Math 181A HW Problems Complete List

General instructions:

- Clearly and thoroughly write your solutions on blank paper, showing all your work. See the syllabus for instructions for uploading to Gradescope. See the calendar for due dates/times.
- 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 as part of your answer. You may take a screenshot of the code/output, or write it by hand.
- Problems tend to focus on content from the two or three previous lectures and never require ideas from a lecture that falls on the day a problem is due. For example, if a problem is due on Friday the 19th, it is likely to use ideas from lectures on Wednesday the 17th, Monday the 15th, and/or Friday the 12th. It will not use ideas from the lecture on Friday 19th. It is possible that a problem requires knowledge from earlier in the course, or from prerequisite courses. 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, impropoer integrals, L'Hôpital's Rule, etc.). Expect prerequisite knowledge to be drawn on frequently.
- At the end of the course calendar, you should see a phrase like "Problems XX-XX not collected". This refers to the problems at the end of this packet that are here to help you learn the material but cannot be collected/graded because of union rules related to UCSD graders. You should work these problems to develop your mastery of topics from the last few lectures in the course as you prepare for the final exam.

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 can be written in this surprising way: $f(x; p) = p^x(1-p)^{1-x}$.

   (b) For many students, the above pmf feels like pure magic. Explain how you can come up with this 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 definitions of mean and variance.

   (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.

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

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

   (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.

3. The "Kernel Technique". 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 $\displaystyle\int_{\text{support}} f(x)\,dx = 1$. For example, if we take $X \sim Exp(\lambda = 4)$, then we know $\displaystyle\int_0^\infty 4e^{-4x}\,dx = 1$ since the pdf is $f(x) = 4e^{-4x}$. 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:

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

   (a) Find $\displaystyle\int_{-\infty}^{\infty} e^{-x^2/2}\,dx$. 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) Find $\displaystyle\int_0^\infty x^4 e^{-3x}\,dx$. (Integration by parts 4 times? Nope. The kernel technique.)

(c) $\sum_{x=0}^{\infty} \frac{2^x}{x!}$ (You could do this problem using Taylor series, but use the kernel technique. Note that because we have sum instead of an integral, you should be thinking about **discrete** random variables here, not **continuous** random variables.)

(d) $\sum_{x=4}^{\infty} \binom{x-1}{3} (0.7)^{x-4}$ (This is very scary without the kernel technique.)

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

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

Find the MME (method of moments estimate/estimator) for θ. (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 $x \geq 1$.)

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 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 know the time 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 6 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.

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.)

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 the first moment of X (which is what people typically use).

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

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 assuming that β is fixed (known). (Hint: **After** setting up an integral, try a u-substitution with $u = \alpha y^\beta$. Remember to switch the bounds to u bounds, and the switch the dy as well. Your answer will have a $\Gamma\left(1 + \frac{1}{\beta}\right)$ in it. Also, 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 $\alpha\beta y^{\beta-1} \exp(-\alpha y^\beta)$, which is a little clearer.)

9. Let Y be a CRV with density $f_Y(y; \theta) = \frac{2y}{\theta} e^{-y^2/\theta}$ 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**.

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 $x_1 = 2, x_2 = 1, x_3 = 6$.

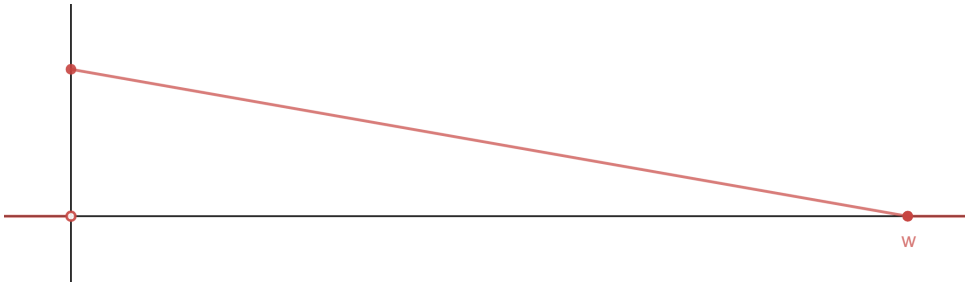
    (b) Using Desmos, draw a graph of the likelihood function (not log-likelihood) for the data $x_1 = 2, x_2 = 1, x_3 = 6$. It should be maximal at the μ value you found in part a. Include a sketch of the graph from Desmos (or a screenshot if you're tech-fancy). (Note: In Desmos, if you click on the wrench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

11. 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 MLE for α and β. (As with all multivariable maximization problems in this class, you need NOT show your MLE is maximal via higher derivatives. Also, as on all MLE 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.".)

