



Stats One-Liner Definitions

Balaji J

Frequency

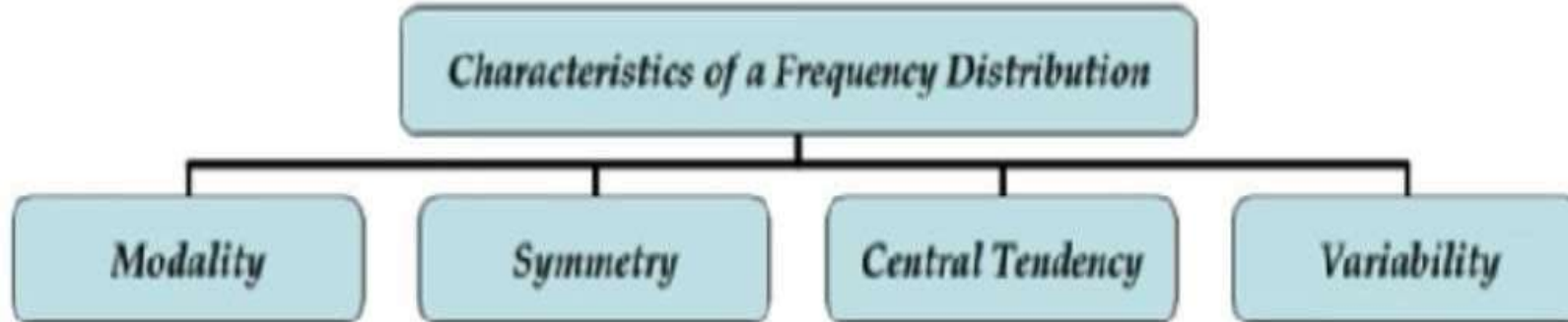
- Frequency is how often something occurs.

Frequency Distribution

- A **frequency distribution** tells how **frequencies** are **distributed** over values.

Summarizing Data

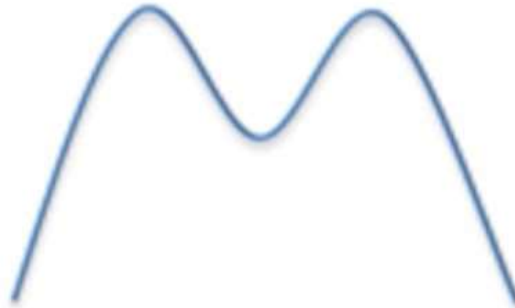
- We can derive Summary about the data with the help of these characteristics of Frequency Distribution, Lets see how!



Unimodal



Bimodal



Multimodal

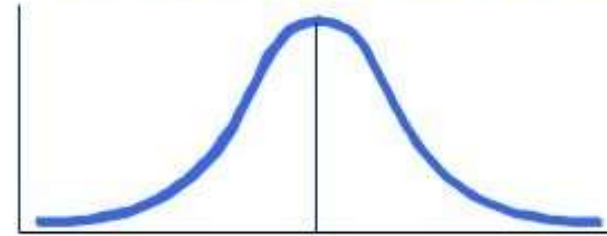


Modality

- Tells you about the peaks in the distribution
- **The modality of a distribution is determined by the number of peaks it contains.**

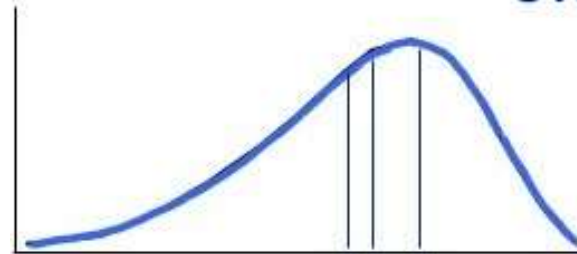
Symmetry

Skewness



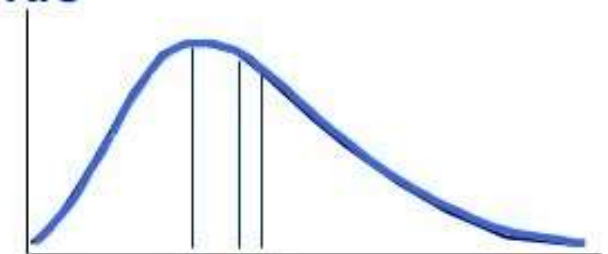
Mode = Mean = Median

SYMMETRIC



Mean Median Mode

SKEWED LEFT
(negatively)

