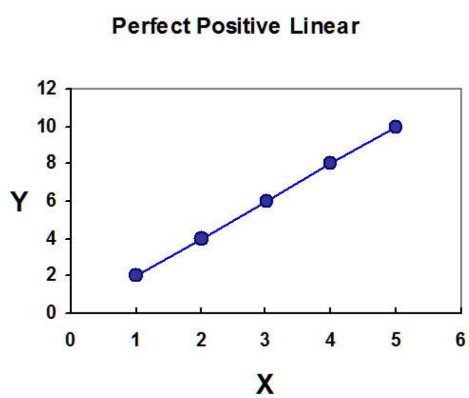
Bivariate Linear Correlation©

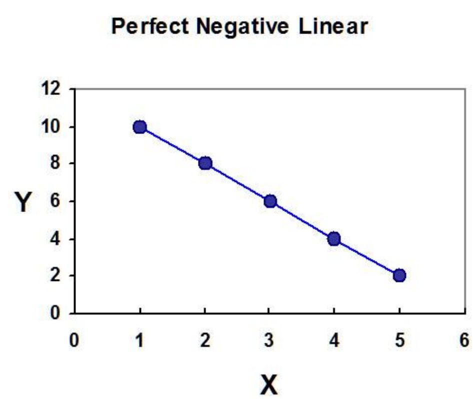
One way to describe the association between two variables is to assume that the value of the one variable is a **linear function** of the value of the other variable. If this relationship is perfect, then it can be described by the slope-intercept equation for a straight line, **Y = a + bX**. Even if the relationship is not perfect, one may be able to describe it as nonperfect linear.

Scatter Plots

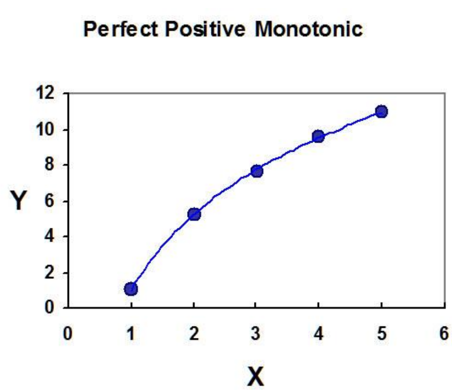
One way to describe a bivariate association is to prepare a **scatter plot**, a plot of all the known paired X,Y values (dots) in Cartesian space. X is traditionally plotted on the horizontal dimension (the abscissa) and Y on the vertical (the ordinate).

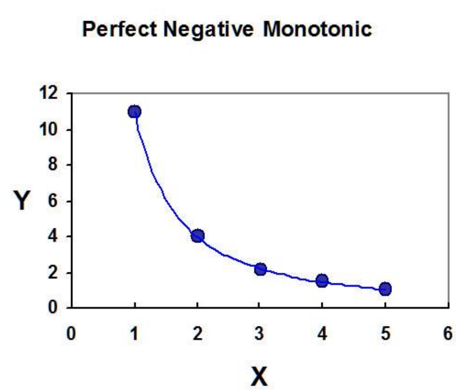
If all the dots fall on a straight line with a positive slope, the relationship is **perfect positive linear**. Every time X goes up one unit, Y goes up *b* units. If all dots fall on a negatively sloped line, the relationship is **perfect negative linear**.





A linear relationship is monotonic (of one direction) – that is, the slope of the line relating Y to X is either always positive or always negative. A monotonic relationship can, however, be nonlinear, if the slope of the line changes magnitude but not direction, as in the plots below:

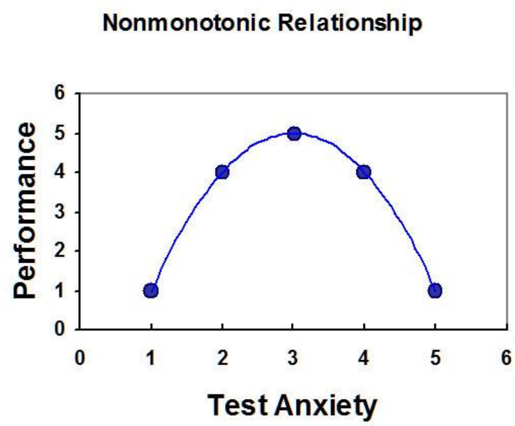




© Copyright 2015, Karl L. Wuensch - All rights reserved.

