

Descriptive Statistics

SPSS

The object of this handout is to give you a brief description of the main summary statistics and how to compute those statistics with the econometric package SPSS. It is certainly not a comprehensive statistics course. It is not a complete review of all the options and possibilities offered by SPSS either. We are simply trying here to give you a feel of how to extract information from your data and how to use the basic commands of SPSS to manipulate your dataset. This class is an initiation to research, which means that your personal contribution will be a large part of your output. Hence, once you have found your dataset, you should not hesitate to play with your data, try to identify patterns, co-movements between variables, compute summary statistics. Most importantly, you should not forget to refer to textbooks and the relevant literature in order to make sure that you use the appropriate methodology. It will also give you insightful ideas on how scholars tackled similar issues.

Note that we will use a dataset about cars, a survey conducted among car dealerships in the DC area. To open this dataset, go to File/open/data/SPSS/cars (the dataset is located in the D: drive, in the folder Program Files).

1 Measures of Central Tendency

Measures of central tendency provide information about the most typical or average values of a variable. We will present three of them: the mean, the average, and the mode.

1.1 The Mean

The **mean** is defined as the sum of a series of observations divided by the number of observations in the series. It is commonly used to describe the central tendency of variables.

To compute the mean using SPSS, go to Analyze/Descriptive Statistics/Descriptives. Then, select the variable you are interested in, say the horsepower, click Options, check the mean box, then Continue and OK.

1.2 The Median

A limitation of the mean as an indicator of central tendency is that its value is greatly affected when a few observations have very large or very low values. The **median** is the middle value in a series of values. It is the observation that divides the sample into two sub-samples of the same size. The median should always be used when your sample contains a relatively small number of observations and/or when a few very large – or small – values affect estimates of the mean.

To compute the mode, go to Analyze/Descriptive Statistics/Frequencies. Then, select the variable you are interested in, say the horsepower, click Statistics, check the median box, then Continue and OK.

1.3 The Mode

The **mode** is defined as the most frequent value of a variable. This indicator might convey more information about the central tendency of a series when variables have certain values that are much more frequent than the others. Assume you analyze a small sample of AU graduate students composed of 50 % of females and 50 % of males. Many of the females have a GRE score of 1800 and many of the males 1600. The mean – around 1700 – is not an appropriate indicator of the central tendency of our series in this case.

To compute the mode, go to Analyze/Descriptive Statistics/Frequencies. Then, select the variable you are interested in, say the horsepower, click Statistics, check the mode box, then Continue and OK.

2 Measures of Dispersion

Measures of dispersion provide information about the distribution of the values of a variable. They tell us how widely values are dispersed around their measures of central tendency. We will present four of them: histograms, the standard deviation, the skewness and the kurtosis.

2.1 Histograms

Histograms show the number of observations in each category. They are very useful because they give a quick visual of the central tendency, the extent of dispersion, and also whether any unusually large or small observations are present.

To draw a histogram, go to Graphs/Histogram... Then, select the variable you are interested in, say the horsepower, and OK.

2.2 The Standard Deviation

The **standard deviation** is a measure of dispersion that is calculated based on the values of the data. It allows us to see how widely the data are dispersed around the mean. The standard deviation has the desirable property that, when the data are normally distributed, 68.3 % of the observations lie within +/- 1 standard deviation from the mean, 95.4% within +/- 2 standard deviations from the mean and 99.7 % within 3 standard deviations from the mean.

To compute the standard deviation, go to Analyze/Descriptive Statistics/Descriptives. Then, select the variable you are interested in, say the horsepower, click Options, check the standard deviation box, then Continue and OK.

2.3 Skewness and Kurtosis

To state that the data are normally distributed simply means that the distribution of the data resembles a bell-shaped curve; in such a case, most of the observations are clustered around the mean. In reality, it is rare to find data that is perfectly normally distributed, but they might appear to be somehow close to a normal distribution. Two statistics will help us determine whether this is the case.

Skewness is a measure of whether the peak is centered in the middle of the distribution. A positive value means that the peak is off to the left, and a negative value suggests that it is off to the right.

Kurtosis is a measure of the extent to which data are concentrated in the peak versus the tail. A positive value indicates that data are concentrated in the peak; a negative value indicates that data are concentrated in the tail.

Values of skewness and kurtosis have little inherent meaning, other than large values indicate greater asymmetry. A rule of thumb is that the absolute value of the ratio of skewness to *its* standard error and of kurtosis to *its* standard error should be less than 2 (N.B: when you compute this ratio, be careful to use the standard error of the skewness and the kurtosis which are computed by SPSS and *not* the standard deviation of the variable). Large ratios indicate departure from symmetry.

To compute the skewness and kurtosis, go to Analyze/Descriptive Statistics/Frequencies. Then, select the variable you are interested in, say the horsepower, click Statistics, check the skewness and kurtosis boxes, then Continue and OK.