30 Hours Online Certificate Course On

"Research & Data Analysis"

- Organised By: **Accounting and Finance Lab, Department of Commerce, Ramanujan College, University of Delhi**
- Resource Person: **Dr. Arnav Kumar**
- Designation: Assistant Professor, Department of Management Studies, Ramanujan College
- Email: arnavkumardse@gmail.com
- Day & Date: **Thursday, June 18, 2020.**

Why “IBM SPSS Statistics” for Data Analysis?

- IBM SPSS Statistics is an integrated family of products that addresses the entire analytical process, from planning to data collection to analysis, and reporting.
- Helps in quickly gaining understanding and insights from datasets in any format using advanced statistical procedures.
- Significantly increases the Analytical Power, and Flexibility (Choice of Methods etc.).
- Ensures high accuracy.
- Easily communicate results via presentation ready output and high quality graphs.

Basic Data Analysis (Frequency & Descriptive Statistics) using IBM SPSS Statistics

- **Variables:**
  - ✓ Grade & Percent;
  - ✓ Gender & Ethnicity.
- **Frequency Distribution:**

Analyze > Descriptive Statistics > Frequencies

- **Descriptive Statistics:**

Analyze > Descriptive Statistics > Frequencies > Statistics

Graphical Analysis using IBM SPSS Statistics

What makes a Good Graph?

- Tufte (2001) points out that graphs should:
  - Show the data.
  - Induce the reader to think about the data being presented (rather than some other aspect of the graph, like how pink it is).
  - Avoid distorting the data.
  - Present many numbers with minimum ink.
  - Make large data sets coherent.
  - Encourage the reader to compare different pieces of data.
  - Reveal the underlying message of the data.

Graphs - Basics

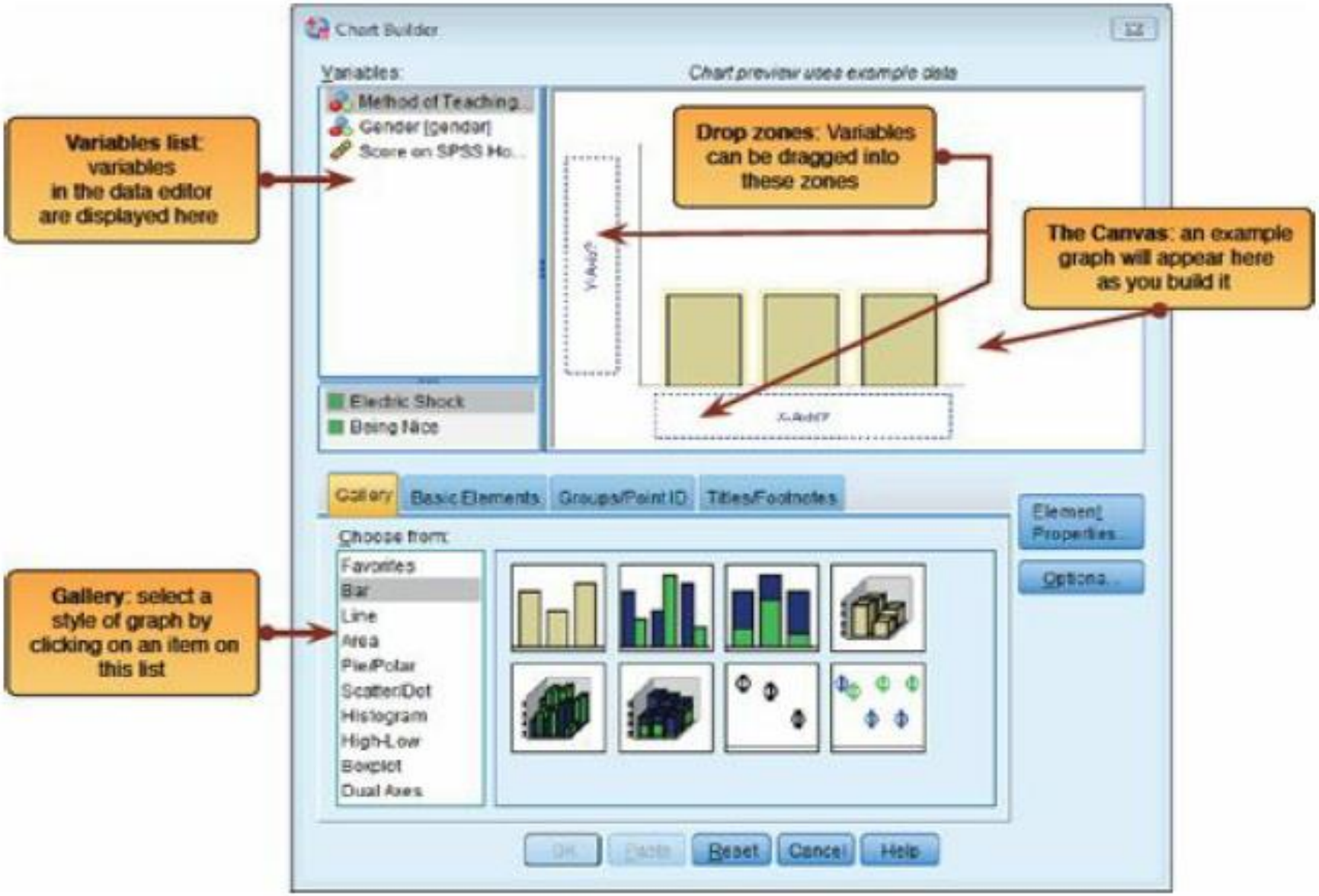
- Vertical Axis of Graph = Y-axis (or Ordinate).
- Horizontal Axis of Graph = X-axis (or Abscissa).
- Some Useful Graphing Tips:

✓ Don't create false impressions of what the data actually show (likewise, don't hide effects) by scaling the y-axis in some weird way.

✓ Abolish Chart-junk: Don't use patterns, 3-D effects, shadows, pictures/photos of anything else.

✓ Avoid excess ink: If you don't need the axes/lines, then get rid of them.

The SPSS Chart Builder



Common Types of Graphs

1. Line Charts
2. Pie Chart
3. Bar Charts
4. Histograms
5. Box Plot – It is commonly used to Check for Outliers.
6. Scatter Plot – It is used to study relationship between two Variables. We will learn more about it in “Correlation Analysis”.

