This problem set covers material from Week 6, dates 10/21 – 10/24.

Instructions: Write or type complete solutions to the following problems and submit answers to the corresponding Canvas assignment. Your solutions should be neatly-written, show all work and computations, include figures or graphs where appropriate, and include some written explanation of your method or process (enough that I can understand your reasoning without having to guess or make assumptions). A general rubric for homework problems appears on the final page of this assignment.

In some the following, you may need to use `R`. If you do, please write down the corresponding code to "show your work". Drawing and labelling curves are also good examples of "showing your work".

Monday 10/21

1. In triathlons, it is common for racers to be placed into age and gender groups. Two friends, Leo and Mary, both completed a triathlon. A better performance in the race corresponds to a faster finishing time. Leo competed into the "Men, Ages 30-34" group. Mary competed in the "Women, Ages 25-29" group. Leo completed the race in 4948 seconds, while Mary completed the race in 5513 seconds. Leo did finish faster, but they are curious about how they did within their respective groups. Below is some information on the performance of their groups:

    - "Men, Ages 30-34" group finishing times have a mean of 4313 seconds with a standard deviation of 583 seconds
    - "Women, Ages 25-29" group finishing times have a mean of 5261 seconds with a standard deviation of 807 seconds
    - The distribution of both groups' finishing times is approximately Normal.

    (a) Write down the short-hand for the two Normal distributions.

    (b) What are the z-scores for Leo and Mary's finishing times? What is the interpretation of the z-scores?

    (c) Did Leo or Mary rank better in their respective group? Explain your reasoning.

    (d) What percent of the triathletes did Leo finish faster than in his group?

    (e) What percent of the triathletes did Mary finish faster than in her group?

    (f) If the distributions of finishing times are not nearly Normal, would your answers to any of (b) - (e) change? Explain your reasoning.

2. Suppose body temperatures are Normally distributed with mean 98.6° F and standard deviation of 0.7° F. Assuming this is true, answer the following:

    (a) Fevers 103° F or higher are considered dangerous. What fraction of people would be expected to have such high a fever?

(b) According to a quick Google search, a range for low-grade fever is between 99.5° F and 100.3° F. What is the probability of having a low-grade fever?

(c) What body temperatures would you consider as unusually low? Briefly explain why.

(d) Provide two intervals that each capture/contain 80% of body temperatures.

3. Find the standard deviation of the distribution in the following situations.

(a) MENSA is an organization whose members have IQs in the top 2% of the population. IQs are Normally distributed with mean 100. The minimum IQ scores required for admission to MENSA is 132.

(b) Cholesterol levels for women ages 20-34 follow an approximately normal distribution with mean 185 milligrams per deciliter (mg/dl). Women with cholesterol levels above 220 mg/dl are considered to have high cholesterol and about 18.5% of women fall into this category.

Wednesday 10/23

Note: if you use the CLT, you should explicitly state something along the lines of "By the CLT, ...". Make sure to check assumptions hold if necessary!

4. The average teacher salary in Vermont is $62,483. Suppose that the distribution of teacher salaries is approximately normal with standard deviation $7000.

(a) What is the probability that a randomly selected Vermont teacher makes less than $60,000 per year?

(b) If we randomly sample 25 Vermont teachers and obtain their salaries, what is the probability that the mean of their salaries is less than $60,000 per year?

(c) Compare the probabilities in (a) and (b), and explain mathematically why one is larger than the other.

(d) How would your answers to (a)-(b) change if the distribution of teacher salaries was not normal?

5. In 2011, a poll found that 25% of young American delayed starting a family due to the economic slump. Determine if the following statements are true of false, and explain your reasoning.

(a) The distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump in random samples of size 12 is right skewed. *(Use your stats intuition/reasoning here!)*

(b) In order for the distribution of sample proportions of young American who have delayed starting a family due to the continued economic slump to be approximately Normal, we need random samples where the sample size is at least 40.

(c) A random sample of 50 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(d) A random sample of 150 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(e) Tripling the sample size will reduce the standard error of the sample proportion by one-third.