Notice that with a perfect positive relationship, every time X increases, Y increases as well, and that with a perfect negative relationship, every time X increases, Y decreases.

A nonlinear relationship may, however, not be monotonic, as shown to the right, where we have a quadratic relationship between level of test anxiety and performance on a complex cognitive task. We shall not cover in this course the techniques available to analyze such a relationship (such as polynomial regression).



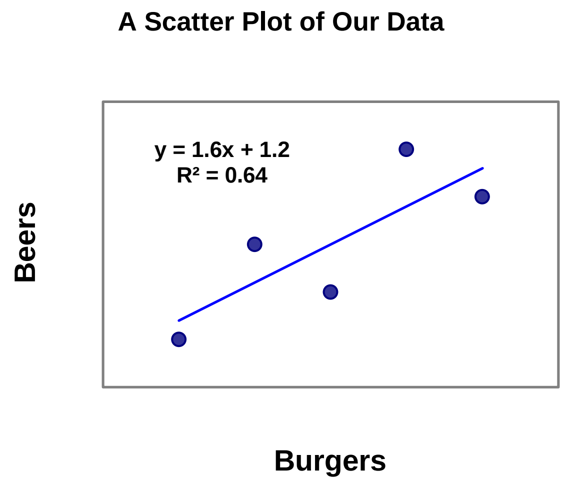
With a perfect nonmonotonic relationship like that pictured above, the linear correlation coefficient (*r*) can be very low (even zero). If you did not see a plot of the data you might mistakenly think that the low *r* means that the variables are not related or only weakly related.

Do note that a linear relationship is a monotonic relationship, but a monotonic relationship is not necessarily a linear relationship. If I tell you that every time X goes up Y also goes up, then you know the relationship is monotonic, but you do not know whether or not it is linear. Please read the document at http://core.ecu.edu/psyc/wuenschk/docs01/If-A-Then-B.doc .

Of course, with real data, the dots are not likely all to fall on any one simple line, but may be approximately described by a simple line. We shall learn how to compute **correlation coefficients** that describe how well a straight line fits the data. If your plot shows that the line that relates X and Y is linear, you should use the **Pearson** correlation coefficient discussed below. If the plot shows that the relationship is monotonic (not a straight line, but a line whose slope is always positive or always negative), you can use the **Spearman** correlation coefficient discussed below. If your plot shows that the relationship is curvilinear but not monotonic, you need advanced techniques (such as polynomial regression) not covered in this class.

Let us imagine that variable X is the number of hamburgers consumed at a cook-out, and variable Y is the number of beers consumed. We wish to measure the relationship between these two variables and develop a regression equation that will enable us to predict how many beers a person will consume given that we know how many burgers that person will consume.

Subject	X (Burgers)	Y (Beers)	X∗Y
1	5	8	40
2	4	10	40
3	3	4	12
4	2	6	12
5	1	2	2
Sum	15	30	106
Mean	3	6	
St. Dev.	1.581	3.162	



$$SS_Y = 4(3.162)^2 = 40$$

Covariance

One way to measure the linear association between two variables is **covariance**, an extension of the unidimensional concept of variance into two dimensions.

The **Sum of Squares Cross Products**,

$$\boxed{SSCP = \sum(X-\bar{X})(Y-\bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{N} = 106 - \frac{15(30)}{5} = 16}.$$

If most of the dots in the scatter plot are in the lower left and upper right quadrants, most of the cross-products will be positive, so *SSCP* will be positive; as X goes up, so does Y. If most are in the upper left and lower right, *SSCP* will be negative; as X goes up, Y goes down.

Just as variance is an average sum of squares, $SS \div N$, or, to estimate population variance from sample data, $SS \div (N\text{-}1)$, **covariance** is an average *SSCP*, $SSCP \div N$. We shall compute covariance as an estimate of that in the population from which our data were randomly sampled. That is, $\boxed{COV = \frac{SSCP}{N-1} = \frac{16}{4} = 4.}$

A major problem with COV is that it is affected not only by degree of linear relationship between X and Y but also by the standard deviations in X and in Y. In fact, the maximum absolute value of COV(X,Y) is the product $\sigma_X\sigma_Y$. Imagine that you and I each measured the height and weight of individuals in our class and then computed the covariance between height and weight. You use inches and pounds, but I use miles and tons. Your numbers would be much larger than mine, so your covariance would be larger than mine, but the strength of the relationship between height and weight should be the same for both of our data sets. We need to standardize the unit of measure of our variables. [Please read this associated document]().

