

## Pearson Product-Moment Correlation

The **Pearson product-moment correlation coefficient** ( $r$ ) can be calculated to determine the nature of a relationship between two numerical (generally interval or ratio scale) variables. The numbers used in the calculations must be inherently paired, either because they come from the same subject or because they come from pairs of matched subjects. It's important throughout the calculations that the numbers remain paired properly, or else the wrong value of  $r$  will be obtained. Thus, when the values are entered into a calculator or computer program, the pairings *must be maintained*.

The variables are often labeled X and Y, where X stands for the **independent** or **predictor variable**, and Y stands for the **dependent** or **criterion variable**.

The correlation coefficient will ALWAYS be a number between -1 and +1. That is, it must be true that

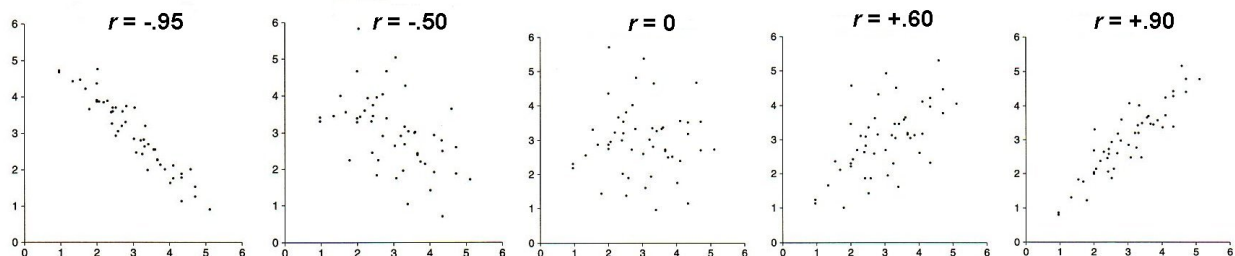
$$-1 \leq r \leq +1$$

No other result is mathematically possible, so any calculated value of  $r$  that falls outside this range has been *calculated incorrectly!*

The value of  $r$  gives three pieces of information:

- 1) It tells if there is a relationship between the two variables. If not, then  $r \approx 0$ .
- 2) If there is a relationship, the sign of  $r$  tells if it is positive or negative.
  - a) A **positive relationship** (also called a direct relationship) is one in which large values of X (i.e., values above the mean of X) are paired with large values of Y, and small values of X (i.e., values below the mean of X) are paired with small values of Y.
  - b) A **negative relationship** (also called an inverse relationship) is one in which large values of X are paired with small values of Y, and *vice versa*.
- 3) The size of  $r$  tells how strong the relationship is.

The same information can be obtained by examining a **scatterplot** of the (X, Y) pairs.



The techniques illustrated here assume a **linear relationship** between X and Y. If the scatterplot shows a **curvilinear relationship**, these techniques must be modified. These modifications will not be dealt with in this course.

Studies involving correlational techniques very often involve no manipulation of an IV by the researcher. Rather, values of X and Y are merely observed and recorded. For this reason, such studies are often referred to as **observational studies** or **correlational studies**. When this is the case, it must always be remembered that *correlation does not imply causality*. Just because two variables are correlated with each other does not mean there is a cause and effect relationship between them. This is illustrated by the following case. There is a strong positive correlation between the number of churches (X) in a community and the number of violent crimes (Y) committed in the community. To propose a cause and effect relationship between these two variables, however, would obviously be absurd. This case illustrates what is sometimes called the **third-variable problem**: a correlation between X and Y is created because both variables are related to a third variable, in this case, population. These "third variables" are also referred to as **lurking variables** and **confounding variables**. Sometimes, the effect of a third variable can be removed statistically by a procedure, not covered in this course, called **partial correlation**.

There are other problems in correlational research which we must also be on the lookout for.

