Name: _____ **Date:** _____

PSTAT 5A: Homework 01

Summer Session A 2023, with *Ethan P. Marzban*

As a reminder, homework is neither collected nor graded. We encourage you to stop by Office Hours to ask any questions you may have about your work, or the problems themselves!

1. Consider the list of numbers $X = \{-3,\ -1,\ 0,\ 0.4,\ 0.7,\ 3.9,\ 6\}$.

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

Solution:

$$\overline{x} = \frac{1}{7}\left[(-3) + (-1) + (0) + (0.4) + (0.7) + (3.9) + (6)\right] = \frac{1}{7} \cdot (7) = 1$$

(b) Compute median(X), the median of X.

Solution:

$$\cancel{-3},\ \cancel{-1},\ \cancel{0},\ 0.4,\ \cancel{0.7},\ \cancel{3.9},\ \cancel{6} \implies \text{median}(S) = 0.4$$

(c) Compute the standard deviation of X.

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

$$\begin{aligned} s_x^2 &= \frac{1}{7-1}\sum_{i=1}^{7}(x_i - \overline{x})^2 \\ &= \frac{1}{6}\Big[(-3-1)^2 + (-1-1)^2 + (0-1)^2 + (0.4-1)^2 + (0.7-1)^2 \\ &\qquad + (3.9-1)^2 + (6-1)^2 + \Big] = \frac{1}{6} \cdot \frac{2743}{50} = \frac{2743}{300} \end{aligned}$$

Therefore, the standard deviation is $\sqrt{\frac{2743}{300}} = \frac{\sqrt{8229}}{30} \approx 3.023795$

(d) Compute the range of X.

Name: _____ **Date:** _____

Solution:

$$\text{range}(X) = \max\{X\} - \min\{X\} = 6 - (-3) = 9$$

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- as such, your explanation/justification will be very important!

(a) An environmental scientist would like to see how (if at all) PM2.5 concentration (which is a measure of air quality) varies with temperature (as measured in Centigrade).

Solution: PM2.5 concentration and temperature are both numerical variables, meaning the best tool to visualize the relationship between them 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) An avid watcher of *Eurovision* has tallied up the number of times each country has won the competition, and would like to visualize their data.

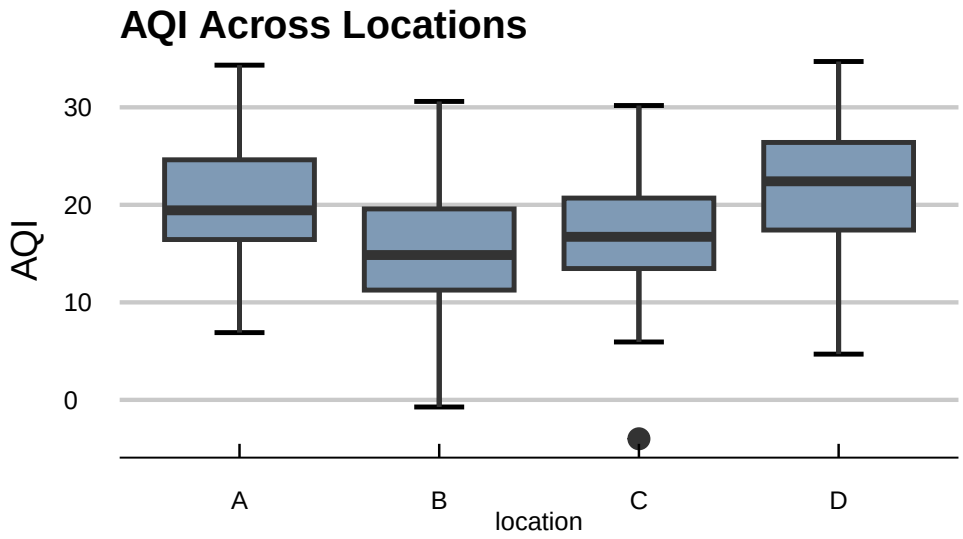
Solution: The "number of times a country wins *Eurovision*" is a numerical (discrete) variable; as such, to visualize its distribution we should use a **histogram**.

