

Representation Chapter Two -- Methodology

Last Modified July 8, 2008

In this chapter, we will discuss some statistical techniques that you will be asked to apply in the exercises in chapter 3. Almost all of these techniques require that variables be measured at least at the interval level. An exception is eta². For this measure, one variable must be at least interval, but the other variable may be at any level of measurement, and is usually nominal.

The Mean

The mean (μ) is a measure of "central tendency." It provides an average for a set of numbers.

The formula for computing the mean is:

$$\mu = \frac{\sum X_i}{N}$$

where

X_i = an individual value of X , and

N = the number of cases.

For example, if we add up the numbers shown in figure 2.1 and divide by 5, the result will be 10.

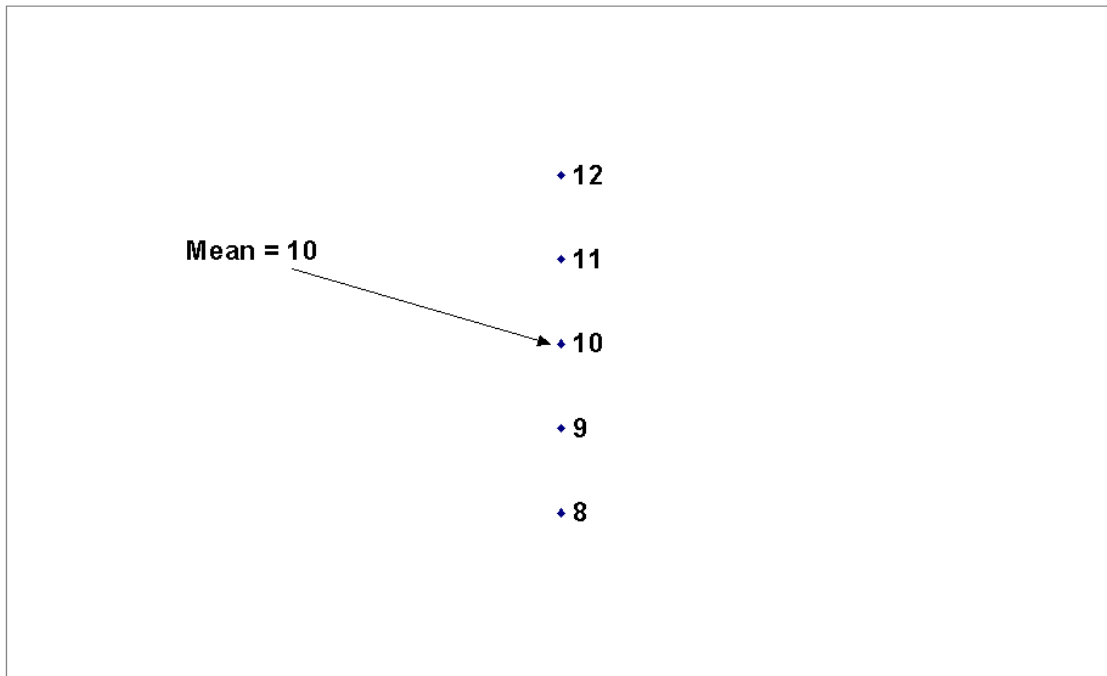


Figure 2.1

The Variance

The variance is a measure of dispersion, that is, its purpose is to show the extent to which the values of a variable are spread out rather than clustered together. Specifically, the variance is the (mean) average squared difference between the individual values of a variable and the mean value.

The formula for computing the variance (σ^2) of variable X is:

$$\sigma^2 = \frac{\sum (X_i - \mu)}{N}$$

Note: For sample data, the denominator of the formula is N-1. In this module, we will be working with the whole population of California state legislators, not with a sample. Even with sample data, the distinction in the formula makes little difference except for small samples.

Figure 2.2 shows numbers for two groups with identical means of 10. In the first group, the numbers are closer to the mean, and the variance is 2. In the second group, although the mean is the same, the numbers are more spread out, and the variance is 8.

