STATISTICS

BY

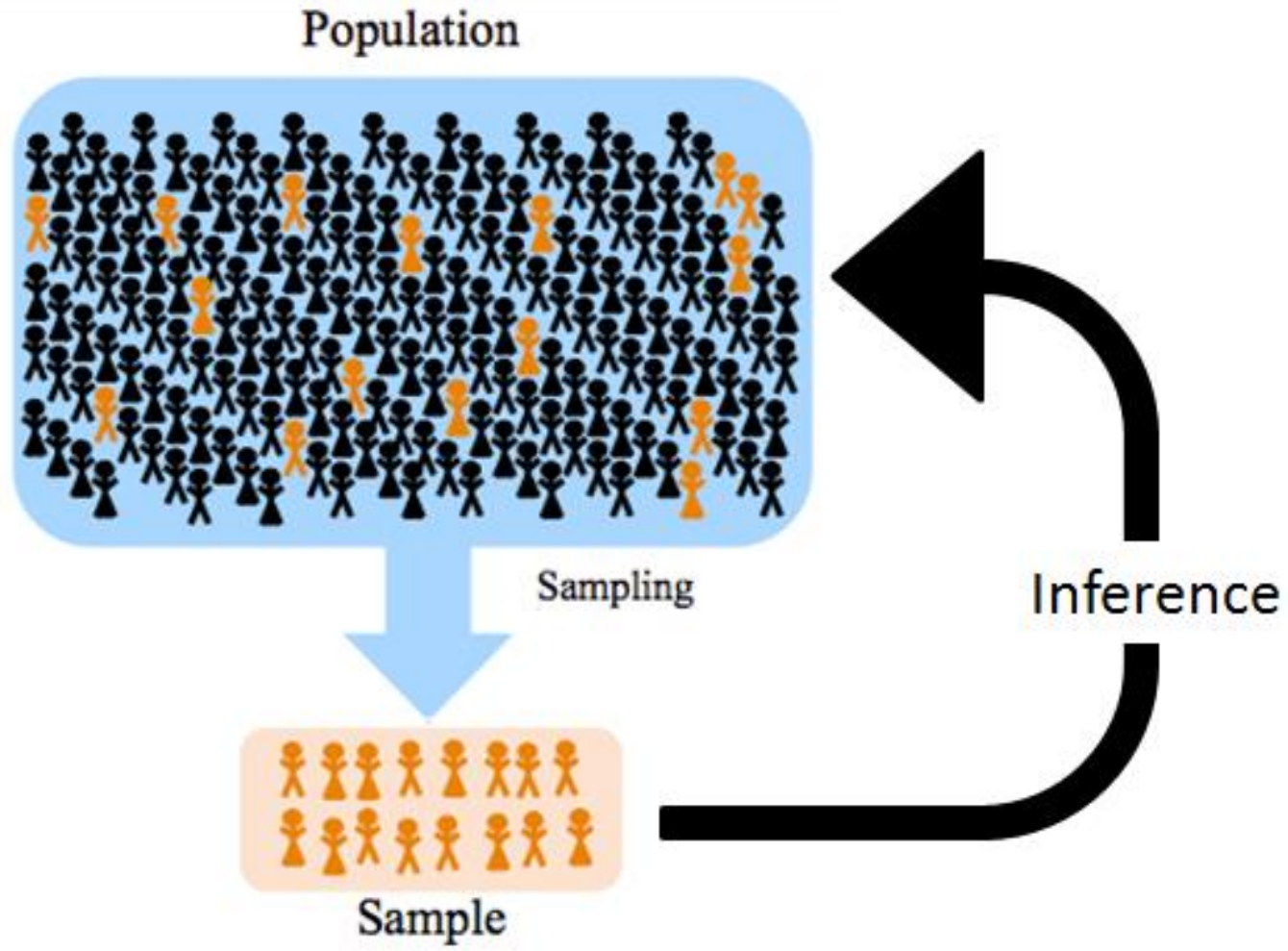
MG ANALYTICS

MG Analytics

STATISTICS

- The science and the art of learning from data.
- It is:
  - the collection, analysis, and interpretation of data
  - The effective communication and presentation of results relying on data.

