School of Oriental and African Studies **September 2013**
Department of Economics

Preliminary Statistics
Lecture 1: Descriptive Statistics (Outline)

1 Introduction

Descriptive Statistics deals with ways of summarising or presenting information from a set of data. The information can be obtained using numerical or graphical techniques.

Data tend to come in three main forms:

- time-series: the same data are collected repeatedly over a number of time periods.
- cross-section: data are collected from the elements of the sample at one point in time.
- panel/longitudinal: the same data are collected from the same elements of the sample over a period of time.

Time-series data have a natural order, 1998 comes after 1997; cross-section data do not have a natural order; the observations (e.g. countries, households) could be ordered alphabetically, by size or any other way.

2 Summary Statistics

2.1 Measures of Central Tendency

Suppose we have a sample of n data, $x_1, x_2, ...x_n$, drawn from a population.

Mean

The arithmetic mean (average) of x_i, usually denoted by a bar over the variable, is defined as

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

In the case of grouped data, the conventional formula needs to be revised since we do not have the values of all the observations, rather we only know the number of observations in each class interval (frequency).

Median

The median is the middle value of the sample (distribution) or the 50th percentile. It is be calculated by ordering the observations in an ascending (or descending) order, and finding the midpoint. $\frac{n+1}{2}$ gives the *location* of the median. When the sample size is odd, then the median is the value of the middle observation, and when the sample size is even, the median is the average of the values of the two middle observations.

A generalisation of the the idea of the median generates the *quartiles*, which divide the distribution into four parts.

Mode

The mode is the most frequently observed value in the sample. A distribution/ sample can have no mode, one mode or many modes.

2.2 Measures of Dispersion

Range and IQR

- The range is the simplest measure of dispersion, which is the difference between the smallest and the largest observation.
- An improvement to the range measure is the inter-quartile range (IQR) which is the difference between the first and third quartiles. IQR is calculated as: Upper Quartile - Lower Quartile= range of the mid 50% of the distribution.

Variance and Standard Deviation

- The variance is a more useful measure of how spread out the observations are, which makes use of all of the information available. It is the average of the squared deviations from the mean. One estimator of the variance, usually denoted $\widehat{\sigma}^2$ is defined as

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 / n$$

  We distinguish between the true value σ^2 (population variance) and our estimate of it $\widehat{\sigma}^2$. You should also be familiar with the unbiased estimator for the true variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ (sample variance). A more dispersed distribution will tend to have larger deviations from the mean and hence a larger variance.

- The Standard Deviation (SD) is the square root of the variance.

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} \text{ and } s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Coefficient of Variation

The coefficient of variation is a relative measure of dispersion, hence unit free. It represents the size of dispersion relative to the mean.

$$\text{Coefficient of Variation } = \frac{\widehat{\sigma}}{\bar{x}}$$

2.3 Measures of Shape

Skewness

Skewness gives a numerical indication of how asymmetric a distribution is. One measure of skewness, known as the coefficient of skewness, reads

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- Coef. of Skewness = 0, the distribution (sample) is symmetrical
- Coef. of Skewness > 0, the distribution is skewed to the right (long right tail)
- Coef. of Skewness < 0, the distribution is skewed to the left (long left tail).

Kurtosis

Kurtosis is a measure for the peakedness of a distribution. The formula for excess kurtosis is

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4-3$$

- Excess Kurtosis = 0, no kurtosis (the distribution is normal)
- Excess Kurtosis > 0, leptokurtic distribution (more acute peak around the mean and fatter tails)
- Excess Kurtosis < 0, platykurtic distribution (smaller peak around the mean and thin tails)

2.4 Measures of Association

Covariance and Correlation

The covariance, which is used to measure how two variables (samples) vary together, is

$$Cov(x,y)=\sum_{i=1}^{n}(x_i-\overline{x})\,(y_i-\overline{y})\,/n$$

- $Cov > 0$, if high values of x are associated with high values of y, that is both variables move to the same direction.
- $Cov < 0$, if high values of x are associated with low values of y, hence the variables move to different directions.
- $Cov = 0$, if there is no *linear* relationship between the variables

The covariance is often standardised to give the (Pearson's) correlation coefficient,

$$r=\frac{Cov(x,y)}{SD(x)SD(y)}$$

The correlation coefficient lies between plus and minus one. A correlation coefficient of -1 means that there is an exact negative linear relation between the variables, $+1$ an exact positive linear relation, and 0 no linear relation.

Useful Transformation

A very useful transformation of data is achieved by substracting the mean and dividing by an estimate of the standard deviation of the sample,

$$z_i=\frac{x_i-\bar{x}}{s}$$

The new (standardised) variable, z_i, has mean zero and variance (and standard deviation) of one. Notice that the correlation coefficient is the covariance between the standardised measures of x and y.

3 Graphical Techniques

Apart from numerical techniques, we can use graphs to describe our data. The most useful type of graph will depend on the nature of the data and the purpose of the exercise.

3.1 Time-series data

- Line (or time-series) graph: The series is plot against time. We can then look for trends (general tendency to go up or down); regular seasonal or cyclical patterns; and outliers (unusual events like wars or crises).

- Histogram: A histogram gives the number (or proportion) of observations which fall in a particular range. It is particularly good for both time-series and cross-sectional data. This type of graph provides useful information regarding the shape of the distribution or sample. We can see if the sample is unimodal or bymodal, if it is symmetric or skewed and so on.

- Scatter (or XY) plot: We can plot one variable against another to see if they are associated, this is a scatter diagram or X-Y Plot. It is used to show if the two variables in question are positively or negatively correlated, and provides similar information to the covariance or correlation measures.

3.2 Cross-section data

- Bar Chart: The bar chart shows the number (frequency) or proportion of the data falling into a particular range or category. The width of each category is either the same or unimportant. So, the height of each column shows the number of the observations in that category.

- Histogram: If the categories or the ranges the data are split are of equal size, then the histogram and the bar chart are the same. However, when this is not the case, the histogram corrects for the different width size.

- Box Plot: A box plot provides information about the median, the quartiles and the minimum and maximum values of a sample.

M. Barrow, *Statistics for Economics, Accounting and Business Studies* Ch. 1 provides a good presentation of numerical and graphical techniques with the use of many examples and graphs.