Measures

R4 Cheng

December 29, 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)

Measures of Non-central Tendency

1. Quartiles: Q1, Q2, Q3
2. Percentiles: P1, P2, ..., P99: E.g. $P_{20} \Rightarrow 20\%$ data <= P_{20}
3. Interquartile Range (IQR): $Q_3 - Q_1$

How to find Quartiles

1. Arrange data in **ascending order**
2. Cal $Q_1 = 0.25 * (n+1)$; $Q_3 = 0.5 * (n+1)$;

3. If Q_1 or Q_3 is not an integer, take the average of the two values around it.

For Bivariate Data

Two metrics to measure the linear relationship between two variables (strength and direction):

1. Covariance (with unit)
   - Population Covariance: $\sigma_{xy} = \frac{\sum(X_i-\mu_x)(Y_i-\mu_y)}{N}$
   - Sample Covariance: $s_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1}$
2. Correlation Coefficient (without unit)
   - Population Correlation Coefficient: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, where $-1 \le \rho_{xy} \le 1$
   - Sample Correlation Coefficient: $r_{xy} = \frac{s_{xy}}{s_x s_y}$, where $-1 \le r_{xy} \le 1$

Remark. *$r = 0$ does not imply no relationship, it only implies no linear relationship.(Maybe strong curvilinear correlation)*

If two variables have a linear relationship, we tend to find it regression equation (a.k.a. least suqare line).

$$\hat{y} = a + bx$$

where $\hat{y}$ is the dependent variable, x is the independent variable, a is the intercept, b is the slope.

$\Rightarrow$

$$b = \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x}$$