

Introduction to Statistics

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



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 quantitative data

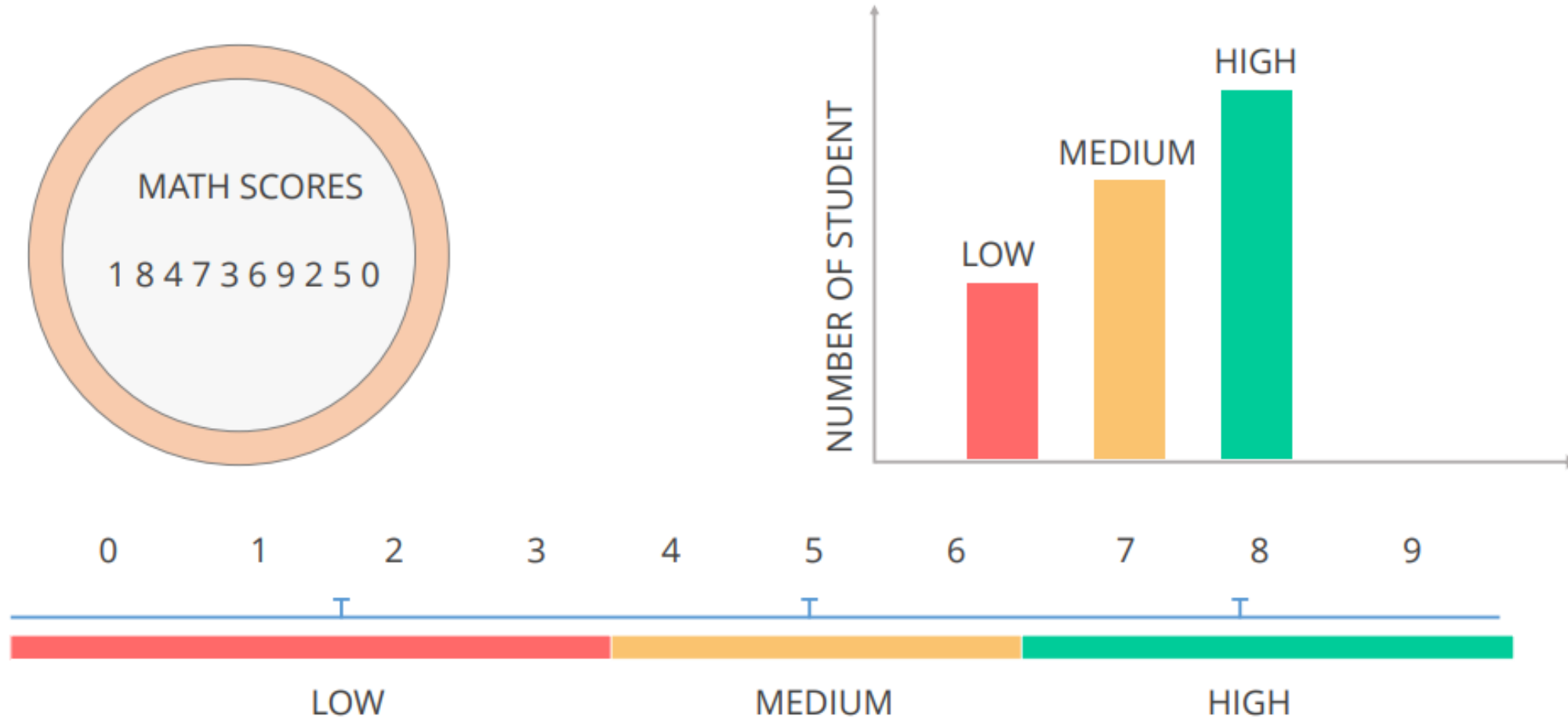


Although both forms of analysis provide results, quantitative analysis provides more insight and a clearer picture. This is why statistical analysis is important for businesses.

Major Categories of Statistics

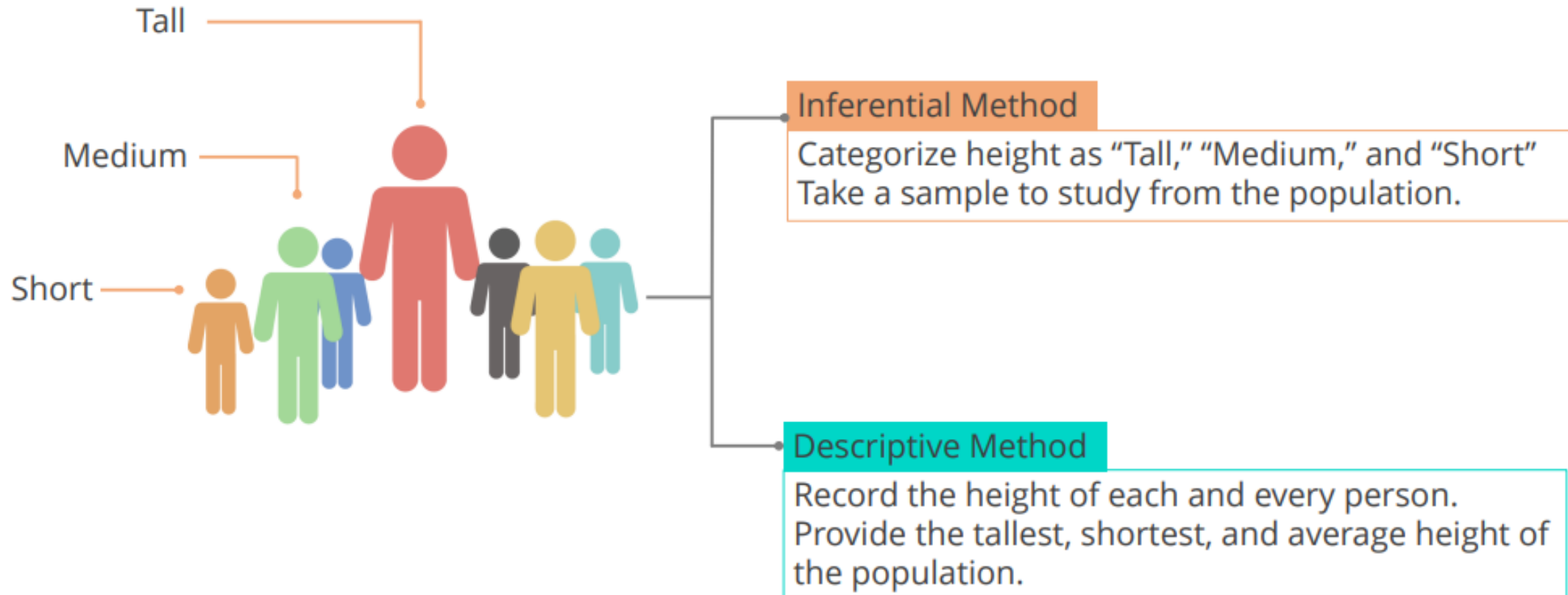
There are two major categories of statistics: Descriptive analytics and inferential analytics

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



Major Categories of Statistics – An Example

Study of the height of the population



Major Categories of Statistics

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



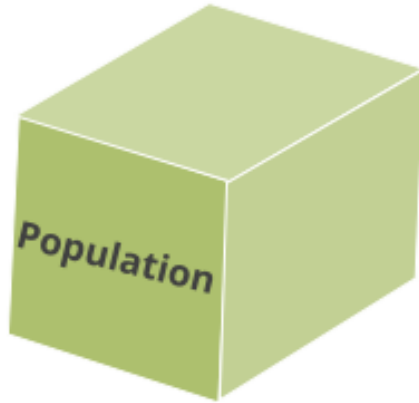
- Random sample is drawn from the population
- Used to describe and make inferences about the population



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

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 $\rightarrow X_0, X_1, X_2, \dots, X_n$



	Population Parameters	Sample Statistics	Formula
Mean	μ	\bar{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:



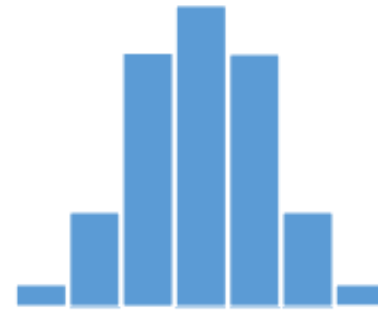
SEARCH

“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.

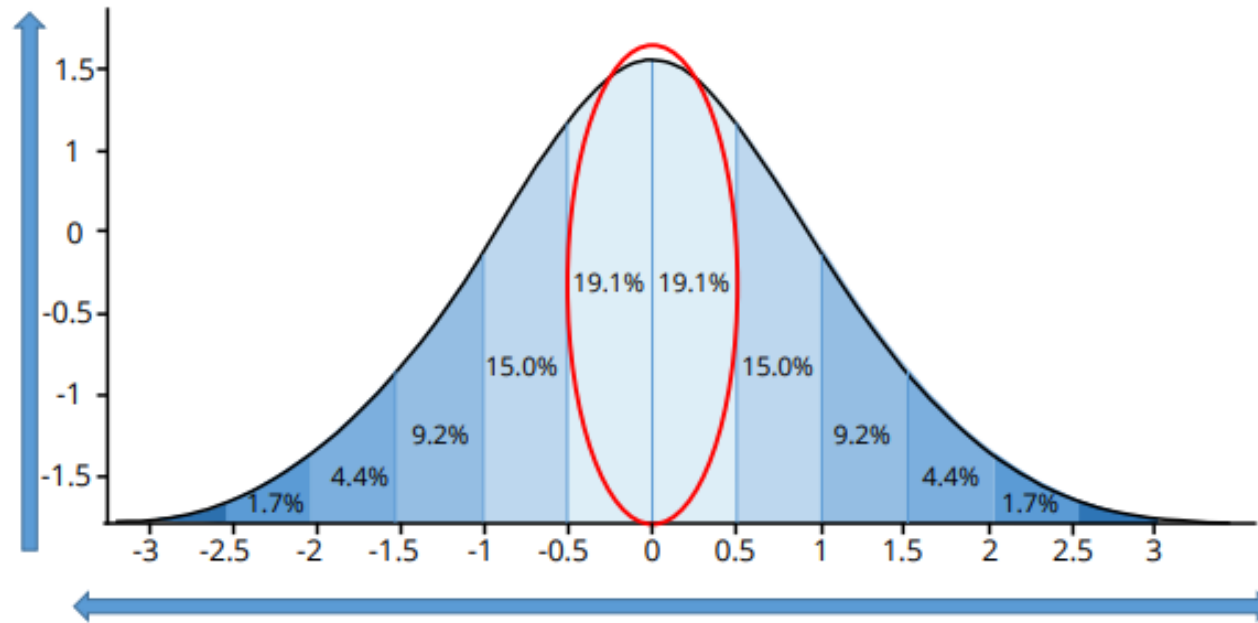
Step 3: Compute sample statistics to describe the spread and shape of the dataset.

