Correlation

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 14, 2013

A Beginning Quotation

The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.
— Stephen Jay Gould

Week 4: Correlation

— "Voodoo Correlations in Social Neuroscience"; the "culling" or search for results in clinical trials and elsewhere, with a subsequent failure to cross-validate what is found; more generally, the problem of "double dipping"

Required Reading:
SGEP (119–140) —
Illusory Correlation
Ecological Correlation
Restriction of Range for Correlations
Odd Correlations
Measures of Nonlinear Association
Intraclass Correlation

Film:

Florence Nightingale (65 minutes)

Snow (22 minutes)

Pearson Product Moment Correlation Coefficient: Definition

The association between two variables measured on the same set of objects is commonly referred to as their correlation and often measured by the Pearson product moment correlation coefficient.

Suppose $Z_{X_1}, \ldots, Z_{X_N}$ and $Z_{Y_1}, \ldots, Z_{Y_N}$ refer to z-scores (that is, having mean zero and variance one) calculated for our original observational pairs, (X_i, Y_i), $i = 1, \ldots, N$.

The correlation between the original variables, r_{XY}, is defined as

$$r_{XY} = (\frac{1}{N}) \sum_{i=1}^{N} Z_{X_i} Z_{Y_i} \ ,$$

or the average product of the z-scores.

The Pearson correlation, r_{XY}, only gives a measure of a *linear* relation that might be present between two variables.

If some nonlinear form of association exists, other measures of correlation should be used.

We will (eventually) discuss four possibilities:
1) Guttman's (weak) monotonicity coefficient (μ_2);
2) Goodman-Kruskal's gamma (γ) coefficient (for a contingency table with ordered classes);
3) Goodman-Kruskal's lambda (λ) coefficient (for a contingency table with unordered classes);
4) Spearman's rank-order correlation coefficient (this is just the Pearson correlation computed using the ranks of the values on the two variables).

The Importance of Scatterplots

1) To assess whether the type of association present might be linear; we could impose scatterplot "smoothers" to evaluate the type of association present;

2) To identify outliers and help figure out why these data points might not be reflective of the general pattern that is present.

3) To help assess the influence of certain data points on the correlation, e.g., by using the size and fill for a plotting symbol to indicate the change in the correlation that would result when the data point was removed.

Correlation Does Not Imply Causation

Latin variants:
post hoc, ergo propter hoc (after this, therefore because of this)
cum hoc, ergo propter hoc (with this, therefore because of this)

Generally, the association we see between two variables might be due to a "lurking" third variable.

A current insidious example: parents who blame children's autism on earlier receiving the MMR vaccine and who therefore now refuse to vaccinate their children – the "herd immunity" levels are dropping badly in some communities.

A Correlation is a Symmetric Measure

Thus, the directionality of any possible causal inference is unknown. Examples:

1) The positive correlation between winning football games and the amount of ground yardage gained: "running the ball" does not necessarily cause winning; winning may cause "running the ball" to wind down the clock;

2) The positive effects seen for moderate drinking may be due to individuals who besides leading healthy lifestyles also drink moderately;

3) The greater the money a political candidate brings in may not be the reason for winning; maybe greater electability leads to more donations.

Algebraic Restrictions on Correlations

It is possible to derive the algebraic restrictions present among any subset of the variables based on the correlations among all the variables.

The simplest case involves three variables, say X, Y, and W. From the basic formula for the partial correlation between X and Y "holding W constant," an *algebraic* restriction is present on r_{XY} given the values of r_{XW} and r_{YW}:

$$r_{XW}r_{YW} - \sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)} \leq$$

$$r_{XY} \leq r_{XW}r_{YW} + \sqrt{(1 - r_{XW}^2)(1 - r_{YW}^2)} \; .$$

An Example of Algebraic Restriction

Suppose X and Y refer to height and weight, respectively, and W is a measure of age. If, say, the correlations, r_{XW} and r_{YW} are both .8, then $.28 \leq r_{XY} \leq 1.00$.

In fact, if a high correlation value of .64 were observed for r_{XY}, should we be impressed by the magnitude of the association between X and Y? Probably not —

if the partial correlation between X and Y "holding W constant" were computed with $r_{XY} = .64$, a value of zero would be obtained. All of the observed high association between X and Y can be attributed to their association with the developmentally related variable.

A Two by Two Contingency Table: The Fourfold Point Correlation

A related type of algebraic restriction for a correlation is present when the distribution of the values taken on by the variables include ties.

