The Math 181A 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*

July 23, 2025

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 too much 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 you'll need to work on using it responsibly. This skill is particularly important in the AI era, where technology can write your essays, design your experiments, summarize your readings, and yes, solve your math problems. It is time for you to learn how to grow academically in a setting where that growth is easily fake-able. If you don't develop actual skills beyond the scope of AI, you will be subsumed by it.

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. **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.

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.
- 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).

Lecture-Problem Mapping

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 those sitting in an actual 181A class, the course calendar will note when each problem is due. So, the below list is most useful for students who are working ahead. For those who have come to this book outside of UCSD's 181A class, this list will help you know when you're ready to tackle each problem. You can find the 181A class lectures on UCSD's podcasting repository website.

- Review of 180A (probability theory): #1-3
- MME: #4-8
- MLE (easier settings): #9-10
- MLE (harder settings): #11-17
- CIs (mean): #18-23
- CIs (proportion): #24-26
- Order Stats: #27-31
- Bias: #32-36
- Efficiency, MSE: #37-39
- FI, CRLB: #40-41
- Consistency, MSE-consistency: #42-45
- MLE Properties and CIs: #46-50
- HT (mean): #51-57
- HT (proportion): #58-61
- Duality, Errors: #62-64
- Power, Sample Size: #65-67
- HT (not mean/proportion): #68-69
- $F_{m,n}, T_n, \chi^2_n$: #70-75
- HT and CIs Using T_n: #76-80
- HT and CIs (variance): #81-83
- LRT, NPL: #84-88
- UMP Tests: #89-92
- GLR Tests: #93-94
- Bayesian Stats (easier settings): #95-97
- Bayesian Stats (harder settings) #98-100

Problems, Hints, Solutions, and Commentaries

1. The simplest random variable (RV) follows the Bernoulli distribution. This is a RV with two possible values: success (which we think of as 1), which appears with probability p, and failure (0), which appears with probability $1 - p$.

   (a) Explain why the pmf for $X \sim Bernoulli(p)$ can be written in this surprising way:

   $$f(x; p) = p^x (1 - p)^{1-x}$$

   with $f(x; p) = 0$ if $x \neq 0, 1$.

   (b) For many students, the above pmf feels like magic. Explain how you can come up with this formula if you happen to know the pmf for the $Binom(n, p)$ distribution.

   (c) Explicitly calculate the mean and variance of the Bernoulli RV with parameter p using the formulas for mean and variance from your probability theory course.

   (d) If $X_1, \ldots, X_n$ are iid (independent and identically distributed) $Bernoulli(p)$ RVs, and $Y \sim Binom(n, p)$ is Binomial, write a formula that relates Y and the X_is. Then, explain how the formula can help you easily remember the mean and variance of a Binomial RV.

Hint List:

- a. What happens when you plug $x = 1$ into the formula?
- b. What value could you set n to in order to transform Binomial into Bernoulli?
- c. If you are struggling, begin by memorizing the mean and variance formulas for a general discrete RV from your probability course.
- d. If you flip a bunch of coins and report out the number of heads, what mathematical operation do you perform to build from the individual results to the full heads count?

Solution:

(a) Notice that $f(1;p) = p^1(1-p)^{1-1} = p$ and $f(0;p) = p^0(1-p)^{1-0} = 1-p$. So, f is really

$$f(x;p) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$$

which is the Bernoulli pmf.

(b) Notice that if $n = 1$, then $Binom(n,p)$ collapses to $Bernoulli(p)$ since we are doing a single trial, rather than many trials. Letting $n = 1$ in the pmf of $Binom(n,p)$ gives

$$g(x;p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{1}{x} p^x (1-p)^{1-x}$$

Since $0 \leq x \leq n = 1$, we see $\binom{1}{x} = 1$, and $g(x;p)$ is simply $f(x;p)$ from part (a).

(c) Let $X \sim Bernoulli(p)$. We have:

$$E[X] = \sum_{x \in \text{support}} x \cdot f(x;p) = 0 \cdot f(0;p) + 1 \cdot f(1;p) = 0 \cdot (1-p) + 1 \cdot p = \boxed{p}$$

$$\begin{aligned} Var[X] &= \sum_{x \in \text{support}} (x-\mu)^2 \cdot f(x;p) = (0-p)^2 \cdot f(0;p) + (1-p)^2 \cdot f(1;p) \\ &= p^2(1-p) + (1-p)^2 p = p(1-p)[p + (1-p)] = \boxed{p(1-p)} \end{aligned}$$

(d) Set $Y = X_1 + X_2 + \cdots + X_n$. Note that each X_i is either a 0 or a 1 (failure or success), so adding the X_is will simply count the total number of successes, which is what a Binomial RV does. Hence,

$$E[Y] = E[X_1] + E[X_2] + \cdots + E[X_n] \overset{\text{id dist}}{=} E[X] + E[X] + \cdots + E[X] \overset{\text{(c)}}{=} \boxed{np}$$

$$Var[Y] \overset{\text{ind}}{=} Var[X_1] + Var[X_2] + \cdots + Var[X_n] \overset{\text{id dist}}{=} Var[X] + Var[X] + \cdots + Var[X] \overset{\text{(c)}}{=} \boxed{np(1-p)}$$

Commentary: Why start Math 181A with this problem? Well, it has a lot to teach us. Part (a) shows that some piecewise functions can be written in a way that avoids the piecewise look. This will be useful in the future, and we'll see this idea again when indicator functions are used to suppress the

piecewise look of continuous random variables. Part (b) reveals an interesting thing we sometimes see in math: For some setups, it is possible to take a more complex formula and "flatten" it (or trivialize it) to something we already know by plugging in a particular value. In this case, we flatten the Binomial RV to a Bernoulli RV by letting $n = 1$. The same idea shows up when you learn the formula for the length of the long diagonal across opposite corners of a rectangular box, $d = \sqrt{a^2 + b^2 + c^2}$. Here, if we (literally) flatten the box by setting the height c to 0, our box becomes a rectangle, and our formula becomes $d = \sqrt{a^2 + b^2}$, a version of the Pythagorean Theorem.

Part (c) reminds you that foundational facts/definitions from probability theory will reappear frequently in this course. Indeed, being successful in Mathematical Statistics requires a very long list of prerequisite knowledge, which is why it is usually first taken by seniors and graduate students. Said differently, if you have any weaknesses in your past, this course will quickly expose them. Finally, part (d) shows you that you can memorize the expectation and variance of the Bernoulli and build up to the Binomial. I find it easier to memorize less but to build ways to quickly re-derive related facts when needed. A similar strategy is useful for memorizing the expectation and variance of the Negative Binomial distribution (building from the Geometric) and the Gamma distribution (building from the Exponential). Seeing connections across the famous distributions pays huge dividends moving forward.

2. Let $X_1, X_2, \ldots, X_n$ be an iid sample from a random variable X with finite mean and finite variance.

   Recall that $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$. Show:

   (a) $E[\,\overline{X}\,] = E[X]$

   (b) $Var[\,\overline{X}\,] = \dfrac{Var[X]}{n}$

   In your solution for each part, explicitly mention where the "independence" (from iid) is actually needed and where the "identically distributed" (from iid) is actually needed. Finally, **memorize these two facts**; we will need them almost every day moving forward.

Hint List:

- a. If A and B are RVs, what is $E[A+B]$?
- b. If A is a RV and c is a constant, double check your simplification of $Var[cX]$.

Solution:

(a)

$$E[\,\overline{X}\,] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{1}{n}\left(E[X_1] + E[X_2] + \cdots + E[X_n]\right)$$

Since the X_is are identically distributed, this means that $E[X] = E[X_1] = \cdots = E[X_n]$. Hence, we get

$$\frac{1}{n}\left(E[X] + E[X] + \cdots + E[X]\right) = \frac{1}{n} \cdot nE[X] = E[X]$$

Notice that we don't need the independence of the X_is to do this part. In particular, $E[A+B] = E[A] + E[B]$ for dependent and independent RVs. Contrast this with part (b).

(b)

$$\begin{aligned} Var[\,\overline{X}\,] &= Var\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{1}{n^2} Var[X_1 + X_2 + \cdots + X_n] \\ &\overset{\text{ind}}{=} \frac{1}{n^2}\left(Var[X_1] + Var[X_2] + \cdots + Var[X_n]\right) \\ &\overset{\text{id dist}}{=} \frac{1}{n^2}\left(Var[X] + Var[X] + \cdots + Var[X]\right) \\ &= \frac{1}{n^2} \cdot nVar[X] = \frac{Var[X]}{n} \end{aligned}$$

Note that we need *independence* to do $Var[A+B] = Var[A] + Var[B]$, and as in part (a), we need *identically distributed* to wash away the subscripts.

Commentary: There are two key aspects of this problem. First, it is important in mathematics to become *detail-oriented*. Most technical disciplines (e.g., lab sciences, engineering, computer science) are very sensitive to small errors, and hence, we must be quite precise when engaging with them. One way to develop this in math is to ask why we need all the things we are assuming. Here, you've been asked to look carefully at how the iid assumption is needed/not needed to establish the facts in parts (a) and (b). One reason this matters is that the iid assumption is a very strong one, and so it is natural to question if it is really needed. Indeed, many polls/studies get things wrong because they require iid data to draw statistical conclusions, and yet, the actual data collected are not truly iid.

The second key aspect has to do with *efficiency*. The end of this problem asks you to memorize the results in parts (a) and (b). Each time you memorize something in life, you must spend the energy required to do this and devote some of your limited brain resources to the memorized content. That's the con of memorization. The pro is that you now have instantaneous access to some piece of information that can make your daily functioning more efficient. For each person, there comes a point where a given piece of info is worth memorizing. Each of you has decided that memorizing $5 \cdot 7 = 35$ is worth the brain space/efficiency trade-off. The reason I added this problem to the 181 Problem Book is that I'm trying to *explicitly* tell you that these two facts are worth the trade-off. If you choose not to memorize them, you will have to "reinvent the wheel" every time you need them, which is massively inefficient.

3. The "Kernel Technique" (KT): One of the most helpful tricks in mathematical statistics is to use the fact that all probability density functions must integrate to 1 over their support. That is $\int_{\text{support}} f(x)\,dx = 1$. For example, if we take $X \sim Exp(\lambda = 4)$, then we know $\int_0^\infty 4e^{-4x}\,dx = 1$ since the pdf is $f(x) = 4e^{-4x}$ and the support is $(0, \infty)$.

   Now, each pdf (or pmf) can be separated into two parts: the constant(s) and the terms with the variable (known as the "kernel"). For the exponential distribution above, the kernel is e^{-4x}. This distinction is useful because:

$$1 = \int_{\text{support}} f(x)\,dx = \int_{\text{support}} \text{constant} \cdot \text{kernel}\,dx \implies \boxed{\int_{\text{support}} \text{kernel}\,dx = \frac{1}{\text{constant}}}$$

   Use this formula to find the following expressions.

   (a) $\displaystyle\int_{-\infty}^{\infty} e^{-x^2/2}\,dx$

   (Do not try to use integration techniques from Calculus. Instead, think of a RV with a pdf whose kernel looks like $e^{-x^2/2}$ and use the above comments to immediately write the answer. Mention the RV in your answer on all parts!)

   (b) $\displaystyle\int_0^\infty x^4 e^{-3x}\,dx$

   (Integration by parts 4 times? Nope. The Kernel Technique.)

   (c) $\displaystyle\sum_{x=0}^{\infty} \frac{2^x}{x!}$

   (You could do this problem using Taylor series, but use the KT.)

   (d) $\displaystyle\sum_{x=4}^{\infty} \binom{x-1}{3} (0.7)^{x-4}$

   (This is very scary without the KT.)

Hint List:

- a. One way to help think of a RV is to look at the zone the integral is over. What continuous RVs have a support of $(-\infty, \infty)$?
- b. Loop through all the pdfs in your mind and see which ones have variable parts (i.e., kernels) that match the integrand.
- c. Because we have a sum instead of an integral, you should be thinking about **discrete** random variables here, not **continuous** random variables.
- d. You don't know many discrete RVs with support $\{4, 5, \dots\}$. This is a give-away for which one to use.

Solution:

(a) Note that $N(0,1)$ has pdf $\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(x-0)^2}{2\cdot 1}\right] = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ and support $(-\infty, \infty)$.
By the given formula: $\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \frac{1}{\text{constant}} = \frac{1}{1/\sqrt{2\pi}} = \boxed{\sqrt{2\pi}}$.

(b) Since the integral is over $(0, \infty)$, we think of CRVs with this support. The Gamma distribution has this support and kernel $x^{r-1}e^{-\lambda x}$. To match this to the integrand, we let $r = 5$ and $\lambda = 3$. The constant on $Gamma(r = 5, \lambda = 3)$ is $\frac{\lambda^r}{\Gamma(r)} = \frac{3^5}{\Gamma(5)} = \frac{3^5}{4!}$. By the given formula, our integral must equal the reciprocal of this value, which is $\frac{4!}{3^5} = \boxed{\frac{8}{81}}$.

(c) Based on the hint, we need a DRV with support $\{0, 1, 2, \dots\}$ because of the index bounds on the sum. The Poisson distribution has this support and pmf $\frac{e^{-\lambda}\lambda^x}{x!}$. To get the kernel to match the summand, we let $\lambda = 2$. This creates the constant e^{-2}, which has reciprocal $\boxed{e^2}$.

(d) Since the index bounds are 4 and ∞, we think of the Negative Binomial distribution which has support $\{r, r+1, \dots\}$. This forces $r = 4$. With this info, the pmf is $\binom{x-1}{4-1}p^4(1-p)^{x-4}$. Matching the variable part to the given summand means that p must be 0.3. So, the pmf is now $\binom{x-1}{3}(0.3)^4(0.7)^{x-4}$. This has a constant of $(0.3)^4$, with reciprocal $\frac{1}{0.3^4} = \boxed{\frac{10000}{81}}$.

Commentary: This problem begins to show you the power that comes with memorizing the support and pmf/pdf of the major distributions. All four of these problems are really rough without this technique, but not especially hard with it.

Looking deeper, note that the integrand in part (a) does not even have an antiderivative using well-known ("elementary") functions. Don't believe me? Go to your favorite integral calculator or AI and ask it to do the indefinite integral $\int e^{-x^2/2}\,dx$. Your tech will either give you a message that it can't be done or use a mystery placeholder like erf (the "error function") as the answer. Typically, the way one evaluates a definite integral (one with bounds) is by first finding the antiderivative and then plugging in

the bounds ("Thanks, Fundamental Theorem of Calculus!"). But, if an antiderivative can't be found, as in part (a), you're in trouble.

Sometimes, the KT helps us be more efficient. As mentioned in part (b), we could use Integration by Parts four times (or Reduction Formulas) to find this integral. Doing so is exhausting and very error-prone. Every time I assign an Integral Calculus student a problem like this, I feel bad for the pain they must endure. Also, when doing very long calculations, the probability that a mistake appears somewhere heads to 100%. So, we need a better, faster approach. The KT offers this. Note, however, that it is not a silver bullet. It only works for integrals and sums whose arguments are the kernel of a well-known distribution, and importantly, we need the bounds to match the support of the distribution. For example, we could do part (a), but if we alter the problem to $\int_2^{17} e^{-x^2/2}\,dx$, then the KT is unhelpful.

Finally, be aware that the KT can be useful if you compete in Integration Bees (UCSD and MIT have famous ones) or math competitions like the Putnam. These tests often have definite integrals that immediately yield to the KT or where the KT is useful after some algebraic massaging.

4. Let $X_1, \ldots, X_n$ be iid from the distribution modeled by

$$f_X(x;\theta) = (\theta^2 + \theta)x^{\theta-1}(1-x) \text{ where } 0 < x < 1 \text{ and } \theta > 0$$

Note: We always assume a pdf is 0 outside of the zone specified. For example, here we assume $f_X(x;\theta) = 0$ if $x \leq 0$ or $x \geq 1$.

(a) Find the MME (method of moments estimate/estimator) for θ.

(b) Since the support of this pdf is $(0, 1)$, it is useful for modeling real-world phenomena, like percentages, that are between 0 and 1. Here is a nice example: When you give a person some medication, you might be interested in how much of the active ingredient is actually absorbed into the bloodstream. Ideally, we would like this value, known as the "bioavailability", near 100% (or 1). For a given drug, this value varies from person to person. Download the file Bioavailability.csv from Canvas/TritonEd which contains the drug absorption percentages for a single drug in 134 different volunteers. Read this data into R and use your result from part (a) to find the best fitting model. Finally, to see how good the model fit is, make a relative frequency histogram of the 134 absorption percentages and then overlay a graph of the model. Use a mixture of internet searches, AI, and R documentation to learn how to do these actions.

Hint List:

- a. This is not a distribution you have memorized, so begin by finding $E[X]$. If you get the answer 1 when doing the calculation, it is because you integrated the pdf (which must give 1) instead of $x \cdot$ pdf, which is how you find the expected value.
- b. Typing hist(..., freq=F) into R will make a histogram that has probability densities, rather than counts, on the vertical axis. This allows you to then overlap a pdf.
- b. To draw a graph in R, you can generate a lot of (x, y) pairs and draw lines between them. With enough points, you get a smooth curve. Keep the x values and y values in separate vectors. Generate the xs using the seq(...) command in R, and generate the ys using the pdf formula. The fact that R is a vectorized language makes it much easier to calculate the ys. Finally, calling the lines(...) command in R adds on top of an existing plot (here, the histogram).

Solution:

(a) We need

$$\begin{aligned} E[X] &= \int_0^1 x \cdot f(x)\,dx = \int_0^1 (\theta^2+\theta)x^\theta(1-x)\,dx = (\theta^2+\theta)\int_0^1 x^\theta - x^{\theta+1}\,dx \\ &= (\theta^2+\theta)\left[\frac{x^{\theta+1}}{\theta+1} - \frac{x^{\theta+2}}{\theta+2}\right]_0^1 = \theta(\theta+1)\left[\frac{1}{\theta+1} - \frac{1}{\theta+2}\right] \\ &= \theta(\theta+1)\left[\frac{1}{(\theta+1)(\theta+2)}\right] = \frac{\theta}{\theta+2} \end{aligned}$$

Setting the first theoretical moment (above) equal to the first sample moment ($\overline{X}$), we get the MME for θ:

$$\frac{\widehat{\theta}}{\widehat{\theta}+2} = \overline{X} \implies \widehat{\theta} = \overline{X}\widehat{\theta} + 2\overline{X} \implies \widehat{\theta} - \overline{X}\widehat{\theta} = 2\overline{X}$$

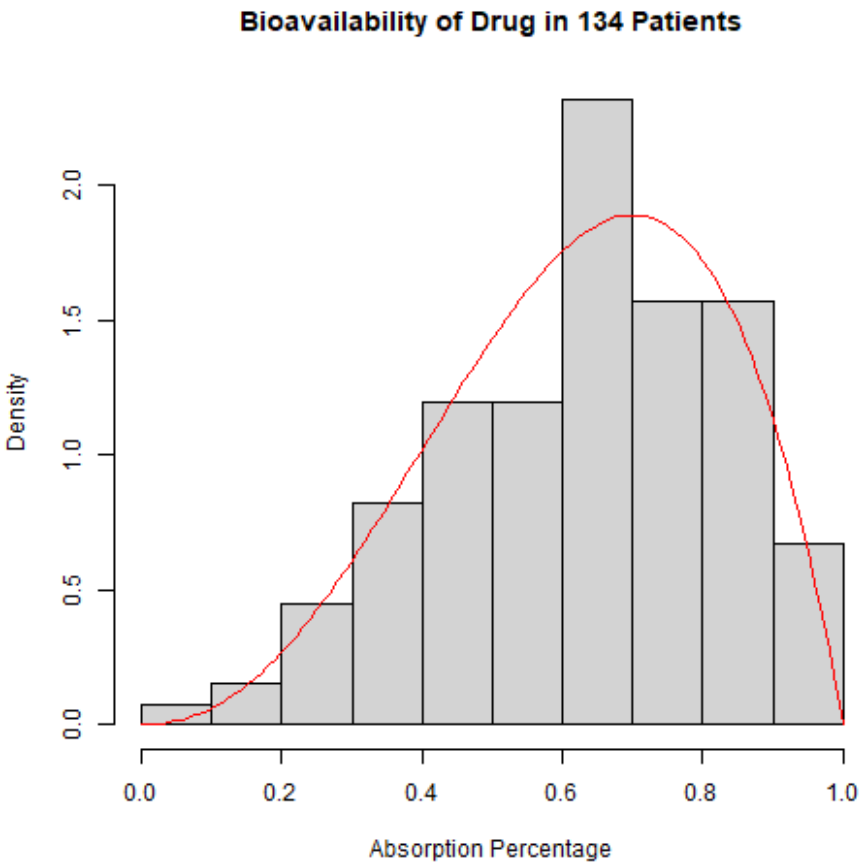
Factoring gives $(1-\overline{X}) \cdot \widehat{\theta} = 2\overline{X}$, or $\boxed{\widehat{\theta} = \dfrac{2\overline{X}}{1-\overline{X}}}$.

(b) We can use the below code:

```
data = read.csv("your file location")
percs = data$percentage

# find MME based on part (a)
thetaMME = 2*mean(percs)/(1-mean(percs))

# view data and model on top
hist(percs, freq=F, xlab = "Absorption Percentage",
     main = "Bioavailability of Drug in 134 Patients")
xs = seq(0, 1, by=0.01)
ys = (thetaMME^2+thetaMME)*xs^(thetaMME-1)*(1-xs)
lines(xs, ys, col="red")
```



Commentary: I assign this problem as an "algebraic warm-up". Here, we follow the standard MME procedure, but along the way, there are many algebraic details to get right (e.g., see the computation of $E[X]$). In addition, we have an algebraic challenge when finding the MME since the estimator $\widehat{\theta}$ appears in two places (numerator and denominator) rather than one. Isolating $\widehat{\theta}$ is similar to solving for y in $\frac{y}{y+2} = 3$. When you teach high school algebra, you have to decide if you'll give students problems of this type. More rigorous textbooks and teachers ask their students to learn this problem type; less rigorous/demanding ones don't. If you ever have the privilege of teaching across large portions of the math curriculum, you'll begin to see that each shortcut or omission at an earlier stage has negative consequences downstream. If you struggled on the last half of part (a), you experienced that directly in this moment. Remember: More than any other class I've taught, Mathematical Statistics reveals weaknesses/gaps from your past.

Another thing we might worry about in this problem is whether the estimator could have a problem since it involves a division, and division by 0 is undefined. Here, that would require $\overline{X} = 1$. Since we know $0 < X_i < 1$ for each i, it isn't possible to get $\overline{X} = 1$, and we are safe. Note that if I had set the support for X as $[0, 1]$ instead of $(0, 1)$, we could get $\overline{X} = 1$ (if $X_1 = X_2 = \cdots = X_n = 1$), but this would be a 0 probability event, and we don't usually worry about such fringe cases.

5. In the 2017 video game hit Legend of Zelda: Breath of the Wild, you must collect star fragments to upgrade your armor to the highest levels. You decide to explore the mechanic behind how these rare items are generated in the game. Suppose you have this partial knowledge:

   - A star fragment will appear once per night, sometime after 9 PM (using the in-game clock).
   - Once the clock reads θ (a particular, unknown time of day), star fragments no longer appear.
   - The game uses a *continuous* random number generator to decide on the spawn time for the star fragment where each time between 9 PM and θ is equally likely.

   It is important for the gaming community to learn what θ is because this helps users understand the game and saves people time: If you see the clock is past θ, you will stop waiting for the star fragment (which you missed) and plan on trying again the next night. To help the community, you plan to record the appearance time for star fragments on 6 random (in-game) nights. You get: x_1 =11:20 PM, x_2 = 1:20 AM, x_3 =12:20 AM, x_4 = 10:00 PM, x_5 = 1:05 AM, and x_6 = 11:55 PM. Find the MME for θ based on these six data points.

Hint List:

- You need to choose a random variable X to represent the spawn time for star fragments.
- The phrase "equally likely" in the third bullet point is useful for choosing your random variable.

Solution: Let $X \sim Unif(0, \theta)$ represent the spawn time of a random star fragment in Legend of Zelda: Breath of the Wild where 0 stands for 9 PM and θ is the number of hours after 0 (9 PM) when stars stop appearing. We know $E[X] = \dfrac{0+\theta}{2}$. So, equating moments gives $\dfrac{\widehat{\theta}}{2} = \overline{X}$, or $\widehat{\theta} = 2\overline{X}$.

If we convert the given data so they match our setup for X, we see that $X_1 = 11{:}20$ PM $= 2\frac{1}{3}, X_2 = 4\frac{1}{3}, X_3 = 3\frac{1}{3}, X_4 = 1, X_5 = 4\frac{5}{60}$, and $X_6 = 2\frac{55}{60}$. Thus, $\overline{X} = \dfrac{18}{6} = 3$ and $\widehat{\theta} = 2 \cdot 3 = 6$.

Converting back to our time of day gives our best guess for θ: 9 PM + 6 = $\boxed{\text{3 AM}}$.

Commentary: For those who are curious, the best current knowledge is that $\theta = 3$ AM. This is a nice example where parameter estimation is our only hope to make progress understanding the game. Nintendo is quite protective of the source code for their games. Buried somewhere in those millions of lines of code is a single line defining the in-game star fragment spawn stop time. I also like this problem because it is a nice example of where some parameter does have a known value, but this knowledge is restricted to some people on Earth (the programmers of the game). Often when we estimate parameters, it is because *no one* on Earth knows the true values. This happens when we try to model some phenomenon in nature (e.g., the mean and variance of normally-distributed fish lengths for a given species).

Also note that the real difficulty in this problem is setting up the RV. In the previous problem, you were simply given a RV to start from; here, you had to generate it yourself. When I first started teaching statistics, I used to walk around and try to "see" distributions behind everything I did: the amount that people were bench-pressing at the gym (probably normal), how many of the 11 lights I'd need to stop at on my way home from school (probably Binomial), how many student emails I'd get on a random day (probably Poisson), etc. In truth, probability distributions are everywhere; you just need to put on your statistical glasses and begin to see the world for what it is.

Finally, note that some people use the setup $X \sim Unif(9, \theta)$. This works as well. When you translate the given data to this setup, you use $X_1 = 11\frac{1}{3}, X_2 = 13\frac{1}{3}$, and so on.

6. Suppose a discrete RV is modeled by

$$p_X(x;\tau) = \begin{cases} \frac{\tau}{3}, & x = 1 \\ \frac{\tau}{6}, & x = 2 \\ \frac{\tau}{4}, & x = 3 \\ 1 - \frac{3\tau}{4}, & x = 4 \end{cases}$$

Suppose you observe the sample $x_1 = 4$, $x_2 = 3$, $x_3 = 4$, $x_4 = 2$, $x_5 = 2$, $x_6 = 2$, and $x_7 = 2$. Find the MME for τ. After this, find the set of values τ could actually take on given that $p_X(x;\tau)$ must be a valid pmf and list your answer in interval notation. (Note: It is possible that some data sets will give rise to an MME value that is outside the set of possible values for the parameter! This is not a problem with the above data, but it is one drawback to MME in general.)

Hint List:

- Why can't $\tau = -2$? Why can't $\tau = 4$?

Solution: We first find $E[X] = \sum_{x \in \text{support}} x \cdot p(x;\tau) = 1 \cdot \frac{\tau}{3} + 2 \cdot \frac{\tau}{6} + 3 \cdot \frac{\tau}{4} + 4 \cdot \left(1 - \frac{3\tau}{4}\right) = 4 - \frac{19\tau}{12}$.

Equating moments gives $\overline{X} = 4 - \frac{19\widehat{\tau}}{12}$, or $\widehat{\tau} = \frac{12}{19}\left(4 - \overline{X}\right)$.

The given data show $\overline{X} = \frac{4+3+4+2+2+2+2}{7} = \frac{19}{7}$. So, $\widehat{\tau} = \frac{12}{19}\left(4 - \frac{19}{7}\right) = \frac{12}{19} \cdot \frac{9}{7} = \boxed{\frac{108}{133}}$.

To be a valid pmf, we need each probability listed to be in $[0,1]$ (or $(0,1)$ if you want all the values listed to actually be possible!). We also need the probabilities listed to sum to 1, but this is guaranteed since $p(4;\tau)$ is $1 - \frac{3\tau}{4}$ and the other probabilities sum to $\frac{3\tau}{4}$. So, we need

$$\begin{aligned}
0 \le \frac{\tau}{3} \le 1 &\implies 0 \le \tau \le 3 \\
0 \le \frac{\tau}{6} \le 1 &\implies 0 \le \tau \le 6 \\
0 \le \frac{3\tau}{4} \le 1 &\implies 0 \le \tau \le \frac{4}{3} \\
0 \le 1 - \frac{3\tau}{4} \le 1 &\implies -1 \le -\frac{3\tau}{4} \le 0 \implies \frac{4}{3} \ge \tau \ge 0
\end{aligned}$$

Since all of these must occur, we need the intersection of these zones, giving the answer $\boxed{\left[0, \frac{4}{3}\right]}$, or $\boxed{\left(0, \frac{4}{3}\right)}$ if you're a pedant.

Commentary: This problem does two things. First, it reminds us that there is a difference between pmfs and pdfs. While both demand a function that is non-negative, pmfs are special in that the heights represent probabilities, and hence, must be less than or equal to 1. The height of a density function represents "likelihood" or "density". This idea is related to probability, but it is measuring how likely something is *relative* to other outcomes, and so the value need not be less than or equal to 1. I always remind myself of this fact by thinking of the density for $X \sim Exp(\lambda = 3)$. The pdf is $3e^{-3x}$ which has a height of $3 > 1$ when $x = 0$.

Second, the answer to this problem reminds us that intuition has its limitations. Did your intuition tell you the answer was $\widehat{\tau} = \frac{108}{133}$?! While there are some settings where intuition is valuable, probability theory/math stats is sometimes not one of these. Indeed, the answers are often horrible fractions that you'll never guess. When I did competition math in high school, our team feared the probability questions because we knew there was little hope of guessing our way into the answer. If you want to learn about settings where intuition is more reliable, check out Malcolm Gladwell's "Blink" or read about Daniel Kahneman's System 1 thinking vs. System 2 thinking.

7. It is important to note that the MME for a parameter is not a unique idea (despite me writing "**the** MME" on problems). Suppose $X_1, X_2, \ldots, X_n$ are iid from $X \sim Pois(\lambda)$.

   (a) Find an MME for λ using first moments of X (which is what people typically use).

   (b) Find an MME using second (!) moments of X. Then, see if the estimators in parts (a) and (b) give the same estimate for λ using the data $X_1 = 1, X_2 = 2, X_3 = 2$.

   Note: On exams, you will be told which moments (first, second, third, etc.) to use. When someone doesn't mention which moments should be/were used, assume first moments.

Hint List:

- a and b. Have you memorized the expectation and variance of a Poisson distribution yet?
- b. Recall that $Var[X] = E[X^2] - E[X]^2$.
- b. You should get a quadratic equation. It can't be solved by factoring, so what do you do then?

Solution:

(a) For $X \sim Pois(\lambda)$, we know $E[X] = \lambda$. Equating moments gives $\boxed{\widehat{\lambda} = \overline{X}}$.

(b) For second moments, we need to find $E[X^2]$. Recall that $Var[X] = E[X^2] - E[X]^2$, so we always have $E[X^2] = Var[X] + E[X]^2$. This fact is so useful, you should memorize it. For $X \sim Pois(\lambda)$, we know $Var[X] = \lambda$, and hence, $E[X^2] = \lambda + \lambda^2$. Equating second moments gives

$$\widehat{\lambda} + \widehat{\lambda}^2 = \overline{X^2} \implies \widehat{\lambda}^2 + \widehat{\lambda} - \overline{X^2} = 0$$

By the quadratic formula, we have $\widehat{\lambda} = \dfrac{-1 \pm \sqrt{1 - 4\left(-\overline{X^2}\right)}}{2}$. Since the Poisson rate parameter, λ, is always positive (to make the pmf non-negative), we must choose the $+$ from the $\pm$, so we get $\boxed{\widehat{\lambda} = \dfrac{-1 + \sqrt{1 + 4\overline{X^2}}}{2}}$.

For the data $X_1 = 1, X_2 = 2, X_3 = 2$, we have $\overline{X} = \dfrac{1+2+2}{3} = \dfrac{5}{3}$ and $\overline{X^2} = \dfrac{1^2 + 2^2 + 2^2}{3} = 3$. Thus, our estimate using (a) is $\widehat{\lambda} = \dfrac{5}{3}$, and our estimate using (b) is $\widehat{\lambda} = \dfrac{-1 + \sqrt{1 + 4 \cdot 3}}{2} = \dfrac{-1 + \sqrt{13}}{2}$. These values are unequal, so we do have different estimators!

Commentary: One component of being a great scholar is learning to ask questions about the world. Surprisingly, we do very little of this in standard schooling. Usually, students are simply handed a list of questions and told to solve them. Rarely, if ever, are you asked to look at a situation and come up with questions about what you see. Said simply, we don't develop curiosity in students. In my opinion, this is one reason kids hate school: We bring in curious, creative children in kindergarten and spend over a decade starving them of this impulse. Did you feel this in your academic journey?

This problem was designed to spark that dormant curiosity. After doing parts (a) and (b), you can't help but ask: Well, if they aren't the same, which one should I use? Answering that question will be our focus for several weeks in this course. Essentially, we need ways to compare different estimators. Just as you might compare different cell phone cases on price, durability, appearance, etc., we will develop a set of *metrics* to gauge the quality of an estimator. Already, this problem hints at a metric: messiness of derivation. Notice how much harder it was to use second moments compared to first moments. In general, the higher the moments used, the more brutal the initial derivation is.

8. Engineers will often use this distribution to model the lifetime of electronics:

$$f_Y(y;\alpha,\beta) = \alpha\beta y^{\beta-1}e^{-\alpha y^\beta}, \text{ where } y>0, \alpha>0, \beta>0$$

Assuming $Y_1,\ldots,Y_n$ are iid from this distribution, find the MME for α using first moments assuming that β is fixed (known).

Note: This problem might be the first time in your life that you've seen exponents inside exponents. Often, such expressions can be tough to read, so mathematicians will use the notation $\exp(a)$ to mean e^a. With this notation, we can write the pdf as $\alpha\beta y^{\beta-1}\exp\left(-\alpha y^\beta\right)$, which is a little clearer.

Hint List:

- After setting up an integral, try a u-substitution with $u = \alpha y^\beta$. Remember to switch the y bounds to u bounds, and to switch from dy to du.

Solution: We must find $E[Y] = \displaystyle\int_0^\infty y \cdot \alpha\beta y^{\beta-1} e^{-\alpha y^\beta}\, dy$.

If we let $u = \alpha y^\beta$, we see $du = \alpha\beta y^{\beta-1}\, dy$. Our new lower bound of integration is $u = \alpha \cdot 0^\beta = 0$, and our new upper bound is $u = \alpha \cdot \infty^\beta = \infty$ (or you can use limits to do this rigorously if you are worried about treating ∞ like a number). The most difficult part of the integral to switch to u is the lone y that comes from the expectation formula. For this, notice that

$$u = \alpha y^\beta \implies \frac{u}{\alpha} = y^\beta \implies \left(\frac{u}{\alpha}\right)^{1/\beta} = y$$

So, $E[Y] = \displaystyle\int_0^\infty \frac{u^{1/\beta}}{\alpha^{1/\beta}} e^{-u}\, du = \frac{1}{\alpha^{1/\beta}} \int_0^\infty u^{\left(\frac{1}{\beta}+1\right)-1} e^{-u}\, du = \frac{\Gamma\left(\frac{1}{\beta}+1\right)}{\alpha^{1/\beta}}$.

For the last equals sign, you can either use the definition of $\Gamma(r)$ (see Lecture 1) or invoke the Kernel Technique on $Gamma\left(r = \frac{1}{\beta}+1, \lambda = 1\right)$.

Finally, equating first moments gives $\overline{Y} = \dfrac{\Gamma\left(\frac{1}{\beta}+1\right)}{\widehat{\alpha}^{1/\beta}}$, or $\boxed{\widehat{\alpha} = \left[\dfrac{\Gamma\left(\frac{1}{\beta}+1\right)}{\overline{Y}}\right]^\beta}$.

Commentary: There are several reasons I assign this problem. First, and most importantly, when we model things in the world, we need to be honest about which distributions are actually appropriate. Students are often told in intro stat classes that lifetimes of electronics are exponentially distributed. But this doesn't mesh with what we experience everyday. The exponential distribution is highest at 0 and then decreases. Are your electronics most likely to break when you first get them or at some point in the future as components begin to fail? If you graph the above pdf for different values of α and β, the typical shape is one that has no height at 0, rises up, and then decays away. This meshes with my personal experience for failure rates.

Second, I want you to know that there's no hiding from complexity. Mathematical Statistics is a no-joke, serious subject. In this problem alone, we need to know the formula for expectation from Probability Theory, the technique of u-substitution from Integral Calculus, formulaic manipulations (to get y) from Algebra, and the Gamma function or Kernel Technique from earlier in the course. On top of this, you have to have the courage to write an answer that looks very messy. I used to have 5 to 10 students per quarter come to office hours who refused to trust the above messy-looking answer!

Finally, this answer should inspire a natural question: How would I evaluate the expression $\Gamma\left(\frac{1}{\beta}+1\right)$ for even a simple value like $\beta = 2$? We see from Lecture 1 that larger arguments can be reduced via the property $\Gamma(x+1) = x \cdot \Gamma(x)$. So $\Gamma\left(\frac{3}{2}\right) = \frac{1}{2} \cdot \Gamma\left(\frac{1}{2}\right)$. But what is $\Gamma\left(\frac{1}{2}\right)$?!

9. Let Y be a CRV with density $f_Y(y;\theta) = \dfrac{2y}{\theta}e^{-y^2/\theta}$ where $y > 0, \theta > 0$. Given a random sample $y_1, \ldots, y_n$:

   (a) Find the MLE for θ. (As always with one parameter, you must check the second derivative condition!)

   (b) Find the MME for θ using first moments. You should get a different answer from part (a), hence showing **the MME and MLE may be different**.

Hint List:

- a. The expression for $\ell''(\theta)$ is quite messy and not always negative (like some class examples). So, you'll need to actually plug in the MLE and see that ℓ'' is negative at that value.
- b. If you follow the standard steps for finding the MME, you will encounter the same integration difficulties as in the previous problem. Try using similar strategies as you did there.

Solution:

(a) Let $L(\theta) = \prod_{i=1}^{n} f(y_i;\theta) = \prod \frac{2y_i}{\theta} e^{-y_i^2/\theta}$. It is ok to omit the bounds on a product or sum once the reader knows what they are.

So, $\ell(\theta) = \sum \left[\ln 2 + \ln y_i - \ln\theta - \frac{y_i^2}{\theta}\right] = C - n\ln\theta - \frac{1}{\theta}\sum y_i^2$.

$$\ell'(\theta) = \frac{-n}{\theta} + \frac{1}{\theta^2}\sum y_i^2$$

$$\frac{n}{\widehat{\theta}} = \frac{1}{\widehat{\theta}^2}\sum y_i^2$$

$$\widehat{\theta}_{\text{MLE}} = \frac{1}{n}\sum y_i^2 = \overline{Y^2}$$

$$\ell''(\theta) = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum y_i^2$$

$$\ell''(\widehat{\theta}_{\text{MLE}}) = \frac{n}{\left(\frac{\sum y_i^2}{n}\right)^2} - \frac{2\sum y_i^2}{\left(\frac{\sum y_i^2}{n}\right)^3}$$

$$= \frac{n^3}{\left(\sum y_i^2\right)^2} - \frac{2n^3}{\left(\sum y_i^2\right)^2} = \frac{-n^3}{\left(\sum y_i^2\right)^2} < 0$$

Since $\ell''(\widehat{\theta}_{\text{MLE}}) < 0$, we have found a max. So, $\boxed{\widehat{\theta}_{\text{MLE}} = \overline{Y^2}}$.

(b) To find the MME, we could do the usual process as the hint suggests, or we could note that the pdf in this problem is the pdf from the previous problem with $\beta = 2$ and $\alpha = \frac{1}{\theta}$. From the last problem, we know

$$\frac{1}{\widehat{\theta}} = \widehat{\alpha} = \left[\frac{\Gamma\left(\frac{1}{\beta}+1\right)}{\overline{Y}}\right]^{\beta} = \left[\frac{\Gamma\left(\frac{1}{2}+1\right)}{\overline{Y}}\right]^{2}$$

Thus, $\boxed{\widehat{\theta}_{\text{MME}} = \frac{\overline{Y}^2}{\Gamma(3/2)^2}}$, which looks different than the answer from (a).

Commentary: The main reason I assign this problem is to answer a question the curious student will naturally raise: Will the MME and MLE always give the same result? This problem reveals that the answer is definitely "no". Now, we could answer that question using this logic: If the MME and MLE always gave the same thing, then it would be inefficient to teach both in school, and since time is limited, schools tend not to waste it. Hence, they're different. This type of logic is not based on the mathematics itself, but rather, on the larger setting in which that mathematics is done. You might

already rely on such ways of thinking to help make decisions: There hasn't had an answer of C on the last 20 multiple choice questions, and since I don't know the math on this one, I'll guess C. Ultimately, as a scholar, you want to avoid these sketchier ways of reasoning and just explore the mathematics itself. In this case, you'd either want to prove that the MME and MLE are always the same, or try to find a single counterexample, which this problem provides.

I also assign this problem and place it immediately after the previous one to see if students will notice similarities between the pdfs and leverage previous work to gain efficiency (as seen in the solution). If you want to spend your life inefficiently re-inventing the wheel, you can do this, but you're placing a ceiling on how high you can academically rise and on how much you can complete in a given time span. Have you ever had trouble finishing a test before? If so, it may be because you have too much inefficiency built into your cognition. Are you a musician who has struggled playing a fast passage on your instrument? Again, inefficiency in how you use your hands or breath might be the reason. While the solution to part (b) is a particularly elegant example of efficiency, it is a reminder that people at the highest ends of human performance (in music, athletics, academia) are incredibly focused on efficiency in every action they take. You must be too.

Finally, for those who read and followed up on the Commentary in the previous problem ("What is $\Gamma\left(\frac{1}{2}\right)$?!"), you were rewarded with an extra simplification in part (b):

$$\widehat{\theta}_{\text{MME}} = \frac{\overline{Y}^2}{\Gamma(3/2)^2} = \frac{\overline{Y}^2}{\left(\frac{1}{2}\cdot\Gamma\left(\frac{1}{2}\right)\right)^2} = \frac{\overline{Y}^2}{\left(\frac{1}{2}\sqrt{\pi}\right)^2} = \frac{4\overline{Y}^2}{\pi}$$

10. One common distribution that appears in branching process theory is a DRV with pmf

$$f_X(x;\mu) = \frac{e^{-\mu x}(\mu x)^{x-1}}{x!} \text{ where } x \in \{1, 2, \ldots\} \text{ and } \mu \in (0, 1)$$

(a) Find the MLE for μ given iid $X_1, \ldots, X_n$. Then, find the MLE for the particular data $x_1 = 2, x_2 = 1, x_3 = 6$.

(b) Using Desmos, draw a graph of the likelihood function (not log-likelihood) for the data $x_1 = 2, x_2 = 1, x_3 = 6$. It should be maximal at the μ value you found in part (a). Include a sketch of the graph from Desmos (or a screenshot if you're tech-fancy). If you click on the wrench icon in the upper-right of Desmos, you can change the bounds on the x and y axes so the graph is clearly visible.

Hint List:

- a. It helps to write $\ln(\mu x_i)$ as $\ln \mu + \ln x_i$ during the log-likelihood steps.
- b. You can define multivariate expressions in Desmos by writing $f(a, b) = \ldots$ and then using this definition in future Desmos entries.

Solution:

(a) First, we have $L(\mu) = \prod_{i=1}^{n} \frac{e^{-\mu x_i} (\mu x_i)^{x_i - 1}}{x_i!}$. Taking logs gives

$$\begin{aligned}
\ell(\mu) &= \sum \left[-\mu x_i + (x_i - 1) \ln(\mu x_i) - \ln(x_i!)\right] \\
&= -\mu \sum x_i + \ln \mu \sum (x_i - 1) + \sum (\ln x_i)(x_i - 1) - \sum \ln(x_i!) \\
&= -\mu \sum x_i + \ln \mu \sum (x_i - 1) + C
\end{aligned}$$

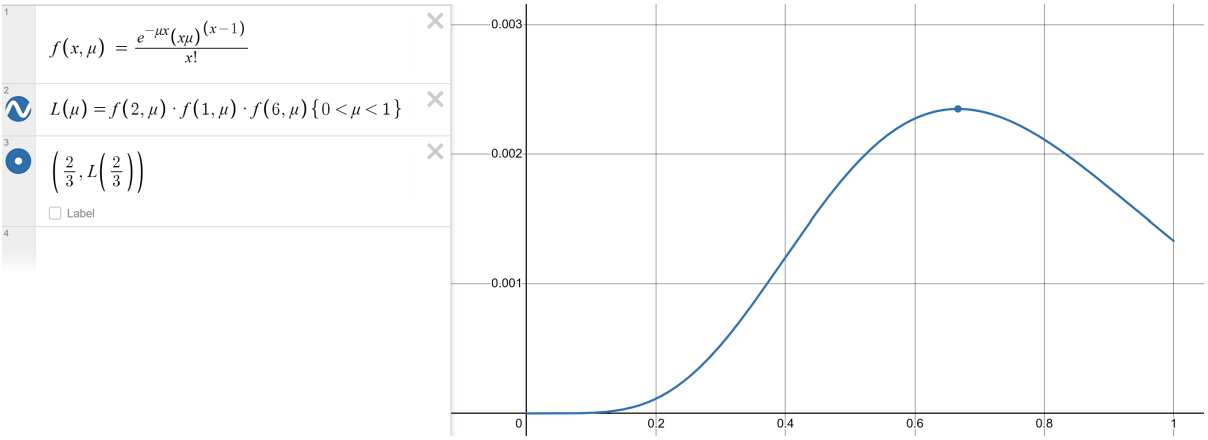
$$\begin{aligned}
\ell'(\mu) &= -\sum x_i + \frac{1}{\mu} \sum (x_i - 1) \\
0 &= -\sum x_i + \frac{1}{\widehat{\mu}} \left[\sum (x_i) - n\right] \\
\widehat{\mu} \sum x_i &= \sum (x_i) - n \\
\widehat{\mu} &= 1 - \frac{n}{\sum x_i} = 1 - \frac{1}{\overline{X}}
\end{aligned}$$

$$\ell''(\mu) = -\frac{1}{\mu^2} \sum (x_i - 1)$$

Note that $\ell'' \leq 0$ for all μ

Since $\ell'' \leq 0$ for all μ, we have found a max, and $\boxed{\widehat{\mu}_{\text{MLE}} = 1 - \frac{1}{\overline{X}}}$ (see fourth point in commentary). For our data, $\overline{X} = \frac{2+1+6}{3} = 3$, so $\widehat{\mu}_{\text{MLE}} = 1 - \frac{1}{3} = \boxed{\frac{2}{3}}$.

(b) See below. You can limit the domain of the likelihood to $(0, 1)$ using the brace notation.



Commentary: I like this problem for several reasons. First, you need to be very detail-oriented to get this correct. In essence, we have an exhausting bookkeeping exercise. The above solution uses eight different letters ($L, \ell, \mu, i, n, e, x, C$), multiple derivatives, and switches between specific data and theoretical data! Second, this problem shows that there different ways to write your answer, and some may be more interpretable than others. In the final steps of finding $\widehat{\mu}$, we could have written our answer as $\dfrac{\sum(x_i) - n}{\sum x_i}$ (which I skipped above), or $1 - \dfrac{n}{\sum x_i}$ (slightly nicer), or $1 - \dfrac{1}{\overline{X}}$. This last interpretation shows that the average is important to the MLE process, and if we knew about branching theory (the source of the pmf), we might even be able to make sense of the reciprocal and "1−" from the formula.

Third, I like this problem as a contrast to the last. Here, we find ℓ'' is always non-positive, so we don't need to plug in the MLE to show it is negative at a particular place (efficiency!). Fourth, we do have a small issue with our answer. Since the data are in the support $\{1, 2, 3, \dots\}$, there is no way for $\overline{X}$ to equal 0, so we can't get a division by 0 in the MLE formula. But, notice that if all your data are 1, then $\overline{X} = 1$, and so $\widehat{\mu} = 1 - \dfrac{1}{1} = 0$. This conflicts with the original pmf definition stating that $0 < \mu$. Here, again, we have a fringe edge case that pedants will get very excited about. In truth, while there is a non-zero probability it occurs (notice we have a DRV here, so $P(X_i = 1) > 0$), if *all* your data from a *random* process are the same value, perhaps you don't have a random process in the first place (or perhaps you need to collect some more data so you actually see variation)!

Finally, one thing that I was not consciously aware of for a long time about MLE is that we take the derivative with respect to the parameter variable (here, μ), not the letter used for the random variable (here, x). Note that our random variable is *discrete*, so $x \in \{1, 2, 3, \dots\}$, and it doesn't even make sense to differentiate with respect to x. In contrast, $\mu \in (0, 1)$, a smooth zone of possibilities, so we can differentiate with respect to μ. Also, note that in the MLE process, the data are known and the parameter is unknown, so it only makes sense to differentiate with respect to the parameter. Because of all this, parameters are almost always defined over smooth zones. Said differently: while a random variable might have discrete or continuous support, a parameter will almost always have continuous support. For a rare example where the parameter has discrete support, see N from the hat-and-tokens example at the start of Lecture 2. To use MLE in such a case requires avoiding the derivative altogether, or acting like the support is continuous and fiddling with the answer at the end to honor the discrete reality of the parameter.

11. Let $X_1, \ldots, X_n$ be a random sample ($n > 1$) from the continuous random variable X with pdf

$$f(x;\theta) = \frac{\theta}{x^2} \text{ where } x \geq \theta \text{ and } \theta > 0$$

(a) Explain why method of moments estimation (MME) is of no help *no matter which moments* (first, second, third, etc.) you try to use!

(b) Show that the maximum likelihood estimator for θ is $\widehat{\theta}_{\text{MLE}} = \min X_i$.

Hint List:

- a. Find $E[X]$ and then separately explore $E[X^k]$ where $k > 1$ is a positive integer.
- b. Notice that the support of the pdf (here, $x \geq \theta$) depends on the parameter θ. Hence, indicator functions should be used.

Solution:

(a) Note that $E[X] = \displaystyle\int_\theta^\infty x \cdot \frac{\theta}{x^2}\,dx = \theta \cdot \ln x \Big|_\theta^\infty = \infty$. Now, let $k > 1$ be an integer. We have

$$E[X^k] = \int_\theta^\infty x^k \cdot \frac{\theta}{x^2}\,dx = \theta \int_\theta^\infty x^{k-2}\,dx = \theta \cdot \frac{x^{k-1}}{k-1}\Big|_\theta^\infty = \infty \text{ since } k-1 > 0$$

Since the method of moments aims to set $\overline{X^k} = E[X^k] = \infty$, we see the technique offers no value for this particular distribution.

(b) If we look at the support, we can write it as $0 < \theta \leq x < \infty$. When we discuss the pdf, we focus on x, so we write $f(x) = \dfrac{\theta}{x^2} \cdot I_{[\theta,\infty)}(x)$ where we use $0 < \theta \leq x < \infty$. When we switch to the likelihood, we think of θ as the variable, so we write $L(\theta) = \dfrac{\theta}{x^2} \cdot I_{(0,x]}(\theta)$ where we use $0 < \theta \leq x < \infty$. Next, we use n pieces of data:

$$L(\theta) = \prod_{i=1}^n \frac{\theta}{x_i^2} \cdot I_{(0,x_i]}(\theta) = \frac{\theta^n}{\prod x_i^2} \cdot \prod I_{(0,x_i]}(\theta) = \frac{\theta^n}{\prod x_i^2} \cdot I_{(0,\min x_i]}(\theta)$$

We want to *maximize* this function. Because of the θ^n term, a max occurs when θ is as large as possible. But, the indicator function turns into 0 if $\theta > \min x_i$. So, we need to choose the largest value of θ for which the indicator gives the value 1. This is at $\boxed{\widehat{\theta}_{\text{MLE}} = \min x_i}$.

Commentary: The given pdf is an example of a random variable where all moments are infinite. This can occur for "heavy-tailed" distributions. As this problem teaches us, MME is of no help in such settings, which is one reason why it is good to have multiple approaches to the problem of estimation. Indeed, the fact that MLE is useful while MME fails shows that these two techniques must be fundamentally different in the information they leverage when building estimates. As someone who studies mathematical problems, I am fascinated by problems which can't be tackled through a given strategy and require other ones instead. This is why I selected this problem in the first place.

This interest in solution pathways is mirrored in my love of video games. When you look at the "roguelike" games Balatro or Hades, the fun is in creating different builds (jokers and cards in the former, boons and aspects in the latter). There are many ways to find success, and the game forces you to uncover these via the random choices you are offered each time you begin anew. In the math classroom, I play the role of the programmer, forcing you to uncover problem solving strategies through the problems I create and select.

12. Suppose X is a uniform distribution on the closed interval $[-\theta, \theta]$ and that we wish to estimate $\theta > 0$ using a random sample $X_1, \ldots, X_n$.

    (a) Explain why the MME based on first moments is not useful in this setting. Then, find a method of moments estimator for θ using second moments.

    (b) Find the MLE for θ.

Hint List:

- b. When you begin using indicator functions, figure out why $\theta \geq |x|$.

Solution:

(a) Notice that $E[X] = \dfrac{-\theta + \theta}{2} = 0$, which doesn't involve the parameter θ! So, setting $\overline{X} = E[X]$ is not productive. Instead, observe that

$$E[X^2] = Var[X] + E[X]^2 = \frac{(\theta - (-\theta))^2}{12} + 0^2 = \frac{4\theta^2}{12} = \frac{\theta^2}{3}$$

Setting $\overline{X^2} = E[X^2] = \dfrac{\widehat{\theta}^2}{3}$ gives $\boxed{\widehat{\theta}_{\text{MME}} = \sqrt{3\,\overline{X^2}}}$.

(b) Notice that the support of our pdf is $-\theta \leq x \leq \theta$. Using indicator functions, our pdf is $f(x;\theta) = \dfrac{1}{2\theta} \cdot I_{[-\theta,\theta]}(x)$. As we switch to the likelihood, our attention turns to θ instead of x. The inequality $-\theta \leq x \leq \theta$ says that $-\theta \leq x$ and $x \leq \theta$, or equivalently, that $\theta \geq -x$ and $\theta \geq x$. This can be written most simply as $\theta \geq |x|$. Hence, the likelihood is $L(\theta) = \dfrac{1}{2\theta} \cdot I_{[|x|,\infty)}(\theta)$. Moving to n pieces of data gives:

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{2\theta} \cdot I_{[|x_i|,\infty)}(\theta) = \frac{1}{(2\theta)^n} \cdot I_{[\max|x_i|,\infty)}(\theta)$$

To maximize this expression, the $\dfrac{1}{(2\theta)^n}$ term suggests making θ as small as possible, but the indicator function won't let us drop below $\max|x_i|$. Hence, $\boxed{\widehat{\theta}_{\text{MLE}} = \max|x_i|}$.

Commentary:

This problem reveals some interesting things that can happen with both MME and MLE. In part (a), we discover that it is possible for a moment (here, $E[X]$) to *not* involve the population parameters (here, θ). Hence, the MME approach of setting $\overline{X} = E[X]$ does not help us estimate θ. Luckily, we are saved by second moments, but at the cost of complexity. It certainly isn't intuitive why $\sqrt{3\,\overline{X^2}}$ is a good way of getting at the value for θ! In contrast, the MLE has an intuitive answer: If our support is $[-\theta, \theta]$, we simply select the x value as far from the origin as possible. If that x happens to be negative, it will be close to $-\theta$, and hence, $|x|$ will be close to θ. If, however, that x is positive, then $|x| = x$, so we are again close to θ. Often, it is useful to look back at your answers and see what they can reveal, as I have above. In my experience, the people who go furthest in life spend time reflecting on what they produce. This might be a musician doing a deep analysis of their playing in an audio recording or an athlete critiquing video of a game. You should do the same for your academic work.

13. Economists frequently use the CRV X with pdf

$$f_X(x;\alpha,\beta) = \beta\alpha^{\beta}x^{-\beta-1} \text{ where } x \geq \alpha > 0 \text{ and } \beta > 1$$

Find the MLEs for α and β.

Note: As with all *multivariate* maximization problems in this class, you need NOT show your MLE is maximal via higher derivatives. Also, as on all estimation problems, if no data values are explicitly given, you should begin by naming them for your use: "Let $x_1, \ldots, x_n$ be a random sample of data".

Hint List:

- Notice that the support depends on α and not β. What did we do in class for pdfs whose support depends on the parameter?
- Deal with α first. After that, switch your focus to β.
- Dealing with α will require indicator functions (treat β like a constant when doing so). Dealing with β will follow the usual analysis pipeline ($L \to \ell \to \ell' = 0$).

Solution: In some problems, you'll need to deal with α and β at the same time by maximizing $L(\alpha, \beta)$. This problem is easier because we can first deal with α and then β.

Note that the support of the pdf depends on α ("$x \geq \alpha > 0$") which suggests that indicator functions will be useful. Begin by rewriting the pdf:

$$f(x; \alpha, \beta) = \beta \alpha^{\beta} x^{-\beta-1} \mathrm{I}_{[\alpha,\infty)}(x)$$

Here, we used the support, $0 < \alpha \leq x < \infty$, to figure out the indicator bounds on x. As we move to the likelihood, we turn our focus to α and switch the indicator to be about α. We again use the support, but explore how α is constrained: $0 < \alpha \leq x < \infty$. This gives us

$$L(\alpha) = \beta \alpha^{\beta} x^{-\beta-1} \mathrm{I}_{(0,x]}(\alpha)$$

If we let $x_1, \ldots, x_n$ be a random sample of data, we can write the likelihood of seeing all n pieces of data:

$$L(\alpha) = \prod_{i=1}^{n} \left[\beta \alpha^{\beta} x_i^{-\beta-1} \mathrm{I}_{(0,x_i]}(\alpha) \right] = \beta^n \alpha^{n\beta} \left(\prod x_i \right)^{-\beta-1} \prod \mathrm{I}_{(0,x_i]}(\alpha) = \beta^n \alpha^{n\beta} \left(\prod x_i \right)^{-\beta-1} \mathrm{I}_{(0,\min x_i]}(\alpha)$$

For any β, notice that this expression gets larger as α gets larger because of the $\alpha^{n\beta}$ term. But, the indicator function becomes zero once α exceeds $\min x_i$. Thus, we maximize $L(\alpha)$ by choosing $\boxed{\widehat{\alpha}_{\mathrm{MLE}} = \min X_i}$. Notice that we don't have to worry about β as we are thinking about α because our choice for α works the same regardless of what β is.

