# Module 2: Data Exploration

2.1
Types of data, Properties of data

**Descriptive Statistics:**
Univariate Exploration: Measure of Central Tendency, Measure of Spread, Symmetry, Skewness: Karl Pearson Coefficient of skewness, Bowley‘s Coefficient, Kurtosis
Multivariate Exploration: Central Data Point, Correlation, Different forms of correlation, Karl Pearson Correlation Coefficient for bivariate distribution

**2.2 Inferential Statistics:**
Overview of Various forms of distributions: Normal, Poisson, Test Hypothesis, Central limit theorem, Confidence Interval, Z-test, t-test, Type-I, Type-II Errors, ANOVA

# Types of Data

Data come in different formats and types.
Understanding the properties of each attribute or feature provides information about what kind of operations can be performed on that attribute.
The temperature in weather data can be expressed as any of the following formats:

- Numeric centigrade (31C, 33.3C) or Fahrenheit (100F, 101.45F) or on the Kelvin scale
- Ordered labels as in hot, mild, or cold
- Number of days within a year below 0C (10 days in a year below freezing)

All of these attributes indicate temperature in a region, but each have different data types.
A few of these data types can be converted from one to another.

2 Types

1. Numeric or Continuous Data
2. Categorical Data

**Continuous Data:**

can be denoted by numbers and take an infinite number of values between digits.

Values are ordered and calculating the difference between the values makes sense

additive and subtractive mathematical operations and logical comparison operators like greater than, less than, and equal to, operations can be applied.

An integer is a special form of the numeric data type which does not have decimals in the value or more precisely does not have infinite values between consecutive numbers

Temperature expressed in Centigrade or Fahrenheit is numeric and continuous

Continuous Data denote a count of something, number of days with temperature less than 0C, number of orders, number of children in a family

## Categorical or Nominal

Categorical data types are attributes treated as distinct symbols or just names.

represent some characteristics or attributes of the data.

The facts and figures depicted by the qualitative data cannot be computed.

These are non-numerical in nature.

The color of the iris of the human eye is a categorical data type

because it takes a value like black, green, blue, gray, etc.

There is no direct relationship among the data values, and hence, mathematical operators except the logical or “is equal” operator cannot be applied.

Not all data science tasks can be performed on all data types.

For example,the neural network algorithm does not work with categorical data.

However,one data type can be converted to another using a type conversion process,but this may be accompanied with possible loss of information.

**credit scores** expressed in poor, average, good, and excellent categories can be converted to either 1, 2, 3, and 4

or average underlying numerical scores like 400, 500, 600, and 700 (scores here are just an example).

In this type conversion, there is no loss of information.

However, conversion from **numeric credit score to categories** (poor, average, good, and excellent) does incur loss of information.

# DESCRIPTIVE STATISTICS

what is meant by statistics?

It means collection, organization, analysis and interpretation of data.

Statistics are mainly used to give numerical conclusions.

For example, if anyone asks you how many people are watching youtube

In this case, we can't say more; many people are watching youtube, we have to answer in numerical terms that give more meaning to you.

# Statistics include

**Design of experiments:** Used to understand Characteristics of the dataset

**Sampling**: Used to understand the samples

**Descriptive statistics:** Summarization of data

**Inferential Statistics:** Hypothesis way of concluding data

**Probability Theory:** Likelihood estimation

Two types: Descriptive and Inferential statistics.

Descriptive statistics refers to the study of the aggregate quantities of a dataset.

| Characteristics of the Dataset | Measurement Technique |
|---|---|
| Center of the dataset | Mean, median, and mode |
| Spread of the dataset | Range, variance, and standard deviation |
| Shape of the distribution of the dataset | Symmetry, skewness, and kurtosis |

Descriptive statistics can be broadly classified into

Univariate and

Multivariate exploration

depending on the number of attributes under analysis

# Arithmetic Mean (Mean)

Definition:
Sum of all the observation s divided by the number of the observations

The arithmetic mean is the most common measure of the central location of a sample.

Population $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$

Sample $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

“N”, the total number of scores in a population

“n”, the total number of scores in a sample

## Univariate Exploration

Univariate data exploration denotes analysis of one attribute at a time

**Measure of Central Tendency or Measures of Location or Measures of Averages**

A measure of central tendency is a descriptive statistic that describes the average, or typical value of a set of scores

A single summary score that best describes the central location of an entire distribution of scores

There are three common measures of central tendency:

- **the mean:** The sum of all scores divided by the number of scores.
- **the median:** The value that divides the distribution in half
- **the mode:** The most frequent score

When the data are arranged or given in the form of frequency distribution *i.e.*, there are $k$ variate values such that a value $X_i$ has a frequency $f_i$ ($i = 1, 2, ..., k$), the formula for the mean is,

$$\mu = \frac{f_1X_1 + f_2X_2 + ... + f_kX_k}{f_1 + f_2 + ... + f_k} \qquad ...(3.2)$$

$$= \frac{\Sigma_i f_i X_i}{\Sigma_i f_i} \qquad i = 1, 2, ..., k \qquad ...(3.2.1)$$

$$= \frac{1}{N}\Sigma_i f_i X_i \qquad ...(3.2.2)$$

where $N = f_1 + f_2 + ... + f_k = \Sigma_i f_i$

**Example :** Calculated the Arithmetic Mean Monthly Users Statistics in the University Library

| Month | No. of Working Days | Total Users | Average Users per month |
|---|---|---|---|
| Sep-2011 | 24 | 11618 | 484.08 |
| Oct-2011 | 21 | 8857 | 421.76 |
| Nov-2011 | 23 | 11459 | 498.22 |
| Dec-2011 | 25 | 8841 | 353.64 |
| Jan-2012 | 24 | 5478 | 228.25 |
| Feb-2012 | 23 | 10811 | 470.04 |
| **Total** | | | |

$$\text{Mean} = \frac{\text{Total number of users}}{\text{Total number of working days}}$$

$$= \frac{\sum X}{N} \quad = \frac{57064}{140} \quad = 407.6$$

Pros
Mathematical center of a distribution.
Good for interval and ratio data.
Does not ignore any information.
Inferential statistics is based on mathematical properties of the mean.

Cons
Influenced by extreme scores and skewed distributions.
May not exist in the data

- 52, 76, 100, 136, 186, 196, 205, 150, 257, 264, 264, 280, 282, 283, 303, 313, 317, 317, 325, 373, 384, 384, 400, 402, 417, 422, 472, 480, 643, 693, 732, 749, 750, 791, 891
- Mean hotel rate:

$$\overline{X} = \frac{\Sigma X}{n}$$

$$\overline{X} = \frac{13005}{35} = 371.60$$

- Mean hotel rate: $371.60

Median: The median is the value of the central point in the distribution.

The median is calculated by sorting all the observations from small to large and selecting the mid-point observation in the sorted list.

If the number of data points is even, then the average of the middle two data points is used as the median

Pros

Not influenced by extreme scores or skewed distributions.

Good with ordinal data.

Easier to compute than the mean.

Cons

May not exist in the data.

Doesn't take actual values into account

Mode: The mode is the most frequently occurring observation.

