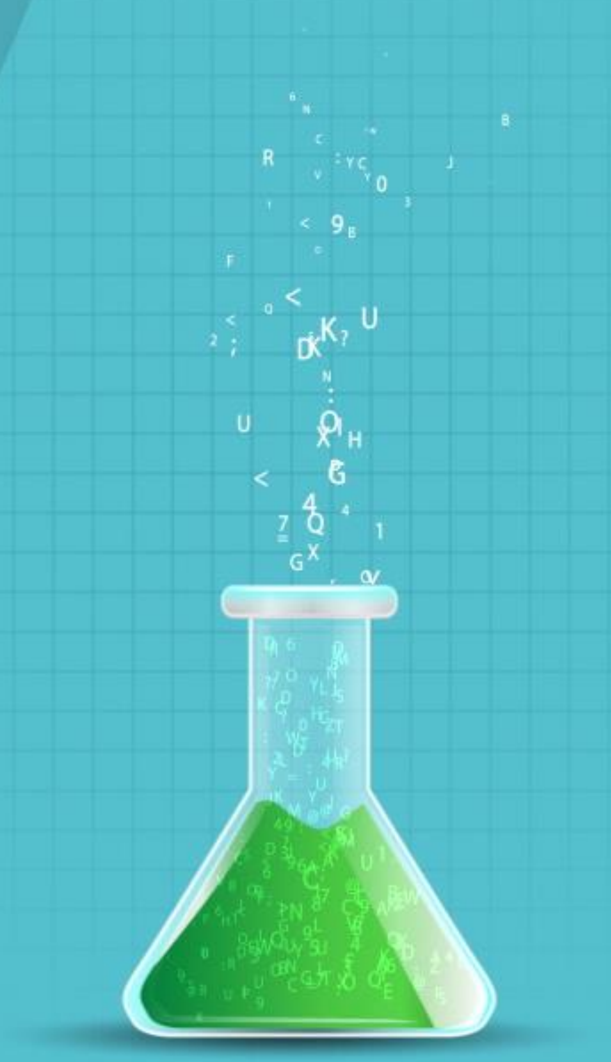
Data Science with Python

Lesson 3 – Statistical Analysis and Business Applications

DATA
SCIENCE

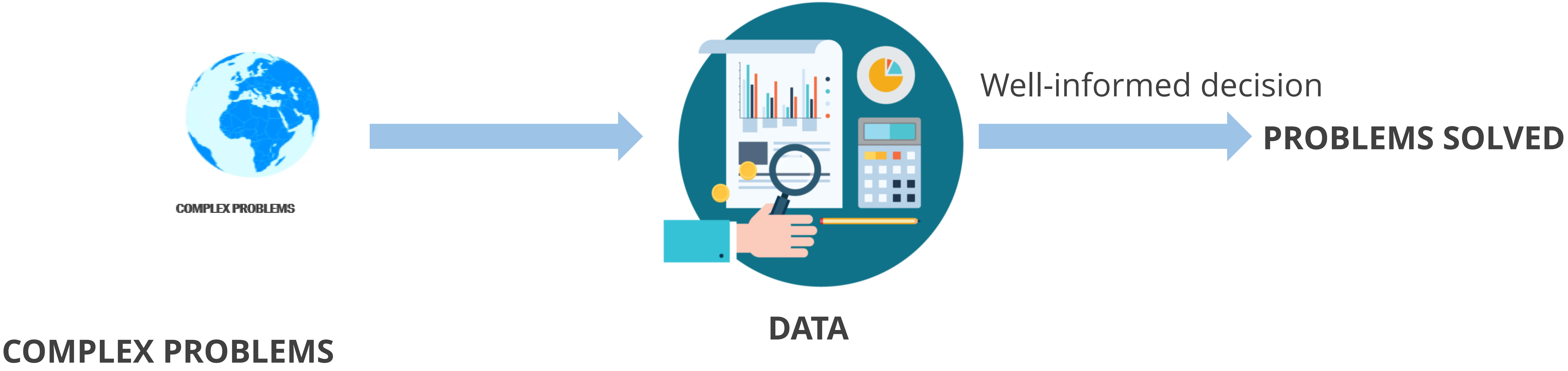
What You'll Learn

- The difference between statistical and non-statistical analysis
- The two major categories of statistical analysis and their differences
- The statistical analysis process
- Mean, median, mode, and percentile
- Data distribution and the various methods of representing it
- Hypothesis testing and the Chi square test
- Types of frequencies
- Correlation matrix and its uses



Introduction to Statistics

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.



COMPLEX PROBLEMS

DATA

Well-informed decision

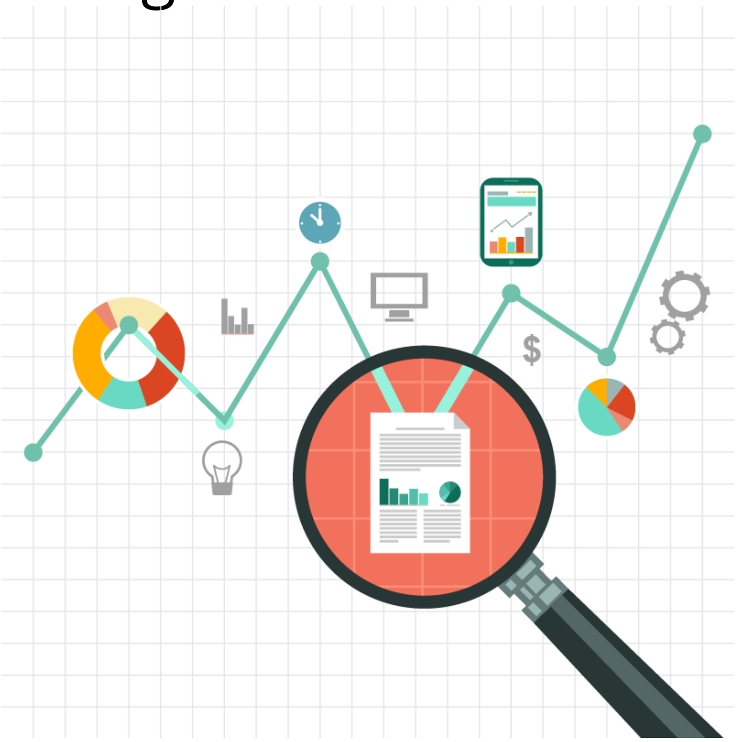
PROBLEMS SOLVED

Introduction to Statistics

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

Tools available to analyze data:

- Statistical principles
- Functions
- Algorithms



What you can do using statistical tools:

- Analyze the primary data
- Build a statistical model
- Predict the future outcome

Statistical and Non-statistical Analysis

Statistical Analysis

Statistical Analysis is:

- scientific
- based on numbers or statistical values
- useful in providing complete insight to the data

Non-statistical Analysis

Non-statistical Analysis is:

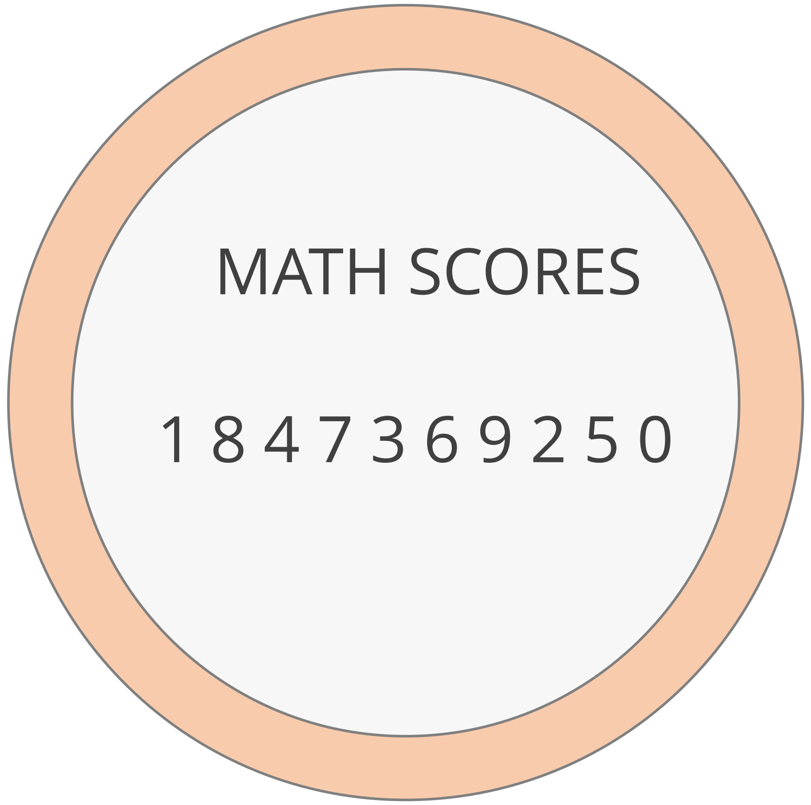
- based on very generic information
- exclusive of statistical or quantitativ

