The Math 183 Problem Book

David Quarfoot
UCSD, Department of Mathematics

"All life is a series of problems which we must try and solve, first one and then the next and then the next, until at last, we die."

– The Dowager Countess, *Downton Abbey*

September 7, 2025

For the Ross Mathematics Program, the Art of Problem Solving, and Project Euler:

You bring great problems to the world.

For Dexter:

You made all problems disappear.

Introduction

I believe in facing reality. In the modern era, that reality includes the internet, unregulated chat rooms, websites devoted to cheating, computer algebra systems, and artificial intelligence that can instantly solve just about any homework problem at the undergraduate level and below. Given these, the temptation for misconduct is higher than ever, and the fairness gap between those that do everything themselves and those who do nothing themselves has never been larger. My goal with this book is to eliminate that gap. On the pages that follow, you will find all the homework problems for this class, including hints, complete solutions (!), and commentary. That's right – **I'm giving you the solutions to all your homework at the start of the quarter**.

Why do this? In addition to eliminating the gap between the ethical and unethical, there are many benefits of this approach:

- You get immediate feedback on your thinking. No more waiting a week to see if your work was correct. Inaccuracy and inefficiency are eliminated quickly.
- You get a clear idea of how in-depth your solutions should be. If you're not showing the steps I include, you're likely under-explaining your solutions. Also, if you write far more than I do, you are likely over-explaining; this trait is problematic for finishing in-class exams and moving to higher levels of academia where solutions continue to grow in length.
- You can spend a predictable amount of time working on your homework. See the below section on how to use this book.
- The nature of office hours changes for the better. Without this approach, 99% of what I do in office hours is help students get through homework problems. While this is important guidance, most of it could be done by simply providing solutions from the start. With this time freed up, students can now use office hours to talk about ideas they are struggling with, discuss career pathways/majors, develop connections with professors, etc.

As Thomas Sowell said: "There are no solutions. There are only trade offs." I know that providing students with complete solutions will tempt some people to game the system. Your natural inclination might be to give up too easily, or to not push as hard as you would in a traditional graded environment. Indeed, some might choose to copy all the solutions and not engage with the content at all. While such strategies will reap short-term rewards (more free time, less exertion), the long-term consequences are serious (poor exam performance, lack of understanding, underdevelopment of grit, weak problem-solving skills, etc.).

As such, I ask you try a new approach to your homework (outlined in the next section). This approach will require *maturity* above all else. One aspect of maturity is understanding the proper use of resources. As an example, when a person has an immature view of the calculator, they see it as a tool that should be used for simple calculations (e.g., $3 \cdot 7$, $99 + 9$). To spend the time to type in such calculations is inefficient. If you build enough inefficient tool use into your solution procedures, you'll soon be unable to complete any large task. In a similar way, people have had to learn to maturely use social media, video recommender algorithms, AI image generation, web-page comments sections, and so on. This class might be the first where you have access to the "solutions included" tool, so let's think about its mature use.

How to Use This Book

In this class, **homework will still be collected and graded**. With solutions at hand, the main challenge is not whether you get the problems correct or not, but rather, whether you are willing to take the time to engage with the problems and complete your work. Here is what I recommend you do. Below you will find a list of class lectures. For each lecture, I have listed the problems that can be completed once you have learned the concepts from that lecture. It is common for problems to also use ideas from earlier lectures to reinforce old concepts.

For each problem, set a time plan based on your schedule. As an example, you might give each homework problem in the class a maximum time of 30 minutes. The first 25 minutes are reserved for an iterative collection of attempts and guidance seeking (e.g., first attempt, first hint, second attempt, second hint, third attempt, consulting part of the solution, etc.). The last 5 minutes are reserved for reading the complete solution and commentary, comparing your solution to mine, thinking about what you learned, etc. By including hints and the full solution, I have tried to create a system that guides you forward but encourages you to provide the intellectual thrust. To help you with your time plan, note that a 4 credit hour class (per UCSD definitions) is designed to engage you in roughly 12 hours of work per week. Removing lecture and discussion, this leaves roughly 8 hours per week for out-of-class growth. Suppose you wish to spend 2 of these hours reading the textbook, previewing lecture slides, asking questions in office hours, etc. This leaves 6 hours/week for homework, or 30 minutes/problem in a week with 12 homework problems.

If your goals include becoming a better problem solver, going to graduate school, or working on the most difficult/interesting problems in the real world, then you should set a longer time value per problem and push yourself to use hints and solutions sparingly.

- Data I: #1
- Data II: #2-4
- Probability Theory I: #5-7
- Probability Theory II: #8-11
- Probability Theory III: #12-16
- Discrete Random Variables (DRVs): #17-21
- Bernoulli, Geometric, Binomial RVs: #22-26
- Negative Binomial, Poisson RVs: #27-29
- Continuous Random Variables (CRVs): #30-31
- Uniform, Exponential RVs: #32-33
- Normal, Standard Normal RVs: #34-36
- RV Expressions/Combos, Central Limit Theorem: #37-42
- Inference Intro: #43
- Confidence Procedures and Intervals (CPs, CIs): #44-47
- Hypothesis Testing (HT) and the PHANTOM Framework: #48-52
- 1 Proportion HT: #53-54
- 1 Proportion CI and the PANIC Framework: #55-57
- 2 Proportion HT and CIs: #58-61
- ≥ 3 Proportion HT (Chi-Squared): #62-64

- 1 Mean HT and CIs: #65
- 2 Mean HT and CIs (Independent and Paired): #66-68
- ≥ 3 Mean HT (ANOVA): #69-72
- Type I/II Errors: #73-74
- Power, Sample Size: #75-76
- Linear Modeling: #77-80

Instructions

Despite having the complete solutions, you will still need to solve the homework problems and write up your answers **by hand** using paper or a tablet.

- It is fine if your solutions exactly match those provided here. **Typed solutions (e.g., in Word or LaTeX) are not accepted and will earn no credit.** The grading team will be checking your handwritten homework solutions and exam solutions to see that the handwriting matches.
- See the course calendar to learn when each problem is due. See the syllabus for instructions for uploading to Gradescope.
- Some problems will have requirements at the end like WISE, VIZ, PHANTOM, and PANIC − these are gradually introduced as the course unfolds and explained in class. If you see these, you must do everything they demand or you will lose credit.
- You may list answers in exact form (e.g., π) or round to three decimal places (e.g., 3.142), unless the problem says otherwise. If rounding to three decimal places would result in the number 0 (e.g., with 0.00012345), instead use scientific notation and write three decimal places (e.g., $1.235 \cdot 10^{-4}$).
- On any problem involving R, you must include your code and output (results, visualizations, etc.) as part of your answer. You may take a screenshot of these or write them by hand.
- If some prerequisite knowledge is required which you have forgotten, you should feel free to consult books/internet to learn this knowledge (e.g., Taylor series, improper integrals, L'Hôpital's Rule, etc.). Expect prerequisite knowledge to be drawn on frequently.
- **Grading:** Each problem submitted is worth 1 point (regardless of its length or number of parts). If you submit a complete and accurate solution, you will earn this point. Some common markdowns: Did not attempt problem (-1), Completed problem but failed to map to a location when uploading to Gradescope (-0.25), Did not hand-write solution (-1), Completed only part of the problem (deduction varies).

Problems, Hints, Solutions, and Commentary

1. On TritonEd/Canvas, under the Files link, you will find a homework folder. In the Data Sets subfolder, you will find COVID19Data.csv, a file with information about the spread of coronavirus, which I downloaded in Summer 2020. Load this into R by using the Import Dataset menu and "From Text (base)" in the upper-right window of R Studio. Make sure the Heading radio button is on "Yes", which tells the computer that the columns are already named in the file. Also, as mentioned above, whenever a problem involves using R, you MUST include your code and any outputs/visualizations as part of your answer. You may either write the code by hand and hand-sketch the output/visualizations, or you can take screenshots of everything and fold these into your PDF.

   (a) State the data type (nominal, ordinal, discrete, or continuous) and the support (listed as a set) for the variables: Recovered, Recovered...100.Cases, and WHO.Region. Respectively, these represent (at the date the data were accumulated): the total number of people who had recovered in a country, the percentage that recover, and which World Health Organization region the country belongs to.

   (b) Often, we don't want to work with an entire data set. In R, you can use the subset command to refine your data set to a smaller set of interest. For example,

```
smallerdataset = subset(COVID19Data, WHO.Region == "Europe")
```

   eliminates all the rows for countries outside Europe. The expression after the comma uses column names and operations like == (equals), != (not equals), >= (greater than or equal to), etc. Limit the data set to countries with fewer than 1000 deaths from COVID-19 and make a histogram of the death counts in the countries that remain. Your histogram should have only 5 bins (from 0 to 200, 200 to 400, 400 to 600, 600 to 800, and 800 to 1000) and the title: Death Counts Among Countries with < 1000 Deaths.

   (c) Describe the number of modes and the skew of your histogram from part b.

   (d) Make a boxplot of the Active case totals in African countries that have fewer than 5000 active cases (build your subset from the original data set, not based on part b). Comment on whether the data are likely skewed left or right based on what you see in the boxplot.

As with all problems in this book, a list of hints, the solution, and a commentary can found on the next page. Never leave a problem without reading the full solution and commentary. The goal of the commentary is to help you understand what you should have learned from the problem and to see larger themes in your education. These larger themes are the messages I wish I had heard when I was a student; instead, I had to discover them myself through years of grinding in academia. A full list of the commentary themes can be found at the back of the book.

Hint List:

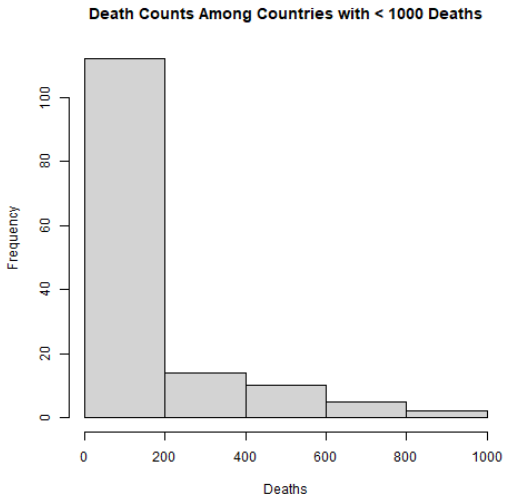
- a: Try searching the internet to learn about WHO Regions. The support is what is *possible*, not just what's in your data set.
- b: By typing ?hist into R, you can see the documentation of the hist function and understand how to alter its parameter list to get what you want. At the end of most documentation files, you can find simple examples showing how things work.
- d: Search "and statement in R" or take a subset of a subset.

Solution:

(a)
- Recovered: Discrete, because we can't have a fractional number of people
  Support: $\{0, 1, 2, \dots\}$ or $\{0, 1, \dots,$ max population of largest country$\}$
- Recovered...100 cases: Continuous
  Support: $[0, 100]$
- WHO.Region: Nominal. The list of values in the data set gives an initial sense for the support, but an internet search gives the complete list of what is possible (which matches what appears in the data set). Note that the alphabetical ordering here does not make the type ordinal since this ordering plays no important role.
  Support: {Africa, Europe, Americas, South-East Asia, Mediterranean, Western Pacific}

(b) The subset operation is useful for reducing the data set:

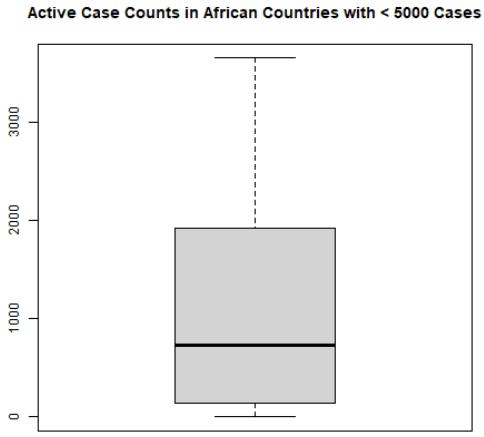
```
tinydata = subset(COVID19Data, Deaths < 1000)
hist(tinydata$Deaths, breaks = 5, xlab = "Deaths",
     main = "Death Counts Among Countries with < 1000 Deaths")
```



(c) The histogram is unimodal and skewed right.

(d) The and operation (&) can speed things up when subsetting:

```
newdata = subset(COVID19Data, WHO.Region == "Africa" & Active < 5000)
boxplot(newdata$Active, main =
        "Active Case Counts in African Countries with < 5000 Cases")
```



Commentary: The support and type of the variable Recovered...100 cases may feel problematic. Notice that this idea is capturing the percent recovered and involves humans, so the answer must be an integer/integer, and hence, a rational number. So, we'll never be able to get an irrational number like π. Because of this, you might think the answer is "Discrete". On the other hand, there certainly isn't space between the possibilities. If one country has a 42.2% recovery rate and another has a 42.3% rate, you could imagine a country with a 42.25% rate. In general, if the support feels like its continuous when you're not being nit-picky, then write "continuous" as your answer.

In fact, you've (possibly) overlooked such nit-pickyness already. If I ask you for the support of how long someone can hold their breath, you might feel like anything non-negative is possible: $[0, \infty)$. But, is there a smallest amount of time? Is time "quantized"? We don't even know! See here for more.

Similarly, if we have a variable for how much water a person consumed yesterday (in grams), you might say $[0, \infty)$ is a sensible support. The ultra-nit-picky person could say: "Water is composed of H_2O molecules. If the mass of a single H_2O molecule is g, then the only possible answers are $\{0, g, 2g, \dots\}$", implying the answer is discrete (although everyone I know would say "continuous").

I find that if we **effectively experience** the support as continuous, then it is sensible to describe it as such and not be pedantic. For example, you experience a movie as a smooth thing, but it is really a series of discrete images shown faster than the human visual complex can individually see.

Finally, in part d, notice how useful it is to do an internet search for a command like and (&). Technical people do this all the time, so you should feel free to as well. Just make sure you recognize the difference between searching for a small syntax element versus searching for the entire solution to a problem.

2. You've been hired as the statistician for a political campaign! Currently, your campaign has received a few donations (in hundreds of dollars): $\{8, 1, 3, 4, 3, 7\}$.

   (a) What are the mean and median donation amounts (in dollars)?

   (b) A wealthy donor would like to make a new donation that raises the *average* donation from its current level to twice that amount. How much should the donor give (in dollars)?

   (c) Using R, find the standard deviation of the original six donations listed above (in dollars).

Hint List:

- Only convert to dollars at the end of each part.

Solution: The donations in order are $\{1, 3, 3, 4, 7, 8\}$.

(a) The median is $\dfrac{3+4}{2} = 3.5$ or $\boxed{\$350}$.

The mean is $\dfrac{1+3+3+4+7+8}{6} = \dfrac{26}{6}$, or \$433.333 or $\boxed{\dfrac{\$1300}{3}}$.

(b) Let x be the new donation (in hundreds). We want

$$\frac{1+3+3+4+7+8+x}{7} = 2 \cdot \frac{13}{3}$$

Thus, $\dfrac{26+x}{7} = \dfrac{26}{3} \implies 78 + 3x = 182$. We get $x = \dfrac{104}{3}$ or \$3466.667 or $\boxed{\dfrac{\$10400}{3}}$.

(c) In R:

```
sd(c(1,3,3,4,7,8))
```

Our answer is about 2.65832 or $\boxed{\$265.832}$.

Commentary: The frustration baked into this problem is that it involves awkward units (hundreds of dollars). You don't always get the units you are comfortable with in life. Notice that I do all the parts in the awkward units and only convert to natural units (dollars) at the end. This problem shows that you may work in the awkward units. It also reveals that certain computations are naturally geared to human computation (e.g., means and medians of small data sets, as in part a), while others are best for technology *even in small data sets* (part c).

3. Let $A = \{-\sqrt{k}, -\sqrt{k-1}, \ldots, -\sqrt{1}, \sqrt{0}, \sqrt{1}, \ldots, \sqrt{k-1}, \sqrt{k}\}$ where k is a positive integer.

   (a) Write all the elements of A when $k = 3$.

   (b) Find the mean of the set A when $k = 3$, and then find the mean of A when k is not known.

   (c) Explain why, in general, A has $2k + 1$ elements.

   (d) Using the results in parts b and c, find the standard deviation of A. Since k is not known, your answer will involve k. Also, simplify your answer so it does not have a summation or ellipsis ($\sum$ or $\ldots$).

   (e) Write a few lines of code in R that create the set A for $k = 100$. Then, have R calculate the standard deviation and compare with the number you get from letting $k = 100$ in your answer from part d. Remember that R is a "vectorized language", which means that if you apply an operation to a vector, it applies it to each spot in the vector. To see this idea, type this code into R:

```
v = c(1,2,3,4)
v^2
```

   (f) Find the IQR of the set A if k is an odd positive integer using the IQR definition shown in class (this definition excludes Q2 from both the lower and upper halves of the data when finding Q1 and Q3).

Hint List:

- c: Just because the expression for A has a k doesn't mean the set has k terms. How many *positive* terms are there?
- d: Try to clean up the numerator under the square root so that it eventually involves the expression $\sum_{i=1}^{k} i$.
- e: Don't try to make the data set A all at once. Can you make the positive terms first? Then, recall that the concatenation operator c(...) can glue together multiple data sets.
- f: If you're struggling to find the middle term of the set $\{\sqrt{1}, \sqrt{2}, \ldots, \sqrt{k}\}$, trying letting k be an odd number (e.g., 5 or 7) and find the middle term by inspection ($\sqrt{3}$ or $\sqrt{4}$). Can you relate 5 to 3? Does the same formula link 7 to 4?

Solution: We have $A = \{-\sqrt{k}, \ldots, 0, \ldots, \sqrt{k}\}$.

(a) When $k = 3$, $A = \boxed{\{-\sqrt{3}, -\sqrt{2}, -\sqrt{1}, 0, \sqrt{1}, \sqrt{2}, \sqrt{3}\}}$.

(b) $\dfrac{-\sqrt{3} - \sqrt{2} - \sqrt{1} + 0 + \sqrt{2} + \sqrt{2} + \sqrt{3}}{7} = 0$. When k is unknown, we get:

$$-\sqrt{k} - \sqrt{k-1} - \cdots - \sqrt{1} + 0 + \sqrt{1} + \cdots + \sqrt{k}$$

in the numerator, which is 0 by symmetry. So, the mean is always $\boxed{0}$.

(c) A will have k *positive* terms: $\sqrt{1}, \sqrt{2}, \ldots, \sqrt{k}$, and k *negative* terms: $-\sqrt{k}, -\sqrt{k-1}, \ldots, -\sqrt{1}$, and 1 zero: 0. This is $k + k + 1 = \boxed{2k + 1}$ terms.

(d)

$$\begin{aligned}
\text{SD} &= \sqrt{\frac{(-\sqrt{k} - 0)^2 + \cdots + (-\sqrt{1} - 0)^2 + (0 - 0)^2 + (\sqrt{1} - 0)^2 + \cdots + (\sqrt{k} - 0)^2}{(2k + 1) - 1}} \\
&= \sqrt{\frac{k + (k-1) + \cdots + 1 + 0 + 1 + \cdots + k}{2k}} \\
&= \sqrt{\frac{2 \cdot \sum_{i=1}^{k} i}{2k}} = \sqrt{\frac{\frac{k(k+1)}{2}}{k}} = \boxed{\sqrt{\frac{k+1}{2}}}.
\end{aligned}$$

(e) In R:

```
k = 100
posValues = sqrt(1:k)
A = c(-posValues, 0, posValues)
sd(A)
sqrt((k+1)/2)
```

Since we get the same value (about 7.106) for the final two lines, it seems like our formula is correct!

(f) 0 is the median of A. So, the upper half of the data is $\{\sqrt{1}, \sqrt{2}, \ldots, \sqrt{k-1}, \sqrt{k}\}$. If k is odd, this list has an odd number of terms, so its median is $\sqrt{\dfrac{k+1}{2}}$.

A simple example can help convince you: If $k = 5$, the upper half is $\sqrt{1}, \sqrt{2}, \sqrt{3}, \sqrt{4}, \sqrt{5}$. We can visually see this has median $\sqrt{3}$, which is $\sqrt{\dfrac{5+1}{2}}$, or $\sqrt{\dfrac{k+1}{2}}$ in general.

So, $\text{Q3} = \sqrt{\dfrac{k+1}{2}}$. By symmetry, $\text{Q1} = -\sqrt{\dfrac{k+1}{2}}$. Thus,

$$\text{IQR} = \text{Q3} - \text{Q1} = \sqrt{\frac{k+1}{2}} - -\sqrt{\frac{k+1}{2}} = \boxed{2\sqrt{\frac{k+1}{2}}}.$$

Commentary: The goal of this problem is to improve your skill with notation and summation formulas. Objects like the set A are common in higher math courses, so it helps to make sense of them. Some things might change about these objects for different values of k (e.g., how many elements are present), but some things might stay the same (e.g., the mean is always 0). Also, try to get comfortable with the types of calculations that appear in part d and the type of thinking in part f (e.g., expressing the middle element of the set $\{\sqrt{1}, \sqrt{2}, \ldots, \sqrt{k}\}$ in terms of k). Also, it is time to memorize these summation formulas:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \qquad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}, \qquad \sum_{i=1}^{n} i^3 = \left(\frac{n(n+1)}{2}\right)^2$$

which can be proved using induction.

4. Let $B = \{1, x, -1, 0, x^2\}$ where x is a real number and the set B is not necessarily in order. What are all the possible values for x that cause the mean and median of B to be equal? Give exact answers.

   (Note: The notation for the set B is meant to suggest that B always has 5 elements and that repeated elements are OK. For example, if x were the value 1, the set would be $\{1, 1, -1, 0, 1\}$ and this would NOT collapse to $\{-1, 0, 1\}$ as some people might be tempted to do.)

Hint List:

- Try cases. One case you might consider is $0 < x < 1$. What is the order of the elements in the set for this case? What are additional cases to consider?

Solution: This problem is all about attention to detail. Let $B = \{1, x, -1, 0, x^2\}$. Squaring a number behaves differently if it is < 1 versus ≥ 1 and positive versus negative. So, we consider 4 cases for x:

Case 1: $x \in (-\infty, -1]$. In order, our set is $\{x, -1, 0, 1, x^2\}$. So, the median is 0. The mean, regardless of order, is $\dfrac{x-1+0+1+x^2}{5} = \dfrac{x+x^2}{5}$. So, in this case, we need

$$\frac{x^2+x}{5} = 0 \implies x^2 + x = 0 \implies x(x+1) = 0 \implies x = 0 \text{ or } x = -1.$$

Since $x \leq -1$ in this case, we toss the $x = 0$ answer and get $\boxed{x = -1}$.

Case 2: $x \in (-1, 0]$. Here, the order is $\{-1, x, 0, x^2, 1\}$. The median is still 0 , so $\dfrac{x^2+x}{5} = 0$ gives $x = 0, -1$ as before, and now we get to keep $\boxed{x = 0}$.

(Notice that since cases 1 and 2 have the same median, we could combine them into a single case.)

Case 3: $x \in (0, 1)$. Now, the order is $\{-1, 0, x^2, x, 1\}$. So,

$$\frac{x^2+x}{5} = x^2 \implies x^2 + x = 5x^2 \implies 4x^2 - x = 0 \implies x(4x-1) = 0 \implies x = 0 \text{ or } x = \frac{1}{4}.$$

Only $\dfrac{1}{4}$ is in the zone $(0, 1)$, so $\boxed{x = \dfrac{1}{4}}$.

Case 4: $x \in [1, \infty)$. The order is $\{-1, 0, 1, x, x^2\}$. We get $\dfrac{x^2+x}{5} = 1 \implies x^2 + x - 5 = 0$. By the Quadratic Formula, $x = \dfrac{-1 \pm \sqrt{1^2 - 4(1)(-5)}}{2} = \dfrac{-1 \pm \sqrt{21}}{2}$. So, $\boxed{x = \dfrac{-1+\sqrt{21}}{2}}$ since we must stay in the interval $[1, \infty)$.

Commentary: I included this problem because it is best solved by using cases. Students are often bothered by this, and some come to me seeking out an elegant solution pathway that avoids this inconvenience. In truth, casework is often needed for more complex problems, and in this problem, it *makes sense* why casework is probably required: The ordering of the set changes based on the nature of x, and so there shouldn't be a one-size-fits-all solution. Note that another solution to this problem involves setting the mean equal to each value v in the data set, getting x, and then checking if this value of x forces v to be the median. This approach also involves casework (one case for each value of v). As the data set grows in size, this alternative solution is worse than the above solution which always has only 4 cases.

5. A box has 7 green balls and 3 red balls. Suppose you draw two balls *with replacement* and record their colors (see lecture slides for the meaning of this term; in this problem, it means we draw one ball, record the color, *return the ball to the box*, draw a second ball, record the color, and return it to the box).

   (a) Why are the colors you see on the two balls you draw independent of one another?

   (b) What is the probability your first draw is red and your second draw is green? [WISE] (See lecture slides for an explanation of what WISE means.)

   (c) What is the probability the colors don't match?

   (d) What is the answer to part c if the balls are drawn *without replacement*?

Hint List:

- b: Part a reveals the independence of draws 1 and 2. What probability rule can be used with independent events?
- c: There are two ways for the colors to not match. What are they?
- d: Note that colors of the draws are now dependent on one another (why?). What rule do we use in such settings?

Solution:

(a) Because you replace the first ball after drawing it, the box is returned to its initial state, and thus, the color of the second ball is unaffected by the first.

(b) Let A_i be the event that draw i is red (WIS).

We want (E) $P\left(A_1 \cap A_2^C\right) = P\left(A_1\right) \cdot P\left(A_2^C\right) = \frac{3}{10} \cdot \frac{7}{10} = \boxed{\frac{21}{100}}$.

(c) $P(\text{don't match}) = P([A_1 \cap A_2^C] \cup [A_1^C \cap A_2]) \overset{\text{disj.}}{=} P(A_1 \cap A_2^C) + P(A_1^C \cap A_2)$

$$\overset{\text{ind.}}{=} P(A_1)P(A_2^C) + P(A_1^C)P(A_2) = \frac{3}{10} \cdot \frac{7}{10} + \frac{7}{10} \cdot \frac{3}{10} = \frac{42}{100} = \boxed{\frac{21}{50}}.$$

(d) Here, the colors of the draws are dependent, so we can't do the above step that requires independence. Instead, we do:

$$P(\text{don't match}) = P(A_1 \cap A_2^C) + P(A_1^C \cap A_2) \overset{\text{dep.}}{=} P(A_2^C \mid A_1)P(A_1) + P(A_2 \mid A_1^C)P(A_1^C)$$
$$= \frac{7}{9} \cdot \frac{3}{10} + \frac{3}{9} \cdot \frac{7}{10} = \frac{42}{90} = \boxed{\frac{7}{15}}.$$

Commentary: The main goal of this problem is to begin setting up events using WISE and to distinguish between independent and dependent events (parts c and d). The uninspired setting of balls-in-a-box was chosen to focus your attention on the probabilistic properties at hand (dependence, independence, disjointness). It is critical to see these ideas from the start; they guide which rules you use and simplify complex problems. You need not write the letters of WISE in your solution or clarifications like "disj.", "ind.", or "dep." over equal signs, but you should be thinking them at all times.

A word of caution: If problems like this one feel easy to you, it is tempting to race through them and start writing numbers all over the page. While such behavior might serve you well on this problem, it won't help you on more difficult problems. You need to spend time carefully transitioning a problem from the written world to the mathematical world. Notice how much of the above solutions *aren't* numbers. WISE is here to help slow down that transition and to focus your attention on the non-numeric aspects of the problem.

6. Suppose you are rolling a fair, eight-sided die (labeled $1 - 8$) twice.

   (a) Find the probability that the absolute value of the difference of the results equals 2. [WISE]

   (b) Find the probability the second roll is at least 2 greater than the first roll. [WISE]

   (c) Suppose the die is a fair, k-sided die (labeled $1 - k$) where $k \geq 3$. If you roll it twice, find the probability the second roll is at least 2 greater than the first roll. Your answer will depend on k and should not have a summation or ellipsis. Also, plug $k = 8$ into your formula and check that it gives your answer from part b.

Hint List:

- a: Try letting A be the event that die 1 gives us a 1, 2, 7, or 8. Then, Let B be the event that die 2 is 2 away from die 1.
- b: Try listing all the outcomes (as ordered pairs) that make up the event we care about. Then, *organize* them so they are easier to count.
- c: Try a listing-and-organization approach as in part b, but use k in your list instead of 8.

Solution:

(a) Let X_1, X_2 be the results of the rolls. Let A is the event that $|X_1 - X_2| = 2$ (WIS). Note that there are $64 = 8 \cdot 8$ outcomes, which we write as (X_1, X_2). We can list all the outcomes in A:

$$A = \{(1,3),(2,4),(3,1),(3,5),(4,2),(4,6),(5,3),(5,7),(6,4),(6,8),(7,5),(8,6)\}$$

So, (E) $P(A) = \dfrac{12}{64} = \boxed{\dfrac{3}{16}}$.

Looking at the above solution, you'll notice the WISE framework isn't doing much. I define A as the very event the problem wants but end up listing everything. WISE is at its best when it *splits up the work of the problem* some. Consider this setup:

Let A be the event that die 1 gives us a $1, 2, 7$, or 8 (WIS).
Let B be the event that die 2 is 2 away from die 1 (WIS).

$$\begin{aligned}\text{We want (E) } P(|\text{difference}| = 2) &= P\left([B \cap A] \cup \left[B \cap A^C\right]\right) = P(B \cap A) + P\left(B \cap A^C\right) \\ &= P(B \mid A) \cdot P(A) + P\left(B \mid A^C\right) \cdot P\left(A^C\right) \\ &= \frac{1}{8} \cdot \frac{4}{8} + \frac{2}{8} \cdot \frac{4}{8} = \frac{12}{64} = \boxed{\frac{3}{16}}, \text{ as above.}\end{aligned}$$

The clever idea here is to separate those cases for die 1 that only have 1 possible match (e.g., a 2 on die 1 must pair with a 4 on die 2) versus those where two matches are possible (e.g., a 5 on die 1 could pair with a 3 or 7 on die 2). In general, if you find yourself using the and/or rules on the events you defined, this is a sign you have deployed WISE in a really helpful way.

(b) Let B be the event that $X_2 \geq X_1 + 2$ (WIS). We list the events in B:

$$\begin{aligned}\{(1,3),(1,4),(1,5),\ldots,(1,8), &\leftarrow 6 \text{ pairs} \\ (2,4),(2,5),\ldots,(2,8), &\leftarrow 5 \text{ pairs} \\ &\cdots\cdots \\ (6,8)\} &\leftarrow 1 \text{ pair}\end{aligned}$$

So, B has $1 + 2 + 3 + 4 + 5 + 6 = 21$ pairs. Thus, (E) $P(B) = \boxed{\dfrac{21}{64}}$.

(c) With a k-sided die, we generalize our approach from part b. Now, B is:

$$\{(1,3),(1,4),(1,5),\ldots,(1,k), \leftarrow k-2 \text{ pairs}$$
$$(2,4),(2,5),\ldots,(2,k), \leftarrow k-3 \text{ pairs}$$
$$\ldots\ldots$$
$$(k-2,k)\} \leftarrow 1 \text{ pair}$$

So, B has $1+2+\cdots+(k-3)+(k-2) = \sum_{i=1}^{k-2} i = \frac{(k-2)(k-1)}{2}$. Since there are $k \cdot k = k^2$ possible pairs, $P(B) = \boxed{\frac{(k-2)(k-1)}{2k^2}}$.

Notice that when $k = 8$, we get $P(B) = \frac{6 \cdot 7}{2 \cdot 64} = \frac{21}{64}$, as expected from part b.

Commentary: This is another problem where cases are at the heart of what is happening. As part a reveals, a roll of 1, 2, 7, or 8 is different than 3, 4, 5, or 6. Don't be afraid of cases when they lie at the heart of the problem's structure. Parts b and c are present to see if you can generalize from a particular value (8) to a fixed unknown (k). This is an important skill, for most of the time we seek general patterns in life, not specific ones. Also, notice that general patterns carry much more information: The answer from part c can not only provide the answer for part b as a special case, but it can also give an answer for any other die one wishes to roll! Finally, this problem also gives you another chance to show your knowledge of summation formulas. When an idea makes repeated appearances, this is a sign you should deeply internalize it. There may be things you saw at the high school level for which you thought "I don't need to memorize this; it feels obscure." As obscurity recedes, let internalization take its place.

7. Let's use R to check your answer from 6a! For each part, explain what this code does in relation to problem 6a by typing the code into R and making sense of the output.

(a)
```
sample(x = 1:8, size = 10000, replace = T)
```

(b)
```
diffs = sample(x = 1:8, size = 10000, replace = T) -
        sample(x = 1:8, size = 10000, replace = T)
```

(c)
```
vals = (abs(diffs) == 2)
```

(d)
```
mean(vals)
```

(e) Programmers often love one-line solutions to problems. We can do all the work above using the single line of code:

```
mean(replicate(10000, abs(diff(sample(1:8, 2, T))) == 2))
```

Hint List:

- Feel free to look up any functions or operations involved. This is a natural part of learning any language. Also, remember that R Studio has the R documentation files built in. For example, type ?replicate if you want to read about the replicate function.
- d: R interprets TRUE as 1 and FALSE as 0. Type TRUE+TRUE in R to be convinced.

Solution:

(a) This simulates 10000 dice rolls of our fair, eight-sided die. I chose 10000 because it was a large number. In general, increasing the sample size by a factor of 100 reduces the effects of randomness by a factor of $\sqrt{100} = 10$, a fact you learn later.

(b) The first sample is imitating 10000 first rolls (values of X_1). The second sample imitates the second rolls (values of X_2). By subtracting the vectors, we take advantage of R's vectorized abilities to create 10000 differences (values of $X_1 - X_2$).

(c) Here, we check which of our differences has an absolute value of 2, returning a list of TRUEs and FALSEs.

(d) This finds the percentage of TRUEs out of the 10000 simulated pairs of differences. I got about $0.187 \approx \frac{3}{16}$ as expect from 6a, but answers will vary.

(e) Here, we roll the die (the 1:8 in sample) twice (the 2 in sample), subtract the results (diff), check if the absolute value is 2, repeat this 10000 times (replicate), and find the proportion of TRUEs (mean).

Commentary: The goal of this problem was to introduce you to/remind you of some of the most useful features of R. My favorite feature of R is its vectorized nature, which didn't exist in early programming languages. Also, I wanted to provide an interesting contrast to you. Parts a through d slowly build toward a solution, allowing the reader to make sense of what is happening. Part e, in contrast, presents a single-line solution. While it may seem impressive to write a one-line solution to a problem, doing so is more about the bravado of the coder than about communicating with the reader. I believe in prioritizing communication (via clarity and cogency) above hubris and hope you will as well. You'll see similar examples of this behavior in other subjects as well (e.g., claiming things are "obvious" or "trivial" in a math class, or believing "big" words improve your essay writing when simple, well-known synonyms exist).

8. In each problem below, you are given two events. First, decide if the events are disjoint or not disjoint. After this, decide if the events are independent or dependent. Explain your reasoning.

   (a) Exactly two candidates are running to be president of the US. A is the event that the first candidate wins the election. B is the event that the second candidate wins the election. (Assume ties do not occur.)

   (b) You roll a fair, four-sided die (labeled $2 - 5$). A is the event you roll an even number. B is the event you roll a prime number.

   (c) Your friend randomly picks an integer in the set $\{1, 2, 3, \ldots, 20, 21\}$, where each choice is equally likely. A is the event that the number is divisible by 3. B is the event that the units digit of the number chosen is "1".

Hint List:

- Return to the definitions of the words "disjoint" and "independent". Have you memorized these definitions?
- c: While you might have an intuitive feeling for (in)dependence, this part is best solved by returning to the definition of independence. Check if $P(A) = P(A \mid B)$.

Solution:

(a) Since it is not possible for A and B to both occur, $A \cap B = \emptyset$, and thus, the events are $\boxed{\text{disjoint}}$.
Now, $P(A) =$ some positive value, but $P(A \mid B) = 0$ because once we know B has won, A cannot win. since $P(A) \neq P(A \mid B)$, the events are $\boxed{\text{dependent}}$.

(b) Note that $A \cap B = \{\text{even prime number}\} = \{2\}$. So, the events are $\boxed{\text{not disjoint}}$.
Now, $P(A) = P(\text{even}) = \dfrac{|\{2,4\}|}{|\{2,3,4,5\}|} = \dfrac{1}{2}$.
But, $P(A \mid B) = P(\text{even} \mid \text{prime}) = \dfrac{|\{2\}|}{|\{2,3,5\}|} = \dfrac{1}{3}$.
Since $P(A) \neq P(A \mid B)$, the events are $\boxed{\text{dependent}}$.

(c) Here, $A \cap B = \{21\} \implies$ the events are $\boxed{\text{not disjoint}}$.
Now, $P(A) = \dfrac{|\{3,6,9,\ldots,21\}|}{21} = \dfrac{7}{21} = \dfrac{1}{3}$, and $P(A \mid B) = \dfrac{|\{21\}|}{|\{1,11,21\}|} = \dfrac{1}{3}$.
Thus, A and B are $\boxed{\text{independent}}$.

Commentary: Notice that when you ask about both disjointness and independence, four $(2 \cdot 2)$ outcomes are possible (disjoint and dependent, disjoint and independent, not disjoint and dependent, not disjoint and independent). This problem shows you examples of *three* of the four possibilities. The reason I created this question was to inspire you to ask about the fourth: "Is it possible to define two events that are disjoint and independent?" One key component of becoming a scholar is learning to ask questions based on what you are seeing. Work to develop curiosity, beginning with this question.

9. Let A and B be events on the same sample space. Suppose you know that $P(A) = 0.3$ and $P(B^C) = \frac{1}{4}$. Can events A and B be disjoint? Give an example of this occurring, or argue that it is not possible.

Hint List:

- Imagine for a moment that A and B are disjoint. What rule is available to us when events are disjoint?

Solution:

If $P\left(B^C\right) = \frac{1}{4}$, then $P(B) = \frac{3}{4} = 0.75$. If A and B were disjoint, then we would have

$$P(A \cup B) = P(A) + P(B) = 0.3 + 0.75 = 1.05,$$

but no event can have probability greater than 1. So, A and B are $\boxed{\text{not disjoint}}$.

Commentary: This problem is most easily solved using counterfactual reasoning (known as "proof by contradiction" in math). Essentially, you assume something (here, that A and B are disjoint), and you see what consequences must logically follow. If you arrive at an absurdity despite using valid logical deduction, then the only possibility is that your initial assumption is wrong. There are many types of proof in mathematics (direct proof, proof by contradiction, proof by induction, etc.). UCSD's Math 109 covers these in more detail. Note that humans informally use counterfactual reasoning all the time: "If the defendant had committed the crime, then they would have been in city X on day Y. However, a witness places them in city Z on day Y. Hence, we have an absurdity." One great reason to learn mathematics is that it formalizes these everyday ways of reasoning and improves your ability to use them.

Indeed, each subject has something to offer you beyond the content we usually think of as the main reason for learning a subject. For example, on the surface, a violin performance major is someone learning to play the violin and working through the major compositions for the instrument. But this field of inquiry teaches you far more. When I was a music major, I learned how to be part of a team (when playing chamber or orchestral music), how to set and achieve long term goals (you don't get good in one month or even one year), and how to connect mental goals with physical actions (when trying to realize a musical vision for a piece). What are the beyond-the-surface skills that your areas of interest are developing?

10. Suppose you have a coin where $P(\text{heads}) = p$.

(a) Four friends – A, B, C, and D – take turns, in order $(A, B, C, D, A, B, C, D, A \ldots)$, flipping the coin until someone finally gets heads and is declared the winner. Find the probability that A wins. Your answer will depend on p and should not have a summation or ellipsis. Also, what value should your formula give if $p = 1$ and why?

(b) Find the probability that A wins if there are k participants instead of 4. Your answer will depend on p and k, and it should not have a summation or ellipsis. Also, what value should should your formula give if $k = 1$ and why?

Hint List:

- a: Let W_i be the event that a winner is declared on flip i (and not before) for $i = 1, 2, 3, \ldots$ Using these events, write all the ways player A can win and go from there.
- a: Note that player A can win in infinitely many ways. A might win immediately. A might win on the fifth flip of the coin (what face comes up on each of the five flips?). What are all the other cases?
- b: You don't need to completely redo your argument from part a. Just explore what pieces of it change when we go from four people to k people.

Solution:

(a) A can win in many ways:

- A immediately flips heads
- A through D all flip tails, then A flips heads
- A through D all flip tails, then all 4 flip tails again, then A flips heads
- etc.

There are multiple ways to set up the events. For example, we might set H_i as the event that flip i is heads. Here, our different cases for A winning are:

- H_1
- $H_1^C \cap H_2^C \cap H_3^C \cap H_4^C \cap H_5$
- $H_1^C \cap \cdots \cap H_8^C \cap H_9$
- etc.

Note that $H_i^C = 1 - p$ and we have calculations like

$$P(H_1^C \cap H_2^C \cap H_3^C \cap H_4^C \cap H_5) = P(H_1^C) \cdot P(H_2^C) \cdot P(H_3^C) \cdot P(H_4^C) \cdot P(H_5) = (1-p)^4 p$$

Adding these gives the probability A wins:

$$p + (1-p)^4 p + (1-p)^8 p + \ldots$$

This is an infinite geometric series with common ratio $r = (1-p)^4$. Hence, the sum is

$$\frac{\text{first term}}{1 - \text{common ratio}} = \boxed{\frac{p}{1-(1-p)^4}}$$

When $p = 1$, the formula gives 1, which makes sense because the game always ends on the first flip with A winning!

We also could have used the events given in the hint. Here, the cases for A winning are:

- W_1
- W_5
- W_9
- etc.

Thus, the probability of A winning is

$$P(W_1 \cup W_5 \cup W_9 \cup \dots) = P(W_1) + P(W_5) + P(W_9) + \dots$$

This notation has the advantage of making the cases look simpler. The downside is that to find $P(W_5)$ we must think carefully about what it means:

$$P(W_5) = P(A \text{ gets tails} \cap B \text{ gets tails} \cap C \text{ gets tails} \cap D \text{ gets tails} \cap A \text{ gets heads})$$

This calculation leads to $(1-p)^4 p$, and similar calculations give the same sum as before.

(b) With k participants, we can follow similar logic and get

$$P(A \text{ wins}) = \frac{p}{1-(1-p)^k}$$

Note that if $k = 1$, $P(A \text{ wins}) = \dfrac{p}{1-(1-p)} = 1$, which makes sense since A is the only player, and hence, must win!

Commentary: There were several goals for this problem. First, we have the casework. Don't be afraid to articulate the cases one-by-one, even if there are infinitely many. Know that when infinity cases are present, you'll have an infinite sum when switching to probabilities. Luckily, your previous coursework in sequences/series has given you a toolkit for adding up infinite series. The easiest such sums are infinite geometric series. These will reappear frequently in this course.

Another goal was to show you the value of "particularization" or "specialization". The answers to parts a and b are very general: part a works for any $p \neq 0$ and part b works for any $p \neq 0$ and integral $k \geq 1$. When you have such formulas, you should always check they make sense when you choose *particular* values of the parameters that reduce the situation to one that is immediately understandable. As another example, in Geometry class you can prove that the length of the diagonal, d, across a box of dimensions a, b, and c satisfies

$$d^2 = a^2 + b^2 + c^2$$

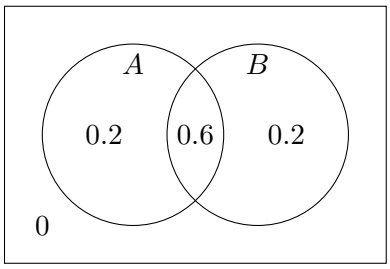
If you restrict this general formula to the specific case where the height, c, of the box is 0, we get $d^2 = a^2 + b^2$, the Pythagorean Theorem. This makes sense because the box has collapsed to a rectangle and the 3D diagonal becomes the hypotenuse of a right triangle with legs a and b. Always check that your general solutions collapse as expected in particular cases that trivialize the setting.

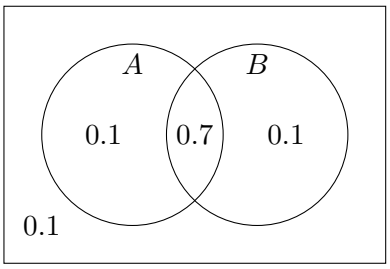
11. Suppose you have two events, A and B, defined on the same sample space with $P(A) = P(B) = 0.8$. What is the smallest value $P(A \cap B)$ could possibly be? [VIZ]

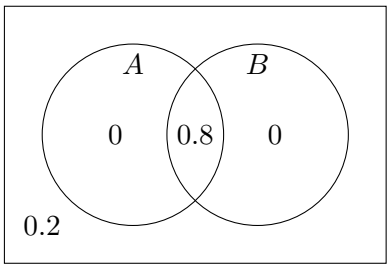
Hint List:

- Use visualizations to help you find the answer, and then write a mathematical argument that shows $P(A \cap B)$ is always greater than or equal to your smallest value.
- Draw a Venn diagram where $P(A \cap B) = 0.7$ and fill in all the area values (including outside the circles). Now create some other pictures. What do you notice?

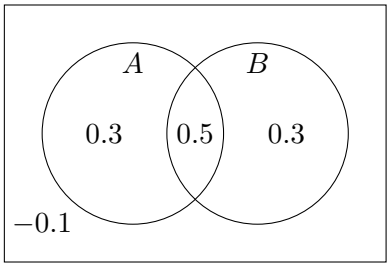
Solution: Consider some different possibilities:







In this last picture A actually equals B. Note that if we try to create less than 0.6 in the intersection, we get a picture like:



Here, the circles have a combined area, $P(A \cup B)$, greater than 1, forcing a negative result for $P((A \cup B)^C)$. Thus, it seems $\boxed{0.6}$ is the smallest $P(A \cap B)$ can be.

If we wish to move beyond pictures, recall that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Since $A \cup B$ is an event, $P(A \cup B) \leq 1$. So,

$$P(A) + P(B) - P(A \cap B) \leq 1 \implies P(A) + P(B) - 1 \leq P(A \cap B)$$

Since $P(A) = P(B) = 0.8$, we have

$$1.6 - 1 \leq P(A \cap B) \implies 0.6 \leq P(A \cap B)$$

yielding the same result as before.

Commentary: While this problem is designed to get you thinking about unions and intersections, it also shows the importance of drawing pictures and exploring during the solving process. When educational researchers study the differences between novices and experts in math, one key finding is that experts are more willing to draw pictures and explore the problem setting in simple cases before trying to solve the problem in its generality.

In addition, this problem is a nice example where you get the answer through a few simple pictures

but may worry whether your answer is right ("Did I miss a case?") or whether pictures qualify as a valid solution (for earning credit in school or convincing others). This is a complex question because the answer changes based on the setting and audience. Roughly speaking, in very rigorous settings (e.g., articles in mathematics journals), pictures can help support a solution, but rarely are sufficient. In very relaxed settings or audiences unfamiliar with proof, pictures can be quite convincing. If you find yourself looking for the tools to write more convincing arguments, Math 109 is your next step.

12. Suppose we have two events, A and B, with $P(A) > 0$ and $P(B) > 0$. We said "A is independent of B" precisely when $P(A \mid B) = P(A)$. Show, using a mathematical argument, that if A is independent of B, then B is independent of A. (This problem shows that we don't have to be careful when discussing independence since the order in which we mention things is irrelevant.)

Hint List:

- We are given that $P(A \mid B) = P(A)$. Write a similar statement that encodes what you are trying to prove.
- What theorems do you know about conditional probability?

Solution: We know A is independent of B. That is, $P(A \mid B) = P(A)$. Now,

$$P(B \mid A) \overset{\text{Bayes}}{=} \frac{P(A \mid B)P(B)}{P(A)} \overset{\text{given}}{=} \frac{P(A)P(B)}{P(A)} = P(B)$$

So, $P(B \mid A) = P(B)$. That is, B is independent of A. Note that we need $P(A) > 0$ to avoid division by 0 and to allow cancellation in the above steps. In general, we tend not to worry about events with 0 probability, so the restrictions that $P(A) > 0$ and $P(B) > 0$ are not concerning.

