EXHIBIT D

TO
PLAINTIFFS' CLAIM CONSTRUCTION BRIEF PURSUANT TO P.R. 4-5(a)

Pearson product-moment correlation coefficient

From Wikipedia, the free encyclopedia

In statistics, the **Pearson product-moment correlation coefficient** (sometimes referred to as the **PMCC**, and typically denoted by r) is a measure of the correlation (linear dependence) between two variables X and Y, giving a value between +1 and −1 inclusive. It is widely used in the sciences as a measure of the strength of linear dependence between two variables. It was first introduced by Francis Galton in the 1880s, and is named after Karl Pearson.[1][2] The correlation coefficient is sometimes called "Pearson's r."

Contents

Definition

The statistic is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom.[3] Based on a sample of paired data (X_i, Y_i), the sample Pearson correlation coefficient can be calculated as

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

where

$$\frac{X_i - \bar{X}}{s_X}, \bar{X}, \text{ and } s_X$$

are the standard score, sample mean, and sample standard deviation (calculated using $n - 1$ in the denominator).[3]

The result obtained is equivalent to dividing the sample covariance between the two variables by the

product of their sample standard deviations:

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

When calculated for an entire population, the Pearson product-moment correlation is typically designated by the Greek letter ρ (rho), while the sample-based estimate is designated by the Latin letter *r*. Other conventions such as the use of ρ to denote the population correlation coefficient and $\hat{\rho}$ to denote the sample correlation coefficient are also in use.

For a finite population, the population correlation coefficient is

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \mu_X}{\sigma_X}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)$$

where

$$\frac{X_i - \mu_X}{\sigma_X}, \mu_X, \text{ and } \sigma_X$$

are the standard score, population mean, and population standard deviation (calculated using *n* in the denominator).

For a population that is not necessarily finite, the population correlation coefficient can be defined in terms of population expected values and population variances (see correlation).

Interpretation

The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between *X* and *Y* perfectly, with all data points lying on a line for which *Y* increases as *X* increases. A value of −1 implies that all data points lie on a line for which *Y* decreases as *X* increases. A value of 0 implies that there is no linear relationship between the variables.[3]

More generally, note that

$$(X_i - \bar{X})(Y_i - \bar{Y}) > 0$$

if and only if X_i and Y_i lie on the same side of their respective means. Thus the correlation coefficient is positive if X_i and Y_i tend to be simultaneously greater than, or simultaneously less than, their respective means. The correlation coefficient is negative if X_i and Y_i tend to lie on opposite sides of their respective means.

Relationship to linear regression and the *r*-squared

Correlation and regression are related in a number of ways. One important way that correlation arises in regression analysis is through the *coefficient of determination*, which is used to explain how well a

regression model fits a particular data set.

The linear equation that best describes the relationship between X and Y can be found by linear regression. This equation can be used to predict the value of Y when only X is observed. We denote this predicted value by Y'.

Any value of Y can be written as the sum of Y' and the difference between Y and Y' (the residual):

$$Y = Y' + (Y - Y').$$

If Y' is estimated using least squares, $Y - Y'$ and Y' are orthogonal. This implies that the sample covariance between $Y - Y'$ and Y' is zero. Since covariances are bilinear, it follows that

$$\operatorname{cov}(Y, Y') = \operatorname{cov}(Y', Y') + \operatorname{cov}(Y - Y', Y') = \operatorname{var}(Y'),$$

where the covariances and variance are sample statistics calculated across the cases in the regression analysis. It follows that we can write the correlation between Y and Y' as

$$\operatorname{cor}(Y, Y') = \frac{\operatorname{var}(Y')}{\operatorname{SD}(Y)\operatorname{SD}(Y')} = \frac{\operatorname{SD}(Y')}{\operatorname{SD}(Y)}.$$

The square of the correlation between Y and Y' is called the "r squared":

$$r^2 = \frac{\operatorname{var}(Y')}{\operatorname{var}(Y)}.$$

The r squared is the proportion of the variability in Y that can be predicted, or explained, from X. It is used as a measure of the association between X and Y. For example, if r^2 is 0.90, then 90% of the variance of Y can be "accounted for" by changes in X through the linear relationship between X and Y.[3]

Inference

Confidence intervals and hypothesis tests relating to ρ are usually carried out using the Fisher transformation:

$$F(r) = \frac{1}{2} \log \frac{1 + r}{1 - r} = \operatorname{arctanh}(r).$$

If $F(r)$ is the Fisher transformation of r, and n is the sample size, then

$$z = \sqrt{n - 3} F(r)$$

is a z-score for r which approximately follows a standard normal distribution under the null hypothesis of no linear association ($\rho = 0$), given the assumption that the sample pairs are independent and identically distributed and follow a bivariate normal distribution. Thus an approximate p-value for the hypothesis $\rho=0$ can be obtained from a normal probability table. For example, if $z = 2.2$ is observed and a two-sided p-value is desired, the p-value is $2\Phi(-2.2) = 0.028$, where Φ is the standard normal cumulative distribution function.

To obtain a confidence interval for ρ, we first compute a confidence interval for z, then invert the Fisher transformation to the correlation scale. For example, suppose we observe $r = 0.3$ with a sample size of n=50, and we wish to obtain a 95% confidence interval for ρ. The z-score is $z = 2.12$, so the confidence interval on the z-scale is 2.12 ± 1.96, or (0.16, 4.08). Converting back to the correlation scale yields (0.02, 0.53).

