1 Introduction and descriptive statistics

1.1 Data

Data is a collection of observations about the world. It is the basis of *empirical* knowledge. There are various ways of classifying data - quantitative and qualitative; cross-sectional, time series and panels; primary, secondary and meta; surveys and censuses and so on. But before we get into these classifications, we can ask the question - why is data important? Data is important to test theories, but it can be useful even before theories are formed. Data can provide some unexplained facts that theories are then formed around. Eg. Phillips curve, Kuznets curve, the speed of light.

1.2 Statistics

Statistics is the systematic study of data. It is a way of getting information out of large and complex sets of data. Eg. getting literacy rate from the census, or using wage data to see if implementation of NREGA led to a change in agricultural wages and if yes, then by how much. Statistics is broadly classified into two categories - descriptive and inferential. In general we can think of descriptive statistics as describing the characteristics of data on hand and inferential statistics as using the data from a sample to make inferences about the larger population.

1.3 Descriptive statistics

Below are some frequently used descriptive statistics. Some of these you would have seen before and some would be new.

Mean: For a series of observations $x_i,\ i = 1, 2, ..., n$, the mean is given by $\frac{\sum_{i=1}^{n} x_i}{n}$.

Median: For a series of observations $x_i,\ i = 1, 2, ..., n$ arranged in ascending order, the median is given by the $\frac{n+1}{2}^{th}$ observation if n is odd and by the average of the $\frac{n}{2}^{th}$ and the $\left(\frac{n}{2}+1\right)^{th}$ observations, if n is even.

Percentile: For a series of observations $x_i,\ i = 1, 2, ..., n$ arranged in ascending order, the p^{th} percentile is calculated as follows.

First we calculate $j = \frac{pn}{100} + 0.5$. If j is an integer, then x_j is the p^{th} percentile. If j is not an integer, then let j^k and j^f be the integer and fractional parts of j respectively. The p^{th} percentile is the weighted average of x_{j^k} and x_{j^k+1} given by $(1 - j^f)x_{j^k} + j^f(x_{j^k+1})$. Note that the 50^{th} percentile is the median.

Variance and standard deviation: For a series of observations $x_i,\ i = 1, 2, ..., n$, the sample variance is given by $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, where $\bar{x}$ is the sample mean. The standard deviation is the positive square root of the variance. It is supposed to represent the average deviation from the mean. Then why are we dividing by $n - 1$ instead of by n?

To understand this, we will have to put one foot into inferential statistics. We will be anticipating things that we are going to be seeing in much more details later. The key point is that when we take a sample of n observations from a much larger population, the sample mean calculated from these observations is not necessarily equal to the actual mean. Variance calculated as the average of the squared deviations from the sample mean (implying division by n) will always be smaller than the actual variance from the population mean.Let us suppose that the mean number of notebooks used by students in Bangalore is μ. Now, if we take a sample of 4 observations - x_1, x_2, x_3, x_4, we get a sample mean $\bar{x}$ of $\frac{1}{4}\sum_{i=1}x_i$, which may be close but is not equal to the actual mean. Now, the sum of the squared deviations from the population mean μ is given by

$$\sum_{i=1}^{n}(x_i - \mu)^2 = (x_1 - \mu)^2 + (x_2 - \mu)^2 + (x_3 - \mu)^2 + (x_4 - \mu)^2$$

We can write each of the four terms as follows.

$$(x_i - \mu)^2 = (x_i - \bar{x} + \bar{x} - \mu)^2 = (x_i - \bar{x})^2 + (\bar{x} - \mu)^2 + 2(x_i - \bar{x})(\bar{x} - \mu)$$

Now if we add them all up

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - \mu)^2 + \sum_{i=1}^{n}2(x_i - \bar{x})(\bar{x} - \mu)$$

In the last term we can take $2(\bar{x} - \mu)$ common and bring it outside the summation sign. Hence,

$$\sum_{i=1}^{n}2(x_i - \bar{x})(\bar{x} - \mu) = 2(x_i - \bar{x})\sum_{i=1}^{n}(\bar{x} - \mu) = 0$$

Therefore,

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - \mu)^2$$

Since, $\sum_{i=1}^{n}(\bar{x} - \mu)^2 > 0$ as long as $\bar{x} \neq \mu$, regardless of whether $\bar{x}$ is more or less than μ, therefore $\sum_{i=1}^{n}(x_i - \mu)^2 > \sum_{i=1}^{n}(x_i - \bar{x})^2$. To compensate for this underestimation of the population variance, the sample variance is inflated by using $n - 1$ instead of n as the denominator. This is called Bessel's correction.