12. 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.)



13. Let's start using R to *see* estimators in action. While an estimator looks like 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 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. (Note: 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.

14. 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.

15. 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.

    (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?

16. 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? (Also, see problem 17.)

    (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.

17. Let's check your answer to 16a. 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 16a 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.

18. 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 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 question 19.)

```r
# 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
```

19. Using the programming skills you gained in question 17, write code to check your answer for question 18b. Include your code in your answer.

20. 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 a 95% CI from a particular random sample, then there is a 95% chance that μ is in this interval.

    (b) If $(3.2, 5.1)$ is a 95% CI from 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.

21. 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 are using decimals). Suppose you change the confidence level to C% and the upper bound of the approximate CI exactly equals 100%. Find C.

22. 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 Google searching to get an estimate for the proportion. At this website, you see 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. What is the new minimal cost estimate?

23. One frustrating issue with proportions is that your mathematics might give a CI like: $(-2\%, 7\%)$ or $(96\%, 103\%)$. This typically happens when you are studying a trait with a proportion near 0% or 100%. This makes it particularly difficult to study rare or hyper-prevelant 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?

Note: Problems 24-28 were created by Jack Determan, UCSD '26.

24. 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 $\lambda = \dfrac{1 \text{ breakdown}}{4 \text{ years}}$. 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?

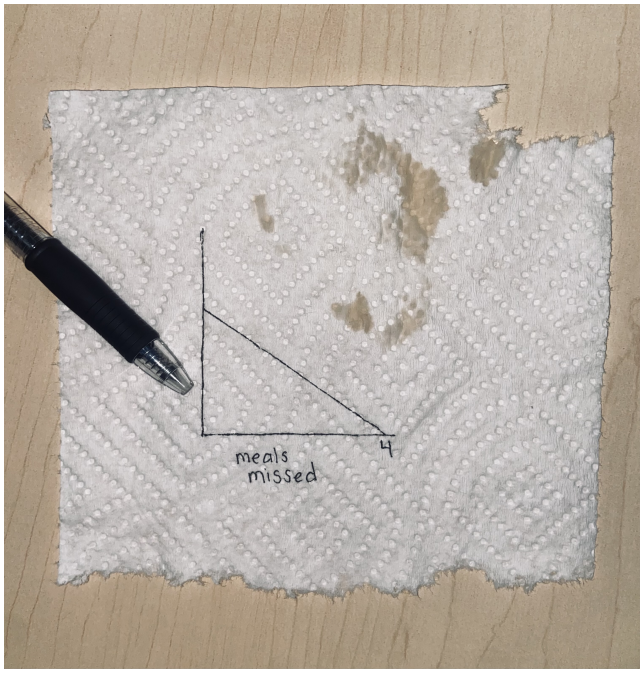
(b) Redo part (a) with n keys and $T \sim \text{Exp}(\lambda)$ instead of $T \sim \text{Exp}\left(\frac{1}{4}\right)$.

(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 the drive train of 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.

25. 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.

26. 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

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

(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?

27. (a) Consider a random variable $M \sim N(10, 2)$.

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.

(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 $Y \sim N(97, 0.6)$. 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 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 process to estimate the distribution of the range.)

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

28. Victoria is a competitive diver, and she is getting ready to compete at a major event. In diving, athletes are often assigned a score 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 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:

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

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

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

ii. Without a calulator/tech (!), find $\mathbb{E}\left[X_{(7)}\right]$.

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

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

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

29. In this problem, we explore more estimators for μ and σ^2 in the distribution $N(\mu,\sigma^2)$.

(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 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.

