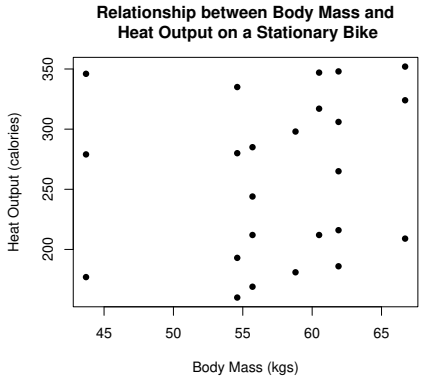
- *So, as work level increases by 1 unit, the heat output is increasing by 3.9212 units which is the slope!*

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

Example 98. *Interpret the slope of our bicycle example in context.*

As the work level increases by one calorie per hour, the heat output is predicted to increase by 3.9212 calories.

Example 99. *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.*



A previously discussed the plot displays a very weak linear form. Thus, as body mass increases, there is no consistent change in heat output. This would result in a slope close to 0. Thus, $b \approx 0$.

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?

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,

$$-1 \leq r \leq 1$$

This implies that,

$$0 \leq r^2 \leq 1$$

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

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

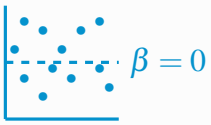
Therefore, r^2 is giving us a proportion. Specifically, r^2 tells us **the proportion of variation in y that is explained by the line with x.** Inference for the slope is not always covered at the author's institution due to time constraints.

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

Hypotheses:

$H_0 : \beta = 0$ (not linearly related)

$H_a : \beta \neq 0$ (linearly related)



X has no impact on Y

Test Statistic:

$$t_c = \frac{b - 0}{SE_b}$$

P-value:

Use `tcdf()` with $df = n - 2$.

p-value $= 2P(t \geq |t_c|) =$ `2tcdf(|tc|,9999,df)`

Example 100. *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?*

From our calculator, we observed $r^2 = 0.9476$. Thus, around 94.76% of the variation in heat output is explained by a linear regression on work level. Since r^2 is close to 1, work level is a useful predictor of heat output.

Example 101. *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?*

Note that $r^2 = (0.1434)^2 = 0.0206$. Thus, around 2.06% of the variation in heat output is explained by a linear regression on body mass. Since r^2 is close to 0, body mass is NOT a useful predictor of heat output.

Example 102. *Previously we 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.*

- ***Hypotheses:*** $H_0 : \beta = 0$ *versus* $H_a : \beta \neq 0$
- ***Test Statistic:***

$$t_c = \frac{b}{SE_b} = \frac{3.9212}{0.1965} = 19.9552$$

- ***Compute the p-value:*** *Since* $n = 24$, $df = 24 - 2 = 22$. *Thus,*

$$\begin{aligned} p\text{-value} &= 2P(t \geq |t_c|) \\ &= 2\texttt{tcdf(19.9552,9999,22)} \\ &= 1.3965 \times 10^{-15} \end{aligned}$$

- ***Interpret the p-value:*** *If the slope is truly 0, we would expect to see results like this or more extreme roughly 0% of the time.*
- ***Conclusion:*** *There is very strong evidence to conclude that a linear relationship exists between work level and heat output* (p-*value* < 0.01).

Example 103. *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.*

- ***Hypotheses:*** $H_0 : \beta = 0$ *versus* $H_a : \beta \neq 0$
- ***Test Statistic:***

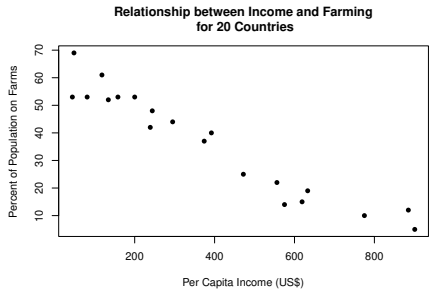
$$t_c = \frac{b}{SE_b} = \frac{1.434}{2.110} = 0.6796$$

- ***Compute the p-value:*** *Since* $n = 24$, $df = 24 - 2 = 22$*. Thus,*

$$\begin{aligned} p\text{-value} &= 2P(t \geq |t_c|) \\ &= \texttt{2tcdf(0.6796,9999,22)} \\ &= 0.5038 \end{aligned}$$

- ***Interpret the p-value:*** *If the slope is truly 0, we would expect to see results like this or more extreme 50.38% of the time.*
- ***Conclusion:*** *There is little to no evidence to conclude that a linear relationship exists between body mass and heat output* (p*-value* > 0.10*).*

Example 104. *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.

- *Regression Equation:* $\widehat{y} = 62.4939 - 0.0675x$
- *In Context:* $\widehat{Farm\ \%} = 62.4939 - 0.0675(Income)$

(b). Interpret the slope in context.

For every \$1 increase in per capita income, the percent of the population living on farms decreases 0.0675% on average.

(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?

