Chapter 1 - Introduction to Statistics:

Key Words -

Data are collections of observations (such as measurements, genders and survey responses).

Statistics is the science of planning studies and experiments, obtaining data, and then organising, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

A population is the complete collection of all individuals (scores, people, measurements, and so on) to be studied. The collection is complete in the sense that it includes all of the individuals to be studied.

A sample is a subcollection of members selected from a population.

A census is the collection of data from every member of the population.

Parameter - is a numerical measurement describing some characteristic of a population.

Eg: Parameter: There are exactly 100 Senators in the 109th Congress of the United States, and 55% of them are Republicans. The figure of 55% is a parameter because it is based on the entire population of all 100 Senators.

Statistic - is a numerical measurement describing some characteristic of a sample.

Eg: Statistic: In 1936, Literary Digest polled 2.3 million adults in the United States, and 57% said that they would vote for Alf Landon for the presidency. That figure of 57% is a statistic because it is based on a sample, not the entire population of all adults in the United States.

Types of Data:

1. Qualitative (or Categorical or attribute) data consists of names or labels that are not numbers representing counts or measurements. (Mathematical operations are meaningless.)

Eg:

- Color, Race, Gender, Religion and Zipcode.
- The political party affiliations (Democrat, Republican, Independent, other) of survey respondents
- The numbers 24, 28, 17, 54, and 31 are sewn on the shirts of the LA Lakers starting basketball team. These numbers are substitutes for names. They don't count or measure anything, so they are Qualitative/categorical data.

2. Quantitative (or numerical) data consist of numbers representing counts or measurements. (Mathematical operations are meaningful.)

Eg:

- The ages (in years) of survey respondents
- Height, Weight, Wages, Kilometer per hour, Temperature, Time.

Quantitative data can be further described by distinguishing between discrete and continuous types.

1. Discrete data result when the number of possible values is either a finite number or a "countable" number. (That is, the number of possible values is 0 or 1 or 2, and so on.)

Eg: The numbers of eggs that hens lay are discrete data because they represent counts.

2. Continuous (numerical) data result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps.

Eg: The amounts of milk from cows are continuous data because they are measurements that can assume any value over a continuous span. During a year, a cow might yield an amount of milk that can be any value between 0 and 7000 litres. It would be possible to get 5678.1234 litres because the cow is not restricted to the discrete amounts of 0, 1, 2, . . . , 7000 litres.

Levels of Measurement:

Another common way of classifying data is to use **four levels of measurement**: nominal, ordinal, interval, and ratio.

1. Nominal level of measurement is characterized by data that consist of names, labels, or categories only. The data cannot be arranged in an ordering scheme (such as low to high).

Eg: (a) Yes no undecided: Survey responses of yes, no, and undecided.

(b) Political Party: The political party affiliations of survey respondents (Democrat, Republican, Independent, other)

2. Data are at the **ordinal level of measurement** if they can be arranged in some order, but differences (obtained by subtraction) between data values either cannot be determined or are meaningless.

Eg: (a) Course Grades: A college professor assigns grades of A, B, C, D, or F. These grades can be arranged in order, but we can't determine the differences between the grades. For example, we know that A is higher than B (so there is an ordering), but we cannot subtract B from A (so the difference cannot be found).

(b) Ranks: U.S. News and World Report ranks colleges. Those ranks (first, second, third, and so on) determine an ordering. However, the differences between ranks are meaningless. For example, a difference of "second minus first" might suggest 2 - 1 = 1, but this difference of 1 is meaningless because it is not an exact quantity that can be compared to other such differences. The difference between Harvard and Brown cannot be quantitatively compared to the difference between Yale and Johns Hopkins.

3. The interval level of measurement is like the ordinal level, with the additional property that the difference between any two data values is meaningful. However, data at this level do not have a natural zero starting point (where none of the quantity is present).

Eg: (a) Temperatures: Body temperatures of 98.2F and 98.6F are examples of data at this interval level of measurement. Those values are ordered, and we can determine their difference of 0.4F. However, there is no natural starting point. The value of 0F might seem like a starting point, but it is arbitrary and does not represent the total absence of heat.

(b) Years: The years 1492 and 1776. (Time did not begin in the year 0, so the year 0 is arbitrary instead of being a natural zero starting point representing "no time.")

4. The ratio level of measurement is the interval level with the additional property that there is also a natural zero starting point (where zero indicates that none of the quantity is present). For values at this level, differences and ratios are both meaningful.

