Descriptive Statistics & Sampling Distributions

LEARNING OBJECTIVES

At the end of this session, you should be able to do the following:

- Basic **descriptive analysis** including summary measures and graphical techniques.
- Basic exploratory data analysis techniques.
- Understand the concept of **sampling distribution**.
- **Estimate** population parameters based on sample statistics.

DESCRIPTIVE ANALYSIS

The elementary transformation of raw data in a way that describes the basic characteristics such as central tendency and variability.

Central Tendency	Variability	Shape	Location	Outliers
Mean	Variance	Skewness	Min, Max	
Median	Standard deviation	Kurtosis	Quartiles	
Mode	Range		Percentiles	
	IQR			

MEASURES OF CENTRAL TENDENCY

Mean

- Average Response

Median

- Midpoint of the distribution
- When the distribution has an even number of observations, median is the average of the two middle scores.

Mode

- Most frequently occurring value
- There may be more than one mode in a distribution. (Bimodal / Multimodal)

Example

Consider the data distribution:

4, 5, 5, 6, 6, 7, 7, 7, 8, 8, 9, 9, 10

- Mean = $\frac{4+5+5+6+6+7+7+7+8+8+9+9+10}{13} = 7$
- Median = 4 5 5 6 6 7 7 7 8 8 9 9 10 = 7
- Mode = 4 5 5 6 6 7 7 7 8 8 9 9 10 = 7

MEASURES OF VARIABILITY

Variance

- Measure of score dispersion about the mean

Standard deviation

- How far away from the average the data values typically are.

Range

- Difference between the largest and the smallest of the distribution

(Interquartile Range) IQR

- Difference between the first and third quartiles of the distribution

Example

Consider the data distribution:

4, 5, 5, 6, 6, 7, 7, 7, 8, 8, 9, 9, 10

- Variance =

$$s^2 = \frac{\begin{array}{c}(4-7)^2+(5-7)^2+(5-7)^2+(6-7)^2+(6-7)^2 \\ +(7-7)^2+(7-7)^2+(7-7)^2+(8-7)^2+(8-7)^2 \\ +(9-7)^2+(9-7)^2+(10-7)^2\end{array}}{13-1} = 3.17$$

- Standard deviation =
- Maximum = 10
- Minimum = 4
- Range = 10 – 4 = 6

4 5 5 6 6 7 7 7 8 8 9 9 10

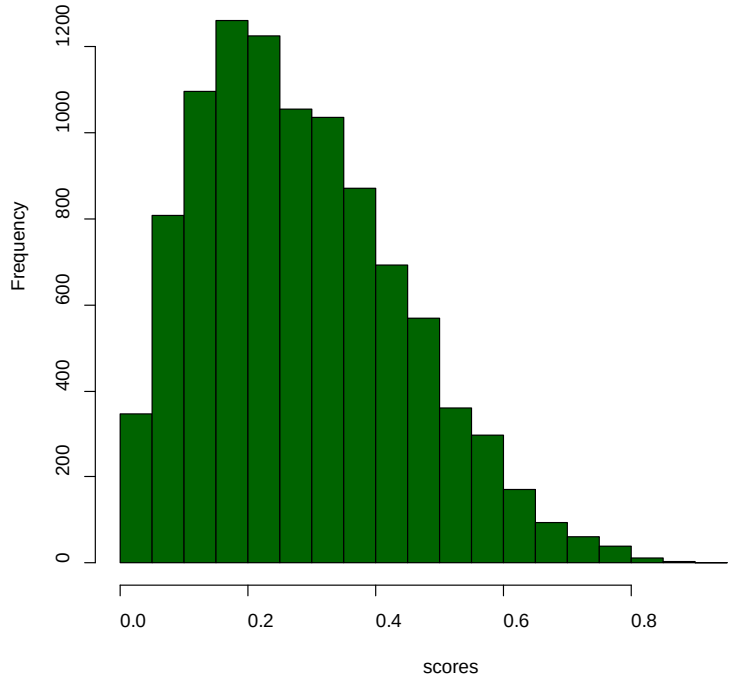
Q1 = 5.5

Q3 = 8.5

IQR = 8.5- 5.5 = 3

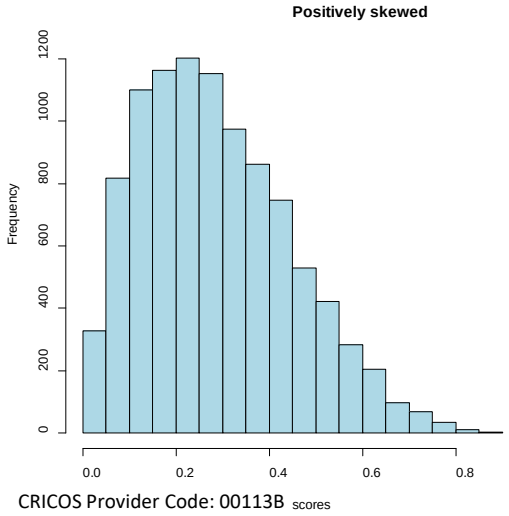
MEASURES OF SHAPE
FREQUENCY DISTRIBUTIONS

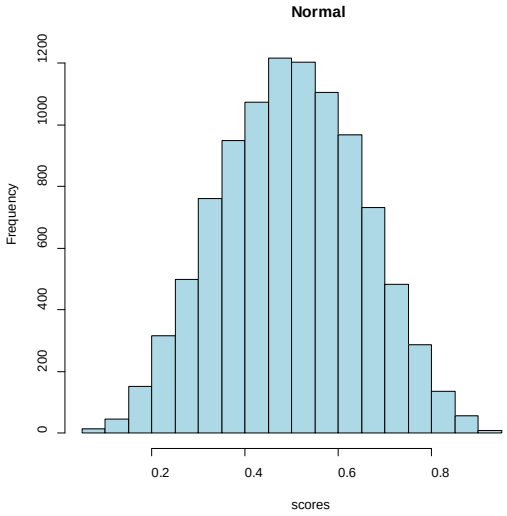
- A graph of how many times each scores occurs
- Horizontal axis: values of observations
- Bars: frequency