In the extreme, consider a 2×2 contingency table, and the fourfold point correlation; this is constructed by using a $0/1$ coding of the category information on the two attributes and calculating the usual Pearson correlation.

Because of the nonuniform marginal frequencies present in the 2×2 table, the fourfold correlation cannot extend over the complete ± 1 range.

The achievable bounds possible can be computed (Carroll, 1961); and it therefore may be of some interest descriptively to see how far an observed fourfold correlation is away from its achievable bounds, and possibly, even to normalize the observed value by such a bound.

Guttman's Coefficient of Monotonicity

The bounds of ± 1 on a Pearson correlation can be achieved only by datasets demonstrating a perfect linear relationship between the two variables.

Another measure that achieves the bounds of ± 1 whenever the datasets have merely consistent rank orderings is Guttman's (weak) monotonicity coefficient, μ_2:

$$\mu_2 = \frac{\sum_{i=1}^{n} \sum_{h=1}^{n} (x_h - x_i)(y_h - y_i)}{\sum_{i=1}^{n} \sum_{h=1}^{n} |x_h - x_i||y_h - y_i|} \ ,$$

where (x_h, y_h) denote the pairs of values being "correlated" by μ_2.

The coefficient, μ_2, expresses the extent to which values on one variable increase in a particular direction as the values on another variable increases, without assuming that the increase is exactly according to a straight line.

It varies between -1 and $+1$, with $+1$ $[-1]$ reflecting a perfect monotonic trend in a positive [negative] direction.

In contrast to the Pearson correlation, μ_2 can equal $+1$ or -1, even though the marginal distributions of the two variables differ from one another.

When the Pearson correlation is $+1.00$ or -1.00, μ_2 will have the same value; in all other cases, the absolute value of μ_2 will be higher than that of the Pearson correlation including the case of a fourfold point correlation.

Illusory Correlation

An illusory correlation is present whenever a relation is seen in data where none exists.

Common examples would be between membership in some minority group and rare and typically negative behavior; the endurance of stereotypes and an overestimation of the link between group membership and certain traits; or the connection between a couple adopting a child and the subsequent birth of their own.

Four decades ago, Chapman and Chapman (1967, 1969) studied such false associations in relation to psychodiagnostic signs seen in projective tests. For example, in the "Draw-A-Person" test, a client draws a person on a blank piece of paper. Although some psychologists believe that drawing a person with big eyes is a sign of paranoia, such a correlation is illusory but very persistent.

Confirmation Bias

Several faulty reasoning relatives exist for the notion of an illusory correlation.

One is *confirmation bias*, where there are tendencies to search for, interpret, and remember information only in a way that confirms one's preconceptions or working hypotheses.

At an extreme, there is the trap of *apophenia*, or seeing patterns or connections in random or meaningless data.

One particular problematic realization of apophenia is in epidemiology when residential cancer clusters are identified that rarely if ever result in identifiable causes.

Texas Sharpshooter Fallacy

What seems to be occurring is sometimes labeled the Texas sharpshooter fallacy, where a Texas sharpshooter fires at the side of a barn and then draws a bullseye around the largest cluster of bullet holes.

In identifying residential cancer clusters, we tend to notice multiple cancer patients on the same street and then define the population base around these.

A particularly well-presented popular article on these illusory associations entitled "The Cancer-Cluster Myth," is by Atul Gawande in the February 8th 1999, *New Yorker*.

Clinician's Fallacy

Illusory relations occur commonly in our day-to-day interactions with others. We have the *clinician's fallacy* due to the self-selected and biased sample of individuals whom a clinician actually sees in practice.

Thus, we have the (incorrect) inference of a uniformity of serious adult trauma for any instance of childhood sexual abuse (see McNally, 2003, *Remembering Trauma*, for a comprehensive discussion).

The representativeness of what is encountered should always be kept in mind. Forms of selection bias appear constantly because what is observed or heard results from its being different than what usually happens.

Making inferences based on out-of-the-ordinary events is generally not a good idea.

Ecological Correlation

An ecological correlation is one calculated between variables that are group averages of some sort; this is in contrast to obtaining a correlation between variables measured at the level of individuals.

Several issues are faced immediately with the use of ecological correlations:

they tend to be a lot higher than individual-level correlations,

and assuming that what is seen at the group level also holds automatically at the level of the individual is so pernicious that it has been labeled the "ecological fallacy" by Selvin (1958).

The specific instance developed by Selvin concerns the 19th century French sociologist Émile Durkheim and his contention that suicide was promoted by the social conditions inherent in Protestantism.

