SFBS Course Notes Part 7: Correlation

Relationships between variables

Psychologists are often interested in describing the relationship between two variables. For instance, personality psychologists may be interesting in how specific tests relate to a person's behavior or feelings; developmental psychologists may be interested in how certain mothering behavior (e.g., responsiveness to an infant) are associated with child behavior (e.g., the infant's attentiveness to the mother); or neuropsychologists may be interested in how specific brain characteristics (measured with an imaging system) relate to psychological functions.

To perform or understand these kinds of studies, we need to add a few more methods to our statistical repertoire—methods that allow us to visualize **bivariate distributions**, and to quantify the relationship between the two variables in these distributions. First, we use **scatter plots** to visualize bivariate distributions. Second, we use ***Pearson r***, often referred to as simply "the correlation coefficient", to measure the strength and direction of linear relationships between variables.

Visualizing Bivariate Distributions: Scatter Plots

All methods we've used so far to characterize distributions dealt with *univariate distributions* – distributions of a single variable. The statistics we've calculated to measure central tendency and variability have been for the distribution of values of one variable. The frequency histograms we've worked with also deal only with one variable—values of a variable are plotted on the X axis, and the number (frequency) of cases with values in a given class interval are plotted on the Y axis. To visualize a *bivariate distribution*—the distribution of two variables per case—we use a *scatter plot*.

Consider the values on the right for two variables measured across ten states of the U.S.: The poverty rate (percentage of families living below the poverty level), and the murder rate (number of murders per 100,000 people in the population in a year). We'll label these variables "*X*" and "*Y*".

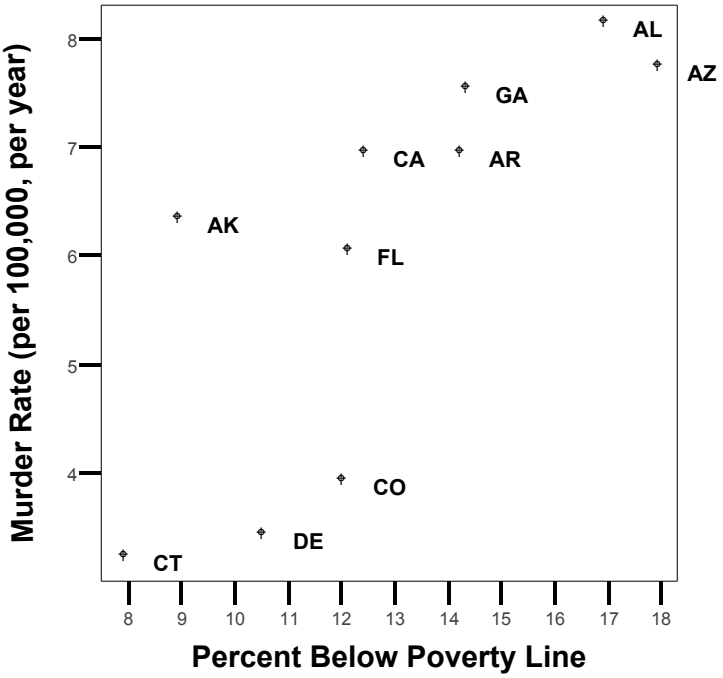
State	Poverty rate (*X*)	Murder rate (*Y*)
AK	8.9	6.3
AL	16.9	8.1
AR	14.2	6.9
AZ	17.9	7.7
CA	12.4	6.9
CO	12	3.9
CT	7.9	3.2
DE	10.5	3.4
FL	12.1	6
GA	14.3	7.5

You might be able to look at this table and tell that a **positive relationship** seems to exist between the two measured variables—states with higher poverty rates seem to have higher murder rates. But to see this more clearly, we can graph these values in a **scatter plot** (next page).

This type of graph uses a single **point** (or "dot") for each case (each state, in this example). The point is positioned at the case's value of variable *X* on the *X* axis, and at its value of variable *Y* on the *Y* axis. So the two-dimensional location of each point tells us the value of *both* variables for each state. *(Compare the two measurements for a state listed on the previous page to the point's X and Y positions on the scatter plot to the right.)*

Again, we can see in this scatter plot that there appears to be a **positive relationship** between the two variables. There are a lot of points with high values on both variables, middle values on both, or low values on both. There aren't many states with high values on one variable and low values on the other.

