Measures of Relationship

Covariance

- Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.

- The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.

- The greater this number, the more reliant the relationship. Positive covariance denotes a direct relationship and is represented by a positive number.

- A negative number, on the other hand, denotes negative covariance, which indicates an inverse relationship between the two variables. Covariance is great for defining the type of relationship, but it's terrible for interpreting the magnitude.

Covariance

Let Σ(X) and Σ(Y) be the expected values of the variables, the covariance formula can be represented as:

$$\text{Covariance}(\mathrm{x},\mathrm{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

where,

- xi = data value of x
- yi = data value of y
- x̄ = mean of x
- ȳ = mean of y
- N = number of data values.

Correlation

Correlation, on the other hand, is a standardized measure that quantifies the strength and direction of the linear relationship between two variables. Unlike covariance, correlation values always fall between -1 (perfect negative correlation) and 1 (perfect positive correlation), with 0 indicating no linear correlation.

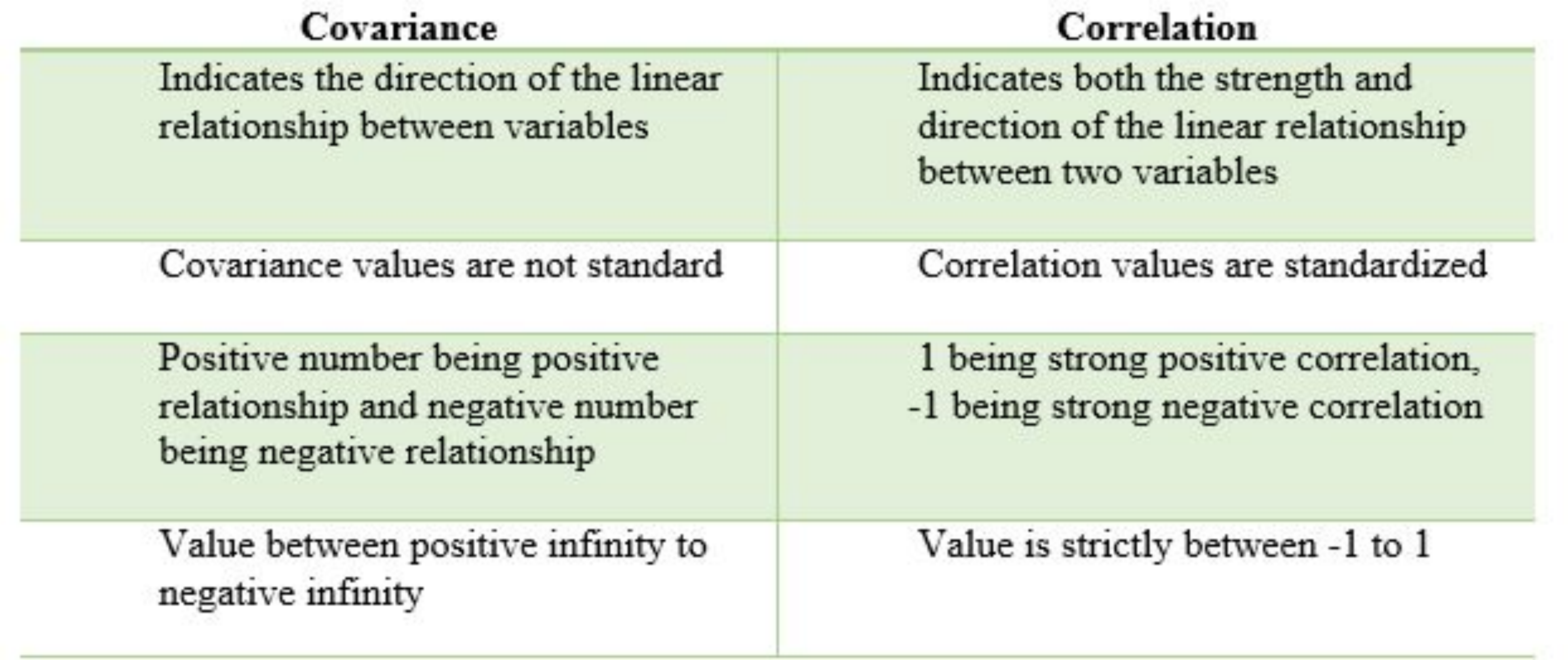
Formula for Correlation (Pearson's Correlation Coefficient):

$$\text{Correlation } (\rho) = \mathrm{Cov}(X, Y) / (\sigma_X * \sigma_Y)$$

Where:

