

Sample Statistics

The sample $\{x_1, x_2, \dots, x_n\}$ is data obtained by taking measurements of some variable from a sample of size n from the total population X .

Definition. The sample mean is $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$

Definition. The sample median is the middle value in the ordered sample. If the sample size is even, the average of the middle two values is taken.

Definition. The concept of median can be generalized to the p -th percentile. Where $p \in (0, 1)$ and \tilde{x}_p will either be the $p(n+1)$ -th value in the ordered sample, or the weighted contribution of the nearest values, relative to how close they are to the integer part of $p(n+1)$

Note: For easier computation of quantiles, just use medians of the lower and upper halves of the ordered sample.

Definition. The sample variance is $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right]$. And the sample standard deviation (SD) is $s = \sqrt{s^2}$.

Note: We like standard deviation because its units match the units of the sample.

Definition. The standardized z -score for a data value is $z_i = \frac{x_i - \bar{x}}{s}$

Definition. If x and y are samples of size n , the sample correlation coefficient is $r = \frac{s_{xy}}{s_x s_y}$ where

$$s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i y_i - n\bar{x}\bar{y})$$

Note: We can also express r as the average of the products of the standardized z -scores:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

When performing linear regression using the *least squares method* we get the line

$$\frac{y - \bar{y}}{s_y} = r \left[\frac{x - \bar{x}}{s_x} \right] \quad \text{or} \quad y = r \frac{s_y}{s_x} x + (\bar{y} - b\bar{x})$$