In the dataset, data points may be repetitive, and the most repetitive data point is the mode of the dataset

- 52, 76, 100, 136, 186, 196, 205, 150, 257, 264, 264, 280, 282, 283, 303, 313, 317, 317, 325, 373, 384, 384, 400, 402, 417, 422, 472, 480, 643, 693, 732, 749, 750, 791, 891
- Mode: most frequent observation
- Mode(s) for hotel rates:
  - 264, 317, 384

**Example 17.** *A candidate obtains the following percentages in an examination. English 46%, Mathematics 67%, Sanskrit 72%, Economics 58%, Political Science 53%. It is agreed to give double weights to marks in English and Mathematics as compared to other subjects. What is the weighted mean?*

**Solution.** Since it is agreed to give double weights to marks in English and Mathematics, we have the following data:

| *Subjects* | *Marks* $(x)$ | *Weights* $(w)$ | $wx$ |
|---|---|---|---|
| English | 46 | 2 | 92 |
| Mathematics | 67 | 2 | 134 |
| Sanskrit | 72 | 1 | 72 |
| Economics | 58 | 1 | 58 |
| Political Science | 53 | 1 | 53 |
| | | $\Sigma w = 7$ | $\Sigma wx = 409$ |

$$\text{Weighted mean} = \frac{\Sigma wx}{\Sigma w} = \frac{409}{7} = 58.43.$$

**Example 8.** *Calculate the mean deviation about the mean for the following series :*
*15, 20, 17, 19, 21, 13, 12, 10, 17, 9, 12.*

**Solution.** Here $n = 11$, and therefore

$$\text{Mean} = \frac{15+20+17+19+21+13+12+10+17+9+12}{11} = \frac{165}{11} = 15 = M \text{ (say)}.$$

Now

| $x$ | $d = x - M = (x - 15)$ | $\lvert d \rvert$ |
|---|---|---|
| 15 | 0 | 0 |
| 20 | 5 | 5 |
| 17 | 2 | 2 |
| 19 | 4 | 4 |
| 21 | 6 | 6 |
| 13 | − 2 | 2 |
| 12 | − 3 | 3 |
| 10 | − 5 | 5 |
| 17 | 2 | 2 |
| 9 | − 6 | 6 |
| 12 | − 3 | 3 |
| | | $\Sigma \lvert d \rvert = 38$ |

$$\therefore \text{ Mean deviation} = \frac{\Sigma |d|}{N} = \frac{38}{11} = 3.455.$$

*examination were as follows:*

### Computation of Mode in a Continuous Frequency Distribution (Or Method of Interpolation)

(*i*) **Modal class.** It is that **class in a grouped frequency distribution in which the mo** The modal class can be determined either by inspection or with the help of grouping table finding the modal class, we calculate the mode by the following formula:

$$\text{Mode} = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i,$$

where

$l$ = **the lower limit of the modal class**

$i$ = **the width of the modal class**

$f_1$ = **the frequency of the class preceding the modal class**

$f_m$ = **the frequency of the modal class**

$f_2$ = **the frequency of the class succeeding the modal class.**

Sometimes, it so happened that the above formula fails to give the mode. In this **modal value lies in a class other than the one containing maximum frequency.** In su we take the help of the following formula:

$$\text{Mode} = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i \text{, where } \Delta_1 = f_m - f_1, \Delta_2 = f_m - f$$

**where $l, f_1, f_2, f_m$, and $l$ have usual meanings.**

The precedure of finding the mode by the above method is called **Method of Inte**

The precedure of finding the mode by the above method is called **Method of In**

**Example 43.** *Find the mode for the following data:*

| *Marks* | *1 – 5* | *6 – 10* | *11 – 15* | *16 – 20* |
|---|---|---|---|---|
| *No. of student* | *7* | *10* | *16* | *32* |

**Solution.** From the above table, it is clear that the maximum frequency is **32** an**d class** 16 – 20. Thus, the **modal class is 16 – 20.**

Here $l = 16,\ f_m = 32,\ f_1 = 16,\ f_2 = 24,\ i = 5.$

$$\textbf{Mode} = l + \frac{f_m - f_1}{2f_m - f_2 - f_1} \times i = 16 + \frac{32 - 16}{64 - 24 - 16} \times 5$$

$$= 16 + \frac{16}{24} \times 5 = 16 + \frac{10}{3} = 16 + 3.33 = \textbf{19.33.}$$

**Example 44.** *Calculate Median and Mode from*

8, 11, 13, 15, $x + 1$, $x + 3$, 30, 35, 40, 43.

Total number of observations = 10 **(even)**

$$\therefore \text{ Median } = \frac{\frac{n}{2}\text{th item} + \left(\frac{n}{2}+1\right)\text{th item}}{2} = \frac{5\text{th item} + 6\text{th item}}{2}$$

$$\Rightarrow 22 = \frac{(x+1)+(x+3)}{2} \Rightarrow 22 = \frac{2x+4}{2} \Rightarrow 22x + 4 \Rightarrow x = 20.$$

**CASE II. When the series is continuous.**

In this case the data is given in the form of a frequency table with class-interval, etc., and the owing formula is used to calculate the Median.

**Median:**

$$M = L + \frac{(n/2) - C}{f} \times i, \text{ where}$$

$L$ = lower limit of the class in which the median lies,

$n$ = total number of frequencies, *i.e.*, $n = \Sigma f$,

$f$ = frequency of the class in which the median lies,

$C$ = cumulative frequency of the class preceding the median class,

$i$ = width of the class-interval of the class in which the median lies.

**Example 32.** *The following table gives the weekly expenditure of 100 families. Find the median weekly expenditure.*

| *Weekly Expenditure (in Rs. )* | *Number of families* |
|---|---|
| 0 – 10 | 14 |
| 10 – 20 | 23 |
| 20 – 30 | 27 |
| 39 – 40 | 21 |
| 40 – 50 | 15 |

**Solution.** Let us prepare a table which gives the frequencies and cumulative frequencies.

**Table: Computation of Median**

| *Weekly Expenditure (in Rs.)* | *Number of families (frequency) f* | *Cumulative frequency* |
|---|---|---|
| 0 – 10 | 14 | 14 |
| 10 – 20 | 23 | 37 |
| 20 – 30 | 27 | 64 |
| 30 – 40 | 21 | 85 |
| 40 – 50 | 15 | 100 |
| | | Here $n = \Sigma f = 100$ |

$\therefore$ Median $= \left(\frac{n}{2}\right)$th value $= \left(\frac{100}{2}\right)$th value $=$ 50th value.

Median class $= 20 - 30$.

Here $\frac{n}{2} = 50$, $L = 20$, $C = 37$, $i = 10$.

$\therefore$ **Median** $= L + \frac{(n/2) - C}{f} \times i = 20 + \frac{50 - 37}{27} \times 10 = 20 + \frac{13}{27} \times 10 = 20 + 4.815 =$ 24.815.

Hence, **median = 24.815.**

**Example 33.** *The following table gives the marks obtained by 50 students in Economics. Find*

**Step V.** The value obtained in Step IV above is the required **median.**

**Example 29.** *Calculate median for the following data:*

| *No. of students* | 6 | 4 | 16 | 7 | 8 | 2 |
|---|---|---|---|---|---|---|
| *Marks* | 20 | 9 | 25 | 50 | 40 | 80 |

**Solution.** Arranging the marks in ascending order and preparing the following table:

**Table: Computation of Median**

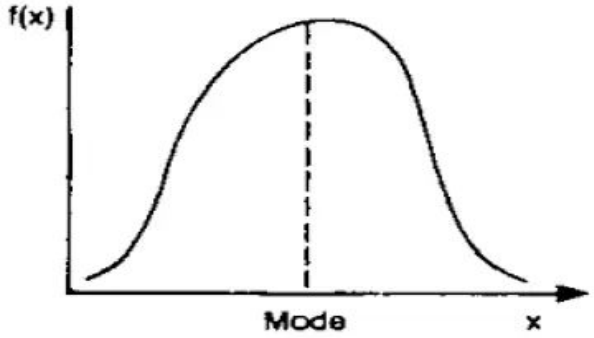
| *Marks* | *Frequency* | *Cumulative Frequency* |
|---|---|---|
| 9 | 4 | 4 |
| 20 | 6 | 10 |
| 25 | 16 | 26 |
| 40 | 8 | 34 |
| 50 | 7 | 41 |
| 80 | 2 | 43 |
| | $n = \Sigma f = 43$ | |

# Types of Mode

Unimodal Mode

A unimodal mode is a set of data with only one mode.

