CORRELATION COEFFICIENTS

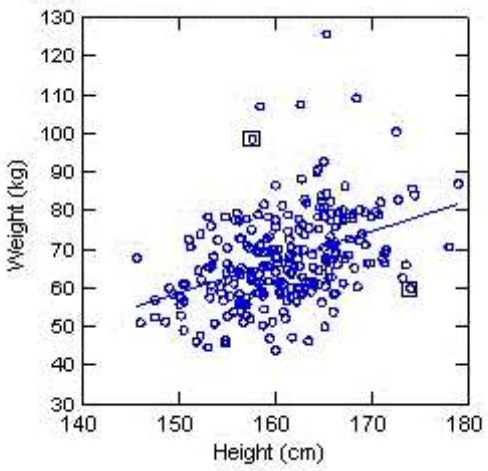
(taken June 7 2007 from http://www.tufts.edu/~gdallal/corr.htm)

We've discussed how to summarize a single variable. The next question is how to summarize a pair of variables measured on the same observational unit--(percent of calories from saturated fat, cholesterol level), (amount of fertilizer, crop yield), (mother's weight gain during pregnancy, child's birth weight). How do we describe their joint behavior?

Scatterplots! Scatterplots! Scatterplots!

The first thing to do is construct a scatterplot, a graphical display of the data. There are too many ways to be fooled by numerical summaries, as we shall see!

The numerical summary includes the mean and standard deviation of each variable separately plus a measure known as the *correlation coefficient* (also the *Pearson correlation coefficient,* after Karl Pearson), a summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.



"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.

Correlations tend to be positive. Pick any two variables at random and they'll almost certainly be positively correlated, if they're correlated at all--height and weight; saturated fat in the diet and cholesterol levels; amount of fertilizer and crop yield; education and income. Negative correlations tend to be rare--automobile weight and fuel economy; folate intake and homocysteine; number of cigarettes smoked and child's birth weight.

The correlation coefficient of a set of observations $\{(x_i,y_i)$: i=1,..,n$\}$ is given by the formula

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

The key to the formula is its numerator, the sum of the products of the deviations.

[Scatterplot of typical data set with axes drawn through (Xbar,Ybar)]

Quadrant	x(i)-xbar	y(i)-ybar	(x(i)-xbar)*(y(i)-ybar)
I	+	+	+
II	-	+	-
III	-	-	+
IV	+	-	-

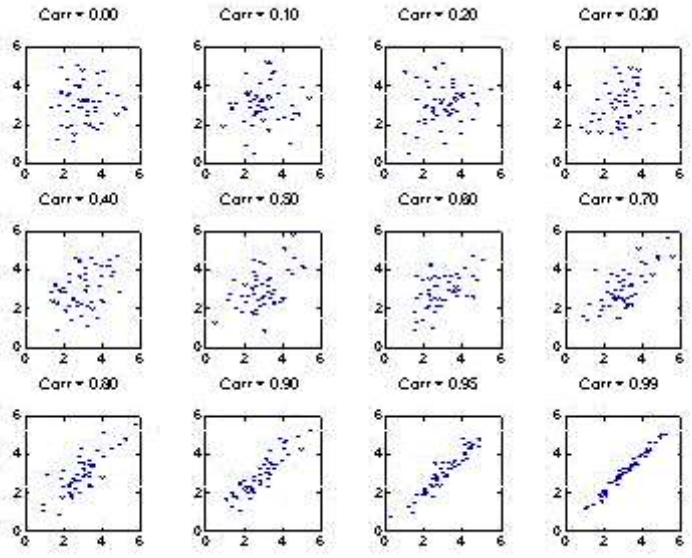
If the data lie predominantly in quadrants I and III, the correlation coefficient will be positive. If the data lie predominantly in quadrants II and IV the correlation coefficient will be negative.

The denominator will always be positive (unless all of the x's or all of the y's are equal) and is there only to force the correlation coefficient to be in the range [-1,1].

Properties of the correlation coefficient, r:

- -l <= r <= +1
- |r| = 1 if and only if the points lie exactly on a straight line.
- If the same constant is added to all of the Xs, the correlation coefficient is unchanged. Similarly for the Ys
- If all of the Xs are multiplied by a constant, the correlation coefficient is unchanged, except that the sign of the correlation coefficient is changed if the constant is negative. Similarly for the Ys.

