Math 3510: Prob/Stats HW #5: due Wednesday, October 25 SOLUTIONS

1. The mean of the first 80 observations in a data set is 12; the mean of the next 20 observations is 17. Find the mean of the 100 observations taken together. Hint: the answer is **not** $(12+17)/2$!!

$$\frac{80\cdot 12+20\cdot 17}{100}=13$$

2. Always, Sometimes, Never. Put an "A," "S," or "N" in the space next to each statement, according to whether the statement is Always, Sometimes, or Never true. Throughout, all data values are real numbers. (You don't need to explain your answers.)

____S____ The mean of a data set *equals* an actual data value.

____S____ The mean of a data set is a positive number.

____N____ The standard deviation of a data set is a negative number.

____A____ Adding a fixed nonzero number d to each data value in a set of data (that is, replacing each data point x by $x+d$) changes the mean.

____N____ Adding a fixed nonzero number d to each data value in a set of data changes the standard deviation.

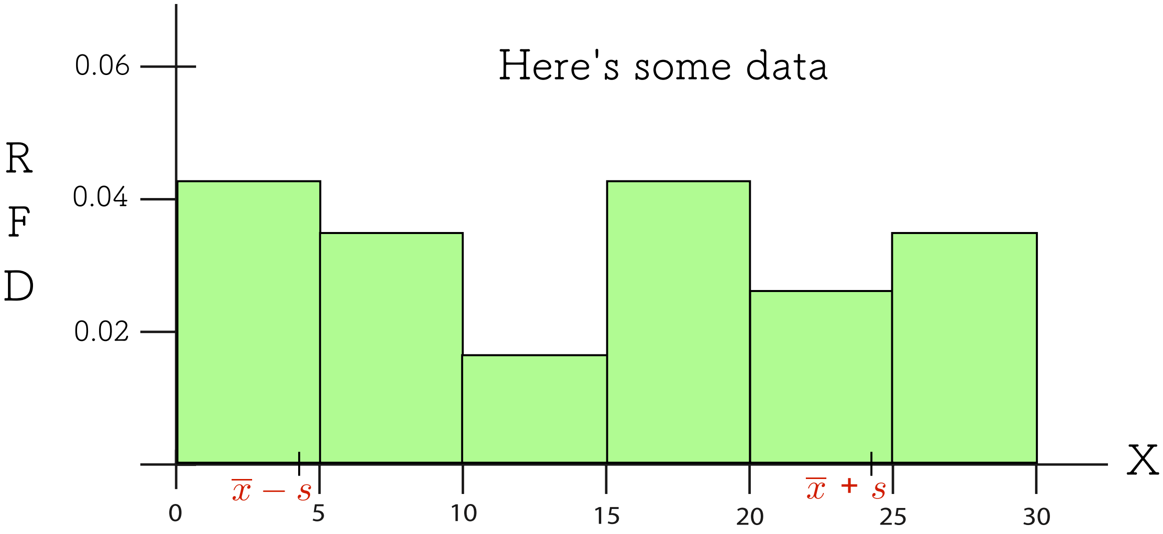
____S____ Adding a data point to a data set (so that the size of the data set increases by one) changes the mean.

____S____ Adding a data point to a data set changes the standard deviation.

3. (This exercise continues on the next page.) Consider the following data set of real numbers:

$$X = \{0.2,\ 0.5,\ 0.6,\ 0.6,\ 2.1,\ 3.7,\ 4.5,\ 5.1,\ 5.6,\ 5.6,\ 7.0,\ 7.4,\ 8.7,\ 10.0,\ 12.1,\ 14.4,\ 16.3,\ 17.4,$$
$$17.5,\ 17.5,\ 17.7,\ 18.1,\ 18.7,\ 20.1,\ 22.2,\ 24.2,\ 24.4,\ 25.0,\ 28.2,\ 28.2,\ 28.9,\ 29.9,\ 27.2\}.$$

(a) Draw a *relative frequency density* histogram for the data, using bins $[0, 5)$, $[5, 10)$, $[10, 15)$, $[15, 20)$, $[20, 25)$, $[25, 30)$. Label your axes as well as the histogram itself. (Don't forget to choose, and mark, a suitable scale on the vertical axis.)



(b) Compute the mean $\overline{x}$ and standard deviation s of the above data.

$\overline{x} = 14.41$, $s = 10.01$.

Label, on the horizontal axis of your above histogram, the points $\overline{x}$, $\overline{x} - s$, $\overline{x} + s$, $\overline{x} - 2s$, $\overline{x} + 2s$, $\overline{x} - 3s$, $\overline{x} + 3s$. (If some of these points lie outside the range of values shown in the histogram, then say so, but you don't need to plot such points.) Use the actual data values, not the histogram in part (a). Note: it's fine if you use a calculator that computes these numbers automatically.