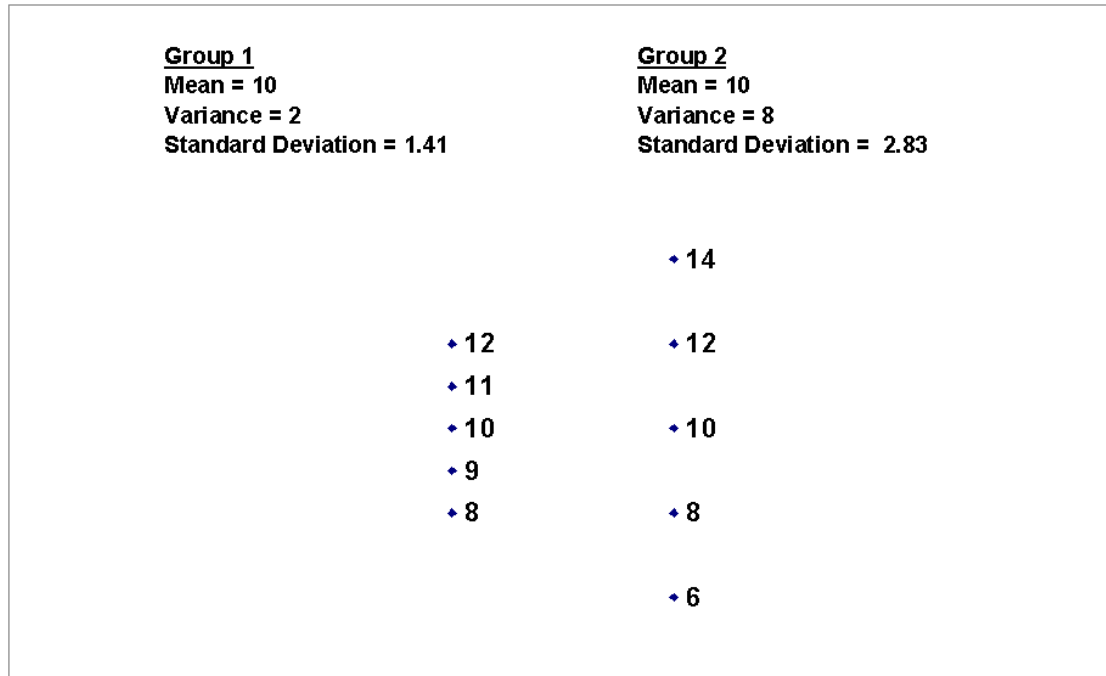


Figure 2.2

The Standard Deviation

The standard deviation (σ), like the variance, is a measure of dispersion, and is the one usually reported. It is simply the positive square root of the variance. In the examples shown in figure 2.2, the standard deviation in the first group of numbers is 1.41, while in the second group it is 2.83.

The variance and the standard deviation usually have no very obvious interpretation. This is the case here: about all we can say based on the results we have obtained is that there is more dispersion in the second group of numbers than in the first. Analyzing variance, however, is extremely important and will be the basis of much of what follows in this chapter.

COMPARING MEANS (η^2)

η^2 (h^2) measures the proportion of the variance in a variable that is "explained" ("determined," "predicted," "accounted for") by dividing cases into two or more groups. η^2 is the proportion of the variance that is *between* groups, with the rest (the "unexplained" variance) being *within* groups. In other words, its purpose is to tell us the degree to which groups are different from one another in terms of some variable.

The formula for η^2 is:

$$\eta^2 = \frac{\sigma_X^2 - \sigma_{Xg}^2}{\sigma_X^2}$$

where

s^2_X = the total variance in X (i.e., the variance about the mean of X), and

s^2_{Xg} = the variance in X about the mean of X within each group (the *within-group* variance).

Total variance is calculated exactly as shown earlier in the chapter. Within-group variance is calculated in the same way, except that, for each case, the group mean is substituted for the overall mean in the variance formula.

Figure 2.3 shows numbers for three groups. There is clearly both between-group variance (the groups have different means) and within-group variance (there are differences among members of the same group).

