HW #9: due Wednesday, November 13

Please **read** Section 1.1 of the "Statistics Notes" on our Canvas page, and please **do** all of the exercises below. You can write your answers directly on these pages (there's plenty of space), or use your own paper.

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 negative number.

   __S__ The standard deviation of a data set is a positive 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.

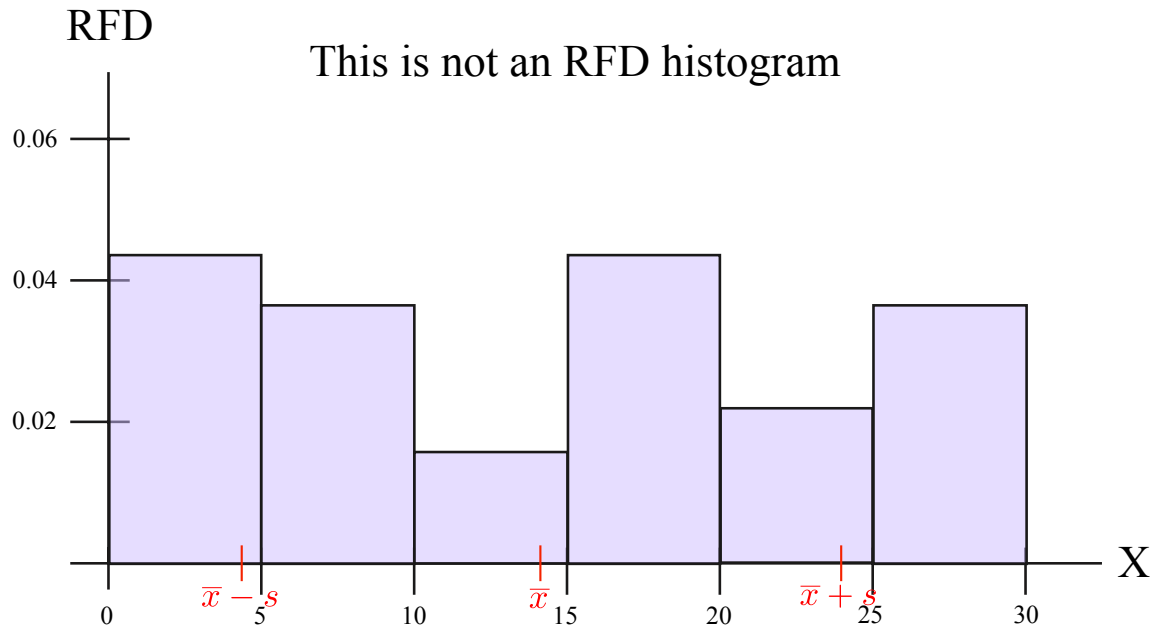
   __S__ Adding a data point to a data set (so that the size of the data set increases by one) changes the mean.

   __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 above data, using bins $[0, 5)$, $[5, 10)$, $[10, 15)$, $[15, 20)$, $[20, 25)$, $[25, 30)$. Label your axes, and give a title to the histogram itself. (Call it anything you want. Be creative – but appropriate.) You can use the axes below for a template, or draw your own.



(b) Compute the mean $\overline{x}$ and standard deviation s of the above data.
$\overline{x} = 14.2303$, $s = 9.70291$.
Label, on the horizontal axis of your 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.) In computing $\overline{x}$ and s, 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.

(c) (Continued from previous page.) What proportion of the data from part (a) lies in the interval $(\overline{x} - s, \overline{x} + s)$? ("Proportion" means: *count* the data points in this range, and divide by the total number of points in the entire data set.) Please express your answer as a decimal to at least four places, *and* as percent to at least two places.

$$\frac{18}{33} = 0.5455 = 54.55\%.$$

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

$$\frac{33}{33} = 1.0000 = 100.00\%.$$

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

$$\frac{33}{33} = 1.0000 = 100.00\%.$$

4. (This exercise continues on the next page.)

(a) MATH 4510 Instructor Dr. Probably rolls a fair, ten-sided die 1,000 times, and each time records the number (from 1 through 10) that comes up. About what will the mean $\overline{x}$ of the resulting data set probably be?

About 5.5, because this is the average of the numbers 1 through 10 (also, if $E[X]$ is the number that comes up on the die, then one computes that $E[X] = 5.5$).

(Continued from the previous page.)

(b) Dr. P. flips 11 fair coins and records the number of heads that come up; Dr. P. repeats this experiment 1,000 times. About what will the mean $\overline{x}$ of the resulting data set probably be?

About 5.5, we would expect half of the coins to come up heads, on average (also, if $E[X]$ is the number of heads that come up, then one computes that $E[X] = 5.5$).

(c) How will the standard deviations of the data sets in parts (a) and (b) of this exercise compare? Please explain.

The standard deviation will probably be larger in the first case, for the following reason. All numbers on the die are equally likely, so numbers FAR from the mean of 5.5 are just as likely as those CLOSE to the mean, so the data is quite spread out. On the other hand, if you flip a fair coin 11 times, then only RARELY will you get zero heads, or 11 heads, or 1 head, or 10 heads: numbers close to the mean of 5.5 are MORE likely than those far away. So the data is less spread out from the mean, so the standard deviation should be smaller.

5. Suppose Dr. P. rolls the die in problem 4(a), above, 50,000 instead of 1,000 times. Is the mean likely to change much? What about the standard deviation? Explain.