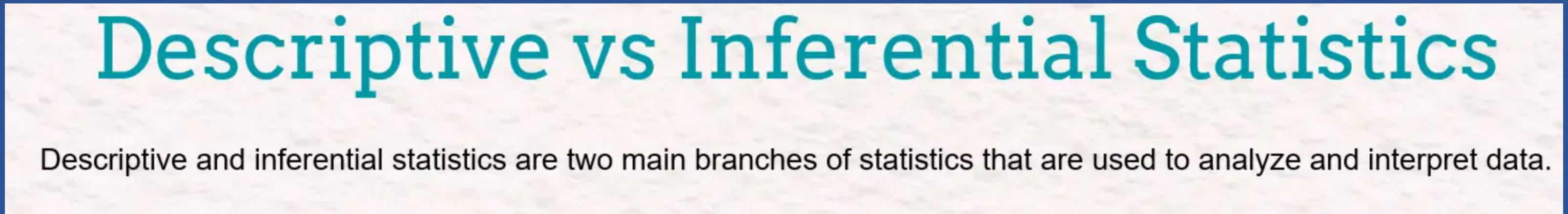
- ρ is the correlation coefficient between variables X and Y.
- Cov(X, Y) is the covariance between X and Y.
- σ_X is the standard deviation of X.
- σ_Y is the standard deviation of Y.

Covariance VS Correlation



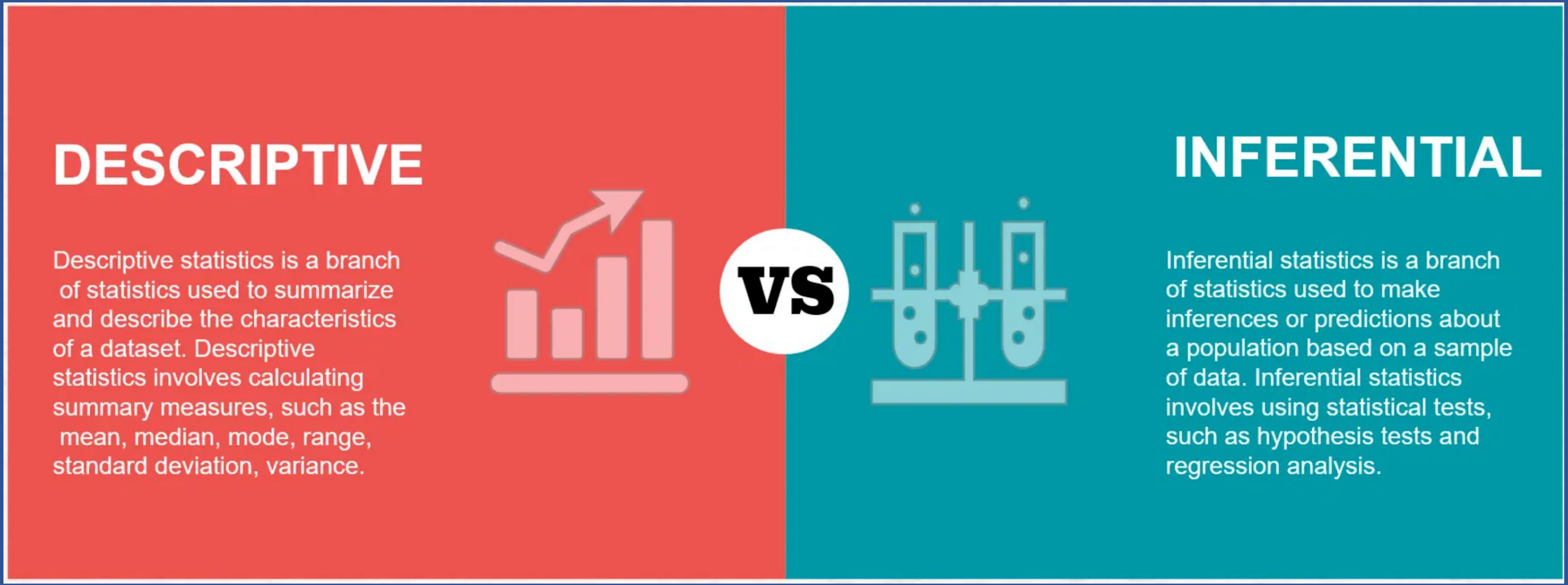
Covariance	Correlation
Indicates the direction of the linear relationship between variables	Indicates both the strength and direction of the linear relationship between two variables
Covariance values are not standard	Correlation values are standardized
Positive number being positive relationship and negative number being negative relationship	1 being strong positive correlation, -1 being strong negative correlation
Value between positive infinity to negative infinity	Value is strictly between -1 to 1

Types of Statistics



Descriptive vs Inferential Statistics

Descriptive and inferential statistics are two main branches of statistics that are used to analyze and interpret data.