This individual-level inference is not justified from the data Durkheim actually had at the aggregated level of country, which did show a relationship between the levels of Protestantism and suicide.

As is true in interpreting any observational study, confounding variables may exist; here, it is that that Protestant countries differ from Catholic countries in many ways other than religion.

Durkheim's data do not link individual level suicide with the practice of any particular religious faith; and to do so is to fall prey to the ecological fallacy.

Ecological Regression

Suppose our interest is in the estimation of support for a specific candidate among Hispanic and non-Hispanic voters.

For each electoral precinct, the fraction, x, of voters who are Hispanic is known, as is the fraction, y, of voters for the candidate.

The problem is to estimate the fraction of Hispanic voters for the candidate, which is unknown because of ballot secrecy.

A regression equation is fitted to the data having the form

$$y_i = a + bx_i + \epsilon_i \ ,$$

where x_i is the fraction of Hispanic voters in precinct i, y_i is the vote fraction for the candidate, and ϵ_i is the error term.

Least squares estimates of a and b are denoted by $\hat{a}$ and $\hat{b}$.

Here, $\hat{a}$ is the height of the regression line at $x = 0$, corresponding to precincts having no Hispanic voters; $\hat{a} + \hat{b}$ is the height of the regression line at $x = 1$, and interpretable as the fraction of Hispanic voters supporting the candidate.

Justifying the statistical procedure requires invoking the "constancy assumption"—voting preferences within ethnic groups do not systematically depend on the ethnic composition of the area of residence.

Strong conditions such as the constancy assumption are generally unverifiable but must be assumed true.

Aggregation causes problems across several disciplines:

Aggregation bias in econometrics refers to how aggregation changes the micro-level structural relationships among the economic variables of interest; explicitly, aggregation bias is a deviation of the macro-level parameters from the average of the corresponding micro-level parameters.

Or in psychology, the models we fit and evaluate at the group level may be very different from what is operative at the level of the individual subject.

The fundamental difficulty with ecological inference is that many different possible relationships at an individual level are capable of generating the same results at an aggregate level.

No deterministic solution exists for the ecological inference problem—

individual-level information is irretrievably lost by the process of aggregation.

Modifiable Areal Unit Problem

A problem related to ecological correlation is the modifiable areal unit problem, where differences in spatial units used in the aggregation can cause wide variation in the resulting correlations, ranging anywhere from plus to minus one.

Generally, the manifest association between variables depends on the size of areal units used, with increases as areal unit size gets larger.

A related "zone" effect concerns the variation in correlation caused by reaggregating data into different configurations at the same scale.

Restriction of Range for Correlations

When a psychological test is used to select personnel based on the achievement of a certain cut-score, an unusual circumstance may occur.

The prediction of job performance after selection is typically much poorer than what one might have expected beforehand.

In one of the more well-known papers in all of Industrial and Organizational Psychology, Taylor and Russell (1939) offered an explanation of this phenomenon by noting the existence of a restriction of range problem:

in a group selected on the basis of some test, the correlation between test and performance must be lower than it would be in an unselected group.

Based on the assumption of bivariate normality between job performance and the selection test, Taylor and Russell provided tables and charts for estimating what the correlation would be in an unselected population from the value seen between test and performance in the selected population.

Odd Correlations

A recent article (Vul et al. 2009) in *Perspectives on Psychological Science*, has the intriguing title, "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" (renamed from the earlier and more controversial "Voodoo Correlations in Social Neuroscience").

These authors comment on the extremely high (for example, greater than .80) correlations reported in the literature between brain activation and personality measures, and point out the fallaciousness of how they were obtained.

Typically, huge numbers of separate correlations were calculated, and only the mean of those correlations exceeding some threshold (based on a very small significance level) are reported.

It is tautological that these correlations selected for size must then be large in their average value.

With no cross-validation attempted to see the shrinkage expected in these measures on new samples, we have sophistry at best.

Any of the usual understanding of yardsticks provided by the correlation or its square, the proportion of shared variance, are inappropriate.

In fact, as noted by Vul et al. (2009), these inflated mean correlations typically exceed the upper bounds provided by the correction for attenuation based on what the reliabilities should be for the measures being correlated.

Measures of Nonlinear Association

An overall concern with the use of the simple correlation coefficient is that it measures linearity only, and then through a rather indirect measure of shared variance defined by the coefficient of determination, the squared correlation.

Specifically, there is no obvious operational measure of strength of relation defined in terms of the given sample at hand, which in turn could be given a transparent probabilistic meaning with respect to the latter.