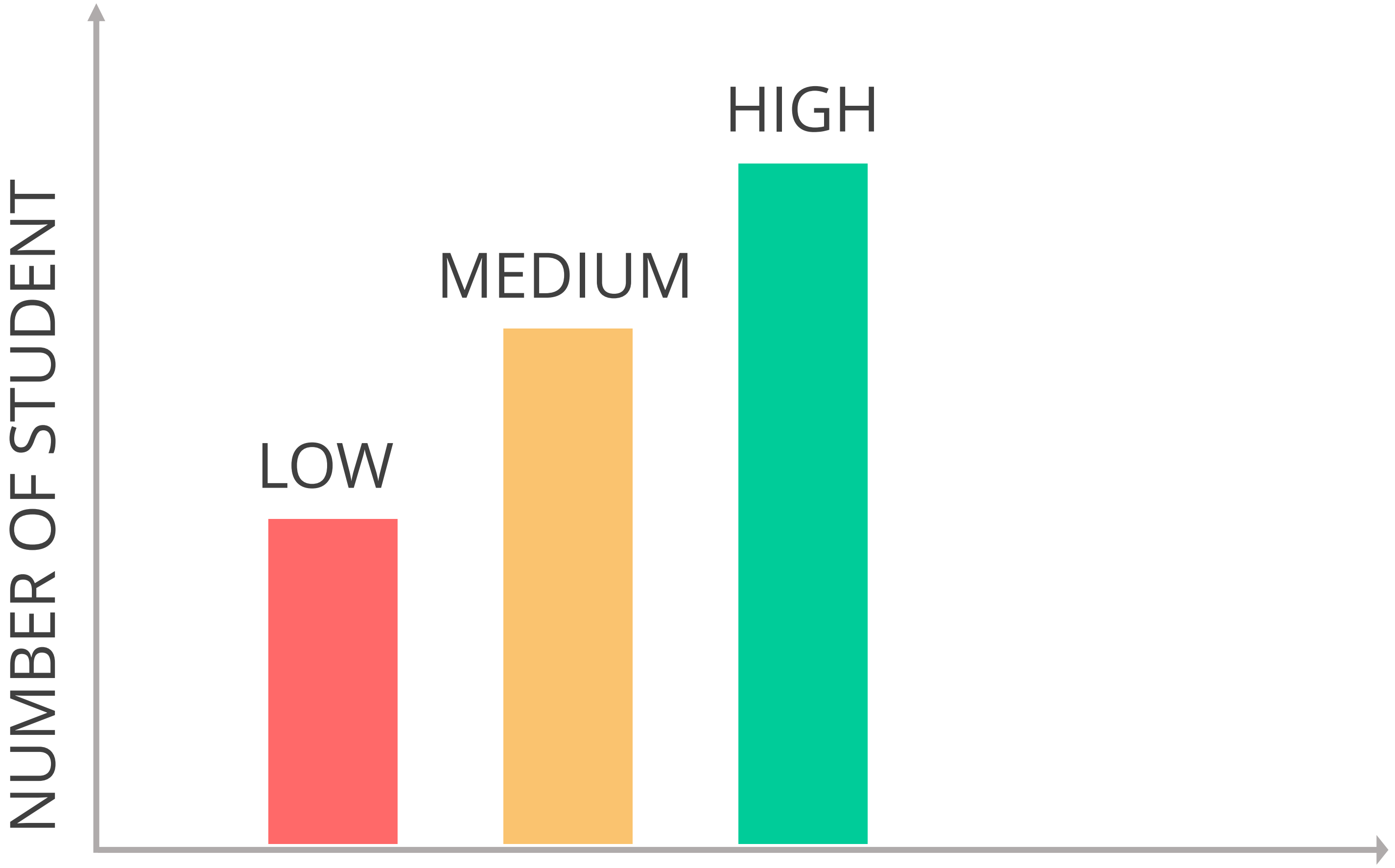
Although both forms of analysis provide results, quantitative analysis provides more insight and a clearer picture. This is why statistical analysis is important for businesses.

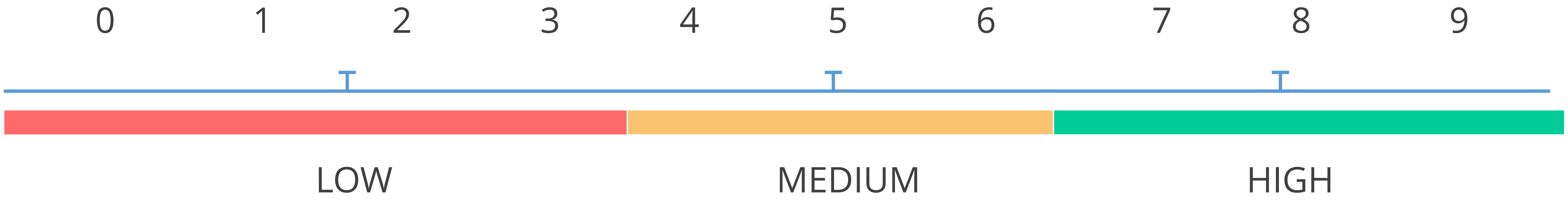
Major Categories of Statistics

There are two major categories of statistics: Descriptive analytics and inferential analytics

Descriptive analysis organizes the data and focuses on the main characteristics of the data.







Major Categories of Statistics

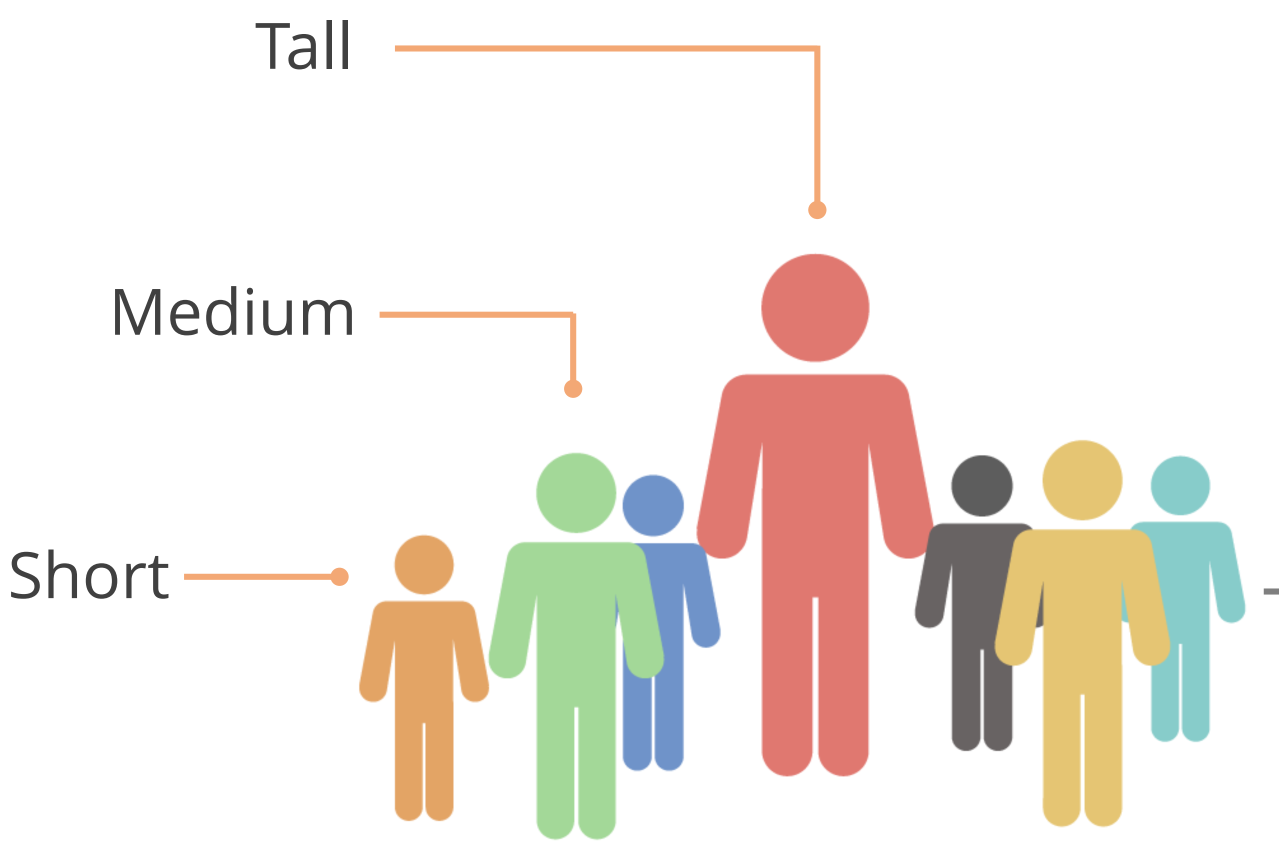
Inferential analytics uses the probability theory to arrive at a conclusion.

- Random sample is drawn from the population
- Used to describe and make inferences about the population

Inferential analytics is valuable when it is not possible to examine each member of the population.

Major Categories of Statistics – An Example

Study of the height of the population



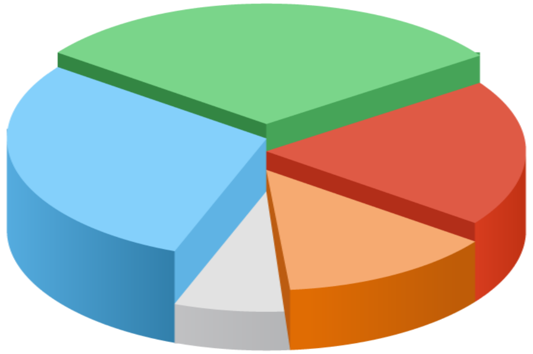
Inferential Method

Categorize height as “Tall,” “Medium,” and “Short”
Take a sample to study from the population.

Descriptive Method

Record the height of each and every person.
Provide the tallest, shortest, and average height of the population.

Statistical Analysis Considerations



Purpose
Clear and well-defined

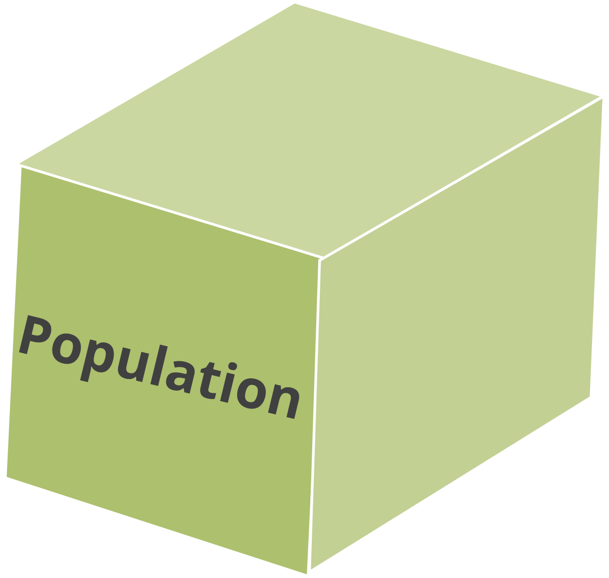
Document Questions
Prepare a questionnaire in advance

Define Population of Interest
Select population based on the purpose of analysis

Determine Sample
Based on the purpose of study

Population and Sample

A population consists of various samples. The samples together represent the population.



A sample is:

- The part/piece drawn from the population
- The subset of the population
- A random selection to represent the characteristics of the population
- Representative analysis of the entire population

Statistics and Parameters

“Statistics” are quantitative values calculated from the sample.

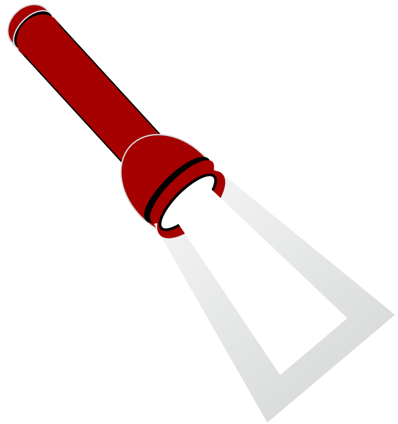
“Parameters” are the characteristics of the population.

