Inferential Statistics and Applied Probability

Assignment # 1

(DEADLINE: 02/10/2023)

REG#: 2023038 NAME: Abdullah Ejaz Janjua

COURSE CODE: DS211 INSTRUCTOR: MUHAMMAD SAJID ALI

TOTAL MARKS: 85

Instructions

- You are free to consult each other for verbal help. However, **copying or sharing the soft/hard copy with each other will not only result in the cancellation of the current assignment, but it may also impact your grade in all the future assignments and exams as well**.
- List your collaborators on the last page of your assignment. Collaborators are any people you discussed this assignment with. This is an individual assignment, so be aware of the course's collaboration policy.
- You must attach this assignment at the top of your solution.

Handwritten Tasks – attempt each of the following task by hand on the A4 page.

Task 1: Prove in general that the standardized dataset $\{\widehat{x_i}\}$ derived from $\{x_i\}$ has a mean of 0 and a standard deviation of 1. **(10 marks)**

Task 2: A tire manufacturer wants to determine the inner diameter of a certain grade of tire. Ideally, the diameter would be 570 mm. The dataset is 572, 572, 573, 568, 569, 575, 565, 570. **(10 marks)**

1. Find the sample mean and median.
2. Find the sample variance, standard deviation, and range.
3. What feature in this data set is responsible for the substantial difference between the two?
4. Using the calculated statistics in parts (1) and (2), can you comment on the quality of the tires?

Task 3: Construct a box plot for the following dataset. These datapoints are the lifetimes, in hours, of fifty 40-watt, 110-volt internally frosted incandescent lamps, taken from forced life tests. What can you comment after constructing the box plot? **(10 marks)**

919 1196 785 1126 936 918 1156 920 948 1067 1092 1162 1170 929 950 905 972 1035 1045 855 1195 1195 1340 1122 938 970 1237 956 1102 1157 978 832 1009 1157 1151 1009 765 958 902 1022 1333 811 1217 1085 896 958 1311 1037 702 923

Task 4: Consider the following data of the measures of the diameters of 36 rivet heads in 1/100 of an inch. **(10 marks)**

6.72 6.77 6.82 6.70 6.78 6.70 6.62 6.75 6.66 6.66 6.64 6.76 6.73 6.80 6.72 6.76 6.76 6.68 6.66 6.62 6.72 6.76 6.70 6.78 6.76 6.67 6.70 6.72 6.74 6.81 6.79 6.78 6.66 6.76 6.76 6.72

1. Compute the sample mean and sample standard deviation.
2. Construct a relative frequency histogram of the data.
3. Comment on whether the histogram graph is symmetric, left or right skewed.

Programming Tasks – for each of the following tasks, you will create a separate Python notebook in Google Colab to analyze the data and answer the specified questions.

Task 5: Consider the nuclear power plants dataset available at Nuclear_plants | DASL (datadescription.com) that includes the cost (in 1976 US dollars), number of megawatts, and year of construction. **(15 marks)**

1. Are there outliers in this data?
2. What is the mean cost of a power plant? What is the standard deviation?
3. What is the mean cost per megawatt? What is the standard deviation?
4. Plot a histogram of the cost per megawatt. Is it skewed? Why?

Task 6: Consider the sodium content and calorie content of three types of hot dog dataset available at Hot dogs. The types are Beef, Poultry, and Meat (a rather disturbingly vague label). Use class conditional histograms to compare these three types of hot dog with respect to sodium content and calories. **(15 marks)**

Task 7: You will find a dataset giving (among other things) the number of 3 or more syllable words in advertising copy appearing in magazines at MegaAds. The magazines are grouped by the education level of their readers; the groups are 1, 2, and 3 (the variable is called GRP in the data). **(15 marks)**

1. Use a box plot to compare the number of three or more syllable words for the ads in magazines in these three groups. What do you see?
2. Use a box plot to compare the number of sentences appearing in the ads in magazines in these three groups. What do you see?

Submission:
Submit both of your handwritten (as a document, i.e. Word or PdF) and programming tasks (as notebooks along the dataset) on *Teams*.