***Scatter Plot** of Poverty Rate and Murder Rate*



Linear Relationships

A *linear relationship* is a relationship between variables that can best be characterized as a straight line. (In other words, the points on a scatter plot are clustered around a straight line.) **A linear relationship can be either *positive* or *negative*; and either *perfect* or *imperfect*.**

The distribution above shows a *positive linear relationship*—a tendency for the values of Y to increase as the values of X increase. (If, instead, the values of Y decreased as the values of X increased—if the points were arranged more closely to a line from the top-left to bottom-right of the scatter plot—this would instead be a *negative* relationship.)

The linear relationship between poverty rates and murder rates is also *imperfect*. A *perfect* linear relationship is one where all of the points in a scatter plot would fall exactly on a straight line. This can only happen in the real world if one variable is a linear transformation of the other—for example, the same temperatures in Fahrenheit and Celsius scales.

The greater the *strength* (also called "degree" or "magnitude") of a linear relationship is between X and Y, the more closely the points seem to fit a straight, diagonal line. Weaker relationships would show more "scatter" of the points around the line of best fit.

Linear Transformations and the Equation of a Line

Here, we have a data table of two variables—the average temperature of each state in degrees Celsius, and in degrees Fahrenheit. As we saw in Handout Set 3 (p. 4-5), degrees Fahrenheit can be found using a *linear transformation* of degrees Celsius:

$$\text{degrees F} = (1.8)(\text{degrees C}) + 32$$

Or in this example, using the variable labels in the table:

$$Y = (1.8)X + 32$$

More generally, the formula for a linear transformation from some variable X to some variable Y is:

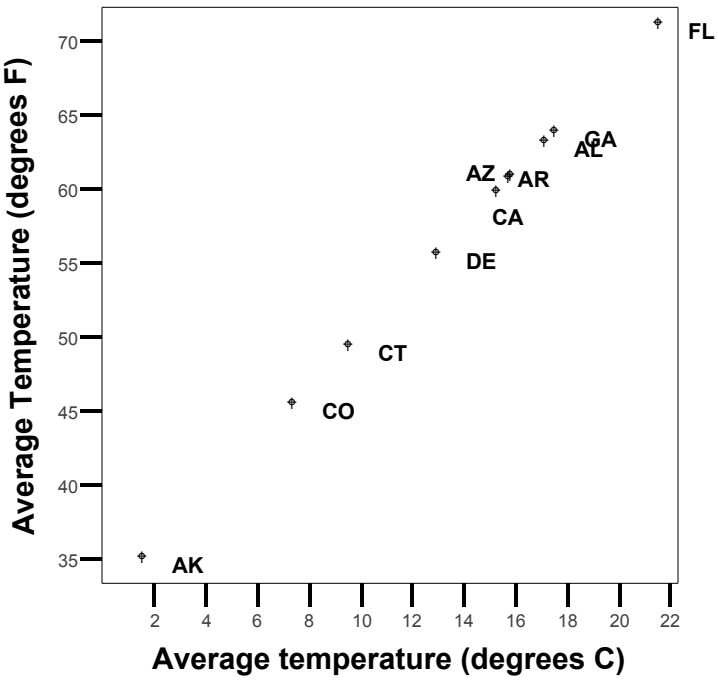
$$Y = bX + a$$

State	Degrees C (*X*)	Degrees F (*Y*)
AK	**1.51**	**34.718**
AL	17.09	62.762
AR	15.79	60.422
AZ	15.73	60.314
CA	15.22	59.396
CO	7.31	45.158
CT	9.47	49.046
DE	12.93	55.274
FL	21.52	70.736
GA	17.51	63.518

This is also called the **equation of a straight line**. The equation uses two constants (***a*** & ***b***) that affect how scores are transformed between X and Y. Once we know these two values, we can figure out what the value of Y is, for any given value of X (or vice versa). We know from the previous equation that *a* in this example is set to 32, and *b* in this example is set to 1.8. So we can get values of Y by plugging in *b, a,* and some value of X we want to convert.