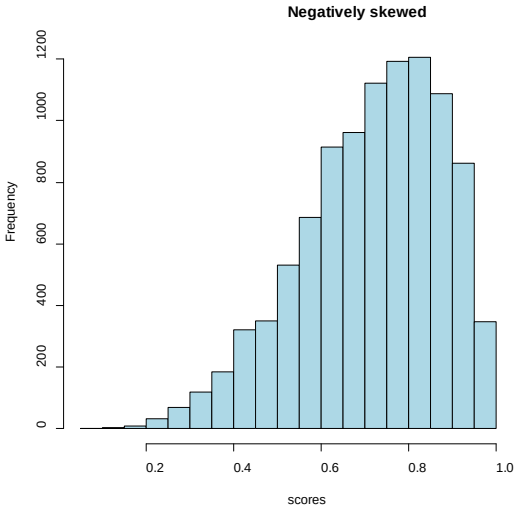


SKEWNESS
MEASURES OF SHAPE

- Measure of distribution's deviation from symmetry
- In a symmetrical distribution, the mean, median and mode are in the same location. (skewness = 0)
- A distribution that has cases stretching towards one tail or the other is called skewed. When the tail stretches to larger values, it is positively skewed. Scores stretching toward smaller values, skew the distribution negatively.

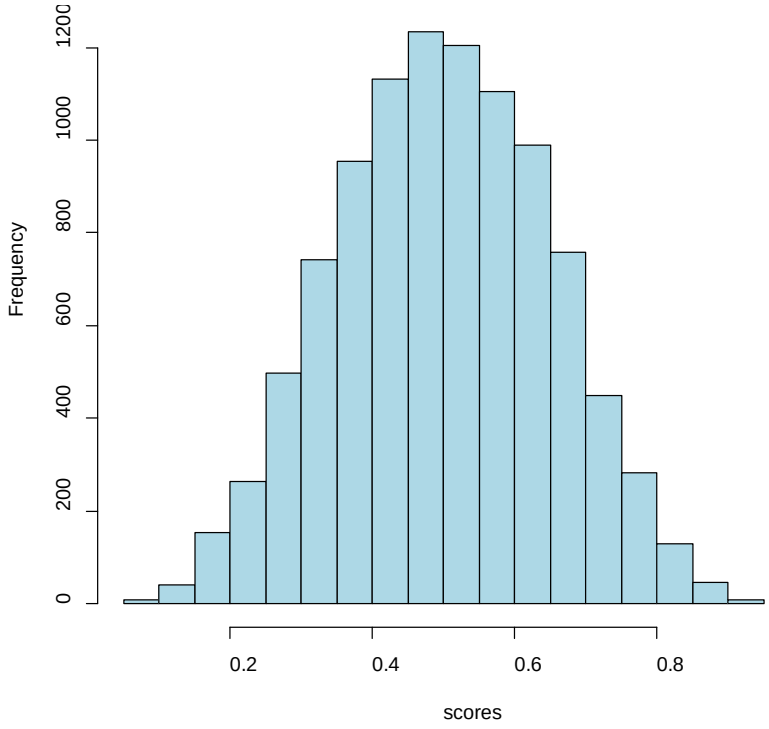






THE NORMAL DISTRIBUTION

- One of the most common frequency distributions in statistics
- Bell-shaped curve



MEASURES OF LOCATION

- Minimum and maximum give the extreme values of a data set
- Quartiles: Split an array into quarters
- Deciles: Split an array into tenths
- Percentiles: Split an array into hundredths
- The p^{th} percentile of a data set is a value such that at least p percent of the items take on this value or less and at least (100 - p) percent of the items take on this value or more.

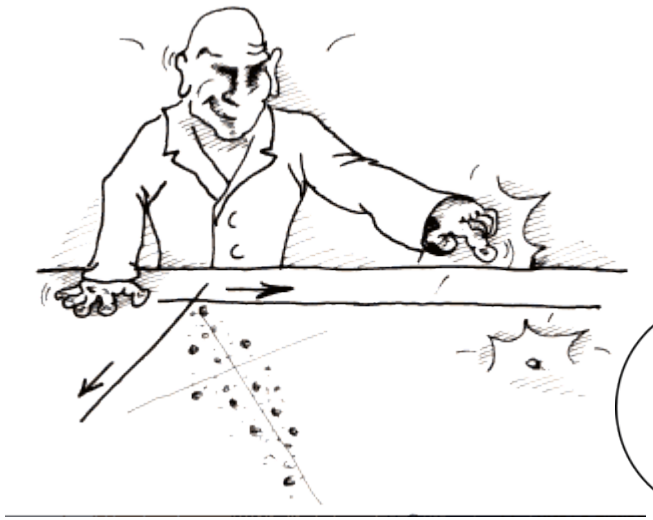
OUTLIER

What is an outlier?

- An observation/response with a unique combination of characteristics identifiable as distinctly different from the other observations/responses.

Issue?

- "Is the observation/response representative of the population?"



WHY DO OUTLIERS OCCUR?

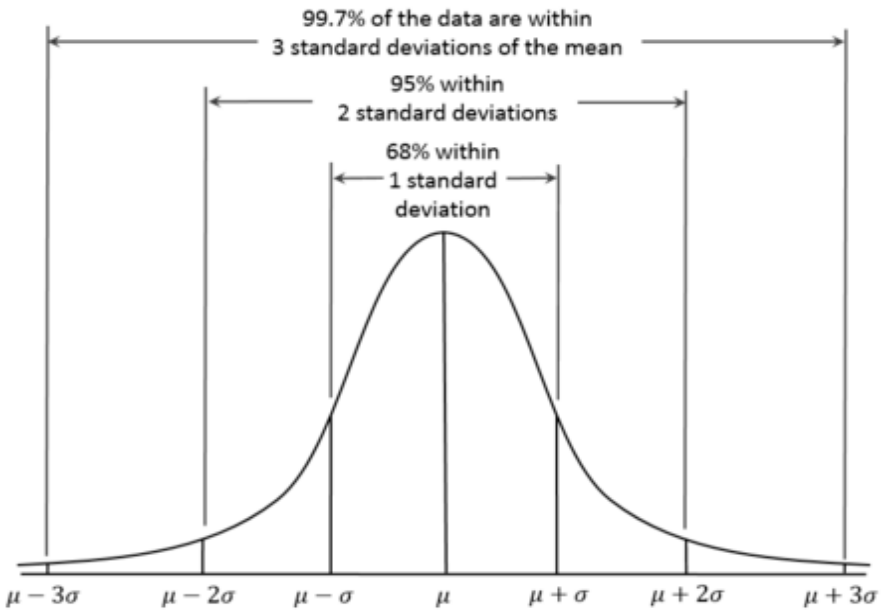
- Procedural Error.
- Extraordinary Event.
- Extraordinary Observations.