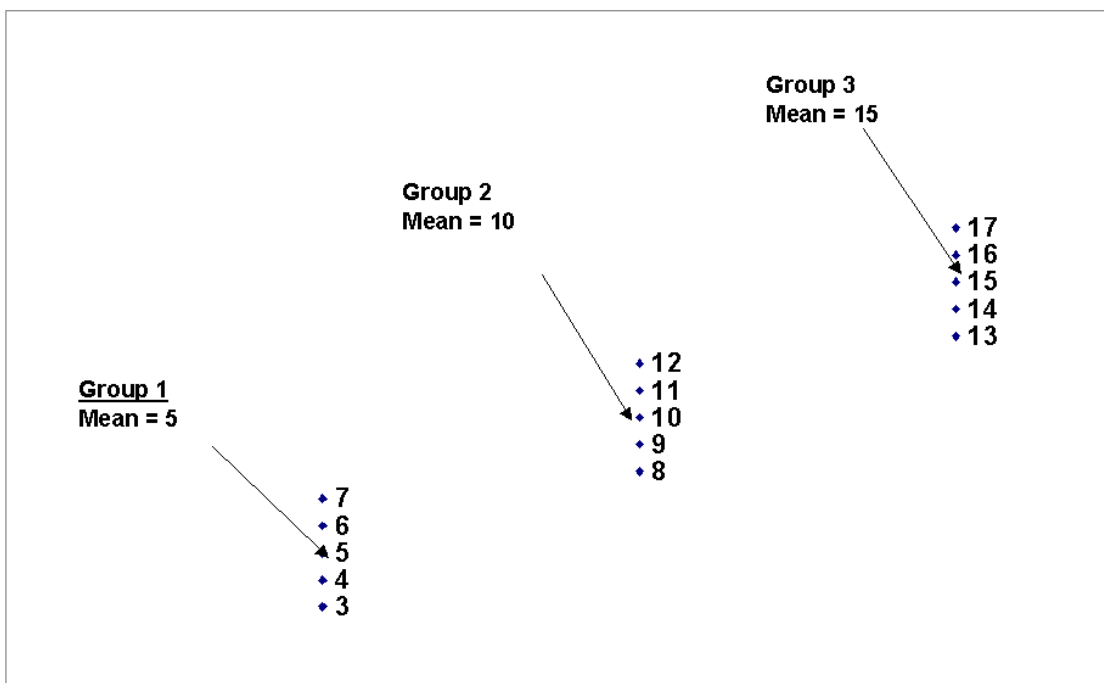


Figure 2.3

When we apply the formula for η^2 , we obtain a total variance of 18.67, and a within-group variance of 2.00. η^2 thus equals .89, indicating that 89 per cent of total variance reflects differences between the three groups, with within-group variance accounting for the remaining 11 per cent.

The Scatterplot (scattergram, scatter diagram)

A scatterplot is used to provide a graphic display of the relationship between two variables. The graph is set up by drawing a horizontal (X) axis along which the values of the independent variable are located, and a vertical (Y) axis along which the values of the dependent variable are located. Each case in a set of data can be placed on the graph by plotting the intersection of its coordinates on the two axes. Each case can be represented by a point. An example is shown in figure 2.4.

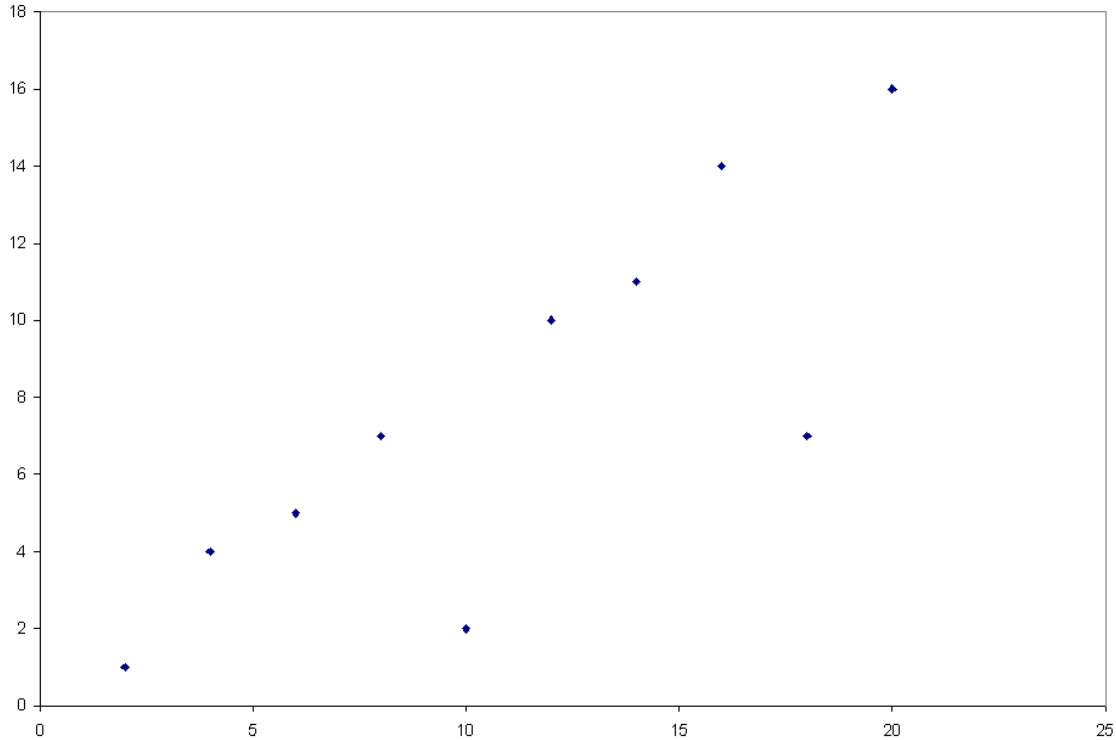


Figure 2.4

The Least Squares Equation (regression equation)

