Name: _____ **Date:** _____

PSTAT 5A: Discussion Worksheet 01

Spring 2023, with *Ethan P. Marzban*

Welcome to the first PSTAT 5A Discussion Section! We encourage you to solve the following problems in groups. Statistics and Data Science are not meant to be lonely fields- we have quite a bit we can learn from each other! Your TA will go over what you need to do in order to receive credit for Discussion Section, so make sure to attend!

1. Consider the list of numbers $X = \{-1,\ 0,\ 1.5,\ 2,\ 3,\ 3,\ 4,\ 5,\ 10\}$.

(a) Compute $\overline{x}$, the mean of X.

Solution:

$$\overline{x} = \frac{1}{9}(-1+0+1.5+2+3+3+4+5+10) = \frac{1}{9}\cdot(27.5) = \frac{55}{18} = 3.0\overline{5}$$

(b) Compute median(X), the median of X. (HW01 provides a formula for how to compute the median of a list of numbers.)

Solution:

$$\cancel{-1},\ \cancel{0},\ \cancel{1.5},\ \cancel{2},\ 3,\ \cancel{3},\ \cancel{4},\ \cancel{5},\ \cancel{10} \implies \text{median}(S) = 3$$

(c) Compute the standard deviation of X.

Solution: We first compute the variance: denoting the elements of X by x_i, we have

$$\begin{aligned} s_x^2 &= \frac{1}{9-1}\sum_{i=1}^{5}(x_i-\overline{x})^2 \\ &= \frac{1}{8}\left[\left(-1-\frac{55}{18}\right)^2+\left(0-\frac{55}{18}\right)^2+\left(1.5-\frac{55}{18}\right)^2+\left(2-\frac{55}{18}\right)^2+\left(3-\frac{55}{18}\right)^2+\right. \\ &\qquad \left.\left(3-\frac{55}{18}\right)^2+\left(4-\frac{55}{18}\right)^2+\left(5-\frac{55}{18}\right)^2+\left(10-\frac{55}{18}\right)^2\right] = \frac{1}{8}\cdot\frac{740}{9} = \frac{185}{18} \end{aligned}$$

Therefore, the standard deviation is $\sqrt{\frac{185}{18}} \approx 3.2059$

Name: _____ **Date:** _____

2. In the parts below, you will be provided with the description of a particular dataset. Identify the type of visualization (e.g. histogram, scatterplot, etc.) that you believe would best achieve the stated goal, and provide a brief justification for your answer. Keep in mind that there are potentially multiple "correct" answers!

(a) A statistician is interested in visualizing the relationship between heights and weights of students at UCSB.

> **Solution:** Height and weight are both numerical variables; as such, the best type of plot to visualize their relationship to one another is a **scatterplot**.

(b) A clinical researcher has administered 4 different dosages of a particular medicine to a large set of volunteers, and would like to visualize how (if at all) the insulin levels of subjects varies across dosages.

> **Solution:** Because we are told that there are only 4 dosage levels, we can treat dosage as a categorical variable (with 4 categories). Insulin level is a continuous variable; therefore, we are trying to determine the relationship between a continuous variable and a categorical one, meaning we should use a **side-by-side boxplot**.
>
> If, however, we treated 'dosage' as numerical, then the answer would be a **scatterplot**. However, because there are so few distinct dosages, it makes more sense to view 'dosage' as a categorical variable.

(c) A soccer fan has tallied up the number of times every country has made it into the World Cup, and would like to visualize their data.

> **Solution:** The "number of times a country makes it into the world cup" is a numerical (discrete) variable; as such, to visualize its distribution we should use a **histogram**.

(d) Morgan has collected information on how long it takes a sample of 100 athletes to complete an obstacle course, and would like to visualize the distribution of completion times.

> **Solution:** "Completion time" is a numerical (continuous) variable; as such, a **histogram** is best suited to visualize its distribution.