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


$$\Sigma [80+70+75+90+80+78+55+60+80]/9$$

Mean = 74.22

Median is the 50th percentile.

55 60 70 75 78 80 80 80 90

Median = 78

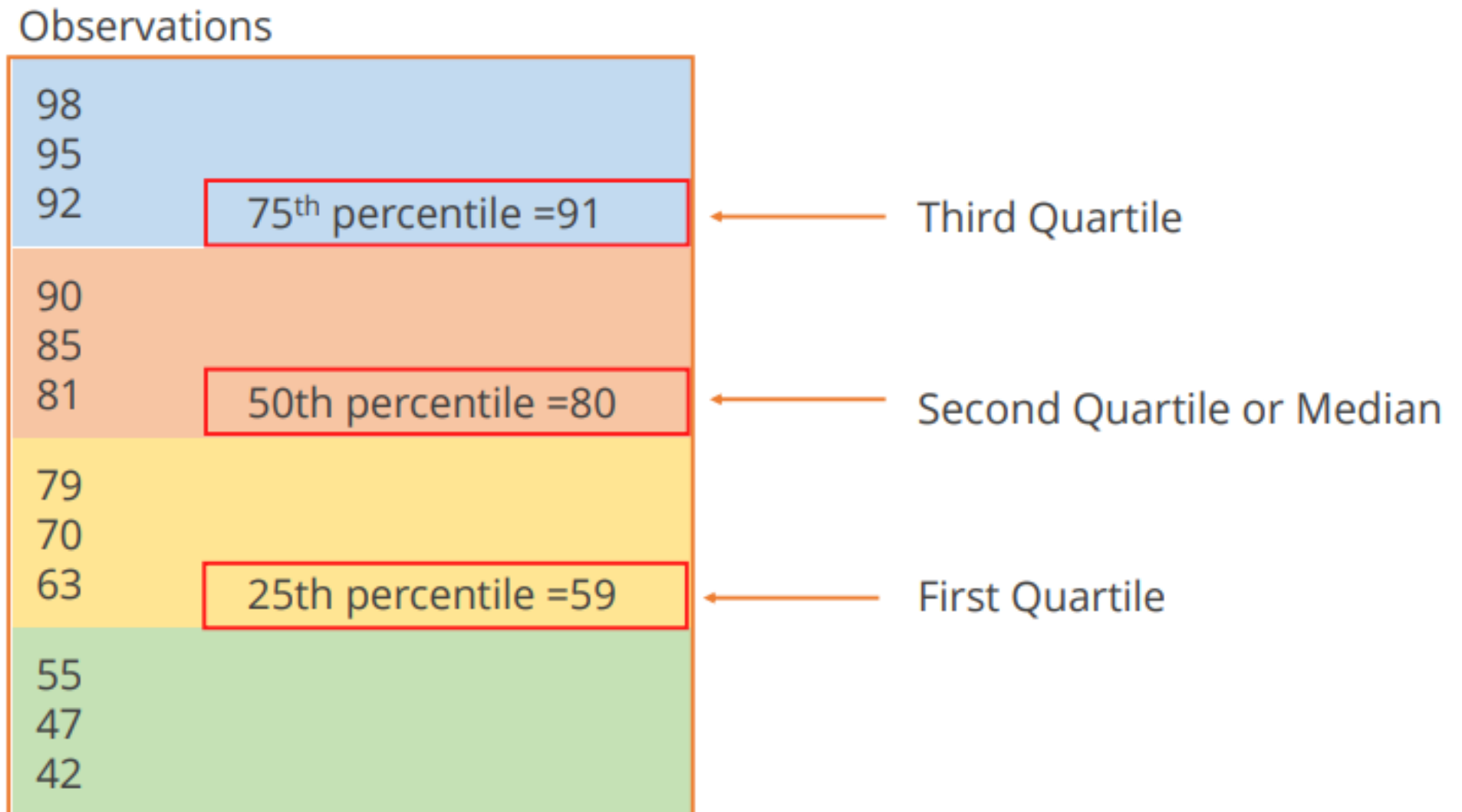
Mode is the most frequent value.

55 60 70 75 78 80 80 80 90

Mode = 80

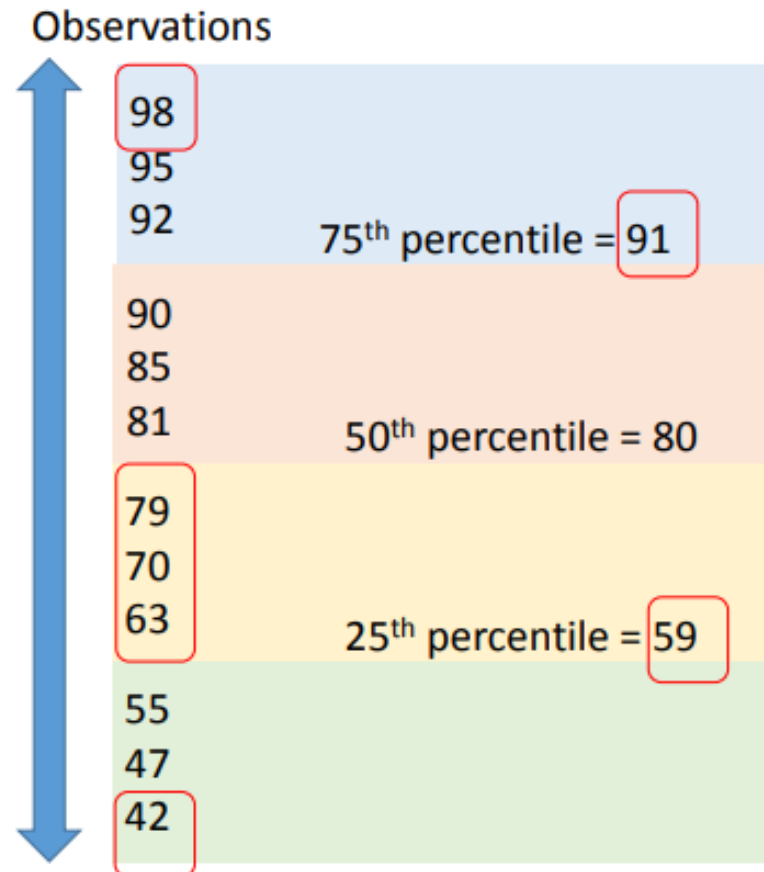
Percentiles in Data Distribution

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



Dispersion

Dispersion denotes how stretched or squeezed a distribution is.



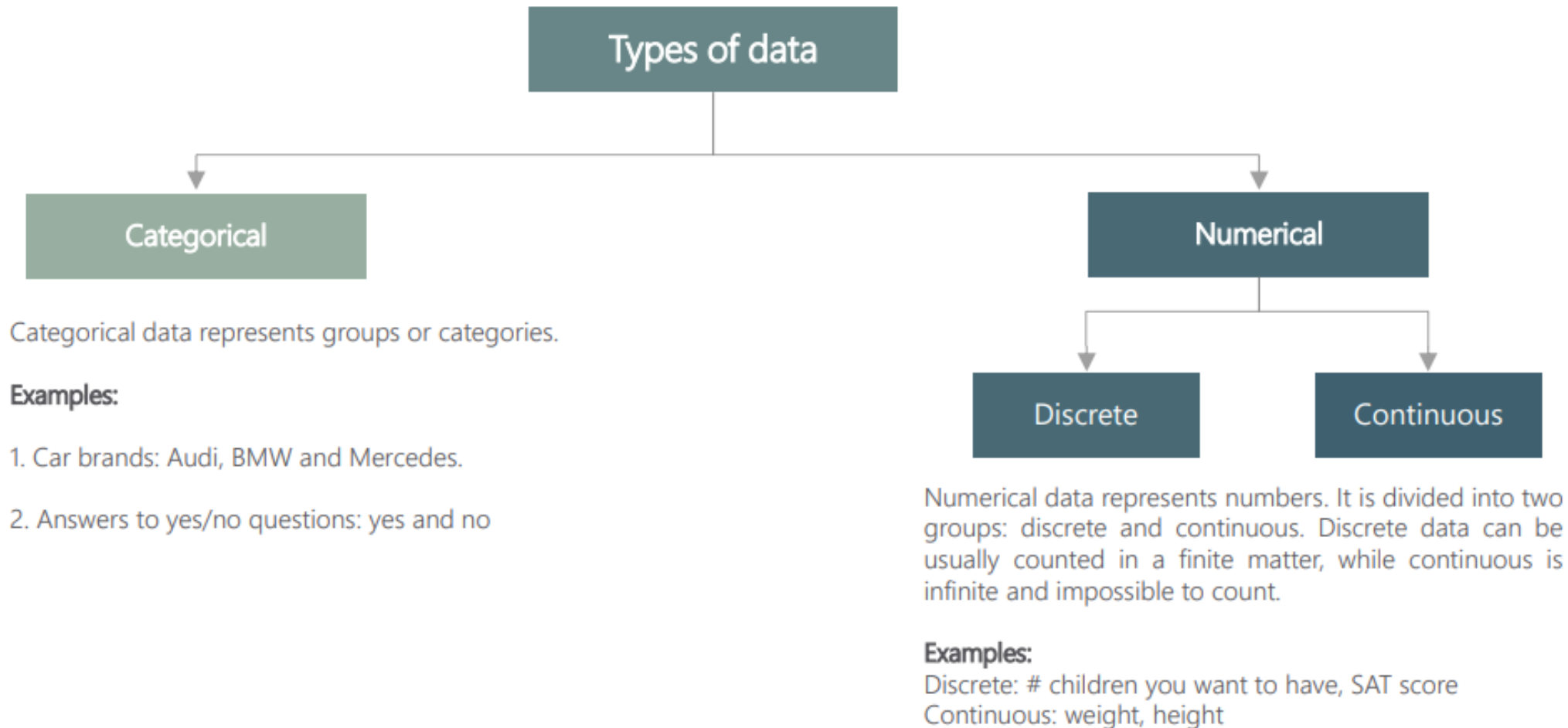
Range: The difference between the maximum and minimum values