Computing Pearson *r*

We can get a standardized index of the degree of linear association by dividing COV by the two standard deviations, removing the effect of the two univariate standard deviations. This index is called the **Pearson product moment correlation coefficient**, ***r*** for short, and is defined as $\boxed{r = \frac{COV(X,Y)}{s_x s_y} = \frac{4}{1.581(3.162)} = .80}$. Pearson *r* may also be defined as a mean, $r = \sum \frac{Z_x Z_y}{N}$, where the *Z*-scores are computed using population standard deviations, $\sqrt{\frac{SS}{n}}$.

Pearson *r* may also be computed as $r = \frac{SSCP}{\sqrt{SS_x SS_y}} = \frac{16}{\sqrt{4(1.581)^2(4)(3.162)^2}} = \frac{16}{\sqrt{10(40)}} = .8$.

Pearson *r* will vary from –1 to 0 to +1. If r = +1 the relationship is perfect positive, and every pair of X,Y scores has $Z_x = Z_y$. If $r = 0$, there is no linear relationship. If r = –1, the relationship is perfect negative and every pair of X,Y scores has $Z_x = -Z_y$.

If we have X,Y data sampled randomly from some bivariate population of interest, we may wish to test H$_\varnothing$: $\rho = 0$, the null hypothesis that the population correlation coefficient (**rho**) is zero, X and Y are independent of one another, there is no linear association between X and Y. This is quite simply done with Student's *t*:

$$\boxed{t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.8\sqrt{3}}{\sqrt{1-.64}} = 2.309}$$, with $df = N - 2$.

You should remember that we used this formula earlier to demonstrate that the independent samples *t* test is just a special case of a correlation analysis – if one of the variables is dichotomous and the other continuous, computing the (point biserial) *r* and testing its significance is absolutely equivalent to conducting an independent samples *t* test. Keep this in mind when someone tells you that you can make causal inferences from the results of a *t* test but not from the results of a correlation analysis – the two are mathematically identical, so it does not matter which analysis you did. What does matter is how the data were collected. If they were collected in an experimental manner (manipulating the independent variable) with adequate control of extraneous variables, you can make a causal inference. If they were gathered in a nonexperimental manner, you cannot.

Interpreting Pearson *r* and r^2

Pearson *r* is the average number of standard deviations that Y increases for every one standard deviation increase in X. For example, if r = +0.5, then Y increases by one half standard deviation for each one standard deviation increase in X. If r = -0.5, then Y decreases by one half standard deviation for each one standard deviation increase in X.

Pearson r^2 tells you what proportion of the variance in Y is explained by the linear relationship between X and Y. For example, if r^2 = .25, then 25% of the differences in the Y scores are explained by the linear relationship between X and Y.

Putting a Confidence Interval on *R* or R^2

It is a good idea to place a confidence interval around the sample value of r or r^2, but it is tedious to compute by hand. Fortunately, there is now available a free program for constructing such confidence intervals. Please read my document Putting Confidence Intervals on R^2 or R.

For our beer and burger data, a 95% confidence interval for r extends from -.28 to .99. This should be reported in the summary statement.

Reporting Pearson *r*

For our beer and burger data, our APA summary statement could read like this: "**The correlation between my friends' burger consumption and their beer consumption fell short of statistical significance, $r(n = 5) = .8$, $p = .10$. A 95% confidence interval for ρ runs from -.28 to .99.**"

For some strange reason, the value of the computed t is not generally given when reporting a test of the significance of a correlation coefficient. You might want to warn your readers that a Type II error is quite likely here, given the small sample size. Were the result significant, your summary statement might read something like this: **"Among my friends, burger consumption was significantly positively related to beer consumption, ..."**

Assumptions When Testing Hypotheses About *r* or Putting a Confidence Interval on *r*.

There are no assumptions if you are simply using the correlation coefficient to describe the strength of linear association between X and Y in your sample. If, however, you wish to use t or F to test hypothesis about ρ or place a confidence interval about your estimate of ρ, there are assumptions.

Bivariate Normality

It is assumed that the joint distribution of X,Y is bivariate normal. To see what such a distribution look like, try the Java Applet at http://ucs.kuleuven.be/java/version2.0/Applet030.html . Use the controls to change various parameters and rotate the plot in three-dimensional space.

