DSC 40A - Group Work Session 2

due January 10, 2022 at 11:59pm

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. **One person** from each group should submit your solutions to Gradescope and **tag all group members** so everyone gets credit.

This worksheet won't be graded on correctness, but rather on good-faith effort. Even if you don't solve any of the problems, you should include some explanation of what you thought about and discussed, so that you can get credit for spending time on the assignment.

In order to receive credit, you must work in a group of two to four students for at least 50 minutes, at one of the scheduled groupwork sessions. You may not do the groupwork alone or meet outside of the scheduled sessions.

1 Correcting Errors in the Data

Problem 1.

You have a data set consisting of the number of millions of pounds of avocados grown annually in each of the 50 states, and from this data, you calculate that on average, each state produces 8.4 million pounds of avocados. Then you learn that your original data set had an error: Florida actually produces 45 million pounds of avocados per year, but you had recorded this as 4.5 million pounds.

a) Can you correct the mistake and find the correct average avocado production per state?

b) Let A be the average number of avocados produced per state. Which of the following statement(s) must be true, based on this information *alone*? Justify your answers.

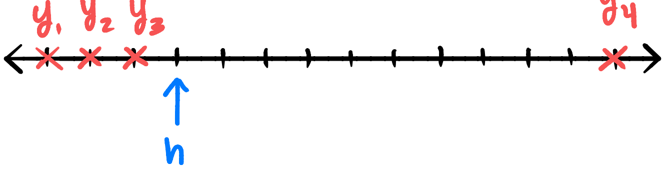
1. At most 25 states produce more than A avocados.
2. At least one state produces less than or equal to A avocados.
3. The number of states that produce more than A avocados is the same as the number of states that produce less than A avocados.

Problem 2.

In general, are you able to correct an average if one of the data values changes? What about the median? The mode?

2 Absolute and Squared Error

As we will see in class, the mean's sensitivity to outliers is due to its role as a minimizer of the mean squared error. Now we'll make this more clear. Suppose we have the data $y_1, \ldots, y_n$ drawn below:



For this problem you do not need to know the exact position of the data points, but, if you like, you can assume that the space between each tick mark is one unit and that $y_1 = 1$, $y_2 = 2$, $y_3 = 3$, $y_4 = 14$ and $h = 4$.

Suppose we start out with the prediction h as shown above. There is a tug-of-war going on in the picture above: y_1, y_2, y_3 want h to move closer to them, while y_4 wants h to move to the right to be closer to it. Who wins depends on the choice of error.

Problem 3.

Suppose that absolute error is used (so that we are trying to minimize mean absolute error). Suppose that h is moved one unit to the left. This increases the error for y_4, but decreases the error for y_1, y_2, y_3. Show that the decrease in $|y_1 - h| + |y_2 - h| + y_3 - h|$ makes up for the increase in $|y_4 - h|$ so that moving h to the left decreases the overall error.

Problem 4.

Now suppose that squared error is used. Again suppose that h is moved one unit to the left. Show that the increase in $(y_4 - h)^2$ is larger than the decrease in $(y_1 - h)^2 + (y_2 - h)^2 + (y_3 - h)^2$, so that moving h to the left increases the overall mean squared error.

Informally, moving h to the left always increases the error associated with y_4, whether the absolute error or squared error is used. That is, y_4 always protests against moving h to the left. This protest isn't strong enough in the case of the absolute error, but if the squared error is used, y_4's voice is amplified, and it wins.