Lecture One:

Descriptive statistics is a process of concisely summarising the characteristics of sets of data.
Inferential statistics involves constructing estimates of these characteristics, and testing hypotheses about the world, based on sets of data.
Modelling and analysis combines these to build models that represent relationships and trends in reality in a systematic way.

Types of Data:
Numerical or *quantitative* data are real numbers with specific numerical values.
Nominal or *qualitative* data are non-numerical data sorted into categories on the basis of qualitative attributes.
Ordinal or *ranked* data are nominal data that can be ranked.

- The *population* is the complete set of data that we seek to obtain information about
- The *sample* is a part of the population that is selected (or sampled) in some way using a *sampling frame*
- A characteristic of a population is call a *parameter*
- A characteristic of a sample is called a *statistic*
- The difference between our estimate and the true (usually unknown) parameter is the *sampling error*
- In a *random sample*, all population members have an equal chance of being sampled

In a population, a perfect *strata* would be a group with:
- individual observations that are similar to the other observations in that strata
- different characteristics from other strata in the population
- stratified sampling can improve accuracy
- may be more costly

In a population, a perfect *cluster* would be a group with:
- individual observations that are different from the other observations in that cluster
- similar characteristics to other clusters in the population
- can reduce costs
- may be less accurate

This is the cost/accuracy trade off

Lecture Two:

Ceteris paribus: the assumption of holding all other variables constant

Cross-sectional data are:
- collected from (across) a number of different entities (such as individuals, households, firms, regions or countries) at a particular point in time
- usually a random sample (but not always)
- not able to be arranged in any “natural” order (we can sort or rank the data into any order we choose)
- often (but not only) usefully presented with histograms

Time series data are:
- collected over time on one particular ‘entity’
- data with observations which are likely to depend on what has happened in the past

- data with a natural ordering according to time
- often (but not only) presented as line charts

Lecture Three:

Measures of Centre:

Mean/Average: population: μ sample: $\bar{x} = \frac{\Sigma x}{n}$

- easy to calculate
- sensitive to extreme observations

Median: middle number, or average of two middle numbers

- not sensitive to extreme observations

Mode: most frequently occurring number

- only used for finding most common outcome

$\bar{x} - \mu$ = *sampling error*

If a distribution is uni-modal then we can show that it is:

- Symmetrical if mean = median = mode
- Right-skewed if mean > median > mode
- Left-skewed if mode > median > mean

Measures of Variation:

Population variance measures an average of the squared deviations between each observation and the population mean: $\sigma^2 = \frac{1}{N}\sum(x_1 - \mu)^2$

Population standard deviation is the square root of population variance: $\sigma = \sqrt{\frac{1}{N}\sum(x_1 - \mu)^2}$

Sample variance measures the average of the squared deviations between each observation and the sample mean: $s^2: \frac{1}{n-1}\sum(x_1 - \bar{x})^2$

Sample standard deviation is the square root of sample variance: $s = \sqrt{\frac{1}{n-1}\sum(x_1 - \bar{x})^2}$

C*oefficient of variation* measures the variation in a sample (given by its standard deviation) relative to that sample's mean, it is expressed as a percentage to provide a unit-free measurement, letting us compare difference samples: $CV = 100 \times \frac{s}{\bar{x}}\%$

Lecture Four:

Measures of Association:

Covariance measures the co-variation between two sets of observations.

With a population size N having observations (x_i, y_i), (x_2, y_2), (x_N, y_N) etc. and having μ_x, μ_y, being the respective means of the x_i and y_i terms, covariance is calculated as,

$$COV(X,Y) = \frac{1}{N}\sum_{i-1}^{N}(x_i - \mu_x)(y - \mu_y)$$