Turning to β, we will follow the usual analysis pipeline ($L \to \ell \to \ell' = 0$) because the support does not depend on β. Here,

$$L(\beta) = \prod_{i=1}^{n} \beta \alpha^{\beta} x_i^{-\beta-1} \implies \ell(\beta) = \sum \left[\ln \beta + \beta \ln \alpha - (\beta+1) \ln x_i \right] = n \ln \beta + n\beta \ln \alpha - (\beta+1) \sum \ln x_i$$

Thus, $\ell'(\beta) = \dfrac{n}{\beta} + n \ln \alpha - \sum \ln x_i \implies 0 = \dfrac{n}{\widehat{\beta}} + n \ln \widehat{\alpha} - \sum \ln x_i$.

Solving for $\widehat{\beta}$, we find $\sum (\ln x_i) - n \ln \widehat{\alpha} = \dfrac{n}{\widehat{\beta}} \implies \widehat{\beta}_{\mathrm{MLE}} = \dfrac{n}{\sum (\ln x_i) - n \ln \widehat{\alpha}}$. Now, we can replace $\widehat{\alpha}$ with our result from above and also massage the answer to have more interpretability:

$$\widehat{\beta}_{\mathrm{MLE}} = \frac{1}{\dfrac{\sum (\ln x_i)}{n} - \ln \min x_i} = \boxed{\frac{1}{\overline{\ln X} - \ln \min X_i}}$$

Commentary: This is a serious problem for mature academics. Beyond the algebraic sophistication, there are many details to attend to, as noted in the above solution. In my mind, the most important takeaway is that this problem shows it is possible to have the "usual" MLE technique ($L \to \ell \to \ell' = 0$) and the "unusual" technique (indicator functions) in the *same* problem. This blend of influences makes for a really special experience. If you'd like to explore more on the power and beauty of blended influences, consider listening to:

(a) This episode of the podcast *Hidden Brain*. The speakers discuss how blends of background knowledge are at the core of scientific creativity.

(b) This composition by current Indian-American composer Reena Esmail. Her blend of Indian and Western classical style has entranced me from the moment I first heard it.

I think this problem and the above links have a greater lesson to teach us about the human experience. For me, my greatest insights have been *interdisciplinary*. That is, they were the blend of two or more knowledge bases coming together to make sense of the world. For example, when playing cello, you might mix the artistry of interpretation with knowledge of acoustics (how sound works) with technical training on the instrument itself. In ice hockey, you could combine the athleticism of skating with behavioral economics to choreograph plays that take advantage of how your opponents will react. If your goal on this Earth is to deeply connect with and understand the human experience, then I think this is best accomplished with an interdisciplinary lens. You need to see the world from multiple angles at the same time.

14. Let $X_1, \ldots, X_n$ be a random sample from $X \sim Gamma(r, \lambda)$ with both r and λ unknown. In class, we found MLE estimators for r and λ by solving a system of two equations and two unknowns. A similar setup can arise when using MME.

    (a) Using first and second moments, find the MMEs for r and λ. Use the bar notation to make the answers less nightmarish.

    (b) Here is an example of me using the estimators from part (a) in my everyday life. Recently, I decided to remodel my bathroom with the help of my dad. One key step was buying tile for the shower walls. After picking out tile I loved, the company sent me a cost estimate where I learned my dream tile was \$10.98 per square foot! Since I needed about 160 square feet, this cost was shocking. After calming down, I wanted to figure out where \$10.98 falls on the distribution of tile costs (per sq ft). I went to my favorite AI and asked it to generate a huge list of different tile products and their costs per square foot. Download TileCosts.csv from Canvas/TritonEd, load it into R, and use your estimators from part (a) to find $\widehat{r}$ and $\widehat{\lambda}$. Then, graph our best-fitting Gamma distribution on top of a histogram of the tile costs mined via AI. Finally, using our Gamma distribution, determine what percent of tiles are cheaper than mine.

Hint List:

- a. If you haven't memorized the mean and variance of the Gamma distribution, start by doing so.
- a. Try replacing $\frac{\widehat{r}}{\widehat{\lambda}}$ by $\overline{X}$ in the equation related to second moments.
- b. See the earlier question on bioavailability to inspire the R code.

Solution:

(a) For the Gamma distribution, we know $E[X] = \frac{r}{\lambda}$ and $Var[X] = \frac{r}{\lambda^2}$. Thus,

$$E[X^2] = Var[X] + E[X]^2 = \frac{r}{\lambda^2} + \frac{r^2}{\lambda^2}$$

Equating moments gives (1) $\overline{X} = \frac{\widehat{r}}{\widehat{\lambda}}$ and (2) $\overline{X^2} = \frac{\widehat{r}}{\widehat{\lambda}^2} + \frac{\widehat{r}^2}{\widehat{\lambda}^2}$. If we substitute (1) into the right-hand side of (2), we get:

$$\overline{X^2} = \frac{\overline{X}}{\widehat{\lambda}} + \overline{X}^2 \implies \boxed{\widehat{\lambda} = \frac{\overline{X}}{\overline{X^2} - \overline{X}^2}}$$

Substituting this result into (1) gives $\boxed{\widehat{r} = \frac{\overline{X}^2}{\overline{X^2} - \overline{X}^2}}$.

(b) In R, we do:

```
tileinfo = read.csv("your file location")
costs = tileinfo$Cost

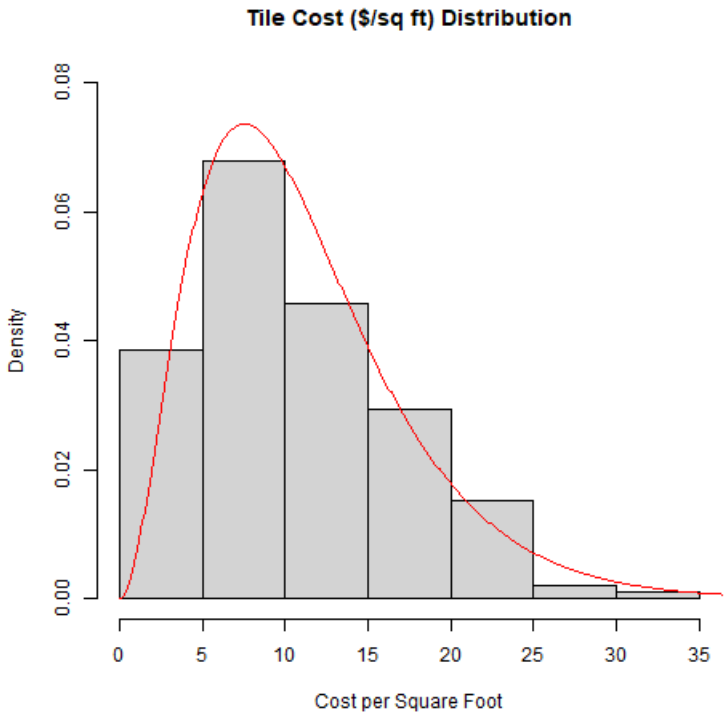
rhat = mean(costs)^2/(mean(costs^2) - mean(costs)^2)
lambdahat = rhat/mean(costs)

xs = seq(0, 40, 0.1)
ys = dgamma(xs, rhat, lambdahat)

hist(costs, freq = F, main="Tile Cost ($/sq ft) Distribution",
     xlab="Cost per Square Foot", ylab="Density", ylim=c(0,0.08))
lines(xs, ys, col="red")

pgamma(10.98, rhat, lambdahat)
```

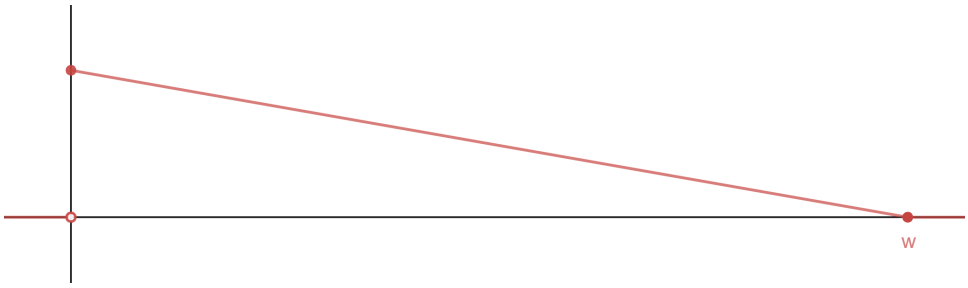
See the image on the next page. The pgamma calculation shows that about $\boxed{56.556\%}$ of tiles are cheaper than mine.



Commentary:

The story described in this problem is 100% true. When I got the answer in part (b), I realized my tile choice wasn't some ultra-expensive indulgence. Instead, it fell near the median of tile prices, and this convinced me to go ahead with the purchase. While tile and remodels may not excite you, hopefully the larger lesson here does: You can make more informed decisions in your life when you know more stuff. Here, having knowledge of random variables and parameter estimation gave me perspective on a new domain (tile costs). One of the most important pieces of advice I give to students is to look at the world through a statistical pair of glasses. This means that whatever you're doing (e.g., running on a treadmill, playing video games, using online dating, buying tile), intentionally think about how to frame that activity using the language of probability and statistics. You'll be surprised by how deeply these subjects underpin the human experience.

15. Suppose Y is a CRV whose pdf is pictured below. Find the MME and MLE for w given the small sample: $y_1 = 1$, $y_2 = 3$. (Technology may be useful on the MLE. Do not try to find a formula for the MLE in the general case of n data; no closed-form solution exists.)



Hint List:

- First, try to write the equation of the pdf.
- To find the y-intercept, answer this: How much area is under *any* pdf?
- For the MLE, try graphing the likelihood function based on the two data points.

Solution: Because the area under any pdf must be 1 and the pdf is shaped like a triangle, we get

$$\frac{1}{2}bh = 1 \implies \frac{1}{2}wh = 1 \implies h = \frac{2}{w}$$

We can now use slope-intercept form to find the pdf:

$$f(y; w) = \frac{-2/w}{w}y + \frac{2}{w} = \frac{-2}{w^2}y + \frac{2}{w}, \text{ where } 0 \le y \le w$$

From this, we find the expectation:

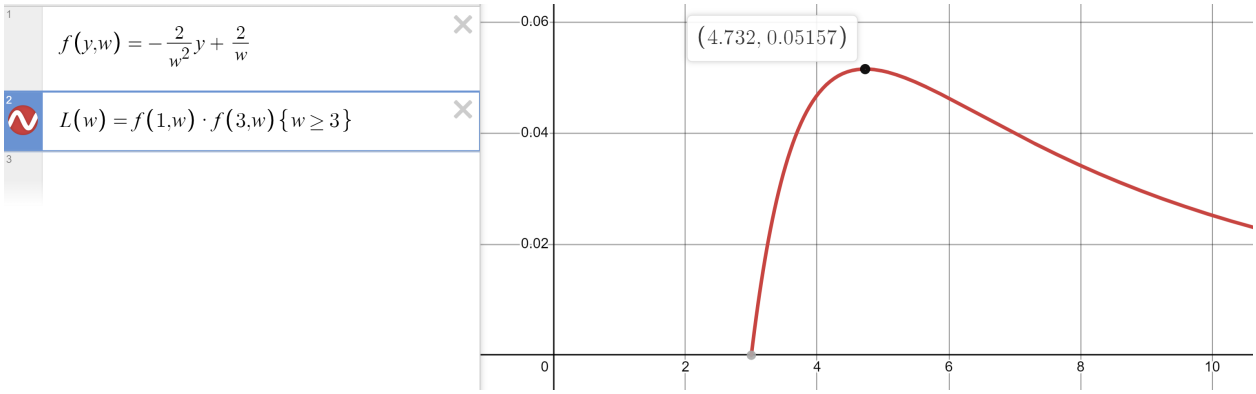
$$E[Y] = \int_0^w y\left(\frac{-2}{w^2}y + \frac{2}{w}\right) dy = \frac{-2y^3}{3w^2} + \frac{y^2}{w}\Big|_0^w = \frac{1}{3}w$$

Equating first moments gives $\overline{Y} = E[Y] \implies \overline{Y} = \dfrac{\widehat{w}}{3} \implies \widehat{w}_{\text{MME}} = 3\overline{Y}$. Our particular data give $\overline{Y} = 2$, so $\boxed{\widehat{w}_{\text{MME}} = 6}$.

For the MLE, we are told not to find the MLE in general, but rather, with our particular data: $y_1 = 1, y_2 = 3$. Here, the support depends on the parameter, so indicators might be helpful. Our support is $0 \le y \le w < \infty$, so focusing on the red helps us write the likelihood:

$$\begin{aligned} L(w) &= \left[\frac{-2}{w^2}\cdot 1 + \frac{2}{w}\right] I_{[1,\infty)}(w) \cdot \left[\frac{-2}{w^2}\cdot 3 + \frac{2}{w}\right] I_{[3,\infty)}(w) \\ &= \left[\frac{-2}{w^2}\cdot 1 + \frac{2}{w}\right]\left[\frac{-2}{w^2}\cdot 3 + \frac{2}{w}\right] I_{[3,\infty)}(w) \quad \text{(combine indicators)} \end{aligned}$$

Often with MLE and indicators, we can look at the formula and see that it is always increasing or decreasing, and so an endpoint of the indicator interval becomes the answer. Here, it isn't clear what the graph of the likelihood looks like, so we might use technology to help:



The image gives us $\boxed{\widehat{w}_{\text{MLE}} \approx 4.732}$.

Commentary: The first reason I assign this problem is that you were required to find the formula for the pdf before you could begin the MME and MLE steps. This can happen if you have a visual idea of what your distribution looks like (and not a famous family it belongs to). Second, we get different answers for the MME and MLE, again showing that these need not be the same. Finally, notice that the answer for the MLE is **not** one of the ends of the interval seen in the indicator function. So far, your experience with indicators has always led to an answer like $\min y_i$ or $\max y_i$. It is tempting after a few consistent examples to believe that a trend must always occur ("If the problem requires an indicator, I will get a max/min answer"), but this isn't the case. In everyday life, we are warned not to generalize the patterns we see in small numbers of people to whole groups of people. Indeed, we have language for this: stereotyping.

The same caution could be offered in mathematics: Do not assume the behavior you observe in a small collection of examples carries over to all examples. As an example, if you ask high school students to list functions they know, they usually come up with continuous functions. Based on that small sample, it might be tempting to claim that all functions are continuous. Alas, they aren't. Examples like $y = \dfrac{1}{x}$ and $y = \tan x$ show this isn't the case. As another example, if you ask Calculus students to think about the differentiability of their favorite *continuous* functions, they'll likely come back and say that their continuous functions are differentiable everywhere (like $\sin x$) or, in rare cases, have a spot where you can't take the derivative (like at the sharp corner of $y = |x|$). From this small set of examples, it is tempting to believe that continuous functions should be differentiable everywhere, or basically everywhere. It turns out this assessment is wrong! In 1872, Karl Weierstrass dreamt up a continuous function that is nowhere differentiable (imagine a continuous function with sharp corners at all places). Later, it was shown that this example is not an obscurity; rather, the "typical" continuous function is nowhere differentiable! This result is so counterintuitive that mathematicians had been assuming the opposite for decades.

The takeaway here is that whether you are generalizing from a small collection of people or a small collection of mathematical examples, danger awaits. This process is known as inductive reasoning. While it does have evolutionary value ("My last three friends who played with lions got eaten, so maybe I should avoid all lions"), it also has problems (see above). This is one reason mathematicians employ deductive reasoning (the act of using logic to prove/derive generalities). In truth, mathematics is often the act of using a small collection of examples to guess at larger truths (inductive reasoning), and then trying to prove these claims (deductive reasoning). And just in case you'd like to deify deductive reasoning as the be-all-end-all type of reasoning, the Incompleteness Theorems show that this approach to thinking has problems as well! Said vaguely, it is possible to make claims whose truth cannot be determined from your axioms no matter how hard you try!

16. Suppose we have a random sample $Y_1, \ldots, Y_n$ from a CRV with density

$$f_Y(y;\theta) = \frac{\theta}{(y+1)^{\theta+1}} \text{ where } y > 0, \theta > 1$$

Find the MME and MLE for θ.

Hint List:

- When finding the expectation, try the substitution $u = y + 1$.

Solution: We first find $E[Y]$. Notice that the integral is improper because of the ∞ bound, not the 0 bound. We use a limit to handle this bound:

$$E[Y] = \int_0^\infty y \cdot \frac{\theta}{(y+1)^{\theta+1}}\, dy = \lim_{t\to\infty} \int_0^t \frac{y\theta}{(y+1)^{\theta+1}}\, dy$$

Setting $u = y + 1$, we find $y = u - 1$, so $dy = du$. Our new bounds are $u = t + 1$ (upper) and $u = 0 + 1 = 1$ (lower). Thus, we get

$$\lim_{t\to\infty} \theta \int_1^{t+1} \frac{u-1}{u^{\theta+1}}\, du = \theta \lim_{t\to\infty} \int_1^{t+1} u^{-\theta} - u^{-\theta-1}\, du = \theta \lim_{t\to\infty} \left[\frac{u^{-\theta+1}}{-\theta+1} - \frac{u^{-\theta}}{-\theta}\right]_1^{t+1}$$

Since $\theta > 1$, both of the exponents on u are negative. Once we substitute $t + 1$ and take the limit, the negative exponents create the value 0 giving us

$$\theta\left[0 - 0 - \frac{1}{-\theta+1} + \frac{1}{-\theta}\right] = \frac{\theta}{\theta-1} - 1 = \frac{\theta-(\theta-1)}{\theta-1} = \frac{1}{\theta-1}$$

Equating moments gives $\overline{Y} = \dfrac{1}{\widehat{\theta}-1}$, or $\boxed{\widehat{\theta}_{\text{MME}} = 1 + \dfrac{1}{\overline{Y}}}$.

For the MLE, note that $L(\theta) = \displaystyle\prod_{i=1}^{n} \frac{\theta}{(y_i+1)^{\theta+1}}$.

Thus, $\ell(\theta) = \sum \left[\ln\theta - (\theta+1)\ln(y_i+1)\right] = n\ln\theta - (\theta+1)\sum \ln(y_i+1)$

$$\ell'(\theta) = \frac{n}{\theta} - \sum \ln(y_i+1)$$

$$\sum \ln(y_i+1) = \frac{n}{\widehat{\theta}}$$

$$\widehat{\theta} = \frac{n}{\sum \ln(y_i+1)} = \frac{1}{\dfrac{\sum \ln(y_i+1)}{n}}$$

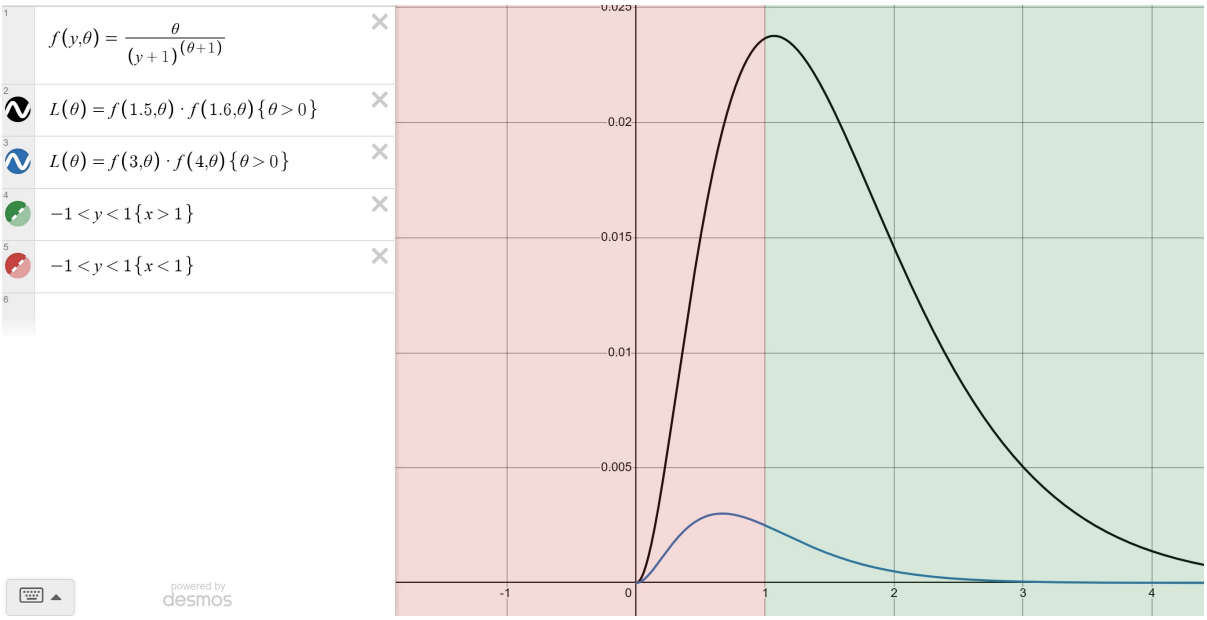
$$\ell''(\theta) = -\frac{n}{\theta^2}$$

Note that $\ell'' < 0$ for all θ

So, we find $\widehat{\theta}_{\text{MLE}} = \dfrac{1}{\overline{\ln(Y+1)}}$ which, again, is not equal to $\widehat{\theta}_{\text{MME}}$.

But is this really the answer? Algebraically, it would appear so. Yet, it always helps to ask yourself: Will this formula give estimates that satisfy the demands of the problem? In this problem, we know $\theta > 1$. As you look at the MME above, you see that the formula $1 + \dfrac{1}{\overline{Y}}$ is indeed greater than 1 (since the support is $y > 0$). But, will the MLE formula give estimates for $\theta > 1$? For example, if each $y_i > e - 1$, then $\ln(y_i + 1) > \ln(e - 1 + 1) = 1$. This forces $\overline{\ln(Y+1)} > 1$, and hence, $\widehat{\theta}_{\text{MLE}} < 1$. Since θ is supposed to be greater than 1, this is a problem. Here, the issue arises because the location we claim the max happens at, $\widehat{\theta}_{\text{MLE}}$, can be *outside* the zone we are maximizing over, $(1, \infty)$.

To illustrate this, consider the below picture.



Here, the black graph is the likelihood for the data set $y_1 = 1.5, y_2 = 1.6$. Notice that the max is attained in the green zone (i.e., allowable values for θ) to the right of $\theta = 1$, so the formula $\widehat{\theta}_{\text{MLE}} = \frac{1}{\ln(Y+1)}$ gives us something reasonable. In contrast, the blue graph is the likelihood for the data set $y_1 = 3, y_2 = 4$. Here, the algebra is pointing us to a location outside of the zone we are maximizing over $(\theta > 1)$, so we are being misled. In this case, simply looking at the graph shows the max is attained (in the green zone) at $\theta = 1$. If the problem statement allowed $\theta = 1$, we could take this as the MLE in such cases. But, the problem says $\theta > 1$, not $\theta \geq 1$, so there actually is no maximum on the open interval $\theta \in (1, \infty)$. So, the correct answer is actually:

$$\boxed{\widehat{\theta}_{\text{MLE}} = \begin{cases} \dfrac{1}{\ln(Y+1)}, & \text{if } \dfrac{1}{\ln(Y+1)} > 1 \\ \text{DNE}, & \text{else} \end{cases}}$$

Commentary: I assign this problem because it is a nightmare of bookkeeping and minutiae. Being good at mathematics often requires that you correctly carry out many steps in sequence and pay attention to exhausting minor things. The higher you go in the subject, the more steps you have to get right, and the more details you must keep track of. One way to outlast these long sequences of steps is to get more accurate at each individual step. By now, you've reached 99.99% consistency on steps like $3+4$ and $2 \cdot 5$. String together a bunch of steps with 99.99% accuracy and you have a high probability of getting to the correct answer. Another approach to enduring long stretches of mathematics is to upgrade to better mathematical machinery that shortens the experience. Metaphorically, this is like taking a plane instead of driving. Earlier in this class, we saw the power of the Kernel Technique to skip over long integration problems.

This problem offers a third approach to dealing with long stretches of algebra. As you read over the MME solution, you might feel that it gets mired in details related to the improper integral. One option, which mathematicians frequently use, is to be a little sloppy with the bounds and carry through

∞ as a bound instead introducing limits. Indeed, I've taken this approach in the solutions to earlier problems. If we do so here, our solution to the MME begins

$$E[Y] = \int_0^\infty \frac{y\theta}{(y+1)^{\theta+1}}\,dy \overset{u=y+1}{=} \theta \int_1^\infty \frac{u-1}{u^{\theta+1}}\,du = \theta\left[\frac{u^{-\theta+1}}{-\theta+1} - \frac{u^{-\theta}}{-\theta}\right]_1^\infty = \theta\left[0 - 0 - \frac{1}{-\theta+1} + \frac{1}{-\theta}\right]$$

When you teach integral calculus, you frequently see students carry infinity through the entire problem as above. Usually they do this because (1) they don't understand they are supposed to convert to limits and/or (2) they are being lazy and don't want to write and think about limits. What I'm suggesting here is that, in time, you consider moving from the longer, official solution to this more compact approach (the lazy student!). The pro of the shorter version is that it has fewer steps, and hence, less chance for error. The con is that it hides the fact that a limit secretly defines every improper integral.

17. Let's start using R to *see* estimators in action. While an estimator looks like a formula, it is actually a random variable because as different random samples come out of a distribution, they combine (via the formula) to make a random value. Different data give rise to different values of the estimator. Let's consider estimating the parameters from $N(\mu, \sigma^2)$. In the Lecture 4 "Additional Practice" section, we see the MLEs based on the data $x_1, \ldots, x_n$ are:

$$\widehat{\mu} = \overline{x} \text{ and } \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

(a) Imagine we are collecting data on IQ scores at UCSD, and suppose these are $N(\mu = 106, \sigma^2 = 14^2)$. I have to give you values for μ and σ^2 so we can run a simulation, but pretend we don't know them! Using the function rnorm in R, generate a random sample of 23 IQ scores from this distribution, and then write code that finds $\widehat{\mu}$ and $\widehat{\sigma^2}$. Include your code and results. When calculating $\widehat{\sigma^2}$, do not use the built-in function var, as this does a slightly different calculation than our formula above and because I want you to see how straight-forward it is to calculate the formula for $\widehat{\sigma^2}$ in a vectorized language like R.

(b) Now, let's imagine that instead of one sample of size 23, we collected 1000 samples, each of size 23. Each sample gives a value for $\widehat{\mu}$, so we have 1000 different values for $\widehat{\mu}$. Using the replicate function in R, create these 1000 values for $\widehat{\mu}$, and then use the hist function to make a histogram. Include your code and a rough sketch (or screenshot) of the histogram. This picture allows you to *see* $\widehat{\mu}$ as a random variable. In R, you can type ?replicate to read the documentation for the replicate function.

Hint List:

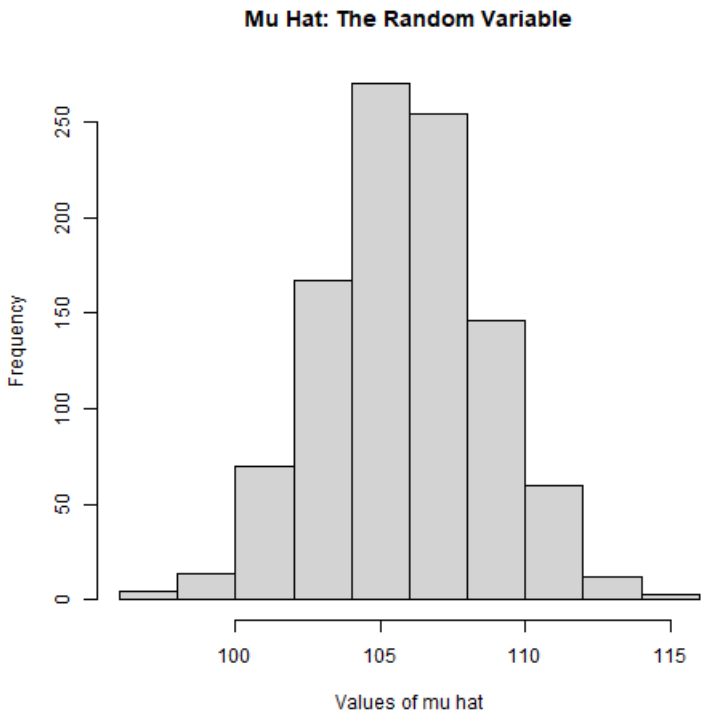
- Feel free to use an internet search or AI to help understand the syntax of specific commands in R. Avoid having tech do the problem for you.

Solution:

(a)
```
IQs = rnorm(23, mean=106, sd=14)

muhat = mean(IQs)
sigma2hat = mean((IQs - muhat)^2)
```

(b)
```
muhats = replicate(1000, mean(rnorm(23, 106, 14)))
hist(muhats, xlab="Values of mu hat", main="Mu Hat: The Random Variable")
```



Commentary: For those with less programming experience, this problem can feel pretty overwhelming. While the prompt points you to the correct functions (e.g., rnorm, replicate, hist), it doesn't hold your hand on every aspect of the R language. As such, you have to look up a lot of stuff we didn't explicitly discuss in order to do this problem. I did this *intentionally*. It is very common in the real world to work with new programming languages/tools. In my career, I've been asked to do things in over 15 different programming languages; it simply isn't possible to be knowledgeable in all these. In addition, tools are constantly changing, going in and out of favor. And, different spheres simply use different tools due to history or preference. For example, it is far more common in academia to use R, while the business world likes Python.

If you currently hold the belief that college will prepare you with all the knowledge you need in your first job, you should stop deluding yourself. Certainly, your time as an undergraduate will lay some of the foundation, but given the diversity of career pathways, it is impossible to prepare 100% of students for 100% of the tasks they'll be asked to do. This reality is especially apparent when it comes to programming languages, for often the language you learn in school is already outdated by the time you graduate. So, it's time you adopted a more realistic view of things. If you plan to work extensively in a language, you'll want to completely memorize its syntax and pros/cons. If you're planning on using a language on a single project or in a single course, it's more common to learn a few basics (e.g., how to make a comment, how to declare variables) and then look up specific syntax (via web searches/AI) along the way. Notice that most people take this approach to foreign languages as well: If you're going to live in Japan for 5 years, you completely learn the language. If you're visiting for 5 days, you learn some key phrases and vocabulary on the flight over.

The key takeaway here is that you need to match how deeply you learn things with how extensively you'll need them moving forward. This idea is especially important in the real world because the demands on your time will be greater than what can be achieved if you do everything as fully as possible. As an example, when I needed to create a website as a UCSD professor, there wasn't time to learn all of web design; instead, I had a half day to pull it together given everything else on my list. Be careful, however, not to overuse this advice. It is tempting to recast laziness as "I won't need this in the future, so I'm not going to learn it thoroughly". Sometimes it can be very difficult to anticipate what will be important moving forward. So, in school, it is better to err on the side of learning most things deeply. Indeed, when schools are choosing what classes will be part of their curricula, they are (usually) trying to choose those topics that will benefit most people in most jobs.

18. Suppose that the time it takes your computer to load R Studio on a random day is normally distributed with unknown mean, μ, and variance 1.2 seconds^2. You'd like to build a 95% confidence interval for μ, so you time your load speeds on 6 random days: 2.1, 1.7, 3.3, 2, 2.1, 1.9 (seconds).

    (a) Find your 95% CI for μ. You should always include units when listing a CI.

    (b) Suppose you had instead created an interval using an 80% confidence procedure. What CI do you get now?

    (c) Suppose you're unhappy with the width of the interval in part (a) and want it to be one-third its current size. Assuming you use the 95% confidence procedure, how many total data would you need to collect to achieve this?

Hint List:

- a. You'll need a Z-table or R to do a calculation related to "95% confidence".
- b. Changing the confidence level alters only one part of the CI formula. Which part?
- c. Here it helps to express the width of the CI in un-simplified form: $2 \cdot 1.96\frac{\sqrt{1.2}}{\sqrt{6}}$. How should you change the number 6 to get the width to be one third of the current length?

Solution:

(a) For 95% confidence, $\alpha = 0.05$ since we know $(1-\alpha)\cdot 100\% =$ confidence level. So, we must find $z_{\alpha/2} = z_{0.05/2} = z_{0.025} \approx 1.96$ using either tables or the R command

```
qnorm(0.025, lower=F)
```

95% CIs are so common that you should memorize the number 1.96 because you will use it frequently. Now, our CI is

$$\overline{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} = \frac{131}{60} \pm 1.96\frac{\sqrt{1.2}}{\sqrt{6}} \approx \boxed{[1.307 \text{ sec}, 3.060 \text{ sec}]}$$

(b) For 80% confidence, $\alpha = 0.20$. So, we must find $z_{\alpha/2} = z_{0.10} \approx 1.281552$ using tables or the R command

```
qnorm(0.10, lower=F)
```

Now, our CI is

$$\overline{x} \pm 1.281552 \cdot \frac{\sigma}{\sqrt{n}} = \frac{131}{60} \pm 1.281552\frac{\sqrt{1.2}}{\sqrt{6}} \approx \boxed{[1.610 \text{ sec}, 2.756 \text{ sec}]}$$

Notice how much is similar from the steps in part (a).

(c) The width in part (a) is $2 \cdot 1.96\frac{\sqrt{1.2}}{\sqrt{6}}$. If we collect more data, only the number 6 will change. Notice that if we collect 9 times as much data, our new width is $2 \cdot 1.96\frac{\sqrt{1.2}}{\sqrt{9\cdot 6}} = \frac{1}{3} \cdot \left[2 \cdot 1.96\frac{\sqrt{1.2}}{\sqrt{6}}\right] = \frac{1}{3} \cdot [\text{old width}]$. So, we need $9 \cdot 6 = \boxed{54}$ pieces of data.

Commentary: There are three main takeaways from this problem. First, parts (a) and (b) reveal that if you want to be less confident in your CIs, they will be narrower. Said differently, if you want to be more confident in your CIs, they will be wider. This aligns with some real-world metaphors: If you want to be more sure you'll catch a fish, increase the size of your net. If you want to be more sure you'll catch a criminal, widen your search radius. If you want to be more sure you'll meet someone you like on a dating app, swipe through more profiles. Second, part (c) reveals there is a *quadratic tradeoff* between sample size (n) and CI width. If want to cut the size by a factor of 3, we need $3^2 = 9$ times as much data. To cut the size by a factor of 10 requires 10^2 times as much data. When people don't understand the nature of tradeoffs, they often make unreasonable requests. For example, administrators in a school district might demand that a school cut its student failure rates in half. In a

quadratic tradeoff, this would require 4 times the resources, staffing, etc. Sadly, people tend to assume linear tradeoff rates, which is often incorrect.

Finally, I wanted to point out that this problem was constructed to be a phony "real-world problem". While it does have a setting (load times for R Studio), that setting is not important for doing the problem, nor are the questions I ask something a person might reasonably ask. Have you ever really cared about the time it takes a common application to load on your computer? Probably not, since most load times are, on average, within the zone of what is reasonable for computer programs. There are plenty of settings where we might actually care about the average of what we are studying; this is not one of them. In addition, this setting is phony because you were given the variance of load times at the population level (σ^2). Rarely do we know this since it requires performing a variance calculation on all data values in the population (here, all possible load times ever for R Studio). If we somehow had access to the entire population, then we could do a simple average calculation to find μ, and there would be no need to create CIs. There are some settings where it is reasonable to assume a value for the population variance while being genuinely interested in the population mean. For example, you might study an indigenous tribe in the Amazon rain forest. If you're curious about how they will do on average (μ) on a standard IQ test, it is reasonable to assume their standard deviation (σ) will match the world-wide standard deviation of 15 IQ points since this number tends to be somewhat consistent in different subpopulations. An authentic real-world problem would use this setup instead. Be on the look-out for fake real-world problems in your schooling.

19. Suppose $X_1, \ldots, X_n$ are iid from $X \sim N(\mu, \sigma^2)$, where σ^2 is known. Consider the confidence procedure that generates intervals of the form $\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}}, \overline{X} + c\frac{\sigma}{\sqrt{n}}\right)$, where c is a constant.

(a) What value must c be for this to be a 70% confidence procedure? Note that this is not our usual formula for making CIs, and our interval may not be symmetric about $\overline{X}$. Also, see Problem 20.

(b) Your friend collects some data in the above setting, builds a 70% CI, and writes in a journal article: "Given our data, we find there is a 70% chance that the unknown μ is in our interval". Critique this statement and offer an improved statement.

Hint List:

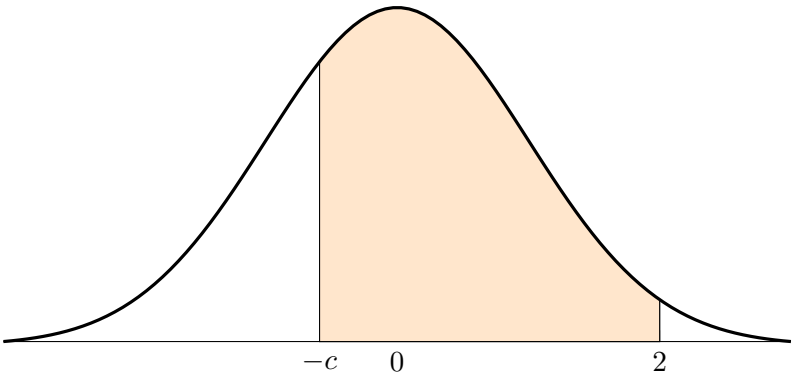
- a. Recall that the confidence level of any confidence procedure is $P(\text{parameter} \in \text{random CI})$. Write what this means for the given setup. Then, start rearranging inequalities.
- b. If I asked "What is the probability you (the person reading these words) attends UCSD?", something should feel weird about that question. Can you figure out what the weirdness is?

Solution:

(a) We have $0.7 = P(\mu \in \text{random CI}) = P\left(\overline{X} - 2 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$.

The first inequality rearranges to $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 2$, while the second inequality is $-c < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$.

Because $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim Z$, we can join our rearrangements together and write $0.7 = P(-c < Z < 2)$. At this point, we can draw a picture of the standard normal and shade the desired region:



We know the shaded region has area 0.7. Since this is a standard normal, we can find the area under the curve to the right of 2 using tables or R:

```
pnorm(2, lower=F)  # we get about 0.0227
```

So, we find the area to the right of $-c$ is $0.7 + 0.0227 = 0.7227$. The qnorm command in R can find a quantile if you give it an area:

```
qnorm(0.7227, lower=F)  # we get about -0.591
```

So, $\boxed{c \approx 0.591}$.

(b) Notice that our friend's claim is about the chance a specific thing ("the unknown μ") is in a specific interval ("our interval"). Once everything in life is specific (rather than random), it doesn't make sense to speak about probabilities anymore: Either μ is in our CI or it isn't. Here's a simpler example: *Before* you flip a coin, it makes sense to say "The probability my flip is heads is 50%". Notice that there is still something unspecific in the process (here, the result of the flip). *After* you flip the coin and are staring at the result, it doesn't make sense to say "The probability my flip is heads is 50%"; you can look at the result and see. It's either heads or tails. CIs work the same way: Before the interval is made, it is correct to say it has a certain probability of capturing the parameter of interest. After it's made, you can't use probability language to discuss things.

So, our friend has two choices to improve the wording:

i. Our CI is from a confidence procedure that has a fairly good confidence level; 70% of the random intervals it makes will capture μ.

ii. **If** μ is actually in our interval, then it is between the lower and upper bounds we calculated.

Both of these are correct but unsatisfying. The first pulls us away from the particular interval we found, which is what we naturally want to talk about. The second feels circular; you're basically saying "If it's in the interval, it's between the bounds". The second also requires using the word "**If**", a frustrating reminder that the CI may not contain the parameter of interest. People hate admitting that their CIs could be fake news.

Commentary: Confidence intervals are radically misunderstood by students, faculty, and working professionals. If we're honest about the statistical training that people are receiving, it's quite limited. Most knowledge workers will take only a single intro-level stats class in their formal education. In that ten- to sixteen-week course, they'll usually spend only a single week on CIs. During that, they'll only have time to learn a few formulas $\left(\text{e.g., } \overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$. The theory behind CIs is rarely developed. The true meaning of CIs is rarely explained. The common misconceptions around CIs are often not addressed.

While one goal of the above problem is to begin reversing those trends, I also want to point out that the educational apparatus itself is pretty broken. Said simply, most people aren't getting enough training in statistical methods. In my math education PhD program, I was only required to take two courses (20 weeks) of statistics coursework! This is on the *higher end* of requirements in PhD programs. As students in Mathematical Statistics, you have moved beyond the intro-stats level, and so you've already had more statistics training than 99.9% of people on Earth. With this expertise comes responsibility. *You* will be the person at your company that will need to step forward and correct the misstatements of colleagues. *You* will be the journal article reviewer that must criticize a paper for misinterpretations of CIs. Part of being a good scholar/citizen is working hard to educate the populace in areas where you have expertise, and reciprocally, actively listening in those areas where you have less training.

20. Let's check your answer to 19a. Below is an outline of some R code that does this. Fill in the missing parts, and then type the code into R and run it to see if your answer from 19a was correct. Our setup will assume the $n = 25$ data come from the distribution $N(\mu = 7, \sigma^2 = 16)$ (we must set a value for μ to run the simulation!). We make 50000 intervals and then see which capture μ and which don't. Finally, we calculate the confidence level.

```
# Code to check

n =                    # sample size
c =                    # value to check
mu =                   # mean value

trials =               # number of intervals to make
capture =              # make a vector to hold trues and falses

for (i in 1:trials) {
  data = rnorm(                          )
  xbar = mean(data)
  LB =                 # the lower bound of the current interval
  UB =                 # the upper bound of the current interval
  capture[i] = (              &&              ) # did we get mu?
}

print(mean(                       )) # calculate and display capture percentage
```

Hint List:

- Remember that you can do an internet search if you don't know the syntax for a given action. Learning what to type into a search engine is an important skill.

Solution:

```
# Code to check

n = 25                    # sample size
c = 0.591                 # value to check
mu = 7                    # mean value

trials = 50000              # number of intervals to make
capture = logical(trials)   # make a vector to hold trues and falses

for (i in 1:trials) {
  data = rnorm(n, mu, 4)
  xbar = mean(data)
  LB = xbar-2*4/sqrt(n)          # the lower bound of the current interval
  UB = xbar+c*4/sqrt(n)          # the upper bound of the current interval
  capture[i] = (LB < mu && mu < UB) # did we get mu?
}

print(mean(capture)) # calculate and display capture percentage
```

Commentary: Because students come to Math 181A with vastly different backgrounds in programming, I provide a rough outline on this problem. While more advanced programmers won't need it, the outline is useful for bringing all students to a similar place by the end of the problem. That is, of course, if students take it upon themselves to understand all the components of the outline. As a thought experiment, contrast these approaches to doing this problem:

(a) The first student feeds the template into an AI program, gives it the prompt, and asks it to fill in the missing parts.

(b) The second student does searches on how to fill in every line and copy-pastes things from the internet. When they don't know what to do (e.g., on the last line), the just try typing in every variable from the problem until one works.

(c) The third student seeks out a short R tutorial for every concept they don't understand. For example, when they need to create a vector of trues/falses, they explore how vectors are created in R and learn about the different data types in the language. When they are asked to do a for-loop, they read about and internalize the syntax/uses of loops in R.

These three archetypes lie along the spectrum of what students do when you ask them to partially build their own understanding. The unethical student (a) learns nothing (that's the con), but finishes with minimal effort/time (pro). On the other end, student (c) devotes time and mentally energy (con) to developing a deep, lasting understanding (pro). Where did you fall on this spectrum when working this problem? As an educator, I want students emulating the final option. I say this for several reasons. First, the real world values actual understanding. While it is sometimes possible to scam the enterprise

of education, this isn't as easy in the outside world. If you don't understand things, there are actual consequences (e.g., deals fall apart, space shuttles explode, bridges collapse). Second, knowledge often builds upon itself. So, if you don't actually develop a strong foundation, you can't build on top of it. This limits how far you can rise in life. Notice how *you* are the one placing a ceiling on your growth. This isn't society, laws, institutions, etc. holding you back. We spend a lot of time in the modern era discussing how *external* forces impede progress, but rarely are we honest about the *internal* forces. Finally, and perhaps most important to me: I believe in honoring the act of learning. Student (a) is clearly disrespecting learning because none is happening while giving the false sense that some has. Student (b) is somewhat better, but the actions taken are really designed to get the problem completed, rather than internalize the core concepts. If you associate "learning" with the "educational industrial complex", it is easy to let your anger/frustration/distaste for the latter corrupt your relationship with the former. Please don't allow this to happen. See learning as a sacred act that is to be respected. If you find yourself making compromises where learning is minimized in favor of efficiency, effort, free time, partying, etc., then *you* (not external forces) are limiting your greatness in life.

21. Suppose $X \sim Unif(0, \theta^2)$ where $\theta > 0$ is unknown. In this problem, we find point and interval estimators for θ.

(a) Find an MME for θ using first moments for a random sample $X_1, \ldots, X_n$.

(b) Suppose your answer from part (a) is called $\widehat{\theta}_{\text{MME}}$. You design an interval estimator to try to capture θ: $\left(\widehat{\theta}_{\text{MME}}, 2 \cdot \widehat{\theta}_{\text{MME}}\right)$. Find the confidence level for this confidence procedure assuming a sample size of $n = 1$. (Also, see Problem 22.)

Hint List:

- b. How does $\overline{X}$ simplify when $n = 1$?
- b. Recall that CL = P(parameter $\in$ random CI). Write out what this means in our setup.

Solution:

(a) Equating first moments gives $\overline{X} = E[X] = \dfrac{0 + \widehat{\theta}^2}{2}$. Solving gives $\boxed{\widehat{\theta}_{\text{MME}} = \sqrt{2\overline{X}}}$.

(b) $\text{CL} = P(\theta \in \text{random CI}) = P(\widehat{\theta}_{\text{MME}} < \theta < 2\widehat{\theta}_{\text{MME}}) = P(\sqrt{2\overline{X}} < \theta < 2\sqrt{2\overline{X}})$

Remember that probability statements are designed to work with random variables in the middle of the inequality chain, not fixed parameters (like θ). Rearranging the first half of the inequality shows that $\overline{X} < \dfrac{\theta^2}{2}$. Rearranging the second half gives $\dfrac{\theta^2}{8} < \overline{X}$. So,

$$\text{CL} = P\left(\frac{\theta^2}{8} < \overline{X} < \frac{\theta^2}{2}\right) = P\left(\frac{\theta^2}{8} < X < \frac{\theta^2}{2}\right) \text{ since } \overline{X} = X \text{ when } n = 1$$

Recalling $P(a < Y < b) = \displaystyle\int_a^b f_Y(y)\,dy$ and the pdf of the uniform distribution, we find that

$$\text{CL} = \int_{\theta^2/8}^{\theta^2/2} \frac{1}{\theta^2}\,dx = \frac{1}{\theta^2} \cdot x\Big|_{\theta^2/8}^{\theta^2/2} = \frac{1}{\theta^2}\left[\frac{\theta^2}{2} - \frac{\theta^2}{8}\right] = \frac{1}{2} - \frac{1}{8} = \boxed{\frac{3}{8} = 0.375 = 37.5\%}$$

Commentary: When you ask students "What do you think math is?", you get different answers, but one common theme is that math is a bunch of formulas you memorize and apply. If this is the view you currently have, I don't blame you; this is what education has largely become. The reason for this is that training students to be formula machines is one of the easiest approaches to teaching. Students and parents have little room for complaining because you can just point to a list of formulas and say "we discussed these, we practiced these, we tested these". Sadly, this view of mathematics is very impoverished. Math is so much more. It's about discovery. It's about personality: mathematics can be surprising, unrelenting, and even funny. It helps you be detail-oriented, logical, creative, and curious. To see how limited your view of math is, just compare how you look at mathematics versus other fields (like poetry or music or athletics).

One reason I assign this problem is that it helps students expand their view of mathematics. Here, we don't have some standard setup with loads of memorized formulas. We have a new setting with a given strange CI. Is this CI any good? That's the point of part (b). Also, there are awkward things you haven't seen before: We don't define the upper range of the uniform support as θ, but rather as θ^2. This might make you curious: Why was this done? Also, we don't end up estimating θ^2, but rather, θ. Why was this done? After you find your answer to part (a), you might notice something: Earlier in the course, when using MME to estimate the upper bound (ub) of a uniform support, we found $\widehat{\text{ub}} = 2\overline{X}$. Here, the upper bound is the square of the parameter we are estimating, and the estimator for θ used the square root function (the inverse function of the square). Is this true in general? How could I use notation to write out what is being said right now? Also, at the end of part (b), you might have noticed that the answer (37.5%) did not involve θ. We've seen the same thing when building a CI for μ in $N(\mu, \sigma^2)$. Does this always/usually/rarely happen? Curiosity beckons.

22. Using the programming skills you gained in Problem 20, write code to check your answer for Problem 21b. Include your code in your answer.

Hint List:

- Revisit the code outline from the previous CL check to get a rough sense of how to proceed. Note that this time we aren't using a normal distribution, so you'll need to change all those references.
- Since the answer to the Problem 21b does not involve θ, you can choose any $\theta > 0$ when running the simulation.

Solution:

```
n = 1          # sample size
theta = 7      # choices will vary

trials = 50000              # number of intervals, choices will vary
capture = logical(trials)   # make a vector to hold trues and falses

for (i in 1:trials) {
  data = runif(n, 0, theta^2)
  MME = sqrt(2 * mean(data))
  LB = MME            # the lower bound of the current interval
  UB = 2 * MME        # the upper bound of the current interval
  capture[i] = (LB < theta && theta < UB) # did we get theta?
}

print(mean(capture)) # calculate and display capture percentage
```

Commentary: I put this problem soon after the previous coding problem so you could get an honest assessment of how deeply you internalized the syntax and ideas from that problem. If your approach was to copy and paste that answer and make small alterations, you missed the chance to see if you could code things from scratch. One of the cons of modern technology is that copy+paste+alter is now possible. While this technique provides greater efficiency, you lose the chance to see if you can stand on your own and build things from the ground up. While I'm not against the use of technology, I am against the unexamined, uncritical use of technology. The critical user of technology looks at both the pros and cons of what a given tool provides, and they don't ignore those cons because they are addicted to the pros. In some cases, it takes time for the cons to be unearthed. When social media first came out, the world could only see its benefit to society. Now, we have numerous studies showing its deleterious effects on young people. The same was true of calculators in math classrooms. What initially seemed like a great tool resulted in a generation of students who couldn't do basic arithmetic without assistance, and as such, were unable to move to higher levels where such calculations must be internalized and automatic. You've had almost two decades of experience with your approach to learning. So, now is the time to look at yourself and ask if you are engaging in any behaviors that help in the short term (pro), but lead to long-term problems (con). Copy+paste+alter might be one of these.

23. Suppose you are drawing a random sample of size $n > 0$ from $N(\mu, \sigma^2)$ where $\sigma > 0$ is known. Decide if the following statements are true or false and **explain your reasoning**. Assume our 95% confidence procedure is $\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$.

(a) If $(3.2, 5.1)$ is the 95% CI based on a particular random sample, then there is a 95% chance that μ is in this interval.

(b) If $(3.2, 5.1)$ is the 95% CI based on a particular random sample, then there is a 95% chance that the mean from our next random sample will be in this interval.

(c) A 95% CI will contain 95% of the possible values from the population distribution we are studying.

(d) If we generate 400 random CIs using our 95% confidence procedure, we expect about 20 intervals to not contain μ.

Hint List:

- Whenever you refer to a probability or percentage, you should be able to fill in a sentence like

  $p\%$ of _____ are _____

  For example, you might write: 50% of coin flips are going to come up heads.
  Or, roughly 51 of every 100 babies born this year will be assigned female at birth.

  Now, try forming this sentence for the number 95% that you see in the problem statement. Then, do the same thing for each part (a) through (d) and see if they have the same structure in the underlines.

Solution:

The 95% in a 95% confidence interval is actually a confidence level (CL). This level is not about the interval itself, but about the confidence procedure (CP) that gave rise to the interval you care about. So, we can correctly say: 95% of random intervals from our CP will capture the parameter μ.

Let's look at each part and see if it is saying the same thing or is some distortion of it.

(a) This part is difficult to fit into the underline strategy since it is actually making a probabilistic claim when no random processes exist. If we try overly hard, we might write: 95% of the time our interval contains μ. People often use the phrase *the time* when they aren't being careful about what precisely they're talking about. For example: "50% of *the time*, a coin returns heads". Here, *the time* should be *all random flips*. In the current problem, *the time* should be *all random CIs*. But we aren't saying what we should, so the given statement is $\boxed{\text{false}}$. This statement is the most common misconception of CIs and for good reason; it shares the second of the two underlines with the correct interpretation and hides the first underline behind the phrase *the time*.

(b) Here, we could write 95% of future means of data sets will be contained in our CI. This is not the same as above, so this statement is $\boxed{\text{false}}$.

(c) This part can be recast as 95% of possible values from the RV X will be in our particular CI. This is not the same as above, so this statement is $\boxed{\text{false}}$.

(d) Equivalently, this statement is claiming that about 380 of every 400 random intervals will contain μ. That is, roughly $\frac{380}{400} = 95\%$ of random CIs will contain μ. This is the same structure as above, so this is a $\boxed{\text{true}}$ interpretation of the 95% CL. Notice that this is moving from the case of all data (e.g., saying there is a 50% chance of getting heads on a fair coin) to a large but finite collection of data ("I expect roughly 500 heads in these 1000 flips"). We use the word "roughly" in this transition to honor the variability present in a finite amount of data.

Commentary:

As mentioned in Problem 19, misconceptions around CIs are widespread. One reason is that many people are not detail-oriented enough. The approach of the hint and solution is to force students to fill in the underlines. This makes you identify the population being discussed (coin flips, babies, random

CIs, etc.) and what feature is being measured (heads, assigned female, capturing μ). When either underline changes, the probability refers to something else. You may recall that a similar attention to detail was needed when discussing conditional probabilities: it matters what comes before and after the conditioning bar. For example: $P(\text{person is aged 18-30} \mid \text{they play in the NFL})$ is roughly 1. But, $P(\text{they play in the NFL} \mid \text{person is aged 18-30})$ is roughly 0. Details!

24. In the modern political era, campaign rallies are being infiltrated by people who do not support the speaker (e.g., to sow dissent, to study those who do support the speaker, etc.). You're curious about this, so you decide to attend a political rally. Your plan is simple: you'll choose 70 random people in the audience and use hidden cameras to videotape them during the rally. When the crowd breaks into a chant, you will use the video footage to see what proportion of your random subjects actually engage in the chant. Suppose you do this and find only 58 of the 70 people took part in the chant.

    (a) Assign notation to and define the population parameter we are trying to study. Then, create an approximate 92% CI for this parameter.

    (b) You may have noticed that your approximate 92% CI from part (a) did not include 100% (or 1, if you were using decimals). Suppose you change the confidence level to C% and the upper bound of the approximate CI exactly equals 100%. Find C.

Hint List:

- b. Trying drawing a picture of $N(0,1)$ once you get an equation involving quantiles. This is my only reliable way of making sense of such equations.

Solution:

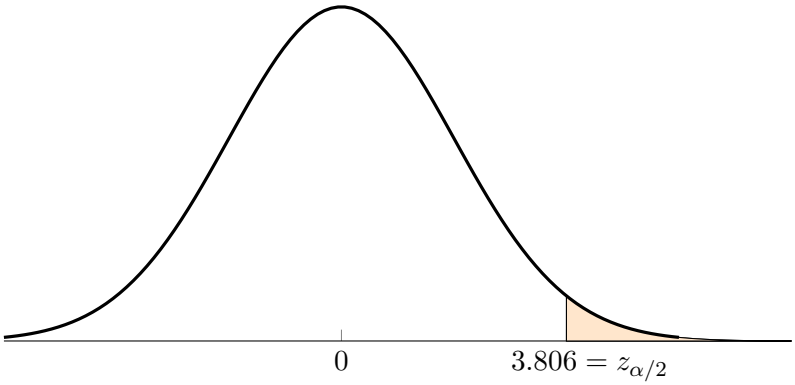
(a) We do inferential statistics because we want to infer from a small collection of data (the sample) that we can study to a large collection (the population) that we can't explore fully. Recall that the parameter is the idea in the population that we are interested in. The parallel idea in the sample is known as the statistic. Remember: a statistic summarizes a sample, while a parameter summarizes a population.

Here, the number $\frac{58}{70}$ is a statistic, a summary of a small sample (70 people). This measures the percent of our small sample that engaged in the chant during the rally. To get the parameter and population, just generalize the idea to a larger, difficult-to-study group. Here, we could let p be the proportion of **all** rally attendees (not just our rally, not just our speaker, not just today) in the modern era who will engage in chants to support their speaker. Using this notation, we know $\widehat{p} = \frac{58}{70}$.

Now, for 92% confidence, $\alpha = 0.08$, so we need $z_{0.08/2} \approx 1.750686$, using qnorm in R. Thus, our CI is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n}} \approx \frac{58}{70} \pm 1.750686\sqrt{\frac{\frac{58}{70}\cdot\frac{12}{70}}{70}} \approx \boxed{(74.974\%, 90.740\%)}.$$

(b) For the upper bound of the CI to be 100% (or 1), we set $\frac{58}{70} + z_{\alpha/2}\sqrt{\frac{\frac{58}{70}\cdot\frac{12}{70}}{70}} = 1$. Solving this gives $z_{\alpha/2} \approx 3.806$. This equation sets up this picture:



From a pnorm calculation, we know the shaded area is $7.06 \cdot 10^{-5}$, but this area is also $\frac{\alpha}{2}$ because of the quantile notation. Setting $\frac{\alpha}{2} = 7.06 \cdot 10^{-5}$ gives $\alpha = 0.000141$, or a confidence level of $100(1-\alpha) \approx \boxed{99.986\%}$.

Commentary: It is tempting to look at your answer in part (b) and get concerned. Doesn't this answer feel too close to 100% confidence? Doesn't the answer look too messy to be correct? When you

start talking to yourself in this way, there are two things you might try. First, stop trying to determine the correctness of your answer by the niceness of your answer. This is an "academic narrative", a story we tell ourselves about how school should operate based on how it has operated in the past. While there is some value in learning how the education system tends to function (in order to more efficiently move through it), you are quickly approaching the point where all your answers are messy because you are nearing the real world, and the real world is messy. This isn't middle school where answers are usually nice whole numbers.