1. Line Charts

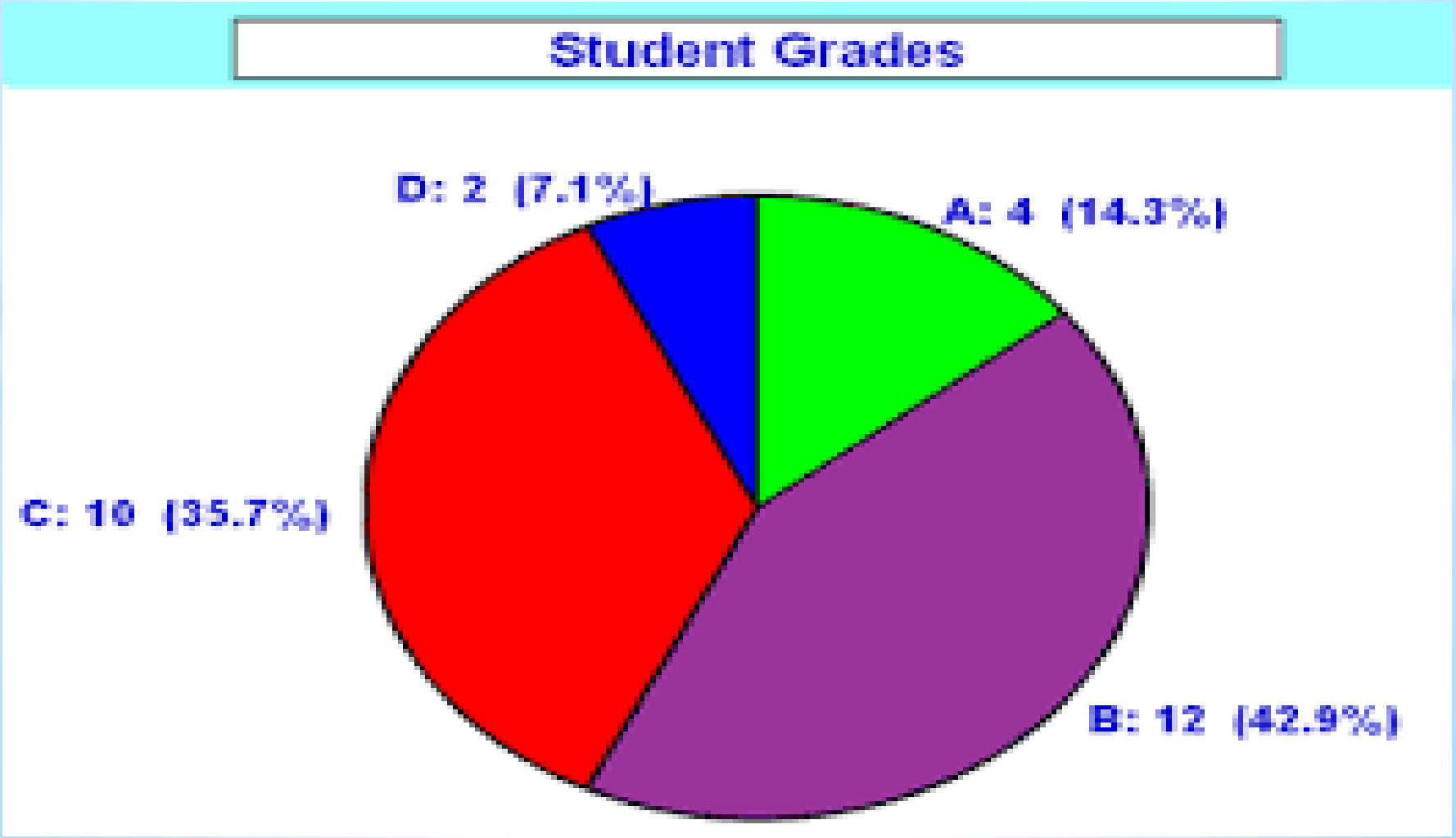
- We can use a Line Chart to summarize categorical variables, in which case it is similar to a bar chart.
- Line charts are also useful for Time Series/Case Wise/Category Wise analysis data.
- How to create a Simple Line Chart:
- ✓ In Chart Builder, click Gallery tab and select Line.
- ✓ Drag the Simple Line icon onto the canvas.
- ✓ Drag a Date/Case/Categorical Variable (*ID*/*Gender*) to the X-axis drop zone.
- ✓ Drag a Scale Variable (*GPA*/*Percentage*) to the Y-axis drop zone. This is the variable whose values were recorded over Categories/Cases/Time.

1. Line Charts

- We can also simultaneously draw multiple line charts.
- How to create a Multiple Line Chart:
- ✓ In Chart Builder, click Gallery tab and select Line.
- ✓ Drag the Multiple Line icon onto the canvas.
- ✓ Drag a Date/Case/Categorical Variable (*Gender/ Ethnicity*) to the X-axis drop zone.
- ✓ Drag a Scale Variable (*Quiz 1, Quiz 2, Quiz 3, Quiz 4 & Quiz 5*) to the Y-axis drop zone.

2. Pie Charts

- Each slice of a pie chart displays the proportion of parts to a whole.



2. Pie Charts

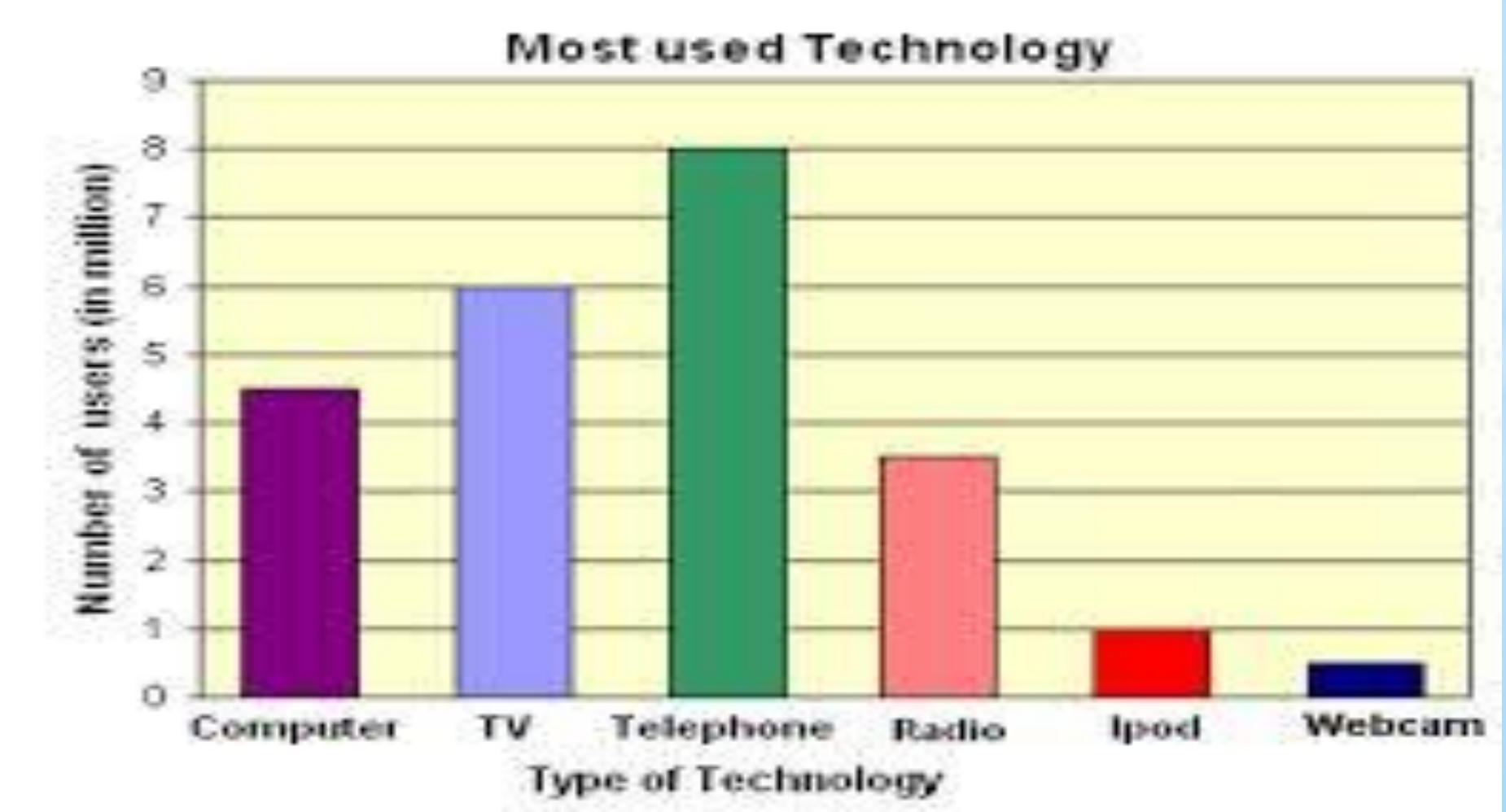
- A pie chart is useful for comparing proportions. Ex: To demonstrate that a greater proportion of women are enrolled in a certain course.
- How to create a Simple Pie Chart:
- ✓ In Chart Builder, click Gallery & select Pie/Polar.
- ✓ Drag the Pie Chart icon onto the canvas.
- ✓ Drag a categorical (nominal or ordinal) variable to the Slice By drop zone.
- ✓ The number of categories in this variable determine the number of slices in the pie chart.