Other descriptive statistics that involve a single variable include measures of inequality like the Gini coefficient, and measures of the skewness of the distribution. Now we will look at a couple of descriptive statistics that involve more than one variable.

Covariance and correlation: For n observations on two variables x and y, the sample covariance is given by $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the respective sample means. Covariance is meant to indicate how to variables move together. A positive covariance indicates that when one variable is higher than its mean, very often the other variable is also higher than its mean. A negative covariance indicates just the opposite, i.e. when one variable is higher than its

mean, very often the other variable is lower than its mean. A covariance close to zero indicates the absence of any strong relationship between the two variables. Note that when the two variables are identical then the covariance between them equals their variance.

The magnitude of the sample covariance depends not only on the relationship between the two variables but also on how spread out the observations of those variables are, i.e. the variance of those variables. Correlation is a measure that tries to correct for this. Correlation is measured by the coefficient of correlation (r). which is given by the following formula -

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

We can also write r in terms of the covariance, denoted as s_{xy} and the standard deviations of x_i and y_i denoted as s_x and s_y respectively. Using this notation, $r = \frac{s_{xy}}{s_x s_y}$.

Note that when thw two variables are identical, the correlation coefficient is 1, indicating perfect correlation. A value of -1 would indicate perfect negative correlation, and 0 would indicate no correlation.

Other descriptive statistics that use more than one variable include composite indices, like the Human Development Index, and the Herfindahl index, that measures market competition.

1.4 Graphical representation of data / data visualisation

Histogram: A histogram is one of the most common way to graphically represent data on a single variable. A histogram divides the range of the variable into a number of 'bins', very often of equal width, and then represents the number of observations in each bin by the height of a vertical bar, whose width represents

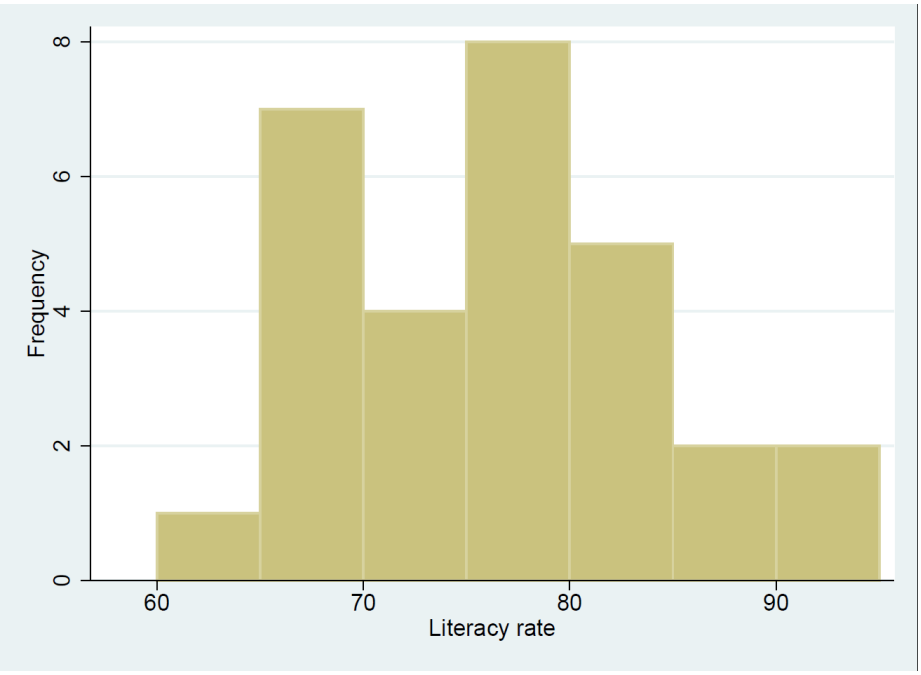

Figure 1: Histogram of literacy rates in the states of India

Source: Census of India, 2011

the width of the bin. For example, Figure 1 shows the histogram of the literacy rates in the various states of India according to the 2011 Census.