One century-old suggestion is to use the Spearman correlation, which is equivalent to the Pearson correlation computed on ranks.

Although it is true that a perfect monotonic relation between two variables turns into one that is perfectly linear when ranks are used, the strength of such an association measure is now a somewhat unsatisfying shared variance between ranks.

Goodman-Kruskal Gamma Coefficient

An alternative notion of rank correlation is based on the number of inversions in rank ordering for the two variables, X and Y, taken over all object pairs.

Suppose (x_i, y_i) and (x_j, y_j) are the observed measures for two objects, i and j. If $x_i > x_j$ but $y_i < y_j$, we have an "inversion"; when $x_i > x_j$ and $y_i > y_j$, a "noninversion" exists.

A simple measure of rank-order association is the Goodman–Kruskal (G-K) γ (gamma) coefficient obtained over the $N(N-1)/2$ object pairs:

the ratio of the number of noninversions (S_+) minus the number of inversions (S_-), all divided by the sum of S_+ and S_- .

The G-K γ coefficient is bounded between plus and minus 1.0, and can be given a convenient and transparent probabilistic meaning with respect to the given sample:

if we choose two objects at random, and consider the ordering provided by untied values on X and Y, γ is the probability of a noninversion minus the probability of an inversion.

Goodman-Kruskal Lambda Coefficient

The G-K γ measure is appropriate only for a contingency table in which the two cross-classification attributes consist of ordered categories.

A more general measure that relates two arbitrary attributes considered to be nominal where both have assumed unordered categories, was also proposed by Goodman and Kruskal and labeled by the Greek letter lambda, λ (the Goodman–Kruskal (G-K) Index of Predictive Association).

We define it in terms of an $R \times C$ contingency table having the following form:

	A_1	A_2	$\cdots$	A_C	Row Sums
B_1	N_{11}	N_{12}	$\cdots$	N_{1C}	$N_{1\cdot}$
B_2	N_{21}	N_{22}	$\cdots$	N_{2C}	$N_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
B_R	N_{R1}	N_{R2}	$\cdots$	N_{RC}	$N_{R\cdot}$
Column Sums	$N_{\cdot 1}$	$N_{\cdot 2}$	$\cdots$	$N_{\cdot C}$	$N_{\cdot\cdot} \equiv N$

Suppose a process is initiated where an object is picked from the table and the row event that occurs is noted (that is, $B_1, \ldots, B_R$).

Then based on this knowledge of the row event, say B_r, we guess the column event by choosing that column with the highest frequency within row B_r.

An error of prediction is made with probability

$$\frac{N_{r\cdot} - \max_{1 \le c \le C} N_{rc}}{N_{r\cdot}},$$

and using the rule of total probability, the overall error of prediction is

$$\sum_{r=1}^{R} \left(\frac{N_{r\cdot} - \max_{1 \le c \le C} N_{rc}}{N_{r\cdot}} \right) \left(\frac{N_{r\cdot}}{N_{\cdot\cdot}} \right) = 1 - \frac{\sum_{r=1}^{R} \max_{1 \le c \le C} N_{rc}}{N_{\cdot\cdot}},$$

and denoted by $P_{error|row}$ = probability of an error in predicting the column category given knowledge of the row category.

If an object is picked from the table and we are asked to make a best prediction of column category without any further information, the column category with the largest frequency would be used.

An error in prediction is made with probability

$$\frac{N_{..} - \max_{1 \leq c \leq C} N_{.c}}{N_{..}} = 1 - \frac{\max_{1 \leq c \leq C} N_{.c}}{N_{..}} \ ,$$

denoted by P_{error} = probability of an error in predicting the column category.

These two probabilities can be used to form a proportional reduction in error measure, $\lambda_{A|B}$ (predicting a column category $(A_1, \ldots, A_C)$ from a row category $(B_1, \ldots, B_R)$):

$$\lambda_{A|B} = \frac{P_{error} - P_{error|row}}{P_{error}} =$$

$$\frac{\left(\sum_{r=1}^{R} \max_{1 \leq c \leq C} N_{rc}\right) - \max_{1 \leq c \leq C} N_{\cdot c}}{N_{\cdot\cdot} - \max_{1 \leq c \leq C} N_{\cdot c}} .$$

If $\lambda_{A|B}$ is zero, then the maximum of the column marginal frequencies, $\max_{1 \leq c \leq C} N_{\cdot c}$, is the same as the sum of the maximum column frequencies within rows.