2. Pie Charts

- ✓ Specify a statistic (Count/Sum/Percentage) in the Element Properties dialog box.
- ✓ For pie charts, you typically want a Count-based Statistic or Sum (for Nominal or Ordinal variable).
- ✓ The result of the statistic determines the relative size of each slice out of the total pie.
- ✓ To see Values/Data Labels: Double left click graph > Chart Editor > Elements > Show Data Labels.
- ✓ Create a Pie Chart of *Grades* (Categorical Variable) based on *Count & Percentages* (Statistics).
- ✓ In Data Label, simultaneously Show *Grade*, *Count & Percentage*.

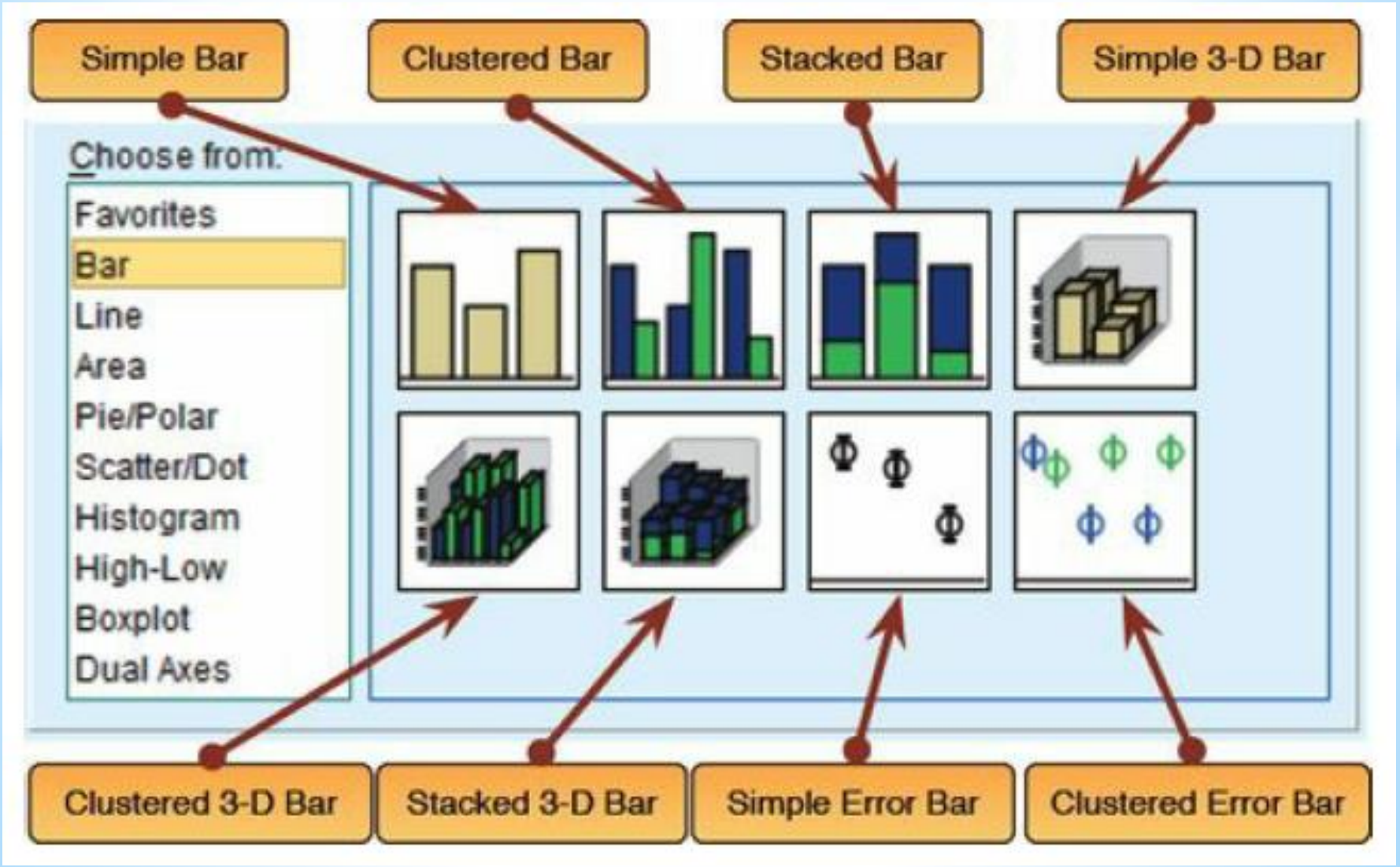
3. Bar Charts

- Displays the count for each distinct value or category as a separate bar.



3. Bar Charts

- Bar Charts are the usual way to display means.



3. Bar Charts

- Simple Bar: To see the Means of Values of Scale Variables across different groups/categories of Categorical Variable.
- Clustered Bar: If we have a second Grouping or Categorical Variable, we could produce a simple bar chart (as above) but with bars produced in different colours for levels of the second grouping variable.
- Stacked Bar: Same as the clustered bar, except that the different coloured bars are stacked on top of each other rather than placed side by side.

