Measures

R4 Cheng

December 26, 2024

For Continuous Data

1. Central Tendency
2. Variability or Dispersion
3. Skewness
4. Kurtosis

Central Tendency

Common central tendency measures: mean, median, mode (the most frequent value)

Mean

Sample Mean:

$$\bar{x} = \frac{\sum x_i}{n}$$

Population Mean:

$$\mu = \frac{\sum X_i}{N}$$

Median

Sample Median: $\tilde{x}$

Population Median: η (eta)

Dispersion or Variability

4 common measures of dispersion:

1. Range: $R = \max - \min$
2. Variance: population $\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$, sample $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$
3. Standard Deviation: population $\sigma = \sqrt{\sigma^2}$, sample $s = \sqrt{s^2}$

4. Coefficient of Variation (CV): population $CV = \frac{\sigma}{\mu} \times 100\%$, sample $CV = \frac{s}{\bar{x}} \times 100\%$ (no unit)

Remark. *Why use $n-1$? becuase it is proved to be more accurate.*

Disadvantages of Range: sensitive to outliers

Variance represents the distance from the mean

Variance and Standard Deviation are absolute measures of dispersion (about mean), while Coefficient of Variation is a relative measure of dispersion (about mean).

Skewness

Aka. shape of the distribution

3 types of skewness:

1. Symmetrical: mean = median = mode
2. Right Skewness or Positive Skewness: mean >> median
3. Left Skewness or Negative Skewness: mean << median

Skewness Coefficient (g_1): $g_1 = \frac{\frac{\sum(X_i - \bar{x})^3}{n-1}}{s^3}$

1. $g_1 = 0$: symmetrical
2. $g_1 > 0$: right skewness
3. $g_1 < 0$: left skewness

Kurtosis

$$g_2 = \frac{\frac{\sum(x_i - \bar{x})^4}{n-1}}{s^4} - 3$$

1. $g_2 = 0$: meso-kurtic
2. $g_2 > 0$: lepto-kurtic (more peaked)
3. $g_2 < 0$: platy-kurtic (less peaked)