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

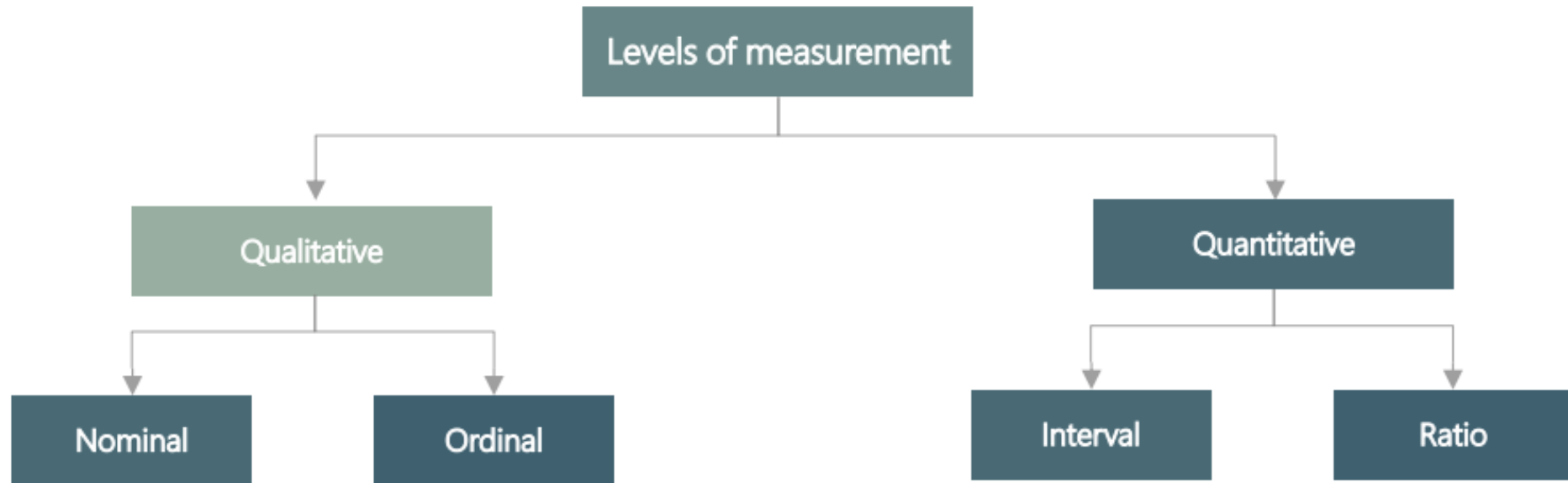
Variance: Data values around the Mean. (74.75)

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

Types of data



Levels of measurement



There are two qualitative levels: nominal and ordinal. The nominal level represents categories that cannot be put in any order, while ordinal represents categories that **can** be ordered.

Examples:

Nominal: four seasons (winter, spring, summer, autumn)

Ordinal: rating your meal (disgusting, unappetizing, neutral, tasty, and delicious)

There are two quantitative levels: interval and ratio. They both represent "numbers", however, ratios **have a true zero**, while intervals don't.

Examples:

Interval: degrees Celsius and Fahrenheit

Ratio: degrees Kelvin, length

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 (f_e)

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

Observed Frequencies (f_o)

- 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 \neq 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

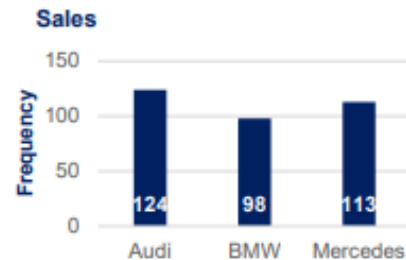
Graphs and tables that represent categorical variables

Frequency
distribution tables

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

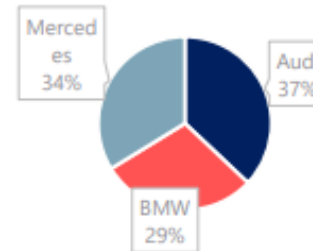
Frequency distribution tables show the category and its corresponding absolute frequency.

Bar charts



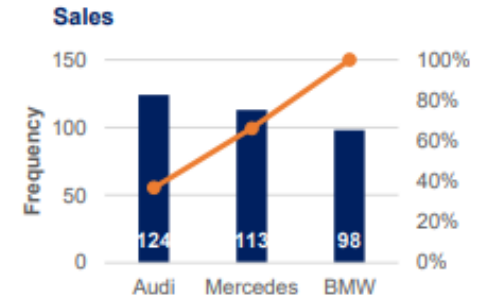
Bar charts are very common. Each bar represents a category. On the y-axis we have the absolute frequency.

Pie charts



Pie charts are used when we want to see the share of an item as a part of the total. Market share is almost always represented with a pie chart.

Pareto diagrams



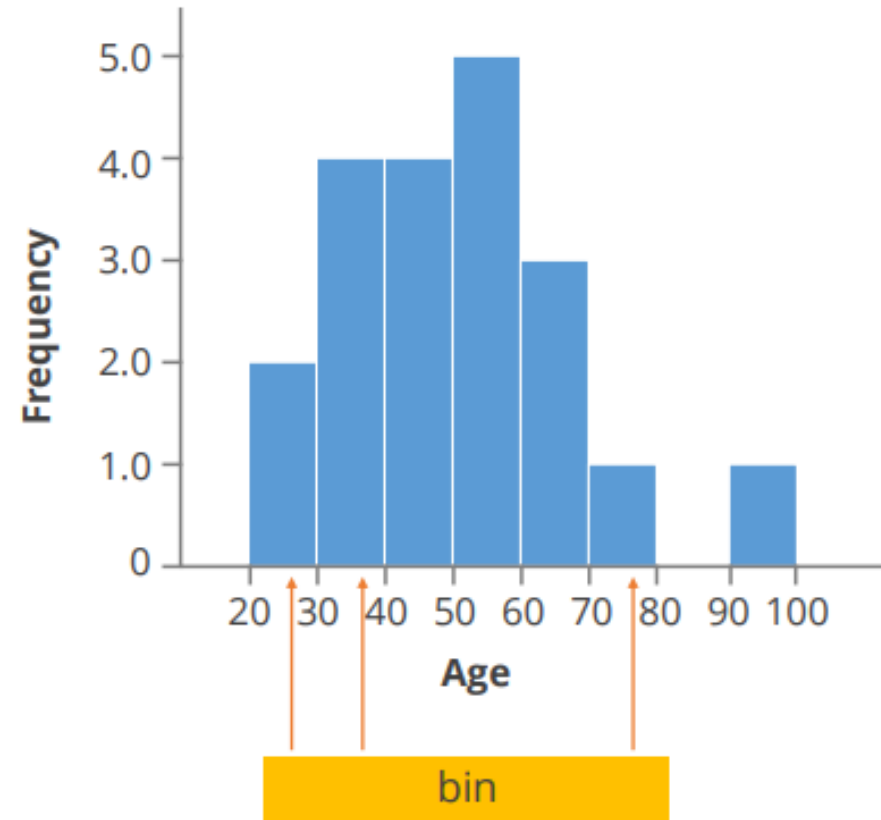
The Pareto diagram is a special type of bar chart where the categories are shown in descending order of frequency, and a separate curve shows the cumulative frequency.

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.

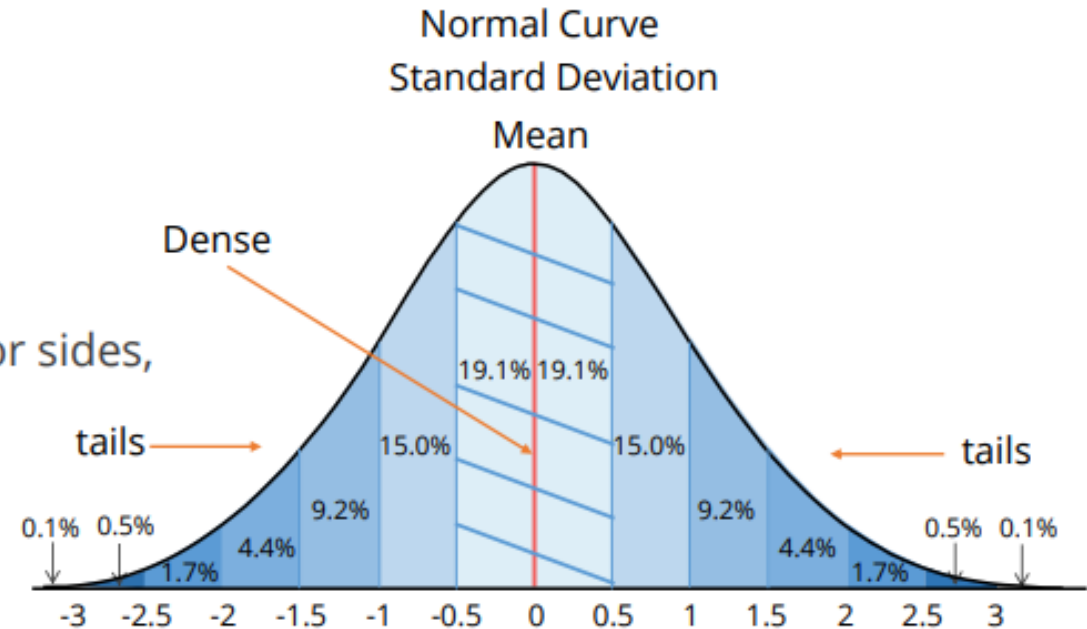


Bell Curve – Normal Distribution

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

Bell curve is:

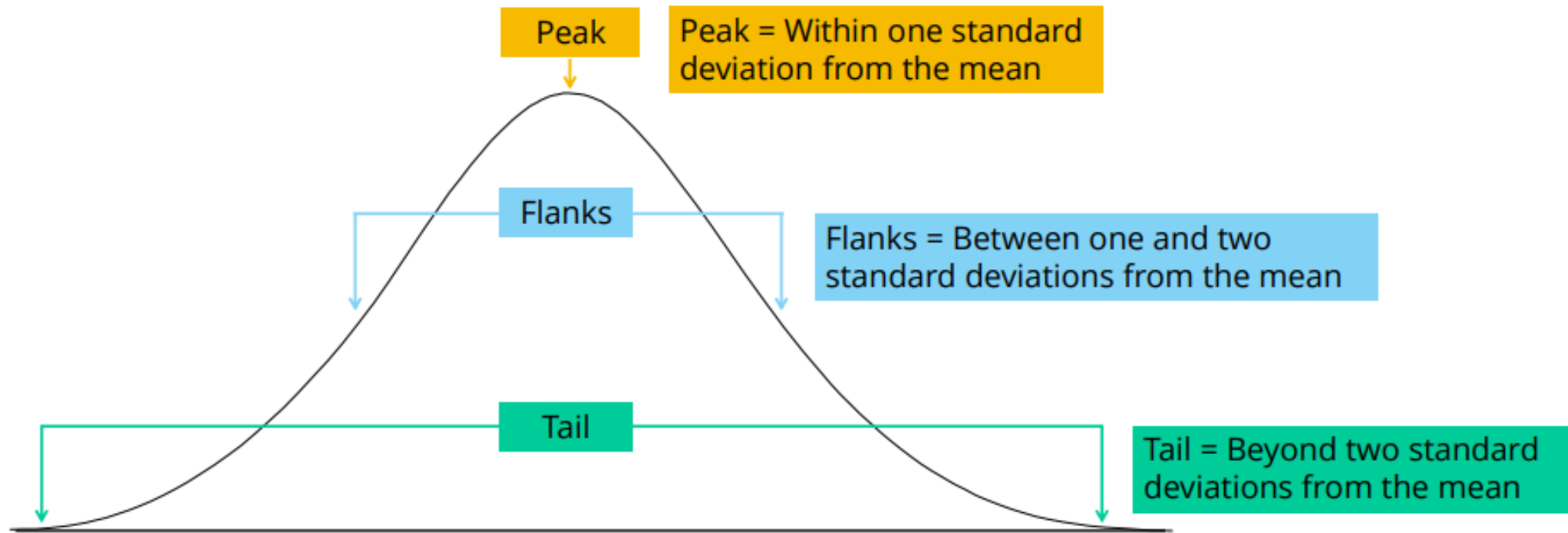
- Symmetric around the mean,
- Symmetric on both sides of the center,
- Having equal mean, median, and mode values,
- Denser in the center and less dense in the tails or sides,
- Defined by mean and standard deviation, and
- Known as the “Gaussian” curve.



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

The Bell Curve

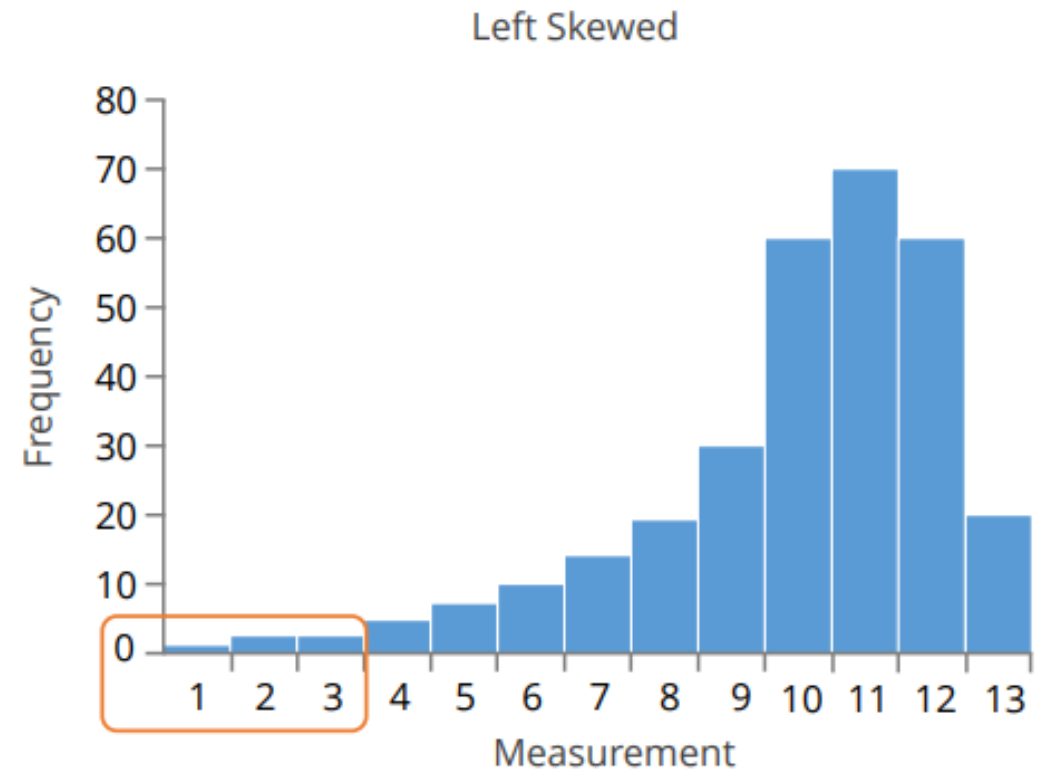
The Bell curve is divided into three parts to understand data distribution better.



Bell Curve - Left Skewed

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

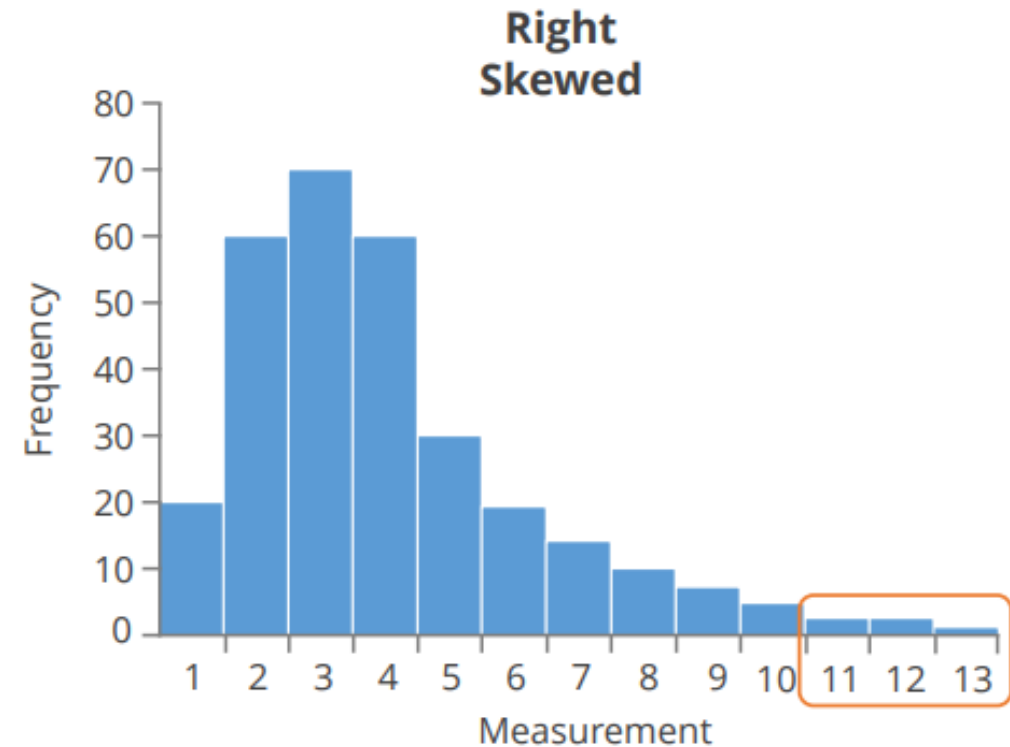
- The data is left skewed.
- $\text{Mean} < \text{Median}$
- The distribution is negatively skewed.
- Left tail contains large distributions.



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

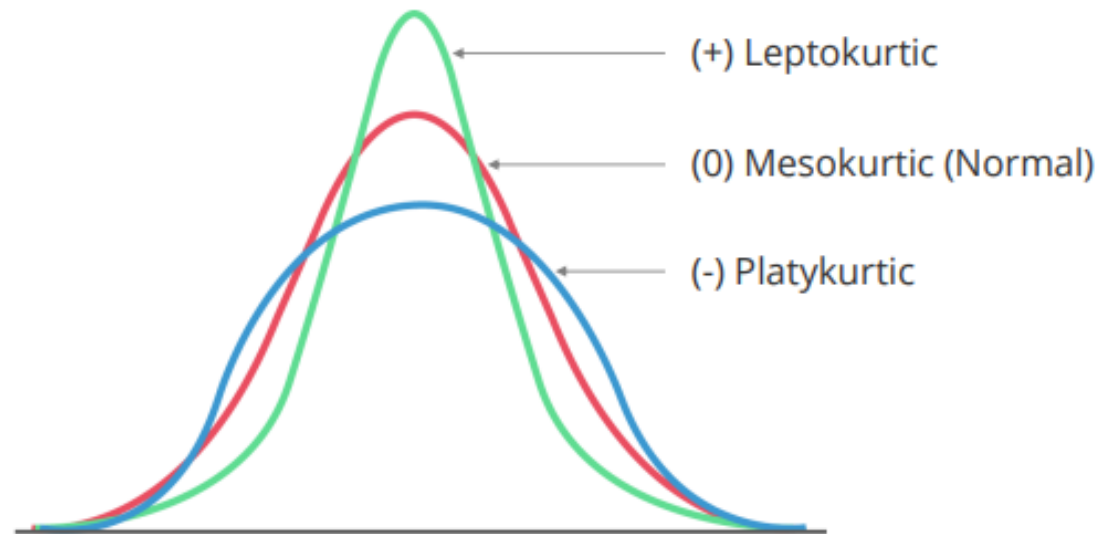
Kurtosis describes the shape of a probability distribution.