3. Bar Charts

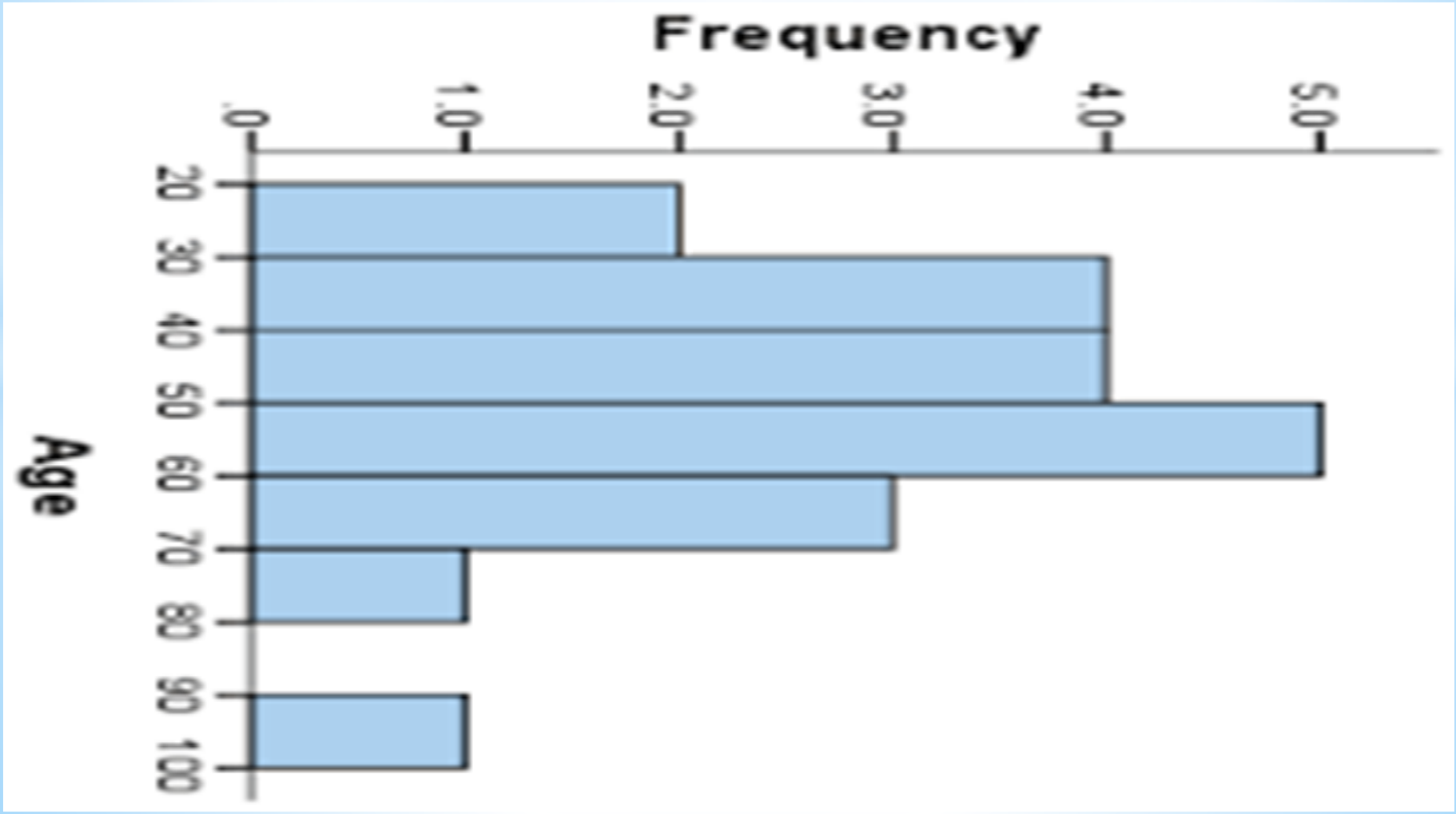
- How to create a Simple Bar Chart:
- ✓ In Chart Builder, click Gallery tab and select Bar.
- ✓ Drag the Simple Bar icon onto the canvas.
- ✓ Drag a categorical (nominal or ordinal) variable (*Grade*) to the x-axis drop zone. You can use a scale variable, but the results will be useful in only a few special cases. A bar chart looks best with a limited number of distinct values.
- ✓ Specify a Statistic (*Cumulative Count*/*Cumulative Percentage*) in the Element Properties dialog box. The result of any statistic determines the height of the bars.

3. Bar Charts

- How to create a Clustered Bar Chart:
- ✓ In Chart Builder, click Gallery tab and select Bar.
- ✓ Drag the Clustered Bar icon onto the canvas.
- ✓ Drag a categorical (nominal or ordinal) variable (*Grade*) to the X-axis drop zone.
- ✓ Drag & drop the second Categorical Variable (*Gender*) to the "Cluster on X: Set Color" box.
- ✓ Specify a Statistic (*Count*/*Percentage*) in the Element Properties dialog box.
- ✓ With all the same settings, just drag the "Stacked Bar" icon on the canvas to get a Stacked Bar Chart.

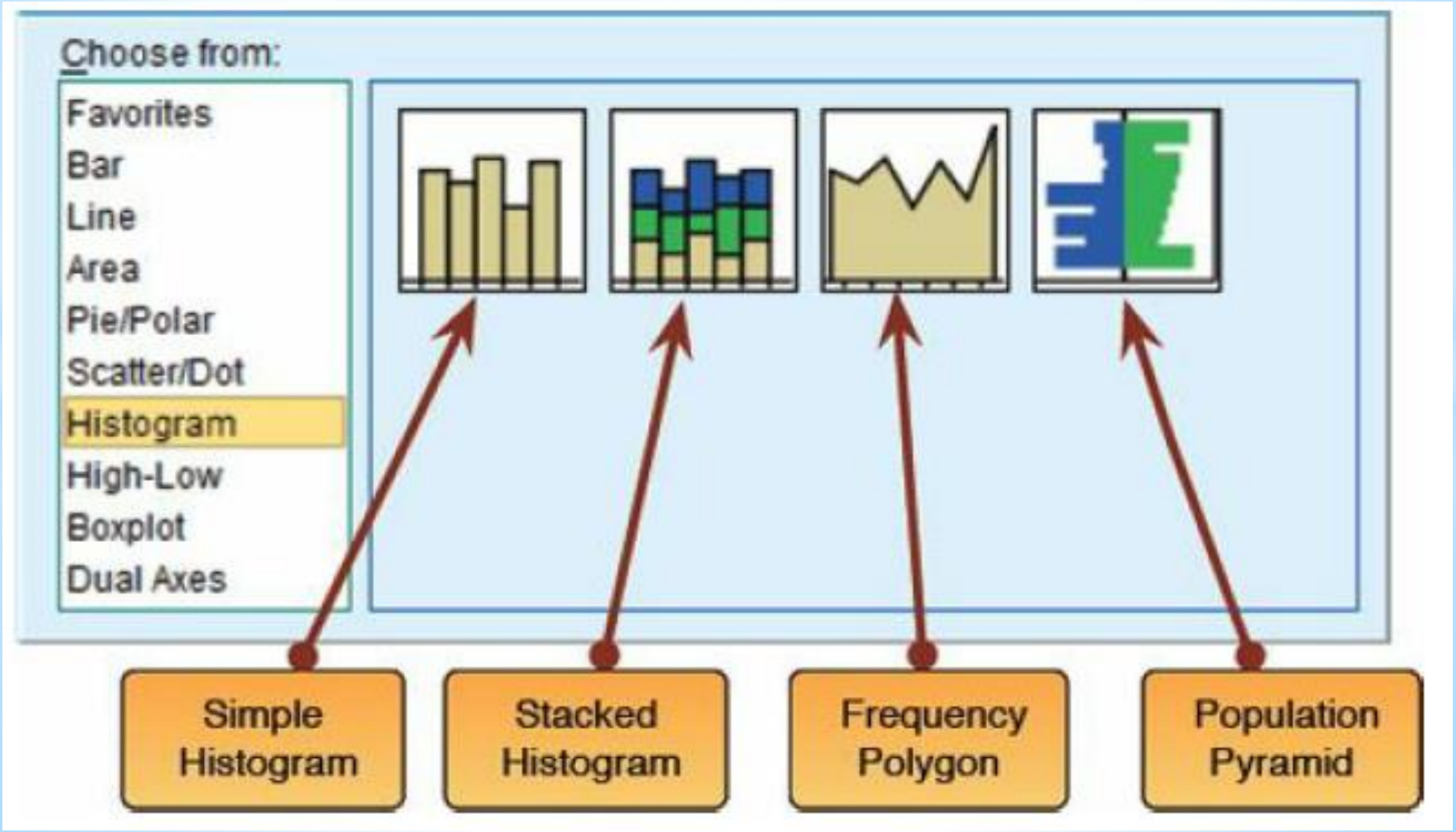
4. Histograms

- Also has bars, but along an equal interval scale.
- Height is Frequency of values.



4. Histograms

- Histograms are useful for showing the distribution or shape of data of a single scale variable.