DEALING WITH OUTLIERS

- Identify outliers.
- Describe outliers.
- Delete or Retain?

EMPIRICAL RULE
DETECTING OUTLIERS

- For symmetrical bell-shaped data almost all data (99.7%) fall within 3 standard deviations of the mean.
- Any data further than 3 standard deviations from the mean is considered a potential outlier.



TUKEY'S RULE
DETECTING OUTLIERS

- Apply to non-bell shaped distributions.
- The rule works by calculating limits (often called fences) which are determined using the quartiles and IQR.
- The lower fence is located 1.5(IQR) below Q1.
- The upper fence is located 1.5(IQR) above Q3.
- Any value outside the fences is considered a potential outlier.

WHICH SUMMARY MEASURES TO USE?

- Describe a variable from all five aspects:

  Central Tendency, Location, Variability, Shape and outliers.
- For a symmetrical distribution we generally use the mean and standard deviation as measures of Central Tendency and Variation.
- What if asymmetrical distribution?
- The median and IQR might be more appropriate.

- Outliers can distort the mean so the median may be more appropriate.
- The mode is appropriate choice for nominal data.

Basic Visualisation Tools For Exploration & Description

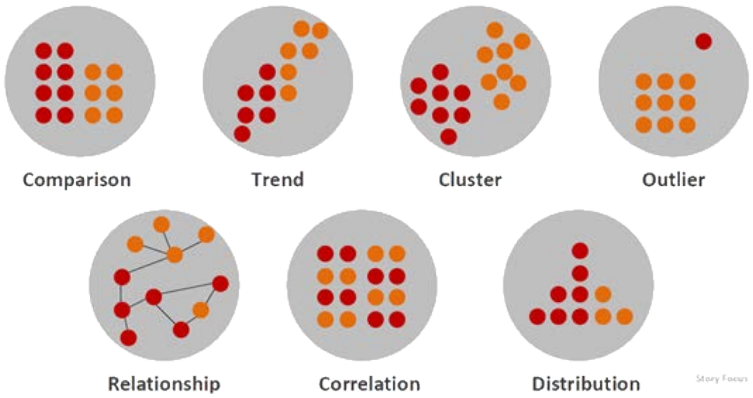
ROLE OF DATA VISUALISATION

As an exploratory tool

- An ability to comprehend huge amounts of data
- Enables problems with data to become immediately apparent

Seeking Analytical patterns

- Visual exploration of data to discover patterns and relationships that are not readily visible



CASE STUDY

- As a part of investigations on recent revenue drop and to better understand the BLITZ customers, a market research was conducted.
- A sample of 300 BLITZ customers (100 customers each from Melbourne, Sydney and Perth) stores were asked a range of questions about BLITZ Products, average spend, shopping frequency and their shopping habits in general.
- All customer responses were recorded and are available to analysis.

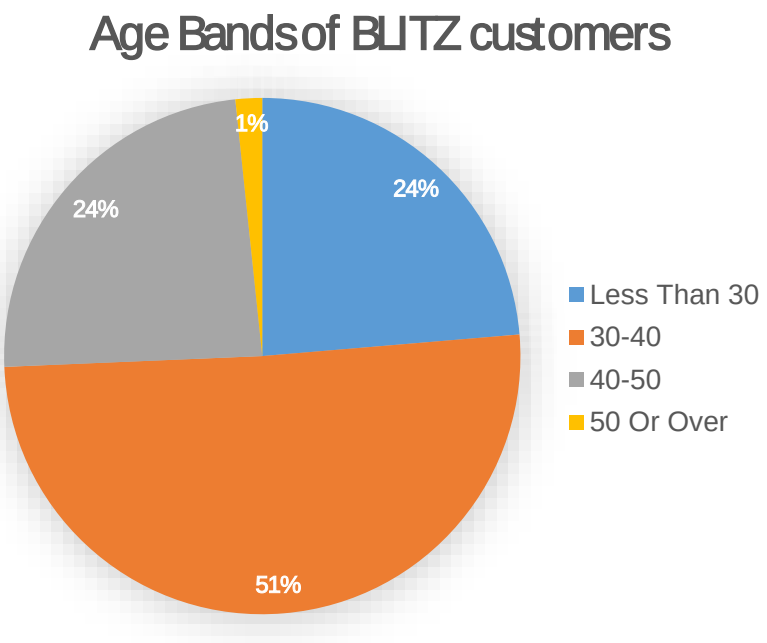
FREQUENCY TABLES
AGE

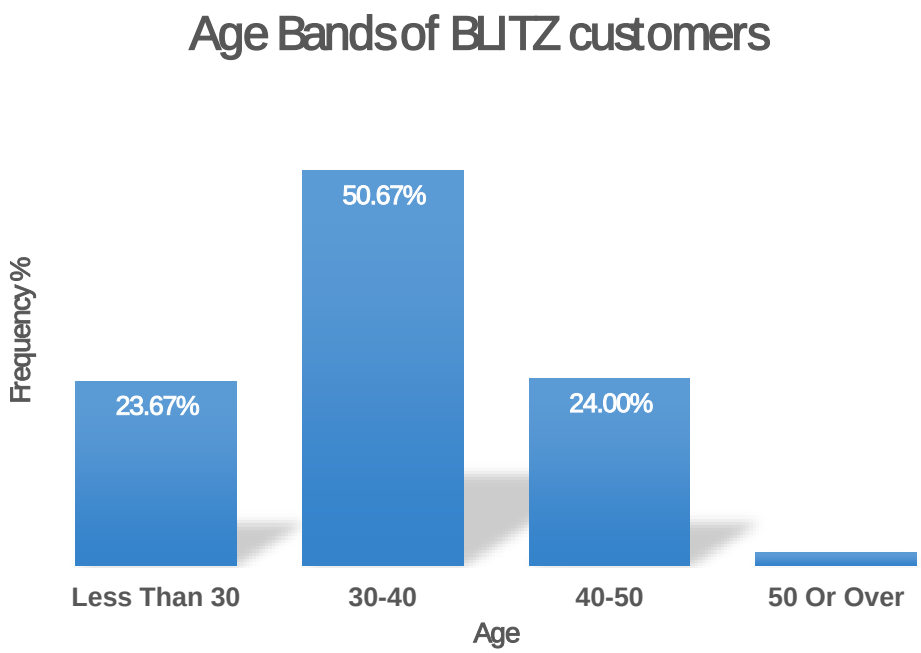
- Simple tool for displaying data.
- It organises data by assigned numerical value, with columns for percent and cumulative percent.