To the extent that there is a linear relationship between two variables, the scatterplot will tend to form a straight line. The least squares (regression) line summarizes this tendency. The line will have an upward slope if the variables are positively related, and a downward slope if they are negatively (inversely) related. If there is a perfect linear relationship between the two variables, the points on the graph will all fall on a straight line.

The general formula for describing any straight line is:

$$Y'_i = a + bX_i$$

where

Y'_i = the calculated (predicted) value of the dependent variable, that is, the value that the dependent variable would have for a given value of the independent variable if it fell precisely on the line,

a=the Y intercept, that is, the point at which the line crosses the Y axis (in other words, the predicted value of the dependent variable when the value of the independent variable is zero),

b=the slope of the line, that is, the increase or decrease in the predicted value of the dependent variable associated with an increase of one unit in the independent variable, and

X_i =any value (real or hypothetical) of the independent variable.

For any scatterplot, there is one and only one "line of best fit" to the data. This is the line that has the smallest variance of points about the line. The concept of variance used here is the same as that described earlier, except that instead of looking at the average squared difference between each value of a variable and the *mean* of that variable, we are now looking at the average squared difference between the actual values of the dependent variable and the values predicted by the line of best fit. Because the line of best fit is the one that has, on average, the smallest squared deviation between the line and the points on the scatterplot (i.e., between the actual and the predicted values of Y), it is called the "least squares" line. It is also called the "regression" line.

The formulas for calculating the least squares line are:

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

For the data in figure 2.4, the least squares equation is:

$$Y'_i = .33 + .67X_i$$

Figure 2.5 shows how this line is drawn on the scatterplot.

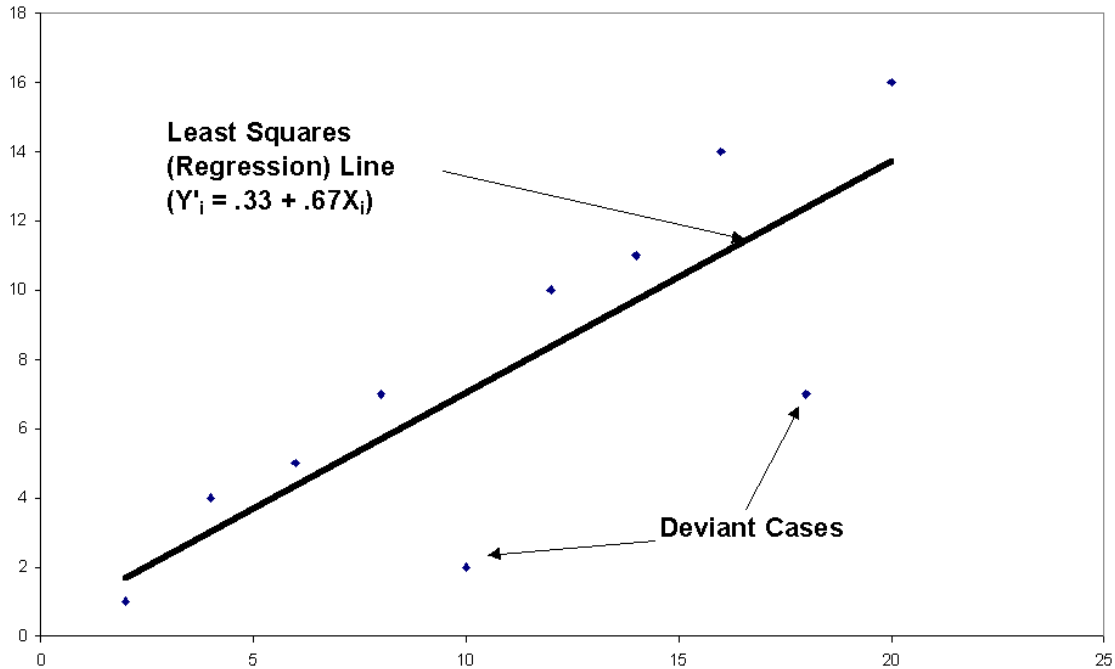


Figure 2.5

In addition to helping to uncover patterns in relationships between variables, least squares analysis can also help in isolating the exceptions to these patterns: these "deviant cases" (shown in figure 2.5) are those that have actual values for the dependent variable that are markedly different from those predicted by the equation.