Task 01:

As $\hat{x}_i = \dfrac{x_i - \text{mean}\{x_i\}}{\text{std}\{x_i\}}$

$$\text{mean}\{\hat{x}_i\} = \sum_i^n \frac{\hat{x}_i}{n}$$

$$= \frac{1}{n} \sum_i^n \frac{x_i - \text{mean}\{x_i\}}{\text{std}\{x_i\}}$$

$$= \frac{1}{n} \cdot \frac{1}{\text{std}\{x_i\}} \sum_i^n (x_i - \text{mean}\{x_i\})$$

$$\sum_i^n (x_i - \text{mean}\{x_i\}) = \sum_i^n x_i - \sum_i^n \text{mean}\{x_i\}$$

$$= \sum_i^n x_i - n \times \frac{1}{n} \sum_i^n x_i$$

$$= 0$$

Thus,

$$\text{mean}\{x_i\} = \frac{1}{n} \cdot \frac{1}{\text{std}\{x_i\}} \cdot 0$$

$$= 0$$

As $\sigma = \sqrt{\frac{1}{n} \sum_i^n (x_i - \mu)^2}$

$$\sigma\{\hat{x}_i\} = \sqrt{\frac{1}{n} \sum_i^n (\hat{x}_i - 0)^2} = \sqrt{\frac{1}{n} \sum_i^n (\hat{x}_i)^2}$$

$$= \sqrt{\frac{1}{n} \sum_i^n \left(\frac{x_i - \mu}{\sigma}\right)^2}$$

$$= \sqrt{\frac{1}{n} \cdot \frac{1}{\sigma^2} (x_i - \mu)^2} = \sqrt{\frac{1}{\sigma^2} \cdot \sigma^2} = \sqrt{1} = 1$$

Task 02:

Q1. $\text{mean} = \frac{572 + 572 + 573 + 568 + 569 + 575 + 565 + 570}{8}$

$= 570.5$

$d = 565, 568, 569, 570, 572, 572, 573, 575$

$\text{median} = \frac{\left(\frac{8}{2}\right)^{th} + \left(\frac{8}{2}+1\right)^{th}}{2}$

$= \frac{570 + 572}{2} = 571 mm$

Q2. $\text{Var} = \frac{(572-570.5)^2 + (572-570.5)^2 + (573-570.5)^2 + (568-570.5)^2 + (569-570.5)^2 + (570-570.5)^2 + (572-570.5)^2 + (572-570.5)^2 + (573-570.5)^2 + (575-570.5)^2}{8}$

$= 8.75 mm^2$

$\text{std} = \sqrt{8.75}$

$= 2.95 mm$

$\text{Range} = 575 - 565$

$= 10 mm$

Q4. The tires seem to be of good quality. Most of the tires lie around 570 with small variability indicated by 2.95.

Q3. The data is relatively balanced but median >

mean shows the presense of smaller values.

Task 3:

1	2	3	4	5	6	7	8	9	10
702,	765,	785,	811,	832,	855,	896,	902,	905,	918,

11	12	13	14	15	16	17	18	19	20
919,	920,	923,	929,	936,	938,	948,	950,	956,	958,

21	22	23	24	25	26	27	28	29
958,	970,	972,	978,	1009,	1009,	1022,	1035,	1037,

30	31	32	33	34	35	36	37	38
1045,	1067,	1085,	1092,	1102,	1122,	1126,	1151,	1156,

39	40	41	42	43	44	45	46	47
1157,	1157	1162,	1170,	1195,	1195,	1196,	1217,	1237,

48	49	50
1311,	1333,	1340

$$P_{25} = \frac{25 \times 50}{100} \qquad P_{75} = \frac{75 \times 50}{100}$$

$$= 12.5 \qquad = 37.5$$

$$Q1 = 920 + (923 - 920) \times 0.5 \qquad Q3 = 1151 + (1156 - 1151) \times 0.5$$