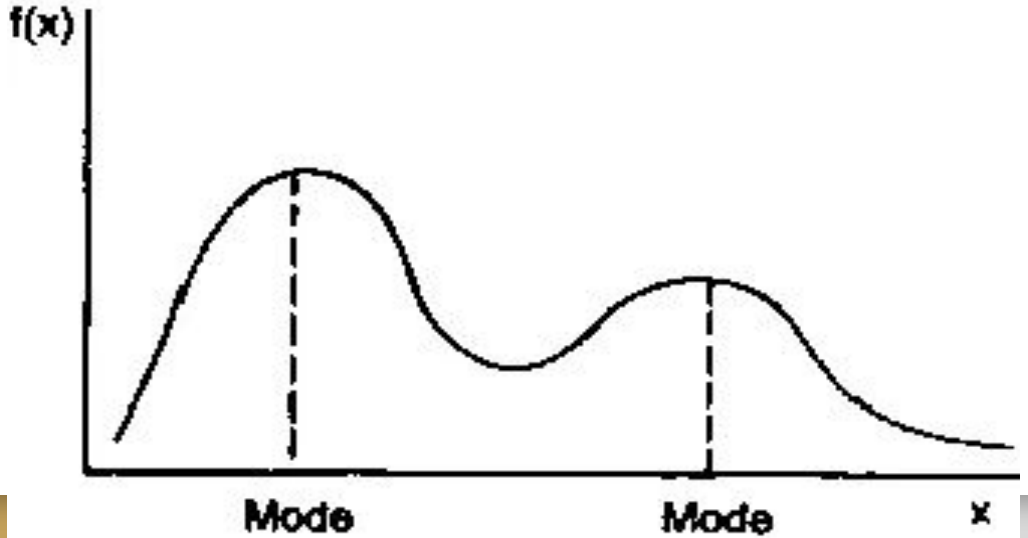
The mode of data set A = {14, 15, 16, 17, 15, 18, 15, 19}, for example, is 15 because just one value repeats itself. As a result, it's a unimodal data set.



# Bimodal Mode

A bimodal mode is a set of data that has two modes. This indicates that the data values with the highest frequencies are two.

Set A = {2,2,2,3,4,4,5,5,5} has a mode of 2 and 5, because both 2 and 5 are repeated three times in the provided set.



## Trimodal Mode

A trimodal mode is a set of data that has three modes. This indicates that the top three data values have the most frequency.

Set A = {2,2,2,3,4,4,5,5,5,7,8,8,8} has a mode of 2, 5, and 8 since all three numbers are repeated thrice in the provided set. As a result, it's a trimodal data collection.

## Multimodal Mode

A multimodal mode is a set of data that contains four or more modalities.

## Pros

Good for nominal data.

Easiest to compute and understand.

The score comes from the data set.

## Cons

Ignores most of the information in a distribution.

Small samples may not have a mode.

**EXAMPLE 3.2.1**

To paraphrase Benjamin Disraeli: "There are lies, darn lies, and DAM STATISTICS."

Compute the mean, median and mode for the following DAM STATISTICS:

| *Name of Dam* | *Height* |
|---|---|
| Oroville dam | 756 ft. |
| Hoover dam | 726 ft. |
| Glen Canyon dam | 710 ft. |
| Don Pedro dam | 568 ft. |
| Hungry Horse dam | 564 ft. |
| Round Butte dam | 440 ft. |
| Pine Flat Lake dam | 440 ft. |

MEAN = (756 + 726 + 710 + 568 + 564 + 440 + 440)/7

= 4204/7

= 600.57 (this has been rounded).

We can say that the typical dam is 600.57 feet tall

In order to find the median we must first list the data points in numerical order
The median is 568.

the number 440 occurs more often than any of the other numbers on this list, the mode is 440

# Measures of Spread

**A measure of spread**, sometimes also called a **measure of dispersion**, is used to describe the **variability** in a sample or population.

Variability: how far apart data points lie from each other and from the center of a distribution.

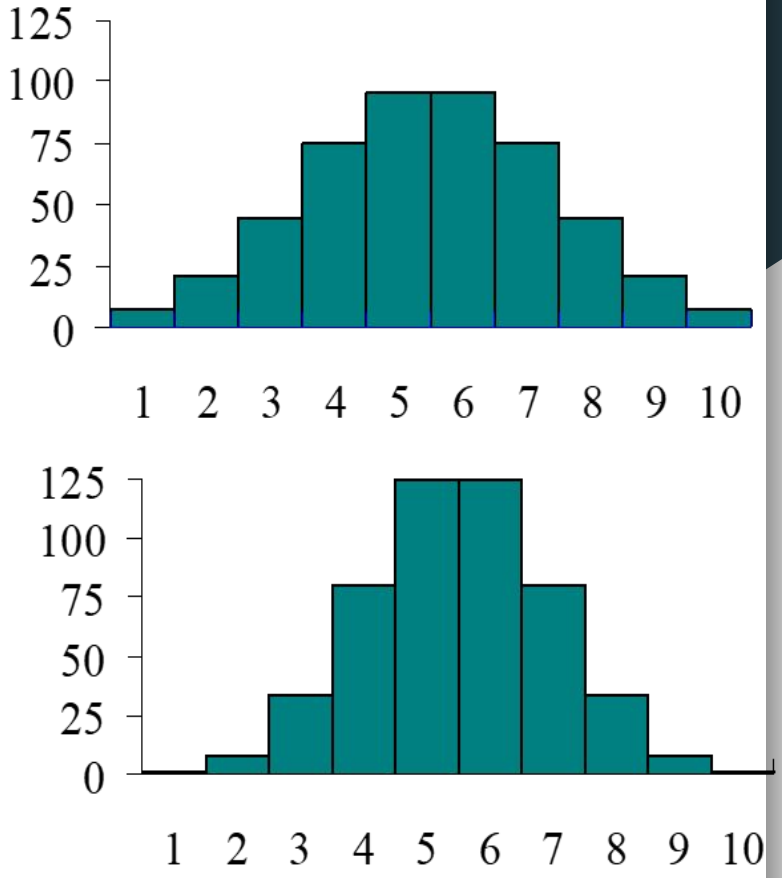
descriptive statistics that describe how similar a set of scores are to each other

- The more similar the scores are to each other, the lower the measure of dispersion will be
- The less similar the scores are to each other, the higher the measure of dispersion will be
- In general, the more spread out a distribution is, the larger the measure of dispersion will be

Which of the distributions of scores has the larger dispersion?

The upper distribution has more dispersion because the scores are more spread out

That is, they are less similar to each other



## Why do we measure spread?

Summarising the dataset can help us understand the data, especially when the dataset is large.

the Measures of Central Tendency, the mode, median, and mean summarise the data into a single value

Representative of all the values in the dataset, but this is only part of the 'picture' that summarises a dataset.

Measures of spread summarise the data in a way that shows how scattered the values are and how much they differ from the mean value.

Measures of spread include the range, quartiles and the interquartile range, variance and standard deviation.

There are two common metrics to quantify spread

1. Range

   The range is the difference between the maximum value and the minimum value of the attribute.