MG Analytics



MG Analytics

Population Vs Sample

Population	Sample
A **population** includes all of the elements from a set of data.	A **sample** consists one or more observations drawn from the population.
A measurable characteristic of a population, such as a mean or standard deviation, is called a **parameter**	A measurable characteristic of a sample is called a **statistic**.
Reports are a true representation of the opinion.	Reports have a margin of error and a confidence interval.

MG Analytics

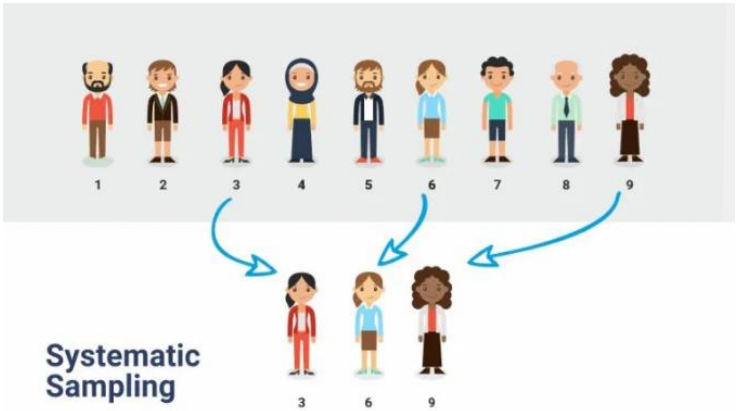
Sampling Techniques

Probability sampling involves random selection, allowing you to make statistical inferences about the whole group.

Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect initial data.

Probability Sampling

Simple Random Sampling

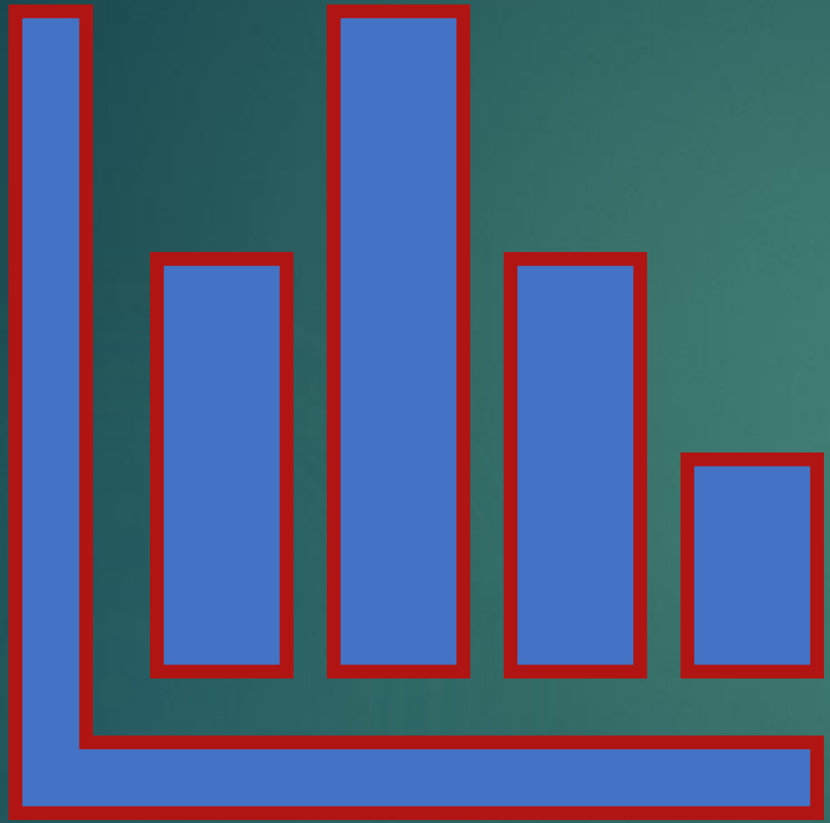


Systematic Sampling

Stratified Random Sampling

Cluster Sampling

MG Analytics



Descriptive Statistics

Descriptive Statistics

- A descriptive statistic is a summary statistic that quantitatively describes or summarizes the data
- while descriptive statistics is the process of using and analyzing those statistics.