A Frequency Table (Age bands of BLITZ customers)

Value Label	Frequency	Frequency Percentage	Cumulative Percentage
Less than 30	71	23.67%	23.67%
30 - 40	152	50.67%	74.33%
40 - 50	72	24.00%	98.33%
50 or More	5	1.67%	100.00%
	300	100%	

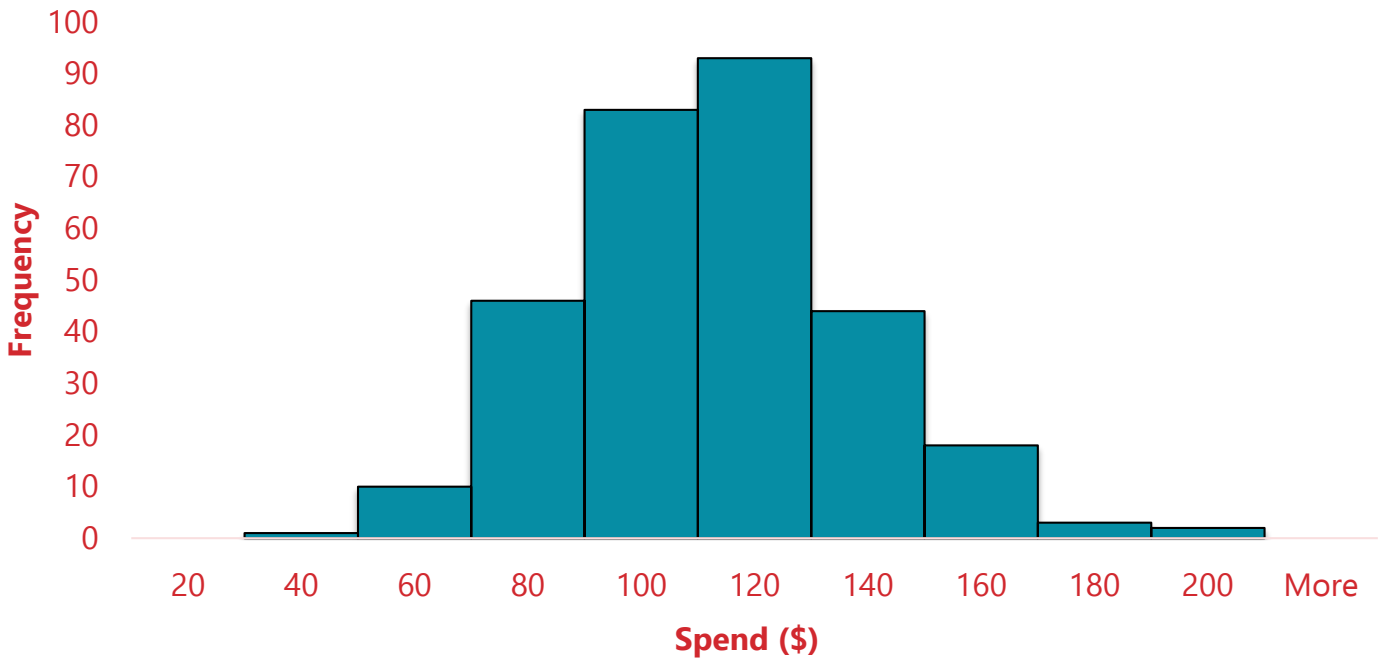
PIE CHART & BAR CHART
AGE





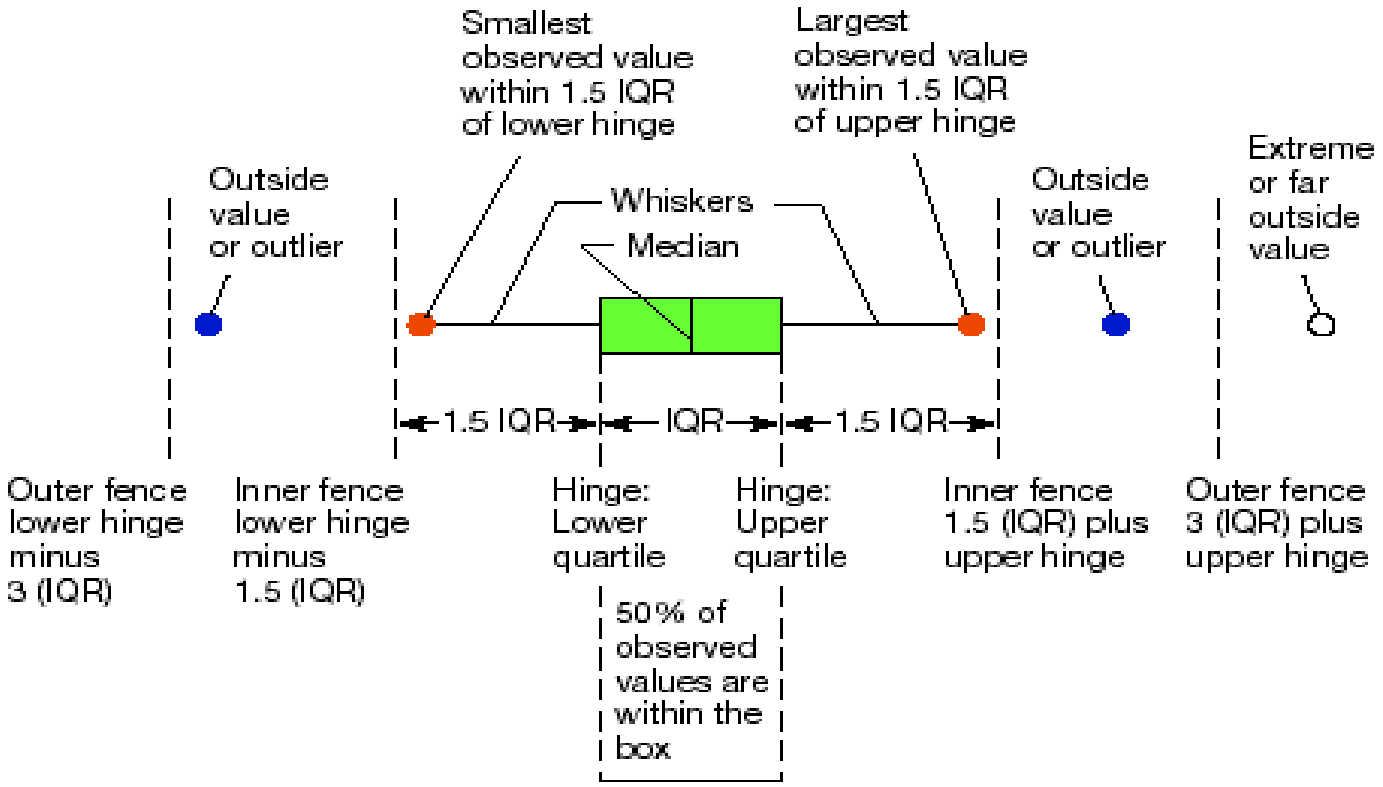
HISTOGRAM
AVERAGE $ SPENT