Sample → Xo, X1,X2……….Xn

	Population Parameters	Sample Statistics	Formula
Mean	µ	$\overline{x}$	$\bar{x} = \frac{1}{n}\sum x_i$
Variance	σ^2	S^2	$S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$
Standard Deviation	σ	S	$S = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$

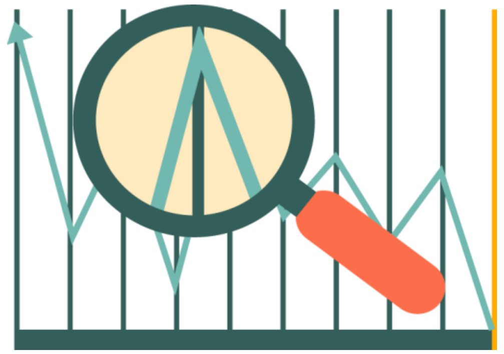
Terms Used to Describe Data

Typical terms used in data analysis are:



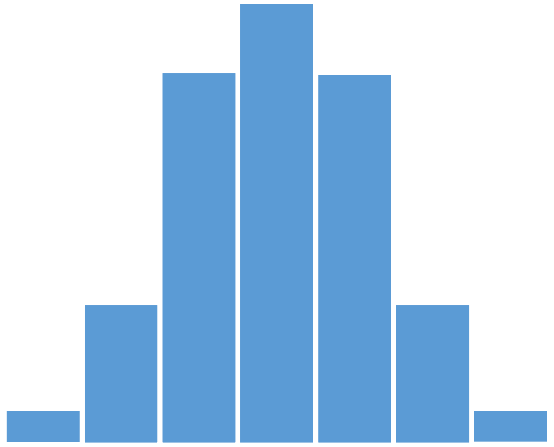
SEARCH

“Search” is used to find unusual data. Data that does not match the parameters.



INSPECT

“Inspect” refers to studying the shape and spread of data.



CHARACTERIZE

“Characterize” refers to determining the central tendency of the data.

CONCLUSION

“Conclusion” refers to preliminary or high-level conclusions about the data.

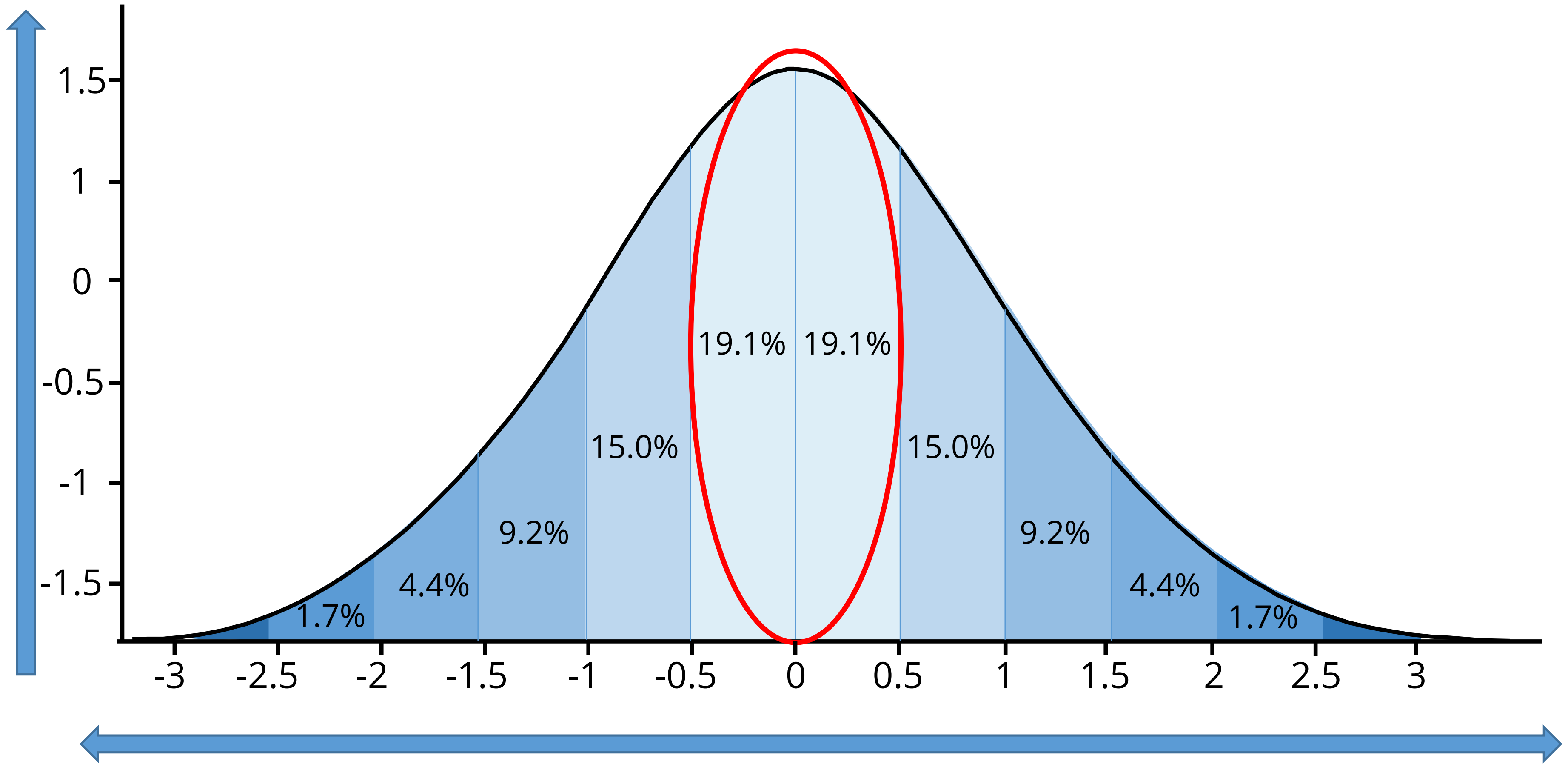
Statistical Analysis Process

There are four steps in the statistical analysis process.

Step 1: Find the population of interest that suits the purpose of statistical analysis.
Step 2: Draw a random sample that represents the population.
Step 3: Compute sample statistics to describe the spread and shape of the dataset.
Step 4: Make inferences using the sample and calculations. Apply it back to the population.

Data Distribution

The collection of data values arranged in a sequence according to their relative frequency and occurrences.



Range of the data refers to minimum and maximum values.

Frequency indicates the number of occurrences of a data value.

Central tendency indicates data accumulation toward the middle of the distribution or toward the end.

Measures of Central Tendency

The measures of central tendency are Mean, Median, and Mode.

Mean is the average.

Determine the mean score of these Math scores.

1. 80
2. 70
3. 75
4. 90
5. 80
6. 78
7. 55
8. 60
9. 80

Σ [80+70+75+90+80+78+55+60+80]/9

Mean = 74.22

Median is the 50th percentile.

55 60 70 75 78 80 80 80 90

Median = 78

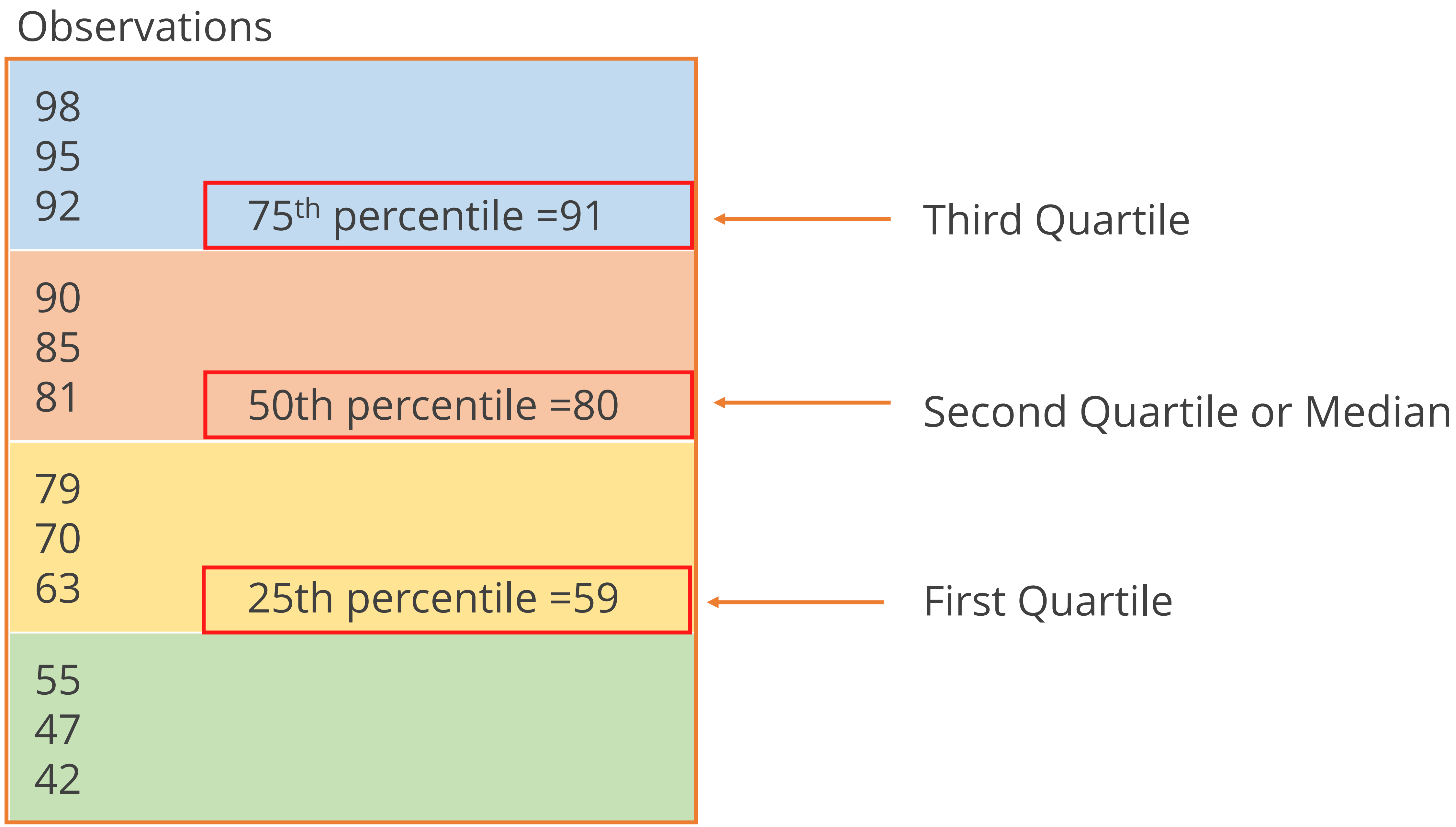
Mode is the most frequent value.

55 60 70 75 78 80 80 80 90

Mode = 80

Percentiles in Data Distribution

A percentile (or a centile) indicates the value below which a given percentage of observations fall.