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.



Numerical variables. Frequency distribution table and histogram

Interval start	Interval end	Frequency	Relative frequency
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25

Frequency distribution tables for numerical variables are different than the ones for categorical. Usually, they are divided into intervals of equal (or unequal) length. The tables show the interval, the absolute frequency and sometimes it is useful to also include the relative (and cumulative) frequencies.

The interval width is calculated using the following formula:

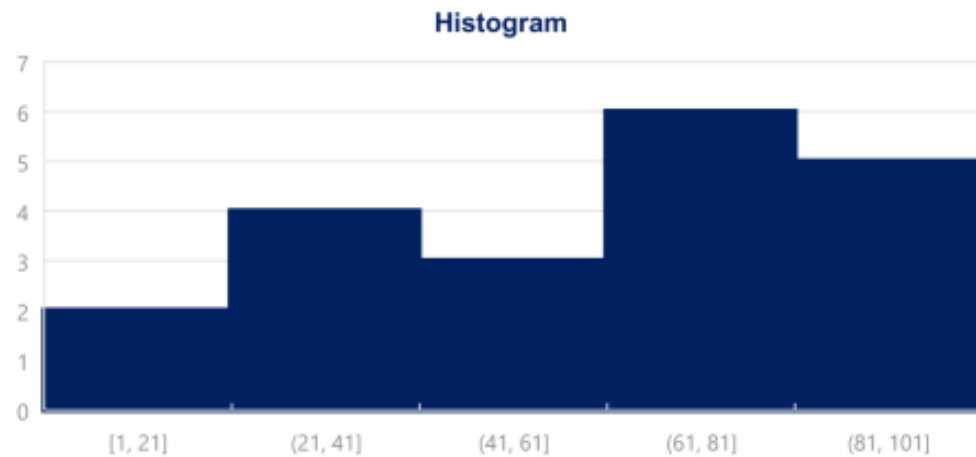
$$\text{Interval width} = \frac{\text{Largest number} - \text{smallest number}}{\text{Number of desired intervals}}$$



Creating the frequency distribution table in Excel:

1. Decide on the number of intervals you would like to use.
2. Find the interval width (using a the formula above).
3. Start your 1st interval at the lowest value in your dataset.
4. Finish your 1st interval at the lowest value + the interval width. (= start_interval_cell + interval_width_cell)
5. Start your 2nd interval where the 1st stops (that's a formula as well - just make the starting cell of interval 2 = the ending of interval 1)
6. Continue in this way until you have created the desired number of intervals.
7. Count the absolute frequencies using the following COUNTIF formula:
=COUNTIF(dataset_range,">="&interval start) -COUNTIF(dataset_range,">"&interval end).
8. In order to calculate the relative frequencies, use the following formula: = absolute_frequency_cell / number_of_observations
9. In order to calculate the cumulative frequencies:
 - i. The first cumulative frequency is equal to the relative frequency
 - ii. Each consecutive cumulative frequency = previous cumulative frequency + the respective relative frequency

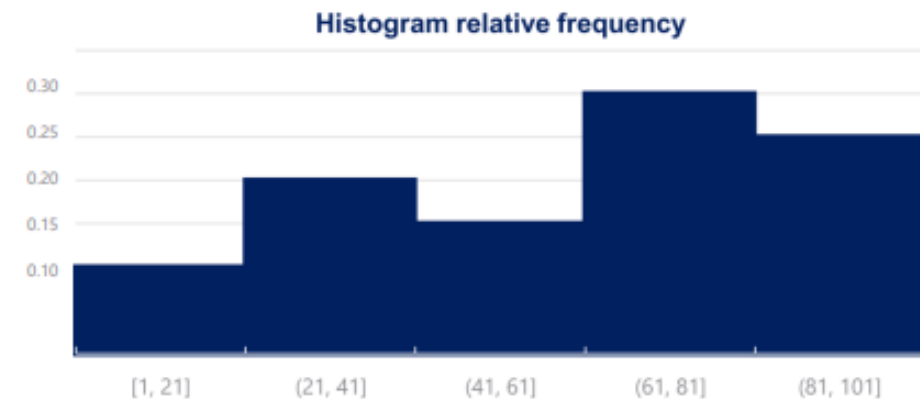
Numerical variables. Frequency distribution table and histogram



Histograms are the one of the most common ways to represent numerical data. Each bar has width equal to the width of the interval. The bars are touching as there is continuation between intervals: where one ends -> the other begins.

Creating a histogram in Excel:

1. Choose your data
2. **Insert -> Charts -> Histogram**
3. To change the number of bins (intervals):
 1. Select the x-axis
 2. Click **Chart Tools -> Format -> Axis options**
 3. You can select the bin width (interval width), number of bins, etc.



Graphs and tables for relationships between variables. Cross tables

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	96	185	39	320
Bonds	181	3	29	213
Real Estate	88	152	142	382
Total	365	340	210	915

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	0.10	0.20	0.04	0.35
Bonds	0.20	0.00	0.03	0.23
Real Estate	0.10	0.17	0.16	0.42
Total	0.40	0.37	0.23	1.00

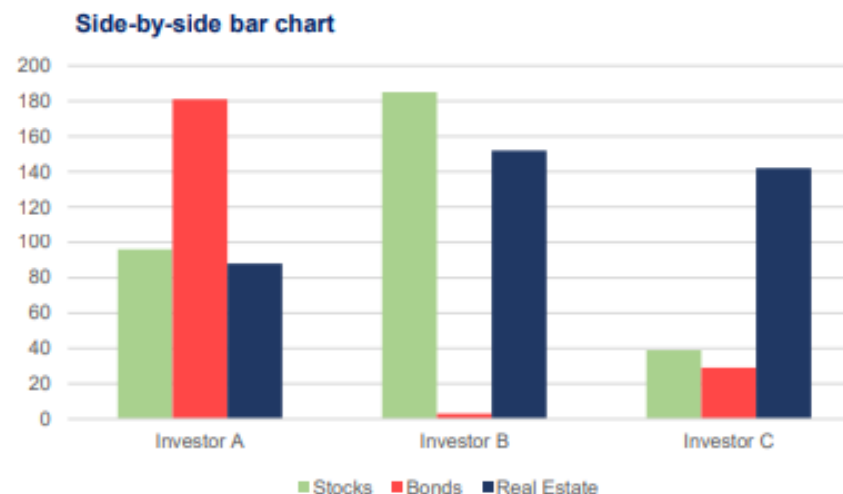
Cross tables (or contingency tables) are used to represent categorical variables. One set of categories is labeling the rows and another is labeling the columns. We then fill in the table with the applicable data. It is a good idea to calculate the totals. Sometimes, these tables are constructed with the *relative frequencies* as shown in the table below.

A common way to represent the data from a cross table is by using a side-by-side bar chart.

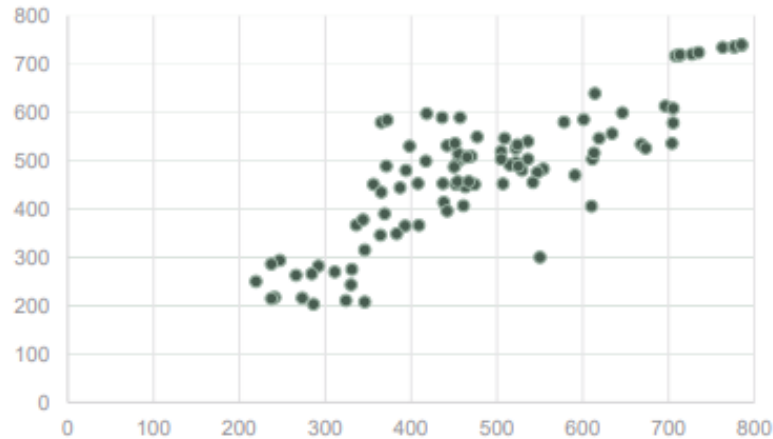
Creating a side-by-side chart in Excel:

1. Choose your data
2. **Insert -> Charts -> Clustered Column**

Selecting more than one series (groups of data) will automatically prompt Excel to create a side-by-side bar (column) chart.



Graphs and tables for relationships between variables. Scatter plots



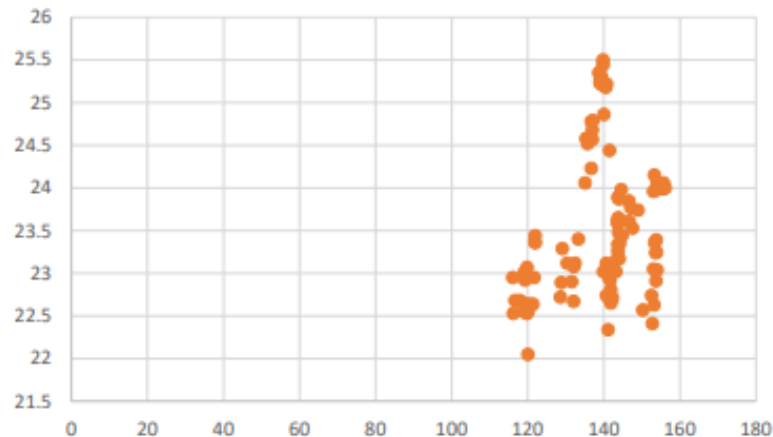
When we want to represent two numerical variables on the same graph, we usually use a scatter plot. Scatter plots are useful especially later on, when we talk about regression analysis, as they help us detect patterns (linearity, homoscedasticity).