30. 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 $\widehat{\theta}_{\text{MLE}}=\max X_i=X_n'$ to estimate θ. Your friend, however, has meganumerophobia, and is afraid to say the maximum number in the random sample. Instead, he'll say the second largest number: $\widehat{\theta}=X_{n-1}'$. 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.

31. 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: 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 known, 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 $\widehat{V} = n\left(\frac{X_1}{n}\right)\left(1 - \frac{X_1}{n}\right)$. 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.

32. The number of times, X, a particular first-year college student calls home during a random week is a Poisson RV with mean λ: $X \sim Poisson(\lambda)$. Curious to find the value for λ, you break into the NSA (!) and access phone records for this student on n random weeks. You record the number of calls home and get the random sample $X_1, \ldots, X_n$.

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

    (b) You're curious how many total minutes, M, these X calls amount to in a week, and you read a recent journal article that suggests the model $M = 2X + 3X^2$. Find the expected number of weekly minutes as an expression involving λ.

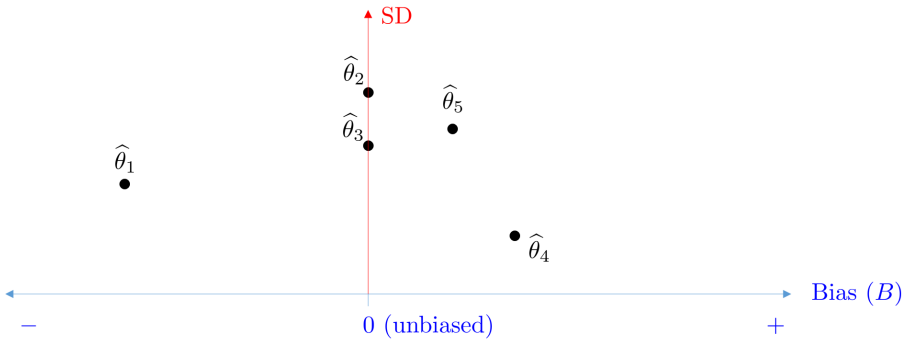
    (c) Find an unbiased estimator of your answer from part b based on the random sample $X_1, X_2, \ldots, X_n$. Note: You don't have a procedure for doing this. You should dream up a guess and then check to see if it works. If not, adjust it.

33. 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 $\frac{1}{\lambda}$ and modify it to create an unbiased estimator. (Hint: During your journey, you'll need the help of the gamma distribution, the gamma function, and the knowledge that $\Gamma(1/2) = \sqrt{\pi}$.)

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

35. 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?

36. Let X be a continuous random variable with $E(X) = \mu$ and $Var(X) = \sigma^2 < \infty$. Suppose we try



to estimate μ using these two estimators from a random sample $X_1, \ldots, X_n$ (where $n \geq 3$):

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

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

37. 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

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

38. 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.

39. Let $X_1, \ldots, X_n$ be iid based on $f(x;\theta) = \dfrac{2x}{\theta} e^{-x^2/\theta}$ where $x > 0$. Show that $\widehat{\theta} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2$ is efficient.

40. 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 $\widehat{\theta} = \max Y_i$. Show $\widehat{\theta}$ is consistent using the ε-definition of consistency.

41. 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 problems 40 and 41b are training two ways to show consistency: via the ε-definition, and through MSE-consistency. Be skilled at both.)

42. 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 the note on exp notation from problem 8.)

43. 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 random test scores of the student: 80%, 92%, 81%, and 96%).

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

(a) Find the MLE for θ based on a 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.

45. One distribution useful for modelling 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. (Recall that $\Gamma(1/2) = \sqrt{\pi}$.)

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.

46. Suppose X is a continuous random variable 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 $\widehat{\mu}_{\text{MLE}}$, the maximum likelihood estimator for μ. (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.)

47. 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 MME or MLE, this reminder is comforting.

48. Calculate the P-values for problems 47b and 47c. Does using these P-values lead you to the same conclusions as the critical regions did?

49. Suppose you wanted to alter problem 47a 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?

50. In December 2017, the J-RPG Xenoblade Chronicles 2 was released for Nintendo Switch. The game is epic in its number of main quests and side quests. Those that try to finish every aspect of the game are known as "completionists". What is the average time for all completionists in the world

(currently)? 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 her 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!]