In a bivariate normal distribution the following will be true:

1. The marginal distribution of Y ignoring X will be normal.
2. The marginal distribution of X ignoring Y will be normal.
3. Every conditional distribution of Y|X will be normal.
4. Every conditional distribution of X|Y will be normal.

Homoscedasticity

1. The variance in the conditional distributions of Y|X is constant across values of X.
2. The variance in the conditional distributions of X|Y is constant across values of Y.

Shrunken r^2

Some researchers prefer to report r^2 instead of r; r^2 estimates the proportion of the variance in the Y variable that is explained by its relationship with the X variable, but sample r^2 overestimates the true value of the population ρ^2, especially with small samles. For a relatively unbiased estimate of population ρ^2, requiring no assumptions, compute the shrunken r^2,

$$\boxed{=1-\frac{(1-r^2)(n-1)}{(n-2)}=1-\frac{(1-.64)(4)}{3}=.52}$$

This corrects for the tendency to get overestimates of ρ^2 from small samples. What is the value of r if n = 2? [How well can you fit any two points in Cartesian space with a straight line? See my document "What is R^2 When $N = p + 1$ (and $df = 0$)?" for the answer to this question.]

Spearman rho

When one's data are ranks, one may compute the **Spearman correlation for ranked data**, also called the **Spearman** ρ, which is computed and significance-tested exactly as is Pearson r (if n < 10, find a special table for testing the significance of the Spearman ρ). The Spearman ρ measures the linear association between pairs of ranks. If one's data are not ranks, but e converts the raw data into ranks prior to computing the correlation coefficient, the Spearman measures the degree of **monotonicity** between the original variables. If every time X goes up, Y goes up (the slope of the line relating X to Y is always positive) there is a perfect positive monotonic relationship, but not necessarily a perfect linear relationship (for which the slope would have to be constant). Consider the following data:

X	1.0	1.9	2.0	2.9	3.0	3.1	4.0	4.1	5
Y	10	99	100	999	1,000	1,001	10,000	10,001	100,000

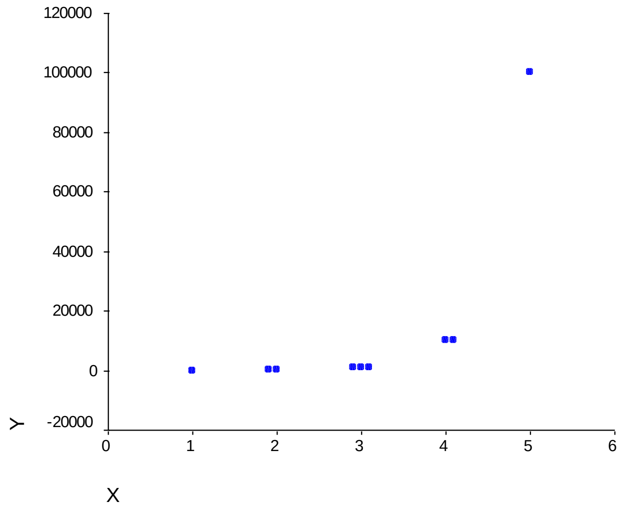
I used SPSS to plot these data and compute the simple Pearson r between X and Y, between X and the base 10 log of Y, and between rank of X and rank of Y (Spearman). Here is the output:

Correlations

		X
Y	Pearson Correlation	.678
	Sig. (2-tailed)	.045
	N	9
LOG10Y	Pearson Correlation	.999
	Sig. (2-tailed)	.000
	N	9

Correlations

			X
Spearman's rho	Y	Correlation Coefficient	1.000
		Sig. (2-tailed)	.
		N	9



As you can see, the relationship between X and Y is perfectly monotonic (and nearly perfect exponential – the correlation between X and the log of Y is almost perfect), so the Spearman coefficient is 1.000. The Pearson (linear) coefficient does not really adequately describe how strongly X and Y are related.

How Do Behavioral Scientists Use Correlation Analyses?

1. to measure the linear **association** between two variables without establishing any cause-effect relationship.

