Explorations in statistics: correlation

Douglas Curran-Everett
Division of Biostatistics and Bioinformatics, National Jewish Health, and Department of Biostatistics and Informatics and Department of Physiology and Biophysics, University of Colorado Denver, Denver, Colorado

Submitted 16 June 2010; accepted in final form 7 August 2010

Curran-Everett D. Explorations in statistics: correlation. Adv Physiol Educ 34: 186–191, 2010; doi:10.1152/advan.00068.2010.—Learning about statistics is a lot like learning about science: the learning is more meaningful if you can actively explore. This sixth installment of Explorations in Statistics explores correlation, a familiar technique that estimates the magnitude of a straight-line relationship between two variables. Correlation is meaningful only when the two variables are true random variables: for example, if we restrict in some way the variability of one variable, then the magnitude of the correlation will decrease. Correlation cannot help us decide if changes in one variable result in changes in the second variable, if changes in the second variable result in changes in the first variable, or if changes in a third variable result in concurrent changes in the first two variables. Correlation can help provide us with evidence that study of the nature of the relationship between \( x \) and \( y \) may be warranted in an actual experiment in which one of them is controlled.

Sir Francis Galton; R; regression; spurious correlation

I can only say that there is a vast field of topics that fall under the laws of correlation, which lies quite open to the research of any competent person who cares to investigate it.
Sir Francis Galton (1890)

This sixth paper in Explorations in Statistics (see Refs. 4–8) explores correlation, a technique that estimates the magnitude of a straight-line relationship between two variables. By its very nature, correlation epitomizes the difference between statistical significance and scientific importance (5, 6, 10). Although basic textbooks of statistics (3, 30, 40) discuss correlation, the value of correlation is limited (23, 25, 40). Nevertheless, correlation has a rich history, and it can help provide us with evidence of a relationship between two variables (30).

A Brief History of Correlation

Sir Francis Galton (Fig. 1) pioneered correlation (21, 35, 36, 39a, 42, 43). Galton, a cousin of Charles Darwin, did a lot: he studied medicine, he explored Africa, he published in psychol-ogy and anthropology, he developed graphic techniques to map the weather (39a, 42). And, like others of his era, Galton strove to understand heredity (13, 14, 17, 20).

In 1877, Galton unveiled reversion, the earliest ancestor of correlation, and described it like this (13):

Reversion is the tendency of that ideal mean type to depart from the parent type, reverting towards what may be roughly and perhaps fairly described as the average ancestral type.

The empirical fodder for this observation? The weights of 490 sweet peas. Nine years later, Galton (14) summarized his sweet pea observations in this way:

It appeared from these experiments that the offspring did not tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small.

In Galton’s subsequent writings (14, 17, 20), reversion evolved into regression.

It was in 1888 that Galton (15) first wrote about correlation:

Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction . . . . It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the co-relation would be perfect, as is approximately the case with the symmetrically disposed parts of the body. If they were in no respect due to common causes, the co-relation would be nil . . . .

The statures of kinsmen are co-related variables; thus, the stature of the father is correlated to that of the adult son . . . ; the stature of the uncle to that of the adult nephew . . . and so on; but the index of co-relation, which is what I there [Ref. 14] called regression, is different in the different cases.

By 1889, Galton was writing co-relation as correlation (42), and he had become fascinated by fingerprints (16, 19). Galton’s 1890 account of his development of correlation (18) would be his last substantive paper on the subject (43).

Karl Pearson, Galton’s colleague and friend, and father of Egon Pearson, pursued the refinement of correlation (33, 34, 37) with such vigor that the statistic \( r \), a statistic Galton called the index of co-relation (15) and Pearson called the Galton coefficient of reversion (36), is known today as Pearson’s \( r \).

With this brief history, we are ready to begin our exploration of correlation.

R: Basic Operations

In the first paper (4) of this series, I summarized R (38) and outlined its installation. For this exploration, there are two additional steps: download Advances_Statistics_Code_Corr.R1 to your Advances folder and install the extra package MASS.2

To install MASS, open R and then click Packages | Install package(s) . . . 3 Select a CRAN mirror4 close to your location.