Commentary: There were two goals with this problem. The first was to provide an opportunity for writing a simple deductive proof (again, consider Math 109 if you enjoy this!).

The second was to show that event independence is a "symmetric property". A property (or, in math, a "relation") R is symmetric if $a\,R\,b$ (a is related to b) forces $b\,R\,a$ (b is related to a). In math/stats, some relations are symmetric, and some are not. For example, equality is symmetric (if $x = y$, then $y = x$), but "less than" is not ($2 < 3$ does not force $3 < 2$). In real life, some relations are symmetric ("is a sibling of") and asymmetric ("is attracted to"). When we think about event independence, it seems to distill the property "has no connection to", which feels like a symmetric property based on our experience in the real world. Here, we prove that symmetry.

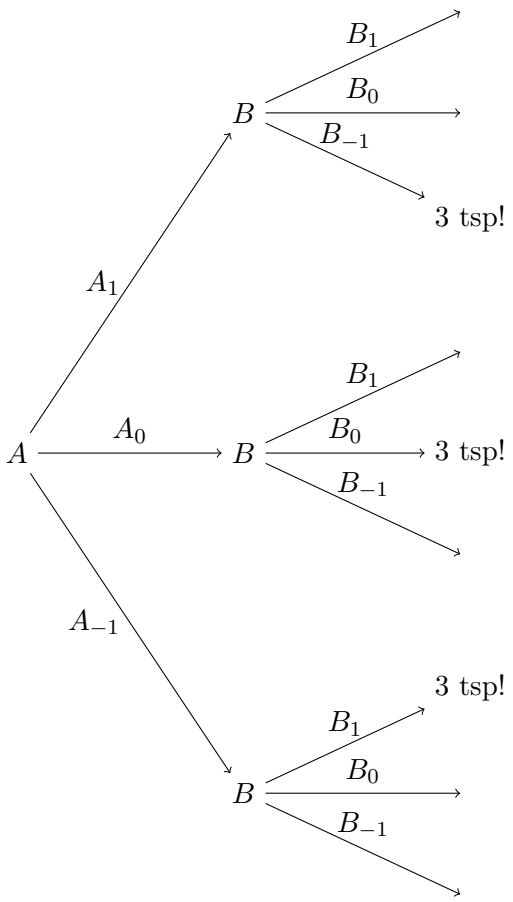
One of my favorite things about math is that it offers language for behaviors that occur frequently. For example, this problem is about proving the "*symmetry* of the independence *relation* for events with non-zero probability". You may have already seen some other examples. When you first learn addition and multiplication, you are taught that $a + b = b + a$ and that $a \cdot b = b \cdot a$. These rules imply that the order in which you do the operation doesn't matter. That is, you may move around the numbers if you want. As such, these operations are said to have the *commutative property* (to "commute" is to move things, like commuting to school). Once you have this language, it begins to structure how you see the world. For example, you might now notice that event union and intersection are *commutative*. That is, $A \cup B = B \cup A$ and $A \cap B = B \cap A$. What other operations are commutative? Do you know any non-commutative operations?

13. Suppose there is a recipe that calls for 3 tsp of cinnamon. Since people have strong feelings about this spice, it is common for people to adjust the recipe after trying it. Suppose that no matter who uses the recipe or how much cinnamon it calls for, there is a 30% chance the user will increase the cinnamon by 1 tsp, a 50% chance they leave the cinnamon unchanged, and a 20% chance they decrease the cinnamon by 1 tsp. Also, assume people's adjustments are independent. Suppose person A uses the recipe and makes their alteration, and then passes their altered recipe to person B. What is the probability that after person B has made their alteration the recipe calls for the original amount of cinnamon, 3 tsp? (Here, the word "alteration" includes the possibility that a person will leave the cinnamon amount unchanged.) [WISE, VIZ]

Hint List:

- Imagine you are person B and your goal is to keep the cinnamon level unchanged. How would you respond to counteract the alteration made by A? List all the cases.

Solution: Let A_1 (A_0, A_{-1}) be the event that person A changes the cinnamon by 1 ($0, -1$) tsp. Define B_1, B_0, B_{-1} similarly. We can represent the possible changes in a tree diagram:



$$\begin{aligned} P(\text{3 tsp at end}) &= P([A_1 \cap B_{-1}] \cup [A_0 \cap B_0] \cup [A_{-1} \cap B_1]) \\ &\overset{\text{disj.}}{=} P(A_1 \cap B_{-1}) + P(A_0 \cap B_0) + P(A_{-1} \cap B_1) \\ &\overset{\text{ind.}}{=} P(A_1)P(B_{-1}) + P(A_0)P(B_0) + P(A_{-1})P(B_1) \\ &= 0.3 \cdot 0.2 + 0.5 \cdot 0.5 + 0.2 \cdot 0.3 = \boxed{0.37}. \end{aligned}$$

Commentary: Some people believe that a tree node can only have two branches because every tree diagram they've ever seen has that appearance. So, one goal of this problem was to erode that false belief. Another goal was to show you that trees can be useful even when the events are *independent*. Typically, trees are the best choice when we have conditional probabilities (usually signifying dependent

events). Here, it makes sense to use a tree because person A passes the result on to person B. The consecutive nature of the cinnamon changes is well-suited to a tree. I was also hoping that you might be bothered by the independence assumption built into the problem. If 100 people before you have increased the cinnamon level, are you really just as likely to increase the cinnamon as if no one before you increased it? Or, if 3 people before you have all decreased the cinnamon (down to 0 tsp), is there still a 20% chance you'll decrease it (to -1 tsp)?! There will be several times in this class where we assume independence to make calculations easier or possible. In some of these cases, you might have genuine concerns about the assumption of independence. This is good; it means you're thinking carefully about the real world.

14. Let's check your answer to question 13. We'll do so via computer simulation (often called a "Monte Carlo" simulation; see the naming history here). For each line of code below, explain what it does, and how it relates to problem 13.

(a)
```
cinnChanges = sample(c(1,1,1,0,0,0,0,0,-1,-1),2, replace = T)
```

(b)
```
netChange = sum(cinnChanges)
```

(c)
```
manyChanges = replicate(100000, sum(sample(c(1,1,1,0,0,0,0,0,-1,-1),2,
                        replace = T)))
```

(d)
```
result = mean(manyChanges == 0)
```

Hint List:

- Type the code into R and see what is produced! Also, consult R's documentation and use internet searches.

Solution:

(a) By sampling from the given vector, there's a $\frac{3}{10} = 30\%$ chance of getting a $+1$, a $\frac{5}{10} = 50\%$ chance of getting a 0, and a $\frac{2}{10} = 20\%$ chance of getting a -1. These are the same odds of the recipe alterations. We sample twice because of A and B, who both adjust, and set the replace parameter to true because of the independence assumption.

(b) Summing the results gives the net cinnamon change.

(c) Here, we simulate 100000 recipe transfers, each across two bakers, keeping track of the cinnamon change totals.

(d) Setting manyChanges equal to 0 returns a vector of TRUEs (1s) and FALSEs (0s) based on whether the pair of bakers got the cinnamon level back to 3 tsp (i.e., a net change of 0). The mean command adds all the 1s and 0s and divides by how many (100000). When you add the 1s and 0s, the result is simply the number of 1s (i.e., the number of unchanged recipes). Thus, the mean finds:

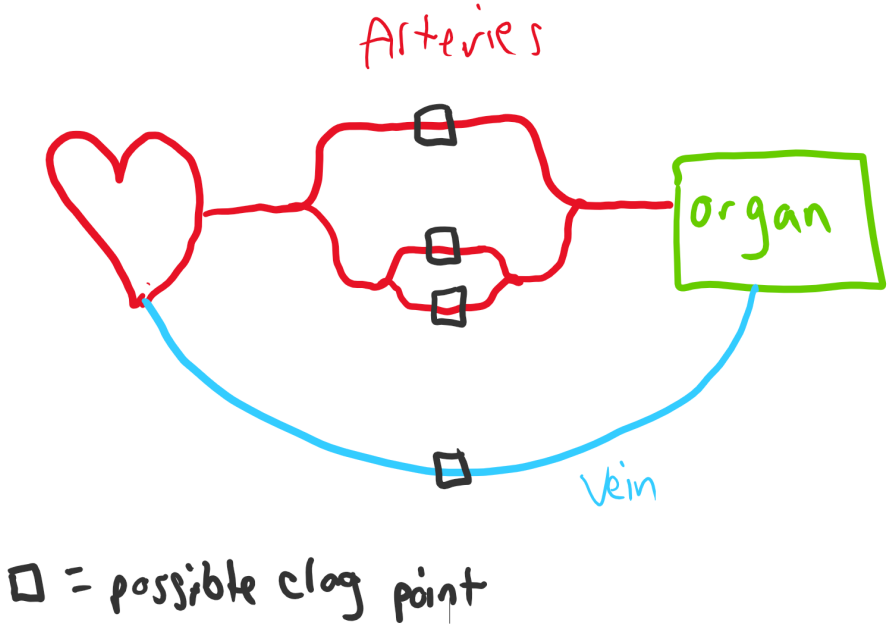
$$\frac{\text{number of unchanged recipes}}{100000} = \text{percent of unchanged in our 100000} \approx P(\text{unchanged recipe})$$

Commentary: Again, we have a problem that allows us to quickly check the result of a probability calculation via simulation. This is an important skill, especially on harder problems. As an example, I recently solved this problem on Project Euler, an online site of interesting math/stat/programming challenges (a huge range of difficulties are present; start with problems at the 5% difficulty!). I had solved the problem theoretically but was unsure of my answer. So, I wrote a simulation and discovered something was wrong. There are entire books of math/stat problems that are well-suited to theoretical analysis *and* computer simulation (e.g., see Paul Nahin's "Duelling Idiots and Other Probability Puzzlers").

Finally, if you're bothered by the makeshift approach of part a, then fear not. R has a better way of doing this that is especially useful when more complicated probabilities are present:

```
sample(c(1, 0, -1), 2, prob=c(0.3, 0.5, 0.2))
```

15. In October 2020, my friend had a heart attack and was in the hospital for several days. This got me thinking about bloodflow in the human body. Below is a simplified model of blood leaving the heart through arteries to an organ, and then returning to the heart via a vein. Suppose that at *each* of the three points marked on the arteries, the chance of a clog that stops bloodflow through that particular path is p_1. Suppose the probability the vein is clogged is p_2.



(a) In this problem, we'll assume that the states of the possible clog points are all independent to make calculations easier. Give a real-world reason why this is a silly assumption.

(b) For the simplied system to work, we just need blood to reach the organ through any of the arterial paths, and for blood to be able to return via the vein. Find the probability the system works. Your answer will involve p_1 and p_2. [WISE]

(c) You can (partially) check your answer to part b by seeing what value it gives in three interesting cases: (i) $p_1 = 0, p_2 = 0$, (ii) $p_1 = 1$, and (iii) $p_1 = 0, p_2 = \frac{1}{2}$. For each case, evaluate your result from part b and then explain why your answer makes sense given what the values say about the simplified system.

Hint List:

- a: This problem is the first of many that will ask you to use your knowledge of the real world. If you don't know much about human biology, feel free to spend a few minutes Googling to learn about artery clogging and its causes. When you teach a stats class, you can either attempt to avoid any situation that requires outside knowledge, or you can embrace the fact that statistics is a science deeply embedded in the real world. I choose the latter option, and recognizing that not all students know about all topics, I allow Google searches to read about ideas (but not search answer sites). In truth, searches are rarely needed; simply thinking about the situation carefully is usually sufficient.
- b: First focus on getting blood to the organ. There are two pathways for this − the top and bottom. What the probability blood will make it through the top path? The bottom path? Then, note that blood makes it to the organ if it makes it via either path (top or bottom).

Solution:

(a) Illness, extreme body stressors, or diet/genetics can affect all the blood vessels similarly, and hence, lead to dependencies in clogs.

(b) Label the artery clog points 1 to 3 from top to bottom. Let A_1 (A_2, A_3) be the event that point 1 (2,3) is clear. Note that

$$\begin{aligned} P(\text{blood to organ}) &= P(\text{top or bottom split works}) \\ &= P(\text{top works}) + P(\text{bottom works}) - P(\text{both work}) \end{aligned}$$

$$\begin{aligned} \text{Now, } P(\text{bottom works}) &= P(A_2 \cup A_3) = P(A_2) + P(A_3) - P(A_2 \cap A_3) \\ &= (1-p_1) + (1-p_1) - (1-p_1)^2 = 2 - 2p_1 - 1 + 2p_1 - p_1^2 = 1 - p_1^2 \end{aligned}$$

$$\begin{aligned} \text{So, } P(\text{blood to organ}) &= P(A_1) + (1-p_1^2) - P(A_1)(1-p_1^2) \\ &= (1-p_1) + (1-p_1^2) - (1-p_1)(1-p_1^2) \overset{\text{alg.}}{=} 1 - p_1^3 \end{aligned}$$

Finally, $P(\text{full circuit works}) = P(\text{blood to organ} \cap \text{vein works}) \overset{\text{ind.}}{=} \boxed{(1-p_1^3)(1-p_2)}$.

(c) We check the cases:

i. $p_1 = 0, p_2 = 0$: Here, we get $(1-0^3)(1-0) = 1$, so blood gets through with 100% probability. This makes sense since $p_1 = p_2 = 0$ means we have a 0% chance of clogs.

ii. $p_1 = 1$: Here, we get $(1-1^3)(1-p_2) = 0$, so blood definitely cannot get through the system. This makes sense because $p_1 = 1$ means that each artery clog point must be clogged, so blood never reaches the organ, and system is doomed (regardless of the vein clog point).

iii. $p_1 = 0, p_2 = \frac{1}{2}$: Now, we have $(1-0^3)\left(1-\frac{1}{2}\right) = \frac{1}{2}$. Since $p_1 = 0$, we know the arteries are definitely clear. When the blood reaches the vein clog point, we have $p_2 = \frac{1}{2}$, meaning it is a coin flip as to whether the vein is clear. Since the arteries are clear, the functioning of the complete system simply mirrors that of the vein, as the math reveals.

Commentary: One reason I assign this problem is that you start to see the power of probability theory to make sense of the world. Many real-life settings involve branching pathways (blood flow, waterways, sewers, electricity, knowledge transmission, etc.) and so it is natural to ask how failures in certain branches affect the system as a whole. You'll also note that part c is exploring the idea of particularization/specialization introduced in (the solution of) the four friends problem from earlier. Would you have tried this without me including part c?

Another important feature is that when solving part b, you may have been surprised when you found that $P(\text{blood to organ}) = 1 - p_1^3$. An answer this simple suggests that an easier analysis may be possible. Note that the system I have drawn is really just blood leaving the heart, branching 3 ways, and then joining together. (If you've studied circuits in physics or engineering, the clog points are "in parallel".) As such, we could do:

$$P(\text{blood to organ}) = P(\text{any path works}) = 1 - P(\text{no path works}) = 1 - P(A_1^C \cap A_2^C \cap A_3^C) = 1 - p_1^3.$$

When you get a clean answer, this may be a sign that the teacher engineered it to be so (recall most of elementary and middle school), that an elegant argument exists to analyze the situation (as seen here), or just by coincidence. Here's another example of this phenomenon that I saw in "The Riddler: Fantastic Puzzles from FiveThirtyEight" by Oliver Roeder (on page 71):

> Suppose 10 married couples are at a retreat. The 20 people are randomly seated in a circle. If we choose two couples at random and tell them to look at their partners, what is the probability that their lines of sight will cross?

When I first solved this, I had an overly complicated solution. When I arrived at the (correct) answer of $\frac{1}{3}$, I figured there might be a simple solution and returned to the problem. Can you find the short, elegant path to $\frac{1}{3}$?

16. One the challenges of election polling is that respondents may lie (because society shames certain choices, to spread disinformation, etc.). Suppose we have exactly two candidates running for office, A and B, and that 51% of people will actually vote for candidate A on election day. Now, suppose that when pollsters ask candidate A supporters who they will vote for, 3% of them lie and say candidate B. To simplify life, we'll suppose that all those who support candidate B will tell the truth when polled.

    (a) Given this setup, what percent support will the polls show for candidates A and B? [WISE, VIZ]

    (b) What is the probability a person will actually vote for candidate B if they said they would to a pollster?

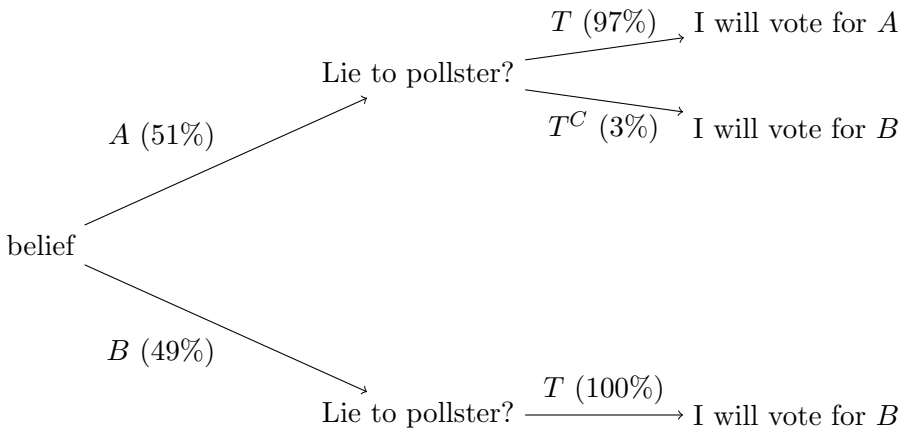
    (c) How powerful is the effect of lying? Instead of the numbers above, suppose that 55% support candidate A and let p be the percent of A voters that will lie to pollsters about their choice. (B voters are still 100% honest.) What values for p will lead to a poll claiming that B has a majority? List your answer in interval notation.

Hint List:

- a: The awkward thing about this problem is that there are two levels going on. We have what people believe in their souls (which is who they will vote for on election day) and what they say to pollsters (which may mirror their beliefs or be a lie). Students sometimes think of a third level to worry about: Will the pollsters lie about the data they collect? Here, we don't analyze this additional level. Try drawing a tree diagram to structure our two levels.
- c: Our VIZ will look similar to part a but with different probabilities. Note that the term "majority" means $> 50\%$.

Solution:

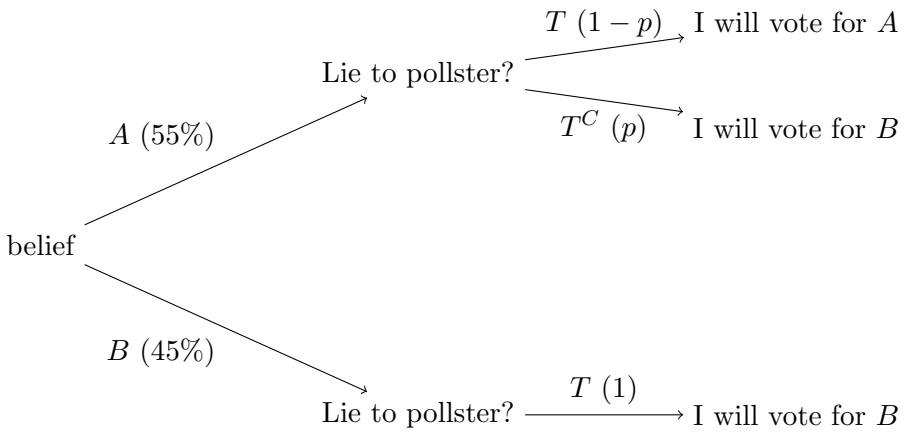
(a) (WIS) Let A (B) be the event that the person, in their heart, wants to vote for for candidate A (B). T is the event that the person will tell the pollsters the truth about their beliefs when asked. We have this picture:



So, (E) $P(\text{I will vote for } A) = 0.51 \cdot 0.97 = \boxed{49.47\%}$. $P(\text{I will vote for } B) = 100 - 49.47 = \boxed{50.53\%}$. You can also do $P([A \cap T^C] \cup [B \cap T])$, which gives the same answer.

(b) $P(B \mid \text{I will vote for } B) \overset{\text{Bayes}}{=} \dfrac{P(\text{I will vote for } B \mid B) \cdot P(B)}{P(\text{I will vote for } B)} = \dfrac{100 \cdot 49}{50.53} = \boxed{96.972\%}$.

(c) Now, our picture is:



We want $P(\text{I will vote for } B) > 50\%$. This means $1 \cdot 45 + p \cdot 55 > 50$, or $p > \dfrac{5}{55} = \dfrac{1}{11}$. So, if p is any value in $\boxed{\left(\dfrac{1}{11}, 1\right]}$, the polls will point in the opposite direction of the true beliefs of people (and hence, of the results on election day)!

Commentary: I created this problem when thinking about the 2016 election. Leading up to election day, virtually all polls showed Clinton (B in the tree diagrams) beating Trump (A). And yet, on election day, Trump won. Some analysts argued that the polls had selection bias (i.e., they had too few Trump supporters and, hence, underrepresented his level of support). This may be true, but I was curious what would happen in the poll results if a small percentage (p) of respondents lied. The finding in part c was surprising to me: If about 1 in 11 Trump supporters decide to lie, then a 10% difference in the candidates can be erased, making it appear that Clinton has greater support in the polls. As the difference shrinks (e.g., in tight races like the 2016 election), the percent that need to lie to reverse the polls gets even smaller than 1 in 11! Finally, note that the tree diagrams in this problem have different branching structures (2 subnodes vs. 1). This is possible and common when there is an asymmetry in how people behave in some setting.

17. In office hours, I tend to get three types of questions. Easy questions (e.g., "What went wrong in my solution here?") require a minute to respond to. Medium questions (e.g., "What good are random variables in the real world?") require 4 minutes. Hard questions (e.g., "How do I become a better student?") take 10 minutes. Based on experience, about 60% of questions are easy, 30% are medium, and 10% are hard.

    (a) Define a random variable for this setting in words and in symbols, state its support, and give its probability model.

    (b) What is the average time I spend responding to a student question, and what is the standard deviation in my response times? (Include units on answers, as always.)

Hint List:

- a,b: Try memorizing the formulas for expectation, variance, and SD before beginning the problem. Also, why is this setting a discrete random variable?

Solution:

(a) Let X be the time (in minutes) of my response to a random student's question.

The problem notes that $X = 1, X = 4$, or $X = 10$. So, the support is $\boxed{\{1, 4, 10\}}$.

The probability model is:

$$\boxed{P(X = x) = \begin{cases} 0.6, & X = 1 \\ 0.3, & X = 4 \\ 0.1, & X = 10 \end{cases}}$$

(b) We break out our formulas for expectation, variance, and standard deviation of a DRV:

$$\begin{aligned} E[X] &= \sum_{x \in \text{ support}} x \cdot P(X = x) \\ &= 1 \cdot 0.6 + 4 \cdot 0.3 + 10 \cdot 0.1 = \boxed{2.8 \text{ minutes}}. \end{aligned}$$

$$\begin{aligned} Var[X] &= \sum_{x \in \text{ support}} (x - \mu)^2 \cdot P(X = x) \\ &= (1 - 2.8)^2 \cdot 0.6 + (4 - 2.8)^2 \cdot 0.3 + (10 - 2.8)^2 \cdot 0.1 \\ &= \boxed{7.56 \text{ minutes}^2} \end{aligned}$$

$$\text{Thus, } SD(X) = \sqrt{Var(X)} = \sqrt{7.56} \text{ minutes } \approx \boxed{2.750 \text{ minutes}}.$$

Commentary: This is your standard "plug-and-chug" problem. You have some formulas (for expectation, variance, and standard deviation) and some numbers that need to go into them. These problems represent a large part of what is asked of students in the American high school system, and you tend not to make it to college if you can't complete such problems. As you move up through undergrad and later grad school, these types of problems are replaced with challenges that require deeper analytic skills.

This transformation is important because we need thinkers that can move beyond simple cognitive tasks. AI is already good enough to do simple plug-and-chug tasks. For example, below is how ChatGPT responds when I give it this problem. I've noted some mistakes it makes in red. As you can see, it is on the verge of doing things correctly. Indeed, on different days, the prompt gets new things right and wrong, so you may get a different response than the one below if you try this yourself.

ChatGPT response:

The random variable for this setting represents the time it takes to respond to a student's question. Let's denote this random variable as X. The support of X is the set of possible response times: $\{1, 4, 10\}$ (in minutes), corresponding to the three types of questions: easy, medium, and hard.

The probability model for X can be described as follows:

$P(X = 1) = P(\text{Easy}) = 0.6$ (60% of questions are easy)
$P(X = 4) = P(\text{Medium}) = 0.3$ (30% of questions are medium)
$P(X = 10) = P(\text{Hard}) = 0.1$ (10% of questions are hard)

To calculate the average time spent responding to a student question, we need to find the expected value or mean of the random variable X.

$$\begin{aligned} E(X) &= (1)(P(\text{Easy})) + (4)(P(\text{Medium})) + (10)(P(\text{Hard})) \\ &= (1)(0.6) + (4)(0.3) + (10)(0.1) = 0.6 + 1.2 + 1 = 2.8 \text{ minutes} \end{aligned}$$

Therefore, the average time spent responding to a student question is 2.8 minutes.

To calculate the standard deviation of the response times, we first need to calculate the variance using the formula:

$$\begin{aligned} Var(X) &= (1 - 2.8)^2(P(\text{Easy})) + (4 - 2.8)^2(P(\text{Medium})) + (10 - 2.8)^2(P(\text{Hard})) \\ &= 1.8^2(0.6) + 1.2^2(0.3) + 7.2^2(0.1) = 5.832 \end{aligned}$$

The standard deviation can be obtained by taking the square root of the variance:

$$\sqrt{Var(X)} = \sqrt{5.832} \approx 2.414 \text{ minutes}$$

18. Researchers interested in human focus will often give test subjects an endless list of boring problems (e.g., What is $2 + 13$? What letter comes after Q in the alphabet?) and count how many problems it takes before they break focus and get a question wrong. For example, if you get 7 questions correct and then mess up, your count-til-failure is 8. Naturally, this count-til-failure will vary from one person to the next. Suppose that extensive experiments have found that unaltered test subjects have an average count-til-failure of 11.3 questions with a standard deviation of 4.2 questions.

    (a) A new drug, Focusizil, has come to market that doubles each person's count-til-failure. What are the average and standard deviation for the count-to-failure of test subjects while on this drug? [WISE]

    (b) Another drug, Extendaron, allows an individual to get 11 more questions correct than they normally would. What are the average and standard deviation for the count-to-failure of test subjects while on this drug (and not Focusizil)? [WISE]

    (c) The news media has been arguing over which drug a person should take, since the drugs can't both be taken. Give an argument in support of each drug using your results from a and b.

Hint List:

- a: If Y is the count-til-failure of a random person on Focusizil, write a random variable expression for Y using X, the count-til-failure of a random unaltered test subject.
- b: Again, try to express what Extendaron does using a transformation of X.
- c: How do the means and standard deviations of the two drugs compare?

Solution:

(a) (WIS) Let X be the count-til-failure of a random unaltered test subject. We are told that $E[X] = 11.3$ and $SD[X] = 4.2$. (WIS) If Y is the count-til-failure of a random person on Focusizil, then $Y = 2X$. Thus (E),

$$E[Y] = E[2X] = 2E[X] = 2 \cdot 11.3 = \boxed{22.6 \text{ questions}}$$
$$SD[Y] = SD[2X] = 2SD[X] = 2 \cdot 4.2 = \boxed{8.4 \text{ questions}}$$

(b) (WIS) Let Z be the count-til-failure of a random person on Extendaron. We see that $Z = X + 11$. So (E),

$$E[Z] = E[X + 11] = E[X] + 11 = 11.3 + 11 = \boxed{22.3 \text{ questions}}$$
$$SD[Z] = SD[X + 11] = SD[X] = \boxed{4.2 \text{ questions}}$$

(c) If the goal is to get the highest average focus, people should take Focusizil ($22.6 > 22.3$). This higher average, however, comes with more variability ($8.4 > 4.2$ for SDs). So, if you want more consistency with nearly the same average, recommend Extendaron.

Commentary: Based on the last commentary, you might recognize parts a and b as plug-and-chug questions: We have formulas to help simplify the expectation and standard deviation of random variable expressions and these need to be invoked. Part c, in contrast, is where you begin to move into higher-order thinking. What should a person do when two (or more) options are present for a situation? One approach is to outline the facts about both options (parts a and b), and then define a decision-making calculus. Here, if you care about expectation, you head one way, and if you care about variability, you head a different way. My friend recently experienced this exact situation when trying to decide between the drugs Xeljanz and Rinvoq. In that case, one had more documented adverse events (AEs) and higher chance of pain relief, while the other had fewer AEs but lower pain relief. While he and his doctor agreed on these facts, they disagreed on how to weigh them when making a decision. As you go about your day, take a look at the debates in your world and try to study if they arise because of disagreements over facts or how to act given those facts (i.e. decision-making calculi).

19. Your computer science friend has written a new app that allows users to pay a fee to win a prize. Here's how it works: Each time the user wants to play, they must pay k pennies (where k is an integer). Then, the user will either win \$4 (with probability $\frac{5}{6}$), \$16 (with probability $\frac{5}{36}$), \$64 (with probability $\frac{5}{216}$), \$256 (with probability $\frac{5}{1296}$), etc.

    (a) Show that the probabilities in this problem sum to 1, as they must.

    (b) What value should your friend choose for k if the goal is to keep the fee as low as possible, but for the programmer to make money on average? [WISE]

Hint List:

- a: There are more than four numbers to sum because of the word "etc." at the end of the problem statement. Can you find a pattern in the numbers and use this to your advantage when summing infinitely-many numbers?
- b: How much money should the programmer charge to break even, on average?

Solution:

(a) We must sum

$$\frac{5}{6}+\frac{5}{36}+\frac{5}{216}+\frac{5}{1296}+\cdots=\frac{5}{6}+\frac{5}{6^2}+\frac{5}{6^3}+\frac{5}{6^4}+\ldots$$

We recognize this as an infinite geometric series with common ratio $\frac{1}{6}$. Thus, the sum is: $\frac{\text{first term}}{1-\text{ common ratio}}$. Hence, we get

$$\frac{5/6}{1-1/6}=\frac{5/6}{5/6}=\boxed{1}.$$

(b) (WIS) Let X be the amount of prize money a random person wins from playing the game. We find (E):

$$\begin{aligned}E[X]&=4\cdot\frac{5}{6}+16\cdot\frac{5}{6^2}+64\cdot\frac{5}{6^3}+\ldots\\&=5\cdot\frac{4}{6}+5\cdot\left(\frac{4}{6}\right)^2+5\cdot\left(\frac{4}{6}\right)^3+\ldots\\&=\frac{5\cdot\frac{4}{6}}{1-4/6}=10.\end{aligned}$$

So, a fee of \$10 causes the programmer to break even. Charging \$10.01, or $\boxed{\text{1001 pennies}}$, leads to the programmer making money and keeps k as small as possible for the users of the app.

Commentary: The main goal of this problem is to offer the solver a setting they have no familiarity with and to ask some natural questions about it. Here, you must decide if the setting is structured properly (part a) and to decode what it means to "keep the fee as low as possible" while the programmer still turns a profit. If the setting (a betting app with a weird pay structure) feels artificial, this is by design. When you are trying to move outside your experience base into new settings, it can be useful at first to do so with domains that do not require a lot of specialized knowledge. In time, as you learn to transfer your skills into new settings, the transference of skills is less of the heavy lifting, while learning the new field has more of your attention. Also, the problem asked you to invoke the infinite geometric series sum formula twice. Be honest with yourself: Did you internalize this result (if you didn't know it) back on the Four Friends problem? When you continually encounter situations that require the same tool, you should put that tool in your intellectual toolbox and learn to use it.

20. One of the most common random variables is known as the "discrete uniform distribution". This is a random variable that can only output the values $1, 2, \ldots, n$, each with equal probability. Here, n is a known, positive integer. For each of the below parts, your answer may involve n.

    (a) Find the probability the random variable outputs an even number. [WISE]

    (b) What are the expected value and variance of the output of the random variable? (Hint: If you haven't yet memorized the formulas for $\sum_{i=1}^{n} i$ and $\sum_{i=1}^{n} i^2$, then you should do so before even starting the problem.)

Hint List:

- a: Consider the cases of n even versus odd separately.
- b: Don't try to use your answer from part a in part b. Also, don't assume that because a problem has multiple parts that the results from earlier parts *must* be used in later parts.
- b: To make progress on the sum, try expanding the expression $(i-\mu)^2 = \left(i - \frac{n+1}{2}\right)^2$. This will temporarily make things worse but, ultimately, allow you to proceed.

Solution:

(a) (WIS) Let X be the output of the discrete uniform random variable If n is even, then half the numbers in $1, 2, 3, \ldots, n$ are even. Since all the numbers are equally likely to be output, we get (E):

$$P(X \text{ is even}) = \frac{n/2}{n} = \frac{1}{2}.$$

If n is odd, the list $1, 2, 3, \ldots, n$ ends in an odd number, so it has one more odd number than even number. So, the number of even numbers is $\dfrac{n-1}{2}$ and thus (E),

$$P(X \text{ is even}) = \frac{\frac{n-1}{2}}{n} = \frac{n-1}{2n}.$$

Hence, we could write

$$\boxed{P(X \text{ is even}) = \begin{cases} \frac{1}{2}, & n \text{ is even} \\ \frac{n-1}{2n}, & n \text{ is odd} \end{cases}}$$

Our answer can be written without cases if you know the round down function:

$$\boxed{P(X \text{ even}) = \frac{\lfloor n/2 \rfloor}{n}}.$$

(b) Here, we use summation formulas to simplify the sums that arise.

$$\begin{aligned} E[X] &= \sum_{i=1}^{n} i \cdot P(X = i) = \sum_{i=1}^{n} i \cdot \frac{1}{n} \\ &= \frac{1}{n} \sum_{i=1}^{n} i = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \boxed{\frac{n+1}{2}}. \end{aligned}$$

Note that $\dfrac{n+1}{2}$ is halfway between 1 and n, which makes intuitive sense.

$$
\begin{aligned}
Var[X] &= \sum_{i=1}^{n}(i-\mu)^2 P(X=i) = \sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2 \cdot \frac{1}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[i^2 - i(n+1) + \left(\frac{n+1}{2}\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} i^2 - \frac{n+1}{n}\sum_{i=1}^{n} i + \frac{1}{n}\underbrace{\sum_{i=i}^{n}\left(\frac{n+1}{2}\right)^2}_{\text{Note: } \sum_{i=1}^{n} c = nc} \\
&= \frac{1}{n}\cdot\frac{n(n+1)(2n+1)}{6} - \frac{n+1}{n}\cdot\frac{n(n+1)}{2} + \frac{1}{n}\cdot n\cdot\left(\frac{n+1}{2}\right)^2 \\
&= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{2} + \frac{(n+1)^2}{4} \\
&= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = (n+1)\left[\frac{2}{2}\cdot\frac{2n+1}{6} - \frac{n+1}{4}\cdot\frac{3}{3}\right] \\
&= (n+1)\left[\frac{4n+2-3n-3}{12}\right] = \frac{(n+1)(n-1)}{12} = \boxed{\frac{n^2-1}{12}}.
\end{aligned}
$$

Those with more background in probability might know the formula:

$$Var[X] = E[X^2] - E[X]^2$$

It is fine to use this if you know it; we learn it later in the course. It can save about 20% of the steps in this problem.

Commentary: Why this problem? Part a is here to remind you that casework is often necessary to understand a situation. It might be tempting to view casework as a substandard solution pathway. I'd like to dissuade you of that belief. Part b is here to help strengthen your algebra skills and to reset your internal threshold on how "messy" a problem can be before you get worried and think you've made a mistake. As you rise through the curriculum, the messiness of your calculations will grow because you will consider less idealized settings, incorporate more variables to analyze, and solve harder problems. The low-hanging fruit of simple problems with five or fewer steps and nice-looking answers has been picked off the tree.

21. Let's check your answers to question 20b.

    (a) Explain how the sample function in R can be used to mimic the behavior of the discrete uniform distribution for a given value of n.

    (b) Write code to check your answers for 20b using $n = 1234$ and 10000 simulated values. When doing this, create a new R script (File $\gg$ New File $\gg$ R Script, on PC), and use set.seed(31415) as your first line of code. This aligns the random number generators on our computers and ensures we will likely get identical answers. (It doesn't work in all cases because of system-specific differences.) When you use a script, make sure to run the whole thing at once, rather than running individual lines of code. In Windows, I do this by pressing Ctrl + Shift + Enter. Find the hotkeys on your computer by selecting the "Code" tab in R Studio and looking for the "Source with Echo" command.

Hint List:

- a: Type ?sample at the command prompt and read the documentation or search the internet.
- b: Recall that R has built-in functions like mean and var, so you need not reinvent the wheel.

Solution:

(a) Typing

```
sample(x, 1, replace=T)
```

will choose a random value from the vector x, giving each choice an equal chance. So,

```
sample(1:n, 1, replace=T)
```

mimics the discrete uniform distribution.

(b)

```
set.seed(31415)

n = 1234
howMany = 10000
vals = sample(1:n, howMany, replace=T)

mean(vals)                # I got 616.3243
print((n+1)/2)            # I got 617.5

var(vals)                 # I got 127398.8
print((n^2-1)/12)         # I got 126896.2
```

Commentary: The goal of this problem is to build your familiarity with R and to show you how useful simulation can be. The length of the calculations leading to $Var(X) = \frac{n^2-1}{12}$ and the exoticness of this formula may have you questioning the accuracy of your answer. Once you run the simulation, however, these fears are allayed. As an example in my life: Recently, I needed to know the average distance between two points randomly chosen inside a unit square. A long calculation by hand gave a nightmarish looking answer: $\frac{2+\sqrt{2}+5\sinh^{-1}(1)}{15}$. Could this possibly be correct?! Five lines of code immediately convinced me that my two-page calculation was accurate.

22. Decide if each random process describes a Bernoulli random variable. If so, give the probability model; if not, explain why not.

    (a) You break a barrel in Legend of Zelda: Breath of the Wild and report to your friend whether you got no arrows (sadness) or some arrows (yay!). (See Lecture 6 for details on this setup.)

    (b) A sociologist is interested in gender identity and asks random college students to identity as "masculine", "feminine", or "non-binary". After extensive testing, the results come back: 40% masculine, 40% feminine, and 20% non-binary.

    (c) You flip six unfair coins ($p_{\text{heads}} = 0.3$) all at once and then report back whether the majority were heads or not.

Hint List:

- Overall, you should be checking if the problem is structured as two possible outcomes and if the outcomes from trial to trial are independent.
- c: Set up one RV that counts the number of heads and a different one that check if a majority were heads. We are trying to check if the *second* RV is Bernoulli or not.

Solution:

(a) If we set a "success" as "some arrows" (yay!) and a failure as no arrows (sadness), then we can combine the cases of 1 arrow and 5 arrows together:

$$P(X=x) = \begin{cases} \frac{3}{10} + \frac{2}{10} = \frac{1}{2}, & x = 1 \text{ (success)} \\ \\ \frac{5}{10} = \frac{1}{2}, & x = 0 \text{ (failure)} \end{cases}$$

The trials are independent because the programmers have coded them as such!

(b) We have 3 outcomes, not 2, as required by a Bernoulli RV.

(c) If $X \sim Binom(6, p = 0.3)$, we are reporting if $X \geq 4$ or not.

Let Y be a RV that is 1 if $X \geq 4$ (a majority were heads!) and 0 otherwise.

$$\begin{aligned} P(Y=1) = P(X \geq 4) &= P(X=4) + P(X=5) + P(X=6) \\ &= \binom{6}{4} \cdot 0.3^4 \cdot 0.7^2 + \binom{6}{5} \cdot 0.3^5 \cdot 0.7^1 + \binom{6}{6} \cdot 0.3^6 \cdot 0.7^0 \\ &= 0.07047 \end{aligned}$$

Thus, the probability model for Y is given by $P(Y=y) = \begin{cases} 0.07047, & y = 1 \\ 0.92953, & y = 0 \end{cases}$

We have two outcomes (1 "majority heads" vs. 0) and independence (the six coins have no memory if you were to flip them all again), so Y is Bernoulli.

Commentary: The goal here is to look carefully at different real-world simulations as see if they can be modelled using a Bernoulli RV. As parts a and c show, it is possible to take complex situations (a: 0, 1, or 5 arrows; c: 0, 1, 2, 3, 4, 5, 6 heads) and recast them into a binary framework (a: No vs. some arrows; c: Majority were heads vs. not). We could have done this in part b if we wished, but I wanted at least one of the answers to be a non-Bernoulli RV. Finally, note that independence can be achieved in different ways. In part a, independence is the result of how video games are coded by developers. In part c, independence results from physical objects having no memory. It's not like the coins are thinking "Last time we gave a majority heads, so let's mix it up and give a minority this time".

23. Suppose the probability that a lightbulb fails on a given day is 0.002.

(a) Find the probability the light bulb first fails on day 80 of use. Write your answer in scientific notation. [WISE]

(b) On average, how many days will it take until the light bulb fails (including the failure day)?

(c) Determine the probability the lightbulb lasts at least one year (at least 365 complete days of use before failure). Give an **exact** answer (as a messy fraction expression) that doesn't involve a summation, ellipsis, or decimals.

(d) What assumption about days and lightbulb failures must be made to do your above caclulations, and why might this assumption be problematic? (Note: There are many possible answers to this question. Don't worry about trying to figure out what answer I wrote; just offer some reasonable, clearly-articulated concern.)

Hint List:

- a: What distribution is useful for counting events until failure?
- c: When is the first possible day the lightbulb could fail? When is the last?

Solution:

(a) (WIS) Let $X \sim Geom(p = .002)$ be the number of days until the lightbulb fails (including the failure day). We want (E):

$$P(X = 80) = (1-p)^{80-1} \cdot p = 0.998^{79} \cdot .002 \approx \boxed{1.707 \cdot 10^{-3}}.$$

(b) We want $E[X] = \dfrac{1}{p} = \dfrac{1}{.002} = \boxed{500 \text{ days}}$.

(c) If we get at least 365 days of use, the failure occurs on day 366 or before. So, we want

$$\begin{aligned} P(X \geq 366) &= P(X = 366) + P(X = 367) + P(X = 368) + \dots \\ &= (1-p)^{365} \cdot p + (1-p)^{366} \cdot p + (1-p)^{367} \cdot p + \dots \text{ (inf. geom. series)} \\ &= \frac{(1-p)^{365} p}{1-(1-p)} = (1-p)^{365} = \left(1 - \frac{1}{500}\right)^{365} = \boxed{\left(\frac{499}{500}\right)^{365}}. \end{aligned}$$

(d) We need the fail/not-fail feature of the lightbulb to be a Bernoulli RV. So, the chance of failure on any day should be *independent* of other days. This is probably not the case since materials get worse with age, so the probability of failure goes up over time. Also, if the bulb fails on day 8, is its failure on day 9 still in question?

Commentary: Parts a and b are your standard plug-and-chug problems. Parts c is here because it provides the chance for an OBOE (off by one error): you must realize that 365 days of success forces the first failure on day 366 or later. To continue growing as an academic, you must learn to be more detail-oriented. OBOE experiences help train this. They are very common in computer science (Should the loop be to $i < n$ or $i \leq n$?). Finally, part d asks you connect with the real world and challenge assumptions. Many statistics teachers use electronics failures as examples of the geometric distribution, but in truth, the Weibull distribution is a better choice (albeit way more complicated).

24. On July 26, 2019, I was watching a Red Sox/Yankees baseball game with a crossword constructor friend from the East Coast. He was explaining that I should know who Mookie Betts is, and that given his fame, his name is fair game to put in a crossword puzzle. Much to our surprise, Betts hit three home runs in that game, an event so strange I wrote this question.

    (a) Assuming Betts has exactly 4 at-bats in a game, what is the probability exactly 3 of these result in home runs? Use this historical data that predates the fateful game we watched: In 3372 career at-bats, Betts hit a total of 153 home runs. [WISE]

    (b) How many standard deviations above the mean is the event we watched?

    (c) What assumption about home runs and at-bats must be made to do your above caclulations, and why might this assumption be problematic? (See the note in 23d.)

Hint List:

- a: I fixed the number of at-bats at 4 to allow us to use a particular probability distribution. Which one involves a fixed number of trials?
- b: If the language in this question is awkward, try an easier one that will show you how to proceed: How many decades older is a 70-year-old compared to a 20-year-old? What operations did you do to find this answer?

Solution:

(a) (WIS) Let $X \sim Binom\left(4, p = \frac{153}{3372}\right)$ be the number of home runs Mookie gets in a game where he has 4 at-bats. We want (E):

$$P(X = 3) = \binom{4}{3} \cdot \left(\frac{153}{3372}\right)^3 \cdot \left(1 - \frac{153}{3372}\right) \approx \boxed{3.567 \cdot 10^{-4}}.$$

(This is about a 1 in 2800 event!)

(b) We know $E(X) = np = 4 \cdot \frac{153}{3372} = \frac{51}{281}$. Also, $SD(X) = \sqrt{npq} \approx 0.416$. To find "how many SDs" our event is above the mean, we find the difference of our event (3) and the mean ($E(X)$) and divide by $SD(X)$:

$$\frac{3 - E(X)}{SD(X)} \approx \frac{3 - \frac{51}{281}}{0.416} \approx \boxed{\text{6.771 SDs above the mean}}.$$

This, again, suggests I saw a strange event.

(c) The Binomial is built on a Bernouli RV, which requires independence. The home run status of an at-bat may be influenced by other at-bats for many reasons: Betts believes he is in a hot or cold streak, other psychological factors, injuries, an easy-to-read pitcher, etc.

Commentary: I see people all the time who believe they have experienced something rare, but when you run the calculations, the event is not rare at all (and vice versa). I created this problem because I was unsure if I should trust my own intuition when it said I saw something really rare. Probability theory is useful because it allows us to *quantify* the strangeness of what we are experiencing while taking into account the random forces at play in almost all aspects of life. One of the great values of education is that it provides you with new tools to moderate and check your own thinking patterns (as this problem shows). Humans are born with many mental tendencies that are unproductive or inaccurate in the modern era. The psychologists Daniel Kahneman and Amos Tversky spent their careers documenting these cognitive biases, giving rise to the field of behavioral economics. Read more here.

25. Each of the below problems involves a random process and an unknown probability p where $0 < p < 1$. Use the information given to find p. Give exact answers.

(a) A friend has a spinner that lands on one of three possibilities: A, B, or C. You are told there is a 50% chance of landing on A, and $p = P(\text{spinner lands on } B)$. The friend also mentions that if you spin twice, the chance of getting both a B and a C in some order is $\frac{3}{32}$.

(b) An unfair coin with $p = P(\text{getting tails})$ has just sold at auction. The auction house mentions that when you flip the coin 8 times and record the number of heads, the variance of the heads count is 3. [WISE]

(c) An unfair coin with $p = P(\text{getting heads})$ is to be flipped 7 times. You are told the chance of getting exactly 1 head is twice the chance of getting all tails. [WISE]

Hint List:

- a: The phrase "in some order" means that the number 3/32 is the probability for both cases (B then C, C then B) combined, not the value for each case separately.

Solution:

(a) We have $P(A) = \frac{1}{2}, P(B) = p$, and $P(C) = 1 - \frac{1}{2} - p = \frac{1}{2} - p$. Thus,

$$P(B \text{ and } C \text{ in some order}) = P(B \text{ then } C \text{ or } C \text{ then } B).$$

By disjointness and independence, this becomes

$$p\left(\frac{1}{2} - p\right) + \left(\frac{1}{2} - p\right)p.$$

So, we have $\frac{3}{32} = 2p\left(\frac{1}{2} - p\right) \implies 64p^2 - 32p + 3 = 0$. Factoring gives

$$(8p - 1)(8p - 3) = 0 \implies \boxed{p = \frac{1}{8} \text{ or } \frac{3}{8}}.$$

(b) (WIS) Let $X \sim Binom(8, 1 - p)$ be the number of heads in 8 flips. We know (E):

$$3 = Var(X) = n(1 - p)p = 8(1 - p)p$$

So, $8p^2 - 8p + 3 = 0$. The discriminant is $(-8)^2 - 4(8)(3) < 0$, so there are $\boxed{\text{no real solutions for } p}$.

