

## Data Analysis: Measures of Dispersion

- Once we understand the measures of central tendency, and check them in our data, we are then concerned with how spread out (or dispersed) the values are. A statistic that conveys this information is a *measure of dispersion*.
- Maybe the simplest measure of dispersion is the *range*, which is the difference between the greatest and the smallest value for a given variable, among the  $n$  observations.
- It is a simple concept often used, but it does not tell us anything about the distribution of the values of the variable of interest, whether they are concentrated around the middle, spread out all over, or concentrated near the endpoints.
- It is also very sensitive to outliers Think of comparing the following samples
  - a) 2, 4, 4, 4, 4, 6
  - b) 2, 2, 2, 6, 6, 6
  - c) 2, 2, 2, 6, 6, 9

- A modified measure is the *interquartile range*: it is the result of the following operation

$$3^{\text{rd}} \text{ quartile value} - 1^{\text{st}} \text{ quartile value} \quad (1)$$

that means it is the range of the middle 50% of observations. It is the difference between the 75<sup>th</sup> percentile and the 25<sup>th</sup> percentile.

- But we can do better. And for that we need a simple concept, the deviation of the values of a variable with respect to its mean. For the  $i^{\text{th}}$  observation, the deviation of  $X$  from the mean is

$$d_i = X_i - \bar{X}, \quad (2)$$

since there are  $n$  observations, there are  $n$  deviations, some of them are positive and some of them are negative.

- We can try to summarize these deviations and compute for example the mean deviation. However, the positive and negative values of the deviations always cancel out, so the mean deviation always equal zero:

$$\sum_{i=1}^n \frac{d_i}{n} = \sum \frac{(X_i - \bar{X})}{n}, \quad (3)$$

but of course we can rewrite that as

$$\sum \frac{X_i}{n} - \sum_{i=1}^n \frac{\bar{X}}{n} = \bar{X} - \frac{n\bar{X}}{n} = 0. \quad (4)$$

- To overcome this problem we can develop other measures, like the *Mean Absolute Deviation*. But other are more useful for us, like for example squaring the deviations, calculating their mean and then taking the square root. It seems like a lot of steps, but it is so commonly used that it is worth the effort:

$$S_X = \sqrt{\frac{\sum_{i=1}^n d_i^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad (5)$$

Notice that we are now dividing by  $n - 1$  instead of  $n$ , this is due to the fact that once we know  $\bar{X}$ ,  $i$  and  $n - 1$  values of a variable, the  $n^{th}$  observations is determined. Here  $n - 1$  is what we call the degrees of freedom.

- The Standard Deviation is usually interpreted as a typical deviation, but it is better to remember that usually 68% of the observations have values within one standard deviation of the mean, that is in the interval  $(\bar{X} - S_X, \bar{X} + S_X)$ . Also about 95% of the observations have values within two standard deviations of the mean, that is in the interval  $(\bar{X} - 2 S_X, \bar{X} + 2 S_X)$ .

- Another typical statistic is the *variance*, which is defined as the square of the standard deviation:

$$S_X^2 = \frac{\sum_{i=1}^n d_i^2}{n-1} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}. \quad (6)$$

Obviously, the standard deviation and the variance carry the same information, we will use both depending on the context. Again, we have to remember that both measures are sensitive to extreme values.

- Notice that the numerator in the definition of the variance is usually called, *total variation* in  $X$ , as it serves as a measure of the overall variability among the values of  $X$ .
- Finally, it is important to mention that although elegant the formulae for the standard deviation and the variance are not very practical to compute by hand, instead it is better to have the following formula for computing the variance (the standard deviation is just its square root):

$$S_X^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}. \quad (7)$$

## Data Analysis: Symmetry

- Sometimes it is useful to have a measure of the symmetry or asymmetry of a distribution of values. For that we have the skewness of a distribution defined as

$$\mu_3 = \frac{\sum (X_i - \bar{X})^3}{n - 2}. \quad (8)$$

Notice that a symmetric distribution has skewness equal to zero.

- A distribution can have positive skew, where the longer tail is to the right, or a negative skew, when the tail is to the left. In the case of positive skew the mode and the median are lower values than the mean. When the skew is negative the mean is lower than the median and the mode.

## Bivariate Statistics

- Our ultimate objective in econometrics is to determine how some variables are related to others, here we take a step in that direction.
- We start with a data set that has  $n$  observations on two variables,  $X$  and  $Y$ .
- We could plot these observations, and observe whether high values of one variable are associated with low values of the other, or with high values, etc.

- One measure that quantifies how much the values of  $X$  and  $Y$  vary together is the *covariance*:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}). \quad (9)$$

Notice that  $X$  and  $Y$  play the same role here, so there is no difference between the covariance of  $X$  and  $Y$ , and that of  $Y$  and  $X$ . The covariance only captures linear relationships, and it is very sensitive to the units used.

- To overcome the used of different units we can compute the *correlation coefficient* that is defined as:

$$r = \frac{S_{XY}}{(S_X)(S_Y)} \quad (10)$$

The interpretation of the correlation coefficient relates to an imaginary line that can be fit through the data points. The sign of  $r$  is the same as the sign of the slope of that line, and the magnitude (in absolute value) measures the degree to which the points lie close to the line.

- If the points in a scatter diagram lie exactly along some straight line with a positive slope, then  $r = 1$ . If they lie exactly along a line with a negative slope then  $r = -1$ . If the points are scattered all over randomly, the correlation will be close to zero. Notice that if they lie along a vertical or horizontal line, then  $r$  will not be defined because either  $S_X$  or  $S_Y$  will be zero.