Ep 3 - Descriptive Statistics 1

Nov 28, 2023

What does Descriptive Statistics encompass?

Descriptive statistics are a set of techniques used to *summarize, organize, and describe the main features of a dataset*. These statistics provide a clear and concise way to understand the essential characteristics of data. Descriptive statistics include the following measures and techniques:

1. Measures of Central Tendency
2. Measures of Dispersion(Variability)
3. Percentiles, Quartiles and Box-plots
4. Covariance and Correlation

Measures of Central Tendency

Mean(Average): The sum of all data points divided by the number of data points. It represents the centre of the data.

- Population Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Sample Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median: The middle value when data is ordered. It's *not affected by extreme values* and is useful for skewed data.

$$\text{Med} = \begin{cases} \frac{n+1}{2} & \text{if } n \text{ is odd} \\ \frac{n}{2} & \text{if } n \text{ is even} \end{cases}$$

Mode: The most frequently occurring value in the dataset. A dataset can have one mode (uni-modal) or multiple modes (multi-modal).

$$\text{Mode} = \arg\max_{x} f(x)$$

- $f(x)$ is a function that returns the frequency of x.
- $\arg\max$ returns value of x which has highest frequency.

Measures of Dispersion

Range: The difference between the maximum and minimum values in a dataset. It gives a sense of how spread out the data is.

$$\text{range} = x_{max} - x_{min}$$

Variance: A measure of how data points deviate from the mean. It provides information about the data's spread.

- Population Variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

- Sample Variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Standard Deviation: The square root of the variance. It indicates the average distance between data points and the mean.

- Population Standard Deviation

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

- Sample Standard Deviation

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

ⓘ Why is sample variance divided by $n-1$?

Bessel's correction is a statistical adjustment used when calculating the sample variance. It's applied to correct for the slight underestimation of the population variance that can occur when estimating from a sample. By dividing by 'n-1' instead of 'n' , it accounts for a loss of one degree of freedom when calculating the sample mean. This correction provides a more accurate estimate of the population variance, particularly with small sample sizes, ensuring unbiased results. It's a standard practice in statistics to improve the reliability of variance estimates.

Calculating Mean, Variance and Standard Deviation
A Mathematical Explanation on The Difference Between Sample and Population Variance
Difference Between Standard Deviation and Standard Error