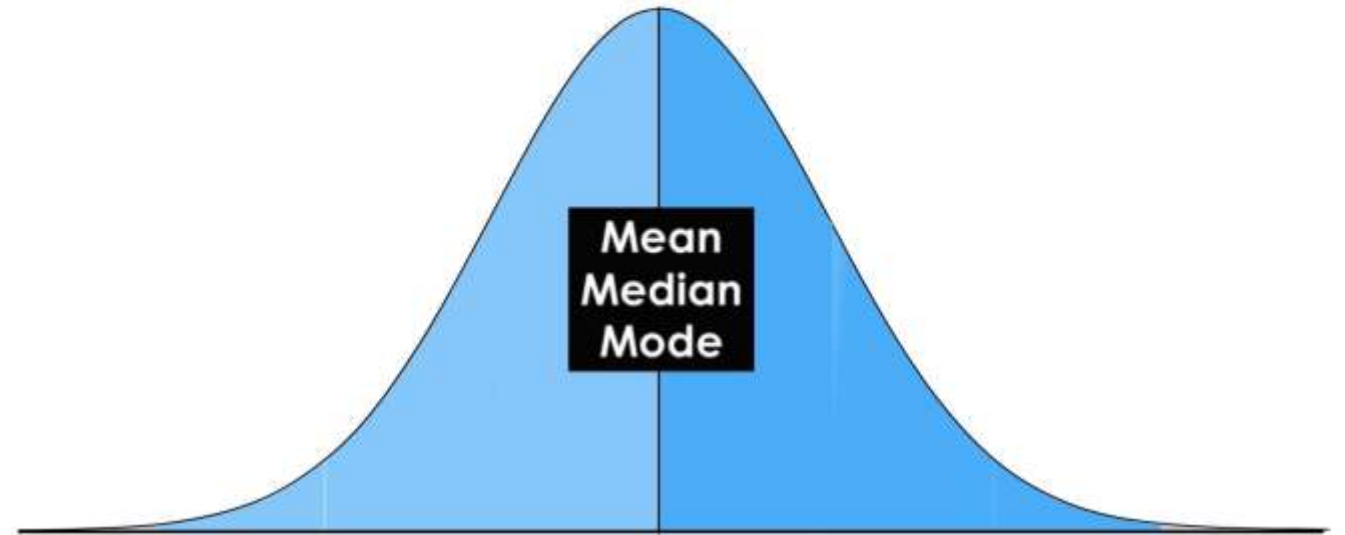


Mode Median Mean

SKEWED RIGHT
(positively)

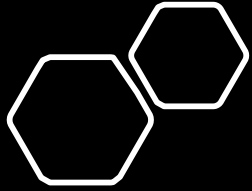
Normal Distribution

**A Data is normally Distributed if
your data is symmetrical, bell-
shaped and unimodal.**



Measures Of Central Tendency (mean, median, mode)

- Describe the “location” of the data
- Fail to describe the “shape” of the data
 - mean = “calculated average”
 - median = “middle value”
 - mode = “most occurring value”



Mean vs. Median

- The mean can be influenced by *outliers*.
- The mean of $\{2,3,2,3,2,12\}$ is 4
- The median is 2.5
- The median is much closer to most of the values in the series!

Measures Of Spread (Range, Variance and SD)

