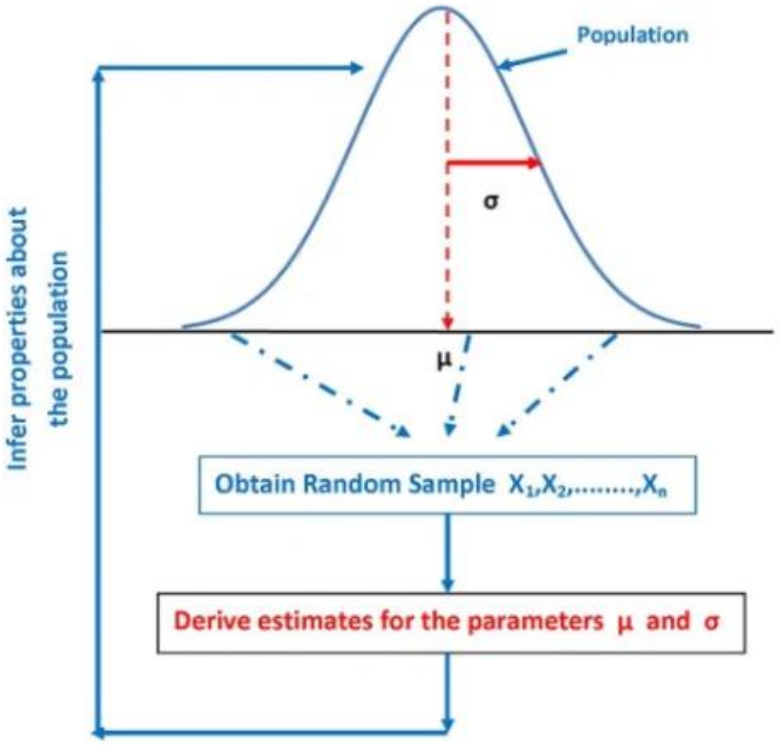
Statistical Inference

Inferential cycle

Schematically the inferential cycle is illustrated below.



The population represents ALL items of interest. It could be the weights of people in the UK, or the ages of such people etc. This population will be characterised by parameters, for example, the mean and variance. Identifying these parameters will enable calculations to be performed on the population such as finding the proportion of the population with weights in excess of 15 stone, or the proportion of people above a certain age say. This process is summarised below

Stage 1:- Obtain an appropriate random sample from the population of interest.

Stage 2:- From your sample estimate the parameters of interest.

Stage 3:- From your estimates infer properties of the population. There could be more than one population and we may wish to compare them.

These stages highlight the three main areas of inference, i.e., Methods of drawing random samples; Estimation procedures for parameters that define the population and; Inference and hypothesis testing.

Sampling

It is clear from the outset that we need to be absolutely clear about the difference between a population and a sample.

Population This is the complete set of members of a group. It could be the population of the UK, if we were interested in demographic trends, it could be the 52 cards in a complete pack of playing cards, if we were interested in the probabilities of certain hands of cards appearing, etc.
A sample This is a subset of the members of a population chosen by some method.

We have to know the rules for sampling before we can establish probabilities of events. We start with some basic methods of sampling.

Census This is when the sample taken is the whole population. It is quite common that the population has a very large number of members which makes it impossible to follow this approach. We do adopt this approach in the UK once every 10 years when the census is taken of all households, but it entails a massive effort at great cost. The benefits from a census is that you have fact, and thereby removes the influence of estimation error.

Simple random sample of size n from a finite population This is a selection of n members of the population made in such a way that all possible selections of size n are equally likely. Equivalently each time a member is to be chosen, all members eligible for selection have an equal probability of being chosen.
Note This method of selection covers sampling both with and without replacement.

Systematic sampling Suppose the number in the population N is reasonably large and that a sample of size n is required. The **sampling interval SI** is defined as $SI = \frac{N}{n}$. Suppose SI is an integer, then we choose an integer at random in the interval [1,SI], J say, and we take as our sample the Jth, (J+SI)th, (J+2SI)th, ……., (J+(n-1)SI)th members of the population. If SI was non-integer then the numbers $kSI \;\; k = 1, (n-1)$ would be rounded down to the nearest integer.

Warning If there is some underlying structure, such as classes in a school, with classes sizes of SI, then the above rule could pick out the second person listed on each class register say. If the class lists are ordered first by gender and then alphabetically this process could result in a sample of the same sex – hardly representative and a disadvantage.

Cluster sampling If a small sample is required from a large population, the cost, including time, of obtaining a list of all members of the population may be prohibitively large. If the population falls naturally into small subsets, called clusters, which are each representative of the population, then a different approach can be used. For example, a secondary school could be a cluster within the population of all secondary schools. The idea then is to select, either at random or systematically, a set of clusters. A random sample from each cluster can then be obtained, either purely randomly or systematically. Sometimes, if the clusters are not too large, all members of each cluster can be used.

Opportunity sampling This is based on a sample selection that is done for convenience, i.e., standing on a street corner and interviewing people as they turn up.

Quota sampling and stratified sampling In both of these techniques the population is divided into exclusive groups(strata), where members within each group have similar

characteristics. The sampling protocol specifies the number required from each group. For stratified sampling the sample from each group is obtained by taking a simple random sample from the group so that each member of the group has an equal probability of being selected. For quota sampling the person selecting the group sample can use opportunity sampling techniques as long as the correct number is sampled.

Sampling from a distribution

This is very important as it underpins many computer simulation models of real word events. Let X be a random variable. A **random sample from the distribution of X** is a set of independent random variables $X_1, X_2, \ldots\ldots, X_n$, each with the same distribution as X, called the **parent distribution**.

Example Computers can generate independent uniform random variables on the interval [0,1]. You could define X_i to be 1 if the generated uniform variable lies in the interval [0,p] and 0 otherwise. Thus

$$P(X_i = 1) = p \quad \text{and} \quad P(X_i = 0) = 1 - p$$

The $X_i's$ are independent and $T = \sum_{i=1}^{n} X_i$ is the total number of successes ($X_i = 1$) in n independent trials with constant probability of success p. This is the condition necessary to generate a Binomial variable and so $T \sim B(n, p)$. Random samples from many other statistical distributions can be constructed in similar ways.