Correlational Research

By Marilyn K. Simon and Jim Goes

Includes excerpts from Simon (2011), *Dissertation and Scholarly Research: Recipes for Success*. Seattle, WA: Dissertation Success LLC

Find this and many other dissertation guides and resources at
www.dissertationrecipes.com

The correlational researcher investigates one or more characteristics of a group to discover the extent to which the characteristics vary together. Descriptive and correlational studies examine variables in their natural environments and do not include researcher-imposed treatments. Correlational studies display the relationships among variables by such techniques as cross-tabulation and correlations. Correlational studies are also known as ex post facto studies. This literally means *from after the fact.* The term is used to identify that the research has been conducted after the phenomenon of interest has occurred naturally. The main purpose of a correlational study is to determine relationships between variables, and if a relationship exists, to determine a regression equation that could be used make predictions to a population. In bivariate correlational studies, the relationship between two variables is measured. Through statistical analysis, the relationship will be given a degree and a direction. The degree of relationship determined how closely the variables are related. This is usually expressed as a number between -1 and +1, and is known as the correlation coefficient. A zero correlation indicates no relationship. As the correlation coefficient moves toward either -1 or +1, the relationship gets stronger until there is a perfect correlation at the end points.

The significant difference between correlational research and experimental or quasi-experimental design is that causality cannot be established through manipulation of independent variables. This leads to the pithy truism: *Correlation does not imply causation*. For example, in studying the relationship between smoking and cancer, the researcher begins with a sample of those who have already developed the disease and a sample of those who have not. The researcher then looks for differences between the two groups in antecedents, behaviors, or conditions such as smoking habits. If it is found that there is a relationship between smoking and a type of cancer, the researcher cannot conclude that smoking *caused* the cancer. Further research would be needed to draw such a conclusion.

Example: The relationship between socioeconomic status and school achievement of a group of urban ghetto children is examined.

Testing a Claim About the Relation Between Two Variables (Correlation and Regression Analysis)

Many real and practical situations demand decisions or inferences about how data from a certain variable can be used to determine the value of some other related variable. For example, researchers of a Florida study of the number of powerboat registrations and the number of accidental manatee deaths confirmed that there was a significant positive correlation. As a result, Florida legislators created coastal sanctuaries where powerboats are prohibited so that manatees could thrive.

Researchers of a study in Sweden found that there was a higher incidence of leukemia among children who lived within 300 meters of a high-tension power line during a 25-year period. This lead Sweden's government to consider regulations that would reduce housing in close proximity to high-tension power lines.

If you can answer yes to both questions below, you can use the identical statistical test described in this section. Are you claiming that

__ 1. There is a relationship or correlation between two factors, two events, or two characteristics?, **and**

__ 2. The data are at least of the interval measure?

To perform regression and correlational analyses:

1. Record the information in table form.
2. Create a scatter diagram see any obvious relationship or trends.
3. Compute the correlation coefficient *r*, also known as the Pearson correlation coefficient factor, to obtain objective analysis that will uncover the magnitude and significance of the relationship between the variables.

4. Determine if *r* is statistically significant. If *r* **is** statistically significant, then regression analysis can be used to determine the relationship between the variables.

Example: Suppose a randomly selected group of teachers is given the Survey on Calculator Use (SOCU) to measure how they integrate calculators in their classrooms and then tested for their levels of math anxiety using the Math Anxiety Rating Scales or MARS test:

1. The results for each participant is recorded in table form (some of these values appear below):

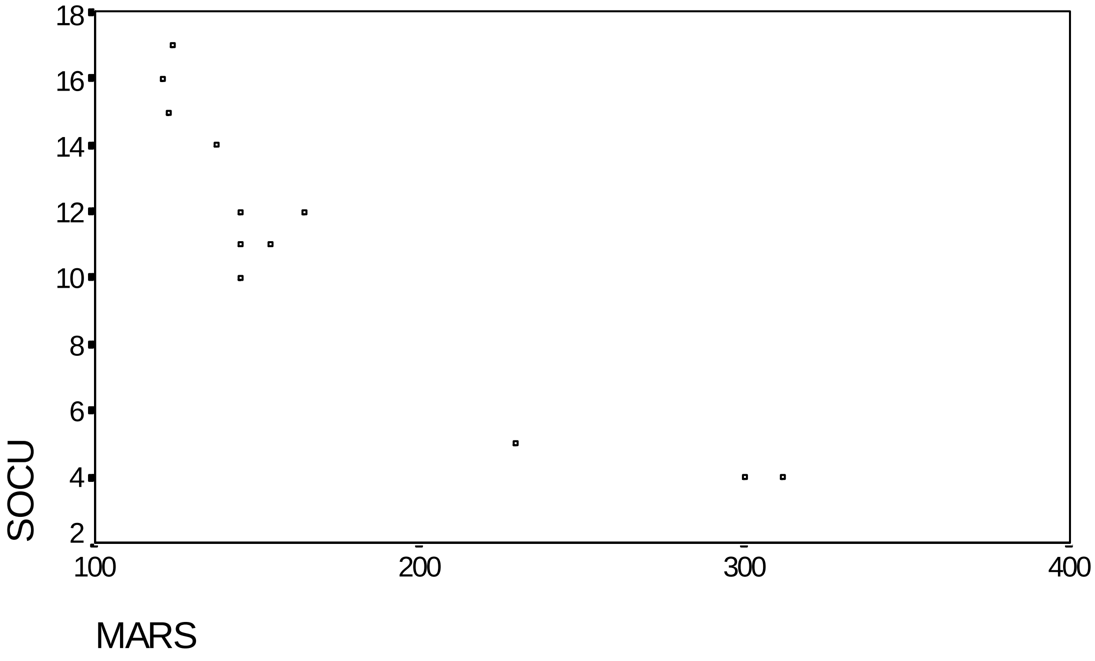
MARS	SOCU
123.00	15.00
145.00	12.00
154.00	11.00
121.00	16.00
230.00	5.00
300.00	4.00
145.00	10.00
124.00	17.00
145.00	11.00
165.00	12.00
138.00	14.00
312.0	4.00

The researcher's hypothesis is that teachers who have lower levels of math anxiety are more likely to use calculators in their classes. (Note: The independent variable (x) is the math anxiety level, determined by MARS, and is being used to predict the dependent variable (y), the use of calculators, as measured by SOCU.)

H_0: $r = 0$ (there is no relationship)
H_1: $r \neq 0$ (there is a relationship)

Note: These will usually be hypotheses in regression analysis.

2. Draw a scatter diagram:



The points in the figure above seem to follow a downward pattern, so we suspect that there is a relationship between level of math anxiety and the use of calculators by teachers surveyed, but this is somewhat subjective.

3. Compute *r*.

To obtain a more precise and objective analysis we can compute the linear coefficient constant, *r*. Computing *r* is a tedious exercise in arithmetic but practically any statistical computer program or scientific calculator would willingly help you along. In our example, the very user-friendly program SPSS determined that *r* = -0.882.

Some of the properties of this number *r* are as follows:

1. The computed value of *r* must be between -1 and +1. (If it's not then someone or something messed up.)

2. A strong positive correlation would yield an *r* value close to +1; a strong negative linear correlation would be close to -1.

3. If *r* is close to 0, we conclude that there is no significant linear correlation between *x* and *y*.

Checking the table, we find that with a sample size of 10 (n = 10), the value r = -0.9169, indicating a strong negative correlation between the use of calculators and measures of math anxiety levels. The r-squared number (0.779) indicates that a person's math anxiety might explain 84% of his or her calculator usage (or nonusage).

4. If there is a significant relation, then regression analysis is used to determine what that relationship is.

5. If the relation is linear, the equation of the line of best fit can be determined. (For two variables, the equation of a line can be expressed as $y = mx + b$, where m is the slope and b is the y–intercept.)

Thus, the equation of the line of best fit would be

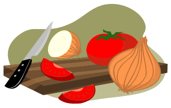
$$S = -.9169\,M + 21.614.$$

The nonparametric counterpart to the Pearson r is the Spearman rank correlation coefficient (r_s), Spearman's rho, or Kendall's tau (τ).