Scatter plots usually represent lots and lots of data. Typically, we are not interested in single observations, but rather in the structure of the dataset.



Creating a scatter plot in Excel:

1. Choose the two datasets you want to plot.
2. Insert -> Charts -> Scatter



A scatter plot that looks in the following way (down) represents data that **doesn't have a pattern**. Completely vertical 'forms' show no association.

Conversely, the plot above shows a linear pattern, meaning that the observations move together.

Mean, median, mode


Mean

The mean is the most widely spread measure of central tendency. It is the simple average of the dataset.

Note: easily affected by outliers

The formula to calculate the mean is:

$$\frac{\sum_{i=1}^N x_i}{N} \quad \text{or} \quad \frac{x_1 + x_2 + x_3 + \dots + x_{N-1} + x_N}{N}$$

 In Excel, the mean is calculated by:


=AVERAGE()

Median

The median is the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science. That is since it is not affected by outliers.

In an ordered dataset, the median is the number at position $\frac{n+1}{2}$.

If this position is not a whole number, it, the median is the simple average of the two numbers at positions closest to the calculated value.


 In Excel, the median is calculated by:

=MEDIAN()

Mode

The mode is the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes.

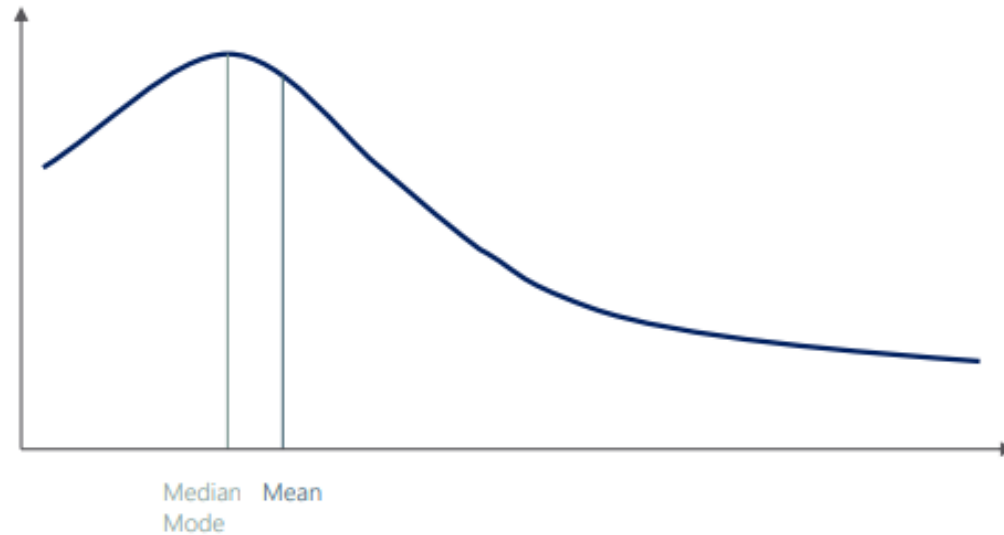
The mode is calculated simply by finding the value with the highest frequency.

 In Excel, the mode is calculated by:

=MODE.SNGL() -> returns one mode

=MODE.MULT() -> returns an array with the modes. It is used when we have more than 1 mode.

Skewness



Skewness is a measure of asymmetry that indicates whether the observations in a dataset are concentrated on one side.

Right (positive) skewness looks like the one in the graph. It means that the **outliers** are to the right (long tail to the right).

Left (negative) skewness means that the outliers are to the left.

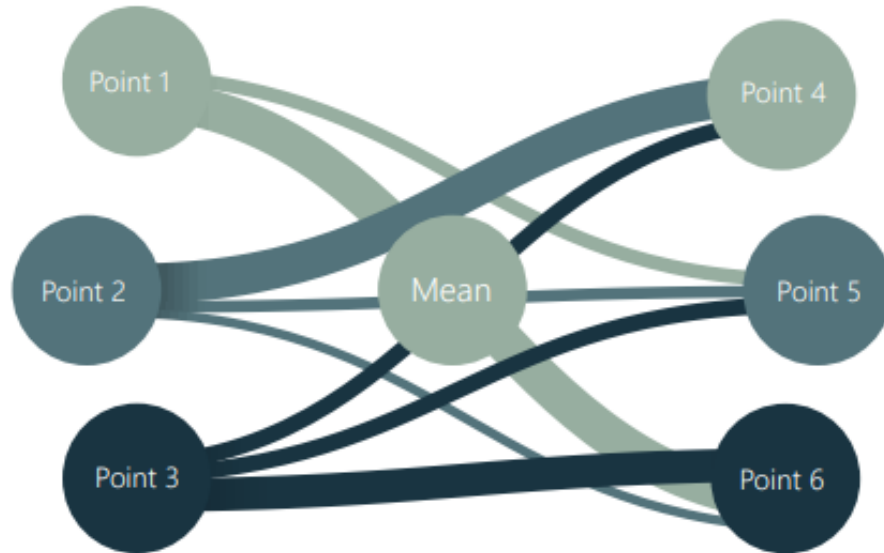
Usually, you will use software to calculate skewness.

 Calculating skewness in Excel:

=SKEW()

Formula to calculate skewness:
$$\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}^3}$$

Variance and standard deviation



Calculating variance in Excel:

Sample variance: `=VAR.S()`

Population variance: `=VAR.P()`

Sample standard deviation: `=STDEV.S()`

Population standard deviation: `=STDEV.P()`

Variance and standard deviation measure the dispersion of a set of data points around its mean value.

There are different formulas for population and sample variance & standard deviation. This is due to the fact that the sample formulas are the unbiased estimators of the population formulas. [More on the mathematics behind it.](#)

Sample variance formula:
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Population variance formula:
$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Sample standard deviation formula:
$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Population standard deviation formula:
$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

Covariance and correlation

Covariance


Covariance is a measure of the joint variability of two variables.

- A positive covariance means that the two variables move together.
- A covariance of 0 means that the two variables are independent.
- A negative covariance means that the two variables move in opposite directions.

Covariance can take on values from $-\infty$ to $+\infty$. This is a problem as it is very hard to put such numbers into perspective.

Sample covariance formula:
$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Population covariance formula:
$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x) * (y_i - \mu_y)}{N}$$

 In Excel, the covariance is calculated by:

Sample covariance: `=COVARIANCE.S()`

Population covariance: `=COVARIANCE.P()`


Correlation

Correlation is a measure of the joint variability of two variables. Unlike covariance, correlation could be thought of as a standardized measure. It takes on values between -1 and 1, thus it is easy for us to interpret the result.

- A correlation of 1, known as perfect positive correlation, means that one variable is perfectly explained by the other.
- A correlation of 0 means that the variables are independent.
- A correlation of -1, known as perfect negative correlation, means that one variable is explaining the other one perfectly, but they move in opposite directions.

Sample correlation formula:
$$r = \frac{s_{xy}}{s_x s_y}$$

Population correlation formula:
$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

 In Excel, correlation is calculated by:

`=CORREL()`

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

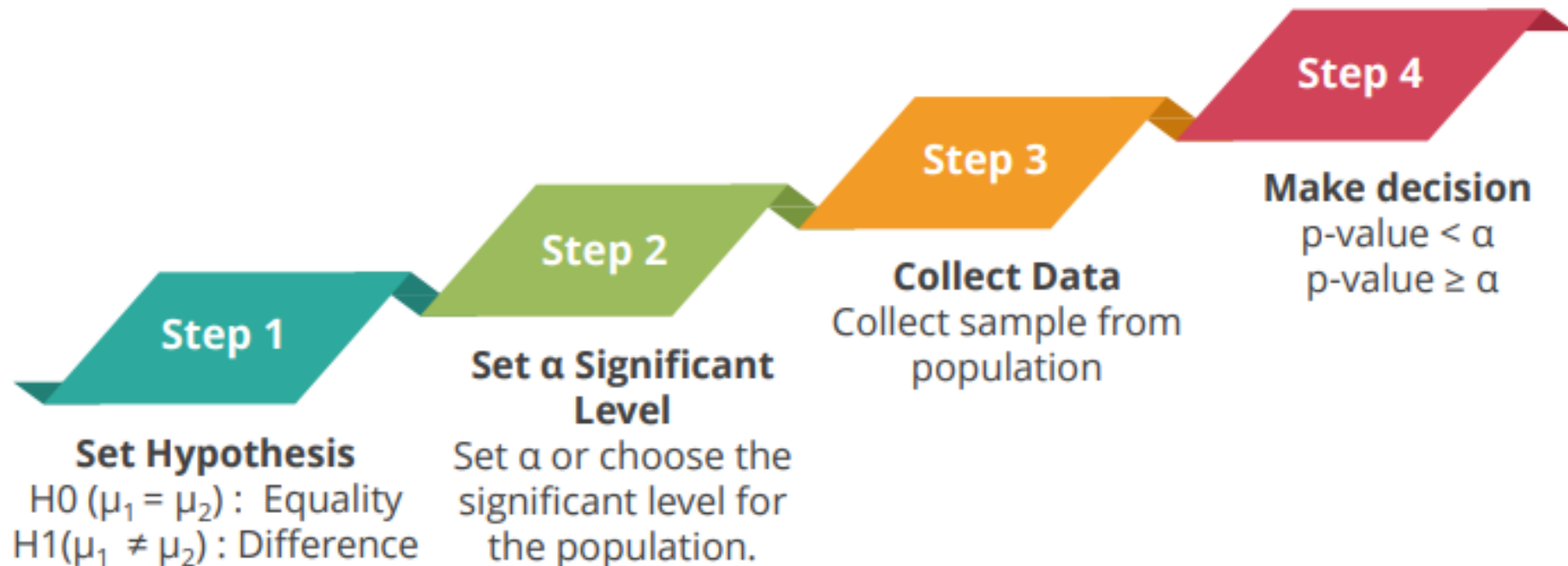
Representation of decision parameters using null hypothesis

Type I Error (α)	<ul style="list-style-type: none">• Rejects the null hypothesis when it is true• The probability of making Type I error is represented by α
Type II Error (β)	<ul style="list-style-type: none">• Fails to Reject the null hypothesis when it false• The probability of making Type II error is represented by β
<i>p-value</i>	<ul style="list-style-type: none">• 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.