Observations

98

95

92

75th percentile =91

Third Quartile

90

85

81

50th percentile =80

Second Quartile or Median

79

70

63

25th percentile =59

First Quartile

55

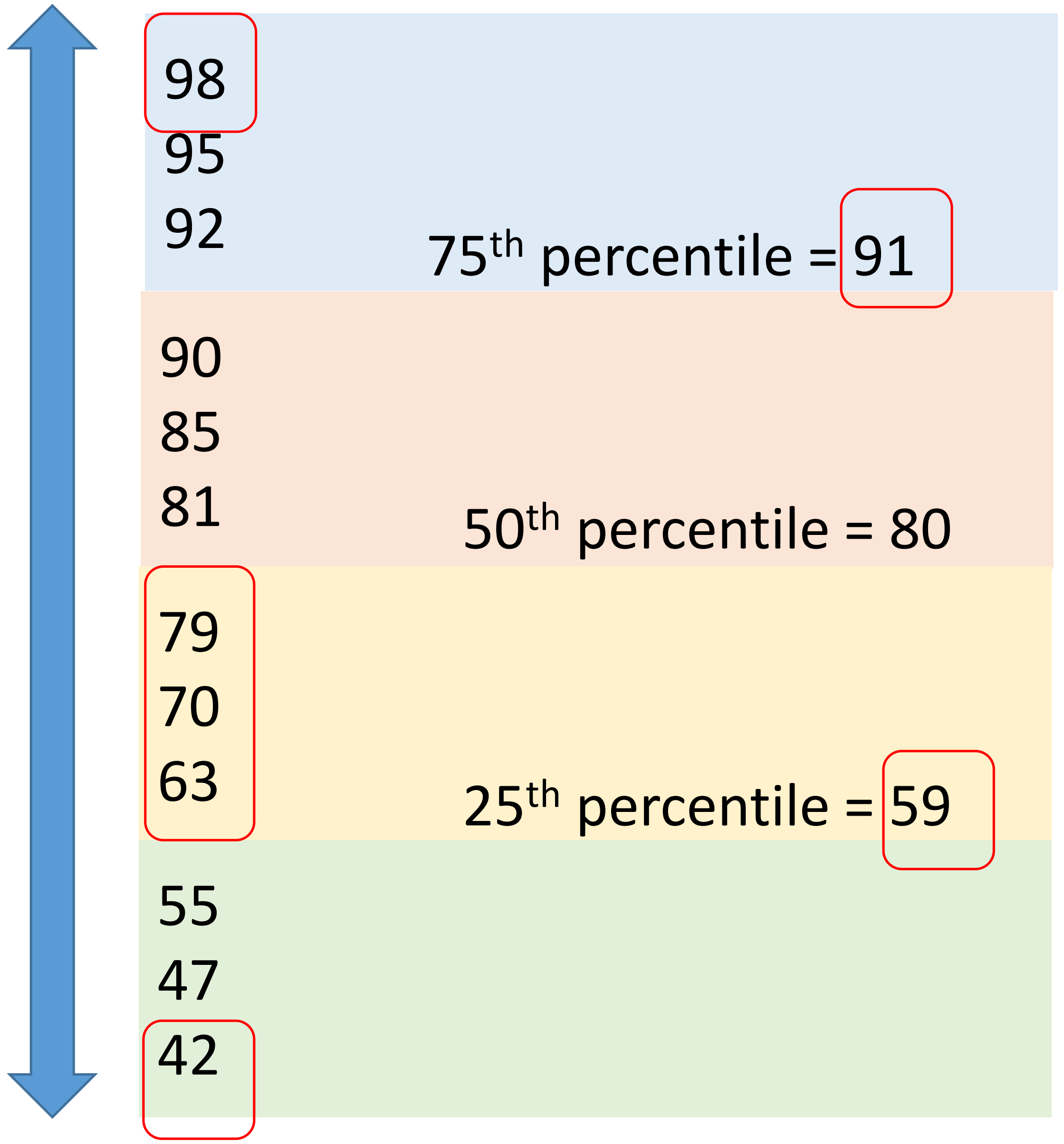
47

42

Dispersion

Dispersion denotes how stretched or squeezed a distribution is.

Observations



98
95
92
75th percentile = 91

90
85
81
50th percentile = 80

79
70
63
25th percentile = 59

55
47
42

Range: The difference between the maximum and minimum values

Inter-quartile Range: Difference between the 25th and 75th percentiles

Variance: Data values around the Mean. (74.75)

Standard Deviation: Square root of the variance measured in small units

Knowledge Check

KNOWLEDGE CHECK

What does frequency indicate?

a. Range of the values present in the dataset

b. Number of occurrences of a particular value in a dataset

c. How spread out the data is

d. Size of the sample drawn from a population

KNOWLEDGE CHECK

What does frequency indicate?

a. Range of the values present in the dataset

b. Number of occurrences of a particular value in a dataset

c. How spread out the data is

d. Size of the sample drawn from a population

The correct answer is **b** .

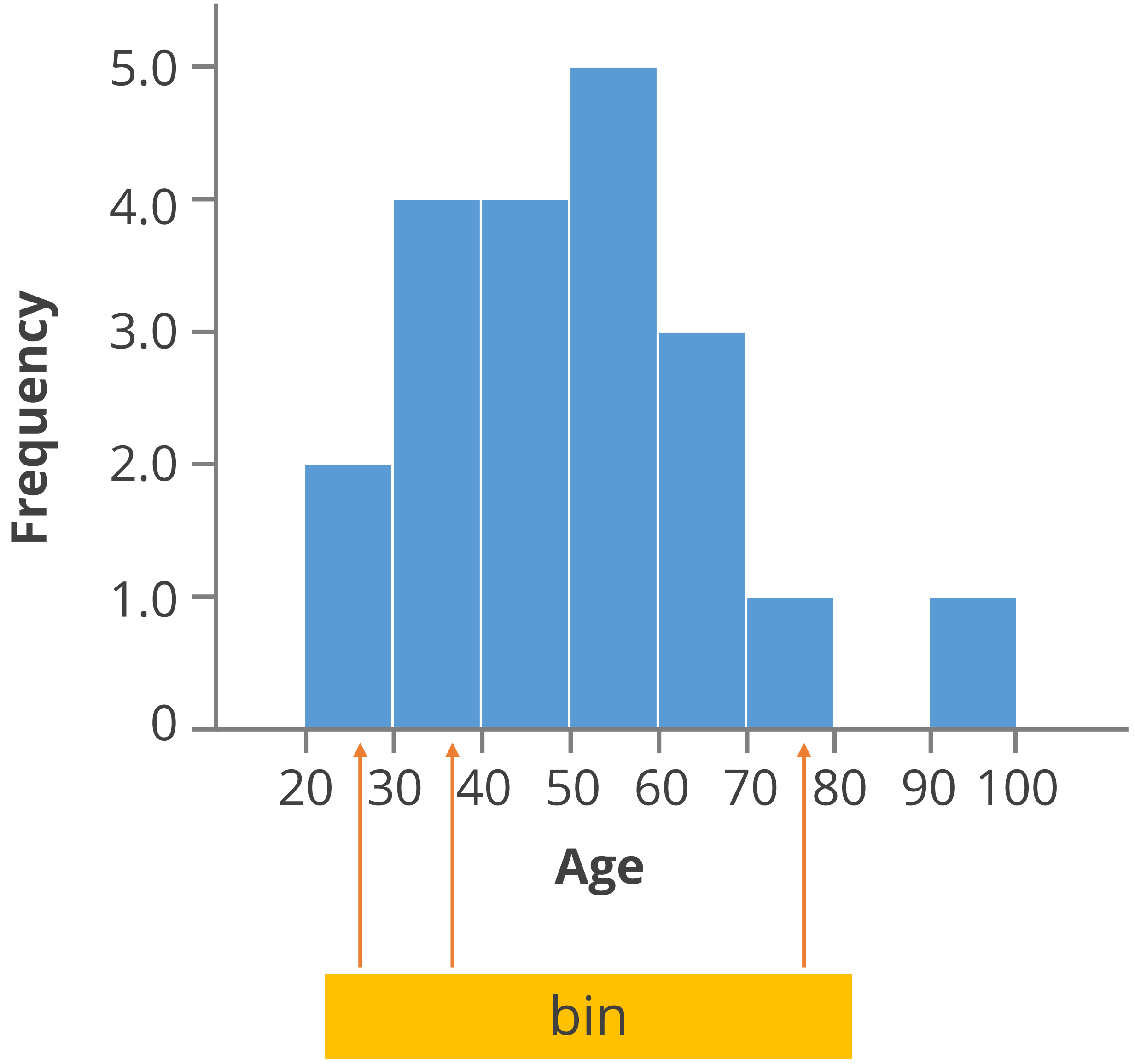
Explanation: Frequency indicates the number of occurrences of a particular value in a dataset.

Histogram

Graphical representation of data distribution

Features of a Histogram:

- It was first introduced by Karl Pearson.
- To construct a Histogram, “bin” the range of values.
- Bins are consecutive, non-overlapping intervals of a variable.
- Bins are of equal size.
- The bars represent the bins.
- The height of the bar represents the frequency of the values in the bin.
- It helps assess the probability distribution of a variable.