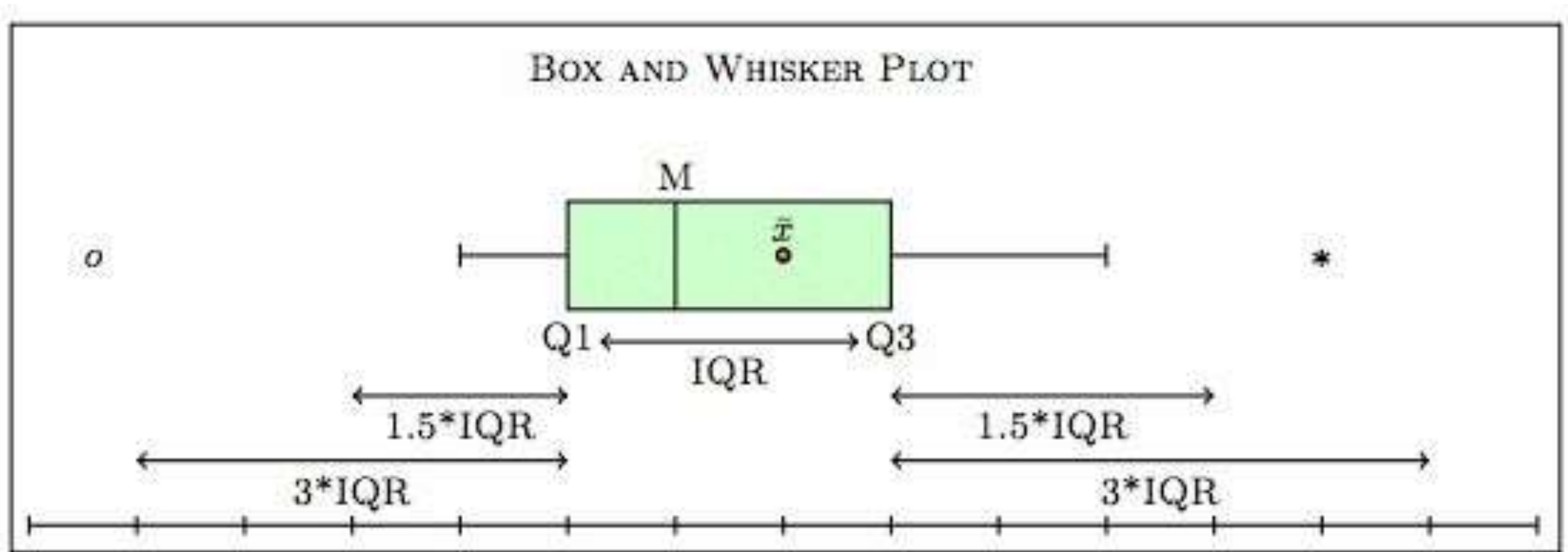
- How wide individuals values are distributed
 - Range – Max – Min
 - **range** should suggest how diversely spread out the values are
- Standard Deviation and Variance – Deviation from the Mean
- How far a set of numbers is spread out from their average value.

Coefficient Of Variation

- It represents the ratio of the standard deviation to the mean.
- Less Coefficient of variance means less risk and more consistency.
- More coefficient of variance means more risk and less consistency.