$$= 921.5 \qquad = 1153.5$$

$$IQR = Q3 - Q1$$

$$= 921.5 - 1153.5$$

$$= 232$$

$$max = 1153.5 + (232 \times 1.5) \qquad min = 921.5 - (232 \times 1.5)$$

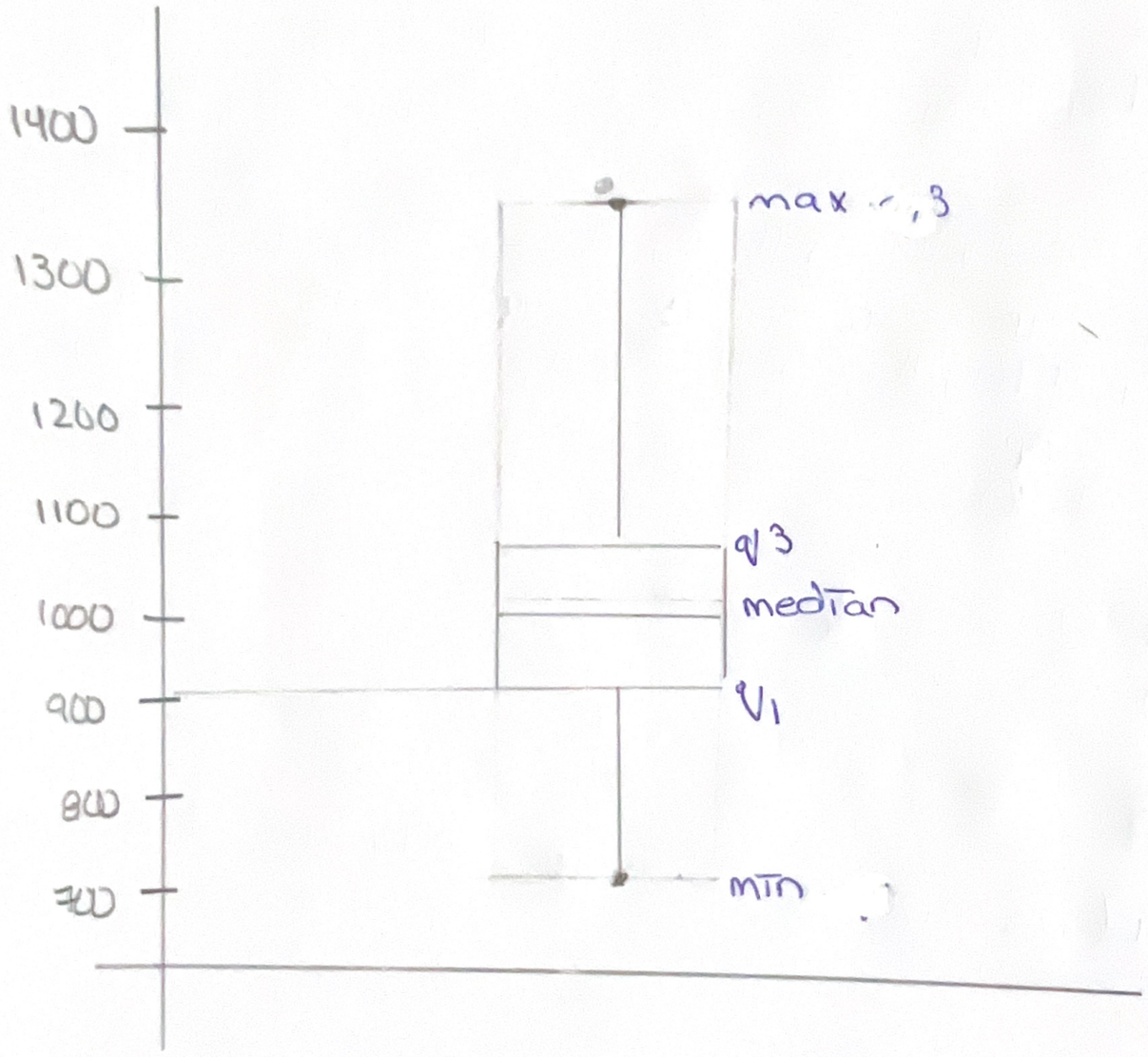
$$= 1501.5 \qquad = 373$$

$$= 1340 \qquad = 702$$

$$\text{Median} = \frac{\left(\frac{50}{2}\right)^{th} + \left(\frac{50}{2}+1\right)^{th}}{2}$$

$$= \frac{1009 + 1009}{2}$$

$$= 1009$$



*The boxplot tells that there is moderate variation in the data.