(d) Alex has collected information on how long it takes a sample of 60 PSTAT 5A students to complete the Final Exam, and would like to visualize the distribution of completion times.

Name: _____ **Date:** _____

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

3. A researcher has collected measurements on the AQI (Air Quality Index) at several locations in Santa Barbara. She has labeled her locations A through D, and collected 100 AQI measurements from each location. The results of her study are displayed graphically below:



(a) What is the (approximate) median AQI at location A?

Solution: We know that, on a boxplot, the median is depicted by the dark bar in the middle of the box. As such, we see that the median AQI at location A is around 20.

(b) Approximately what percent of AQI readings at Location D were less than 16?

Solution: We know that, on a boxplot, the first quartile is depicted by the bottom of the box. This means that the first quartile of AQI readings at location D is around 16, meaning around 25% of AQI readings at Location D were less than 16.

(c) Are there any outliers at any of the locations?

Solution: Outliers are depicted by points/circles outside of the reach of the whiskers. Only Location C has such points, meaning only Location C had outliers.

(d) From the plot, does it appear that the different locations have different AQI readings? Explain briefly.

> **Solution:** It does appear that the average AQI readings in the different locations are different. Specifically, it seems that Locations B and C have slightly lower average AQI readings than Locations A and D.

4. Consider a list of numbers $X = \{x_i\}_{i=1}^n$, and another list of numbers $Y = \{x_i + b\}_{i=1}^n$ where b is a fixed constant. In other words, the elements of Y are found by taking the elements of X and adding b.

(a) Show that $\overline{y} = \overline{x} + b$.

> **Solution:**
>
> $$\begin{aligned}
> \overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i && \text{[Definition of mean]} \\
> &= \frac{1}{n}\sum_{i=1}^{n} (x_i + b) && \text{[Definition of } y_i\text{]} \\
> &= \frac{1}{n}\left(x_1 + b + x_2 + b + \cdots + x_n + b\right) && \text{[Definition of Sigma Notation]} \\
> &= \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n + \underbrace{b + b + \cdots + b}_{n \text{ times}}\right) && \text{[Commutativity of Addition]} \\
> &= \frac{1}{n}\left[(x_1 + x_2 + \cdots + x_n) + n \cdot b\right] && \text{[Factorization]} \\
> &= \frac{1}{n} \cdot (x_1 + x_2 + \cdots + x_n) + \frac{1}{\cancel{n}} \cdot \cancel{n} \cdot b && \text{[Factorization]} \\
> &= \overline{x} + b && \text{[Definition of mean]}
> \end{aligned}$$

(b) What is s_Y^2 in terms of s_X^2?

> **Solution:**
>
> $$\begin{aligned}
> s_Y^2 &= \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \overline{y})^2 && \text{[Definition of } s_Y^2\text{]} \\
> &= \frac{1}{n-1}\sum_{i=1}^{n} \left[(x_i + b) - (\overline{x} + b)\right]^2 && \text{[Definition of } y_i\text{; answer to (a)]}
> \end{aligned}$$

Name: _____ **Date:** _____

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i + \cancel{b} - \overline{x} - \cancel{b})^2 \qquad \text{[Expansion]}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{[Simplifying]}$$

$$= s_X^2 \qquad \text{[Definition of } s_X^2\text{]}$$

There is a bit of intuitive sense to this answer as well. Recall that variance is a measure of spread. If we take a set of numbers $\{x_i\}_{i=1}^n$ and shift them all the same number of units (b), the spread will not change!

5. A recent survey revealed that 40% of UCSB students own an *X-Box*, 70% own a *Playstation*, and 30% own both an *X-Box* and a *Playstation*. A UCSB student is selected at random.