   The range is simple to calculate but has **shortcomings** as it is severely impacted by the presence of **outliers**

   So it  fails to consider the distribution of all other data points in the attributes

   Range = maximum value - minimum value

## When To Use the Range

- The range is used when
  - you have ordinal data
    - Ordinal data is a kind of categorical data with a set order or scale to it.
  - or
  - you are presenting your results to people with little or no knowledge of statistics

- The range is rarely used in scientific work as it is fairly insensitive
  - It depends on only two scores in the set of data, XL and XS

Two very different sets of data can have the same range:

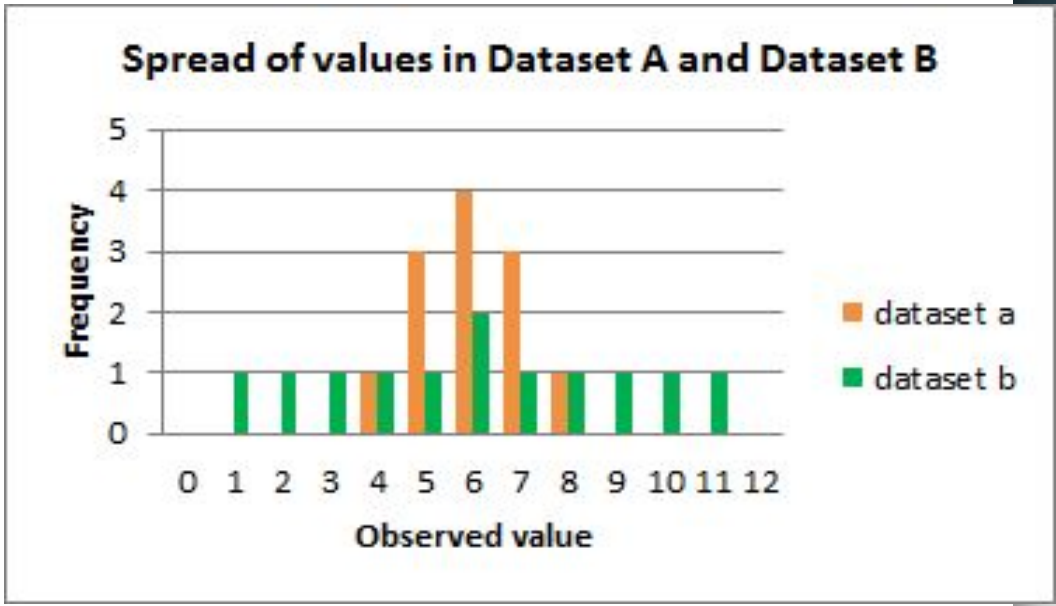
1 1 1 2 4 5 6 1 9 vs 1 3 2 4 5 6 7 5 7 9

Dataset A

4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

Dataset B

1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11




if we look at the spread of the values in the graph,
we can see that Dataset B is more dispersed than Dataset A.

Used together,

- **the measures of central tendency and**
- **measures of spread**

help us to better understand the data

Variance is defined as the average of the square deviations:measures of the spread of the data around the mean.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

population Variance

where:

$X_i$ represents the *ith* unit, starting from the first observation to the last
$\mu$ represents the population mean
$N$ represents the number of units in the population

The Variance of a sample

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

where:

$x_i$ represents the *ith* unit, starting from the first observation to the last
$\bar{x}$ represents the sample mean
$n$ represents the number of units in the sample

The standard deviation is the square root of the variance. The standard deviation for a population is represented by $\sigma$, and the standard deviation for a sample is represented by $s$.

Dataset A

4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Dataset B

1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

Calculate the population mean ($\mu$) of Dataset A.

(4 + 5 + 5 + 5 + 6 + 6 + 6 + 6 + 7 + 7 + 7 + 8) / 12

mean ($\mu$) = 6

Calculate the population mean ($\mu$) of Dataset B.

(1 + 2 + 3 + 4 + 5 + 6 + 6 + 7 + 8 + 9 + 10 + 11) / 12

mean ($\mu$) = 6

Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset

$X_i - \mu$ = -2, -1, -1, -1, 0, 0, 0, 0, 1, 1, 1, 2

$X_i - \mu$ = -5, -4, -3, -2, -1, 0, 0, 1, 2, 3, 4, 5,

Square each individual deviation value $(X_i - \mu)^2$

Calculate the mean of the squared deviation values $\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$

(4 + 1 +1 +1 + 0 + 0 + 0 + 0 +1 +1 +1 + 4) / 12

Variance $\sigma^2$ = 1.17

(25 + 16 + 9 + 4 + 1 + 0 + 0 + 1 + 4 + 9 + 16 + 25) / 12

Variance $\sigma^2$ = 9.17

Calculate the square root of the variance σ = 1.08

σ = 3.03

The larger Variance and Standard Deviation⇒ more dispersion

The variance and the standard deviation are measures of the spread of the data around the mean.

They summarise how close each observed data value is to the mean value.

In datasets with a small spread all values are very close to the mean, resulting in a small variance and standard deviation.

Where a dataset is more dispersed, values are spread further away from the mean, leading to a larger variance and standard deviation.

if all values of a dataset are the same, the standard deviation and variance are zero.

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a sample of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; Calculate Standard Deviation

$$\overline{x} = 9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)20 = 10.525$$

The average age is $10.53$ years, rounded to two places.

| Data | Freq. | Deviations | $Deviations^2$ | (Freq.)( $Deviations^2$) |
|---|---|---|---|---|
| $x$ | $f$ | $(x - \overline{x})$ | $(x - \overline{x})^2$ | $(f)(x - \overline{x})^2$ |

| Data | Freq. | Deviations | $Deviations^2$ | (Freq.)( $Deviations^2$) |
|---|---|---|---|---|
| $x$ | $f$ | $(x-\bar{x})$ | $(x-\bar{x})^2$ | $(f)(x-\bar{x})^2$ |
| 9 | 1 | $9-10.525=-1.525$ | $(-1.525)^2=2.325625$ | $1\times 2.325625=2.325625$ |
| 9.5 | 2 | $9.5-10.525=-1.025$ | $(-1.025)^2=1.050625$ | $2\times 1.050625=2.101250$ |
| 10 | 4 | $10-10.525=-0.525$ | $(-0.525)^2=0.275625$ | $4\times 0.275625=1.1025$ |
| 10.5 | 4 | $10.5-10.525=-0.025$ | $(-0.025)^2=0.000625$ | $4\times 0.000625=0.0025$ |
| 11 | 6 | $11-10.525=0.475$ | $(0.475)^2=0.225625$ | $6\times 0.225625=1.35375$ |
| 11.5 | 3 | $11.5-10.525=0.975$ | $(0.975)^2=0.950625$ | $3\times 0.950625=2.851875$ |
| | | | | The total is $9.7375$ |

The sample variance, $s^2$, is equal to the sum of the last column $(9.7375)$ divided by the total number of data values minus one $(20-1)$:

$s^2=\frac{9.7375}{20-1}=0.5125$

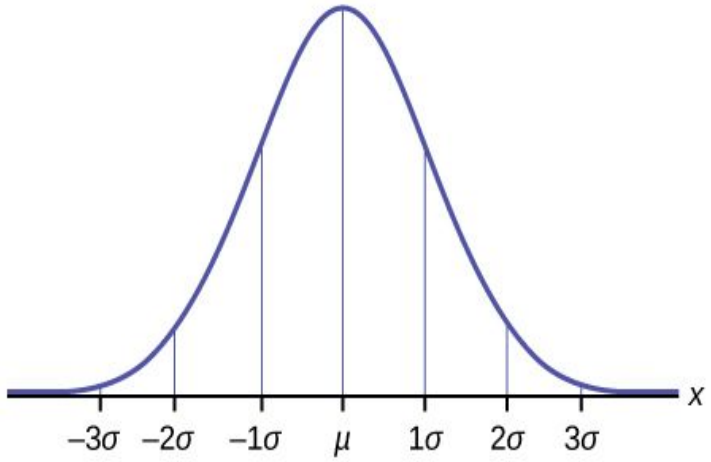
The **sample standard deviation** $s$ is equal to the square root of the sample variance: $s=\sqrt{0.5125}=0.715891$ which is rounded to two decimal places, $s$ = 0.72.