Eg: (a) Distances: Distances (in km) travelled by cars (0 km represents no distance travelled, and 400 km is twice as far as 200 km.)

(b) Prices: Prices of college textbooks ($0 does represent no cost, and a $100 book does cost twice as much as a $50 book.)

Ratio:	There is a natural zero starting point and ratios are meaningful.	Example: Distances, Prices.
Interval:	Differences are meaningful, but there is no natural zero starting point and ratios are meaningless.	Example: Body temperatures in degrees Fahrenheit or Celsius
Ordinal:	Categories are ordered, but differences can't be found or are meaningless.	Example: Ranks of colleges in U.S. News and World Report
Nominal:	Categories only. Data cannot be arranged in an ordering scheme.	Example: Eye colours, Political party, Yes/No

The Statistical methods are driven by the data that we collect. We typically obtain data from two distinct sources: observational studies and experiments.

Observational Study: In an observational study, we observe and measure specific characteristics, but we don't attempt to modify the subjects being studied.

Eg: A good example of an observational study is a poll in which subjects are surveyed, but they are not given any treatment. The Literary Digest poll in which respondents were asked who they would vote for in the presidential election is an observational study. The subjects were asked for their choices, but they were not given any type of treatment.

Experiment: In an experiment, we apply some treatment and then proceed to observe its effects on the subjects. (Subjects in experiments are called experimental units.)

Eg: In the largest public health experiment ever conducted, 200,745 children were given a treatment consisting of the Salk vaccine, while 201,229 other children were given a placebo.

The Salk vaccine injections constitute a treatment that modified the subjects, so this is an example of an experiment.

Collecting Sample Data:

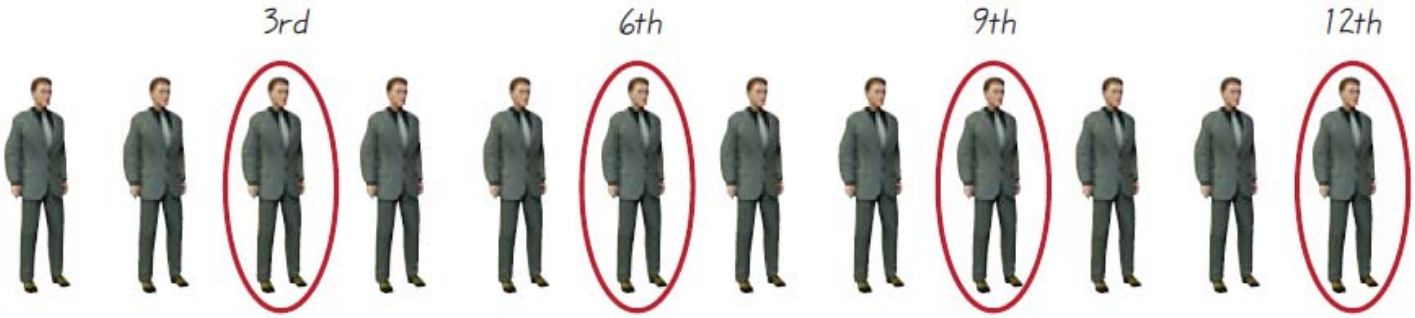
Simple random sample: A simple random sample of n subjects is selected in such a way that every possible sample of the same size n has the same chance of being chosen.

Random sample: In a random sample members of the population are selected in such a way that each individual member in the population has an equal chance of being selected.