51. 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 suggest 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?

52. 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 force us 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 force us to reject H_0 if we change H_1 to $\mu \neq \mu_0$? Assume α remains the same.

53. 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 using R's subset command. 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/reviewed by users.

54. 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 (type ?prop.test into the R Studio prompt) 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).

55. 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." Respond to this claim from a statistical perspective.

56. People often look down on machine learning because in some settings, it can only improve things in small increments. While this might be true, in many settings a slight change can have a huge impact. As an example, Americans spend about 3 trillion dollars per year spread across 30 billion 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. If machine learning could help reduce the percentage of fraudulent claims even slightly, this would save companies billions of dollars! Researchers at Visa have designed a new algorithm to predict fraud and are curious if it has reduced illegal 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? (Continue to think about the Larsen & Marx criterion as in part a.)

57. What hypothesis test has duality with the CI $\left(-\infty, \overline{X} + 1.3\frac{\sigma}{\sqrt{n}}\right)$? Assume $X \sim N(\mu, \sigma^2)$ with σ known and μ unknown.

58. 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 verify 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.

59. 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. 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 $\alpha = 0.05$.

60. 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 this is our go-to belief about sounds in life. Describe the alternative hypothesis for this setup, explain what Type I and II errors would mean, 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.

61. 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.

62. 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.

63. 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. Now, my student claims to have a new manufacturing process that raises this average (he is 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.

64. Before being de-platformed from Twitter in January 2021 (only to be re-platformed in November 2022), the number of tweets that former President Trump sent on a random day might be modeled by $X \sim Poisson(\lambda)$. You've heard his average tweet rate was 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.)

65. Suppose $X \sim Exp(\lambda)$ where X is modeled by $f(x;\lambda) = \lambda e^{-\lambda x}$, where $x, \lambda > 0$. You draw a sample of size n and plan to use the statistic $X_{\min}$ to decide between two hypotheses.

    (a) Show that $X_{\min}$ also has an exponential distribution and determine what parameter it is based on (instead of λ).

    (b) You plan to test H_0: $\lambda = 2$ vs. H_1: $\lambda > 2$ via the rule: If $X_{\min} < c$, reject H_0; else keep H_0. 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 $n = 5$. On this problem, you must write hypotheses and include a beautifully-labeled diagram in your answer with two pdfs and a rejection fence. (Include your code if you use R to create the diagram.)

66. Let $Y \sim \chi^2_n$.

    (a) Show that $E[Y^k] = \dfrac{2^k \Gamma\left(\frac{n}{2} + k\right)}{\Gamma\left(\frac{n}{2}\right)}$ 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: Think of T_n as a product of two RVs.)

67. Prove these two useful facts about the F distribution:

    (a) $E(F_{m,n}) = \dfrac{n}{n-2}$ (when $n > 2$)

    (b) $F_{m,n} = \dfrac{1}{F_{n,m}}$

68. Students often find it hard to believe that $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. So, let's simulate this situation and see if the data agree. To do so, generate $n = 7$ 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 $\dfrac{(n-1)S^2}{\sigma^2}$. Then, in red, overlay the pdf for χ^2_{n-1}. Include your code and a sketch of your plot. (Note: 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!)

69. It is also surprising that $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim T_{n-1}$. To empirically convince you of this, 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.

70. 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 $P(|T_{n+1}| < 1)$

71. Let $X_1, \ldots, X_{16}$ be a random sample from a normal distribution with mean 0. For what k is the below inequality true? Explain your reasoning.

$$P\left(\left|\frac{4\overline{X}}{S}\right| > k\right) = 0.08$$

72. 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. She believes that the coffee dispensing amounts are normally distributed, and her seven data points give $\overline{x} = 12.9$ 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.

73. In question 55, 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 one-sided 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.

74. The most brutal algebra moment in 181A: Show that the pdf of T_n converges to the pdf of $N(0,1)$ as $n \to \infty$. When doing this problem you may use Stirling's Formula: $n! \approx \sqrt{2\pi n} \cdot n^n e^{-n}$ and a helpful fact from Calculus I: $\lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = e^a$. Also, it is fine if you assume $\Gamma(r) = (r-1)!$, even when r is not an integer.