DESCRIPTIVE

Descriptive statistics is a branch of statistics used to summarize and describe the characteristics of a dataset. Descriptive statistics involves calculating summary measures, such as the mean, median, mode, range, standard deviation, variance.

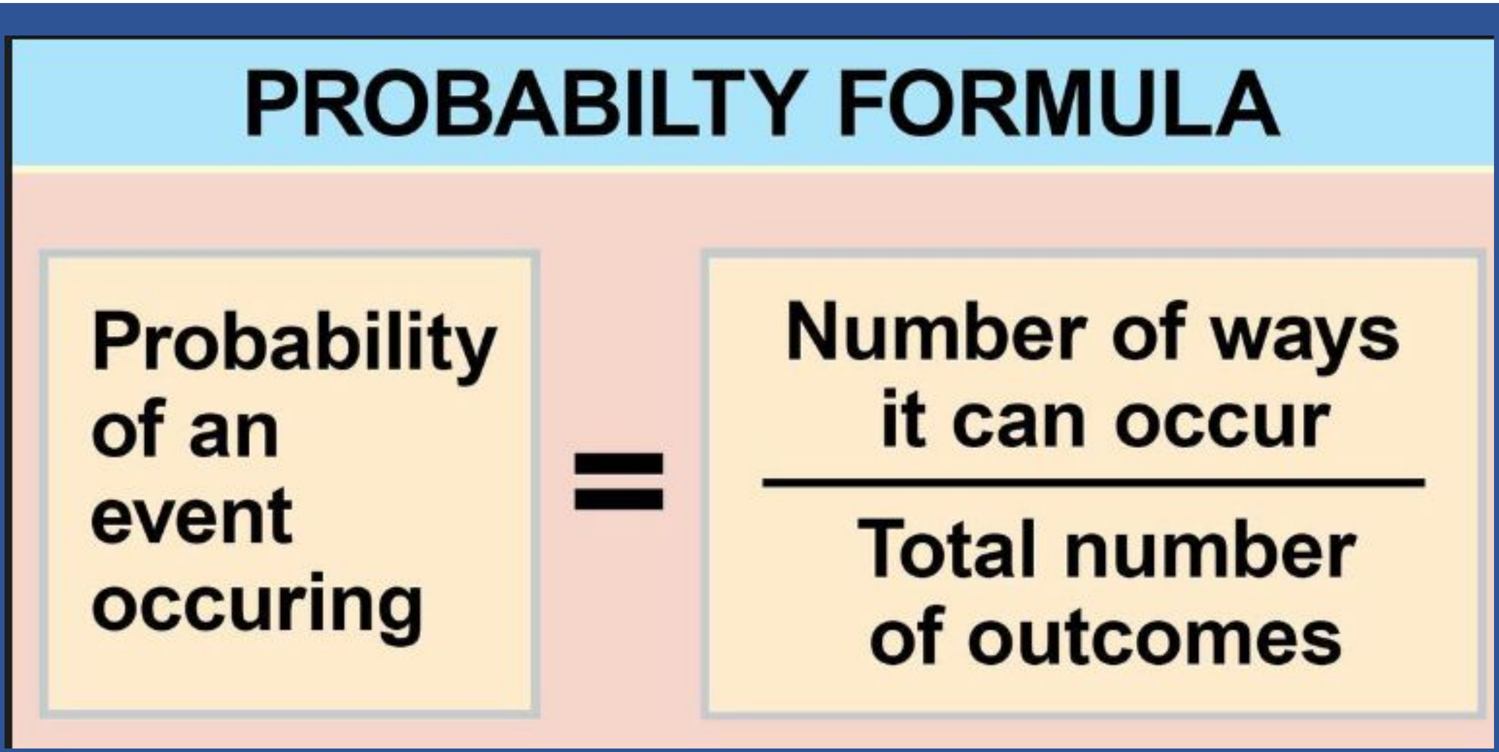
INFERENTIAL

Inferential statistics is a branch of statistics used to make inferences or predictions about a population based on a sample of data. Inferential statistics involves using statistical tests, such as hypothesis tests and regression analysis.