- Conventional solution for the display of interval-ratio data.
- Histogram of the average spend per visit

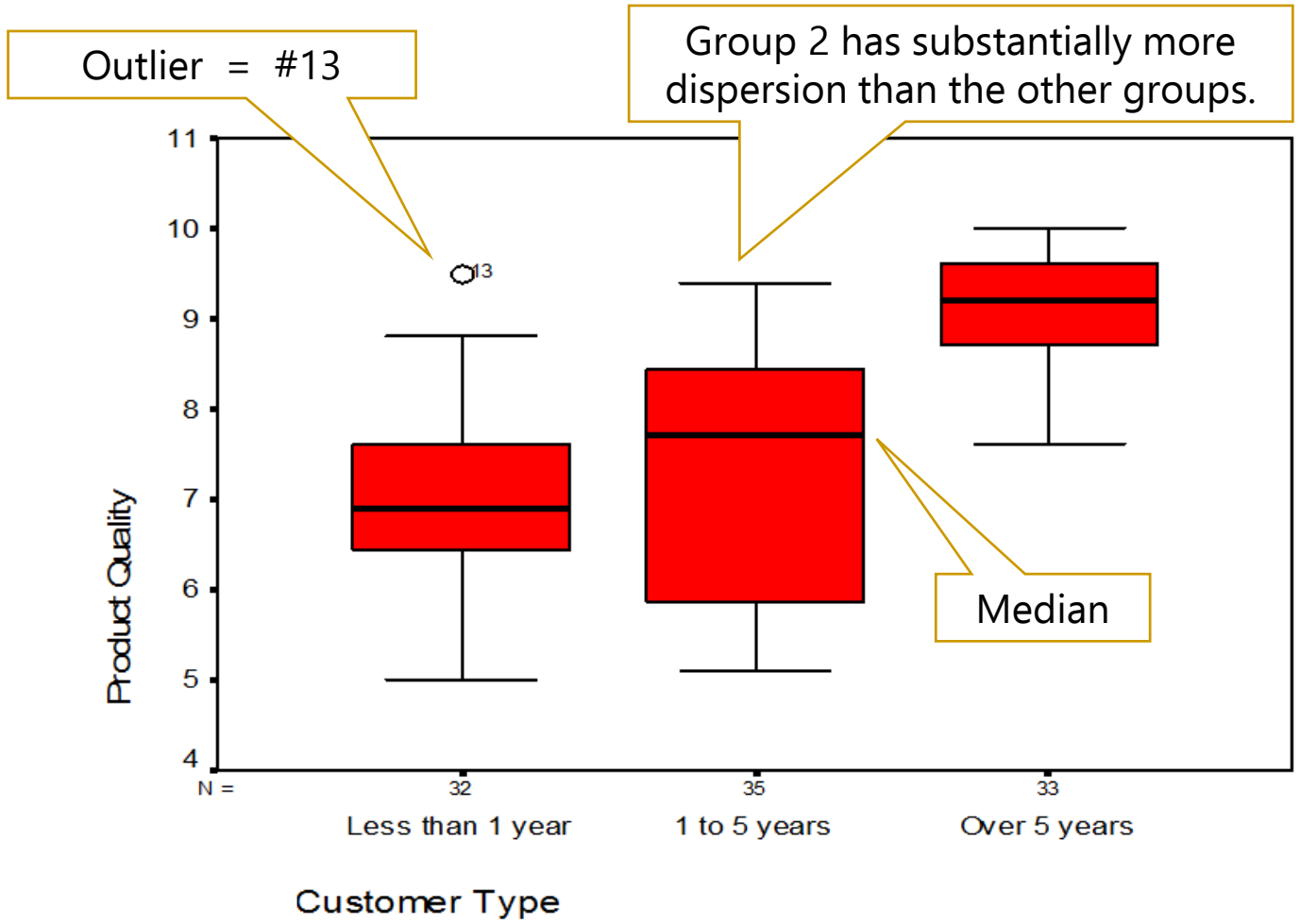


BOXPLOT

- Extensions of the five-number summary of a distribution.



MULTIPLE BOX PLOTS
QUALITY V. CUSTOMER TYPE



CROSS-TABULATION (PIVOT TABLES)
GENDER V. LOYALTY PROGRAM

- Technique for comparing data from two or more categorical variables.