Second, you can use technology to check your answer is reasonable. In part (a), we built an interval based on a 92% confidence level. What if we repeat the calculation for 95% and then 98%, 99%, 99.9%, and so on, heading to our answer from part (b)? As we increase the confidence level, we'll see the upper bound of the interval heading to 1. Here is some code to do this:

```
CLs = c(92, 95, 98, 99, 99.9, 99.986)  # different levels

for (CL in CLs) {
   alpha = 1 - CL/100
   quantile = qnorm(alpha/2, lower=F)
   UB = 58/70 + quantile * sqrt(58*12/70^3)
   print(100 * UB)
}
```

Running this gives us:

```
 90.74331  # same UB as in part (a)
 91.68602
 93.33644
 94.46026
 97.67969
 100.0115  # hoping for UB of about 100%
```

25. One thing that often surprises San Diego newcomers is how present the U.S. military is here. As a statistician at DQ Industries, your boss has tasked you with finding the proportion of San Diego jobs that are connected to the military (this includes jobs at bases, contractors, military R&D, etc.). You are required to draw a large-enough sample so that the sample proportion will be within 1.5% of the true value 90% of the time.

    (a) Suppose you have no information about this proportion and that it costs $5 to contact each person in your sample. What is the least amount of money you can spend to meet your requirements?

    (b) Your boss is horrified by the cost estimate in part (a). You decide to do some online searching to get an initial estimate for the proportion. At a military advising website, you see a claim that 1 in 5 jobs in San Diego is linked to the military sector. Since this is a pro-military group, you figure this number can serve as an upper bound on the true proportion. Given this info, what is the new minimal cost estimate?

Hint List:

- If you're having trouble starting this problem, it is because your knowledge of the class lectures is very thin. This may be because your approach to doing homework is to hunt for something in the lectures that will be helpful. This is the wrong direction to go. Instead, you should start by developing a deep understanding of the content and then head to the homework problems.

Solution:

(a) In class, we saw $n \geq \dfrac{z_{\alpha/2}^2}{4d^2}$. Here, 90% confidence forces $\alpha = 0.1$. If we want to "be within 1.5%", this means $d = 0.015$. In R, we find $z_{\alpha/2} = z_{0.05} = \text{qnorm}(0.05, \text{lower} = \text{F}) \approx 1.645$.

So, $n \geq \dfrac{1.645^2}{4 \cdot 0.015^2} \approx 3006.69$. Rounding up guarantees our requirements (since more data improves any measures of quality). Thus, our cost is $3007 \cdot \$5 = \boxed{\$15035}$.

(b) Let p be the true proportion of current jobs in San Diego connected to the military. Our research suggests $p \leq 0.20$. Since we have this info, we can use the formula $n \geq \dfrac{z_{\alpha/2}^2 \cdot p(1-p)}{d^2}$. This is maximized (i.e., the worst case happens) when p is as close to $\dfrac{1}{2}$ as possible while honoring the info that $p \leq 0.20$. So,

$$n \geq \frac{1.645^2 \cdot 0.2 \cdot 0.8}{0.015^2} \approx 1924.28$$

Rounding up and multiplying by the cost per participant gives a new, cheaper cost: $1925 \cdot \$5 = \boxed{\$9625}$.

Commentary: This problem is nothing special from a cognitive perspective. It does ask you to make sense of a real-world situation and do a few calculations that might actually arise in a job. Nevertheless, at its core, the problem is just plugging numbers into formulas. If anything, the problem should feel like the type of exercise you might come across in a high school course. As you head into college-level work, these types of exercises become less common because it is expected students can use formulas fluently with less practice.

When you find yourself not being intellectually pushed by a problem, use the opportunity to search for additional ways to grow. For example, on this problem, you might:

- Decide you'll commit both of the formulas to memory after finishing.
- Explore how the formulas are related: Why does one have a $\dfrac{1}{4}$ while the other has a $p(1-p)$?
- Try to derive a general formula for the percent reduction in cost when moving from no information [part (a)] to some info [part (b)].

When you study the behaviors of great athletes, musicians, and thinkers, they are constantly trying to push themselves, especially when external forces do not.

26. One frustrating issue with proportions is that your calculations might give a CI like $(-2\%, 7\%)$ or $(96\%, 103\%)$. This sometimes happens when you are studying a trait with a proportion near 0% or 100% (i.e., rare or hyper-prevalent phenomena).

    (a) Suppose you are building an approximate 95% CI for a parameter p using a sample of size 100. If the lower bound of your CI exactly equals 0%, what is your sample proportion? (Note: It may not be possible to actually get this sample proportion because with 100 people the only possible sample proportions are $0, 0.01, 0.02, \ldots, 0.99$, and 1.)

    (b) Suppose you are trying to decide on a sample size for a study to determine what percentage of Americans self-identify as transgender. Based on previous studies, this proportion is somewhere around 1%, so you decide you'd like to be within 0.1% of the true proportion 95% of the time. If you take 7% as an upper bound on the transgender proportion, what is the smallest sample size you can draw?

Hint List:

- a. Try replacing $\widehat{p}$ with x so the equation is less intimidating.
- a. You don't need the quadratic formula or fancy tools to solve the equation. Dividing by x simplifies things.
- b. If you're trying to figure out how the number 1% gets used, it doesn't. See the commentary.

Solution:

(a) We set $0 = \widehat{p} - z_{\alpha/2}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{100}}$, thinking of $\widehat{p}$ as a decimal instead of a percentage. For 95% confidence, we find $z_{\alpha/2} \approx 1.96$ giving $1.96\sqrt{\dfrac{x(1-x)}{100}} = x$, where we have replaced $\widehat{p}$ with x for notational convenience. Squaring gives $1.96^2 \cdot \dfrac{x(1-x)}{100} = x^2$. If $x = 0$, our CI is simply $[0, 0]$, which works. If $x \neq 0$, divide by x to get $\dfrac{1.96^2}{100}(1-x) = x$. Solve to get $x = 0.03699$. So, our possible answers are $\boxed{\widehat{p} = 0\% \text{ or } \widehat{p} \approx 3.7\%}$.

(b) Let p be the percentage of all Americans who identify as transgender. We are told to assume $p \leq 0.07$, $d = 0.001$, and $\alpha = 0.05$. So, in the inequality $n \geq \dfrac{z_{\alpha/2}^2 \cdot p(1-p)}{d^2}$, we replace p with 0.07 since this is the value closest to 0.5 on the zone $[0, 0.07]$. Thus, we have $n \geq \dfrac{1.96^2 \cdot 0.07 \cdot 0.93}{0.001^2} \approx 250088.16$. Rounding up gives $\boxed{n = 250089}$ (or $\boxed{n = 250079}$ if you used a more accurate version of 1.96 in your calculations).

Commentary:

This problem has two lessons to teach us. First, on the statistical side of things, you might be surprised by the very large answer to part (b). The reason this happened is that we are studying a rare trait (current estimates place the transgender percentage around 0.5%). With ultra-rare (near 0%) or ultra-common (near 100%) traits, you need a very small value of d to keep your CI bounds from spilling past 0% or 100% percent (if this matters to you). Once d is small, squaring it and placing it in the denominator leads to a huge value for n. So, be prepared if you want to precisely study ultra-rare or ultra-common phenomena.

Second, part (b) has an interesting piece of information: "Based on previous studies, this proportion is somewhere around 1%". Notice that this number is never used in our calculations. I included it because it motivates the choice of $d = 0.1\%$. Consider this: if you think your proportion of interest is about 40%, then setting $d = 2\%$ isn't problematic. If your proportion is instead closer to 1%, then being within 2% is awkward since you are entertaining negative percentages. So, while this information didn't get used in the calculations, it added some value. When a piece of information is not central to solving a problem, it is called "extraneous" information. As an example, consider this prompt:

A boy has 7 apples and gives 3 to his friend. How many apples remain?

Here, the fact that we care about a boy is not important. Also, it doesn't really matter that we are discussing apples (they could be oranges, candies, etc.) or that a friend received the apples (it could

have been anyone else). When you solve problems in school, there is usually very little extraneous information. In real life, however, you often need to begin by deciding which information is really relevant to the problem and which is irrelevant. Criminal investigators and lawyers spend much of their time deciding on the importance of information in a particular case (i.e., their problem to solve). As a student, avoid the temptation to believe that every piece of info you see must be used somewhere in a formula.

27. Caleb is ordering a piano for his new house. A standard piano has 88 keys, and the lifetime of a key, in years, before it detunes or breaks is independently modeled by an exponential distribution, T, with parameter $\lambda = \frac{1 \text{ breakdown}}{4 \text{ years}}$. Once any one key detunes or breaks, Caleb will have to call the piano tuner to repair his piano.

    (a) When should Caleb expect to call the piano tuner?

    (b) Redo part (a) with n keys and $T \sim Exp(\lambda)$ instead of $T \sim Exp\left(\frac{1}{4}\right)$.

    (c) What we've calculated in parts (a) and (b) is an extremely common value in mechanical engineering called "Time to First Failure" (TTFF). Give 2 ways that TTFF is useful beyond this problem, considering the quality and number of components in a system.

    (d) Give 1 other real-world setting besides piano keys where TTFF is important. Be careful to preserve the iid nature of the components in the system. For example, the components in a car (gas pedal, engine, antifreeze) are (roughly) independent, but not identically distributed. Conversely, ten smoke detectors in the same room are identically distributed, but they are not independent. A suitable answer for this question is the springs on a circular trampoline. These springs are independent and identically distributed, and when one breaks, we might stop using the complete system.

Note: This problem and the next four were co-created with Jack Determan (UCSD '25).

Hint List:

- a. Why is the min order statistic helpful here?
- c. Explore how changing n and λ affect your answer from part (b).

Solution:

(a) Looking ahead to part (b), we will need to explore the case of a general λ and n keys, so we will solve the problem in general, and then let $\lambda = 1/4$ and $n = 88$ to complete this part.

Let $T \sim Exp(\lambda)$ be the breakdown/detuning time for a single piano key. We can model a piano with n keys as $T_1, T_2, \ldots, T_n$ where T_i is the time until key i has a problem. We need to call a tuner once *any* key has a problem, so we need to understand $T_{\min}$, which tells us when we first discover a key with a problem (i.e., the smallest time until a problem arises). Note first that

$$F_T(t) = \int_0^t \lambda e^{-\lambda x}\, dx = -e^{-\lambda x}\Big|_0^t = 1 - e^{-\lambda t}$$

The order statistics pdf formula with $i = 1$ gives

$$f_{T_{\min}}(t) = \frac{n!}{(1-1)!(n-1)!} \cdot F_T(t)^{1-1} \cdot f_T(t) \cdot [1 - F_T(t)]^{n-1} = n \cdot \lambda e^{-\lambda t} \cdot \left[e^{-\lambda t}\right]^{n-1} = n\lambda e^{-n\lambda t}$$

Notice that this looks like the pdf of an exponential distribution, but with rate parameter $n\lambda$ instead of λ. Indeed, we have discovered that $T_{\min} \sim Exp(n\lambda)$. So,

$$E[T_{\min}] = \frac{1}{n\lambda} = \frac{1}{88 \cdot \frac{1}{4}} = \boxed{\frac{1}{22} \text{ years} \approx 16.591 \text{ days}}$$

(b) From part (a), we have $\boxed{E[T_{\min}] = \frac{1}{n\lambda} \text{ years}}$.

(c) If we look at the answer from part (b), we notice:

- As $n \to \infty$, TTFF $\to 0$. This means, the more components you add to a system, the faster the system will break down. This is one argument in favor of electric vehicles. According to recent estimates, gas-powered cars have about 2000 parts in the powertrain (the set of components needed to make the car go), while this number is around 20 for EVs.
- As $\lambda \to \infty$, TTFF $\to 0$. If λ is getting larger, this means there are more breakdowns per unit time (think about the units on λ). Thus, a larger λ is associated with crummier parts. Said simply, if you put bad parts into a system, the system will easily break down.

(d) Answers will vary.

Commentary:

The steps in this problem model a common way of making sense of the world. Often, we begin with a particular object and want to understand it. Here, that was a standard piano in part (a). Then, we begin to generalize but stay in the same setting. Here, that was changing the number and quality of

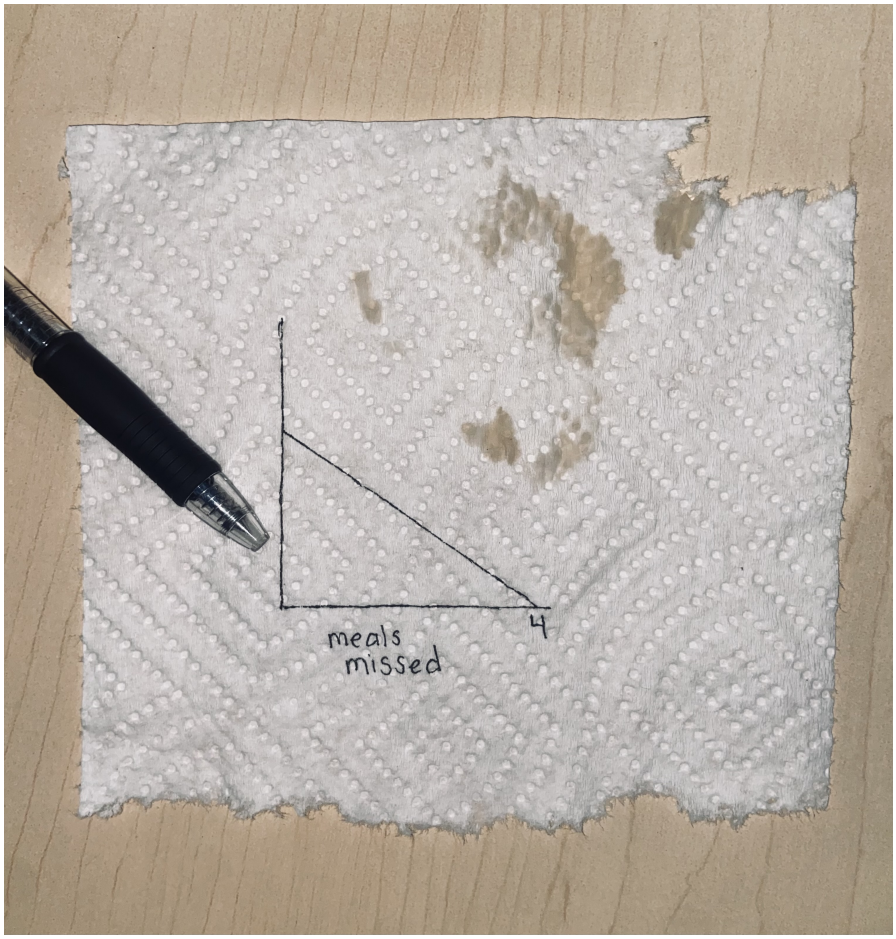
the keys in part (b). After this, we leave our setting completely and think about how pieces of what we're studying relate at a general level. In part (c), we look at TTFF and make sense of this in relation to the number of components in a system (n) and the quality of those components (λ). Finally, in part (d) we return to the real world and think about other objects (besides a piano) that share the same structure (here, iid components). Because we have done part (c), we can carry information we learned from (a) through (c) to these new areas.

This cognitive arc (i.e., specific to general to specific) is one of the ways we study the human experience. Some people enjoy the first part of the arc more (specific to general). For example, you might notice one day that there are many specific pairs of primes with a difference of two (3 and 5, 11 and 13, etc.). This might lead you to conjecture that, in general, there are infinitely many pairs of primes with this property. This is known as the Twin Prime Conjecture and is currently unsolved.

Others enjoy the second part of the arc more (general to specific). For example, you might take the general CI formula $\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and apply it to the specific example of salmon weights in an Alaskan river. Or, you might take the general concept of a Uniform distribution and start seeing it as a model for specific things in your everyday life (e.g., star fragment spawn times in Breath of the Wild, the time of day a comet enters Earth's atmosphere, the direction a spinner points after a hard flick, etc.). As you finish up your time as an undergrad, ask yourself what part of the arc excites you the most. This awareness can be useful in choosing your future direction.

28. Allison is at a party with you, and in conversation, develops an ingenious new method to test whether a community is suffering from food scarcity. She samples n random people from the community, and observes how many meals each misses per week. Then, if the median observation misses more than 2 meals per week, the community is deemed "food scarce".

    In a scrambled haste to model the number of meals per week missed by a random individual in your community, Allison sketches the following image on the back of a napkin:



    In this *continuous* representation, 2.4 meals missed would mean that an individual missed 2 entire meals, and missed out on 0.4 of what they should have eaten during a third meal. Find the probability that Allison would deem your community food scarce given that she takes a sample of 5 people. Use an integral calculator to finish this problem.

Hint List:

- Start by writing the equation of the pdf. Recall that the area under any pdf is 1.
- Which order stat is the median?

Solution: Since the area under a pdf is 1, the height of the line is $\frac{1}{2}$. Thus, the pdf has equation

$$y = \frac{-1/2}{4}x + \frac{1}{2} = -\frac{1}{8}x + \frac{1}{2}$$

So, our pdf is $f_X(x) = -\frac{1}{8}x + \frac{1}{2}$ where $0 \le x \le 4$. From this, we find

$$F_X(x) = \int_0^x -\frac{1}{8}t + \frac{1}{2}\,dt = \left[-\frac{1}{16}t^2 + \frac{1}{2}t\right]_0^x = -\frac{x^2}{16} + \frac{x}{2}$$

With 5 observations, the median is the third order stat, $X_{(3)}$, which has support $[0, 4]$ and pdf

$$\begin{aligned} f_{X_{(3)}}(x) &= \frac{5!}{(3-1)!(5-3)!}F(x)^{3-1}f(x)[1-F(x)]^{5-3} \\ &= \frac{120}{2\cdot 2}\left(-\frac{x^2}{16} + \frac{x}{2}\right)^2 \cdot \left(-\frac{1}{8}x + \frac{1}{2}\right) \cdot \left(1 + \frac{x^2}{16} - \frac{x}{2}\right)^2 \end{aligned}$$

Now, $P(\text{community is food scarce}) = P(X_{(3)} > 2) = \int_2^4 f_{X_{(3)}}(x)\,dx$. Using our order stat pdf above and an integration tool, we find the probability is about $\boxed{0.104}$.

Commentary: At first, this problem might feel a little silly to you.

- A conversation at a party and a napkin sketch? In truth, many interesting ideas arise in this way.
- Conceptualizing food scarcity through meals missed and a *continuous* distribution? Well, how would you do it?
- Defining a community as food scarce if the median member is experiencing something? Can you think of a better definition?

When we talk about issues in the real world, we tend to use vague terms that don't have a single definition. For example, what does "climate change" actually mean? What about "housing insecurity" or "food scarcity"? One job of a statistician/data scientist is to help people interested in such issues come up with a way to measure and analyze them. While different people may use different setups, definitions, and analyses, the point is not that a single way of thinking should exist, but rather, when desired, we should have the ability to *quantify* and *rigorously study* the real world using statistics. To look at the world with a statistical pair of glasses is to help people transition from vague concepts to precise measures. There is no way to get all parties to agree (we can't even get the world to use a common system for temperature measurement!), but this shouldn't be the goal. Rather, you should seek out *reasonable choices* as you make the transition and *responsibly communicate them* to others.

29. Over the next month, Jack will run the 800 meter dash 10 times to try to run a time that is faster than or equal to his current fastest time of 112 seconds. His times are iid from the distribution

$$f_X(x) = \frac{1}{3}\left(\frac{x}{120}\right)^{39} \exp\left[-\left(\frac{x}{120}\right)^{40}\right] \text{ where } x > 0$$

(a) Use a graphing calculator or Desmos to graph $f_X(x)$, and decide whether the pdf is reasonable for a runner who generally runs around 120 seconds in the 800 meter dash. Include a screenshot or sketch of the pdf as part of your answer.

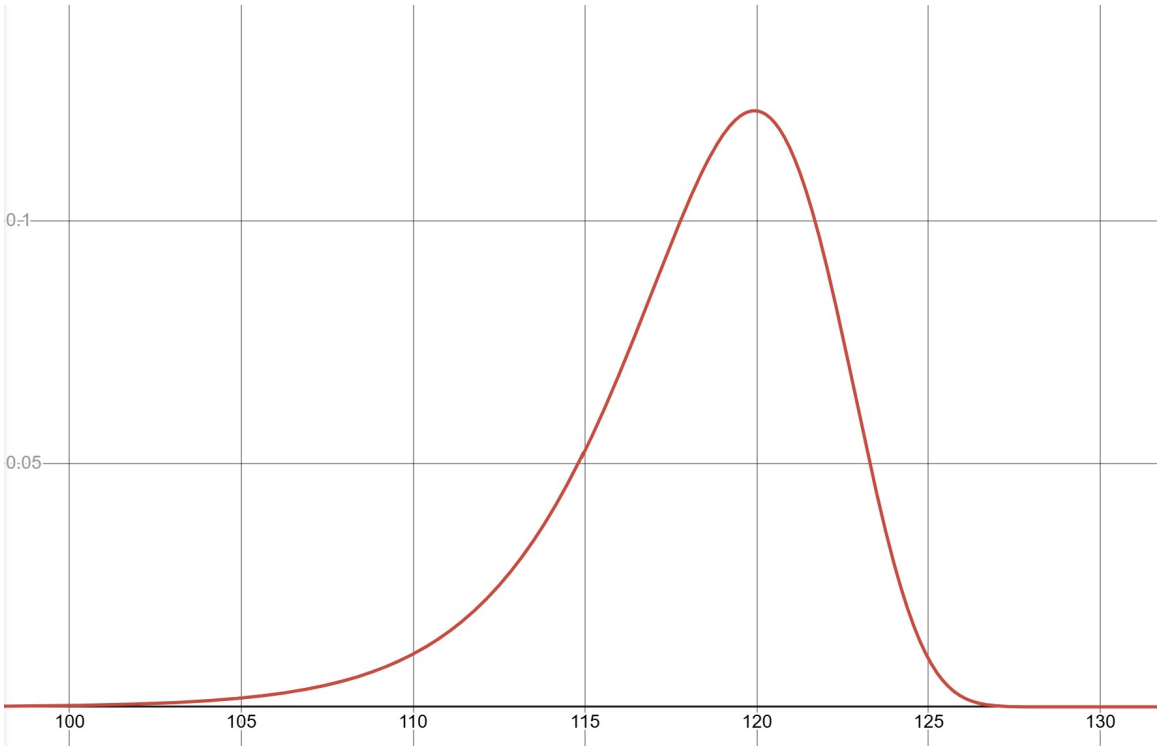
(b) What is the probability that Jack runs a time that is faster than or equal to his current fastest time in the 800 meter race during these 10 attempts?

Hint List:

- b. Which order stat captures your best attempt when we are discussing running times?
- b. Use a u-substitution to dramatically simplify the cdf calculation.

Solution:

(a) Typing the pdf into a graphing tool gives us this picture:



This seems like a reasonable pdf because it dies off beyond 100 and 130 (Jack will never be extremely different than his typical pace) and it peaks near 120, his most common time. It also allows for times a little above and below his typical pace, which captures the variability inherent in running (due to weather conditions, energy level, training level, etc.).

(b) Let $X_1, \ldots, X_{10}$ be the times for Jack's ten 800 meter attempts over the month. If Jack is "faster than or equal to" his best time of 112 seconds, this would mean that the minimum of these times is at or below 112. So, this problem wants $P(X_{(1)} \leq 112)$. To find the pdf of $X_{(1)}$, we first find

$$F_X(x) = \int_0^x \frac{1}{3}\left(\frac{t}{120}\right)^{39} \exp\left[-\left(\frac{t}{120}\right)^{40}\right] dt$$

Letting $u = \left(\frac{t}{120}\right)^{40}$, we find $du = 40\left(\frac{t}{120}\right)^{39} \cdot \frac{1}{120} = \frac{1}{3}\left(\frac{t}{120}\right)^{39}$. Our new lower bound is $u = \left(\frac{0}{120}\right)^{40} = 0$, and our new upper bound is $u = \left(\frac{x}{120}\right)^{40}$. Substituting gives

$$F_X(x) = \int_0^{(x/120)^{40}} e^{-u}\, du = -e^{-u}\Big|_0^{(x/120)^{40}} = 1 - e^{-(x/120)^{40}}$$

From here, we use the order stat formula with $i = 1$ and $n = 10$:

$$\begin{aligned} f_{X_{(1)}}(x) &= \frac{10!}{(1-1)!(10-1)!} F_X(x)^{1-1} f_X(x) \left(1 - F_X(x)\right)^{10-1} \\ &= 10 \left[\frac{1}{3}\left(\frac{x}{120}\right)^{39} e^{-(x/120)^{40}}\right] \left[1 - \left(1 - e^{-(x/120)^{40}}\right)\right]^9 \\ &= \frac{10}{3}\left(\frac{x}{120}\right)^{39} e^{-10(x/120)^{40}} \end{aligned}$$

Finally, we find $P\left(X_{(1)} \leq 112\right) = \displaystyle\int_0^{112} \frac{10}{3}\left(\frac{x}{120}\right)^{39} e^{-10(x/120)^{40}}\, dx$. We do a u-substitution with $u = 10(x/120)^{40}$ to get

$$P\left(X_{(1)} \leq 112\right) = \int_0^{10\cdot(112/120)^{40}} e^{-u}\, du = \boxed{1 - e^{-10\cdot(112/120)^{40}} \approx 0.469}$$

Commentary: There are several reasons why this is an excellent problem. First, it uses an honest distribution to model something you might care about (your 800 meter times if you're a runner). Second, it asks a natural question: If I do a bunch of time trials, what's the chance I break my record? Since speed records represent a minimum, order statistics naturally arise. If the question had a sports setting where records are maxima (like weightlifting totals or archery scores), a maximum would appear. So, the problem has authenticity.

Third, the problem is notation-heavy, so you must be very detail-oriented to see it through. We have exponents inside exponents, so even displaying our work is a challenge. One thing that might have surprised you was this: despite the nightmarish notation, it is actually possible to do the problem completely by hand, and moreover, the order stat pdf wasn't really worse than the starting distribution (compare the previous problem). In fact, if you do some clever rearranging, we can write

$$f_{X_{(1)}}(x) = \frac{\sqrt[40]{10}}{3}\left(\frac{x}{120/\sqrt[40]{10}}\right)^{39} \exp\left[-\left(\frac{x}{120/\sqrt[40]{10}}\right)^{40}\right]$$

This looks a lot like the pdf of X with 120 replaced by $120/\sqrt[40]{10}$. Said differently, whatever distribution family X belongs to, it appears that $X_{\min}$ belongs to the same family but with an adjustment to the starting parameters. With some internet searching, you can discover that $X \sim Weibull(k = 40, \lambda = 120)$, and so we have a new conjecture based on this problem:

$$\text{If } X \sim Weibull(k = 40, \lambda = 120), \text{ then } \min(X_1, \ldots, X_{10}) \sim Weibull\left(k = 40, \lambda = 120/\sqrt[40]{10}\right)$$

From this, we could generalize our conjecture:

$$\text{If } X \sim Weibull(k, \lambda), \text{ then } X_{(1)} = \min(X_1, \ldots, X_n) \sim Weibull\left(k, \lambda/\sqrt[k]{n}\right)$$

Indeed, you'll find this conjecture is true and listed halfway down the link above. The point here is that the algebra we did in this problem was trying to tell us something. If you keep your ears open, mathematics will whisper its truths. Listen hard enough to reach glory.

30. (a) Consider a random variable $M \sim N(10, \sigma^2 = 2)$.

i. Simulate 10,000 samples of size $n = 25$, and plot a histogram displaying the distribution of $M_{(1)}$. Then, repeat this process for $M_{(10)}, M_{(13)}$, and $M_{(25)}$. Finally, plot an overlaid histogram of all four distributions.

ii. Simulate 1,000 samples of size 5, and plot a histogram displaying the distribution of $M_{(1)}$. Then, repeat this process for samples of size 20 and 100, and plot an overlaid histogram of all three distributions.

On both sub-parts, submit only the overlaid histogram and code.

(b) Peter is trying to become an MLB pitcher, so he attends the Padres tryouts. As part of his tryout, Peter must throw 20 fastballs at a consistent speed. More specifically, the difference between his fastest and slowest fastballs must be no more than 3 miles per hour.

i. Using a simulation in R, plot the distribution of the difference between Peter's fastest and slowest fastballs if the speed of a throw is $Y \sim N(97, \sigma^2 = 0.6)$. Use 5,000 groups of 20 fastballs. (Note: The statistic we are simulating is $Y_{(20)} - Y_{(1)}$, or the range of the speeds of Peter's pitches. To calculate this by hand, we would need to first calculate the pdfs of $Y_{(20)}$ and $Y_{(1)}$, which involve the cdf of the normal distribution. This cdf does not have a closed form, making simulation a far easier way to visualize the distribution of the range.)

ii. In what proportion of your simulations did Peter satisfy the requirement that his fastest and slowest fastballs be within 3 miles per hour of each other?

Hint List:

- a. It helps to make a function that outputs a particular order statistic (across several simulations) from a normal distribution based on some inputs. I declared my function this way:

```
makeorderstats = function(sims, size, i, mu, variance) {
     # code in here
}
```

- a. When making overlapping histograms using the base graphics package in R (i.e., not in ggplot), I like to use the first hist command to set the labels on the axes and the main title. On subsequent calls to hist, use the add=T parameter to add atop an existing histogram. It also helps to use a common set of x bounds and bin partition locations each time you call the hist command.
- b. Push yourself by seeing if you can generate all the range statistics in a single line of code! When you go to plot the *distribution* (instead of the histogram), the density command in R is very useful.
- b. For part ii, again, push yourself to do this in one line of code.

Solution:

(a) i.

```
makeorderstats = function(sims, size, i, mu, variance) {
    ostats = numeric(sims)

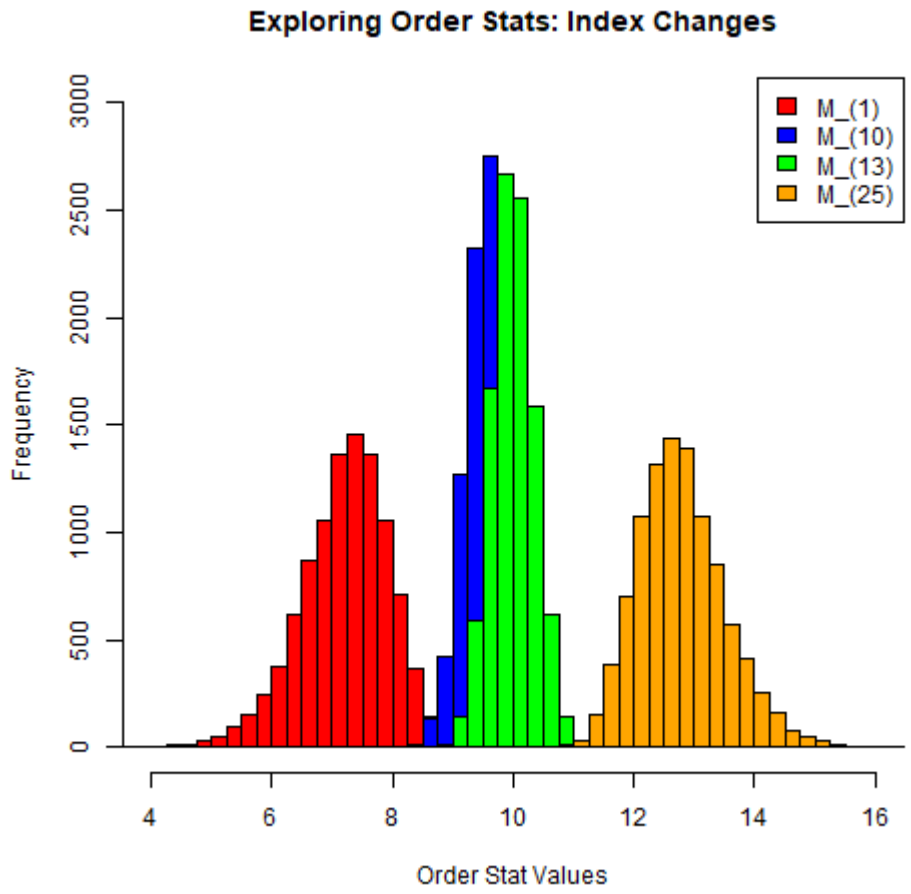
    for (j in 1:sims) {
      dataset = rnorm(size, mu, sqrt(variance))
      sorted = sort(dataset)
      ostats[j] = sorted[i]
    }

    return(ostats)
}

# Make first histogram, set boundaries and title
data = makeorderstats(10000, 25, 1, 10, 2)
hist(data, xlim=c(4, 16), ylim=c(0, 3000),
     xlab="Order Stat Values",
     breaks=seq(-20, 30, 0.25), col="red",
     main="Exploring Order Stats: Index Changes")

# Add more histograms
cols = c("blue", "green", "orange")
indices = c(10, 13, 25)
for (j in 1:3) {
  data = makeorderstats(10000, 25, indices[j], 10, 2)
  hist(data, breaks=seq(-20, 30, 0.25), col=cols[j], add=T)
}

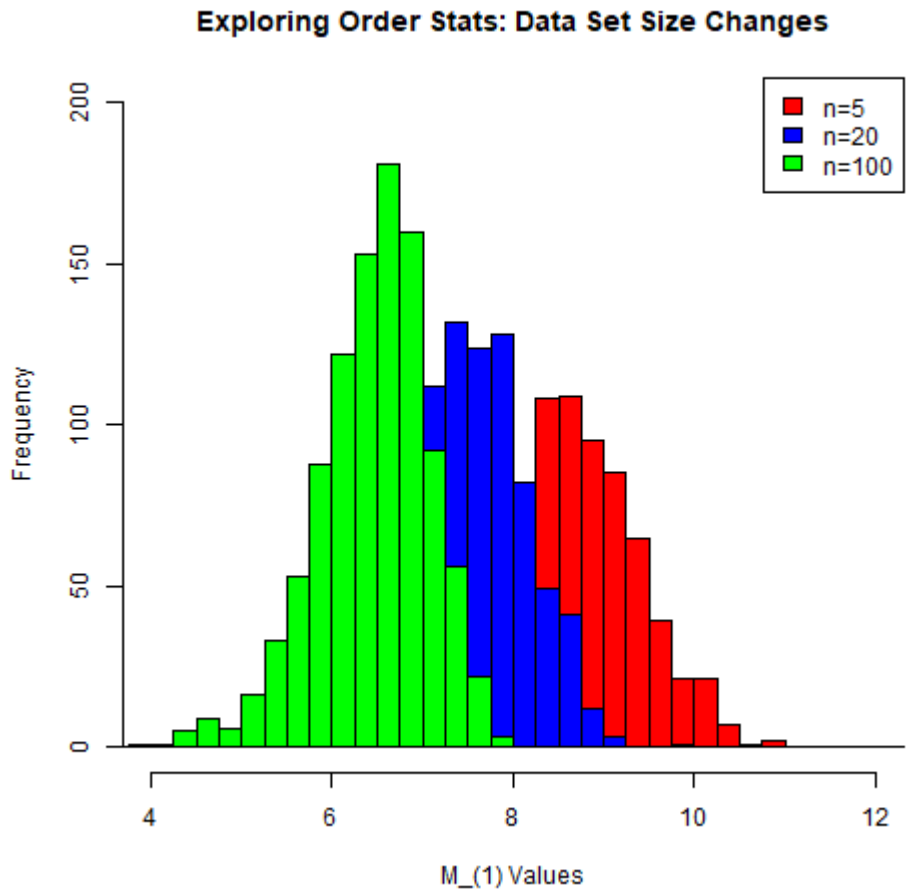
legend("topright", c("M_(1)", "M_(10)", "M_(13)", "M_(25)"),
       fill=c("red", cols))
```



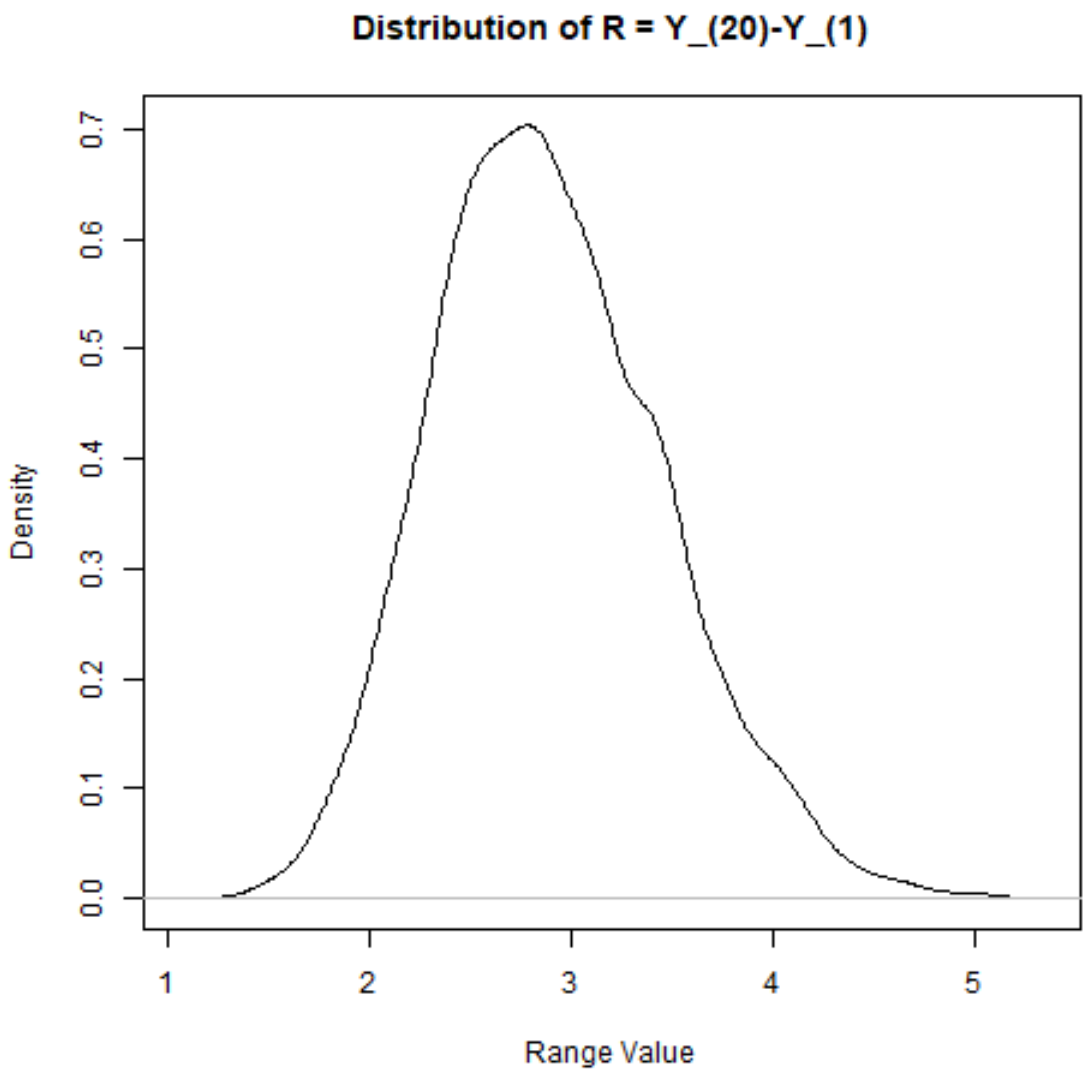
ii.
```
# Make first histogram, set boundaries and title
data = makeorderstats(1000, 5, 1, 10, 2)
hist(data, xlim=c(4, 12), ylim=c(0, 200),
     xlab="M_(1) Values",
     breaks=seq(-20, 30, 0.25), col="red",
     main="Exploring Order Stats: Data Set Size Changes")

# Add more histograms
cols = c("blue", "green")
sizes = c(20, 100)
for (j in 1:2) {
  data = makeorderstats(1000, sizes[j], 1, 10, 2)
  hist(data, breaks=seq(-20, 30, 0.25), col=cols[j], add=T)
}

legend("topright", c("n=5", "n=20", "n=100"),
       fill=c("red",cols))
```



(b) i.
```
ranges = replicate(5000, diff(range(rnorm(20, 97, sqrt(0.6)))))
plot(density(ranges), xlab="Range Value",
     main="Distribution of R = Y_(20)-Y_(1)")
```



ii. `mean(ranges <= 3)`

This code show about 60% of 20-fastball-sets will have a narrow enough range for Peter to make the team.

Commentary: This problem was created because statistics can quickly become so complex that algebraically deriving results is impossible or nightmarish. In such cases, simulation is really valuable, since you can see what is going on with relative ease. Indeed, you probably spent very little time writing the code to simulate the desired statistics in parts (a) and (b) while most of your time was consumed with creating graphics (which are usually exhausting and frustrating). So, the main point of this problem was to let you experience the simulation of complicated statistics.

As usual, though, the problem has much more to offer. After you created your images, did you spend time examining them and wondering about things? It's hard to become curious in life if you never give yourself the space to ask questions or make original observations. For example, in (a)(ii),

can you explain why the histograms shift to the left as the sample size increases? Why do you think the histograms are getting more peaked as the sample size increases? In (a)(i), did you notice the shape of $M_{(1)}$ and $M_{(25)}$? In the lecture animation for order stats, we saw different shapes for the min and max order stats than we got here. Why do we get a different shape here? Also, $M_{(13)}$ is actually the median order stat (when $n = 25$). When you look at this histogram, it appears to be centered at 10 (like the original distribution) and *appears to be* normal. Is it normal? Interestingly, the first few professors I asked weren't sure of the distribution of $M_{(13)}$. Also, how does the variance of the green histogram relate to the variance of the original distribution $M \sim N(10, \sigma^2 = 2)$? These questions can lead you to some internet searches where you'll discover that if we take $n = 2m + 1$ (an odd number!) pieces of data from $X \sim N(\mu, \sigma^2)$, then the median $X_{(m+1)} \approx N\left(\mu, \frac{\pi\sigma^2}{4m}\right)$. If you don't believe this, you might be inspired to write some R code to check it out:

```
n = 25
m = (n-1)/2
sigmasq = 2 # variance
data = makeorderstats(10000, n, m+1, 10, sigmasq)
qqnorm(data)
qqline(data, col="red")
var(data)
pi*sigmasq/(4*m)
```

Here I use a tool known as a qqplot to determine if the data are roughly normal and I check the variance claim at the end. Run the code yourself to see the results.

31. Victoria is a competitive diver getting ready to compete at a major event. In diving, athletes are often assigned an overall score in the following way:

- 7 judges each give the diver a score.
- The highest and lowest scores are removed.
- The remaining 5 scores are averaged.

This is called the "trimmed mean" of the scores and is used to minimize the impact of extreme scoring and biased judging. Assume Victoria's scores from each judge are independently distributed according to a uniform distribution on the interval $[0, 10]$. Victoria is interested in her expected overall score and asks you to help her calculate it.

(a) **[Least elegant, most tech-dependent approach]** Using an integral calculator (!), find Victoria's expected score by calculating this nightmare:

$$\frac{E\left[X_{(2)}\right] + E\left[X_{(3)}\right] + E\left[X_{(4)}\right] + E\left[X_{(5)}\right] + E\left[X_{(6)}\right]}{5}$$

(b) **[More elegant, non-tech approach]**

i. Show why $\dfrac{E\left[X_{(2)}\right] + E\left[X_{(3)}\right] + E\left[X_{(4)}\right] + E\left[X_{(5)}\right] + E\left[X_{(6)}\right]}{5}$ is the same as $\dfrac{7 \cdot E\left[\overline{X}\right] - E\left[X_{(7)}\right] - E\left[X_{(1)}\right]}{5}$.

ii. Find $E\left[X_{(7)}\right]$ by hand (i.e., without technology or a calculator).

iii. Explain how you can infer $E\left[X_{(1)}\right]$ from $E\left[X_{(7)}\right]$ without actually calculating $E\left[X_{(1)}\right]$.

iv. Find Victoria's expected overall score using parts (i) through (iii).

(c) **[Most elegant, no-work-needed approach]** Now that you've seen the answer from parts (a) and (b), explain in words how to intuitively get Victoria's expected score without doing any real work.

Hint List:

- a. Try writing the pdf of $X_{(i)}$ without specifying what i is, and again, write $E[X_{(i)}]$ without a particular value of i. Then, let $i = 2, 3, \ldots, 6$ as you do calculations.
- b(i). Adding a list of data before and after sorting always gives the same result. Can you express this fact using random variables?

Solution:

(a) If $X \sim Unif(0, 10)$, then $F_X(x) = \displaystyle\int_0^x \frac{1}{10}\,dt = \frac{x}{10}$. So, the general order stat has pdf

$$f_{X_{(i)}}(x) = \frac{7!}{(i-1)!(7-i)!}\left(\frac{x}{10}\right)^{i-1}\frac{1}{10}\left(1-\frac{x}{10}\right)^{7-i} = \frac{7!}{(i-1)!(7-i)!}\cdot\frac{1}{10^7}\cdot x^{i-1}\,(10-x)^{7-i}$$

So, we must evaluate $E[X_{(i)}] = \displaystyle\int_0^{10} x \cdot \frac{7!}{(i-1)!(7-i)!}\cdot\frac{1}{10^7}\cdot x^{i-1}\,(10-x)^{7-i}\,dx$ for $i = 2, \ldots, 6$.

We use a calculator for each integral and eventually get $\dfrac{2.5 + 3.75 + 5 + 6.25 + 7.5}{5} = \boxed{5}$.

(b) i. Notice that $X_1 + \cdots + X_7 = X_{(1)} + \cdots + X_{(7)}$. That is, it doesn't matter if you add your data up or sort and then add, the sum will be the same. Hence,

$$E\left[\,\overline{X}\,\right] = E\left[\frac{X_1 + \cdots + X_7}{7}\right] = E\left[\frac{X_{(1)} + \cdots + X_{(7)}}{7}\right]$$

So, $\dfrac{7\cdot E\left[\overline{X}\right] - E\left[X_{(7)}\right] - E\left[X_{(1)}\right]}{5} = \dfrac{E\left[X_{(1)} + \cdots + X_{(7)}\right] - E[X_{(7)}] - E[X_{(1)}]}{5}$.

Breaking the expectation into pieces and canceling gives

$$\frac{E\left[X_{(2)}\right] + E\left[X_{(3)}\right] + E\left[X_{(4)}\right] + E\left[X_{(5)}\right] + E\left[X_{(6)}\right]}{5}$$

ii. From part (a), we must find

$$\begin{aligned} E[X_{(7)}] &= \int_0^{10} x \cdot \frac{7!}{(7-1)!(7-7)!}\cdot\frac{1}{10^7}\cdot x^{7-1}\,(10-x)^{7-7}\,dx \\ &= \frac{7}{10^7}\int_0^{10} x^7\,dx = \frac{7}{10^7}\cdot\frac{1}{8}10^8 = \frac{70}{8} = \boxed{8.75}. \end{aligned}$$

iii. Note that $X \sim Unif(0, 10)$ is symmetric about the value 5. As such, the pdfs of $X_{(1)}$ and $X_{(7)}$ are reflections across the line $x = 5$. Hence, their expectations are on opposite sides of and equidistant from 5. So $E[X_{(1)}] = 5 - \big(E[X_{(7)}] - 5\big) = 5 - (8.75 - 5) = \boxed{1.25}$.

iv. Since $E[\,\overline{X}\,] = E[X] = \dfrac{0+10}{2} = 5$, our result from part (b)i becomes:

$$\frac{7\cdot 5 - 8.75 - 1.25}{5} = \frac{35 - 10}{5} = \boxed{5}.$$

(c) Continuing the logic of (b)iii, we see that the pdfs of $X_{(2)}$ and $X_{(6)}$ will also be reflections about 5, as will $X_{(3)}$ and $X_{(5)}$. Once you put expectations on these random variables, you create values that pair and are symmetrically placed around 5. Finding the average of numbers that symmetrically pair about a given value will always yield that given value. Here, it is $\boxed{5}$.

Commentary: As the parts of this problem clearly demonstrate, there are often multiple ways to do a problem. Here, these parts arose during the problem construction process. Originally, the goal was to simply ask part (a). But, when the integrals were unpleasant to do by hand, a natural question arose: Is there a way to do this that avoids technology? Part (b) achieves this, but requires some cleverness to change which integrals must be calculated.

Another thing to notice is that this problem is simultaneously very authentic and inauthentic. The authenticity relates to the use of order stats. In many human-judged sports (like diving), judges provide scores and the most extreme scores are thrown out. The inauthenticity comes from the assumption that $X \sim Unif(0, 10)$. This choice was made to simplify the mathematics present. In school, teachers often make this choice so that the difficulty level will be appropriate for students. In the real world, however, you will not be shown such kindness.

32. In this problem, we explore more estimators for μ and σ^2 in the distribution $N(\mu, \sigma^2)$ using a random sample $X_1, \ldots, X_n$.

(a) Typically, people use $\widehat{\mu}_1 = \overline{X}$ as an estimator for μ. You might also use $\widehat{\mu}_2 = \dfrac{2X_1 + X_2}{3}$ or $\widehat{\mu}_3 = \dfrac{2\left(X_1 + 2X_2 + 3X_3 + \cdots + nX_n\right)}{n^2 + n}$. Show that all three of these estimators are unbiased.

Note: $\widehat{\mu}_2$ might be used if you didn't trust the data $X_3, \ldots, X_n$, while $\widehat{\mu}_3$ might be used if you wanted to give increasing importance to data collected later in the process.

(b) In class, we showed that $\widehat{\sigma_1^2} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is a biased estimator for σ^2 when both μ and σ are unknown. Suppose, however, that μ is known and so we can use $\widehat{\sigma_2^2} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(X_i - \mu\right)^2$. Show that $\widehat{\sigma_2^2}$ is actually unbiased.

Hint List:

- a. For the unbiasedness of $\widehat{\mu}_3$, you'll need to recall how to sum the first n integers. Have you ever seen a formula for $\sum_{i=1}^{n} i$?
- b. There are several ways to do this problem. If you want the fastest way, ask yourself: How was the variance of a random variable X *defined* in my probability course?

Solution:

(a) The first estimator is unbiased because $E[\widehat{\mu}_1] = E[\,\overline{X}\,] = E[X] = \mu$.

Next, $E[\widehat{\mu}_2] = \dfrac{2E[X_1] + E[X_2]}{3} = \dfrac{2\mu + \mu}{3} = \mu$.

Finally, $E[\widehat{\mu}_3] = \dfrac{2\left(E[X_1] + 2E[X_2] + \cdots + nE[X_n]\right)}{n(n+1)} = \dfrac{2\left(\mu + 2\mu + \cdots + n\mu\right)}{n(n+1)}$.

Pulling out μ gives $\dfrac{2\mu}{n(n+1)}\left(1 + 2 + \cdots + n\right) = \dfrac{2\mu}{n(n+1)} \cdot \dfrac{n(n+1)}{2} = \mu$.

(b) $E[\widehat{\sigma_2^2}] = \dfrac{1}{n} E\left[\sum_{i=1}^{n} (X_i - \mu)^2\right] = \dfrac{1}{n} \sum E\left[(X_i - \mu)^2\right] \overset{\text{id dist}}{=} \dfrac{1}{n} \sum E\left[(X - \mu)^2\right]$.

The variance of a random variable is defined as $E[(X-\mu)^2]$, so the above expression is

$$\frac{1}{n} \sum Var[X] = \frac{1}{n} \sum \sigma^2 = \frac{1}{n}\left(n\sigma^2\right) = \sigma^2.$$

Commentary: This is your entry-level problem for showing estimators are unbiased. I tried to add some spice by including some strange estimators ($\widehat{\mu}_2$ and $\widehat{\mu}_3$) and refreshing knowledge from your past $\left(\text{e.g., } \sum_{i=1}^{n} i = \dfrac{n(n+1)}{2} \text{ and the definition of variance}\right)$. Notice that part (a) shows that multiple unbiased estimators might exist for a given parameter, and hence, we will want additional criteria (besides unbiasedness) to help choose among estimators. Part (b) shows that with a little extra info (i.e., the value of μ), we can transform a biased estimator ($\widehat{\sigma_1^2}$) into an unbiased estimator ($\widehat{\sigma_2^2}$). The real trouble with $\widehat{\sigma_1^2}$ is that it has *layers* of estimation. Not only is it trying to estimate σ^2 (the primary layer), but it also needs to estimate μ (via $\overline{X}$, the second layer) in the process. In general, it is difficult to maintain unbiasedness when you do such layering.

This problem is another good example where you could miss these observations if you are just "getting the homework done" and pushing through the algebra. You should always look back over problems you complete and try to imagine the commentary the instructor would have offered if they were using the Problem Book approach to teaching. The best question to facilitate this is: Why was this problem assigned?

33. Suppose you decide to randomly generate numbers from $X \sim Unif(0, \theta)$. Your friend will ask for n numbers and then use this information to guess what value you (secretly) chose for θ. Typically, one might use $\widehat{\theta}_{\text{MLE}} = \max X_i = X'_n = X_{(n)}$ to estimate θ. Your friend, however, has meganumerophobia, and is afraid to say the maximum number in the random sample. Instead, they'll say the second largest number: $\widehat{\theta} = X'_{n-1} = X_{(n-1)}$. Determine the bias of this estimator by carefully finding the density function for X'_{n-1} and continuing from there. If the estimator is biased, check if it is asymptotically unbiased, and also modify it to create a new unbiased estimator.

Hint List:

- Once you get to the expected value, break the integral into two power functions.

Solution:

Note that $f(x;\theta) = \dfrac{1}{\theta}$ on $[0,\theta]$. So, $F(x) = \displaystyle\int_0^x \frac{1}{\theta}\,dt = \frac{x}{\theta}$. Letting $i = n-1$ in the order stat pdf gives

$$f_{X_{(n-1)}}(x) = \frac{n!}{((n-1)-1)!(n-(n-1))!}F^{(n-1)-1}\cdot f\cdot(1-F)^{n-(n-1)} = n(n-1)\left(\frac{x}{\theta}\right)^{n-2}\frac{1}{\theta}\left(1-\frac{x}{\theta}\right)$$

Recall that the support for an order stat always matches that of the original distribution. Thus,

$$\begin{aligned}
E[X_{(n-1)}] &= \int_0^\theta x\cdot n(n-1)\left(\frac{x}{\theta}\right)^{n-2}\frac{1}{\theta}\left(1-\frac{x}{\theta}\right)\,dx = \frac{n(n-1)}{\theta^{n-1}}\int_0^\theta x^{n-1}-\frac{x^n}{\theta}\,dx \\
&= \frac{n(n-1)}{\theta^{n-1}}\left[\frac{x^n}{n}-\frac{x^{n+1}}{\theta(n+1)}\right]_0^\theta = \frac{n(n-1)}{\theta^{n-1}}\left[\frac{\theta^n}{n}-\frac{\theta^n}{n+1}\right] \\
&= \frac{n(n-1)}{\theta^{n-1}}\left[\frac{\theta^n}{n(n+1)}\right] = \frac{n-1}{n+1}\theta \neq \theta.
\end{aligned}$$

So, $X_{(n-1)}$ is biased, but since $\displaystyle\lim_{n\to\infty}\frac{n-1}{n+1} = 1$, it is $\boxed{\text{asymptotically unbiased}}$.

Note that $\boxed{\widehat{\theta}_{\text{new}} = \dfrac{n+1}{n-1}X_{(n-1)}}$ is unbiased since $E[\widehat{\theta}_{\text{new}}] = \dfrac{n+1}{n-1}E[X_{(n-1)}] = \dfrac{n+1}{n-1}\cdot\dfrac{n-1}{n+1}\theta = \theta$.

Commentary: The reason I assign this problem is because it requires you to use multiple skills at once. Specifically, you need to find an order statistic, calculate bias, understand asymptotic unbiasedness, and alter estimators to build unbiased ones. My experience is that the most interesting (and rewarding) problems in life demand that multiple skills be brought together to solve a single challenge. For really large challenges (e.g., getting to the moon or building a self-driving car), no single person has all these skills, so different people or teams must be assembled where each does its part. In the classroom, which is a smaller and more well-defined setting, you can ask a single student to leverage multiple knowledge sets.

Sadly, much of education has become "atomized". Teachers see their courses as a long list of individual skills to be trained and tested. The internet is full of videos on how to use a single technique (e.g., "Understand u-sub in 3 Easy Steps!") or even do a particular type of problem within a technique (e.g., "Defeating the Integral $\displaystyle\int e^x\sin x\,dx$ via IBP"). While there is certainly a place for focusing on a single skill at once, education must do better. When possible on homework and exams, we need to bring together ideas from across a course and, when appropriate, across many courses. As a metaphor, imagine if the music world did nothing but assign musicians etudes (pieces that usually focus on a single technique). We would miss the great masterpieces which bring together numerous skills to create a larger, and more rewarding, experience. The takeaway here is this: If you find yourself in an educational or vocational space that focuses too much on atomized skills, seek out richer experiences.

34. Suppose you have $X \sim Binom(n, p)$ where n is known and p is unknown. Typically, people use $\widehat{p} = \dfrac{X_1}{n}$ to estimate p, where $X = X_1$ is simply a sample of size 1. Note that a sample of size 1 from a Binomial RV is equivalent to n Bernoulli trials. This might represent simultaneously flipping n coins (just once!) and counting the number of heads you see, where each coin has $p_{\text{heads}} = p$. Now, if both n and p are given, we know the variance, V, of X is just $np(1-p)$. If p is unknown, you might want to estimate V using the estimator $\widehat{V} = n\left(\dfrac{X_1}{n}\right)\left(1 - \dfrac{X_1}{n}\right)$. Find the bias of $\widehat{V}$, and if it is biased, determine if it is asymptotically unbiased, and also modify $\widehat{V}$ to create a new unbiased estimator.

Hint List:

- You'll need to do some factoring to make progress on this problem. Remember that you are trying to compare to $np(1-p)$, so work toward factorizations that have pieces of this goal.