Introduction to Probability

Probability is the likelihood of an event occurring

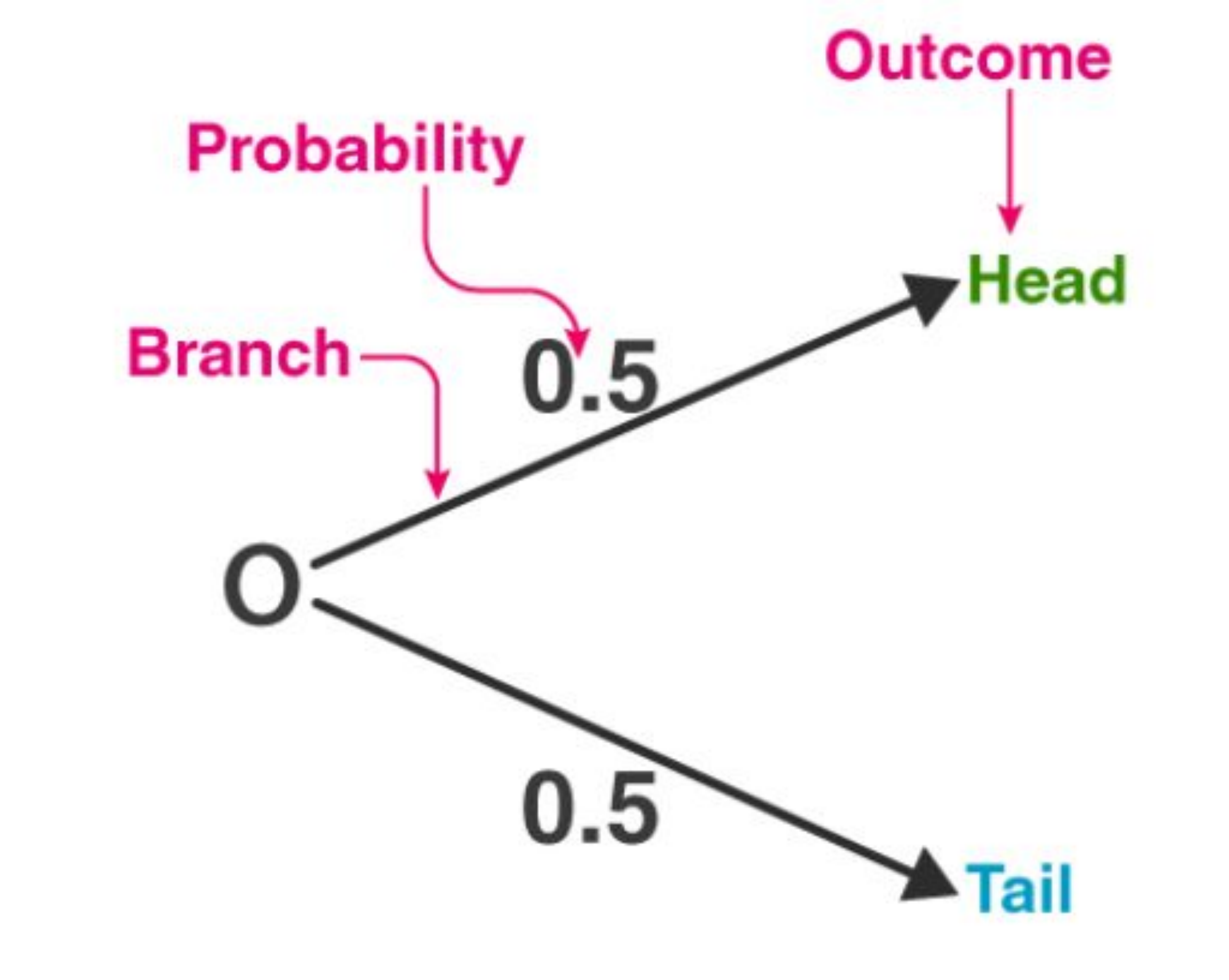
- It is typically expressed as a number between 0 and 1.
- 0 represents an impossible event (certainty it won't happen), and 1 represents a certain event (certainty that it will happen).
- A probability of 0.5, for example, represents a 50% chance or a 50-50 probability of an event occurring.