The difference between the actual value and the predicted value is called the "residual." The predicted value for each case is calculated by plugging the value of the independent variable (X_i) into the least squares equation and solving for Y'_i . The residual is then computed as $Y_i - Y'_i$.

The Coefficient of Determination (Pearson's r^2)

Pearson's r^2 measures how good a job the best fitting line does. Specifically, it measures the degree to which the "residual" or "unexplained" variance (the variance about the least squares line) is smaller than the total variance (the variance about the mean) in the dependent variable. Pearson's r^2 has a range of from zero to one, with its value indicating the proportion of the variance in the dependent variable that is "explained," ("determined," "predicted," "accounted for") by variance in the independent variable.

The formula that we will use for r^2 is:

$$r^2 = \frac{\sigma_Y^2 - \sigma_{Y'}^2}{\sigma_Y^2}$$

where

σ_Y^2 = the total variance in Y (i.e., the variance about the mean of Y), and

$\sigma_{Y'}^2$ = the variance about Y', that is, about the least squares line (also called the *residual* variance, because it is the variance that is left over after accounting for the portion of the variance that is associated with that of the independent variable).

Computing the variance about the mean was covered earlier. Computing the residual variance is similar, except that, in the formula, the mean value of Y is replaced by the predicted value of Y (Y') for each value of X. For the data in figure 2.5, total variance is 22.41 and residual variance is 7.61, and so $r^2 = .66$.

The Correlation Coefficient (Pearson's r)

Pearson's r is the positive square root of Pearson's r^2 if the least squares line has an upward slope (that is, if the relationship is positive). It is the negative square root if the line slopes downward (that is, the relationship is negative [inverse]). Pearson's r has a range of from zero to plus or minus one. In our example, $r = .81$.

Pearson's r can be calculated directly, without having to first obtain Pearson's r^2 . (This is, in fact, the more common procedure. For approaches similar to that used here, see William Buchanan, *Understanding Political Variables*, 4th Ed. (N.Y.: Macmillan, 1988), especially pp. 267 and 288, and Susan Ann Kay, *Introduction to the Analysis of Political Data* (Englewood Cliffs, N.J.: Prentice-Hall, 1991), pp. 47ff.) The formula for Pearson's r is:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{[\Sigma(X - \bar{X})^2][\Sigma(Y - \bar{Y})^2]}}$$

Multiple Least Squares (Multiple regression)

Least squares analysis can fairly straightforwardly be extended to cover nonlinear least squares (which will not be covered in this module), and multiple least squares. Multiple least squares, or multiple regression, is used when there is more than one independent variable and we wish to measure the impact of each one controlling for the others, as well as the combined impact of all independent variables taken together. The *unstandardized* multiple regression equation for "n" independent variables is:

$$Y' = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

It is called the unstandardized equation because the regression (b) coefficients are expressed in terms of the original units of analysis of the variables. For example, if we were trying to predict legislators' ratings by an interest group (on a scale from 0 to 100) based on the per capita income of their districts (X_1), the percent Democratic party registration in their districts (X_2), and their ages (X_3), then b_1 , for example, would measure the increase or decrease in rating that would result from an increase of one dollar in per capita income, while controlling for Democratic registration and age. Similarly, b_2 and b_3 would measure (holding other variables constant) the change in rating resulting from an increase of one percent in Democratic registration and one year in legislator's age respectively.

It is also useful to calculate the *standardized* multiple regression equation:

$$Y' = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

In this equation, standardized regression coefficients (b, or "beta") measure the change in standard deviations produced in the dependent variable resulting (other variables again held constant) from an increase of one standard deviation in an independent variable. It is thus possible to compare the relative importance of each independent variable by comparing their beta coefficients. In other words, you *can* compare apples and oranges! Finally, R^2 , the multiple coefficient of determination, measures the proportion of variance in the dependent variable explained by all independent variables together.

Multiple least squares analysis requires that all variables be at least interval. Dichotomous variables (those with only two values, such as male-female or, for certain purposes, Democrat-Republican) are a special case. Because such variables contain only one interval (which is equal to itself) they meet the requirement for interval level measurement, and so can be introduced as independent variables in a multiple least squares equation. Such variables are called *dummy* variables. (If the *dependent* variable is a dummy variable, ordinary least squares cannot be used. Instead, logistic regression, which is beyond the scope of this module, is required.)