Solution:

$$\begin{aligned} E[\widehat{V}] &= E\left[X_1 - \frac{X_1^2}{n}\right] = E[X_1] - \frac{1}{n}E[X_1^2] = E[X] - \frac{1}{n}\left(Var[X] + E[X]^2\right) \\ &= np - \frac{1}{n}\left(np(1-p) + (np)^2\right) = np - p(1-p) - np^2 \\ &= p[n - 1 + p - np] = p(n-1)(1-p) = (n-1)p(1-p) \end{aligned}$$

$$\begin{aligned} \text{Bias}[\widehat{V}] &= E[\widehat{V}] - V = (n-1)p(1-p) - np(1-p) \\ &= (n - 1 - n)p(1-p) = -p(1-p) \neq 0 \end{aligned}$$

So, $\widehat{V}$ is $\boxed{\text{biased}}$. Also, $\lim_{n\to\infty} \text{Bias}[\widehat{V}] = \lim_{n\to\infty} -p(1-p) = -p(1-p) \neq 0$, so $\widehat{V}$ is $\boxed{\text{asymptotically biased}}$. Your first inclination to fix the bias might be to create $\widehat{V}_{\text{new}} = \widehat{V} + p(1-p)$. The only trouble is that this estimator uses p, which we don't know, so it is not well defined. As an alternative, if we look at $E[\widehat{V}]$, we can set $\widehat{V}_{\text{new}} = \dfrac{n}{n-1}\widehat{V}$. Then, we have

$$E[\widehat{V}_{\text{new}}] = \frac{n}{n-1}E[\widehat{V}] = \frac{n}{n-1} \cdot (n-1)p(1-p) = np(1-p) = V$$

Commentary: There are two interesting things to notice here. First, note that we do the most natural thing possible in this problem and it doesn't yield niceness. In particular, since we don't know p, we estimate it first via $\dfrac{X_1}{n}$, and then we estimate the variance, $np(1-p)$, by using the estimate we just built: $n\left(\dfrac{X_1}{n}\right)\left(1 - \dfrac{X_1}{n}\right)$. This is another example where *layering* estimation (recall the Commentary on Problem 32) doesn't work as we might hope.

A second take-away is this: one way to understand a subject is through a list of key examples and counterexamples. This is your first example of an asymptotically *biased* estimator. You might start a stack of notecards with important examples from 181A. For example, you could make cards for:

- A RV whose MME and MLE are (not) the same
- An asymptotically (un)biased estimator
- An MLE that is (not) found via the standard pipeline ($L \to \ell \to \ell' = 0$)

Add cards as you encounter important examples. Search "counterexamples in math" online to see entire books based on this idea.

35. Your family says you don't call home enough times per week. To better understand your own behavior, you decide to model your calls-home-per-week via $X \sim Poisson(\lambda)$. To estimate λ, you pick n random weeks from the last year and look up your number of calls home, getting the random sample $X_1, \ldots, X_n$.

    (a) Find an unbiased estimator of λ and prove it is unbiased.

    (b) Perhaps what really matters is how many total minutes, M, these X calls amount to in a week, not the number of calls. You find a journal article that suggests the relationship $M = 2X + 3X^2$. Find the expected number of weekly minutes as an expression involving λ.

    (c) Find an unbiased estimator of your answer from part (b) based on $X_1, X_2, \ldots, X_n$. Note: You don't have a procedure for doing this. You should dream up a guess and then check to see if it works. If not, adjust it.

Hint List:

- a. If you need to dream up an estimator, you know two ways from earlier in the class to create estimators.
- c. If you had an estimator that was unbiased for the parameter θ, what would you do to it to estimate θ^2?

Solution:

(a) We might try $\widehat{\lambda} = \overline{X}$, which is both the MME and MLE. Here, $E[\widehat{\lambda}] = E[\overline{X}] = E[X] = \lambda$. So, our guess works. Use $\boxed{\widehat{\lambda} = \overline{X}}$. Note that other answers could work here but may be non-optimal based on metrics other than bias (e.g., $\widehat{\lambda} = X_1$).

(b) We simply take the expectation of M:

$$E[M] = E[2X+3X^2] = 2E[X]+3E[X^2] = 2\lambda+3\left(Var[X] + E[X]^2\right) = 2\lambda+3(\lambda+\lambda^2) = \boxed{3\lambda^2 + 5\lambda}.$$

(c) Here, we are trying to dream up an expression that uses $X_1, \ldots, X_n$ and has expectation equal to $3\lambda^2 + 5\lambda$. Since $\overline{X}$ was unbiased for λ, we might try the expression $3\left(\overline{X}\right)^2 + 5\overline{X}$. Does this have the correct expectation?

$$\begin{aligned} E[3\left(\overline{X}\right)^2 + 5\overline{X}] &= 3E[\left(\overline{X}\right)^2] + 5E[\overline{X}] = 3\left(Var[\overline{X}] + E[\overline{X}]^2\right) + 5E[\overline{X}] \\ &= 3\left(\frac{Var[X]}{n} + E[X]^2\right) + 5E[X] = 3\left(\frac{\lambda}{n} + \lambda^2\right) + 5\lambda \\ &= 3\lambda^2 + 5\lambda + \frac{3\lambda}{n} \end{aligned}$$

This is a little off, so we could make an adjustment to the previous guess and try: $\boxed{3\left(\overline{X}\right)^2 + 5\overline{X} - \frac{3\overline{X}}{n}}$.

$$E\left[3\left(\overline{X}\right)^2 + 5\overline{X} - \frac{3\overline{X}}{n}\right] = E\left[3\left(\overline{X}\right)^2 + 5\overline{X}\right] - \frac{3}{n}E[\overline{X}] = 3\lambda^2+5\lambda+\frac{3\lambda}{n}-\frac{3\lambda}{n} = 3\lambda^2+5\lambda, \text{ as desired.}$$

Other answers are possible. Can you find one that uses $\overline{X^2}$ instead of $\overline{X}^2$?

Commentary: This problem provides a nice example of what can happen in the real world: You begin by estimating the parameters of some model (here, λ from $Poisson(\lambda)$), and then you start to care about more complex aspects of the model (here, the number of minutes rather than the number of phone calls). Said more generally, you move from estimating θ in $f_X(x;\theta)$ to estimating $g(\theta)$, where g is some function relevant to the situation. In part (c), we have $g(x) = 3x^2 + 5x$ and $\theta = \lambda$. As we discover, moving from θ to $g(\theta)$ can be tough, especially if you want to keep unbiasedness. This is another example of the "layering issue" mentioned in other commentaries: When you add extra layers to what you are doing in the estimation process ($g(\theta)$ instead of θ), it isn't so easy to simply build on your original results. Here, going from $\overline{X}$ to $3\left(\overline{X}\right)^2 + 5\overline{X}$ didn't immediately work. We've seen this trend in other places in the course. For example, when trying to maximize a function of one variable for the MLE, the maximum check is fairly gentle: take the second derivative and ensure it is negative at the supposed location. Things get much harder when you add one more layer and create a two-variable maximization problem. Now we would have to check the determinant of the Hessian matrix (which I let you skip to preserve your sanity)!

36. Let $X \sim Exp(\lambda)$ with λ unknown, and suppose X_1, X_2 is a random sample of size 2. Show that $M = \sqrt{X_1 \cdot X_2}$ is a biased estimator of $\dfrac{1}{\lambda}$ and modify it to create an unbiased estimator. Use the fact that $\Gamma(1/2) = \sqrt{\pi}$ to completely simplify any expressions that arise.

Hint List:

- Recall that the iid assumption gives us $E[X_i \cdot X_j] = E[X_i] \cdot E[X_j]$ if $i \neq j$.
- Recall that $E[g(X)] = \int_{\text{support}} g(x) \cdot f_X(x)\, dx$.

Solution: Note that $E[M] = E[\sqrt{X_1} \cdot \sqrt{X_2}] \overset{\text{ind}}{=} E[\sqrt{X_1}]E[\sqrt{X_2}] \overset{\text{id dist}}{=} E[\sqrt{X}]E[\sqrt{X}] = E[\sqrt{X}]^2$. Here, we use the fact that if X_1 and X_2 are independent, then so are $\sqrt{X_1}$ and $\sqrt{X_2}$. This should feel believable: If the outputs from X_1 and X_2 have no connection to or affect on each other, then sticking a square root on these unrelated values won't suddenly make them related.

Now, $E[\sqrt{X}] = \int_0^\infty \sqrt{x} \cdot \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty x^{(3/2)-1} e^{-\lambda x}\, dx$. Here, the integrand is the kernel of $Gamma(r = 3/2, \lambda)$, so the Kernel Technique helps us continue:

$$E[\sqrt{X}] = \lambda \cdot \frac{\Gamma\left(\frac{3}{2}\right)}{\lambda^{3/2}} = \frac{\frac{1}{2}\Gamma\left(\frac{1}{2}\right)}{\sqrt{\lambda}} = \frac{\sqrt{\pi}}{2\sqrt{\lambda}}$$

So, $E[M] = \left(\frac{\sqrt{\pi}}{2\sqrt{\lambda}}\right)^2 = \frac{\pi}{4\lambda} \neq \frac{1}{\lambda}$. Since our estimator is off by a multiplicative constant, we can easily build an unbiased estimator by adjusting it: $\boxed{M_{\text{new}} = \frac{4}{\pi} M}$.

Commentary: I like this problem for several reasons. First, it requires that you bring together a lot of different ideas: expectations are multiplicative for independent RVs, finding $E[\sqrt{X}]$, the Gamma distribution, properties of the Gamma function, and bias. I feel it is quite rewarding when you are able to cook a dish using so many different ingredients. Second, the problem invites you to be curious:

- How would one dream up M in the first place? Answer: For the exponential distribution, we know $E[X] = \frac{1}{\lambda}$, so $E[X_1 \cdot X_2] = E[X_1]E[X_2] = \frac{1}{\lambda^2}$. Thus, if we stick a square root on things, we'll likely get something about equal to $\sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda}$. Here, we say "about" because $E[\sqrt{X}] \neq \sqrt{E[X]}$, so the expectation can't just be moved around wherever we want in an expression.
- How could we extend M to use n pieces of data? Answer: We might try $\sqrt[n]{X_1 \cdot X_2 \cdots X_n}$ using the nth root. This is known as the geometric mean and is part of a famous trio: the arithmetic, geometric, and harmonic means. Once you've seen this idea, you might be tempted to revisit the method of moments and attempt to adjust it so geometric means make an appearance instead of arithmetic means.
- This problem required knowing the value of $\Gamma\left(\frac{1}{2}\right)$. How is this calculated? Are there other values of $r \in (0, 1)$ for which we know the exact value of $\Gamma(r)$? Answer: This and this help.

37. Suppose that $X \sim Unif(0, 3\theta)$ and we draw a random sample $X_1, \ldots, X_n$ where $n \geq 2$. Find the MME and compute its relative efficiency to $\widehat{\theta}_2 = 2X_1 - \frac{4}{3}X_2$.

Hint List:

- Recall that $Var[X-Y] = Var[X] + Var[Y]$ if X and Y are independent.

Solution: For $X \sim Unif(0, 3\theta)$, $E[X] = \dfrac{0+3\theta}{2}$. Equating moments gives $\overline{X} = \dfrac{3\widehat{\theta}}{2}$, or $\boxed{\widehat{\theta}_{\text{MME}} = \dfrac{2}{3}\overline{X}}$.

$$E[\widehat{\theta}_{\text{MME}}] = \frac{2}{3}E[\,\overline{X}\,] = \frac{2}{3}E[X] = \frac{2}{3}\cdot\frac{3}{2}\theta = \theta$$
$$Var[\widehat{\theta}_{\text{MME}}] = \frac{4}{9}Var[\,\overline{X}\,] = \frac{4}{9}\frac{Var[X]}{n} = \frac{4}{9n}\frac{(3\theta-0)^2}{12} = \frac{\theta^2}{3n}$$

Turning to $\widehat{\theta}_2 = 2X_1 - \dfrac{4}{3}X_2$, we find

$$E[\widehat{\theta}_2] = 2E[X_1] - \frac{4}{3}E[X_2] = 2\cdot\frac{3\theta}{2} - \frac{4}{3}\cdot\frac{3\theta}{2} = \theta$$
$$\begin{aligned} Var[\widehat{\theta}_2] &= Var\left[2X_1 - \frac{4}{3}X_2\right] \overset{\text{ind}}{=} Var[2X_1] + Var\left[\frac{4}{3}X_2\right] \\ &= 4Var[X_1] + \frac{16}{9}Var[X_2] \overset{\text{id dist}}{=} \frac{52}{9}Var[X] \\ &= \frac{52}{9}\frac{(3\theta-0)^2}{12} = \frac{13\theta^2}{3} \end{aligned}$$

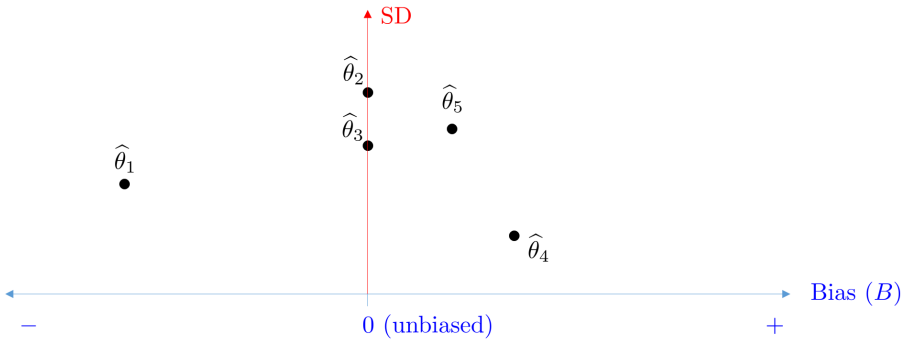
Since both $\widehat{\theta}_{\text{MME}}$ and $\widehat{\theta}_2$ are unbiased, the relative efficiency explores variance (rather than MSE). So,

$$\text{eff}\left(\widehat{\theta}_{\text{MME}}, \widehat{\theta}_2\right) = \frac{\frac{13\theta^2}{3}}{\frac{\theta^2}{3n}} = \boxed{13n}$$

Commentary: This is a "grunt work" problem. The goal is to establish base-level proficiency with a concept (relative efficiency). Essentially, if you can do a problem like this, you have reached the "C student" designation. To become a "B student", you must solve harder versions of the same problem type (e.g., see Problem 39). "A students" can do the hardest problem types, but they can also tackle new challenges that don't look similar to homework problems. This is because they are focused on *understanding the ideas* of a course, not simply memorizing a list of problem types in the hope that similar problems will appear on exams. Be honest with yourself: Do you achieve success in school by cataloging problem types or deeply learning concepts?

My personal view of learning mathematical subjects is this: Some of what we do is internalizing standard problem types that arise frequently. For example, I need to solve a lot of quadratic equations, so I've memorized how to do this (via a mix of factoring and the Quadratic Formula). But, there is so much more to learning math. For example, I memorize a lot of definitions (e.g., Gamma pdf, Poisson pmf) and results (e.g., $3+4=7$, the expectations/variances of distributions) to help expedite my mathematical work. I also spend a lot of time trying to develop visual understandings of the ideas in front of me (e.g., the next problem and the animations that appear in class). There are many more things on my list, but the key point here is this: Understanding isn't a single thing; it is a constellation of overlapping skill sets that allow you to successfully make sense of the world. Demand this higher bar of yourself and you'll see your grades rise as well.

38. In class, I showed the below picture. Here, I have changed the vertical axis from variance to SD. In this new picture, how can we visualize the MSE? How does this way of seeing the MSE help us decide which of two (possibly biased) estimators is more efficient?



Hint List:

- Write down the formula for the MSE of an estimator. What operations to you see? Where have you seen similar operations in Geometry?
- In Geometry, we care about location, distance, angles, area, volume, etc. These are the things you should be trying to "see" as the MSE.

Solution: When we place an estimator $\widehat{\theta}$ in the picture, its coordinates are $(\text{Bias}(\widehat{\theta}), \text{SD}(\widehat{\theta}))$ because of the labels on the axes. Thus, the distance from this point to the origin is $\sqrt{\text{Bias}(\widehat{\theta})^2 + \text{SD}(\widehat{\theta})^2}$ via the Pythagorean Theorem or distance formula. But this distance equals $\sqrt{\text{Bias}(\widehat{\theta})^2 + \text{Var}(\widehat{\theta})} = \sqrt{\text{MSE}(\widehat{\theta})}$.

Thus, we have

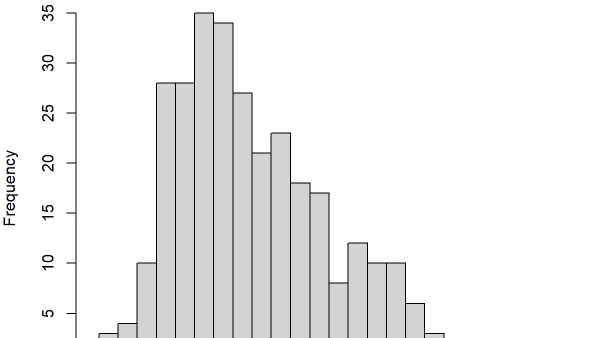
$$\begin{aligned} \widehat{\theta}_1 \text{ is more efficient than } \widehat{\theta}_2 &\iff \text{MSE}(\widehat{\theta}_1) < \text{MSE}(\widehat{\theta}_2) \\ &\iff \sqrt{\text{MSE}(\widehat{\theta}_1)} < \sqrt{\text{MSE}(\widehat{\theta}_2)} \\ &\iff \widehat{\theta}_1 \text{ is closer to the origin than } \widehat{\theta}_2 \end{aligned}$$

So, proximity to the origin shows relative efficiency in the visualization.

Commentary: The human visual complex is incredibly powerful. Part of this has to do with information density. Written words on the page are low density. Images are medium density. Videos are high density. Look at file sizes for text, images, and video on your computer to be convinced on this.

Visual media also have the power to reveal trends that may not be recognized in non-visual formats. As an example, below you'll see the start of a long list of data. Looking at the values on the left gives little sense of what distribution they may have come from. But, if we make a quick visualization, your brain immediately thinks of a Gamma distribution because of the support and the overall shape.

```
  [1] 10.0042069 14.3025447  9.0042736 13.8198548 16.4042301  4.3499295  1.4318095
  [8]  6.5982424 16.8672522  9.5204032  6.9001577  9.6576640 10.5473865  3.6002232
 [15]  7.9177223  5.0161286 16.1639828 10.3218169  9.9465756 15.6986632  5.9182905
 [22]  5.1996047  8.1089914  9.2548987 13.2090332  8.0908371  8.2583178  3.5950980
 [29]  6.2956216 11.8168731  3.0224331  5.3032616  3.6386451 16.8837615  7.1286142
 [36] 15.2738988 10.7113781  6.4086606  6.0832376 20.7763981 11.6237489  4.7737094
 [43]  6.1066132  7.7292781 10.3554208 14.8471314  8.1319982  3.9409208  5.8105584
 [50]  3.5917512  6.1240842  9.2090178 17.7704722  7.6864694  3.4785455  6.0316085
 [57]  5.8925747  4.3396475  6.7771480 11.3687666 12.8782546 11.4168729  2.2633986
 [64]  5.2680378 11.6182535 16.0976879 16.0752673  3.3273192  4.2865978  4.6897770
 [71] 14.1246844  7.0467837  2.8293846 13.7773705  1.4666472  4.6917989  5.0099000
 [78] 10.6389052  9.1181111  6.9680489  9.0384145  3.1434285  3.5917028 13.5060233
 [85]  7.6811364  5.9752389 10.5965239  2.5470668  1.4164814  9.2499674  7.5726364
 [92]  9.2216526  5.6470182  4.4662712 13.6796509  8.6298388  6.9063605  7.9145242
 [99]  8.6314072  8.3989999 14.6554819 10.9376493 13.8541704 10.3406588 14.9609966
[106] 14.6392999 11.8370099 11.3070723  7.5924681  3.9874888  2.0501504  5.3033125
[113]  6.9688549  5.9573957 11.9478749  5.7265690  8.6525903  3.2607203 10.8378498
[120]  6.6885425  3.6814075 11.4543260  4.5585651  8.1458029  8.6322421  5.4566545
[127]  4.3966066  5.0358559  5.6537996 20.6156815  9.9976305  5.2906833 12.9430187
[134]  3.0028435  7.2776897  5.9121127 14.9529371  3.5750790  6.6204017  5.7939032
[141]  3.5184320  0.4920073  6.1574841  2.8066964  5.4323246 10.3842436  8.9256948
[148]  4.7274486  7.2238175 11.8104432  2.5740494 11.0532817  3.8030318  0.5789092
[155] 13.8583232 12.1349948  6.4910835  9.6605402  6.9172941  5.5402940  8.1416480
[162]  2.7096979  7.4350946  4.8086142  9.6635684 13.0359410 13.4883074  6.7755060
[169]  9.2394204  8.1697182  4.2799062  7.3818119 10.7197500 10.5935259 15.9801194
[176]  7.7965496 12.0868395 14.0898608  6.0949052  4.6893931  6.0320207 10.6101692
[183]  5.8584025  9.4443108  2.0430136  9.4309035  6.5586735 14.6829526 11.3036647
```



The goal of this problem was to help you see the power of visualization in a setting as bland as the MSE. When I am working on hard problems of my own, the first thing I *always* do is try to visualize what is going on. In some cases, like the MSE above, the setting is not inherently visual, and so we must work hard to find a visual interpretation. Some content creators, like 3Blue1Brown, are masters of leveraging the power of visualization.

39. Let X be a continuous random variable with $E[X] = \mu \neq 0$ and $Var[X] = \sigma^2 < \infty$. Suppose we try to estimate μ from a random sample $X_1, \ldots, X_n$ $(n \geq 3)$ using these two estimators:

$$\widehat{\mu}_1 = \overline{X}$$
$$\widehat{\mu}_2 = 2X_1 + aX_2 + bX_3$$

For what a and b are both estimators unbiased *and* the relative efficiency of $\widehat{\mu}_1$ to $\widehat{\mu}_2$ is $45n$?

Hint List:

- We have two demands to meet: (1) both estimators are unbiased, and (2) the relative efficiency is $45n$. Each can give us an equation involving a and b.

Solution: Note that $\widehat{\mu}_1$ is unbiased since $E[\widehat{\mu}_1] = E[\overline{X}] = E[X] = \mu$. If $\widehat{\mu}_2$ is unbiased, we need $\mu = E[\widehat{\mu}_2] = 2E[X_1] + aE[X_2] + bE[X_3] = 2\mu + a\mu + b\mu$. Since $\mu \neq 0$, this simplifies to $-1 = a + b$. This is the first of two equations we'll need.

Now, $Var[\widehat{\mu}_1] = Var[\overline{X}] = \dfrac{Var[X]}{n} = \dfrac{\sigma^2}{n}$. Also, $Var[\widehat{\mu}_2] \overset{\text{ind}}{=} 4Var[X_1] + a^2Var[X_2] + b^2Var[X_3] \overset{\text{id dist}}{=} (4 + a^2 + b^2)\sigma^2$. We are told that $45n = \text{eff}(\widehat{\mu}_1, \widehat{\mu}_2) = \dfrac{(4 + a^2 + b^2)\sigma^2}{\sigma^2/n} \implies a^2 + b^2 = 41$. This is our second equation.

Rearranging the first equation to $a = -1 - b$ and substituting this into the second equation gives $(-1 - b)^2 + b^2 = 41$. This simplifies to $b^2 + b - 20 = 0$, or $(b - 4)(b + 5) = 0 \implies b = 4$ or $b = -5$. If we take each of these values to the first equation, we can find a.

When $b = 4$, $a = -1 - b = -1 - 4 = -5$.
When $b = -5$, $a = -1 - b = -1 + 5 = 4$.

Thus, we have two answers: $\boxed{(a, b) = (-5, 4) \text{ or } (a, b) = (4, -5)}$.

Commentary: In contrast to the previous problem, this exercise is all symbol pushing. Part of doing mathematics is symbol pushing, so I make sure you get enough exposure to this. At the same time, I want other interesting things to arise as those symbols are being pushed. Here, we get a system of two equations and two unknowns, but they aren't both linear equations. For most students, it's been a long time since you've seen such a system, and I wanted you to refresh your skills in solving these. Interestingly, why did we get a non-linear system here? Pause and look back before reading ahead. In this problem, when finding $Var[\widehat{\mu}_2]$, we were faced with $Var[2X_1 + aX_2 + bX_3]$ which becomes $Var[2X_1] + Var[aX_2] + Var[bX_3]$ due to independence. Then, recalling the variance rule $Var[aX] = a^2Var[X]$, we got the nasty-looking expression $4Var[X_1] + a^2Var[X_2] + b^2Var[X_3]$. I skipped these steps in the above solution so you'd have to revisit them when looking back just now. So, rules that introduce non-linearity (like the $Var[aX]$ rule) have the ability to create systems of non-linear equations.

Another interesting facet of this problem is that we get two solutions. Students are often worried about this and come to office hours wanting me to calm their nerves. The next time you have this feeling, try to *calm your own nerves*. Here are two ways to do so in this problem:

- First, our system of equations is a line ($-1 = a + b$) and a circle ($a^2 + b^2 = 41$). These can have two intersections.
- Second, notice that our two answers give rise to $\widehat{\mu}_2 = 2X_1 - 5X_2 + 4X_3$ and $\widehat{\mu}_2 = 2X_1 + 4X_2 - 5X_3$. It make senses that if one of these estimators has the desired properties, then so will the other. The only difference is the coefficient switch on X_2 and X_3. But note that X_2 and X_3 are themselves essentially interchangeable. It doesn't really matter what you call the second and third data observations, and in advance of working the problem, we know we could switch their roles if we wanted another solution.

40. Find the Fisher Information (FI) and the Cramer-Rao lower bound (CRLB) for the variance of an unbiased estimator of θ given a random sample $X_1, \ldots, X_n$ from the density

$$f(x;\theta) = \frac{x^3}{6\theta^4} e^{-x/\theta} \text{ where } x > 0 \text{ and } \theta > 0$$

Hint List:

- The Kernel Technique can help with $E[X]$.

Solution:

Here, $\ln f(X;\theta) = 3\ln X - \ln 6 - 4\ln\theta - \dfrac{X}{\theta}$. Note that we are using capital X here, not lowercase x. This is one of the strange aspects of finding the FI and CRLB. It is important to do this because, in a moment, we will take an expectation, and this requires the presence of a random variable, which X is.

Next, $\dfrac{\partial \ln f}{\partial\theta} = \dfrac{-4}{\theta} + \dfrac{X}{\theta^2} \implies \dfrac{\partial^2 \ln f}{\partial\theta^2} = \dfrac{4}{\theta^2} - \dfrac{2X}{\theta^3}$.

Now, $E[X] = \displaystyle\int_0^\infty x \cdot \frac{x^3}{6\theta^4} e^{-x/\theta}\,dx = \frac{1}{6\theta^4}\int_0^\infty x^{5-1} e^{-\frac{1}{\theta}x}\,dx \overset{\text{KT}}{=} \frac{1}{6\theta^4} \cdot \frac{\Gamma(5)}{\left(\frac{1}{\theta}\right)^5} = \frac{4!}{6}\theta = 4\theta$, where we have used the Kernel Technique at KT with $Gamma\left(r = 5, \lambda = \dfrac{1}{\theta}\right)$.

Thus, $I(\theta) = -E\left[\dfrac{\partial^2 \ln f}{\partial\theta^2}\right] = \dfrac{-4}{\theta^2} + \dfrac{2}{\theta^3}E[X] = \dfrac{-4}{\theta^2} + \dfrac{2}{\theta^3}4\theta = \boxed{\dfrac{4}{\theta^2}}$. So, the CRLB $= \left[n \cdot \dfrac{4}{\theta^2}\right]^{-1} = \boxed{\dfrac{\theta^2}{4n}}$.

Commentary:

This is your basic "C student" practice problem with some KT flare to keep the "A student" from being too bored. One thing you might ask yourself after finishing the problem is: How many supports did I need to finish this problem?. For example, you might have looked back on:

- The definition of FI
- The definition of the CRLB
- A worked example of the CRLB
- The pdf of the Gamma distribution
- Facts about evaluating $\Gamma(r)$ when r is an integer

It is tempting to think that if you get a homework problem correct that you understand what is going on, or that you could stand alone (say, on an exam) and perform the same skill. If you used many of the above supports, then you're not ready to stand alone in mathematical battle. When students feel frustrated by how differently they perform on homework and exams, it might be because they aren't honest with themselves about the supports they used when doing the homework. In fact, you may have used many other supports in the making of your solution: instructors, TAs, fellow students, the calculator, AI, internet cheating sites, the textbook, large amounts of time, etc. If you want to perform better on exams, try to better align *the supports you use when doing homework* with *those allowed on exams.*

41. Find the Fisher Information and the Cramer-Rao lower bound for the variance of an unbiased estimator of θ given a random sample $X_1, \ldots, X_n$ from the density

$$f(x;\theta) = \frac{1}{\pi \cdot [1 + (x-\theta)^2]} \text{ where } -\infty < x < \infty \text{ and } -\infty < \theta < \infty$$

You should use WolframAlpha.com to evaluate the complicated integral that will arise.

Hint List:

- Don't forget that $E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx$, even if g is really complicated.
- Before you head to Wolfram, let $w = x - \theta$ to reduce the number of variables.

Solution:

Here, $\ln f(X;\theta) = -\ln \pi - \ln[1+(X-\theta)^2]$. Thus, $\dfrac{\partial \ln f}{\partial \theta} = \dfrac{-1}{1+(X-\theta)^2} \cdot 2(X-\theta)(-1) = \dfrac{2(X-\theta)}{1+(X-\theta)^2}$.

Since we get to use Wolfram, we can skip finding $\dfrac{\partial^2 \ln f}{\partial \theta^2}$ and just make a huge mess:

$$I(\theta) = E\left[\left(\frac{\partial \ln f}{\partial \theta}\right)^2\right] = E\left[\frac{4(X-\theta)^2}{(1+(X-\theta)^2)^2}\right] = \int_{-\infty}^{\infty} \frac{4(x-\theta)^2}{(1+(x-\theta)^2)^2} \cdot \frac{1}{\pi(1+(x-\theta)^2)}\, dx$$

We have used $E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx$ for the last step. To make this easier to enter into Wolfram, set $w = x - \theta$. This forces $dw = dx$ and the bounds remain $-\infty$ and ∞. Thus,

$$I(\theta) = \int_{-\infty}^{\infty} \frac{4w^2}{\pi(1+w^2)^3}\, dw = \boxed{\frac{1}{2}}, \text{ from Wolfram Alpha.}$$

So, $\boxed{\text{CRLB} = \dfrac{2}{n}}$, a surprisingly simple result, and one that doesn't depend on θ.

Commentary:

The distribution that appears here is known as the Cauchy distribution. It is an important example in higher level classes because both its expectation and variance do not exist (!). I have included it here because of the integral that arises and the surprising value of that integral. Were you stunned when that nightmare evaluated to $\dfrac{1}{2}$? Did something in your soul say: I want to know how to get that surprisingly simple answer? If you want to give it a go by hand, let $w = \tan \alpha$ and see what happens.

One of the things that bothers me most about society is how incurious people are. People seem totally satisfied with understanding little about how the world works. How does a computer function? What about a cell phone? What about AI? In the modern age, it is possible to be entertained at every moment so that you need not ever look critically at the world or wonder about anything. This problem is *begging you* to be curious by giving a simple answer to a horrible integral. How could that be?

On top of this, you may have noticed that the FI (and, hence, CRLB) doesn't depend on the parameter θ. Why is that the case? Well, if you sit down and play with the Cauchy pdf in Desmos, you'll find that changing θ just shifts the pdf left or right. The actual shape of the pdf is unaffected, and so you don't get any more or less information out of a datum for different θ values. Can you think of another distribution where you found the same thing? Typically, the FI does depend on the parameter. Compare what we saw here with how the λ in $Exp(\lambda)$ and the p in $Binom(n,p)$ reshape their distributions as λ and p change.

42. Let $X_1, \ldots, X_n$ be iid based on $f(x;\theta) = \dfrac{2x}{\theta} e^{-x^2/\theta}$ where $x > 0$ and $\theta > 0$.

(a) Show that $E[X^k] = \theta^{k/2} \cdot \Gamma\left(\frac{k}{2} + 1\right)$.

When you work parts (b) and (c), you'll see how useful it is to know $E[X], E[X^2], E[X^3], E[X^4], \ldots$. These expressions are just $k = 1, 2, 3, 4, \ldots$ in this formula. This problem is having you derive a formula for the "moments of X", which can arise when calculating MME, bias, variance, FI, etc.

(b) In Problem 9, we found $\widehat{\theta}_{\text{MLE}} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} X_i^2$. Show this estimator is efficient.

(c) Show that $\widehat{\theta}_{\text{MLE}}$ is MSE-consistent.

Hint List:

- a. Do not use induction. Instead, just directly evaluate $E[X^k]$ by recalling that

$$E[g(X)] = \int_{\text{support}} g(x) f_X(x)\, dx \text{ where } g(x) \text{ is } x^k$$

Solution:

(a) We find $E[X^k] = \displaystyle\int_0^\infty x^k \cdot \frac{2x}{\theta} e^{-x^2/\theta}\, dx$. This looks related to a Gamma distribution or Gamma function, inspiring the substitution $u = \dfrac{x^2}{\theta}$. We have $du = \dfrac{2x}{\theta}\, dx$ and u bounds of $u = \dfrac{0^2}{\theta} = 0$ (lower) and ∞ (upper). To convert x^k, notice that $u\theta = x^2$, so $x^k = (u\theta)^{k/2}$. Thus,

$$E[X^k] = \int_0^\infty (u\theta)^{k/2} e^{-u}\, du = \theta^{k/2} \int_0^\infty u^{(k/2+1)-1} e^{-u}\, du = \theta^{k/2} \cdot \Gamma\left(\frac{k}{2}+1\right)$$

(b) An estimator is efficient if it is unbiased and its variance equals the CRLB. We use part (a) for particular values of k in the below steps. To show unbiasedness, we find

$$E[\widehat{\theta}_{\text{MLE}}] = E[\overline{X^2}] \overset{\text{HW\#2}}{=} E[X^2] \overset{k=2}{=} \theta^{2/2} \cdot \Gamma\left(\frac{2}{2}+1\right) = \theta \cdot 1! = \theta$$

To show the variance equals the CRLB, we must find both of these.

$$Var[\widehat{\theta}_{\text{MLE}}] = Var[\overline{X^2}] \overset{\text{HW\#2}}{=} \frac{Var[X^2]}{n} = \frac{E[(X^2)^2] - E[X^2]^2}{n} \overset{k=4}{=} \frac{\theta^2 \cdot \Gamma(3) - \theta^2}{n} = \frac{\theta^2}{n}$$

For the FI, note that $\ln f(X;\theta) = \ln(2X) - \ln\theta - \dfrac{X^2}{\theta}$. So, $\dfrac{\partial \ln f}{\partial \theta} = \dfrac{-1}{\theta} + \dfrac{X^2}{\theta^2} \implies \dfrac{\partial^2 \ln f}{\partial \theta^2} = \dfrac{1}{\theta^2} - \dfrac{2X^2}{\theta^3}$. Thus, $I(\theta) = -E\left[\dfrac{\partial^2 \ln f}{\partial \theta^2}\right] = \dfrac{-1}{\theta^2} + \dfrac{2E[X^2]}{\theta^3} \overset{k=2}{=} \dfrac{-1}{\theta^2} + \dfrac{2\theta}{\theta^3} = \dfrac{1}{\theta^2}$.

So, the CRLB $= \dfrac{\theta^2}{n} = Var[\widehat{\theta}_{\text{MLE}}]$, from above, and thus, $\boxed{\widehat{\theta}_{\text{MLE}} \text{ is efficient}}$.

(c) For MSE-consistency, we must show $\text{MSE}(\widehat{\theta}_{\text{MLE}}) \to 0$ as $n \to \infty$. Here,

$$\text{MSE}(\widehat{\theta}_{\text{MLE}}) = Var(\widehat{\theta}_{\text{MLE}}) + [\text{Bias}(\widehat{\theta}_{\text{MLE}})]^2 \overset{\text{(b)}}{=} \frac{\theta^2}{n} + 0^2 \to 0 \text{ as } n \to \infty$$

Commentary: Part (a) of this problem has us finding the kth moments of the RV X. This idea, which I call the "Moments Technique", might seem time-consuming (or overkill) at first but can ultimately save a lot of time in the long run. If you had to actually evaluate both $E[X^2]$ and $E[X^4]$ in part (b) without using part (a), it would have taken longer. After years of experience, I now have this rule for myself: If I can look at a problem and see that I'll need two or more moments along the way (e.g., both $E[X]$ and $E[X^3]$), I usually just start by finding $E[X^k]$. Doing so will save time, and if more moments are needed, I have immediate access to them and save even more time! Try using this idea in problems where I don't force you to do so.

43. Let $X \sim Unif(\theta, \theta + 1)$ and suppose $X_1, \ldots, X_n$ is a random sample from X. Show that the estimator $\widehat{\theta} = \min X_i = X_1' = X_{(1)}$ is consistent for θ using the ε-definition of consistency.

Hint List:

- Begin by finding the cdf of X and the pdf of $\widehat{\theta}$.

Solution: Note that $f(x;\theta) = 1$ on the support $[\theta, \theta+1]$ and

$$F(x) = \int_{\theta}^{x} 1\, dt = t\Big|_{\theta}^{x} = x - \theta \text{ if } \theta \leq x \leq \theta + 1$$

Using the order stat pdf with $i = 1$ gives

$$f_{X_{\min}}(x) = \frac{n!}{(1-1)!(n-1)!} F(x)^{1-1} f(x) \left[1 - F(x)\right]^{n-1} = n(1 - x + \theta)^{n-1} \text{ where } \theta \leq x \leq \theta + 1$$

We must show $\lim_{n\to\infty} P(|\widehat{\theta} - \theta| < \varepsilon) = 1$. Recall that $P(|\widehat{\theta} - \theta| < \varepsilon) = P(\theta - \varepsilon < \widehat{\theta} < \theta + \varepsilon)$.

Now, if $\varepsilon \geq 1$, then this probability calculation includes the entire zone $[\theta, \theta+1]$ (the support of $X_{\min}$) and hence equals 1. Taking the limit as $n \to \infty$ still gives us 1.

For $0 < \varepsilon < 1$, the probability becomes

$$\int_{\theta}^{\theta+\varepsilon} n(1 - x + \theta)^{n-1}\, dx = -(1 - x + \theta)^{n}\Big|_{\theta}^{\theta+\varepsilon} = -(1 - \varepsilon)^{n} + 1 \to 1 \text{ as } n \to \infty$$

Note that our integral starts at θ instead of $\theta - \varepsilon$ because of the support of $X_{\min}$.

Commentary:

Students often ask me how to improve their exam performance. One way to do this is to explicitly focus on solving speed. When you are first solving this problem, you can essentially take as long as you need; your only bound is completing the assignment before the due date. If this problem were to appear on an exam (and it did, in the past), then you'd need to be able to solve it in about 5 minutes. For a 50-minute exam at this level, you should aim for: 4-5 minutes for easier problems, 7-8 minutes for medium problems, and 10-11 minutes for the hardest problems. One thing you can try is this: categorize each homework problems as easy/medium/hard and then rework these problems as you prepare for the exams. Use your phone to time yourself and see how close you come to the above time recommendations. Students rarely do this type of speed work and yet, it is extremely common in other settings: sports training (e.g., track and field, swimming), practicing music with a metronome, etc. Get serious about speed to up your exam game.

44. Let $Y_1, \ldots, Y_n$ be a random sample from Y with pdf

$$f_Y(y;\theta) = \frac{3(\theta - y)^2}{\theta^3} \text{ where } 0 < y < \theta$$

Let $\widehat{\theta} = \max Y_i = Y_n' = Y_{(n)}$. Show $\widehat{\theta}$ is consistent using the ε-definition of consistency.

Hint List:

- Don't be tempted to expand expressions like $(\theta - t)^2$ or $(\theta - y)^3$.
- For the tough integral, use a big u-substitution.
- Separating the cases of $\varepsilon \geq \theta$ and $0 < \varepsilon < \theta$ is useful.

Solution:

First, $F(y) = \displaystyle\int_{-\infty}^{y} f(t)\,dt = \int_0^y \frac{3(\theta - t)^2}{\theta^3}\,dt = \frac{-(\theta - t)^3}{\theta^3}\Big|_0^y = 1 - \frac{(\theta - y)^3}{\theta^3}$.

Thus, $f_{\widehat{\theta}}(y) = \dfrac{n!}{(n-1)!(n-n)!} F^{n-1} f (1 - F)^{n-n} = n \cdot \left[1 - \dfrac{(\theta - y)^3}{\theta^3}\right]^{n-1} \cdot \dfrac{3(\theta - y)^2}{\theta^3}$, where $0 < y < \theta$.

Now, let $\varepsilon > 0$. We must show $\lim\limits_{n\to\infty} P(|\widehat{\theta} - \theta| < \varepsilon) = 1$.

$$\begin{aligned} \text{So, } P(|\widehat{\theta} - \theta| < \varepsilon) &= P(-\varepsilon < \widehat{\theta} - \theta < \varepsilon) = P(\theta - \varepsilon < \widehat{\theta} < \theta + \varepsilon) = P(\theta - \epsilon < \widehat{\theta} < \theta) \quad \text{(because } \widehat{\theta} < \theta\text{)} \\ &= \int_{\theta-\varepsilon}^{\theta} n \cdot \left[1 - \frac{(\theta - y)^3}{\theta^3}\right]^{n-1} \cdot \frac{3(\theta - y)^2}{\theta^3}\,dy \end{aligned}$$

Let $u = 1 - \dfrac{(\theta - y)^3}{\theta^3}$, so $du = \dfrac{-3(\theta - y)^2(-1)}{\theta^3}\,dy$. The new lower bound is $u = 1 - \dfrac{\varepsilon^3}{\theta^3}$; the new upper bound is $u = 1$. Thus, our integral becomes

$$\int_{1-\varepsilon^3/\theta^3}^{1} n u^{n-1}\,du = u^n\Big|_{1-\varepsilon^3/\theta^3}^{1} = 1 - \left(1 - \frac{\varepsilon^3}{\theta^3}\right)^n$$

Notice that if $\varepsilon \geq \theta$, then our original probability expression simplifies: $P(\theta - \varepsilon < \widehat{\theta} < \theta) = P(0 < \widehat{\theta} < \theta)$ because $0 < \widehat{\theta}$. This probability equals 1 since it is simply asking for the probability $\widehat{\theta}$ is in its support.

If, however, $0 < \varepsilon < \theta$, then $0 < 1 - \dfrac{\varepsilon^3}{\theta^3} < 1$, and so $\left(1 - \dfrac{\varepsilon^3}{\theta^3}\right)^n \to 0$ as $n \to \infty$.

So, $P(|\widehat{\theta} - \theta| < \varepsilon) \to 1 - 0 = 1$ for all $\varepsilon > 0$, as required.

Commentary:

It takes a lot of practice to write up detail-oriented arguments like seen above. When you sit back and look at all the different letters and ideas, you should feel proud of the sophistication you've spent years building. One interesting thing about all those letters $(F, f, y, t, \theta, n, \widehat{\theta}, \varepsilon, u)$ is that they behave in *different* ways. So, while you might be tempted to call them all "variables", they don't all act the same. Here's an easier example of what I mean. Below are five statements that use the letter c. For each, decide what c stands for:

$$\int x\,dx = \frac{1}{2}x^2 + C, \quad E = mc^2, \quad C + O_2 \to CO_2, \quad P = R - C, \quad \lim_{c\to\infty} \frac{1}{c}$$

Now, revisit each statement and decide what values c can take on. Try this before reading on.

This example, which uses the *same* letter, shows that a single letter might act in really different ways depending on the context of the statement. For example, the c in $E = mc^2$ (the mass-energy

equivalence from Physics) stands for a **constant**: $c \approx 3 \cdot 10^8$ m/s. The C in the third statement stands for carbon in a chemistry class, so it is acting as a **label**. The C in the first statement is suggesting a constant, but not a specific one as in $E = mc^2$. Instead, it can be anything, and in doing so, we get a family of functions as the general antiderivative. Educational researchers call this a **parameter**, and you've seen this idea for c when defining a quadratic equation: $ax^2 + bx + c = 0$. In the fourth example, C stands for cost (P for profit and R for revenue, from Economics), and here it might be any value depending on the situation. Because of this, it is known as a **varying quantity**. In the final statement, c is playing a role in a more general process (the limit to infinity), so I call this a **process variable**.

These boldfaced terms describe the *ontological status* of variables. Ontology is the study of the nature of being. Said simply, it answers the question: What is that thing and how is it behaving? You may have never given thought to the ontology of variables before (I didn't until my PhD program), but the nature of their existence is always there in the background whether you notice it or not. As a simpler example, compare the x in $7 = 2x + 1$ and $y = 2x + 1$. In the first, we recognize that x is simply hiding the value 3, so it's not actually varying in any way. In the second, we recognize that x could be any value (it can vary!), and as it does so, the value of y is forced, and we get a line as the solution set. So, the x in each $2x + 1$ has a different ontology.

If we return to the solution above, you might try seeing which letters fit into which boldfaced ontologies. For example, F is just the name we give to the cdf, so it is behaving like a **label**. This is its nature of being. How about the other letters? Do any letters not fit into the boldfaced categories above? This discussion of variable ontologies has a larger lesson: It is possible to go through life without looking carefully at what is all around you. Socrates warned against this with his (supposed) quote: "The unexamined life is not worth living". Are you carefully examining what you're doing, or is your goal simply to get something done so you can then be entertained by screens? There is so much depth to uncover if you're willing to start digging.

45. Let $Y_1, \ldots, Y_n$ be iid based on Y with pdf

$$f_Y(y;\beta) = \frac{1}{\beta} e^{-(y-3)/\beta} \text{ where } y > 3, \beta > 0$$

(a) Find $\widehat{\beta}_{\text{MME}}$ for β using first moments.

(b) Show $\widehat{\beta}_{\text{MME}}$ is MSE-consistent, and hence, consistent.

Note that the last few problems are training two ways to show consistency: via the ε-definition, and through MSE-consistency. Be skilled at both.

Hint List:

- a. Try a u-substitution. Then, split the integral into two pieces.
- a and b. If you want to do both parts with as little work as possible, see if you can re-express Y as a constant plus a well-known random variable.

Solution:

(a) $E[Y] = \int_3^\infty y \cdot \frac{1}{\beta} e^{-(y-3)/\beta}\, dy$. Let $u = y - 3$, so $du = dy$ and the new bounds are $u = 3 - 3 = 0$ (lower) and ∞ (upper). After substitution, we get

$$\begin{aligned} E[Y] &= \int_0^\infty (u+3)\frac{1}{\beta} e^{-u/\beta}\, du = \int_0^\infty u\frac{1}{\beta} e^{-u/\beta}\, du + 3\int_0^\infty \frac{1}{\beta} e^{-u/\beta}\, du \\ &= E\left[Exp\left(\frac{1}{\beta}\right)\right] + 3\left(\text{Area under the } Exp\left(\frac{1}{\beta}\right) \text{ pdf}\right) = \frac{1}{1/\beta} + 3 \cdot 1 = \beta + 3 \end{aligned}$$

Equating moments gives $\widehat{\beta} + 3 = \overline{Y}$, or $\boxed{\widehat{\beta}_{\text{MME}} = \overline{Y} - 3}$.

A more clever way to do this part is to notice that the given pdf looks a lot like the Exponential pdf where $\lambda = \frac{1}{\beta}$. The support of Y is $(3, \infty)$ instead of the usual $(0, \infty)$, so maybe this is a "shifted Exponential distribution". Indeed, $Y = 3 + Exp\left(\lambda = \frac{1}{\beta}\right)$. So, $E[Y] = 3 + E[Exp(1/\beta)] = 3 + \frac{1}{1/\beta} = 3 + \beta$, as above, but with less work.

(b) First, $E[\widehat{\beta}] = E[\overline{Y} - 3] = E[Y] - 3 \overset{\text{(a)}}{=} (\beta + 3) - 3 = \beta$.

Next, $Var[\widehat{\beta}] = Var[\overline{Y} - 3] = Var[\overline{Y}] = \frac{Var[Y]}{n}$. This variance can be found by first doing $E[Y^2]$ as in part (a). Or, we could use the clever observation above:

$$Var[Y] = Var[3 + Exp(1/\beta)] = Var[Exp(1/\beta)] = \frac{1}{(1/\beta)^2} = \beta^2$$

So, $\text{MSE}[\widehat{\beta}] = Var[\widehat{\beta}] + \text{Bias}[\widehat{\beta}]^2 = \frac{\beta^2}{n} + (\beta - \beta)^2 = \frac{\beta^2}{n} \to 0 \text{ as } n \to \infty$. Thus, $\boxed{\widehat{\beta} \text{ is MSE-consistent}}$, and hence, consistent.

Commentary:

They say that the best coders spend 80% of their time thinking/organizing/planning and 20% of their time actually writing code. Those who teach introductory programming classes will add to this claim: Beginner programmers *switch* these percentages. While these percentages may not be completely accurate, I find there truth in both claims. When I taught computer science at the high school level, students always wanted to start typing as soon as possible. They reasoned that they'd just "figure things out as they went". One semester (in a post-AP Computer Science class), our goal was to write

a program that could simulate millions of hands in blackjack and return the odds of various betting strategies. After explaining this on the first day, I told the students to "get to work". Almost everyone in the room just opened their computers and started typing code. After a week of getting nowhere, I stopped the class, had them delete all their code, and explained the 80/20 rule. We then spent weeks simply talking about the problem and how we would organize our work. Along the way, they realized several things that seem obvious in retrospect:

- They needed to completely understand the rules of the game before they could code it. (Blackjack has an endless collection of rule variants used by different casinos to change the player's odds.)
- They would have to work as a team to complete the project. It was too large to do individually.
- They should assign different roles within the team. Some people would only test small bits of code, others would write the code, others would manage the order of the tasks, etc.
- Some people were better at some jobs than others. Eventually, they leaned into people's talents to give the project the highest chance of being successful.

The key point here is that putting in time thinking about the problem *before* beginning work on it can be very useful (or absolutely necessary, as in the above story). I like this problem because the pdf looks so similar to an Exponential distribution, that it's screaming at you to pause and re-frame things via the clever solution pathway in the solution to part (a). If you have the courage to do this, you can skip several integrals (which are error-prone and exhausting). You might also be tempted to make a conjecture based on what you're seeing with the pdfs:

If the RV X has pdf $f(x;\theta)$ and support $[a,b]$, then $X+c$ has pdf $f(x-c;\theta)$ and support $[a+c,b+c]$.

Notice how much can come out of time spent analyzing/planning instead of mindlessly plowing ahead with algebra.

46. Let $X_1, \ldots, X_n$ be a random sample from the discrete RV X with pmf

$$f(x;\alpha) = \frac{\exp(\alpha x - e^{\alpha})}{x!} \text{ where } x = 0, 1, 2, \ldots$$

Find the MLE for α and use it to create a formula for an approximate 82% MLE CI for α. (Recall that $\exp(y)$ is an alternative notation for e^y.)

Hint List:

- This problem was designed to be straightforward. So, if you are struggling, it means that some foundational procedure is not well-established. Most likely, you need to review (1) finding an MLE, (2) finding the FI, or (3) using the MLE CI formula.

Solution:

$$L(\alpha) = \prod_{i=1}^{n} \frac{\exp\left(\alpha x_i - e^{\alpha}\right)}{x_i!} \implies \ell(\alpha) = \sum_{i=1}^{n} \left[\alpha x_i - e^{\alpha} - \ln(x_i!)\right] = \alpha \sum x_i - n e^{\alpha} + C$$

$$\ell'(\alpha) = \sum x_i - n e^{\alpha}$$

$$n e^{\widehat{\alpha}} = \sum x_i$$

$$e^{\widehat{\alpha}} = \overline{X} \implies \widehat{\alpha}_{\text{MLE}} = \ln \overline{X}$$

$$\ell''(\alpha) = -n e^{\alpha} < 0 \text{ for all } \alpha$$

Next, we find the FI:

$$\ln f(X;\alpha) = \alpha X - e^{\alpha} - \ln(X!) \implies \frac{\partial \ln f}{\partial \alpha} = X - e^{\alpha} \implies \frac{\partial^2 \ln f}{\partial \alpha^2} = -e^{\alpha}$$

So, $I(\alpha) = -E\left[-e^{\alpha}\right] = e^{\alpha}$. For 82% confidence, we have $\alpha = 0.18$, so $z_{\alpha/2} = \text{qnorm}(0.09, \text{lower} = \text{F}) \approx 1.340755$. Thus, the approximate MLE CI has the form:

$$\widehat{\alpha}_{\text{MLE}} \pm z_{\alpha/2} \frac{1}{\sqrt{n I(\widehat{\alpha}_{\text{MLE}})}} = \ln \overline{X} \pm 1.340755 \frac{1}{\sqrt{n e^{\ln \overline{X}}}} = \boxed{\ln \overline{X} \pm 1.340755 \frac{1}{\sqrt{n \overline{X}}}}$$

Commentary: I chose this problem because every part of the MLE CI is unproblematic. The MLE is as easy as possible: the first derivative comes out nicely, and the second derivative is always negative. Then, when finding the FI, the second partial doesn't involve X, so the expectation is easy to calculate. Finally, when doing the MLE CI formula, we even get simplifications ($e^{\ln \overline{X}} = \overline{X}$). Problems like this are the exception, not the rule. It is tempting to think that problems in school should come out nicely, or that the algebra should always "feel good". Remember that this is just an academic narrative you tell yourself because teachers have been engineering this experience for you all your life. A better default view is to expect horrible answers and messy algebra, especially at the Mathematical Statistics level and beyond.

47. Suppose we try to model the test-taking abilities of a given student by the CRV X with pdf

$$f(x;\theta) = (\theta+3)x^{\theta+2} \text{ where } 0 < x < 1 \text{ and } \theta > -3$$

Here, the constant θ is unknown and is determined by the work ethic and background training of the student. Design an approximate 93% MLE CI for θ and use it to build a CI for the data $x_1 = 0.8, x_2 = 0.92, x_3 = 0.81, x_4 = 0.96$ (which represent four random test scores of the student: 80%, 92%, 81%, and 96%).

Hint List:

- This problem was also designed to be straightforward. Do you have the foundational skills: (1) finding an MLE, (2) finding the FI, and (3) using the MLE CI formula?

Solution:

$$L(\theta) = \prod_{i=1}^{n}(\theta+3)x_i^{\theta+2} \implies \ell(\theta) = \sum_{i=1}^{n}[\ln(\theta+3) + (\theta+2)\ln x_i] = n\ln(\theta+3) + (\theta+2)\sum \ln x_i$$

$$\ell'(\theta) = \frac{n}{\theta+3} + \sum \ln x_i$$

$$\frac{-\sum \ln x_i}{n} = \frac{1}{\widehat{\theta}+3}$$

$$\widehat{\theta}+3 = \frac{-1}{\overline{\ln X}}$$

$$\widehat{\theta}_{\text{MLE}} = -3 - \frac{1}{\overline{\ln X}}$$

$$\ell''(\theta) = \frac{-n}{(\theta+3)^2} < 0 \text{ for all } \theta$$

Now, $\ln f(X;\theta) = \ln(\theta+3) + (\theta+2)\ln X \implies \dfrac{\partial \ln f}{\partial \theta} = \dfrac{1}{\theta+3} + \ln X \implies \dfrac{\partial^2 \ln f}{\partial \theta^2} = \dfrac{-1}{(\theta+3)^2}$.

Thus, $I(\theta) = -E\left[\dfrac{-1}{(\theta+3)^2}\right] = \dfrac{1}{(\theta+3)^2}$.

With a 93% CL, $\alpha = 0.07 \implies z_{\alpha/2} = z_{0.035} = \text{qnorm}(0.035, \text{lower} = \text{F}) \approx 1.811911$.

Our MLE CI is $-3 - \dfrac{1}{\overline{\ln X}} \pm 1.811911 \dfrac{1}{\sqrt{n \cdot \dfrac{1}{\left(-3 - \frac{1}{\overline{\ln X}} + 3\right)^2}}} = -3 - \dfrac{1}{\overline{\ln X}} \pm 1.811911 \dfrac{1}{\sqrt{n}|\overline{\ln X}|}$.

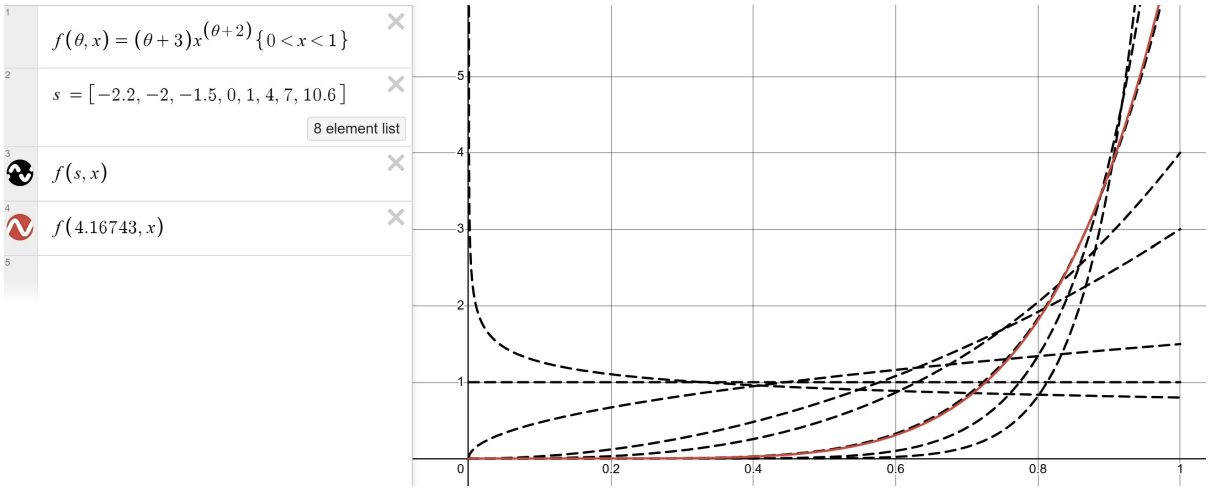
Our data give $n = 4$, $\overline{\ln X} \approx -0.13952$, and thus, our approximate MLE CI for θ is $\boxed{(-2.326, 10.661)}$.

Commentary:

This problem, like the last, is another example of a smooth MLE CI experience. I do this because there are a lot of details to get right, and most students need a little more practice before they're ready for harder challenges. I also assign this problem because the answer is a bit surprising. Here, we get an interval estimate that contains a really big range for θ. If you look at the exponent on x in the pdf and think about the CI range for θ, you'll notice that the CI is suggesting both positive exponents on x in the pdf (e.g., when $\theta = 4$) and negative exponents (e.g., when $\theta = -2.2$) are reasonable. You might be curious what all this looks like. I went to Desmos and graphed the pdf for some different values of θ (listed in the variable s) in black on the next page. In red, you'll see the pdf for $\widehat{\theta}_{\text{MLE}} \approx 4.167$; this is our best guess for the pdf given the data.