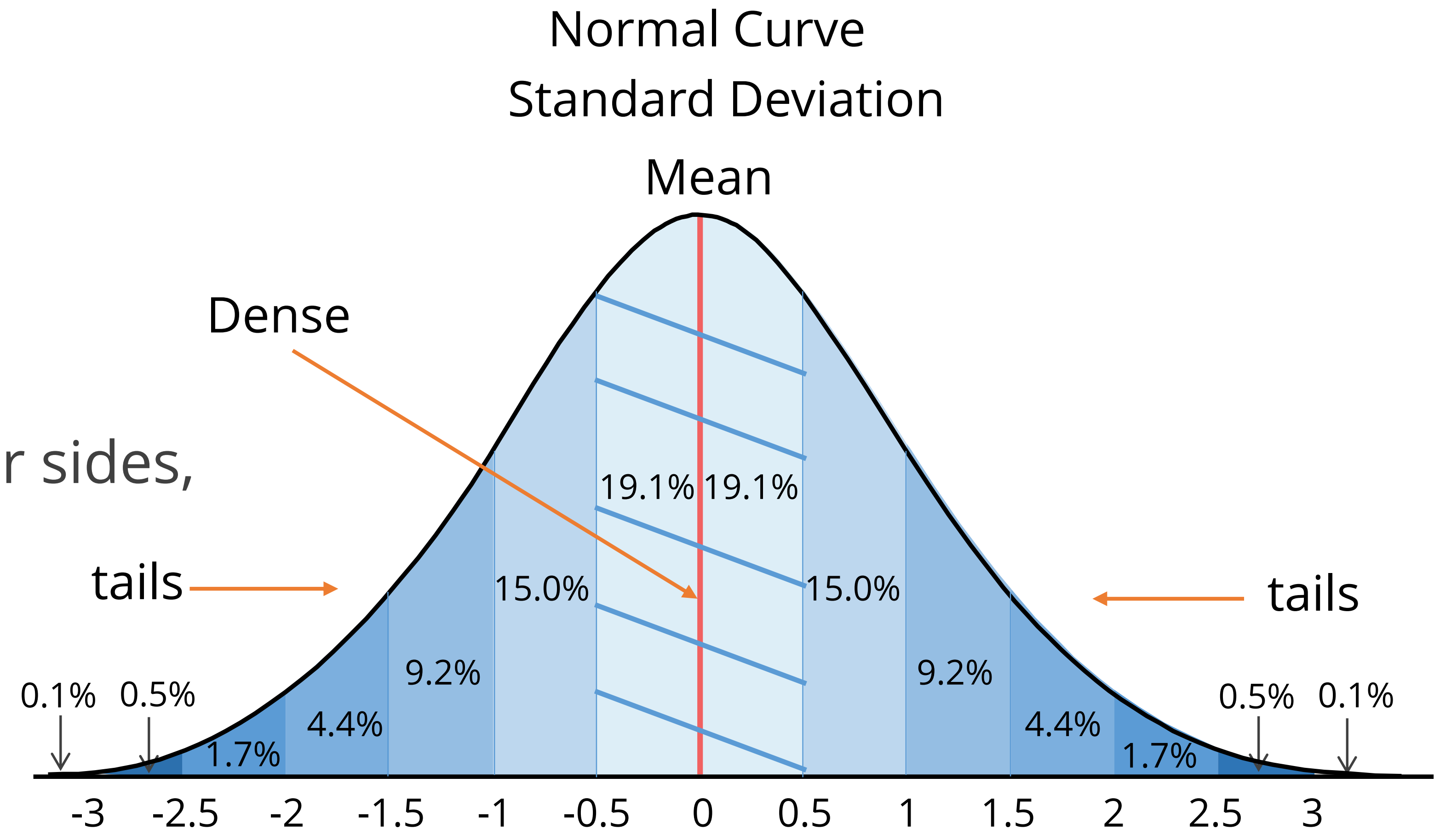


Bell Curve – Normal Distribution

The bell curve is characterized by its bell shape and two parameters, mean and standard deviation.

Bell curve is:

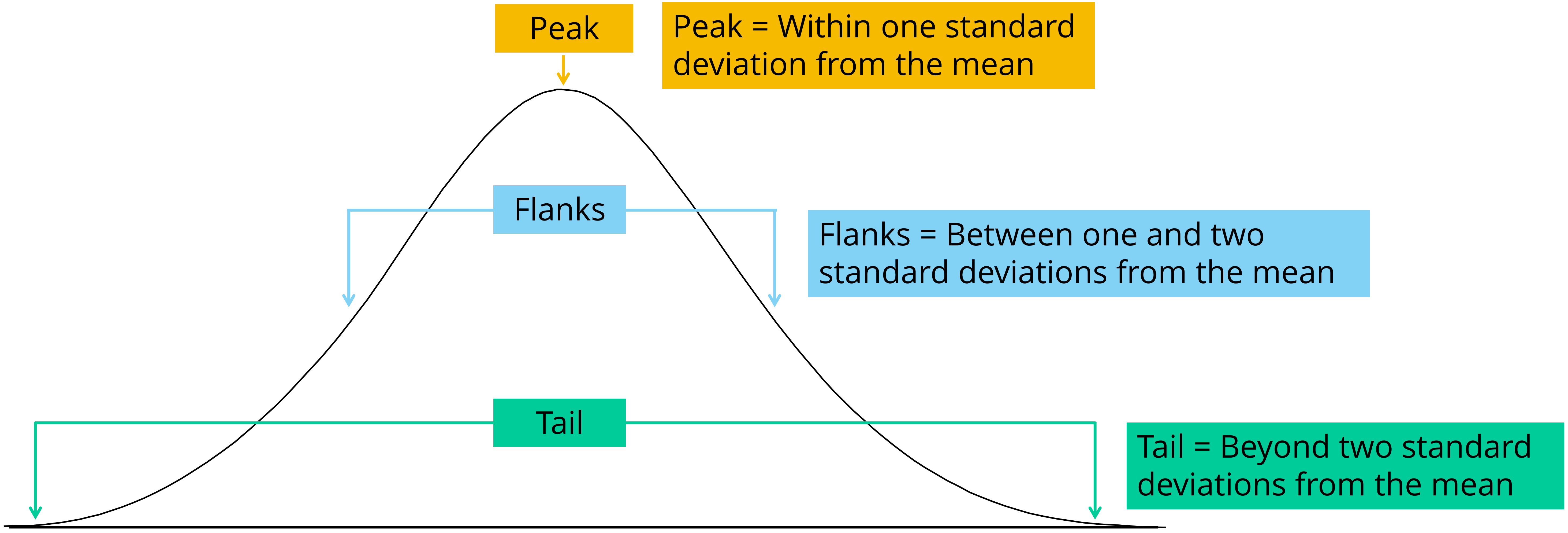
- Symmetric around the mean,
- Symmetric on both sides of the center,
- Having equal mean, median, and mode values,
- Denser in the center and less dense in the tails or sides,
- Defined by mean and standard deviation, and
- Known as the “Gaussian” curve.



The Bell curve is fully characterized by the mean (μ) and standard deviation (σ).

The Bell Curve

The Bell curve is divided into three parts to understand data distribution better.



Peak = Within one standard deviation from the mean

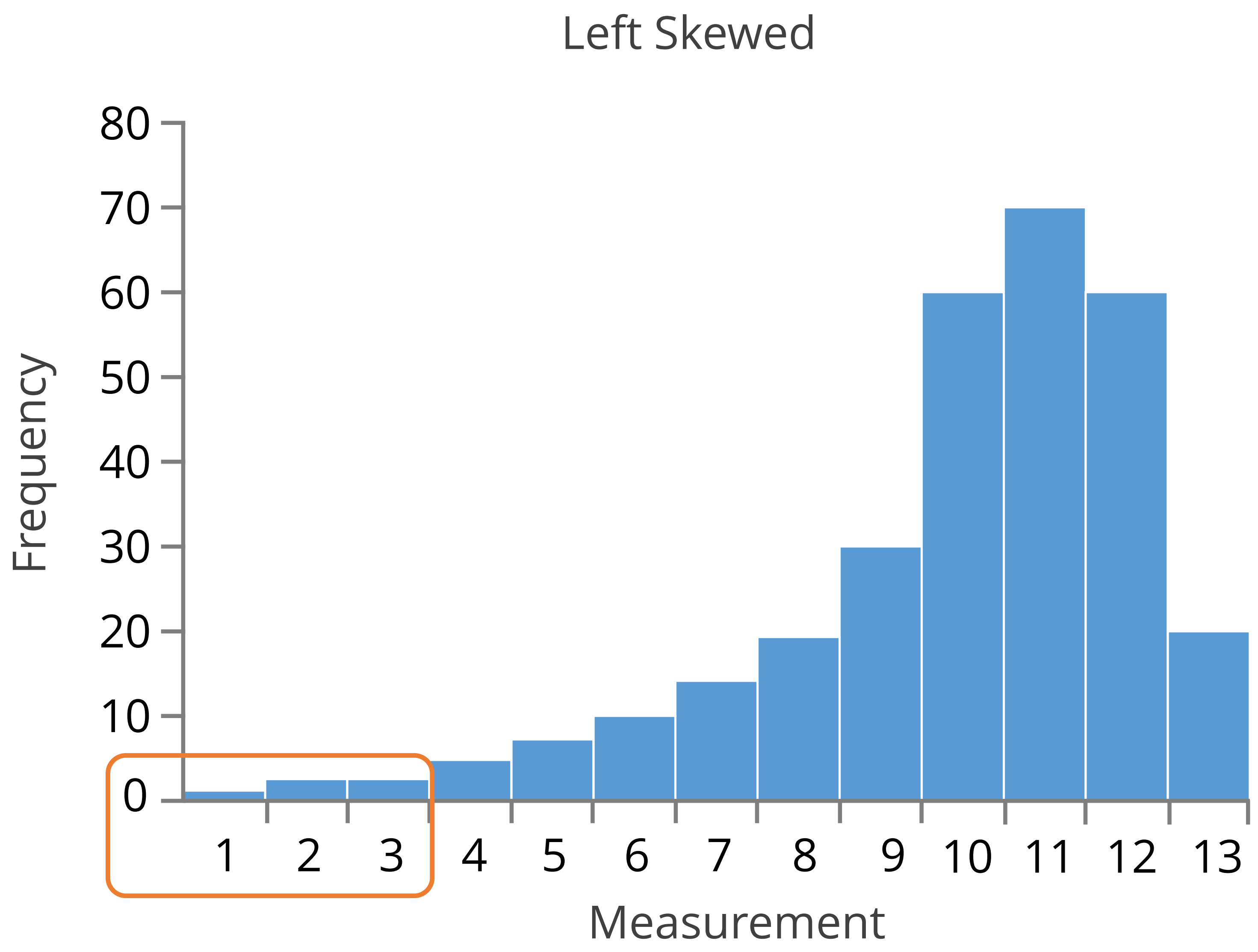
Flanks = Between one and two standard deviations from the mean

Tail = Beyond two standard deviations from the mean

Bell Curve – Left Skewed

Skewed data distribution indicates the tendency of the data distribution to be more spread out on one side.

