This problem set covers material from Week 2, dates 9/15 – 9/18.

Instructions: Write or type complete solutions to the following problems and submit answers to the corresponding Gradescope 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.

Monday 9/15

1. Answer the following. Show your work by setting up the calculations of the mean and standard deviations. You should use R or a calculator to actually evaluate them!

   (a) A school offers several classes. We sample two classes, one with 10 students and one with 100 students. What is the average class size? What is the standard deviation of the class sizes?

   (b) A school offers several classes. We sample two classes, one with 10 students and one with 100 students. What is the average size of the class that a student is enrolled in, along with the corresponding standard deviation?

   (c) Provide an intuitive explanation for why the smaller standard deviation occurs in the part that it does.

   (d) Are the quantities you found above population parameters or sample statistics? Make sure your notation in (a) and (b) are consistent with your answer here!

2. The average score on a history exam (scored out of 100 points) was 85, with a standard deviation of 15.

   (a) If we plotted a histogram of the scores, would you *expect* the distribution to be symmetric or skewed (and if skewed, in which direction)? Explain your reasoning.

   (b) If you answered skewed above: is it possible for the distribution to be symmetric? If so, describe the scenario(s) where this would be possible.

3. Consider the following three sets of sampled data:

   $\mathbf{x} = (-2,\ 0,\ 1,\ 2,\ 4)$ $\mathbf{y} = (23,\ 25,\ 26,\ 27,\ 29)$ $\mathbf{z} = (-6,\ 0, 3, 6, 12)$

   (a) For each set of data $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, find the mean and the variance. Please use the proper symbols/notation! Show your work by setting up the calculations. You can use R or a calculator to actually evaluate them!

   (b) Come up with an equation that relates x_i to y_i for each index $i = 1, \ldots, 5$. Do the same for relating x_i to z_i. How does this relationship carry over/affect the mean and the variance of $\mathbf{y}$ in comparison to those of $\mathbf{x}$, if at all? What about the mean and variance of $\mathbf{z}$ in comparison to those of $\mathbf{x}$? Be as specific as possible!

(c) Using what you learned in parts (a) and (b), how does adding the same value c to each element of a dataset affect the mean and variance? How does multiplying each element of the dataset by the same factor d affect the mean and variance? (Note this isn't a rigorous "proof"; come take MATH/STAT 310 to prove this!).

Wednesday 9/17

4. We have the wait times (in minutes) of 17 different customers at Heymaker:

$$x = (2, 2, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 8, 8, 9, 10, 20)$$

   (a) Draw out a boxplot of the data.

   (b) If Heymaker wants to advertise "90% of our customers wait under X minutes", based off this sample, what should X be? What is this quantity more generally called?

   (c) Are there any potential outliers? If so, what might be an explanation for their value(s)?

   (d) What does this boxplot suggest about customer experience?

5. Robust statistics.

   (a) The mode (or a mode) is a sample is defined as the most frequent value. Do you believe the mode is a robust statistic? Why or why not?

   (b) Let's consider another measure of spread: for a sample, define a new measure of spread s^* as the median of absolute (value) deviations from the median. Suppose a teacher records the number of hours that seven students spent studying for a test:

$$x = (4, 5, 5, 6, 6, 7, 20)$$

   Find the IQR, sample standard deviation s, and the new measure of spread s^*.

   (c) Based on your work above, does s^* appears to be a robust statistics? Why or why not?

   (d) Why might we prefer to use this new statistic s^* over IQR?

Thursday 9/18

6. Problems in the associated .qmd file. For grading purposes, R problems 0-3 will be treated as one problem, as will 5-7.

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.