The standard deviation of a normal distribution enables us to calculate confidence intervals.

The normal distribution is a continuous probability distribution that is **symmetrical around its mean**,

In a normal distribution, about 68% of the values are within one standard deviation [–1σ and +1σ of the mean μ] either side of the mean

about 95% of the scores are within two standard deviations [–2σ and +2σ of the mean μ] of the mean.



The key difference between **percentage and percentile** is

the percentage is a mathematical value presented out of 100

[(Given Value)/(Total value)] × 100

**percentile** is the percent of values below a specific value.

The percentage is a means of comparing quantities.

A percentile is used to display position or rank.

It has quartiles.

Percentile(x) = (Number of values fall under ‘x’/total number of values) × 100

$P = (n/N) \times 100$

P is percentile
n – Number of values below ‘x’
N – Total count of population

# How to calculate Percentile?

Let's find the 90th percentile

| Sorted Marks | |
|---|---|
| 43 | 75 |
| 45 | 77 |
| 45 | 78 |
| 50 | 81 |
| 50 | 87 |
| 53 | 89 |
| 58 | 92 |
| 66 | 94 |
| 69 | 94 |
| 73 | 97 |

$$P_{90} = \frac{90(20 + 1)}{100}$$

$$P_{90} = \frac{1890}{100}$$

$$P_{90} = 18.9 \sim 19$$

$$P_{90} = 94$$

# How to calculate Percentile?

Let's find the 90th percentile

| Sorted Marks | |
|---|---|
| 43 | 75 |
| 45 | 77 |
| 45 | 78 |
| 50 | 81 |
| 50 | 87 |
| 53 | 89 |
| 58 | 92 |
| 66 | 94 |
| 69 | 94 |
| 73 | 97 |

$$P_{90} = 18.9 \sim 19$$

$$P_{90} = 94$$

$P_{90}$ = 94 means that 90% of students got less than 94 and 10% of students got more than 94

# How to calculate Percentile?

Let's find the percentile for marks 78

| Sorted Marks | |
|---|---|
| 43 | 75 |
| 45 | 77 |
| 45 | 78 |
| 50 | 81 |
| 50 | 87 |
| 53 | 89 |
| 58 | 92 |
| 66 | 94 |
| 69 | 94 |
| 73 | 97 |

$$P = \frac{n}{N} * 100$$

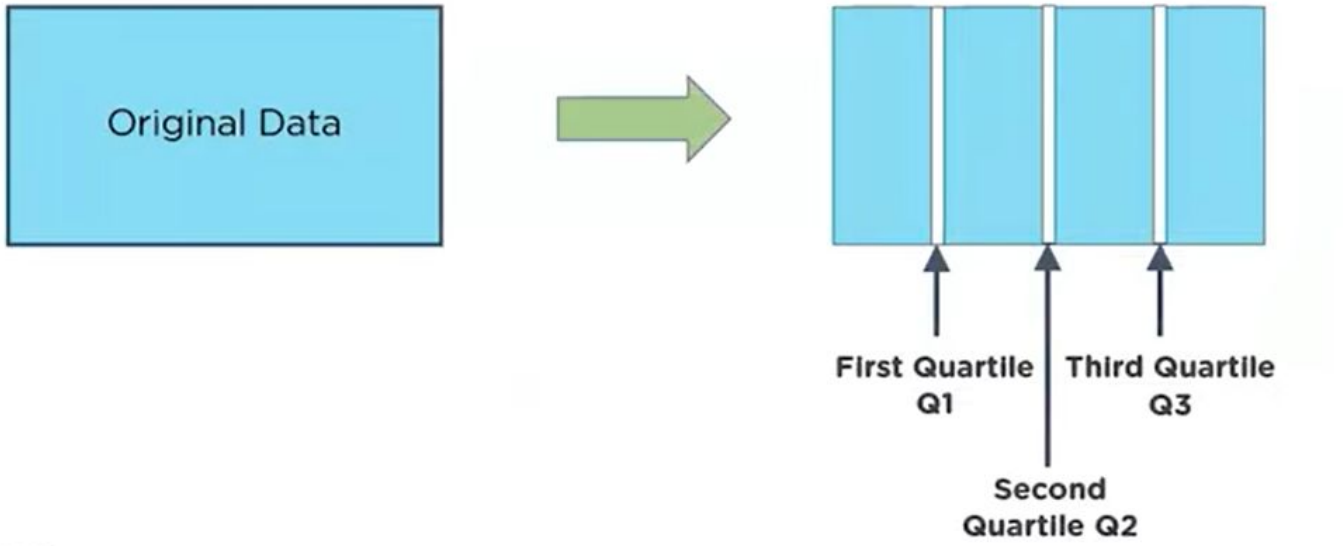
n = Ordinal rank of values
N = Total values in the dataset

$$P = \frac{12 * 100}{20}$$

$$P = 60$$

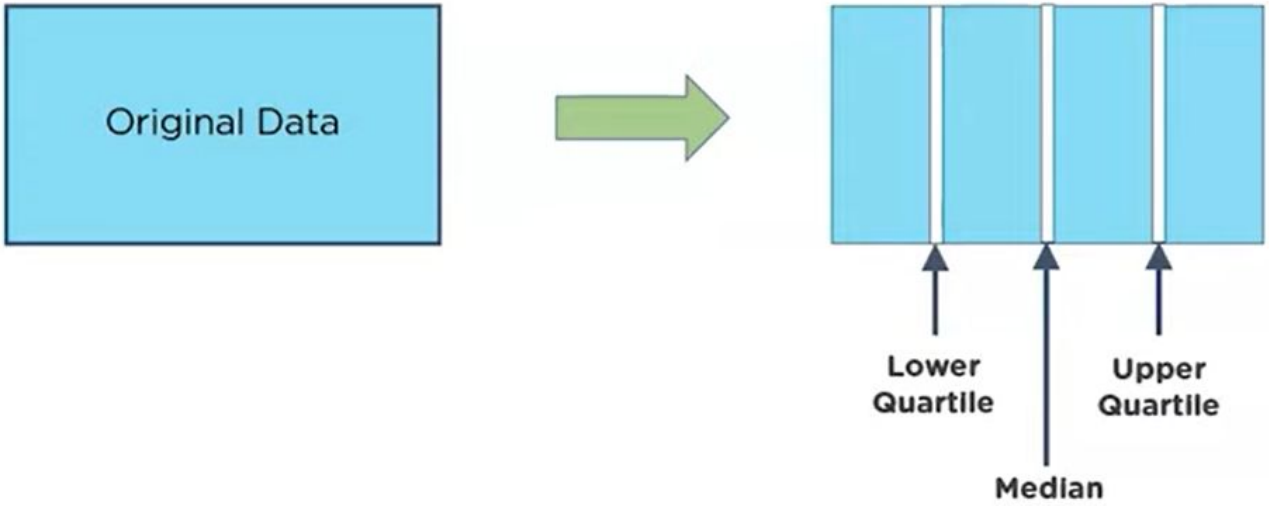
# What are Quartiles?