First, we need to compute the predicted value for Greece:

$$\widehat{y} = 62.4939 - 0.0675(134) = 53.4489$$

Next, we compute the residual:

$$e = y - \widehat{y} = 52 - 53.4489 = -1.4489$$

The percent of the population living on farms for Greece is 1.4489% below the average based on their per capita income.

(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.

From our calculator, we observed $r^2 = 0.9271$. Thus, around 92.71% of the variation in percent of the population living on farms is explained by a linear regression on per capita income. Since r^2 is close to 1, the regression equation is useful for predictions.

(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.

- ***Hypotheses:*** $H_0 : \beta = 0$ *versus* $H_a : \beta \neq 0$
- ***Test Statistic:***

$$t_c = \frac{b}{SE_b} = \frac{0.0675}{0.0044} = 15.3409$$

- ***Compute the p-value:*** *Since $n = 20$, $df = 20 - 2 = 18$. Thus,*

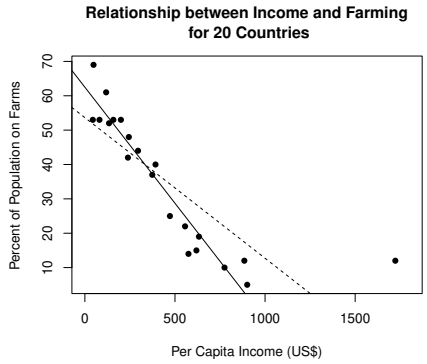
$$\begin{aligned} p\text{-value} &= 2P(t \geq |t_c|) \\ &= \texttt{2tcdf(15.3409,9999,18)} \\ &= 8.8451 \times 10^{-12} \end{aligned}$$

- ***Interpret the p-value:*** *If the slope is truly 0, we would expect to see results like this or more extreme roughly 0% of the time.*
- ***Conclusion:*** *There is very strong evidence to conclude that a linear relationship exists between per capita income and the percent of the population living on farms (p-value < 0.01).*

Influential Observations and Extrapolation

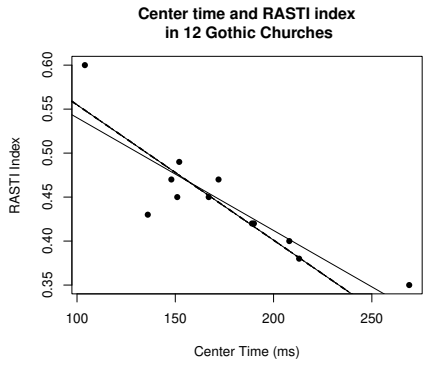
Influential points in regression are those that have a very large or small x-value compared to the majority of the data. In addition, influential points do not follow the overall pattern of the data. It is important to note and carefully examine any influential points in the data because they will influence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

Example 105. *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.*



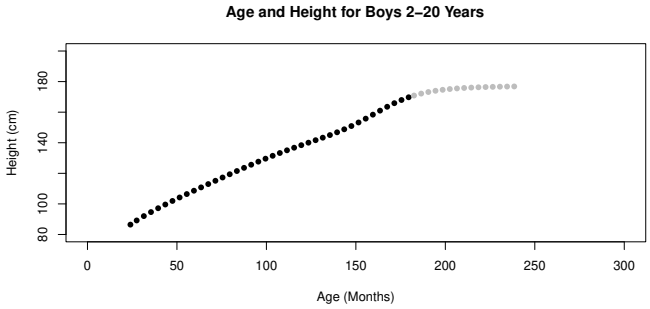
Data	*Slope*	*Intercept*	*r*	r^2
Original	*-0.068*	*62.494*	*-0.963*	*0.927*
With US	*-0.041*	*53.635*	*-0.828*	*0.686*

Example 106. *Recall our example of acoustics in Gothic churches. An influential observation existed there as well.*

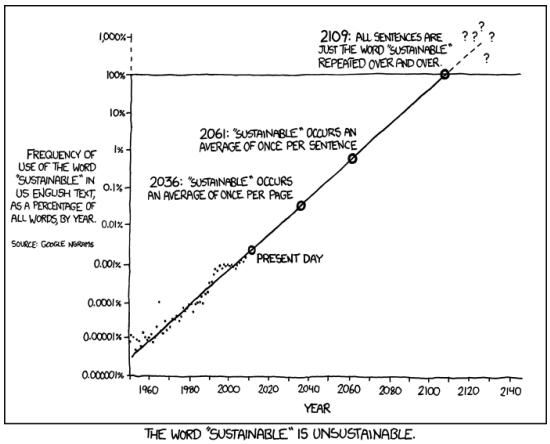


Extrapolation is the practice of using your line of best fit to make predictions for values of x that were never modeled. It is bad practice to use the line to make predictions for values of x that are smaller or larger than the observed data because the trend may change beyond the observed data.

Example 107. *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 108. *To end on a comic note (*`https://xkcd.com/1007/`*). The word "sustainable" is unsustainable."*



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.