(c) (WIS) Let $X \sim Binom(7, p)$ be the number of heads in 7 flips. We have (E):

$$P(X = 1) = 2 \cdot P(X = 0) \implies \binom{7}{1}p^1(1 - p)^6 = 2 \cdot \binom{7}{0}p^0(1 - p)^7 \implies 7p(1 - p)^6 = 2(1 - p)^7$$

Since $p \neq 1$, we can divide by $(1 - p)^6$ to get $7p = 2(1 - p)$. This gives $\boxed{p = \frac{2}{9}}$.

Commentary: The main goal of this problem is to provide settings for the use of WISE. I created the WISE mnemonic to help students bridge the worlds of language and mathematical notation. Most statistics problems are presented in a textual setting but are asking about mathematical ideas (e.g., probabilities, statistics). A good solution helps the reader navigate this semiotic transition, and WISE is the scaffolding that facilitates it.

I also designed these problems to push back on academic narratives you might have. For example, on part a, some students are bothered when they get two answers (Is this possible? Did I make a mistake?). If the story you tell yourself about school is "I should get one clean answer in roughly 3-8 steps", then it's time to update your narrative. It might be better to tell yourself: I can get many (part a), one (part c), or no answers (part b), which may or may not be pretty (see problems throughout this problem book), and which might take between 1 and 50 steps.

Parts a and b also have the potential to get you asking interesting questions. In part a, you might wonder if it makes sense to get two answers. Notice that when $p = \frac{1}{8}$, the probabilities for the spinner landing on B and C are $\frac{1}{8}$ and $\frac{3}{8}$ respectively. When $p = \frac{3}{8}$, we get the probabilities of $\frac{3}{8}$ and $\frac{1}{8}$ for B and C. Since the problem is symmetric with respect to B and C, these two cases inject the same two numbers into our problem, and hence, should lead to the same results. On part b, you might ask: "There is no *real* solution for p, but what if we allow complex numbers? Can a probability be a complex number?!" I leave this for you to ponder.

Note that the act of asking questions is critically important. When the physics Nobelist Isidor Rabi was asked why he became a scientist, he cited his mother. Instead of asking how his day was upon returning from school, she instead said: "Izzy, did you ask a good question today?". Spend time with your answers thinking about their reasonableness and what directions they might lead you. Question what is happening around you all the way to a Noble Prize.

26. Most people who work jigsaw puzzles begin by assembling the border (the edge and corner pieces) because these pieces are easy to notice. If you have lots of space to assemble the puzzle, you might dump out all the pieces and begin separating border and non-border pieces. If you don't have space, one approach is to blindly reach into the bag of pieces and pull out a random piece. If your selection isn't a border piece, you return it to the bag and try again. Suppose you are working a 1000-piece puzzle that is organized as a 40 by 25 rectangular grid.

    (a) How many pulls must you make, on average, in order to get your first border piece? [WISE]

    (b) How many additional pulls do you expect to make to get the second border piece (don't count pulls related to the first border piece)? [WISE]

    (c) How many *total* pulls will you make, on average, to get all the border pieces? Write your answer as a summation and then write some R code to evaluate the summation. [WISE]

    (d) In real life, when you pull out a non-border piece, it's unlikely you'd return it to the bag. In this problem, however, we do return the piece. From a statistical perspective, why do you think the problem was set up this way?

    (e) Suppose you work for the jigsaw puzzle company and want your customers to spend as much time as possible with your puzzles. Assuming the puzzle must be a rectangular grid of 1000 pieces, but not necessarily 40 by 25, what dimensions lead to the largest possible answer for part c? Use R to help you! Assume that both the length and height of the puzzle must be at least 3 (so the puzzle actually has some non-border pieces).

Hint List:

- a: There are NOT 130 border pieces.
- b: Notice that you can't use the same RV to model the second border piece because the probability of drawing a border piece has changed slightly.
- c: Look at the probabilities from parts a and b. They are the beginning of a pattern. Try writing the probability involved in the RV for the 3rd and 4th pieces. Now write the probability for the ith piece. Note that getting all the border pieces is just a succession of RVs.
- e: The coding part of this is intentionally designed to offer little direction. There are many ways to go, and for some, you'll need to do some searching to learn additional syntax.

Solution:

(a) (WIS) Let $X_1 \sim Geom(p_1)$ be the number of pulls to get the first border piece. With a 40 by 25 puzzle, there are $40+40+23+23 = 126$ border pieces. You could also do $40+40+25+25-4 = 126$; this intentionally double-counts the corner pieces and then adjusts. Thus, $p_1 = \dfrac{126}{1000}$. So (E),

$$E(X_1) = \frac{1}{p_1} = \frac{1000}{126} = \boxed{\frac{500}{63} \approx 7.937 \text{ pulls}}.$$

(Recall that expected values should NOT be rounded to the nearest integer.)

(b) (WIS) Let $X_2 \sim Geom\left(p_2 = \dfrac{125}{999}\right)$ be the number of pulls to get the second border piece. The bag now has one less piece and one less border piece because of our actions in part a. So (E),

$$E(X_2) = \frac{1}{p_2} = \boxed{\frac{999}{125} \approx 7.992 \text{ pulls}}.$$

(c) (WIS) Let $X_i \sim Geom\left(p_i = \dfrac{127-i}{1001-i}\right)$ be the number of pulls to get the ith border piece for $i = 1, \ldots, 126$. The total number of pulls is $X_1 + X_2 + \cdots + X_{126}$. This has expectation (E):

$$E(X_1) + \cdots + E(X_{126}) = \frac{1001-1}{127-1} + \frac{1001-2}{127-2} + \cdots + \frac{1001-126}{127-126} = \sum_{i=1}^{126} \frac{1001-i}{127-i}$$

There are no known summation rules that can help us create a simpler version for this expression, so we need technology to help us evaluate it. In R, we can do:

```
# colon notation can make decreasing sequences
nums = (1001 - 1):(1001 - 126)
denoms = (127 - 1):(127 - 126)
sum(nums/denoms)
```

So, we get about $\boxed{4860.861 \text{ pulls}}$.

(d) By drawing with replacement, we get the independence needed for the Bernoulli trials, allowing for the Geometric model to be used.

(e) The only possible sizes (since the length and height are at least 3) are 4 by 250, 5 by 200, 8 by 125, 10 by 100, 20 by 50, and 25 by 40 (and symmetries of these: 40 by 25, 50 by 20, etc.). For each possible size, we find the number of border pieces: $b = \ell + \ell + h + h - 4$ and do a sum like in part c: $\displaystyle\sum_{i=1}^{b} \frac{1001 - i}{b + 1 - i}$. In R, we use the below code. The modulus operator ($a \,\%\%\, b$) gets the remainder when a is divided by b and will return 0 when b is a divisor of a.

```
for (length in 3:sqrt(1000)) {
  if (1000 %% length == 0) {
    height = 1000/length
    b = 2*length + 2*height - 4      # number of border pieces
    pulls = sum(1000:(1001-b)/b:1)   # expected pulls for current size
    print(c(length, height, b, pulls))
  }
}
```

The printout shows that the optimal dimensions for maximizing solver border pulls is a $\boxed{\text{10 by 100}}$ puzzle with an average of about 4884.569 pulls:

```
 4.000  250.000  504.000 3877.189
 5.000  200.000  406.000 4317.371
 8.000  125.000  262.000 4798.831
10.000  100.000  216.000 4884.569
20.000   50.000  136.000 4882.421
25.000   40.000  126.000 4860.861
```

Commentary: I believe a problem should be a journey (when possible). Here, the journey involves a complex, real-world setup that strings together multiple Geometric models with changing success probabilities. In addition, we reach an expression in part c which cannot be simplified, and so we *must* turn to technology to help us calculate the final result. This scenario becomes increasingly common the closer you get to the real world; clean little answer nuggets are an artifice perpetuated by the American educational system.

I also included part e as an example of asking good questions. After completing parts a through c, it tempting to wonder about other sizes. Will we get the same answer with different dimensions? Apparently not. What size maximizes the expected number of pulls? Based on part e, the result is counterintuitive. Some people believe the answer will be the size with the most border pieces (4 by 250) since we get a lot of terms in the sum from part c. Others think the answer might involve the fewest border pieces (25 by 40) because this means they are hard to find, making each term in the summation larger. As part e shows, the answer is between these extremes, striking a balance between more terms and bigger-valued terms.

27. I recently read that, on average, 10 different species on Earth go extinct each year.

(a) What is the probability no species go extinct next year? [WISE]

(b) An animal rights organization labels a year as "disastrous" if 4 or more species are lost that year. What is the probability of a "disastrous" year?

(c) A newspaper has asked you to make a statement of the form: "Next year, the Earth will most likely lose either m different species or n different species." What values should you give for m and n? Use R to help, if you wish.

(d) What value must we change the number 10 to in order for the answer in part a to get doubled? Give an exact answer, not a decimal.

Hint List:

- b: Instead of evaluating an infinite sum, consider a complementary approach.
- c: Consider a simpler situation. Suppose you have a RV X with this model:

$$P(X = x) = \begin{cases} 0.2, & X = 1 \\ 0.3, & X = 2 \\ 0.1, & X = 3 \\ 0.4, & X = 4 \end{cases}$$

What two outcomes are most likely from this RV? Why did you give the answer you did?

Solution:

(a) (WIS) Let $X \sim Pois\left(\lambda = \frac{10 \text{ species}}{\text{year}}\right)$ be the number of species that go extinct in a random year on Earth. We find (E):

$$P(X = 0) = \frac{e^{-10} \cdot 10^0}{0!} = \boxed{e^{-10} \approx 4.540 \cdot 10^{-5}}.$$

(b) $P(\text{ “disastrous year”}) = P(X \geq 4) = 1 - P(X < 4)$

$$= 1 - P(X = 0) - P(X = 1) - P(X = 2) - P(X = 3)$$

$$= 1 - e^{-10} - \frac{e^{-10} \cdot 10}{1} - \frac{e^{-10} \cdot 10^2}{2} - \frac{e^{-10} \cdot 10^3}{6} \approx \boxed{0.990}.$$

(c) We must find the two values for X (m and n) that give the largest $P(X = x)$. In R, try:

```
dpois(0:20, lambda = 10)
```

This uses the vectorized nature of R to plug in each value from $x = 0$ to $x = 20$ into the probability model for the Poisson and see how likely the values are. R then packages the results back into a vector. Note that you could also use the table feature of a calculator and explore $\frac{e^{-10} \cdot 10^x}{x!}$ for $x = 0, 1, 2, \ldots$ We get the largest probabilities for $\boxed{m = 9}$ and $\boxed{n = 10}$, which make sense given the average rate.

(d) Let $Y \sim Pois(\lambda = c)$ be our new extinction model. We need $P(Y = 0) = 2 \cdot P(X = 0)$. So,

$$\frac{e^{-c} \cdot c^0}{0!} = 2e^{-10} \implies \frac{1}{e^c} = \frac{2}{e^{10}} \implies e^c = \frac{e^{10}}{2} \implies c = \ln\left(\frac{e^{10}}{2}\right) = \boxed{10 - \ln 2}.$$

Commentary: Parts a and b are here to check for basic fluency with the Poisson distribution. The questions are phrased in the ways you are used to. Part c is here to show you that the way people talk about things may not be what you're used to. Here, you must decode “most likely” as “highest probability”, and you must recognize if two possibilities are requested, then these must be the two outcomes in the model with the high probabilities (which are the locations of the two tallest bars in the visualization of the model).

28. For a brief period of time, TV shows and movies about Mars were all the rage. In these films, the system designed to provide water is often breaking down, so let's build a model for such a setting. Suppose that the water system has a 1 in 100 chance of breaking down on a given day, and that breakdowns are independent from day to day. When a breakdown occurs, you fix it using a brilliant idea (as in the movies!) and then restart the system the next day. Sadly, your brilliant ideas are limited in number; you can only fix the system 4 times.

    (a) Find the probability you'll have to use up all your brilliant fixes within the first 7 days of the journey. [WISE]

    (b) How many days do you expect you'll go until you finally exhaust your last brilliant idea?

    (c) Thinking in terms of the mean and standard deviation, how strange would it be for you to use your final fix on day 1400?

    (d) Once the water system breaks down and you're out of ideas, you'll have enough water for 20 more days. How long should you expect to survive on Mars? [WISE]

Hint List:

- a: If you're struggling to know which probability model to use, take a moment to memorize what each one models in the real world and then return to this part.
- c: Recall the recent problem involving Mookie Betts and the way we measured strangeness there.
- d: If you feel bothered using the same random variable as part a, you are right. Note that in part a, we have used our fourth brilliant idea and the water system is now functional again. What has to happen next?

Solution:

(a) (WIS) Let $X \sim NegBinom\left(k = 4, p = \frac{1}{100}\right)$ be the number of days until we use all 4 brilliant ideas. There's a $\frac{1}{100}$ chance we'll deploy an idea because we only do so when the water system breaks. Here (E):

$$\begin{aligned} P(X \leq 7) &= P(4 \leq X \leq 7) = P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) \\ &= \binom{3}{3} 0.01^4 0.99^0 + \binom{4}{3} 0.01^4 0.99^1 + \binom{5}{3} 0.01^4 0.99^2 + \binom{6}{3} 0.01^4 0.99^3 \\ &\approx \boxed{3.417 \cdot 10^{-7}}. \end{aligned}$$

(b) $E(X) = \frac{k}{p} = \frac{4}{1/100} = \boxed{400 \text{ days}}$.

(c) Note that $SD(X) = \frac{\sqrt{kq}}{p} = \frac{\sqrt{4 \cdot 0.99}}{0.01} \approx 199$. So $\frac{1400 - 400}{199} \approx 5.025$, meaning that 1400 is about 5 standard deviations above what we expect on average. This is pretty strange.

(d) Since we can fix 4 breakdowns, it's the fifth breakdown that's trouble.

(WIS) Let $Y \sim NegBinom\left(k = 5, p = \frac{1}{100}\right)$ be the number of days til the fifth breakdown. We'll survive $Y + 20$ days, and so (E),

$$E[Y + 20] = E[Y] + 20 = \frac{k}{p} + 20 = \frac{5}{1/100} + 20 = \boxed{520 \text{ days}}.$$

Commentary: This problem reiterates many of the themes of recent problems. In part a, you need to view the world through a statistical lens and recognize the Negative Binomial situation. Then, you must develop confidence in what looks like an absurd answer. Part b happens so quickly that you might have the opposite reaction: "I did something wrong because the problem was so short and clean." Toss out all these narratives and ask yourself if you did what the problem asked. Also note that part b should make you feel better about your answer in part a. If the average time for the four breakdowns is 400 days, then it is quite unlikely to get $X \leq 7$. Part c is here to see if you've internalized the idea of measuring how far apart two numbers are (a given value and the expectation) in terms of some measuring stick (SDs). See the Mookie Betts problem for the same idea. Part d offers the chance to make an off-by-one error (OBOE) ($k = 5$ vs. $k = 4$).

29. For any discrete RV, the probabilities specified by the probability formula must add up to one as you consider all values in the support. That is, $1 = \sum_{x \in \text{support}} P(X = x)$. For each of the below distributions, show that the probabilities add up to 1. In each case, you will need to use some important fact/result from your *previous* mathematics coursework. Make sure you **explicitly note the fact as you use it**; you don't need to re-derive/prove the fact.

(a) $X \sim Geom(p)$

(b) $X \sim Binom(n, p)$

(c) $X \sim Pois(\lambda)$

Hint List:

- a: Fact usually seen in Algebra 2 and Precalculus
- b: Fact usually seen when learning induction in Algebra 2 or Precalculus
- c: Fact first seen near the end of integral calculus

Solution:

(a) For $X \sim Geom(p)$, we know $P(X = x) = (1-p)^{x-1} \cdot p$.

$$\text{So, } \sum_{x=1}^{\infty} P(X = x) = \sum_{x=1}^{\infty} (1-p)^{x-1} p \overset{!}{=} \frac{(1-p)^{1-1} p}{1-(1-p)} = \frac{p}{p} = 1.$$

At $\overset{!}{=}$, we used the **sum formula for an infinite geometric series**.

(b) For $X \sim Binom(n, p)$, we know $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$.

$$\text{So, } \sum_{x=0}^{n} P(X = x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \overset{!}{=} (p + (1-p))^n = 1^n = 1.$$

At $\overset{!}{=}$, we used the **Binomial Theorem**: $(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$ with $a = p$ and $b = 1-p$.

(c) For $X \sim Pois(\lambda)$, we know $P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$.

$$\text{So, } \sum_{x=0}^{\infty} P(X = x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \overset{!}{=} e^{-\lambda} e^{\lambda} = 1.$$

At $\overset{!}{=}$, we used the **Taylor series** for e^x: $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ with $n = x$ and $x = \lambda$.

Commentary: Mathematics is a massively complex web of conceptual interconnectivity. College is often the place that this becomes harshly clear: you have a history with the subject that spans more than a decade, and new knowledge might build on *any* element from this long past. If you've had inadequate instruction at any point in this span, these holes might prevent future progress. This problem is designed to help you see this. Many previous homework problems have reminded you of the infinite geometric series formula, allowing for a seamless connection to a previous idea. The results in parts b and c may or may not have been taught by previous instructors (and may or may not have been forgotten, if taught). The intricacy of this web of connectivity makes math/stat especially unforgiving as a subject: holes in your previous understanding generate holes in your future understanding at an exponentially growing rate.

30. Most people don't last long on diets. In fact, you can model how long a person will last before giving up on a diet. Let $q(t) = \begin{cases} c \cdot e^{-t/4}, & t \geq 0 \\ 0, & \text{else} \end{cases}$ be the pdf of a continuous random variable for the time (in days) until a random person quits a diet.

(a) Find the particular number c must be if $q(t)$ is to be a valid pdf. (You'll need this value below.)

(b) Find the probability a random person lasts exactly 6.283185 days before quitting. [WISE]

(c) Find the probability a random person lasts at least a week before quitting.

(d) A clickbaity e-mag is looking for a quote of the form "90% of all dieters quit within x weeks!". Find x. (x need not be an integer.)

(e) On average, how long does a person last on a diet? (Carefully show every step! This will take a good deal of work, which should not be skipped, even if you know some quick approach.)

Hint List:

- a: The area under a pdf must be a probability, and probabilities must be in the interval $[0, 1]$. This forced two conditions on the pdf, which, in turn, will force a single value for c.
- c: Remember to switch improper integrals to proper integrals with limits. You could also use complementary probability to avoid the improper integral in this problem.
- e: Try using integration by parts on the integral that arises. Also, you'll need L'Hôpital's Rule to complete the problem.

Solution:

(a) If $c \geq 0$, we will have $q(t) \geq 0$ for all t, which is one of the two requirements of a pdf. The other condition is that $\int_{-\infty}^{\infty} q(t)\,dt = 1$. Given that the support is is $[0, \infty)$, this creates a simpler integral:

$$1 = \int_0^{\infty} ce^{-t/4}\,dt = c \cdot \lim_{n\to\infty} \int_0^n e^{-t/4}\,dt = c \cdot \lim_{n\to\infty} \left[-4e^{-t/4}\right]_0^n = c \cdot \lim_{n\to\infty} \left(4 - 4e^{-n/4}\right) = 4c$$

So, $\boxed{c = \frac{1}{4}}$.

(b) (WIS) Let X be the quitting time (in days) of a random person on a diet. We want (E) $P(X = 6.283185) = \boxed{0}$, as is the case with any continuous RV.

(c) Here, we find:

$$P(X \geq 7) = \int_7^{\infty} \frac{1}{4}e^{-t/4}\,dt = \lim_{n\to\infty} -e^{-t/4}\Big|_7^n = \lim_{n\to\infty} -e^{-n/4} + e^{-7/4} = \boxed{e^{-7/4} \approx 0.174}.$$

(d) We are trying to create the setup on the next page where the shaded area is 0.9 and the graph is $q(t)$: Thus, we need

$$0.9 = \int_0^x \frac{1}{4}e^{-t/4}\,dt = -e^{-t/4}\Big|_0^x = 1 - e^{-x/4}.$$

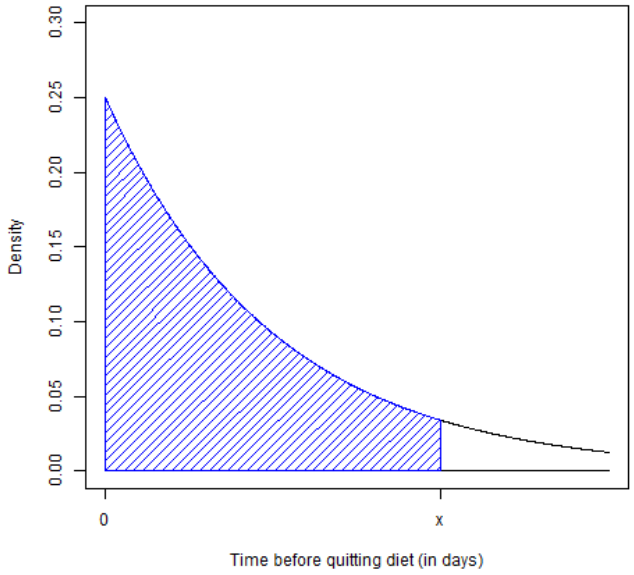
So, $e^{-x/4} = 0.1 \implies -\frac{x}{4} = \ln 0.1 \implies x = -4\ln 0.1 = 4\ln 10$. Since x is in days, we divide by 7 to get the time in weeks: $\boxed{\frac{4}{7}\ln 10 \approx 1.316 \text{ weeks}}$.

(e) We want $E(X) = \int_{-\infty}^{\infty} t \cdot q(t)\,dt = \frac{1}{4}\int_0^{\infty} te^{-t/4}\,dt$. We solve the indefinite integral first using integration by parts. Let $u = t$ and $dv = e^{-t/4}\,dt$. This gives $du = dt$ and $v = -4e^{-t/4}$. So we see:

$$\int te^{-t/4}\,dt = -4te^{-t/4} + \int 4e^{-t/4}\,dt = -4te^{-t/4} - 16e^{-t/4} + C$$

Thus,

$$E(X) = \frac{1}{4}\lim_{n\to\infty}\left[-4te^{-t/4} - 16e^{-t/4}\right]_0^n = \frac{1}{4}\lim_{n\to\infty}\left(\frac{-4n}{e^{n/4}} - 16e^{-n/4} + 4\cdot 0 + 16\cdot 1\right)$$



The limit of the first piece requires L'Hôpital's Rule to proceed:

$$\lim_{n\to\infty} \frac{-4n}{e^{n/4}} \overset{\text{LH}}{=} \lim_{n\to\infty} \frac{-4}{e^{n/4}\cdot\frac{1}{4}} = 0$$

Thus, $E(X) = \frac{1}{4}(0 - 0 + 0 + 16) = \boxed{4 \text{ days}}$.

Note: Once you learn the Exponential Distribution, you'll see this problem as an application of that model where $\lambda = \frac{1}{4}$. You'll also memorize that $E(\text{Exponential Model}) = \frac{1}{\lambda}$, giving $\frac{1}{1/4} = 4$ with no work!

Commentary: This problem takes you through the most common questions asked about CRVs. Part a focuses on the pdf. Part b checks if you understand the probability is 0 for a single value. Part c checks if you can find the probability over a given interval. Part d reverses this: given a probability, find the interval. Part e asks for a fact (here, expectation) about the CRV. Any time you learn a new idea/topic, there will be a set of *standard questions* asked about that topic. Much of academia, especially at the undergraduate level and below, is simply internalizing these standard problem types and developing high accuracy with their solutions. Top students naturally look for these common questions, anticipating what will show up on assessments. When you begin teaching a subject, these common question types become immediately clear. Since you are no longer struggling in the weeds of the concepts, it is easy to step back see greater levels of organization. Seek out and catalog the common problem types asked for every new topic you encounter.

31. The **median** of a CRV X is a place m on the horizontal axis such that $P(X \leq m) = \frac{1}{2} = P(X \geq m)$. Find the median of the CRVs with the below pdfs. In addition, draw a picture of each pdf (by hand or in Desmos), mark the median you find, and see if about half the area is on each side of the median. Include your sketch.

    (a) $f(x) = 20x^3(1-x)$ where $0 < x < 1$, and 0 elsewhere.

    (b) $g(x) = \dfrac{1}{\pi(x^2+1)}$ where x is any real number. (Once you're done with the problem, your answer will make perfect sense based on your picture.)

Hint List:

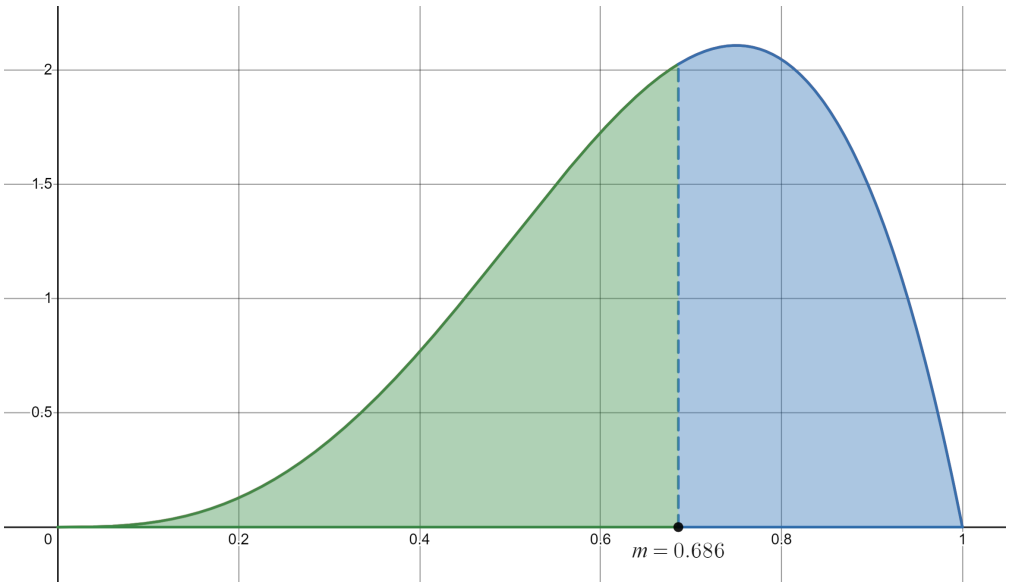
- a: Set up an integral that goes from 0 to m under the pdf. How much area should this integral have?
- b: Memorize the derivatives of inverse trig functions.

Solution:

(a) We want

$$\frac{1}{2} = \int_0^m 20x^3(1-x)\,dx = \int_0^m 20x^3 - 20x^4\,dx = 5x^4 - 4x^5\Big|_0^m$$

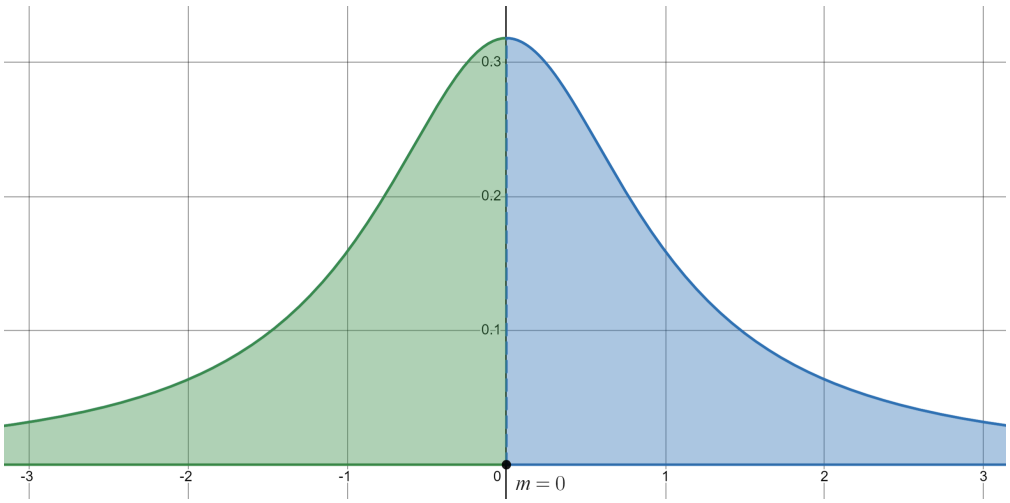
So, $\frac{1}{2} = 5m^4 - 4m^5$. This requires (see commentary) a calculator to solve: $\boxed{m \approx 0.686}$.



(b) We want

$$\begin{aligned}\frac{1}{2} &= \int_{-\infty}^m \frac{1}{\pi\,(x^2+1)}\,dx = \frac{1}{\pi}\cdot\lim_{n\to-\infty}\int_n^m \frac{1}{1+x^2}\,dx \\ &= \frac{1}{\pi}\cdot\lim_{n\to-\infty}\tan^{-1}x\Big|_n^m = \frac{1}{\pi}\left[\tan^{-1}m - \lim_{n\to-\infty}\tan^{-1}n\right] \\ &= \frac{1}{\pi}\left[\tan^{-1}m - \left(-\frac{\pi}{2}\right)\right] = \frac{1}{\pi}\left(\tan^{-1}m + \frac{\pi}{2}\right).\end{aligned}$$

Thus, $\frac{\pi}{2} = \tan^{-1}m + \frac{\pi}{2} \implies \tan^{-1}m = 0$. So, $\boxed{m = 0}$. Note that if you plot the pdf, it is symmetric across the y-axis. So, half the area is to each side of $x = 0$ and we immediately get the answer:



Commentary: I included this problem to expose you to an important concept (the median), to refresh your integration skills, and to raise an important issue: When can a problem be solved "by hand" (i.e., without the use of any technology, as in part b), and when must we use a calculator to help (as in part a)? I find that most people have no sense for this because they were either trained in an environment that always allowed a calculator or in one that never allowed it. Both of these pedagogical approaches are non-optimal, for they don't help students understand the boundaries of tool usage (in this case, the calculator). Here is what's in my head: You can always solve linear and quadratic equations by hand, although if the numbers are messy, you might use a calculator. Degree 3 and higher polynomials usually require technology except in special cases (e.g., $x^3 = 18$ or $4x^4 = x^2$). When we only have one kind of function in our equation (e.g., $2\ln x = 7$ or $\sin x = 1$), we can often solve this by hand, but when functions are mixed, we usually need a calculator (e.g., $2\sin x = 1 - x$ or $e^x - 3 = \cos x$). It is important to think about *how and when* you use tools in life. Some examples of misusing tools include: calculator over-reliance (makes mathematical work inefficient or inexact), failing to use comments when programming (generates hard-to-follow code), and always using a tuning device to check intonation for a musical instrument (leads to an underdeveloped musical ear).

32. How long does it take to recover from COVID-19? Based on this paper focused on patients from India, the average recovery time for people under 60 years old was 21 days. Let IPU60 stand for a random "Indian Patient Under 60".

For each part in this problem, do it two ways: 1) by hand, getting an exact answer, and 2) by writing a single line of code in R that checks your answer.

(a) Find the probability an IPU60 recovers in a week or less. [WISE, VIZ]

(b) Based on this paper with mostly US patients, it seems that Remdesivir can shorten recovery times to 2/3 of what they are without the drug. If this finding carries over to IPU60s, find the answer to part a if all patients are given Remdesivir. [WISE]

(c) The news often shares stories of patients who have recovered after long battles with the illness. If the media wants to focus on those who are among the longest 1% recovery times (without Remdesivir), how long must an IPU60 have taken to recover (at least)? [VIZ]

Hint List:

- a: Here, we are waiting for someone to get better. What distribution models the *time* needed *until* something occurs?
- a,b: If you're frustrated because the question is about weeks while the setup mentions days, just convert my question to an equivalent one with the units you want.
- c: The visualization is really helpful for me when doing problems like this.

Solution:

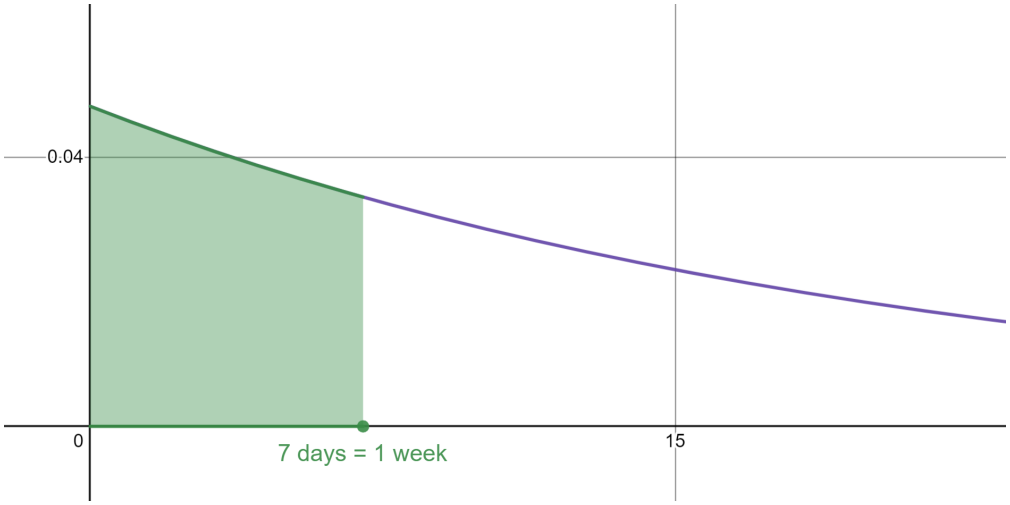
(a) (WIS) Let $X \sim Exp\left(\lambda = \frac{1 \text{ recovery}}{21 \text{ days}}\right)$ be the recovery time of a random IPU60 in days.

We want (E) $P(X \leq 7) = \int_0^7 \frac{1}{21} e^{-t/21}\, dt = -e^{-t/21}\Big|_0^7 = \boxed{1 - e^{-1/3}} \approx 0.283.$

In R, we write:

```
pexp(7, rate = 1/21, lower = T)
```

which gives the same answer (the last parameter is optional since true is the default value).



(b) (WIS) Let $Y \sim Exp\left(\lambda = \frac{1 \text{ recovery}}{14 \text{ days}}\right)$ be the recovery time of a random IPU60 on Remdesivir. Note that $21 \cdot \frac{2}{3} = 14$.

Again, we want (E) $P(Y \leq 7) = \int_0^7 \frac{1}{14} e^{-t/14}\, dt = -e^{-t/14}\Big|_0^7 = \boxed{1 - e^{-1/2}} \approx 0.393.$

In R, we write:

```
pexp(7, rate = 1/14)
```

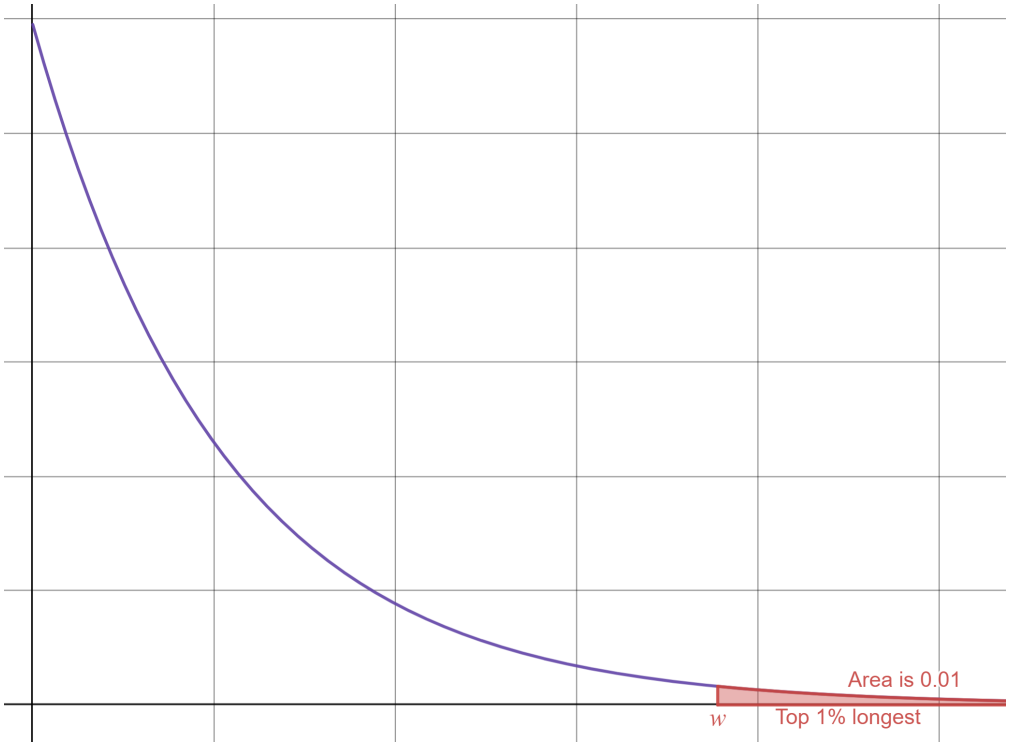
(c) We need $0.01 = \int_w^\infty \frac{1}{21} e^{-t/21}\, dt$ or, using complementary probability:

$$0.99 = \int_0^w \frac{1}{21} e^{-t/21}\, dt = -e^{-t/21}\Big|_0^w = 1 - e^{-w/21}$$

Rearranging gives $e^{-w/21} = 0.01 \implies \boxed{w = -21 \ln 0.01} \approx 96.709$ days.

In R, we write:

```
qexp(0.01, rate = 1/21, lower = F)
```



Commentary: One sign of deep understanding is the ability to express and coordinate what is desired across multiple formats. In part c, we have the desired idea in **words** ("the cutoff to be among the longest 1% recovery times"), in **symbols** ($0.01 = \int_w^\infty \frac{1}{21} e^{-t/21}\, dt$, and the complementary statement), in a **visualization** (the graph of the pdf with shading), and in computer **code** (the R statement). In this class, you'll often find me forcing you to do re-expression and coordination across formats. At times, this will feel unnecessary because the different formats bring equivalent insight to the solution pathway. However, this is not always the case, and many of my own greatest intellectual insights have come from switching between formats. An extreme example of this is the "proof without words". This is a visualization that immediately convinces you of a mathematical statement. So, instead of trudging through a nightmarish argument using the format of mathematical language, the format of visualization offers such profound insight that the solution is immediate. Consider the sum of cubes formula:

$$\sum_{i=1}^{n} i^3 \stackrel{?}{=} \left(\frac{n(n+1)}{2}\right)^2 = (1 + 2 + \cdots + n)^2$$

One can prove the $\stackrel{?}{=}$ using an exhausting induction argument or, by switching formats, see the result immediately:

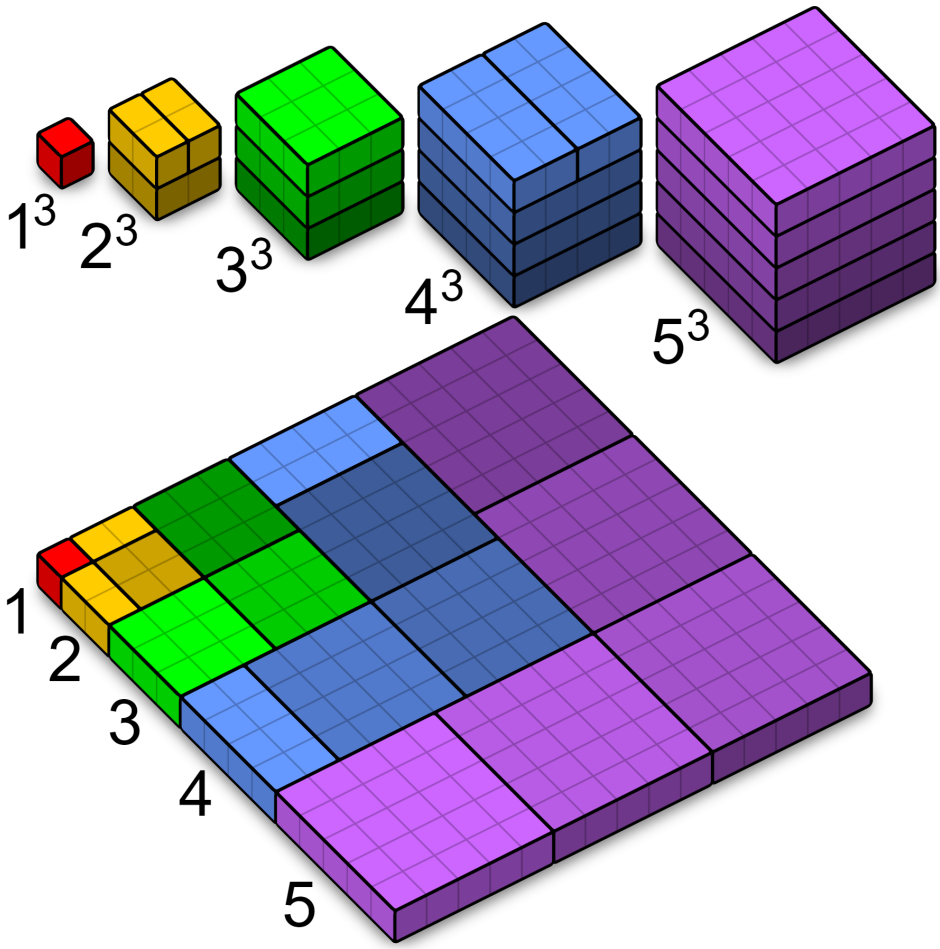


Image by Cmglee on Wikipedia's Proof Without Words page

As you move forward in academia, spend time switching between representational formats (known as "semiotics"). In some cases, a different format might provide an easier way forward.

33. Let $X \sim Unif(a, b)$ where $a < b$.

(a) Prove that $SD(X) = \dfrac{b-a}{\sqrt{12}}$ by first finding $Var(X)$ using the formula for the variance of a CRV.

(b) Check that the above formula is correct by running a simulation in R that generates 10000 random values from $Unif(3, 7)$.

Hint List:

- a: Start by finding $E[X^2]$.
- a: Recall the difference of cubes formula: $x^3 - y^3 = (x-y)(x^2+xy+y^2)$. Don't be afraid to push through the messy algebra.

Solution:

(a) Let $X \sim Unif(a,b)$. First, we find

$$E[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{x^3}{3}\Big|_a^b = \frac{b^3-a^3}{3(b-a)} \stackrel{!}{=} \frac{(b-a)(b^2+ab+a^2)}{3(a-b)} = \frac{b^2+ab+a^2}{3}$$

where we have used the difference of cubes formula at the $\stackrel{!}{=}$.

Since we know $E[X] = \dfrac{a+b}{2}$, we have

$$\begin{aligned} Var[X] = E[X^2] - E[X]^2 &= \frac{4}{4} \cdot \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} \cdot \frac{3}{3} \\ &= \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} = \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}. \end{aligned}$$

Thus, $SD[X] = \sqrt{\dfrac{(b-a)^2}{12}} = \dfrac{|b-a|}{\sqrt{12}} = \boxed{\dfrac{b-a}{\sqrt{12}}}$ since $b > a$.

(b)
```
vals = runif(10000, 3, 7)
sd(vals)            # I got 1.151255 (answers will vary)
(7-3)/sqrt(12)      # I got 1.154701
```

Commentary: As you venture deeper into mathematics/statistics, you will see more claims made without proof that are left as "exercises to the reader". This happens when the author does not want to spend the space fleshing out the details of a claim they believe are routine and follow from straightforward application of definitions or procedures. There are often good reasons for omitting the details of simpler claims: space may be limited in a publication, you may want to keep the reader's focus on a particular idea, you may not have the time to present the details (e.g., in a lecture), etc. So, the point of this problem was for you to experience the ideal "exercise to the reader" cycle: 1) I omit the details of a claim (e.g., $SD[Unif(a,b)] = \dfrac{b-a}{\sqrt{12}}$) for a good reason and suggest you provide these, 2) You flesh out those details. In the process of 2), you learn the result, strengthen skills from your past (e.g, difference of cubes formula), and gain insight into pieces of the claim (e.g., the mysterious 12 in the denominator comes from creating a common denominator of 3 and 4).

34. I was recently at the gym using the elliptical machine and noticed a Data menu, from which I got the below picture using my cell phone. When you're running, it is possible to measure "stride length". This is the distance, in inches, between two consecutive foot strikes. For each successive step I take, there will be a slightly different stride length based on my energy level, the terrain I face (in real life or on the machine), etc. Suppose we model my stride length via a normal distribution with mean 23 inches and standard deviation 1.8 inches.



(a) Based on my observations, the elliptical will highlight the image of the guy running (the rightmost figure in the picture) when my stride length goes above 26 inches. What is the probability of this? Do the problem two ways: on the given distribution using R, and on a standard normal distribution using tables. [WISE, VIZ x 2, Tech and tables] Note: When you see VIZ on a problem involving the normal distribution, this means you should be drawing – either by hand (easiest) or in Desmos/R (nightmarish) – a bell-shaped curve and shading an area under it. Sometimes students will take screenshots of probability tables and think this is a visualization; this is NOT what I want. VIZ x 2 implies one picture should be on the original normal distribution from the problem, and one picture should be on the standard normal distribution.

(b) What is the probability my stride length will be exactly 23 inches?

(c) A trainer wants a benchmark for deciding when I'm not pushing myself very hard. Any time my stride length is among the bottom 10% of stride lengths, they'll encourage me to work harder. What length demarcates the bottom 10% of stride lengths? [VIZ, Tables only (no R)]

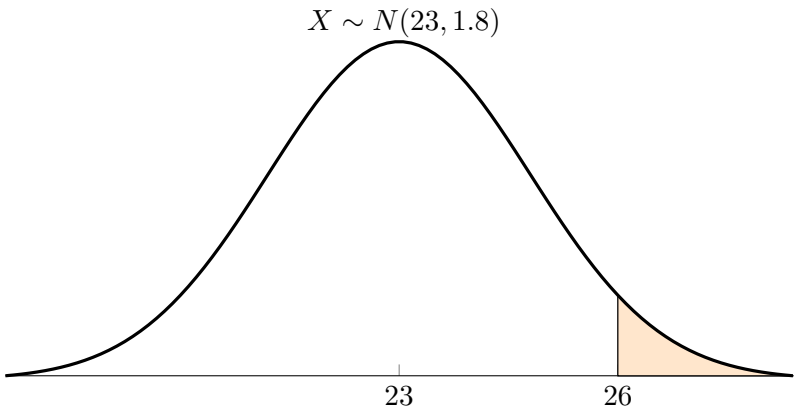
(d) I've found that if I stretch my legs before using the elliptical, then my stride lengths all increase by 2 inches. Write a random variable expression for this new situation, and then re-find the probability in part a using a single line of R code. [WISE]

Hint List:

Solution:

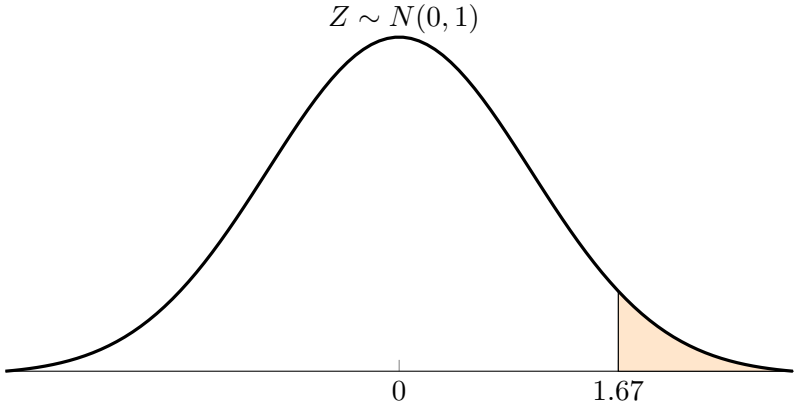
(a) (WIS) Let $X \sim N(23, 1.8)$ be the length in inches of a random stride I take. We want (E):

$$P(X \geq 26) = \text{pnorm}(26,\ 23,\ 1.8,\ \text{lower} = \text{F}) \approx \boxed{0.048}.$$



We can also do the problem using the standard normal, Z:

$$P(X \geq 26) = P\left(\frac{X-23}{1.8} \geq \frac{26-23}{1.8}\right) \approx P(Z \geq 1.67) \overset{\text{table}}{=} 1 - 0.9525 = \boxed{0.0475}.$$

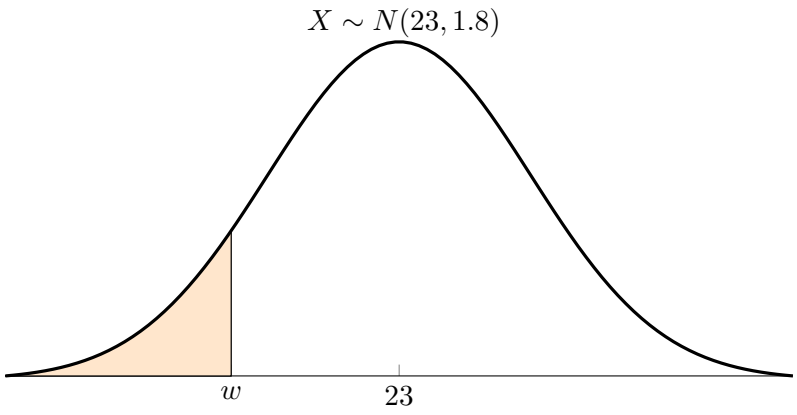


(b) $P(X = 23) = \boxed{0}$, as with any CRV.