MG Analytics

Measures Of central tendency / Dispersion

- Mean: A calculated "central" value of a set of numbers. Add all the numbers and divide by the count of number added.
- Median: The middle number in a sorted, ascending or descending, list of The **median** is sometimes used as opposed to the mean when there are outliers in the sequence that might skew the average of the values.
- Mode: The mode of a set of data values is the value that appears most often. It is the value that is most likely to be sampled.

Odd Values:

1,2,3,
4,5,5,6,7,8,
9,12

- Minimum : 1
- Maximum : 12
- Range : 12-1 = 11
- Mode : 5
- Median : 5
- Mean : (1 + 2 + 3 + 4 + 5 + 5 + 6 + 7 + 8 + 9 + 12)/11 = 5.63

Even values:

1,2,3,
4,5,5,6,7,8,
9,12,100

- Minimum : 1
- Maximum : 100
- Range : 100-1 = 99
- Mode : 5
- Median : (5+6)/2
- Mean : (1 + 2 + 3 + 4 + 5 + 5 + 6 + 7 + 8 + 9 + 12 + 100)/12 = 13.5

Which value to use when?

MG Analytics

Mean: Is highly impacted by outliers

Median: Is robust against outliers

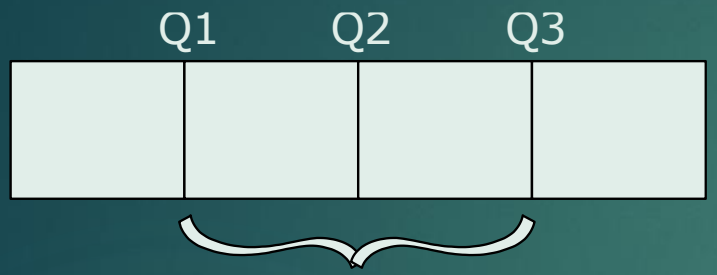
Mode: Is the value that is most likely to be sampled.

Measures of Variability (Spread)

- Mean , Median , Mode can provide information about the central points of the data but do not tell about how data varies.
- Range
- IQR
- Variance
- Standard Deviation

MG Analytics

Quartiles &Inter Quartile Range (IQR)



Q3 –Q1 = IQR = 17-7 =10

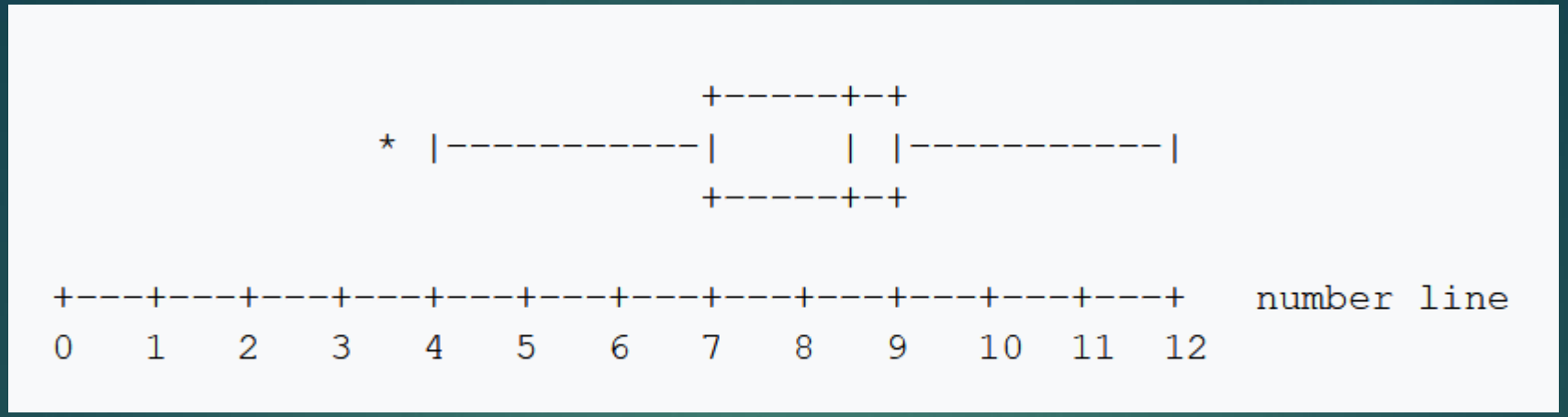
Outliers :