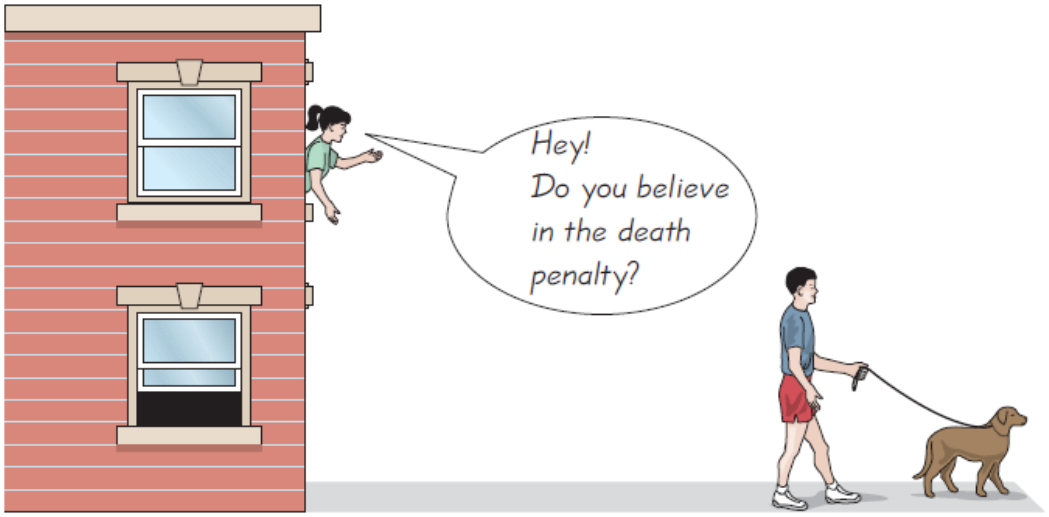
Probability sample: A probability sample involves selecting members from a population in such a way that each member of the population has a known (but not necessarily the same) chance of being selected.

Other Sampling Methods:

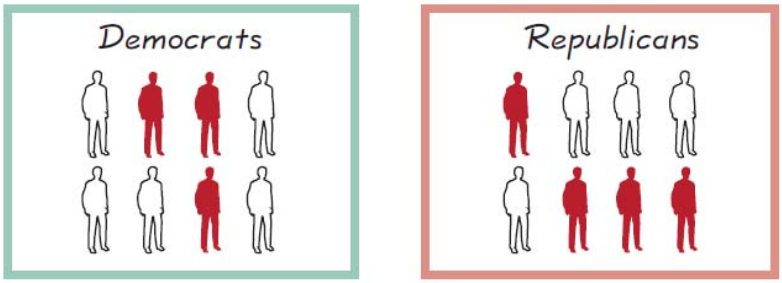
Systematic sampling: In systematic sampling, we select some starting point and then select every kth (such as every 50th) element in the population.



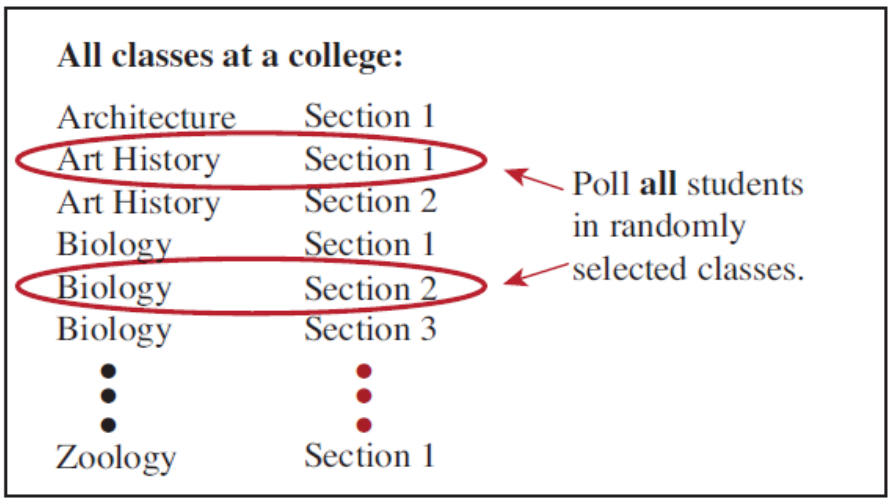
Convenience sampling: With convenience sampling, we simply use results that are very easy to get.



Stratified sampling: With stratified sampling, we subdivide the population into at least two different subgroups (or strata) so that subjects within the same *subgroup share the same characteristics* (such as gender or age bracket), and then we draw a sample from each subgroup (or stratum).



Cluster sampling: In cluster sampling, we first divide the population area into sections (or clusters), then randomly select some of those clusters, and then choose all the members from those selected clusters.



Sampling Errors: No matter how well you plan and execute the sample collection process, there is likely to be some error in the results. For example, randomly select 1000 adults, ask them if they graduated from high school, and record the sample percentage of "yes" responses. If you randomly select another sample of 1000 adults, it is likely that you will obtain a different sample percentage.

Sampling error: A sampling error is a difference between a sample result and the true population result; such an error results from chance sample fluctuations.

Non-Sampling error: A nonsampling error occurs when the sample data are incorrectly collected, recorded, or analysed (such as by selecting a biased sample, using a defective measurement instrument, or copying the data incorrectly).