Address for reprint requests and other correspondence: D. Curran-Everett, Div. of Biostatistics and Bioinformatics, M222, National Jewish Health, 1400 Jackson St., Denver, CO 80206 (e-mail: EverettD@NJHealth.org).

1 This file is available through the Supplemental Material link for this article at the Advances in Physiology Education website.

2 MASS was so named because its functions support Modern Applied Statistics with S (44).

3 This file is available through the Supplemental Material link for this article at the Advances in Physiology Education website.

2 MASS was so named because its functions support Modern Applied Statistics with S (44).
As we know, correlation estimates the magnitude of the straight-line relationship between two variables. The usual correlation statistic is Pearson’s $r$, defined as

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2\sum (y_i - \bar{y})^2}} \ , \quad (I)$$

where $x_i$ and $y_i$ are the values of $x$ and $y$ for observation $i$ and where $\bar{x}$ and $\bar{y}$ are the sample means of $x$ and $y$ (11, 30, 39, 40).\(^5\) By virtue of the Cauchy-Schwarz inequality (28), the value of correlation coefficient $r$ can vary from $-1$ to $1$.\(^6\)

If $r = -1$, then the observations fall on a straight line whose slope is negative. If $r = +1$, then the observations fall on a straight line whose slope is positive. Figure 4 depicts the straight-line relationship between $x$ and $y$ for correlation coefficients that range from $-0.14$ to $-0.91$.

In our first exploration, we generated theoretical distributions for two sample statistics: the standard deviation and the mean (Ref. 4, Figs. 3 and 5). We can do the same thing for the sample correlation coefficient $r$ (Fig. 5).

If we restrict $x$ or $y$ in some way, then the correlation coefficient will reflect not only the relationship between $x$ and $y$ but also our restriction (30). For example, if we decrease the variability of $x$, then the magnitude of the correlation coefficient will decrease (Fig. 6); this is a general phenomenon (Fig. 7). The message: correlation is meaningful only when $x$ and $y$ are random variables (23, 30, 40).

\(^5\) The notation $\text{click A | B}$ means $\text{click A}$, then $\text{click B}$.

\(^6\) The Cauchy-Schwarz inequality dictates that $|x \cdot y| \leq |x||y|$. In Eq. 1, this means the magnitude of the numerator is less than or equal to the magnitude of the denominator. As a result, $|r| \leq 1$. 

---

**Correlation**


With these 9 sets of sample observations and statistics, we are ready to explore correlation.

**The Simulations: Observations and Sample Statistics**

In our early explorations (4–6) we drew 1000 random samples of 9 observations from a standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. For this exploration, we want to draw instead a single sample of 100 observations from 9 bivariate normal distributions (Figs. 2–4).

For each sample of 100 observations, we calculate the sample statistics listed in Table 1. These are the statistics for the 9 samples:
Moreover, the apparent meaning of a correlation can be distorted by a single observation (Fig. 8). Applets (22, 29, 41) are useful tools with which to explore this effect.

**Relationship to Regression**

Just as we can use correlation to explore a relationship between two variables, so too can we use regression. Regression estimates a straight-line relationship between two variables, so too can we use regression. Regression is a useful tool with which to explore this effect.

The relationship between the correlation coefficient $r$ and the slope $b_1$ is

$$b_1 = r \frac{s_y}{s_x},$$

where $s_y$ and $s_x$ are the sample standard deviations of $y$ and $x$ (11, 40). In this exploration, we defined the standard deviation $\sigma$ of $X$ and $Y$ to be 1 (see Fig. 3). This means we expect $s_y$ and $s_x$ to be similar. As a result, the slope $b_1$ will approximate the correlation coefficient $r$. We can use our sample statistics to confirm this empirically. Suppose we use the statistics from when the correlation $\rho$ was −0.2:

$$b_1 = r \frac{s_y}{s_x} = -0.2599 \cdot \frac{0.9296}{0.9516} = -0.2539.$$

This derived value matches the value of the slope we obtained from regression (Table 1, column 9). Regardless of $s_y$ and $s_x$, if the correlation coefficient $r$ is 0, then the slope $b_1$ will be 0.

The relationship between $r$ and the regression statistic $R^2$ is direct (11, 40):

$$R^2 = r^2,$$

We can use our sample correlation coefficient $r = −0.2599$ to confirm this:

$$R^2 = (-0.2599)^2 = 0.0675.$$

This derived value matches the value we obtained from regression (Table 1, column 10).