Solution: Let X denote the event "student owns an *X-box*" and P denote the event "student owns a *Playstation*". From the information provided, we have

$$\mathbb{P}(X) = 0.4; \quad \mathbb{P}(P) = 0.7; \quad \mathbb{P}(X \cap P) = 0.3$$

(a) What is the probability that this student owns an *X-Box* or a *Playstation* (or both)?

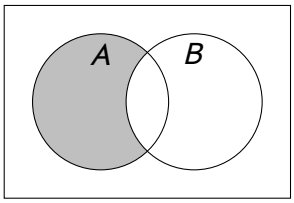
Solution: We seek $\mathbb{P}(X \cup P)$, which can be computed using the **Addition Rule**:

$$\begin{aligned}\mathbb{P}(X \cup P) &= \mathbb{P}(X) + \mathbb{P}(P) - \mathbb{P}(X \cap P) \\ &= (0.4) + (0.7) - (0.3) = 0.8\end{aligned}$$

(b) What is the probability that this student owns a *Playstation* but not an *X-Box*?

Solution: We seek $\mathbb{P}(P \cap X^C)$. Let's see first derive a formula for $\mathbb{P}(A \cap B^C)$ for arbitrary sets A and B, using a Venn Diagram:

Name: _____ **Date:** _____



We see that, in general,

$$\mathbb{P}(A \cap B^C) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$$

Therefore, we have

$$\mathbb{P}(P \cap X^C) = \mathbb{P}(P) - \mathbb{P}(X \cap P) = 0.7 - 0.3 = \boxed{0.4}$$

(c) What is the probability that this student owns neither an *X-Box* nor a *Playstation*?

Solution: We seek $\mathbb{P}(X^C \cap P^C)$. The **Complement Rule** tells us that

$$\mathbb{P}(X^C \cap P^C) = 1 - \mathbb{P}\left[(X^C \cap P^C)^C\right]$$

DeMorgan's Laws tell us that

$$(X^C \cap P^C)^C = X \cup P$$

meaning, putting everything together,

$$\mathbb{P}(X^C \cap P^C) = 1 - \mathbb{P}(X \cup P) = 1 - (0.8) = \boxed{0.2}$$

6. Two numbers are to be selected at random and with replacement from the set $\{1, 2, 3, 4, 5\}$.

(a) Express the outcome space for this experiment using a table.

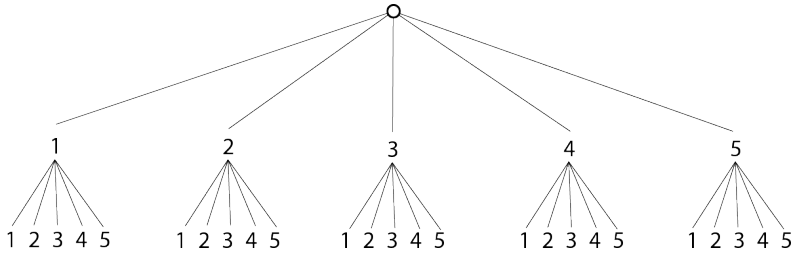
Name: _____ **Date:** _____

Solution: Letting rows denote the first number selected and columns denote the second number selected, we have

	1	2	3	4	5
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)

(b) Express the outcome space for this experiment using a tree.

Solution:



(c) How many elements are in the outcome space?

Solution: Counting either the elements in our table from part (a) or the number of leaves in our tree from part (b), we see that the total number of outcomes in Ω is 25.

(d) For each of the following events, write down the mathematical formulation of the event described (i.e. list out the outcomes comprised in each event):

i. A = the first number selected was 3

Solution: $A = \{(3,1),\ (3,2),\ (3,3),\ (3,4),\ (3,5)\}$

ii. B = the second number selected was even

Solution:

$$B = \{(1,2),\ (2,2),\ (3,2),\ (4,2),\ (5,2)$$
$$(1,4),\ (2,4),\ (3,4),\ (4,4),\ (5,4)\}$$

iii. C = the first number selected was 3 and the second number selected was even.