*There are No outliers in the data

sk 4:

Q1. $\frac{6.72 + 6.77 + 6.82 + 6.70 + \cdots + 6.72}{36}$

mean = 6.726

Standard deviation = $\sqrt{\frac{(6.72-6.726)^2 + (6.77-6.726)^2 + \cdots + (6.72-6.726)^2}{36}}$

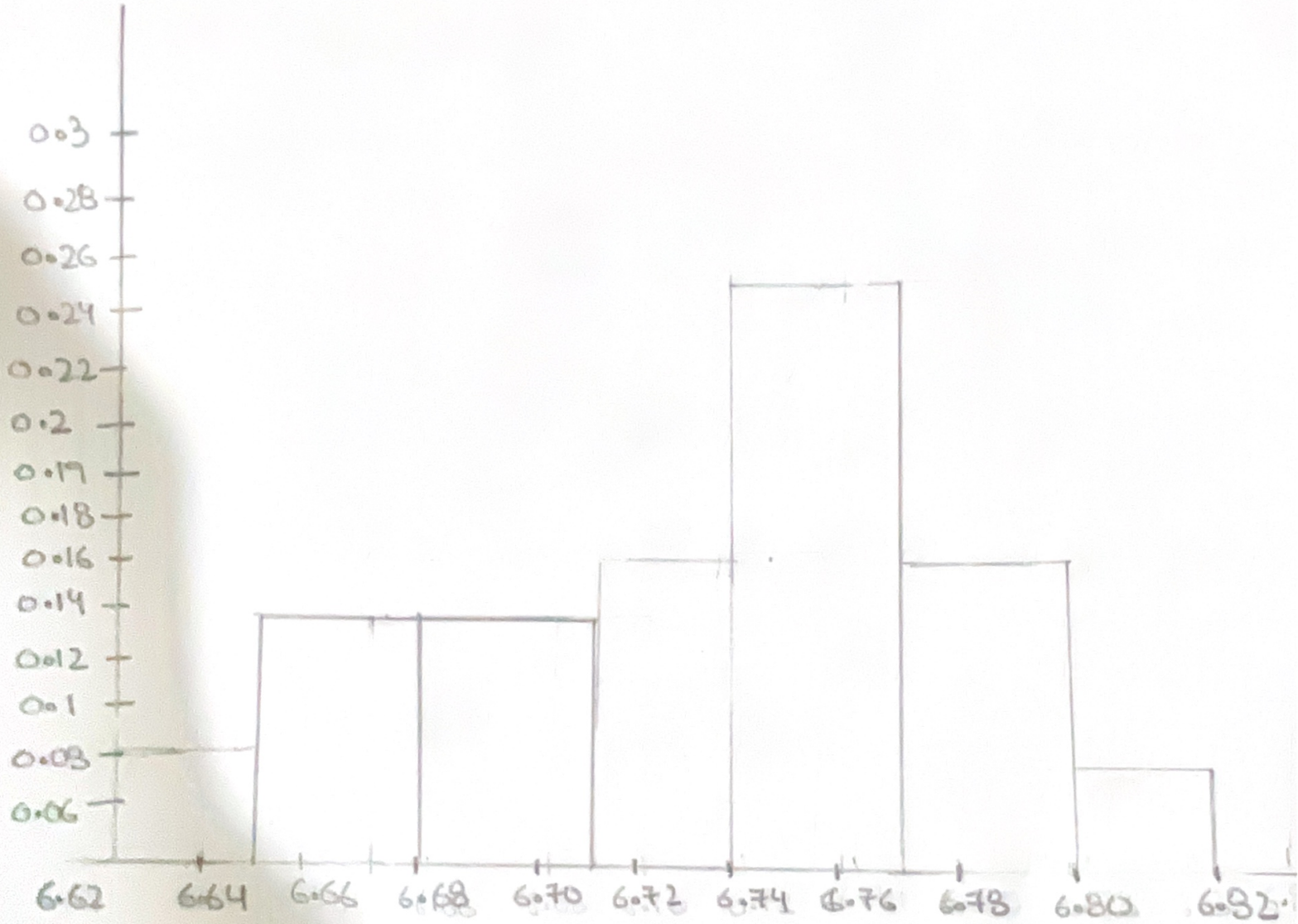
$= 0.0536$

Class width = $\frac{6.82 - 6.64}{6}$, Bins = 6

$= 0.03$

Intervals:	frequency	Relative frequency
6.62 – 6.64	3	3/36 = 0.0833
6.65 – 6.67	5	5/36 = 0.138
6.68 – 6.70	5	5/36 = 0.138
6.71 – 6.73	6	6/36 = 0.167
6.74 – 6.76	9	9/36 = 0.25
6.77 – 6.79	5	5/36 = 0.138
6.80 – 6.82	3	3/36 = 0.0833

Q2.



Q3. The graph is mostly symmetric with peak at the end graph indicating left skew