With this background, the next question is, how do we interpret a correlation?

**Table 1. Sample statistics calculated for each random sample**

<table>
<thead>
<tr>
<th>Column</th>
<th>Heading</th>
<th>Sample Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\rho$</td>
<td>Population correlation $\rho$</td>
</tr>
<tr>
<td>2</td>
<td>Ave$_X$</td>
<td>Average of the $x$ values, $\bar{x}$</td>
</tr>
<tr>
<td>3</td>
<td>SD$_X$</td>
<td>Standard deviation of the $x$ values, $s_x$</td>
</tr>
<tr>
<td>4</td>
<td>Ave$_Y$</td>
<td>Average of the $y$ values, $\bar{y}$</td>
</tr>
<tr>
<td>5</td>
<td>SD$_Y$</td>
<td>Standard deviation of the $y$ values, $s_y$</td>
</tr>
<tr>
<td>6</td>
<td>$r$</td>
<td>Sample correlation $r$</td>
</tr>
<tr>
<td>7</td>
<td>$P$</td>
<td>$P$ value associated with $H_0: \rho = 0$</td>
</tr>
<tr>
<td>8</td>
<td>Intcptrt</td>
<td>Intercept $b_0$ from the fitted regression</td>
</tr>
<tr>
<td>9</td>
<td>Slope</td>
<td>Slope $b_1$ from the fitted regression</td>
</tr>
<tr>
<td>10</td>
<td>R_Sqrd</td>
<td>$R^2$ from the fitted regression</td>
</tr>
</tbody>
</table>
Interpretation

Suppose we want to estimate $\rho$, the correlation between two physiological things we care about. We define the null and alternative hypotheses, $H_0$ and $H_1$, as

$$H_0: \rho = 0 \quad H_1: \rho \neq 0.$$  

That is, the sample observations are consistent with having come from a bivariate population that has a correlation $\rho$ of 0, or the sample observations are consistent with having come from a bivariate population that has a correlation $\rho$ other than 0 (6).

To test this null hypothesis, suppose we draw at random 100 observations from a bivariate normal distribution with a correlation $\rho = -0.2$ (see Fig. 3, middle left). The scatterplot in Fig. 6, bottom right, depicts the relationship we observe between $x$ and $y$, the things we care about.

When we estimate the sample correlation coefficient $r$, we obtain $r = -0.2599 \approx -0.26$. Associated with this value of $r$ is a $P$ value of 0.009. What do we do with $r = -0.26$? We interpret it within the context of a true null hypothesis: if the true correlation is 0, how unusual is this value of $r$? If the null hypothesis is true, we expect to observe a value of $|r|$ at least as big as 0.26 just 9 times in 1000 ($P = 0.009$). Suppose we established beforehand a benchmark of $\alpha = 0.01$ (9). If the null hypothesis is true, the sample correlation $|r| = 0.26$ is more unusual than our benchmark. As a result, we reject the null hypothesis and conclude that the sample observations are consistent with having come from a bivariate population that has a correlation $\rho$ other than 0.

The statistical conclusion is clear: changes in $y$ are related to changes in $x$. The scientific question is, how well do changes in $x$ account for changes in $y$? We can answer this question if we square $r$ to get $R^2$:

$$R^2 = r^2.$$  

One test statistic with which we can assess $H_0: \rho = 0$ is the familiar $t$ statistic. Because the distribution of the sample correlation coefficient $r$ is complex (Fig. 5), we compute $t$ as

$$t = \sqrt{r^2(n-2)/(1-r^2)},$$  

where $n$ is the number of sample observations (30, 40). The degrees of freedom for this $t$ distribution is $n - 2$. 