Q1 – 1.5 * IQR = 7-15 = -8

Q3 + 1.5 * IQR = 17+15 = 32

				Q1				Median				Q3			
Values	**3**	**4**	**5**	**7**	**8**	**8**	**9**	**11**	**12**	**14**	**15**	**17**	**18**	**19**	**19**
Count	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

MG Analytics



For the data set in this box plot:

- lower (first) quartile $\mathbf{Q_1} = 7$
- median (second quartile) $\mathbf{Q_2} = 8.5$
- upper (third) quartile $\mathbf{Q_3} = 9$
- interquartile range, **IQR** = $Q_3 - Q_1 = 2$
- **lower 1.5*IQR whisker** = $Q_1 - 1.5 * IQR = 7 - 3 = 4$. (If there is no data point at 4, then the lowest point greater than 4.)
- **upper 1.5*IQR whisker** = $Q_3 + 1.5 * IQR = 9 + 3 = 12$. (If there is no data point at 12, then the highest point less than 12.)

This means the 1.5*IQR whiskers can be uneven in lengths.

MG Analytics

Standard Deviation and Variance

number	Deviation from first number	Deviation from last number	Deviation from mean	Squared Deviation from mean
1	0	15	7.444444444	**55.41975308**
2	-1	14	6.444444444	**41.53086419**
4	-3	12	4.444444444	**19.75308642**
7	-6	9	1.444444444	**2.086419752**
7	-6	9	1.444444444	**2.086419752**
12	-11	4	-3.555555556	**12.64197531**
12	-11	4	-3.555555556	**12.64197531**
15	-14	1	-6.555555556	**42.97530865**
16	-15	0	-7.555555556	**57.08641976**
8.444444444	**-7.444444444**	**7.555555556**	**-4.44445E-10**	**27.35802469**
			Variance	**5.230489909**

MG Analytics

number	Deviation from first number	Deviation from last number	Deviation from mean	Squared Deviation from mean
1	0	58	12.22222222	**149.382716**
2	-1	57	11.22222222	**125.9382716**
4	-3	55	9.222222222	**85.04938272**
7	-6	52	6.222222222	**38.71604938**
7	-6	52	6.222222222	**38.71604938**
12	-11	47	1.222222222	**1.49382716**
12	-11	47	1.222222222	**1.49382716**
15	-14	44	-1.777777778	**3.160493827**
59	-58	0	-45.77777778	**2095.604938**
13.22222222	**-12.22222222**	**45.77777778**	**0**	**282.1728395**
			Variance	**16.79800106**

MG Analytics

number	Deviation from first number	Deviation from last number	Deviation from mean	Squared Deviation from mean
-55	0	72	57.33333333	**3287.111111**
2	-57	15	0.333333333	**0.111111111**
4	-59	13	-1.666666667	**2.777777778**
7	-62	10	-4.666666667	**21.77777778**
7	-62	10	-4.666666667	**21.77777778**
12	-67	5	-9.666666667	**93.44444444**
12	-67	5	-9.666666667	**93.44444444**
15	-70	2	-12.66666667	**160.4444444**
17	-72	0	-14.66666667	**215.1111111**
2.333333333	**-57.33333333**	**14.66666667**	**1.97373E-15**	**432.8888889**
			Variance	**20.80598205**

MG Analytics

Variance

- In probability theory and statistics, variance is The average of the squared differences from the Mean.
- Informally, it measures how far a set of numbers are spread out from their overage value.