The last two properties mean the correlation coefficient doesn't change as the result a linear transformation, aX+b, where 'a' and 'b' are constants, except for a change of sign if 'a' is negative. Hence, when investigating height and weight, the correlation coefficient will be the same whether height is measured in inches or centimeters and the weight is measured in pounds or kilograms.

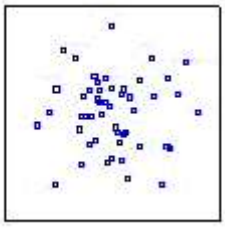


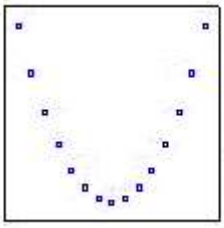
How do values of the correlation coefficient correspond to different data sets? As the correlation coefficient increases in magnitude, the points become more tightly concentrated about a straight line through the data. Two things should be noted. First, correlations even as high as 0.6 don't look that different from correlations of 0. I want to say that correlations of 0.6 and less don't mean much if the goal is to predict individual values of one variable from the other. The prediction error is nearly as great as we'd get by ignoring the second variable and saying that everyone had a value of the first variable equal to the overall mean! However, I'm afraid that this might be misinterpreted as suggesting that all such associations are worthless. They have important uses that we will discuss in detail when we consider linear regression. Second, although the correlation can't exceed 1 in magnitude, there is still a lot of variability left when the correlation is as high as 0.99.

[(American Statistician article) conducted an experiment in which people were asked to assign numbers between 0 and 1 to scatterplots showing varying degrees of association. They discovered that people perceived association not as proportional to the correlation coefficient, but as proportional to 1 - $\sqrt{}$(1- r^2).

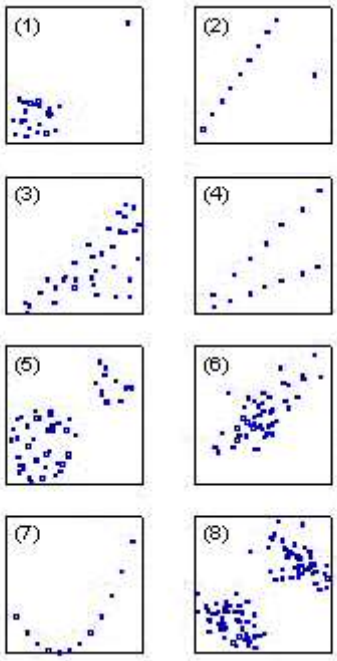
r	$1-\sqrt{(1-r^2)}$
0.5	0.13
0.7	0.29
0.8	0.40
0.9	0.56
0.99	0.86
0.999	0.96

Trouble!



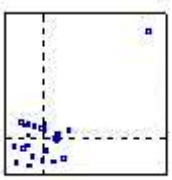


The pictures like those in the earlier displays are what one usually thinks of when a correlation coefficient is presented. But the correlation coefficient is a single number summary, a measure of linear association, and like all single number summaries, it can give misleading results if not used with supplementary information such as scatterplots. For example, data that are uniformly spread throughout a circle will have a correlation coefficient of 0, but so, too, will data that is symmetrically placed on the curve $Y = X^2$! The reason the correlation is zero is that high values of Y are associated with both high and low values of X. Thus, here is an example of a correlation of zero even where there is Y can be predicted perfectly from X!



To further illustrate the problems of attempting to interpret a correlation coefficient without looking at the corresponding scatterplot, consider this set of scatterplots, which duplicates most of the examples from pages 78-79 of *Graphical Methods for Data Analysis* by Chambers, Cleveland, Kleiner, and Tukey. **Each data set has a correlation coefficient of 0.7.**

What to do:

1. 

The correlation is 0 within the bulk of the data in the lower left-hand corner. The outlier in the upper right hand corner increases both means and makes the data lie predominantly in quadrants I and III. Check with the source of the data to see if the outlier might be in error. Errors like these often occur when a decimal point in both measurements is accidentally shifted to the right. Even if there is no explanation for the outlier, it should be set aside and the correlation coefficient or

the remaining data should be calculated. The report must include a statement of the outlier's existence. It would be misleading to report the correlation based on all of the data because it wouldn't represent the behavior of the bulk of the data.

As discussed below, correlation coefficients are appropriate only when data are obtained by drawing a random sample from a larger population. However, sometimes correlation coeficients are mistakenly calculated when the values one of the variables--X, say--are determined or constrained in advance by the investigator. In such cases, the message or the outlier may be real, namely, that over the full range of values, the two variables tend to increase and decrease together. It's poor study design to have the answer determined by a single observation and it places the analyst in an uncomfortable position. It demands that we assume thr association is roughly linear over the entire range and that the variability in Y will be no different for large X from what it is for small X. Unfortunately, once the study is performed, there isn't much that can be done about it. The outcome hinges on a single obsrevation.

