Biostatistics: *Correlations*

One of the most common errors we find in the press is the confusion between *correlation* and *causation* in scientific and health-related studies. In theory, these are easy to distinguish — an action or occurrence can *cause* another (such as smoking causes lung cancer), or it can *correlate* with another (such as smoking is correlated with alcoholism). If one action causes another, then they are most certainly correlated. But just because two things occur together does not mean that one caused the other, even if it seems to make sense.

One way to get a general idea about whether or not two variables are related is to plot them on a "scatterplot". If the dots on the scatterplot tend to go from the lower left to the upper right it means that as one variable goes up the other variable tends to go up also. This is a called a "direct (or positive) relationship." On the other hand, if the dots on the scatterplot tend to go from the upper left corner to the lower right corner of the scatterplot, it means that as values on one variable go up values on the other variable go down. This is called an "indirect (or negative) relationship."

Correlations between X and Y

Weak Positive
Correlation

Strong Positive
Correlation

No
Correlation

Weak Negative
Correlation

Strong Negative
Correlation

No linear
Correlation

Spurious Correlations:
http://www.huffingtonpost.com/2014/05/15/spurious-correlations-graphs-make-no-sense-video_n_5325835.html

Example 1: ____

Is there a direct or indirect correlation?

Is it fair to conclude there is a causal relationship?

Example 2: ____

Is there a direct or indirect correlation?

Is it fair to conclude there is a causal relationship?

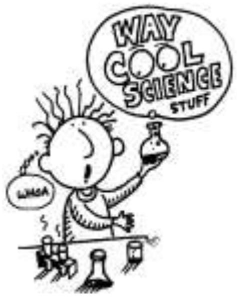
Example 3: ____

Is there a direct or indirect correlation?

Is it fair to conclude there is a causal relationship?

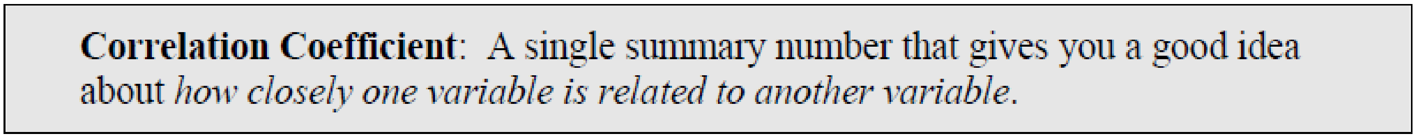
How can one best determine if there is a causal relationship between two variables?

The most effective way of doing this is through a controlled study. In a controlled study, two groups of people who are comparable in almost every way are given two different sets of experiences (such one group watching soap operas and the other game shows), and the outcome is compared. If the two groups have substantially different outcomes, then the different experiences may have caused the different outcome.



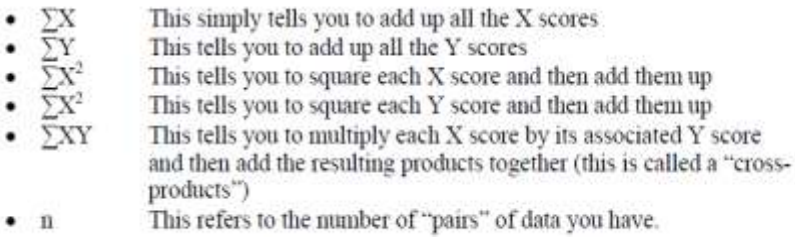
Biostatistics: *Correlation Coefficient (r)*

A *really* smart guy named Karl Pearson figured out how to calculate a summary number that allows you to answer the question "How strong is the relationship of a correlation?" In honor of his genius, the statistic was named after him. It is called Pearson's Correlation Coefficient (r).



Correlation Coefficient: A single summary number that gives you a good idea about *how closely one variable is related to another variable.*

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left[\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n_x}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n_y}\right)\right]}}$$



- $\sum X$ This simply tells you to add up all the X scores
- $\sum Y$ This tells you to add up all the Y scores
- $\sum X^2$ This tells you to square each X score and then add them up
- $\sum X^2$ This tells you to square each Y score and then add them up
- $\sum XY$ This tells you to multiply each X score by its associated Y score and then add the resulting products together (this is called a "cross-products")
- n This refers to the number of "pairs" of data you have.

Example of a way to set up data to make sure you don't make mistakes when using the computational formula to calculate Pearson's r

X	X^2	Y	Y^2	XY
5		45		
15		32		
18		37		
20		33		
25		24		
25		29		
30		26		
34		22		
38		24		
50		15		
$\sum X=$	$\sum X^2=$	$\sum Y=$	$\sum Y^2=$	$\sum XY=$

- Step 1: calculate and fill in the X^2 and Y^2 values

- Step 2: multiply each X score by its paired Y score which will give you the cross-products of X and Y.

- Step 3: fill in the last row of the table which contains all of you "Sum Of" statements. In other words, just add up all of the X scores to get the ΣX, all of the X^2 scores to get the ΣX^2 and etc.