(c) Below. we draw a picture where the bottom shaded area is 0.1. Now,

$$0.1 = P(X \leq w) = P\left(\frac{X-23}{1.8} \leq \frac{w-23}{1.8}\right) = P\left(Z \leq \frac{w-23}{1.8}\right)$$

Using a Z-table, we see $0.1 \approx P(Z \leq -1.28)$, so we set $-1.28 \approx \dfrac{w-23}{1.8} \implies \boxed{w \approx 20.696 \text{ inches}}$.



(d) (WIS) Let $Y = X+2$ be my random stride lengths after stretching. Note that $Y \sim N(23+2, 1.8) = N(25, 1.8)$. So (E),

$$P(Y \geq 26) = \text{pnorm}(26,\ 25,\ 1.8,\ \text{lower} = \text{F}) \approx \boxed{0.289}.$$

Commentary: Take a moment to compare and contrast this problem to the one a few back about quitting a diet. Like that problem, this one walks you through some of the standard prompts: finding areas given quantiles (part a), finding quantiles given areas (part c), and knowing $P(X = a) = 0$ (part b). Did you recognize these standard skills now that they are starting to recur? Unlike that problem, you'll notice that finding areas and quantiles for the normal distribution *cannot* be done algebraically. You must either use tables or a computer to estimate what you want.

This problem also highlights the idea of "seeing the world through a pair of statistical glasses". When you learn a subject, you don't just internalize facts about that topic. You also begin to *look at and interact with the world differently.* You start deploying the ways of thinking common in that discipline (e.g., studying math makes you think more deductively). You begin framing the world with structures from that discipline (e.g., studying statistics lets an athlete see their performance as a normal distribution with the goal of pulling the mean upward or downward, or reducing the SD). As I go through my day, my previous coursework influences the way I experience *every* moment. As an example, here are three thoughts I had the morning I wrote this commentary: "the amount of coffee in my Starbucks order was on the high end of the normal distribution" (statistics), "that politician is using ad hominem attacks rather than answering the question" (debate), and "the claim that a caterpillar's weight doubles every day in its infancy is an example of a Geometric sequence" (math). At its best, college is giving you many sets of glasses through which to look at the world. And the better you see the world, the more effective you can be when navigating it.

35. If you ask a classroom of students to write an essay on some topic, the responses you get vary in length quite dramatically. Suppose we create a random variable L for the length of a random student's essay (in words) on a given topic. We might try to model L using a normal distribution.

(a) Give **two** reasons why the normal distribution is *not* a perfect choice for this setting.

(b) Ignoring the issues in part a, we proceed anyway. Suppose a given essay prompt leads to lengths distributed as $L \sim N(1500, 200)$. As a student, you might want to avoid writing an essay that is too short (the bottom 15%; the prof might think you have little to say) or too long (the top 15%; the prof might mark you down for not being concise), so you decide to land somewhere in the middle 70% of the distribution. Between what word counts should you keep your essay? Do the problem on the original distribution using R to help, and again on the standard normal using tables. [VIZ, decimal values for word counts OK]

(c) Suppose for a given essay prompt the lengths are $L \sim N(2200, \sigma)$ where σ is unknown. Find σ if the prof tells the class "You were among the shortest 20% of essays if you wrote 1700 words or fewer". [VIZ, Tables not tech]

Hint List:

- a: When you are trying to decide if a particular model is a good fit for some phenomenon, you might consider the continuous/discrete nature of the model, the support of the model, or the shape of the model. Do these fit the phenomenon?
- c: Since we don't know σ, you should try to transition the problem to an equivalent one on the standard normal.

Solution:

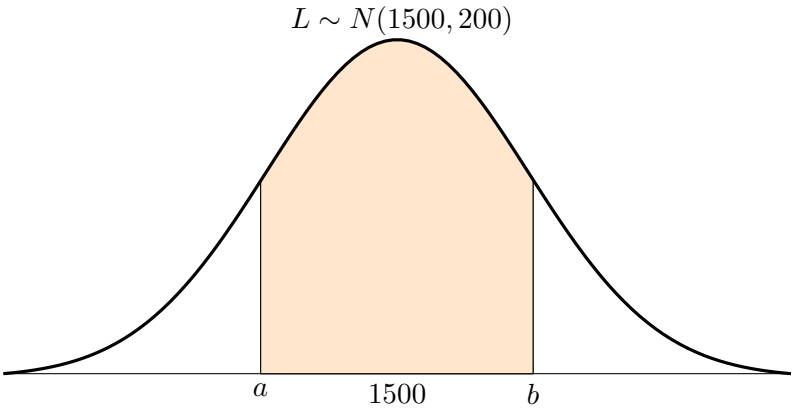
(a) In this problem, you could note:

- "Number of words" is a discrete idea, but the normal distribution is continuous.
- The support of the normal is $(-\infty, \infty)$, but the number of words can't be negative.
- Some classes set word/page maximums for essays, so counts can't be arbitrarily large, conflicting with the support of the normal model.
- If some students don't do the essay, their word counts will be zero, creating a bimodal graph (with peaks near 0 and the value most common for those completing the essay), conflicting with the shape of the normal.

(b) Let $L \sim N(1500, 200)$ be the number of words in an essay. We shade the middle 0.70 area leaving 0.15 in each tail. If we use this distribution, we do:

```
a = qnorm(0.15, 1500, 200, lower = T)
b = qnorm(0.15, 1500, 200, lower = F)
```

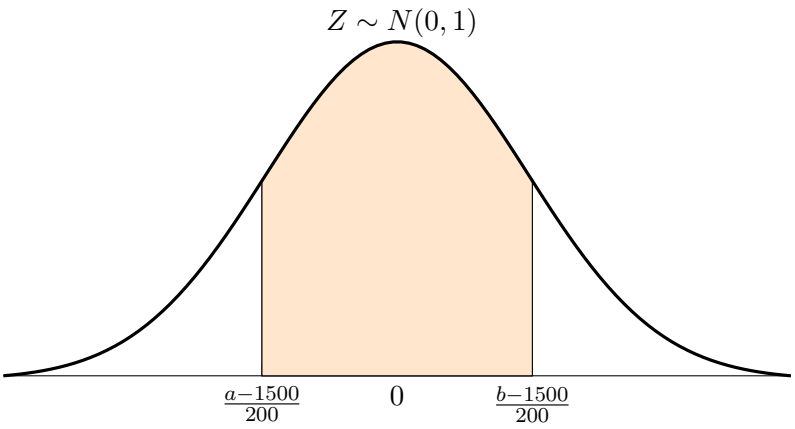
Thus, the middle 70% is between $\boxed{a = 1292.713 \text{ and } b = 1707.287 \text{ words}}$.



If we plan to use tables for the standard normal, we find that $P(Z < -1.03) \approx 0.15$ via a reverse area lookup (OK to have -1.04). So we set:

$$-1.03 \approx \frac{a - 1500}{200} \implies a \approx 1294 \text{ words}$$

$$+1.03 \approx \frac{b - 1500}{200} \implies b \approx 1706 \text{ words}$$

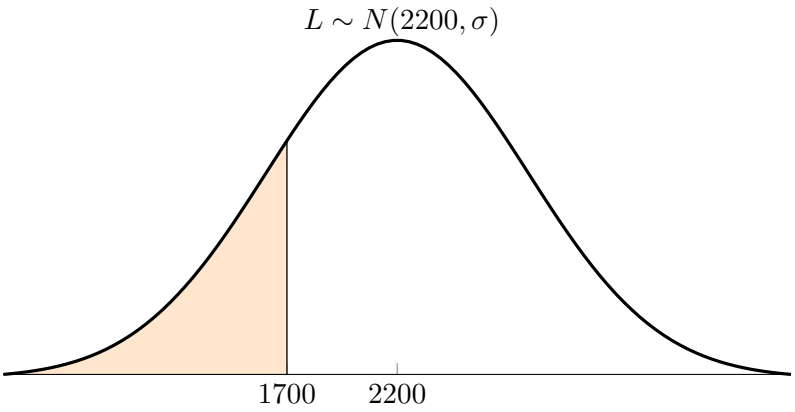


(c) Now, we let $L \sim N(2200, \sigma)$. The goal is to find σ given the below shaded area is 0.2. Since we can't work on a distribution whose SD is unknown, we transition to Z:

$$0.2 = P(L \leq 1700) = P\left(\frac{L - 2200}{\sigma} \leq \frac{1700 - 2200}{\sigma}\right) = P\left(Z \leq \frac{-500}{\sigma}\right).$$

Based on a Z-table, $0.20 \approx P(Z \leq -0.84)$. So,

$$\frac{-500}{\sigma} \approx -0.84 \implies \boxed{\sigma \approx 595.238 \text{ words}}$$



Commentary: We often use models that are "close enough" but not perfect. This can maddening to some when they first see it. Indeed, this problem was constructed specifically to bring out that discomfort. Part a addresses it head-on, and part b even suggests you give your answer as a decimal number of words (!). Why use "close enough" things? Sometimes it's because we don't have a good discrete model for what we are studying, but a continuous one is really close. Indeed, the list of common continuous models is longer than those that are discrete, as seen on this Wikipedia page. Sometimes, a continuous model might offer you more flexibility than a discrete alternative. For example, the normal distribution allows you to decide the mean and standard deviation (how kind!), whereas other models let you choose different parameters (e.g., λ in a Poisson distribution), but then you're stuck with whatever mean and SD follow from that choice (e.g., λ and $\sqrt{\lambda}$, respectively, for the Poisson). Finally, sometimes the places where the model and your phenomenon differ are unimportant or negligible. For example, when you read that "human heights can be modelled by a Normal distribution", this is not exactly true; the Normal distribution allows for negative heights, which are impossible. But think how close to

0 the graph of the pdf would be above the negative values. While the pdf is suggesting the possibility of negative heights, the area above $(-\infty, 0)$ might be $< 10^{-50}$ (or some tiny number). Remember that the act of modelling is *always* an imperfect action. That is the downside. The upside is that the model allows you to easily make calculations and predictions. In many cases, this is a trade-off worth making. As a stark example, if someone could predict the date of your death, but would be off by at most a week, would you value this imperfect model?

36. We have see several cases of finding the expectation of a CRV X using the formula

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Sometimes, however, we wish to find the expectation after a function, $g(x)$, has been to our RV: $E[g(X)]$. There are several ways to do this, which we explore in this problem. Here, let $X \sim Unif(0, 1)$, which has pdf $f(x) = 1$ for $0 < x < 1$, and 0 otherwise. Suppose we want to find the expectation of a new CRV $Y = -3 \ln X$. Here, $g(x) = -3 \ln x$.

(a) One approach is to use a definition from probability theory:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx$$

Find $E[Y]$ using this formula. As always, show all your steps when evaluating the integral.

(b) Another approach is to use simulation. In R, generate 10000 values from X, push each value through the formula given by $g(x)$, and then take the average of your 10000 transformed values. Include your code and final result. Your answer should be close to that in part a.

(c) In some cases, the RV $Y = g(X)$ might be recognizable as a distribution you already know. Continuing your work from b, make a histogram of your 10000 transformed values and use this to guess what distribution $Y = -3 \ln X$ secretly is. (Thinking about supports is useful here!) Include your code and a sketch/screenshot of your histogram. Then, write down your distribution guess, and find its expectation. Your answer should match parts a and b. (You don't need to prove your guess; you learn how to do so in a class like Math 180A.)

Hint List:

- a: This problem will require three important ideas from integral calculus: Type II improper integrals, techniques of integration, and L'Hôpital's Rule – so you should refresh these if needed.
- c: If $x \in (0,1)$, what zone will $\ln x$ be in? What about $-3\ln x$?

Solution: Let $X \sim Unif(0,1)$ and set $Y = -3\ln X$.

(a) By the given formula, $E(Y) = E[g(X)] = \int_0^1 -3\ln x \cdot 1\, dx$. Since $\ln x$ has a vertical asymptote at $x = 0$, this a Type II improper integral. We use a limit to help and also proceed with integration by parts (IBP: $u = \ln x$, $dv = dx$)

$$\lim_{t\to 0^+} \int_t^1 -3\ln x\, dx \overset{\text{IBP}}{=} -3 \lim_{t\to 0^+} \left[x\ln x \Big|_t^1 - \int_t^1 \frac{1}{x}\cdot x\, dx \right] = -3 \lim_{t\to 0^+} (0 - t\ln t - (1-t))$$

L'Hôpital's Rule is needed to resolve this limit:

$$\lim_{t\to 0^+} t\ln t = \lim_{t\to 0^+} \frac{\ln t}{t^{-1}} \overset{\text{LH}}{=} \lim_{t\to 0^+} \frac{1/t}{-t^{-2}} = \lim_{t\to 0^+} -t = 0$$

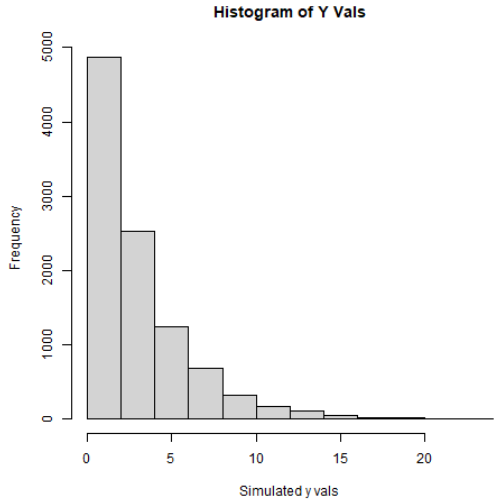
Putting this all together gives $E(Y) = -3(0 - 0 - (1-0)) = \boxed{3}$.

(b) In R, we do:

```
vals = runif(10000, 0, 1)
transformedvals = -3 * log(vals)  # vectorized operations
mean(transformedvals)
```

My random result was about $2.992882 \approx 3$.

(c) Note that if $x \in (0,1)$, then $\ln x \in (-\infty, 0)$, and thus, $-3\ln x \in (0,\infty)$. So, the support of Y must be $(0,\infty)$. The only CRV you know with this support is $Exp(\lambda)$. Making a histogram of the data from part b confirms that the data have the correct shape:



The only question is: What is the value for λ if this really is an exponential distribution? Here, the result from part a helps. We know $E(Y) = 3$, but we also know that $E[Exp(\lambda)] = \dfrac{1}{\lambda}$. This forces $3 = \dfrac{1}{\lambda}$, or $\lambda = \dfrac{1}{3}$. Thus, it appears $\boxed{Y \sim Exp(\lambda = 1/3)}$.

Commentary: This problem was constructed to do several things:

- Push hard on ideas from previous coursework. See part a.
- Explore an idea ($Y = -3 \ln X$ and its expectation) using many different representations: formulaic (part a), empirical (part b), and graphical (part c). Hopefully, by the end you are fairly convinced that $Y \sim Exp(\lambda = 1/3)$. Note that we have not *proven* this fact. To do so requires some tools from Math 180A.
- Offer you a chance to make a conjecture. After seeing the final result of part c, my hope is that you might ask: In general, is it true that $-\lambda \cdot \ln [Unif(0, 1)] \sim Exp(1/\lambda)$?

37. Let $A \sim N(\mu = -2, \sigma = 3)$, $B \sim N(\mu = 4, \sigma = \sqrt{5})$, and $C \sim Unif(a = -2, b = 4)$ be independent random variables. For the below problems, **find the distribution** of the random variable expression given. To "find a distribution" means to decide its shape (e.g., Normal, Exponential, Geometric, etc.) and the parameters that define it (e.g., μ, σ, etc.). In some cases, you may only be able to find an *approximate* distribution that the random variable expression follows.

(a) $2A - 4$

(b) $A + 3B$

(c) $C_1 + C_2 + \cdots + C_{1000}$ where $C_1, C_2, \ldots, C_{1000}$ are independent values from the C distribution.

(d) Write a few lines of code in R that convince you that your answer to part c is correct.

Hint List:

- a,b,c: Recall formulas like $E(X+c) = E(X) + c$ and $Var(X+Y) \overset{\text{ind.}}{=} Var(X) + Var(Y)$. Think carefully about *why* the shape of the answer is what you claim.
- d: This is intentionally open-ended. Don't let its lack of specificity stop you from acting.

Solution:

(a) $2A - 4$ will be normal because A is normal and multiplying, dividing, adding, or subtracting constants keeps the normal shape (this is not true for other distributions!). Now,

$$E[2A-4] = 2E[A] - 4 = 2(-2) - 4 = -8$$

$$SD[2A-4] = SD[2A] = |2|SD[A] = 2 \cdot 3 = 6$$

So, $\boxed{2A - 4 \sim N(-8, 6)}$.

(b) $A+3B$ will be normal because A and $3B$ are normal, and adding/subtracting independent normal distributions keeps the normal shape (again, this is not true for other distributions!). Now,

$$E[A+3B] = E[A] + 3E[B] = -2 + 3 \cdot 4 = 10$$

$$Var[A+3B] \overset{\text{ind.}}{=} Var(A) + Var(3B) = Var(A) + 3^2 Var(B) = 3^2 + 9(\sqrt{5})^2 = 54$$

So, $\boxed{A + 3B \sim N(10, \sqrt{54})}$.

(c) By the Central Limit Theorem, $C_1 + \cdots + C_{1000}$ is *approximately* normal. Since $C \sim Unif(-2, 4)$, we find that $E(C) = \dfrac{-2+4}{2} = 1$ and $Var(C) = \dfrac{[4-(-2)]^2}{12} = 3$. Thus,

$$E[C_1 + \cdots + C_{1000}] = E[C_1] + \cdots + E[C_{1000}] = 1000 \cdot E(C) = 1000$$

$$Var[C_1 + \cdots + C_{1000}] = Var(C_1) + \cdots + Var(C_{1000}) = 1000 \cdot Var(C) = 3000$$

So, $\boxed{C_1 + \cdots + C_{1000} \approx N(1000, \sqrt{3000})}$.

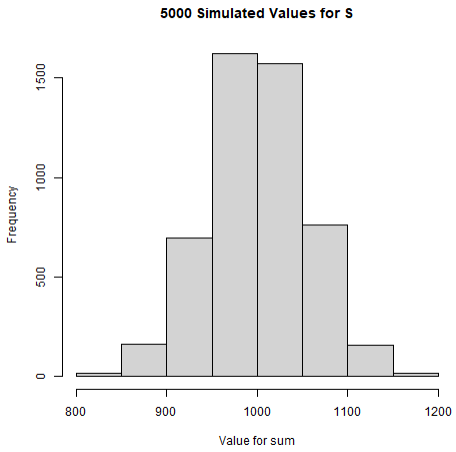
(d) Answers will vary. You might generate a bunch of values for the expression $S = C_1 + \cdots + C_{1000}$ and use R to see if these values have the claimed shape, mean, and variance (or SD):

```
svals = replicate(5000, sum(runif(1000, -2, 4)))

# hoping for approximate normality in histogram
hist(svals, main="5000 Simulated Values for S", xlab="Value for sum")

# hoping for an expectation of 1000
mean(svals)

# hoping for a variance of 3000
var(svals)
```



Commentary: Parts a and b reveal two interesting things. First, notice in part a, the solution directly finds the SD (rather than going through the variance and taking the square root as in part b). This is possible because part a only has *one* random variable, and so everything else present is a constant. Luckily, we have formulas for SD and constants: $SD(X+c) = SD(X)$ and $SD(aX) = |a|SD(X)$. We have yet to introduce a rule for SD when two independent RVs are present. Such a rule exists, but it is actually just the variance sum rule in disguise:

$$SD(X+Y) = \sqrt{Var(X+Y)} \overset{\text{ind.}}{=} \sqrt{Var(X)+Var(Y)} = \sqrt{SD(X)^2+SD(Y)^2}$$

Some people call this the "Pythagorean Theorem for Standard Deviations", but I don't bring this up in class or give it a sexy name because I think this is a pedagogical mistake. The variance operator is the one that artfully deals with the sum of independent random variables (see the red text above!), so I don't want to overlay an inelegant operator (SD) atop this.

The second important distinction between parts a and b is how they get their normality. Part a takes a normal distribution and manipulates it with *constants*. Part b takes a normal distribution and manipulates it with *other normal distributions*. These are different changes, and you might land back in the normal universe or not. One of the nice properties of a normal distribution is that we land back in the normal universe in *both* of these cases. Other distributions aren't so kind to you. For example, constant manipulations to a Uniform will keep us Uniform (e.g., $Unif(a,b)+c \sim Unif(a+c, b+c)$), but adding two Uniforms doesn't give us a Uniform back!

Finally, part d was intentionally designed to be very open-ended. How exactly should you write some code to convince someone of your answer? By the end of high school, students have come across a wide range of tasks. Some are very well-structured ("Solve $x^2 - 4x + 3 = 0$ using factoring.") while others are ill-structured ("Discuss the theme of heroism in book X."). It is tempting to hope that STEM classes will only give you really well-structured problems; in fact, students are often drawn to STEM because this is so common. But, it is useful to recognize that as you proceed deeper into STEM coursework, ill-structured problems are more and more common. Indeed, the top end of these fields is mostly ill-structured: "Discover something interesting about mathematical object Y", "Build me a better cell phone case", "Figure out how to colonize Mars". If you're going to have the courage to tackle these types of problems, you have to start now with problems like part d.

38. Suppose $X \sim N(\mu, \sigma)$ and let $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ where $X_1, X_2, \ldots, X_n$ are independent values from X and n is some fixed positive integer. Find the exact distribution of $\overline{X}$ (your answer will involve the variables in the problem). Then, memorize this result because we will need it for the rest of the class.

Hint List:

- Try following steps similar to part c of the last problem. This is tougher because we don't get to use specific numbers, but the logic is the same.

Solution: Let $X \sim N(\mu, \sigma)$. We know $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ is normal because the numerator is a sum of independent normal distributions (and hence normal) and dividing by a constant maintains normality. Now,

$$E[\overline{X}] = E\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{E[X_1] + \cdots + E[X_n]}{n} = \frac{nE[X]}{n} = \mu$$

$$Var[\overline{X}] = \frac{1}{n^2} Var[X_1 + \cdots + X_n] \overset{\text{ind.}}{=} \frac{1}{n^2}\left(Var[X_1] + \cdots + Var[X_n]\right) = \frac{1}{n^2}(n \cdot VarX) = \frac{\sigma^2}{n}$$

Thus, $SD(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$, and we have $\boxed{\overline{X} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)}$.

Commentary: It took me a long time to understand what the symbol $\overline{X}$ really means. As a RV, it is trying to convey that different values are possible for an average, and that the different values are not equally likely, but follow a normal distribution (in this case). In real life, we usually work so hard to collect a *single* data set and, from this, to find a *single* average, that it is hard to imagine we could have gotten a different data set, and hence, a different average. The essence of $\overline{X}$ became clearer when I started to see situations in life where the data $(X_1, \ldots, X_n)$ kept changing on me. Each time I get on an elevator, I think of the people's weights as the X_is and notice how $\overline{X}$ is a new value. Or, at the end of each school week, I think about my commute times for that week $(X_1, \ldots, X_5)$ and note how $\overline{X}$ is my average commute time for the week, which changes weekly due to random forces.

This problem is also difficult for some because everything is so general. We don't know the actual values for the mean and SD of the RV X, just that it's centered at μ with spread σ. Also, we don't know precisely how many terms are going into the random averages, just that n pieces of data are used, whatever n actually is. People often say that "there are no numbers in higher level math". While this statement is overly strong, it does capture an important transition for the subject: Mathematics often tries to uncover *general* truths, not *specific* results. As an example, if I note that 31 can be written as the sum of four squares ($31 = 1^2 + 1^2 + 2^2 + 5^2$), this comment is so specific that you might be underwhelmed. If, however, I note that **any** natural number can be written as the sum of four squares ($n = a^2 + b^2 + c^2 + d^2$), this general result is quite surprising. I find that general claims (like this problem or the Four Squares Theorem) are harder to initially grasp because they carry so much weight. Indeed, this problem shows that **for any** μ, **for any** $\sigma > 0$, and **for any** natural number n, we have $\overline{X} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$ when $X \sim N(\mu, \sigma)$; that's a lot of "for any"s!

39. I recently read "The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century" by Steven Pinker (one of my favorite authors). He cites an interesting paper that explores the amount and kinds of errors that college students make in their writing and how these behaviors have changed over time. Based on this paper, Pinker notes that students make about 4 errors per page related just to commas!

    (a) What's the probability a random page of student writing will contain 3 or more comma errors? [WISE]

    (b) Suppose that a teacher has this policy for a five-page assigment: "If more than one page of your assignment has 3 or more comma errors, you must rewrite the entire assignment." What percentage of students will have to rewrite their assignments? [WISE]

Hint List:

- a: This part is about the number of comma errors. Is this a discrete or continuous idea? What values are possible for the number of comma errors? When struggling, always start with the type (discrete vs. continuous) and support of the random idea. This will quickly get you to the correct model. If you feel like the prototypical description of the model does not match what is happening in this problem, see the below commentary.
- b: You'll need your answer from part a when setting up the new RV used here. This part is about pages (in a five-page essay) that meet the condition from part a. Ask yourself the two questions in the above hint.

Solution:

(a) (WIS) Let $X \sim Pois\left(\lambda = \frac{4 \text{ comma errors}}{1 \text{ page}}\right)$ be the number of comma errors a random student makes on a random page. We want (E):

$$\begin{aligned} P(X \geq 3) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - \frac{e^{-4}4^0}{0!} - \frac{e^{-4}4^1}{1!} - \frac{e^{-4}4^2}{2!} \\ &= 1 - e^{-4}(1 + 4 + 8) = \boxed{1 - 13e^{-4} \approx 0.762} \end{aligned}$$

(b) (WIS) Let $Y \sim Binom\left(n = 5, p = 1 - 13e^{-4}\right)$ be the number of pages out of 5 that have 3 or more comma errors for a random essay. Here (E),

$$\begin{aligned} P(\text{“rewrite”}) &= P(Y > 1) = 1 - P(Y = 0) - P(Y = 1) \\ &= 1 - \binom{5}{0}\left(1 - 13e^{-4}\right)^0\left(13e^{-4}\right)^5 - \binom{5}{1}\left(1 - 13e^{-4}\right)^1\left(13e^{-4}\right)^4 \approx \boxed{0.987} \end{aligned}$$

Commentary: Models often have a prototypical description. For example, a Poisson model counts the number of occurrences of an event (e.g., a car crash) over a **time span** assuming the event could happen at any moment with an average rate of λ. Notice that this problem isn't about time but about pages, which can upset some people. But what really is time? If you view time as a fixed span in which things occur, then a "page" is quite similar. Instead of car crashes over an hour, we have comma errors over a page. Basically, if we relax the prototypical description of a model (without losing its essence), we can use models to describe far more settings than we originally intended. For example, we might use Poisson to model the number of hurricanes during the summer, the number of injuries during a baseball season, the number of gaffes during a speech, the number of main characters in a novel, and so on.

40. Let $X \sim Exp(\lambda = 1)$. Suppose we set Y as the value of $X + 1$ rounded down: $Y = \lfloor X + 1 \rfloor$. For example, if $X = 3.7$, then $Y = \lfloor 3.7 + 1 \rfloor = \lfloor 4.7 \rfloor = 4$.

(a) Find $P(Y = 2)$.

(b) Find $P(Y = y)$ and state the support of Y.

(c) Using your result from part b, show that Y is a Geometric RV, $Geom(p)$, and state what value p is.

Hint List:

- a: What range of X values leads to $Y = 2$?
- b: Try to repeat your argument from part a, but use y instead of 2.
- c: The probabilities of a Geometric RV form a geometric sequence. What are these values? What number do you multiply by to get from one term to the next? Does your result from part b have such a pattern?

Solution: Let $X \sim Exp(\lambda = 1)$ and $Y = \lfloor X + 1 \rfloor$.

(a) Note that if a number rounds down to 2, it is in the interval $[2, 3)$. So,

$$P(Y = 2) = P(\lfloor X + 1 \rfloor = 2) = P(2 \leq X + 1 < 3) = P(1 \leq X < 2)$$
$$= \int_1^2 1e^{-1x}\, dx = -e^{-x}\Big|_1^2 = \boxed{e^{-1} - e^{-2}}.$$

(b) As in part a, if a number rounds down to y, it is in $[y, y + 1)$. So,

$$P(Y = y) = P(\lfloor X + 1 \rfloor = y) = P(y \leq X + 1 < y + 1) = P(y - 1 \leq X < y)$$
$$= \int_{y-1}^{y} 1e^{-1x}\, dx = -e^{-x}\Big|_{y-1}^{y} = \boxed{e^{-(y-1)} - e^{-y}}.$$

Since $X \geq 0 \implies X + 1 \geq 1 \implies \lfloor X + 1 \rfloor \in \boxed{\{1, 2, 3, \ldots\}}$.

(c) To show a distribution is Geometric, you must show its probabilities match those of $W \sim Geom(p)$ for some p. Since these probabilities are

$$P(W = 1) = p, P(W = 2) = qp, P(W = 3) = q^2p, \ldots$$

we need only show our probabilities are a Geometric sequence where the common ratio will equal q. Using the support and probability formula from part b, we get:

$$P(Y = 1) = e^{-0} - e^{-1}, P(Y = 2) = e^{-1} - e^{-2}, P(Y = 3) = e^{-2} - e^{-3}, \ldots$$

Notice you can get from one term to the next by multiplying by $q = e^{-1}$. Thus, we find that $\boxed{Y \sim Geom(p = 1 - q = 1 - e^{-1})}$.

Commentary: This problem had two goals. First, I wanted you to see a connection between a discrete RV (Geometric) and a continuous RV (Exponential). Since the supports of these are quite different, you need some operation that moves from the land of continuity to the land of discreteness. The round down function does this beautifully. When I was in school, I remember disliking the round down function; it has a weird graph and requires rounding in a different way than we normally do (rounding up or down depending on how a digit compares to 5). Only after I saw uses like in this problem did I begin to appreciate this function. I expect you'll have more moments like this in the years ahead; it is common to not completely see the value in something when you first encounter it. This is true for mathematics but also people, rules, arguments counter to what you believe, etc. Also, if you think the linkage between RVs seen here (thanks round down!) is rare, you'd be wrong. Many connections exist between the RVs you've been studying. You can see visualizations and read more here.

Second, I wanted to use parts a to c as a way to outline *how* mathematicians think. When a mathematician approaches a new object (here, $Y \sim \lfloor X + 1 \rfloor$), they usually start by trying to understand it in particular cases (part a). Often, the logic involved in understanding such cases can generalize (part b). Finally, as you prove results about an object (part b), you might be able to completely describe it (part c), sometimes reducing it to an object you already understand. Notice that I omitted a part d, which could have been: "At the start of the problem, if we change λ to a new value, how does this affect the results in parts a to c?". This final step of conjecturing is important, for it is how the field moves forward. In other problems, I explicitly forced you to do this conjecturing, but here, it was left off to see if I could tempt you into doing so without prompting. While some people are naturally curious about the world, I believe others need to foster this skill through opportunities that invite conjecturing.

41. Suppose you doubt your answers to parts a and c of problem 40. Write code in R that convinces you.

Hint List:

- What would convince you? Revel in the open-endedness of the problem.

Solution: For part a, we might generate 10000 values of Y and see what fraction of these equal 2:

```
vals = floor(rexp(10000, rate = 1) + 1)
mean(vals == 2)
```

When I ran this, I got 0.2363 which is about $e^{-1} - e^{-2} \approx 0.233$.

Note: Writing vals == 2 returns a vector of booleans (e.g., [TRUE,TRUE,FALSE,...]) which the computer interprets as 0s and 1s (e.g., $[1, 1, 0, \ldots]$). Thus. the mean $\left(\text{e.g., } \frac{1+1+0+\ldots}{10000}\right)$ is just the proportion of TRUEs.

For part c, we might compare our 10000 values from Y above and 10000 generated values from $Geom(p = 1 - e^{-1})$ using histograms or summary tables. Remember that when you generate random data using rgeom, you must add 1 because R's implementation of the Geometric distribution only counts the failures leading to the success, not the success and failures, as we do:

```
geomvals = 1 + rgeom(10000, 1 - exp(-1))
table(vals)
table(geomvals)
```

When I ran this code, I got:

```
     vals
   1    2    3    4    5    6    7    8    9
6357 2363  812  295  108   39   20    5    1

     geomvals
   1    2    3    4    5    6    7    8    9
6379 2304  863  282  110   39   16    5    2
```

Commentary: Continuing in the trend of open-ended (or ill-structured) problems, here you must *decide how* to create a convincing argument. I've presented one approach above, but answers will vary. Note that when you couple part c of the last problem with this problem, you have two approaches to showing $Y \sim Geom(p = 1 - e^{-1})$. The first, a deductive proof, is the bread-and-butter of mathematicians. The second, an empirical argument, is also common in society. How did these two approaches make you feel? Which actually convinced you more? While it is true that a deductive proof is the only way to be *logically* certain of a result, one surprising result from psychology research is that people put a lot of value in empirical arguments − sometimes *more* than a deductive argument. For problems that are difficult to mathematize (e.g., "Is climate change real?", "Will my partner cheat on me?"), data-driven arguments are often the best alternative. Of course, such arguments are imperfect, sometimes surprisingly so.

42. Suppose your lab has designed a new vaccine and is ready to enroll participants in trials. Volunteers come to your lab and take a quiz that returns their "risk score", a real number between 0 and 1 that indicates whether a person has a higher risk of contracting the disease (a value near 1) or not (a value near 0). Based on previous data, you know risk scores follow the distribution with density

$$f(x) = \begin{cases} \frac{15}{4}\sqrt{x}(1-x), & 0 \leq x \leq 1 \\ 0, & \text{else} \end{cases}$$

To speed up the race for a cure, your funding agency will only allow you to enroll participants with a risk score above 0.8. How many people will your lab need to screen, on average, in order to find 100 patients that meet this criterion? [WISE]

Hint List:

- Focus first on finding the probability that a single random person will have a high enough risk score to be included in the study. Then, set up a random variable that is concerned with getting 100 into the study.

Solution: (WIS) Let X be the risk score of a random patient. We have (E):

$$P(\text{accepted into study}) = P(X > 0.8) = \int_{0.8}^{1} \frac{15}{4}\sqrt{x}(1-x)\,dx = \frac{15}{4}\int_{0.8}^{1} x^{1/2} - x^{3/2}\,dx$$
$$= \frac{15}{4}\left(\frac{2}{3}x^{3/2} - \frac{2}{5}x^{5/2}\right)\Big|_{0.8}^{1} \approx 0.0698.$$

Now, (WIS) let $Y \sim NegBinom(k = 100 \text{ acceptances}, p \approx 0.0698)$ be the number of patients we must screen to get 100 who can be accepted into the study. We find (E):

$$E[Y] = \frac{k}{p} \approx \frac{100}{0.0698} \approx \boxed{1432.753 \text{ patients}}.$$

Commentary: The goal of this problem was to present a real-world scenario with a combination of random variables (here, X and Y). I wanted one of these to be a model you learn about in future courses ($X \sim Beta(3/2, 2)$) and the other to be one you should already have memorized ($Y \sim NegBinom(k, p)$). Notice that once I give you the pdf of X, you can get whatever you need (a probability, the expectation, the variance) with some work. Here, we found a probability with a messy integration. When you have a known distribution (e.g., Y), we can skip messy calculations if we're asked for the expectation or variance. This is one important reason to internal facts in life: they save you time on calculations. This savings is especially important as you move on to more advanced coursework. You need to free up some of your mental focus to direct toward the current challenges at hand. Notice that the cognitive demand of this problem involves reading comprehension and mapping the problem description onto a combo of RVs. When you study how humans learn, one idea you come across is Cognitive Load Theory (CLT). This suggests each person has a bandwidth for how much cognitive work they can process at a given moment (like the bandwidth for sending info over the internet). When you overload this, the act of understanding or problem-solving shuts down. In this exercise, my goal was for the majority of the cognitive load to be wrapped up in re-phrasing the prompt into RVs using WISE. If, however, you struggle with integration or have yet to internalize the Negative Binomial distribution, these might add to the cognitive load and overwhelm your bandwidth. If you struggled with this problem, ask yourself why. Some possible answers:

- "My reading skills are weak, so I had trouble understanding what the problem was saying."
- "I don't really remember how to integrate, so I was afraid to do/couldn't do the integration step."
- "I haven't learned the Negative Binomial well, so it didn't come to mind as a modelling choice."

Whatever the reason, CLT suggests your bandwidth was overloaded. The best way to avoid this is to ensure deep understanding of prerequisite material when tackling more advanced topics.

43. For each part below, identify the population of interest, the parameter I wanted to study, the sample I drew, and the statistic I calculated.

(a) I recently returned to swimming after some time away from the sport. I was curious what my new average time was for swimming a distance of 50 yards, so I recorded 13 random times with the help of the clock that swimmers use to keep pace. My 13 times averaged 45.31 seconds.

(b) I recently wrote code in R that sorted 200 different random lists of numbers, each with 100000 numbers. The minimum sort time in my data set was 0.07 seconds. I was curious what the fastest time possible is when R uses the quicksort function to sort a list of 100000 random numbers on my computer.

Hint List:

- a,b: Remember that a parameter always summarizes a population while a statistic summarizes a sample.
- The population should feel like the sample, just infinite or very large; it is the extension of the idea seen in the sample to all possible cases/values.

Solution:

(a) Population: The times for all 50-yard swims I might do after returning to swimming
Parameter: The average of the population times, μ
Sample: The times of the 13 random 50-yard swims
Statistic: The average of the sample times, $\overline{x} = 45.31$ seconds

(b) Population: All possible times when sorting a list of 100000 numbers in R using quicksort
Parameter: The minimum of the population
Sample: The 200 quicksort times for the 200 lists (each of 100000 numbers)
Statistic: The minimum of the sample, 0.07 seconds

Commentary: There are two things to note here. First, when describing the population (or sample), some people will list the *things you are observing* rather than the numbers you draw from those (e.g., the 50-yard swims rather than the times for those swims). This is fine, and I often do this. One reason for this approach is that there is something dehumanizing about gathering a sample of people and seeing them only as their weights, or how they will vote, etc.

Second, I included part b because you do not know the notational standard for the minimum in a population and in a sample as you do for the mean in part a. I was hoping this might encourage you to ask "How does the field typically denote these?". Part of learning a field is asking how the subject structures and notates various ideas. Often, these notations are the result of centuries of refinement (e.g., the way we present calculus today looks quite different than initial efforts). As such, notation tends to be concise and clear, once you understand it. For example, the notation $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ is a masterclass in reducing a very complicated expression into just a few symbols. Embracing the notation of a field is a nice way to better know and use a subject, and it provides a stepping off point for taking the discipline in new notational directions (e.g., see the pi vs. tau debate).

44. I wrote this question the day before Thanksgiving! Have you ever been curious about the average number of calories a person consumes at their Thanksgiving meal? To figure this out, you might suppose Thanksgiving calorie-intakes are normally distributed with mean μ and standard deviation 400 (a value determined by studying other events with big meals).

    (a) One approach to thinking about μ is to record the calorie intake for a single random person X and then to claim μ is in the interval $[X - 200, X + 200]$. What is the probability your claim is correct? [WISE, Tables not tech]

    (b) Another approach is to collect data on 16 random people, calculate the average $\overline{X}$, and then give the confidence interval $[\overline{X} - 100, \overline{X} + 100]$. What is the confidence level of this interval? [Memorized areas not tables or tech]

    (c) How is it that your answer to part b is larger than part a when the interval in part b is half as wide as that in part a? Shouldn't larger intervals have a better chance of containing a parameter of interest?!

    (d) A third approach is to collect calorie counts for 7 random people and take the median, M, of these values. Suppose we report the interval $[M - 10, \infty)$ for μ. Using R, write code that finds the confidence level for this confidence procedure using 100000 simulated intervals of this form. (Do not try to find the confidence level algebraically. Assume that the true value of μ is 1500 in your code.)

Hint List:

- a: The confidence level of any confidence procedure is defined as

$$P(\text{parameter of interest} \in \text{CI expression})$$

  Here, this gives us $P(\mu \in [X - 200, X + 200])$ as a starting point.
- b: Use a previous homework problem to find the distribution of $\overline{X}$.
- c: If both the intervals were based on X, then the wider interval would have the higher confidence level. Here, one interval is based on X and the other on $\overline{X}$. What do you notice about the SDs of these RVs?
- d: R has a built-in median function to help. The replicate function is also useful.

Solution:

(a) (WIS) Let $X \sim N(\mu, 400)$ be the Thanksgiving calorie count of a random person. We find (E):

$$P(\mu \in [X - 200, X + 200]) = P(X - 200 \leq \mu \leq X + 200) = P(\mu - 200 \leq X \leq \mu + 200)$$

Probabilities statements are most useful when the random variable appears in the *middle*, as in the last step. Since the problem demands tables, we standardize:

$$P(\mu - 200 \leq X \leq \mu + 200) = P\left(\frac{\mu - 200 - \mu}{400} \leq \frac{X - \mu}{400} \leq \frac{\mu + 200 - \mu}{400}\right) = P\left(-\frac{1}{2} \leq Z \leq \frac{1}{2}\right)$$

From a Z-table, we find this area is about $0.6915 - 0.3085 \approx \boxed{0.383}$.

(b) Proceeding as before, we find:

$$P(\mu \in [\overline{X} - 100, \overline{X} + 100]) = P(\overline{X} - 100 \leq \mu \leq \overline{X} + 100) = P(\mu - 100 \leq \overline{X} \leq \mu + 100)$$

Recall from a previous homework problem that if $X \sim N(\mu, \sigma)$, then $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Here, we find $\overline{X} \sim N\left(\mu, \frac{400}{\sqrt{16}}\right) = N(\mu, 100)$. So, the above probability is just the area within one standard deviation of the center of $\overline{X}$. We've memorized this area as about $\boxed{0.68}$.

(c) While the interval in part b is half the size of that in part a, it has a higher confidence level because it is based on more data (16 data points instead of 1). $\overline{X}$ has a better chance of being near μ than X does (compare the SDs!), so even with less padding (± 100), $\overline{X}$ outperforms X with more padding (± 200).

(d) I used the below code to get a confidence level of about $\boxed{0.521}$.

```
medians = replicate(100000, median(rnorm(7, 1500, 400)))
lbs = medians - 10    # the lower bounds of the intervals
capture = lbs < 1500  # is mu = 1500 in each interval?
mean(capture)
```

Commentary: One of the best ways to understand confidence intervals is to see a variety of different examples. In time, we'll use a small collection of common CI formulas (e.g., you may have seen expressions like $\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}\cdot\widehat{q}}{n}}$ in previous courses). The idea of a CI, however, is quite general: it is just a rule that spits out intervals (based on data) that try to contain some parameter of interest. Such rules might be quite good at their job, or quite bad; the confidence level tries to quantify this. In this problem, I proposed three different CPs, and we got CLs of 0.383, 0.68, and 0.521. While it is tempting to prefer the CP with the highest CL, this may not be the "best" or most useful. For example, the CP that ignores the data and always gives the interval $(-\infty, \infty)$ will have a CL of 1; it always captures the parameter μ. But ask yourself, how useful is the knowledge that μ is somewhere between $-\infty$ and ∞?!

At first blush, the notion of a CI may seem quite exotic. In reality, though, we use CIs all the time. For example, when heading to a friend's house, you might text "I'll be there between 5 and 5:15". Here, you're trying to capture t, the exact time of your arrival using the interval [5 PM, 5:15 PM]. You likely took your best guess (perhaps provided by GPS) and padded it to the nearest times one can say without seeming socially awkward. As another example, your toothpaste might advertise "Whitens teeth up to 33% brighter!". Here, the company is trying to capture the average whitening improvement, μ, seen by product users. To put as large a number as possible on the tube, they report the interval $[0\%, \max(X_1, \ldots, X_n)\%]$ where $X_1, \ldots, X_n$ are the whitening improvements in a sample of n subjects. The confidence level also has a meaning in these real-world settings. For the first, it measures the fraction of ETA windows that are correct. We all know people who provide reliable time estimates and those who don't. For the second example, the CL is a measure of how reliable a company's advertising is. Look for other real-world CIs with your new statistical glasses!

45. Visualizing the sampling distribution! Typically, when doing a study, you only collect one sample and use this to get one value for a statistic. It is useful, however, to imagine more samples being collected, and from these, more values we might have gotten for the statistic. Let's explore this.

(a) You're going to try to figure out the average download rate you get over your internet connection. Go to Google and type in "speed test". The first search result will let you run a speed test on your computer to measure your download speed. Do this three times and record your *download* speeds in Megabits per second (Mbps). I got 259.1, 249.7, and 222.6. Now average these to get one value for the $\overline{X}$ statistic. State the population of interest, the parameter, the sample, and the statistic for this setup (using your data).

(b) If you did the steps in part a on a different day, you'd get 3 new rates and a new value for $\overline{X}$. Write a single line of R code that simulates finding a new value for $\overline{X}$. Just so we all have the same setup, assume your download rates come from the distribution $N(\mu = 250, \sigma = 20)$.

(c) Write R code that simulates 10000 averages (based on samples each of size 3) from the distribution in part b. Then make a histogram of these values to see an approximation of the sampling distribution. Finally, state the exact shape, center, and spread of the sampling distribution if we could create infinitely-many versions of the statistic. (Include a screenshot or hand sketch of the histogram.)

Hint List:

- c: Recall a previous homework problem to find the exact distribution of $\overline{X}$. Also, there are many ways to see if the sampling distribution matches the theoretical result. If you want to go down the route the solution will present, search how to create overlapping histograms in R.

Solution:

(a) Answers will vary. My times give an average of $\frac{259.1 + 249.7 + 222.6}{3} \approx 243.8$ Mbps.