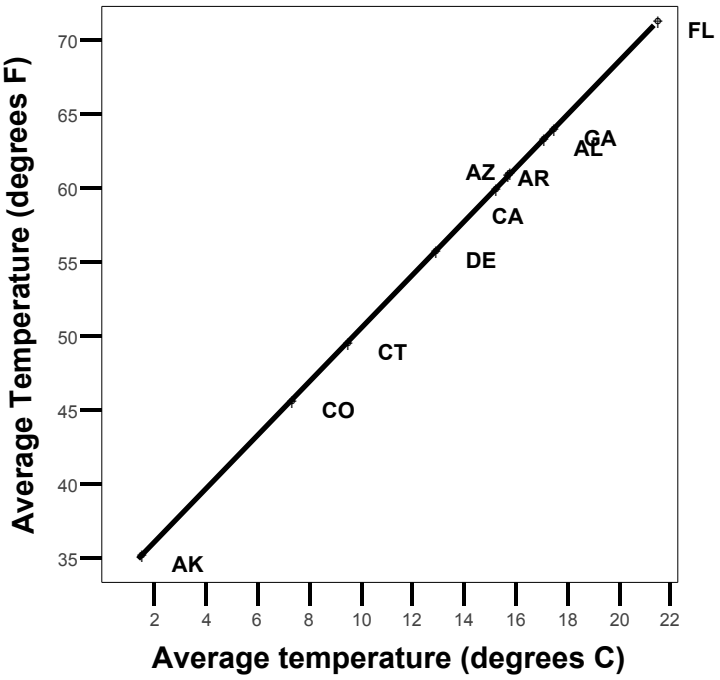
E.g., Alaska: $Y = bX + \text{a} = (1.8)X + 32 = (1.8)(\mathbf{1.51}) + 32 = \mathbf{34.718}$

If we convert all of the values of X into values of Y, we can produce the 2-variable table above. On the right is a scatter plot of these two variables for the 10 states—they fall perfectly onto a straight line, and we can therefore say there is a *perfect linear relationship* between the two.



*The Meaning of Constants **a** and **b** in the Equation*

To understand the meaning of the constants in the equation of a line, let's eliminate the data points and just look at the line they fall on.



a **is referred to as the *Y intercept*.**

This is because the value of *a* is what we would get for *Y*, if we plugged (*X* = 0) into the equation of a line. In other words, it's the place where the line "intercepts" the *Y* axis, because we draw the *Y* axis vertically at (*X* = 0). Looking at this figure, we can see that when *X* = 0, the value of *Y* should be about 30. In fact, we know ***a*** = 32 (in other words, when *X* = 0, *Y* = 32) because we've seen the equation used to produce this line (on the previous page).

b **is referred to as the *slope*.**

Because each value of *X* is multiplied by *b* as we solve for *Y* (see equation on the previous page), the *slope (b)* tells us how much Y changes for any given change in X along the line. (In other words—how far the line goes up or down as we move from left to right along the line.) For *any two points* on the line:

$$b = \frac{\text{change in } Y}{\text{change in } X} = \frac{(Y \text{ for point 2}) - (Y \text{ for point 1})}{(X \text{ for point 2}) - (X \text{ for point 1})}$$

Although we already know the slope for the temperatures is 1.8 (see equation on previous page), we could also find it by plugging in any two points on the line. Using Alaska ("1") and Alabama ("2"):

$$b = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{62.762 - 34.718}{17.09 - 1.51} = \frac{28.044}{15.58} = 1.8$$

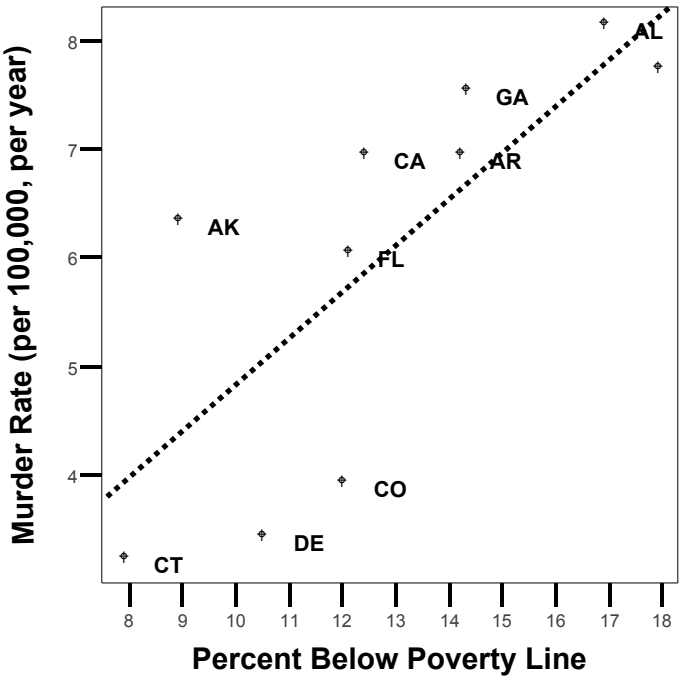
If *b* is negative, then Y values *decrease* as X values increase. So the sign of *b* (i.e., whether it is positive or negative) tells us the *direction of the relationship* – positive or negative. *b* also tells us how "steep" or "shallow" the line is, because it determines how quickly the values of *Y* change as we go from left to right. The uses of this equation for predictive purposes will be discussed in more detail when we cover the topic of linear *regression*.

