Correlation Coefficient

Correlation coefficient is a measure of association between two variables, and it ranges between –1 and 1. If the two variables are in perfect linear relationship, the correlation coefficient will be either 1 or –1. The sign depends on whether the variables are positively or negatively related. The correlation coefficient is 0 if there is no linear relationship between the variables. Two different types of correlation coefficients are in use. One is called the Pearson product-moment correlation coefficient, and the other is called the Spearman rank correlation coefficient, which is based on the rank relationship between variables. The Pearson product-moment correlation coefficient is more widely used in measuring the association between two variables. Given paired measurements (X_1,Y_1), (X_2,Y_2), . . . ,(X_n,Y_n), the Pearson product-moment correlation coefficient is a measure of association given by

$$r_P = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

where $\bar{X}$ and $\bar{Y}$ are the sample mean of $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, respectively.

Case Study and Data

The following 25 paired measurements can be found at http://lib.stat.cmu.edu/DASL/Datafiles/SmokingandCancer.html:

77	84
137	116
117	123
94	128
116	155
102	101
111	118
93	113
88	104
102	88
91	104
104	129
107	86
112	96
113	144
110	139
125	113
133	146
115	128
105	115
87	79
91	85
100	120
76	60
66	51

For a total of 25 occupational groups, the first variable is the smoking index (average 100), and the second variable is the lung cancer mortality index (average 100). Let us denote these paired indices as (X_i,Y_i). The Pearson product-moment correlation coefficient is computed to be $r_p = 0.69$ Figure 1 shows the scatter plot of the smoking index versus the lung

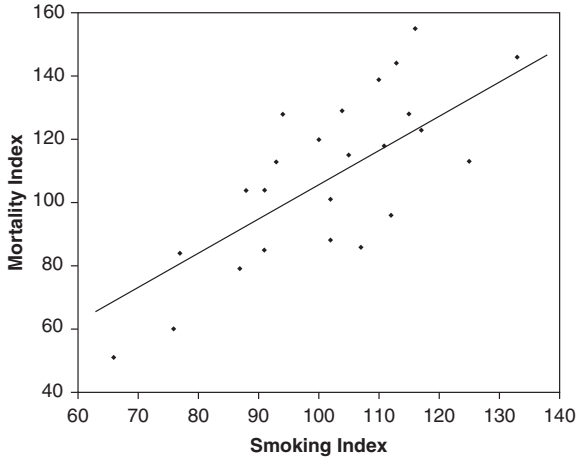

Figure 1 Scatter Plot of Smoking Index Versus Lung Cancer Mortality Index

Source: Based on data from Moore & McCabe, 1989.

Note: The straight line is the linear regression of mortality index on smoking index.

cancer mortality index. The straight line is the linear regression line given by $Y = \beta_0 + \beta_1 \cdot X$.

The parameters of the regression line are estimated using the least squares method, which is implemented in most statistical packages such as SAS and SPSS. The equation for the regression line is given by $Y = -2.89 + 1.09 \cdot X$. If (X_i, Y_i), are distributed as bivariate normal, a linear relationship exists between the regression slope and the Pearson product-moment correlation coefficient given by

$$\beta_1 \simeq \frac{\sigma_Y}{\sigma_X} r_P,$$

where σ_X and σ_Y are the sample standard deviations of the smoking index and the lung cancer mortality index, respectively ($\sigma_X = 17.2$ and $\sigma_Y = 26.11$). With the computed correlation coefficient value, we obtain

$$\beta_1 \simeq \frac{26.11}{17.20} \cdot 0.69 = 1.05,$$

which is close to the least squares estimation of 1.09.

Statistical Inference on Population Correlation

The Pearson product-moment correlation coefficient is the underlying population correlation ρ. In the smoking and lung cancer example above, we are interested in testing whether the correlation coefficient indicates the statistical significance of relationship between smoking and the lung cancer mortality rate. So we test. $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$..

Assuming the normality of the measurements, the test statistic

$$T = \frac{r_P \sqrt{n-2}}{\sqrt{1 - r_P^2}}$$

follows the t distribution with n–2 degrees of freedom. The case study gives

$$T = \frac{0.69\sqrt{25-2}}{\sqrt{1-0.69^2}} = 4.54.$$

This t value is compared with the 95% quantile point of the t distribution with n–2 degrees of freedom, which is 1.71. Since the t value is larger than the quantile point, we reject the null hypothesis and conclude that there is correlation between the smoking index and the lung cancer mortality index at significance level $\alpha = 0.1$. Although r_p itself can be used as a test statistic to test more general hypotheses about ρ, the exact distribution of ρ is difficult to obtain. One widely used technique is to use the Fisher transform, which transforms the correlation into

$$F(r_P) = \frac{1}{2} \ln \left(\frac{1 + r_P}{1 - r_P} \right).$$

Then for moderately large samples, the Fisher transform is normally distributed with mean $\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$ and variance $\frac{1}{n-3}$. Then the test statistic is $Z = \sqrt{n-3}\,(F(r_P) - F(\rho))$, which is a standard normal distribution. For the case study example, under the null hypothesis, we have

$$Z = \sqrt{25-3} \left(\frac{1}{2} \ln \left(\frac{1+0.69}{1-0.69} \right) - \frac{1}{2} \ln \left(\frac{1+0}{1-0} \right) \right) = 3.98.$$

The Z value is compared with the 95% quantile point of the standard normal, which is 1.64. Since the Z value is larger than the quantile point, we reject the null hypothesis and conclude that there is correlation between the smoking index and the lung cancer mortality index.

—Moo K. Chung

See also Coefficients of Correlation, Alienation, and Determination; Multiple Correlation Coefficient; Part and Partial Correlation

Further Reading

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, *10*, 507–521.

Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics*. New York: W. H. Freeman. (Original source: Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales, 1970–1972, Her Majesty's Stationery Office, London, 1978.)

Rummel, R. J. (n.d.). *Understanding correlation*. Retrieved from http://www.mega.nu:8080/ampp/rummel/uc.htm