Sample correlation coefficient $r$ for $(x_1, y_1), \ldots, (x_n, y_n)$

$$r = \frac{1}{n-1} \sum \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n \sum x_i^2 - (\sum x_i)^2)(n \sum y_i^2 - (\sum y_i)^2)}}$$

where $\bar{x}$ and $s_x$ are mean and standard deviation for $x_1, \ldots, x_n$ (similarly, $\bar{y}$ and $s_y$ are mean and standard deviation for $y_1, \ldots, y_n$)

**Theorem:** Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sample with correlation coefficient $r$ from a population with correlation coefficient $\rho$. Assume that $X$ and $Y$ are random variables with the bivariate normal distribution. Then the quantity

$$W = \frac{1}{2} \ln \frac{1+r}{1-r}$$

is approximately normal with mean $\mu_W = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ and variance $\sigma^2_W = \frac{1}{n-3}$.

**Remark:** $\rho = \frac{e^{2\mu_W - 1}}{e^{2\mu_W + 1}}$

**Significance test for the correlation $\rho$**
- can we estimate $\rho$ using $r$?
- to test the hypothesis $H_A : \rho \neq 0$ (or $H_A : \rho < 0$ or $H_A : \rho > 0$), we use the student’s $t$-distribution with $n-2$ degrees of freedom and test statistic

$$u = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$$

**Regression analysis**
- population simple linear regression model is of the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

- sample simple linear regression model is of the form

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

where

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$