- 1) **Restricted range.** A well-known researcher at Harvard University attempted to find if there is a relationship between scores on the GRE and performance in grad school, as measured by grad school GPA. He found no relationship, which is hardly surprising. Do you see the problem with his study?
- 2) **Person-who arguments.** Correlations are rarely perfect, and in the social sciences are rarely even particularly strong. There is a distinct positive correlation between score on the SAT and college GPA. However, everyone knows a "person who" bucks this trend, doing well on the SAT and bombing in college, or *vice versa*. This does not invalidate the relationship. When correlations are not perfect, and especially when they are weak, there will always be persons who buck the trend.
- 3) **Sneaky cause-and-effect.** Once a correlation is established, cause and effect arguments can sneak in no matter how cautious we are. For example, there is a distinct positive relationship between fat intake per capita and death rate. Does this mean that reducing fat intake will reduce the death rate? Not necessarily.
- 4) **Extrapolations.** Relationships, even strong ones, don't go on forever. Once we are outside the range of the original data, it is risky to assume the relationship we saw still exists. Population is increasing rapidly over time and has been for several decades now. Is it reasonable to predict what the population will be 50 or 100 years from now by looking at this relationship. Probably not. The availability of essential resources will eventually cause population numbers to level off or even decrease again. This is one trend that will definitely not continue indefinitely!
- 5) **Outliers.** Outliers can have a strong effect of correlation and regression coefficients.

A linear relationship between X and Y can further be described by calculating the equation of a line that comes as close as possible to all the points on the scatterplot. This technique is called **regression analysis** and the equation is called the **regression equation**. Two numbers, called regression coefficients, must be calculated: the **slope** of the regression line, and the **y-intercept** of the regression line. The most common method of calculating these values is called the **method of least squares**, which is the method illustrated here. The regression equation can then be used to **predict** values of Y when only an X value is known. For example, college GPA can

be predicted from an SAT score. CAUTION: Predictions cannot be made in the reverse direction. You CANNOT predict values of X from a value of Y, by solving the equation for X.

**Example.** An industrial psychologist wishes to determine if there is a statistical relationship between the number of weeks experience in a job involving the wiring of electrical components and the number of workers' components rejected as being defective. A sample of 10 randomly selected workers is taken, and the number of components rejected is counted during a one-week period. The data are shown below.

<b>Weeks of experience (X)</b>	7	9	6	14	8	12	10	4	2	11
<b>Number of rejected components (Y)</b>	26	20	28	16	23	18	24	26	38	22

- 1) Transfer the data to the supplied worksheet.
- 2) Plot the scatterplot.
- 3) Calculate the summary statistics, the cross-products, and the sum of products (SP).
- 4) Calculate the correlation coefficient ( $r$ ) and interpret it. Describe the nature of the relationship.
- 5) Calculate the coefficient of determination ( $r^2$ ) and interpret it.
- 6) Calculate the regression coefficients, write out the regression equation, and interpret it.
- 7) Use the regression equation to make the following predictions:
  - a) The number of rejects for a worker with 6 weeks of experience: \_\_\_\_\_
  - b) The number of rejects for a worker with 8 weeks of experience: \_\_\_\_\_
  - c) The number of rejects for a worker with 10 weeks of experience: \_\_\_\_\_
  - d) The number of rejects for a worker with 30 weeks of experience: \_\_\_\_\_
- 8) Set alpha at .05 and test the following hypotheses, following the five customary steps:
 

$H_0$ : there is no correlation between weeks of job experience and number of rejected components  
 $H_1$ : there is a correlation between weeks of job experience and number of rejected components

-or-

$H_1$ : there is a negative correlation between weeks of job experience and number of rejected components

Note: for these tests,  $df = n - 2$ , where  $n$  is the number of pairs of scores. The critical values of  $r$  come from a table of critical values in the back of the book.

**Another example.** In the state of Florida there is concern that the ever increasing number of powerboats licensed to private citizens is taking its toll on the endangered manatee population. A wildlife biologist wishes to test the hypothesis that as powerboat registrations increase, manatee deaths from powerboat accidents also increase.

**Step 1)** State the null and alternative hypotheses.

$H_0$ :

$H_1$ :

**Step 2)** Establish a decision criterion (set alpha).

**Step 3)** Go out and collect the data, calculate summary statistics, and calculate the value of the test statistic.

year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
<b>powerboats registered (in thousands)</b>	498	513	512	526	559	585	614	645	675	711
<b>manatee deaths</b>	16	24	20	15	34	33	33	39	43	50

**Step 4)** Make a statistical decision concerning the null hypothesis.

**Step 5)** Write a conclusion describing the result of the statistical analysis.

**Question)** In 1990, there were 719,000 powerboats registered. Use this figure to predict the number of manatee deaths from powerboat accidents. You'll need to calculate the regression equation to do this. How does your prediction agree with the actual figure of 47 deaths?