**2 Statistics**

**Sampling**

* In statistics, quality assurance, and survey methodology, sampling is the selection of a subset (a statistical sample) of individuals from within a statistical population to estimate characteristics of the whole population.



**Primary data**

* This is data collected by an individual (eg yourself, a company, a business, etc)
* An advantage of using primary data is that researchers are collecting information for the specific purposes of their study. In essence, the questions the researchers ask are tailored to elicit the data that will help them with their study. Researchers collect the data themselves, using surveys, interviews and direct observations.

**Secondary data**

* This is data collected by a third party and used by an individual. (eg CSO website, published data in magazines, online, etc)
* There are several types of secondary data. They can include information from the national population census and other government information collected by Central Statistics Office (www,cso.ie). There are any number of examples: motor vehicle registrations, hospital intake and discharge records, workers’ compensation claims records, and more.

**Raw or Unordered data**

* Raw or unordered data is usually collected randomly in a survey, questionnaire etc.
* Raw data is the data that is collected from a source, but in its initial state. It has not yet been processed, cleaned, or organized. You can find raw data in a variety of places, including databases, files, spreadsheets, etc.

**Ordered Data**

* We normally organise raw data in numerical or alphabetical order, so it is easier to see standout characteristics immediately. For example, when you see 10 numbers in numerical order one can identify the smallest and largest. Ordered data is easier to read which may help make better decisions based on the set of data.

**Mean** (arithmetic mean or average)

Example: Add all numbers and divide by the number of numbers

**Mode** (most common value or the value with the biggest frequency)

Example: Most common age of students in the college.

**Median** (this is the "middle" value in the list of numbers)

* To find the median, your numbers have to be listed in numerical order from smallest to largest, so you may have to rewrite your list before you can find the median.

Example: 1, 4, 5, 7, 9 The median is 5 as it is the number in the middle.

Example: 2, 4, 6, 6, 9, 16 When there are two numbers in the middle.

The media is calcuated as follows: $\frac{6+6}{2}=6$

**Range** (highest x value – lowest x value)

Example: If the oldest student in the college is 58 and the youngest is 18

Then the range is 58 – 18 = 40

Range = 40

**Standard Deviation** (this is a quantity expressing by how much the members of a group differ from the mean value for the group)

Example: Explain based on the age of students in the college for example.

**Formula**



**Variance**

The sample variance, (standard deviation)$$, is used to calculate how varied a sample is. A sample is a select number of items taken from a population. For example, if you are measuring American people’s weights, it wouldn’t be feasible (from either a time or a monetary standpoint) for you to measure the weights of every person in the population. The solution is to take a sample of the population, say 1000 people, and use that sample size to estimate the actual weights of the whole population. The variance helps you to figure out how spread out your weights are.



**Statistics Basics**

* The most common basic statistics terms you’ll come across are the mean, mode and median. These are all what are known as “Measures of Central Tendency.” Also important in this early chapter of statistics is the shape of a distribution. This tells us something about how data is spread out around the mean or median. Perhaps the most common distribution you’ll see is the normal distribution, sometimes called a bell curve. Heights, weights, and many other things found in nature tend to be shaped like this:



Example of a **Normal Distribution**

* A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme. Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.





Measure of Central Tendency





**Bias in Statistics**

* Bias is the tendency of a statistic to overestimate or underestimate a parameter/value/data. Sampling error is the tendency for a statistic not to exactly match the population.
* For example, let’s say you have a population in the United States with an average height of 5 feet 9 inches. If you take a sample, even a fairly sizable sample of say, 10,000 people, it’s unlikely that you’ll get exactly 5 feet 9 inches. You might get very close, perhaps to within a fraction of an inch. If you repeat the experiment, you might get another very close result.
* For example, in experiment 1 you might get 5 feet 8.9 inches and in experiment 2 you might get 5 feet 9.1 inches. The tendency for statistics to get very close, but not exactly right, is called sampling error.
* **Note:** If the statistic is unbiased, the average of all statistics from all samples will average the true population parameter/value/data.

**Example:** Excel file with ages of sample 1000 students in Tramore Road Campus.