75. 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 $T_{n-1} \approx \frac{\overline{X} - \mu}{S/\sqrt{n}}$. 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 $\frac{\overline{X} - \mu}{S/\sqrt{n}}$. Repeat this process to get a total of 50000 t-scores. 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.

76. If you look at a bottle of ibuprofen, it will likely list the amount of medicine per pill (usually, 200 mg). Of course, this is only an average, and if you carefully measured 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 dosage 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 and randomly select some to measure the dosage. You get these values: 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.

77. 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. You decide to explore the effects of various rule changes that occurred in the NHL

(ice hockey!). Prior to a new rule set lauched 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 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.

78. Suppose that $X \sim N(\mu, 1)$. We plan to test H_0: $\mu = 0$ vs. H_1: $\mu = 4$ using a random sample of size n. Show that the BCR will take the form $C = \{\mathbf{X} | \sum_{i=1}^{n} X_i > c\}$ where c is a constant.

79. Suppose that $X \sim N(\theta_1, \theta_2)$, both unknown. We plan to test H_0: $\theta_1 = 0$, $\theta_2 = 1$ vs. H_1: $\theta_1 = 1$, $\theta_2 = 3$. Find the simplest expression you can for a BCR in this setting. Assume a random sample of size n.

80. Suppose you take a random sample of size 12 from $X \sim N(0, \sigma^2)$, whose variance is unknown. You plan to test H_0: $\sigma^2 = 1$ against H_1: $\sigma^2 = 3$. Find the BCR of size $\alpha = 0.08$.

81. Let $X \sim Bernoulli(p)$. We wish to test H_0: $p = \frac{1}{3}$ vs. H_1: $p < \frac{1}{3}$ using a random sample of size 11. Let $C = \{\mathbf{X} | \sum_{i=1}^{11} X_i < c\}$. Show that C is a UMPCR and find its α when $c = 2.3$.

82. You and a friend are arguing about what model X better explains 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 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 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.

83. 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: $\lambda = \lambda_0$ vs. H_1: $\lambda < \lambda_0$.

(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.

84. According to 83b, there is no UMP test for H_0: $\lambda = \lambda_0$ vs. H_1: $\lambda \neq \lambda_0$. Find the GLRT for this setup, simplifying your answer as much as you can.

85. 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 mean 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 for 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.

86. Let X be a RV with pdf $f(x;\theta) = \frac{m}{\theta}x^{m-1}e^{-x^m/\theta}$, 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: $\theta = \theta_0$ against H_1: $\theta > \theta_0$.

(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 X^m. To do so, use this result (sometimes taught in Math 180A): If Y is a RV with pdf $f_Y(y)$ and h is either increasing or decreasing, then $U = h(Y)$ has pdf $f_U(u) = f_Y(y) \cdot \left|\frac{dy}{du}\right|$ where $y = h^{-1}(u)$.

87. 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.

88. 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 (answers will vary!)?

89. 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?

90. 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 that hard if you're willing to bend your coin so it has a U-shape. Read this blog 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 $Beta(a, b)$?

91. Suppose you are trying to fit a Weibull distribution to your data where $X \sim Weibull(b)$ has

$$f_X(x; b) = 4bx^3 \cdot \exp(-bx^4) \text{ where } x > 0, b > 0$$

You're unsure what value b should have, and your friend suggests encoding your beliefs about b into a $Gamma(r, \alpha)$ prior. Given the data $x_1, \ldots, x_n$, 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.

92. Once again, your friend is behind her 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:

$$g(\theta; \pi, \alpha) = \frac{\alpha \pi^{\alpha}}{\theta^{\alpha+1}} \text{ where } \theta \geq \pi$$

Here, α is a hyper-parameter that controls how fast the likelihood of θ decays as θ gets bigger. Show this distribution is a conjugate prior for the uniform distribution using the data $x_1, \ldots, x_n$ and give the updating rule. (Hint: Use indicator functions throughout.)

93. 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 for a 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 above log-normal distribution, and explain how the values r and λ get updated in the posterior distribution.