Reject the null hypothesis if $p\text{-value} < \alpha$
Fail to reject the null hypothesis if $p\text{-value} \geq \alpha$

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.

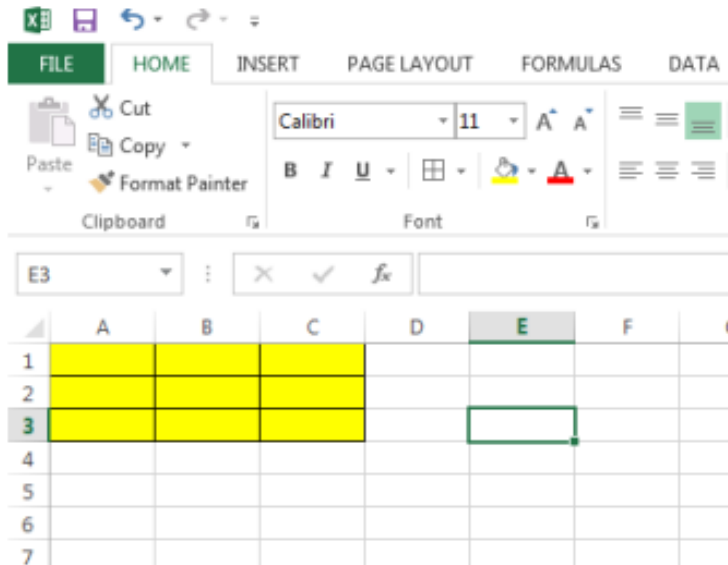


Poisson Data

Evaluate rate of occurrence or frequency.

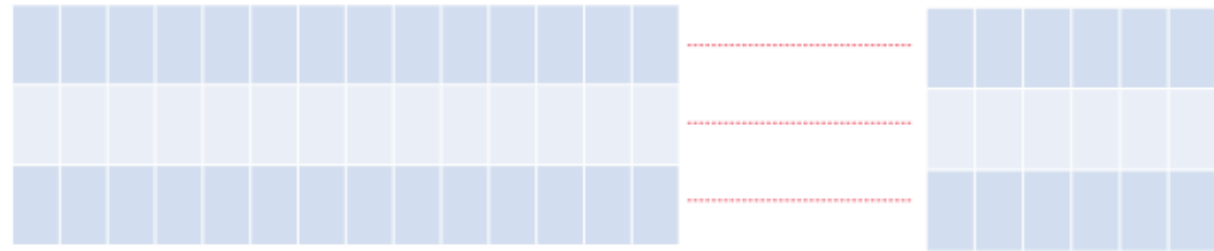
Correlation Matrix

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



(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 matrix – a square matrix

$n \times n$ Matrix

(very large number of rows and columns)

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.
	<ul style="list-style-type: none">• 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		fx =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.

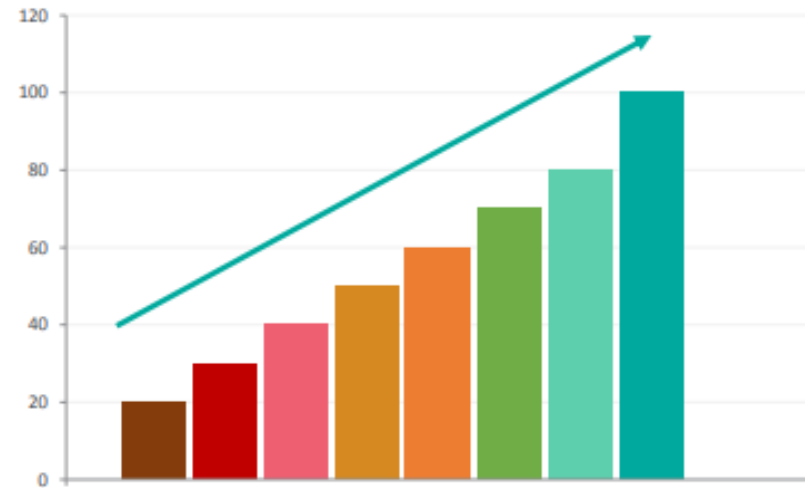
- It can help determine the strength of the relationships within the sample.
- If it is very difficult to obtain a population list and draw a random sample, do the best you can with what you have.