2. Similar to 1. Check the outlier to see if it is in error. If not, report the correlation coefficient for all points except the outlier along with the warning that the outlier occurred. Unlike case 1 where the outlier is an outlier in both dimensions, here the outlier has a reasonable Y value and only a slightly unreasonable X value. It often happens that observations are *two-dimensional outliers*. They are unremarkable when each response is viewed individually in its histogram and do not show any aberrant behavior until they are viewed in two dimensions. Also, unlike case 1 where the outlier increases the magnitude of correlation coefficient, here the magnitude is decreased.
3. This sort of picture results when one variable is a component of the other, as in the case of (total energy intake, energy from fat). The correlation coefficient almost always has to be positive since increasing the total will tend to increase each component. In such cases, correlation coefficients are probably the wrong summaries to be using. The underlying research question should be reviewed
4. The two nearly straight lines in the display may be the result of plotting the combined data from two identifiable groups. In might be as simple as one line corresponding to men, the other to women. It would be misleading to report the single correlation coefficient without comment, even if no explanation manifests itself.
5. The correlation is zero within the two groups; the overall correlation of 0.7 is due to the differences between groups. Report that there are two groups and that the within group correlation is zero. In cases where the separation between the groups is greater, the comments from case 1 apply as well. It may be that the data are not a simple random sample from a larger popultion and the division between the two groups may be due to a conscious decision to exclude values in the middle of the range of X or Y. The correlation coefficient is an inappropriate summary of such data because its value is affected by the choice of X or Y values.
6. What most researcher think of when a correlation of 0.7 is reported.
7. A problem mentioned earlier. The correlation is not 1, yet the observations lie on a smooth curve. The correlation coefficient is 0.70 rather than 0 because here the curve is not symmetric. Higher values of Y tend to go with higher values of X. A correlation coefficient is an inappropriate numerical summary of this data. Either (i) derive an expression for the curve, (ii) transform the data so that the new variables have a linear relationship, or (iii) rethink the problem.
8. This is similar to case 5, but with a twist. Again, there are two groups, and the separation between them produces the postive overall correlation. But, here, the within-group correlation is negative! I would do my best to find out why there are two groups and report the within group correlations.

The moral of these displays is clear: **ALWAYS LOOK AT THE SCATTERPLOTS**!

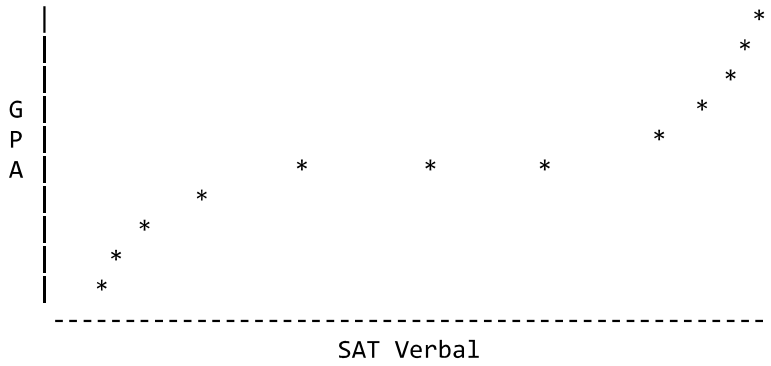
The correlation coefficient is a numerical summary and, as such, it can be reported as a measure of association for any batch of numbers, no matter how they are obtained. Like any other statistic, its proper interpretation hinges on the sampling scheme used to generate the data.

The correlation coefficient is most appropriate when both measurements are made from a simple random sample from some population. The sample correlation then estimates a corresponding quantity in the population. It is then possible to compare sample correlation coefficients for samples from different populations to see if the association is different within the populations, as in comparing the association between calcium intake and bone density for white and black postmenopausal females.

If the data do not constitute a simple random sample from some population, it is not clear how to interpret the correlation coefficient. If, for example, we decide to measure bone density a certain number of women at each of many levels of calcium intake, the correlation coefficient will change depending on the choice of intake levels.

This distortion most commonly occurs in practice when the range of one of the variables has been restricted. How strong is the association between MCAT scores and medical school performance? Even if a simple random sample of medical students is chosen, the question is all but impossible to answer because applicants with low MCAT scores are less likely to be admitted to medical school. We can talk about the relationship between MCAT score and performance only within a narrow range of high MCAT scores.