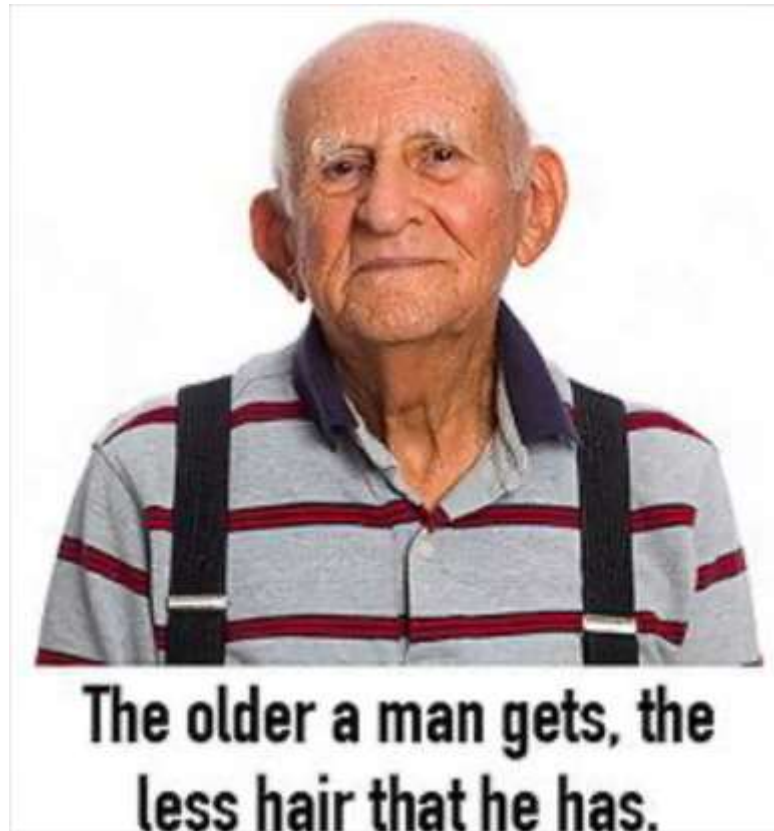
Box and Whisker plots

Helps you find Outliers



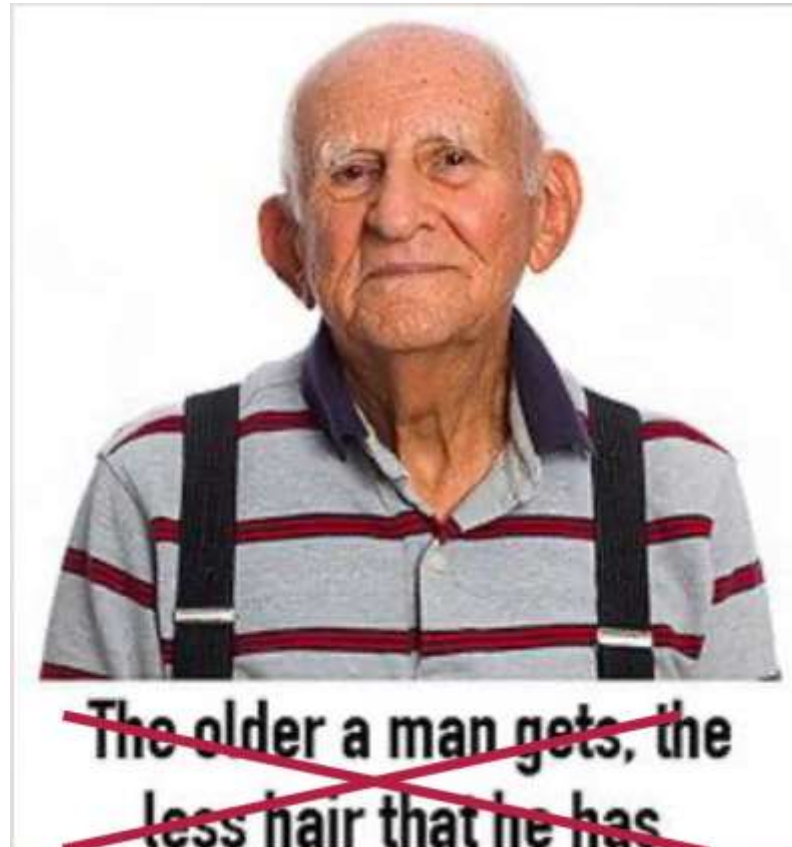
Causation

- Does Age 'cause' Hair loss?