Solution:

$$C = A \cap B = \{(3,2),\ (3,4)\}$$

iv. D = the first number selected was odd, and the first number selected was even.

Solution: There are no outcomes in which the first number is simultaneously even and odd; hence $D = \varnothing$.

v. E = the first number was strictly greater than the second.

Solution:

	1	2	3	4	5
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)

(e) Are we justified in using the Classical Approach to Probability in this problem? Why or why not?

Solution: We are justified in using the Classical Approach, because the numbers are stated to be selected at random.

(f) Use the Classical Approach to Probability to compute the probabilities of the events listed out in part (d) above.

Solution: For each probability, we simply count the number of elements in the event and divide by the total number of outcomes in Ω (which we already computed to be 25).

(i) $\mathbb{P}(A) = \frac{1}{5} = 0.2 = 20\%$

(ii) $\mathbb{P}(B) = \frac{2}{5} = 0.4 = 40\%$

(iii) $\mathbb{P}(C) = \frac{2}{25} = 0.08 = 8\%$

(iv) $\mathbb{P}(D) = 0$

(v) $\mathbb{P}(E) = \frac{2}{5} = 0.4 = 40\%$

7. On a particular website, passwords must be exactly 7 characters long and consist of 3 letters (A through Z), followed by 2 digits (0 through 9), followed by another letter (A through Z), followed by a special character (!, @, #, $, %).

(a) How many passwords can be created using this scheme, assuming repeated letters, digits, and characters *are* allowed?

Solution: We use a Slot Diagram, with 7 slots (one for each of the characters in the password):

$$\underbrace{26 \times 26 \times 26}_{\text{3 letters}} \times \underbrace{10 \times 10}_{\text{2 digits}} \times \underbrace{26}_{\text{1 letter}} \times \underbrace{5}_{\text{1 special char.}} = (26)^4 \cdot (10)^2 \cdot (5)$$

which amounts to 228,488,000 total possible passwords.

(b) How many passwords can be created using this scheme, assuming repeated letters, digits, and characters *are not* allowed?

Solution: We again use a Slot Diagram, with 7 slots (one for each of the characters in the password); this time our slots need to be modified slightly to account for the fact that repeated characters are not allowed:

$$\underbrace{26 \times 25 \times 24}_{\text{3 letters}} \times \underbrace{10 \times 9}_{\text{2 digits}} \times \underbrace{23}_{\text{1 letter}} \times \underbrace{5}_{\text{1 special char.}} = (26)_4 \cdot (10)_2 \cdot (5)$$

which amounts to 161,460,000 total possible passwords. Intuitively, we know that our answer should have been lower than our answer to part a (which it is!) because we have imposed an additional condition passwords must satisfy, thereby decreasing the total number of possible passwords.

(c) Suppose now that the letters must still appear together, the digits must still appear together, and the special characters must still appear together, but the order in which these three categories of characters appear is now free to vary. For example, `%A122AAB` is now a valid password. (Again assume that repeated letters/digits/characters *are* allowed.) How many passwords can be created using this new scheme?

Solution: Here is how I like to think about this: let's start by fixing the order of characters to be `LLLDDLS` (where `L` denotes "letter", `D` denotes "digit", and `S` denotes "special character".) Finding the number of such passwords is easy- we actually did that in part (a)!

Of course, that is not our final answer- we still need to take into account the fact that the types of characters (letters, digits, special characters) can appear in any order. The fact that the character types must remain together allows us to think of four "blocks": one for the first group of letters, one for the group of digits, one for the second group of letters, and one for the special characters.

Here is what I mean: instead of focusing on the individual letters/digits/characters, we can now think of our password as consisting of 4 blocks:

LLL	DD	L	S

The order in which these blocks appear next to each other is not fixed; so, we need to multiply our answer from part (a) by the number of ways to reorder these 4 blocks in a line. We know that the number of ways to reorder 4 blocks in a line is 4!, meaning our final answer is

$$(26)^4 \cdot (10)^2 \cdot (5) \times (4!)$$