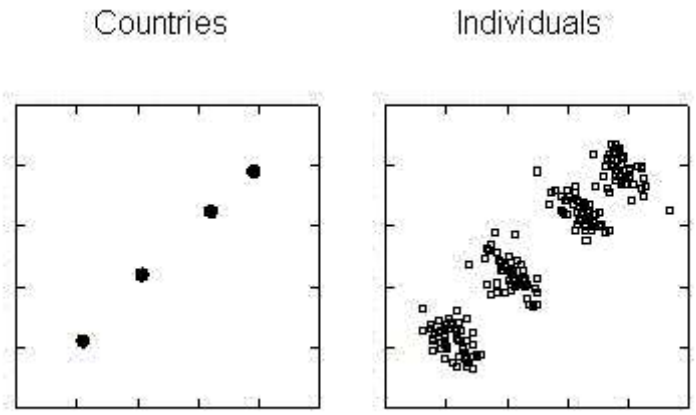
[One major New York university with a known admissions policy that prohibited penalizing an applicant for low SAT scores investigated the relationship between SAT scores and freshman year grade point average. The study was necessarily non-scientific because many students with low SAT scores realized that while the scores wouldn't hurt, they wouldn't help, either, and decided to forego the expense of having the scores reported. The relationship turned out to be non-linear. Students with very low SAT Verbal scores (350 or less) had low grade point averages. For them, grade point average increased with SAT score. Students with high SAT Verbal scores (700 and above) had high grade point averages. For them, too, grade point average increased with SAT score. But in the middle (SAT Verbal score between 350 and 700), there was almost no relationship between SAT Verbal score and grade point average.



Suppose these students are representative of all college students. What if this study were performed at another college where, due to admissions policies, the students had SAT scores only within a restricted range?

- How would the results of that study differ from the results here?
- What would be the effect on the correlation coefficient?
- Could a valid comparison of the relationship between SAT scores and grade point average in the two schools be made by comparing correlation coefficients? If not, then how?]

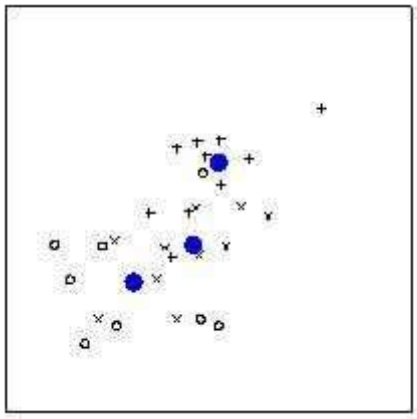
Ecological Fallacy



Another source of misleading correlation coefficients is *the ecological fallacy*. It occurs when correlations based on grouped data are incorrectly assumed to hold for individuals.

Imagine investigating the relationship between food consumption and cancer risk. One way to begin such an investigation would be to look at data on the country level and construct a plot of overall cancer risk against per capita daily caloric intake. The display shows cancer increasing with food consumption. But it is people, not countries, who get cancer. It could very well be that within countries those who eat more are less likely to develop cancer. On the country level, per capita food intake may just be an indicator of overall wealth and industrialization.

The ecological fallacy was in studying countries when one should have been studying people.



When the association is in the same direction for both individuals and groups, the ecological correlation, based on averages, will typically overstate the strength of the association in individuals. That's because the variablity within the groups will be eliminated. In the picture to the left, the correlation between the two variables is 0.572 for the set of 30 individual observations. The large blue dots represent the means of the crosses, plus signs, and circles. The correlation for the set of three dots is 0.902

Spurious Correlations

Correlation is not causation. The observed correlation between two variables might be due to the action of a third, unobserved variable. Yule (1926) gave an example of high positive correlation between yearly number of suicides and membership in the Church of England due not to cause and effect, but to other variables that also varied over time. (Can you suggest some?) Mosteller and Tukey (1977, p. 318) give an example of aiming errors made during World War II bomber flights in Europe. Bombing accuracy had a high positive correlation with amount of fighter opposition, that is, the more enemy fighters sent up to distract and shoot down the bombers, the more accurate the bombing run! The reason being that lack of fighter opposition meant lots of cloud cover obscuring bombers from the fighters and the target from the bombers, hence, low accuracy.

Last modified: 11/03/2022 22:38:43.Last modified: 03/19/2007 17:45:35.