An Informational Measure of Correlation

E. H. Linfoot

The Observatories, University of Cambridge, Cambridge, England

In a recently published paper "Una teoria de la certidumbre" M. Castaños (1955) defines the certitude of a discrete probability distribution as the quantity

$$\log n + \sum_{i=1}^{n} p_i \log p_i,$$

where \(n\) is the number of discrete, mutually exclusive possibilities and \(p_1, p_2, \ldots, p_n\) their respective probabilities. Writing this expression in the form

$$-\sum_{i=1}^{n} p_i \log \frac{1}{p_i} + n \frac{1}{n} \log n,$$

we see that it represents the amount by which the entropy of the probability distribution \(p_1, p_2, \ldots, p_n\) of \(n\) discrete cases falls below its greatest possible value \(\log n\), which is assumed when every \(p_i\) has the same value \(1/n\). It is therefore the amount of information conveyed, to an individual who previously supposed that the \(n\) possible discrete values \(x_1, x_2, \ldots, x_n\) of a discrete variable \(x\) were all equally likely, by the statement that their respective probabilities are \(p_1, p_2, \ldots, p_n\) (Shannon, 1948, pp. 379, 623).

In a later paper, Castaños Camargo and Medina e Isabel (1956) consider two sets of discrete values, with probabilities \(p_1, p_2, \ldots, p_n\) and \(q_1, q_2, \ldots, q_m\), respectively, and with joint probabilities

$$p_{ij}(i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m).$$

Here

$$p_i = \sum_j p_{ij}, \quad q_j = \sum_i p_{ij}$$

and it may be shown (Goldman, 1953) that

$$\sum_{ij} p_{ij} q_{ij} \log (p_{ij} q_{ij}) \leq \sum_{ij} p_{ij} \log p_{ij},$$

85
with equality only if \( p_{ij} = p_i q_j \) for all \( i \) and \( q \). They then define the logarithmic index of correlation

\[
r_0 = \sum_{ij} \left( p_{ij} \log p_{ij} - p_i q_j \log p_i q_j \right);
\]

by (4), \( r_0 \geq 0 \).

It will be seen that \( r_0 \) also has a simple informational interpretation. It has been discussed from this point of view by McGill (1954). To an individual who previously supposed all the possible discrete values \( (x_i, y_j) \) of a pair of variables \( (x, y) \) to be equally likely, the statement that the probability distribution of \( (x_i, y_j) \) is \( p_{ij} (i = 1, 2, \ldots, n; j = 1, 2, \ldots, m) \) conveys an amount

\[
\sum_{ij} p_{ij} \log p_{ij} + \log mn
\]

of information. This is greater than the amount of the information which he received on being told only the separate probability distributions \( p_1, \ldots, p_n \) and \( q_1 \ldots, q_m \) of \( x \) and of \( y \); and the former amount exceeds the latter by

\[
\sum_{ij} p_{ij} \log p_{ij} + \log mn - \left( \sum_i p_i \log p_i + \log n \right) - \left( \sum_j q_j \log q_j + \log m \right)
\]

\[
= \sum_{ij} p_{ij} \log p_{ij} - \sum_{ij} p_i q_j \log p_i - \sum_{ij} p_i q_j \log q_j
\]

\[
= r_0.
\]

It is easy to show, by applying a well known property (Shannon, 1948, sect. 6) of information, that the value of the information gain \( r_0 \) is unchanged if the prior opinion of equiprobable discrete values \( (x_i, y_j) \) is replaced by the prior opinion that \( x \) and \( y \) are statistically independent \( (p_{ij} = p_i q_j) \). Thus \( r_0 \) can be interpreted as an information gain which provides a measure of the correlation between \( x \) and \( y \).

For continuous variables \( x \) and \( y \) with joint probability density distribution \( p(x, y) \) the corresponding quantity \( r_0 \) is given by the equation

\[
r_0 = \iint \{ p(x, y) \log p(x, y) - p(x)q(y) \log [p(x)q(y)] \} \, dx \, dy,
\]

where \( p(x) \) and \( q(y) \) are the probability density distributions of \( x \) and \( y \) taken separately. This is the amount of information conveyed, to any individual who previously supposed \( x \) and \( y \) to be independent, by the statement that their joint probability density distribution is \( p(x, y) \).
It is independent of the probability density distributions \( p_0(x), q_0(y) \) which express his prior opinions about the values of \( x \) and \( y \).