Sensitivity to the data distribution

The correlation coefficient is defined in terms of moments and does not require the data to be either marginally or jointly normally distributed [1]. Some distributions such as the Cauchy distribution have undefined variance and hence ρ is not defined if X or Y follows such a distribution. In some practical applications, such as those involving data suspected to follow a heavy-tailed distribution, this is an important consideration. However, the existence of the correlation coefficient is usually not a concern; for instance, if the range of the distribution is bounded, ρ is always defined.

Like many commonly-used statistics, r is not robust[4], so its value can be misleading if outliers are present[5][6]. Specifically, the PMCC is neither distributionally robust, nor outlier resistant[4] (see Robust statistics#Definition). Inspection of the scatterplot between X and Y will typically reveal such a situation, and in such cases it may be advisable to use a robust measure of association.

If X and Y are assumed to follow a bivariate normal distribution, the sampling distribution of r can be explicitly obtained, which is not possible in general. However, an extended version of the result noted above for the Fisher transformation allows the asymptotic (large sample size) distribution of the sample correlation coefficient to be obtained under weaker assumptions on the distribution of X and Y.

Computing correlation accurately in a single pass

The following algorithm (in pseudocode) will calculate **Pearson** correlation with good numerical stability[7] in a single pass.

```
sum_sq_x = 0
sum_sq_y = 0
sum_coproduct = 0
mean_x = x[1]
mean_y = y[1]
for i in 2 to N:
    sweep = (i - 1.0) / i
    delta_x = x[i] - mean_x
    delta_y = y[i] - mean_y
    sum_sq_x += delta_x * delta_x * sweep
    sum_sq_y += delta_y * delta_y * sweep
    sum_coproduct += delta_x * delta_y * sweep
    mean_x += delta_x / i
    mean_y += delta_y / i
pop_sd_x = sqrt( sum_sq_x/N )
pop_sd_y = sqrt( sum_sq_y/N )
cov_x_y = sum_coproduct/N
correlation = cov_x_y / (pop_sd_x * pop_sd_y)
```

Calculating a weighted correlation

Suppose observations to be correlated have differing degrees of importance that can be expressed with a weight vector *w*. To calculate the correlation between vectors *x* and *y* with the weight vector *w* (all of length *n*),[8][9]

- Weighted mean:

$$\mathrm{m}(x;w) = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

- Weighted covariance

$$\mathrm{cov}(x,y;w) = \frac{\sum_i w_i (x_i - \mathrm{m}(x;w))(y_i - \mathrm{m}(y;w))}{\sum_i w_i}$$

- Weighted correlation

$$\mathrm{corr}(x,y;w) = \frac{\mathrm{cov}(x,y;w)}{\sqrt{\mathrm{cov}(x,x;w)\mathrm{cov}(y,y;w)}}$$

Removing correlation

It is always possible to remove the correlation between zero-mean random variables with a linear transformation, even if the relationship between the variables is nonlinear. Suppose a vector of *n* random variables is sampled *m* times. Let *X* be a matrix where $X_{i,j}$ is the *j*th variable of sample *i*. Let $Z_{m,m}$ be an *m* by *m* square matrix with every element 1. Then *D* is the data transformed so every random variable has zero mean, and *T* is the data transformed so all variables have zero mean and zero correlation with all other variables - the moment matrix of *T* will be the identity matrix. This has to be further divided by the standard deviation to get unit variance. The transformed variables will be uncorrelated, even though they may not be independent.

$$D = X - \frac{1}{m} Z_{m,m} X$$

$$T = D(D^T D)^{-\frac{1}{2}}$$

where an exponent of -1/2 represents the matrix square root of the inverse of a matrix. The covariance matrix of *T* will be the identity matrix. If a new data sample *x* is a row vector of *n* elements, then the same transform can be applied to *x* to get the transformed vectors *d* and *t*:

$$d = x - \frac{1}{m} Z_{1,m} X$$

$$t = d(D^T D)^{-\frac{1}{2}}$$

See also

- Linear correlation (wikiversity)
- Spearman's rank correlation coefficient
- Association (statistics)
- Tests of Association Calculator (http://www.meta-numerics.net/Samples/BivariateSampleCalculator.aspx)

References

1. ^ [a] [b] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient (http://www.jstor.org/stable/2685263) . The American Statistician, 42(1):59–66, Feb 1988.
2. ^ Stigler, Stephen M. (1989). "Francis Galton's Account of the Invention of Correlation". *Statistical Science* **4** (2). http://www.jstor.org/stable/2245329.
3. ^ [a] [b] [c] [d] Moore, David (August 2006). "4". *Basic Practice of Statistics* (4 ed.). WH Freeman Company. pp. 90–114. ISBN 0-7167-7463-1.
4. ^ [a] [b] Wilcox, Rand R. (2005). *Introduction to robust estimation and hypothesis testing*. Academic Press.
5. ^ Devlin, Susan J; Gnanadesikan, R; Kettenring J.R. (1975). "Robust Estimation and Outlier Detection with Correlation Coefficients". *Biometrika* **62** (3): 531–545. http://www.jstor.org/stable/2335508.
6. ^ Huber, Peter. J. (2004). *Robust Statistics*. Wiley.
7. ^ Ronald A. Thisted (1988). *Elements of Statistical Computing: Numerical Computation*, pp. 84-91
8. ^ http://sci.tech-archive.net/Archive/sci.stat.math/2006-02/msg00171.html
9. ^ A MATLAB Toolbox for computing Weighted Correlation Coefficients (http://www.mathworks.com/matlabcentral/fileexchange/20846)

Retrieved from "http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient"
Categories: Covariance and correlation | Parametric statistics
Hidden categories: Statistics articles linked to the portal | Statistics articles with navigational template