Causation

- Does Age 'cause' Hair loss? - NO



Correlation vs Covariance

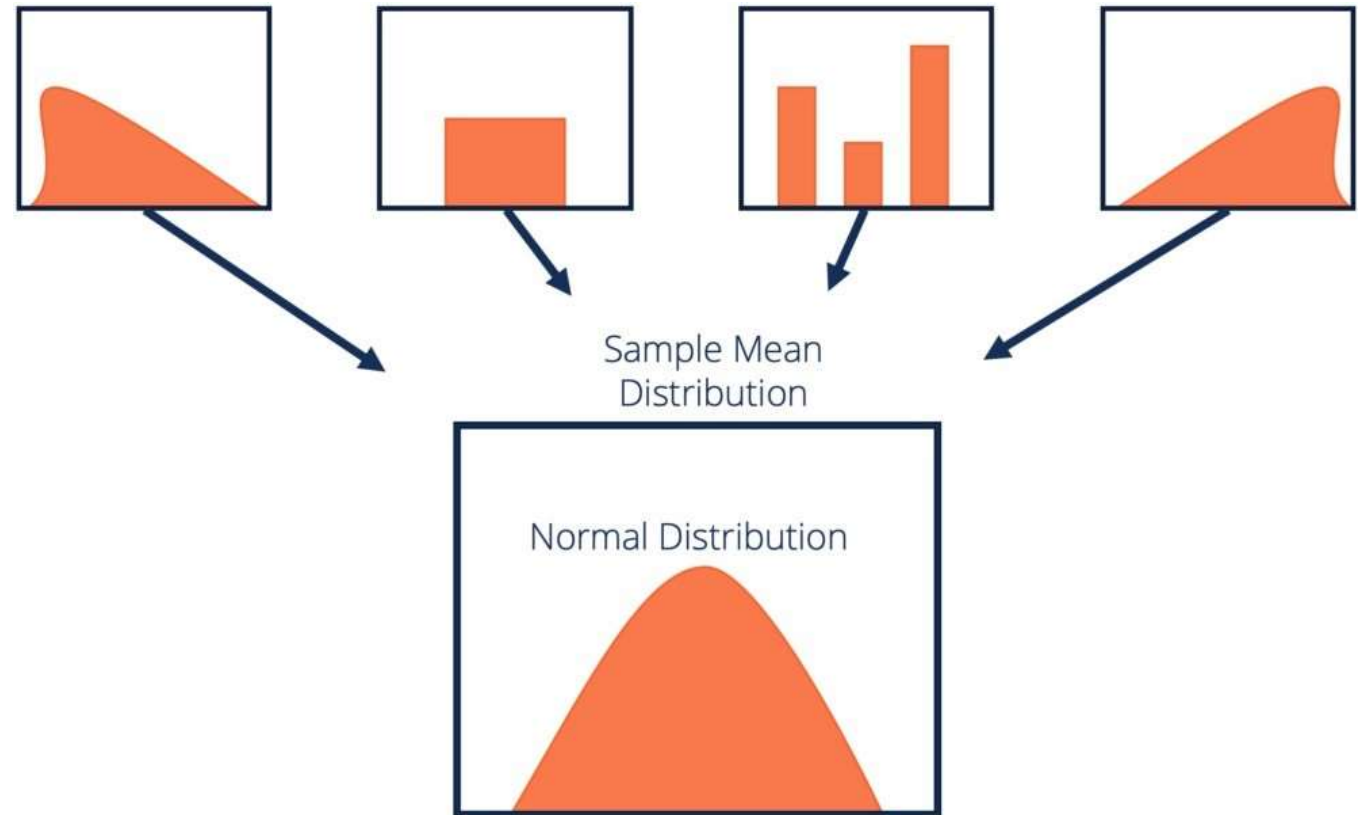
		<u>Relationship</u>	<u>Definition</u>	
	+		Many people who smoke also drink.	Correlation
	=		Smoking has been proven to cause lung cancer	Causation

Covariance vs Correlation

BASIS FOR COMPARISON	COVARIANCE	CORRELATION
Meaning	Covariance is a measure indicating the extent to which two random variables change in tandem.	Correlation is a statistical measure that indicates how strongly two variables are related.
Values	Lie between $-\infty$ and $+\infty$	Lie between -1 and $+1$

Central Limit theorem

- Population mean – μ
Population SD – σ
- Take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed

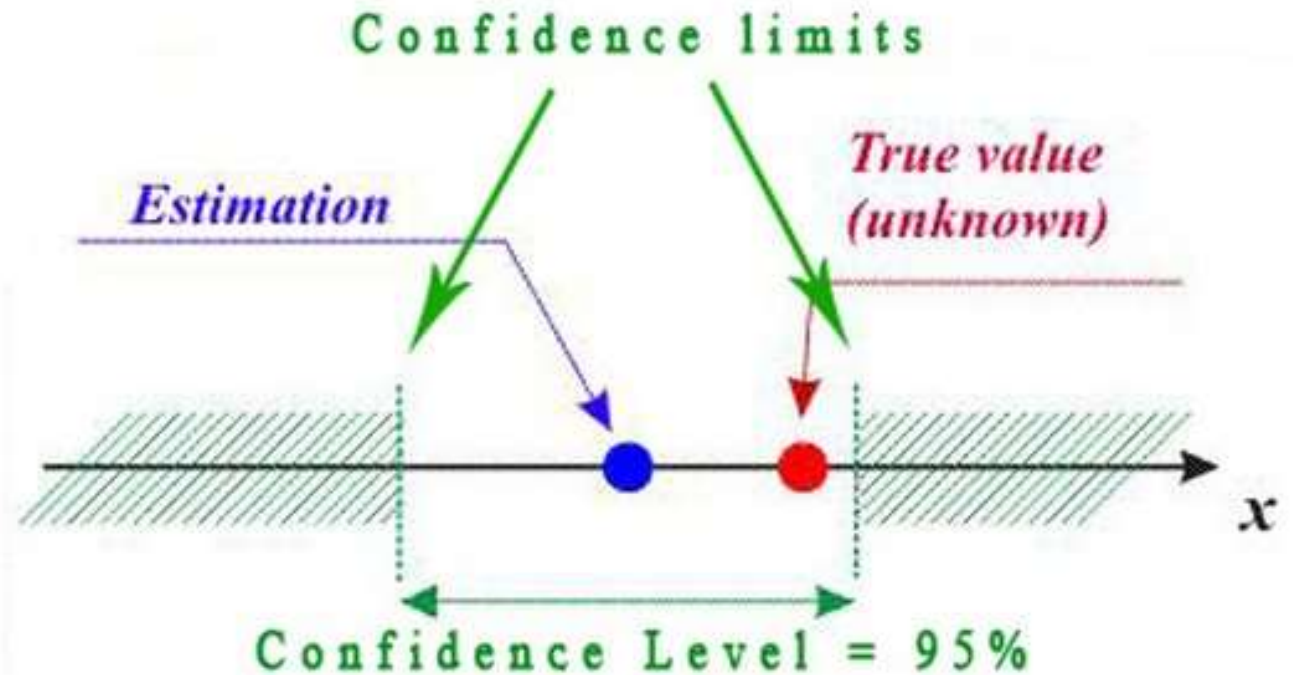


Standard Error

- Standard error of the mean describes how far the sample mean may be deviated from the population mean

Confidence Interval

A **Confidence Interval** is a range of values we are fairly sure our true value lies in



Margin of Error

- A margin of error tells you how many percentage points your results will differ from the real population value
- $ME = z * SE$
Where z is the z score, and SE is the Standard Error

HYPOTHESIS TESTING STEPS

1. State the Null Hypothesis (H_0) and Alternate Hypothesis (H_1)
2. Choose the Level of Significance
3. Find Critical Values
4. Find test Statistic
5. Draw your conclusion

Null and Alternate Hypothesis

Null Hypothesis

$$H_0$$

A statement about a population parameter.

We test the likelihood of this statement being true in order to decide whether to accept or reject our alternative hypothesis.

Can include =, \leq , or \geq sign.

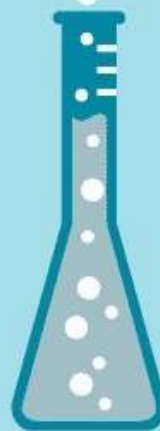
Alternative Hypothesis

$$H_a$$

A statement that directly contradicts the null hypothesis.

We determine whether or not to accept or reject this statement based on the likelihood of the null (opposite) hypothesis being true.

Can include a \neq , $>$, or $<$ sign.