Applications of Inferential Statistics

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

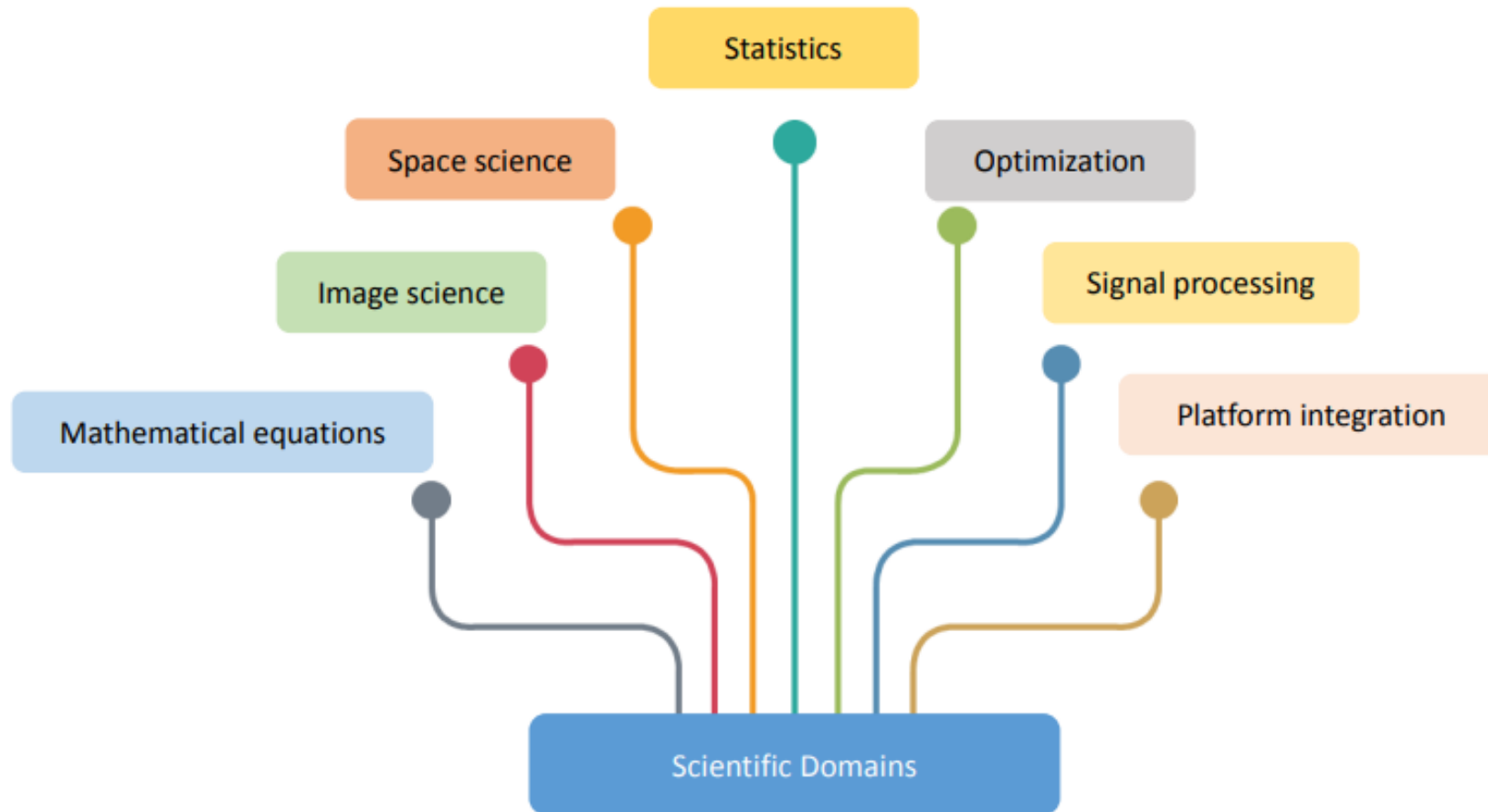
Inferential Statistics

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



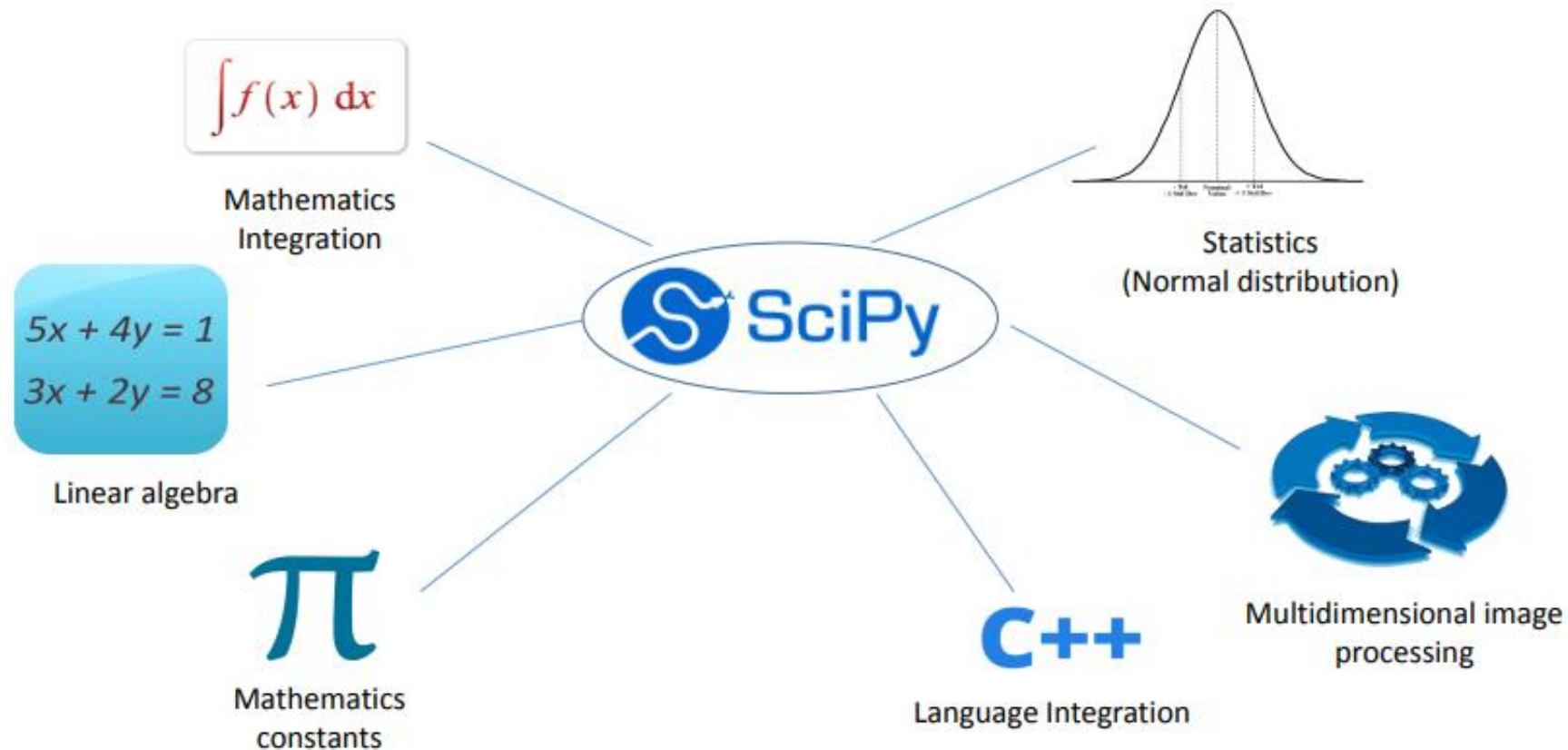
The Real World: Multiple Scientific Domains

How to handle multiple scientific domains? The solution is SciPy.



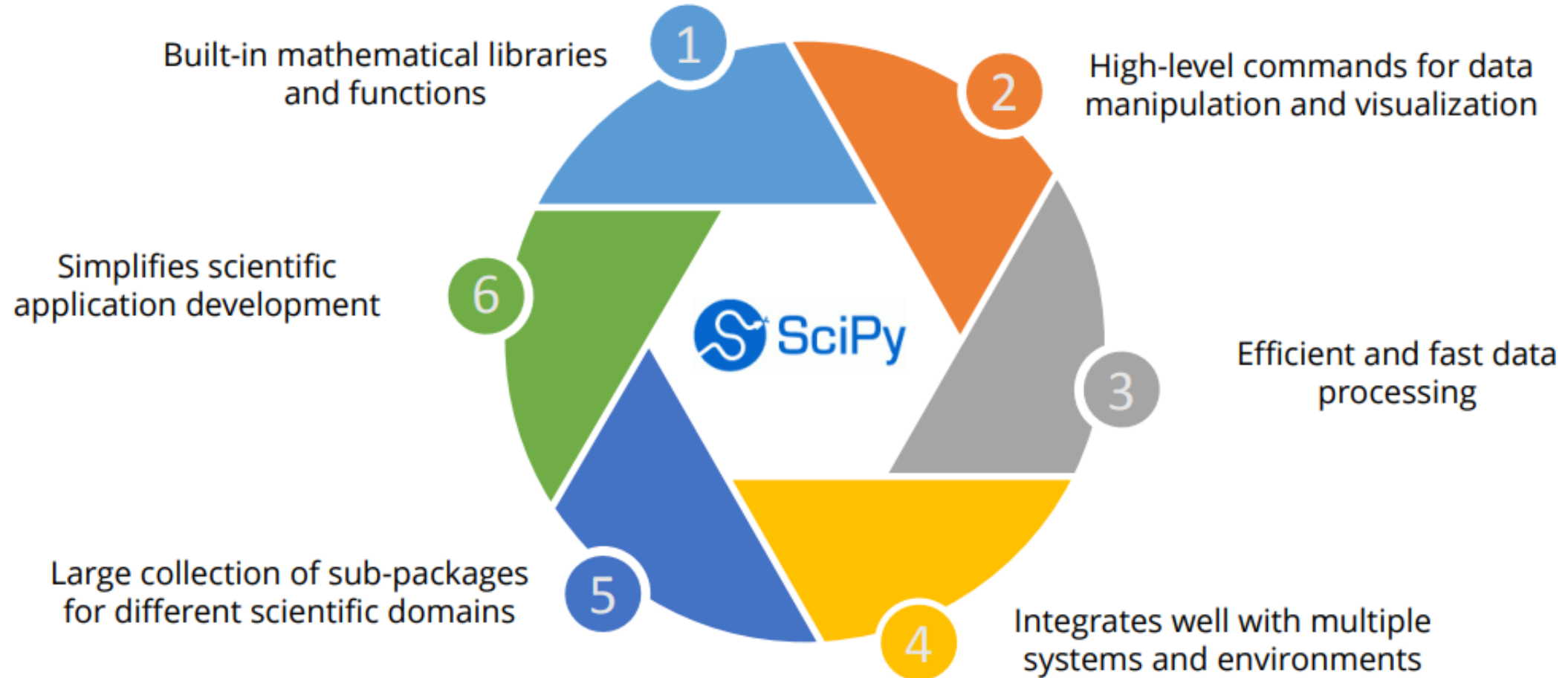
SciPy: The Solution

SciPy has built-in packages that help in handling the scientific domains.



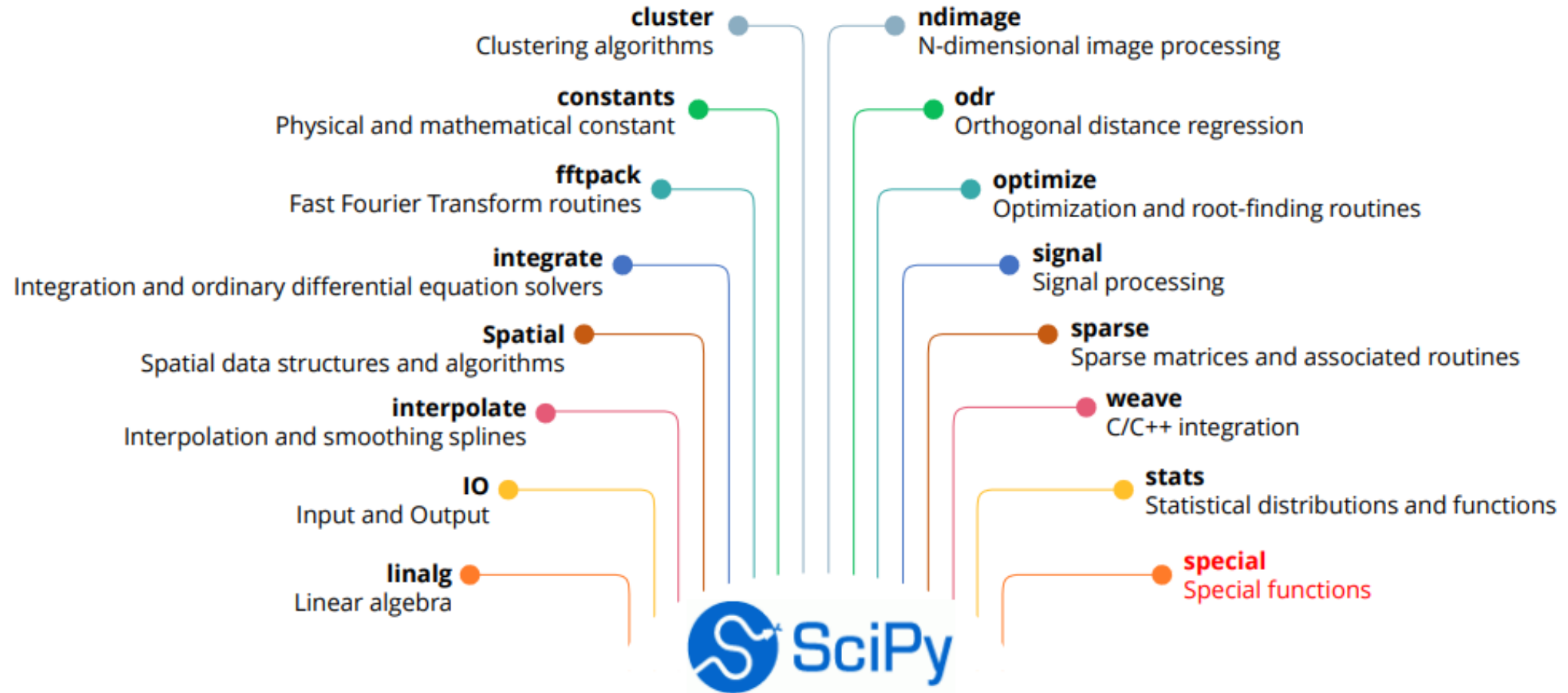
SciPy and its Characteristics

Characteristics of SciPy are as follows:



SciPy Sub-package

SciPy has multiple sub-packages which handle different scientific domains.

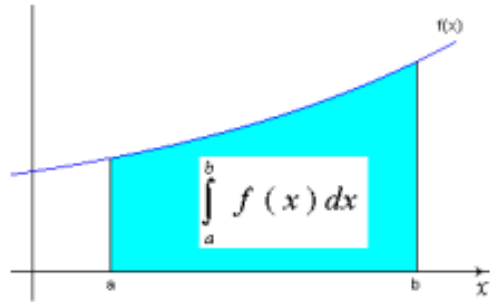


SciPy Sub-package: Integration

SciPy provides integration techniques that solve mathematical sequences and series, or perform function approximation.

General integration (quad)

- `integrate.quad(f, a, b)`



General multiple integration (dblquad, tplquad, nquad)

- `integrate.dblquad()`
- `integrate.tplquad()`
- `integrate.nquad()`

The limits of all inner integrals need to be defined as functions.