Population: The download rates (in Mbps) for all speed tests you could run
Parameter: μ, the average download rate of all speed tests
Sample: The download rates for the 3 tests you ran
Statistic: $\overline{X} = 243.8$ Mbps (for my data)

(b) This takes only one line of code:

```
mean(rnorm(3, 250, 20))
```

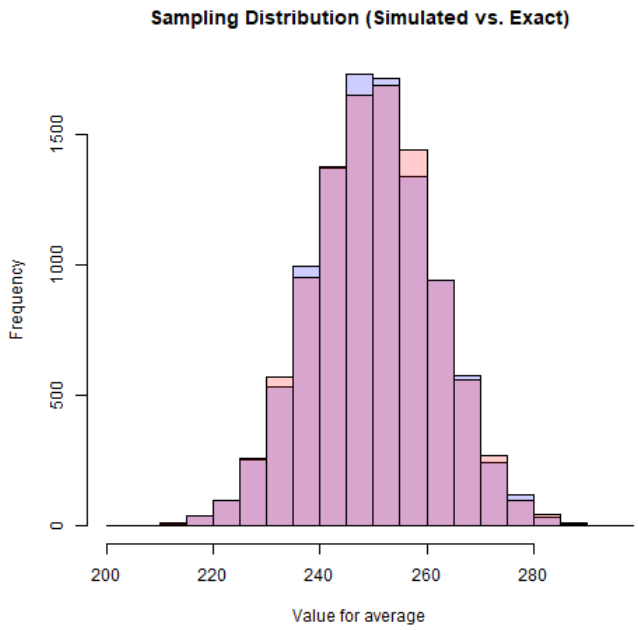
(c) Here, $X \sim N(250, 20)$, so $\overline{X} \sim N\left(250, \frac{20}{\sqrt{3}}\right)$ by a previous homework problem. If I make a lot of values for $\overline{X}$ using code as in part b, will these values follow this distribution? Let's find out. Below I create overlapping histograms. The colors are set up as blue and red with some see-through-ness (the alpha value). Then, I simulate 10000 averages as in part b and pull 10000 values from the true distribution for $\overline{X}$. Finally, I plot both to see if they look similar.

```
color1 = rgb(0, 0, 255, max = 255, alpha = 50)
color2 = rgb(255, 0, 0, max = 255, alpha = 50)

xbars = replicate(10000, mean(rnorm(3, 250, 20)))
simulatedhist = hist(xbars, plot = FALSE)

actualhist = hist(rnorm(10000, 250, 20/sqrt(3)), plot = FALSE)

plot(simulatedhist, xlab = "Value for average", col = color1,
     main = "Sampling Distribution (Simulated vs. Exact)")
plot(actualhist, col = color2, add = TRUE)
```



Commentary: People often think that when you leave college, you'll have all the skills you need to do the job you're hired for. Nothing could be further from the truth, especially in the quantitative and tech spaces where things change rapidly. When I left undergrad, people used basic calculators, animated tools like Desmos and Manim did not exist, and people programmed mostly in C++ and Java (rather than R, Python, and the many other languages now popular). Obviously, the current landscape is dramatically different, and *my usefulness is a reflection of my willingness to adapt to the changing landscape*. Part of this willingness is having the initiative to look things up when I don't know how to do them. Part c of this problem offers the chance to look things up and go beyond what you currently know how to do in R. Along the way, you might learn how to save histograms for future plotting or how computers structure colors (as a mixture of red, blue, and green, and a transparency value).

See your undergraduate years as an opportunity to transition from the broken middle-school paradigm to the reality you'll face post-graduation:

- Broken middle-school paradigm: "I, the teacher, will train you, the student, in a very particular skill. You will faithfully reproduce it, and I promise to never ask you to go outside this skill!"
- Post-graduation reality: "The skills, tools, and knowledge you'll need will change constantly. You may have to navigate this change without much help. If you can't adapt, you'll be left behind."

46. If you study heat conduction in physics, you are likely to encounter the CRV Y with pdf

$$f(y) = \begin{cases} \sqrt{\frac{2}{\pi}} e^{-y^2/2}, & y > 0 \\ 0, & \text{else} \end{cases}$$

(a) Find $E(Y)$ by evaluating an integral by hand. As always, use beautiful notation to deal with the improperness of the integral. It's time to get serious about presenting your work.

(b) Suppose your friend wasn't able to do part a and suggested, instead, to use a single value, Y_1, from the distribution to build a confidence interval $[Y_1, 2 \cdot Y_1]$ that would hopefully contain the answer from part a. What is the confidence level of this confidence procedure? (You will need to use your answer from part a.) [Tables not tech]

Hint List:

- a: One common mistake when writing the expression for $E[Y]$ is to forget that the integrand is $y \cdot (\text{pdf of } Y)$. That y is very important in this problem because the integral is much easier when you remember it. Note that if you do forget the y, then you're simply integrating the pdf over the support, so you will get the answer 1. This is a nice thing to remember: If you ever find $E[\text{RV}] = 1$, do a quick check to see if you forgot the y in the integrand!
- b: About halfway through the problem, you will be faced with a difficult integral. If you write down the pdf of $Z \sim N(0,1)$, you'll notice your integral is very similar to the pdf of Z, and this observation offers a way forward.

Solution:

(a) We use the definition of expectation and a limit for the improper integral:

$$E[Y] = \int_0^\infty y \cdot \sqrt{\frac{2}{\pi}} e^{-y^2/2}\, dy = \lim_{t\to\infty} \sqrt{\frac{2}{\pi}} \int_0^t y e^{-y^2/2}\, dy = \lim_{t\to\infty} \sqrt{\frac{2}{\pi}} \left(-e^{-y^2/2}\right)\Big|_0^t$$
$$= \sqrt{\frac{2}{\pi}} \lim_{t\to\infty} \left(-e^{-t^2/2} + 1\right) = \sqrt{\frac{2}{\pi}} \cdot (0+1) = \boxed{\sqrt{\frac{2}{\pi}} \approx 0.798}.$$

(b) We must find $P\left(Y_1 \le \sqrt{\frac{2}{\pi}} \le 2Y_1\right)$. Pivoting the RV to the center gives:

$$P\left(\frac{1}{\sqrt{2\pi}} \le Y_1 \le \sqrt{\frac{2}{\pi}}\right) = \int_{1/\sqrt{2\pi}}^{\sqrt{2/\pi}} \sqrt{\frac{2}{\pi}} e^{-y^2/2}\, dy = \sqrt{\frac{2}{\pi}} \cdot \sqrt{2\pi} \int_{1/\sqrt{2\pi}}^{\sqrt{2/\pi}} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}}_{\text{pdf of } Z}\, dy$$
$$\approx 2 \int_{0.40}^{0.80} f_Z(y)\, dy \overset{\text{tables}}{\approx} 2(0.7881 - 0.6554) \approx \boxed{0.265}.$$

Commentary: Mathematics can be a *sensitive* subject. Notice that the integral in part a is straightforward. You can either use u-substitution ($u = y^2/2$) or skip this and do it in your head. In contrast, when the y is not present in the integrand of part b, the integral is suddenly a nightmare to deal with. Indeed, we must cleverly massage the integrand to the pdf of Z and get rescued by tables. There are so many examples like this in math: Make a slight change and you go from something pleasant to something very difficult. When I teach integral calculus, I show the sensitivity of integration through a collection of "small change, big difference" problems. Try some if you're looking for a challenge; the first integral in each pair is much easier to do!

$$\int x^2 e^{x^3}\, dx \qquad \text{vs.} \qquad \int x^3 e^{x^2}\, dx$$
$$\int \arcsin x\, dx \qquad \text{vs.} \qquad \int (\arcsin x)^2\, dx$$

47. Take a look at this classic commercial. For many years, the phrase "How many licks ..." was quite popular (today it would be a meme). Suppose you wanted to find out by having 30 random people actually count their licks $(L_1, \ldots, L_{30})$ and calculating the average of these counts, $\overline{L}$.

(a) Describe the population, parameter, sample, and statistic for this experiment.

(b) Explain why a reasonable model for these lick counts is $L \sim Pois(\lambda)$. Also, explain why $\overline{L}$ is a good estimate for the value λ.

(c) Algebraically, find the *approximate* shape, center, and spread of $\overline{L}$.

(d) Using R, run a simulation that finds 10000 values of $\overline{L}$, plots them in a histogram, and checks your answer to part c. Assume $\lambda = 120$ licks in this part and in part e. [No histogram sketch needed]

(e) Suppose you plan to use the interval $[\overline{L} - 6, \overline{L} + 6]$ as a confidence interval for λ. What is the approximate confidence level of this confidence procedure? [Memorized areas not tables or tech]

Hint List:

- c: Recall the Central Limit Theorem.
- e: Try to pivot the RV to the center of the inequality chain.

Solution:

(a) Population: The licks needed to reach the center of a Tootsie pop for all people trying
Parameter: μ, the average of all possible lick counts
Sample: The lick counts for our 30 people $(L_1, \ldots, L_{30})$
Statistic: $\overline{L}$, the average of our 30 lick counts

(b) Lick counts are a discrete idea and $\frac{\text{licks to center}}{\text{Tootsie pop}}$ feels like a rate. So, Poisson seems like a good choice. Also, if $L \sim Pois(\lambda)$, then we know $E[L] = \lambda$. Since $\overline{L} \approx E[L]$, we have $\overline{L} \approx \lambda$.

(c) By the CLT, $\overline{L} = \dfrac{L_1 + \cdots + L_{30}}{30}$ is approximately normal.

$$E[\overline{L}] = \frac{1}{30}\left(E[L_1] + \cdots + E[L_{30}]\right) = \frac{30 \cdot E[L]}{30} = \lambda$$

$$Var[\overline{L}] = \frac{1}{30^2} Var[L_1 + \cdots + L_{30}] \overset{\text{ind.}}{=} \frac{1}{900}\left(Var[L_1] + \cdots + Var[L_{30}]\right) = \frac{30 \cdot Var[L]}{900} = \frac{\lambda}{30}$$

Thus, $\overline{L} \approx N\left(\lambda, \sqrt{\dfrac{\lambda}{30}}\right)$.

(d) I used the below code:

```
avgs = replicate(10000, mean(rpois(30, lambda = 120)))
hist(avgs)  # hoping for approximate normality
mean(avgs)  # hoping for 120
var(avgs)   # hoping for 4
```

(e) We want

$$\begin{aligned} P(\overline{L} - 6 \leq 120 \leq \overline{L} + 6) &= P(120 - 6 \leq \overline{L} \leq 120 + 6) \\ &= P\left(\frac{120 - 6 - 120}{\sqrt{120/30}} \leq \frac{\overline{L} - 120}{\sqrt{120/30}} \leq \frac{120 + 6 - 120}{\sqrt{120/30}}\right) \\ &\approx P\left(\frac{-6}{2} \leq Z \leq \frac{6}{2}\right) = P(-3 \leq Z \leq 3) \overset{\text{memory}}{\approx} \boxed{0.997} \end{aligned}$$

Note: If you notice that $6 = 3 \cdot 2 = 3 \cdot SD(\overline{L})$, you can skip the standardization steps and immediately see the answer is 0.997 since we are asking for the area within three SDs of the mean on a normal distribution (here, $\overline{L}$).

Commentary: As an educator, my favorite problems are those that ask students to bring together *many* skills. This problem requires you to think about and model the world (parts a and b), recall the

CLT and use expectation/variance rules (part c), run an R simulation (part d), and do a confidence level calculation (part e).

Why are these types of problems so useful? First, they have the ability to reveal weaknesses, misconceptions, or gaps in understanding across large swaths of a subject. In essence, they demand *intellectual breadth*. If you struggled with any part of this problem, you now have a sense of what you can work on. We see the value of breadth challenges in other areas as well. The individual medley in swimming requires you be good at backstroke, butterfly, breast stroke, and freestyle. A concerto in music will demand proficiency in many techniques (not just one or two, like an etude). Being a good parent or partner requires a long list of disparate skills (e.g., patience, kindness, etc.).

Second, having strong intellectual breadth is critically important in the modern workplace. As an example, consider what goes into creating an app like Uber. Beyond writing code in some programming language, you must understand how to adapt your interface for different sizes of devices (tablets vs. phones) and different types of devices (Apple vs. Samsung). You also need to understand geolocation tools (to connect riders and drivers), mapping technology (to secure routes), finance systems (for payments), and language and localization (to go international). Then, you need to deal with laws (Does city X allow ridesharing?), customs (Should I include a tipping interface in country Y?), marketing, etc. Naturally, you would split these tasks across multiple employees at large companies, but the stunning demand for breadth is unavoidable in complex modern endeavors. Said simply, the greater breadth you bring to these problems, the greater your relevance and probability of success.

As a personal example, on my first day of college, I was pursuing a triple major: math, music education, and electrical engineering. Within a year, this had become a double major: math and violin performance. By the end of junior year, I was taking the MCAT to head to medical school, adding an economics minor, and thinking about computer science and chemistry as possible directions. In my senior year, I confidently settled on a future in education. People often ask me if this non-linear route was a waste of time. Surprisingly and luckily, it was one of the best things I ever did for my education. By pursuing so many different academic directions, I developed intellectual breadth without realizing it. This diversity of academic experience affects my current life in innumerable ways. For example, I can better advise students on the pros/cons of medical school. I can also write better statistics problems that genuinely pull from a wealth of domains. When one of our Ph.D. programs needed a new website, I could write it from scratch and save the school tons of money. Said simply, you never know what life is going to demand of you. So, learn as much as you can, from as many different subjects as you can.

48. In the below problems find the PH of the PHANTOM mnemonic. Use the notational conventions for parameters established in class.

    (a) You want to poll Americans and ask if they agree with the statement "Donald Trump actually won the 2020 election". You're curious whether self-declared Democrats and self-declared Republicans will agree with this statement in equal percentages or not.

    (b) You've designed a new economic reform that is supposed to reduce income inequality. You collect data on 1000 people after the reform and explore the standard deviation of incomes compared to a historical standard deviation value of $32000 before the reform.

    (c) Your friend is using a Poisson model to explore how often NFL quarterbacks get sacked in the NFC vs. AFC. As a fan of NFC teams, you hope the average rate is higher in the AFC than the NFC, but your friend is convinced the rates will be the same.

Hint List:

- Try to identify a very large population(s) that is being studied and a numerical summary of that population(s). These are the population(s) and parameter(s). Sometimes a sample might be described; this is only useful in that it is a small piece of your population. If you are given a statistic, this is only useful in that it indicates what the numeric summary is. The sample and statistic (if given), should be lifted up to a larger setting to create the population and parameter.

Solution:

(a) (P) Let p_R (p_D) be the percentage of **all** self-declared Republicans (Democrats) who agree that "Donald Trump actually won the 2020 election".

(H) We test H_0: $p_R = p_D$ vs. H_1: $p_R \neq p_D$ ("or not").

(b) (P) Let σ be the standard deviation of **all** incomes in the population you are studying.

(H) We test H_0: $\sigma = \$32000$ vs. H_1: $\sigma < \$32000$ ("reduce income inequality")

(c) (P) Let λ_{NFC} (λ_{AFC}) be the average rate at which NFC (AFC) quarterbacks get sacked across **all** games.

(H) We test H_0: $\lambda_{\text{NFC}} = \lambda_{\text{AFC}}$ vs. H_1: $\lambda_{\text{NFC}} < \lambda_{\text{AFC}}$ ("hope ... higher in the AFC")

Commentary: There are several challenges when finding the PH of PHANTOM from a block of text. First, you must avoid the temptation to encode what someone wants as H_0. In part c, for example, you "hope the average rate is higher" (note: this signals an inequality). Just because you hope for this to happen does not mean you encode the statement as H_0. H_0 should always involve an equals sign, and H_1 should always have an inequality (the "always" will be relaxed in future coursework). What you're hoping for might be H_0 or H_1 depending on whether you hope for an equality or inequality in your setting.

Second, a prompt might have extraneous or missing information. For example, part b mentions the number 1000. Since this is about the sample, it is irrelevant when finding the PH. If a problem wanted you to compare the average radius of planets in the Milky Way galaxy to the radius of the Earth, you might have to go look up that $r_{\text{Earth}} \approx 3960$ miles if it were missing.

Finally, written language is context dependent. Suppose we want to "reduce a disparity" for group G in some setting. If the group is over-represented in that setting, then we hope to reduce the percentage of people from group G in that setting. So, H_1: $p_G <$ current percentage. If, however, the group is under-represented, then reducing the disparity means that the percentage should increase instead. So, H_1: $p_G >$ current percentage. So, you need to look beyond the surface-level of the words in the problem and interpret them in context.

(Note: When I write an alternative hypothesis, you'll see me quoting the words that indicated the nature of the inequality sign. You need not do this; I include it to help students transition between written language and mathematical notation.)

49. How do you decide when someone has gotten better at something (e.g., how long it takes to swim a lap, how strong someone is in the gym, etc.) when there is natural variability in their performance? Suppose your friend's one-rep max for bench press (the most weight that can be pushed in a single trial) is normally distributed with unknown mean and standard deviation of 6 pounds. Last year, you know the mean was 162 pounds, but your friend claims to be stronger now.

(a) Complete the PH of PHANTOM.

(b) Your friend proposes a test to decide between the hypotheses: They will perform a one-rep max and you'll use this to determine which hypothesis to support moving forward. If your friend bench-presses 180 pounds, what is the probability of seeing this event or something more extreme assuming the null hypothesis is true? [WISE, VIZ, Tables not tech]

(c) What conclusion should you draw for part b if a strange event in your mind is something that happens only 1 in a 100 times (or less often)?

(d) Another approach to testing your friend's claim is to have them perform 4 one-rep maxes (on different days) and use the **average** of the values to decide which hypothesis to believe. If your friend's results are 180, 160, 170, and 164, find the average. Then, assuming the null hypothesis is true, find the distribution (shape, center, spread) the average follows, and calculate the probability of getting the observed average or something more extreme. [Tech not tables]

(e) What conclusion should you draw for part d if a strange event in your mind is something that happens only 1 in a 100 times (or less often)?

Hint List:

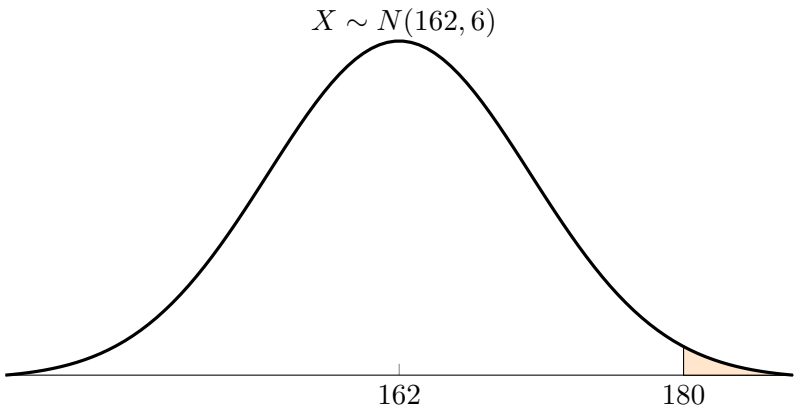
- It helps to note that parts b and c are separate from parts d and e.

Solution:

(a) (P) Let μ be the current average for our friend's one-rep max on bench press.

(H) We test H_0: $\mu = 162$ vs. H_1: $\mu > 162$ ("stronger now")

(b) Assuming H_0 is true, we set (WIS) $X \sim N(\mu = 162, 6)$, where X is the result of a random one-rep max bench press attempt at the gym.



We must find (E):

$$P(X \geq 180) = P\left(\frac{X - 162}{6} \geq \frac{180 - 162}{6}\right) = P(Z \geq 3) \overset{\text{tables}}{\approx} 1 - 0.9987 = \boxed{0.0013}.$$

(c) Based on part b, we either saw a 1 in 769 (1/0.0013) event if H_0 is true, or H_1: $\mu > 162$ is responsible for our large value of X. Since our event was rarer than we can tolerate, we reject H_0 in favor of H_1. We can process this more simply by noting that $0.0013 < 0.01 = \frac{1}{100}$. It is possible that our friend is getting stronger.

(d) We have $\frac{180 + 160 + 170 + 164}{4} = 168.5$. If H_0 is true, then as before, $X \sim N(162, 6)$, which forces $\overline{X} \sim N\left(162, \frac{6}{\sqrt{4}}\right) = N(162, 3)$. So, is our average of 168.5 strange to see from $\overline{X} \sim N(162, 3)$? Using R, we have:

$$P(\overline{X} \geq 168.5) = \text{pnorm}(168.5,\ 162,\ 3,\ \text{lower} = \text{F}) \approx \boxed{0.015}.$$

(e) Since $0.015 > 0.01$, we keep H_0. While getting 168.5 (or something greater) had a fairly low probability, it doesn't quite meet our cutoff as a rare event. We keep H_0 as it is still a reasonable belief system.

Commentary: Hypothesis testing (HT) can feel very strange. I remember the day I first saw it (a genetics class in college) and my response: "Why is this so awkward? That's not how people think!" In time, I've come to see that HT is actually a nice distillation of how we often approach life: Adopt a belief system (i.e., assume H_0), collect some data (i.e., $X = 180$ in part b or $\overline{X} = 168.5$ in part d), and then decide if our belief system makes sense given the data coming in (i.e., part c or e). In this problem, we go through the HT process twice, once on X and once on $\overline{X}$. I wanted to show you that there can be multiple ways to gauge how you feel about a belief system. The same is true in real life. If you adopt the belief that your partner is faithful to you, then the data you collect to assess this might be: 1) their words and actions over the next 3 months, 2) a survey of their phone messages last week, 3) a lie detector test, etc. Whatever your data source, you must measure the data you get against what that data source could have provided. In this problem, we compared the value 180 against $X \sim N(162, 6)$ in part b, but we compared 168.5 against $\overline{X} \sim N(162, 3)$ in part d (because 168.5 was itself an average).

The second takeaway from this problem is that we might get conflicting messaging when running more than one HT. Notice that the answers to parts c and e disagree (reject H_0 vs. keep H_0). This might not be earth-shattering because these parts use different data sets (180 vs. $\{180, 160, 170, 164\}$). Indeed, you could argue that the first attempt of 180 was artificially high and that the future attempts (160, 170, 164) revealed the "true" strength of our friend. Maybe so. What's more shocking is that even with *the same data set*, it is possible to run two different HTs and get disagreement about whether to believe in H_0 or H_1! This can be upsetting when you first hear it. After all, isn't statistics a quantitative subject, and don't quantitative subjects help us avoid the ambiguity in non-quant subjects? When an English teacher asks you to write an essay on "Is character X in book Y a hero?", you expect that you and your classmates might disagree despite reading *the same book*. Likewise, you expect ambiguity when two people experience the same poetry, cooking, or sporting event. The truth is that statistics is somewhere between the non-quant subjects and the cleanness of theoretical mathematics. You've actually already come across this reality earlier in the course. Suppose I give everyone the same data set of prices for recent San Diego home sales and ask "What did the typical person recently pay for a home in SD?". Since there are many choices for a measure of center (mean, median, mode, trimmed mean, etc.), you can get conflicting answers. Because of this reality, some believe "You can say whatever you want with statistics!". A more accurate and nuanced conclusion is: "You can say different things with statistics. As such, it is important to provide context about the choices you made and why you believe those choices are prudent." This statement won't go viral online like the first, which is perhaps an indication of its veracity.

50. A casino claims to be using fair, six-sided dice (labeled $1 - 6$). One way to check this claim is by watching gamblers play the game of Craps. In this game, a single person rolls a pair of dice over and over until they "crap out", a term that means the pair of dice on a given roll totaled either 2, 3, or 12. Suppose you walk up to a table and begin watching people roll the pair of dice. If the casino's claim is true:

    (a) How many rolls, on average, should it take to crap out? (A "roll" refers to throwing both dice at once, which is a requirement of the game.) [WISE]

    (b) What is the probability it takes 20 or more rolls to crap out? (Give an exact answer as a fraction expression, no technology is needed.)

    (c) Suppose a rare event in your mind is something that happens 1 out of 50 times (or less). What is the smallest value I could change 20 to in part b to make a rare event? (Only a simple calculator is needed to do this.)

Hint List:

- a: There are 4 ways to crap out.
- b: Evaluate an infinite sum, not a finite sum.
- c: Repeat the style of calculation in part b but use the constant c. Also, be careful of dividing by a negative number at one point.

Solution:

(a) Let (WIS) X be the number of rolls to crap out with the fair dice. Note that $X \sim Geom(p)$ where $p = P(2, 3, \text{ or } 12 \text{ showing})$. So, $p = P((1,1) \text{ or } (1,2) \text{ or } (2,1) \text{ or } (6,6)) = \frac{4}{36} = \frac{1}{9}$.

We want (E) $E[X] = \frac{1}{p} = \boxed{9 \text{ rolls}}$.

(b) We want:

$$P(X \geq 20) = P(X = 20 \text{ or } X = 21 \text{ or } \ldots) = \left(\frac{8}{9}\right)^{19} \cdot \frac{1}{9} + \left(\frac{8}{9}\right)^{20} \cdot \frac{1}{9} + \cdots$$

$$= \frac{\left(\frac{8}{9}\right)^{19} \cdot \frac{1}{9}}{1 - \frac{8}{9}} = \boxed{\left(\frac{8}{9}\right)^{19}}.$$

(c) Here, we need $P(X \geq c) < \frac{1}{50}$ to make a rare event. Thus:

$$\frac{1}{50} > P(X \geq c) = \left(\frac{8}{9}\right)^{c-1} \cdot \frac{1}{9} + \left(\frac{8}{9}\right)^{c} \cdot \frac{1}{9} + \cdots = \frac{\left(\frac{8}{9}\right)^{c-1} \cdot \frac{1}{9}}{1 - \frac{8}{9}} = \left(\frac{8}{9}\right)^{c-1}$$

Taking the natural log of the ends of this inequality gives $\ln \frac{1}{50} > (c-1) \ln \frac{8}{9}$.

Since $\ln \frac{8}{9} < 0$, dividing reverses the inequality: $\frac{\ln 1/50}{\ln 8/9} < c - 1$. Thus, $34.2 < c \implies \boxed{c = 35}$.

Commentary: Here, I wanted to mix together many ideas from the course and your past: finding probabilities by listing outcomes in the event, the Geometric RV, expectation, infinite series, determining if an event is rare, solving inequalities, and log rules. For me, the most interesting element is part c. Typically, you work from known data and calculate the strangeness of that data, and then compare this to a known bound (e.g., see the last problem). Here, you have to work backwards, taking the strangeness cutoff and back-deriving what data could trigger it. In life, procedures tend to mostly operate in one direction. I believe we can better understand our procedures by using them in reverse as well.

51. While R has plenty of distributions built in (e.g. exp, norm, pois), it can't do everything. Suppose you need to use the *discrete* random variable with probability model

$$P(X = k) = \frac{-1}{\ln(1-p)} \cdot \frac{p^k}{k} \text{ where } k = 1, 2, 3, \ldots \text{ and } 0 < p < 1$$

This is known as the logarithmic series distribution, and we'll write $X \sim LogSeries(p)$.

(a) Show that the probabilities for this distribution add to 1, as they must.

(b) Let's get to visually know $LogSeries(p)$. Using R's vectorized abilities, make a vector for the integers from 1 to 10, and then find the heights of the probability function at these values when $p = \frac{1}{2}$. Plot points in red using the plot(...) command to get a picture of the probability model on $\{1, 2, \ldots, 10\}$. Now, re-find the heights of the probability function when $p = \frac{9}{10}$ and add blue points to your existing plot using the points(...) command. Include your code and a screenshot/sketch of the plot. (Note: The goal of this problem is NOT to dig through R documentation or online libraries to find an implementation of the logarithmic series distribution. Instead, your goal is to write simple mathematical formulas in R based on the given probability formula that find any probabilities you need.)

(c) Suppose your friend has correctly written R code that allows for use of this distribution in the usual way (d-p-q-r quartet!). Your friend proposes a challenge: To decide between H_0: $p = \frac{1}{2}$ and H_1: $p = \frac{9}{10}$ using a single value, X, generated from the distribution. Using your picture from part b, explain why getting a large value of X would support H_1 instead of H_0.

(d) Using R, find the probability of getting $X = 8$, or something more extreme, assuming the null hypothesis is true. Using this probability, make a decision about which hypothesis to trust assuming that results rarer than a probability of 0.05 cause you to switch hypotheses.

Hint List:

- a: What functions do you know the Taylor series for?
- b, d: What do you get when running this code?

```
xs = 1:3
vals = 2^xs/xs
```

Solution:

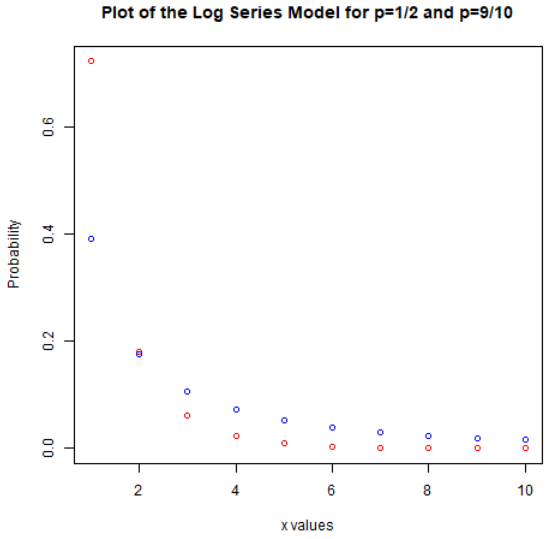
(a) We must find $\sum_{k=1}^{\infty} \frac{-1}{\ln(1-p)} \cdot \frac{p^k}{k}$ where $0 < p < 1$. When you study Taylor series, you learn that we can write $\ln(1-x)$ as $-\sum_{n=1}^{\infty} \frac{x^n}{n}$ for $x \in [-1, 1)$. We rearrange our sum and then use this fact:

$$\frac{1}{\ln(1-p)} \cdot -\sum_{k=1}^{\infty} \frac{p^k}{k} \overset{\text{Taylor}}{=} \frac{1}{\ln(1-p)} \cdot \ln(1-p) = 1.$$

(b) Using the vectorized nature of R, we can do this quite easily:

```
xvals = 1:10
p = 1/2
yvals = -1/log(1-p) * (p^xvals)/xvals
plot(xvals, yvals, col = "red",
     xlab = "x values", ylab = "Probability",
     main = "Plot of the Log Series Model for p=1/2 and p=9/10")

p = 9/10
yvals = -1/log(1-p) * (p^xvals)/xvals
points(xvals, yvals, col = "blue")
```



(c) Based on part b, it seems that $p > \frac{1}{2}$ (e.g., $p = 9/10$) tends to shift the probability from small values of X (like 1 and 2) to larger values of X. We see this because the blue dots ($p = 9/10$) are above the red dots ($p = 1/2$) for larger values of X in the picture.

(d) If $X \sim LogSeries(p = 1/2)$, we want $P(X \geq 8) = 1 - P(X \leq 7)$. In R, we don't have an existing implementation of the log series (although there exist packages with this that one can download), so we do the work ourselves:

```
xvals = 1:7
p = 1/2
yvals = -1/log(1-p) * (p^xvals)/xvals
print(1 - sum(yvals))
```

We get the value 0.00128 from this code, which is less than 0.05, so we reject H_0 in favor of H_1.

Commentary: When you were writing code for parts b and d, did you have a nagging feeling that you kept writing the same code over and over? This frustration is relieved when you learn how to write your own function in R. Functions are sections of code that you find yourself using frequently. The idea is to hide that repeated code behind a name (the function name) and to pass it any info (the parameters) that might change when you use the function. You've actually been working with functions already: mean, sd, print, plot, hist, and log are all functions. Hopefully this problem made you feel like you should write *your own* function. Inspired by the d-p-q-r quartet for other distributions, we might write:

```
dlogseries = function(xvals, p)
{
   return(-1/log(1-p) * (p^xvals)/xvals)
}
```

Once you compile this code, you can use it to simplify parts b and d. For example, part d is simply:

```
1 - sum(dlogseries(1:7, 1/2))
```

How do you simplify the code in part b? Note that the use of user-created functions improves the clarity and simplicity of your solutions.

52. Read this short article and watch the included video from data journalism website 538. Then, write a single paragraph for how you would respond to "What is a p-value?" if asked this at a conference.

Hint List:

- Return to the lecture slides where we first define a p-value. Also, ask yourself "How do you know when something is true?".

Solution: Answers will vary. Here's what I might say:

A p-value is a number, in fact, a probability, that helps researchers decide between two competing hypotheses. To find a p-value, you must first set up two competing hypotheses. Then, you must collect data. Next, you find the p-value, which is the probability of seeing your observed data or data more extreme (where "more extreme" is influenced by the alternative hypothesis) assuming the null hypothesis is true. This value captures the strangeness/rarity of what you have observed. This value must then be compared to some arbitrarily chosen cutoff for rareness (often 0.05). If p is less than this cutoff, we have seen an insane (to us) event in a world governed by the null, so we throw out this belief system and adopt the alternative. If p is greater than or equal to the cutoff, our event wasn't strange and we continue to live in a world governed by the null. Nowhere in this decision-making process do we prove or disprove the null or alternative hypotheses. Science has no way of proving anything is true or false. The hypothesis testing framework is a tool for using data to triangulate researchers to belief structures that harmonize with the data the universe is giving them.

Commentary: It is tempting, indeed reasonable, to believe that experts know what they are talking about. After all, experts appear on news programs, are invited to give lectures, and make important decisions within the government. I'm here to tell you that when it comes to the statistics underpinning much of their work, experts are often quite confused or simply wrong. From psychologists to geneticists to social scientists to chemists, we have an academy of people who are required to use statistical methods in their work but who have a tenuous grasp on these ideas. Most of this is the result of inadequate training. The typical Ph.D. student in a research-based discipline likely takes only 1 or 2 courses in statistics (often called a quantitative methods course/sequence). Such students usually have disparate backgrounds, so it is challenging to pitch such courses at a high level or to maintain rigorous standards. Indeed, it is easier to train people how to hit buttons in computer programs (like SPSS, SAS, Tableau, Minitab, JMP, Excel, or Stata) than to develop a deep understanding of statistical ideas. People have been aware of this problem for quite some time. In the 1986 book "Statistical inference: A commentary for the social and behavioural sciences", Michael Oakes describes an experiment in which 70 academic psychologists were given 6 statements about p-values and asked to identify which were true. Despite all being false, respondents, on average, marked about 2.5 as true! Further work has shown similar troubles with understanding of confidence intervals, and researchers have demonstrated that these misconceptions occur at roughly the same (high!) rates in undergraduates, graduate students, and professors. The training you are getting in Math 183 is quite thorough and rigorous. Many students who take this class and join research groups on campus find they have the best statistical training in their groups!

53. Check out the last minute of this video by Twitch streamer and Mario Maker legend PangaeaPanga. In it, you'll see him playing a user-created level for Super Mario Maker 2 (SMM2). At the end, Panga has the option to Boo the level (blue broken heart icon) or not Boo the level, which includes Liking it (red heart icon) or not rating it at all (what he calls "indifferent"). This got me thinking: What percentage of levels would Panga Boo if he played all the levels in SMM2? Since there are millions of SMM2 levels, we need statistical inference to help us out.

    (a) Given that I've played thousands of SMM2 levels, I can personally attest that many of the levels are bad (that's user-created content for you). Suppose we want to explore if Panga will Boo 40% of levels or if the landscape of levels is actually worse. Complete the PH of PHANTOM.

    (b) One way that Panga plays levels is to have them randomly chosen by the system from all possible levels. Suppose we watch him play 100 levels in this way (on 4 separate days, 25 levels per day) recording his Boos and non-Boos. Explain why we might have a problem with the independence assumption in the A of PHANTOM. Also discuss the other assumptions in the A of PHANTOM. As always with Assumptions, you should not only state what the Assumptions are, but also whether each is or isn't met and why.

    (c) Ignoring the troubles of part b, we go ahead and find that Panga Boo-ed 47 of the 100 levels in our sample. Conduct a HT with $\alpha = 0.08$. [NTOM of PHANTOM, VIZ, Tech not tables]

Hint List:

- b: Can you see any problem with collecting 25 pieces of data all on the same day?

Solution:

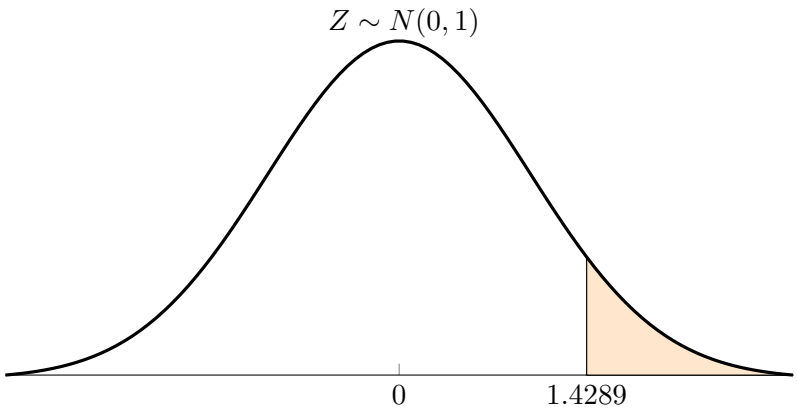
(a) (P) Let p be the true percentage of **all** SMM2 levels Panga would Boo.

(H) We test H_0: $p = 40\%$ (or 0.40) vs. H_1: $p > 40\%$ ("landscape worse")

(b) (A) After assuming H_0, we must check the two assumptions:

- Independence: While this is frequently checked via randomization and the $< 10\%$ condition, notice that Panga is playing 25 levels on *each of 4 days*. If he's in a bad mood (e.g., breakup, bad test score, etc.), he might Boo more levels than normal. This is one problem with concentrating data collection in one time or location.
- 10 predicted successes and failures: Under H_0, we expect $100 \cdot 0.4 = 40$ successes ("Boos") and $100 \cdot (1 - 0.4) = 60$ failures, both of which are ≥ 10.

(c) (N) We run a 1-sample z-test. Here, $\widehat{p} = \dfrac{47}{100} = 47\%$. So, (T) $z = \dfrac{47 - 40}{\sqrt{40 \cdot 60/100}} \approx 1.4289$. Thus, (O) $p = \text{pnorm}(1.4289, \text{lower} = \text{F}) \approx 0.077$. (M) Given that $0.077 < 0.08$, we reject H_0 in favor of H_1. It appears that Panga's Boo percentage might be higher than 40%.



Commentary: This is our first standard hypothesis testing question. Note how much work goes into properly running a test! My goal is for you to do this work and to avoid the tendency to just have technology do things for you. Many students who have taken statistics before (e.g., AP Stats in high school) have developed the bad habit of believing a hypothesis test is just about the data and P-value, and that you find these by hitting buttons on a calculator. In truth, a test is so much more. You want to take the time to carefully think about the population/parameter (P), what hypotheses the researchers care about (H), whether you should even run a test (A), and communicating what you are doing to others (N, M). In time, you'll use technology for the T and O steps like most people do, but don't forget the bread (P, H, A, N, and M) of the testing sandwich. Finally, you need not label the P through M as I did in red above, but it is helpful to keep you on track.

Another thing to consider is that even with this overarching structure, pitfalls abound. For example, when we talk about the "successes" and "failures" that make up a proportion, remember that "successes" need not be a positive or pleasant idea. Here, Boo-ing a level is a success because it is the focus of our study. Similarly, if you study the prevalence of a disease, your success might be someone who has the disease. Also, be aware that checking independence can ban be tricky. Even if you know your sample is $< 10\%$ of the population size and see the word "random" in the problem, you need to think carefully about how the data were actually collected. If I wanted to know what percentage of days you set an alarm to wake up, I could randomly sample from *your weekends* and get a result. Since you're more likely to use an alarm during the week, we have a problem. Here, the issue is that we didn't sample randomly *from all days of the week*, which is the true population. In general, randomly sampling from *a subset* of the population is a bad practice as your data will be dependent; they are all linked since they belong to the same subset.

54. I recently read "The Coddling of the American Mind" by Lukianoff and Haidt. On page 110, they mention an interesting study in which people can supposedly guess, better than chance, the political party a person supports by looking at a picture of just their desk (where any obvious signs of political affiliation have been removed). Suppose we plan on replicating this study, but we're too lazy to do the actual work, so we'll make up data (!!). We'll pretend we brought in 103 random Americans and showed each a random picture of a desk from someone who was either a Democrat or a Republican (our pictures would be evenly split between the two parties). We'll pretend to record the number of correct identifications, x, and then run a HT to show that people can, in fact, guess affiliations more often than by pure chance. What is the smallest value we can make up for x that gives a statistically significant result assuming we use $\alpha = 0.0228$? Do not use guess-and-check. Recall that the term "statistically significant" means that you get a P-value $<$ the significance level α. [PHANT, VIZ, Tables not tech]

Hint List:

- When setting up the null, how often will a person guess correctly if they are randomly guessing? This is the standard we should set up as a baseline, not 0%.
- Try working backward from the P-value. What quantile (i.e., location on the horizontal axis) cuts off an area of 0.0228?

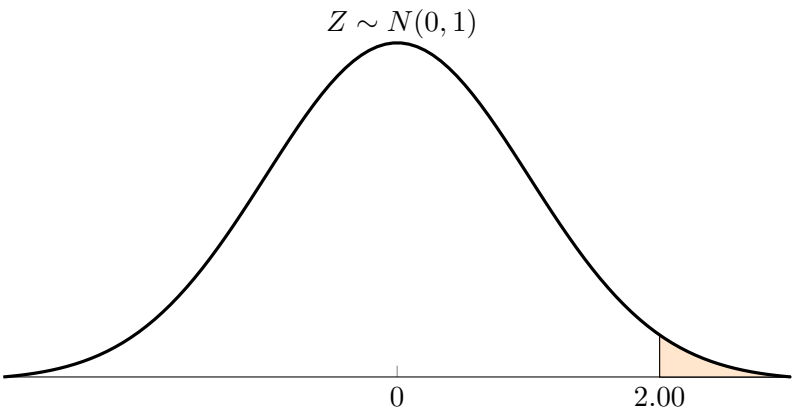
Solution: (P) Let p be the percentage of political affiliations people can correctly identify when shown a picture of a Democrat's or Republican's desk. If the photos are half Republican/half Democrat, then with random guessing, you'll get things correct half the time.

(H) We test H_0: $p = 0.5$ vs. H_1: $p > 0.5$ ("better than random chance")

(A) Assuming H_0, to run our fake study, we must check:

- Independence: Yes, since we randomize people and pictures, and there are more than $103 \cdot 10 = 1030$ trials we could run.
- 10 predicted successes and failures: We expect $103 \cdot 0.5 = 51.5$ successes and failures, which are ≥ 10.

We run a (N) one-sample z-test. Since $\widehat{p} = \dfrac{x}{103}$, our test stat (T) is $z = \dfrac{x/103 - 0.5}{\sqrt{0.5 \cdot 0.5/103}}$.



Based on a Z-table, the area beyond 2.00 is $1 - 0.9772 = 0.0228$. Thus, if our test stat is to the right of 2.00, we will get an area of less than 0.0228, and hence, a statistically significant result. So, we set:

$$\frac{x/103 - 0.5}{\sqrt{0.5 \cdot 0.5/103}} > 2.00 \overset{\text{alg.}}{\Longrightarrow} x > 61.2$$

Since x is an integer, we get $\boxed{x = 62}$.

Commentary: The finding mentioned by Lukianoff and Haidt is a classic psychology result. It suggests that our political affiliation isn't just whom we vote for – it is correlated with other aspects of who we are, how we live, and how we present ourselves (e.g., the look of our desks). This is not shocking; people are complex webs of behavior, with most traits affecting other traits. If you were really paying attention, however, you would have noticed a very important (intentional) omission on

my part. What was the percentage of correct guesses in the study the authors cite? 51%? 90%? While researchers and the media are often obsessed with statistical significance, a more important metric is the effect size. For a simple proportion problem, this is just the percentage you are trying to measure. In more complex settings, the effect size is more complex as well. For example, if you're measuring what percentage of patients report feeling relief using your drug versus a placebo, the effect size might be $p_{\text{drug}} - p_{\text{placebo}}$. If you've designed a new educational intervention, the effect size might be the average amount students improve by on an important exam. At its heart, the effect size tries to answer: How *practically* important is what I'm seeing? To be honest, I no longer care about statistical significance, for this is simply a measure of whether the results surpass what we reasonably expect from random chance. Even with that knowledge, I still have no sense of what is really going on. Does your new drug cure 0.1% more people or 70% more people? Does your approach to teaching logarithms improve student understanding by 1 point or 15 points, on average? The abuse of "statistical significance" vs. "practical importance" is especially prevalent in the pharmaceutical market. How much does your asthma medication actually help? What percent of people show improvement from a depression medication (beyond a placebo)? We all take medication because "it helps" (i.e., statistical significance has been achieved in a large randomized controlled trial). But by *how much*?!

55. Recently, a friend and I were watching a soccer game on TV. I noticed how international the players were, sparking the question: "What percentage of current Major League Soccer (MLS) players were born in the United States?". One approach to estimating this is to look at the roster for the Portland Timbers and count how many aren't labeled as "International" under the "Player Category" column. When I did this recently (don't use more up-to-date data), I found 9 of the 27 players were born in the US.

(a) State the population, parameter, sample, and statistic in this problem.

(b) Find an 80% CI for the parameter mentioned in part a. [PANIC, Tables not tech]
Note: In this class, if you ever fail to meet assumptions, you should continue with the problem. The goal is to practice skills. In the real world, you should stop the analysis and collect a better data set.

(c) Suppose the MLS wanted a narrower CI for the parameter in part a. One approach is to change the confidence level of the interval. What value should we change 80% to in order to have a 1% margin of error (MOE)? [Tables not tech]

(d) Another approach to getting a 1% MOE is to keep the confidence level at 80% but to collect a new, larger data set. Using your first data set as a guide in sample-size planning, how large should n be to ensure a 1% MOE (or smaller)?

(e) Explain why the results you get from parts c and d are silly solutions to the goal of getting a 1% MOE.

Hint List:

- c: Solve for $z_{\alpha/2}$. Then, draw a picture of Z and try to label where $z_{\alpha/2}$ and $\alpha/2$ are in the picture.

Solution:

(a) Population: The birthplace status (US/International) of all current MLS players
Parameter: p, the proportion of all current MLS players born in the US
Sample: The birthplace status of 27 players from the Portland Timbers
Statistic: $\widehat{p} = \dfrac{9}{27}$, the percentage of our sample born in the US

(b) (P) Let p be the percentage of **all** current MLS players born in the US. To calculate a (N)one-sample z-interval, we must check (A):

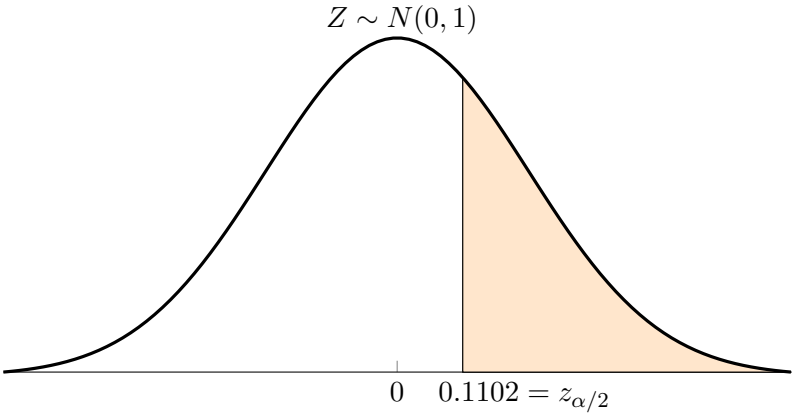
- Independent data: Failed, since all our data come from the same team, which may exhibit atypical trends.
- At least 10 observed successes and failures: Failed, only 9 successes.

Foolishly, we move ahead. For 80% confidence, $\alpha = 0.2$. So, $z_{\alpha/2} = z_{0.1} \approx 1.28$ (or 1.29) using a Z-table. Thus,

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}\,\widehat{q}}{n}} = \frac{9}{27} \pm 1.28\sqrt{\frac{\frac{9}{27}\cdot\frac{18}{27}}{27}} \approx [21.721\%, 44.946\%] \text{ (I)}$$

(C) **If** $p \in [21.721\%, 44.946\%]$, then the proportion of all MLS players born in the US is between about 21.7% and 44.9%.

(c) The MOE $= z_{\alpha/2}\sqrt{\dfrac{\widehat{p}\,\widehat{q}}{n}}$, or $0.01 = z_{\alpha/2}\sqrt{\dfrac{\frac{9}{27}\cdot\frac{18}{27}}{27}} \implies z_{\alpha/2} \approx 0.1102$. Remember that $z_{\alpha/2}$ is a quantile – a place on the horizontal axis – and the subscript $\alpha/2$ shows how much area is to the right of this place. We draw a picture noting that the shaded area is $\alpha/2$:



But we can also find the shaded area using a Z-table. The area to the right of 0.1102 is about 0.4562. So, $\dfrac{\alpha}{2} \approx 0.4562 \implies \alpha \approx 0.9124$. Thus, our CL is $100 \cdot (1 - \alpha) \approx \boxed{8.76\%}$.