Variance Formula

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

The "**Population** Standard Deviation": $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

The "**Sample** Standard Deviation**":** $$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Looks complicated, but the important change is to divide by **N-1** (instead of **N**) when calculating a Sample Variance.

MG Analytics

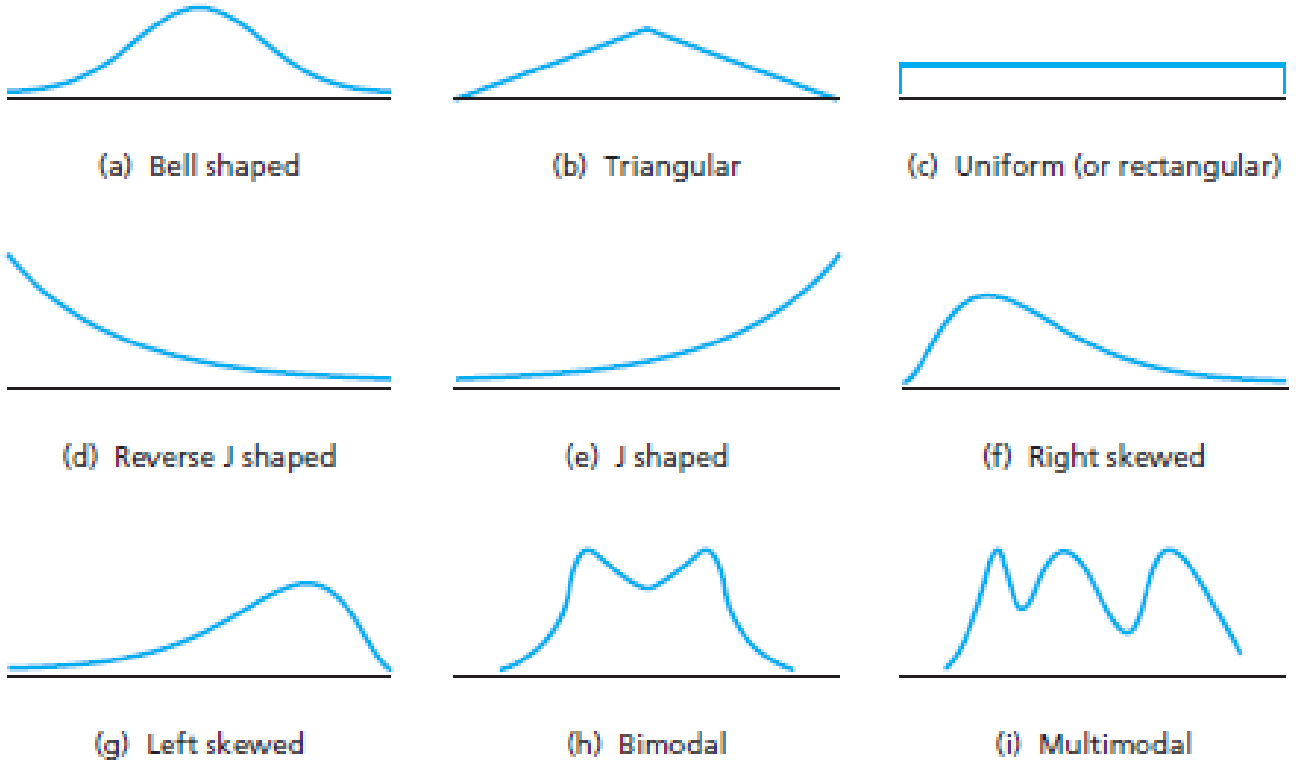
Standard Deviation

Bessels' Correction

- *Bessels' correction* refers to the "n-1" found in several formulas, including the sample variance and sample standard deviation formulas. This *correction* is made to *correct* for the fact that these sample statistics tend to underestimate the actual parameters found in the population.

Shapes of distributions of Data

- A data **distribution** shows all the possible values of the data and how often each value occurs.
- For **Normally distributed**, the mean, median and mode are all equal, and therefore are all appropriate measure of centre central tendency.
- For **skewed**, the median may be a more appropriate measure of central tendency.