4. Histograms

- Simple Histogram: Use this option to see the frequencies of scores for a single variable.
- Stacked Histogram: If you had a grouping variable you could produce a histogram in which each bar is split by group. This is a good way to compare the relative frequency of values of the Scale variable across groups of the Categorical Variable.
- Frequency Polygon: This option displays the same data as the simple histogram, except that it uses a line instead of bars to show the frequency, and the area below the line is shaded.

4. Histograms

- Population Pyramid:
- Like a stacked histogram, this shows the relative frequency of scores in two populations.
- It plots the variable on the vertical axis and the frequencies for each population on the horizontal: the populations appear back to back on the graph.
- If the bars either side of the dividing line are equally long then the distributions have equal frequencies.

4. Histograms

- How to create a Simple Histogram:
- ✓ In Chart Builder, click Gallery and select Histogram.
- ✓ Drag the Simple Histogram icon onto the canvas.
- ✓ Drag a scale variable (*GPA*) to the X-axis drop zone.

- How to create a Stacked Histogram:
- ✓ In Chart Builder, click Gallery and select Histogram.
- ✓ Drag the Stacked Histogram icon onto the canvas.
- ✓ Drag a scale variable (*GPA*) to the X-axis drop zone.
- ✓ Drag a Categorical Variable (*Gender*) to the "Stack: Set Color" box.

4. Histograms

- How to create a Frequency Polygon:
- ✓ In Chart Builder, click Gallery and select Histogram.
- ✓ Drag the Frequency Polygon icon onto the canvas.
- ✓ Drag a scale variable (*Percent*) to the X-axis drop zone.

4. Histograms

- How to create a Population Pyramid:
- ✓ In Chart Builder, click Gallery and select Histogram.
- ✓ Drag the Population Pyramid icon onto the canvas.
- ✓ Drag a Scale Variable (*Percentage*) to the Distribution Variable drop zone.
- ✓ Drag a Categorical Variable (*Gender*) to the Split Variable drop zone. Although it is possible to use a split variable with many categories, it is recommend to use a variable that has only two categories.
- ✓ The split variable acts as a paneling variable in that it creates multiple graphs.

Inferential Statistics

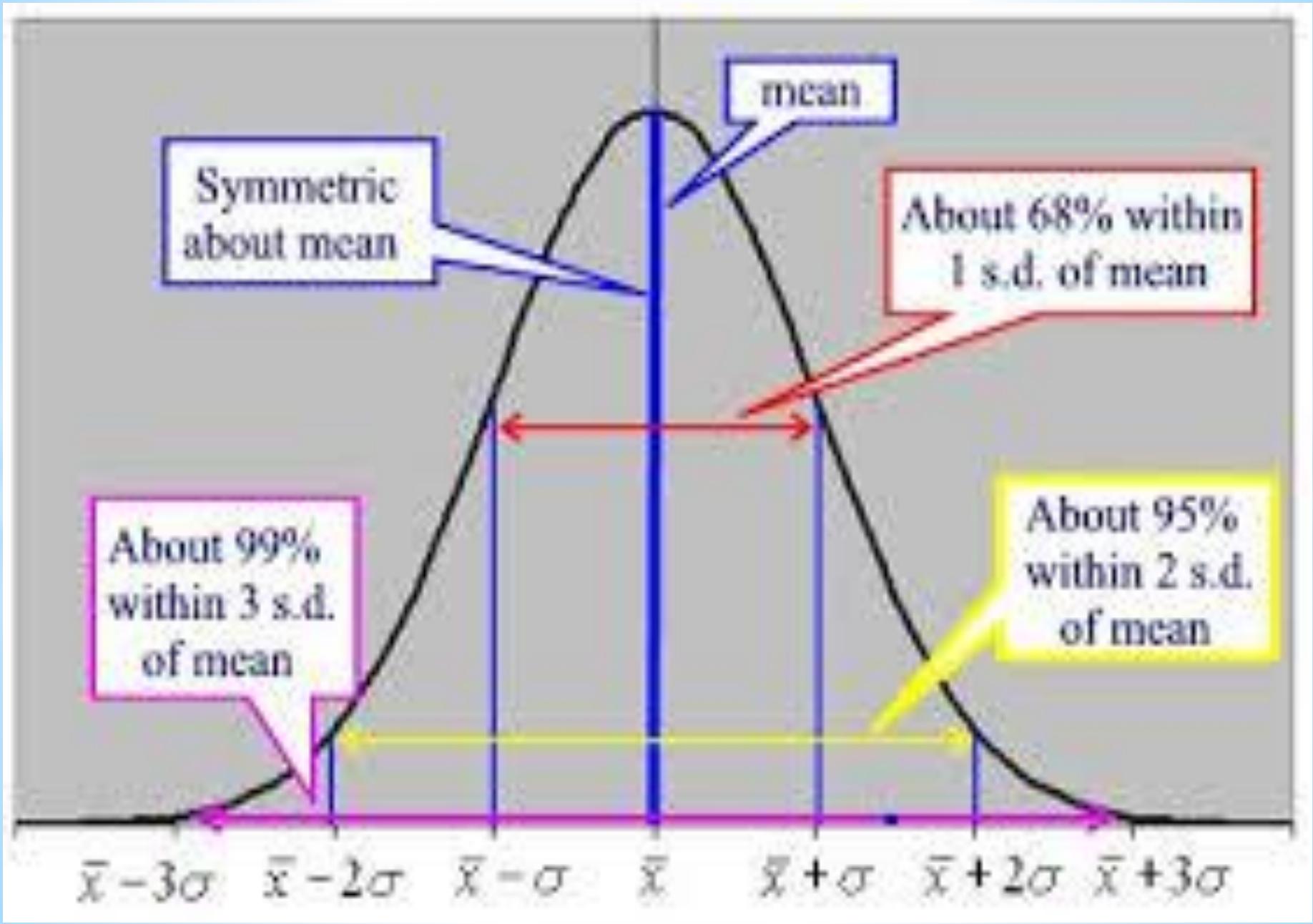
Types of Probability Distributions

- **Discrete Probability Distributions** - The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. Examples: Binomial (Only 2 Outcomes), Poisson.
- **Continuous Probability Distributions** - These describe an "unbroken" continuum of possible occurrences. A random variable is continuous if it can take any value in an interval. Example: Normal, Student's t, Chi-square & F.
- In essence, we make predictions about the Population Parameter from the Sample Statistic.

Normal Distribution

- Bell Shaped;
- Mean = Median = Mode.
- Symmetric about Mean, so no Skewness (Positive or Negative);
- Depends on only two parameters – Mean (μ), and Standard Deviation (σ).
- Standard Normal Variate (Z) = $(x - \mu)/\sigma$.
- Standard Normal Variate is Normally Distributed with Mean = 0 and Standard Deviation = 1.
- We convert any variable (X) to a Standard Normal Variable (Z) using above formulae and give it same property ($\mu = 0$ & $\sigma = 1$).

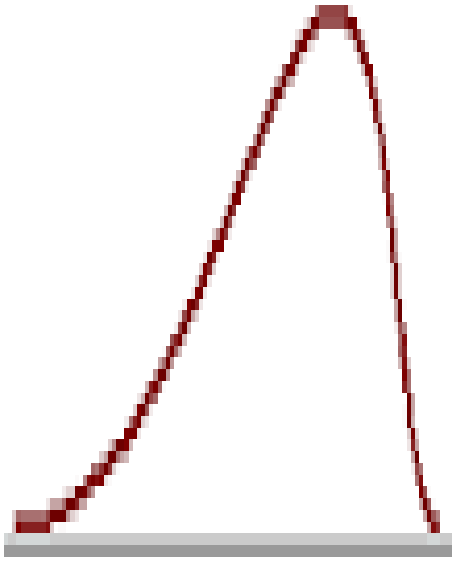
Normal Distribution



Skewness

- Skewness is a measure of symmetry, or more precisely, the lack of symmetry.
- A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point.
- Skewed distributions are not symmetrical, Normal Distribution has Zero Skewness or is Un-Skewed.
- If the distribution is positively skewed, then the probability density function has a long tail to the right, and if the distribution is negatively skewed then the probability density function has a long tail to the left.

