10.2: The Correlation Coefficient r

As we begin this section we note that the type of data we will be working with has changed. Perhaps unnoticed, all the data we have been using is for a single variable. It may be from two samples, but it is still a single variable (**univariate**) . The type of data described in the examples above and for any model of cause and effect is **bivariate** data ("bi" for two variables). In reality, statisticians use **multivariate** data, meaning they use many variables in their analyses.

For our work we can classify data into three broad categories, time series data, cross-section data, and panel data. We met the first two very early on. Time series data measures a single unit of observation; say a person, or a company or a country, as time passes. What are measured will be at least two characteristics, say the person's income, the quantity of a particular good they buy and the price they paid. This would be three pieces of information in one time period, say 1985. If we followed that person across time we would have those same pieces of information for 1985, 1986, 1987, etc. This would constitute a times series data set. If we did this for 10 years we would have 30 pieces of information concerning this person's consumption habits of this good for the past decade and we would know their income and the price they paid.

A second type of data set is for cross-section data. Here the variation is not across time for a single unit of observation, but across units of observation during one point in time. For a particular period of time we would gather the price paid, amount purchased, and income of many individual people.

A third type of data set is panel data. Here a panel of units of observation is followed across time. If we take our example from above we might follow 500 people, the unit of observation, through time, ten years, and observe their income, price paid and quantity of the good purchased. If we had 500 people and data for ten years for price, income and quantity purchased we would have 15,000 pieces of information. These types of data sets are very expensive to construct and maintain. They do, however, provide a tremendous amount of information that can be used to answer very important questions. As an example, what is the effect on the labor force participation rate of women as their family of origin, mother and father, age? Or are there differential effects on health outcomes depending upon the age at which a person started smoking? Only panel data can give answers to these and related questions because we must follow multiple people across time. The work we do here however will not be fully appropriate for data sets such as these.

Beginning with a set of data with two independent variables we ask the question: are these related? One way to visually answer this question is to create a scatter plot of the data. We could not do that before when we were doing descriptive statistics because those data were univariate. Now we have bivariate data so we can plot in two dimensions.

To provide mathematical precision to the measurement of what we see we use the correlation coefficient. The correlation tells us something about the co-movement of two variables, but nothing about why this movement occurred. Formally, correlation analysis assumes that both variables being analyzed are **independent** variables. This means that neither one causes the movement in the other. Further, it means that neither variable is dependent on the other, or for that matter, on any other variable. Even with these limitations, correlation analysis can yield some interesting results.

The correlation coefficient, ρ (pronounced rho), is the mathematical statistic for a population that provides us with a measurement of the strength of a linear relationship between the two variables. For a sample of data, the statistic, r, developed by Karl Pearson in the early 1900s, is an estimate of the population correlation and is defined mathematically as:

$$r_{XY} = \frac{\sum(X_i - \overline{X}) * (Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 * \sum(Y_i - \overline{Y})^2}}$$

or

$$r_{XY} = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sqrt{\left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right] * \left[\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}\right]}}$$

where $\overline{X}$ and $\overline{Y}$ are the sample means of the two independent variables X and Y, X_i and Y_i are the individual observations of X and Y, and n is our sample size. The correlation coefficient r ranges in value from -1 to 1. The second equivalent formula is often used because it may be computationally easier. As scary as these formulas look, they are really just the ratio of the covariance between the two variables and the product of their two standard deviations. That is to say, it is a measure of relative variances.

To visualize any **linear** relationship that may exist review the plot of a scatter diagrams of the standardized data. Figure 10.2.1 presents several scatter diagrams and the calculated value of r. In panels (a) and (b) notice that the data generally trend together, (a) upward and (b) downward. Panel (a) is an example of a positive correlation and panel (b) is an example of a negative correlation, or relationship. The sign of the correlation coefficient tells us if the relationship is a positive or negative (inverse) one. If all the values of two variables X_1 and X_2 are on a straight line the correlation coefficient will be either 1 or -1 depending on whether the line has a positive or negative slope and the closer to one or negative one the stronger the relationship between the two variables.

Figure 10.2.1

Remember, all the correlation coefficient tells us is whether or not the data are linearly related. In panel (d) the variables obviously have some type of very specific relationship to each other, but the correlation coefficient is zero, indicating no **linear** relationship exists. In panel (c) the variables are not related and the correlation coefficient is understandably equal to zero.

If you suspect a linear relationship between two variables X and Y, then r can measure how strong the linear relationship is.

What the VALUE of r tells us:

- The value of r is always between -1 and $+1$: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the **linear** relationship between variables X and Y. Values of r close to -1 or to $+1$ indicate a stronger linear relationship between the two variables.
- If $r = 0$ there is absolutely no linear relationship between variables X and Y **(no linear correlation)**.
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us

- A positive value of r means that when X increases, Y tends to increase and when X decreases, Y tends to decrease **(positive correlation)**.
- A negative value of r means that when X increases, Y tends to decrease and when X decreases, Y tends to increase **(negative correlation)**.

Note

Strong correlation does not suggest that X causes Y or that Y causes X. We say **"correlation does not imply causation."**