(d) Here, we need

$$1.28 \cdot \sqrt{\frac{\frac{9}{27} \cdot \frac{18}{27}}{n}} \leq 0.01 \implies \frac{1.28}{0.01}\sqrt{\frac{9}{27} \cdot \frac{18}{27}} \leq \sqrt{n} \implies 60.3397 \leq \sqrt{n} \implies 3640.9 \leq n$$

Since n must be an integer, this forces $\boxed{3641 \leq n}$.

(e) The solutions in parts c and d are both silly. In part c, we have a CL so low that the true parameter is very unlikely to be in the interval we report. In part d, the required sample size is larger than the current number of players in the MLS! The point of sampling is to draw a small subset of a large population.

Commentary: Buried in the middle of this problem is a standard CI exercise (the I of part b). Around that exercise, I have layered mounds and mounds of nonsense to show you some of the ways that people abuse statistics. Here are some of those abuses:

- From the start, using a US/International binary is too simplistic. While a person might be born in a location, it is common in the modern era to move to a new location and receive most of your training outside your birthplace. The world is obsessed with over-simplified binaries.
- Our sample is trash. We see this in the failure of both conditions in the A of PANIC. This is depressingly common: academic psychologists study college students at their universities, educational researchers work in atypical classrooms with teachers most excited about change efforts, and medical trials often enroll volunteers. Garbage samples in, garbage inference out.
- While it is fine to think about your desired MOE (as in parts c and d), this should be done *in advance of data collection* since it can affect your sample size (part d) or your confidence level (part c), both of which must be considered carefully.
- Most importantly, this setting is one in which you should NOT be doing statistical inference. "Inference" refers to using a small subset of an enormous population to draw conclusions about that population (via HT or CIs). Here, we want to study the MLS, which is small enough (29 teams $\cdot$ 30 players/team = 870 total players) to study without the need for inference. Just get the data you want on **all** the players and directly find p.

One of strongest temptations students have is to break out technology/formulas and then churn away: You see a problem about topic X, you have some formulas for topic X, so you get to work (here, the I of PANIC). While the act of doing is important, so are all the related activities around it. One goal of PANIC is to force you to think outside the calculation part of statistics (the I). You need to better communicate what you are doing (the P, N, and C) and to check if you should be doing it (the A). This problem shows that there are even issues to consider outside of PANIC: Is inference even necessary in my setting? Are the goals I have for PANIC reasonable (e.g., desiring a certain MOE)? The theme of this commentary is also relevant outside of statistics. We need to look beyond the churning (e.g., creating a social media site, introducing a cryptocurrency, releasing a new AI) and ask larger questions like "Should I do this?" and "How do I communicate this effectively?".

56. Those Pesky Assumptions! What happens when you don't meet the assumptions when running a HT? Here, we run a simulation to explore life when the assumptions are violated. Suppose you plan to interview 50 random people and ask if they have some less-common trait (e.g., left-handedness or green eyes). We think the trait appears 10% of the time, but we're not sure, and it might be different. Write code in R that simulates asking those 50 people and reports a single value for $\widehat{p}$. Now convert this to the appropriate z-stat. Next, repeat this entire process to create 10000 total z-stats. Make a histogram of the z-stats and overlap a histogram of 10000 random values from $Z \sim N(0, 1)$. Your answer should include the PHA steps of PHANTOM, your code, a sketch of the overlapped histograms, and some commentary on what you notice in the picture and why it occurs.

Hint List:

- How could the Binomial RV be used to help run the simulation?
- In a previous solution, we saw you can overlap histograms by simply including add = T in the parameter list of the second usage of the function hist. You also learned how to specify a color by its red, green, and blue (RGB) values and its opacity.

Solution: (P) Let p be the percentage of all people with our less-common trait.

(H) We test H_0: $p = 0.10$ vs. H_1: $p \neq 0.10$ ("might be different").

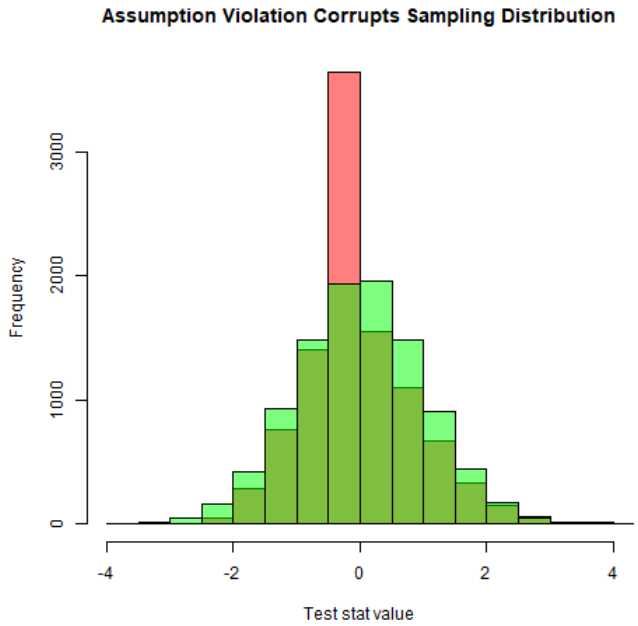
For our test to be well-founded, we must meet these assumptions upon assuming H_0 (A):

- Independence: Since we simulate data, we'll have randomness. Our sample size is also $< 10\%$ of the population size.
- At least 10 expected successes and failures: We expect $np = 50 \cdot 0.1 = 5$ successes (violation!) and $nq = 45$ failures.

When your setup violates the assumptions, this will result in a sampling distribution for the test stat that does NOT follow the curve it should. We can see this using the below R code:

```
simulated_phats = rbinom(10000, 50, p = 0.1)/50
simulated_zstats = (simulated_phats - 0.1)/sqrt(0.1 * 0.9/50)
hist(simulated_zstats, col = rgb(1, 0, 0, 0.5),
     main = "Assumption Violation Corrupts Sampling Distribution",
     xlab = "Test stat value", xlim = c(-4, 4))

actual_zvals = rnorm(10000)
hist(actual_zvals, col = rgb(0, 1, 0, 0.5), add = T)
```



Why does our distribution in pink have the shape it does? Since we don't expect at least 10 successes, the histogram of rbinom(10000, ...) gets bunched up on the left side of the support near 0. Once these

values are fed through the standardizing formula $z = \dfrac{\widehat{p} - p}{\sqrt{pq/n}}$, this compression results in more negative values than we should have when compared to the green histogram of values from $Z \sim N(0, 1)$.

Commentary: One bit of sloppiness I see in the real world regarding statistics is the failure to check the assumptions when running a hypothesis test. Indeed, I have *never* seen a single journal article where the authors discuss the assumptions required for their tests and verify they were met. As this simulation reveals, when you don't meet the assumptions, your test stat is one of many which, together, create a sampling distribution that does not have the shape it is supposed to. As such, if you begin finding areas until a shape you have not achieved (e.g., when finding a P-value), your result will not be an accurate reflection of what is happening, and you could be led in an incorrect direction.

One reason scholars omit this discussion is they simply don't know statistics well enough to realize these things matter. A more sobering reason for their omission is that researchers don't want to face the stark reality that assumption provide. In essence, the assumptions are telling you whether you have a good enough data set to study the thing you want to study. If the data were collected in a suspect way (e.g., recruiting college students when your population is all Americans), you'll fail the independence assumption. If there's a mismatch between the rarity of your trait and the amount of data you have (as in this problem), you'll fail the ≥ 10 condition. Since many studies have inadequate data sets, ignoring assumptions has become an unfortunate de facto practice in the research community.

57. What is the best dessert to serve at Thanksgiving? One way to decide this is to ask millions of people to make a binary choice: Dessert Option 1 vs. Dessert Option 2 (e.g., apple pie vs. pecan pie). If you randomize what options people must compare, then every dessert will have a "win percentage". This is the percentage of times the dessert beats its (random) opponent when humans were forced to make a binary (!) decision. Luckily, the data journalism website 538 did this very thing! According to their data, pumpkin pie reigns supreme, winning 81% of head-to-heads with other desserts. Since the poll was from 2016, you might be curious if things are different now. So, you conduct a new round of head-to-heads and find that in 632 of these, people preferred pumpkin pie. Sadly, you forgot to write down the total number of head-to-heads, n. You do know that your hypothesis test came back statistically significant at the $\alpha = 0.101$ level. What are all the possible values n could have been? Don't use R, trial-and-error, or a computer algebra system to help solve this. [PH of PHANTOM, VIZ, Tables and basic calculator operations OK]

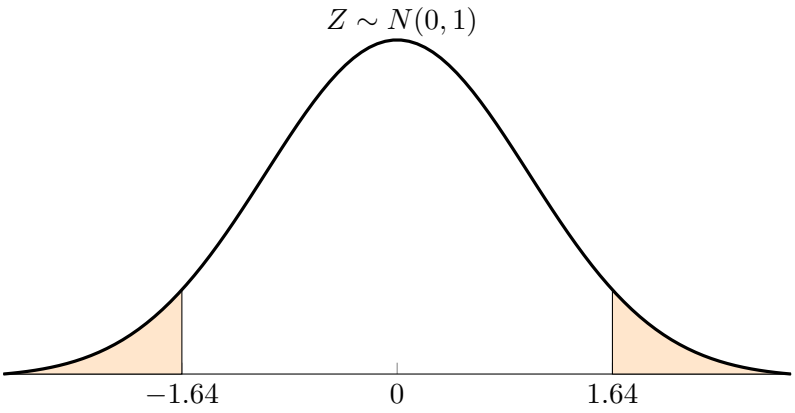
Hint List:

- Begin by writing an expression for the z-stat.
- Because of the two-sided alternative, we are excited about either very small or very large z-stats. What are the cutoffs for that excitement?
- Notice that $a < -12$ and $a > 12$ can be written more simply as $|a| > 12$, or even more simply as $a^2 > 144$.
- To solve a quadratic inequality, start by finding its roots. Then draw a picture of the quadratic and label the roots to solve the inequality.

Solution: (P) Let p be the current percentage of **all** dessert head-to-heads that pumpkin pie wins.

(H) We test H_0: $p = 0.81$ vs. H_1: $p \neq 0.81$ ("things are different now").

We have $\widehat{p} = \dfrac{632}{n}$, so $z = \dfrac{\frac{632}{n} - 0.81}{\sqrt{0.81 \cdot 0.19/n}}$. Because of our two-sided alternative hypothesis, P-values will involve areas from both the lower and upper tails. If we split $\alpha = 0.101$ into two equal pieces (0.0505 each), we can find the zones along the horizontal axis that would lead to statistically significant results. Using a Z-table, we find $P(Z \leq -1.64) \approx 0.0505$:



Thus, we need our z-stat to satisfy either $z < -1.64$ or $z > 1.64$. These two cases can be combined into a single expression if we use absolute value bars: $|z| \geq 1.64$. So, we solve

$$\left| \frac{\frac{632}{n} - 0.81}{\sqrt{0.81 \cdot 0.19/n}} \right| \geq 1.64$$

We can get rid of the absolute value by squaring each side, and after rearrangement, we have:

$$\frac{632^2}{n^2} - \frac{1023.84}{n} + 0.6561 \geq \frac{0.41392944}{n}$$

Multiply by n^2 and simplify to get the quadratic inequality $0.6561n^2 - 1024.253929n + 399424 \geq 0$. We use the Quadratic Formula to find where the quadratic equals 0:

$$n = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \approx 802.75 \text{ or } 758.37$$

So, we have an upward-opening quadratic with roots at 758.37 and 802.75. This quadratic will be positive to the left of the first root and to the right of the second root (draw a picture!). Since n must be an integer, this forces $n \leq 758$ and $n \geq 803$. Now, note that $n \geq 632$ because you can't have 632 successes in fewer than 632 trials! So, our final answer is $\boxed{n \in \{632, 633, \ldots, 758, 803, 804, \ldots\}}$.

Commentary: One thing I like to do when creating problems is to change what is known and unknown. In a typical proportion HT setup, we know: the population of interest (for the P of PHANTOM), the value of p_0 in H_0: $p = p_0$ (for the H), the sided-ness of the alternative (H), how the data were collected (A), the values of x and n in $\widehat{p} = \dfrac{x}{n}$ (T), and the significance cutoff α (M). The goal in such setups is to find the P-value (O). In the above problem, all I did was change n from a known to an unknown, and switch the P-value from an unknown to a known (by requiring statistic significance, we force the P-value $< \alpha$). As you probably noticed, this small change created a big challenge and, in the process, revealed the *sensitivity* of the HT apparatus to changes in its components. Having n as an unknown is more troubling for the HT infrastructure than having the P-value as an unknown. This is somewhat expected since the HT infrastructure is designed to move forward *toward* the P-value. So, to go from the P-value back to n is like swimming against a current.

Computer science majors have some experience with the sensitivity of solution pathways to the adjustment of setups. As an example, suppose you have this programming problem:

How many positive integers less than 1000 are multiples of 3 or 5?

One solution pathway is to write a loop that goes from 1 to 999. While looping, we'll increment a counter by 1 each time we encounter a multiple of 3 or 5. This approach works well when the setup uses the number $1000 = 10^3$. What happens if we alter the setup from 10^3 to 10^{50}? This small change doesn't just make the solution pathway harder (as in the pumpkin pie problem), it actually makes the pathway unusable (a standard modern computer could not finish looping before you died). Computer scientists use something called big O notation to describe solution pathways; this notation essentially quantifies the sensitivity of your solution to changes in certain inputs. If you're looking for a collection of math-CS problems that really help you feel this sensitivity phenomenon, head to Project Euler. For problems at the 5-20% difficulty, just about any solution pathway works. As you move into the 25-60% zone, your first few solution pathways are usually too fragile to survive the demands of the problem (e.g., a solution might work for 10^3 but the problem demands 10^{12}). Above 65%, your solution must be a speedboat to overcome the upstream torrent you face.

58. What should we do about peanut allergies in our society? In America, a common approach in recent times has been to create nut-free products, nut-free classrooms, nut-free dining halls, etc. The idea was to create safe spaces through avoidance. This thinking was challenged by a 2015 study that suggested nut avoidance causes *more* children to develop peanut allergies, not fewer. Here's a piece of what they did: Researchers took 542 babies with no nut allergies and divided them into two groups. The "avoidance group" ($n = 270$) was kept away from peanuts, while the "consumption group" ($n = 272$) was fed a specific amount of peanut powder per week. At the end of the time frame, the groups were compared to see what percentage had developed peanut allergies. 37 babies in the avoidance group and 5 in the consumption group developed peanut allergies.

(a) Conduct a hypothesis test using a significance level of 0.01 to see if (shockingly!) avoiding peanuts actually leads to more peanut allergies among children. [PHANTOM, VIZ, Tables not tech]

(b) How would your p-value from part a change if the test were two-sided, not one-sided? Is there *any* difference in our conclusion from the test? [VIZ]

(c) Construct a 98% CI for the difference in true peanut allergy percentages that develop using the two approaches. The subtraction order is your choice; also, see part d. [PNIC of PANIC, Tables not tech]

(d) How does your answer from part c change if you use the other order of subtraction?

(e) In R, the prop.test command can run hypothesis tests and make CIs for this setting. Type ?prop.test in R to help understand this function, then write code to check your answer to part c. Make sure to always use correct = F to turn off a feature known as continuity correction. (Your answer will differ slightly because at one point you used tables which are less precise.)

Hint List:

- a: Don't forget to pool proportions in a two-sample z-test.
- b: Here, the VIZ is especially useful.
- c: Why don't we pool here?
- d: Can you use inequalities to show that our final result is the same upon switching the subtraction order? Remember that $a \in [b, c] \iff b \leq a \leq c$.

Solution:

(a) (P) Let p_A (p_C) be the percentage of **all** no-nut-allergy babies that develop peanut allergies when avoiding (consuming) peanuts in youth.

(H) We test H_0: $p_A = p_C$ (or $p_A - p_C = 0$) vs. H_1: $p_A > p_C$ (or $p_A - p_C > 0$) ("avoiding ... leads to more").

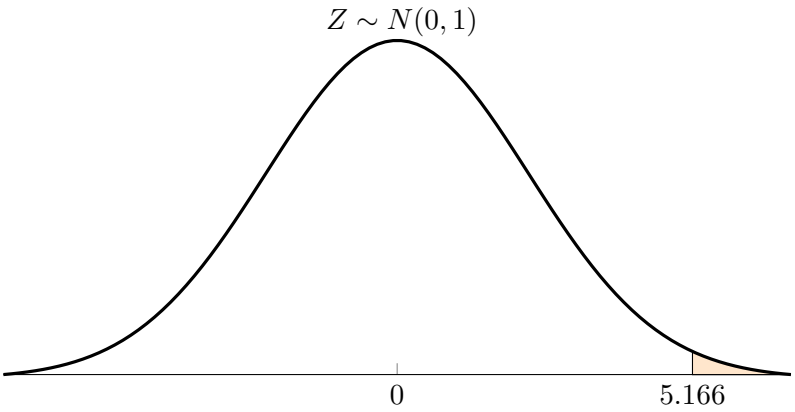
(A) Assuming H_0, to run a (N) two-sample z-test, we check:

- Independent samples: Yes, we draw our samples from two populations that are not connected because we randomly assign babies to the two groups.
- Within each sample, we explore:
  - Independent data: Yes, via randomization and $< 10\%$.
  - At least 10 observed successes and failures: No, in the consumption group we only have 5 successes (babies who develop allergies).

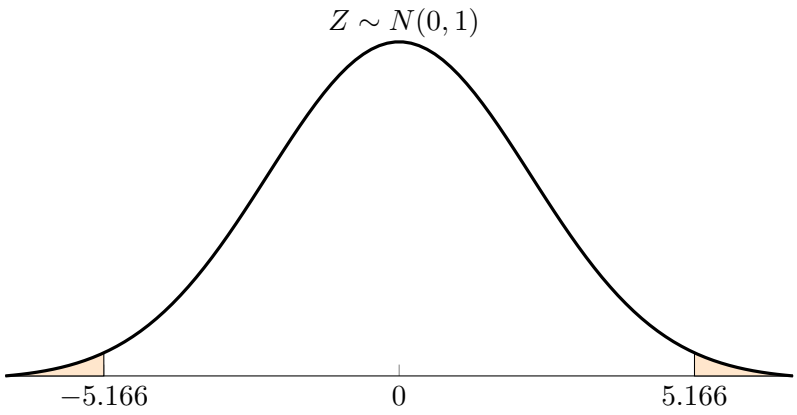
Disregarding the A failure, we move ahead. Recall that we pool proportions when running a two-sample z-test. We have: $p_* = \dfrac{37+5}{270+272} = \dfrac{42}{542}$. Thus,

$$\text{(T) } z = \frac{\widehat{p}_A - \widehat{p}_C - (p_A - p_C)}{\sqrt{\frac{p_* q_*}{n_A} + \frac{p_* q_*}{n_C}}} = \frac{\frac{37}{270} - \frac{5}{272} - 0}{\sqrt{\frac{\frac{42}{542} \cdot \frac{500}{542}}{270} + \frac{\frac{42}{542} \cdot \frac{500}{542}}{272}}} \approx 5.166$$

(O) This z-stat is off most tables, so our P-value is basically 0:



Since $0 < 0.01$, we (M) reject H_0 in favor of H_1. It appears that peanut avoidance leads to greater allergy development than peanut consumption!



(b) With a two-sided alternative (H_1: $p_A \neq p_C$) our VIZ shades both tails as seen above. So, our P-value is twice the area from before (technology gives about $2.39 \cdot 10^{-7}$). Note that we reject H_0 as before, but now our alternative is H_1: $p_A \neq p_C$, so our belief after the test is that there is a *difference between the two groups*. This is a different conclusion than in part a where we leave the test believing that avoidance leads to *greater allergy rates*.

(c) We explore the subtraction order (P) $p_A - p_C$ (other order OK since the problem does not force a certain order) using a (N) two-sample z-interval. Here, the CL is 98%, so $\alpha = 0.02$. Thus, $z_{\alpha/2} = z_{0.01} \approx 2.33$ (or 2.32) using a Z-table. We find (I):

$$\widehat{p}_A - \widehat{p}_C \pm \sqrt{\frac{\widehat{p}_A\widehat{q}_A}{n_A} + \frac{\widehat{p}_C\widehat{q}_C}{n_C}} \approx \frac{37}{270} - \frac{5}{272} \pm 2.33\sqrt{\frac{\frac{37}{270}\cdot\frac{233}{270}}{270} + \frac{\frac{5}{272}\cdot\frac{267}{272}}{272}} \approx [6.633\%, 17.098\%]$$

(C) **If** $p_A - p_C \in [6.633, 17.098]$, then $p_A \in [p_C + 6.633, p_C + 17.098]$, which says that avoiding peanuts increases the rate of peanut allergy development in babies by between 6.633% and 17.098%.

(d) If we repeat the above approach but use $p_C - p_A$, we get the negative version of our interval: $[-17.098\%, -6.633\%]$. Note that this says the same thing because:

$$p_C - p_A \in [-17.098, -6.633] \iff -17.098 \leq p_C - p_A \leq -6.633$$

Multiplying by -1 changes the terms and reverses the inequalities:

$$17.098 \geq -p_C + p_A \geq 6.633 \iff p_A - p_C \in [6.633, 17.098]$$

(e) To check part c, we do:

```
prop.test(x = c(37, 5), n = c(270, 272), conf.level = 0.98,
          correct = F)
```

We get the interval $[6.641\%, 17.090\%]$, as above.

Commentary: This problem is an all-in-one adventure for proportions in two samples. As mentioned in previous commentaries, what you see here are the *standard problem types*: running a hypothesis test

and building a CI. I also tried to include parts that address questions students commonly have (e.g., What changes between a one-sided and two-sided alternative? Does the subtraction order matter for the CI?). I also wanted you to see the stark difference between engaging with the problem by hand (parts a and c) vs. using technology (part e). After doing parts a and c yourself, the result of part e feels terse and devoid of meaning/context.

This problem also provides a nice opportunity to organize your knowledge of proportion inference so far. When we look at HT and CI-building for one- and two-sample setups, you might notice some subtle differences. For example, the A of PHANTOM and PANIC requires you check there are at least 10 successes and failures (S and F). Sometimes we use *expected* S and F, but other times, we use *observed* S and F. Also, sometimes we calculate a pooled proportion, and sometimes not. Below is my internal summary of these small differences:

	HT	CI
One-sample	No pooling, ≥ 10 **expected** S and F	No pooling, ≥ 10 observed S and F
Two-sample	**Pooling**, ≥ 10 observed S and F	No pooling, ≥ 10 observed S and F

Why do we have these annoying differences? The simple answer is that we always want to bring the most appropriate information we can to an inferential process. For example, when it comes to pooling, this is not possible in the one-sample setup because it requires two samples that must be considered next to one another. In the two-sample case, only the HT has us assume H_0: $p_A = p_B$, and hence, pooling makes sense in the HT, but not in the CI. With regard to expected vs. observed S and F, only the one-sample HT has us assume H_0: $p = \#$. Because we believe in a *specific value* for p, it is possible to calculate np and nq, the expected S and F. CIs don't involve assuming H_0, and the two-sample HT only has us assume H_0: $p_A = p_B$, which doesn't give us any actual values to use when calculating expected S and F.

When you encounter many similar techniques in a given field, spend time organizing the techniques in your mind. Look for similarities and differences and try to understand why the differences exist.

59. Do you think your future employers will look at your social media accounts in the hiring process? If so, do you think things like religious affiliation, which is protected under hiring law, will influence their decision to invite you for an interview? In a 2015 paper, researchers explored this question. They created two identical resumes that differed only in the name at the top. Each name was searchable online and led to a fake Facebook account that clearly matched the resume and revealed the candidate to either be Christian or Muslim. After posting the two resumes on various hiring platforms, they explored what percentage of resume views by employers led to requests for an interview. At the end of the study, the resume linked to the Muslim Facebook account got 112 interview requests from 1022 views, while the Christian resume got 134 requests from 1060 views.

    (a) Find a 92% CI for the difference in interview invitation proportions for the two resumes. Use the subtraction order: Muslim − Christian. [PANIC, Tables or tech]

    (b) Figure out how to use prop.test in R to get the same results you found in part a.

Hint List:

- This problem is quite straight-forward. Why do you think I assigned it?

Solution:

(a) (P) Let p_M (p_C) be the proportion of **all possible** views of the resume linked to the Muslim (Christian) Facebook account that would lead to interview requests via online job-hunting services. To build a (N) two-sample z-interval for $p_M - p_C$ using $\widehat{p}_M - \widehat{p}_C$, we check (A):

- Independent populations: Yes, since we have two separate Facebook pages.
- In each sample, we check:
  - Independent data: Maybe. We meet the $< 10\%$ criterion, but it's unclear if the "various hiring platforms" were randomly chosen. For example, imagine if the Muslim Facebook account/resume was posted to sites dedicated to STEM hiring, but the resume was not applying for such a job. This might give an inaccurate assessment of the Muslim resume that has nothing to do with the linked Facebook account.
  - At least 10 observed successes and failures: Yes.

Now, we find (I):

$$\widehat{p}_M - \widehat{p}_C \pm z_{0.08/2}\sqrt{\frac{\widehat{p}_M\widehat{q}_M}{n_M} + \frac{\widehat{p}_C\widehat{q}_C}{n_C}} \approx \frac{112}{1022} - \frac{134}{1060} \pm 1.75\sqrt{\frac{\frac{112}{1022}\cdot\frac{910}{1022}}{1022} + \frac{\frac{134}{1060}\cdot\frac{926}{1060}}{1060}}$$
$$\approx [-0.04156, 0.007911] \approx [-4.156\%, 0.791\%]$$

(C) **If** $p_M - p_C$ is in our interval, then having a Muslim (rather than a Christian) designation on your Facebook page alters your interview percentage by between -4.2% and $+0.8\%$.

(b) The below code gives our answer from part a.

```
prop.test(x = c(112, 134), n = c(1022, 1060), conf.level = 0.92,
          correct = F)
```

Commentary: There are three important things I wanted to do with this problem. First, I wanted an example where the bounds on the CI were small, like the number 0.007911. Notice that this number is a decimal, not a percentage, because I used decimals instead of percentages throughout my calculation (i.e., $\frac{112}{1022} \approx 0.10959$ instead of 10.959%). Said simply: what you put in is what you get out. Now, for those who converted to percentages after finding the CI, you get 0.791%, which is correct, but feels terrible since we rarely encounter percentages less than 1. Did you have this feeling if you converted to percentages (or used them throughout your calculation)?

Second, I wanted (via the A of PANIC) to reveal just how careful you have to be when doing research. Suppose this study happened in a research group where one grad student was tasked with uploading only the Muslim resume to job hiring sites, and a different grad student handled only the Christian resume. If these grad students choose sites they are personally familiar with, you might have a huge disparity between the two populations that reflects the job-site familiarity of the grad students, not the religion of the linked Facebook profile.

Finally, and most importantly, what should one make of a CI like we get in this problem? It contains negative values (suggesting bias in favor of Christians), 0 (suggesting no difference), and positive

values (suggesting bias in favor of Muslims). The correct conclusion is that any of those three beliefs are reasonable; the data from this study do not generate a CI that can help us choose one of the three possibilities. Now, it's natural to ask: "Since *most* of the CI is negative, doesn't this mean that $p_M - p_C$ is most likely to be negative, and hence, we have a bias in favor of Christians?". Sadly, this logic is flawed. First, $p_M - p_C$ may not even be in the interval (see my **If** above). Second, the parameter expression $p_M - p_C$ is *not* a random variable. It is a fixed, unknown-to-us number, so it doesn't make sense to speak of it using probability language. In the end, you have no choice but to leave the problem feeling dissatisfied with the conclusion, and this feeling was the goal of the problem.

60. In this problem, we consider what effect changing the CL has on our answer to question 59.

(a) Will decreasing the confidence level from 92% to 50% make a CI narrower or wider? Explain your choice by referring to a particular part of the CI formula.

(b) If you did problem 59 correctly, you'll notice that 0 should be in your CI. This suggests that the true difference in proportions might be 0, and hence, that religious affiliation on Facebook profiles may not matter in the hiring process. If you rerun prop.test using a CL of 50%, you'll notice the CI does not contain 0. Indeed, your interval will have all negative values, suggesting that Christian candidates are more likely to be invited for an interview. What is the largest value you could use for the CL that suggests a preference for the Christian profile? Solve algebraically, not via trial and error.

Hint List:

- a: The width of the CI is determined by the expression after the $\pm$ in the CI formula.
- b: Based on part a, as the CL increases, the CI width increases. So, as we increase the CL from 50% to 92%, we start with an interval completely in the negatives, and it grows to an interval that has both negative and positive values. What do you notice about the upper bound of the CI during this growth?

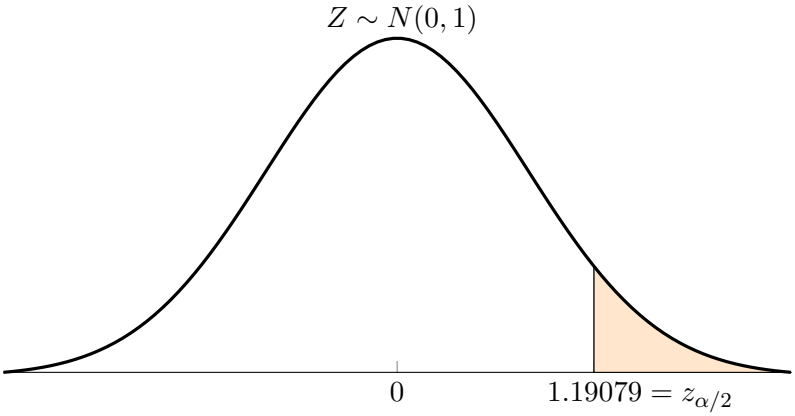
Solution:

(a) As the CL decreases (from, say, 92% to 50%), the CI width decreases. In our case, with a 92% CL, we have $\alpha = 0.08$, so $z_{\alpha/2} = z_{0.04} \approx 1.75$. When the CL is 50%, $\alpha = 0.5$, so $z_{\alpha/2} = z_{0.25} =$ qnorm(0.25, lower=F) ≈ 0.67. Since $z_{\alpha/2}$ got smaller, we don't add/subtract as much padding in the formula $\widehat{p}_M - \widehat{p}_C \pm z_{\alpha/2}\sqrt{\dfrac{\widehat{p}_M\widehat{q}_M}{n_M} + \dfrac{\widehat{p}_C\widehat{q}_C}{n_C}}$, and hence, the CI is narrower.

(b) If the CI is all negative numbers, this suggests a preference for the Christian resume. If $0 \in$ CI, then $p_M - p_C$ may be 0, and we lose the preference. So, we want the largest CL that gives an all-negative CI. Thus, we set the upper bound of the CI to 0 and solve for α:

$$0 = \widehat{p}_M - \widehat{p}_C + z_{\alpha/2}\sqrt{\frac{\widehat{p}_M\widehat{q}_M}{n_M} + \frac{\widehat{p}_C\widehat{q}_C}{n_C}} \implies z_{\alpha/2} = \frac{\dfrac{134}{1060} - \dfrac{112}{1022}}{\sqrt{\dfrac{134 \cdot 926}{1060^3} + \dfrac{112 \cdot 910}{1022^3}}} \approx 1.19079$$

Now, recall that $z_{\alpha/2}$ is a quantile, or place on the horizontal axis, to the right of which the area is $\alpha/2$. We draw a picture to help where the shaded area is $\alpha/2$:



Since pnorm(1.19079, lower = F) ≈ 0.117, we have $\alpha/2 \approx 0.117 \implies \alpha = 0.234$.

Thus, our CL is $(1-\alpha)100\% = (1 - 0.234)100\% \approx \boxed{76.626\%}$.

Commentary: There are three important takeaways from this problem. First, part a is an example of "conceptual understanding". The problem helps you create a mental link between these ideas:

$$\text{CL} \longleftrightarrow \alpha \longleftrightarrow z_{\alpha/2} \longleftrightarrow \text{width of the CI}$$

Ultimately, we want to know that increasing the CL will increase the width of the CI, and vice versa. Some students like to internalize this idea through real world metaphors: "Want to be more sure you'll catch a fish? Throw a bigger net!" or "To increase your chances of catching the criminal, widen your search radius." I find these metaphors are strengthened by uncovering the formulaic underpinnings of the trend.

Second, part b is another example of me switching up what is known and unknown in a standard procedure. The typical procedure has a known CL, known data, and an unknown CI we must find. Here, we have a known CI (by setting the upper bound to 0), known data, and an unknown CL. Thus, we must reverse the normal solving architecture and work backwards from the typical end point (the CI). One nice thing about such problems is that we can usually check our answer at the end by simply following our usual forward steps. For efficiency, you might use R to check:

```
prop.test(x = c(112, 134), n = c(1022, 1060), conf.level = 0.76626,
          correct = F)
```

This gives a CI with an approximate upper bound of 0.

Finally (and most concerning!), this problem shows how simply changing the CL can give a different result in a study. This is another lever that corrupt people can pull on to create the outcome they desire. To be clear, this is an abuse of statistics, and you should always choose your CL *before* looking at your data. Remember that all tools (calculators, statistics, social media, physical objects, etc.) can be used and abused; statistics (and quantitative methods more generally) cannot escape this reality. One goal in this class is to deeply internalize proper use (hence, PHANTOM and PANIC) and to see examples of statistical abuse. Hopefully, this *complete* picture will allow you to better use and assess the statistics you encounter daily.

61. In Super Mario Maker 1 (SMM1), users create their own levels and upload them to the cloud for others to play. Each level is assigned a difficulty level and a tag so players will know what type of level it is.

    (a) Load the file SMMLevelsHW.csv (from the HW Data Sets folder on Convas/TritonEd) into R using the "Import Dataset" button in the upper-right window of R Studio. Then, explore the difficulty and tag columns. For each, state its data type (nominal, ordinal, discrete, or continuous) and its support.

    (b) What is the median clear rate for levels with the Super Expert difficulty?

    (c) *Among tagged levels rated with an Expert or Super Expert difficulty*, you might wonder if the percentage of Expert **Puzzle** levels is the same as the percentage of Super Expert **Puzzle** levels or not. If you're told that the data set is a random sample of all the levels in the game, conduct a hypothesis test with $\alpha = 0.05$ using prop.test to answer this inquiry. (Note: prop.test uses a different test stat than we do (so ignore that), but arrives at the correct p-value.) [PHANOM of PHANTOM]

Hint List:

- If you're having trouble with the coding side of things, try searching these R commands: table, summary, is.na, the ! operator, and the & operator.

Solution:

(a) The table command is useful here because we have categorical data:

```
table(data$difficulty)
table(data$tag)
```

We see difficulty is $\boxed{\text{ordinal}}$ with support $\boxed{\{\text{Easy, Normal, Expert, Super Expert}\}}$. The tag variable is $\boxed{\text{nominal}}$ with support $\boxed{\{\text{Automatic, Autoscroll}, \ldots, \text{Yoshi}\}}$.

(b) We find a median clear rate of $\boxed{0.68\%}$ using this code:

```
seLevels = subset(data, difficulty == "Super Expert")
summary(seLevels$clear_rate)
```

(c) To run our HT, we need to first find how many tagged Expert levels there are, and how many of those have the "Puzzle" tag. Then we repeat for Super Expert levels:

```
taggedExpert = subset(data, !is.na(tag) &
                              difficulty == "Expert")
print(nrow(taggedExpert))
sum(taggedExpert$tag == "Puzzle")

taggedSuperExpert = subset(data, !is.na(tag) &
                                 difficulty == "Super Expert")
print(nrow(taggedSuperExpert))
sum(taggedSuperExpert$tag == "Puzzle")
```

We find 1415 tagged Expert levels, 218 of which are Puzzle levels. We also find 901 tagged Super Expert levels, 121 of which are Puzzle levels. Now, (P) let p_{E} (p_{SE}) be the percentage of **all** tagged (Super) Expert levels in SMM1 that have the Puzzle tag.

(H) We test H_0: $p_{\text{E}} = p_{\text{SE}}$ vs. H_1: $p_{\text{E}} \neq p_{\text{SE}}$ ("the same ... or not").

To run our (N) two-sample z-test we (A) assume H_0 and check:

- Independent populations: Yes, we are exploring Expert and Super Expert levels which have no overlap.
- Within each sample, we need:
  - Independent data: Yes, because the data set is a random subset of SMM1 levels and there are millions of SMM1 levels.
  - At least 10 observed successes and failures: Yes, by the above numbers.

We use R to finish the test:

```
prop.test(c(218,121), c(1415,901), correct = F)
```

(O) This gives a P-value of $0.1895 > 0.05$, so (M) we keep H_0. It appears the percentage of Puzzle levels in similar among all tagged Expert and Super Expert levels in SMM1.

Commentary: I wanted this problem to give you the *feel* of being a data scientist/statistician. People who do this work are constantly moving between exploration (part a), communication/summary (part b), and testing claims (part c). During this movement, they might do work on paper, write code, discuss findings with others, create reports, etc. Furthermore, this work is often in a domain you have no experience with, which is why I chose an older video game only available on a single platform (Nintendo). In one month as a data scientist, you might have to research domains as diverse as economic indicators, golf ball aerodynamics, and food additives in commercial products. Some data scientists work in companies that focus on particular fields (e.g., pharmaceuticals), so the diversity of what they encounter is less, but still significant.

Another consideration in this problem arises in part a. What should you do if NA (Not Applicable/Available) appears in a data set (e.g., for the tag variable)? Is NA part of the support or simply an indicator that we are missing data? Well, this depends on how the variable was conceptualized in the first place. Suppose you design a survey for college students and ask how they feel about their major giving these options: Very positive, Somewhat positive, Somewhat negative, Very negative. You intend for the support to be one of these four possibilities, and if someone skips the question, you might use NA to show this in your data set. If, however, you create a question that reads: "How do you feel about being a **STEM** major?", then you might list the same four options and include a "Not applicable" option for the non-STEM students taking the survey. Here, NA is part of the support and indicates a non-STEM major, rather than a person who skipped the question. In part a, I would not list NA as part of the support because Nintendo did not list this as one of the tags to choose from. If I were at Nintendo, I would have included an Other tag for levels that didn't fit nicely into any of Nintendo's tags. Given this context, the NA designation might indicate a person who mindlessly skipped the tagging process, or a person who really wanted to tag their level but felt none of the labels was appropriate, and ultimately, didn't tag their level. Finally, it is important to note that the support is the *theoretical* list of all possible values as conceptualized by the data collector. So, asking the data set what values are present (as we do in part a) only gives those values from the support that happen to appear in the data set. In this problem, the data set is large enough to contain all values from the support, but this may not always happen.

62. Recently, I heard this claim on NPR: "Coronavirus hospitalizations in the US disproportionately affect communities of color". Let's explore this using different analyses.

(a) One approach to this issue is to see if US "people of color" (POC, which often includes African American, Asian American, Latinx, Native American, and additional communities) make up a greater percentage of hospitalizations than you would expect given their overall percentage in the population (39.9% based on a recent US Census). Suppose you call up random hospitals around the US and collect their data on race for COVID-19 hospitalizations, finding that 150 of 320 patients self-identify as POC. What is the smallest significance level for which NPR's claim would be reasonable? [PHNTO of PHANTOM, Tech not tables]

(b) One problem with the above approach is that the number 39.9% might be outdated, and hence, you might want to draw an additional sample (of the entire US populace) to find the proportion of Americans who self-identify as POC. Suppose you reach out to 1009 random Americans and find x identify as POC. What values of x would lead to a hypothesis test that rejects H_0 using $\alpha = 0.05$? Use a calculator or computer to solve the nasty mathematics that arises. Assume the 320 patients in part a are an independent sample from the new sample and that we will pursue a one-sided alternative similar in feel to part a. [PHANT of PHANTOM]

(c) A final approach for exploring the NPR claim is to look at several racial categories of interest and see if their percentages in the general populace are close to patient counts at hospitals. Suppose we explore the four groups: White (Non-Hispanic), Latinx, African American, and Multiple/Other (these are similar to labels currently used by the CDC), and that these groups make up 60.1%, 18.5%, 13.4%, and 8% of the American populace respectively. In a random sample of 840 COVID-19 hospitalizations, you find these self-identifications: 445 White, 199 Latinx, 116 African American, and 80 Multiple/Other. Conduct a hypothesis test with significance level 0.01 to see if the population percentages are a good fit for the COVID-19 hospitalization percentages. After doing the problem by hand, write a single line of R code that reproduces your results. [PHANTOM, Tech not tables]

(d) Explain why the approach taken in part c isn't the best way to support or refute the NPR claim.

For those interested in language related to race, I recently read this article on the term "Latinx".

Hint List:

- a: For the test to support NPR's claim, we need the P-value $< \alpha$. If you find the P-value, the minimal α follows immediately.
- b: First, work toward a pooled proportion of $\frac{150+x}{1329}$. Then, get a very messy expression for the z-stat and use technology to solve an inequality involving the z-stat.
- c: Run a chi-squared test with four categories.
- d: If you try to write the alternative hypothesis from part c in symbols rather than words, you can better see the inadequacy of part c vis-a-vis NPR's claim.

Solution:

(a) (P) Let p_{H} be the percentage of US coronavirus hospitalizations who identify as POC.

(H) We test H_0: $p_{\text{H}} = 39.9\%$ vs. H_1: $p_{\text{H}} > 39.9\%$ ("greater percentage", also see the commentary).

We run a (N) one-sample z-test with $\widehat{p} = \frac{150}{320}$ getting (T) $z = \frac{\frac{150}{320} - 0.399}{\sqrt{0.399 \cdot 0.601/320}} \approx 2.547977$.

So, our (O) P-value is pnorm(2.547977, lower = F) ≈ 0.005417. NPR's claim is reasonable if we move to H_1, which happens if $0.005417 < \alpha$. So, the smallest α is about $0.005417 \approx \boxed{0.005}$.

(b) (P) We focus on p_{H}, as above, but also p_{US}, the percentage of US residents who identify as POC.

(H) We test H_0: $p_{\text{H}} = p_{\text{US}}$ vs. H_1: $p_{\text{H}} > p_{\text{US}}$.

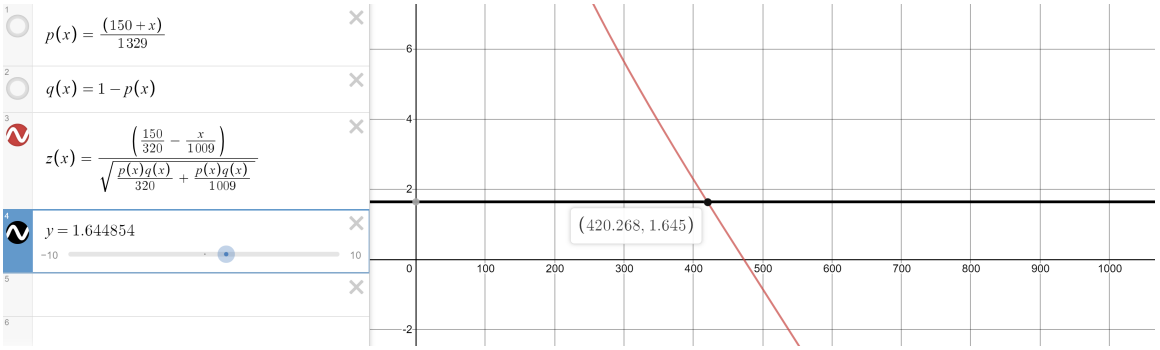
(A) Assuming H_0, we must check:

- Independent populations: Yes, since our populations are different ideas. One is about hospitalizations, the other is about US residents.
- In each sample, we need:
  - Independent data: Yes, via random sampling and $< 10\%$ criterion.
  - At least 10 observed successes and failures: Yes, for hospitalizations by the given data. To get this for US residents, we force $x \geq 10$ and $1009 - x \geq 10$ below.

Our (N) two-sample z-test uses a pooled proportion $p_* = \frac{150+x}{320+1009} = \frac{150+x}{1329}$. Thus, $q_* = \frac{1179-x}{1329}$. Our test stat (T) is:

$$z = \frac{\frac{150}{320} - \frac{x}{1009} - 0}{\sqrt{\frac{p_* q_*}{320} + \frac{p_* q_*}{1109}}}$$

To reject H_0 and support NPR's claim, we need z to land in the upper tail of $Z \sim N(0,1)$. Using R, we find that qnorm(0.05, lower = F) ≈ 1.644854 cuts off the top 0.05 area, so we set $z \geq 1.644854$. This is a perfect setting for the productive use of technology. Using a graphing calculator or Desmos, we solve the inequality:



From the image, we find $x \leq 420.268$. Given that x is an integer and we are forced to have $x \geq 10$ from our test assumptions, we get $\boxed{10 \leq x \leq 420}$. If you don't trust your answer, technology can check it:

```
prop.test(c(150, 420), c(320, 1009), alt = "greater", correct = F)
prop.test(c(150, 421), c(320, 1009), alt = "greater", correct = F)
```

The P-values from these tests are 0.04913 and 0.05243 respectively, showing we did correctly find the x value that serves as the upper bound.

(c) (P) Let p_W, p_L, p_{AA}, and p_O be the percentage of all COVID-19 hospitalizations who identify as White, Latinx, African American, and Multiple/Other respectively. We test if these proportions are a good fit with representation in the larger US populace (H):

H_0: $p_W = 0.601, p_L = 0.185, p_{AA} = 0.134, p_O = 0.08$ vs. H_1: Some proportion is different.

To run a (N) chi-squared goodness-of-fit test, we (A) assume H_0 and check:

- Counts: Yes, counts are present in the problem.
- Independent data: Yes, we used a random sample and there are more than $840 \cdot 10 = 8400$ COVID hospitalizations.
- At least 5 expected counts per category: Yes, our rarest category should have $840 \cdot p_O = 840 \cdot 0.08 = 67.2 \geq 5$.

Our (T) test stat is:

$$\chi^2 = \frac{(445 - 840 \cdot 0.601)^2}{840 \cdot 0.601} + \cdots + \frac{(80 - 67.2)^2}{67.2} \approx 21.869$$

We look this number up on $\chi^2_{4-1} = \chi^2_3$ via pchisq(21.869, df = 3, lower = F) $\approx 6.95 \cdot 10^{-5}$ (O). Since our P-value < 0.01, we (M) reject H_0 in favor of H_1. It does appear something is different in the racial makeup of COVID hospitalizations and the US populace. We can repeat this analysis in one line of code:

```
chisq.test(c(445, 199, 116, 80), p = c(0.601, 0.185, 0.134, 0.08))
```

(d) If we move to H_1 in part c, we only know *something* is off with the percentages, but we don't know which percentages are off and whether they are too high or too low. Since NPR's claim is essentially a one-sided hypothesis (that POC hospitalizations are *too high*), you shouldn't run a test with a two-sided feel. Note that H_1 for part c could be written

$$p_W \neq 0.601 \text{ or } p_L \neq 0.185 \text{ or } p_{AA} \neq 0.134 \text{ or } p_O \neq 0.08$$

Commentary: The primary goal of this problem was to reveal that there are many ways to analyze a given situation. Indeed, the more tests you know, the more flexibility you have in analysis. This is a double-edged sword, however, for you must be careful not to conduct test after test until you get the outcome you desire. Such behavior is known as a "questionable research practice" (QRP), an action which, while not a clear example of research misconduct, erodes the essence of scientific inquiry.

As you worked parts a and b, my hope is that you saw these were examples of manipulating the knowns and unknowns in our standard HT frameworks for one- and two-sample z-tests. In contrast, part c was a standard chi-squared test; I made no alterations since this was your first usage of this framework.