Skewness



Negatively skewed distribution
or Skewed to the left
Skewness <0

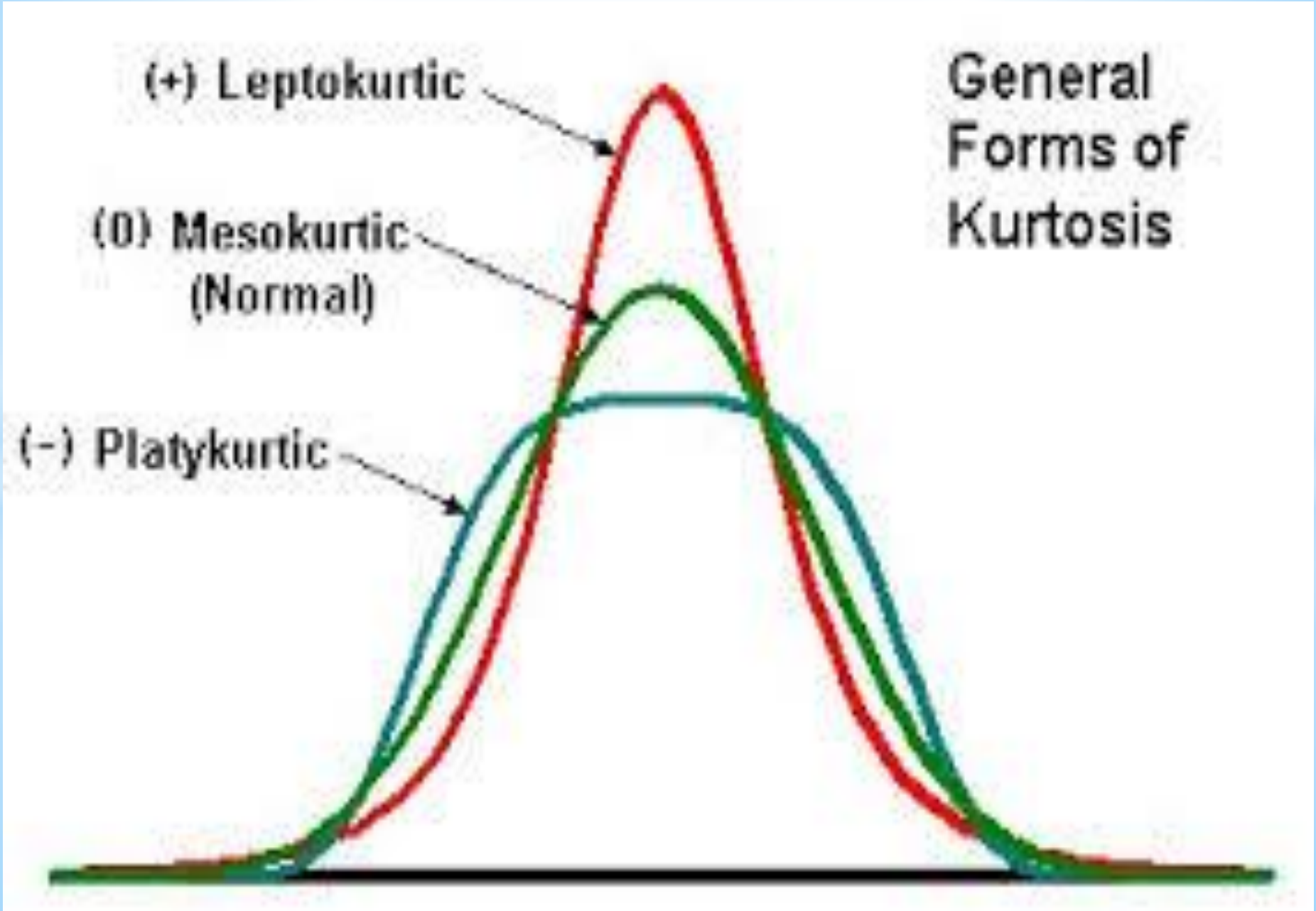
Normal distribution
Symmetrical
Skewness = 0