- Step 4: Enter the numbers you have calculated in the spaces where they should go in the formula.

- Step 5: Multiply the (ΣX)(ΣY) in the numerator (the top part of the formula) and do the squaring to $(\Sigma X)^2$ and $(\Sigma Y)^2$ in the denominator (the bottom part of the formula).

- Step 6: Do the division by n parts in the formula.

- Step 7: Do the subtraction parts of the formula

.

- Step 8: Multiply the numbers in the denominator.

- Step 9: Take the square root of the denominator.

- Step 10: Take the last step and divide the numerator by the denominator and you will get the Correlation Coefficient!

What Good Is A Correlation Coefficient?
As can see above, we just did a whole lot of calculating just to end up with a single number. How ridiculous is that? Seems kind of like a waste of time, huh? Well, guess again! It is actually very cool! ("Yeah, right!" you say, but let me explain.)

Important Things Correlation Coefficients Tell You

1. ____*:* If your correlation coefficient is a negative number you can tell, just by looking at it, that there is an indirect, negative relationship between the two variables. As you may recall, a negative relationship means that as values on one variable increase (go up) the values on the other variable tend to decrease (go down) in a predictable manner. If your correlation coefficient is a positive number, then you know that you have a direct, positive relationship. This means that as one variable increases (or decreases) the values of the other variable tend to go in the same direction. If one increases, so does the other. If one decreases, so does the other in a predictable manner.

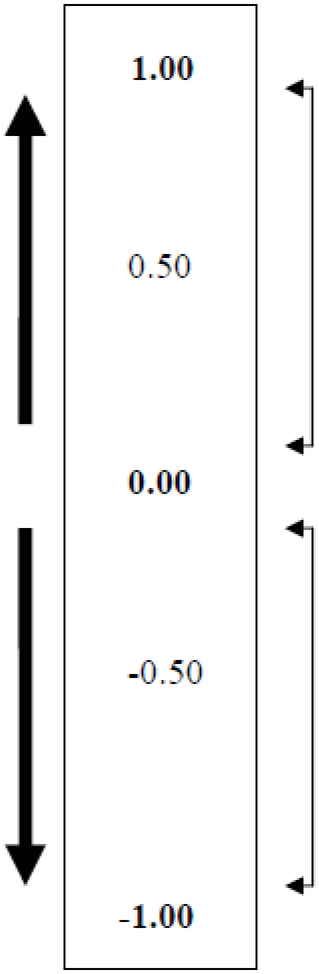
2. ____*:*
   a. A correlation coefficient of -1.00 tells you that there is a perfect negative relationship between the two variables. This means that as values on one variable increase there is a perfectly predictable decrease in values on the other variable. In other words, as one variable goes up, the other goes in the opposite direction (it goes down).
   b. A correlation coefficient of +1.00 tells you that there is a perfect positive relationship between the two variables. This means that as values on one variable increase there is a perfectly predictable increase in values on the other variable. In other words, as one variable goes up so does the other.
   c. A correlation coefficient of 0.00 tells you that there is a zero correlation, or no relationship, between the two variables. In other words, as one variable changes (goes up or down) you can't really say anything about what happens to the other variable.

3. ____

a. Most correlation coefficients (assuming there really is a relationship between the two variables you are examining) tend to be somewhat lower than plus or minus 1.00 (meaning that they are not perfect relationships) but are somewhat above 0.00. Remember that a correlation coefficient of 0.00 means that there is no relationship between your two variables based on the data you are looking at.

b. The closer a correlation coefficient is to 0.00, the weaker the relationship is and the less able you are to tell exactly what happens to one variable based on knowledge of the other variable. The closer a correlation coefficient approaches plus or minus 1.00 the stronger the relationship is and the more accurately you are able to predict what happens to one variable based on the knowledge you have of the other variable.

Values of "r"



Making Statistical Inferences from Pearson's r.

How do you determine whether or not your correlation is simply a chance occurrence or if it really is true of the population? You will need three things in order to determine whether you can infer that the relationship you found in your sample also is true (in other words, "is generalizable" in the larger population:

1. The Correlation Coefficient that you calculated

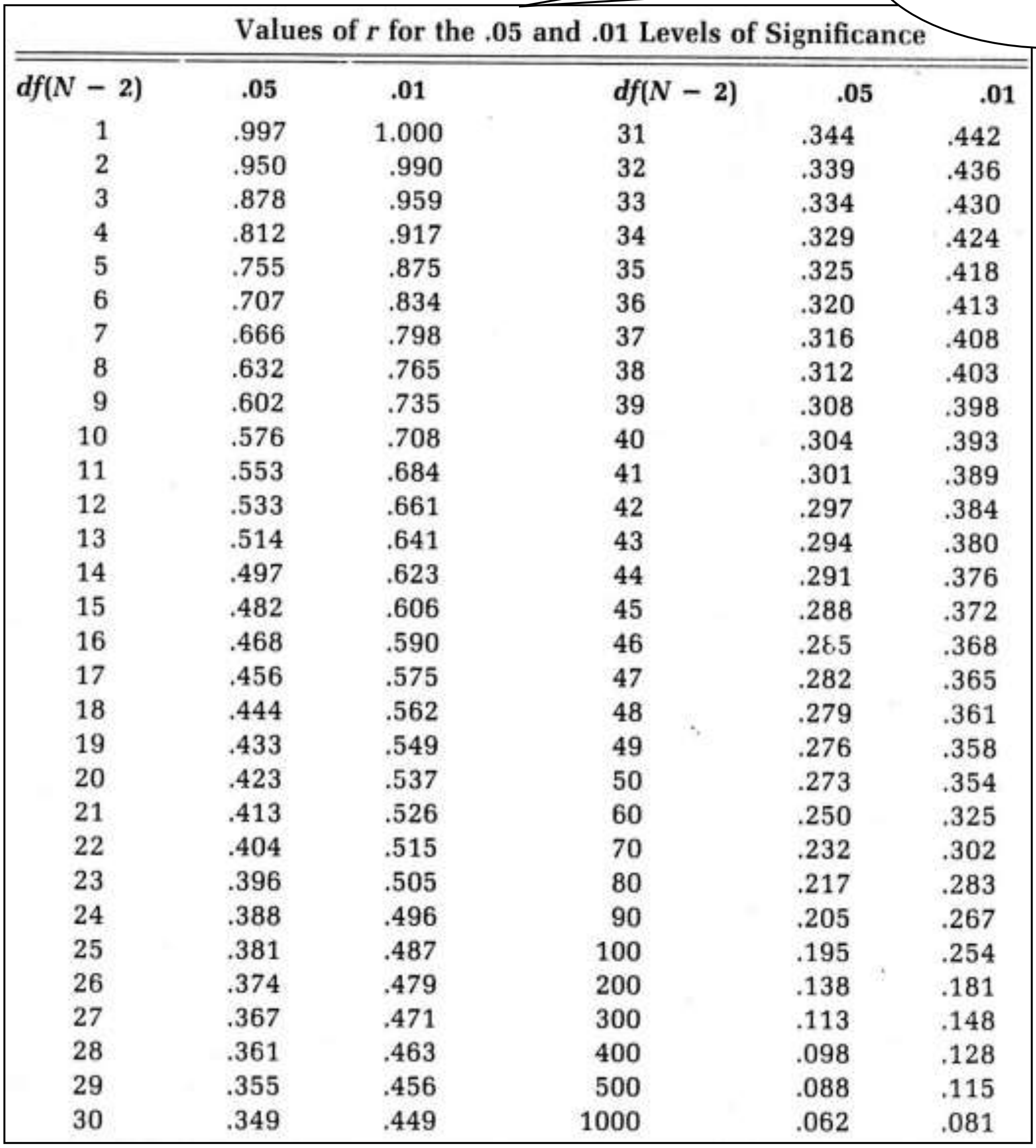
2. Something called the **"degrees of freedom" which is simply the number of pairs of data in your sample minus 2.**

DF =

3. A table of “Critical Values” of the correlation coefficient.

Pearson’s R Critical Values

We’ll always use the 0.05 level of significance



Values of *r* for the .05 and .01 Levels of Significance

df(*N* − 2)	.05	.01	*df*(*N* − 2)	.05	.01
1	.997	1.000	31	.344	.442
2	.950	.990	32	.339	.436
3	.878	.959	33	.334	.430
4	.812	.917	34	.329	.424
5	.755	.875	35	.325	.418
6	.707	.834	36	.320	.413
7	.666	.798	37	.316	.408
8	.632	.765	38	.312	.403
9	.602	.735	39	.308	.398
10	.576	.708	40	.304	.393
11	.553	.684	41	.301	.389
12	.533	.661	42	.297	.384
13	.514	.641	43	.294	.380
14	.497	.623	44	.291	.376
15	.482	.606	45	.288	.372
16	.468	.590	46	.285	.368
17	.456	.575	47	.282	.365
18	.444	.562	48	.279	.361
19	.433	.549	49	.276	.358
20	.423	.537	50	.273	.354
21	.413	.526	60	.250	.325
22	.404	.515	70	.232	.302
23	.396	.505	80	.217	.283
24	.388	.496	90	.205	.267
25	.381	.487	100	.195	.254
26	.374	.479	200	.138	.181
27	.367	.471	300	.113	.148
28	.361	.463	400	.098	.128
29	.355	.456	500	.088	.115
30	.349	.449	1000	.062	.081

The first thing you need to do is look down the degrees of freedom column until you see the row with the number of degrees of freedom that matches your sample degrees of freedom. Look across to the number listed under .05. This number is called “the critical value of r”.

Critical r =

SO WHAT?

If your calculated r value is ______________________________ the number in the table, you conclude that the correlation is

________ ____

____

____

If your calculated r value is _____________________ than the number in the table, you conclude that the correlation is

________ ____

____

____