In other words, no differential predictions of a column event are made based on knowledge of what row an observation belongs to.

The G-K λ measure is asymmetric, and $\lambda_{A|B}$ is not necessarily the same as $\lambda_{B|A}$; for example, one measure could be zero and the other positive.

In general, there is no necessary relation between $\lambda_{A|B}$ and $\lambda_{B|A}$ and the usual chi-square association statistic.

The latter is a nontransparent measure of relationship in a contingency table with unordered attributes that increases proportionately with increasing sample size;

a λ measure has the transparent interpretation in terms of the differential predicability of one attribute from another.

Given that only nominal attributes are required, it is universally appropriate for just about any task of relating two variables, irrespective of the levels of measurement they might have.

Intraclass Correlation

A different type of correlational measure, an intraclass correlation coefficient (ICC), can be used when quantitative measurements are made on units organized into groups, typically of the same size.

It measures how strongly units from the same group resemble each other.

Here, we will emphasize only the case where group sizes are all 2, possibly representing data on a set of N twins, or two raters assessing the same N objects.

The basic idea generalizes, however, to an arbitrary number of units within each group.

Early work on the ICC from R. A. Fisher and his contemporaries conceptualized the problem as follows:

let $(x_i, x_i^{'})$, $1 \leq i \leq N$, denote the N pairs of observations (thus, we have N groups with two measurements in each).

The usual correlation coefficient cannot be computed, however, because the order of the measurements within a pair is unknown (and arbitrary).

As an alternative, we first double the number of pairs to $2N$ by including both $(x_i, x_i^{'})$ and $(x_i^{'}, x_i)$. The Pearson correlation is then computed using the $2N$ pairs to obtain an ICC.

Random Effects Model

As a more convenient and generalizable version of the ICC computations, we adopt the Model II (random effects) analysis-of-variance model:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \ ,$$

where Y_{ij} is the jth observation in the ith group, μ is the overall mean, α_i is a random variable indicating an effect shared by all values in group i, and ϵ_{ij} is a random variable representing error.

The two random variables, α_i and ϵ_{ij}, are assumed uncorrelated within and between themselves with expected values of zero, and variances of σ_α^2 and σ_ϵ^2, respectively.

The population ICC parameter is given by

$$\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2},$$

and estimated by a ratio:

(Mean Square Between $-$ Mean Square Within) divided by (Mean Square Between $+$ Mean Square Within).

Heritability Coefficient

In studying heritability, we need two central terms:

Phenotype: the manifest characteristics of an organism that result from both the environment and heredity; these characteristics can be anatomical or psychological, and are generally the result of an interaction between the environment and heredity.

Genotype: the fundamental hereditary (genetic) makeup of an organism; as distinguished from (phenotypic) physical appearance.

Based on the random effects model (Model II) for describing a particular phenotype, symbolically we have:

Phenotype(P) = Genotype(G) + Environment(E), or in terms of variances, Var(P) = Var(G) + Var(E), assuming that the covariance between G and E is zero.

The ICC in this case is the heritability coefficient,

$$H^2 = \frac{\mathrm{Var}(G)}{\mathrm{Var}(P)} \ .$$

Heritability estimates are often misinterpreted, even by those who should know better.

In particular, heritability refers to the proportion of variation between individuals in a population influenced by genetic factors.

Thus, because heritability describes the population and not the specific individuals within it, it can lead to an aggregation fallacy when one tries to make an individual-level inference from a heritability estimate.

For example, it is incorrect to say that because the heritability of a personality trait is, say, .6, that therefore 60% of a specific person's personality is inherited from parents and 40% comes from the environment.

The term "variation" in the phrase "phenotypic variation" is important to note.

If a trait has a heritability of .6, it means that of the observed phenotypic variation, 60% is due to genetic variation.

It does not imply that the trait is 60% caused by genetics in a given individual.

Nor does a heritability coefficient imply that any observed differences between groups (for example, a supposed 15 point I.Q. test score difference between blacks and whites) is genetically determined.

As noted explicitly in Stephen Jay Gould's *The Mismeasure of Man* (1996), it is a fallacy to assume that a (high) heritability coefficient allows the inference that differences observed between groups must be genetically caused.

As Gould succinctly states: "[V]ariation among individuals within a group and differences in mean values between groups are entirely separate phenomena. One item provides no license for speculation about the other."

For an in-depth and cogent discussion of the distinction between heritability and genetic determination, the reader is referred to Ned Block, "How Heritability Misleads About Race" (*Cognition*, 1995, *56*, 99–128).