Finally, I wanted to underscore the vagueness of language. In part a, notice that when we speak of a negative idea (like hospitalization), a group is "disproportionately represented" if its percentage is too high. For positive ideas (like college attendance), the same phrase means that the group's representation is too low! This is just the tip of the language disaster iceberg. When I listen to popular media, I hear an endless stream of vague statements (e.g., "The economy is improving", "Support for the president is waning") based on unspoken or abstruse metrics (e.g., the "core personal consumption expenditure index") built upon uncited data sources (e.g., a "recent study").

63. As of March 2023, Mario Kart 8 (MK8) Deluxe was the best-selling Nintendo Switch game. When playing MK8, you race around in karts and compete with other human players and computer AI to finish the fastest. Along the way, you hit item boxes that randomly decide what item you will be awarded to help tilt the race in your favor. If you've played, you'll know that you get crummy items while leading, and amazing items when in last place. Let's explore this setting.

(a) The image below (source) shows the weights given to each possible item for a human-controlled player currently in first place in Grand Prix mode. Explain how to use these weights to find the *probability* that you'll get a coin from an item box while in first place. (Any item without a weight is impossible to get, and you can ignore it.)



(b) Write some R code that finds the probabilities for each of the possible items (while in first place) using the weights in the image. (Check that your coin probability from part a is correct.)

(c) Suppose you got a Japanese version of the game and were curious if the first-place item box percentages were the same as the US version summarized by the image (assume the *possible* items are identical). To find out, you are "forced" to play MK8 for hours while your friend records the items you get from all item boxes while in first place. What is the fewest number of first-place item box results your friend should record to meet the assumptions of running a goodness-of-fit test?

(d) Run a goodness-of-fit test using R using these counts from your friend: 136 banana peels, 107 green shells, 9 red shells, 11 mushrooms, 10 super horns, 153 coins. [PHOM of PHANTOM]

(e) When you're in last place and really far behind, you get great items based on these weights:



As long as you remain in last place, how many item boxes will you need to collect, on average, before you finally get a blue shell? [WISE]

(f) How can you use the last-place item weights to code what item a player gets when an item box is collected? One approach is to use $Unif(0,1)$ and set up a collection of "cutpoints": $0 = c_1 < c_2 < c_3 < c_4 < c_5 < c_6 < c_7 = 1$. Then, you generate a random X from $Unif(0,1)$. If $c_1 \leq X < c_2$, the player gets a blue shell. If $c_2 \leq X < c_3$, the player gets the triple-mushroom, and so on, in order. The important feature of the cutpoints is that they get the probabilities correct. So, $P(c_1 \leq X < c_2)$ must equal P(getting a blue shell), and similarly for the other items. Based on this setup, what cutpoints will demarcate the star powerup?

Hint List:

- a: Recall how we used weights to find probabilities in the Zelda arrow-barrel problem from earlier in the quarter.
- b: This only needs a few lines of code if you use the vectorized nature of R.
- c: Recall the A from a goodness-of-fit test.
- d: Can you write the null and alternative hypotheses using words and symbols? Why is our P-value sensible?
- e: Think back to the discrete RVs we studied.
- f: Can you figure out that $c_2 = 0.025$? Now, work from there.

Solution:

(a) Note that the weights sum to 200 $(65 + 50 + 5 + 5 + 5 + 70)$. So, the image conveys that out of every 200 items awarded, 70 of these will be coins. Thus, $P(\text{coin}) = \dfrac{70}{200} = \boxed{0.35}$.

(b) The vectorized nature of R helps:

```
weights = c(65, 50, 5, 5, 5, 70)
probs = weights/sum(weights)
print(probs)
```

(c) We certainly have counts and independent data (because Mario Kart 8 randomly decides on items, like most video games). We just need to ensure $E_i \geq 5$ for all i. Our rarest items have weight 5, or probability 0.025 from part b. So, we need $n \cdot 0.025 \geq 5$, or $n \geq 200$. Thus, $\boxed{n = 200}$ is the fewest number of first-place item boxes your friend should record.

(d) Let (P) $p_1, \ldots, p_6$ be the true percentages found in the game's code for our six different items while in first place in Mario Kart 8 on the Japanese version. (H) We test H_0: The US percentages are a good fit for the Japanese percentages $(p_1 = 0.325, \ldots, p_6 = 0.35)$ vs. H_1: They are not a good fit (some p_i is different). In R, we do (using probs from part b):

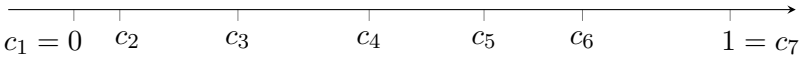
```
chisq.test(c(136, 107, 9, 11, 10, 153), p = probs)
```

(O) This gives a P-value of 0.9937, suggesting our data behave almost exactly as we would expect. So, (M) we keep H_0; it appears the two versions of the game share the same coding for this feature.

(e) Let (WIS) $X \sim Geom(p)$ be the number of item boxes to get a blue shell while in last place. Here $p = P(\text{blue shell}) = \dfrac{5}{5 + 35 + 35 + 60 + 10 + 55} = 0.025$. Thus (E),

$$E(X) = \frac{1}{p} = \frac{1}{0.025} = \boxed{40 \text{ item boxes}}.$$

(f) We draw this picture to help make sense of things:



Here, the zone $[c_1, c_2]$ demarcates the blue shell, $[c_2, c_3]$ is the triple mushroom, $[c_3, c_4]$ is the star, and so on. For a uniform distribution, $P(a \leq X \leq b) = \int_a^b 1\,dx = b - a$, the length of the interval. So, we choose cutpoints so that the length of each interval matches the probability. Thus, $c_2 = 0.025$ (since $c_1 = 0$). Using R or a calculator, we find $P(\text{triple mushroom}) = 0.175$ and $P(\text{star}) = 0.175$. So,

$$c_3 - c_2 = 0.175 \implies c_3 = 0.175 + 0.025 = 0.2$$

$$c_4 - c_3 = 0.175 \implies c_4 = 0.175 + 0.2 = 0.375$$

So, the star has cutoffs of $\boxed{0.2 \text{ and } 0.375}$.

Commentary: In this problem, I wanted to show just how useful statistics can be in understanding just about any setting. The weights given in parts a and e are believed to be the weights present in Nintendo's code, and our hypothesis test in part d supports these values. Part e is a question that many MK8 players have asked themselves while lagging in last place, although few people know the machinery of the Geometric distribution to answer their own question (or frame it using this language). Finally, note that part d is a nice example of a setting in which you, as the researcher, might be hoping to end up at the *null hypothesis* after running the test. It's very unlikely developers would change the foundational probabilities for items from one country to the next, so we expect the null to make the most sense. You should not think that researchers are *always* rooting for the alternative hypothesis. This might be true in vaccine trials (where you want to outperform a placebo), but there are many settings where the null hypothesis is desired. For example, in our COVID hospitalization percentage problem, you probably prefer the null (hospitalizations match population percentages) rather than the alternative (POC are suffering more from the disease). As another example: If you've designed a new manufacturing process, you might hope the rate of defects under the new process is the same as before (not higher, which the alternative might propose).

These examples should not just reveal that the null is desirable at times, but more generally, they should serve as an intellectual warning. Stop trying to create "always" or "never" guideposts in your reasoning patterns. Black-and-white rules might serve you fairly well in high school and below, but most advanced reasoning is about nuance, or shades of gray. For example, Calculus students will often ask me for a flowchart of reasoning that will solve *any* integral. Such a set of rules does not exist, and I do not believe turning students into if-then machines is a good idea. A better approach to creating scholars is to alert them to strong patterns that exist (e.g., the alternative hypothesis *often* encodes the desires of the researcher), but to present them with situations that reveal the boundaries of these patterns (e.g., the desirability of the null hypothesis in part d). You must maintain your flexibility as a thinker; do not let academia rigidify your mental pathways into binary, black-and-white rule systems.

64. Stats, Teacher Edition! Suppose you're a stat prof and want to create a goodness-of-fit problem for your class. Here's the setup: A person recently bought a 5-sided die (labeled $1-5$) that is supposedly weighted so that faces 2, 3, and 4 come up equally often, but face 1 comes up half as often as face 2. Face 5 comes up twice as often as face 2. You'll have the person roll the die 550 times to explore these weightings, and you plan to give your students these observed counts (the blank cell for face 5 is not a typo):

Face	1	2	3	4	5
Num rolls	45	105	100	x	

You want your students to run a goodness-of-fit test with the claimed die weights that results in keeping H_0 with $\alpha = 0.02$. What are all the possible values x could be? Do not use trial-and-error. Use a χ^2-table and basic calculator, not R/fancy tech. [PHNT of PHANTOM]

Hint List:

- To find the blank cell, ask an easier question: "If I flip a coin 100 times and get 30 heads, how many tails did I get?" Apply the same logic to our setting.
- If we keep H_0, what inequality is true about the P-value from our problem? Can you now write an inequality involving the test statistic?

Solution: (P) Let $p_1, \ldots, p_5$ be the true probabilities of the five faces. Based on the problem statement, the weights of these faces have the ratio $1:2:2:2:4$. We can get from weights to probabilities using the typical approach: $p_1 = \dfrac{1}{1+2+2+2+4} = \dfrac{1}{11}$, and so on.

Thus, (H) we test H_0: $p_1 = \dfrac{1}{11}, p_2 = p_3 = p_4 = \dfrac{2}{11}, p_5 = \dfrac{4}{11}$ vs. H_1: Some p_i is off.

We assume H_0 and run a (N) chi-squared goodness-of-fit test. Note that $E_1 = 550 \cdot \dfrac{1}{11} = 50, E_2 = E3 = E_4 = 100$, and $E_5 = 200$. So (T),

$$\chi^2 = \frac{(45-50)^2}{50} + \frac{(105-100)^2}{100} + \frac{(100-100)^2}{100} + \frac{(x-100)^2}{100} + \frac{(300-x-200)^2}{200}$$

Here, we used the fact that the face counts must sum to 550, so

$$45 + 105 + 100 + x + (\text{Face 5 count}) = 550 \implies \text{Face 5 count} = 300 - x$$

Thus, $\chi^2 = \dfrac{1}{2} + \dfrac{1}{4} + 0 + \dfrac{2}{2} \cdot \dfrac{(x-100)^2}{100} + \dfrac{(100-x)^2}{200} = \dfrac{3}{4} + \dfrac{3(x-100)^2}{200}$, since $(a-b)^2 = (b-a)^2$.
Based on the χ^2-table for $\chi^2_{5-1} = \chi^2_4$, the cutoff for the area 0.02 is 11.67. To keep H_0, we just need

$$\chi^2 < 11.67 \implies \frac{3(x-100)^2}{200} < 10.92 \implies (x-100)^2 < 728$$

Note that $(x-100)^2 = 728$ when $x - 100 = \pm\sqrt{728}$, or $x = 100 \pm \sqrt{728} \approx 73.019, 126.982$. Since we have an upward-opening quadratic, it will be below 728 between these values, and the integral nature of x forces $\boxed{x \in \{74, 75, \ldots, 126\}}$.

Commentary: By now, I hope you see this problem as one in which the knowns/unknowns have switched. Usually, the counts are known and we work forward to an unknown P-value. Here, the P-value is known (more precisely, since H_0 is kept, we have an inequality constraint on the P-value) and the counts are unknown (more precisely, one of the counts, x, is unknown). Because of this switch, we partly work forward from the counts to the test stat, and partly backward from the P-value inequality to a test-stat inequality.

Also, even though the problem prohibits tech during the solving process, we can use it to check our answer. Below, we explore what P-value we get for some xs just inside the correct range (74 and 126) and just outside (73 and 127). If we have done the problem correctly, the P-value should hop over the $\alpha = 0.02$ threshold as we leave the correct zone:

```
probs = c(1/11, 2/11, 2/11, 2/11, 4/11)
for (x in c(73, 74, 126, 127)) {
    counts = c(45, 105, 100, x, 300-x)
```

```
    test = chisq.test(x = counts, p = probs)
    print(paste(x, test$p.value))
}

# output
73 0.0198540584253909
74 0.0278284788491952
126 0.0278284788491952
127 0.0198540584253909
```

65. In music, perfect pitch is the ability to hear a note and state its pitch (e.g., "That's a C sharp!"). My cello teacher is helping me develop this skill via an activity: Each day, at random times, I have to get out my phone and sing the note A (which is what string players tune off of). The Soundcorset app tells me how far above or below an A my singing is in a unit called "cents" (100 cents equals a half step). Here is the amount I was off by (in cents) in thirteen random recent trials:

$$-103, -115, -85, -97, -80, -90, -100, 26, -18, -6, -70, -94, -5$$

On average, can I pull the pitch for an A out of thin air?

(a) Conduct a hypothesis test using a significance level of 0.05. After doing so, write a single line of R code that checks your analysis. [PHANTOM, VIZ, Tech not tables]

(b) Create an 80% CI (with units) for the parameter of interest. Again, check your answer with a single line of R code. [PNIC of PANIC, Tech not tables]

(c) How does the confidence level in part b have to change if we want a CI that is twice as wide? Assume other aspects of the problem remain the same. [VIZ, Tech not tables]

Hint List:

- a: For the CLT balance condition, try visualizing the data.
- b: Use R to get the summary info you need.
- c: What part of the CI formula dictates the width of the CI?

Solution:

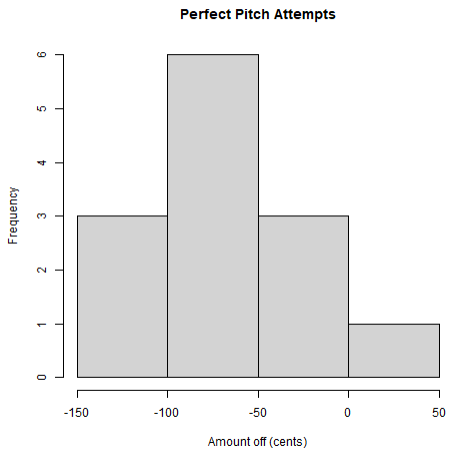
(a) (P) Let μ_A be the true average amount I am off by (in cents) when singing an A (440 Hz) with no musical reference points.

(H) We test H_0: $\mu_A = 0$ vs. H_1: $\mu_A \neq 0$. Here, we use a two-sided alternative because before any data are collected, it is not clear if I will be too high ($\mu_A > 0$) or too low ($\mu_A < 0$), on average.

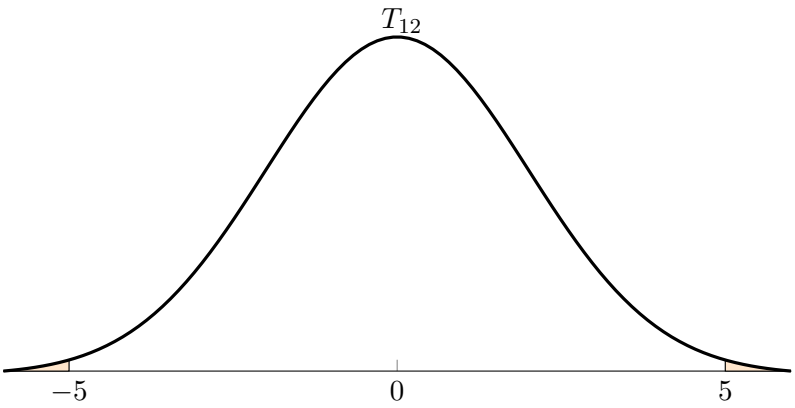
Assuming H_0, we run a (N) one-sample t-test after checking (A):

- Independent data: Yes, we have random attempts at singing an A, and I could have made more than $13 \cdot 10 = 130$ attempts.
- CLT balance condition: Using the below R code, we can make a histogram of our sample. Since our data are independent, this also gives a sense of what our population (all attempts to sing an A) looks like. With a low-skew sample/population and a sample size of 13, we probably meet the CLT balance criterion.

```
data = c(-103,-115,-85,-97,-80,-90,-100,26,-18,-6,-70,-94,-5)
hist(data, xlab = "Amount off (cents)",
     main = "Perfect Pitch Attempts")
mean(data)    # I get -64.38462
sd(data)      # I get 46.43551
```



The above code helps us find our (T) test stat of $t = \dfrac{\overline{x} - \mu_A}{s/\sqrt{n}} = \dfrac{-64.38462 - 0}{46.43551/\sqrt{13}} \approx -5.0$. This value lives on $T_{n-1} = T_{12}$, and because of our two-sided alternative, we label both our test stat and its negative, shading both tails of the distribution:



(O) Our P-value is $2 \cdot \text{pt}(-5, \text{df} = 12) \approx 0.000309 < 0.05$. So, (M) we reject H_0 in favor of H_1; it seems my average A pitch is not accurate yet. We check parts of this reasoning using R:

```
t.test(data)  # gives the same test stat and P-value as above
```

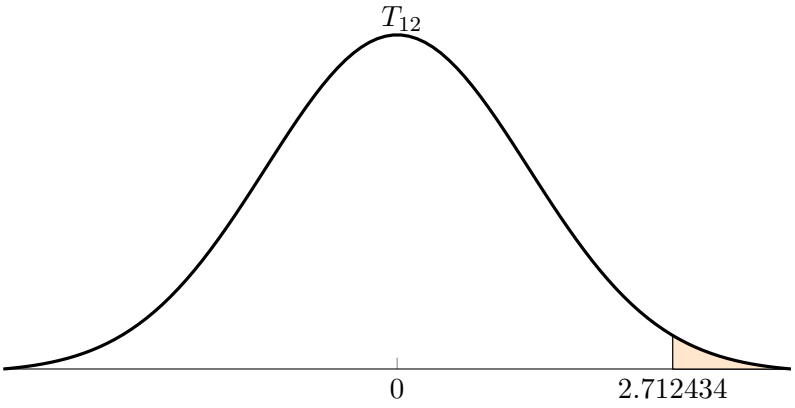
(b) For 80% confidence, $\alpha = 0.2$, so we need $t_{0.1,12} = \text{qt}(0.1, \text{df} = 12, \text{lower} = \text{F}) \approx 1.356217$.
(P) Our parameter is as above and we create a (N) one-sample t-interval. We have (I),

$$\overline{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} = -64.38462 \pm 1.356217 \cdot \frac{46.43551}{\sqrt{13}} \approx [-81.851 \text{ cents}, -46.918 \text{ cents}]$$

(C) **If** $\mu_A \in$ our interval, it means, on average, I am between 46.918 and 81.851 cents below the true pitch for an A. We can check this in R by changing the confidence level to 80%:

```
t.test(data, conf.level = 0.8)  # gives the same CI as above
```

(c) Our CI width is $2 \cdot t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$. If n and s are unchanged, we need $t_{\alpha/2,n-1}$ to be double in order to get an interval that is twice as wide. So, we solve $t_{\alpha/2,12} \approx 2 \cdot 1.356217 \approx 2.712434$. We have the below picture where the shaded area is $\alpha/2$:



In R, we find $\text{pt}(2.712434, \text{df} = 12, \text{lower} = \text{F}) \approx 0.009435$, so $\alpha \approx 2 \cdot 0.009435 \approx 0.0188$. Thus, our CL is $(1 - \alpha) \cdot 100\% \approx \boxed{98.113\%}$.

Commentary: There are two important takeaways from this problem. The first is that once you have the toolkit of statistics, you have the ability to collect and analyze data about yourself. Without your

knowledge from this class, you might not know how to summarize, display, build CIs, or run HTs on your personal data. As such, any work you would do would be made up on the fly. It is better to have an organized approach to the study of data. Statistics offers this, and it allows you to account for the random variability that arises in data collection when checking for deeper patterns that are present. You may have noticed that the collection and analysis of personal data has become a huge industry recently. You regularly see people wearing devices like the Apple Watch or Whoop in order to better understand themselves.

The second takeaway comes in part a. At the end of the hypothesis test, you land at H_1: $\mu_A \neq 0$. Because we settle on this, you see me writing "my average A pitch is *not accurate* yet", suggesting the average might be higher or lower. This conclusion is frustrating to some when viewed against the value $\overline{x} = -64.38462$, suggesting I'm *below* the pitch on average. Indeed, you might be tempted to conclude that $\mu_A < 0$. This, however, would be a mistake. Remember that the hypotheses are established *before* any data are collected, and they reflect the two directions you want to contrast once data are collected. As such, you must end up at one of those two options (and not a third option you create because of what the data are hinting at!). If, after seeing the data in this study, you would like to set up the new hypotheses H_0: $\mu_A = 0$ vs. H_1: $\mu_A < 0$, then you are welcome to do so. But, you must collect a new data set to explore these hypotheses because the last data set was collected to explore a different set of hypotheses. If you incorrectly use the data set from part a on these new hypotheses, then (unsurprisingly!) you'll end up at H_1: $\mu_A < 0$. Notice that this practice would use the data to inform the hypotheses which are then confirmed by the same data. This is not how science should work. Indeed, if you keep forming hypotheses based on a single data set, then all you do is create a belief structure about the world that is deeply tied to the particulars of that data set. While every data set contains some signal (which we want to capture), every data set also contains some noise (which we don't want to capture). Creating hypotheses from a data set and then using that same data set to corroborate your hypotheses creates a situation in which you overfit your beliefs about the world to the signal **and** noise of a data set.

66. One of my favorite podcasts is The Happiness Lab with Dr. Laurie Santos. She teaches one of Yale's most popular courses, "Psychology and the Good Life". As a happiness researcher, she has recently made podcasts focused on improving the holiday season. This got me thinking: For those who celebrate Christmas, do Christmas presents increase happiness levels *after* the excitement of Christmas day? To find out, you could imagine this study: We recruit 45 random Americans who celebrate Christmas and ask them to rate their happiness on a 1 to 10 scale at two separate times: Five days before xmas, and five days after xmas. Suppose our results are summarized below:

	Before xmas	After xmas	Difference of after − before
Average Happiness Rating	7.8	8.22	0.42
SD	1.2	1.6	1.44
n	45	45	45

(a) Conduct a hypothesis test for this setting using $\alpha = 0.01$. [PHANTOM, Tables not tech]

(b) Suppose you wanted to fake your data so that you get a statistically significant result using $\alpha = 0.01$. What is the smallest value you could choose for the number 8.22 (hence changing 0.42) to achieve your goal? (Assume you leave the other values unchanged.) [Tables not tech]

(c) Explain why the number 1.44 **cannot** be found from the values 1.2 and 1.6, while the value 0.42 **can** be found from the values 7.8 and 8.22.

(d) If you get a statistically significant result from this study, it is tempting to claim that "For Christmas celebrators, presents raise happiness levels many days beyond Christmas." Give an alternate reason for this apparent increase that has nothing to do with presents. (Every time I ask you a question like this, it should serve as a reminder that any observed difference may be the result of a theory you weren't originally considering!)

Hint List:

- a: Remember that paired data are analyzed using the one-sample framework from before. The only change is that our data represent differences.
- b: What test stat would result in a P-value of exactly 0.01?
- c: Try thinking about this problem from the perspective of random variables, rather than a data set. For example, let X be the happiness score of a random person 5 days before Christmas. Notice that the number 7.8 is just an approximation of $E[X]$. Rewrite the table using RV expressions and work from there.
- d: Why do you feel better *after* a major holiday happens?

Solution:

(a) (P) Let μ_d be the average happiness difference (after − before) of **all** people's self-reports 5 days after Christmas compared to 5 days before Christmas.

(H) We test H_0: $\mu_d = 0$ vs. H_1: $\mu_d > 0$ ("presents increase happiness levels").

Assuming H_0, we run a (N) paired t-test (i.e., one-sample t-test on differences) after checking (A):

- Differences are independent: Yes, the differences come from different people, who were randomly chosen, and more than $45 \cdot 10 = 450$ people celebrate Christmas.
- CLT balance condition: Most likely met. We only have a summary of the data here, but usually $n = 45$ is large enough to balance out any reasonable skew in the population distribution (i.e., happiness difference scores).

Our (T) test stat is $t = \dfrac{\overline{d} - \mu_d}{s_d/\sqrt{n}} = \dfrac{0.42 - 0}{1.44/\sqrt{45}} \approx 1.95656$. This value lives on $T_{n-1} = T_{44}$. Using a T-table, we see $1.68 < 1.95656 < 2.02$ in the row for df $= 44$, so (O) our P-value is between 0.025 and 0.05. Since this range is above 0.01, we (M) keep H_0; it is possible that happiness levels, on average, are unchanged five days before and after Christmas.

(b) Based on a T-table, $t_{0.01,44} \approx 2.41$. If our test stat lands above this number, our P-value would be smaller than 0.01, and we would have a statistically significant result. So, we set

$$2.41 < \frac{\overline{d} - \mu_d}{s_d/\sqrt{n}} = \frac{\overline{d} - 0}{1.44/\sqrt{45}} \implies \overline{d} > 0.517$$

Since $\overline{d}$ = (average 5 days after) − (average 5 days before), we see

$$\text{(average 5 days after)} > 0.517 + \text{(average 5 days before)} = 0.517 + 7.8 = 8.317$$

If we are choosing the minimal possible value, we use $\boxed{8.317}$.

(c) While the given table focuses on data from a particular sample, it is easier to think about this problem using the language of random variables. Let X be the happiness score of a random person 5 days before Christmas. Let Y be the happiness score of *that same person* 5 days after Christmas. Notice that the numbers in our table are approximations of these expressions:

	Before xmas	After xmas	Difference of after − before
Average Happiness Rating	$E[X]$	$E[Y]$	$E[Y-X]$
SD	$SD[X]$	$SD[Y]$	$SD[Y-X]$

We know form earlier in the course that $E[Y-X] = E[Y] - E[X]$, which shows how to get 0.42 using 7.8 and 8.22. Sadly, we can't get $SD[Y-X]$ from $SD[X]$ and $SD[Y]$. If we try, we run into a problem: First, recall that

$$Var[Y-X] = Var[Y] + Var[X], \text{ if } X \text{ and } Y \text{ are independent}$$

Taking the square root of both sides gives

$$SD[Y-X] = \sqrt{Var[Y] + Var[X]} = \sqrt{SD[Y]^2 + SD[X]^2}$$

This formula does relate $SD[Y-X]$ to $SD[Y]$ and $SD[X]$, but we don't have the independence between X and Y. Indeed, by using *paired* data, we have an important *dependency* built in. Here, if X is really high, you can guess that Y will be really high too. In later stat classes, you'll learn how to relate all these measures of spread when you have *dependent* variables, as is this case here. The formula uses a new idea we don't cover called covariance (Cov):

$$Var[Y-X] = Var[Y] + Var[X] - 2 \cdot Cov[X,Y]$$

(d) Perhaps people are relieved the holiday season is over, or that visiting relatives have left, or they are back to eating healthier, or etc. Any of these theories would also account for higher happiness scores after the holidays, so the Christmas presents theory may not be the reason.

Commentary: Part a is standard practice with PHANTOM in our new setting of paired data. I find that students mostly struggle with recognizing that the analysis happens on differences (which adds a layer of complexity in understanding). As such, in the A step, we are checking if the *differences* are independent, not the data that make up the differences (which are definitely dependent because of the pairing). Part b switches up the knowns and unknowns so you see the testing procedure from a new angle. Part c builds a connection to previous coursework with random variables.

To me, part d is the most important question, and one I never see addressed in textbooks or journal articles. In science, we might have some theory about how the world works (e.g., Christmas presents increase happiness). To explore our theory, we choose something quantitative and measurable that ties to our theory (e.g., self reports five days before/after Christmas). Notice that the statistical test is about the quantitative idea (μ_d), not the theory. At the end of the test, you decide which hypothesis to believe, and the hypotheses are stated using the quantitative idea. If, in this problem, we were to adopt H_1: $\mu_d > 0$, it is unbelievably tempting to think this conclusion supports or proves your theory (It doesn't!). As you see in the solution, there are many other theories that give rise to H_1: $\mu_d > 0$, and you'll never know which is responsible (if any of them!). Similarly, if you end up keeping H_0: $\mu_d = 0$, this does *not* mean your theory is false. Perhaps presents really do increase happiness, but Christmas brings lots of added stress which *offsets* this increase with an equal decrease, resulting in no change in happiness levels! The simple and disheartening takeaway is that linkages between our theories and our hypothesis tests are more tenuous than people are willing to admit.

67. Here's a question for those interested in legal reform: If a jury is asked to disregard something they have heard in a courtroom, can they? Recently, I was reading "Theory Construction and Model-Building Skills" when I came across an interesting reference to work by researcher Robert Cialdini. In one of his studies, randomly-selected jurors studied a case where a person was injured by a reckless driver. The jurors were to decide how much money to award the injured person. One group of jurors (Group A) simply heard a tape of the case and were not notified whether the reckless driver had liability insurance (money paid out by an insurance company to a person hurt in an accident). A different group of jurors (Group B) heard the same tape with one addition: they learned the reckless driver did have liability insurance, but then the judge in the case told jurors to ignore this fact, ruling the information was inadmissible. So, do people ignore information in the US legal system when they are told to do so by the court? Below are simulated data that closely follow the results of Cialdini's study.

	Group A Jurors	Group B Jurors
Average Payout to Injured Person	\$33000	\$46000
SD	\$3000	\$3200
n	24	27

(a) Based on the above data, conduct a hypothesis test with a two-sided alternative using $\alpha = 0.01$. [PHANTOM, Tables not tech]

(b) Build a 99% CI for this setting. [PNIC of PANIC, Tables not tech]

Hint List:

- a: Because our sample sizes are unequal, our data can't be paired.

Solution:

(a) (P) Let μ_A (μ_B) be the average amount awarded by all jurors who could be placed in Group A (B).

(H) We test H_0: $\mu_A = \mu_B$ vs. H_1: $\mu_A \neq \mu_B$ ("with a two-sided alternative").

Assuming H_0, we run a (N) two-sample t-test after checking (A):

- Independent populations: Yes, we have completely separate groups and no person will be assigned to both.
- In each sample, we need:
  - Independent data: Unsure. There is no discussion of how people were chosen for the groups.
  - CLT balance condition: Unsure. We only have summarized data here and $n_A = 24$ and $n_B = 27$ may not be large enough to counteract skew present in payout amounts from our groups.

Proceeding anyway (!), we find (T):

$$t = \frac{\overline{B} - \overline{A} - (\mu_B - \mu_A)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{46000 - 33000 - 0}{\sqrt{\frac{3000^2}{24} + \frac{3200^2}{27}}} \approx 14.97$$

This value lives on $T_{\min(24-1,27-1)} = T_{23}$. Using a T-table, we find $14.97 > 2.81$, so (O) our P-value is < 0.01. Thus, (M) we reject H_0 in favor of H_1; it appears our means are different. This might be the result of our theory (e.g., that people do not disregard info they are told to disregard) or not (see last question's commentary).

(b) We build an interval for (P) $\mu_B - \mu_A$ ($\mu_A - \mu_B$ order also OK). This (N) two-sample t-interval is:

$$\overline{B} - \overline{A} \pm t_{0.01/2,23} \cdot \sqrt{\frac{S_A^2}{n_A} + \frac{s_B^2}{n_B}} = (46000 - 33000) \pm 2.81 \cdot \sqrt{\frac{3000^2}{24} + \frac{3200^2}{27}}$$
$$\approx [\$10559.568, \$15440.432] \text{ (I)}$$

(C) **If** $\mu_B - \mu_A$ is in our CI, this suggests that group B jurors award, on average, between about \$10560 and \$15440 more than group A jurors. This may suggest jurors were attending to the info they were told to ignore (or another theory might explain this).

Commentary: This problem is a standard two-sample setup. Often, you will be given data summarized in a table, and this problem reveals that to run a test or build a CI, we only need the six given numbers. This is pretty impressive! To understand two populations (payout amounts for Group A and B jurors) which might be incredibly large, we can collect some small samples and then reduce these even more to six summary values!

With this reduction, however, comes frustration (recall that all life is trade-offs). In the A of PHANTOM in part a, you hopefully felt frustration over how little you actually knew about the data sets underpinning our six values. I get this feeling all the time when reading journal articles: I see nice summary tables like the one given in this problem but have no sense of the *quality* of the data being summarized. When you don't have a clear sense for your sample(s), you can't trust the inferential step of drawing conclusions at the population level. For example, if someone says that in a sample of 1000 people, they found that 42% support politician X, this is too vague to be meaningful. If that sample was a random sample of people who will actually vote, then great. If, however, the sample is racially homogeneous, or just contained college students, or only involved people who would answer a cell phone call from an unknown number, etc., then inference is completely inappropriate. You can't determine any of this from a summary table.

Finally, remember to look deeply at what is in front of you. For the legal system to be well-founded, you need jurors to ignore things that they were not supposed to hear. As such, we were hoping to end up at the *null* hypothesis in this problem. The data, however, strongly suggest the alternative, and this means that jurors have trouble looking past information they are not supposed to consider. While this problem is about a single, unemotional bit of knowledge (whether the reckless driver had liability insurance), you can imagine the consequences for cases where multiple, highly provocative pieces of information are forced into the trial, only to be later deemed inadmissible.

68. You've now seen 6 different types of hypothesis tests: one-sample z-test, two-sample z-test, chi-squared goodness-of-fit test, one-sample t-test, paired t-test, and two-sample t-test. Each has its own hypothesis setup, assumptions to check, and test statistic. While these details can be challenging to remember, they are easy to look up if you forget. A more difficult task is to decide which test to deploy given a brief description of a research project. For each description below, decide which of the six tests is appropriate and write the hypotheses the problem suggests. If none is a good fit, write "none of the above" and skip the hypotheses. Do not worry about test assumptions.

(a) You want to determine whether students at Harvard or Princeton spend more time studying. You survey 50 randomly chosen students at each university and ask them how many hours they spent studying in the last week.

(b) You want to determine whether, in heterosexual marriages since the year 2000, the average age of the husband exceeds the average age of the wife by 3 years. For 175 randomly chosen heterosexual couples who were married since the year 2000, you obtain the husband's age and the wife's age.

(c) In a particular town, you want to study the number of girls present in families that have exactly three children. You are curious if the number of girls in such families follows a binomial distribution with $n = 3$ and $p = \frac{1}{2}$. You draw a small random sample and conduct your study.

(d) You're curious if the average IQ of Americans living in Texas, Ohio, Florida, and New York are all similar or not. You'll collect random samples from all four states and run a test to analyze all the averages at once.

(e) Out of 200 voters surveyed, 108 support withdrawing US troops from a particular foreign country. You want to determine whether a majority of all voters support withdrawing troops.

(f) You're curious if graduate students and undergraduates students will use the same number of lines of Python code, on average, to solve an introductory programming task. You recruit random students around the world and explore the data.

(g) You plan to explore the effect of advertising by seeing what percentage of Californians support a proposition before and after an extensive social media campaign. You'll reach out to 1000 random Californians before the campaign, and your colleague will handle the post-campaign calls with 1000 new Californians.

(h) You take samples of rain water from 40 different places in a forest, and record the pH level in each sample. You want to determine whether the average pH level of rain in the forest is below 5.6, which is considered to be the threshold for acid rain.

(i) You've read online that at most universities, the breakdown of students by year is 30% freshmen, 25% sophomores, 25% juniors, and 20% seniors. You're curious if this holds at UCSD, so you ask the registrar for information on 400 random current students.

(j) You're curious if students change their opinions about an instructor during a course. So, you pick random Math 183 students and see what portion approve of DQ in week 1 of the class. Then, you return to the same students in week 10 and again ask what portion approve.

Hint List:

- On all parts, it helps to first decide if the problem is about proportions (z-tests and chi-squared) or means (t-tests).

Solution:

(a) Two-sample t-test; H_0: $\mu_{\text{H}} = \mu_{\text{P}}$ vs. H_1: $\mu_{\text{H}} \neq \mu_{\text{P}}$

(b) Paired t-test; H_0: $\mu_d = 3$ vs. H_1: $\mu_d > 3$

(c) Chi-squared goodness-of-fit test; H_0: $p_i = \text{dbinom(i, 3, 1/2)}$ vs. H_1: Some p_i is different

(d) None; this problem is about means but involves more than two samples. We will soon learn this design is handled by a technique known as ANOVA (ANalysis Of VAriance).

(e) One-sample z-test; H_0: $p_{\text{withdraw}} = 0.5$ vs. H_1: $p_{\text{withdraw}} > 0.5$

(f) Two-sample t-test; H_0: $\mu_{\text{grad}} = \mu_{\text{undergrad}}$ vs. H_1: $\mu_{\text{grad}} \neq \mu_{\text{undergrad}}$

(g) Two-sample z-test; H_0: $p_{\text{pre}} = p_{\text{post}}$ vs. H_1: $p_{\text{pre}} \neq p_{\text{post}}$

(h) One-sample t-test; H_0: $\mu_{\text{pH}} = 5.6$ vs. H_1: $\mu_{\text{pH}} < 5.6$

(i) Chi-squared goodness-of-fit test; H_0: $p_{\text{F}} = 0.3, p_{\text{SO}} = 0.25, p_{\text{J}} = 0.25, p_{\text{SE}} = 0.2$ vs. H_1: Some p_i is different

(j) None; this should feel like paired two-sample *proportion* data. For means, we created differences to move to a one-sample design, but we can't do that here because the data look like Yeses and Nos. One approach to this design is called McNemar's Test.

Commentary: With the completion of this problem, you have your first pair of "experimental design" glasses. Experimental design is the study of how to structure experiments and the statistical tools that analyze different setups. As an example, when you want to study the measure of center in a single population, you use the one-sample t-test and focus on the mean in a population that is normally distributed (or, by the CLT, not normally distributed but still well-behaved enough that the average, $\overline{X}$, is well-behaved enough to run our test). Interestingly, there are other analytical choices for this setup with names like the "Sign Test" and "Wilcoxon Signed Rank Test". There are entire books/courses in experimental design for those who work with complex data setups, but in practice, most people are fine with the tests we cover in this class.

I describe your training as "a pair of glasses" because you will now see the world differently than before. In the past, if you heard about a poll that compared how much Democrats vs. Republicans support some piece of legislation, you likely gave no thought to the *statistical structure* of this survey, only the take-home message. Now, you see this as a two-sample z-test with the implicit null that $p_{\text{D}} = p_{\text{R}}$. Furthermore, you're likely only hearing about the poll because the null was ultimately rejected by the researchers, so something surprising is supposedly going on. Interestingly, all the studies where the null is kept, and Democrats and Republicans agree, are unlikely to make the news, inflating the apparent divide between the parties. This is known as the "file drawer problem" and is also present in academic publishing where unexciting results (usually, keeping the null hypothesis in a study) are unlikely to be printed, so scholars throw their "unpublishable" papers into a drawer.

69. On Canvas, you'll find the file DQPitch.csv which has the actual data of me working on my ability to sing an A440 pitch without any references (recall question 65). Now that you know more, we'll explore this topic in more depth by looking at three distinct time periods: trials from initial attempts (labeled "early" in DQPitch.csv), those a few months later ("later"), and finally, my most recent attempts ("latest"). For each of the below parts, do the PHNOM of PHANTOM *using R* to run an appropriate HT. Let $\alpha = 0.05$ and do not discussion assumptions.

    (a) My early attempts were "flat" (below the correct pitch) on average (which we discovered in problem 65). Did my flatness problem improve by the second round of attempts? Note: When you use t.test here, you should always set the parameter var.equal to F (which is the default, so you need not type it), and be forewarned that the df value listed in the print-out will not be what you are expecting (because R uses a more precise, but painful formula for finding df).

    (b) Perhaps my perfect pitch skills have not changed at all over the three time spans in question. Decide if this claim is reasonable.

Hint List:

- a: Try using the subset command in R two times to get the data you want.

Solution:

(a) (P) Let μ_1 (μ_2) be my average cents off of an A440 in all possible early (later) attempts I could have made. (H) We test H_0: $\mu_2 = \mu_1$ (or $\mu_2 - \mu_1 = 0$) vs. H_1: $\mu_2 > \mu_1$ (or $\mu_2 - \mu_1 > 0$, "flatness problem improve"). Using a (N) two-sample t-test in R, we have:

```
DQPitch = read.csv("...your location...")
early = subset(DQPitch, timeSpan == "early")$centsOff
later = subset(DQPitch, timeSpan == "later")$centsOff
t.test(later, early, alt = "greater")
```

We get (O) a P-value of $0.0009126 < 0.05$, so we (M) reject H_0 in favor of H_1. It appears my flatness problem has improved.

(b) (P) To μ_1 and μ_2, we add μ_3, the average of all possible attempts in the latest time span. We use an (N) ANOVA to compare multiple means by testing (H) H_0: $\mu_1 = \mu_2 = \mu_3$ vs. H_1: Some μ_i is different. In R, we do:

```
summary(aov(centsOff ~ timeSpan, data = DQPitch))
```

This gives (O) a P-value of $0.00173 < 0.05$ suggesting (M) we reject H_0 in favor of H_1; it appears my average amount off is changing across the three time spans.

Commentary: One goal of this problem was to strengthen your HT skills for means in R. You'll notice that the way we structure our data in part a is different than in part b. One main task a data scientist does is restructuring data to fit the requirements of running various tests. In fact, R has packages designed exclusively to aid in data restructuring (e.g., dplyr, which has the fun description: "A Grammar of Data Manipulation").

Another goal was to leave you with a sense of unease. Did you leave this problem feeling like you didn't really do much? We didn't spend any time *actually looking at, visualizing, or understanding* our data. Indeed, the numbers in the data set could all be nonsensical values, and most people would never even know. When I come across a new data set, the first thing I do is spend time getting to know each column of data. I think about what types of values I should see. I find out how many values are missing and try to understand why. I try to visualize the distribution of responses for each variable. The time you spend getting your bearings with a data set is worth its weight in gold. The second bit of unease comes in the H of part a. If my flatness problem has improved, this does mean that $\mu_2 > \mu_1$, but is this *really* the alternative we want to set up? Ideally, we'd like to see if I'm getting *closer* to singing an A440. This suggests an average nearer 0, regardless of whether the average is positive (sharp) or negative (flat). So, ideally we want to test H_1: $|\mu_2| < |\mu_1|$. Unfortunately, we don't know how to test for this alternative which is why I forced you to use the awkward "flatness problem improve" alternative, H_1: $\mu_2 > \mu_1$. In sequences like Math 181A/B and Math 281A/B, you develop the skills to create tests for scenarios you've never seen before.

70. Does the amount of sleep random US college students get vary by year (Freshman, Sophomore, Junior, Senior)? To find out, you and a colleague will collect data from students. As a senior yourself, you decide to tackle data collection for Seniors and reach out to your friends around the country who are also seniors. To speed up data collection, your colleague decides to reach out to random families in the US that simultaneously have 3 kids in college − one a freshman, one a sophomore, and one a junior. Together, you and your colleague end up collecting data on 1088 students.

(a) Complete the PHAN of PHANTOM. As always, for the A step, carefully state the assumptions and whether they are met/not met and why, noting cases where not enough information is known to draw a conclusion.

(b) After running your analysis, you get the below table. Fill in the missing boxes (also show how you arrive at your numbers) and make a decision on the HT using $\alpha = 0.05$.

Source	*df*	*SS*	*MS*	*F*	*P*
Groups			8.048	2.127	
Errors					

Hint List:

- a: The A step is the hardest. Whenever you see a strange description of how the data are collected, see if you can come up with some reason why it may be problematic. Also, there may be assumptions you can't check because of the info available to you.
- b: The *df* column is always found from the description of the problem itself. The MS and SS columns have a connection once *df* is known.

Solution:

(a) (P) Let $\mu_{\text{F}}, \mu_{\text{SO}}, \mu_{\text{J}}, \mu_{\text{SE}}$ be the average hours of nightly sleep for all US college students who are Freshmen, Sophomores, Juniors, and Seniors. We wish to test (H) H_0: All the averages are the same ($\mu_{\text{F}} = \mu_{\text{SO}} = \mu_{\text{J}} = \mu_{\text{SE}}$) vs. H_1: Some average is different. We assume H_0 and run an (N) ANOVA after checking (A):

- Data are independent across groups: No. By picking families with a Freshman, Sophomore, and Junior all in school, their responses are likely to be similar because genetics influence sleep requirements.
- Roughly equal spreads across groups: Unclear, not enough info to decide.
- Data are independent within each group. No. In the Seniors group, you asked your friends. Since friends tend to share some things in common, this might link their responses in an unintended way.
- Data are nearly normal within each group. Possible, but we need more information. Sleep times are a biometric variable, and many of these are normally distributed, including height, weight, and IQ.

(b) We find $df_G = \#\text{ groups} - 1 = 4 - 1 = 3$. Also, $df_E = n - \#\text{ groups} = 1088 - 4 = 1084$.

Since $F = \dfrac{MSG}{MSE} \implies MSE = \dfrac{MSG}{F} = \dfrac{8.048}{2.127} \approx 3.784$.

Since $MSG = \dfrac{SSG}{df_G} \implies SSG = df_G \cdot MSG = 3 \cdot 8.048 = 24.144$.

Similarly, $SSE = df_E \cdot MSE \approx 4101.567$.

Finally, our P-value is pf(2.127, df1 = 3, df2 = 1084, lower = F) ≈ 0.095. Since this is greater than 0.05, we keep H_0; it seems sleep averages are similar across the four years of college.

Commentary: At first blush, this problem might feel artificial. After all, as a scientist you usually work in a forward direction: from your data to the P-value. There are times, however, when we must work backward. For example, when I am reading or refereeing a journal article, it is common that the authors have not included an entire ANOVA table. If I want to double check their work, I do the types of steps in this problem. As another example, when you conduct a meta-analysis (a study that combines the data across multiple other studies), you often need access to statistical details not present in the original articles. In many cases, these details can back-derived from the available information. So, the type of reasoning in this problem is not as strange as you might think.

71. I recently told a student who was considering a career in teaching that education is one of the most complex and interesting systems I've ever seen. Here, we explore a simple study that shows how even the smallest changes can affect academic performance. Ask yourself this:

> Does the ink color a teacher chooses for their exam directions change student performance on simple tasks, like finding anagrams (i.e., recognizing that IONAP rearranges to PIANO)?

Researchers tested this using 3 colors (red, green, and black). Of the 19 participants in the red group, people got an average of 4.4 of the anagrams correct. For the 27 participants whose directions were green, the average was 5.7. For the 25 black-directions participants, the average was 5.9.

(a) Run a test to determine if the color of the directions affects anagramming ability using $\alpha = 0.01$. You should construct an ANOVA table given that the sum of squares for groups was 27.7 and the sum of squares for errors was 57. [PHNTOM]

(b) In part a, I didn't actually have to give you the value 27.7. Explain how this value can be found using the information given in the problem. (Note: You should get a value close to 27.7, but it won't match exactly because there are rounded values all over this problem.)

Hint List:

- a: You'll need R to find the P-value.
- b: Start by finding the MSG using the formula from class.
- b: If you know 5 test scores had an average of 80, what is the sum of those test scores?

Solution:

(a) Let μ_R, μ_G, μ_B be the average number of anagrams subjects can successfully rearrange when they are given directions in either red, green, or black ink. We test H_0: $\mu_R = \mu_G = \mu_B$ vs. H_1: Some mean is different using an ANOVA. Using the given info, we get this table:

Source	df	SS	MS	F	P
Groups	$3-1=2$	27.7	13.85	16.523	$1.418 \cdot 10^{-6}$
Errors	$19+27+25-3=68$	57	0.838		

Since our P-value is < 0.01, we reject H_0; it appears the color of the directions can influence anagramming ability.

(b) In class, we saw MSG is defined as $\frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2$ where $\overline{x}$ is the average of all the data (ignoring group labels). We'd like to add all the data values and divide by $19+27+25=71$, but we don't have the individual data values. Notice, though, that if you know an average $\overline{y} = \frac{\sum \text{data values}}{n}$, then $n \cdot \overline{y} = \sum \text{data values}$. So,

$$\overline{x} = \frac{\sum \text{all data}}{71} = \frac{\sum \text{red data} + \sum \text{green data} + \sum \text{black data}}{71}$$
$$= \frac{n_R \cdot \overline{r} + n_G \cdot \overline{g} + n_B \cdot \overline{b}}{71} = \frac{19 \cdot 4.4 + 27 \cdot 5.7 + 25 \cdot 5.9}{71} \approx 5.423$$

Now, we find that $MSG = \frac{1}{2}\left[19(4.4-5.423)^2 + 27(5.7-5.423)^2 + 25(5.9-5.423)^2\right] \approx 13.82$. Finally, $SSG = 2 \cdot MSG \approx 27.644$, or about 27.7, as given.