2. as a necessary (and suggestive) but not sufficient condition to establish **causality** [see the online document When Does Correlation Imply Causation?]. If changing X causes Y to change, then X and Y must be correlated (but the correlation is not necessarily linear). X and Y may, however, be correlated without X causing Y. It may be that Y causes X. Maybe increasing Z causes increases in both X and Y, producing a correlation between X and Y with no cause-effect relationship between X and Y. For example, smoking cigarettes is well known to be correlated with health problems in humans, but we cannot do experimental research on the effect of smoking upon humans' health. Experimental research with rats has shown a causal relationship, but we are not rats. One alternative explanation of the correlation between smoking and health problems in humans is that there is a third variable, or constellation of variables (genetic disposition or personality), that is causally related to both smoking and development of health problems. That is, if you have this disposition, it causes you to smoke and it causes you to have health problems, creating a spurious correlation between smoking and health problems – but the disposition that caused the smoking would have caused the health problems whether or not you smoked. No, I do not believe this model, but the data on humans cannot rule it out.

As another example of a **third variable problem**, consider the strike by PATCO, the union of air traffic controllers back during the Reagan years. The union cited statistics that air traffic controllers had much higher than normal incidence of stress-related illnesses (hypertension, heart attacks, drug abuse, suicide, divorce, etc.). They said that this was caused by the stress of the job, and demanded better benefits to deal with the stress, no mandatory overtime, rotation between high stress and low stress job positions, etc. The government crushed the strike (fired all controllers), invoking a third variable explanation of the observed correlation between working in air traffic control and these illnesses. They said that the air traffic controller profession attracted persons of a certain disposition (Type A individuals, who are perfectionists who seem always to be under time pressure), and these individuals would get those illnesses whether they worked in air traffic or not. Accordingly, the government said, the problem was the fault of the individuals, not the job. Maybe the government would prefer that we hire only Type B controllers (folks who take it easy and don't get so upset when they see two blips converging on the radar screen)!

3. to establish an instrument's **reliability** – a reliable instrument is one which will produce about the same measurements when the same objects are measured repeatedly, in which case the scores at one time should be well correlated with the scores at another time (and have equivalent means and variances as well).

4. to establish an instruments (criterion-related) **validity** – a valid instrument is one which measures what it says it measures. One way to establish such validity is to show that there is a strong positive correlation between scores on the instrument and an independent measure of the attribute being measured. For example, the Scholastic Aptitude Test was designed to measure individuals' ability to do well in college. Showing that scores on this test are well correlated with grades in college establishes the test's validity.

5. to do **independent groups *t*-tests**: if the independent variable, X, groups, is coded 0,1 (or any other two numbers) and X is correlated with the dependent variable, Y, a significance-test of the hypothesis that $\rho = 0$ will yield exactly the same *t* and *p* as the traditional pooled-variances independent groups *t*-test. In other words, the independent groups *t*-test is just a special case of correlation analysis, where the X variable is dichotomous. The *r* is called a **point-biserial *r***. It can also be shown that the 2 x 2 Pearson Chi-square test is a special case of *r*. When both X and Y are dichotomous, the *r* is called **phi (ϕ)**.

6. One can measure the correlation between Y and an optimally weighted set of two or more X's. Such a correlation is called a **multiple correlation**. A model with multiple predictors might well predict a criterion variable better than would a model with just a single predictor variable. Consider the research reported by McCammon, Golden, and Wuensch in the *Journal of Research in Science Education*, 1988, *25*, 501-510. Subjects were students in freshman and sophomore level Physics courses (only those courses that were designed for science majors, no general education <football physics> courses). The mission was to develop a model to predict performance in the course. The predictor variables were CT (the Watson-Glaser Critical Thinking Appraisal), PMA (Thurstone's Primary Mental Abilities Test), ARI (the College Entrance Exam Board's Arithmetic Skills Test), ALG (the College Entrance Exam Board's Elementary Algebra Skills Test), and ANX (the Mathematics Anxiety Rating Scale). The criterion variable was subjects' scores on course examinations. Our results indicated that we could predict performance in the physics classes much better with a combination of these predictors than with just any one of them. At Susan McCammon's insistence, I also separately analyzed the data from female and male students. Much to my surprise I found a remarkable sex difference. Among female students every one of the predictors was significantly related to the criterion, among male students none of the predictors was. A posteriori searching of the literature revealed that Anastasi (*Psychological Testing*, 1982) had noted a relatively consistent finding of sex differences in the predictability of academic grades, possibly due to women being more conforming and more accepting of academic standards (better students), so that women put maximal effort into their studies,

whether or not they like the course, and according they work up to their potential. Men, on the other hand, may be more fickle, putting forth maximum effort only if they like the course, thus making it difficult to predict their performance solely from measures of ability.

