JOURNAL OF RESEARCH of the National Bureau of Standards — B. Mathematical Sciences
Vol. 72B, No. 1, January–March 1968

The Distribution of the Sample Correlation Coefficient With One Variable Fixed

David Hogben

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(November 30, 1967)

For the usual straight-line model, in which the independent variable takes on a fixed, known set of values, it is shown that the sample correlation coefficient is distributed as Q with $(n-2)$ degrees of freedom and noncentrality $\theta=(\beta/\sigma)\sqrt{\Sigma(x_i-\bar{x})^2}$. The Q variate has been defined and studied elsewhere by Hogben et al. It is noted that the square of the correlation coefficient is distributed as a noncentral beta variable.

Key Words: Analysis of variance, calibration, correlation coefficient, degrees of freedom, distribution, fixed variable, noncentral beta variable, noncentrality, Q variate.

1. Introduction

Consider the straight-line model

$$Y_i=\alpha+\beta x_i+\epsilon_i, \qquad i=1, 2, \ldots, n \tag{1}$$

where

(i) the ϵ_i are assumed to behave as normally and independently distributed random variables with mean zero and common variance σ^2,

(ii) α and β are unknown parameters, and

(iii) lowercase italic letters denote fixed, known constants and uppercase italic letters denote random variables, i.e., x is fixed and Y is random. This and other straight-line models are discussed in detail by Acton [1959].

The sample correlation coefficient r_{xY} is defined by

$$r_{xY}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}, \tag{2}$$

where

$$\bar{x}=\sum_{i-1}^{n} x_i/n \text{ and } \bar{Y}=\sum_{i-1}^{n} Y_i/n.$$

It is of some interest in the calibration problem where a fitted straight-line is used in reverse for estimating an unknown x_0 corresponding to an observed Y_0. The distribution of r_{XY} where X and Y follow the bivariate normal is well known; see for example Kendall and Stuart [1961, pp. 383–390]. In the present paper the distribution of r_{xY} as defined by (2) is derived for x fixed. This distribution is well known for the special case with all Y_i identically distributed (i.e., $\beta=0$), in which case it is the same as the distribution of r_{XY} for X, Y independent and normal. See, e.g., Hotelling [1953, p. 196]. J. N. K. Rao and an unidentified person have pointed out that the distribution of r_{xY}^2 can be obtained as a special case of the conditional distribution of the multiple correlation coefficient for the multi-variate normal; see, e.g., C. R. Rao [1965, p. 509].

2. Derivation

In an analysis of variance for the model (1) the (corrected) total sum of squares with $(n-1)$ degrees of freedom may be partitioned into two independent components; the first being the sum of squares due to the slope with 1 degree of freedom and the second being the residual sum of squares with $(n-2)$ degrees of freedom. This partition can be expressed by

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\frac{\left[\sum(x_i-\overline{x})(Y_i-\overline{Y})\right]^2}{\sum(x_i-\overline{x})^2}+\sum(Y_i-\hat{Y}_i)^2, \tag{3}$$

where

$$\hat{Y}_i=\overline{Y}+\hat{\beta}(x_i-\overline{x})$$

and

$$\hat{\beta}=\sum_{i-1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})\Big/\sum_{i-1}^{n}(x_i-\overline{x})^2$$

is the usual least squares estimator for β. Let the random variables W and X^2 be defined by

$$W=\frac{\sum(x_i-\overline{x})(Y_i-\overline{Y})}{\sigma\sqrt{\sum(x_i-\overline{x})^2}}, \tag{4}$$

and

$$X^2=\sum(Y_i-\hat{Y}_i)^2/\sigma^2. \tag{5}$$

Using (4) and (5) and dividing both sides of eq (3) by σ^2 we have

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2/\sigma^2=W^2+X^2. \tag{6}$$

If both the numerator and denominator of r_{xY} are divided by σ^2 and the first factor of the denominator is combined with the numerator, the correlation coefficient may be written as

$$r_{xY}=\frac{W}{\sqrt{W^2+X^2}}. \tag{7}$$

Since the Y_i are normally distributed, it is easily shown that W is normally distributed with mean $\theta=(\beta/\sigma)\sqrt{\Sigma(x_i-\overline{x})^2}$ and variance 1. Further, it is well known from the theory of the general linear hypothesis that under model (1) W and X^2 are independently distributed and X^2 is distributed as chi-squared with $(n-2)$ degrees of freedom. Therefore, r_{xY} is equal to the random variable Q defined and studied in Hogben et al., [1964a] and [1964b]. Hence, the following theorem is proved.

THEOREM: The correlation coefficient r_{xY}, defined by (2) under model (1), is distributed as Q with $(n-2)$ degrees of freedom and noncentrality $\theta=(\beta/\sigma)\sqrt{\Sigma(x_i-\overline{x})^2}$.

Various properties of Q are given in the previous two references, including analytic expressions and recurrence relations for the moments about zero, numerical values for the first four central moments and an approximation to the distribution of Q by that of a linearly transformed beta variable. It follows from (7) that r_{xY}^2 is distributed as noncentral beta; see for example Seber [1963], where in his notation $n_1=1$, $n_2=n-2$ and $\lambda=\theta^2/2$. Furthermore, $t=\sqrt{(n-2)r^2/(1-r^2)}$ is distributed as noncentral t with noncentrality θ and $(n-2)$ degrees of freedom. The distribution of r_{xY} also follows from the interesting and easily derived relation

$$r_{xY}=\frac{\hat{\beta}}{\sqrt{\hat{\beta}^2+(n-2)s_{\hat{\beta}}^2}}.$$

Thanks go to Joan Rosenblatt for pointing out the looseness of an earlier proof of the theorem and to her and Edwin L. Crow for constructive suggestions.

3. References

Acton, F. (1959), Analysis of Straight-Line Data (John Wiley & Sons, Inc., New York, N.Y.).

Hogben, D., Pinkham, R. S. and Wilk, M. B. (1964a), The moments of a variate related to the non-central t. Ann. Math. Statist., **35**, 298–314.

Hogben, D., Pinkham, R. S. and Wilk, M. B. (1964b), An approximation to the distribution of Q (a variate related to the non-central t). Ann. Math. Statist. **35**, 315–318.

Hotelling, H. (1953), New light on the correlation coefficient and its transforms. J. Roy. Statist. Soc. Ser. B, **15**, 193–225.

Kendall, M. G. and Stuart, A. (1961), The Advanced Theory of Statistics, **I.** (Hafner Publishing Co., New York).

Rao, C. R. (1965), Linear Statistical Inference and Its Application (John Wiley & Sons, Inc., New York, N.Y.).

Seber, G. A. F. (1963), The non-central chi-squared and beta distributions. Biometrika **50**, 542–544.

(Paper 72B1–257)