

Chi-Squared χ^2 Tests

Definition 1.1

A random variable X is said to follow a χ^2 -distribution, with degree of freedom ν , iff its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

General Information

- Properties of the chi-squared distribution.
 - $E(X) = \nu$ and $\text{Var}(X) = 2\nu$.
 - The $\chi^2_{(\nu)}$ distribution tends to a normal distribution as $\nu \rightarrow \infty$.
 - Suppose $Z_i \sim N(0, 1)$ are independent. Then, $Z_1^2 + \cdots + Z_n^2 \sim \chi^2_{(n)}$.
 - If $X \sim \chi^2_{(\nu)}$ and $Y \sim \chi^2_{(v)}$, then $X + Y \sim \chi^2_{(\nu+v)}$.

Correlation and Linear Regression

Note

A good scatter diagram should follow the guidelines below.

- The relative position of each point on the scatter diagram should be clearly shown.
- The range of values for the set of data should be clearly shown by marking out the extreme x and y values on the corresponding axis.
- The axes should be labeled clearly with the variables.

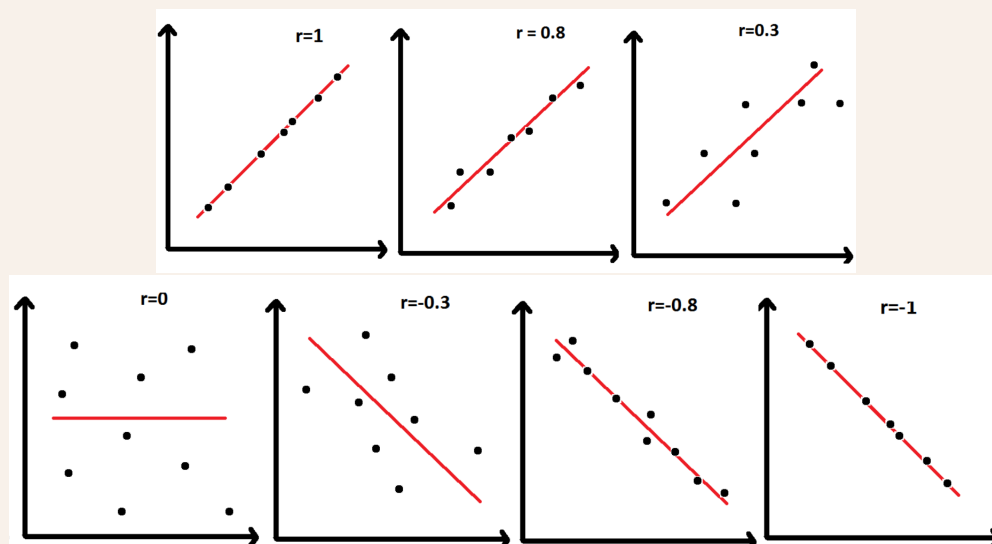
General Information

- The Product Moment Correlation Coefficient is a measure of the linear correlation between two variables. It is defined by

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]}}$$

which takes on a value from 0 to 1.

- When $r = 0$, there is no linear relationship. But, a nonlinear relationship may be present. Additionally, the regression lines are perpendicular.
- The closer the value of r is to 1 (or -1), the stronger the positive (or negative) linear correlation. Furthermore, the regression lines coincide.



- The regression line of y on x minimises the sum of squares deviation (error) in the y -direction. (i.e. we are assuming x is the independent variable whose values are known exactly.) It is given by

$$y = \bar{y} + b(x - \bar{x}), \quad \text{where} \quad b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}.$$

- The point (\bar{x}, \bar{y}) always lies on both the regression lines of y on x , and x on y .
- Say we are given the value of one variable, and asked to approximate the value of the other variable. Then, we should always use the line of the *dependent* variable on the *independent*.
- Estimations should not be taken for data outside the range of the sample provided, even if the value of r is close to 1.