Sampling Error

- A **statistical error** to which an analyst exposes a model simply because he or she is working with sample data rather than population or census data.
- Using sample data presents the risk that results found in an analysis do not represent the results that would be obtained from using data involving the entire population from which the sample was derived.

Sampling Error

- The use of a sample relative to an entire population is often necessary for practical or budgetary reasons.
- Although there are likely to be some differences between sample analysis results and population analysis results, the degree to which these can differ is not expected to be substantial.

Sampling Error

- Methods of reducing sampling error include increasing the sample size and ensuring that the sample adequately represents the entire population.

Non-Sampling Error

- A statistical error caused by human error to which a specific statistical analysis is exposed.
- These errors can include, but are not limited to, data entry errors, biased questions in a questionnaire, biased processing/decision making, inappropriate analysis conclusions and false information provided by respondents.

Non-Sampling Error

- Non-sampling errors are part of the total error that can arise from doing a statistical analysis.
- The remainder of the total error arises from sampling error.
- Unlike sampling error, increasing the sample size will not have any effect on reducing non-sanpling error.
- Unfortunately, it is virtually impossible to eliminate non-sampling errors entirely.

Probability Rules

There are two rules which are very important.

- All probabilities are between 0 and 1 inclusive

$$0 \leq P(E) \leq 1$$

- The sum of all the probabilities in the sample space is 1

Probability Rules

- The probability of an event which cannot occur is 0.
- The probability of any event which is not in the sample space is zero.
- The probability of an event which must occur is 1.
- The probability of the sample space is 1.

Probability Rules

The Complement Rule

- The probability of an event not occurring is one minus the probability of it occurring.

$$P(E^C) = 1 - P(E)$$

Probability: Addition Rule for Any Two Events

- For any two events A and B, the probability of A or B is the sum of the probability of A and the probability of B minus the probability of both A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- We subtract the probability of $A \cap B$ to prevent it getting counted twice.
- *($A \cup B$ and $A \cap B$ denotes "A or B" and "A and B" respectively)*

Probability: Addition Rule for Any Two Events

- If events A and B are **mutually exclusive**, then the probability of A or B is the sum of the probability of A and the probability of B:

$$P(A \cup B) = P(A) + P(B)$$

- If A and B are mutually exclusive, then the probability of both A and B is zero.

Coefficient Of Variation (CV)

- This is a statistical measure of the dispersion of data points in a data series around the mean.
- It is calculated as follows:

$$\mathrm{CV} = \frac{\sigma}{\mu}$$

$$\mathrm{CV} = \frac{s}{\bar{x}}$$

- The coefficient of variation represents the ratio of the standard deviation to the mean
- The coefficient of variation is a useful statistic for comparing the degree of dispersion from one data set to another, even if the means are very different from each other.

Data

- Data can be defined as groups of information that represent the qualitative or quantitative attributes of a variable or set of variables, which is the same as saying that data can be any set of information that describes a given entity. Data in statistics can be classified into grouped data and ungrouped data.
- Any data that you first gather is ungrouped data.
- Ungrouped data is data in the raw. An example of ungrouped data is a any list of numbers that you can think of.

Grouped Data

- Grouped data is data that has been organized into groups known as classes.
- Grouped data has been 'classified' and thus some level of data analysis has taken place, which means that the data is no longer raw.
- A data class is group of data which is related by some user defined property.
- For example, if you were collecting the ages of the people you met as you walked down the street, you could group them into classes as those in their teens, twenties, thirties, forties and so on. Each of those groups is called a class.

Class Intervals

- Each of those classes is of a certain width and this is referred to as the **Class Interval**.
- This class interval is very important when it comes to drawing Histograms and Frequency diagrams.
- All the classes may have the same class size or they may have different classes sizes depending on how you group your data.
- The class interval is always a whole number.

Below is an example of grouped data where the classes have the same class interval.

Age (years)	Frequency
0 - 9	12
10 - 19	30
20 - 29	18
30 - 39	12
40 - 49	9
50 - 59	6
60 - 69	0

Solution:

Below is an example of grouped data where the classes have different class interval.

Age (years) Frequency Class Interval 0 - 9 15 10 10 - 19 18 10 20 - 29 17 10 30 - 49 35 20 50 - 79 20 30

Calculating Class Interval Given a set of raw or ungrouped data, how would you group that data into suitable classes that are easy to work with and at the same time meaningful?

The first step is to determine how many classes you want to have.

Next, you subtract the lowest value in the data set from the highest value in the data set and then you divide by the number of classes that you want to have:

Example 1:
Group the following raw data into ten classes.
Solution:
The first step is to identify the highest and lowest number

Class interval should always be a whole number and yet in this case we have a decimal number. The solution to this problem is to round off to the nearest whole number.

In this example, 2.8 gets rounded up to 3. So now our class width will be 3; meaning that we group the above data into groups of 3 as in the table below.

Number Frequency 1 - 3 7 4 - 6 6 7 - 9 4 10 - 12 2 13 - 15 2 16 - 18 8 19 - 21 1 22 - 24 2 25 - 27 3 28 - 30 2

Class Limits and Class Boundaries

- Class limits refer to the actual values that you see in the table.
- Taking an example of the table above, 1 and 3 would be the class limits of the first class.
- Class limits are divided into two categories: lower class limit and upper class limit.
- In the table above, for the first class, 1 is the lower class limit while 3 is the upper class limit.

Class Limits and Class Boundaries

- On the other hand, class boundaries are not always observed in the frequency table.
- Class boundaries give the true class interval, and similar to class limits, are also divided into lower and upper class boundaries.
- The relationship between the class boundaries and the class interval is given as follows:

Class boundaries are related to class limits by the given relationships:
As a result of the above, the lower class boundary of one class is equal to the upper class boundary of the previous class.
Class limits and class boundaries play separate roles when it comes to representing statistical data diagrammatically as we shall see in a moment.

Sampling Distribution

- A probability distribution of a statistic obtained through a large number of samples drawn from a specific population.
- The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

Measures of Centrality

```
4.10, 4.10, 4.25, 4.25, 4.25,
4.35, 4.40, 4.53, 4.90, 5.20,
5.26, 5.35, 5.45, 5.71, 6.09,
6.10, 6.30, 6.50, 6.80, 7.11.
```

- There are 20 values in this data set.
- The sum of the values is 105.

The Mean Absolute Deviation

- The mean absolute deviation, or MAD, is based on the absolute value of the difference between each value in the data set and the mean of the group.
- It is sometimes called the average deviation.
- The mean average of these absolute values is then determined.
- The absolute values of the differences are used because the sum of all of the plus and minus differences (rather than the absolute differences) is always equal to zero.
- Thus the respective formulas for the population and sample MAD are

$$\text{Population MAD } = \frac{\sum |x_i - \mu|}{N}$$

$$\text{Sample MAD } = \frac{\sum |x_i - \bar{x}|}{n}$$