PROBABILTY FORMULA

$$\text{Probability of an event occuring} = \frac{\text{Number of ways it can occur}}{\text{Total number of outcomes}}$$

Introduction to Probability



Marginal
&
Conditional
Probability

Marginal Probability: Marginal probability refers to the probability of an event occurring without considering any other events. It focuses on a single event in isolation.

Conditional Probability: Conditional probability is the probability of an event occurring given that another event has already occurred or is known to have occurred. It takes into account a specific condition or context for the event.

	MALE	FEMALE	TOTAL
GAME OF THRONES	80	120	200
WEST WORLD	100	25	125
OTHERS	50	125	175
TOTAL	230	270	500

	MALE	FEMALE	TOTAL
GAME OF THRONES	80	120	200
WEST WORLD	100	25	125
OTHERS	50	125	175
TOTAL	230	270	500

Probability distribution

	MALE	FEMALE	TOTAL
GAME OF THRONES	0.16	0.24	0.4
WEST WORLD	0.2	0.05	0.25
OTHERS	0.1	0.25	0.35
TOTAL	0.46	0.54	1

Joint Probability

P(Female and GOT)
=0.24

P(Female ∩ GOT)
=0.24

	MALE	FEMALE	TOTAL
GAME OF THRONES	0.16	0.24	0.4
WEST WORLD	0.2	0.05	0.25
OTHERS	0.1	0.25	0.35
TOTAL	0.46	0.54	1

Joint Probability Distribution

SUMS to 1

Marginal Probability Distribution

SUMS to 1

	MALE	FEMALE	TOTAL
GAME OF THRONES	0.16	0.24	0.4
WEST WORLD	0.2	0.05	0.25
OTHERS	0.1	0.25	0.35
TOTAL	0.46	0.54	1

Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Q. Noni(F) just got an HBO subscription . What is the chance that her favorite show will be Game Of Thrones?