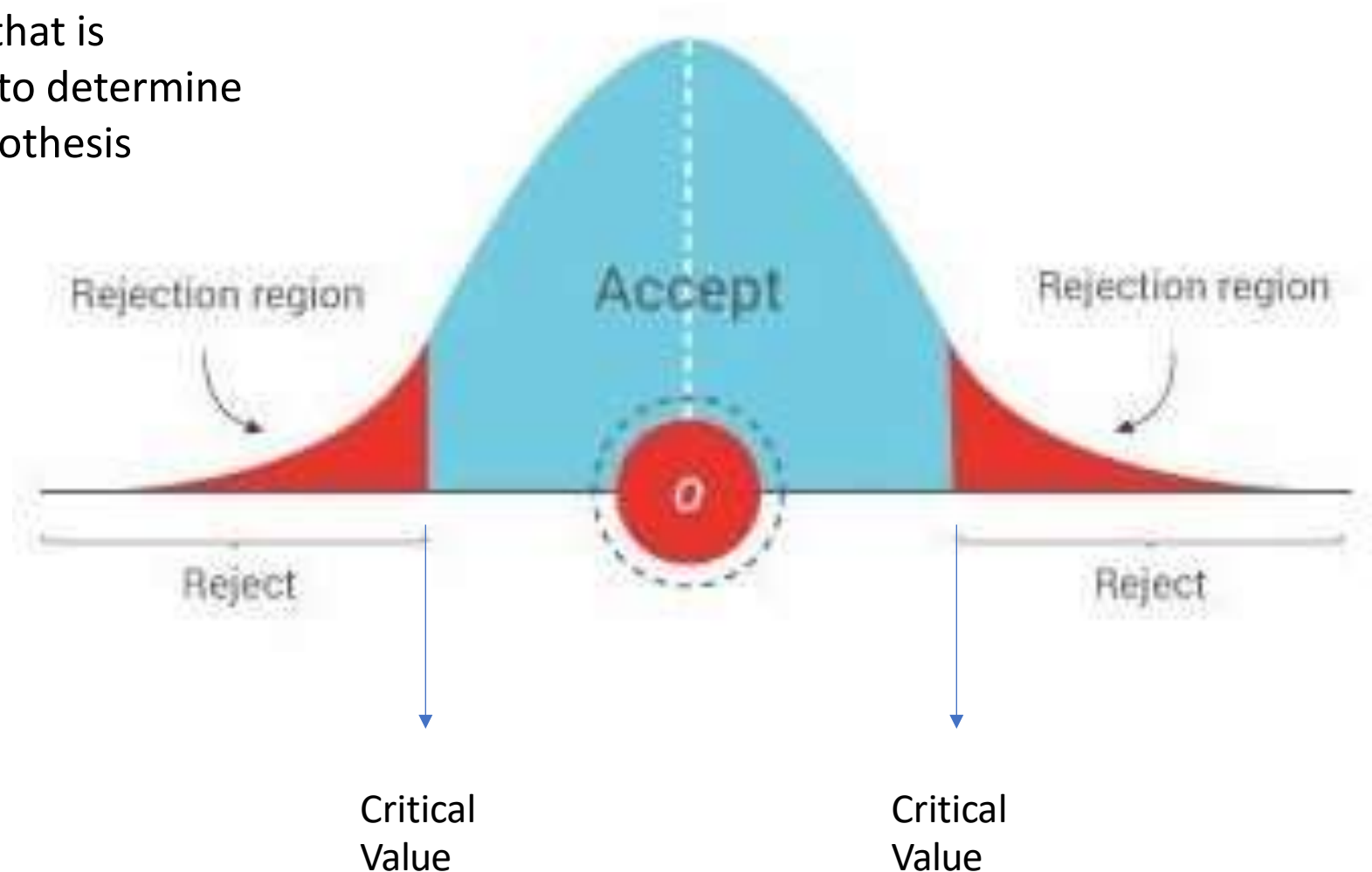
Level of Significance

- The **significance level**, also denoted as alpha or α , is the **probability** of rejecting the null hypothesis



Critical Values

In hypothesis testing, a **critical value** is a point on the test distribution that is compared to the test **statistic** to determine whether to reject the null hypothesis



Test Statistic

- A **test statistic** is a random variable that is calculated from sample data and used in a hypothesis **test**.
- You can use **test statistics** to determine whether to reject the null hypothesis.

Test statistic	Associated test	Sample size	Information given	Distribution	Test question
z-score	z-test	large samples ($n > 30$)	<ul style="list-style-type: none"> Standard deviation of the population (this will be given as σ) Population mean or proportion 	Normal	Do these two populations differ?
t-statistic	t-test	Two small samples ($n < 30$)	<ul style="list-style-type: none"> Standard deviation of the sample (this will be given as s) Sample mean 	Normal	Do these two samples differ?
f-statistic	ANOVA	Three or more samples	<ul style="list-style-type: none"> Group sizes Group means Group standard deviations 	Normal	Do any of these three or more samples differ from each other?
chi-squared	chi-squared test	Two samples	<ul style="list-style-type: none"> Number of observations for each categorical variable 	Any	Are these two categorical variables independent?