Direction and Strength of a Linear Relationship

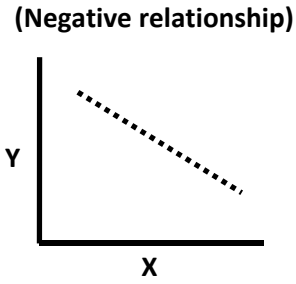
We've seen that a *perfect linear relationship* produces points that all fall exactly on a straight line in a scatter plot. An *imperfect linear relationship* produces points that are arranged roughly in a straight line, perhaps very close to the line, or perhaps clustered very loosely around it.

Coming back to our cheerful "poverty & murder" example, we can take a guess at about what this line would look like.

(We can't actually calculate exactly where this "best fit" line is yet—that process is part of *linear regression*.)



We can see that we have a *positive* relationship above, because the line goes from low values of both variables to high values of both variables. If this were a *negative* relationship, the "line of best fit" for the bivariate distribution would look more like the illustration on the right: (Note: this type of relationship is also sometimes referred to as an *inverse* relationship. Positive relationships are sometimes referred to as *direct* relationships.)

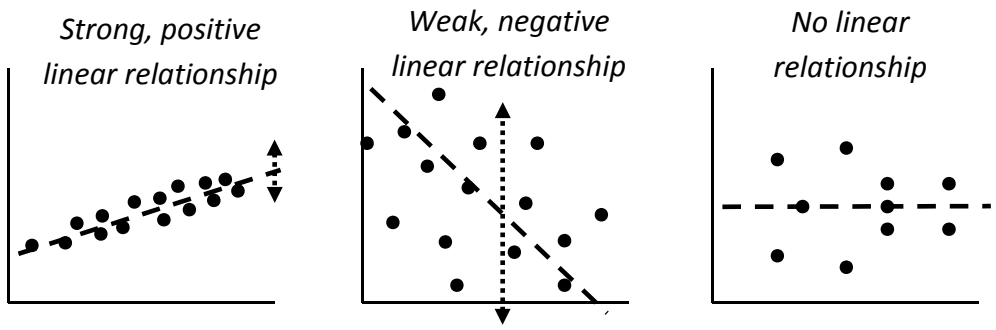


The most important things a researcher needs to know about a relationship between two variables are the relationship's *strength* (also referred to as the *degree* of the relationship) and *direction*.

Direction (positive or negative) can be visualized as the direction that the "line of best fit" moves in as you follow it from left to right—either up (positive) or down (negative).
Strength (or degree) can be visualized as the extent to which the points cluster tightly around the line, or loosely, or don't cluster around a line at all.

We can measure both direction and degree with the correlation coefficient: **Pearson *r***.

Examples:



Pearson r

Imperfect linear relationships can differ in how imperfect they are. On the one hand, some relationships are nearly perfect, with the points in a scatter plot almost falling on a single line. On the other hand, they can be quite imperfect, with some relationship present, but weak. To measure imperfect relationships, we use a statistic called *the Pearson product-moment correlation coefficient*, or **Pearson *r***, or just ***r***.

r can range from -1.0 to +1.0. The *sign* of *r* (i.e., whether *r* is a positive or negative number) indicates the direction of relationship (a negative or positive relationship). The magnitude of r indicates the strength of relationship. The closer *r* is to 1 or -1, the stronger the relationship. The closer *r* is to zero, the weaker the relationship.