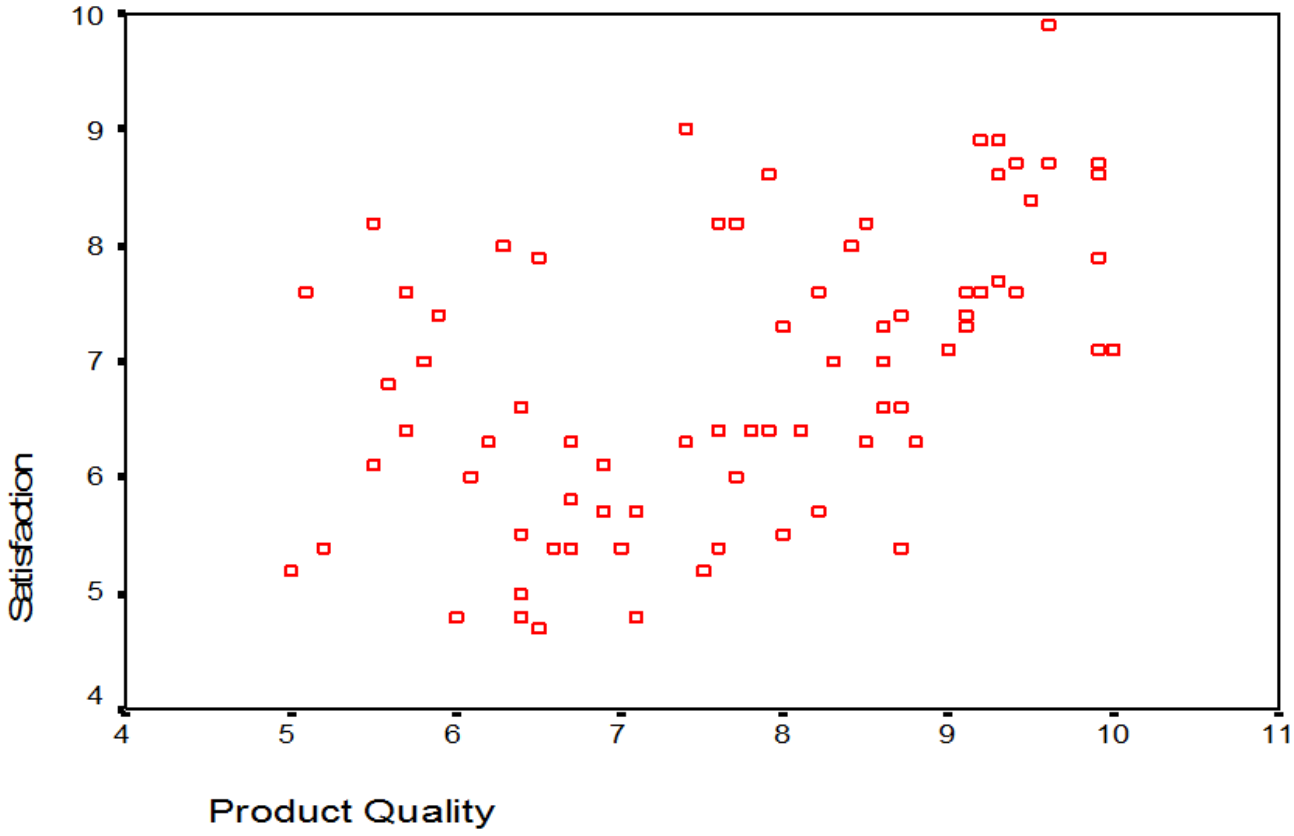
	Loyalty Program Awareness		
	Yes	No	Total
Male	51	71	122
Female	63	115	178
Total	114	186	300

- Percentages make comparisons easier

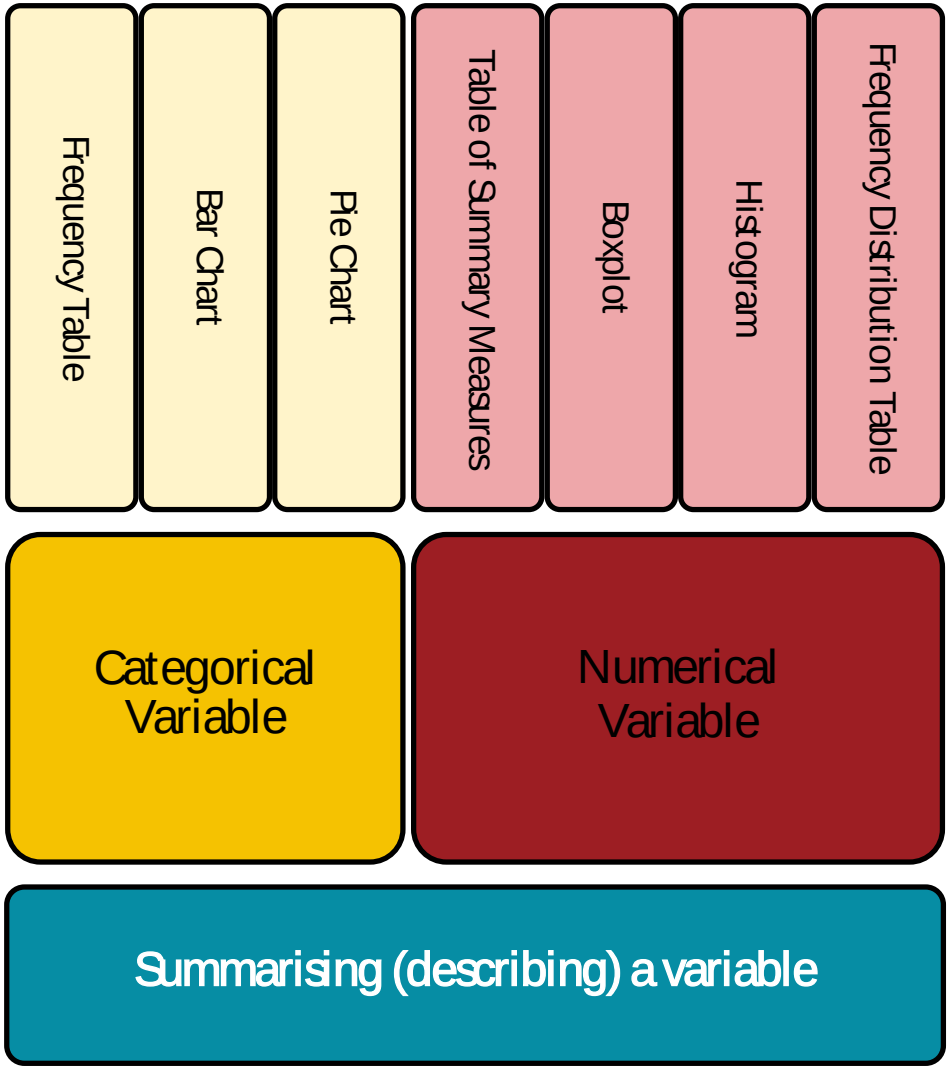
	Loyalty Program Awareness		
	Yes	No	Total
Male	41.80%	58.20%	100%
Female	35.39%	64.61%	100%
Total	38.00%	62.00%	100%