FOR YOUR INFORMATION AND EDUCATION

The full name of the Pearson r is the Pearson product-moment correlation coefficient. It is named for Karl Pearson (1857–1936), who originally developed it. It is called product-moment because it is calculated by multiplying the z scores of two variables by one another to get their product and then calculating the average or mean value, which is called a moment of these products.

Also check out http://www.socialresearchmethods.net/kb/statcorr.php



Cutting Board

How alike are two people's tastes in television shows? The following activity will employ the nonparametric, Spearman rank correlation coefficient test to help determine the answer to this question. You will need a friend or a relative to perform this activity.

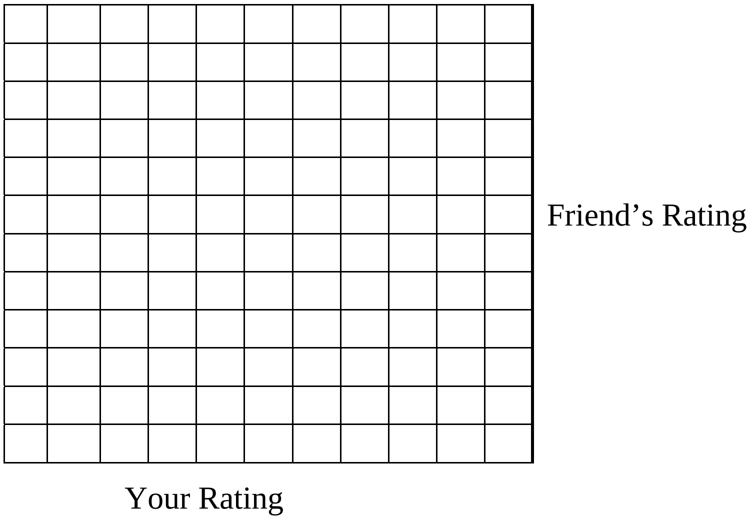
1. In Column I of the chart provided in Step 3, list 10 different TV shows that you and a friend or relative are familiar with. Try to have at least one news show, a

situation comedy, a mystery, a variety show, a talk show, and a drama. Include shows that you like as well as those that you dislike.

2. In Column II, rank the shows that are listed, where 1 is your favorite (the one you would be most inclined to watch) and 10 is your least favorite (the one you would be least inclined to watch).
3. Have your friend or relative do a similar ranking in Column III.

I	II	III IV		V
TV Shows	Your Ratings	F/R Ratings	*d*	d^2
A				
B.				
C.				
D.				
E.				
F				
G.				
H.				
I.				
J.				

4. Use the graph below to plot the ordered pairs consisting of the two rankings. Label the points with the letters corresponding to the shows in the list. If the two rankings were identical, the points would be on a starting line pointing northeast and forming a 45-degree angle with both axes. If you were in total disagreement, then the points would be on a straight line pointing southeast and also form a 45-degree angle with both axes.



5. Although the scattergram you created might give you an impression of how the two ratings match or correlate with each other, it is probably not very definitive. To determine how closely correlated these rankings are, we can use the r statistics

and the Spearman rank correlation coefficient, which we will compute in the steps that follow.

6. Go back to the chart in step 3 and compute d, the difference between the two ratings for each show, and d2, that is, (d)(d). After you have all the d2, add them up.
7. The formula for finding the rank correlation is:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}.$$

8. To do this on your calculator, multiply the sum of your d2 numbers by 6. Divide this product by 990, which is the denominator, (10)(99). Store this number in memory +. Compute 1 minus memory recall. The number in the display is your r number. It should be between -1 and +1.
9. A Spearman table indicates that for your sample size of 10, an r value of .564 or greater would indicate a positive correlation with an alpha of 0.10, or a negative value less than -.564 would indicate a negative correlation with an alpha value of 0.10. The closer r is to 1 or -1, the stronger the relation. An r value close to 0 indicates no particular relation. What can you conclude from this test? Should you and this other person turn on the tube when you are together or would it be better to find a different activity?

Critical Values of Spearman's Rank Correlation Coefficient: $r_{s\ (\text{rho})}$

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$
10	.564	.648	.745	.794

More on Correlational Statistics

Warning: Correlation does not imply CAUSATION!