Normal Distribution

- A normal distribution, sometimes called the bell curve, is a distribution that occurs naturally in many situations. For example, the bell curve is seen in scores of students. The bulk of students will score the average (C), while smaller numbers of students will score a B or D. An even smaller percentage of students score an F or an A. This creates a distribution that resembles a bell.
- The bell curve is symmetrical. Half of the data will fall to the left of the mean; half will fall to the right.
- Many groups follow this type of pattern. That's why it's widely used in business, statistics

MG Analytics

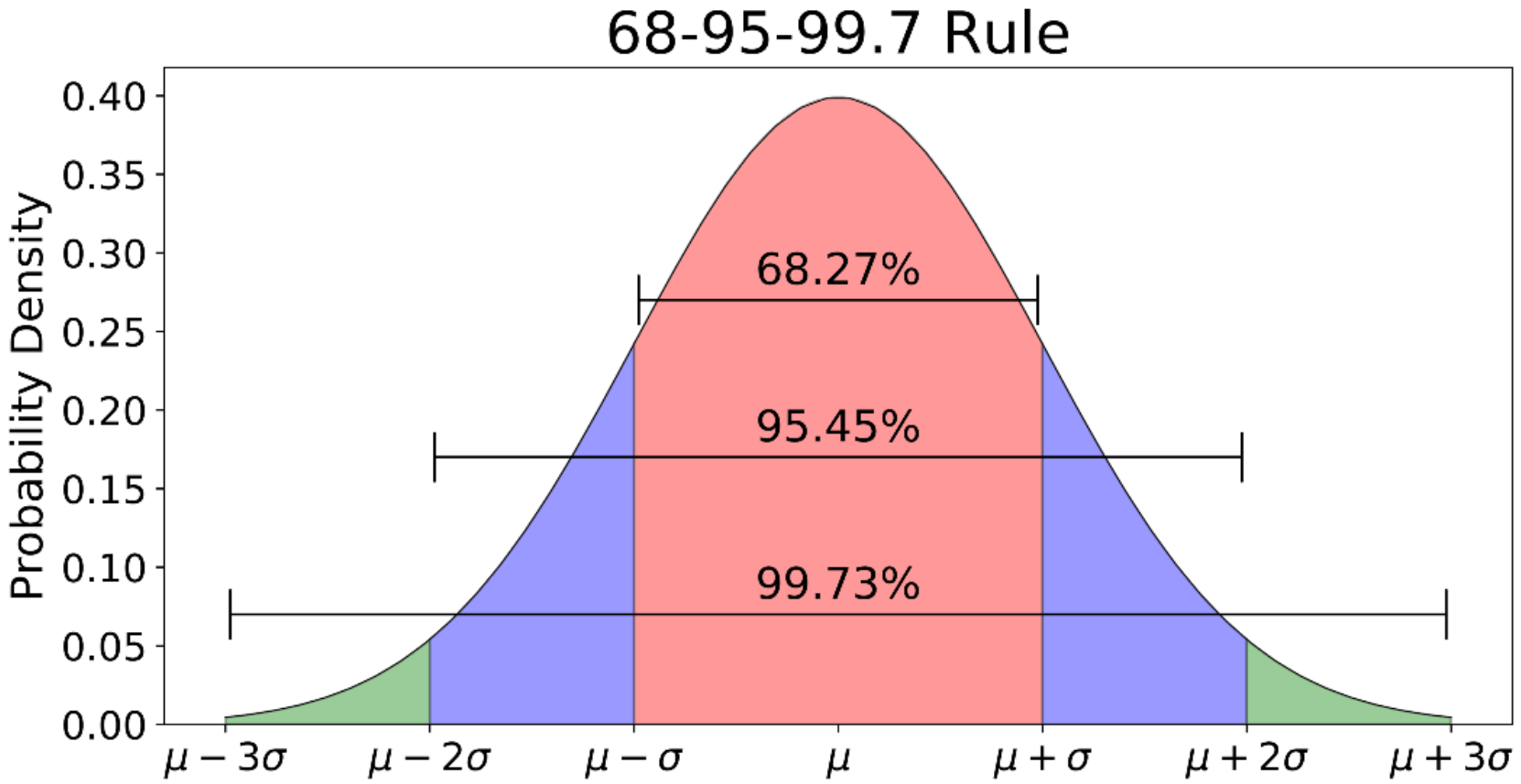
Bell Curve

Created from the parameters :