Scatter plot: Scatter plots are used to show the relationships between two variables. The two variables are plotted as the axes of a two-dimensional graphs and individual observations are plotted as points on the graph. The sign of the slope of a straight line fitted through these points indicates the sign of the correlation coefficient. How the line is to be fitted is the subject of subsequent lectures. Figure 3 shows a scatter plot between urbanisation and literacy rates.

Trends: Trends are a particular kind of scatter plots, where one of the variables is time. The points indicating the values of the other variable, which is the variable of interest, are often joined with lines to indicate the movement of the variable with time. We ca often show two or more trends on the same graph to compare the movements of different variables over time.

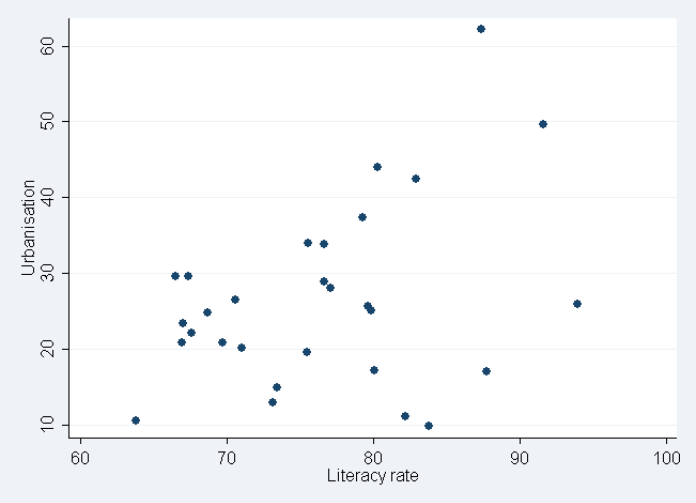

Figure 2: Scatter plot of urbanisation versus literacy rates in the states of India

Source: Census of India, 2011

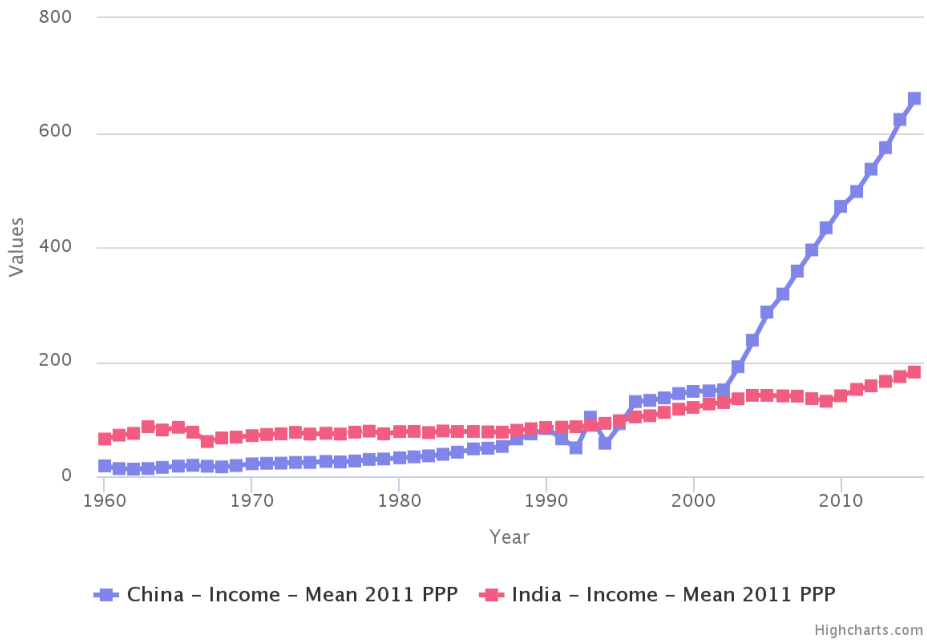

Figure 3: Trends in mean per capita income in India and China

Source: Global consumption and income project, gcip.info

Data visualisation using techniques more advanced than the ones described here, has become a very useful tool to convey large amounts of complicated data in simple and easy to understand forms. Some examples will be shared in class.

Resources

1. Global Consumption and Income Project - gcip.info
2. OECD better life index - www.oecdbetterlifeindex.org
3. Census of India - censusindia.gov.in