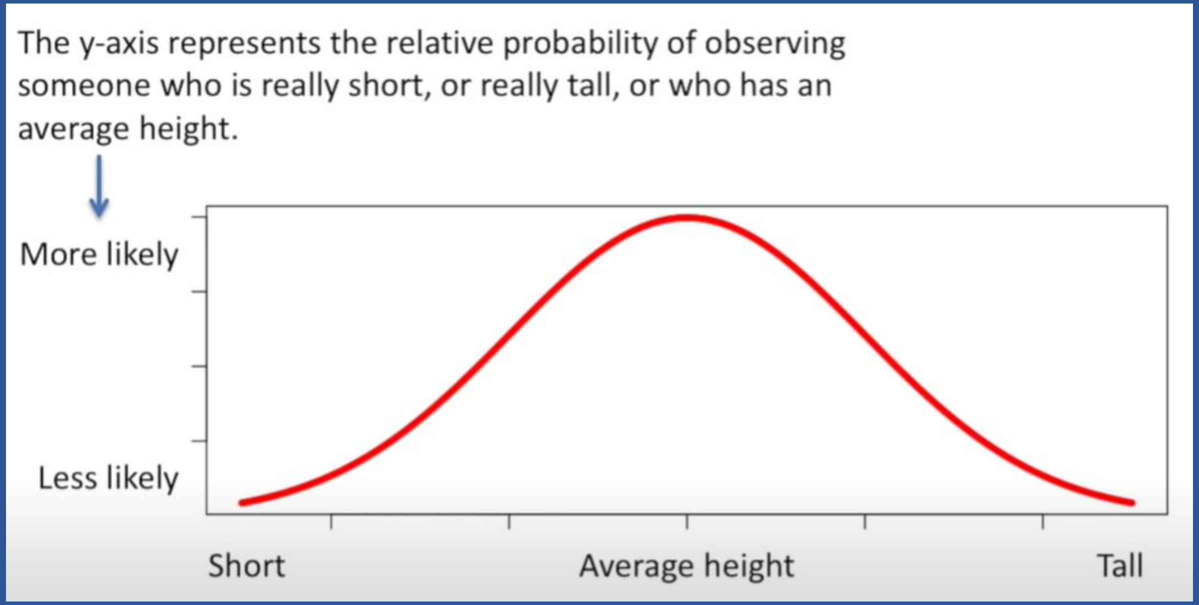
P(GOT|Female)=0.24/0.54=0.4444

Q. Given that a subscriber's favorite show is West World. What is the probability that they are male?

P(Male|West World)= 0.2/0.25=0.80

Normal Distribution

Normal Distribution

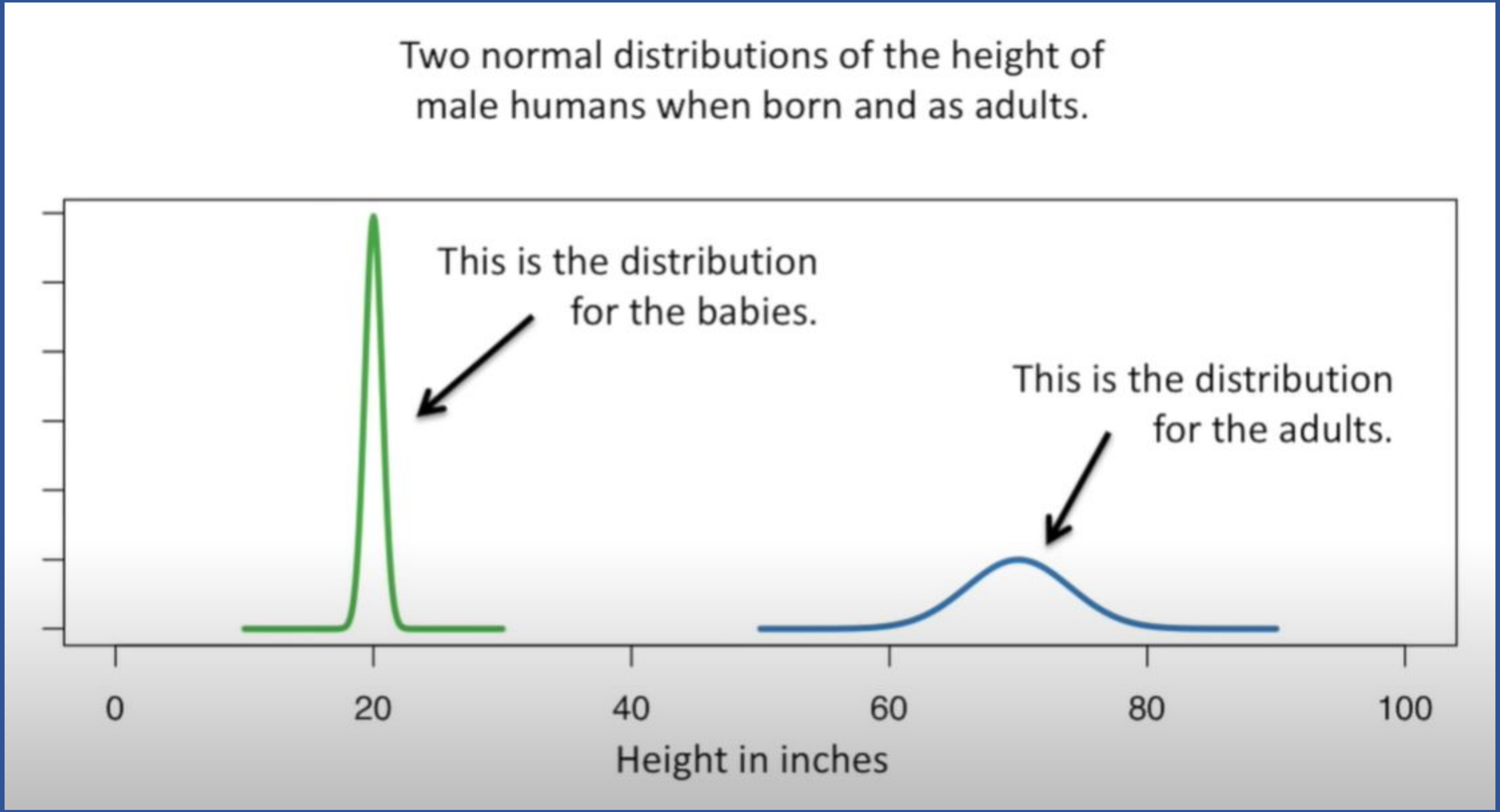


The y-axis represents the relative probability of observing someone who is really short, or really tall, or who has an average height.