A quartile divides a sorted data set into 4 equal parts, so that each part represents ¼ of the data set

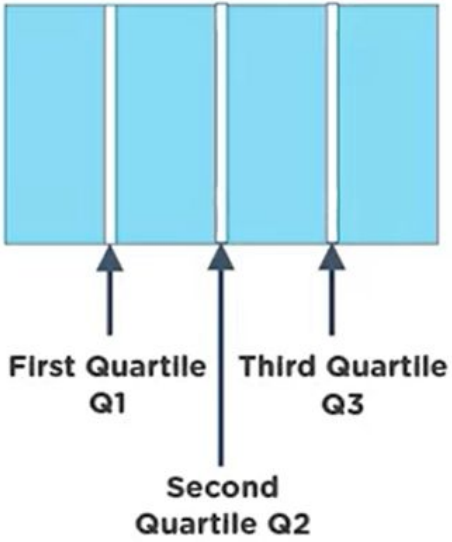


# What are Quartiles?

A quartile divides a sorted data set into 4 equal parts, so that each part represents ¼ of the data set



# What are Quartiles?



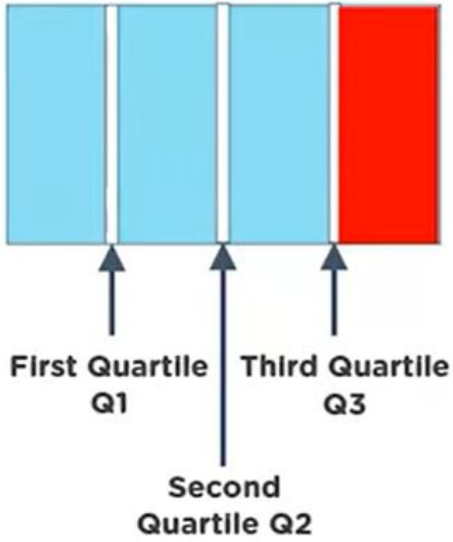

25% of all the data has a value ≤ Q1

50% of all the data has a value ≤ M

75% of all the data has a value ≤ Q2

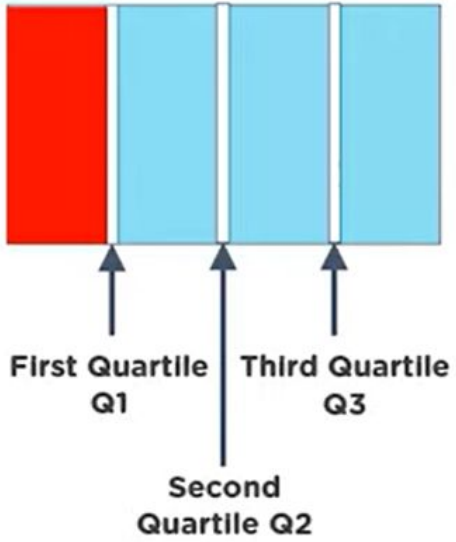
50% of all the data lies between Q1 and Q3

# What are Quartiles?




If a measurement falls to the right of the upper quartile of a set of data, then we know that it is in the top 25% of the data

# What are Quartiles?




If a measurement falls to the left of the lower quartile of a set of data, then we know that it is in the bottom 25% of the data

# How to calculate Quartiles?

To find $Q_1$, $Q_2$ and $Q_3$, we need to first sort the da
ascending order

| Petal Lengths in cm |
|---|
| 1.1 |
| 2.0 |
| 3.3 |
| 5.6 |
| 1.5 |
| 2.7 |
| 4.5 |
| 4.0 |
| 3.8 |
| 5.1 |
| 3.8 |
| 2.9 |
| 1.3 |
| 2.6 |
| 4.2 |

sort

| Petal Lengths in cm |
|---|
| 1.1 |
| 1.3 |
| 1.5 |
| 2.0 |
| 2.6 |
| 2.7 |
| 2.9 |
| 3.3 |
| 3.8 |
| 3.8 |
| 4.0 |
| 4.2 |
| 4.5 |
| 5.1 |
| 5.6 |

# How to calculate Quartiles?

Find the Median = Middle value in the dataset, whi
the second quartile

| Petal Lengths in cm |
|---|
| 1.1 |
| 1.3 |
| 1.5 |
| 2.0 |
| 2.6 |
| 2.7 |
| 2.9 |
| 3.3 |
| 3.8 |
| 3.8 |
| 4.0 |
| 4.2 |
| 4.5 |
| 5.1 |
| 5.6 |

n = 15

Since n is odd,
median = $(n + 1)/2$

Median = 16/2 = 8

$Q_2$ is the 8th
observation = 3.3

# How to calculate Quartiles?

Find the First Quartile value = $Q_1$

| Petal Lengths in cm |
|---|
| 1.1 |
| 1.3 |
| 1.5 |
| 2.0 |
| 2.6 |
| 2.7 |
| 2.9 |
| 3.3 |
| 3.8 |
| 3.8 |
| 4.0 |
| 4.2 |
| 4.5 |
| 5.1 |
| 5.6 |

$Q_1 = (n + 1) \times (1/4)$

$Q_1 = 16/4 = 4$

$Q_1$ is the $4^{th}$ observation = 2.0

# How to calculate Quartiles?

Find the Third Quartile value = Q3

| Petal Lengths in cm |
|---|
| 1.1 |
| 1.3 |
| 1.5 |
| 2.0 |
| 2.6 |
| 2.7 |
| 2.9 |
| 3.3 |
| 3.8 |
| 3.8 |
| 4.0 |
| 4.2 |
| 4.5 |
| 5.1 |
| 5.6 |

$Q_3 = (n + 1) \times (3/4)$

$Q_3 = 16x(3/4) = 12$

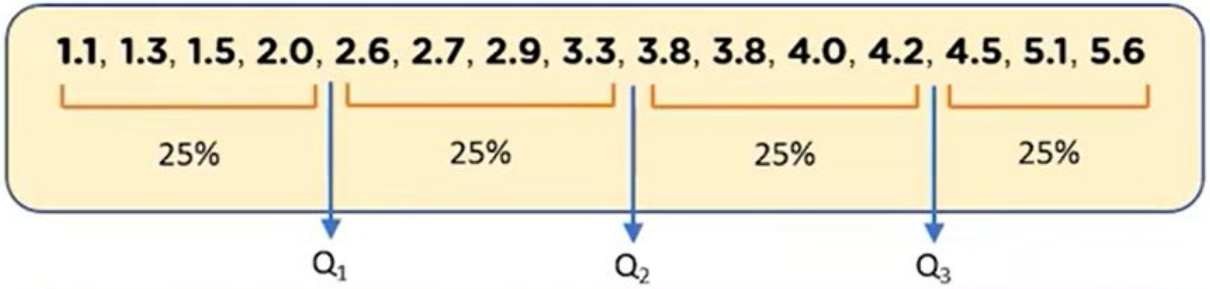
Q3 is the $12^{th}$ observation = 4.2

# How to calculate Quartiles?

Find the Third Quartile value = Q3

| Petal Lengths in cm | |
|---|---|
| 1.1 | |
| 1.3 | |
| 1.5 | |
| 2.0 | → $Q_1$ = First Quartile |
| 2.6 | |
| 2.7 | |
| 2.9 | |
| 3.3 | → $Q_2$ = Second Quartile |
| 3.8 | |
| 3.8 | |
| 4.0 | |
| 4.2 | → $Q_3$ = Third Quartile |
| 4.5 | |
| 5.1 | |
| 5.6 | |

# Quartile Summary




- ¼ or 25% of the flowers have a petal length that is less than or equal to 2.0 cm
- ½ or 50% of the flowers have a petal length that is less than or equal to 3.3 cm
- ¾ or 75% of the flowers have a petal length that is less than or equal to 4.2 cm
- ½ or 50% of the flowers have a petal length between 2.6 cm and 4.2 cm

# Quartiles and Interquartile Range

Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.

Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters.

A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).

Quartiles

25% of values **Q1** 25% of values **Q2** 25% of values **Q3** 25% of values

The lower quartile (Q1) is the point between the lowest 25% of values and the highest 75% of values. It is also called the 25th percentile.

The second quartile (Q2) is the middle of the data set. It is also called the 50th percentile, or the median.

The upper quartile (Q3) is the point between the lowest 75% and highest 25% of values. It is also called the 75th percentile.

If we have even no of scores say 100 scores

The first quartile (Q1) lies between the **25th and 26th** observation,

the second quartile (Q2) between the **50th and 51st** observation,

the third quartile (Q3) between the **75th and 76th** observation

if we had an odd number of scores (say, 99 observations), we would only need to take one score for each quartile (that is, the 25th, 50th and 75th scores)

## Quartile Formula

The Quartile Formula for Q1 = $\frac{1}{4}(n+1)^{th}$ term

The Quartile Formula for Q3 = $\frac{3}{4}(n+1)^{th}$ term

The Quartile Formula for Q2 = Q3 – Q1 (Equivalent to Median)

**Quartiles** are a useful measure of spread because they are much **less affected by outliers** than the equivalent measures of mean and standard deviation.

For this reason, quartiles are often reported along with the median as the best choice of measure of spread and central tendency, respectively, when dealing with outlier.

A common way of expressing quartiles is as an interquartile range.

The interquartile range describes the difference between the third quartile (Q3) and the first quartile (Q1)

Interquartile range = Q3 - Q1

# A five-number summary

It is especially useful in descriptive analyses or during the preliminary investigation of a large data set.

A summary consists of five values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median.

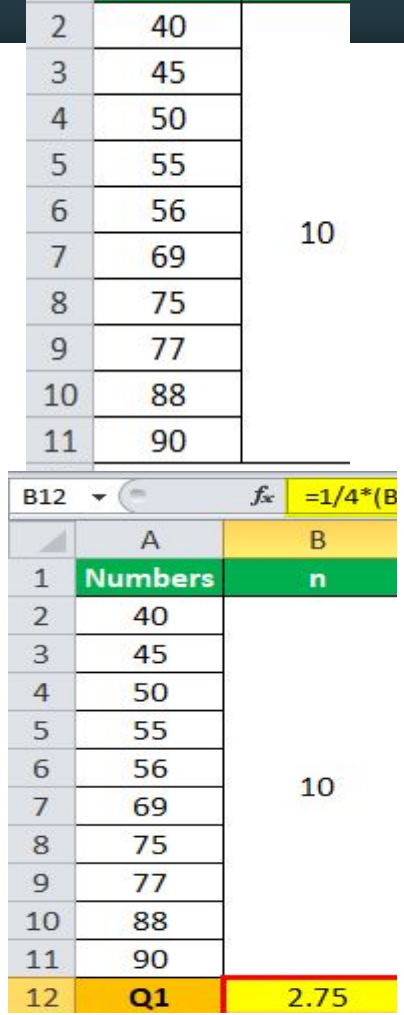
**Simple Ltd. is a clothing manufacturer working on a scheme to please their employees for their efforts. The management is in discussion to start a new initiative which states they want to divide their employees as per the following:**

- Top 25% lying above Q3- $25 per cloth
- Greater than the middle one but less than Q3 – $20 per cloth
- Greater than Q1 but less than Q2 – $18 per cloth
- The management has collected its average daily production data for the last 10 days per (average) employee.
- 55, 69, 88, 50, 77, 45, 40, 90, 75, 56.
- Use the quartile formula to build the reward structure.
- What rewards would an employee get if he produced 76 clothes ready

25% of values **Q1** 25% of values **Q2** 25% of values **Q3** 25% of values

for the calculation of quartile arrange dataset in order.

Q1 = ¼ (n+1)th term
= ¼ (10+1)
= ¼ (11)
Q1 = 2.75 th Term
Here, the average must be taken, which is of **2nd and 3rd terms**, which are 45 and 50. The average formula
of the same is (45+50)/2 = 47.50
The Q1 is 47.50, which is bottom 25%



| | | |
|---|---|---|
| 2 | 40 | |
| 3 | 45 | |
| 4 | 50 | |
| 5 | 55 | |
| 6 | 56 | 10 |
| 7 | 69 | |
| 8 | 75 | |
| 9 | 77 | |
| 10 | 88 | |
| 11 | 90 | |

B12 =1/4*(B

| | A | B |
|---|---|---|
| 1 | Numbers | n |
| 2 | 40 | |
| 3 | 45 | |
| 4 | 50 | |
| 5 | 55 | |
| 6 | 56 | 10 |
| 7 | 69 | |
| 8 | 75 | |
| 9 | 77 | |
| 10 | 88 | |
| 11 | 90 | |
| 12 | Q1 | 2.75 |

Q3 = ¾ (n+1)th term

= ¾ (11)

**Q3 = 8.25 Term**

Here, the average needs to be taken, which is of 8th and 9th terms, which are 88 and 90. The average of the same is (88+90)/2 = 89.00.

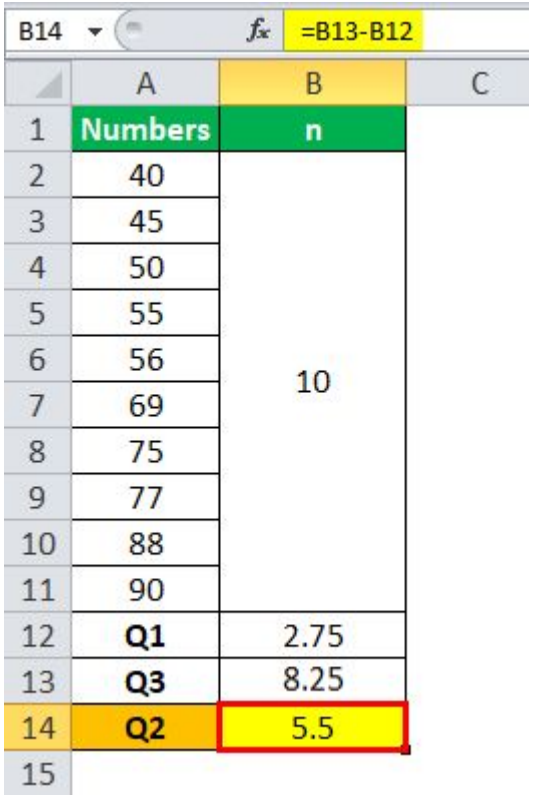
The Q3 is 89, which is the top 25%.

The Median Value (Q2) = 8.25 – 2.75

**Median or Q2= 5.5 Term**

Here, the average needs to be taken, which is of 5th and 6th 56 and 69. The average of the same is (56+69)/2 = 62.5.

The Q2 or median is 62.5



**The Reward Range would be:**

47.50 – 62.50 will get $18 per cloth

>62.50 – 89 will get $20 per cloth

>89.00 will get $25 per cloth

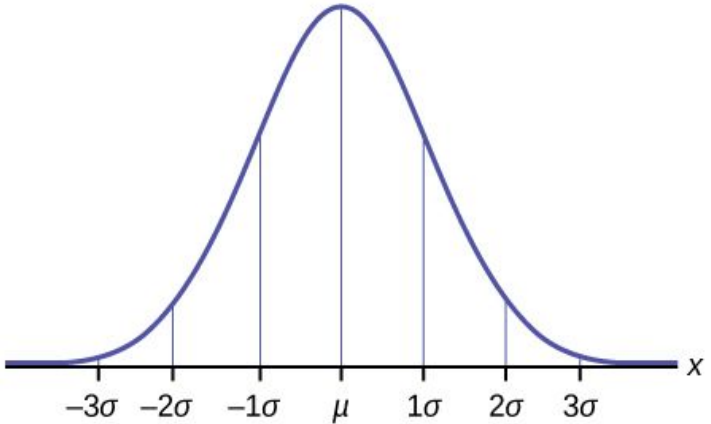
If an employee produces 76, he will lie above Q1. Hence, would be eligible for a $20 bonus.