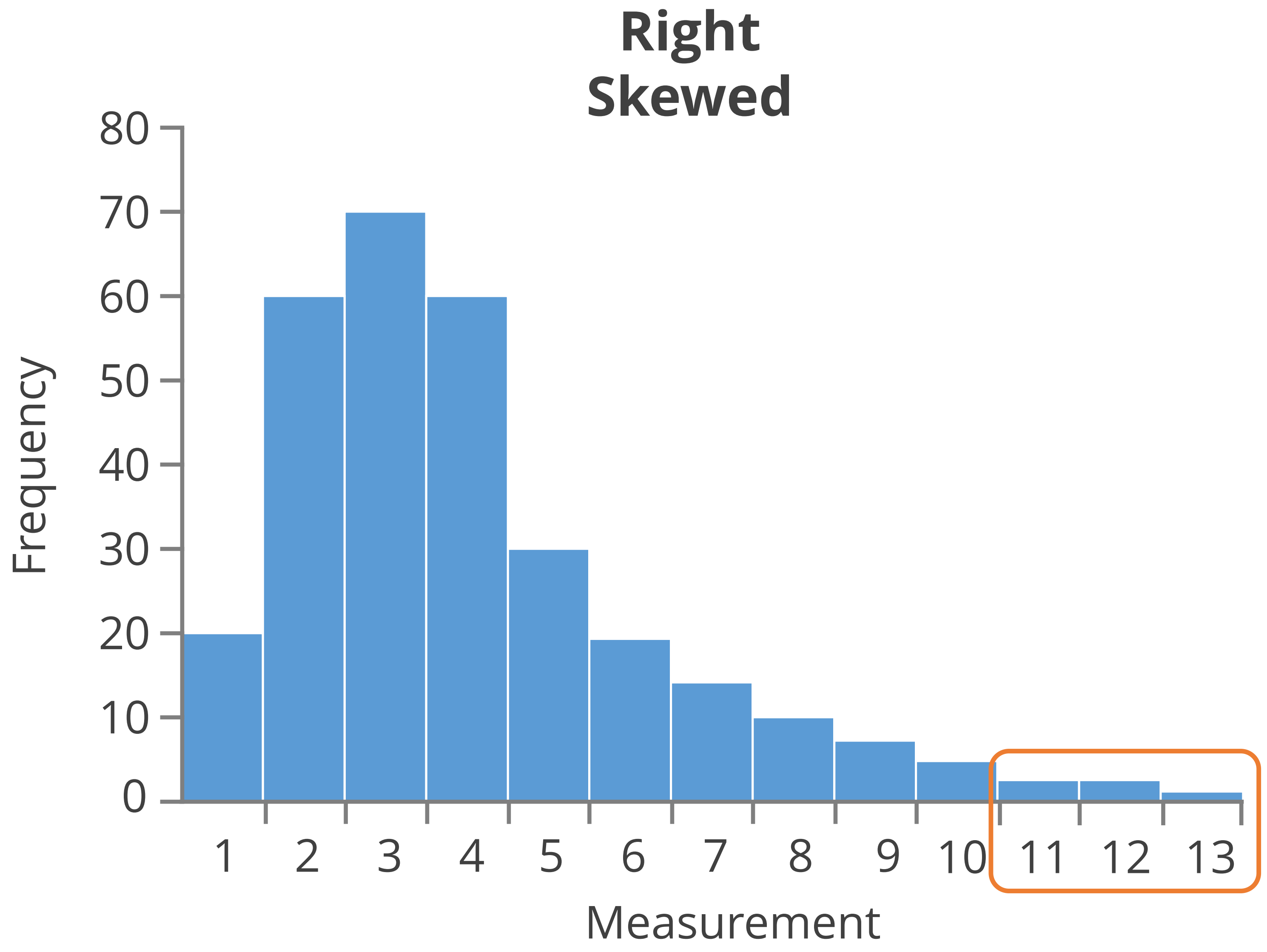
- The data is left skewed.
- Mean < Median
- The distribution is negatively skewed.
- Left tail contains large distributions.



Bell Curve – Right Skewed

Skewed data distribution indicates the tendency of the data distribution to be more spread out on one side.

- The data is right skewed.
- The distribution is positively skewed.
- Mean > Median
- Right tail contains large distributions.



Kurtosis

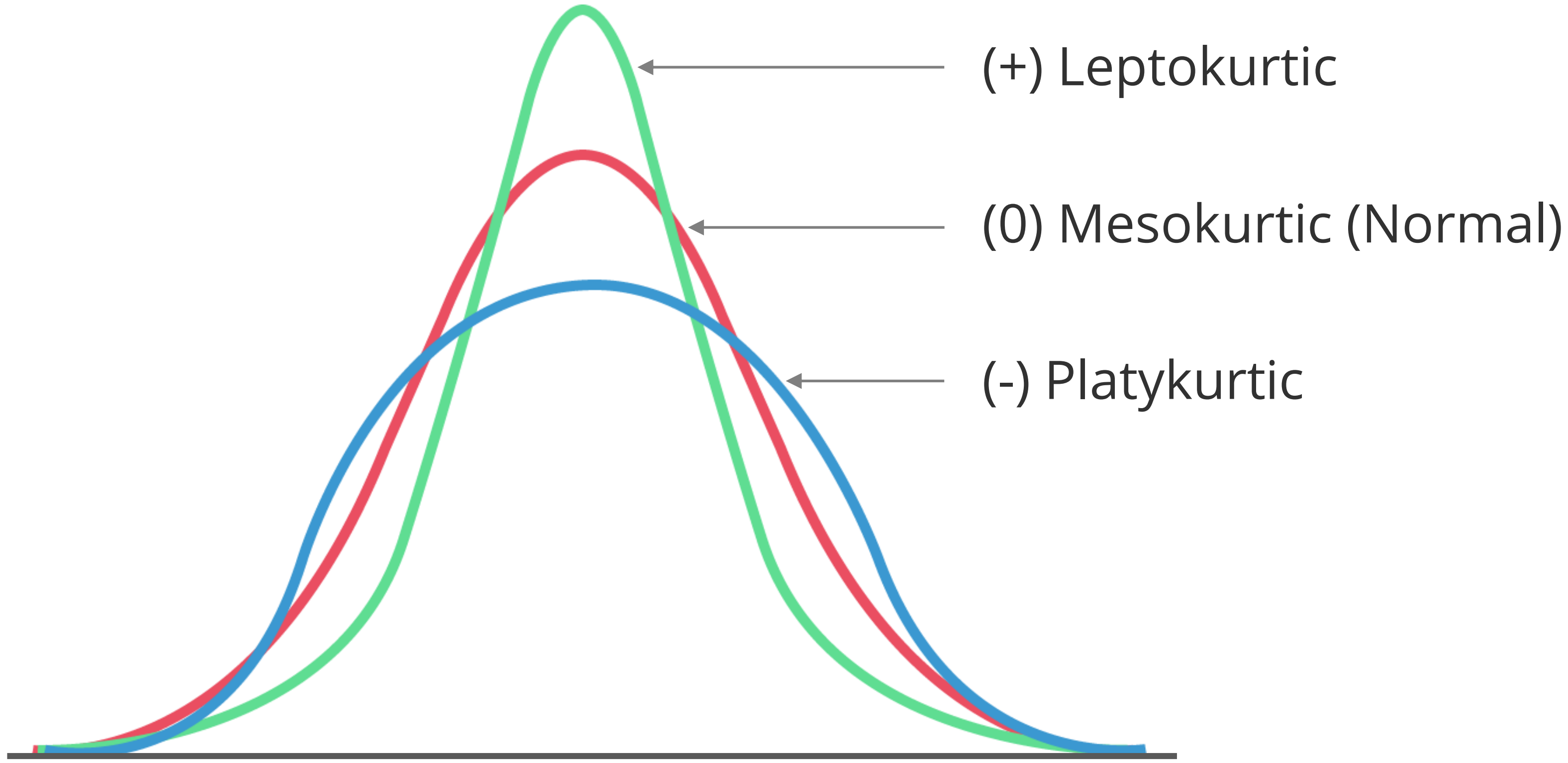
Kurtosis describes the shape of a probability distribution.

Kurtosis measures the tendency of the data toward the center or toward the tail.

Platykurtic is negative kurtosis.

Mesokurtic represents a normal distribution curve.

Leptokurtic is positive kurtosis.



Knowledge Check

KNOWLEDGE CHECK

Which of the following is true for a normal distribution?

a. Mean and median are equal

b. Mean and mode are equal

c. Mean, median, and mode are equal

d. Mode and median are equal

KNOWLEDGE CHECK

Which of the following is true for a normal distribution?

a. Mean and median are equal

b. Mean and mode are equal

c. Mean, median and mode are equal

d. Mode and median are equal

The correct answer is · c

Explanation: for Bell curve mean, median, and mode are equal.

Hypothesis Testing

Hypothesis testing is an inferential statistical technique that determines if a certain condition is true for the population.

Alternative Hypothesis (H1)	Null Hypothesis (H0)
A statement that has to be concluded as true.	A statement of “no effect” or “no difference”.
It’s a research hypothesis.	It’s the logical opposite of the alternative hypothesis.
It needs significant evidence to support the initial hypothesis.	It indicates that the alternative hypothesis is incorrect.
If the alternative hypothesis garners strong evidence, reject the null hypothesis.	Weak evidence of alternative hypothesis indicates that the null hypothesis has to be accepted.

Hypothesis Testing – Error Types

Representation of decision parameters using null hypothesis

Type I Error (α)

- Rejects the null hypothesis when it is true
- The probability of making Type I error is represented by α

Type II Error (β)

- Fails to Reject the null hypothesis when it false
- The probability of making Type II error is represented by β

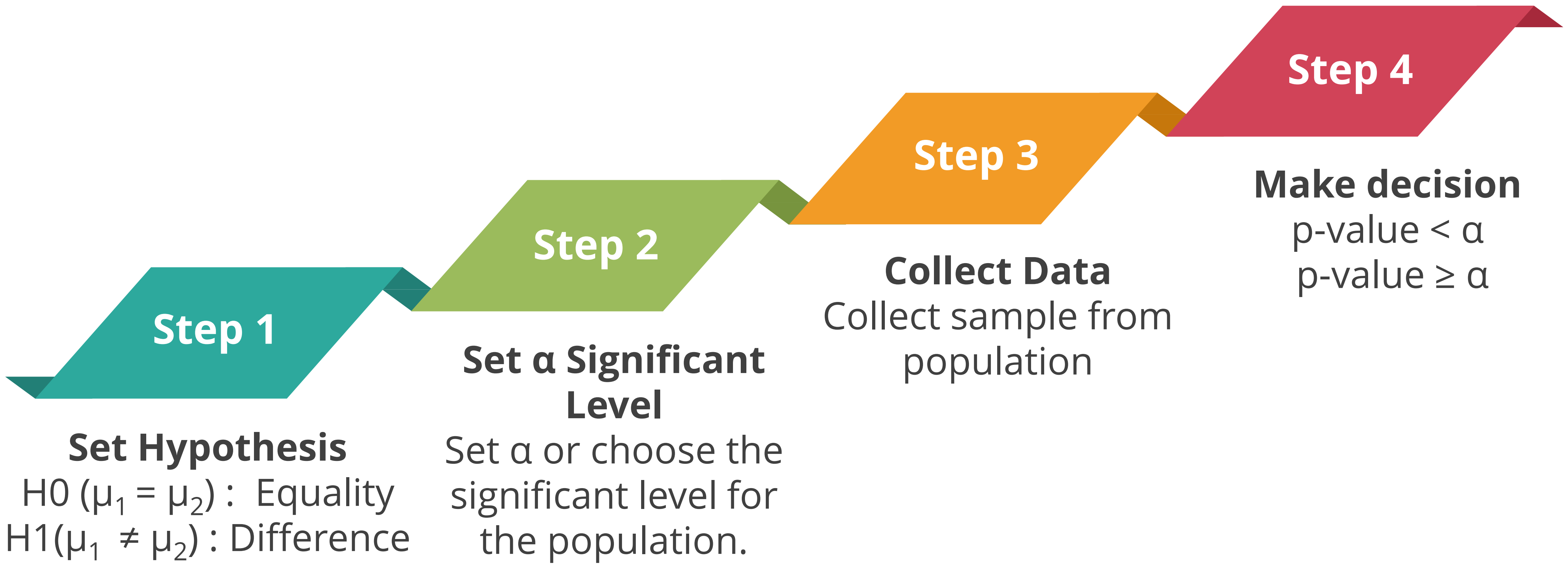
p-value

- The probability of observing extreme values
- Calculated from collected data

Decision	Ho is True	Ho is False
Fail to Reject Null	Correct	Type II Error
Reject Null	Type I Error	Correct

Hypothesis Testing - Process

There are four steps to the hypothesis testing process.