The formula for Pearson *r* is:

$$r = \frac{\sum z_X z_Y}{N - 1}$$

…where $\boldsymbol{z_X}$ and $\boldsymbol{z_Y}$ (think "***z* of *X***," "***z* of *Y***") are z-transformed versions of scores for the *X* variable and the *Y* variable, respectively. If we're familiar with calculating *z* scores *(see Handout Set 3, p. 6)*, and with summations using two variables *(see Handout Set 1, p. 9)*, we can calculate *r.*

r is a simple but incredibly useful statistic that summarizes both **strength** and **direction** of a linear relationship between two variables. The idea is that we boil the relationship down to a number (*r*) that is *higher if the z scores are very similar* for each pair of X and Y values, and that is *lower if the scores are dissimilar*.

Put another way, the value of *r* is higher if each case's placement in the distribution of one of the measured variables (*X*) is similar to its placement in the distribution of the other measured variable (*Y*). The *r* formula evaluates this by simply multiplying each pair of *z* scores together.

To understand how the Pearson *r* works, consider the following data:

X	***Y***	z_X	z_Y	$z_X z_Y$
2.25	.75	-1.3363	-1.3363	1.7857
3.00	1.00	-.8018	-.8018	.6429
3.75	1.25	-.2673	-.2673	.0714
4.50	1.50	.2673	.2673	.0714
5.25	1.75	.8018	.8018	.6429
6.00	2.00	1.3363	1.3363	1.7857
				Sum = **5.0000**

Here we're given a set of *X* and *Y* values. We solve the summation $\sum z_X z_Y$ above by:

(1) Finding the **means** and **standard deviations** of each of the two variables. (Calculations not shown—the values would come to $\bar{X} = 4.125,\ s_X = 1.4031,\ \bar{Y} = 1.375,\ s_Y = .4677$.)

(2) Using those values to produce **z-transformed scores** for each value of *X* and *Y* in the lists.

(3) Finally, finding $\sum z_X z_Y$ by multiplying z_X and z_Y for each row, and summing across the rows. We now have the top half of the formula: the **sum of the values of $z_X z_Y$**. In this case, the sum is **5**.

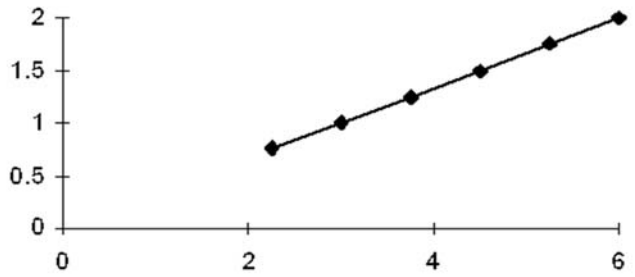
Next, find the value of ***N*** to plug into the formula. It's important to note that *N* is the number of cases—pairs of X & Y values—in this example, **6**. It is NOT the total number of measurements (which would be twice as many, as there are two variables measured per case).

Plugging in what we've found:

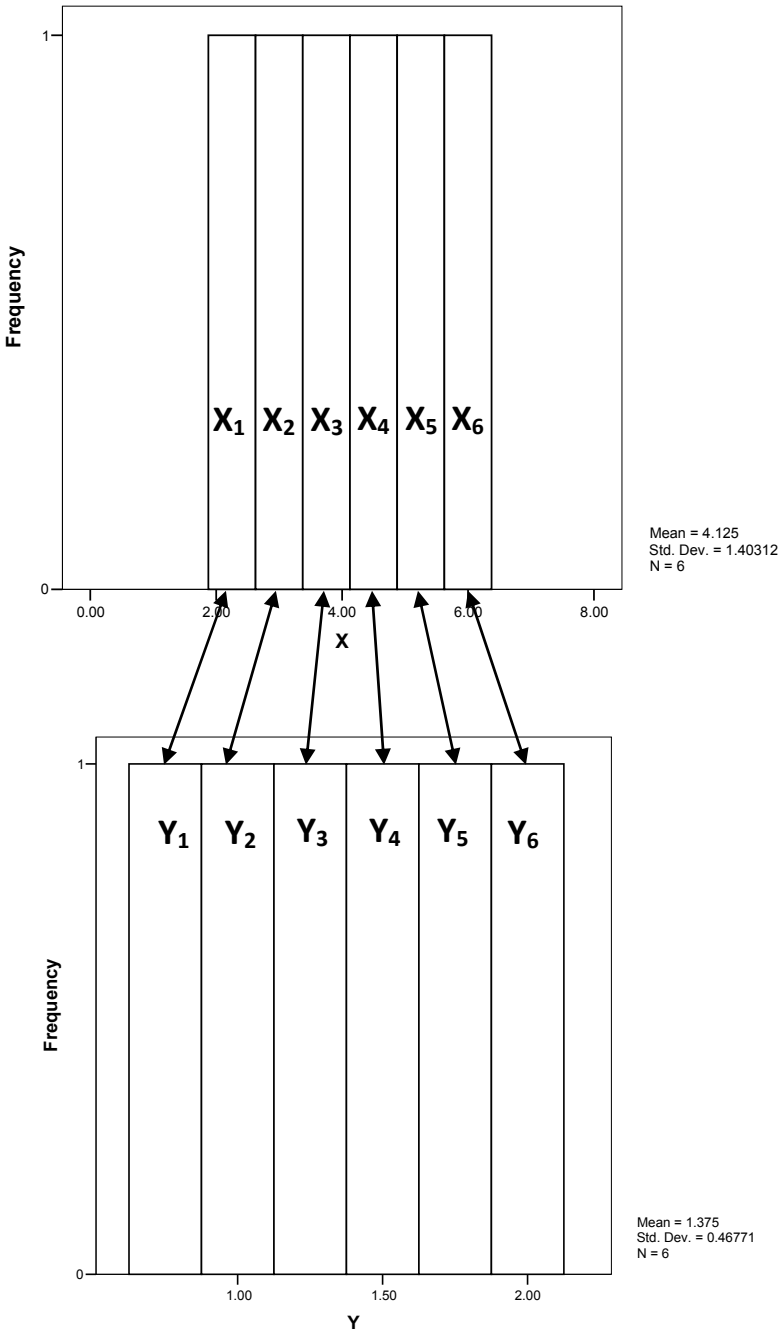
$$r = \frac{\sum z_X z_Y}{N - 1} = \frac{5}{6 - 1} = 1$$

The value of *r* we've found, to measure the relationship between these two variables, is 1. Because r only ranges from -1 to 1, this is the greatest possible value of *r*. This means that we have the strongest positive relationship between two variables we can possibly have. This value will be closer to zero if the scores are more weakly related. If *r* is close to (-1.0), the relationship is very *strong*, but in the *negative* direction. So *r* gives us two pieces of information—the *magnitude* (AKA strength or degree) and the *direction* of the relationship.

What would this look like on a scatter plot? If *r* comes out to exactly (–1) or exactly 1, we know it's a *perfect linear relationship*. And the scatter plot does indeed look like a straight line, with a positive slope (*Y* increases as *X* increases):



Another way to think about two *perfectly related* variables is to notice that each case occupies precisely the same relative position in the distributions of both *X* and *Y*. That is, in a positive, perfect linear relationship, the *z* score for *X* will be the same as the z score for *Y* for every case (which they were in the above example). Again, *r* tells us how similar the pairs of scores are, within their own distributions—i.e., how similar, on average, the *z* scores are between individuals' values of *X* and *Y*. In this example, the paired scores' places within their distributions are identical, so the value of *r* indicates a perfect, positive relationship (*r* = 1):



Additional example of calculation of Pearson r

Steps for calculating the formula for r: $\Sigma z_X z_Y / (N - 1)$

(1) Find the *mean* and *standard deviation* of both of the variables.

(2) Standardize (convert to z scores) each value of the first variable (within that distribution!)

(3) Standardize (convert to z scores) each value of the second variable (within that distribution!)

(4) Find $z_X z_Y$ within each case (each data table row) by multiplying the two *z* scores together.

(5) Find $\Sigma z_X z_Y$: Sum the result of step 4 across all cases.

(5) Divide this sum by ($N - 1$). Remember that *N* is the number of cases ("rows"), not the total number of measurements! *(e.g., below, N = 6, not 12.)*

Side note: The book demonstrates both this method and another method (which it refers to as the "computational equation") for calculating r. Because this method is more conceptually enlightening, it is the only one I'm demonstrating. You will not be required to know or use the other method.