The purpose of correlational research is to find co-relationships between two or more variables with the hope of better understanding the conditions and events we encounter and with the hope of making predictions about the future. (From the Annals of Chaos Theory: Predictions are usually very difficult—especially if they are about the future; Predictions are like diapers, both need to be changed often and for the same reason!)

As was noted previously, the linear correlation coefficient, *r*, measures the strength of the linear relationship between two paired variables in a sample. If there is a linear correlation, that is, if *r* is large enough between two variables, then regression analysis is used to identify the relationship with the hope of predicting one variable from the other.

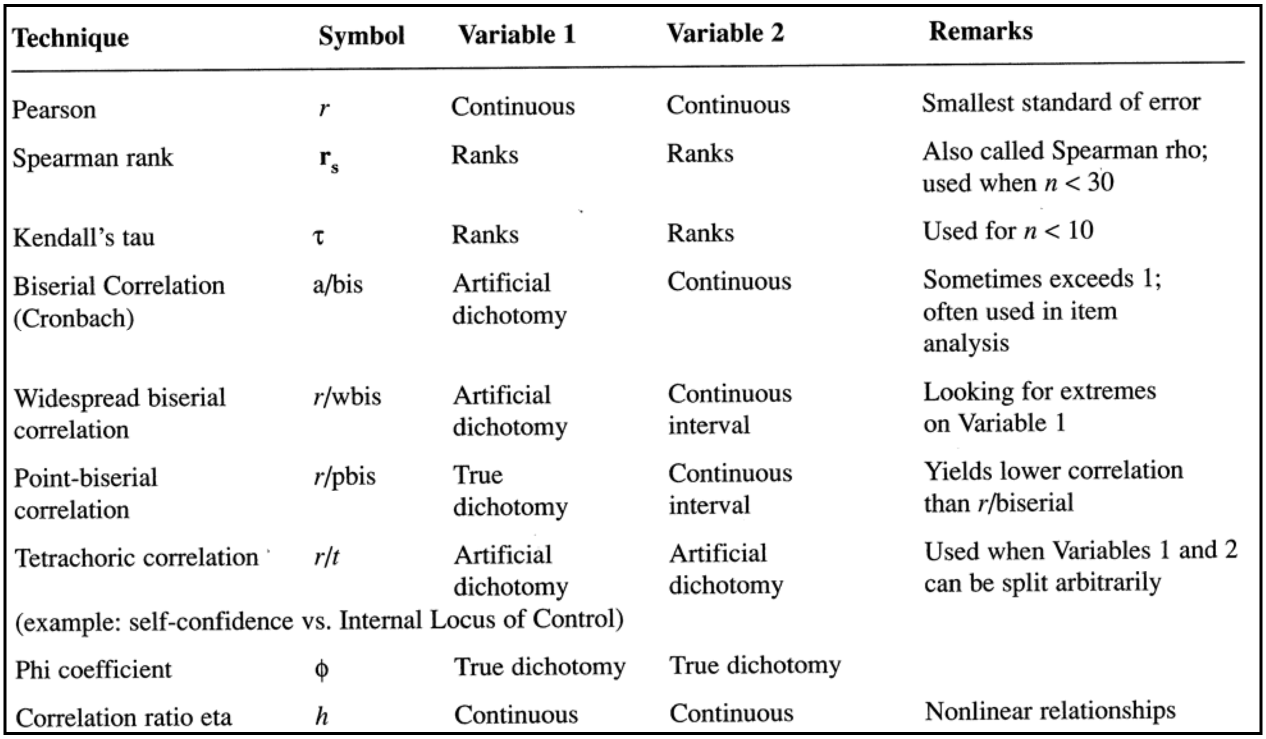
Note: If there is no significant linear correlation, then a regression equation *cannot* be used to make predictions.

A regression equation based on old data is not necessarily valid. The regression equation relating used car prices and ages of cars is no longer usable if it is based on data from the 1960s. Often a scattergram is plotted to get a visual view of the correlation and possible regression equation.