Although \( r_0 \) itself provides a logically very satisfactory measure of correlation, applicable whatever the form of \( p(x, y) \), it is natural to ask whether something more closely resembling the classical coefficient of correlation can be derived from informational considerations. In the second paper referred to above (Castaños Camargo and Medina e Isabel, 1956), the two authors consider the quantity

\[
-2r_0 \left| \sum_{ij} p_i q_j \log (p_i q_j) \right|^{-1},
\]

which they call the "logarithmic coefficient of correlation." It appears on examination that this coefficient cannot be interpreted as an informational measure of correlation.

However, it is a simple matter to obtain the desired result in the following way. Consider the probability density distribution

\[
p(x, y) = \frac{1}{2\pi} \sqrt{ab - h^2} e^{-\frac{(ax^2 + 2hxy + by^2)}{2}},
\]

where \( a > 0, ab - h^2 > 0 \). As is well known (Whittaker and Robinson, 1944), the classical correlation coefficient \( r \) is given in this case by the equation

\[
r = -h / \sqrt{ab}.
\]

To calculate the informational measure \( r_0 \), we first note that, in the notation already used above,

\[
p(x) = \int_{-\infty}^{\infty} p(x, y) \, dy = \sqrt{\frac{\alpha}{\pi}} e^{-ax^2}, \quad q(y) = \int_{-\infty}^{\infty} p(x, y) \, dx = \sqrt{\frac{\beta}{\pi}} e^{-by^2},
\]

where

\[
\alpha = (ab - h^2) / 2b, \quad \beta = (ab - h^2) / 2a.
\]

Equation (6) then gives, after a short calculation (Shannon, 1948, p. 54),

\[
r_0 = -\log \frac{2\pi e}{\sqrt{ab - h^2}} + \log \frac{\pi e}{\alpha} + \log \frac{\pi e}{\beta}
\]

\[
= \frac{1}{2} \log \frac{ab}{ab - h^2};
\]
and from (8) and (10) we see that, for the distribution (7)

$$r = \sqrt{1 - e^{-2r_0}}$$

(11)

It is easy to verify that the same result follows when $p(x, y)$ is given the more general form

$$p(x, y) = \frac{1}{2\pi} \sqrt{ab - h^2} \exp \left\{-\frac{1}{2}[a(x - x_0)^2 + 2h(x - x_0)(y - y_0) + b(y - y_0)^2]\right\}.$$  

(12)

We can now define the informational coefficient of correlation $r_1$ by the equation

$$r_1 = \sqrt{1 - e^{-2r_0}},$$  

(13)

where $r_0$ is given by (6). This coefficient reduces to the classical correlation coefficient in the case (12); it lies between 0 and 1 whatever the distribution $p(x, y)$. It is zero whenever $x$ and $y$ are statistically independent, since then $r_0 = 0$, and it is 1 whenever $x$ and $y$ are fully correlated, in the sense that each determines the value of the other uniquely.

An important advantage of the informational measures of correlation $r_0$ and $r_1$ in physical applications is that they are independent of the particular manner in which the measure numbers $x$ and $y$ are assigned to the two physical quantities under examination; in mathematical terms $r_0$ and $r_1$ are invariant under a transformation $x' = f(x), \ y' = g(y)$ of the variables $x$ and $y$ into new variables $x'$ and $y'$ respectively. The invariance of $r_0$ was pointed out by Jeffreys (1946) many years ago. In fact, since

$$\int \int p(x, y) \log \frac{p(x, y)}{p(x)q(y)} \, dx \, dy$$

$$= \int \int p(x)q(y) \log \frac{p(x)}{p(x)q(y)} \, dx \, dy,$$

Eq. (6) can be written in the equivalent form

$$r_0 = \int \int p(x, y) \log \left\{ \frac{p(x, y)}{p(x)q(y)} \right\} \, dx \, dy.$$  

(14)

Here $\log \{p(x, y)/p(x)q(y)\}$ is invariant under the above transformation, and hence its mathematical expectation $r_0$ is invariant; the invariance of $r_1$ follows immediately by (13).
Because of its interpretation in terms of quantity of information, \( r_0 \) seems to provide a more natural measure of correlation than \( r_1 \), but \( r_1 \) has the advantage that it is an informational measure of correlation which can be regarded as a generalization of an already familiar concept, viz. the ordinary correlation coefficient of a normal distribution.

**SUMMARY**

Informational considerations lead to a natural generalization of the classical correlation coefficient of a normal distribution. The generalized coefficient, here called the *informational coefficient of correlation*, is a function of the joint probability density distribution \( p(x, y) \) of the two variables \( x \) and \( y \), is invariant under a change of parameterization \( x' = f(x), y' = g(y) \), and reduces to the classical correlation coefficient when \( p(x, y) \) is normal.

**RECEIVED:** April 8, 1957

**REFERENCES**