Positively skewed distribution
or Skewed to the right
Skewness > 0

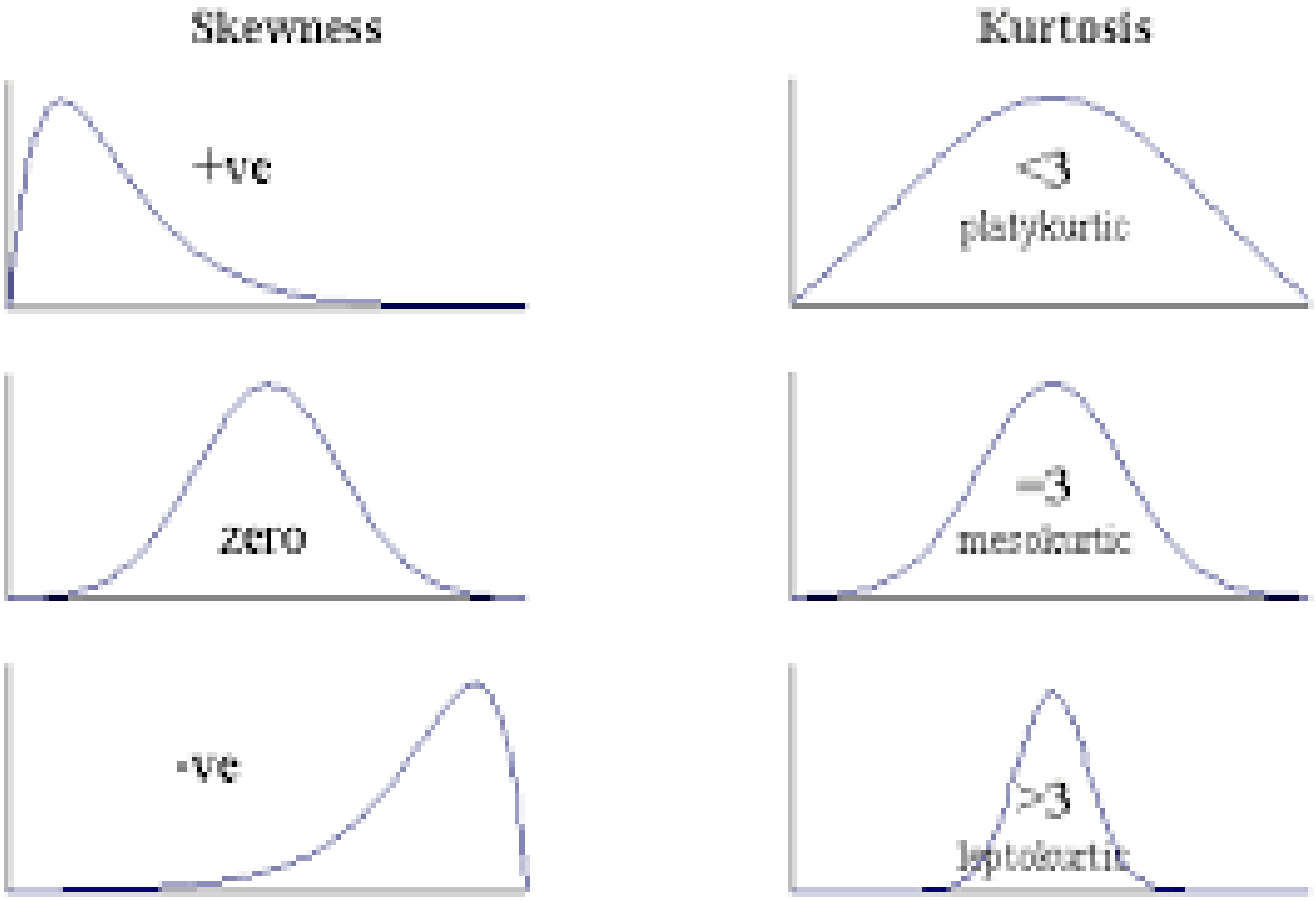
Kurtosis

- Kurtosis is a measure of whether the data are heavy-tailed (more outliers) or light-tailed (less outliers) relative to a normal distribution.
- Kurtosis of the standard normal distribution is 3 and it is called Mesokurtic.
- Variables with low kurtosis (<3) are called Platykurtic (fat or short tailed).
- Variables with high kurtosis (>3) are called Leptokurtic (slim or long tailed).
- Using the standard normal distribution as a benchmark, the Excess Kurtosis of a random variable X = Kurtosis of X − 3.

Forms of Excess Kurtosis



Skewness & Kurtosis



Sampling

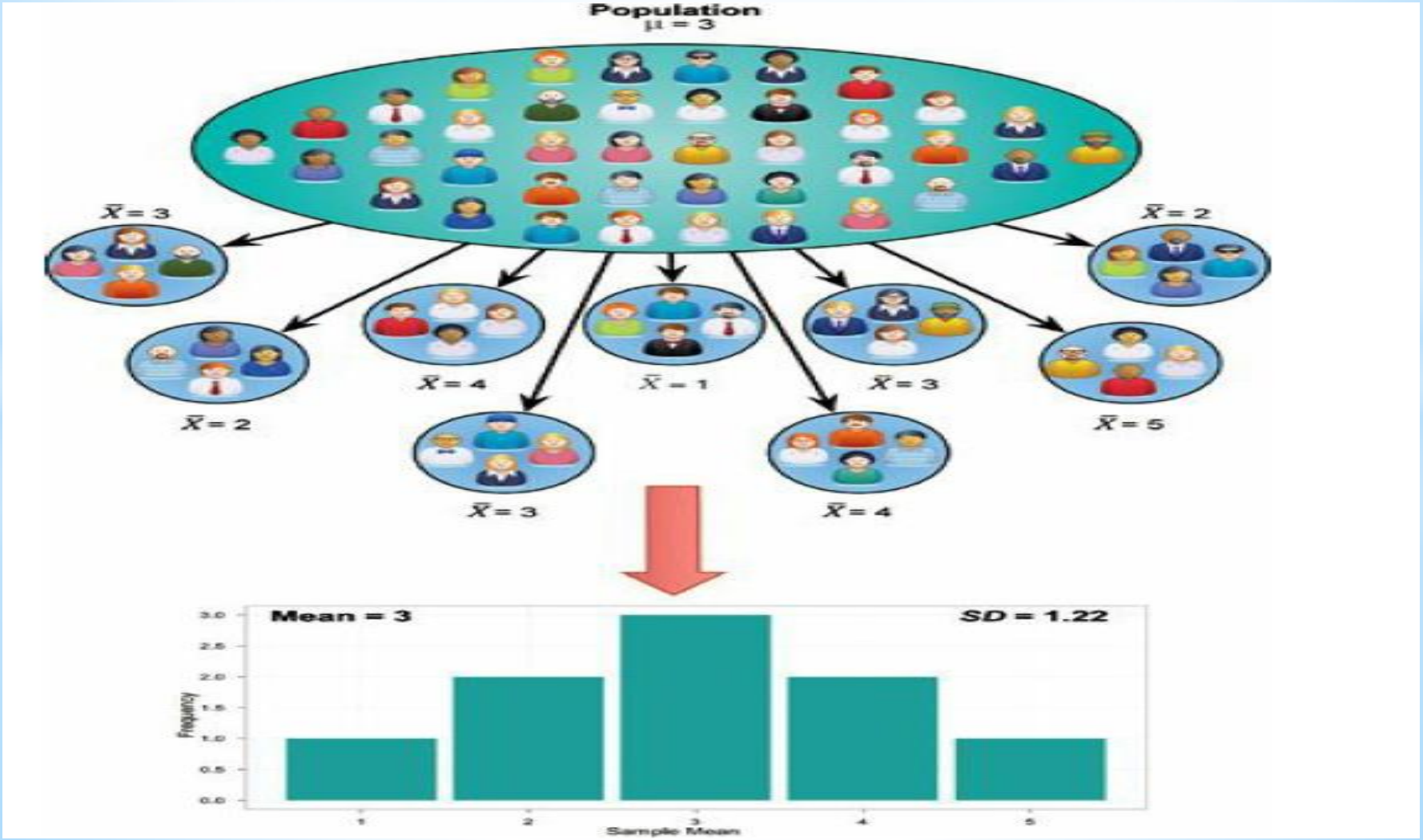
- Sampling is an integral part of inferential statistics that allows drawing conclusions about a population from a relatively small number of observations drawn from it.
- Reasons: Time, Cost, Effort and No Choice
- Probability and Non Probability Sampling.
- Conditions for Sample to be reliable:
- ✓ Random Sampling method
- ✓ Sample Size should be adequate in relation to Population size
- ✓ Sample should be representative of Population

Estimation

- Estimation: Process that involves using limited information obtained from a sample to draw conclusions about population from which sample is taken. It is used to estimate value of population parameter.
- Estimate: An educated guess about some value of a population parameter.
- Estimator: The rule or procedure used to obtain that guess.
- Point Estimate: Specific Value; Interval Estimate: Range of Values.

Sampling Distributions

- It is the frequency distribution of sample means from the same population.