Since the four given test scores were pretty high, we anticipate a pdf that likes giving high scores (like the red curve). As seen in black, though, some θ values in the CI give pdfs that make very little sense. The reason this is happening is that we have a high confidence level (93%) and a low amount of data ($n = 4$). Basically, these two realities cannot be reconciled by the MLE CI formula, and the

statistical infrastructure is trying to tell you that, if you're willing to spend some time making sense of your answer. And here again, we face the reality of tradeoffs: If you can't get a large data set, you can't set a high confidence level unless you want an interval that is so large that it is effectively worthless.



48. Consider a RV X modeled by the density $f(x;\theta) = \dfrac{1}{\theta}x^{(1-\theta)/\theta}$ where $0 < x < 1$ and $\theta > 0$.

(a) Find the MLE for θ based on a random sample $X_1, \ldots, X_n$.

(b) According to MLE theory, $\widehat{\theta}_{\text{MLE}}$ should be asymptotically unbiased and consistent. Explicitly show that both of these are true for your result from part (a).

Hint List:

- a. The second derivative is not always negative, so you'll need to actually substitute the MLE. It helps to remember that $\sum \ln x_i = n \cdot \overline{\ln X}$, a rearrangement of the formula for an average.
- b. Try finding $E[(\ln X)^k]$ first. You may need several u-substitutions.

Solution:

(a) $L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} x_i^{(1-\theta)/\theta} \implies \ell(\theta) = \sum_{i=1}^{n} \left[-\ln\theta + \frac{1-\theta}{\theta} \ln x_i \right] = -n\ln\theta + \left(\frac{1}{\theta} - 1\right) \sum \ln x_i$

$$\ell'(\theta) = \frac{-n}{\theta} - \frac{1}{\theta^2} \sum \ln x_i$$

$$\frac{n}{\widehat{\theta}} = \frac{-1}{\widehat{\theta}^2} \sum \ln x_i$$

$$\widehat{\theta}_{\text{MLE}} = \frac{-\sum \ln x_i}{n} = -\overline{\ln X}$$

$$\ell''(\theta) = \frac{n}{\theta^2} + \frac{2}{\theta^3} \sum \ln x_i$$

$$\ell''(\widehat{\theta}_{\text{MLE}}) = \frac{n}{\left(-\overline{\ln X}\right)^2} + \frac{2}{\left(-\overline{\ln X}\right)^3} \cdot n \cdot \overline{\ln X}$$

$$= \frac{n}{\left(\overline{\ln X}\right)^2} - \frac{2n}{\left(\overline{\ln X}\right)^2} = \frac{-n}{\left(\overline{\ln X}\right)^2} < 0$$

(b) Based on the MLE formula and our end goal, we'll likely need $E[\ln X]$ and $Var[\ln X]$ (and, hence, $E[(\ln X)^2]$) moving forward. To save time, we'll find $E[(\ln X)^k]$, the Moments Technique we saw in Problem 42. We simplify the exponent in the pdf as $\frac{1-\theta}{\theta} = \frac{1}{\theta} - 1$ as we proceed:

$$E[(\ln X)^k] = \int_0^1 (\ln x)^k \cdot \frac{1}{\theta} x^{1/\theta} x^{-1}\, dx$$

Letting $u = \ln x$, we find $du = \frac{1}{x}\, dx$ and $x = e^u$. Our new bounds are $-\infty$ (lower) and $u = \ln 1 = 0$ (upper). This gives us

$$E[(\ln X)^k] = \int_{-\infty}^0 u^k \frac{1}{\theta} e^{u/\theta}\, du$$

Now, set $y = -u/\theta$, so $dy = \frac{-1}{\theta} du$ and $u = -\theta y$. Our new bounds are ∞ (lower) and $y = -0/\theta = 0$ (upper). Finally, we have

$$\int_\infty^0 (-\theta y)^k e^{-y} (-dy) = (-\theta)^k \int_0^\infty y^{(k+1)-1} e^{-y}\, dy = (-\theta)^k \Gamma(k+1) = (-\theta)^k k!$$

Thus, $E[\ln X] \overset{k=1}{=} -\theta$ and $E[(\ln X)^2] \overset{k=2}{=} 2\theta^2$. Now,

$E[\widehat{\theta}_{\text{MLE}}] = -E[\overline{\ln X}] = -E[\ln X] = -(-\theta) = \theta$, so the MLE is (asymptotically) unbiased.

$$Var[\widehat{\theta}_{\text{MLE}}] = Var[-\overline{\ln X}] = (-1)^2 \frac{Var[\ln X]}{n} = \frac{E[(\ln X)^2] - E[\ln X]^2}{n} = \frac{2\theta^2 - (-\theta)^2}{n} = \frac{\theta^2}{n}$$

Thus, $\text{MSE}(\widehat{\theta}_{\text{MLE}}) = \frac{\theta^2}{n} + (\theta - \theta)^2 = \frac{\theta^2}{n} \to 0$ as $n \to \infty$. So, the MLE is MSE-consistent, and hence, consistent.

Commentary: Why do we prove things in mathematics? One reason is so that we know which statements follow from our axioms and which don't. This information helps us make sense of the world and move forward with confidence. This problem reveals another reason. When we proved in class that (under reasonable assumptions) all MLEs are asymptotically unbiased and consistent, this means we don't need to *individually* establish these facts for each MLE that might arise in life. As this problem shows, deriving these facts for a particular RV can be quite tedious. Indeed, if you really pushed yourself on this problem, you are likely tired right now; it is a computational and bookkeeping nightmare. When I was focused on theoretical math, I felt a real sense of accomplishment in proving statements like "All X have properties Y and Z". These results were both theoretically rewarding (in that I better understood the universe) and saved me future time (in that I never needed to show properties Y and Z when I encountered a particular example of X). Hopefully, the pain of this problem helped you see the value of theoretical results.

49. One distribution useful for modeling frequencies in the field of spectroscopy has the pdf

$$f(x;c) = \sqrt{\frac{c}{2\pi}} \cdot \frac{\exp\left(\frac{-c}{2x}\right)}{x^{3/2}} \text{ where } x > 0, c > 0$$

a. Show that for any $c > 0$, the area under this pdf is 1, as it must be.

b. Given iid data $X_1, X_2, \ldots, X_n$ from this distribution, find a formula for a 92% approximate MLE CI for c.

Hint List:

- a. Try a u-substitution to push the integral toward the Gamma function/distribution.

Solution:

(a) We find $\displaystyle\int_0^\infty \sqrt{\frac{c}{2\pi}}\frac{e^{-c/(2x)}}{x^{3/2}}\,dx$. Let $u=\dfrac{1}{2x}$, so $x=\dfrac{1}{2u}$ and $dx=\dfrac{-1}{2u^2}du$. Our new bounds are ∞ (lower) and 0 (upper). So, our integral becomes

$$\sqrt{\frac{c}{2\pi}}\int_\infty^0 \left(\frac{1}{2u}\right)^{-3/2} e^{-cu}\left(\frac{-1}{2u^2}\right)du = \frac{\sqrt{c}}{\sqrt{2\pi}}\int_0^\infty 2^{3/2}u^{3/2}2^{-1}u^{-2}e^{-cu}\,du = \frac{\sqrt{c}}{\sqrt{\pi}}\int_0^\infty u^{(1/2)-1}e^{-cu}\,du$$

This integrand is the kernel of $Gamma(\lambda=c, r=1/2)$. So the Kernel Technique gives us

$$\frac{\sqrt{c}}{\sqrt{\pi}}\cdot\frac{\Gamma\left(\frac{1}{2}\right)}{c^{1/2}} = 1, \text{ as required.}$$

(b) Here, $L(c)=\displaystyle\prod_{i=1}^n \sqrt{\frac{c}{2\pi}}\cdot\frac{e^{-c/(2x_i)}}{x_i^{3/2}}$. Taking the natural log and simplifying, we get

$$\ell(c)=\sum_{i=1}^n\left[\frac{1}{2}\ln c-\frac{1}{2}\ln(2\pi)-\frac{c}{2x_i}-\frac{3}{2}\ln x_i\right]=\frac{n}{2}\ln c-\frac{c}{2}\sum\frac{1}{x_i}+C$$

$$\ell'(c)=\frac{n}{2c}-\frac{1}{2}\sum\frac{1}{x_i} \qquad \ell''(c)=\frac{-n}{2c^2}<0 \text{ for all } c$$

$$\widehat{c}=\frac{n}{\sum\frac{1}{x_i}}=\frac{1}{\overline{\frac{1}{X}}}$$

Now we turn to the FI: $\ln f(X;c)=\dfrac{1}{2}\ln c-\dfrac{1}{2}\ln(2\pi)-\dfrac{c}{2X}-\dfrac{3}{2}\ln X$.

So, $\dfrac{\partial \ln f}{\partial c}=\dfrac{1}{2c}-\dfrac{1}{2X} \implies \dfrac{\partial^2 \ln f}{\partial c^2}=\dfrac{-1}{2c^2} \implies I(c)=-E_X\left[\dfrac{-1}{2c^2}\right]=\dfrac{1}{2c^2}$.

For 92% confidence, $\alpha=0.08$, so $z_{\alpha/2}=z_{0.04}=\text{qnorm}(0.04, \text{lower}=\text{F})\approx 1.751$. So, our MLE CI is

$$\widehat{c}_{\text{MLE}}\pm z_{\alpha/2}\sqrt{\frac{1}{nI\left(\widehat{c}_{\text{MLE}}\right)}}=\widehat{c}_{\text{MLE}}\pm z_{\alpha/2}\sqrt{\frac{2\left(\widehat{c}_{\text{MLE}}\right)^2}{n}}=\widehat{c}_{\text{MLE}}\pm z_{\alpha/2}\sqrt{\frac{2}{n}}|\widehat{c}_{\text{MLE}}|$$

Since $X_i>0$, $\widehat{c}_{\text{MLE}}>0$, and we can ignore the absolute value bars. Thus, we get $\boxed{\dfrac{1}{\overline{\frac{1}{X}}}\pm 1.751\sqrt{\dfrac{2}{n}}\dfrac{1}{\overline{\frac{1}{X}}}}$.

Commentary: This is your third MLE CI problem. At this point, the overall flow should be routine, and you should be focused on the efficiency of your steps. Compare your solution to that above and find places where you can cut unnecessary steps. Never minimize the importance of efficiency.

50. Suppose X is a CRV with pdf

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \frac{\exp\left[-(\ln x - \mu)^2/(2\sigma^2)\right]}{x}$$

where $x > 0$, μ is an unknown real number, and $\sigma^2 > 0$ is known. Suppose we have a random sample $X_1, \ldots, X_n$ and wish to estimate μ.

(a) Find $\widehat{\mu}_{\text{MLE}}$. (If you have some clever, fast way to do this part, don't use it. Show all the typical steps for finding an MLE.)

(b) Use your answer in part (a) to help find an approximate (two-sided) 92% MLE CI for μ given the sample $X_1 = 1, X_2 = e, X_3 = e^2, X_4 = e$ and $\sigma = 2$. (Again, don't take any shortcuts when answering this part. Show all the usual steps.)

Hint List:

- Don't let your fear of messy algebra derail you emotionally.

Solution:

(a) First, $L(\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \frac{\exp\left[-(\ln x_i - \mu)^2/(2\sigma^2)\right]}{x_i}$.

This gives $\ell(\mu) = \sum \left[-\ln(\sigma\sqrt{2\pi}) - \frac{(\ln x_i - \mu)^2}{2\sigma^2} - \ln x_i \right] = C - \frac{1}{2\sigma^2} \sum (\ln x_i - \mu)^2$.

$$\ell'(\mu) = -\frac{1}{2\sigma^2} \sum 2(\ln x_i - \mu)(-1) = \frac{\sum (\ln x_i) - n\mu}{\sigma^2}$$

$$n\widehat{\mu} = \sum \ln x_i$$

$$\widehat{\mu} = \frac{\sum \ln x_i}{n} = \overline{\ln X}$$

$$\ell''(\mu) = \frac{-n}{\sigma^2} < 0 \text{ for all } \mu$$

So, we have $\boxed{\widehat{\mu}_{\text{MLE}} = \overline{\ln X}}$.

(b) We find the FI: $\ln f(X;\mu) = -\ln(\sigma\sqrt{2\pi}) - \frac{(\ln X - \mu)^2}{2\sigma^2} - \ln X$

Thus, $\frac{\partial \ln f}{\partial \mu} = -\frac{2(\ln X - \mu)(-1)}{2\sigma^2} = \frac{\ln X - \mu}{\sigma^2} \implies \frac{\partial^2 \ln f}{\partial \mu^2} = \frac{-1}{\sigma^2}$.

So, $I(\mu) = -E_X\left[\frac{-1}{\sigma^2}\right] = \frac{1}{\sigma^2}$. Our CI is $\overline{\ln X} \pm z_{0.08/2}\sqrt{\frac{1}{nI(\widehat{\mu}_{\text{MLE}})}} = \overline{\ln X} \pm 1.751\sqrt{\frac{\sigma^2}{n}}$.

Here, we have $\overline{\ln X} = \frac{\ln 1 + \ln e + \ln e^2 + \ln e}{4} = \frac{0+1+2+1}{4} = 1$.

So, our 92% MLE CI is $1 \pm 1.751\sqrt{\frac{4}{4}} \approx \boxed{(-0.751, 2.751)}$.

Commentary: Did you notice that in both parts (a) and (b), the problem ends by hinting there may be a much faster way to do the problem but to not use it? I did this intentionally to see if it would inspire you to be curious. If you want a chance to be curious, stop reading now and spend time thinking (and internet searching) how you might do these parts much faster. Then, return and read on.

If you look at the given pdf, it looks an awful lot like the pdf of the normal distribution. But, there are some differences: (1) we see $(\ln x - \mu)^2$ instead of $(x-\mu)^2$ in the exponent, (2) there is an extra x in the denominator, and (3) the support here is $(0,\infty)$ instead of $(-\infty,\infty)$. Despite these differences, maybe this random variable X is related to a normal distribution $Y \sim N(\mu,\sigma^2)$. Based on the different supports, you might guess that $X = e^Y$ since e raised to a power is always positive and we see a natural log (which is related to e^x). Indeed, this guess turns out to be true and can be proven if you know this result:

Let X and Y be continuous random variables and h be a monotonic function where $X = h(Y)$. Then, the pdf of X is $f_X(x) = f_Y(h^{-1}(x)) \cdot \left| \frac{d}{dx} h^{-1}(x) \right|$.

In our setting, $h(x) = e^x$, so $h^{-1}(x) = \ln x$ and $\left| \frac{d}{dx} h^{-1}(x) \right| = \left| \frac{1}{x} \right| = \frac{1}{x}$ (since $x > 0$), which is how the $\ln x$ and $\frac{1}{x}$ make an appearance in our pdf. You can fully check the pdf result above to see everything else checks out.

At this point, you might ask: Given how closely-related X is to a normal distribution, does this mysterious distribution have a name? Indeed, it does. Since $X = e^Y$, this means $\ln X = Y$, or that the log of our distribution is the normal distribution. For this reason, it is called the "log-normal distribution". In fields where the log-normal is used commonly, people will have memorized the MLE (making part (a) immediate) and the FI (making part (b) almost immediate). This is why I included the comments about showing all your work.

51. State the decision rule (i.e., test) that would be used to test the following hypotheses *for the specific test statistic mentioned*. Then, make a decision using the data provided and write a conclusion. Assume the data come from a normal distribution with unknown μ and known σ. Include a picture (OK to draw by hand, doing this in R is inefficient) of the sampling distribution for the test statistic and label the critical region.

(a) H_0: $\mu = 20, H_1$: $\mu < 20$, $n = 16$, $\sigma = 3$, and $\alpha = 0.06$. Test stat: $\overline{x}$. Data: $\overline{x} = 18.5$

(b) H_0: $\mu = 20, H_1$: $\mu < 20$, $n = 16$, $\sigma = 3$, and $\alpha = 0.06$. Test stat: $\dfrac{\overline{x} - 20}{\sigma/\sqrt{n}}$. Data: $\overline{x} = 18.5$

(c) H_0: $\mu = 10, H_1$: $\mu \neq 10$, $n = 100$, $\sigma = 0.4$, and $\alpha = 0.12$. Test stat: $\overline{x}$. Data: $\overline{x} = 11$

(d) H_0: $\mu = 50, H_1$: $\mu > 50$, $n = 60$, $\sigma = 4$, and $\alpha = 0.08$. Test stat: $3\overline{x}$. Data: $\overline{x} = 50.5$

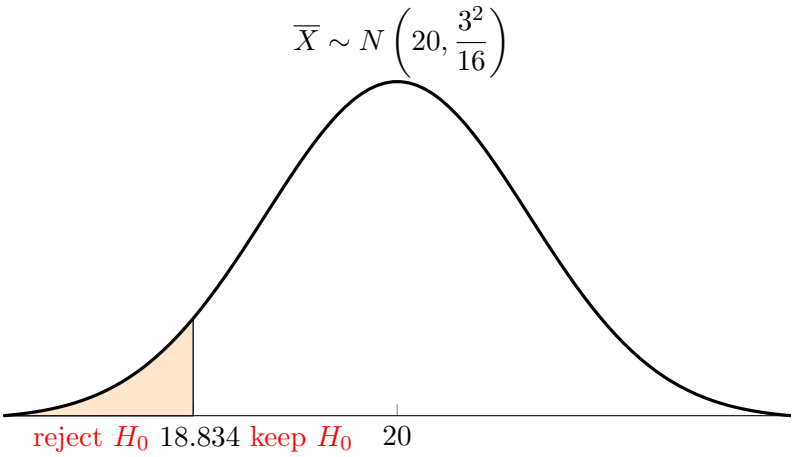
Note: Life is about trade-offs. This problem helps you see this. For example, part (a) has a nicer-looking test stat, but the distribution it follows is a little messier. Part (b) has a messy test stat, but the distribution it follows is very nice. Part (d) is here to remind you that just about any expression can act as a test stat, as long as you can determine its distribution. Since we never know what expression might arise from an MME or MLE, this reminder is comforting.

Hint List:

- If you are stuck, read the answer to (a) and then try to do parts (b) through (d) without assistance.

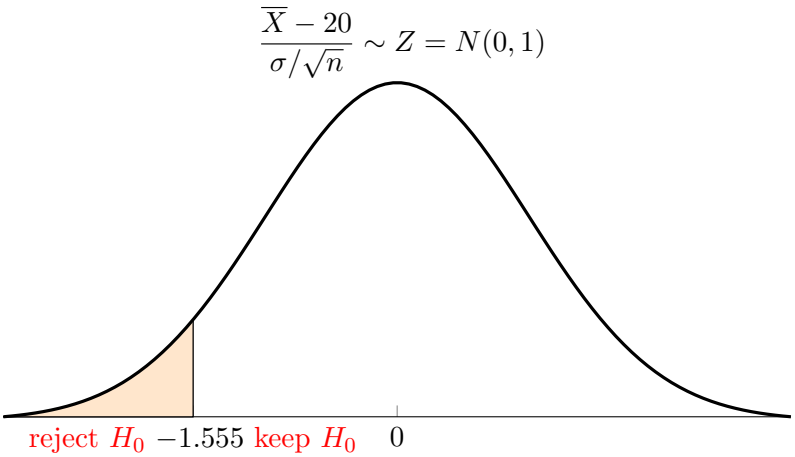
Solution:

(a) We want to reject H_0 when $\overline{x}$ gets too small since H_1: $\mu < 20$ believes μ should be smaller than 20. The bottom $0.06 = 6\%$ of strange values are demarcated by qnorm(0.06, 20, 3/4) ≈ 18.834.



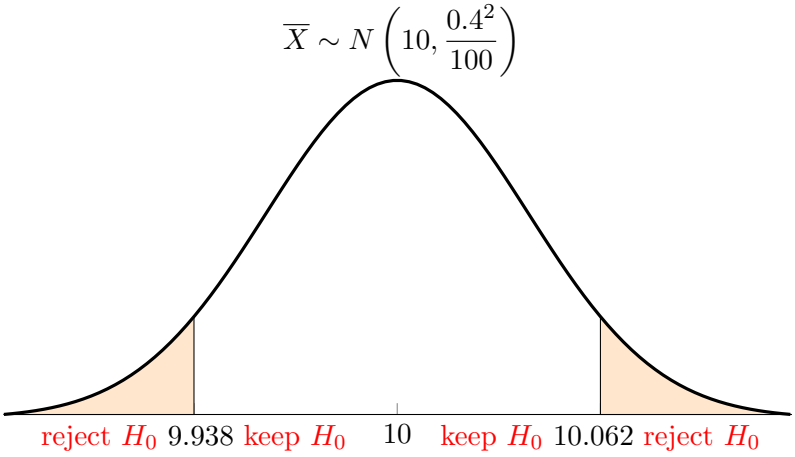
Thus, our decision rule (i.e., test) is: If $\overline{x} < 18.834$, reject H_0; else keep H_0. Using the test on our data, we see $18.5 < 18.834$, hence we reject H_0 in favor of H_1. Note that the test above and the picture both give the critical region: $(-\infty, 18.834)$.

(b) As in part (a), we want to reject H_0 when our test stat $\dfrac{\overline{x} - 20}{\sigma/\sqrt{n}}$ gets small because the alternative hypothesis is focused on small values of μ, which would appear as small values of $\overline{x}$, and hence, small values of $\dfrac{\overline{x} - 20}{\sigma/\sqrt{n}}$. Here, the test stat has a simple distribution: $\dfrac{\overline{X} - 20}{\sigma/\sqrt{n}} \sim Z$ because $\overline{X}$ is normal (because X is) and has undergone a standardizing process.



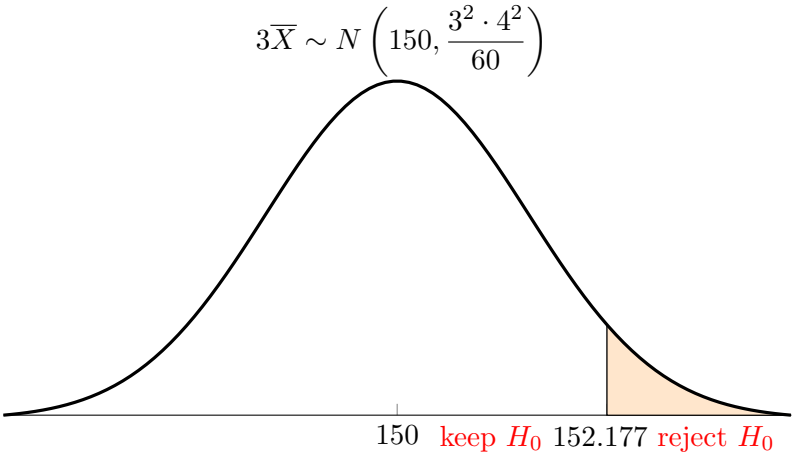
Here, we use qnorm(0.06) ≈ -1.555 to find the demarcation of the bottom 0.06 area. So, our test is: If $\dfrac{\overline{x} - 20}{\sigma/\sqrt{n}} < -1.555$ reject H_0; else keep H_0. Our data give $\dfrac{18.5 - 20}{3/4} = -2 < -1.555$, so we reject H_0, as in part (a).

(c) With a two-sided alternative (H_1: $\mu \neq 10$), we split α into two equal pieces and find qnorm(0.06, 10, 0.4/10) ≈ 9.938 and qnorm(0.06, 10, 0.4/10, lower=F) ≈ 10.062.



Our test: If $\overline{x} < 9.938$ or $\overline{x} > 10.062$, reject H_0; else keep H_0. Here, our data give $\overline{x} = 11 > 10.062$, so we reject H_0 in favor of H_1.

(d) This part has decided $3\overline{x}$ is the way to compress our sample down to a single number, so we need to know the distribution of $3\overline{X}$. Note that $3\overline{X} \sim 3\,N\left(50, \frac{4^2}{60}\right) = N\left(150, \frac{3^2 \cdot 4^2}{60}\right)$ because $E[aX] = aE[X]$ and $Var[aX] = a^2 Var[X]$. Here, the alternative is H_1: $\mu > 50$, so large values μ (and hence, $3\overline{x}$) are strange to someone who believes in H_0. Thus, we set the demarcation line (AKA "rejection fence") by finding qnorm(0.08, 150, 3*4/sqrt(60), lower=F) ≈ 152.177.



Our test is: If $3\overline{x} > 152.177$, reject H_0; else keep H_0. Our data give $3\overline{x} = 151.5 < 152.177$, so we keep H_0.

Commentary:

I think the hypothesis testing infrastructure is awkward and counterintuitive. I have felt this way since I first learned it as an undergraduate, and I continue to feel so today. After all that time, here is how I try to explain the idea to the masses of people who are understandably confused: When you are conducting a hypothesis test, you are always focused on a parameter. This is a value that underpins a random variable, like μ from $N(\mu, \sigma^2)$ or p from $Binom(n, p)$. We write two competing hypotheses

about this parameter and then must decide which is better to believe in. To make this decision, we collect a data set related to the random variable that has the parameter. For example, if you care about the average one-rep-max bench press for UCSD students, you are excited about μ from $X \sim N(\mu, \sigma^2)$. Your data $x_1, \ldots, x_n$ are one-rep-max bench press values from n UCSD students. To use these values to make a choice between the hypotheses, we must collapse the n values to a single number. There are many ways you might do this. For example, you could find $\overline{x}$ or $\dfrac{\overline{x} - \mu}{\sigma/\sqrt{n}}$ or $3\overline{x}$ or etc.

Once you have this number (known as the test stat), you have to decide if it lends more support to H_0 or H_1. To make this decision, you have to know what values you could have gotten for the test stat. This means you have to know the distribution of $\overline{X}$ or $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ or $3\overline{X}$ or etc. To figure out this distribution, you usually need to assume a value for the parameter, and since the null looks like H_0: parameter = number, we temporary assume this to get a value for the parameter. For example, in part (a), by assuming H_0 is true, we know $\mu = 20$, and hence $X \sim N(20, 3^2)$, and hence $\overline{X} \sim N\left(20, \dfrac{3^2}{16}\right)$.

Now we can decide if our particular data compression of $\overline{x}$ is weird by seeing how unlikely it is on $\overline{X}$. Oh, but wait. We know $P(\overline{X} = \overline{x}) = 0$ since $\overline{X}$ is a CRV. So, instead, we need to shade a tail area to determine that our $\overline{x}$ is "among the ...% strangest results we could have gotten" (P-value calculated). And, oh wait, we need to decide an arbitrary cutoff (the significance level α) for when something is too strange to hold our original belief structure. And after we compare these values, we can make a decision about what to believe moving forward. So, said "simply": we used a numerical compression of our data set (the test stat) and the distribution of that test stat (based on the assumption of H_0) to gauge the strangeness of our data via tail areas as determined by H_1 (the P-value), which we compared to an arbitrary threshold (α) to make a decision between two hypotheses about a parameter from a starting distribution. You can forgive people for being confused!

To make things worse, at the end of this nightmarish logic, all we do is choose which hypothesis we prefer to believe in moving forward. We have not *proven* either hypothesis is true or false, and we might have made a mistake when running the test because any data set can mislead us due to randomness. Also, there is usually a theory underpinning the alternative (e.g., you might believe the average one-rep-max bench press for UCSD students is higher than a historical value because the dining hall has added more protein options to the menu). If you end up believing the alternative hypothesis, it doesn't imply your theory was correct. Maybe bench presses are improving because better training information is available on the internet, not your original theory.

My experience with null hypothesis significance testing (NHST) is that (1) students rarely understand it because it is a nightmare, (2) understanding does not improve over time (studies confirm this), and (3) basically everyone wants it to do things it cannot do (e.g., prove hypotheses are true/false and offer support to theories). Sadly, people refuse to take the time to truly understand NHST, and indeed, many leverage its abstruse nature to mislead themselves and others about what conclusions can actually be drawn from a statistical hypothesis test. Please do not be part of the confused masses. There are many confusing tools in the intellectual space (e.g., NHST, delta-epsilon arguments in analysis, academic jargon); do not abuse their complex nature to delude yourself or others.

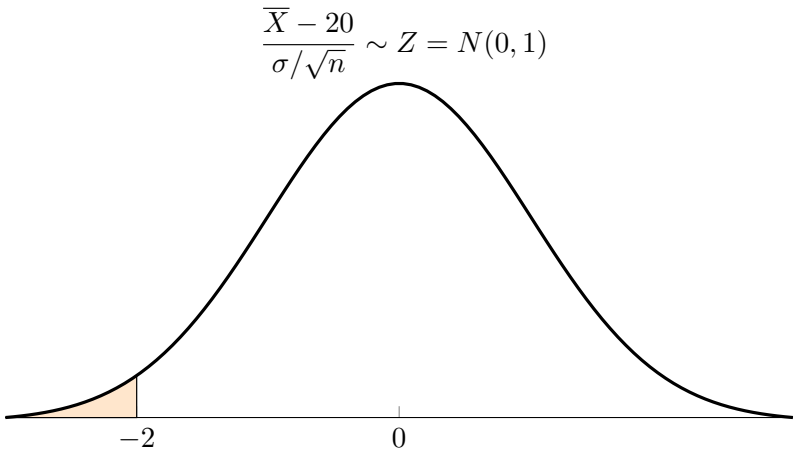
52. Calculate the P-values for Problems 51b and 51c. Does using these P-values lead you to the same conclusions as the critical regions/decision rules did?

Hint List:

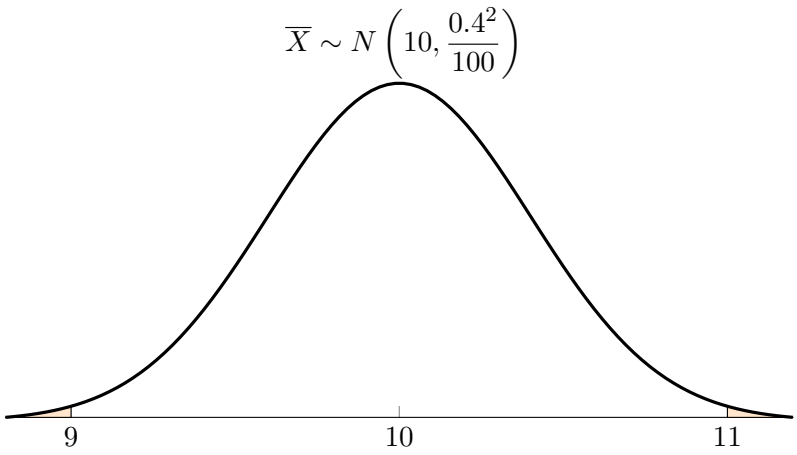
- You'll need tables or technology to do this calculation. You cannot find these areas through integration.

Solution:

For part (b), we shade to the left of the test stat $\dfrac{\overline{x} - 20}{\sigma/\sqrt{n}} = \dfrac{18.5 - 20}{3/4} = -2$. Since we are on the standard normal, this area is easily found in R using pnorm$(-2) \approx 0.023$ (the shaded area below). Since $0.023 < 0.06$, we reject H_0 in favor of H_1, as before.



For part (c), we shade to the right of our test stat $\overline{x} = 11$ **and** to the left of its symmetric spot on the distribution. Since 11 is one unit above the mean of 10, the symmetric spot is one unit below the mean. In R, we can find one of these tail areas and double it to get the P-value: 2 * pnorm(9, 10, $0.4/10) \approx 0 < 0.12$, so we reject H_0 in favor of H_1, as before.



It's important to know that the answer you get (keep vs. reject H_0) is the same whether you use (1) critical regions/decision rules as in the last problem or (2) P-values as in this problem. They are just different frameworks for the same logic flow.

Commentary: When calculating the P-value for part (c), you get a very small value (indeed, the above solution says it's approximately 0). Many students get worried at this point and think they've

done the problem wrong. I intentionally chose the test stat value ($\overline{x} = 11$) so this would happen and you would be tempted to get worried. While strange-looking answers can be a sign you've done something wrong, this is not always the case. The trouble here is that you've spent the last 10+ years in a setting (K through 12 education) where all your answers were nice. This was part of the secret contract you and your teachers developed but never said out loud. Well, that ends now. Moving forward, you should expect nightmare answers. The real world is going to be full of them, and it doesn't care about making your life easy-breezy.

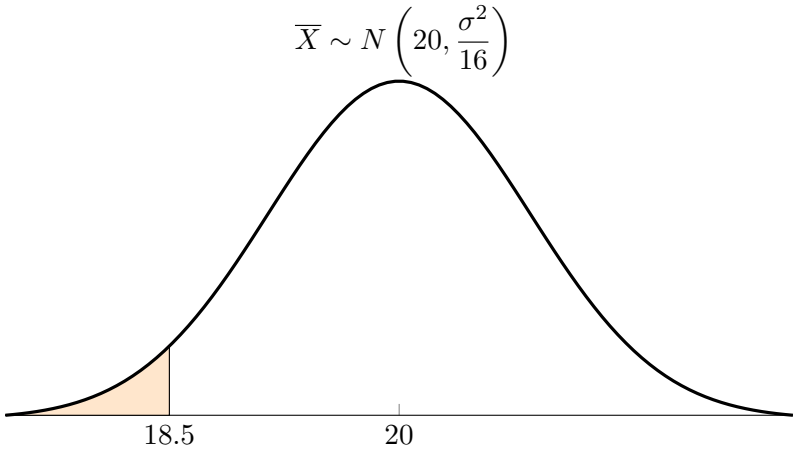
To put this all simply: Stop making judgments about the accuracy of your answers using silly narratives from 6th grade. Instead, actually think about your setting and see if the absurdity of the answer can be justified. In part (c), $\overline{X}$ has a standard deviation of $\frac{0.4}{10} = 0.04$. So, if we go 3 SDs from the mean, we reach $10 - 3 \cdot 0.04 = 9.88$ and $10 + 3 \cdot 0.04 = 10.12$. The locations based on our data (9 and 11) are 25 SDs (!!) away from the mean. So, we would expect basically no area in the tails beyond these insane locations. Now our answer makes perfect sense.

53. Suppose you wanted to alter Problem 51a so that the P-value, when calculated, would equal 0.04. If you could only change σ, what value would it need to equal to get the P-value to be 0.04?

Hint List:

- Draw a picture of $\overline{X}$ that includes 18.5 and a shaded area. Since you can't answer the question directly on this distribution (since σ is unknown), can you answer it on a related distribution?

Solution: We are exploring the test stat $\overline{x} = 18.5$ on $\overline{X} \sim N\left(20, \frac{\sigma^2}{16}\right)$ and trying to get a P-value of 0.04. We draw a picture below where the shaded area is 0.04.

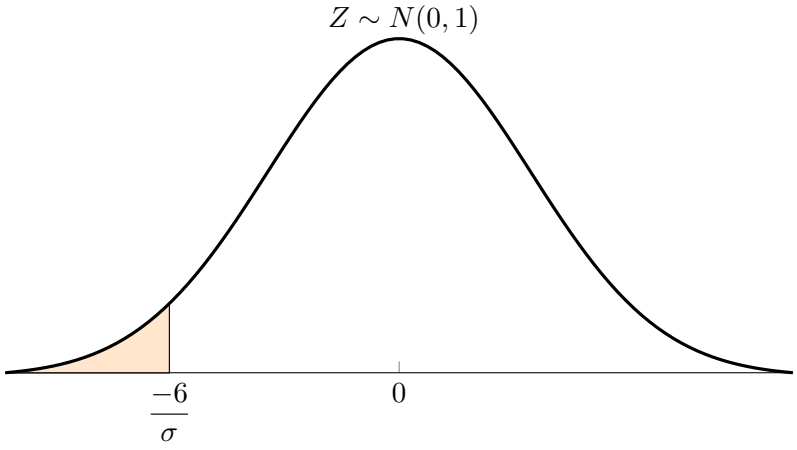


So, $0.04 = P\left(\overline{X} < 18.5\right) = P\left(\dfrac{\overline{X} - 20}{\sigma/4} < \dfrac{18.5 - 20}{\sigma/4}\right) = P\left(Z < \dfrac{-6}{\sigma}\right)$.

On Z, we know the bottom 4% area is demarcated by the value qnorm(0.04) ≈ -1.751.

So, $\dfrac{-6}{\sigma} \approx -1.751 \implies \boxed{\sigma \approx 3.427}$.

Commentary: The key to this problem is creating an equivalent question on a different distribution: $Z \sim N(0, 1)$. When we do so, we actually get this picture:



What's great is that the algebraic act of standardizing *relocates the unknown.* In the top picture, the unknown is part of the distribution (the variance $\sigma^2/16$); in the bottom, the unknown is a place on the horizontal axis (the quantile $-6/\sigma$). This movement is useful because it means we completely know the distribution in the bottom picture and can begin answering area questions about it. Amazing!

54. In December 2017, the J-RPG Xenoblade Chronicles 2 was released for the Nintendo Switch. The game is epic in its number of main quests and side quests. Those who try to finish every aspect of the game are known as "completionists". Given the massive scope of this game, you might ask: "What is the (current) average time for all completionists on Earth?" Assume completion times are normally distributed with unknown mean and standard deviation 50 hours (a reasonable estimate for J-RPGs). Before you collect data, your friend claims this average time is 250 hours (based on their personal experience). You think the value is something different and go to HowLongToBeat.com to find some data. Based on when I looked at this page (don't use more recent data!), 96 completionists had submitted their times for an average of 254 hours. Define parameter(s), write hypotheses, draw a sampling distribution, and decide which hypothesis to support using $\alpha = 0.01$ (and any one of the three methods shown in class).

    (For those curious, my completion time was around 225 hours, and my current play time is around 700 hours because of expansion pass content.)

Hint List:

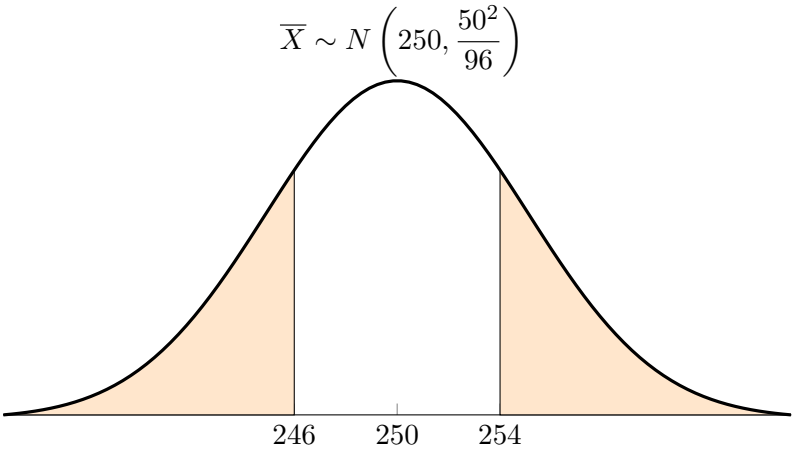
- Start by setting up a random variable for the time of a random Xenoblade Chronicles 2 completionist. What is it distribution, center, and spread?

Solution:

Let $X \sim N(\mu, 50^2)$ be the time for a random completionist of Xenoblade Chronicles 2 (XC2). We focus on μ, the average time of **all** XC2 completionists. We test

$$H_0\text{: } \mu = 250 \text{ vs. } H_1\text{: } \mu \neq 250 \text{ (“something different”)}$$

We have $\overline{x} = 254$ and $n = 96$. Here, we find a P-value rather than use critical regions. Assuming H_0, we know $\overline{X} \sim N\left(250, \dfrac{50^2}{96}\right)$, and we explore how strange $\overline{x} = 254$ is on this distribution. Because of the two-sided alternative, we shade both tails as below:



Thus, our P-value is 2 * pnorm(254, 250, 50/sqrt(96), lower=F) $\approx 0.433 > 0.01$. So, we $\boxed{\text{keep } H_0}$. Our friend's claim may be correct.

Commentary:

Here, again, it is tempting to think you did something wrong because the P-value isn't even close to the significance level α. If you've taken stats before, your teachers may have artificially engineered problems so that these two numbers were often close. This served two purposes: (1) it trained you to expect this and offered a subtle way for you to check your answer (hence, inflating grades), and (2) it created a false sense of drama (e.g., “We just made the cut and get to reject H_0!” or “Awww, we just missed having a statistically significant result”). As you might guess, I think this “answer engineering” is bad for students because it creates an academic narrative that doesn't exist in the real world. So, dump the narrative that the P-value and α must be close.

55. Students often wonder what to do if you get a P-value of **exactly** 0.05 when $\alpha = 0.05$. In truth, it doesn't matter if you define this as rejecting H_0 or keeping it, because the probability your P-value exactly equals α is 0 (since the P-value is actually a continuous random variable). Let's say you wanted to be evil and design a problem for your next statistics exam where the P-value would exactly equal 0.05. You plan to make a problem where we study μ from $X \sim N(\mu, 3^2)$ with data $\overline{x} = 7$ and $n = 28$. What value(s) should you have students use for the null hypothesis to get your P-value to be 0.05 assuming a two-sided H_1?

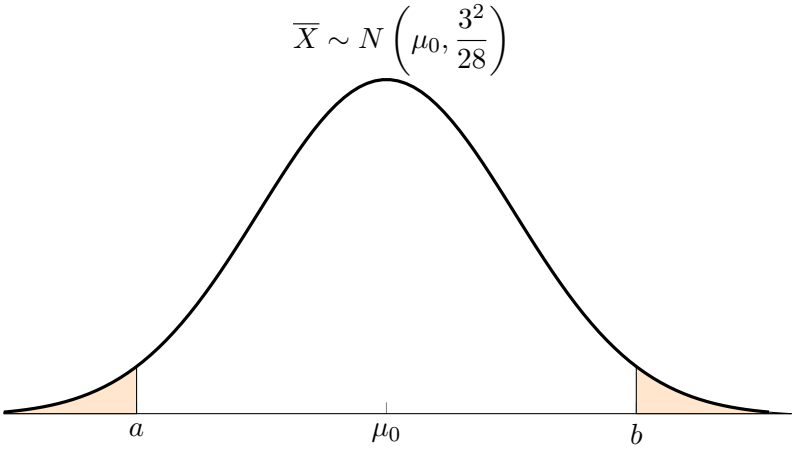
Hint List:

- You'll get two answers because of the two-sided nature of the problem.
- Start with $\overline{X}$ (what is its distribution?) and then move to Z.

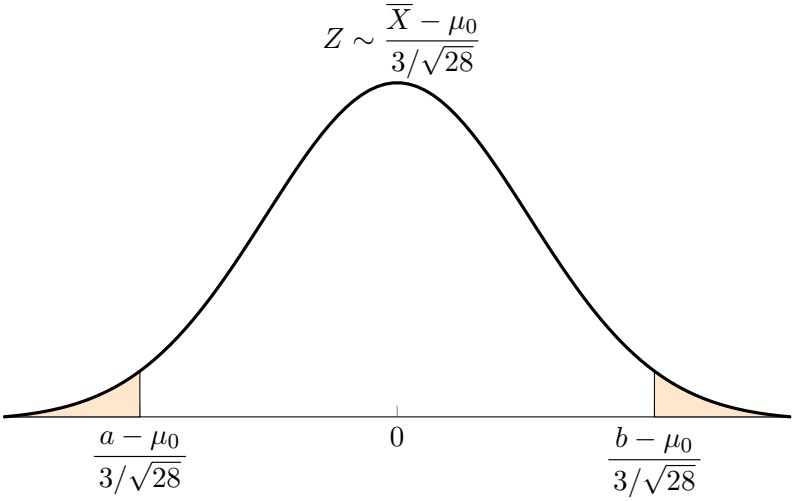
Solution: We study $X \sim N(\mu, 3^2)$ with the competing hypotheses

$$H_0\text{: } \mu = \mu_0 \text{ vs. } H_1\text{: } \mu \neq \mu_0 \text{ (“two-sided } H_1\text{”)}$$

We assume H_0 and know $\overline{x} = 7, n = 28$, and $\sigma = 3$. To get a P-value of exactly 0.05, we need $\overline{x}$ to land at a or b in the below picture, where each tail has area 0.025:



Since we don't know μ_0 (and hence, the distribution of $\overline{X}$), we standardize to shift the location of the unknown from the distribution to quantiles:



On Z, we can find the demarcation spots for tail areas of 0.025 using R: qnorm(0.025) ≈ -1.96. So, we have $-1.96 = \dfrac{a - \mu_0}{3/\sqrt{28}}$. If $\overline{x} = 7$ is at a, we substitute this and find that $\boxed{\mu_0 \approx 8.111}$. Similarly, if $\overline{x} = 7$ falls at b, we get $+1.96 = \dfrac{b - \mu_0}{3/\sqrt{28}} = \dfrac{7 - \mu_0}{3/\sqrt{28}} \implies \boxed{\mu_0 \approx 5.889}$.

Commentary: I like this question because it does a few things really well. First, it preempts common student questions ("Can I get a P-value of 0.05?" and "What should I do if this happens?"). One thing you should expect of your instructors is that they know, in advance, the things you might ask and plan accordingly. When educators can do this, it means they have moved beyond thinking about the particular mathematical content of the course and have begun thinking about how students experience this content. Great teachers think about questions like: What common misconceptions will arise? What metaphors do students find useful for understanding ideas? And so on. If you'd like to be a great instructor some day, learning the core content knowledge is just the beginning.

The second interesting aspect to this question is that it's easiest to solve when you use a mixture of different *semiotics*. I think of a semiotic as a way we communicate meaning. For example, the English alphabet is a collection of symbols that convey meaning when ordered in just the right way. Other semiotics include: sheet music notation, mathematical notation, graphs of functions, hieroglyphics, programming language syntax, and American Sign Language. In the above solution, I use mathematical notation, graphing, R code, and written language to weave a step-by-step explanation that helps the reader understand a complex statistical situation. My experience is that the best problem solvers and communicators are experts in many different semiotics.

As an example, in Calculus you learn about the derivative. You could present this solely in the mathematical semiotic: $f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$. Doing so, though, leaves out much of the story. Moving to the visual semiotic, we can recast $f'(a)$ as the slope of the tangent line to $f(x)$ at $x = a$. This helps us understand why $f(x) = |x|$ is not differentiable at $a = 0$. We might even move to the kinesthetic semiotic (i.e., use of the body to convey meaning) and have someone think about derivatives in terms of body position, velocity, and acceleration. In the language semiotic, we might build a mental dictionary of phrases related to the derivative: rate of change, instantaneous change, maximum, minimum, change over time, marginal (cost, revenue), etc. Notice how a rich understanding is built *across* multiple semiotics rather than *within* a single semiotic. You should aim for this ideal when trying to understand or explain a concept.

56. Change is Coming!

(a) Suppose a problem has H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$. If a given data set causes us to reject H_0 when $\alpha = 0.02$, would the same data set force us to reject H_0 if $\alpha = 0.05$?

(b) Suppose a problem has H_0: $\mu = \mu_0$ and H_1: $\mu > \mu_0$. If a given data set causes us to reject H_0 for some α, would the same data set force us to reject H_0 if we change H_1 to $\mu \neq \mu_0$? Assume α remains the same.

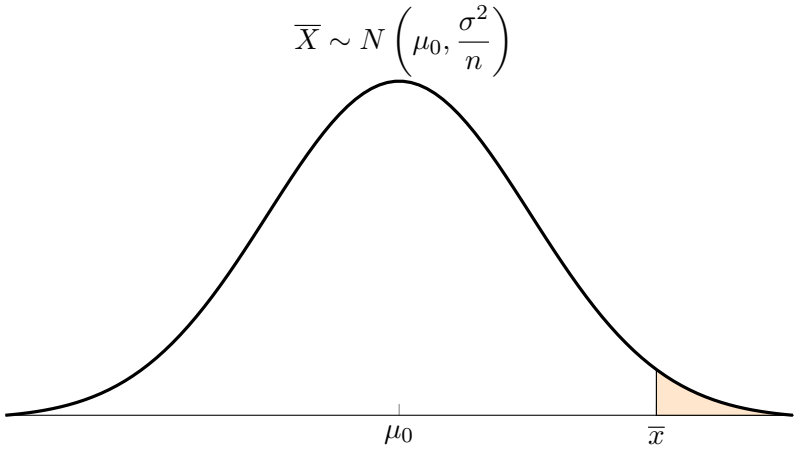
Hint List:

- b. Try different numbers for the P-value and draw pictures to determine if/how the test conclusion changes.

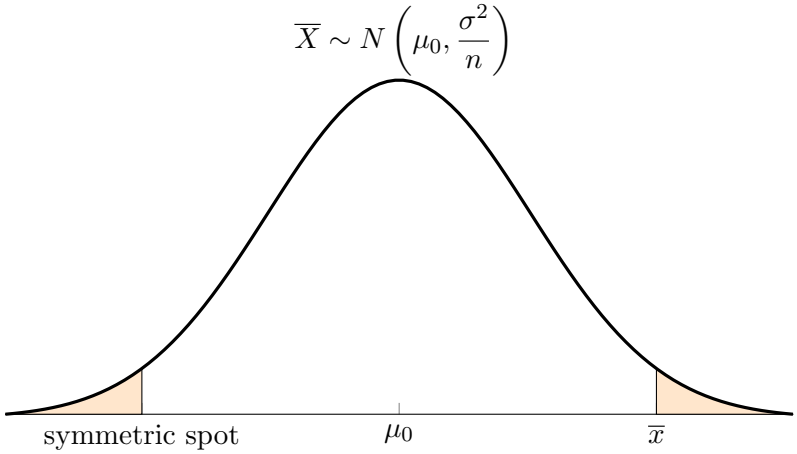
Solution:

(a) If we reject H_0, this means our P-value is less than $\alpha = 0.02$. If α is changed to 0.05, we must have P-value $< 0.02 < 0.05 = \alpha$, so we still reject H_0.

(b) We can't be sure what will happen when changing a one-sided alternative to a two-sided alternative. Before the switch we have the below setup, where the shaded area is the P-value and H_1 is $\mu > \mu_0$:



After the switch we have the below setup, where the shaded area is the P-value and H_1 is $\mu \neq \mu_0$:



Here, our P-value is twice as much. Different results are possible for the hypothesis test after the switch. For example, if the P-value was initially 0.02 and α was 0.05, we would reject H_0 before the switch (since $0.02 < 0.05$), and we would reject H_0 after as well (since $2 \cdot 0.02 = 0.04 < 0.05$). If, however, our P-value was initially 0.03, then we would reject H_0 before the switch (since $0.03 < 0.05$) but keep H_0 after (since $2 \cdot 0.03 = 0.06 > 0.05$).

Commentary: This problem reveals that there are ways to adjust the hypothesis testing framework that alter the conclusion of the test. In part (b), switching from a one- to a two-sided alternative can shift the result from statistically significant (P-value $< \alpha$) to not statistically significant (P-value $\geq \alpha$). The same is true in part (a): if you make α smaller (rather than larger), you can change the end result as well. These are examples of "questionable research practices" (QRPs). A QRP is a way of manipulating an experiment to achieve the result you wish to achieve. The list of QRPs is very long since you can corrupt the ideals of science at any point in an experiment. Some examples include: intentionally picking a biased sample, tossing "outliers" that are counter to your pre-held belief, manipulating choices in the hypothesis test framework (see above), re-running experiments when the outcomes don't match your desired outcomes, creating false data sets, choosing new statistical tests when the original tests don't give you what you want, etc. If you think these are crazy and no one is doing them, you're wrong. Indeed, studies of QRP prevalence in different disciplines suggest at least 50% of researchers use them. To learn more, search "QRP prevalence" + your field of interest. As you might guess, getting people to admit their unethical behavior is difficult, so most of these estimates are believed to be underestimates. The opposite of a QRP is an ORP (open research practice). ORPs are behaviors that promote transparency and honesty. Some examples include: pre-registering your study on a public forum (including how data will be collected and what analyses will be run), making all your data available to others, and, for journal editors, publishing articles not based on statistical significance, but based on the quality of the study design.

57. You've just made the best app ever! You plan to upload it to the app store and are curious how many reviews you might get from users. The histogram of review counts for various apps in the Apple store is very right-skewed: most apps get a small number of reviews, but some apps − like Pandora, PayPal, and LinkedIn − get millions. It turns out that ln(review count) is roughly normally distributed with $\sigma = 2.6$ (for those apps with more than 5 reviews). In this problem, we'll explore $Y \sim N(\mu, 2.6^2)$ where $Y = \ln X$ is the natural log of the review counts. Your friend claims that $\mu = 6.5$, but you think it's higher: people love rating stuff in the modern era! Using the data set AppleStore.csv (found on Canvas/TritonEd in the Homework folder), conduct a hypothesis test to determine whom to momentarily believe in life. This data set contains information on 7197 random apps from Apple's app store. Load this into R using the "Import Dataset" button in the upper right window of R studio. Make sure to remove rows with 5 or fewer reviews. Your answer will be a mix of R code and written work. Use $\alpha = 0.01$. The rating_count_tot column lists how many times a given app has been rated by users.

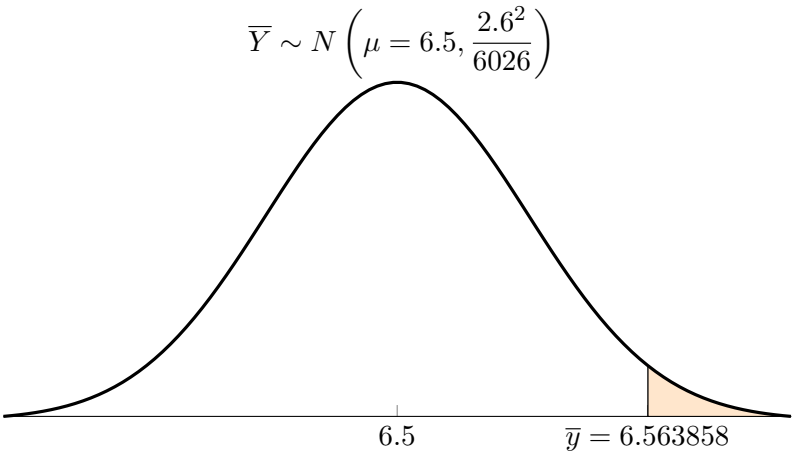
Hint List:

- Typing ?subset in R can help you learn how to take a subset of a data frame. Scroll to the bottom of the documentation to see examples.
- If you're struggling, take a moment to explore why. This problem requires coding, visualizing, and mathematical notation.

Solution: We want to test the hypotheses H_0: $\mu = 6.5$ vs. H_1: $\mu > 6.5$ ("you think it's higher") To do so, we must load, clean, and summarize our data:

```
AppleStore = read.csv("file location on your computer")
cleanData = subset(AppleStore, rating_count_tot > 5)
cleanedRatings = cleanData$rating_count_tot
logRateCounts = log(cleanedRatings)
hist(logRateCounts)   # checking problem's claim of normality
mean(logRateCounts)   # getting ybar
```

This code gives $\overline{y} = 6.563858$ and $n = 6026$. Based on this, we can draw a picture of $\overline{Y}$ and gauge how strange our data are via $\overline{y}$ (or a standardized version):



Here, our P-value is pnorm(6.563858, 6.5, 2.6/sqrt(6026), lower=F) $\approx 0.028 > 0.01$, so we keep H_0.

Commentary: This problem, like many recent ones, requires you to work across many semiotics. The solution begins with written language, jumps to R code, shifts to a visualization, and finishes with more code and mathematical notation. If you're interested in applied mathematics, statistics, data science, machine learning, or AI, then you'll be constantly switching between semiotics. As such, it is important to begin identifying and attacking your weaknesses. You don't get to say "I'm not good at coding" or "I don't think visually". You can be good at all of these things if you spend time working on them (through coursework, reading on your own, doing projects, etc.). In addition, there is a lot of competition in these fields, so if you can't bring expertise across all these semiotics, companies will find someone who can (or use AI). Other domains help cement this point: Would you hire an NFL quarterback that was good at passing but struggled with handoffs and was a poor verbal communicator? Do you want a musician in your orchestra that is great at interpreting Romantic era music but is stylistically lost with Baroque and Classical compositions? High-level jobs demand excellence across semiotics; prepare accordingly.

58. Most manufacturing processes create defective items. Often these are tolerated up to a certain point, after which machines must be replaced, a costly and time-consuming process. Suppose you are working at a company that will permit 6% of all items to be defective. Your boss is curious if things have gotten worse and asks you to inspect 230 random items.

    (a) If you find 23 items are defective, what advice should you give your boss based on an approximate hypothesis test with $\alpha = 0.01$? As always, follow the steps from class.

    (b) In R, you can conduct the test quite easily with the prop.test command. Read the documentation for this by typing ?prop.test in R and write a single line of code that reproduces your results from part (a). Note: Set the "continuity correction" to false. We don't discuss this in class, but you can read about it here if you're curious.

    (c) Your boss is wondering if an exact test for this situation would give different results. Find the P-value based on an exact test without using binom.test (you'll still need the computer at one point) and then with binom.test in R (you'll get the same answer).

Hint List:

- c. When you call a function like pnorm(x) in R, it finds $P(X \leq x)$, not $P(X < x)$. For CRVs, the difference between $\leq$ and $<$ doesn't matter, but for DRVs, it does. So, if you call pbinom(q=4, size=9, prob=0.5), this finds $P(X \leq 4)$ where $X \sim Binom(n = 9, p = 0.5)$, not $P(X < 4)$. If you read through the documentation on pbinom, the inequality being used is not clear!

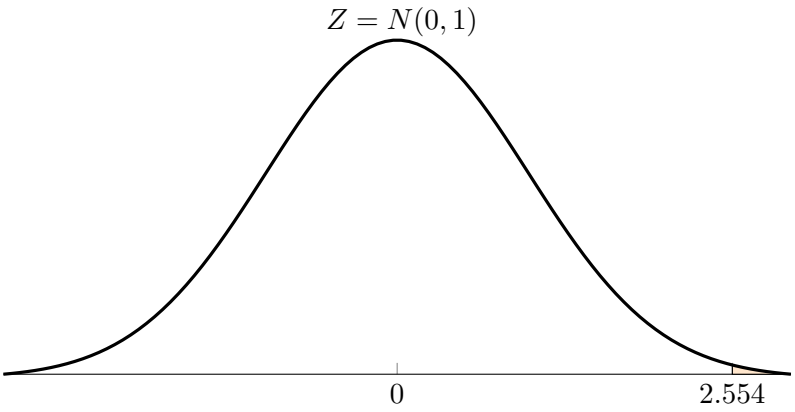
Solution:

(a) Let p be the percentage of **all** currently-produced items that are defective. We test

$$H_0\text{: } p = 0.06 \text{ vs. } H_1\text{: } p > 0.06 \text{ (“things have gotten worse”)}$$

We can use an approximate test if we meet the Larsen-Marx criterion or the $np, nq \geq 10$ criterion. We use the latter inequalities. After assuming H_0, we have $np = 230 \cdot 0.06 = 13.8 \geq 10$ and $nq = 230 \cdot 0.94 = 216.2 \geq 10$, so we have met the assumptions required to run the test.