Neither should change much. The mean should be about 5.5 after a large number of rolls; it shouldn't change much as you roll more and more. Nor should the spread of the data, so the standard deviation shouldn't change much either. (It's perhaps not as obvious that the standard deviation doesn't change much. But this may be seen by considering the formula for standard deviation of grouped data that we saw in class. If you roll your die 50,000 instead of 1000 times, then each of the frequencies in the numerator (inside the square root) should increase by roughly a factor of 50. Moreover, the denominator (inside the square root) will change from $1{,}000-1$ to $50{,}000-1$, which is a change by roughly a factor of 50. So overall, the standard deviation shouldn't change much.)

6. (This exercise continues on the next page.) Three fair, six-sided dice (with sides numbered 1 through 6) are tossed, and the *sum* on the dice is recorded. This experiment is repeated 100,000 times. (Actually this experiment was simulated; no actual dice were tossed.) Let X be the dataset consisting of all 100,000 recorded sums.

   Here's a frequency table for the sums that came up:

Sum on dice	Frequency	Sum on dice	Frequency
3	435	11	12430
4	1428	12	11482
5	2811	13	9920
6	4515	14	6939
7	7052	15	4619
8	9710	16	2696
9	11539	17	1422
10	12493	18	509

   (a) Compute the mean $\overline{x}$ and standard deviation s of this data set.

$$\overline{x} = \frac{435 \cdot 3 + 1428 \cdot 4 + 2811 \cdot 5 + \cdots + 509 \cdot 18}{100000} = 10.5036;$$

$$s = \sqrt{\frac{435 \cdot (3-\overline{x})^2 + 1428 \cdot (4-\overline{x})^2 + 2811 \cdot (5-\overline{x})^2 + \cdots + 509 \cdot (18-\overline{x})^2}{99999}}$$

$$= \sqrt{8.77175} = 2.96171.$$

(Continued from the previous page; also, continued on the next page.)

(b) This time let Y be the *random variable* given by the sum of the numbers when three fair, six-sided dice are rolled. Compute the expected value $E[Y]$ and the standard deviation $\mathrm{SD}[Y] = \sqrt{\mathrm{Var}[Y]}$ of Y. Assume that the dice land independently of each other. Hint: $Y = Y_1 + Y_2 + Y_3$, where Y_i is the number on the ith die. Compute $E[Y_1]$ and $\mathrm{Var}[Y_1]$ using the usual formulas. Of course, you'll get the same numbers for $E[Y_2]$ and $\mathrm{Var}[Y_2]$, and for $E[Y_3]$ and $\mathrm{Var}[Y_3]$, respectively.

Now, since the dice are independent, you can use the sum rules for expected value and variance. See, for example, the formula sheet for Exam 2 for these sum rules.

$$E[Y_1] = E[Y_2] = E[Y_3] = \frac{1}{6}\sum_{i=1}^{6} i = 3.5;$$

$$\mathrm{Var}[Y_1] = \mathrm{Var}[Y_2] = \mathrm{Var}[Y_3] = \frac{1}{6}\sum_{i=1}^{6} (i - 3.5)^2 = 2.91667.$$

So

$$E[Y] = E[Y_1] + E[Y_2] + E[Y_3] = 10.5;$$

$$\mathrm{Var}[Y] = \mathrm{Var}[Y_1] + \mathrm{Var}[Y_2] + \mathrm{Var}[Y_3] = 8.75;$$

$$\mathrm{SD}[Y] = \sqrt{8.75} = 2.95804.$$

(Continued from the previous page; also, continued on the next page.)

(c) How do $\overline{x}$ and s compare with $E[Y]$ and $\sqrt{\text{Var}[Y]}$?

$\overline{x}$ and s are quite close to $E[Y]$ and $\sqrt{\text{Var}[Y]}$, respectively.

(d) For the actual data set X, compute the proportion of the time a sum of 7 arises. Do the same for a sum of 16.

7 arises $\frac{7052}{100000} = .07052 = 7.052\%$ of the time; 16 arises $\frac{2696}{100000} = .02696 = 2.696\%$ of the time.

(e) For the random variable Y of part (d) above, compute $P(Y = 7)$, by actually counting the outcomes that add up to 7. Do the same for $P(Y = 16)$. How do $P(Y = 7)$ and $P(Y = 16)$ compare with your empirical proportions from the previous part of this problem?

$$P(Y = 7) = \frac{\big|\{115, 124, 133, 142, 151, 214, 223, 232, 241, 313, 322, 331, 412, 421, 511\}\big|}{6^3}$$
$$= \frac{15}{6^3} = 0.0694444;$$
$$P(Y = 16) = \frac{\big|\{466, 556, 565, 646, 655, 664\}\big|}{6^3} = \frac{6}{6^3} = 0.0277778.$$

$P(Y = 7)$ and $P(Y = 16)$ are not that far off from the empirical proportions from the previous part of this problem.

(Continued from the previous page; also, continued on the next page.)

(f) Finally, if you were to draw a histogram of the data given at the start of this problem, what shape would it have? What theorem suggests this? What would happen if you did a similar experiment for the sum on 6 dice, or 10 dice, or 100 dice?

It would look something like a normal curve, more and more so the larger the number of dice. This is all suggested by the Central Limit Theorem.