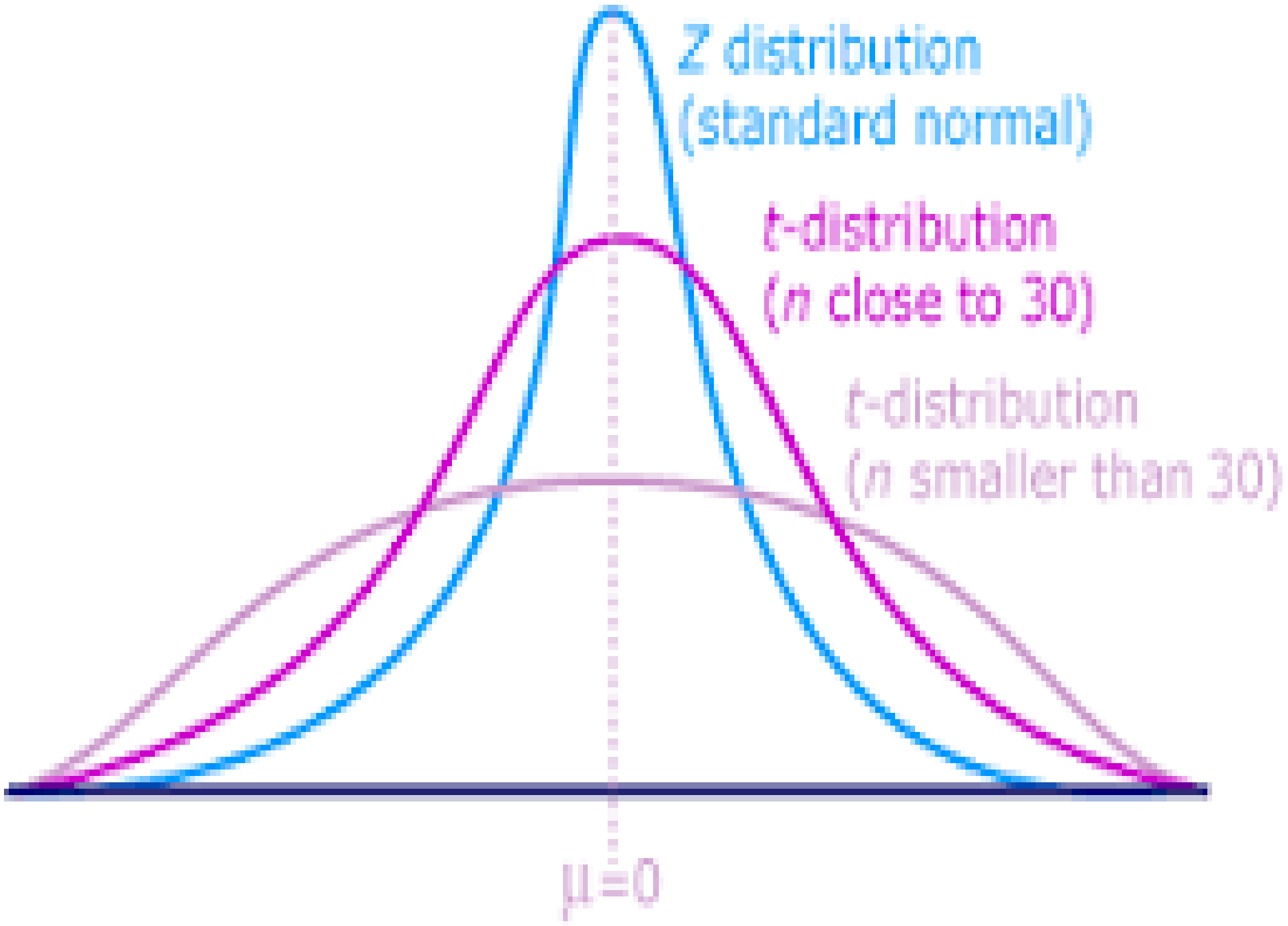
Standard Error (S.E.)

- If our data is that of the sample means, then the standard deviation of these sample means would tell us how widely spread (i.e., how representative) sample means are around their average. i.e., population mean.
- It tells whether sample means are typically representative of the population mean.
- The standard deviation of sample means is known as Standard Error of the mean (SE).
- A small Standard Error indicates that most sample means are similar to the population mean and so our sample is likely to be an accurate reflection of the population.

Normal & Student's t Distributions

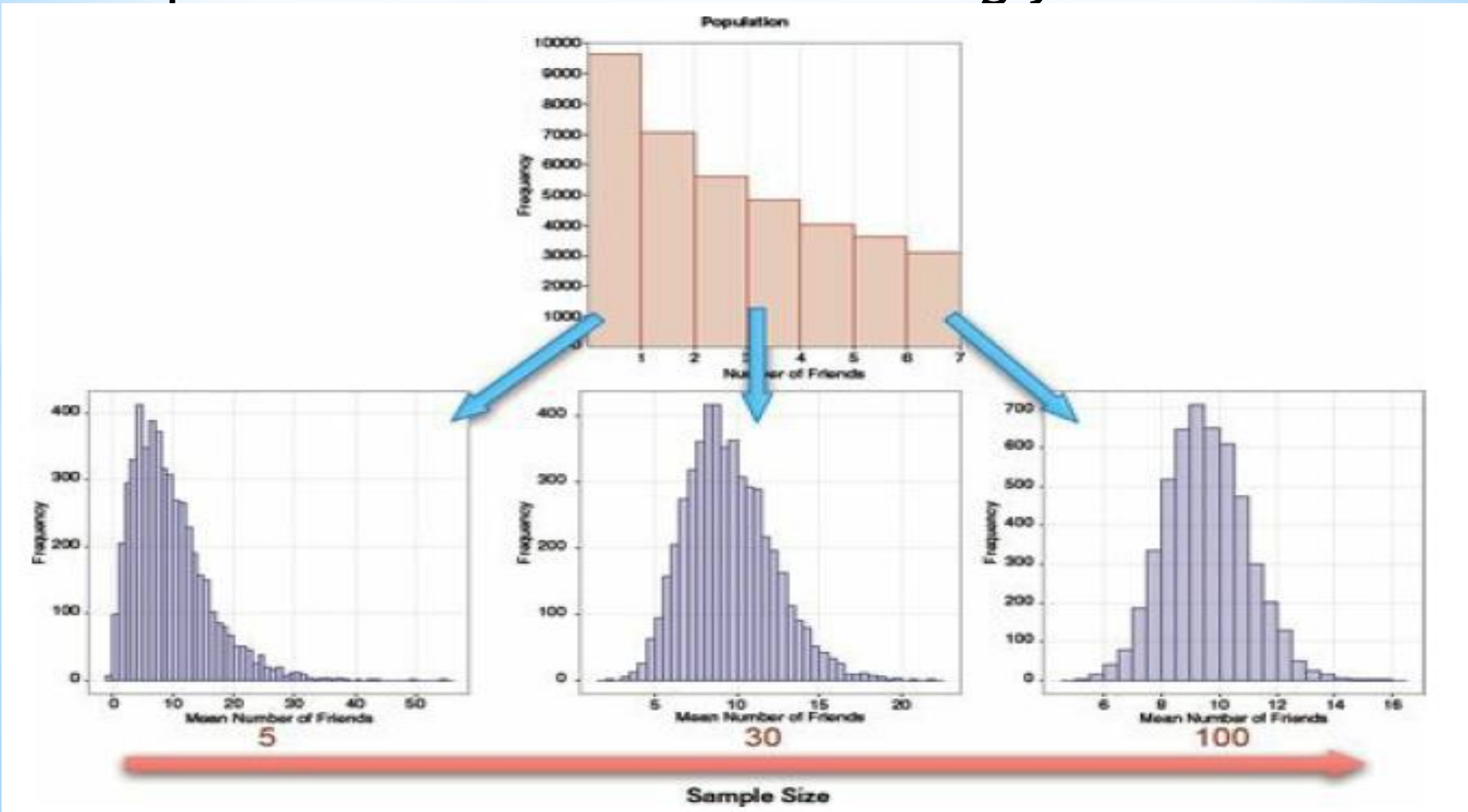
- There is a specific relationship between t and the standard normal distribution. Like Normal distribution, t distribution is symmetrical, but is flatter than normal distribution.
- The t-distribution has another parameter, its degrees of freedom. As the degrees of freedom increase, t dist. becomes standard normal dist.
- Degress of Freedom are number of observations that are free to vary, i.e., can take any values. It is equal to total Number of Observations (N) – Number of Parameters to be Estimated or Restrictions (K).

Normal & Student's t Distributions



The Central Limit Theorem Revisited

- Parameter estimates sampled from a Non-Normal population: As Sample size increases, Distribution of parameters becomes increasingly Normal.



The end!
Thank you for viewing and listening!