Now, $\widehat{p} = \dfrac{23}{230} = 0.10$. So, $z = \dfrac{\widehat{p} - p}{\sqrt{pq/n}} = \dfrac{0.10 - 0.06}{\sqrt{0.06 \cdot 0.94/230}} \approx 2.554$. We place this test stat on the test distribution, Z, and shade the upper tail because of the alternative hypothesis:



In R, we find the P-value is pnorm(2.554, lower=F) $\approx 0.0054 < 0.01$, so we reject H_0 in favor of H_1. It seems the percentage of defective products has increased.

(b) In R, we do:

```
prop.test(23, 230, p=0.06, alt="greater", correct=F)
```

(c) Let $X \sim Binom(n = 230, p)$ be the number of defective items among 230. Here,

$$P\text{-value} = P(\text{data}^+ \mid H_0) = P(X \geq 23 \mid p = 0.06) = \sum_{i=23}^{230} \binom{230}{i} 0.06^i \cdot 0.94^{230-i}$$

We can evaluate this sum in R (see hint above) by doing

```
1 - pbinom(22, 230, 0.06)
```

This gives us a P-value $\approx 0.0118 > 0.01$, so we keep H_0. It seems the percentage of defective items is still 0.06. We can write one line of R code to do the entire test:

```
binom.test(23, 230, p=0.06, alt="greater")
```

Commentary: This problem shows that we can get different results based on which type of statistical test we use. As such, we can get different conclusions based on the same data! As mentioned in a previous commentary, this provides an opportunity to engage in a questionable research practice (QRP): simply try different analytic frameworks until you get the result you want! While I hope this will not be your approach to life, it does raise a worrisome issue. If different analytic frameworks lead to contradictory results, what are we to do?

My first response is that you must accept that when two people have the same data ("the same facts", as the media likes to say) they may reach different conclusions. It is natural to dislike this reality, especially after years of mathematical training where everyone gets the same answers from the same starting prompts. Real life, however, is far more complex. Second, when different answers are coming from different frameworks, you need to start critiquing those frameworks. Here, we used an approximate test and an exact test. The approximate test trades imprecision for a gentler distribution, Z, to model the test stat. Since we had access to a computer and could run the exact test, this trade-off wasn't really necessary, so we might favor the exact test.

Third, when you get conflicting results from a hypothesis test, often your two P-values are quite close and on opposite sides of α (here, we have P-values of 0.0118 and 0.0054). In truth, the tests are basically trying to say the same thing, but α has wedged itself between them. In such cases, collecting a little more data can move the P-values to the same side of α and you no longer have a conflict. Finally, you could try stepping away from harsh binaries in life. Here, α is a line in the sand: on one side you're an H_0 person, on the other, you're an H_1 person. Instead of sacrificing yourself to the tyranny of the binary, you could return to your boss and say this:

"I ran a study to explore the rate of defective items. The old rate was 6% and my small sample suggested a higher rate of 10%. This might be because the rate has actually increased, or my higher value may be due to randomness when collecting data, or both. If the rate really is still 6%, my 10% is very strange. According to one analysis, it's among the 1.18% strangest high things I could have gotten. According to another analysis, it's among the 0.54% strangest high things. So, either the rate is now higher, or it's the old rate and I saw something with odds around 1 in a 100 or 1 in 200. I'll let you decide what actions to take moving forward."

59. Does the idea known as "home-field advantage" (HFA) actually exist? HFA suggests that in a given sport, the home team will beat the away team more often than half the time. Several theories have been offered for why this might occur: familiarity with your arena/playing space, support of the home crowd, refereeing that favors the home team, etc. To explore HFA, researchers looked at 1000 random NFL games in the last 40 years and found that in 574 cases, the home team won.

    (a) Draw a conclusion about the idea of HFA using an approximate test with $\alpha = 0.02$, and show that you have met the conditions necessary for using this test.

    (b) After publication of the findings from part (a), you read on an NFL blog that "Data show the existence of HFA, likely the result of biased refereeing." Critique this claim from a statistical perspective.

Hint List:

- b. Focus on the words "show" and "likely" in your critique.

Solution:

(a) Let p be the percentage of **all** NFL games where the home team wins. We test

$$H_0\text{: } p = 0.5 \text{ vs. } H_1\text{: } p > 0.5 \text{ ("more often than half")}$$

When we assume H_0, we have $np = nq = 1000 \cdot \frac{1}{2} = 500 \geq 10$, so we can use an approximate test. Here, $\widehat{p} = \frac{574}{1000} = 0.574$, so $z = \frac{\widehat{p} - p}{\sqrt{pq/n}} = \frac{0.574 - 0.5}{\sqrt{0.5 \cdot 0.5/1000}} \approx 4.68$. Placing this test stat on Z and shading the upper tail gives a P-value of pnorm(4.68, lower=F) $\approx 0 < 0.02$. So, we reject H_0 in favor of H_1. HFA might be a real thing.

(b) While the data suggest it is better to believe in H_1: $p > 0.5$, this does not *prove* H_1 is true. The phrase "Data **show** the existence of HFA" is too strong. Also, when you decide to believe H_1, this does not prove or support any particular theory underpinning H_1. If HFA is indeed real, it may be due to many factors: biased refereeing (unjustly cited as the "likely" reason), players giving extra effort to please home-town fans, players knowing their local arenas better, support of the home crowd, less exhaustion from traveling, and so on.

Commentary: Let's be really honest for a moment about how most scientists, researchers, and academics operate. It all begins with a theory based on some anecdotal evidence ("I think referees favor the home team"). This theory suggests something about a parameter from a statistical model ("$p > 0.5$, if I'm right"). So, they go collect a data set (usually crummy and non-representative) to try to prove (!) they're right. After a mix of honest statistical work and questionable research practices (QRPs), they arrive at the statistical conclusion they hoped for ("We reject H_0!"). Then, they claim this provides support for their original idea (it doesn't). Finally, they wrap all this in a series of research articles, popular books, media appearances, social media posts, podcast appearances, and sexy marketing lingo (e.g., "home-field advantage").

As you might guess, I am bothered by almost every step above. The most important thing you should remember is that *statistics is theory agnostic*. This means that the statistical test doesn't care why you originally wrote down H_1. All it tries to do is answer this question: Are the data we see (as compressed into a test stat) beyond what's reasonable to expect given the ever-presence of randomness? If so, we move to H_1; if not, we keep H_0. Because of this agnosticism, when you move to H_1, this does **not** lend support to the theory you had in mind when initially creating H_1. For example, it might actually be the case that referees *punish* the home team (because they're afraid of appearing biased), and this *lowers* the home-team win percentage from 50% to 48%. But then, maybe knowledge of the home arena and not traveling actually help the home team, raising the percentage from 48% to 57.4%. So, it's possible you correctly predicted which side of 0.5 that p will be on, but you gave credit to the wrong forces under the hood.

60. Americans spend about 5.4 trillion dollars per year on credit card transactions. Suppose that 0.4% (0.004, as a decimal) of these transactions are fraudulent, and hence, credit card companies lose money reimbursing their users. Researchers at Visa have designed a new algorithm to predict fraud and are curious if it has reduced fraudulent card usage. You are tasked with determining whether this claim is statistically reasonable using an approximate test with $\alpha = 0.03$.

    (a) What is the smallest number of transactions you could look at and meet Larsen & Marx's requirements for using the approximate test?

    (b) Suppose you end up looking at 2400 claims. What is the largest number of fraudulent claims that could appear among those 2400 claims that would cause a move to the alternative hypothesis?

Hint List:

- b. What test stat do we use when running a hypothesis test for a proportion? What place on the horizontal axis demarcates a left-tail area of 0.03?

Solution:

(a) Let p be the true percentage of fraudulent claims among **all** credit card transactions after the new algorithm is deployed. We test

$$H_0\text{: } p = 0.004 \text{ vs. } H_1\text{: } p < 0.004 \text{ (“reduced fraudulent card usage”)}$$

To run an approximate test, Larsen and Marx require that $0 \overset{(1)}{<} np - 3\sqrt{npq} \overset{(2)}{<} np + 3\sqrt{npq} \overset{(3)}{<} n$. Note that inequality (2) is always true, and here, inequality (3) is easily met because p is so small. We explore inequality (1) after assuming H_0:

$$3\sqrt{n}\sqrt{0.004 \cdot 0.996} < n \cdot 0.004 \implies \frac{3\sqrt{0.004 \cdot 0.996}}{0.004} < \sqrt{n} \implies 2241 < n$$

So, we round up to get $\boxed{n = 2242}$ (or 2241 if you use $\leq$ instead of $<$ in the Larsen and Marx inequalities).

(b) Since $2400 > 2242$, we can run the approximate test. Let x be the number of fraudulent claims among the 2400. So, $\widehat{p} = \dfrac{x}{2400}$ and $z = \dfrac{\dfrac{x}{2400} - 0.004}{\sqrt{0.004 \cdot 0.996/2400}}$. To reject H_0 in favor of H_1, we need our test stat z to land in the bottom $\alpha = 0.03$ strangest results on $Z \sim N(0, 1)$. We use R to find the quantile that demarcates the bottom 3% of test stats: qnorm(0.03) ≈ -1.880794. Thus, we solve:

$$\frac{\dfrac{x}{2400} - 0.004}{\sqrt{0.004 \cdot 0.996/2400}} < -1.880794 \implies x < 3.78$$

Since we are asked for the “largest number” that causes a move to H_1, we take x as large as possible, giving $\boxed{x = 3}$.

Commentary: This problem is non-standard in that you don't know the results from the sample, and yet, want to run the test anyway. In a typical problem, you work from the data set to the test stat to the P-value to the decision. Here, we work backwards, from the decision to the critical region to the possibilities for the data set. Great thinkers have the ability to think forward (drawing conclusions from givens) and think backward (“If we know we're ending here, then what must be true at the previous step? And the step before?”). Note that making progress in the forward and backward directions may not be *equally easy*. For example, taking a derivative (forward direction) is not especially hard after learning a few basic rules and some practice. Integration (backward direction) is much harder, and students struggle with techniques like partial fractions and trig substitution. Here are some if-and-only-if mathematical statements where the forward direction (left to right implication) is easier to prove than the backward direction (right to left implication):

$$a = 0 \text{ or } b = 0 \iff ab = 0$$

$p|a$ or $p|b \iff p|(a \cdot b)$ where p is prime, a and b are integers ($x|y$ means x divides evenly into y)

61. A real number is said to be "normal" if, when written in any base b, the digits $0, 1, 2, \ldots, b-1$ all appear with equal frequency. That is, there should be an equal proportion of 0s, and 1s, and 2s, etc. For example, in base 10, the number 1/3 is not normal because the decimal expansion is 0.3333..., and so the proportion of 3s is 100%, while the other digits appear 0% of the time. An example of a normal number in base 10 is 0.01234567890123456789012... Here, each digit appears 10% of the time, by construction.

    One of the strangest results in mathematics is this: It can be shown that nearly all real numbers are normal, and yet, proving any particular number is normal is very difficult. Indeed, mathematicians do not know if π or e are normal!

    On Canvas/TritonEd, you can find the file pibinary.csv which contains the first 10242 digits of pi in binary (base 2). Notice it starts with 11, which means 3 in binary. If π really is normal, what percentage of 1s do you expect in its binary expansion? Assuming the first 10242 digits are representative of all of π, conduct a hypothesis test on the proportion of 1s in π by randomly selecting 314 digits without replacement. You should set up variable(s) and hypotheses, write code to select the digits and get the sample proportion, draw a picture of a sampling distribution, shade an area, calculate a P-value, and reach a conclusion about your hypotheses using $\alpha = 0.05$.

Hint List:

- In binary, only two digits (0 and 1) are possible. If these are to appear with equal frequency, each should appear $1/2 = 50\%$ of the time.

Solution:

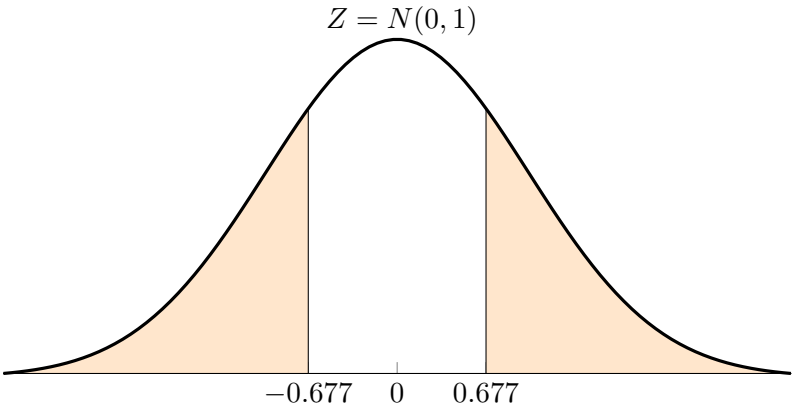
Let p be the true proportion of 1s in the binary expansion of π. We test

$$H_0\text{: } p = \frac{1}{2} \text{ vs. } H_1\text{: } p \neq \frac{1}{2} \text{ (not normal)}$$

We must use R to draw a sample and find the sample proportion, $\widehat{p}$:

```
p = 1/2  # Assume H_0
n = 314
pibinary = read.csv("location on your computer")
digits = pibinary$digits
oursample = sample(digits, n)
phat = mean(oursample)
z = (phat-p)/sqrt(p*(1-p)/n)
```

Since the sample command chooses random data when building the sample, you will get different results from one run to the next, and from one student to the next (unless we use set.seed to align our randomness). When I ran this code, I got $\widehat{p} \approx 0.481$ and $z \approx -0.677$. Given our two-sided alternative, we need to shade both tails on Z to find our P-value:



So, our P-value is $2 \cdot \text{pnorm}(-0.677) \approx 0.498 > 0.05$, so we keep H_0. π might be normal in base 2.

Commentary:

Math is full of unproven claims that are so complex you need a PhD to even understand what they are saying. For example, consider:

Yang-Mills Existence and Mass Gap Conjecture: For any compact simple gauge group G, a non-trivial quantum Yang-Mills theory exists on $\mathbb{R}^4$ and has mass gap $\Delta > 0$. (You can read more on the Wikipedia page, but it doesn't help much!)

Amazingly, math also has a few easy-to-state conjectures that are unproven. This problem is an

example:

Normality of Pi Conjecture: When the constant π is written in binary, each digit (0 through 1) appears $1/2$ of the time. Similarly, when π is written in base b, each digit (0 through $b-1$) appears $1/b$ of the time.

The interesting thing for me is that we can gather statistical evidence in support of this claim. Indeed, here the normality of π is built into the null hypothesis, so this is a rare example of a hypothesis test where you secretly are hoping to *keep* H_0 rather than reject H_0. Some other examples where you secretly are hoping to keep H_0: if you are measuring the rate of side effects from a new drug compared to a placebo, or if you are measuring the test scores of students after reducing resources to a school.

To be clear, the statistical evidence we gather here does not *prove* π is normal in base 2. The statistical test just says its more reasonable to believe in normality vs. non-normality based on our data and the randomness inherent in data collection. Indeed, the normality of π is an open question, meaning it hasn't been proved true or false as I type this.

In fact, we must be very careful to distinguish "strong trends" from "absolute truth". When you have lots and lots of data pointing in the same direction, what you have is a strong trend. For example, if you go out in nature and the first 1000 swans you encounter are white, you have strong trend. But you can't yet claim that "All swans are white" (an absolute truth because of the word "all"). Indeed, it's possible the 1001st swan you see will be black!

To clarify the distinction between strong trends and absolute truths, here's an example that I first saw on Project Euler. This is a programming challenge website that focuses on math and statistics puzzles:

Goldbach's Other Conjecture: Every odd composite number (> 1) can be written as the sum of a prime and twice a square.

If you start collecting data on this conjecture, you'll notice that $9 = 7 + 2 \cdot 1^2$ and $15 = 7 + 2 \cdot 2^2$ and $21 = 3 + 2 \cdot 3^2$. Each of these shows an odd composite number (on the left of the $=$) equaling a prime plus double a square (on the right). In fact, you could keep going through all the odd composite numbers, and you would have hundreds of examples that all support the conjecture (strong trend). But, it turns out the conjecture is false (not an absolute truth!), and the goal of the Project Euler problem is to write code that finds the smallest counterexample (the black swan!). Can you do it?

62. Suppose we are studying μ in $X \sim N(\mu, \sigma^2)$ where σ is known. What hypothesis test has duality with the CI $\left(-\infty, \overline{X} + 1.3\frac{\sigma}{\sqrt{n}}\right)$?

Hint List:

- Rewrite the statement $\mu_0 \in \left(-\infty, \overline{X} + 1.3\frac{\sigma}{\sqrt{n}}\right)$ using inequalities and try to rearrange it so you see the test stat used in hypothesis testing for means. Then, think.

Solution:

The CI is trying to tell us what values for μ make the most sense. So, if $\mu_0 \in \left(-\infty, \overline{X} + 1.3\frac{\sigma}{\sqrt{n}}\right)$, this is a value of μ we might trust. Unraveling this statement, we find

$$\mu_0 \in \left(-\infty, \overline{X} + 1.3\frac{\sigma}{\sqrt{n}}\right) \iff \mu_0 < \overline{X} + 1.3\frac{\sigma}{\sqrt{n}} \iff -1.3\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu_0 \iff -1.3 < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = z$$

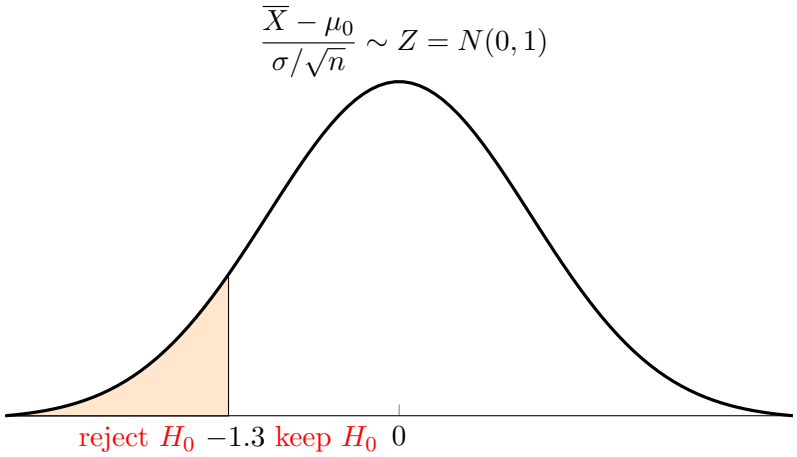
So, μ_0 is reasonable $\iff$ $z > -1.3$. We can switch this to the language of hypothesis testing:

$$H_0\text{: } \mu = \mu_0 \text{ should be kept} \iff z > -1.3$$

Or, equivalently:

$$H_0 \text{ should be rejected} \iff z < -1.3$$

Note that $P(Z = -1.3) = 0$, so we can ignore this case. Thus, we have the below picture:



Since pnorm$(-1.3) \approx 0.0968$, we could create this dual test:

$$H_0\text{: } \mu = \mu_0 \text{ vs. } H_1\text{: } \mu < \mu_0 \text{ where } \alpha = 0.0968$$

Commentary:

The most important result of this problem is that the one-sided CI we were given ultimately has duality with a HT that uses a one-sided alternative hypothesis. So, we can remember: A two-sided (normal) CI has duality with a two-sided alternative in a HT; a one-sided CI pairs with a one-sided H_1. Great, history got its language correct!

The next thing to note is that because there is always a duality between CIs and HTs, this implies we don't ever need to run both. The duality means that these two analytic frameworks are trying

to tell us the same thing. Essentially, they communicate the same information but say it in different ways (like saying "heads" or "not tails"). In settings more complicated than exploring the mean in a single population, you might find that there is no known duality or that one of the frameworks (HT vs. CI) is more commonly used. You may also discover that certain communities or publications prefer one framework over the other. In cases where a duality is known, these preferences are not because one system carries more analytic value than the other. Rather, they might be the result of historical precedent or the ease of understanding one of the frameworks.

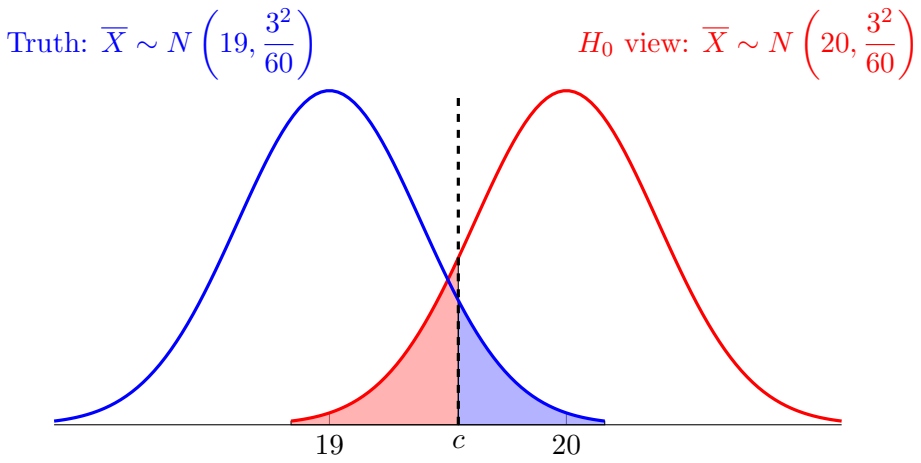
63. Suppose you are studying a phenomenon that is well-modeled by $N(\mu, 3^2)$. Existing research claims that $\mu = 20$, but you think μ might be lower based on recent changes in society. You'd like to collect some data to support your claim and plan to use a sample of size 60. If μ is actually 19, what should you set α to in your hypothesis test if you want a Type II error rate of 0.08? Include a beautiful picture in your answer.

Hint List:

- If you assume H_0 is true, what do you believe is the distribution of $\overline{X}$?
- Draw a single picture with two versions of $\overline{X}$: what you believe it is based on H_0, and what it actually is based on the truth of the universe.

Solution:

Let $X \sim N(\mu, 3^2)$ and test H_0: $\mu = 20$ vs. H_1: $\mu < 20$ ("might be lower"). We know $n = 60$ and μ is actually 19. When we assume H_0 is true, we expect an average, $\overline{x}$, from the distribution $\overline{X} \sim N\left(20, \frac{3^2}{60}\right)$ (in red below), but the universe gives us an average, $\overline{x}$, from $\overline{X} \sim N\left(19, \frac{3^2}{60}\right)$ (in blue). Because the alternative hypothesis uses a less-than sign (H_1: $\mu < 20$), we are looking for values in the left tail (red shaded area) to reject H_0. So, we have this picture:



Our blue shaded area is the Type II error rate, $\beta = 0.08$. From this, we find the rejection fence:

```
c = qnorm(0.08, 19, 3/sqrt(60), lower=F)    # about 19.54418
```

We use c to find the Type I error rate, α (the red shaded area):

```
alpha = pnorm(19.54418, 20, 3/sqrt(60))  # about 0.120
```

So, we have $\boxed{\alpha \approx 0.120}$.

Commentary:

Here we have another problem that requires using multiple semiotics: written language, mathematical notation, visualization, and computer code. Which of these four is most difficult for you, and what steps are you taking to improve your fluency in this semiotic?

64. I recently read Pete Buttigieg's book "Shortest Way Home", a beautiful autobiography discussing his life growing up and time as mayor of South Bend, Indiana. While most people thought the book would focus on his sexuality (he was one of the first openly-gay men to run for the US presidency), instead he spends most of the book discussing the hard challenges he faced as mayor. In one section, he describes implementing a technology known as Shotspotter, which listens for gun shots in communities and automatically dispatches police when it believes a shot has been fired. Naturally, the technology makes mistakes because slamming car doors and dropped objects might sound like a gun shot. Suppose you work at Shotspotter and run a bunch of tests to determine the accuracy of your technology, getting the below table.

	You fire a gun	You create gun-like sounds in creative ways
Shotspotter says "Bullet fired"	413	32
Shotspotter says "No bullet fired"	46	312

Suppose we set the null hypothesis H_0: "No bullet was fired", since most of the sounds we hear each day are not guns firing bullets. Describe the alternative hypothesis in words for this setup, explain what Type I and II errors would mean in this context, and discuss the consequence of making each type of error if this technology were used in an actual city. Then, find the Type I and II error rates in this data set. Finally, discuss which type of error you *personally* think is worse *in your hometown* and explain why.

Hint List:

- The best starting point is memorizing the general definitions of Type I and Type II error.

Solution:

We have H_0: "No bullet was fired" and H_1: "A bullet was fired". A Type I error (in general) happens when H_0 is true, but the data lead us to H_1. In the context of this problem, that means: No bullet was fired, but Shotspotter is mistaken and believes a bullet was fired. Some consequences of making this error include: wasting resources addressing a non-threat and potential misunderstanding of the danger present in a community.

A Type II error (in general) happens when H_1 is true, but our data lead us to keep H_0. In the context of this problem, this means that a bullet actually was fired, but Shotspotter misses this and thinks the noise was not a shot. Some consequences of making this error include: not addressing a potential danger and misgauging the danger of a community.

In this problem, we can actually calculate the probability of making each type of error:

$$P(\text{Type I error}) = P(\text{Shotspotter says bullet fired} \mid \text{no bullet fired}) = \frac{32}{32+312} \approx \boxed{0.093}.$$

$$P(\text{Type II error}) = P(\text{Shotspotter says no bullet fired} \mid \text{bullet fired}) = \frac{46}{413+46} \approx \boxed{0.100}.$$

As for which type of error is worse, answers will vary (see Commentary).

Commentary:

Students sometimes get frustrated when they are asked a question where there is not a single, clear answer which can appear in the solutions. Here, I have asked you to decide which type of error is worse, and both Type I and Type II may seem like reasonable answers. If you are running a city and don't have enough resources to get by, then a Type I error is really frustrating because you likely send the police to a situation that doesn't need it, wasting resources. If you prioritize reducing crime and saving lives, then a Type II error is very problematic: you've missed the chance to identify a criminal and/or save a life if someone was shot.

In the modern age, I feel that we spend a lot of time angry that people are giving different answers. "You support Proposition X? Well, I oppose it, you fool!" Remember, though, that what you ultimately support is a reflection of your deeper belief structures. Here, we support Type I or II because we place different importance on ideas like resource management, crime reduction, and saving lives. Virtually every complex decision in life is a trade-off among these deeper beliefs. For example, during the COVID pandemic, most school districts in the US closed down for an extended period, moving to online instruction. Without focusing on whether you like this choice or not, spend some time thinking about what things are improved because of this action, and what things are worsened. This thought experiment is a useful principle when exploring any tough decision in life. Instead of obsessing over what answer someone gives (and populating your life with people who believe the same answers), look at each possible answer and try to understand the trade-offs and belief structures that underpin them.

65. Congratulations! You've just gotten a job at the most popular museum in America: The Air and Space Museum in Washington DC. Your first task as the resident statistician is to decide if a recent exhibit change has increased the average number of visitors per day. Before the change, the number of visitors per day was $N(24000, 2000^2)$. Your plan is to check attendance numbers on n random days in the next year, but need n to be as small as possible because it is costly and intrusive to count visitors. Your boss would be excited by a new daily average of 25000 (assume the spread is unchanged by the exhibit) and wants you to use $\alpha = 0.04$ in your test. If you demand a power of (at least) 0.85, what sample size should you use? You MUST include a picture with your answer. Feel free to use R/calculator to do some of the calculations. If you do, include your code/commands.

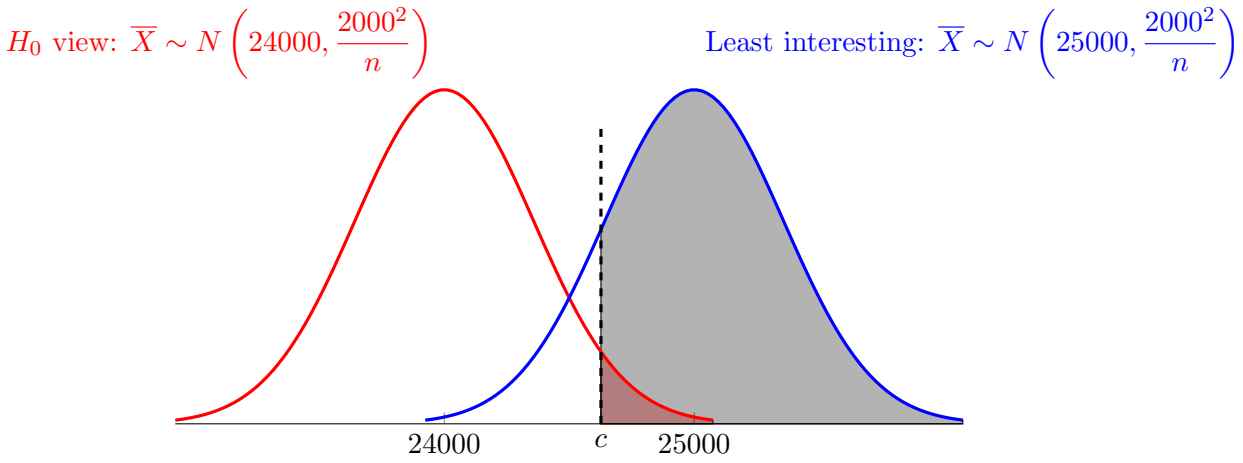
Hint List:

- If you assume H_0 is true, what do you believe is the distribution of $\overline{X}$?
- Draw a single picture with two versions of $\overline{X}$: what you believe it is based on H_0 and what it is in the least interesting analysis, as specified by your boss.

Solution: Let μ be the new average visitor count per day for the Air and Space Museum where $X \sim N(\mu, 2000^2)$ is the new visitor count on a random day. We test

$$H_0\text{: } \mu = 24000 \text{ vs. } H_1\text{: } \mu > 24000 \text{ (“increased the average”)}$$

Here, your boss has stated the least interesting new μ value is 25000, which we use as the center of the “least interesting” distribution:



On $Z \sim N(0,1)$, the top $\alpha = 0.04$ area is demarcated by $z_{0.04} = \text{qnorm}(0.04, \text{lower=F}) \approx 1.750686$. So, looking at the red shaded area, we have

$$\frac{c - 24000}{2000/\sqrt{n}} = 1.750686 \implies c = 24000 + 1.750686 \cdot \frac{2000}{\sqrt{n}}$$

Similarly, on Z, we find $z_{0.85} = \text{qnorm}(0.85, \text{lower=F}) \approx -1.036433$. Looking at the black area, we know

$$\frac{c - 25000}{2000/\sqrt{n}} = -1.036433 \implies c = 25000 - 1.036433 \cdot \frac{2000}{\sqrt{n}}$$

Setting our two equations for c equal, we have:

$$24000 + 1.750686 \cdot \frac{2000}{\sqrt{n}} = 25000 - 1.036433 \cdot \frac{2000}{\sqrt{n}} \implies \frac{5574.238}{\sqrt{n}} = 1000 \implies 5.574238 = \sqrt{n}$$

Squaring gives $n \approx 31.0721$. Since n must be an integer, we round up to ensure we meet (and exceed) the requirements for α and power: $\boxed{n = 32}$.

Commentary: You might be surprised by the smallness of this answer. It only takes 32 participants to run our study which has great Type I and II error rates? As mentioned in previous commentaries,

when you doubt an answer, try to resolve the doubt yourself, rather than seeking an external source. Here, the difference between what we think is going on (24000) and what gets our boss excited (25000) is fairly large, especially when you note that the SD of X is 2000, and hence, the SD of $\overline{X}$ is $\frac{2000}{\sqrt{n}}$ As n gets larger, the two normal distributions above really start to separate given the distance between their means and their small SDs. This separation allows for high power without a huge sample size.

This problem is an example of a general pattern in statistics (and life): It's easy to detect large changes, and hard to detect small changes. The simple reason for this is that large changes lead to distributions (see red and blue above) that have little overlap, and hence, the randomness present in all data sets has less of a chance to hide the true change from us. When a small change is present, it's difficult to sort out if a small change we see is a real effect or just randomness causing trouble. An analogy: If you're given two coins, one with $p_{\text{heads}} = 0.1$ and the other with $p_{\text{heads}} = 0.9$, it doesn't take many flips to be convinced the coins are different. If, however, the coins have $p_{\text{heads}} = 0.49$ and $p_{\text{heads}} = 0.51$, uncovering this difference takes a huge number of flips. While randomness is present in both scenarios, its influence is muted in the first because of the large p_{heads} difference in the coins.

Some scholars take this idea one step further. They argue that whenever you compare a parameter across two populations (e.g., the percentage of people that experience relief from a drug versus a placebo), it's *always* the case that the parameter value differs in the two populations (if even by a tiny amount) since everything in life affects everything else. That is, they argue the null hypothesis is **never actually true**. As such, you always have a real effect in front of you, and whether you detect it is really just a function of your sample size, which is just a function of the resources behind your study. The most cynical people with this viewpoint will say "Show me a statistically significant result, and all you've done is shown me a study of a large parameter difference and small sample size, or a small parameter difference and large sample size."

66. Every 4 years (or so), we get a leap day (2/29). This raises an interesting question: If you're born on leap day (a "leap-day baby") but we're in a non-leap year, would you rather celebrate your birthday on 2/28 or 3/1? On 2/29/2020, I was listening to NPR and a guest claimed that leap-day babies opt for the two options in equal proportion. Naturally, I doubted this (I wasn't sure if the preference would be for 2/28 or 3/1), and so let's think about a study you could conduct. Imagine we'll ask 500 random leap-day babies which day they prefer and record the percentage that choose 2/28. If the true percentage that opt for 2/28 is 49%, find the Type II error rate and power of our test assuming we use a significance level of 0.10. You must define a parameter, write hypotheses, and include a beautiful picture in your answer.

Hint List:

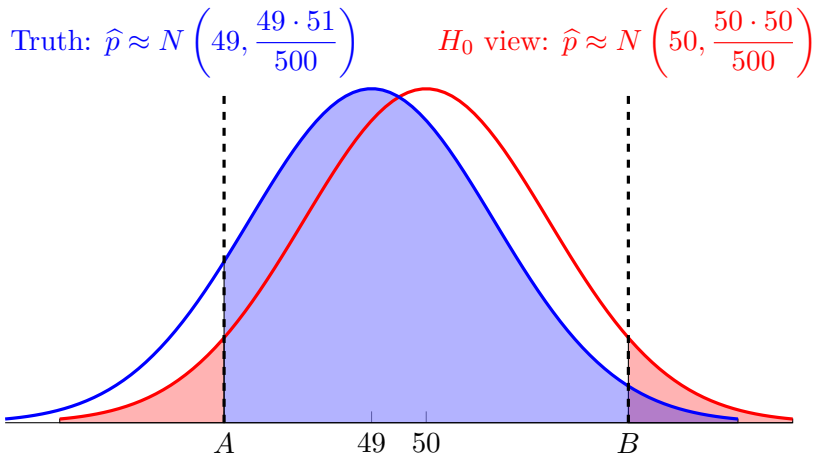
- Recall that $\widehat{p} \approx N\left(p, \frac{pq}{n}\right)$. Draw this curve for H_0 and for the true value of p.

Solution:

Let p be the percentage of **all** leap day babies that answer 2/28 when asked if they prefer 2/28 or 3/1 for birthday celebrations in non-leap years. We test

$$H_0\text{: } p = 50\% \text{ vs. } H_1\text{: } p \neq 50\% \text{ (“wasn't sure if the preference would be”)}$$

Recall that $\widehat{p} \approx N\left(p, \frac{pq}{n}\right)$. We assume H_0 to draw the red distribution below and include the truth of the universe in blue:



Since $\alpha = 0.1$, this means each red shaded area is 0.05, so:

$$A = \text{qnorm(0.05, 50, sqrt(50*50/500))} \approx 46.322$$
$$B = 50 + (50 - 46.322) = 53.678 \text{ (by symmetry, or use qnorm)}$$

So, $P(\text{Type II error}) = \text{pnorm(53.678, 49, sqrt(49*51/500))} - \text{pnorm(46.322, 49, sqrt(49*51/500))} \approx \boxed{0.866}$. Thus, power $= 1 - 0.866 = \boxed{0.134}$.

Commentary:

You might be concerned about your answers. A Type II error rate of 0.866? A power of 0.134?! As previously discussed, when you get an answer that seems off, spend time trying to make sense of it. Here, we are trying to detect a 1% change (from our belief of 50% to the truth of 49%) using only 500 people. This may seem like a large sample, but it simply isn't large enough to detect such a small difference. For example, in 100 coin flips, you can move from a 50-50 even split to a 49-51 split by changing the result of a single flip. For 500 coins, only 5 coins must be flipped to switch from 50% to 49%. Visually, if you look at the picture above (which is drawn to scale), you'll notice that the distributions have a lot of overlap; this is what causes the high Type II error rate. To get separation in these curves, we need either (1) the true value (49%) and hypothesized value (50%) of p to be more different, which we have no control over, or (2) a larger sample size to decrease the variance $\left(\frac{pq}{n}\right)$ of the curves, thus creating less overlap.

67. This problem is inspired by one of my best students, who went on to get his PhD in materials science at MIT and his MD from Columbia. If you look at car tires, they tend to average about 30000 miles before being replaced. In general, tire lifetimes are known to be normally distributed with SD 4500 miles. Suppose my former student claims to have a new manufacturing process that raises this average (he was right!) and keeps the spread the same. Let μ be the true average lifespan of tires with the new manufacturing process. When you draw a sample of size n, the power of this HT is 0.2 when $\mu = 32000$, and the power is 0.85176 when $\mu = 35000$. Find α and n for this HT.

Hint List:

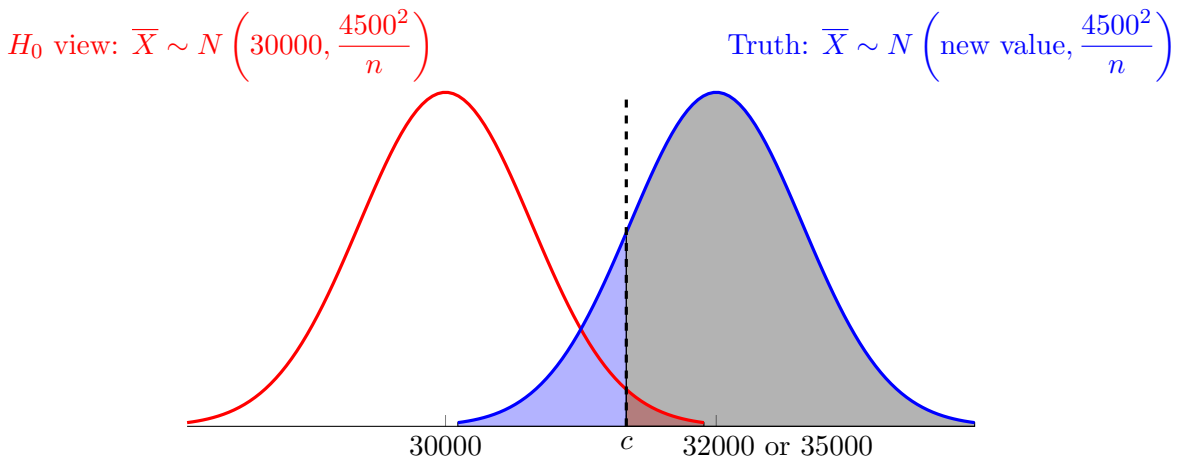
- As ever, a well-labeled picture is your main hope moving forward. Try using each given power (and associated new mean) to write an equation for the rejection fence.

Solution:

Let $X \sim N(\mu, 4500^2)$ be the lifetime of tires with the new manufacturing process. We test

$$H_0\text{: } \mu = 30000 \text{ vs. } H_1\text{: } \mu > 30000 \text{ (“raises this average”)}$$

We can draw this picture:



Here, the red shaded area is α, the blue shaded area is β, and the black shaded area is the power (0.2 or 0.85176, depending on the new value for μ).

When $\mu = 32000$, $\beta = 1 - \text{power} = 1 - 0.2 = 0.8$. So, $\dfrac{c - 32000}{4500/\sqrt{n}} = \text{qnorm}(0.8) \approx 0.8416212$.

When $\mu = 35000$, $\beta = 1 - \text{power} = 0.14824$. So, $\dfrac{c - 35000}{4500/\sqrt{n}} = \text{qnorm}(0.14824) \approx -1.044012$.

(Note that $\mu = 35000$ is the value I actually used for the above picture.)

Solving each equation for c and setting these equal to one another gives $\boxed{n = 8}$ after some algebra.

If we let $n = 8$ in either equation above, we find $c \approx 33339.01113$, and thus

$$\alpha = \text{pnorm}(33339.01113,\ 30000,\ 4500/\text{sqrt}(8),\ \text{lower=F}) \approx \boxed{0.018}$$

Commentary:

One of the questions I researched early in my career was this: What are all the common ways to alter an existing math problem to create a new one? As a student, you may have noticed that some teachers take problems from the homework and simply *change the numbers* when making exam problems. This is the most basic of the “transformation techniques”. A more complex technique is to *change the knowns and unknowns*. For example, here is what is known and unknown in the last few problems (c stands for the location(s) of the rejection fence(s)):

Problem	Knowns	Unknowns
Space Museum	New value for parameter, α, power	c, n
Leap Year	New value for parameter, α, n	$c, \beta,$ power
Tire Lifetimes	New values for parameter, powers	c, n, α

One of the interesting things about this transformation technique is that altering what is known and unknown can dramatically change the difficulty of a problem. Indeed, the three problems mentioned above may not have felt equally hard to you. One reason for this difference is that changing knowns and unknowns can alter the direction you take with a procedure. Usually, procedures flow in a certain direction and to use them backward feels awkward. For example, if I give you a quadratic equation, you probably feel great going in the "forward direction" to find its roots. If, however, I give you the roots and ask for a quadratic equation with those roots, this "backward direction" is more awkward (and, there are multiple possible answers!).

68. Some people love sending tweets on X. Here, we study the number of tweets that Donald Trump sends on a random day using $X \sim Poisson(\lambda)$. You hear a claim on a podcast that his average tweet rate is 8 tweets/day, but you believe it might be lower. You plan to collect a sample of size 1 and reject H_0 if $X_1 \leq 3$. Find the Type I Error rate and the Type II Error rate if $\lambda = 6.4$. On this problem, you must write hypotheses and include a beautifully-labeled diagram in your answer with two pmfs and a rejection fence. Include your code if you use R to create the diagram.

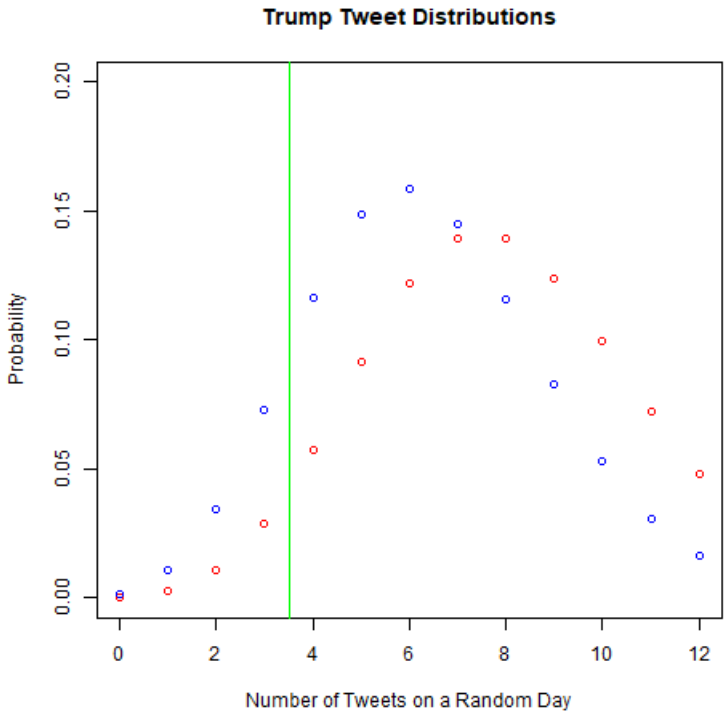
Hint List:

- Because the RV is discrete, our picture has two pmfs, not pdfs. You show this via plotting points, not graphing curves.
- In R, use plot(...) to plot points. Then call points(...) to add more points to the existing plot.

Solution: Let λ be Trump's true average tweet rate per day on X. We test

$$H_0\text{: } \lambda = 8 \text{ vs. } H_1\text{: } \lambda < 8 \text{ (“might be lower”)}$$

After assuming H_0, we believe $\lambda = 8$, and so our test stat has distribution $X_1 \sim X \sim Poisson(\lambda = 8)$ (red pmf below). When we go measure his tweet count on a random day, however, the universe gives us data based on $X_1 \sim Poisson(\lambda = 6.4)$ (blue pmf below).




```
xvals = 0:12  # show some of the pmf
plot(xvals, dpois(xvals, lambda=8), col="red", ylim=c(0, 0.2),
     ylab="Probability", main="Trump Tweet Distributions",
     xlab="Number of Tweets on a Random Day")
points(xvals, dpois(xvals, lambda=6.4), col="blue")
abline(v=3.5, col="green")
```

Now, we find the error rates:

Type I error rate $= P(\text{reject } H_0 \mid H_0 \text{ true}) = P(X_1 \leq 3 \mid \lambda = 8) = \text{ppois}(3, 8) \approx \boxed{0.042}$.

Type II error rate $= P(\text{keep } H_0 \mid H_1 \text{ true}) = P(X_1 > 3 \mid \lambda = 6.4) = \text{ppois}(3, 6.4, \text{lower=F}) \approx \boxed{0.881}$.

Commentary:

It's easy to complete a problem without thinking about what it's trying to tell us. Here, we have a great Type I error rate ($\alpha = 0.042$), but a terrible Type II error rate ($\beta = 0.881$). In essence, we have a really bad hypothesis test that we should never run. At this point, you should ask yourself: Why is this happening? One reason is that our test statistic, X_1, is based on a single piece of data. If you think one piece of data can do a good job of distinguishing the red and blue plots above, then think again. If the red and blue plots were really different (e.g., if $\lambda = 8$ (red) and $\lambda = 30$ (blue)), then sure, one piece of data is useful. Another option is to use more than one piece of data: $X_1, \ldots, X_n$. Now the issue is: How do you collapse those n pieces of data into a single number *whose distribution is known*? This last point is critical. In this problem I artificially chose the test stat $T = X_1$ because $X_1 \sim X \sim Pois(\lambda)$. With n pieces of data, you might try $T = \sum_{i=1}^{n} X_i$ or $T = \overline{X}$ or $T = X_{(1)}$ or $T = X_{(n)}$. If any of these has a nice distribution (do they?), you could make progress and (most likely) get better Type I/II error rates.

69. Suppose $X \sim Exp(\lambda)$. You draw a random sample of size n and plan to use the statistic $X_{\min}$ to decide between two hypotheses about λ:

$$H_0\text{: } \lambda = 2 \text{ vs. } H_1\text{: } \lambda > 2$$

You plan to test via the rule: If $X_{\min} < c$, reject H_0; else keep H_0. Recall from Problem 27 that $X_{\min} \sim Exp(n\lambda)$.

(a) Typically, when the alternative hypothesis (here, H_1: $\lambda > 2$) has a $>$ sign, the decision rule also uses a $>$ sign. In this problem, things are reversed: "If $X_{\min} < c$ ...". Using a picture, explain why *larger* values for λ are more likely to result in *smaller* values for $X_{\min}$ (and hence, why the inequalities make sense as given).

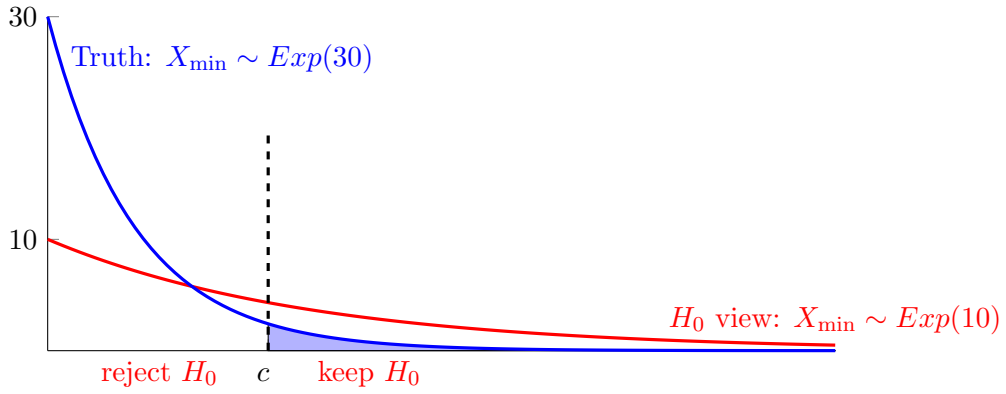
(b) What must c be if you want your Type II error rate to be 0.08 when the true value of λ is 6? Assume that $n = 5$.

Hint List:

- a. Draw the pdf of $X_{\min}$ with the values in part (b) for both H_0 and the true value of λ. What does the height of the pdf convey?

Solution:

(a) While we could explore this question in general, it is easiest to use the values in part (b) so we can draw a single picture for both parts. Below is a diagram showing the pdf of the test statistic ($X_{\min}$) under the null hypothesis ($\lambda = 2$) and using the true value ($\lambda = 6$) when $n = 5$. These distributions are $Exp(n\lambda) = Exp(5 \cdot 2) = Exp(10)$ and $Exp(5 \cdot 6) = Exp(30)$, respectively.



If you look at the two pdfs, the blue one wants to give small values on the horizontal axis (because the graph is higher there). Thus, if we're getting small values of our test stat, this is a sign the blue curve (i.e., alternative hypothesis) is our reality. This is why the test rule reads $X_{\min} < c$.

(b) Now, we know the Type II error rate is $0.08 = P(\text{keep } H_0 \mid H_1 \text{ is true}) = P(X_{\min} \geq c \mid \lambda = 6)$

Evaluating this gives $0.08 = \displaystyle\int_c^{\infty} 30e^{-30x}\,dx = -e^{-30x}\Big|_c^{\infty} = e^{-30c} \implies c = \boxed{\frac{-\ln 0.08}{30} \approx 0.084}$.

Commentary:

The main reason I assign this problem is that it is different than what we've done previously. As such, you start to see the high level structure of how hypothesis testing works. In general, we compress a data set via a test stat. In the past, these stats were $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$ and $\widehat{p} = \dfrac{x}{n}$; currently, we are using $X_{\min}$. After that, we find the distribution of the test stat ($Exp(n\lambda)$, here), and then draw both versions of the test stat pdf (the distribution we think it follows after assuming H_0 and the truth of the universe). Finally, we do what we need to (e.g., find a P-value, calculate the Type II error rate, etc.). As you are studying different ideas in school, try to look beyond the surface level to find patterns in the underlying deep structures.

70. Let $Y \sim \chi^2_n$.

(a) Show that $E[Y^k] = \dfrac{2^k \cdot \Gamma\left(\dfrac{n}{2} + k\right)}{\Gamma\left(\dfrac{n}{2}\right)}$ if k is a real number with $k > -\dfrac{n}{2}$.

(b) Using part (a), prove that $E[Y] = n$ and $Var[Y] = 2n$, as claimed in class.

(c) Using part (a), prove that $E[T_n] = 0$ and $Var[T_n] = \dfrac{n}{n-2}$ when $n > 2$.

Hint List:

- a. Recall that $\chi^2_n \sim Gamma\left(r = \frac{n}{2}, \lambda = \frac{1}{2}\right)$.
- c. Think of T_n as a *product* of two RVs.

Solution:

(a) Using the hint:

$$E[Y^k] = \int_0^\infty y^k \cdot \frac{\left(\frac{1}{2}\right)^{n/2}}{\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} e^{-y/2}\, dy = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \int_0^\infty y^{k+\frac{n}{2}-1} e^{-\frac{1}{2}y}\, dy$$
$$\overset{\text{KT}}{=} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \cdot \frac{\Gamma\left(k+\frac{n}{2}\right)}{\left(\frac{1}{2}\right)^{k+\frac{n}{2}}} = \frac{2^k \cdot \Gamma\left(k+\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

When we use the Kernel Technique (KT), it is with $Gamma\left(r = k + \frac{n}{2}, \lambda = \frac{1}{2}\right)$. Since $r > 0$ in the definition of Gamma, this means that $k + \frac{n}{2} > 0$, or $k > -\frac{n}{2}$, as the problem mentions.

(b) Letting $k = 1$ and $k = 2$ in part (a) gives us:

$$E[Y] = E[Y^1] = \frac{2^1 \cdot \Gamma\left(1+\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \frac{2 \cdot \frac{n}{2} \cdot \Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = n$$
$$E[Y^2] = \frac{2^2 \cdot \Gamma\left(2+\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \frac{4 \cdot \left(1+\frac{n}{2}\right) \cdot \frac{n}{2} \cdot \Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = n^2 + 2n$$

Above, we use the fact that $\Gamma(x+1) = x \cdot \Gamma(x)$ to simplify the terms involving the Gamma function.

Now, our first equation shows $E[Y] = n$. When both equations are used, we see

$$Var[Y] = E[Y^2] - E[Y]^2 = (n^2 + 2n) - (n)^2 = 2n$$

(c) Write $T_n = \frac{Z}{\sqrt{\chi^2_n/n}} = \sqrt{n} \cdot Z \cdot \left(\chi^2_n\right)^{-1/2}$ where Z and χ^2_n are independent. Recall that if A and B are independent RVs, then $E[A \cdot B] = E[A] \cdot E[B]$. Thus, we have:

$$E[T_n] = \sqrt{n}E[Z]E\left[\left(\chi^2_n\right)^{-1/2}\right] = 0, \text{ since } E[Z] = 0$$

$$E[(T_n)^2] = nE[Z^2]E\left[\left(\chi^2_n\right)^{-1}\right] \overset{\bigstar}{=} n\left[Var(Z) + E(Z)^2\right] \cdot \frac{2^{-1}\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{n}{2}\right)}$$

$$= n[1+0^2]\frac{2^{-1}\Gamma\left(\frac{n}{2}-1\right)}{\left(\frac{n}{2}-1\right)\Gamma\left(\frac{n}{2}-1\right)} = \frac{n}{2\left(\frac{n}{2}-1\right)} = \frac{n}{n-2}$$

So, $Var[T_n] = E[T_n^2] - E[T_n]^2 = \dfrac{n}{n-2} - 0^2 = \dfrac{n}{n-2}$, as claimed.

Note that at $\bigstar$, we need $-1 > -\dfrac{n}{2}$ to use part (a). This is equivalent to $n > 2$, which is given in the problem.

Commentary: I assign this problem for two reasons. First, the problem is very algebraically rewarding. Despite the complexity of the RVs we are working with, the algebra leads us to nice answers at the end. I, like most people, enjoy when this happens. Note, however, that this emotion is just an artifact of the education system. Beginning in early grades, we give students problems that come out nicely. As noted in other commentaries, real-world answers are rarely nice. So, for now, bask in the academic narrative that answers are nice, but expect carnage ahead.

Second, this problem reveals just how important the Gamma distribution is. Part (a) shows that the moments of chi-squared, $E\left[\left(\chi^2_n\right)^k\right]$, can be found via the Kernel Technique with the Gamma distribution. In parts (b) and (c), we find that the important facts about χ^2_n and T_n rest on part (a). Because of this, the Gamma distribution is central to the Mathematical Statistics architecture.

Finally, you may have noticed that part (c) only considers $n > 2$. This raises a natural question: What are the expectation and variance of T_1 and T_2? Several previous commentaries have promoted the idea of being academically curious. Here, you have another opportunity to develop that trait. Don't just ask AI to give you the answers, try to discover them *yourself*.

71. Let X be a normal distribution with equal mean and standard deviation: $X \sim N(\theta, \theta^2)$. Assume $\theta > 0$ and let $X_1, \ldots, X_n$ be iid from X.

(a) Use AI to find a real-world setting where this type of setup might make sense.

(b) Find the Fisher Information (FI) and Cramer-Rao lower bound (CRLB) for the variance of an unbiased estimator of θ.

(c) Since θ is the mean of X, it is natural to estimate θ using $\widehat{\theta}_1 = \overline{X}$. Determine if this estimator is efficient.

(d) Since θ is the standard deviation of our normal distribution, it is natural to estimate θ by using S, the sample standard deviation. Recall that the sample variance, S^2, is unbiased for the population variance of a normal distribution (here: θ^2). Surprising, $S = \sqrt{S^2}$ is a *biased* estimator for the population standard deviation (here: θ). Let $\widehat{\theta}_2 = cS$. Find c so that $\widehat{\theta}_2$ is unbiased.

Hint List:

- b. Avoid nightmarish derivatives by writing $\dfrac{(X-\theta)^2}{\theta^2}$ as $\left(\dfrac{X}{\theta}-1\right)^2$ before differentiating.
- d. The original formula for S^2 is usually too difficult to work with. If you multiply S^2 by some particular constants, however, you can get a simple distribution.
- d. Use the result from Problem 70a.

Solution:

(a) Based on AI searches, such a model might be appropriate for the time estimate to complete certain types of projects. When projects are quick (small values of θ) like writing an email, the amount of variability might also be small. For long projects (large values of θ) like remodeling a kitchen, the variability might be much larger due to unforeseen problems, difficulty sourcing materials/labor, etc. Answers will vary; be skeptical of all AI claims.

(b) Given that both the mean and variance depend on θ, the pdf is simpler than usual:

$$f(x;\theta)=\frac{1}{\theta\sqrt{2\pi}}\exp\left[-\frac{1}{2}\cdot\frac{(x-\theta)^2}{\theta^2}\right]$$

Thus, $\ln f(X;\theta)=-\dfrac{1}{2}\ln(2\pi)-\ln\theta-\dfrac{1}{2}\dfrac{(X-\theta)^2}{\theta^2}=C-\ln\theta-\dfrac{1}{2}\left(\dfrac{X}{\theta}-1\right)^2$.

So, $\dfrac{\partial \ln f}{\partial\theta}=-\dfrac{1}{\theta}-\dfrac{1}{2}\cdot 2\left(\dfrac{X}{\theta}-1\right)\left(\dfrac{-X}{\theta^2}\right)=\dfrac{-1}{\theta}+\dfrac{X^2}{\theta^3}-\dfrac{X}{\theta^2}$.

