STAR 511 HW #3

See Canvas Calendar for due date.
79 points total, 2 points per problem unless otherwise noted.

Questions 1 through 8 (Diet): Healthy Eating Index (HEI) is a measure of diet quality used to assess how well a set of foods aligns with established dietary guidelines. HEI scores are recorded for a sample of n = 23 women receiving SNAP benefits (food stamps) and living in Denver, CO. Assume this is a random sample (but in practice that could be difficult to achieve). The data is available from Canvas as Diet.csv.

Reminders: (1) Use read.csv() to import the data. (2) Check the data after importing. (3) Use $ to access the HEI column.

1. Given this sample, describe the population to which (statistical) inferential statements can reasonably be made.
2. Report a histogram of the data. Use `xlab=` (or `xlab()` in ggplot) to give the x axis the title "Health Eating Index".
3. Report the sample mean and standard deviation.
4. Report a 95% confidence interval for μ (population mean HEI).
5. What does the confidence interval from the previous question tell you?
6. A colleague (incorrectly) says that the confidence interval is not valid because sample size is less than 30. Explain why the confidence interval is valid. Hint: Consider the histogram from above.

Questions 7 and 8 (Diet continued): We continue with the diet data. But now we test H0: $\mu = 60$ versus HA: $\mu \neq 60$ (using $\alpha = 0.05$). (US mean HEI is 60 based on a very large, diverse sample.)

7. Test this null hypothesis using the confidence interval from question 4. State the "statistical decision" (reject or fail to reject H_0), and briefly explain how you used the confidence interval to reach this decision. (**4 pts**)
8. Calculate an appropriate p-value, and write a sentence explaining what this p-value tells you. Your sentence should interpret the p-value itself; don't just say "reject" or "fail to reject". (**4 pts**)

Questions 9 through 11 (CI): Describe how the following affect the width of the confidence interval (assuming everything else is held constant). Answer should be one of **increase, decrease** or **stays the same,** and provide a brief explanation for why.

9. Sample size increases
10. Standard deviation increases
11. Confidence level decreases

Questions 12 through 15 (Oxygen): Suppose the mean oxygen level of a certain lake is of interest. A total of **n=10** samples were taken (from randomly selected locations) and oxygen level was measured in ppm. The sample mean oxygen level is 4.62 and the sample standard deviation is 0.58. Use $\alpha = 0.05$.

Notes:

- Because we are working from summary statistics (instead of raw data), these questions should be done "by hand" (but using R as a "calculator"). This is for practice and for illustration.
- The "rejection region" or "rejection rule", should be of the form Reject H0 if… and should include a numeric value.
- The decision can be brief: Reject H0 or Fail to Reject H0.
- Watch out for **sign**, direction, and absolute value. It may help to make a sketch.
- Answers should be organized and labeled such they can be easily read and understood.

12. Test H0: $\mu = 5$ vs HA: $\mu \neq 5$.
    A. Define the rejection region. (**2 pts**)
    B. Calculate the test statistic. (**2 pts**)
    C. State your decision. (**1 pt**)
13. Test H0: $\mu \geq 5$ vs HA: $\mu < 5$.
    A. Define the rejection region. (**2 pts**)
    B. Calculate the test statistic. (**2 pts**)
    C. State your decision. (**1 pt**)
14. Now suppose that the summary statistics were based on a sample of size **n=40**. Rerun the hypothesis test from question 12 (H0: $\mu = 5$ vs HA: $\mu \neq 5$) based on this larger sample size.
    A. Define the rejection region. (**2 pts**)
    B. Calculate the test statistic. (**2 pts**)
    C. State your decision. (**1 pt**)
15. Considering the results of question 12 (n = 10) vs question 14 (n = 40), make a brief statement summarizing how increased sample size impacts hypothesis testing.

Questions 16 through 19 (Pills): Manufacturers must test the amount of the active ingredient in medications before releasing the batch of pills. The data Pills.csv (available from Canvas) represents the content (in mg) of the active ingredient in n = 24 pills (from a random sample of the same large batch). Use $\alpha = 0.05$.

16. Create a histogram and qqplot of the data. Based on this evidence, briefly discuss whether the data appear to have come from a normal distribution.
17. Give an estimate of the mean content and corresponding 95% confidence interval.
18. For this question, suppose that if there is evidence that the mean is different from 20mg, the batch of pills will be destroyed. Is there evidence that the batch of pills has a mean amount different from 20mg? (**2 pts per part**)
    A. State your hypotheses.
    B. Provide the test statistic and p-value.
    C. Make a conclusion in context of this study.
19. For this question, suppose that if there is evidence that the mean is less than 20mg, the batch of pills will be destroyed. Is there evidence that the batch of pills has a mean amount less than 20mg? (**2 pts per part**)
    A. State your hypotheses.
    B. Provide the test statistic and p-value.
    C. Make a conclusion in context of this study.

Questions 20 and 21 (CUE): An ecologist is planning a study to estimate mean carbon use efficiency (CUE) in a certain region. They will measure CUE on some number of randomly selected soil samples. They want to estimate the average CUE (in the region) within 0.03 units of the true population mean, using a 95% confidence interval.

20. Suppose that based on the previous experience, they expect almost all (>99%) CUE values to fall within the range 0.36-0.90. Use this information to "estimate" the standard deviation. Hint: Use an approach based on the empirical rule. For more detail, see the Ch5.4 notes.
21. Using the standard deviation from above, find the (minimum) sample size required to achieve a 95% ME < 0.03.

Questions 22 through 27 (Zinc): A national agency sets recommended daily allowances for many supplements. In particular, the allowance for zinc for adult men is 15 mg/day. The agency would like to determine if the average intake of zinc for adult men is greater than 15 mg/day. Suppose from a previous study they estimate the standard deviation to be 1.5 mg/day and they conjecture that the true population mean is 15.3 mg/day. The investigators plan to use a one-sample t-test with α=0.05.

22. Find the power with n = 85 for the scenario above. (**4 pts**)

For questions 23 through 26, give a brief justification for your answer.

23. If the sample size was larger (more than 85), would the power be higher or lower than that calculated in Q22?

24. If we use $\alpha = 0.01$ (instead of 0.05), would the power be higher or lower than that calculated in Q22?
25. If we use a conjectured mean of 15.6 mg/day (instead of 15.3), would the power be higher or lower than calculated in Q22?
26. If the standard deviation was larger (more than 1.5), would the power be higher or lower than that calculated in in Q22?

27. Return to the original scenario and find the sample size required to achieve 80% power. Remember to "round" up to an integer value. (**4 pts**)