ANOVA, which we shall cover later, can be shown to be a special case of multiple correlation/regression analysis.

7. One can measure the correlation between an optimally weighted set of Y's and an optimally weighted set of X's. Such an analysis is called **canonical correlation**, and almost all inferential statistics in common use can be shown to be special cases of canonical correlation analysis. As an example of a canonical correlation, consider the research reported by Patel, Long, McCammon, & Wuensch (*Journal of Interpersonal Violence*, 1995, *10*: 354-366, 1994). We had two sets of data on a group of male college students. The one set was personality variables from the MMPI. One of these was the PD (psychopathically deviant) scale, Scale 4, on which high scores are associated with general social maladjustment and hostility. The second was the MF (masculinity/femininity) scale, Scale 5, on which low scores are associated with stereotypical masculinity. The third was the MA (hypomania) scale, Scale 9, on which high scores are associated with overactivity, flight of ideas, low frustration tolerance, narcissism, irritability, restlessness, hostility, and difficulty with controlling impulses. The fourth MMPI variable was Scale K, which is a validity scale on which high scores indicate that the subject is "clinically defensive," attempting to present himself in a favorable light, and low scores indicate that the subject is unusually frank. The second set of variables was a pair of homonegativity variables. One was the IAH (Index of Attitudes Towards Homosexuals), designed to measure affective components of homophobia. The second was the SBS, (Self-Report of Behavior Scale), designed to measure past aggressive behavior towards homosexuals, an instrument specifically developed for this study.

Our results indicated that high scores on the SBS and the IAH were associated with stereotypical masculinity (low Scale 5), frankness (low Scale K), impulsivity (high Scale 9), and general social maladjustment and hostility (high Scale 4). A second relationship found showed that having a low IAH but high SBS (not being homophobic but nevertheless aggressing against gays) was associated with being high on Scales 5 (not being stereotypically masculine) and 9 (impulsivity). This relationship seems to reflect a general (not directed towards homosexuals) aggressiveness – in the words of one of my graduate students, "being an equal opportunity bully."

Factors Which Can Affect the Size of *r*

Range restrictions. If the range of X is restricted, *r* will usually fall (it can rise if X and Y are related in a curvilinear fashion and a linear correlation coefficient has inappropriately been used). This is very important when interpreting criterion-related validity studies, such as one correlating entrance exam scores with grades after entrance.

Extraneous variance. Anything causing variance in Y but not in X will tend to reduce the correlation between X and Y. For example, with a homogeneous set of subjects all run under highly controlled conditions, the *r* between alcohol intake and reaction time might be +0.95, but if subjects were very heterogeneous and testing conditions variable, *r* might be only +0.50. Alcohol might still have just as strong an effect on reaction time, but the effects of many other "extraneous" variables (such as sex, age, health, time of day, day of week, etc.) upon reaction time would dilute the apparent effect of alcohol as measured by *r*.

Interactions. It is also possible that the extraneous variables might "interact" with X in determining Y. That is, X might have one effect on Y if Z = 1 and a different effect if Z = 2. For example, among experienced drinkers (Z = 1), alcohol might affect reaction time less than among novice drinkers (Z = 2). If such an interaction is not taken into account by the statistical analysis (a topic beyond the scope of this course), the *r* will likely be smaller than it otherwise would be.

Power Analysis

Power analysis for *r* is exceptionally simple: $\boxed{\delta = \rho\sqrt{n-1}}$, assuming that *df* are large enough for *t* to be approximately normal. Cohen's benchmarks for effect sizes for *r* are: **.10** is small but not necessarily trivial, **.30** is medium, and **.50** is large (Cohen, J. A Power Primer, *Psychological Bulletin*, 1992, *112*, 155-159).

For our burger-beer data, how much power would we have if the effect size was large in the population, that is, ρ = .50? $\delta = .5\sqrt{4} = 1.00$. From our power table, using the traditional .05 criterion of significance, we then see that power is 17%. As stated earlier, a Type II error is quite likely here. How many subjects would we need to have 95% power to detect even a small effect? Lots: $n = \left(\frac{\delta}{\rho}\right)^2 + 1 = 1297$. That is a lot of burgers and beer!

Copyright 2015, Karl L. Wuensch - All rights reserved.