Thus, $\dfrac{\partial^2 \ln f}{\partial\theta^2}=\dfrac{1}{\theta^2}-\dfrac{3X^2}{\theta^4}+\dfrac{2X}{\theta^3}\implies I(\theta)=\dfrac{-1}{\theta^2}+\dfrac{3}{\theta^4}E[X^2]-\dfrac{2}{\theta^3}E[X]$.

Recalling that $E[X^2]=Var[X]+E[X]^2$ and $X\sim N(\theta,\theta^2)$, we get

$$I(\theta)=\frac{-1}{\theta^2}+\frac{3}{\theta^4}\left(\theta^2+\theta^2\right)-\frac{2}{\theta^3}\theta=\frac{3}{\theta^2}$$

Thus, CRLB $=[nI(\theta)]^{-1}=\boxed{\dfrac{\theta^2}{3n}}$.

(c) For an estimator to be efficient, it must first be unbiased, and then it must achieve the CRLB. The unbiasedness is immediate:

$$E[\widehat{\theta}_1]=E[\,\overline{X}\,]=E[X]=\theta$$

To explore efficiency, we find

$$Var[\widehat{\theta}_1]=Var[\,\overline{X}\,]=\frac{Var[X]}{n}=\frac{\theta^2}{n}\neq\text{CRLB}$$

So, $\widehat{\theta}_2$ is $\boxed{\text{not efficient}}$.

(d) The original formula for $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is too difficult to work with. Luckily, we now know that $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, when X is normally distributed. Thus,

$$E[\widehat{\theta}_2] = E[cS] = cE[\sqrt{S^2}\,] = \frac{c\sqrt{\theta^2}}{\sqrt{n-1}}E\left[\sqrt{\frac{(n-1)S^2}{\theta^2}}\right] = \frac{c\theta}{\sqrt{n-1}}E\left[\sqrt{\chi^2_{n-1}}\right]$$

Using Problem 70a with $k = \dfrac{1}{2}$, we get

$$E[\widehat{\theta}_2] = \frac{c\,\theta}{\sqrt{n-1}}\,\frac{2^{1/2}\Gamma\left(\dfrac{n-1}{2}+\dfrac{1}{2}\right)}{\Gamma\left(\dfrac{n-1}{2}\right)}$$

From this, we see $\widehat{\theta}_2$ is unbiased when $\boxed{c = \sqrt{\dfrac{n-1}{2}}\cdot\dfrac{\Gamma\left(\dfrac{n-1}{2}\right)}{\Gamma\left(\dfrac{n}{2}\right)}}$.

Commentary:

I assign this problem because it offers a nice journey to a new land: $N(\theta, \theta^2)$. In part (a), we learn why someone might be interested in this setting. In parts (b) through (d), we look at some natural places to visit and explore their pros/cons. Here, we discover that $\widehat{\theta}_1 = \overline{X}$ is unbiased but *not* efficient. I think it is important to see non-examples in mathematics; we often focus too much on examples of certain behavior. In part (d), we encounter another natural site in $N(\theta, \theta^2)$ and work to make it as clean as possible. By choosing c appropriately, we can get a second unbiased estimator, so now we have a choice of estimators if we return to this setting in the future. I particularly like part (d) because it reveals the value of hierarchical knowledge. In Problem 70a, the Kernel Technique helped us evaluate $E[\left(\chi^2_n\right)^k]$. On top of that, we used the relationship of S^2 to χ^2_{n-1}, ultimately allowing us to find the desired expectation. This problem was also designed to leave you at the precipice of curiosity. After we find c in part (d) to make $\widehat{\theta}_2$ unbiased, you might, in the spirit of part (c), ask if this new estimator is efficient. If you check, you'll find it isn't, which will raise the next question: Does this new world have an efficient estimator? A graduate-level math stats class can help answer this.

One final observation is this: If you calculate the efficiency of $\widehat{\theta}_1$ in part (c), you get

$$\text{eff}(\widehat{\theta}_1) = \frac{\text{CRLB}}{Var(\widehat{\theta}_1)} = \frac{\dfrac{\theta^2}{3n}}{\dfrac{\theta^2}{n}} = \frac{1}{3} < 1$$

This calculation reveals that our estimator is inefficient, but also provides a nice example where the efficiency is a *constant*! Often the efficiency depends on the parameter (θ) and/or the sample size (n). This is a nice example where neither affects the efficiency, so collecting more data doesn't get us closer to efficiency.

72. Let X_1, X_2, X_3 be a random sample from the standard normal distribution. Package these three values together to get a random point in three-dimensional space: (X_1, X_2, X_3). Find the probability this random point lies inside (but not on) the sphere of radius 1 in two ways: (1) use a mathematical argument that ends with a single line of R code, and (2) write a simulation involving 100,000 simulated points.

Hint List:

- Using the variables x, y, z, and r, what is the equation of a sphere?

Solution:

Recall that a sphere of radius r has equation $x^2 + y^2 + z^2 = r^2$. If our point lies inside the sphere of radius 1, we need $X_1^2 + X_2^2 + X_3^2 < 1^2$. Since the X_is are a random sample, they are independent and come from the distribution $Z \sim N(0, 1)$. Thus, we know $X_1^2 + X_2^2 + X_3^2 \sim \chi_3^2$, the chi-squared distribution with df $= 3$. Therefore, we want $P(\chi_3^2 < 1) = \boxed{\text{pchisq(1, df=3)} \approx 0.199}$.

We can also simulate 100,000 points in space and see what fraction lie inside the sphere:

```
numsim = 100000

inSphere = function() {
  coords = rnorm(3, 0, 1)
  dist = sqrt(sum(coords^2))
  return(dist < 1)
}

mean(replicate(numsim, inSphere()))
```

This code gives a similar result to the above value. Note that we can compress the above code into a single line if we don't care about the reader at all:

```
mean(replicate(100000, sum(rnorm(3)^2) < 1))
```

Commentary:

The two types of probability you find in this problem are known as "theoretical" (i.e., exact and based on distributions) and "empirical" (i.e., approximate and based on data in a simulation). Which do you like better? If you ask "pure" (i.e., theoretical) math majors, most prefer the former, pchisq(1, df=3). If you ask "applied" folks, you tend to get the latter, the approximation from the simulation. As with most binaries, there are pros and cons to both options. When we examine pchisq(1, df=3), this represents an exact number (the area to the left of 1 under χ_3^2). Unfortunately, we have no sense of what that number is unless we go estimate it. This raises an interesting question: What algorithm is R using to estimate the area under the chi-squared distribution? And more generally, what algorithms are used by modern software programs to estimate areas?

Turning to our applied probability, it has the benefit of offering a clear numeric value. We can also see how the code creates a 3D coordinate, measures its distance to the origin, and determines if it's in the sphere. That is, the reader can directly see what is going on. Unfortunately, the accuracy of the answer is problematic. Without a large number of simulated points, the estimated answer can be far off. For me, the choice of an exact vs. approximate answer comes down to factors like communication and accuracy. When we define new symbols like π (as the ratio of a circle's circumference and diameter) and $\sqrt{2}$ (as the positive solution to $x^2 = 2$), we are creating symbols that hold *exact* values. At times, we use those symbols. But, in other settings, we use approximations like $\pi \approx 3.14$ (to make sense of Pi Day, March 14) or $\sqrt{2} \approx 1.414$ (for a calculation during a construction project).

73. Prove these two useful facts about the F distribution:

(a) $E[F_{m,n}] = \dfrac{n}{n-2}$ (when $n > 2$)

(b) $F_{m,n} = \dfrac{1}{F_{n,m}}$

Hint List:

- a. Recall the definition of $F_{m,n}$ and previous homework problems.
- b. Surprisingly, you don't have to work very hard.

Solution:

(a)

$$\begin{aligned} E[F_{m,n}] &= E\left[\frac{\chi^2_m/m}{\chi^2_n/n}\right] = \frac{n}{m} E\left[\chi^2_m \cdot (\chi^2_n)^{-1}\right] \overset{\text{ind}}{=} \frac{n}{m} E[\chi^2_m] \cdot E[(\chi^2_n)^{-1}] \\ &= \frac{n}{m} \cdot m \cdot \frac{2^{-1} \cdot \Gamma\left(\frac{n}{2} - 1\right)}{\Gamma\left(\frac{n}{2}\right)} \text{ (see Problem 70a)} \\ &= \frac{n \cdot 2^{-1} \cdot \Gamma\left(\frac{n}{2} - 1\right)}{\left(\frac{n}{2} - 1\right) \cdot \Gamma\left(\frac{n}{2} - 1\right)} = \boxed{\frac{n}{n-2}}. \end{aligned}$$

To use Problem 70a, we need $k > -\frac{n}{2}$. Here, $k = -1$, giving $-1 > -\frac{n}{2}$, or $2 < n$, as the problem stipulates.

(b) Notice that $\frac{1}{F_{n,m}} = \frac{1}{\frac{\chi^2_n/n}{\chi^2_m/m}} = \frac{\chi^2_m/m}{\chi^2_n/n} = F_{m,n}$.

Commentary:

I assign this problem for three reasons. First, it builds on Problem 70a. Since the F distribution is defined in terms of chi-squared RVs, this implies you can, in part, understand the F distribution via your chi-squared knowledge. This is a more general life lesson: When new systems are built atop old ones, you can understand the higher-level constructions based on their lower-level components.

Second, the solutions to both (a) and (b) are surprising. In (a), you might notice that the expectation only involves n (not m). This is somewhat counterintuitive, but possible; recall that $E[N(\mu, \sigma^2)] = \mu$, and σ^2 does not appear. In (b), the solution feels so easy that you might think it is wrong. It turns out there are several ways to show two RVs are equal. Sometimes we show they have the same pdf (or cdf). Here, we show both have the same RV expressions defining them.

Finally, stop listening to the academic narratives in your head. Here, part (b) goes against many of the things you might believe: If you think "the solution to part (b) will always need part (a)", this isn't true. If you think "the difficulty should increase in multi-part problems", this isn't true. If you think "short solutions are likely incorrect", this isn't true. Put your trust in the veracity of your logical deduction, not in socially-constructed rules about how academia usually works.

74. Students often find it hard to believe that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. So, let's simulate this situation and see if the data agree. To do so, generate $n = 30$ numbers from $N(4, \sigma^2 = 6)$. Find the variance of these n numbers, and replicate this process a total of 10000 times. Make a density plot of the 10000 values for $\frac{(n-1)S^2}{\sigma^2}$. Then, in red, overlay the pdf for χ^2_{n-1}. Include your code and a sketch of your plot.

Hint List:

- You may need to look up how to do several of these steps in R. This is totally fine and how your life might look in the future. I've programmed in many languages: Pascal, C, C++, Java, Python, R, SQL, html, Matlab, etc. It simply isn't possible to remember the syntax of so many languages. Learning *how to effectively search* for syntax-related questions is an important skill too!

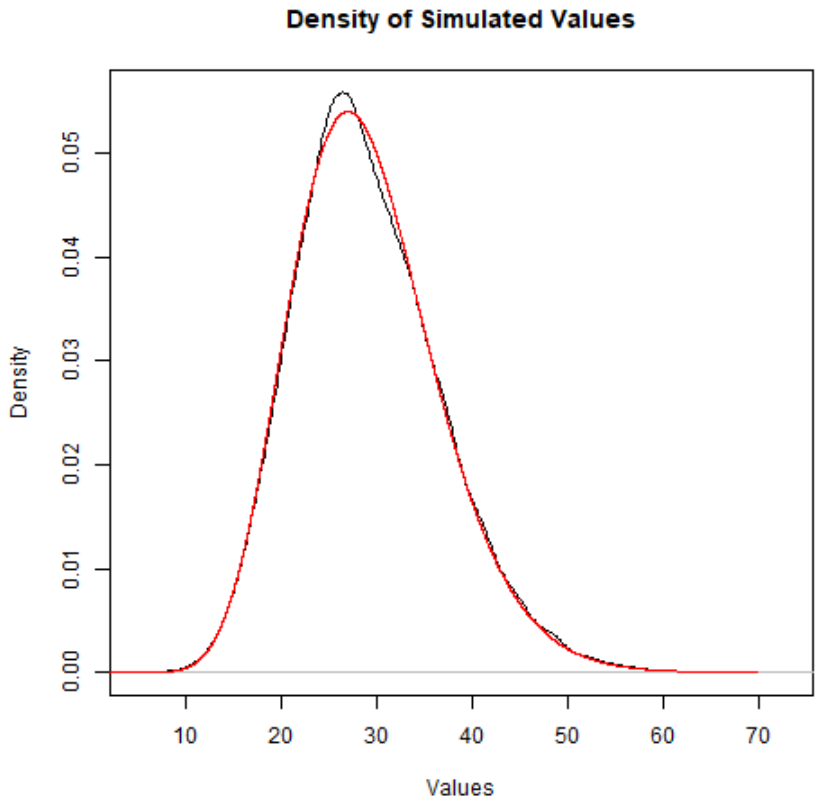
Solution:

```
# constants
n = 30
trials = 10000
mu = 4
sigma = sqrt(6)

# make the data
s2vals = replicate(trials, var(rnorm(n, mu, sigma)))
vals = (n-1)*s2vals/sigma^2

plot(density(vals), main="Density of Simulated Values", xlab="Values")

# add chi-squared (df = n-1) to existing plot
xvals = seq(0, max(vals), 0.01)
yvals = dchisq(xvals, df=n-1)
lines(xvals, yvals, col="red")
```



Commentary:

It may be difficult to believe, but the idea of simulating complex probabilistic scenarios is fairly new. Computers needed to be discovered and then developed enough to handle large numbers of iterations quickly. We didn't really reach that point (for the average consumer) until the early 2000s. When you run such a simulation, it is known as a "Monte Carlo simulation". The naming history and some nice examples are on the Wikipedia page. There are many situations in which simulations are incredibly useful. For example, if you want to know your long-run odds of winning a blackjack hand, this is easiest to simulate since the theoretical probability calculations are nightmarish (given the number of cases to consider). Indeed, professional blackjack players make their decisions based on charts that were built through simulations, not probability calculations. Other questions practically scream out "Run a simulation!". For example, you might wonder how many times you need to shuffle a deck of cards before it appears to be in random order. There is an entire branch of mathematics that looks at this question. Once you decide what it means to "shuffle a deck" and "appear random", you can code things up and get your answer pretty easily via simulation (Spoiler alert: It's about 6 or 7 shuffles using common definitions of "apparent randomness").

75. It is also surprising that $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim T_{n-1}$. To empirically convince yourself, do these steps:

Let $X \sim N(3, \sigma^2 = 5^2)$. Draw an iid sample of size $n = 3$. Using this sample, calculate the ratio $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$. Replicate this process 10000 times. Draw the density, and then in red, overlay the density of T_{n-1}. Include your code and a sketch of the plot.

Hint List:

- See the R code from the last exercise for inspiration. Some interesting changes may be required.

Solution:

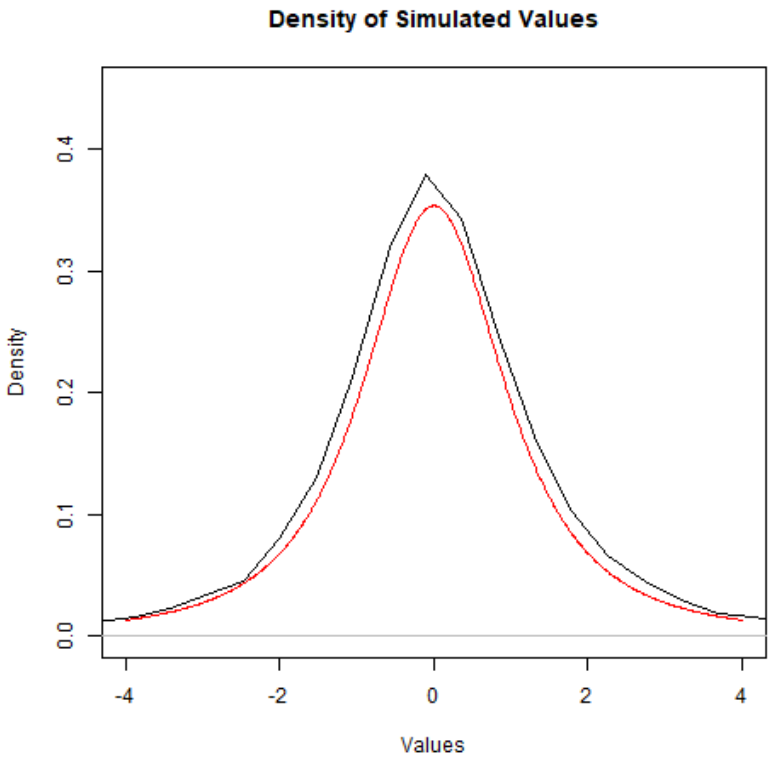
```
# constants
n = 3
trials = 10000
mu = 3
sigma = 5

# a user-defined function to make a t-stat
tstat = function() {
  data = rnorm(n, mu, sigma)
  t = (mean(data) - mu)/(sd(data)/sqrt(n))
  return(t)
}


tstats = replicate(trials, tstat())

plot(density(tstats), main="Density of Simulated Values",
     xlab="Values", xlim=c(-4, 4), ylim=c(0, 0.45))

# add T (df = n-1) to existing plot
xvals = seq(-4, 4, 0.01)
yvals = dt(xvals, df=n-1)
lines(xvals, yvals, col="red")
```



Commentary:

This problem is quite similar to the previous. The main differences are (1) that you likely wrote your own function to make the t stats and (2) you probably limited the viewing window on the plot (via xlim and/or ylim in the plot command). One of the best ideas to come out of educational research is the notion of necessitation. In academia, if an idea is "necessitated", it means that you have provided a setting where the student will feel the need for a certain object, action, procedure, etc. For example, you could ask young students to solve $x^2 = 4$ and then $x^2 = 9$ and then $x^2 = 7$. With the right guidance, you can get them to see that there is a solution between $x = 2$ and $x = 3$, and that we *need* a symbol that represents the solution ($\sqrt{7}$), even if we cannot simplify it.

In this problem, you realize that we *need* to perform a few calculations on the data set (mean, sd) before we can calculate the t stat. So, it is difficult to write a single line of code that makes a t stat and then surround this with replicate (as in the previous problem's one-line solution). As such, we feel it is necessary to write our own function (or use a for loop or learn about "anonymous functions", which are alternative approaches). Also, when you went to plot the density of the simulated t stats, it's likely your viewing window was far too wide to clearly see things. This necessitates the idea of limiting the plot bounds, and a search with ?plot in R reveals that xlim and ylim are useful parameters for this.

One of my biggest gripes with math education is that educators do a poor job of *necessitating* important concepts. Virtually all mathematical ideas were necessitated by some moment in history. For example, when early humans needed notation to symbolize debt ("I owe you 4 sheep"), negative numbers were born. When scholars asked if $x^2 = -1$ might have a solution, $i = \sqrt{-1}$ came into being. And so on. Too often, we forget the powerful moments in history that made humans feel like they *had to* create a new concept (or notation or programming syntax). Hopefully this problem made you feel a bit of necessitation.

76. Decide which is bigger and explain why. Only use R to confirm your answer.

(a) $t_{0.07,n} - t_{0.14,n}$ or $t_{0.14,n} - t_{0.21,n}$

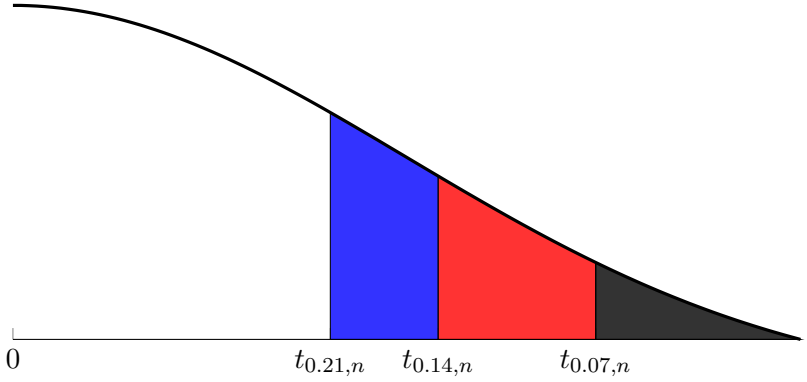
(b) $P(|T_n| < 1)$ or $P(|T_{n+1}| < 1)$

Hint List:

- a. Draw a picture of T_n and label the quantiles mentioned. Now shade areas and think.
- b. How does the shape of T_n change as $n \to \infty$?

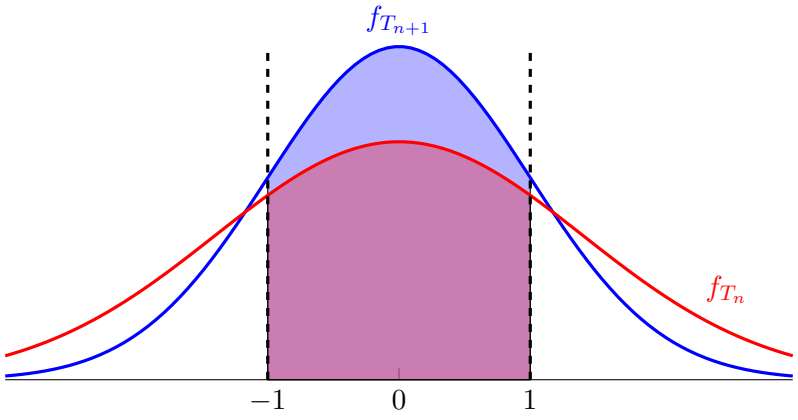
Solution:

(a) Recall that $t_{\alpha,n}$ is the place on the horizontal axis of T_n to the right of which the area is α. Thus, we can draw this picture:



Each of the different colors above represents an area of 0.07 because the labels on the horizontal axis demarcate right-tailed areas of 0.21, 0.14, and 0.07 (hence, the in-between areas are $0.21 - 0.14 = 0.07$ and $0.14 - 0.07 = 0.07$). Notice that when the graph of T_n is lower, it takes more horizontal distance to create the same area of 0.07. So, the width of the red area is more than the width of the blue. That is: $\boxed{t_{0.07,n} - t_{0.14,n} > t_{0.14,n} - t_{0.21,n}}$.

(b) If we rewrite the inequality, we are exploring $P(-1 < T_n < 1)$, which is the area under the pdf of T_n on $(-1, 1)$.



As $n \to \infty$, $T_n \to Z$. During the journey to Z, the peak at 0 gets higher and the tails get lower. Because of this, more area is concentrated near 0 and thus $\boxed{P(|T_{n+1}| < 1) > P(|T_n| < 1)}$.

Commentary:

How did this problem *feel* to you? Notice that there's no procedure for doing this problem. Part (a) requires that you understand the shape of the pdf of T_n for a fixed n, and part (b) requires that you understand the sequence of pdfs $T_1, T_2, \ldots$. Problems like this develop *conceptual understanding* (CU). That is, they focus on concepts (e.g., the shape of a pdf, the limiting behavior of T_n) rather than demand you know a procedure (e.g., "Differentiate $\sin x \cdot \cos x$"). It is easy to see your mathematics training as nothing but a set of procedures. *Procedural understanding* (PU) is the easiest thing to teach, and the safest thing to put on exams.

Indeed, the typical classroom cycle is this: 1) a teacher shows a procedure in class (e.g., how to find the roots of a quadratic equation using the quadratic formula), 2) students do 5-30 problems using that procedure on homework, and 3) an exam contains a few problems with that procedure using new numbers. Sadly, this approach leaves out much of what makes math so interesting. When you add CU to a classroom, the story becomes far richer. Students learn how things are defined, why things are true, how different ideas relate, and so on. As a challenge to you, spend your next day at school constantly asking yourself this: Is the thing I'm learning right now an example of PU or CU? Also, when you study for exams, do you focus exclusively on PU? The best students naturally prioritize *both* types of understanding even if they don't have names for these categories.

77. Each day, Zelda starts the morning with a cup of coffee from her Keurig machine. Lately, she has begun to think the machine is malfunctioning because the amount dispensed is different than the advertised average amount, 12 oz. To explore this, she picks 7 random days from the next month and actually weighs her coffee using a calibrated kitchen scale. Assume that the coffee dispensing amounts are normally distributed and that her seven data values give $\overline{x} = 12.9$ oz and $s = 0.7$ oz.

(a) Conduct a hypothesis test for Zelda's predicament using $\alpha = 0.06$.

(b) Find a confidence interval for the true average coffee dispensing amount that has duality with the hypothesis test from part (a).

Hint List:

- This is a basic problem using the t distribution for hypothesis tests and CIs. If you're struggling, it is best to return to the notes and review these before starting.

Solution:

(a) Let $X \sim N(\mu, \sigma^2)$ represent the current coffee-dispensing distribution for Zelda's Keurig where σ^2 is unknown. We focus on μ, the current average amount dispensed. We test

$$H_0\text{: } \mu = 12 \text{ vs. } H_1\text{: } \mu \neq 12 \text{ (“is different than”)}$$

Assuming H_0, we have: $\dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} = \dfrac{12.9 - 12}{0.7/\sqrt{7}} \approx 3.402$. Given our two-sided alternative, the P-value is $2 \cdot \text{pt}(3.402, \text{df=6}, \text{lower=F}) \approx 0.0145$. Since $0.0145 < 0.06$, we reject H_0 in favor of H_1. It appears the average dispensing amount may now be different.

(b) A two-sided alternative hypothesis creates a test that has duality with our usual “two-sided” CI. For $\alpha = 0.06$, our $100(1-\alpha) = 94\%$ CI is

$$\overline{x} \pm t_{\alpha/2, n-1}\frac{s}{\sqrt{n}} = 12.9 \pm \text{qt}(0.03, \text{df=6}, \text{lower=F}) \cdot \frac{0.7}{\sqrt{7}} \approx \boxed{(12.288 \text{ oz}, 13.512 \text{ oz})}$$

Note that $12 \notin$ CI and that we rejected $H_0\text{: } \mu = 12$ in part (a), as the duality demands.

Commentary:

While this is a straightforward exercise using the t infrastructure for HT and CIs, it does point to a nice use of statistics. In manufacturing processes, we work with machines that were built to certain specifications. For example, a Keurig may be designed to dispense 12 oz of coffee, allowing only a small variation. Over time, machines may need to be re-calibrated to attain the original mean. To determine when this re-calibration must be made, a hypothesis test can help us. This is another example where we might actually be excited by the null hypothesis, since it says the machine is still working properly, and we need not pay to have things fixed.

78. In question 59, we explored home-field advantage (HFA) using win percentages. Now we revisit the question using the "margin of victory", which is amenable to means. Looking at data from 317 college (American) football games involving top-25-ranked teams, researchers found an average margin of victory (home team score − away team score) of $\overline{y} = 4.57$ with $s = 18.29$. Do these data support the notion of HFA?

(a) Conduct an appropriate hypothesis test using $\alpha = 0.03$. Then, create a CI that has duality with this hypothesis test.

(b) After publishing your findings, a rival academic argues that you have failed to establish the normality of the population being considered (margins of victory for college football games involving top-25-ranked teams). Respond to this criticism.

Hint List:

- a. For the CI, return to the HT-CI duality lecture and adapt the formula there to our setting involving the t distribution.

Solution:

(a) Let μ be the average difference (home team score − away team score) in **all** football games involving top-25-ranked teams. We test

$$H_0\text{: } \mu = 0 \text{ vs. } H_1\text{: } \mu > 0 \text{ (HFA is real)}$$

After assuming H_0, we have $t = \dfrac{\overline{y} - \mu_0}{s/\sqrt{n}} = \dfrac{4.57 - 0}{18.29/\sqrt{317}} \approx 4.449$. Since we have a one-sided alternative hypothesis, our P-value is pt(4.449, df=316, lower=F) $\approx 5.984 \cdot 10^{-6} < 0.03$, so we reject H_0 in favor of H_1. It appears home teams are out-performing away teams.

In our lecture on HT-CI duality, we saw that HTs like H_0: $\mu = \mu_0$ vs. H_1: $\mu > \mu_0$ have duality with CIs like $\left(\overline{x} - z_\alpha \dfrac{\sigma}{\sqrt{n}}, \infty\right)$ where σ^2 is known. Here, σ^2 is unknown, so we adapt our CI to have the form $\left(\overline{x} - t_{\alpha,n-1} \dfrac{s}{\sqrt{n}}, \infty\right)$ which, for our data, gives

$$\left(4.57 - \text{qt}(0.03, \text{ df=316, lower=F}) \cdot \frac{18.29}{\sqrt{317}}, \infty\right) = \boxed{(2.631, \infty)}$$

Note that $0 \notin$ CI, so by the duality, we should reject H_0: $\mu = 0$, which we did above.

(b) We began our study of the T_n distribution with μ in $X \sim N(\mu, \sigma^2)$ and σ^2 unknown. After doing so, we learned that even if X is *not* normally distributed, we still have $\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \approx T_{n-1}$ if n is large enough (or X has low skew). Here, $n = 317$ is large enough to establish the $\approx$ sign that we need. So, your rival is incorrect.

Commentary:

Hopefully by now, you've gotten a message that I repeat over and over: Statistics does not prove your *theories* are correct. This problem shows that the data strongly push us toward $\mu > 0$. But, the data don't reveal what is really behind that $>$ sign. Indeed, there are many reasons why home teams may outperform away teams: biased refereeing, players giving extra effort, key knowledge about one's local playing field, travel exhaustion/time zone changes for the away team, etc. In fact, almost all phenomena in the real world are multivariate in nature. That is, there are many factors that contribute to an overall effect you might observe. Social media and standard news outlets love to push *univariate* reasoning for complex issues (e.g., "gas prices went up this week because the President did X"). Resist this way of thinking unless you want to be a one-dimensional individual.

79. The most brutal algebra moment in 181A:

Show that the pdf of T_n converges to the pdf of $N(0,1)$ as $n \to \infty$.

To help, you should use three tools:

- Stirling's Formula: $n! \approx \sqrt{2\pi n} \cdot n^n \cdot e^{-n}$
- A fact from Calculus: $\lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = e^a$
- Allow yourself to write $\Gamma(r) = (r-1)!$, even if r is not an integer

Hint List:

- Try to create the expression $\left[\frac{n-1}{n-2}\right]^{n/2}$ in the middle of your argument.
- You might be tempted to claim that $\lim_{n\to\infty}\left[\frac{n-1}{n-2}\right]^{n/2} = 1$, but this isn't true. This expression heads to 1^∞, but $1^\infty \neq 1$. Instead, this is an indeterminate power (IP), which you first encountered when studying L'Hôpital's Rule in Calculus. Another example of an IP is $\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n$, which equals e^a, not 1, as you are reminded in the problem statement.

Solution:

Note first that $f_{T_n}(t) = \dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}$.

First, $\left(1+\frac{t^2}{n}\right)^{(n+1)/2} = \sqrt{\left(1+\frac{t^2}{n}\right)^n}\cdot\left(1+\frac{t^2}{n}\right)^{1/2} \to \sqrt{e^{t^2}}\cdot 1^{1/2} = e^{t^2/2}$ as $n\to\infty$.
Here, we use the Calculus fact to evaluate the first limit (with $a = t^2$), but we don't need it for the second limit since the exponent is $1/2$, not n.

Now, we simplify the red expression by switching to factorials and using Stirling's Formula:

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)} = \frac{\left(\frac{n-1}{2}\right)!}{\sqrt{n\pi}\left(\frac{n-2}{2}\right)!} \approx \frac{\sqrt{2\pi\cdot\frac{n-1}{2}}\left(\frac{n-1}{2}\right)^{(n-1)/2}e^{-(n-1)/2}}{\sqrt{n\pi}\sqrt{2\pi\cdot\frac{n-2}{2}}\left(\frac{n-2}{2}\right)^{(n-2)/2}e^{-(n-2)/2}}$$

We combine the purple expressions and rework the green ones to have a common exponent of $\frac{n}{2}$:

$$\frac{\sqrt{n-1}}{\sqrt{n\pi}\sqrt{n-2}}\cdot\left[\frac{\frac{n-1}{2}}{\frac{n-2}{2}}\right]^{n/2}\cdot\frac{\left(\frac{n-1}{2}\right)^{-1/2}}{\left(\frac{n-2}{2}\right)^{-1}}\cdot e^{-\frac{n}{2}+\frac{1}{2}+\frac{n}{2}-1} = \frac{\sqrt{n-1}}{\sqrt{n\pi}\sqrt{n-2}}\cdot\left[\frac{n-1}{n-2}\right]^{n/2}\cdot\frac{\sqrt{2}(n-2)}{2\sqrt{n-1}}\cdot e^{-1/2}$$

Canceling and reorganizing gives: $\frac{\sqrt{n-2}}{\sqrt{n}}\cdot\frac{1}{\sqrt{2\pi}}\cdot\left[\frac{n-1}{n-2}\right]^{n/2}\cdot e^{-1/2}$. $(\bigstar)$

As $n\to\infty$, the orange term $\to 1$ and the green term goes to $\sqrt{e}$ because:

$$\left[\frac{n-1}{n-2}\right]^{n/2} = \sqrt{\left(1+\frac{1}{n-2}\right)^n} = \sqrt{\left(1+\frac{1}{n-2}\right)^{n-2}\cdot\left(1+\frac{1}{n-2}\right)^2} \to \sqrt{e^1\cdot 1^2} = \sqrt{e}$$

Here, we have used the Calculus fact (replace n with $n-2$) to evaluate the first limit; the fact is not needed on the second limit since the exponent is 2, not n. So $(\bigstar) \to 1\cdot\frac{1}{\sqrt{2\pi}}\cdot\sqrt{e}\cdot e^{-1/2} = \frac{1}{\sqrt{2\pi}}$.

Returning to the start, we now see $f_{T_n}(t) \to \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{e^{t^2/2}} = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} = f_Z(t)$, as desired.

Commentary:

Oof! What a mess. The first time you encounter something with this much algebra, it can be really overwhelming. I remember having that feeling when I was a student and first saw something like this. The reason I assign this problem is three-fold. First, I want you to see how your past (the Calculus fact), your present (the Gamma function relationship to factorials), and your future (Stirling's Formula) can come together to create something beautiful. I think it's really rewarding when a lot of pieces work in harmony, like the musicians in an orchestra. Second, I want you to feel proud of how far you've come mathematically. The percentage of people on Earth who can make sense of this argument is incredibly small, and it's valuable to pause at key moments on your climb toward greatness and admire the view that you've earned through years of effort. Finally, I suspect many in this class are considering graduate school in the future. To be successful at that level, you have to be prepared for long, detail-oriented challenges like this. For many, the jump to graduate school feels huge because their undergraduate professors did not do a good job of dovetailing the content and expectations of the current level with the next level. (Some students also feel this in the transition from high school to undergrad because their high school teachers failed to navigate the transition.) I want to avoid such leaps by offering problems like this at select moments in the course when you are best positioned to strive for the next level of excellence. If even a small part of you thinks grad school might be in your future, remember to give extra effort on the hardest challenges your courses provide.

80. In class, I claimed that if a population distribution has moderate or severe skew, then as long as the sample size was about 30 or more, we could count on $T_{n-1} \approx \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$. Our goal here is to empirically show this. To begin, let $n = 4$ and draw the T_{n-1} density on the interval $(-4, 4)$ using a dashed line (use the lty parameter in the plot function to make a dashed line). Now, let $X \sim Exp(\lambda = 3)$, draw a sample of size n, and compute $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ for this data set to get a single t-stat. Repeat this process to get a total of 50000 t-stats. Plot the density of these atop T_{n-1} using a solid line. Both plots should be in the same color. Next, repeat the above process for $n = 10$, 30, and 60 (use a new plot and new color for each n value). Include your code and a sketch of the four plots.

    Note: This example will also show that $n \approx 30$ is not some absolute rule. You might be quite bothered by the difference in your two graphs when $n = 30$. In general, the larger the skew in the population, the larger n should be to overwhelm its effect. There is no perfect guideline here: data analysis and statistics involve *human, imperfect* decision making. Sorry to break your quantitative heart.

Hint List:

- The code par(mfrow = c(2,2)) will set up a 2 by 2 array of plots, if you wish to put the four plots in a single image.

Solution:

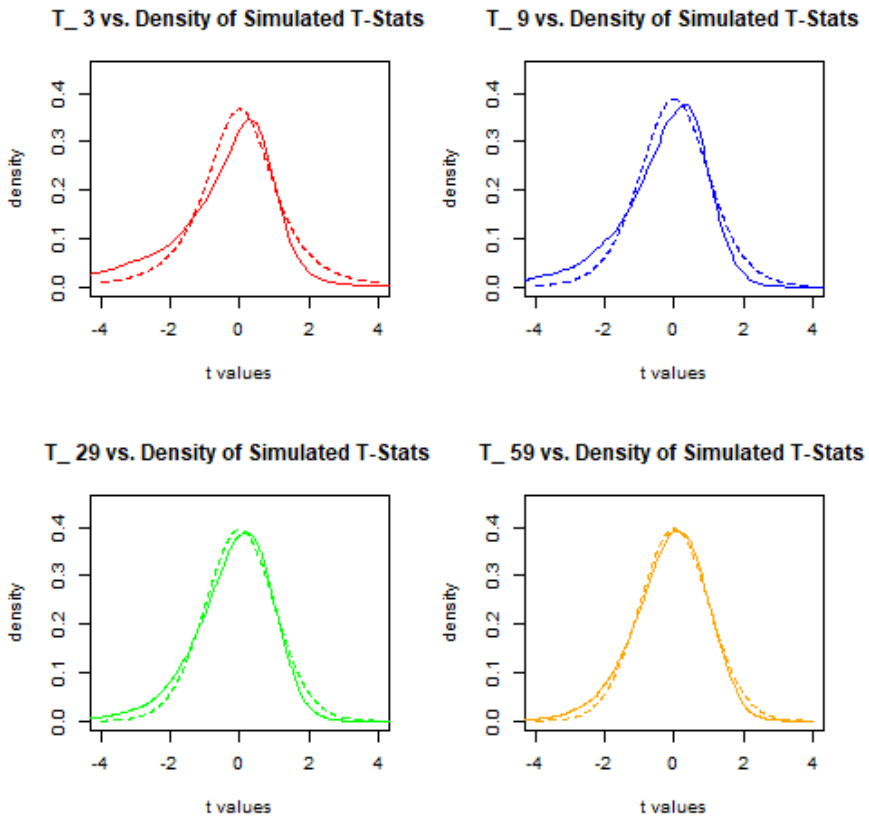
```
tstat = function(data, mu) {
  return((mean(data) - mu)/(sd(data)/sqrt(length(data))))
}

xs = seq(-4, 4, 0.01)
cols = c("red", "blue", "green", "orange")  # colors for pic
nvals = c(4, 10, 30, 60)

par(mfrow = c(2,2)) # make a 2x2 array of plots (optional)

for (i in 1:4) {
  # set up plot and add appropriate T distribution
  plot(xs, dt(xs, df=nvals[i]-1), type="l", col=cols[i], lty=2,
       xlim=c(-4, 4), ylim=c(0, 0.45), xlab="t values", ylab="density",
       main=paste("T_",nvals[i]-1,"vs. Density of Simulated T-Stats"))

  # simulate T stats and add to existing plot
  tstats = replicate(50000, tstat(rexp(nvals[i], 3), 1/3))
  lines(density(tstats), col=cols[i])
}
```



Commentary:

This problem highlights a trade-off in statistics. If you're lucky enough to study something with a normal distribution, $X \sim N(\mu, \sigma^2)$, then you get $\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim T_{n-1}$. That last statement is an exact, not approximate, equivalence. If, however, you study something that isn't normally distributed, then $\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$ does not exactly have a T distribution. If you want to have a *single* testing infrastructure that works in virtually all cases, then you must settle for that ratio having an *approximate* T distribution. And, you must try to put some guardrails in place for when the approximate sign is "good enough" to functionally use. These guardrails look like: "the $\approx$ sign is good enough when $n \geq 60$ for high-skew in X, $n \geq 30$ for low to moderate skew in X, and small n is OK for no skew in X". So, if you're willing to endure this vagueness and only be approximately correct, then you get to use a single tool to study means.

Note that this trade-off is not unique to statistics. In the real world, the Swiss army knife is a *single* tool that contains many smaller tools (e.g., scissors, corkscrew, rasp, paring knife, etc.). The smaller tools are an *approximation* of their larger version, and if you're OK with only getting their approximate functionality, then all is well. Similarly, in the English language, we have *single* words (e.g., "gadget") that convey an *approximate* meaning (e.g., an electronic/mechanical device or tool) if you only need an approximate description of something.

81. If you look at a bottle of ibuprofen, it will likely list the dosage of medicine per pill (usually, 200 mg). Of course, this is only an average, and if you carefully modeled the amount from pill to pill, you would get a normal distribution. The spread of this distribution is very important because giving too much or too little medicine can be dangerous. Suppose that the standard deviation in dosages is 10 mg based on current manufacturing processes. You've come up with a new way to create the pills that you believe will increase the precision of the dosage. To check this claim, you produce a bunch of pills with the new process and randomly select some to measure the dosage. You get these values (in mg):

$$206.5, 198.9, 205.2, 205.8, 192.0, 199.5, 182.5, 191.9, 197.6, 190.7, 186.8, 187.3, 192.0$$

(a) Conduct a hypothesis test with $\alpha = 0.04$.

(b) Construct a 97% two-sided CI for the standard deviation in pill dosages for the new manufacturing process.

Hint List:

- Study the lecture on variance if you are struggling to begin this problem.

Solution:

(a) Let σ^2 be the variance in pill dosages using the new manufacturing process. We test

$$H_0\text{: } \sigma^2 = 10^2 \text{ vs. } H_1\text{: } \sigma^2 < 10^2 \text{ (increased precision means less variability)}$$

We enter the data into R to get $s^2 \approx 60.31731$. Assuming H_0, we have

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{12 \cdot 60.31731}{10^2} \approx 7.238$$

To decide which tail to shade on $\chi^2_{n-1} = \chi^2_{12}$, imagine if H_1 were actually true. If so, the universe would give us data that would have a smaller s^2, and hence a smaller χ^2 test stat (because s^2 is in the numerator). Since small χ^2 values support the alternative, we shade toward the lower tail. Thus, our P-value is pchisq(7.238, df=12) ≈ 0.159. Since $0.159 > 0.04$, we keep H_0. Our process may have a similar variance as the previous process.

(b) Since $\alpha = 0.03$, we must find $\chi^2_{\alpha/2,n-1}$ = qnorm(0.015, 12) and $\chi^2_{1-\alpha/2,n-1}$ = qnorm(0.985, 12). With these, we get our CI for σ:

$$\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2.n-1}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2.n-1}}}\right) \approx \boxed{(5.385 \text{ mg}, 13.605 \text{ mg})}$$

Commentary:

When you worked (a) and (b), did any part of you get bothered by the fact that the HT and CI do not have duality? Part (a) is a one-sided HT with $\alpha = 0.04$; part (b) is a two-sided CI with $\alpha = 0.03$. I did this intentionally to inspire students to ask: How would I create a CI in part (b) to have duality with the HT in part (a) (or vice versa)? If you asked this question, kudos. If you went exploring, kudos factorial.

I also assign this question because it presents a real-world situation where the variance really matters. You simply can't have pills with wildly different amounts of medicine in them or people might die. One thing I love about education is that after you learn an idea, you begin to see the world differently. For example, after being introduced to the concept of an inequality, you are able to mathematize statements like "Customers saved up to 30% by switching to Company X". Here, the company is saying "savings $\leq 30\%$", which is not great, since it leaves open the possibility that you'll save 0% by switching. Once you've taken some statistics, you also recognize that the company is sharing the *max* savings. As a consumer, you probably want to know a different statistic, like the mean, median, or mode. The company intentionally chooses a misrepresentative statistic to lure you in. Always remember that knowledge lets you better understand the world; you can view life with a pair of statistical glasses, or philosophical glasses, or whatever content you've studied. As you begin to see things more clearly, you can also actively seek out situations that are best viewed with certain glasses. For example, now that you've seen this problem, can you think of other examples in life where we care most about the *standard deviation* or *variance* of a distribution (as opposed to its mean)?

82. I recently attended a Padres (baseball) game that was on pace to be the shortest game in the modern era (all hope was ruined when people started scoring in the 8th inning). I also happened to be watching TV in 2010 when the longest tennis game ever was played (11 hours, 5 minutes). All this got me thinking about the times of sporting events. For the sake of fans, commentators, and marketing departments, it is helpful to have low variability in the time it takes to complete an event, and an average game time that is long enough to entertain fans, but not so long that people get exhausted.

    In this vein, you decide to explore the effects of various rule changes that occurred in the NHL (ice hockey). Prior to a new rule set launched in the early 2000s, hockey game times were known to be normally distributed with an average time of 2 hours and 36 minutes and a standard deviation of 19.2 minutes. Using a random sample of 24 games from the 2012 season (these occurred after rule changes; we'll assume they are normally distributed), you find an average time of $\overline{x} = 2.316$ hours with $s_x = 18.3$ minutes. If the goal of these changes was to decrease the average game time but keep the variation the same, do you think the new rules have done it? Argue using two hypothesis tests, each with $\alpha = 0.02$. Do the variance test first, and then the mean test.

Hint List:

- Note that the results of the first test definitely influence how the second test is run. The first test is either telling you that σ^2 has a known value, or that it doesn't, and we have different infrastructures for studying μ in $N(\mu, \sigma^2)$ depending on whether σ^2 is known or unknown.

Solution: We begin by exploring variances and use this result in our mean test. We first test

$$H_0\text{: } \sigma^2 = 19.2^2 \text{ min}^2 \text{ vs. } H_1\text{: } \sigma^2 \neq 19.2^2 \text{ (“keep the variation the same”)}$$

Assuming H_0, $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{23 \cdot 18.3^2}{19.2^2} \approx 20.894$, which lives on $\chi^2_{n-1} = \chi^2_{23}$. Because the alternative is two-sided, we can't use P-values because the χ^2 distribution is asymmetric, so there is no "other tail" in which to draw a "symmetric spot" before shading. Instead, we use critical regions. R can find the spots that demarcate the bottom 0.01 and top 0.01 (here, $\alpha/2 = 0.02/2 = 0.01$) strangest results: qchisq(0.01, 23) ≈ 10.196 and qchisq(0.99, 23) ≈ 41.638.

Since $10.196 < 20.894 < 41.638$, we keep H_0. It seems like the variance is unchanged, so we can assume $\sigma^2 = 19.2^2$ moving forward.

Let $X \sim N(\mu, 19.2^2)$ be the time for a random NHL game after the rule changes. We now test

$$H_0\text{: } \mu = 156 \text{ min vs. } H_1\text{: } \mu < 156 \text{ (“decrease the average time”)}$$

Because $\sigma^2 = 19.2^2$ is known, we can use $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$ to run our test instead of $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$.

We get $z = \dfrac{138.96 - 156}{19.2/\sqrt{24}} \approx -4.348$. Here, our P-value is pnorm$(-4.348) \approx 6.869 \cdot 10^{-6} < 0.02$, so we reject H_0. It appears the average game time has decreased. Again, this might be due to the rule changes (our current theory) or some other reason (e.g., fewer television breaks because advertisers lost interest in pro hockey, changes in refereeing led to fewer stoppages in play, etc.).

Commentary: I like this question because it does a few things. First, it provides a nice real-world setting where you would care about *both* the variance and mean for some idea (here, lengths of sporting matches). Indeed, the MLB (baseball) changed several rules in the early 2020s to reduce the time for its games. Second, the problem provides an example where you would actually run a mean test using the z-stat instead of the t-stat. We began our study of hypothesis testing with the z-stat mainly because the math is easier when you assume more from the outset. Later, we got honest and admitted that it's unlikely we'll ever know σ^2. Well, here is a rare case where we do! Finally, this problem inspires a natural question: Instead of running two HTs in sequence, is it possible to run a *single* HT with both parameters? Currently, we don't have the infrastructure for such explorations, but perhaps this problem is necessitating a future tool for our toolkit.

83. Here we develop an alternative way of creating confidence intervals for the variance of a normal distribution. Let $X \sim N(\mu, \sigma^2)$ where μ and σ^2 are unknown.

(a) Show that $\chi^2_{n-1} \approx N(n-1, 2(n-1))$.

(b) Show that $\dfrac{\dfrac{(n-1)S^2}{\sigma^2} - (n-1)}{\sqrt{2(n-1)}} \approx N(0,1)$, and then use this fact to create a formula for an approximate two-sided $(1-\alpha)100\%$ CI for σ^2.

Hint List:

- a. Recall the definition of the chi-squared RV and results from earlier problems.
- b. Try the pivoting trick seen in the development of other CI formulas.

Solution:

(a) Recall that $\chi^2_n = \sum_{i=1}^{n} Z_i^2$ where $Z_i \overset{\text{iid}}{\sim} N(0,1)$. By the Central Limit Theorem, the sum of iid RVs (here, Z_i^2) is approximately normal. Now, from Problem 70b, we know $E[\chi^2_{n-1}] = n-1$ and $Var[\chi^2_{n-1}] = 2(n-1)$, completing the argument.

(b) We standardize our finding from part (a): $\dfrac{\chi^2_{n-1} - (n-1)}{\sqrt{2(n-1)}} \approx Z$. Now, recall from class that $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Together, these facts allow us to write

$$1 - \alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P\left(-z_{\alpha/2} < \frac{\frac{(n-1)S^2}{\sigma^2} - (n-1)}{\sqrt{2(n-1)}} < z_{\alpha/2} \right)$$

Pivoting the inequality to get σ^2 in the middle eventually reveals that

$$1 - \alpha = P\left(\frac{(n-1)S^2}{(n-1) + z_{\alpha/2}\sqrt{2(n-1)}} < \sigma^2 < \frac{(n-1)S^2}{(n-1) - z_{\alpha/2}\sqrt{2(n-1)}} \right)$$

Hence, our approximate two-sided $(1-\alpha)100\%$ variance CI is

$$\frac{(n-1)S^2}{(n-1) \pm z_{\alpha/2}\sqrt{2(n-1)}} \cdot \frac{1/(n-1)}{1/(n-1)} = \boxed{\frac{S^2}{1 \pm z_{\alpha/2}\sqrt{2/(n-1)}}}$$

Commentary:

After boxing your answer on this problem, you might naturally ask: Now that I have two variance CI formulas, which should I use? This is a reasonable question once you have multiple ways to do something. The answer always is: It depends on the setting and your goals. For example, if you want an exact CI and have access to a chi-squared table/computer, use the result from class. If you only need an approximate CI or only have access to a Z table, use the result above. In the modern era, it is tempting to prize exactness and believe you'll always have access to any resource you need, but this isn't always the case. Also, there are few downsides to adding more tools to your mental toolkit, and now you have one more resource at your disposal. Moreover, this problem suggests a general pathway for building approximate CIs: If an expression is approximately normal, it may be possible to craft a CI formula. In fact, this is precisely what we did for proportion CIs near the beginning of class. Now that you've seen additional examples of this approach, it might put that algebra in context.

84. As we began to study hypothesis tests in general, we decided to compare two tests by first equalizing their Type I error rates and then determining which has the higher power. Here is a natural example of what that looks like in the setting of manufacturing.

A plant produces a product (e.g., boxes of cereal, metal rods) that is sold by weight (more accurately, by mass). These weights are normally distributed with variance 800 g^2. Suppose that the product is supposed to have a mean weight of 56 g, but managers are concerned the mean may now be higher. To study this, they collect the weights $(X_1, \ldots, X_{100})$ of a random sample of 100 products. Then, to decide between the hypotheses H_0: $\mu = 56$ vs. H_1: $\mu > 56$, the managers propose two decision rules:

- Rule 1: Reject H_0 if $\overline{X} > 60$
- Rule 2: For each X_i, declare the product to be "defective" if $X_i > 70$. Reject H_0 if the number of defective products is ≥ 38.

(a) Show that the Type I error rate for each rule is about 8%.

(b) Assuming the true mean weight is now 62 g, determine which rule is better by comparing their powers.

Hint List:

- In both parts, use a normal distribution to help analyze Rule 1 and a Binomial distribution to analyze Rule 2.

Solution:

(a) Assuming H_0, we have $X \sim N(56, 800)$.

- Rule 1: Note that $\overline{X} \sim N\left(56, \frac{800}{100}\right) = N(56, 8)$. Thus,

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true}) = P(\overline{X} > 60) = \text{pnorm(60, 56, sqrt(8), lower=F)} \approx 0.08$$

- Rule 2: First, notice that

$$P(\text{“defective”}) = P(X_i > 70) = \text{pnorm(70, 56, sqrt(800), lower=F)} \approx 0.310$$

Let $Y \sim Binom(n = 100, p = 0.310)$ be the number of defective products in a sample of 100. Here,

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true}) = P(Y \geq 38) = \text{pbinom(37, 100, 0.310, lower=F)} \approx 0.08$$

(As a reminder, when you call the p function on a *discrete* random variable, like binom, the lower tail calculation is $\leq$ the place in question. So pbinom(37, ...) finds the sum of the probabilities up to, and including, 37. If you swap to an upper tail (via lower=F), the calculation begins at 38 and goes to the right. This nuance is not important for continuous random variables since the probability of any particular outcome is 0.)

(b) Here, we are given that $\mu = 62$, so $X \sim N(62, 800)$. Now, we must recalculate the distributions for both rules.

- Rule 1: Here, $\overline{X} \sim N(62, 8)$, and we have

$$\text{power} = P(\text{reject } H_0 \mid H_1 \text{ true}) = P(\overline{X} > 60) = \text{pnorm(60, 62, sqrt(8), lower=F)} \approx 0.760$$

- Rule 2: Now, $P(\text{“defective”}) = P(X_i > 70) = \text{pnorm(70, 62, sqrt(800), lower=F)} \approx 0.3886$.

Let $Y \sim Binom(n = 100, p = 0.3886)$ be the number of defective products in a sample of 100.

$$\text{power} = P(\text{reject } H_0 \mid H_1 \text{ true}) = P(Y \geq 38) = \text{pbinom(37, 100, 0.3886, lower=F)} \approx 0.607$$

Based on parts (a) and (b), we should prefer Rule 1 since it has the same Type I error rate as Rule 2, but has greater power as a test.

Commentary:

When you've finished this problem, it's natural to ask: Intuitively, why is Rule 1 better than Rule 2? Both are looking for evidence of a new, higher mean. The first does this by exploring the average in a sample. The second looks for overly-heavy products ($X_i > 70$) and raises a red flag when there are too many of these. These are different, but natural ways, to suss out a larger true mean. The trouble with Rule 2 is that it contains a *loss of information*. Notice that $X_i = 70.5$ and $X_i = 103$ are treated

the same way (as defective products). In Rule 1, these two values would influence the sample average differently.

In life, there is a trade-off between information richness and ease of communication. In general, the less information contained in something (e.g., a sentence, a podcast, a mathematical argument), the easier it is to communicate. As an example, when you are looking up the quality of an item on Amazon, the site could have listed every reviewer and what score they gave. Or, they could lose the information about who scored what and just report how many people gave each rating (1000 5 star reviews, 223 4 star reviews, etc.). Or, they could reduce this down to a single number (4.7 stars, on average). Always consider both the pros and cons as you collapse information-rich sources.

85. Suppose that $X \sim N(\mu, 1)$. We plan to test

$$H_0\text{: } \mu = 0 \text{ vs. } H_1\text{: } \mu = 4$$

using a random sample of size n. Show that the BCR will take the form $C = \left\{\mathbf{X} \middle| \sum X_i > c\right\}$ where c is a constant.

Hint List:

- Start by memorizing the statement of the Neyman-Pearson Lemma (NPL) and the pdf of the normal distribution. After that, everything is algebra.

Solution: Because $X \sim N(\mu, 1)$, we have

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{1 \cdot \sqrt{2\pi}} \exp\left[\frac{-(x_i - \mu)^2}{2 \cdot 1}\right] = (2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum (x_i - \mu)^2\right]$$

By the NPL, we find the BCR by setting $\dfrac{L(\mu = 0)}{L(\mu = 4)} < k$. Substituting, this gives

$$k > \frac{(2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum (x_i)^2\right]}{(2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum (x_i - 4)^2\right]} = \exp\left[-\frac{1}{2}\sum \left(x_i^2 - x_i^2 + 8x_i - 16\right)\right] = \exp\left[-\frac{1}{2}\sum (8x_i - 16)\right]$$

Taking the natural log gives

$$\ln k > -\frac{1}{2}\sum (8x_i - 16) \implies k_2 = -2\ln k < 8\sum (x_i) - 16n$$

Rearranging gives

$$c = \frac{k_2 + 16n}{8} < \sum x_i$$

Thus, $C = \left\{\mathbf{X} \,\middle|\, \sum X_i > c\right\}$ is a BCR.

Commentary: This is your typical first BCR problem. To make things gentler on students, I have provided the answer in the problem: $\left\{\mathbf{X} \,\middle|\, \sum X_i > c\right\}$. To make things harder, teachers could omit this, use more difficult pdfs for X, or find examples where the algebra is more challenging. Here, however, we are simply establishing basic procedural fluency. Make sure you are not showing too many steps (or these problems can grow out of control) and that you continue to introduce new constants $(k, k_2, k_3, \ldots, c)$ to hide messy expressions behind simple facades. People tend to start this lettering with k because of the NPL and end with c because this constant is related to a critical region. Following these conventions can help the reader.

86. Suppose that $X \sim N(\theta_1, \theta_2)$, both unknown, where θ_2 is the variance. We plan to test

$$H_0\text{: } \theta_1 = 0,\ \theta_2 = 1 \text{ vs. } H_1\text{: } \theta_1 = 1,\ \theta_2 = 3$$

Find the simplest expression you can for a BCR using a random sample $X_1, \ldots, X_n$.

Hint List:

- This problem is a more complex version of the last one. So, start there if you're stuck.