A normal distribution, also known as a Gaussian distribution or a bell curve, is a probability distribution that is symmetric and follows a specific mathematical shape. It is characterized by the following key properties:

1. **Symmetry:** The normal distribution is symmetric around its mean (average) value. The mean, median, and mode are all equal and located at the center of the distribution.
2. **Bell-shaped:** The distribution is bell-shaped, with the highest point (peak) at the mean, and it tapers off gradually on both sides.

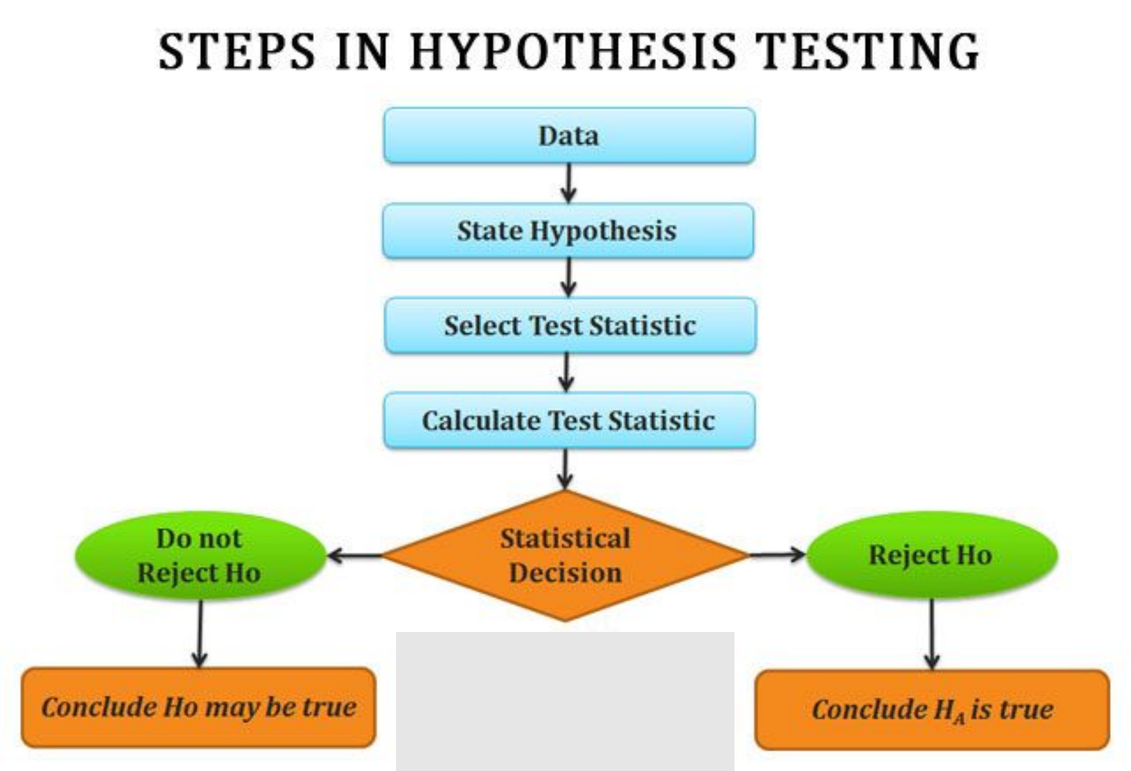
Normal Distribution


Two normal distributions of the height of male humans when born and as adults.

Hypothesis Testing

Hypothesis testing is a fundamental concept in statistics used to make decisions or draw conclusions about population parameters based on sample data. It involves formulating two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (H1 or Ha) and conducting statistical tests to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

- Null Hypothesis (H0): This is the default or initial assumption that there is no effect, no difference, or no relationship. It often represents the status quo or a null condition.
- Alternative Hypothesis (H1 or Ha): This is the statement we want to test. It represents a specific claim or hypothesis about a population parameter that we are interested in.



STEPS IN HYPOTHESIS TESTING

Data

State Hypothesis

Select Test Statistic

Calculate Test Statistic

Statistical Decision

Do not Reject Ho

Reject Ho

Conclude Ho may be true

Conclude H_A is true

A random sample of 50 items gives the mean of 6.2 and variance 10.24. Can it be regraded as drawn from normal population with mean 5.4 at 5% level of significance?