	Loyalty Program Awareness		
	Yes	No	Total
Male	44.74%	38.17%	40.67%
Female	55.26%	61.83%	59.33%
Total	100%	100%	100%

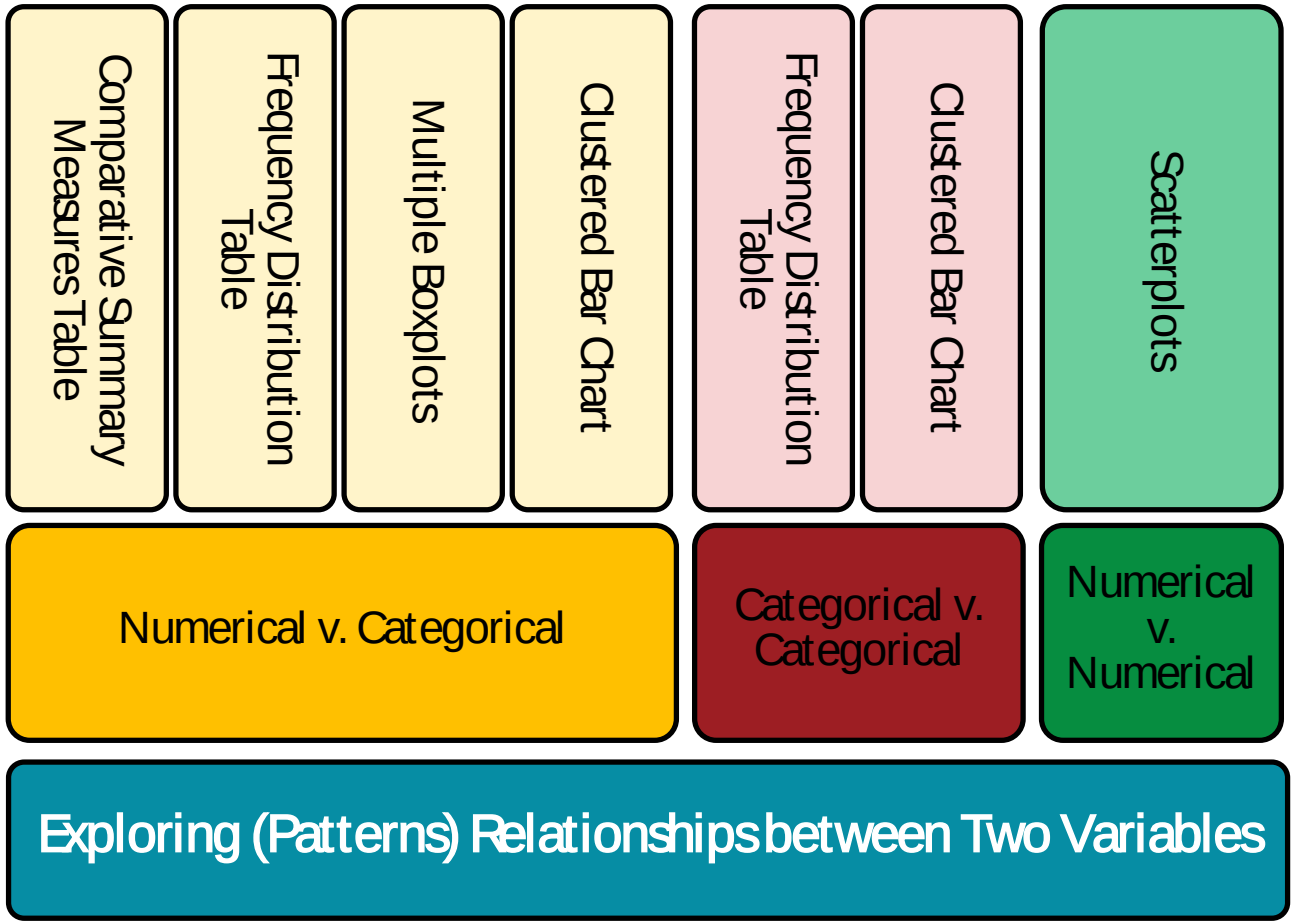
SCATTERPLOT
SATISFACTION V. PRODUCT QUALITY



WHICH VISUALISATION WHEN?



WHICH VISUALISATION WHEN?



Going Beyond Sample Data

ESTIMATION AT A GLANCE…

- How to use a sample mean to estimate true population mean?
- Let's say you are a tyre manufacturer developing a new brand of tyre. For marketing purposes, you intent to measure the average distance travelled by this new brand of tyre…
- So far, 16 new tyres are manufactured.
- How do you go about addressing this issue?

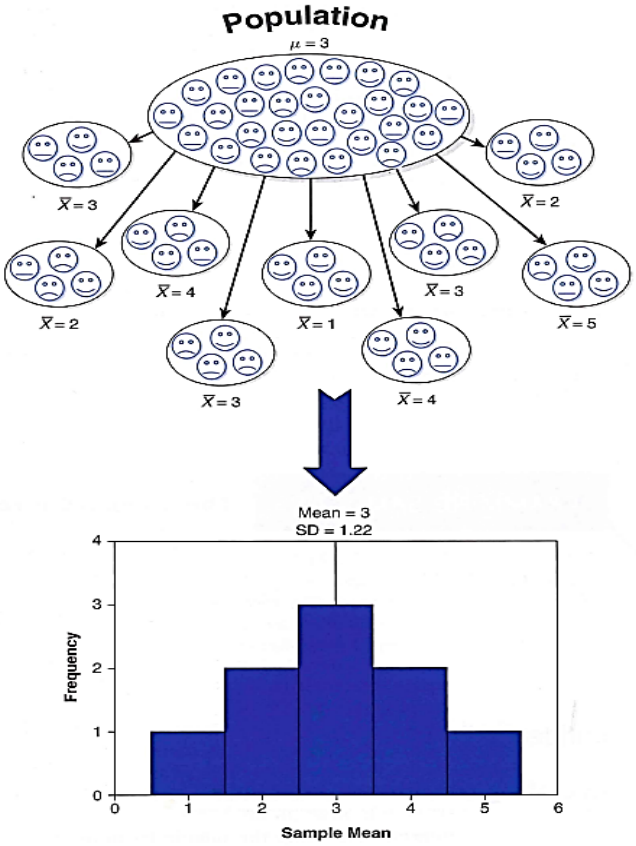
Microsoft Excel Worksheet

STANDARD ERROR

- We collect data from samples because we don't have access to the entire population.
- It is important to know how well a particular sample represents the population.
- Sampling Distribution of Sample Means:

  frequency distribution of all possible sample means of size ***n*** from the same population
- Standard Error:

  The standard deviation of the Sampling Distribution of Sample Means



CENTRAL LIMIT THEOREM

- If the population distribution is NORMAL, the sampling distribution of $\bar{x}$ will be normal no matter what the size of n.
- If the population distribution is NOT NORMAL, the sampling distribution of $\bar{x}$ will be normal or approximately normal when n is sufficiently large enough, generally 30 or more.
- In both cases:

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Confidence Interval Estimation

ESTIMATION OF PARAMETERS

Point estimate

- An estimate of the population parameter in the form of a single value, usually the sample statistic.

