CORRELATION

Proxy variables

Variables used in place of some unmeasurable qty of interest; highly correlated with this unmeasurable qty

PEARSON'S CORRELATION COEFFICIENT

$$\boxed{r = \frac{\sum_{i=1}^{n} Z_x Z_y}{n} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n\,\sigma_x \sigma_y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}}$$

$$\boxed{r_{\text{sample}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}}$$

- $r(A+BX, C+DY) = \begin{cases} -r(X,Y) & \text{if } \mathrm{sign}(B) \neq \mathrm{sign}(D) \\ r(X,Y) & \text{otherwise} \end{cases}$
- $r^2 = R^2$ score (coefficient of determination)
- $r \in [-1, 1]$
- r may be zero even when there is a strong non-linear relationship

HYPOTHESIS TEST

Let ρ be the population corr. coefficient

$H_0 : \rho = 0$

$H_A : \rho \neq 0$

Sampling distribution of r:

approx. t-distribution

$(n-2)$ dof } we are estimating two means

$\mu = \rho$

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

Test statistic:

$$\boxed{t = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}}} \xrightarrow[H_0]{\rho=0} \boxed{t = r\sqrt{\frac{n-2}{1-r^2}}}$$

SPEARMAN'S RANK CORRELATION

ordinal.

$$\boxed{r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}}$$

$D_i = x_i - y_i$ difference in rank for case i under x, y

Same sampling distribution:

$\mu = \rho_s \qquad \sigma = \sqrt{\frac{1-\rho_s^2}{n-2}} \qquad df = n-2$

POINT BI-SERIAL CORRELATION

continuous ⟷ binary

X: continuous $\qquad Y$: dichotomous (binary)

1. group X based on Y → X_0, X_1
2. Calculate means → $\bar{X}_0$, $\bar{X}_1$
3. Let $n_0 = |X_0|$, $n_1 = |X_1|$, S_x → std deviation of whole X

$$\boxed{r_b = \frac{\bar{X}_0 + \bar{X}_1}{S_x}\sqrt{\frac{n_0 n_1}{n(n-1)}}} \qquad S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

Same results as if you used Pearson's

Sampling distribution same

PHI COEFFICIENT

binary ⟷ binary

Create contingency table

	Y = 0	Y = 1	Total
X = 0	N_{00}	N_{01}	N_{0Y}
X = 1	N_{10}	N_{11}	N_{1Y}
Total	N_{X0}	N_{X1}	

$$\boxed{\varphi = \frac{\text{prod(Agree)} - \text{prod(disagree)}}{\sqrt{\text{prod(row/col totals)}}}}$$

$$\boxed{\varphi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{0Y}N_{1Y}N_{X0}N_{X1}}}} \qquad \frac{11{:}00 - 10{:}01}{\text{root(totals)}}$$