Tutorial 1 (Descriptive Statistics)

Question 1: Descriptive Statistics - Central Tendency

(a) Define and explain the concept of "central tendency" in descriptive statistics. What are the common measures of central tendency, and under what circumstances is each measure most appropriate?

Central tendency refers to the measure that represents the center or typical value of a dataset. It provides a single value around which data tend to cluster. The common measures of central tendency are:

Mean: The mean is calculated by summing all values in a dataset and dividing by the total number of values. It is suitable for continuous data and provides a balanced representation of the dataset.

Median: The median is the middle value when the data is ordered. If there is an even number of data points, it's the ‘average’ (mean) of the two middle values. The median is a robust measure and less affected by outliers.

Mode: The mode is the value that appears most frequently in the dataset. It is suitable for categorical and nominal data, but a dataset can have zero or multiple modes.

(b) You are given the following dataset representing the ages of 11 individuals in a sample: [25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75]. Calculate the mean, median, and mode of the dataset. Interpret what each of these measures tells you about the distribution of ages in the sample.

Calculating Mean, Median, and Mode:
Dataset: [25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75]

Mean (Average):
(25 + 30 + 35 + 40 + 45 + 50 + 55 + 60 + 65 + 70+ 75) / 11 = 550 / 11 = 50 years

Median:
Since the data is already in ascending order, the median is the middle value, which is the 6th value, therefore:

Median = 50 years.

Mode:
There is no mode in this dataset as all values occur only once.

Interpretation:

The mean age of the sample is 50 years, which provides a measure of the average age in the group.
The median age is 50 years, which indicates that half of the individuals in the sample are older than 50, and half are younger. Since both are the same it shows the data is very symmetric (uniform).
Since no age value repeats, there is no mode in this dataset, but since data appears to be symmetric we have found the measure of centrality using mean and median.

Question 2: Descriptive Statistics - Variability

(a) Define and explain the concept of "variability" in descriptive statistics. How is variability measured, and why is it important in data analysis?

Variability refers to the spread or dispersion of data points in a dataset. It quantifies how much individual data points deviate from the central tendency. Variability is important because it helps us understand the range and consistency of the data. If there was no variability there would be nothing else to say about the data. Common measures of variability include range, interquartile range, variance, and standard deviation.

(b) Consider two datasets: Dataset A with values [10, 20, 30, 40, 50, 60] and Dataset B with values [10, 10, 10, 50, 60, 60]. Calculate the range, interquartile range, variance, and standard deviation for both datasets. Compare the variability in these two datasets and discuss which dataset exhibits more variability and why.

Dataset A: [10, 20, 30, 40, 50, 60]

Range: Range is the difference between the maximum and minimum values.
Range = 60 (maximum) - 10 (minimum) = 50

Interquartile Range (IQR): IQR is the difference between the upper quartile and lower quartile values.
Median is half way point between 30 and 40
Median = (30 + 40) / 2 = 35
Lower half of data is 10, 20, 30
Lower quartile is middle of lower half: 20
Upper half of data is 40, 50, 60
Upper quartile is middle of upper half: 50
IQR: 50-20 =30

Variance: Variance measures the average squared difference between each data point and the mean.
Mean = (10 + 20 + 30 + 40 + 50 + 60) / 6 = 35
Variance = [(10-35)^2 + (20-35)^2 + (30-35)^2 + (40-35)^2 + (50-35)^2 + (60-35)^2] / 5
Variance = (625 + 225 + 25+ 25 + 225 + 625) / 5 = 350

Standard Deviation: Standard deviation is the square root of the variance.
Standard Deviation = √(Variance) = √(350) ≈ 18.71

Dataset B: [10, 10, 10, 50, 60, 60]

Range: Range = 60 (maximum) - 10 (minimum) = 50

Interquartile Range (IQR):.
Median is half way point between 10 and 50
Median = (10 + 50) / 2 = 30
Lower half of data is 10, 10, 10
Lower quartile is middle of lower half: 10
Upper half of data is 50, 60, 60
Upper quartile is middle of upper half: 60
IQR: 60-10 =50

Variance:
Mean = (10 + 10 + 10 + 50 + 60+ 60) / 6 = 33.33
Variance = [(10-33.33)^2 + (10-33.33)^2 + (10-33.33)^2 + (50-33.33)^2 + (60-33.33)^2 + (60-33.33)^2] / 5
Variance = 3333.333/ 5 =666.667

Standard Deviation: Standard Deviation = √(Variance) = √(666.667)≈ 25.82

Comparison:

Dataset B exhibits more variability than Dataset A. This is evident in both the larger interquartile range and the higher standard deviation for Dataset B, indicating that the values in Dataset B are more spread out from the mean compared to Dataset A. Dataset A has less variation and is more clustered around the mean and median. Can see here the range was the same for both so although gives a maximum spread between any two values, it does not tell us anything about the spread in between the minimum and maximum value.

Question 3: Odds Ratios and Relative Risks

A double-blind study that took a random sample of people who suffer from nose bleeds were split into two groups, a treatment group and a placebo group, to see if a new drug would improve their symptoms (Response). The results are given in the table below.

	Response	No Response	Total
Treatment	53	22	75
Placebo	44	31	75
Total	97	53	150

(a) Calculate the relative risk for responding to the treatment compared to placebo. Interpret this relative risk.

$$\widehat{RR} = \frac{\hat{\boldsymbol{\pi}}_1}{\hat{\boldsymbol{\pi}}_2} = \frac{\frac{53}{75}}{\frac{44}{75}} \cong \frac{0.7067}{0.5867} = \frac{53}{44} \cong 1.2045$$

Estimated that you are 1.2045 times more likely to respond if you have given the treatment as opposed to been given the placebo. More likely to respond if given the treatment than placebo.

(b) Calculate the odds ratio for responding to the treatment compared to placebo. Interpret this odds ratio.

$$\hat{\theta} = \frac{\frac{53}{22}}{\frac{44}{31}} = \frac{53(31)}{22(44)} \cong 1.697314$$

Estimated that the odds of responding when given the treatment is 1.697 times the odds of responding when given the placebo. More likely to respond if given the treatment than placebo.