Interval Estimate

- A specified range of numbers within which a population parameter is expected to lie.
- An estimate of the population parameter based on the knowledge that it will be equal to the sample statistic plus or minus a small sampling error.

Confidence Interval

- A percentage that indicates the long run probability results will be correct; it states the long run percentage of confidence intervals that will include the true population parameter.

MARGIN OF ERROR AND AN INTERVAL ESTIMATE

- An interval estimate is constructed by subtracting and adding the margin of error (ME), to a point estimate:

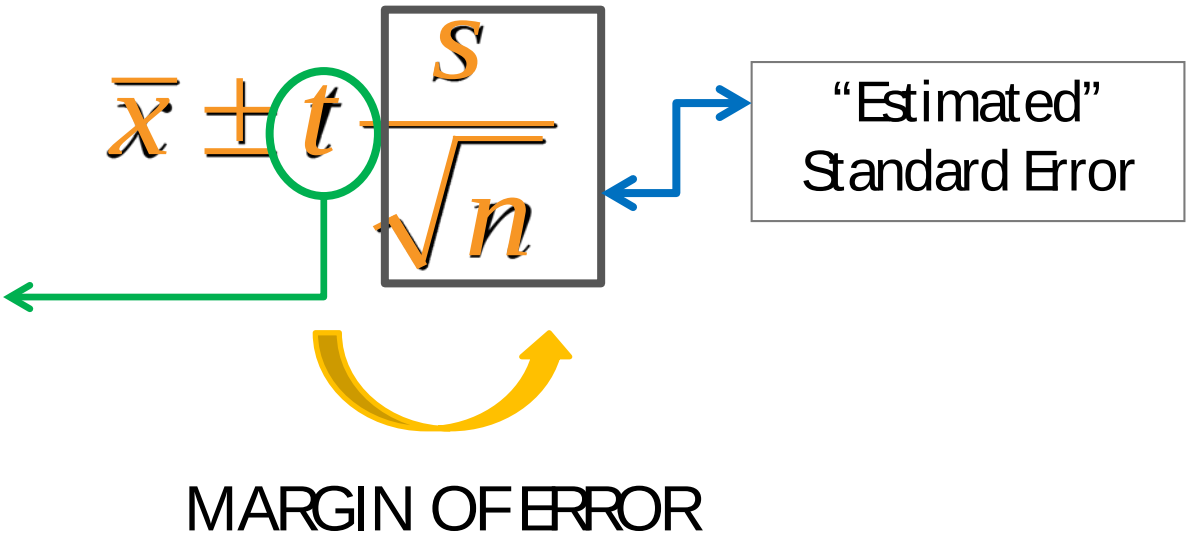
$$\text{Sample Statistic} \pm \text{ME}$$

- An interval estimate of the population mean is:

$$\bar{x} \pm \text{ME}$$

CONFIDENCE INTERVAL FOR μ

t indicates how **wide** the confidence interval is in terms of **standard errors**. The *t* **value** ties back to the level of confidence and degrees of freedom (n-1).

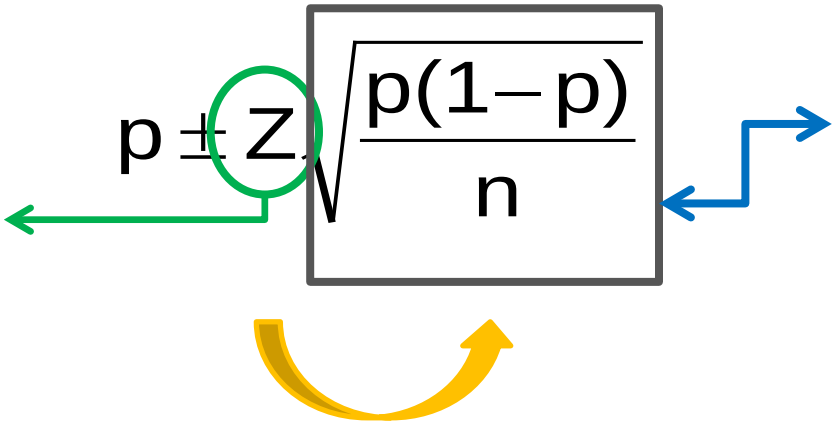


SAMPLING DISTRIBUTION OF PROPORTIONS

- Sampling distribution of a proportion can be thought of as the theoretical distribution that we would observe from taking repeated samples and each time computing the sample proportion.
- Statistical theory tells us the sampling distribution of a sample proportion, p is:
- Approximately Normally distributed when sample size, n, is large enough: $np \geq 5$; and $n(1-p) \geq 5$
- Has mean of $\bar{x}_p = \boldsymbol{\pi}$ (population proportion)
- Has standard deviation (SD) of $\frac{\sqrt{\pi(1-\pi)}}{n}$

CONFIDENCE INTERVAL FOR PROPORTION π

z indicates how **wide** the confidence interval is in terms of **standard errors**. The ***z* value** ties back to the level of confidence.



SUMMARY

- Descriptive statistics, exploratory data analysis can be used to better understand the business problem and data.
- Uncertainty is involved in almost all common business scenarios.
- Many people believe they have an intrinsic understanding of error however, good decision makers are systematic to their approach in solving problems and don't rely on just their 'gut feel' for a situation
- Use of statistical thinking (and data) provides evidence for a decision maker to justify the chosen outcome
- Sampling distributions are used to manage sample error by using probability.
- In practice, sampling distributions are used to calculate confidence interval estimates.

QUESTIONS?

DEAKIN UNIVERSITY AUSTRALIA
Worldly