Gitel Panwar STB03-T0036

**STATISTICAL ANALYSIS**

**1.What is the purpose of descriptive statistics?**

**Ans:** Descriptive statistics summarizes our data and gives its description. It can be used to describe an entire population or a sample.

**2.Can you explain the difference between mean, median, and mode?**

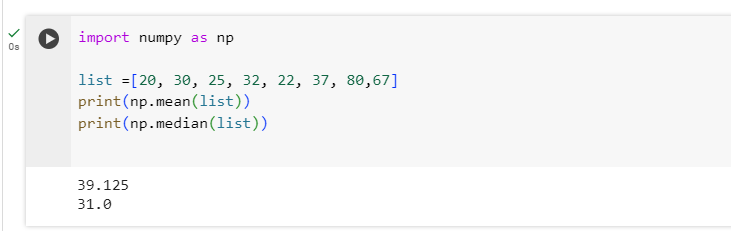
**Ans:**  All the three are measures of central tendency, but have the following differences:

**Mean:** average of all values. Calculated by (sigma) xi/n, where xi are the observations and n is the number of observations.

* Mean can be used to find the central value in case of numerical data.
* The problem with mean arises when there are outliers in our data, in that case, the mean is not the true representative of our data, since it becomes biased with the outlier:
* Example: weights of students are: 20, 30, 25, 32, 22, 37, 80,67. Now here, all values mostly lie between 20-30, except for the last two values, in this case the mean will be 39,125, which is shifted towards these outliers, which does not give us the correct description of our data.

**Median:** Median is the middle value obtained after sorting the data. In case of even number of terms, the median is the average of the two middle values.

* In case of outliers, median gives us a more accurate description of our data, since, the skewed values shall remain at the end and the median will not get affected.
* Example: in the above example, if we take median instead of mean, we get 31, which can be used to see if we want the common weight range.

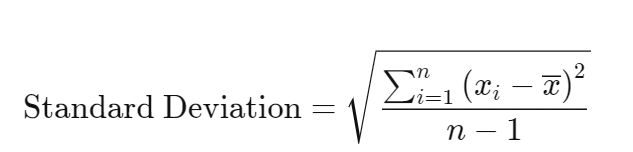


**Mode:** Mode is the measure of central tendency (usually useful in categorical data). It gives us the value of the data which has the maximum frequency/ occurrence in our dataset.

* Example: In a basket of fruits, the most occurring fruit can be found out using mode.

**3.How do you interpret the standard deviation of a dataset?**

**Ans: formula:**

****

Since we are subtracting every value from the mean, it is a statistic that measures the dispersion of a dataset relative to the mean. It shows how the data is centered around our mean, and how far the values are from it.A large standard deviation shows that the data is dispersed, whereas a small std deviation indicates that the data points are close to each other.

**4.Describe the concept of skewness in statistics.**

**Ans:** Skewness indicates the amount of symmetry in our data. In a normal distribution, when we plot the data, we get a bell shaped curve. This data is said to be balanced/ centered.

Skewness arises when the data is not symmetrically arranged around the center, a positive/negative skewness value indicates that the values on both sides of the mean are not equal.

On the bases of skewness, data can be seen as:

1. No skew: symmetric distribution, mean=median=mode
2. Left skewed/ negatively skewed: Tail towards starting, mean shifted towards left, mean<median<mode
3. Right skewed/positively skewed: Tail towards end, mean shifts towards right, mode<median<mean

**5.What is the main goal of inferential statistics?**

**Ans:** Inferential statistics is used to estimate population parameters from sample parameters. We apply operations on a small sample and from the results from the sample, we conclude things about the whole population

**6.Explain the difference between a population and a sample.**

**Ans:** Population is the entire group that we want to draw conclusions about whereas, sample is a subset of this population, which is a representative of the entire population. A sample is smaller and hence easier to manage while analysis, it is chosen from the population such that it has the characteristics of the population that we want to focus upon.

**7.What is a confidence interval, and how is it useful in inferential statistics?**

**Ans:** Confidence interval displays the probability that a parameter will fall between a pair of values around the mean.

* Confidence intervals measure the degree of uncertainty or certainty in a sampling method.
* For. example: If there are elections in a country, and a report has a 95% confidence interval of (0.52, 0.58) for the proportion of voters supporting Candidate A.
* This would mean "We are 95% confident that the true proportion of voters supporting Candidate A is between 52% and 58%."

**8.Define p-value**

**Ans:** A p-value is a statistical measurement used to validate a hypothesis against observed data. It measures the probability of obtaining the observed results, assuming that the null hypothesis is true.

* The lower the p-value, the greater the statistical significance of the observed difference.
* A p-value of 0.05 or lower is generally considered statistically significant.