Only $\overline{x} - s = 4.40$ and $\overline{x} + s = 24.42$ lie on the range of values of X shown. See above histogram.

(c) (Continued from previous page.) What percentage of the data from part (a) lies in the interval $(\overline{x} - s, \overline{x} + s)$? ("Percentage" means: *count* the data points in this range, and divide by the total number of points in the entire data set.)

$\frac{21}{33} = 63.64\%$ of the data lies within one standard deviation of the mean.

(d) Repeat part (c) for the interval $(\overline{x} - 2s, \overline{x} + 2s)$.

100% of the data lies in the interval $(\overline{x} - 2s, \overline{x} + 2s)$.

(e) Repeat part (d) for the interval $(\overline{x} - 3s, \overline{x} + 3s)$.

100% of the data lies in the interval $(\overline{x} - 3s, \overline{x} + 3s)$.

4. (This exercise continues on the next page.) The purpose of this exercise is to compare expected value and mean.

   Consider a data set $X = \{x_1, x_2, \dots, x_n\}$ of real numbers.

   (a) Suppose the real number y occurs exactly f times among the data points in X. (That is, exactly f of the numbers $x_1, x_2, \dots, x_n$ equal y.) Find $P(X = y)$, meaning the probability that a point in X, selected at random, has the value y. (Your answer should be in terms of f and n.) Hint 1: just count and divide, using the "equally likely" formula for probabilities. Hint 2: Note that X has n points.

   $P(X = y) = \dfrac{f}{n}$.

   (b) Here's a more general look at part (a). Suppose that the data in X can only take the distinct values $y_1, y_2, \dots y_k$. Suppose the value y_i happens f_i times in X, for each $i = 1, 2, \dots, k$. For each one of these values y_i, find $P(X = y_i)$. Hint: it's just like part (a).

   $P(X = y_i) = \dfrac{f_i}{n}$.

(c) (Continued from the previous page.) Suppose X is as in part (b) above. (That is, the value y_i occurs in X f_i times, for $1 \leq i \leq k$.) Find the expected value of X, meaning the value that you would expect a randomly selected data point in X to have, on average. (Your answer should be in terms of $y_1, y_2, \ldots, y_k$, $f_1, f_2, \ldots, f_k$, and n.) Hint: Use part (b), together with our known formula for expected value:

$$E(X) = y_1 \cdot P(X = y_1) + y_2 \cdot P(X = y_2) + \cdots + y_k \cdot P(X = y_k).$$

$$\begin{aligned} E(X) &= y_1 \cdot P(X = y_1) + y_2 \cdot P(X = y_2) + \cdots + y_k \cdot P(X = y_k) \\ &= y_1 \cdot \frac{f_1}{n} + y_2 \cdot \frac{f_2}{n} + \cdots + y_k \cdot \frac{f_k}{n} \\ &= \frac{y_1 f_1 + y_2 f_2 \cdots + y_k f_k}{n} = \overline{x}, \end{aligned}$$

the last step by the formula for the mean of grouped data. (See, for example, Definition 6.1.1(b) on page 276 of the "Statistics Notes" on our Canvas page.)

(d) How does the value you found for $E(X)$ in part (c) of this exercise compare with the mean $\overline{x}$ of the data in X? Hint: review the "grouped data" formula for the mean that we did in class.

They're equal.

5. (a) A fair, six-sided die (with sides numbered 1 through 6) is tossed 12,000 times. Consider the data set X that consists of all the numbers that come up on these 12,000 trials.

About what would you expect the mean $\overline{x}$ of this data set to be? Please explain.

The mean should be about 3.5, since we would expect the numbers 1, 2, and 3 to come up as often as the numbers 4, 5, and 6.

(b) Now suppose an *unfair* six-sided die, which heavily favors threes and fours (that is, each of these numbers is more likely to come up than is a one, two, five, or six), is tossed 12,000 times. Suppose the mean of these tosses is the same as in part (a) of this exercise. Is the standard deviation of these 12,000 tosses likely to be smaller than, about the same as, or larger than the standard deviation that came from the fair die? Explain.

The standard deviation should be smaller. Standard deviation measure spread, and if the data is concentrated near the mean, the data is less spread out.

(c) Answer the question from part (b) above, but this time with an unfair die that heavily favors ones and sixes.

The standard deviation should be larger. Again, standard deviation measure spread, and if the data is concentrated away from the mean, the data is less spread out.