Example data:

X	***Y***	→	z_x	z_y	$z_x z_y$
66	72		0	1.1859	0
64	68		-0.9535	-0.3953	0.3769
66	70		0	0.3953	0
65	68		-0.4767	-0.3953	0.1884
70	71		1.9069	0.7906	**1.5076**
65	65		-0.4767	-1.5811	**0.7537**

Note: When calculating r from raw data, means and standard deviations for X and Y would be required as the first step to calculating z scores, where each $z_X = \frac{X - \bar{X}}{s_X}$ *and each* $z_Y = \frac{Y - \bar{Y}}{s_Y}$*...*

For X: mean = 66, standard deviation = 2.10.
For Y: mean = 69, standard deviation = 2.53.

(Try completing this—compute r by finding the sum of $z_x z_y$ and dividing it by (N – 1). As always, be sure take your intermediate values out to 4 decimal places, or your final answer will often be greatly thrown off by rounding error! By this procedure, using the rounded standard deviation values above, you should find the z scores shown in the table, and a final value of r = 0.5653.)

Looking at the values in each row above, we can get a clearer understanding of what *r* is telling us. The bottom two cases increase the final value of *r* substantially, because their $z_x z_y$ values are high:

- The 5th case's $z_x z_y$ value is high because its values of *X* and *Y* are both on the high ends of the *X* and *Y* distributions, which means there are **positive *z* scores** for both values.
- The 6th case's $z_x z_y$ value is high because its values of *X* and *Y* are both on the low ends of the *X* and *Y* distributions, which means there are **negative *z* scores** for both values.

Interpretation of Correlations

*When **r** Is an Inappropriate Measure of a Relationship*

r is the most frequently used correlation coefficient, and the only one we're calculating for this class. It is used for *linear* relationships between quantitative variables.
The Spearman rank order correlation coefficient (r_s), also called *rho*, provides a more accurate characterization of a relationship when one or both of the variables is a rank order—e.g., the relationship between SAT scores and high school class ranking.

As mentioned previously, *r* is only appropriate for measuring the linear relationship between two variables. To measure *curvilinear* relationships, Eta (η) is used instead. Pearson *r* would underestimate the degree of the relationship in these cases.

Theoretical Interpretation of Correlations in Research

Once a correlation is found, we must be careful about interpretations of ***causality*.** The statistical demonstration of a relationship between two variables does not prove that one has *caused* the other. There are three reasons why relationships between two variables ("X" and "Y") might be consistently found:

- ***Differences in X might cause differences in Y.*** If a correlation between mothering behavior (e.g., time spent holding a baby) and infant temperament (e.g., frequency of crying) is found—a negative relationship, in this case—maybe the mothering behavior influences temperament…
- ***Differences in Y might cause differences in X.***
  …or maybe infants' crying causes mothers to spend less time holding them…
- ***Differences in a third variable might cause differences in both X and Y.***
  …or maybe both are influenced by something else—genetic influences on personality that affect both mother and child? The stress of the home environment?

Another example of this: ice cream consumption is correlated with drowning (because both occur when the weather is hot).

This is probably the most persistent misunderstanding of statistical literature in popular media. It can be tempting to assume that (A) one of the measured variables causes the other, and therefore (B) manipulation of one of the measured variables will effectively manipulate the other. Neither can be assumed based only on the existence of a correlation.

For example:

> *"Compared to teens that have frequent family dinners, those who rarely have family dinners are three-and-a-half times more likely to have abused prescription drugs or an illegal drug other than marijuana. …Eating dinner together proves to be a simple, effective way to reduce the risk of youth substance abuse."*
>
> --U.S. Department of Health and Human Services

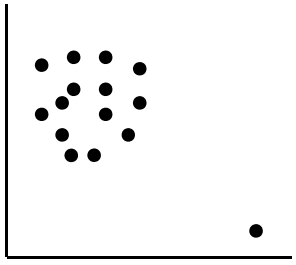
Only with an *experiment*—a type of research design in which conditions are systematically manipulated in ways that rule out alternative causes—can a single study demonstrate a causal relationship between one variable and another.