6. In the US, businesses and schools shut down during the COVID-19 pandemic in March 2020, and a vaccine became publicly available for the first time in April 2021. That month, a Gallup poll surveyed a random sample of 3731 US adults, asking them how they felt about a COVD-19 vaccine requirement for air travel. The poll found that 57% said they would favor it.

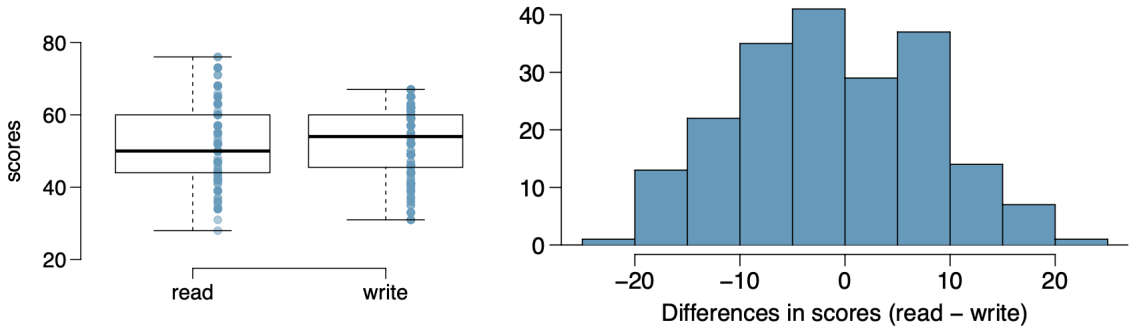
(a) Identify/define the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Using a mathematical model, construct a 95% confidence interval for the proportion of US adults who favor requiring proof of the COVD-19 vaccination for travel by airplane. Be sure to check conditions/assumptions necessary for your answer. If you use `R` in any way, please also provide the code.

(c) Based on your CI, would it be appropriate to claim that the majority of Americans were in favor of the COVID-19 requirement? Briefly explain why.

Thursday 10/24

7. The National Center of Education Statistics conducted a survey of high school seniors, collecting standardized test data on their performance in reading, writing, and several other subjects.. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of each student's reading and writing scores as well as a histogram of the differences in scores are shown below.



(a) Are the data paired? Briefly explain why or why not.

(b) We want to obtain a 95% confidence interval for the average difference between the reading and writing scores of all high school seniors. Check the conditions required to obtain this interval.

(c) The observed mean difference in scores is $\bar{x}_{\text{read}-\text{write}} = -0.545$ and the observed standard deviation of differences is 8.887 points. Calculate a 95% confidence interval for the average difference between the reading and writing scores of all students.

(d) Does your confidence interval provide convincing evidence that there is a real difference in the average scores? Explain.

8. Problems in associated `.Rmd` template. Note: `R` questions 1 and 2 will be treated as one problem, and 3-5 will be treated as one problem. Once again, there will be points associated with reproducible and clean code, along with informative axis labels!

General rubric

Points	Criteria
5	The solution is correct *and* well-written. The author leaves no doubt as to why the solution is valid.
4.5	The solution is well-written, and is correct except for some minor arithmetic or calculation mistake.
4	The solution is technically correct, but author has omitted some key justification for why the solution is valid. Alternatively, the solution is well-written, but is missing a small, but essential component.
3	The solution is well-written, but either overlooks a significant component of the problem or makes a significant mistake. Alternatively, in a multi-part problem, a majority of the solutions are correct and well-written, but one part is missing or is significantly incorrect.
2	The solution is either correct but not adequately written, or it is adequately written but overlooks a significant component of the problem or makes a significant mistake.
1	The solution is rudimentary, but contains some relevant ideas. Alternatively, the solution briefly indicates the correct answer, but provides no further justification.
0	Either the solution is missing entirely, or the author makes no non-trivial progress toward a solution (i.e. just writes the statement of the problem and/or restates given information).
Notes:	For problems with multiple parts, the score represents a holistic review of the entire problem. Additionally, half-points may be used if the solution falls between two point values above.
Notes:	For problems with code, well-written means only having lines of code that are necessary to solving the problem, as well as presenting the solution for the reader to easily see. It might also be worth adding comments to your code.