- Mean
- Standard Deviation

Can be followed for various purposes like :

- scholarship distribution
- finding the best performers or worst performers.
- Finding a segment of data



Kurtosis

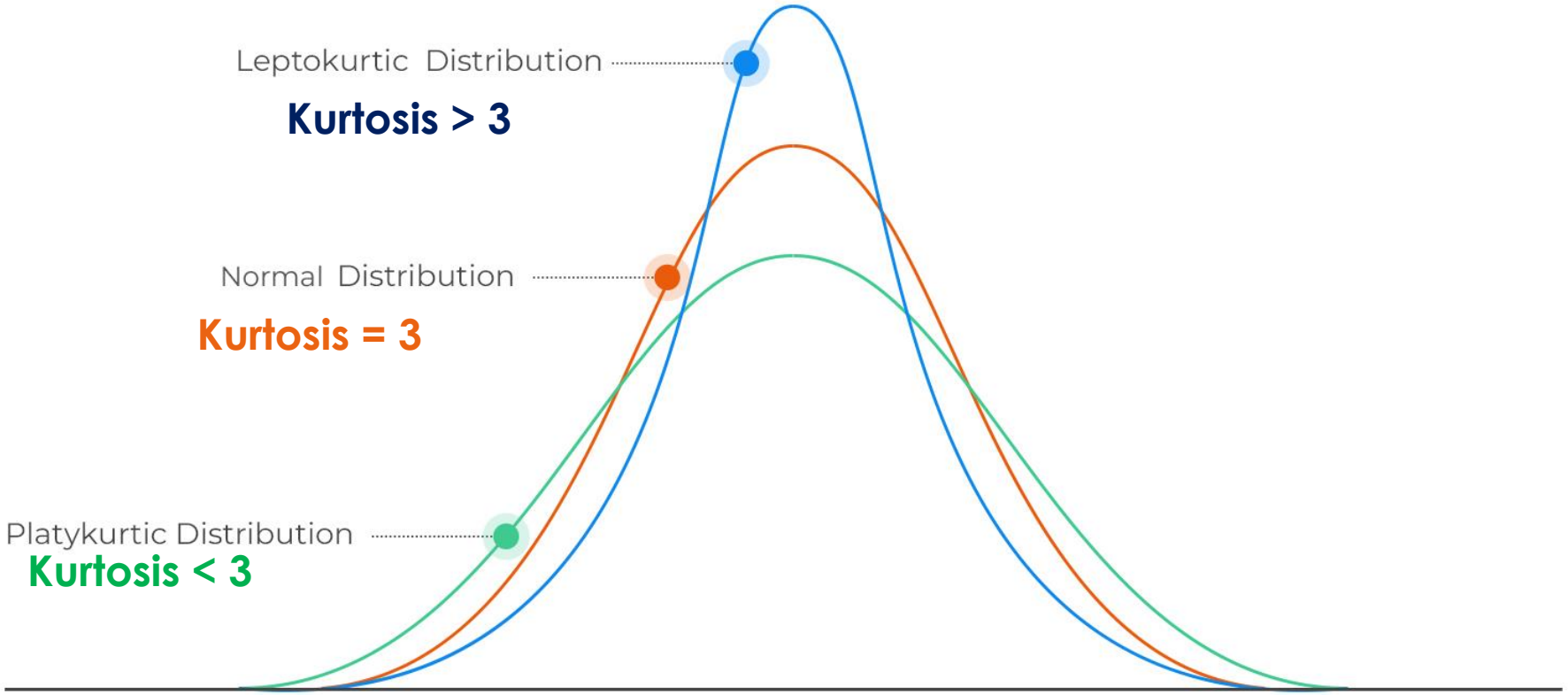
MG Analytics

Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution.

kurtosis identifies whether the tails of a given distribution contain extreme values.