---

The text provides a detailed explanation of statistical concepts, including hypothesis testing, correlation coefficients, and the calculation of $R^2$ and $t$-statistics. It includes figures illustrating theoretical distributions and sample correlation coefficients, as well as code snippets for generating these simulations.
Fig. 8. The impact of a single point on the correlation coefficient. For the 11 original observations (black circles), the sample correlation coefficient \( r = +0.56 \) \((P = 0.07)\). Data are a subset of those reported by Pearson and Lee (37) and used by Snedecor and Cochran (40). If the point \((65, 59)\) becomes instead \((65, 69)\), the correlation coefficient \( r \) decreases to \(-0.02 \) \((P = 0.96)\). The commands in lines 138–163 of Advances_Statistics_Code_Corr.R generate this simulation. To generate this data graphic, highlight and submit the lines of code from Figure 8: first line to Figure 8: last line.

\[
R^2 = (-0.2599)^2 = 0.0675 = 0.07
\]

How well do changes in \( x \) account for changes in \( y \)? Not well: changes in \( x \) account for 7% of the variation in \( y \). The scatterplot in Fig. 4, top middle, confirms this weak relationship.

Bear in mind that a correlation may not be meaningful even when changes in \( x \) account for most of the changes in \( y \) (Fig. 9). The message: a scatterplot is essential if you want to interpret a correlation.

A correlation is not meaningful if \( x \) and \( y \) are related through computation. For some physiological response, suppose we consider the initial value \( y_i \) and the subsequent change \( \Delta y \) to be drawn at random from a normal distribution with mean \( \mu = 0 \) and standard deviation \( \sigma = 1 \). There is no relationship between the change \( \Delta y \) and the initial value \( y_i \) (Fig. 10).

\[
\begin{align*}
\hat{y} &= 3 + 0.5x \\
r &= 0.92 \\
\end{align*}
\]

calculate the final value \( y_f \) as \( y_f = y_i + \Delta y \), we create an obvious relationship between the final value \( y_f \) and the initial value \( y_i \). This phenomenon is one example of spurious correlation, recognized by Pearson (34) and condemned by Neyman (31): “Spurious correlations have been ruining empirical statistical research from times immemorial.”

A correlation is also not meaningful if \( x \) and \( y \) are values that result from different ways of measuring the same thing (2, 23, 26, 27, 40). In this situation, \( x \) and \( y \) will be related simply because they estimate the same quantity.

Imagine this: you discover that changes in \( y \) are related to changes in \( x \) \((P = 0.01)\) and, even better, that changes in \( x \) account for 90% of the variation in \( y \). You like this. A lot. But what do you make of this result? You have three choices:

1. Changes in \( x \) result in changes in \( y \).
2. Changes in \( y \) result in changes in \( x \).
3. Changes in \( z \) result in changes in \( x \) and \( y \).

From correlation alone, you have no idea what is going on.

When I teach my statistics course, I use these examples (24) to illustrate the tenuous nature of correlation:

Lice make you healthy. You say you want evidence? On islands in the South Pacific, healthy people have lice. Sick people do not. What more do you need? The explanation? When someone gets sick, they often develop a fever. Lice do not like higher temperatures.

There is a tight relationship between the salaries of Presbyterian ministers in Massachusetts and the price of rum in Havana. The difficult question is, do the ministers support the rum trade, or do they profit from it? The answer is neither: worldwide inflation affects salaries and the price of rum.

In the first example, it is not that lice \((x)\) make you healthy \((y)\). Rather, it is that being sick with a fever \((y)\) drives off lice \((x)\). In the latter example, worldwide inflation \((z)\) affects salaries \((x)\) and the price of rum \((y)\).
Summary

As this exploration has demonstrated, correlation estimates the magnitude of a simple straight-line relationship between two variables, and it embodies the fundamental concepts of statistical significance and scientific importance (5, 6, 10). Still, correlation cannot help us decide if changes in x result in changes in y, if changes in y result in changes in x, or if changes in a third variable result in simultaneous changes in x and y. Correlation can help provide us with evidence that study of the nature of the relationship between x and y may be warranted in an actual experiment in which one of them is controlled.

In the next installment of this series, we will explore regression, a technique that, like correlation, also estimates the nature of different kinds of relationships between two variables. Unlike correlation, however, regression can estimate the nature of different kinds of relationships between two variables. Because of this versatility, regression is a far more useful technique.

ACKNOWLEDGMENTS

I thank John Ludbrook (Department of Surgery, The University of Melbourne, Melbourne, Victoria, Australia) and Lori Silveira and Matthew Strand (National Jewish Health, Denver, CO) for their helpful comments and suggestions.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

REFERENCES

33. Pearson K. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond 60: 489–498, 1897.