Name: _____ **Date:** _____

3. (**A Bit of Math**) Consider two sets of data $X = \{x_i\}_{i=1}^{n}$ and $Y = \{y_i\}_{i=1}^{m}$ (note that X and Y have different sizes!) Define the set $Z = X \cup Y$ (i.e. $Z = \{x_1, \cdots, x_n, y_1, \cdots, y_m\}$); show that

$$\overline{Z} = \left(\frac{n}{n+m}\right) \cdot \overline{X} + \left(\frac{m}{n+m}\right) \cdot \overline{Y}$$

How does the formula for $\overline{Z}$ simplify if $n = m$ (i.e. if the two sets of data have the same size)?

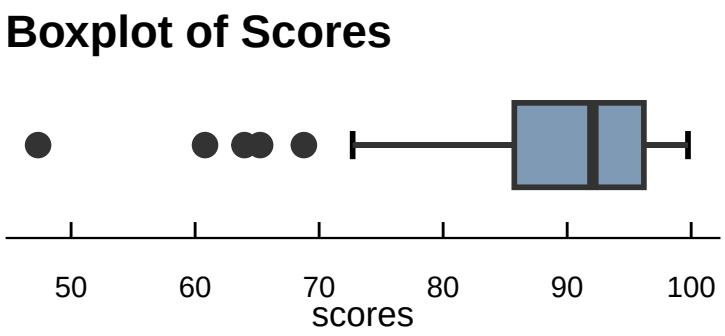
Solution: We begin with the definition of $\overline{Z}$:

$$\begin{aligned}
\overline{Z} &= \frac{1}{n+m} \sum_{i=1}^{n+m} z_i \\
&= \frac{1}{n+m} \left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i \right) \\
&= \frac{1}{n+m} \left(n \cdot \underbrace{\frac{1}{n} \sum_{i=1}^{n} x_i}_{:=\overline{X}} + m \cdot \underbrace{\frac{1}{m} \sum_{i=1}^{m} y_i}_{:=\overline{Y}} \right) \\
&= \frac{1}{n+m} \left(n \cdot \overline{X} + m \cdot \overline{Y} \right) = \left(\frac{n}{n+m}\right) \cdot \overline{X} + \left(\frac{m}{n+m}\right) \cdot \overline{Y}
\end{aligned}$$

If $n = m$, then

$$\overline{Z} = \frac{1}{n+n} \left(n \cdot \overline{X} + n \cdot \overline{Y} \right) = \frac{1}{2n} \left(n \cdot \overline{X} + n \cdot \overline{Y} \right) = \frac{1}{2}\overline{X} + \frac{1}{2}\overline{Y}$$

4. The boxplot of scores on a particular exam is given below:



Name: _____ **Date:** _____

(a) What score was the median score on this exam?

> **Solution:** The median is denoted by the dark line in the middle of the boxplot; for this particualr plot, it appears to be near 93% .

(b) Kara knows that she performed better than 25% of the class. What was Kara's score?

> **Solution:** If Kara performed better than 25% of the class, then her score is at the 25^{th} percentile of scores. The 25^{th} percentile is the same as the **first quartile**, which is denoted by the left boundary of the box on the boxplot- for this particular plot, this appears to be near 85.5% .

(c) How many outliers are there?

> **Solution:** Outliers appear on boxplots as points to either the left or the right of the whiskers- on this particular plot, we see 5 such points (all in the negative direction).

(d) Provide the 5-number summary for the distribution of exam scores.

> **Solution:** Recall that the 5-number summary consists of: the minimum, the first quartile, the median, the third quartile, and the maximum. The median appears to be around 48; the first and third quartiles are the ends of the box (which appear to be around 85.5 and 97.5); the median we found in part (a) to be 93; the maximum appears to be 100. Hence, the 5-number summary takes the form
>
> | Min. | Q_1 | Median | Q_3 | Max |
> |---|---|---|---|---|
> | 48 | 85.5 | 93 | 97.5 | 100 |

(e) Compute the Interquartile Range (IQR) of exam scores.

> **Solution:** We have IQR $= Q_3 - Q_1 \approx 97.5 - 85.5 \approx$ 12