Step 1

Set Hypothesis
H0 ($\mu_1 = \mu_2$) : Equality
H1($\mu_1 \neq \mu_2$) : Difference

Step 2

Set α Significant Level
Set α or choose the significant level for the population.

Step 3

Collect Data
Collect sample from population

Step 4

Make decision
p-value < α
p-value ≥ α

Reject the null hypothesis if p-value < α
Fail to reject the null hypothesis if p-value ≥ α

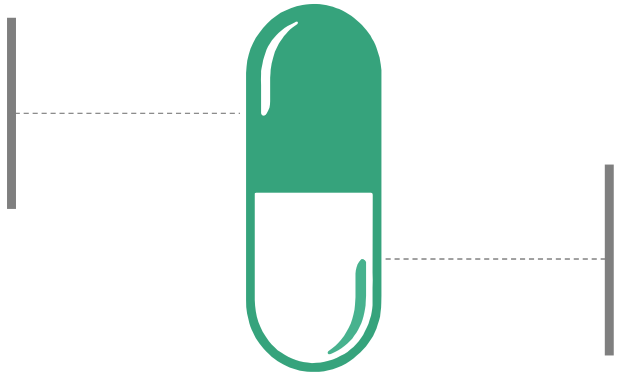
Perform Hypothesis Testing

An example of clinical trials data analysis.

Company A

Company B

Null Hypothesis:
Both medicines are
equally effective.



Alternative Hypothesis:
Both medicines are
NOT equally effective.

Data for Hypothesis Testing

There are three types of data on which you can perform hypothesis testing.

Continuous Data

Evaluate the mean, median, standard deviation, or variance.

Binomial Data

Evaluate the percentage, general classification of data.

Poisson Data

Evaluate rate of occurrence or frequency.

Types of Variables

There are three types of variables in categorical data.

Nominal Variables

- Values with no logical ordering
- Variables are independent of each other
- Sequence does not matter

Ordinal Variables

- Values are in logical order
- Relative distance between two data values is not clear

Association

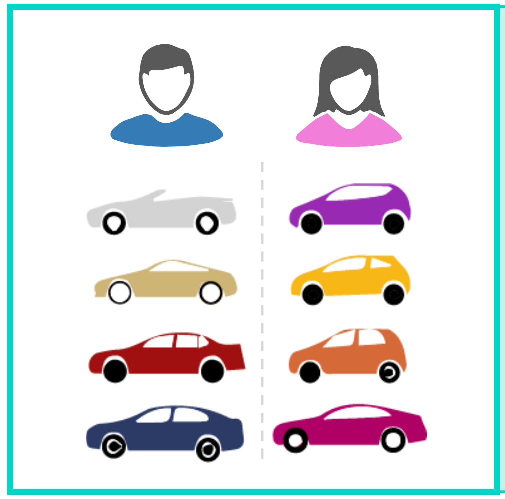
Two variables are associated or independent of each other.

	🚆		🚆	
🌧	85%	15%	68%	32%
☀	85%	15%	95%	55%

Chi-Square Test

It is a hypothesis test that compares the observed distribution of your data to an expected distribution of data.

Test of Association:
To determine whether one variable is associated with a different variable. For example, determine whether the sales for different cellphones depends on the city or country where they are sold.



Test of Independence:
To determine whether the observed value of one variable depends on the observed value of a different variable. For example, determine whether the color of the car that a person chooses is independent of the person's gender.

Test is usually applied when there are two categorical variables from a single population.

Chi Square Test - Example

An example of Chi-Square test.

Null Hypothesis

- There is no association between gender and purchase.
- The probability of purchase does not change for 500 dollars or more whether female or male.

Alternative Hypothesis

- There is association between gender and purchase.
- The probability of purchase over 500 dollars is different for female and male.

		<$500	>$500
👩	fo	.55	.45
👨	fo	.75	.25

Types of Frequencies

Expected and observed frequencies are the two types of frequencies.

Expected Frequencies (fe)

The cell frequencies that are expected in a bivariate table if the two tables are statistically independent.

Observed Frequencies (fo)

- There is association between gender and purchase.
- The probability of purchase over 500 dollars is different for female and male.

		Purchases	
		<$500	>$500
👩	fo	.55	.45
👨	fo	.75	.25

No Association
Observed Frequency = Expected Frequency

Association
Observed Frequency ≠ Expected Frequency

Features of Frequencies

The formula for calculating expected and observed frequencies using Chi Square:

$$\sum \frac{(f_e - f_o)^2}{f_e}$$

Features of Expected and Observed frequencies:

- Requires no assumption of the underlying population
- Requires random sampling

Knowledge Check

KNOWLEDGE CHECK

In Chi-Square test, there is no association of variables if _____.

a. Observed Frequency ≠ Expected Frequency

b. Observed Frequency = Expected Frequency

c. Independent of observed frequencies

d. Independent of expected frequencies

KNOWLEDGE CHECK

In Chi-Square test, there is no association of variables if:

a. Observed Frequency ≠ Expected Frequency

b. Observed Frequency = Expected Frequency

c. Independent of observed frequencies

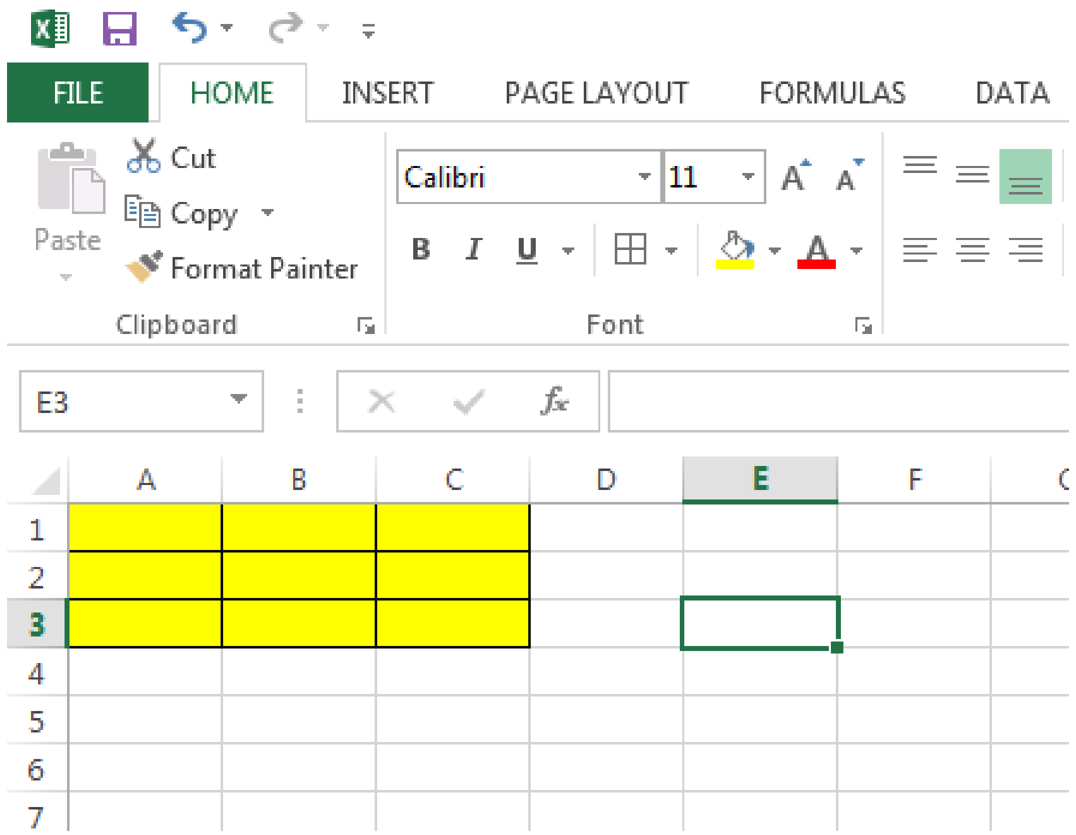
d. Independent of expected frequencies

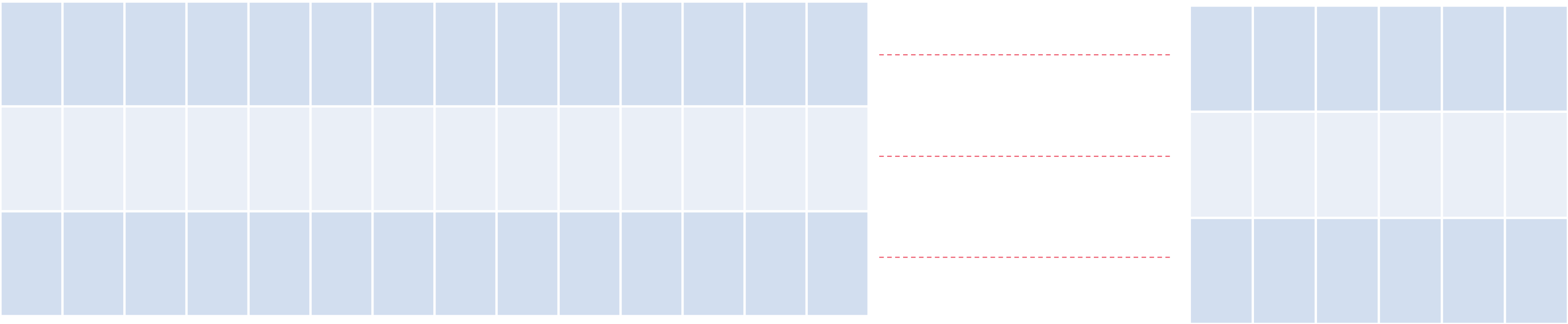
The correct answer is **b** .

Explanation: Observed Frequency = Expected Frequency indicates no association.

Correlation Matrix

A Correlation matrix is a square matrix that compares a large number of variables.