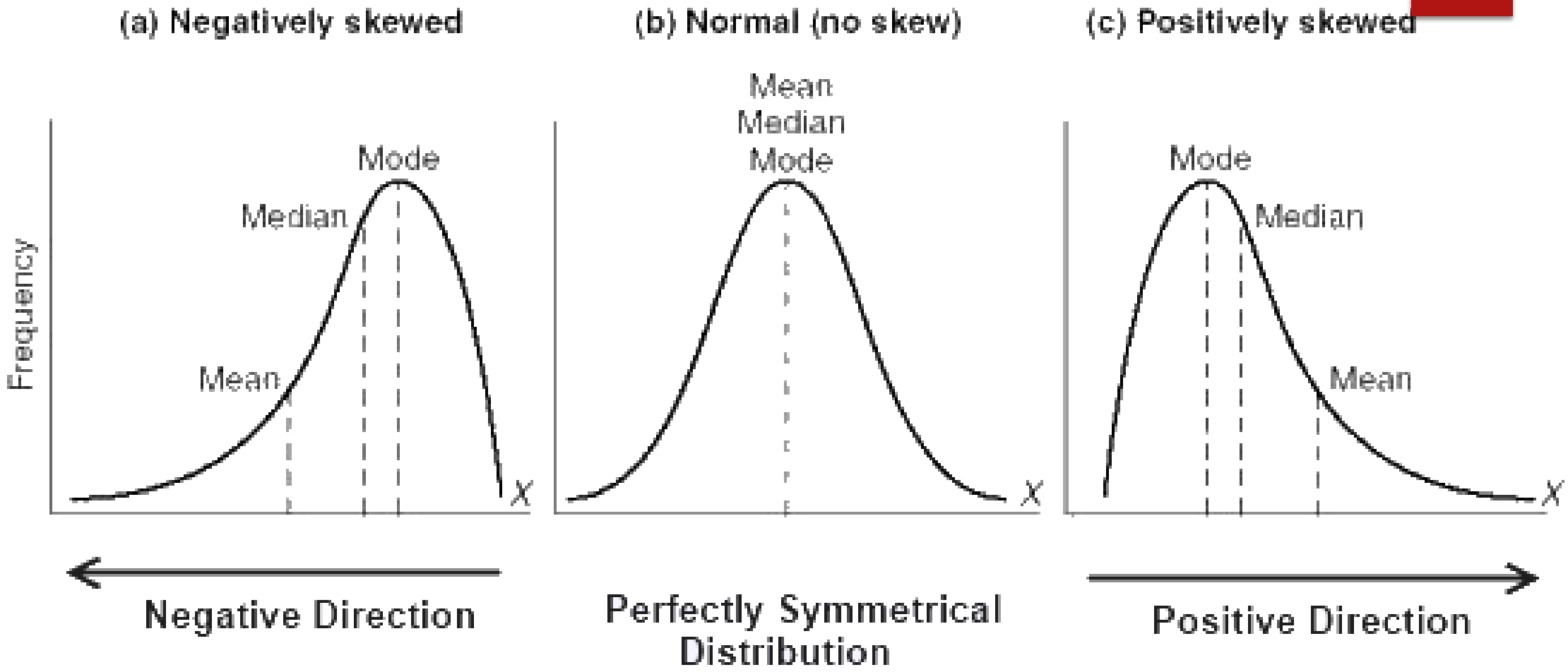
Kurtosis is a measure of the combined weight of the tails relative to the rest of the distribution.

Kurtosis



Skewness

- Skewness refers to distortion or asymmetry in a symmetrical bell curve, or normal distribution, in a set of data.
- If the curve is shifted to the left or to the right, it is said to be skewed.
- Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.
- A normal distribution has a skew of zero



MG Analytics

Correlation vs covariance

Both measure the **relationship and** the dependency between two variables.

"**Covariance**" indicates the direction of the linear **relationship** between variables.

"**Correlation**" on the other hand measures both the strength **and** direction of the linear **relationship** between two variables.

MG Analytics