Note: Nonlinear relationships can also be determined, but due to the fact that more complex mathematics is used to describe and interpret data, they are used considerably less often. The following are characteristics of all linear correlational studies:

1. Main research questions are stated as null hypotheses, i.e., no relationship exists between the variables being studied.
2. In simple correlation, there are two measures for each individual in the sample.
3. To apply parametric methods, there must be at least 30 individuals in the study.
4. Can be used to measure the degree of relationships, not simply whether a relationship exists.
5. A perfect positive correlation is 1.00; a perfect negative (inverse) is -1.00.
6. A correlation of 0 indicates no linear relationship exists.
7. If two variables, x and y, are correlated so that r = .5, then we say that (0.5)(2) or 0.25 or 25% of their variation is common, or variable x can predict 25% of the variance in y.

TABLE 2. Types of Correlations



Technique	Symbol	Variable 1	Variable 2	Remarks
Pearson	r	Continuous	Continuous	Smallest standard of error
Spearman rank	$\mathbf{r}_s$	Ranks	Ranks	Also called Spearman rho; used when $n < 30$
Kendall's tau	τ	Ranks	Ranks	Used for $n < 10$
Biserial Correlation (Cronbach)	a/bis	Artificial dichotomy	Continuous	Sometimes exceeds 1; often used in item analysis
Widespread biserial correlation	r/wbis	Artificial dichotomy	Continuous interval	Looking for extremes on Variable 1
Point-biserial correlation	r/pbis	True dichotomy	Continuous interval	Yields lower correlation than r/biserial
Tetrachoric correlation	r/t	Artificial dichotomy	Artificial dichotomy	Used when Variables 1 and 2 can be split arbitrarily
(example: self-confidence vs. Internal Locus of Control)				
Phi coefficient	ϕ	True dichotomy	True dichotomy	
Correlation ratio eta	h	Continuous	Continuous	Nonlinear relationships

Bivariate correlation is when there are only two variables being investigated. These definitions help us determine which statistical test can be used to determine correlation and regression.

Continuous scores: Scores can be measured using a rational scale
Ranked data: Likert-type scales, class rankings
Dichotomy: Participants classified into two categories—Republican versus Democrat
Artificial – Pass/ fail (arbitrary decision); true dichotomy (male/female).

The Pearson product-moment correlation coefficient (that is a mouthful!), or Pearson *r*, is the most common measure of the strength of the linear relationship between two variables. The Spearman rank correlation coefficient, or Spearman *r* (which we performed above), used for ranked data or when you have a sample size less than 30 (*n* < 30), is the second most popular measure of the strength of the linear relationship between two variables. To measure the strength of the linear relationship between test items for reliability purposes, Cronbach alpha is the most efficient method of measuring the internal consistency. Table 2 below can be used to determine what statistical technique is best used with respect to the type of data the researcher collects.

Multivariate Correlational Statistics

If you wish to test a claim that multiple independent variables might be used to make a prediction about a dependent variable, several possible tests can be constructed. Such studies involve *multivariate correlational statistics*.

Discriminant Analysis – This is a form of regression analysis designed for classification. It is used to determine the correlation between two or more predictor variables and a dichotomous criterion variable. The main use of discriminant analysis is to predict group membership (e.g., success/nonsuccess) from a set of predictors. If a set of variables is found that provides satisfactory discrimination, classification equations can be derived, their use checked out through hit/rate tables, and if good, they can be used to classify new participants who were not in the original analysis. In order to use discriminant analysis, the following ingredients, or assumptions (conditions), are needed:

1. At least twice the number of participants as variables in study
2. Groups have the same variance/covariance structures
3. All variables are normally distributed

For more information, check out http://tinyurl.com/33snxtz

Canonical Correlation – This is also a form of regression analysis used with two or more independent variables and two or more dependent variables. It is used to predict a combination of several criterion variables from a combination of several predictor variables. For example, suppose a researcher was interested in the relationship between a student’s conation and school achievement. She or he may wish to use several measures of conation (number of hours spent on homework, receiving help when needed, class participation) and several measures of achievement (grades, cores on achievement tests, teacher evaluation). The two clusters of measurement could be studied with canonical correlations.