Solution:

We begin with the likelihood:

$$L(\theta_1,\theta_2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\theta_2}}\exp\left[\frac{-(x_i-\theta_1)^2}{2\theta_2}\right]=(2\pi)^{-n/2}\left(\theta_2\right)^{-n/2}\exp\left[-\frac{1}{2\theta_2}\sum(x_i-\theta_1)^2\right]$$

From the NPL, we set

$$\begin{aligned}
&\frac{L(\theta_1=0,\theta_2=1)}{L(\theta_1=1,\theta_2=3)}<k\\
\iff&\frac{1^{-n/2}\exp\left[-\frac{1}{2}\sum(x_i-0)^2\right]}{3^{-n/2}\cdot\exp\left[-\frac{1}{6}\sum(x_i-1)^2\right]}<k\\
\iff&\exp\left[-\frac{1}{6}\left(3\sum x_i^2-\sum(x_i-1)^2\right)\right]<k\cdot 3^{-n/2}=k_1\\
\iff&\sum\left(3x_i^2-x_i^2+2x_i-1\right)>-6\ln k_1=k_2\\
\iff&2\sum x_i^2+2\sum x_i-n>k_2\\
\iff&\sum x_i^2+\sum x_i>\frac{k_2+n}{2}=c
\end{aligned}$$

Thus, our BCR takes the form $C=\boxed{\left\{\mathbf{X}\,\middle|\,\sum X_i^2+\sum X_i>c\right\}}$.

Commentary:

One thing you should do as a student is study how teachers/textbooks make standard problem types (like finding the BCR) harder. If we compare this problem to the last, there are two small difficulty increases. First, we move from one parameter (μ) to two (θ_1,θ_2). This requires thinking about a multivariate likelihood. Notice that because the hypotheses are simple, they completely describe the pdfs in question, so all the parameters vanish once we look at the likelihood ratio. So, this difficulty is only temporary (unlike finding the MLE of multiple parameters at once). Second, unlike the last problem, this one does not tell us the destination in advance. Because of this, we can't check our answer and find mistakes as in the last problem. Thus, we must have the courage to put a box around $\left\{\mathbf{X}\,\middle|\,\sum X_i^2+\sum X_i>c\right\}$.

This problem also presents a nice opportunity to discuss how we display our answer. Remember that the goal in writing the BCR is to present the *simplest* expression possible. One idea is to divide the BCR by n, so we could write it as $\left\{\mathbf{X}\,\middle|\,\overline{X^2}+\overline{X}>\frac{c}{n}=c_2\right\}$. This is a correct and alternate form for the answer, but it isn't any simpler mathematically. While the bar notation might make things *appear* prettier, the same complexity lies beneath. Another faux simplification is to complete the square in the BCR to get $\left\{\mathbf{X}\,\middle|\,\sum\left(X_i+\frac{1}{2}\right)^2>\frac{n}{4}+c=c_2\right\}$. While some people might think this looks simpler,

it still contains $\sum X_i^2$ and $\sum X_i$ if you expand the square. Note that taking the square root of both sides does *not* simplify the left-hand side because of the sum.

Indeed, no matter what we try to do, all of our attempts at simplifying the answer retain two key expressions: $\sum X_i^2$ and $\sum X_i$. You might be curious if $\sum X_i^2$ is somehow related to $\sum X_i$, allowing us to get a simpler answer. Ultimately, the answer is no, for there are many samples that share the same value of $\sum X_i$ but have different values for $\sum X_i^2$. Take $\mathbf{X} = (-1, 1)$ and $\mathbf{X} = (-2, 2)$, for example. In light of this, our answer cannot be simplified further. In addition, our answer actually makes perfect sense: Since H_1 has a larger mean (which is related to $\sum X_i$) and larger variance (which is related to $\sum X_i^2$) than H_0, our BCR should prioritize H_1 when $\sum X_i^2 + \sum X_i$ gets big.

87. Suppose you take a random sample of size 12 from $X \sim N(0, \sigma^2)$, where σ^2 is unknown. You plan to test

$$H_0\text{: } \sigma^2 = 1 \text{ vs. } H_1\text{: } \sigma^2 = 3$$

Find the BCR of size $\alpha = 0.08$.

Hint List:

- Because we are given α, we can actually find the value of c that appears in the BCR.
- Notice that upon assuming H_0, $X \sim N(0,1)$. What does the RV expression in your BCR simplify to when $X \sim Z$?

Solution:

$$L(\sigma^2) = \prod_{i=1}^{12} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x_i - 0)^2}{2\sigma^2}\right] = (2\pi)^{-12/2} \left(\sigma^2\right)^{-12/2} \exp\left[\frac{-\sum x_i^2}{2\sigma^2}\right]$$

From the NPL, we explore

$$\begin{aligned}
&\frac{L(\sigma^2 = 1)}{L(\sigma^2 = 3)} = \frac{1^{-6}}{3^{-6}} \cdot \exp\left[\frac{-\sum x_i^2}{2 \cdot 1} + \frac{\sum x_i^2}{2 \cdot 3}\right] < k \\
\iff &\frac{-\sum x_i^2}{2} + \frac{\sum x_i^2}{6} < \ln\left(k \cdot 3^{-6}\right) = k_2 \\
\iff &\frac{-2\sum x_i^2}{6} < k_2 \\
\iff &\sum x_i^2 > \frac{-6k_2}{2} = c
\end{aligned}$$

So, our BCR is $C = \left\{\mathbf{X} \,\middle|\, \sum X_i^2 > c\right\}$. In this problem we know $\alpha = 0.08$, so we can actually find c. When we assume H_0, this forces $X \sim N(0,1)$, so $\sum_{i=1}^{12} X_i^2$ (seen in C) is just χ^2_{12}. Now, we know $\alpha = P_0(\mathbf{X} \in C)$, so we must solve $0.08 = P(\chi^2_{12} > c)$. In R, qchisq(0.08, df=12, lower=F) gives $c \approx 19.369$. Thus, the BCR with $\alpha = 0.08$ is

$$\boxed{C = \left\{\mathbf{X} \,\middle|\, \sum X_i^2 > 19.369\right\}}$$

Commentary:

Now that we're in our third BCR problem, the invocation of the NPL and subsequent algebra are becoming easier. To increase the difficulty, I have not given you the form of the BCR, and this problem asks you to use some given info ($\alpha = 0.08$) to find the c in the BCR. In general, finding c from α is difficult. Here, we get lucky because the assumption of H_0 collapses the BCR expression, $\sum X_i^2$, to a well-known distribution, χ^2_{12}. This will occasionally occur because some of the distributions we learn about were specifically created to summarize expressions that arise frequently. Finally, note that the BCR expression actually makes intuitive sense. The expression $\sum X_i^2$ is a "computational relative" of σ^2 (notice the role it plays in $\widehat{\sigma^2}_{\text{MLE}} = \frac{1}{n}\sum(X_i - \overline{X})^2$), so if we are deciding between H_0: $\sigma^2 = 1$ and H_1: $\sigma^2 = 3$, we should move to the alternative when this "relative" starts getting big.

88. You have decided to become an educational researcher and want to find a good model to describe test scores in a given teacher's classroom. You're choosing between two different cdfs: H_0: $F_0(x) = x^3$ and H_1: $F_1(x) = x^4$ where $0 \leq x \leq 1$ (here, $x = 0.79$ would represent a test score of 79%). Assume a sample size of 1 is drawn for all parts of this problem.

(a) For what x is the likelihood ratio less than 1?

(b) Find the general form of a BCR for this test.

(c) Find the critical region of a best test with significance level α and the power of such a test.

Hint List:

- a. Recall the relationship between the pdf and cdf of a RV.
- c. Find c in terms of α, then proceed.

Solution:

(a) Since the cdfs are given in H_0 and H_1, we can differentiate to find the pdfs: $f_0(x) = 3x^2$ and $f_1(x) = 4x^3$. Now, we use the NPL:

$$\frac{3x^2}{4x^3} < 1 \iff \frac{3}{4x} < 1 \iff \boxed{x > \frac{3}{4}}$$

Note that when $n = 1$ and $k = 1$, we are simply discovering where $\dfrac{f_0}{f_1} < k = 1$, or $f_0(x) < f_1(x)$. That is, we move to the alternative when its pdf is higher.

(b) Repeating the logic above with k instead of 1 gives:

$$\frac{3x^2}{4x^3} < k \iff \frac{3}{4x} < k \iff x > \frac{3}{4k} = c$$

So, our BCR is $\boxed{C = \{x \,|\, x > c\}}$.

(c) Since $\alpha = P_0(x \in C) = P_0(x > c) = \displaystyle\int_c^1 3x^2\,dx = 1 - c^3$, we find that $c = \sqrt[3]{1-\alpha}$.

So, $\boxed{C = \{x \,|\, x > \sqrt[3]{1-\alpha}\}}$.

Now, power $= P_1(x \in C) = \displaystyle\int_{\sqrt[3]{1-\alpha}}^1 4x^3\,dx = \boxed{1 - (1-\alpha)^{4/3}}$.

Commentary: I like this problem because everything feels slightly off when you do it. First, there are no parameters being discussed. Here, we are comparing two different distributions. On top of that, cdfs were given, not pdfs, so you had to differentiate before using the NPL. Second, the math is much easier when $n = 1$, and this bothers students. Indeed, despite this being one of the algebraically easiest questions in the whole course, it is one of the most asked-about in office hours because students simply don't trust that so little work could be required. As mentioned before, stop with the academic narrative that your work should invoke a certain level of pain. Let the problem dictate how much elbow grease will be needed, not some voice from your soul based on high school trends.

Third, part (c) is interesting in that we can express c (from the BCR) in terms of α, and then find the power in terms of α. Notice in the answer that as α gets bigger, so does the power. We first saw this relationship when power was introduced. This problem makes that relationship algebraically visible. Finally, you may have been bothered by the similarity in your steps in parts (a) and (b). Indeed, I simply copied, pasted, and changed my solution from one part to the next. The reason this should bother you is that part (a) is just part (b) with $k = 1$. Because of this, you could do the general case first (giving the answer to part (b)), and then simply plug in $k = 1$ to get the answer to (a). Indeed, this is how I did parts (a) and (b) of Problem 27. In the current problem, I "re-invented the wheel" with all the same steps so you could sit in that inefficiency and be bothered.

89. You and a friend are arguing about what RV X better models your data. You claim H_0: $X \sim Geom(3/4)$, while your friend is excited about some obscure distribution H_1: $X \sim Yule{-}Simon(\rho = 2)$. Using Wikipedia, you see that the pmf for the Yule-Simon distribution is

$$f(x;\rho) = \frac{\rho \cdot \rho! \cdot (x-1)!}{(x+\rho)!} \text{ where } x = 1, 2, 3, \ldots \text{ and } \rho \text{ is some integer}$$

(a) Using R, draw a single plot of the pmfs for each hypothesis on $\{1, \ldots, 10\}$. Use black filled-in circles for the H_0 distribution, and red ones for H_1. Include your code and a sketch/screenshot of both pmfs on the same set of axes. Note that the Geometric distribution built into R only counts the number of failures leading to the success, which is different than how we defined the Geometric distribution, so you'll need to fiddle with it to give the results we're expecting. Type ?dgeom to learn more. Also, you should not hunt down a package with the Yule-Simon distribution. Instead, use R's vectorized abilities to create any probabilities you need.

(b) What BCR do you get when drawing a sample of size 1 and using $k = 1$ in the NPL? In addition to doing the math, you should explain how you can use your picture from part (a) to check your answer.

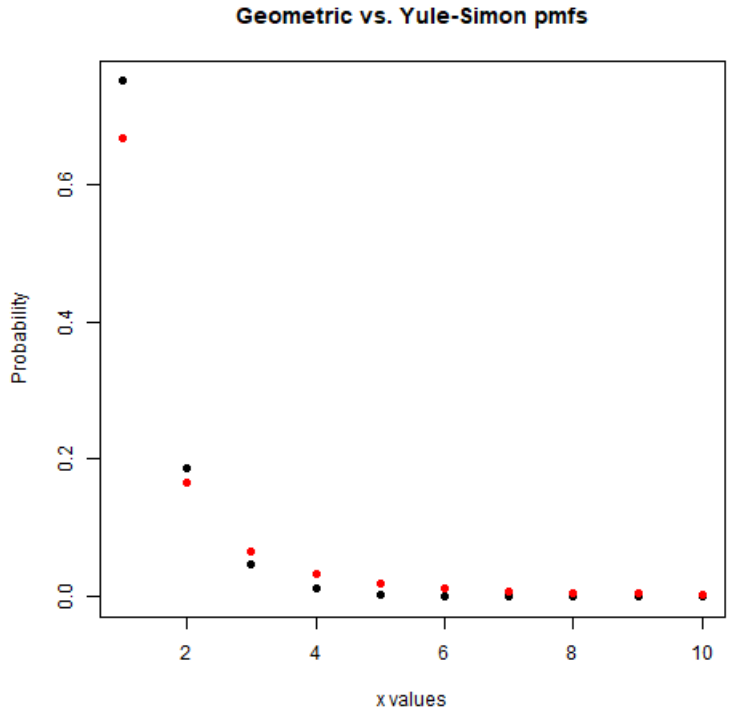
Hint List:

- a. Use factorial(...) in R.

Solution:

(a)
```
xs = 1:10
geomys = dgeom(xs-1, prob=0.75)
plot(xs, geomys, pch=16, main="Geometric vs. Yule-Simon pmfs",
     xlab="x values", ylab="Probability")

rho = 2
yuleys = rho*factorial(rho)*factorial(xs-1)/factorial(xs+rho)
points(xs, yuleys, pch=16, col="red")
```



(b) When $n = 1$ and $k = 1$, the NPL gives us

$$\frac{\left(\frac{1}{4}\right)^{x-1} \cdot \frac{3}{4}}{\frac{2 \cdot 2! \cdot (x-1)!}{(x+2)!}} < 1$$

$$\iff \frac{3(x+2)!}{4^{x-1} \cdot 4 \cdot 2 \cdot 2! \cdot (x-1)!} < 1$$

$$\iff \frac{3(x+2)(x+1)x}{4^{x+1}} < 1$$

Since the numerator is a cubic polynomial and the denominator is an exponential function, we know as x gets larger, the ratio will eventually be less than 1. We can plug in several values of

x to discover that this happens when $x \geq 3$. So, our BCR is $\boxed{\{x \,|\, x \geq 3\}}$. This makes sense in our picture. Since $n = 1$ and $k = 1$, we really just checked where $\dfrac{\text{Geom pmf}}{\text{Y-S pmf}} < 1$, or Geom pmf $<$ Y-S pmf. We can see the Yule-Simon pmf (red dots) is higher than the Geometric pmf (black dots) precisely when $x \geq 3$.

Commentary: This problem does a few things for us. First, it is example of us running a hypothesis test to explore two different distributions. In the past, our tests compared different parameter values for the *same* distribution. If you've ever wondered how you might decide between competing model choices, this problem offers a way forward. Second, I wanted to introduce you to yet another distribution (Yule-Simon) and have you realize that you don't need R to have everything built in. If you want probabilities from new distributions, you can write a line of code to make these yourself. Indeed, it actually takes *more* time to hunt down a package with Yule-Simon and learn the syntax for using it. Finally, I wanted you to see that the NPL is actually quite intuitive when $n = 1$ and $k = 1$. It is asking you to move to H_1 when the pmf/pdf for H_1 is higher than that of H_0. This should feel comforting: When tools collapse to sensible things in special cases, it's a sign that they are reasonable generalizations.

90. Let $X \sim Bernoulli(p)$. We wish to test

$$H_0\text{: } p = \frac{1}{3} \text{ vs. } H_1\text{: } p < \frac{1}{3}$$

using a random sample of size 11. Let $C = \left\{ \mathbf{X} \middle| \sum_{i=1}^{11} X_i < c \right\}$. Show that C is an UMPCR and find its α when $c = 2.3$.

Hint List:

- It helps to hide $\sum x_i$ behind the letter s when doing this problem.
- Recall that $b^{11-s} = b^{11}b^{-s}$.
- What distribution describes the sum of Bernoulli RVs?

Solution: Let $0 < a < \frac{1}{3}$. We test

$$H_0\text{: } p = \frac{1}{3} \text{ vs. } H_1\text{: } p = a$$

Here, $L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{11-\sum x_i} = p^s(1-p)^{11-s}$ where $s = \sum x_i$. Invoking the NPL, we get

$$\begin{aligned}
&\frac{L\left(p=\frac{1}{3}\right)}{L\left(p=a\right)} = \frac{\left(\frac{1}{3}\right)^s\left(\frac{2}{3}\right)^{11-s}}{a^s(1-a)^{11-s}} < k \\
\iff &\left(\frac{1}{3a}\right)^s\left(\frac{2}{3(1-a)}\right)^{11-s} < k \\
\iff &\left(\frac{1}{3a}\right)^s\left(\frac{2}{3(1-a)}\right)^{-s}\left(\frac{2}{3(1-a)}\right)^{11} < k \\
\iff &\left(\frac{1-a}{2a}\right)^s < k\left(\frac{3(1-a)}{2}\right)^{11} = k_1 \\
\iff &s \cdot \ln\left[\frac{1-a}{2a}\right] < \ln k_1 = k_2
\end{aligned}$$

Since $0 < a < \frac{1}{3}$, then $1-a > \frac{2}{3}$ and $2a < \frac{2}{3}$. Thus, $\frac{1-a}{2a} > \frac{2/3}{2/3} = 1$. So, $\ln\left[\frac{1-a}{2a}\right] > 0$. Dividing by this above keeps the inequality direction, and we get $s < \frac{k_2}{\ln\left[\frac{1-a}{2a}\right]}$, or $\sum x_i < c$, as claimed.

If $c = 2.3$, we have: $\alpha = P_0(\mathbf{X} \in C) = P\left(\sum X_i < 2.3 \,\middle|\, H_0 \text{ true}\right) = P(Binom(11, 1/3) < 2.3) =$ pbinom(2, 11, 1/3) $\approx \boxed{0.234}$.

Commentary: Here, we increase the difficulty a few more notches. In Problem 87, you were given α and asked to find c. Here, you are given c and asked to find α. As before, we are saved because the test stat, $\sum X_i$, has a well-known distribution (Binomial). Asking for this extra step at the end adds to the cognitive burden. On top of that, this problem has more difficult algebra than previous examples, and we are finding an UMPCR, not just a BCR (because the alternative is composite, rather than simple). The most important step, which is commonly overlooked, is the division by $\ln\left[\frac{1-a}{2a}\right]$. When you are finding an UMPCR, you will have chosen a variable (here, a) to stand for an arbitrary value in the alternative hypothesis. If you ever divide by an expression that contains a, you must be sure of the positive/negative sign of that expression because this affects the direction of the inequality. Often, it can be difficult to determine this at first glance, and a careful analysis is *always* needed.

91. Suppose X is a RV with pdf $f(x;\lambda) = \frac{1}{2}\lambda e^{-\lambda|x|}$ where $x \in \mathbb{R}$ and $\lambda > 0$. Let $X_1, \ldots, X_n$ be a random sample collected to test

$$H_0\text{: } \lambda = \lambda_0 \text{ vs. } H_1\text{: } \lambda < \lambda_0$$

(a) Find a general form for the UMP test.

(b) Explain why no UMP test exists for H_0: $\lambda = \lambda_0$ vs. H_1: $\lambda \neq \lambda_0$.

(c) Show that $|X| \sim Exp(\lambda)$.

(d) Using parts (a) and (c), find the UMPCR with $\alpha = 0.03$ when $n = 16$ and $\lambda_0 = 2$. Then, find the power of the UMP test based on this UMPCR if λ is actually 1.

Hint List:

- c. Try showing the CDFs of $|X|$ and $Exp(\lambda)$ are the same.
- d. The UMPCR from part (a) should get *much* simpler once you have the result from part (c).

Solution:

(a) Let $0 < a < \lambda_0$. We test

$$H_0\text{: } \lambda = \lambda_0 \text{ vs. } H_1\text{: } \lambda = a$$

Here, $L(\lambda) = \prod_{i=1}^{n} \frac{1}{2}\lambda e^{-\lambda|x_i|} = \frac{\lambda^n}{2^n} \exp\left[-\lambda \sum |x_i|\right]$. The NPL gives us

$$\frac{L(\lambda = \lambda_0)}{L(\lambda = a)} = \frac{\frac{\lambda_0^n}{2^n} \exp\left[-\lambda_0 \sum |x_i|\right]}{\frac{a^n}{2^n} \exp\left[-a \sum |x_i|\right]} = \left(\frac{\lambda_0}{a}\right)^n \exp\left[(a - \lambda_0) \sum |x_i|\right] < k$$

$$\iff (a - \lambda_0) \sum |x_i| < \ln\left[k \cdot \left(\frac{a}{\lambda_0}\right)^n\right] = k_1$$

$$\iff \sum |x_i| > \frac{k_1}{a - \lambda_0} = c, \text{ since } a - \lambda_0 < 0 \text{ because } a < \lambda_0$$

Since we get the same form for the BCR regardless of the particular value for a, we have an UMPCR of $\boxed{C = \left\{\mathbf{X} \,\middle|\, \sum |X_i| > c\right\}}$. Thus, our UMP test is: If $\mathbf{X} \in C$, reject H_0; else keep H_0.

(b) At the end of part (a), we had to divide both sides of the inequality by $a - \lambda_0$. Notice that in a two-sided alternative hypothesis (H_1: $\lambda \neq \lambda_0$), we would also need to consider the case of $a > \lambda_0$. In that case, $a - \lambda_0 > 0$, and after division by $a - \lambda_0$, the inequality is now reversed from part (a), giving $\sum |x_i| < c$. Thus, it isn't possible to get a consistent form for the BCR. Hence, there is no UMPCR/UMP test.

(c) One way to show two RVs have the same distribution is to show they have the same CDF. Since $|X| \geq 0$, fix a value of $x \geq 0$.

$$\begin{aligned}
F_{|X|}(x) &= P(|X| \leq x) = P(-x \leq X \leq x) \\
&= \int_{-x}^{x} \frac{1}{2}\lambda e^{-\lambda|t|}\, dt \\
&= 2\int_{0}^{x} \frac{1}{2}\lambda e^{-\lambda|t|}\, dt \text{ (integrand is an even function)} \\
&= \int_{0}^{x} \lambda e^{-\lambda t}\, dt \text{ (absolute value unneeded since } 0 \leq t \leq x) \\
&= F_{Exp(\lambda)}(x)
\end{aligned}$$

Thus, $|X| \sim Exp(\lambda)$.

(d) Using our result from part (c), we can simplify the UMPCR in part (a):

$$C = \left\{ \mathbf{X} \,\middle|\, \sum |X_i| > c \right\} = \left\{ \mathbf{X} \,\middle|\, \sum Exp(\lambda) > c \right\} = \{ \mathbf{X} \,|\, Gamma(n, \lambda) > c \}$$

Assuming H_0, we get $\lambda_0 = 2$, and thus

$$0.03 = \alpha = P_0(\mathbf{X} \in C) = P(Gamma(n = 16, \lambda = 2) > c)$$

So, $c =$ qgamma(0.03, 16, 2, lower=F) ≈ 12.16027.

Now, power $= P_1(\mathbf{X} \in C) = P(Gamma(n = 16, \lambda = 1) > 12.16027)$. Using R, we find that pgamma(12.16027, 16, 1, lower=F) $\approx \boxed{0.833}$.

Commentary:

I love problems with multiple parts because they can take you on interesting journeys. Often, they show new directions you can take a concept that you're studying in class. While I don't share this in the above problem statement, the given random variable X is known as a Laplace distribution, which can be useful when studying Brownian motion (i.e., the random movement of a particle in a gas or liquid). For this distribution, λ is a rate parameter which can be useful in understanding that random movement. If you wanted to create a hypothesis test for λ, the UMP test in part (a) is the first natural starting point. This, then, inspires part (b). In general, developing a test for a one-sided hypothesis is slightly easier than for a two-sided alternative, which is why we proceed from (a) to (b). As you look at the answer to (a), it is natural to ask: What is the distribution of $\sum |X_i|$? Here, knowing some probability theory really helps (i.e., the absolute value of Laplace is Exponential; the sum of Exponentials is Gamma). Once we put everything together, we have a really good testing infrastructure for a brand new setting! Part (b) continues to nag at us about how to test two-sided alternatives, but we remedy that in Problem 93. While it's easy to get lost in the algebra/details of this problem, you should take a moment to look back and reflect on what we just accomplished. Essentially, we took a RV you have never seen before and built an entire testing infrastructure in about one page of work. This is the power of the NPL/UMP system. My hope is that after we do this a few more times, you'll start to believe that you could do this with almost *any* new setting. When you reach that point, you've come a long way from the intro-level student who only knows how to study μ in $N(\mu, \sigma^2)$ and p in $Binom(n, p)$.

92. Let X be a RV with pdf $f(x;\theta) = \dfrac{m}{\theta} x^{m-1} e^{-x^m/\theta}$, where $x > 0$, $\theta > 0$, and m is some known, positive constant. Draw a random sample $X_1, \ldots, X_n$ to test

$$H_0\text{: } \theta = \theta_0 \text{ vs. } H_1\text{: } \theta > \theta_0$$

(a) Find the UMPCR.

(b) Suppose $n = 10$ and $\theta_0 = 8$. Find the UMPCR that has significance level $\alpha = 0.02$.

Hint: It helps to find the pdf for $Y = X^m$. To do so, use this result (sometimes taught in Math 180A): If X is a RV with pdf $f_X(x)$ and h is either always increasing or decreasing, then $Y = h(X)$ has pdf $f_Y(y) = f_X(u) \cdot \left| \dfrac{du}{dy} \right|$ where $u = h^{-1}(y)$.

Hint List:

- b. Try $y = h(x) = x^m$, so $h^{-1}(y) = y^{1/m}$.

Solution:

(a) Let $a > \theta_0$ and use the NPL to test H_0: $\theta = \theta_0$ vs. H_1: $\theta = a$. Now,

$$L(\theta) = \prod_{i=1}^{n} \frac{m}{\theta} x_i^{m-1} \exp\left(\frac{-x_i^m}{\theta}\right) = \frac{m^n}{\theta^n}\left(\prod x_i\right)^{m-1} \exp\left(\frac{-1}{\theta}\sum x_i^m\right)$$

So, we get $\dfrac{L(\theta_0)}{L(a)} = \dfrac{a^n}{\theta_0^n}\exp\left(\left[\dfrac{1}{a} - \dfrac{1}{\theta_0}\right]\sum x_i^m\right) < k \iff \left[\dfrac{1}{a} - \dfrac{1}{\theta_0}\right]\sum x_i^m < \ln\left(\dfrac{k\theta_0^n}{a^n}\right) = k_2$

Since $a > \theta_0$, $\dfrac{1}{a} < \dfrac{1}{\theta_0}$, and thus, $\dfrac{1}{a} - \dfrac{1}{\theta_0} < 0$. Dividing gives $\sum x_i^m > \dfrac{k_2}{\dfrac{1}{a} - \dfrac{1}{\theta_0}} = c$.

Since we get this inequality for any $a > \theta_0$, the form of the BCR is consistent, and hence we have an UMPCR: $\boxed{C = \left\{\mathbf{X} \,\middle|\, \sum X_i^m > c\right\}}$.

(b) Let $Y = X^m$. Since $m > 0$, we know $h(x) = x^m$ is an increasing function, and we can use the given hint. Here, $y = x^m$, so $x = y^{1/m} = h^{-1}(y)$. Thus,

$$\begin{aligned} f_Y(y) &= f_X(h^{-1}(y))\left|\frac{d}{dy}h^{-1}(y)\right| = \frac{m}{\theta}\left(y^{1/m}\right)^{m-1}\exp\left[\frac{-(y^{1/m})^m}{\theta}\right]\cdot\left|\frac{1}{m}y^{(1/m)-1}\right| \\ &= \frac{1}{\theta}y^{1-(1/m)}\cdot\exp\left[\frac{-y}{\theta}\right]y^{(1/m)-1} = \frac{1}{\theta}e^{-y/\theta} \sim Exp\left(\frac{1}{\theta}\right) \end{aligned}$$

So, $\displaystyle\sum_{i=1}^{n} X_i^m = \sum_{i=1}^{n} Exp\left(\frac{1}{\theta}\right) \sim Gamma\left(n, \frac{1}{\theta}\right)$.

Thus, $0.02 = P_0(\mathbf{X} \in C) = P(Gamma(10, 1/8) > c)$.

We get $c = \text{qgamma}(0.02,\ 10,\ 1/8,\ \text{lower=F}) \approx 140.079$. So, our UMPCR is

$$\boxed{C = \left\{\mathbf{X} \,\middle|\, \sum_{i=1}^{10} X_i^m > 140.079\right\}}$$

Commentary: The real reason I assign this problem is part (b). Once you get the answer to part (a), it raises the natural question: Does X_i^m have a nice distribution? Here, the answer is yes, and because $X_i^m \sim Exp\left(\dfrac{1}{\theta}\right)$, the sum in the UMPCR also has a nice distribution. Once again, the Gamma distribution is our friend. How many times has it trivialized a horrible integral (via the Kernel Technique) or simplified a messy looking critical region? Love the Gamma.

93. According to 91b, there is no UMP test for

$$H_0\colon\ \lambda = \lambda_0 \text{ vs. } H_1\colon\ \lambda \neq \lambda_0$$

when studying λ from $f(x;\lambda) = \frac{1}{2}\lambda e^{-\lambda|x|}$. Find the generalized likelihood ratio test (GLRT) for this setup, simplifying your answer as much as possible.

Hint List:

- Remember to find the MLE first. Also, your answer will be a complicated expression where $\sum |X_i|$ appears more than once.

Solution:

Here, $\omega = \{\lambda_0\}$ and $\Omega = (0, \infty)$. Because Ω contains all possible values for λ, we'll need the MLE in the generalized likelihood ratio.

$$L(\lambda) = \prod_{i=1}^{n} \frac{1}{2}\lambda \cdot \exp\left(-\lambda|x_i|\right) = \frac{\lambda^n}{2^n} \cdot \exp\left(-\lambda \sum |x_i|\right) \implies \ell(\lambda) = n \ln \lambda - \lambda \sum |x_i| + C$$

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum |x_i|$$

$$\sum |x_i| = \frac{n}{\widehat{\lambda}}$$

$$\widehat{\lambda} = \frac{n}{\sum |x_i|} = \frac{1}{\overline{|X|}}$$

$$\ell''(\lambda) = -\frac{n}{\lambda^2}$$

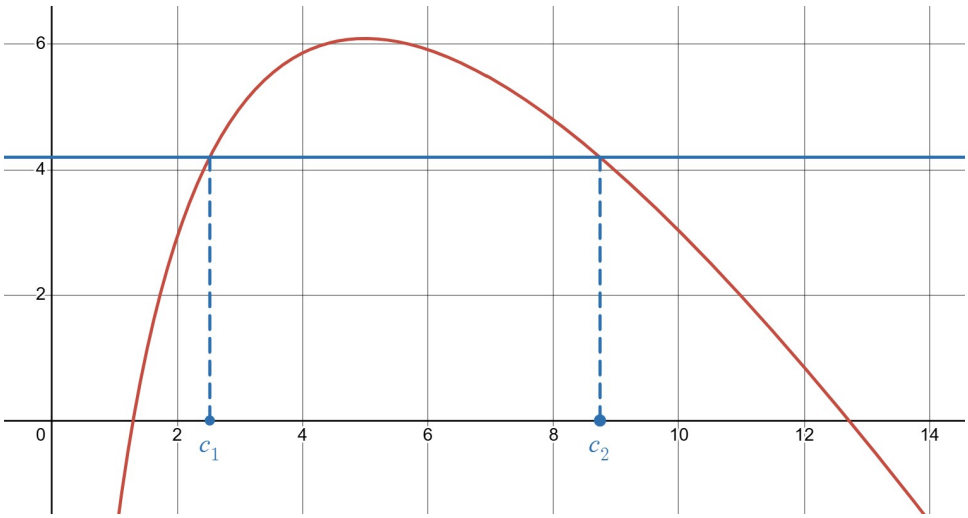
Note that $\ell'' < 0$ for all $\lambda > 0$

$$\text{Now, GLR} = \frac{\max\limits_{\omega} L(\lambda; \vec{x})}{\max\limits_{\Omega} L(\lambda, \vec{x})} = \frac{L(\lambda_0)}{L(\widehat{\lambda}_{\text{MLE}})} = \frac{\left(\frac{\lambda_0}{2}\right)^n \cdot \exp\left(-\lambda_0 \sum |x_i|\right)}{\left(\frac{n/\sum |x_i|}{2}\right)^n \cdot \exp\left(-\frac{n}{\sum |x_i|} \sum |x_i|\right)} < k$$

Simplifying gives $\dfrac{\lambda_0^n \left(\sum |x_i|\right)^n}{n^n} \cdot \exp\left(-\lambda_0 \sum |x_i|\right) \cdot e^n < k \iff \left(\sum |x_i|\right)^n \cdot \exp\left(-\lambda_0 \sum |x_i|\right) < k_2.$

Taking the natural log gives: $n \ln\left(\sum |x_i|\right) - \lambda_0 \sum |x_i| < \ln k_2 = c$. For this class, you can leave the answer like this. So, our critical region is $\boxed{C = \left\{\mathbf{X} \,\middle|\, n \ln\left(\sum |X_i|\right) - \lambda_0 \sum |X_i| < c\right\}}$.

In graduate school, you may be asked to simplify this further. We won't do so in this class, but here is what that looks like if you're curious. If we let $T = \sum |X_i|$, then we're trying to get $n \ln T - \lambda_0 T < c$. If we switch to familiar letters, we are trying to get $y = n \ln x - \lambda_0 x$ to fall below $y = c$. On the next page is a picture of what that looks like for $n = 10, \lambda_0 = 2$, and $c = 4.2$ (arbitrarily chosen). As you can see, if $x < c_1$ or $x > c_2$, then the graph falls below $y = c$. So, we could list the critical region as $C = \left\{T = \sum |X_i| \,\middle|\, T < c_1 \text{ or } T > c_2\right\}$. Listing your answer this way makes some intuitive sense. Generally, for a two-sided alternative hypothesis, we start moving to the alternative if our data get too big (supporting the $>$) or too small (supporting the $<$), which is what C is basically telling us. The frustrating thing about this answer is that c_1 and c_2 depend on n, λ_0, and c. So, they aren't simple constants, like they might appear. A more honest answer is: $C = \left\{T = \sum |X_i| \,\middle|\, T < c_1(n, \lambda_0, c) \text{ or } T > c_2(n, \lambda_0, c)\right\}$. This reminds the reader of the dependency of c_1 and c_2 on those other values (by using function notation). At this point, you probably prefer the original answer boxed above, which is why I let you stop there.



Commentary:

There are few classes in the math department quite like Mathematical Statistics. As you have likely noticed by this point, the course requires proving/deriving results, working with actual data and real-world settings, writing code, and creating visualizations. These four skills are the focus of four different majors: theoretical (AKA pure) math/statistics, applied math/statistics, computer science, and information visualization. Since these are four of my favorite things in the world, it's no surprise this is my favorite class to teach. But the class also provides *you* with a special opportunity. What aspect(s) of this class have you enjoyed the most? The above problem, for example, is largely in the pure math-/stats bucket. Essentially, we're deriving the rule for a test (H_0: $\lambda = \lambda_0$ vs. H_1: $\lambda \neq \lambda_0$) about the parameter (λ) of a distribution (Laplace) that you've likely never seen or used before. There's no real-world setting tied to what we're doing and no work behind the computer to process or display actual data based on the distribution. If this problem felt great, it's a sign you should be a pure math major or probability/statistics major. If you need data sets about pH levels, toilet paper thicknesses, and football passing totals, then applied math or data science is likely a better fit. If you enjoy the problems that require the computer, then CS is your wheelhouse. In truth, there is a lot of overlap in the Venn diagram for these disciplines, which is why choosing a major can be so difficult. Hopefully, seeing all these skill sets in the *same* class provides a unique opportunity to find your life direction.

94. Suppose you draw a random sample $X_1, \ldots, X_n$ from a CRV with density

$$f(x;\theta) = \frac{x}{\theta^2}\exp\left(\frac{-x^2}{2\theta^2}\right), \text{ where } x \geq 0, \theta > 0$$

(a) Find $\widehat{\theta}_{\text{MLE}}$.

(b) Find the GLRT to test H_0: $\theta = 1$ vs. H_1: $\theta \neq 1$, simplifying as much as you can.

Hint List:

- b. Don't be afraid of the algebra nightmare. Indeed, you should expect this on GLRT problems.

Solution:

(a) One approach to this problem is to go through the usual steps. If you do, you'll get

$$\widehat{\theta}_{\text{MLE}} = \sqrt{\frac{\sum x_i^2}{2n}} = \sqrt{\frac{\overline{X^2}}{2}}$$

Another approach is to look back at Problem 9 and notice the similarity of the pdf there and this pdf. Indeed, if we replace θ in Problem 9 with $2\theta^2$, we convert that pdf into this one. Equivalently, if we replace θ in this one with $\sqrt{\frac{\theta}{2}}$, we get the pdf from Problem 9. So, you might wonder if there is a way to adjust the answer from Problem 9 to quickly get the answer here. In graduate-level math stats, you will learn the **Invariance Property (IP) of MLEs:**

> If $\widehat{\theta}$ is the MLE of θ, then for any function g, the MLE of $g(\theta)$ is $g(\widehat{\theta})$ (i.e., $\widehat{g(\theta)} = g(\widehat{\theta})$)

In Problem 9,we showed that the MLE is $\widehat{\theta}_9 = \overline{X}$. Combining this with the above observations, we can find the MLE here:

$$\widehat{\theta}_{94} = \widehat{\sqrt{\frac{\theta_9}{2}}} \overset{\text{IP}}{=} \sqrt{\frac{\widehat{\theta}_9}{2}} = \sqrt{\frac{\overline{X^2}}{2}}$$

(b) Here, $L(\theta) = \prod_{i=1}^{n} \frac{x_i}{\theta^2} \cdot \exp\left[\frac{-x_i^2}{2\theta^2}\right] = \frac{\prod x_i}{\theta^{2n}} \cdot \exp\left[-\frac{1}{2\theta^2}\sum x_i^2\right]$.

Thus, the GLR is

$$\Lambda = \frac{\max\limits_{\theta\in\{1\}} L(\theta)}{\max\limits_{\theta\in(0,\infty)} L(\theta)} = \frac{L(1)}{L(\widehat{\theta}_{\text{MLE}})} = \frac{\prod x_i \cdot \exp\left[-\frac{1}{2}\sum x_i^2\right]}{\frac{\prod x_i}{\left(\frac{\sum x_i^2}{2n}\right)^n} \cdot \exp\left[-\frac{1}{2\cdot\frac{\sum x_i^2}{2n}} \cdot \sum x_i^2\right]}$$

$$= \frac{\left(\sum x_i^2\right)^n}{(2n)^n} \cdot \exp\left[n - \frac{1}{2}\sum x_i^2\right]$$

We have $\Lambda < k \iff n\ln\left(\sum x_i^2\right) + n - \frac{1}{2}\sum x_i^2 < \ln\left[(2n)^n k\right] = k_2$. Thus, our critical region is

$$\boxed{C = \left\{\mathbf{X} \,\middle|\, n\ln\left(\sum x_i^2\right) - \frac{1}{2}\sum x_i^2 < c = k_2 - n\right\}}$$

Commentary:

One thing I often tell my classes is that you can't "re-invent the wheel" in math too often if you want to do things at a high level. This is the reason we have young students memorize their times tables. If your adult approach to evaluating $5 \cdot 6$ is to return to the original definition of multiplication as repeated addition $(5+5+5+5+5+5)$, then you'll never be able to finish a single Calculus problem because you'll be too busy re-inventing every basic fact you need along the way. One reason I include this problem on the homework is that you've seen similar pdfs on many previous homework problems. (Don't believe me? Go back and look; you might be surprised.) This awareness will lead to some natural questions:

- Have I already found the MLE for this exact pdf? (answer: no)
- Is there anything close enough? (answer: maybe Problem 9)
- How do you adapt MLEs? (answer: internet search to find the IP!)

The second reason I assign this problem is that part (b) is bringing you right to the doorway of graduate level math stats. The next level will demand this algebraic sophistication. If you could do this problem, you're ready.

95. Using this Desmos link, experiment to find the answers to these questions related to the Beta distribution:

(a) For what pair(s) (a, b) does $Beta(a, b)$ become $Unif(0, 1)$?

(b) For what pair(s) (a, b) will the mode be at $\theta = 0.5$?

(c) If you had a *very strong* belief that $\theta = 0.3$, but would consider other values, what a and b might be reasonable to use when setting up a distribution to model these prior beliefs using the mode?

Hint List:

- a and b. Play around in Desmos. Then, see if you can convince yourself *algebraically*.

Solution:

(a) The only values of a and b that create a uniform distribution are $\boxed{(a = 1, b = 1)}$. You can discover this through trial and error using the Desmos link, or you can think about the pdf of the Beta distribution: $f(\theta) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$. Since the uniform has a constant pdf, we need the terms with θ to vanish. This forces $a - 1 = 0$ and $b - 1 = 0$, giving the answer above.

(b) Trial and error in Desmos reveals the mode is at $\theta = 0.5$ when $\boxed{a = b \geq 1}$ (or $a = b > 1$, depending on how you define the "mode" for constant functions). We can derive this algebraically, if desired. According to the internet, if $a, b > 1$, then the mode is at $\dfrac{a-1}{a+b-2}$. Setting this equal to $\dfrac{1}{2}$ gives $2(a-1) = a + b - 2 \iff 2a = a + b \iff a = b$.

(c) Here, we set $\dfrac{a-1}{a+b-2} = \dfrac{3}{10} \implies a = \dfrac{3b+4}{7}$. Exploring in Desmos, we find that the larger we make b, the more concentrated the distribution is around the mode of 0.3. Because the problem mentions we have a "very strong" belief in 0.3, we choose b large. For example, we might choose $b = 40$, so $a = \dfrac{3 \cdot 40 + 4}{7} = \dfrac{124}{7}$. Answers will vary.

Commentary: For some, this problem is very unsettling. In parts (a) and (b), you were asked to play around with a pdf. It's been a while since you were in a sandbox; how did it feel to go back? If you look at how we build scholars in this world, we tend to remove all elements of *mathematical play* from academia. Instead of asking students to explore what can be done with some definition, theorem, or tool, we just begin building out the major things that follow from that definition, theorem, or tool. For example, in a number theory class, as soon as students learn the "evenly-divisible" operator, $a|b$, teachers begin presenting theorems like: If $a|b$ and $a|c$, then $a|(b+c)$. In Geometry class, when students first gain exposure to the straight-edge and compass, teachers immediately begin covering the major constructions (bisecting a segment, bisecting an angle, etc.). What would education look like if we let students play with these things for a while before rattling off the major results that follow from them?

The worst part of this trend is that children are amazingly curious *as a default*. They want to play around with anything you give them (toys, food, language, ideas, etc.). When we push these naturally-curious people through education, all that goes away. We graduate high school seniors who want to be shown procedures and then reproduce them on exams. If these exams diverge from the standard problem types drilled endlessly, the unspoken academic contract between students, teachers, and parents has been broken, and chaos ensues. On top of this, it is critical that all problems have a common answer and that students should not have to make any value judgments when arriving at said answer. You may have noticed that part (c) forces you to decide what a "very strong" belief means in terms of the parameters a and b. Obviously, this term is not well-defined, and hence, answers will vary. As you might guess, students are upset when any element of a problem involves subjective choices. Did this bother you? Also, has your world devolved from play to procedures?

96. When information (coded as 0s and 1s) is sent across the internet, it can occasionally become corrupted (i.e., a 0 sent becomes a 1 received, or a 1 sent becomes a 0 received). Suppose that over a given connection, the ratio of 0s to 1s sent out is 2:7. Also, suppose that a 0 will be corrupted to a 1 with probability 1/3, and a 1 will be corrupted to a 0 with probability 1/5. If a 1 is received by a server, what is the probability that a 1 was actually sent by the sender?

Hint List:

- Try writing some of the given info using conditional probabilities. Use phrases like "0 received" and "1 sent".

Solution:

We know $P(0 \text{ sent}) = \dfrac{2}{2+7} = \dfrac{2}{9}$ and $P(1 \text{ sent}) = \dfrac{7}{9}$. Also, we are given that

$$P(1 \text{ received} \,|\, 0 \text{ sent}) = \frac{1}{3} \text{ and } P(0 \text{ received} \,|\, 1 \text{ sent}) = \frac{1}{5}$$

Here, we want

$$\begin{aligned} P(1 \text{ sent} \,|\, 1 \text{ received}) &\overset{\text{Bayes}}{=} \frac{P(1 \text{ received} \,|\, 1 \text{ sent}) \cdot P(1 \text{ sent})}{P(1 \text{ received})} \\ &= \frac{\left(1 - \dfrac{1}{5}\right) \cdot \dfrac{7}{9}}{P(1 \text{ received} \,|\, 1 \text{ sent}) \cdot P(1 \text{ sent}) + P(1 \text{ received} \,|\, 0 \text{ sent}) \cdot P(0 \text{ sent})} \\ &= \frac{\dfrac{4}{5} \cdot \dfrac{7}{9}}{\dfrac{4}{5} \cdot \dfrac{7}{9} + \dfrac{1}{3} \cdot \dfrac{2}{9}} = \boxed{\frac{42}{47} \approx 0.894}. \end{aligned}$$

Commentary:

I liked this problem the moment I saw it. When you're using conditional probabilities, it's very easy to confuse $P(A \,|\, B)$ with $P(B \,|\, A)$. Indeed, many people (roughly 80% in some studies) incorrectly think these probabilities are the same (known as the "Inverse Probability Fallacy"). If they were the same, Bayes's Theorem would not exist since we would just have $P(A \,|\, B) = P(B \,|\, A)$. The reason this problem is great is that it features two "inverse" probabilities:

$$P(1 \text{ received} \,|\, 1 \text{ sent}) \text{ and } P(1 \text{ sent} \,|\, 1 \text{ received})$$

What's neat in this problem is that the *human being* thinking of each of these probabilities is different. The person sending the message is worried about the first: "I sent a 1. Will they receive it?" The person receiving the message will care about the second: "I got a 1. Is that what was actually sent?" Once you notice that different people care about each of these, you start to *feel* the probabilities may be different.

Math and stats are full of common incorrect ways of thinking, known as fallacies or misconceptions. One thing you should do as a student is make note of these as faculty, textbooks, and problems bring them up. Moreover, you should seek out examples that help eradicate them from your brain for good. The above problem is one such example for me. I also use a simple example to remind myself that inverse probabilities need not be equal. Compare:

$$P(\text{person in California} \,|\, \text{person in San Diego}) \text{ and } P(\text{person in San Diego} \,|\, \text{person in California})$$

97. We often speak about unfair coins in this class, and you might wonder if it is possible to create such a monster. It turns out that it's not so hard if you're willing to bend your coin so it has a U-shape. Read this post to get started.

    Suppose I've bent a coin and want to explore p_{heads} using $Beta(a, b)$ to create a non-informative prior. If we flip the coin 200 times and get 140 tails, what is the posterior distribution? What is the posterior distribution if the prior is $Beta(10, 10)$? If the prior is $Beta(a, b)$?

Hint List:

- Try first deriving the general updating rule. That is, if our prior is $Beta(a, b)$, then find the ?s in the posterior: $Beta(?, ?)$.

Solution:

In several problems now, we've seen that if we do the general case first, we can then plug in particular values the problem asks about. This will avoid re-inventing the wheel.

So, let $p \sim Beta(a, b)$ model our prior belief about the heads probability. Also, let $X \sim Binom(200, p)$ be the number of heads when flipping the unfair coin 200 times. We update our beliefs:

$$g(p \mid x) \propto f(x; p) \cdot g(p) \propto \binom{200}{x} p^x (1-p)^{200-x} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} \propto p^{a+x-1}(1-p)^{200-x+b-1}$$

We recognize this last expression as the kernel of $Beta(a + x, 200 - x + b)$. Thus, our updating rule is

$$Beta(a, b) \to Beta(a + x, 200 - x + b)$$

Since we got 140 tails, we know $x = 60$ (heads). This info and the above rule allow us to particularize to the cases we are asked about. Recall that a non-informative prior is simply $Unif(0, 1) = Beta(1, 1)$:

$$\begin{aligned} Beta(1, 1) &\to Beta(1 + 60, 200 - 60 + 1) = Beta(61, 141) \\ Beta(10, 10) &\to Beta(70, 150) \\ Beta(a, b) &\to Beta(a + 60, 140 + b) \end{aligned}$$

Commentary:

The most important example of conjugate priors in Bayesian statistics is the relationship of the Beta and Binomial distributions. As we've seen in this course, most of what people care about is either means or proportions, so this problem reveals how the Bayesian will work with proportions. On top of this, the problem begins to answer a natural question you may have had for some time: We talk frequently about unfair coins, but do such objects actually exist? The post linked in this problem shows to create these, if you're willing to expand your definition of a "coin". Once you open this door a crack, you'll notice other unfair "coins". For example, a board game might have you roll a twenty-sided die (labeled 1 to 20), and you win a battle if the result is less than or equal to 12. This "coin" has heads probability $\frac{12}{20} = 0.60$ where the "heads side" of the coin is the combined surface area of the sides labeled 1 to 12. You may have also seen spinners that have two zones, one comprising 240° and the other 120°. This is a "coin" with $p_{\text{heads}} = \frac{2}{3}$ when the goal is to land in the larger zone. It's time to re-imagine the definition of a coin!

98. Suppose you are trying to fit a Weibull distribution to your data where $X \sim Weibull(b)$ has pdf

$$f_X(x; b) = 4bx^3 \cdot \exp(-bx^4) \text{ where } x > 0, b > 0$$

You're unsure what value b should have, and your friend suggests encoding your beliefs about b into a $Gamma(r, \alpha)$ prior. Given a random sample $x_1, \ldots, x_n$, show that $b \sim Gamma(r, \alpha)$ is a conjugate prior for this Weibull distribution, and find the updating rule that converts the prior hyper-parameters into posterior hyper-parameters.

Hint List:

- When using the Bayes updating rule, remember that b is the variable we care about. Any expression without a b can be hidden behind the proportionality symbol.

Solution:

We find the posterior as usual:

$$\begin{aligned} g(b \mid \vec{x}) &\propto f(\vec{x}; b) \cdot g(b) \propto \prod_{i=1}^{n} \left[4bx_i^3 \exp\left(-bx_i^4\right)\right] \cdot \frac{\alpha^r}{\Gamma(r)} b^{r-1} \exp(-\alpha b) \\ &\propto b^n \exp\left(-b \sum x_i^4\right) \cdot b^{r-1} \exp(-\alpha b) \propto b^{r+n-1} \exp\left[-\left(\alpha + \sum x_i^4\right) b\right] \\ &\sim Gamma\left(r+n, \alpha + \sum x_i^4\right) \end{aligned}$$

Thus, our updating rule is $\boxed{Gamma(r, \alpha) \to Gamma\left(n+r, \alpha + \sum x_i^4\right)}$.

Commentary:

The Weibull distribution is named for the Swedish mathematician Waloddi Weibull. People tend to pronounce this as "Vay-Bull" or "Way-Bull" or "Way-ee-bull". As with many things in math, despite being named for Weibull, the distribution was discovered by someone else. It appears René Fréchet used this distribution to model the size of particles six years before Weibull. As such, this naming is an example of Stigler's Law of Eponymy, which states that scientific discoveries are never/rarely named for their true inventor.

Naming aside, one thing to notice in this solution is just how sophisticated you've become. Only a small percentage of people on Earth can understand these few lines of notation. One reason for this is the use of *specialized* symbols to hide complex ideas. Here, scribbles like $\sum, \prod, \sim, \propto$, and exp only make sense if you've spent time in mathematics. Furthermore, our symbols are incredibly *compact*. The notation $\vec{x}$ actually stands for the vector $(x_1, x_2, \ldots, x_n)$, which is our random sample packaged together in one place. Similarly, when we write $Gamma(r, \alpha)$, this stands for a complicated random variable with a pdf that relies on a function, $\Gamma(x)$, that is defined as an improper integral which only makes sense as a limit, and on and on and on. When you sit back and look at how far you've come, you should be quite proud of your progress.

While it is valuable to honor the growth you have achieved, it is also important to remember the responsibility you have to *communicate clearly* with different audiences. Never forget how dense and specialized your language has become. When you are speaking to audiences well-versed in the mathematical semiotic, feel free to write things as you have in this problem. For less quantitative audiences, adjust accordingly. Most importantly, never weaponize abstruse language to seem more profound, to get crummy research published, to sell products, etc. In recent times, scholars have begun studying this phenomenon, podcasting about it, and exposing it in our institutions.

99. Your friend is behind the computer using a random number generator modeled on $Unif(0, \theta)$. You don't know θ, so you decide to model your belief in what value your friend would choose for θ. You've got a hunch that θ won't be less than π, and that really large values for θ are unlikely (would your friend *really* choose $\theta = 1234567890$?). One option for this belief structure is the Pareto distribution, which has pdf

$$g(\theta; \pi, \alpha) = \frac{\alpha \pi^{\alpha}}{\theta^{\alpha+1}} \text{ where } \theta \geq \pi$$

Here, α is a hyper-parameter that controls how fast the likelihood of θ decays as θ gets bigger. Also π is a hyper-parameter demarcating where the support begins (not the constant 3.14...).

Show the Pareto distribution is a conjugate prior for the uniform distribution assuming a random sample $x_1, \ldots, x_n$ and give the updating rule.

Hint List:

- Use indicator functions throughout.
- The problem tells you that Pareto should be the conjugate prior for the Uniform. Use this to your advantage in the algebra. The kernel of g is $\frac{1}{\theta^{\#}} \cdot I_{[\#,\infty)}(\theta)$ where the # signs are just (possibly different) numbers. Head toward that as you calculate the posterior.

Solution:

We first write g using indicator functions: $g(\theta;\pi,\alpha) = \frac{\alpha\pi^{\alpha}}{\theta^{\alpha+1}} I_{[\pi,\infty)}(\theta)$.

Doing the same for f gives $f(x;\theta) = \frac{1}{\theta} I_{[x,\infty)}(\theta)$.

Using Bayes Theorem, we have

$$\begin{aligned} g(\theta \mid \vec{x}) &\propto f(\vec{x};\theta) \cdot g(\theta) \propto \prod_{i=1}^{n} \frac{1}{\theta} I_{[x_i,\infty)}(\theta) \cdot \frac{\alpha\pi^{\alpha}}{\theta^{\alpha+1}} I_{[\pi,\infty)}(\theta) \\ &\propto \frac{1}{\theta^{n+\alpha+1}} I_{[\max x_i,\infty)}(\theta) \cdot I_{[\pi,\infty)}(\theta) \\ &\propto \frac{1}{\theta^{n+\alpha+1}} I_{[M,\infty)}(\theta) \text{ where } M = \max(\max x_i, \pi) \\ &\propto Pareto(M, \alpha+n) \end{aligned}$$

So, our updating rule is $\boxed{Pareto(\pi,\alpha) \to Pareto(\max(\max x_i, \pi), \alpha+n)}$.

Commentary:

This problem is interesting for two reasons. The first is that it answers a question you might naturally have: What is the conjugate prior for each of the distributions I've worked with this quarter? Here, we learn that Pareto is the conjugate prior for the Uniform. The best students will ask these types of questions and go hunt down the answers. Second, the problem teaches us that when the support of the likelihood depends on the parameter we are studying (here, θ), it can be useful to do the problem with indicator functions. We also saw this when studying MLEs.

100. Another famous distribution is the log-normal which, when indexed by the parameter τ, has the pdf

$$f_X(x;\tau) = \frac{\sqrt{\tau}}{x\sqrt{2\pi}}\exp\left[-\frac{\tau}{2}(\ln x)^2\right] \text{ where } x > 0, \tau > 0$$

(a) Find the Cramer-Rao lower bound (CRLB) on the variance of an unbiased estimator of τ based on a random sample of size n.

(b) Suppose you wish to do Bayesian analysis on τ, so you assign it the prior distribution $\tau \sim Gamma(r, \lambda)$. Given a random sample of data $x_1, \cdots, x_n$, show that the Gamma distribution is a conjugate prior for the log-normal distribution, and explain how the values r and λ get updated in the posterior distribution.

Hint List:

- b. This part does not use the result from (a). Do you still get uncomfortable when I violate the academic narrative that "each part must use the previous parts"?

Solution:

(a) First, $\ln f(X;\tau) = \frac{1}{2}\ln\tau - \ln X - \frac{1}{2}\ln(2\pi) - \frac{\tau}{2}(\ln X)^2$.

Thus, $\frac{\partial \ln f}{\partial \tau} = \frac{1}{2\tau} - 0 - 0 - \frac{1}{2}(\ln X)^2 \implies \frac{\partial^2 \ln f}{\partial \tau^2} = -\frac{1}{2\tau^2}$.

So, $I(\tau) = -E\left[\frac{\partial^2 \ln f}{\partial \tau^2}\right] = \frac{1}{2\tau^2}$. Hence, the CRLB $= \boxed{\frac{2\tau^2}{n}}$.

(b) We have

$$\begin{aligned} g(\tau \mid \vec{x}) &\propto f(\vec{x};\tau)\cdot g(\tau) = \prod_{i=1}^{n} \frac{\tau^{1/2}}{x_i\sqrt{2\pi}} \exp\left[-\frac{\tau}{2}(\ln x_i)^2\right] \cdot \frac{\lambda^r}{\Gamma(r)}\tau^{r-1}\exp(-\lambda\tau) \\ &\propto \tau^{n/2}\exp\left[-\frac{\tau}{2}\sum(\ln x_i)^2\right]\cdot \tau^{r-1}\exp(-\lambda\tau) \\ &\propto \tau^{r+n/2-1}\exp\left[-\tau\left(\lambda + \frac{1}{2}\sum(\ln x_i)^2\right)\right] \\ &\sim Gamma\left(r+\frac{n}{2}, \lambda + \frac{1}{2}\sum(\ln x_i)^2\right) \end{aligned}$$

Thus, our updating rule is $\boxed{Gamma(r,\lambda) \to Gamma\left(r+\frac{n}{2}, \lambda + \frac{1}{2}\sum(\ln x_i)^2\right)}$.

Commentary:

After doing a few conjugate prior problems, you are likely getting comfortable with the algebra involved. When this happens, your brain is freed up to notice other things going on. Did you notice in this problem just how powerful the Gamma distribution is? Because its kernel contains both a power function and an exponential function, it easily acts as a conjugate prior for the log-normal since this likelihood also has power and exponential functions (which can join forces with their partners in the Gamma kernel). Indeed, this is the beautiful essence of the Gamma distribution: it is *the* general tool that marries power and exponential functions. As such, the Gamma is the conjugate prior for many different likelihood functions: log-normal, Weibull, Poisson, Exponential, Gamma, inverse Gamma, and others. When you find yourself working many problems of the same type, try to find the cognitive core of what is going on. For conjugate prior problems, we are essentially trying to show that the algebraic structure of the likelihood can be absorbed in the algebraic structure of the prior. Since the Gamma has such a complex algebraic structure, it can absorb many a likelihood.