Other Issues of Interpretation of ***r***

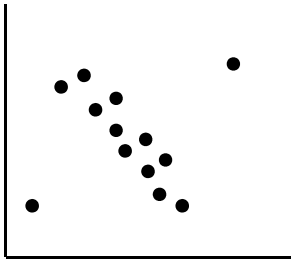
Depending on the distribution of bivariate data, the value of *r* may not be a useful summary of the relationship between two variables. Two of these problems—extreme scores and dissimilar groups—can be detected by inspecting the scatter plot for a bivariate distribution.

Extreme Scores ("Outliers").

Values far from the center of the distribution, either on the *X* variable or the *Y* variable, can have a very strong influence on the value of *r*. In some cases, the value of *r* would change dramatically if one or a few cases were removed. Such values can usually be identified in a scatter plot, because they are far from the main "cluster" of the bivariate distribution (as in the example on the right).



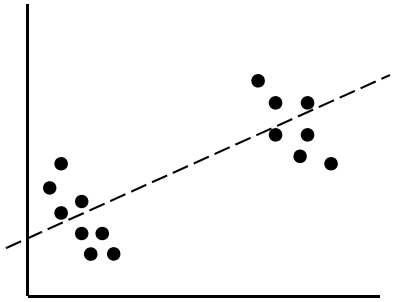
This can work in either direction—it may appear that a relationship exists (when in fact the high magnitude of *r* is due entirely to outliers), or it may appear that there is *no* relationship (when in fact the magnitude of *r* would be high, if it weren't for the outliers—as in the example on the right). *Pearson r fails to usefully measure the general relationship between X and Y in these situations (where outliers strongly affect the value of r).*



This effect can be observed mathematically while we're calculating *r.* You may have noticed that, when calculating the summation $\sum z_X z_Y$ in the formula, some scores contribute much more than others to the sum, and therefore to the eventual value of *r.* Cases far from the mean on *X*, *Y*, or both variables (i.e., those with large negative or positive z scores) tend to affect the value of *r* more than cases closer to the mean.

Combining Dissimilar Groups

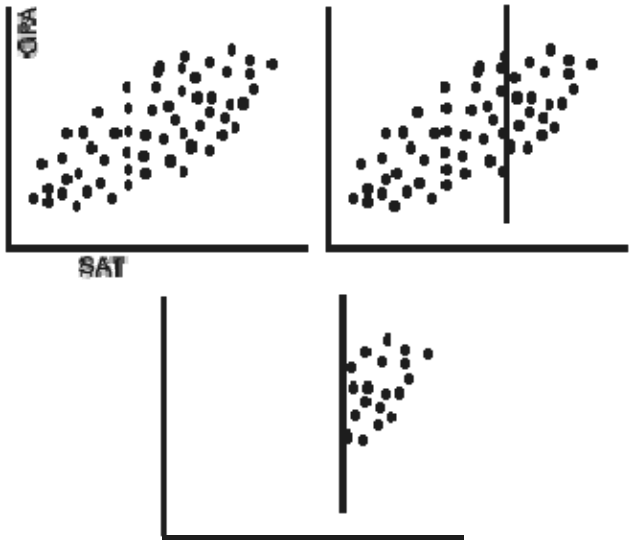
If scores are taken from two groups with distinct distributions, the relationship between X and Y may be mischaracterized by the value of *r* for the combined sample. On the right is an example of this. We can visually identify two groups—each by itself would be negatively related, but combining the groups results in a positive correlation. Again, this situation means that the value of *r* does not meaningfully represent the relationship between the two variables.



Restriction of Range

Another problem of interpretation—one that cannot be detected by looking at a scatter plot of your bivariate data—can come from measuring an inadequate range of values. This typically has the effect of reducing the magnitude observed for a correlation.

Example: The correlation of aptitude test scores (*X*) with college grade point average (*Y*) is reduced if the range of X values is restricted to the higher range of test scores. (This is what would happen if we used a college sample—college students are selected to have higher aptitude test scores than the average.)



From http://www.gseis.ucla.edu/courses/ed230bc1/notes1/cor2-1.gif

Spurious correlations

An additional possibility, for any given research study, is that the relationship found in the data occurred *by chance*—in this case, the relationship can be called a "spurious correlation."

We can't be *certain* whether this is the case just by looking at a single study—but we can manage the risk of this in the design and analysis of the study. (Management of the risk of finding spurious relationships is a fundamental topic in *inferential statistics*, and we will cover the methods of managing this risk later in the course.)