Path Analysis – A type of multivariate analysis in which causal relations among several variables are represented by graphs or path diagrams showing how causal influences traveled. It is used to test theories about hypothesized causal links between variables that are correlated. Researchers can calculate direct and indirect effects of independent variables that are not usually done with ordinary multiple regression analysis.

Factor Analysis – Used to reduce a large number of variables to a few factors by combining variables that are moderately or highly correlated with one another. Factor analysis is often used in survey research to see if a long series of questions can be grouped into shorter sets of questions, each of which describes an aspect or factor of the phenomenon being studied.

Differential Analysis – Used to examine correlation between variables among homogeneous subgroups within a sample; can be used to identify moderator variables that improve a measure's predictive validity.

Multiple Linear Regression – Used to determine the correlation between a criterion variable and a combination of two or more predictor variables. The coefficient for any particular predictor variable is an estimate of the effect of that variable while holding constant the effects of the other predictor variables. As in any regression method we need the following conditions to be met: We are investigating linear relationships; for each x value, y is a random variable having a normal distribution. All of the y variables have the same variance; for a given value of x, the distribution of y values has a mean that lies on the regression line.

Note: Results are not seriously affected if departures from normal distributions and equal variances are not too extreme.

The following example illustrates how a researcher might use different multivariate correlational statistics in a research project:

Example: Suppose a researcher has, among other data, scores on three measures for a group of teachers working overseas:

1. Years of experience as a teacher
2. Extent of travel while growing up
3. Tolerance for ambiguity

Research Question: Can these measures (or other factors) predict the degree of adaptation to the overseas culture they are working on?

Discriminant Analysis - Hypothesis 1: The outcome is dichotomous between those who adapted well and those who adapted poorly based on these three measures. Hypothesis 2: Knowing these three factors could be used to predict success.

Multiple Regression - Hypothesis: Some combination of the three predictor measures correlates better with predicting the outcome measure than any one predictor alone.

Canonical Correlation - Hypothesis: Several measures of adaptation could be quantified, i.e., adaptation to food, climate, customs, etc. based on these predictors.

Path Analysis - Hypothesis: Childhood travel experience leads to tolerance for ambiguity and desire for travel as an adult, and this makes it more likely that a teacher will score high on these predictors, which will lead them to seek an overseas teaching experience and adapt well to the experience.

Factor Analysis - Suppose there are five more (a total of eight) adaptive measures that could be determined. All eight measures can be examined to determine whether they cluster into groups such as education, experience, personality traits, etc.

Example: To compute the correlation between gender (male/female) and employment status (employed/unemployed), you could use a phi coefficient. You couldn't use it for age and income, however, because these are not dichotomous variables.

Example: Kendall's tau could be used to compute the correlation between feelings about a new health plan (not in favor/in favor/highly in favor) and health of a patient (unhealthy/healthy/very healthy).
Example: A point biserial correlation can be used when females and males applying for a job report the total number of years of education they have had and we want to know whether there is any correlation between gender and years of education.

Analysis of Covariance

A "yes" on one of the questions below will lead you to a different type of statistical test involving bivariate data. Are you claiming that

___1. Two groups being compared come from the same population and contain similar characteristics?

___2. There is a covariate source of variation that is not controlled for in the design of the experiment, but which does affect the dependent variable?

If you are planning to divide your participants into two groups (perhaps a control group and an experimental group), or if you are planning to use two different treatments on two different groups, then problems in randomization and matching the groups might be a concern.

A statistical process called analysis of covariance (ANCOVA) has been developed to equate the groups on certain relevant variables identified prior to the investigation. Researchers often use pretest mean scores as covariates.

The following guidelines should be used in the ANCOVA:

1. The correlation of the covariate (some variable different than the one you are testing) and the response variable should be statistically and educationally significant.
2. The covariate should have a high reliability.
3. The covariate should be a variable that is measured prior to any portions of the treatments.
4. The conditions of homogeneity of regression should be met. The slopes of the regression lines describing the linear relationships of the criterion variable and the covariate from cell to cell must not be statistically different.
5. All follow-up procedures for significant interaction or post-hoc comparisons should be made with the adjusted cell or adjusted marginal means.