Correlation matrix – a square matrix

n × n Matrix

(very large number of rows and columns)

(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

3 × 3 matrix (simple square matrix)

Correlation coefficient measures the extent to which two variables tend to change together.

The coefficient describes both the strength and direction of the relationship.

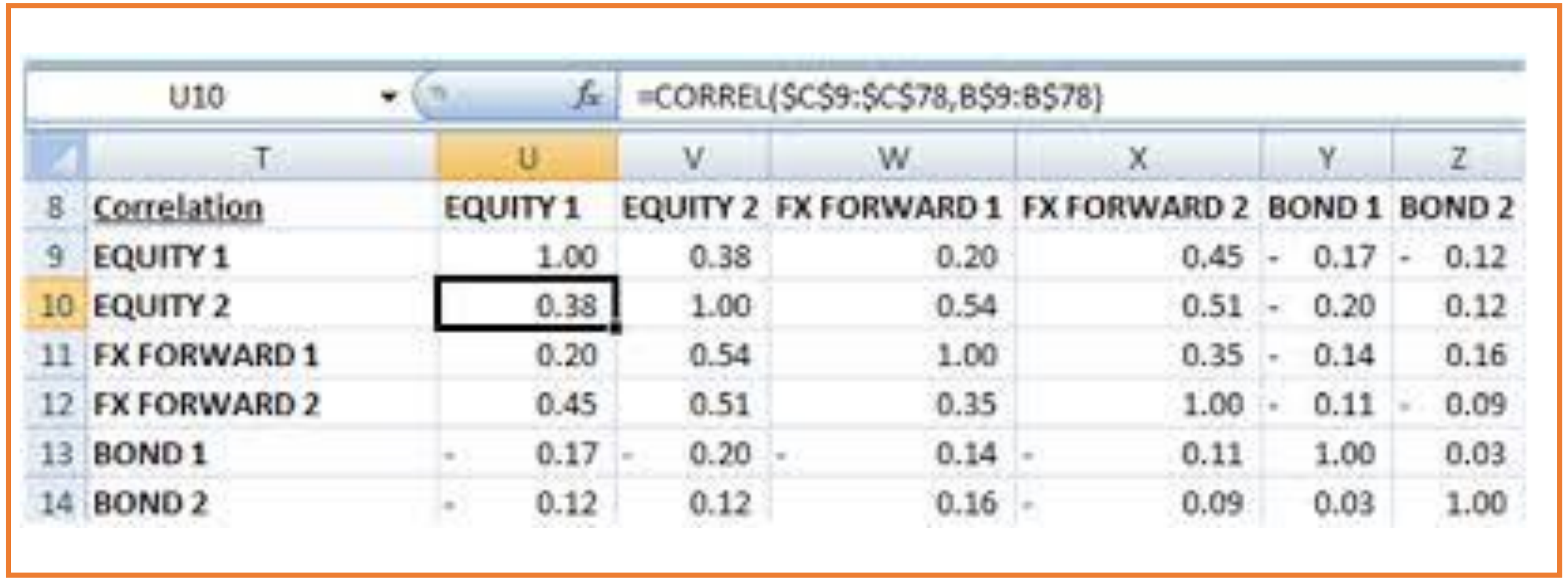
Correlation Matrix

A Correlation matrix is a square matrix that compares a large number of variables.

Pearson product moment correlation	It evaluates the linear relationship between two continuous variables. Linear relationship means that a change in one variable results in a proportional change in the other.
Spearman rank order correlation	It evaluates the monotonic relationship between two continuous or ordinal variables. • Monotonic relationship means that the variables tend to change together though not necessarily at a constant rate. • The correlation coefficient is based on the ranked values for each variable rather than the raw data.

Correlation Matrix - Example

An example of a correlation matrix calculated for a stock market.



U10 =CORREL($C$9:$C$78,B$9:B$78)

	T	U	V	W	X	Y	Z
8	Correlation	EQUITY 1	EQUITY 2	FX FORWARD 1	FX FORWARD 2	BOND 1	BOND 2
9	EQUITY 1	1.00	0.38	0.20	0.45	- 0.17	- 0.12
10	EQUITY 2	0.38	1.00	0.54	0.51	- 0.20	0.12
11	FX FORWARD 1	0.20	0.54	1.00	0.35	- 0.14	0.16
12	FX FORWARD 2	0.45	0.51	0.35	1.00	- 0.11	- 0.09
13	BOND 1	- 0.17	- 0.20	- 0.14	- 0.11	1.00	0.03
14	BOND 2	- 0.12	0.12	0.16	- 0.09	0.03	1.00

A correlation matrix that is calculated for the stock market will probably show the short-term, medium-term, and long-term relationship between data variables.

Inferential Statistics

Inferential statistics uses a random sample from the data to make inferences about the population.

Inferential statistics can be used only under the following conditions:

- A complete list of the members of the population is available.
- A random sample has been drawn from the population.
- Using a pre-established formula, you determine that the sample size is large enough.

Inferential statistics can be used even if the data does not meet the criteria.

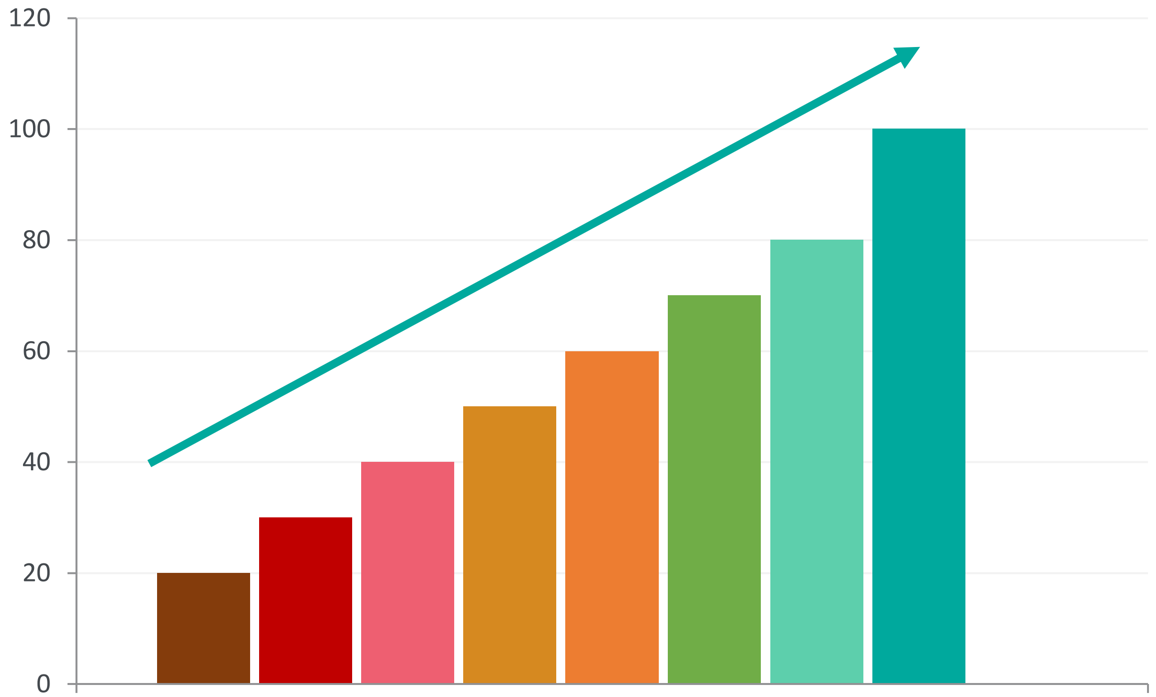
- It can help determine the strength of the relationships within the sample.
- If it is very difficult to obtain a population list and draw a random sample, do the best you can with what you have.

Applications of Inferential Statistics

Inferential Statistics has its uses in almost every field such as business, medicine, data science, and so on.

Inferential Statistics

- Is an effective tool for forecasting.
- Is used to predict future patterns.



Quiz

QUIZ 1

If a sample of five boxes weigh 90, 135, 160, 115, and 110 pounds, what will be the median weight of this sample?

a. 160

b. 115

c. 90

d. 135

QUIZ 1	If a sample of five boxes weigh 90, 135, 160, 115, and 110 pounds, what will be the median weight of this sample?

a. 160

b. 115

c. 90

d. 135

The correct answer is **b** .

Explanation: Arrange in a sequential order and the middle number will be the median. If the set of numbers is even then take the average or mean of the two numbers in the middle.

QUIZ

2

Identify the parameters that characterize a bell curve. *Select all that apply.*

a. Variance

b. Mean

c. Standard deviation

d. Range

QUIZ

2

Identify the parameters that characterize a bell curve. *Select all that apply.*

a. Variance

b. Mean

c. Standard deviation

d. Range

The correct answer is **b,c** .

Explanation: Bell Curve is completely characterized by mean and standard deviation.

QUIZ

3

Identify the accurate statement about the relationship between standard deviation and variance.

a. Standard deviation is the square root of variance.

b. Variance is the square root of standard deviation.

c. Both are inversely proportional.

d. Both are directly proportional.

QUIZ

3

Identify the accurate statement about the relationship between standard deviation and variance

a. Standard deviation is the square root of variance.

b. Variance is the square root of standard deviation.

c. Both are inversely proportional.

d. Both are directly proportional.

The correct answer is a.

Explanation: Standard deviation is the square root of variance.

QUIZ

4

Identify the hypothesis decision rules. *Select all that apply.*

a. Reject the null hypothesis if *p*-value < α

b. Is independent of *p*-value

c. Fail to reject the null hypothesis if *p*-value ≥ α

d. Is independent of α

QUIZ

4

Identify the hypothesis decision rules. ***Select all that apply.***

a. Reject the null hypothesis if *p*-value < α

b. Is independent of *p*-value

c. Fail to reject the null hypothesis if *p*-value ≥ α

d. Is independent of α

The correct answer is **a, c**.

Explanation: A hypothesis decision rule :

•Reject the null hypothesis if *p*-value < α

•Fail to reject the null hypothesis if *p*-value ≥ α

Key Takeaways

- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.
- Statistical analysis is more reliable when compared to non-statistical analysis.
- Descriptive and inferential are the two major categories of statistics.
- Mean, median, and mode are measures of central tendency, while variance and standard deviation measure the spread of data.
- The spread of distribution is called dispersion and is graphically represented by a histogram and a bell curve.
- Hypothesis testing is an inferential statistical technique that is useful for forecasting future patterns.
- Chi-Square test is a hypothesis test that compares observed distribution to an expected distribution.
- The correlation coefficient or covariance is measured with the help of correlation matrix.

This concludes “Statistical Analysis and Business Applications”

The next lesson is “Data Analytics Overview”