Commentary: If this problem was interesting, psychology might be your field. Note that there is a large literature exploring the effect of color on human performance. Researchers have shown you're more likely to lose to an opponent wearing red and perform poorly on a test when exposed to the color red.

On a different note, it might be surprising that it's possible to solve part b. As mentioned, usually $\overline{x}$ is found by combining *individual* data values to form the average. Here, we just have $\overline{r}, \overline{g}$, and $\overline{b}$, not the individual values. Recall that since a statistic is a summary, it represents a *loss* of information. So, $\overline{r} = 4.4$ hides the 19 values that went into this average; we lost that information. Amazingly, while $\overline{r}$ loses that info, it maintains the sum of those 19 numbers ($19 \cdot 4.4$) which is all we need in this problem! Would we have been able to find MSE instead of MSG with the given info?

72. The irrational number $\pi = 3.14159265358979\ldots$ has infinitely-many digits. Here's a question I asked myself the other day: Suppose we break the digits into three groups, taking every third digit:

- Group A: $3, 1, 2, 3, 9, \ldots$ (first digit, and then every third digit thereafter)
- Group B: $1, 5, 6, 5, 7, \ldots$ (second digit, and then every third digit thereafter)
- Group C: $4, 9, 5, 8, 9, \ldots$ (third digit, and then every third digit thereafter)

(For those who know number theory, these are digits grouped by their placement modulo 3.) Will the average digit in each of the three groups be the same? To figure this out, we'll use the file pi.csv on TritonEd/Canvas which contains the first 10000 digits of π. Conduct a HT in R using a significance level of 0.05 to answer my question. You must include a side-by-side boxplot as part of the A of PHANTOM. [PHANTOM, VIZ]

Hint List:

- There are some interesting R issues in this problem. First, this is a data frame with only one column, so be careful to use dataFrameName$columnName to get at the digits (instead of just dataFrameName). Second, you need to create the three groups using the first 10000 digits. This process involves extracting only certain values from a vector. This is most easily done with the bracket notation. As inspiration, observe what this code does in R:

```
DQFavoriteColors = c("green", "blue", "black", "orange", "purple")
DQFavoriteColors[c(2,3,5)]
```

  Finally, you'll need to build a new data frame using your three groups that is structured correctly for the R test you want to run. Making data frames is not too tough. Observe what this code does in R:

```
funData = data.frame(colors = c("red","blue","gold","purple"),
                     numLetters = c(3,4,4,6))
View(funData)
```

- The functions seq and rep may also be useful as you explore this problem.

Solution: (P) Let μ_A (μ_B, μ_C) be the average of **all** digits in group A (B, C). (H) We test H_0: $\mu_A = \mu_B = \mu_C$ vs. H_1: Some mean is different using an (N) ANOVA. We (A) assume H_0 and check:

- Independence within and across groups: Possibly not. Notice that the groups A, B, and C are defined using *all* digits of π. When we draw our samples, we use only the first 10000 digits, not a random sample from *all* digits. To see how this might be problematic, just consider a number w that begins with 10000 3s before switching to an alternating pattern of 3s and 5s: $w = 0.333\ldots353535353\ldots$ Ideally, we'd like to be able to sample randomly from all digits of π, but no known algorithm can do this in a reasonable amount of time.
- Data nearly normal in each group: This fails also. We can also use R to create a table or histogram of our data:

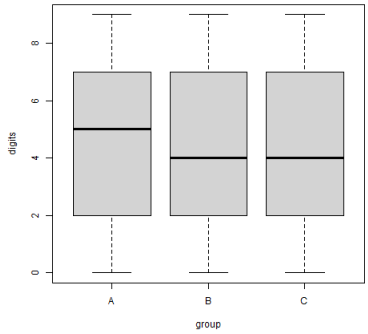
```
pirep = read.csv("...location...")
digits = pirep$digits
groupA = digits[seq(1, 10000, by = 3)]
table(groupA)

# I got:
#   0   1   2   3   4   5   6   7   8   9
# 321 366 353 326 334 336 331 320 323 324
```

- Roughly equal spreads: Yes. Once we restructure our data to run the ANOVA, we can use a side-by-side boxplot to explore spreads:

```
groupB = digits[seq(2, 10000, by = 3)]
groupC = digits[seq(3, 10000, by = 3)]
group = c(rep("A", 3334), rep("B", 3333), rep("C", 3333))
piData = data.frame(group = group,
                    digit = c(groupA, groupB, groupC))
boxplot(digits ~ group, data = piData)
```

  If you prefer the SD (instead of the IQR) as a measure of spread, you can take the SD of each group and see they are all about 2.8.



To run the ANOVA, we use:

```
summary(aov(digit ~ group, data = piData))
```

We get an (T) F-stat of 0.836 and a (O) P-value of 0.433. Since $0.433 > 0.05$, we (M) keep H_0; it appears the "average digit" in the three groups is similar.

Commentary: There are three takeaways for this problem. First, notice how the side-by-side boxplot *fails* to reveal the lack of normality. People often believe that if the line (Q2) in the boxplot is roughly centered, then the data will be normal. This is false, as these data show. Indeed, any symmetric data set will have a roughly-centered Q2, normal or not. Second, this problem is another example of a setting where the null is actually the interesting hypothesis. Why? Since π is an irrational number, the digits should appear to be random. So, when we divide them into these three groups, there should be no discernible patterns. Thus, we would expect the means of the three groups to be roughly equal.

Finally, I wanted you to see an example of what it looks like to pose an original question, explore it using data, and reach a statistical conclusion. This is the essence of research, and there are endless questions out there that you are ready to ask and tackle right now. I hope when you look back at this problem, you think "That question is something I could have thought up." Notice that it doesn't use any fancy terms or mathematical machinery. Many students at UCSD want to get involved in research. One option for this is getting connected with an existing research team at the school. The pros of this are that you get to be part of a team, you have an expert guiding your work, and you might get your name on a publication. The major con is that you're always working on *someone else's* question. As such, you don't learn *how* to ask questions, which is the first and, sometimes, most important component of advancing knowledge. If you're looking for a way to spend a summer, try this: Think of a simply-stated question that you've never seen anyone discuss. Then, design a plan to answer this. Carry it out, and then share your exploration with others (via GitHub, YouTube, a write-up on your website, etc.). One former student, Owen Shen, did this with the question: "Do the causes of death mentioned in news articles match the rates at which these causes actually occur?". It's a simple question, and the answer may surprise you. Read more here.

Why do I suggest this? One reason is that school is almost exclusively about working other's problems (the professor, the textbook, the AP exam, etc.). And yet, at some point, we need people who can ask new questions. These questions might be in academia, business, or government. Wherever they are, the issue is the same: To advance society/knowledge, we need people who have the courage to ask: Is this true? What would happen if we changed this? How do we make this process more efficient? Now is the time to begin the courageous act of asking original questions.

73. If you watch much soccer ("football" outside the US), you'll quickly learn that players will fake injuries to get referees to make calls in their team's favor. During a recent game, I reached my breaking point when I saw a player clearly fake an injury [video replay showed no contact with other players] which resulted in his team getting a penalty kick, which resulted in a goal. I suggested to my friend that the sport could be improved by establishing a review board that would study supposed infractions after the fact and use all available video footage to support the referee's call from the field or contradict this call. If the committee adopts a null hypothesis that a referee's call during the game was accurate, describe what Type I and Type II errors mean in this context, and discuss the implications of the review board making each type of error.

Hint List:

- Set up hypotheses using words rather than numbers/variables.

Solution:

We set H_0: The ref's call was accurate vs. H_1: The ref's call was inaccurate.

Recall that a Type I error occurs when H_0 is true, but our HT leads us to believe in H_1. Here, the HT is the review board looking at video evidence and keeping the ref's on-field decision or not. If we make a Type I error, this means the ref made the correct call, but the review board disagrees. This could result in a reversal of the on-field decision and possible fines for the referee. In this setting, a Type I error erodes spectator trust in referees, which diminishes the integrity of the sport. Note that this kind of error might occur if the referee had a view of the play that led to their decision but which was not available for video review.

A Type II error occurs when H_1 is correct, but the HT leads us to keep H_0. Here, this means the ref's call was incorrect, but the video footage is not helpful for revealing this to the review board, and they think the ref's call was accurate. One trouble with this error is that we've created another layer of bureaucracy that isn't doing what it was designed for (to reverse incorrect referee calls). This wastes time and money, creating a system people can't trust.

Commentary: In any complex system (e.g., sports matches, legal disputes, etc.), it is useful to establish an arbiter, or entity responsible for having the final say on contested issues. These decision-makers specialize in the rules of the system. Sadly, they are capable of making mistakes, and with the advancement of civilization, many systems have established a secondary layer of checks-and-balances. In sports, for example, it is common to review all scoring situations and to allow for team challenges to on-field calls. In the US legal system, cases can be appealed to higher courts (ultimately leading to the Supreme Court). Always remember that the arbitration of systems is imperfect. We are constantly making Type I and II errors in situations that can be framed using a set of binary hypotheses. As such, I find it beneficial to approach life with the expectation that errors will be made. I will make them in my actions, and others will make them in relation to me. One benefit of this knowledge is that you move through life with greater generosity of spirit. If I misplay a musical passage, miscalculate when solving a problem, or misremember a name, these are not moments that reveal my inadequacy as a person. Rather, they reflect the reality of the human experience: we are awash in Type I and Type II errors. We should certainly work on minimizing these, but along the way, we'll make many ourselves and have many made that affect us directly. I suggest kindness and humility when these occur, rather than the current trend of outrage and accusation.

74. Each January, some percentage of Americans make resolutions to better themselves. You might wonder if that percentage changed in January 2021 because of COVID-19. This problem was inspired by ideas I saw in this journal article.

(a) Give a real-world reason why the percentage of Americans making resolutions might have been higher in January 2021 than in January 2020. Then, give a reason why the January 2021 resolution percentage might be lower than that in January 2020. (Because of your competing ideas, we'll explore a two-sided alternative hypothesis below.)

(b) Suppose you know that in 2020, 58.8% of all Americans made New Year's resolutions. To understand 2021, you plan to interview 300 random Americans and see what percentage made resolutions. Describe what a Type II error means in this context, and find the probability of making such an error if the true 2021 New Year's resolution percentage is 74%. Assume your hypothesis test uses $\alpha = 0.04$. You must include a picture with two labeled curves as part of your answer. Give your answer in scientific notation and explain why you should get a very tiny answer based on the problem setup. [PHAN of PHANTOM, VIZ with two well-labeled curves, Tech not tables]

Hint List:

- b: A two-sided alternative requires we divide α in half and set up two rejection fences.
- b: Recall that $\widehat{p} \approx N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$. When we assume H_0, we believe in a certain value for p. That's one of the curves. The other is the truth of the universe: regardless of what we believe, the universe knows the true value for p and gives us a value for $\widehat{p}$ based on this truth, not our belief!

Solution:

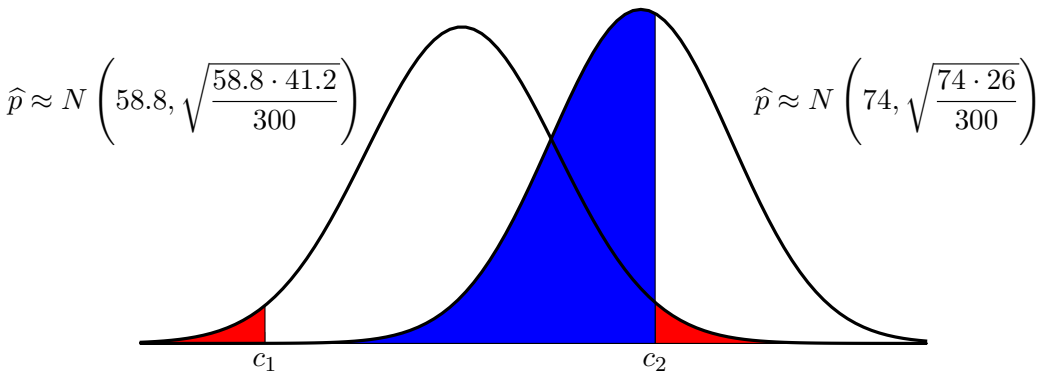
(a) Higher: Perhaps the pandemic made people think more about life and death, resulting in the desire to live life to the fullest. This could inspire more resolution-making.

Lower: Making and enacting resolutions requires time, mental energy, and resources. People may have had less of these because of the pandemic.

(b) (P) Let p be the percentage of **all** Americans who made New Year's resolutions in 2021. (H) We test H_0: $p = 58.8\%$ vs. H_1: $p \neq 58.8\%$, and we secretly know that $p = 74\%$. Assuming H_0, we run a (N) one-sample z-test after checking (A):

- Independent data: Yes, since we choose random Americans and our sample size is less than 10% of the population size.
- At least 10 predicted successes and failures: Yes, since $300 \cdot 0.588 = 176.4 \geq 10$ and $300 \cdot 0.412 = 123.6 \geq 10$.

We create the below picture showing the distribution we expect to give us a value for $\widehat{p}$ (based on our assumption of H_0) and the distribution that actually does so:



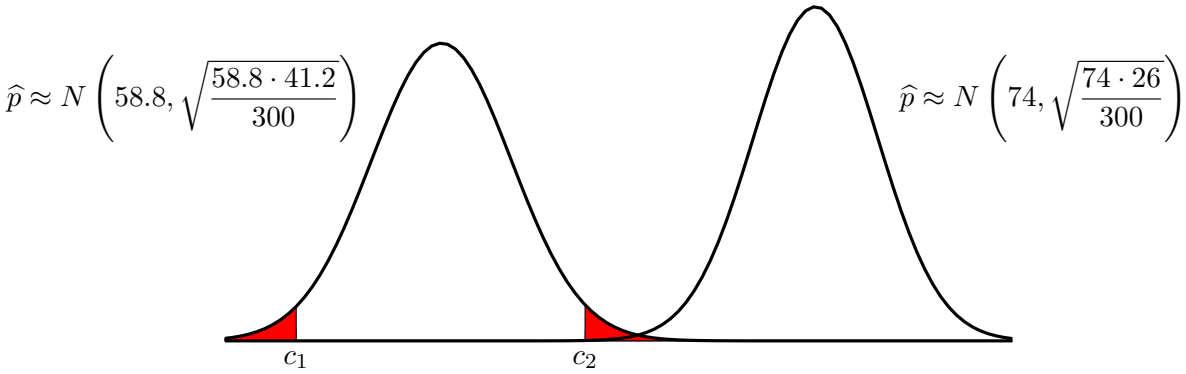
R helps find the locations of the rejection fences:

```
c1 = qnorm(0.02, 58.8, sqrt(58.8 * 41.2 / 300))
c2 = qnorm(0.02, 58.8, sqrt(58.8 * 41.2 / 300), lower = F)
```

Here, a Type II error means that while there is a different percentage of people making resolutions in 2021, our data mislead us, and we end up thinking the rate is the same as in 2020. Thus, $P(\text{Type II error}) = P(c_1 \leq \widehat{p} \leq c_2) =$ blue area. We find this in R:

```
sd = sqrt(74 * 26 / 300)
pnorm(c2, 74, sd) - pnorm(c1, 74, sd)
```

Our answer is $\boxed{1.088 \cdot 10^{-4}}$, which we expect to be tiny because it is easy to detect a 15% change (74% − 58.8%) using 300 people. It would be very unlikely (i.e., low Type II error rate) to miss such a glaring difference. Indeed, if we actually draw the above picture to scale, it looks like the below image. We can't even see the blue shading!



Commentary: One of my favorite mental activities is the type of thinking seen in part a. When you look at something complicated (e.g., the percentage of people who enact some behavior), it is useful to think about the competing forces that want to push that value up or down. Indeed, most things of interest in the world are influenced by many forces. For example, see Wikipedia's lengthy article on the causes of unemployment in the US. When you try to understand and incorporate this multitude of influences, you are engaging in *multivariate thinking*. Sadly, this is not how most media presents current issues. Today, the tendency is to oversimplify, blaming a single factor for an observed change. This is known as the *univariate fallacy*. This thinking occurs when a politician credits a single bill for fixing a problem, when a person blames a single cause for climate change, or when a student cites a single reason for performing poorly on a test. Life is simply more complex than this.

75. In early 2021, President Biden signed several executive orders focused on racial inequity in America. In the briefing that followed, PBS journalist Yamiche Alcindor asked the press secretary what metrics the administration would use to decide if their efforts were working. Here's one idea that could have been offered in response: Given a random person of color (POC) in a particular job at a particular company, you can randomly choose a non-POC (with the same job/age/gender/experience/etc. at the same company) and look at the difference in their salaries (non-POC salary − POC salary). When you find this difference for lots of pairs of workers and see what is happening on average, this is known as the wage gap (based on POC status with matched job/age/gender/experience/etc.). Assume these differences are normally distributed with SD \$450, and that in 2020, the wage gap was \$1750. Suppose the administration wanted to run a study in 2022 to see if their interventions had decreased the wage gap. They'll assume the SD of pay differences remains at \$450 and would feel they had been successful if the wage gap had dropped by at least \$100. If they want their test, which uses $\alpha = 0.03$, to have power of (at least) 0.9, how many random pairs (POC and non-POC sharing the same job/age/gender/experience/etc.) should they plan to recruit? [PHA of PHANTOM, VIZ with two well-labeled curves, Tech not tables]

    If you want to read more about this topic, and consider the effect of other variables, see here.

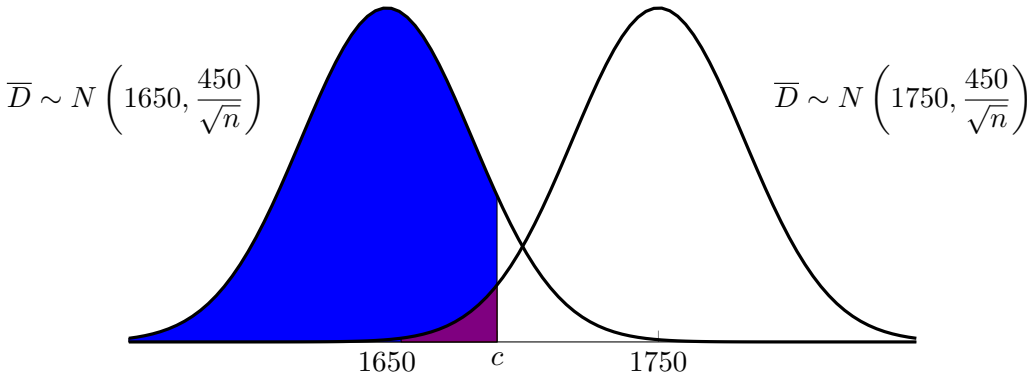
Hint List:

- The number \$100 doesn't get encoded in the hypotheses. Rather, it reflects the "least interesting change" the administration would care about.
- Recall that if $X \sim N(\mu, \sigma)$, then $\overline{X} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$. Here, X represents a salary difference between a non-POC and POC who have been matched on other attributes.

Solution: (P) Let μ_d be the average of wage difference (order: non-POC − POC) between all POC salaries and non-POC salaries when matched for job, company, gender, age, experience, etc. (H) We test H_0: $\mu_d = 1750$ vs. H_1: $\mu_d < 1750$ ("decreased wage gap") where we claim success if we get μ_d below $1750 - 100 = 1650$ (the "least interesting" new mean). (A) Assuming H_0, we check:

- Independent differences: Yes. We are choosing random pairs and are likely to sample $< 10\%$ of all possible pairs.
- CLT balance condition on differences: Yes. We are told the differences are normally distributed, so there is no skew to worry about balancing out!

Let $D \sim N(\mu_d, 450)$ be the pay difference for a random non-POC/POC pair matched on non-racial attributes. We draw the distribution of $\overline{D} \sim N\left(\mu_d, \dfrac{450}{\sqrt{n}}\right)$ under the assumption of H_0 ($\mu_d = 1750$) and the least interesting μ_d for which the administration could claim success ($\mu_d = 1650$):



On $Z \sim N(0,1)$, we see qnorm(0.03) ≈ -1.880794. Thus, on the H_0 distribution above, we have

$$\frac{c - 1750}{450/\sqrt{n}} \approx -1.880794 \implies c = 1750 - 1.880794 \cdot \frac{450}{\sqrt{n}}$$

Using the blue area which represents the power, we see qnorm(0.9) ≈ 1.281552, so

$$\frac{c - 1650}{450/\sqrt{n}} \approx 1.281552 \implies c = 1650 + 1.281552 \cdot \frac{450}{\sqrt{n}}$$

Setting these equal and simplifying gives

$$100 = 3.162346 \cdot \frac{450}{\sqrt{n}} \implies n \approx 202.509$$

Since n must be an integer, we round this up to ensure the power is at least 0.9 (rounding down will drop the power slighly below 0.9). Thus, $\boxed{n \geq 203}$ random pairs.

Commentary: When you compare two groups (e.g., POC and non-POC) on some variable (e.g., job salary), it is very important to consider other variables that may have an influence on your study. Such variables are known as *confounding variables* (to "confound" means to cause confusion) because they cause confusion in the analysis. In a perfect world, if you find a difference in your variable, you'd like to point to your group membership as the reason why. Confounding variables make this problematic. As an example, suppose you want to compare Democrats and Republicans (the groups) looking at average house price (the variable). Lurking under the surface is the variable of location: Democrats are more likely to be found in cities where house costs are higher (as are salaries to compensate). So, if you run your analysis, you might conclude that Democrats have an average higher home cost than Republicans, but this may have more to do with location than political affiliation.

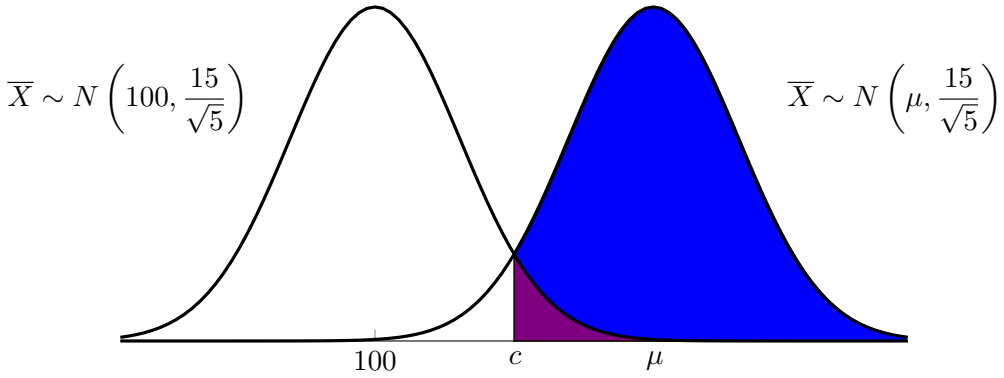
When researchers try to deal with the influence of confounding variables, they say they have "controlled for" these variables. One way to do this is by matching up members of your two groups so they are identical on your confounding variables (as we do in the above problem when I mention "job, company, gender, age, experience, etc."). Another approach is to use a multivariate statistical technique that considers *all* the variables in question (the ones you care about *and* the confounding variables) and acts accordingly. These approaches are limited by the ability of researchers to dream up and measure confounding variables. For example, you might care about variable V for groups G_1 and G_2, and you might realize that variables $C_1, \ldots, C_k$ are confounding (and hence must be controlled for in the analysis). But, it might be the case there are addition confounding variables $D_1, \ldots, D_\ell$ that you didn't think of, and that these variables actually drive any differences you see between G_1 and G_2 on measure V. This observation is sobering. Even if we are able to move past the univariate fallacy and consider the multitude of factors that influence an object of study, there will always be variables we haven't considered that confound our analysis. If those overlooked variables are the true drivers of differences we observe, then we may misattribute the causes of observed differences.

76. There are some research settings where you'll never be able to collect much data. This might happen if data collection is time consuming, expensive, or dangerous to participants. In these settings, it can be useful to do a "reality check" on whether moving forward with the research is sensible or not. Here's an example of this. Suppose you have a product you believe can raise human IQ. You'd like to run a test to check if it increases IQ from the known average of 100, but the product is expensive to make, so you can only run the test on 5 participants. You'd like to publish the findings in a journal that demands a Type I error rate of 0.05 and a power of (at least) 95%. What is the smallest IQ increase for the drug that would meet these requirements given that you only have 5 participants? Assume that the drug does not alter the standard deviation of individual IQ scores, which is known to be 15 points, or the shape of the distribution, which is known to be normal. [PH of PHANTOM, VIZ with two well-labeled curves, Tables not tech]

Hint List:

- If X is the random IQ of someone one our drug, what are the distributions for X and $\overline{X}$?
- Areas and quantiles are easy to find on Z. How do you move between a given normal distribution and the standard normal?

Solution: (P) Let μ be the average IQ of people using your product. (H) We test H_0: $\mu = 100$ vs. H_1: $\mu > 100$ ("raise human IQ"). If $X \sim N(\mu, 15)$ is the IQ of a random person, then $\overline{X} \sim N\left(\mu, \frac{15}{\sqrt{5}}\right)$ is the average IQ of 5 random people. When we assume H_0, we believe $\mu = 100$, thus forcing $\overline{X} \sim N\left(100, \frac{15}{\sqrt{5}}\right)$. However, when we collect the IQs of 5 random people on our drugs, this yields one average from the true distribution $\overline{X} \sim N\left(\mu, \frac{15}{\sqrt{5}}\right)$. Thus, we have this setup:



On $Z \sim N(0,1)$, notice that $P(Z > 1.65) \approx 0.05$ (1.64 also OK). This means that

$$\frac{c - 100}{15/\sqrt{5}} = 1.65 \implies c \approx 111.069$$

If we want the blue area to be 0.95, we start first on $Z \sim N(0,1)$. Here, we find $P(Z > -1.65) \approx 0.95$ (-1.64 also OK). So, in our setting, we see:

$$\frac{111.069 - \mu}{15/\sqrt{5}} = -1.65 \implies \boxed{\mu \approx 122.137}$$

Commentary: In school, it is tempting to quickly box your answer and move on at the end of a problem. After all, you're usually graded on the accuracy of your final answer and, in classes that show kindness, the steps that led to that answer. How often, though, are you graded on what comes *after* getting an answer? If you haven't thought past the answer above, take a moment to do so now. What do you make of this answer? What does it suggest? After you've considered this, continue reading below.

To think beyond the answer is to explore what your finding means for/in the world. This problem says that we need at least a $122.137 - 100 = 22.137$ IQ point improvement for our sample of 5 to reveal a statistically significant result with 0.95 (or higher) probability. That is, if you want a really great chance (i.e., ≥ 0.95) of discovering an increase (i.e., a statistically significant result using $\alpha = 0.05$) with $n = 5$ people, then that increase better be about 22 IQ points (or more). If you've never explored changes in academic performance outside of trying to improve your own grades, let me help. Said simply, changes this large in the academic space are basically non-existent. Even small changes in IQ/academic performance are difficult to create, time-intensive, and expensive. For our setting above, there's little point in running the study because the drug is unlikely to raise IQ scores this much. Indeed, this problem shows more generally that if your sample must be *small* (for whatever reason), then the change you are trying to detect (here, IQ improvement) must be *huge*. For this reason, almost all *quantitative* educational studies with fewer than 10 subjects can be trashed. As such, you'll often see small-sample educational studies focus instead on *qualitative* methods (e.g., interviews, video analysis) and explore different lines of inquiry than above.

77. My friend recently bought a home in east county, and I drove out there to watch *House of the Dragon*. As I headed away from the Pacific coast, I noticed the temperature on my car's outside thermometer continually rose. I recorded a few data points as (distance from coast in miles, temperature in °F). When I got home, I created a linear model in R and got the equation $\widehat{\text{Temp}} = 65.3 + 1.2 \cdot (\text{Distance})$.

(a) Interpret the numbers in this equation given our setting. Also state what units they have.

(b) How far does the model predict I would need to drive east from the coast for the temperature to reach 100°F? Using a map app, approximate what city this would be along Interstate 8.

(c) Because of elevation changes, microclimates, and other local phenomena, the relationship between distance and temperature does not have a perfect correlation of 1. Indeed, my data gave a correlation of 0.82. If my friend had been making the reverse of my journey when I collected the data, roughly what would his correlation have been? (Assume he measures distance traveled west from his house.)

Hint List:

- c: Try sketching what my data might look like. Then sketch my friend's data on a different set of axes.

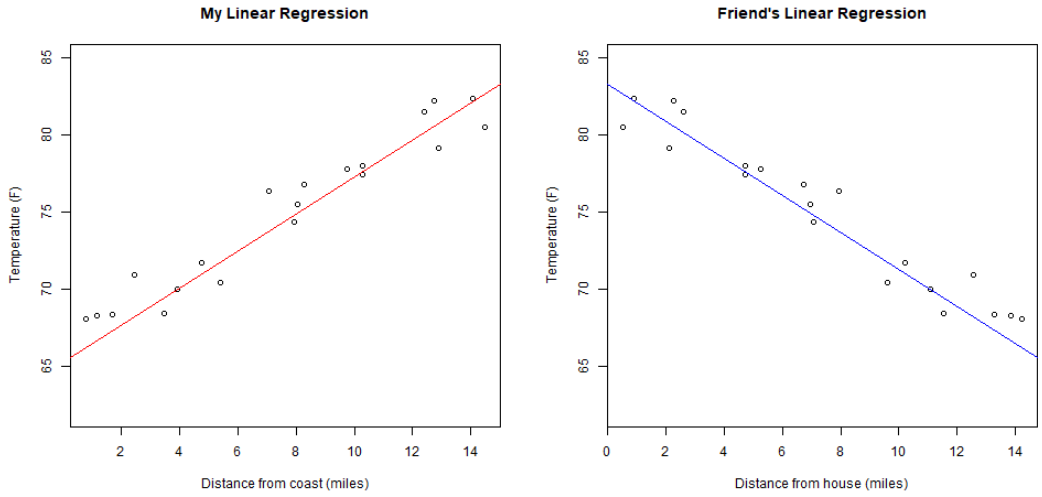
Solution:

(a) 65.3 °F (y-intercept): This is temperature the model predicts when $x = 0$ (i.e., at the coast) on the day/time of my journey.

1.2 $\frac{\text{°F}}{\text{mile}}$ (slope): This means that for each mile I head east away from the coast, the model predicts a temperature increase of 1.2 °F.

(b) Set $100 = 65.3 + 1.2 \cdot \text{distance} \implies \text{distance} \approx \boxed{28.917 \text{ miles}}$. Using a map app, this is near $\boxed{\text{Alpine, CA}}$.

(c) Imagine if you plotted my data relating the distance from the coast in miles to the temperature in °F. Since the temperature rises as we head east, the slope is +1.2. If my friend took the same roads westward, he would experience roughly the same temperatures but in reverse (see images). So, his temperatures would decrease by 1.2 °F per mile. Thus, his regression has slope −1.2, which will make his correlation about $\boxed{-0.82}$.



Commentary: Whether you're aware of it or not, the world is constantly being governed by rules that generate patterns. The unobservant person might never notice that the weather gets hotter as you head east from the coast. The observant person will be aware of this but is unable to precisely quantify it. The statistically trained person can quantify the trend and use it to their advantage (e.g., in choosing what clothes to bring, in deciding where to buy a home in SD, etc.). The more you understand about the world, the better choices you can make in every moment.

78. The correlation coefficient is supposed to be unaffected by rescaling. To show this, suppose we have data points $(x_1, y_1), \ldots, (x_n, y_n)$ with correlation r. Show that doubling all the y values: $(x_1, 2y_1), \ldots, (x_n, 2y_n)$, gives the same correlation as the original data.

Hint List:

- Write out the formula for r using the un-doubled data: $(x_1, y_1), \ldots, (x_n, y_n)$.
- How is the mean of the y values in the doubled data related to the mean of the y values in the un-doubled data? What about the SDs?

Solution: With our un-doubled data set, we know $\overline{x} = \dfrac{\sum x_i}{n}$ and $\overline{y} = \dfrac{\sum y_i}{n}$. Also,

$$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \overline{x})^2} \text{ and } s_y = \sqrt{\frac{1}{n-1}\sum (y_i - \overline{y})^2}$$

With this notation, we know r is defined as $\dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}$.

When the y values are all doubled, how does this affect the mean and SD? Note that

$$\overline{y}_{\text{doubled}} = \frac{\sum 2y_i}{n} = 2\overline{y}$$

Using this, we find that

$$s_{y \text{ doubled}} = \sqrt{\frac{1}{n-1}\sum (2y_i - \overline{y}_{\text{doubled}})^2} = \sqrt{\frac{1}{n-1}\sum (2y_i - 2\overline{y})^2} = \sqrt{\frac{4}{n-1}\sum (y_i - \overline{y})^2} = 2s_y$$

Thus, the correlation of our doubled data is

$$r_{\text{doubled}} = \frac{\sum (x_i - \overline{x})(2y_i - \overline{y}_{\text{doubled}})}{(n-1)s_x s_{y \text{ doubled}}} = \frac{\sum (x_i - \overline{x})(2y_i - 2\overline{y})}{(n-1)s_x \cdot 2s_y} = \frac{2\sum (x_i - \overline{x})(y_i - \overline{y})}{2(n-1)s_x s_y} = r$$

Commentary: This is another example of a gentle mathematical proof. If this type of argument resonates with you, courses like Math 109 (Mathematical Reasoning) and Math 181A/B (Mathematical Statistics I and II) are good future choices. I knew I should major in mathematics when I consistently felt an empty feeling upon being told *that* something was true without being told *why* it was true. For some, the feeling of understanding/contentment only comes when they have drilled down to the details at the very base of a claim. In this problem, we get that satisfaction on the final line: Because the new y mean is twice the old, we can factor a 2 from the numerator. Because the new y SD is twice the old, we can do the same in the denominator, eventually leading to a cancellation. The correlation coefficient is defined so cleverly: The numerator and denominator scale similarly as the data scale, and hence, the quotient allows for the erasure of scaling alterations! You miss this observation if you simply accept that "The correlation coefficient is unaffected by scaling". More generally, you miss many opportunities to deeply understand your tools if you spend all your time using them rather than studying how/why they work.

79. You've just landed an amazing statistics internship at Marvel Comics for the summer. Since everything they create is fictional, your boss has asked you to make up data. They want a plot showing the popularity of Marvel (0 to 100 scale) in the two decades following the release of *X-Men* in 2000. The goal is for the made-up data to roughly follow the line $\widehat{y} = 50 + 2.1 \cdot (\text{Years since 2000})$. You use the below code:

```
set.seed(12345)
tvals = runif(60, 0, 20)
yvals = 50 + 2.1 * tvals + rnorm(60, 0, 2)
```

(a) Explain, in words, what this code does.

(b) Using R, find the correlation between these two vectors of data.

(c) In a and b, you generated 60 data points and found the correlation. Imagine, instead, that infinitely many points were generated. Without R, find the exact correlation using ideas seen earlier in the course.

(d) What number must 2 (in rnorm above) be changed to if you want the correlation to be roughly 0.6?

Hint List:

- c: Set $T \sim Unif(0, 20), E \sim N(0, 2)$, and $Y \sim 50 + 2.1 \cdot T + E$. Find $Var[Y]$ and proceed.
- d: Set $E \sim N(0, \sigma)$ and build on your steps from part c.

Solution:

(a) When you generate random data using R, the computer uses features particular to your computer (e.g., the exact time you run the code and specific features of your operating system) to help give the appearance of randomness. By calling set.seed, we override this behavior and ensure that (almost) everyone will get the same sequence of random numbers regardless of the time or type of computer you run the code on. After this, we generate 60 values between 0 and 20 (i.e., the years 2000 to 2020). These will act as values on the horizontal axis and symbolize possible moments we could have surveyed the public on Marvel's popularity. Finally, we generate popularity scores that follow the desired model $(50 + 2.1 \cdot t)$ but which have some variability above/below that line (because of the word "roughly" in the problem). The rnorm command adds an "error" making the data noisier, and hence, more believable.

(b) After typing in the given code, we add:

```
cor(tvals, yvals)
```

Our correlation coefficient is $\boxed{r \approx 0.983}$.

(c) Whenever you are thinking about "infinitely many data points", you are really talking about random variables. Set $T \sim Unif(0, 20), E \sim N(0, 2)$ (the error), and $Y \sim 50 + 2.1 \cdot T + E$. Recall that $Var[Unif(a, b)] = \dfrac{(b-a)^2}{12} \implies SD(T) = \dfrac{20-0}{\sqrt{12}} = \dfrac{10}{\sqrt{3}}$.

$$Var[Y] = Var[50 + 2.1 \cdot T + E] = 2.1^2 \cdot Var[T] + Var[E] = 2.1^2 \cdot \frac{(20-0)^2}{12} + 2^2 = 151$$

Since $b = r \cdot \dfrac{s_y}{s_x}$, we have $r = \dfrac{bs_x}{s_y} \approx \dfrac{2.1 \cdot 10/\sqrt{3}}{\sqrt{151}} \approx \boxed{0.987}$ (see commentary).

(d) Set $E \sim N(0, \sigma)$. Looking at the steps in part c, we can write

$$0.6 = r = \frac{bs_x}{s_y} \approx \frac{2.1 \cdot 10/\sqrt{3}}{\sqrt{2.1^2 \cdot 400/12 + \sigma^2}}$$

Rearranging gives

$$2.1^2 \cdot \frac{400}{12} + \sigma^2 \approx \left(\frac{2.1 \cdot 10/\sqrt{3}}{0.6}\right)^2 \implies \boxed{\sigma \approx 16.166}.$$

Commentary: While the setting for this problem is artificial, it gives you the chance to feel statistics at *two levels*. The problem begins (parts a and b) at the *sample level*: You have 60 data points that are constructed to follow a particular line. In parts c and d, we move to the *population level*, considering

all possible data points and exploring what is happening among that multitude of points. Never forget that one main goal of statistics is to use information in a *sample* to draw mathematically-responsible inferences about the *population*. We usually model what's happening at the population by random variables since they have the potential to represent infinitely many outcomes; this is why I introduced the RVs in part c.

One subtle issue with working across these two levels is that it's tempting to treat them as interchangeable when, in fact, they aren't. For example, in a sample, s_x means the standard deviation of the n data points that are present and is given by

$$s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{x})^2}$$

In contrast, $SD(X)$ represents the standard deviation of the infinitely many values in the support of X and is given by

$$SD(X) = \sqrt{Var(X)} = \sqrt{\int_{-\infty}^{\infty}(x-\mu)^2 \cdot f_X(x)\,dx}, \text{ if } X \text{ is a CRV}$$

These formulas are designed so that $s_x \approx SD(X)$, which is why you see me use the $\approx$ sign in the middle of the solutions to parts c and d. I did not use an $=$ sign because there is a difference between the two levels (sample vs. population). I believe that if you train yourself to pay attention to the smallest details of a subject, then, like a fine-meshed sieve, you're sure to catch the medium and large matter as well.

80. Below is a table showing a variety of countries (first column), the number of cigarettes consumed per adult per year (second column), and the rate at which people die from coronary heart disease (CHD) per 100,000 people (third column).

United States	3900	256.9
Canada	3350	211.6
Australia	3220	238.1
New Zealand	3220	211.8
United Kingdom	2790	194.1
Switzerland	2780	124.5
Ireland	2770	187.3
Iceland	2290	110.5
Finland	2160	233.1
West Germany	1890	150.3
Netherlands	1810	124.7
Greece	1800	41.2
Austria	1770	182.1
Belgium	1700	118.1
Mexico	1680	31.9
Italy	1510	114.3
Denmark	1500	144.9
France	1410	59.7
Sweden	1270	126.9
Spain	1200	43.9
Norway	1090	136.3

(a) Put these data into a data frame in R. Then run a linear regression that examines how CHD rates (response variable) are predicted by cigarette consumption (predictor variable). What is the equation of the regression line you get? What are the units on the slope? Interpret the slope in this context.

(b) In Mauritania, the annual cigarette consumption is 30.3 cigarettes per adult per year. Predict the CHD rate using your model, and then look up the CHD rate online. Discuss what you find and why it might be happening.

Hint List:

- a: As you type the data into R using brute force, try to create an organized layout.

Solution:

(a) We begin by entering the data and running the linear model in R:

```
cig = c(3900, 3350, 3220, 3220, 2790, 2780, 2770,
        2290, 2160, 1890, 1810, 1800, 1770, 1700,
        1680, 1510, 1500, 1410, 1270, 1200, 1090)
CHD = c(256.9, 211.6, 238.1, 211.8, 194.1, 124.5, 187.3,
        110.5, 233.1, 150.3, 124.7, 41.2,  182.1, 118.1,
        31.9,  114.3, 144.9, 59.7,  126.9, 43.9,  136.3)
smoking = data.frame(cig = cig, CHD = CHD)
model = lm(CHD ~ cig, data = smoking)
summary(model)
```

From the readout, we find $\widehat{\text{CHD}} = 15.77115 + 0.06010 \cdot \text{cig}$. The slope of 0.06010 is $\dfrac{\Delta y}{\Delta x}$, so it's units are $\dfrac{y \text{ units}}{x \text{ units}}$. Here, that is $\dfrac{\text{deaths from CHD per 100000 people per year}}{\text{cigarettes consumed per person per year}}$. This means that if each person in a country were to consume 1 more cigarette per year, then 0.06010 more people would die from CHD each year for every 100000 people in the country.

(b) $\widehat{\text{CHD}} = 15.77115 + 0.06010 \cdot 30.3 \approx \boxed{17.592 \text{ deaths/100000 people per year}}$. Online, I get around 97.58 CHD deaths per 100000 people per year. Our result may be off because our model is built using data where the cigarette consumption is in the range $[1090, 3900]$. Since 30.3 is far outside this zone, we are extrapolating. Extrapolation rests on the assumption that any observed trend continues outside the range of your data, which may or may not be true. Another possibility is that we are using a current value for the CHD rate of Mauritania, but the given data are from a much different time. This is clear because the cigarette usage rate of the US feels far too high and because the table mentions "West Germany" which hasn't been a country since 1990.

Commentary: There were a few goals for this problem. First, I wanted you to feel frustrated by the data entry in part a. For me, there were several causes of that frustration. For starters, I don't enjoy the act of copying numbers from one location to another. If this describes you too, it is important you find a job that engages your brain more deeply. Additionally, I recognized that I was using an approach (brute-force data entry) that wouldn't work with a larger data set (i.e., my approach lacked scalability). Computer science majors might know this feeling as well. If you've ever run an inefficient algorithm overnight (instead of dreaming up a more efficient algorithm), you know what I'm talking about. If our data set were much bigger, we'd need to use some form of optical character recognition (OCR) to automate the image-to-data conversion process.

Second, when you look at how I typed in the data values to R, you'll notice I organized my values into three rows of seven values. By doing this, I made it easier to ensure I copied all the values. It is important to build safeguards and checks into every cognitive act you perform (even something as simple as copying numbers from one location to another). For example, when I'm using the quotient rule on $\dfrac{f(x)}{g(x)}$ in Calculus, I always write the $\dfrac{}{g(x)^2}$ part of the answer first, as it is the thing I know I'm most likely to forget (especially when the functions $f(x)$ and $g(x)$ get complicated, making the

numerator of the quotient rule especially attention-grabbing).

Third, I wanted a setting where the units were quite complicated. When you talk about countries, you can't just discuss total cigarette consumption or the number of people who die from CHD. Since countries have different numbers of people, you must convert these ideas into rates (e.g., cigarettes per person), and in some cases, you want to scale those rates so the numbers you are working with are more easily expressed. Here, the CHD rate is per 100000 people instead of per person so that we can use the number 256.9 (CHD deaths per 100000 people per year) for the US instead of 0.002569 (CHD deaths per person per year). To make things even more confusing, the units on the slope in part a are actually a ratio of these two rates. Demographic units can be quite strange, and you need to spend time working in these unit systems to get comfortable with them.

Finally, I wanted to create a problem that asked you to build a linear model but never explicitly asked you to *graph* the data. Do the data I've given even follow a linear pattern?! While it might seem unbelievable, it is quite common for people to run analyses without ever visualizing their data. The exploration of a data set is a cognitive act with great freedom. Creating a linear model, in contrast, is very routinized. Many people prefer routines over exploration, for they know exactly how to proceed. You can see this phenomenon in cooking (following a recipe vs. inventing a meal using the food in the fridge) and even TV viewing (re-watching a show for the tenth time instead of searching for something new). In statistics, both exploration and routine analysis are necessary.

Commentary Themes

Throughout the commentaries in this book, I have shared ideas that I wish someone had explicitly told me during my own academic journey. Below, I reiterate these themes and list the problems that best highlight them. While it's natural to focus on the particular content a course teaches you, there's far more gold in them thar hills. Courses can teach you about how to be a better scholar and, more generally, a better citizen. As such, I've grouped the themes based on their focus: subject, scholar, or citizen.

Themes about the subject itself:

- Probability theory is great at modeling the world: 15, 16, 17, 22, 24, 28, 30, 34, 44, 47, 63, 65
- Don't demand that modeling be perfect: 1, 35, 39
- Look at the world through a pair of statistical glasses: 34, 44, 68, 77
- Statistics isn't a perfect quantitative subject: 49, 60, 61, 62, 73, 75
- Think carefully about assumptions and the real world: 13, 15, 23, 56, 59, 66
- Use computer simulation to model complexity and check solutions: 14, 21, 33, 37, 64
- Most people, even experts in other fields, know very little statistics: 52, 55
- Effect size is usually more important than statistical significance: 54
- Learn to spot statistical abuse: 55, 56, 60, 62, 65
- Statistics does not prove the theories behind your CIs and HTs: 59, 65, 66, 67, 73, 75
- Sometimes we want to end up at the null hypothesis: 72

Themes about becoming a better scholar:

- Don't fear casework and grindy solutions: 4, 6, 10, 20, 33
- Get good at algebra and prerequisite material: 3, 6, 10, 19, 20, 29, 31, 33, 36, 42, 50, 51
- Internalize standard problem types for each new topic you learn: 30, 34, 58, 66
- Search online for clarifications: 1, 7, 14, 15, 45, 61
- Embrace yucky units and yucky answers: 2, 28, 62, 80
- Look for problems that excite you to determine your career direction: 61, 71, 80
- Solutions are about clarity and communication: 5, 7, 25, 51, 53, 55
- Particularize your general solutions to check they make sense: 10, 15
- Use technology and other tools appropriately: 2, 26, 31, 53, 62
- Flesh out details left to the reader: 33
- Reasons to take Math 109: 9, 11, 12, 78
- The nature of problems changes as you progress through school: 17, 20, 37
- Push back on academic narratives: 25, 28
- Don't fear open-ended/ill-structured problems: 18, 37, 41
- Simple answers may come from elegant solutions: 15
- Beware Off-By-One Errors (OBOEs): 28

- Embrace memorization: 42
- Immerse yourself in a subject's notation: 43
- Mathematical calculations can be very sensitive to small changes: 46, 57
- Seek out intellectual breadth in your academic training: 47
- Use procedures both forward and backward to deeply understand them: 50, 57, 60, 62, 64, 66, 70
- Organize the knowledge of the field you are studying: 58
- Work on both conceptual and procedural understanding: 60
- Build safeguards/checks into your cognitive architecture: 80
- Focus on the smallest details of a subject to ensure the large ones are definitely captured: 79

Themes about becoming a better citizen:

- Embrace multivariate thinking and beware confounding variables: 74, 75
- Avoid "always" and "never" ways of thinking: 63
- Work to be curious and ask good questions: 8, 25, 26, 36, 40, 72
- Harness the power of visualizations: 11, 13, 32, 80
- Learn to coordinate and switch between representational formats: 32, 36
- Look beyond the surface and seek deep understanding: 9, 38, 55, 67, 69, 76
- See new things you learn in greater mathematical contexts: 12
- Language is a mess: 27, 48, 62
- The value of things may be unclear at first: 40
- Be prepared for missing or extraneous information: 48
- Use metaphors liberally to improve understanding/communication: 60