What Is **Symmetrical Distribution?**

A symmetrical distribution occurs when **the values of variables appear at regular frequencies** and often the **mean, median, and mode all occur at the same point.**

If a line were drawn dissecting the middle of the graph, it would reveal two sides that mirror one other.

In graphical form, symmetrical distributions may appear as a normal distribution (i.e., bell curve).



# Skewness

Data skewness is one of the important challenges that data scientists often face in real-time case studies.

**Definition**

**Skewness** is the measure of symmetry or asymmetry of data distribution.

A distribution or data set is said to be symmetric if it looks the same to the left and right points of the center.

A distribution is said to be 'skewed' when the **mean and the median** fall at **different points in the distribution**, and the balance (or centre of gravity) is shifted to one side or the other-to left or right.

Measures of skewness tell us the direction and the extent of Skewness.

In symmetrical distribution the mean, median and mode are identical.

The more the mean moves away from the mode, the larger the asymmetry or skewness

Types of skewness

Skewness is generally classified into 2 broad categories-

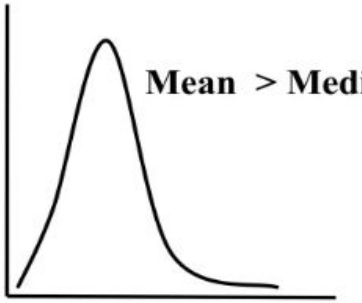
· Right skewness or Positive skewness

· Left skewness or Negative skewness

‘L’ shaped positively skewed

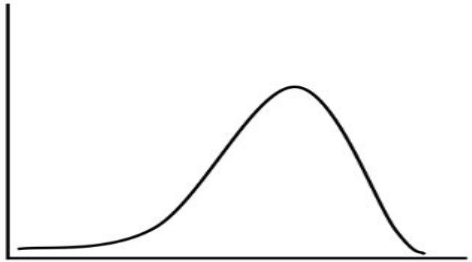
‘J’ shaped negatively skewed

## Positively skewed



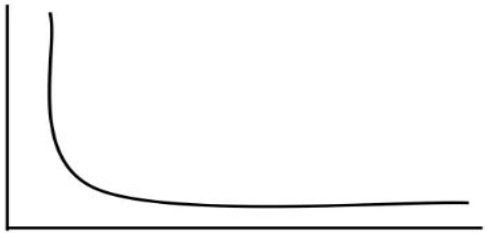
Mean > Median > Mode

## Negatively skewed



Mean < Median < Mode

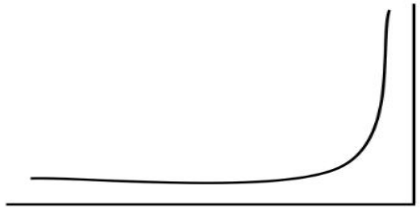
## ‘L’ Shaped Positively skewed



Mean < Mode

Mean < Median

## ‘J’ Shaped Negatively Skewed



Mean > Mode

Mean > Median

# Right skewness

A right-skewed distribution will have a long tail in the right direction on the number line


the tail is lengthier and flattened towards the right side of the distribution; it is often called right-skewed distribution.

For a right-skewed distribution,

Mean >= Median >= Mode

Here, Mean will be getting overrated because the tail is getting more flat and long at the right side

It is very easy to calculate the type of skewness in MS Excel through a metric called “skew”. his metric will always produce a positive number as output.

| ID | Marks |
|---|---|
| 1 | 20 |
| 2 | 30 |
| 3 | 60 |
| 4 | 60 |
| 5 | 50 |
| 6 | 50 |
| 7 | 80 |
| 8 | 40 |
| 9 | 10 |
| 10 | 100 |
| Skew | 0.377667 |

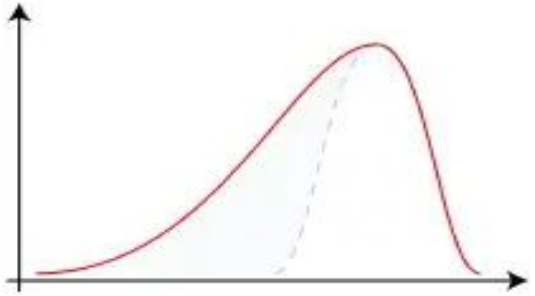
# Left skewness

A left-skewed distribution will have a long tail in the left direction on the number line such that the mean of the all data points will eventually go down.

For a left skewed distribution

Mode >= Median >=Mean

Here, Mean will be getting underrated because the tail is getting more flat and long on the left side



Negative Skew

Contrary to the right-skewed distribution, here the skew metric will be negative. This is because the tail of the distribution is flattened along the left side of the distribution

For a **perfect normal distribution,** skewness will be zero.

Here, Mean will be equal to median and mode.

Which means Mean=Median=Mode.

| ID | Marks |
|---|---|
| 1 | 20 |
| 2 | 30 |
| 3 | 60 |
| 4 | 60 |
| 5 | 50 |
| 6 | 50 |
| 7 | 80 |
| 8 | 95 |
| 9 | 100 |
| 10 | 100 |
| Skew | -0.08313 |

# Statistical Measures of Skewness

| Absolute Measures of Skewness | Relative Measures of Skewness |
| --- | --- |
| Following are the absolute measures of skewness:<br><br>• Skewness (Sk) = Mean – Median<br><br>• Skewness (Sk) = Mean – Mode<br><br>• Skewness (Sk) = (Q3 - Q2) - (Q2 - Q1) | There are four measures of skewness:<br><br>• β and γ Coefficient of skewness<br><br>• Karl Pearson's Coefficient of skewness<br><br>• Bowley's Coefficient of skewness<br><br>• Kelly's Coefficient of skewness |

## Karl Pearson Coefficient of Skewness

Pearson's coefficient of skewness is a method developed by Karl Pearson to find skewness in a sample using **descriptive statistics like the mean and mode.**

There are two different ways of calculating skewness through using Karl Pearson's Skewness formula.

1- Karl Pearson's Skewness formula through the usage of mode

**SKp= (Mean−Mode)/ S.D = (μ-Z)/σ**

2- Karl Pearson's Skewness formula through the usage of the median

**SKp= 3(Mean–Median)/ S.D =3(μ-M)/σ**

The value of coefficient of skewness is **zero**, when the distribution is **symmetrical**.

The value of coefficient of skewness is **positive**, when the distribution is **positively skewed**.

The value of coefficient of skewness is **negative**, when the distribution is **negatively skewed**.

From the marks secured by 120 students in Section A and B of a class, the Following measures are obtained:

Section A:X= 46.83; S.D = 14.8; Mode = 51.67

Section B:X= 47.83; S.D = 14.8; Mode = 47.07

Determine which distribution of marks is more skewed.

$$\text{For Section A: } Sk_p = \frac{\overline{X} - Z}{\sigma} = \frac{46.83 - 51.67}{14.8} = \frac{-4.84}{14.8} = -0.3270$$

$$\text{For Section B: } Sk_p = \frac{\overline{X} - Z}{\sigma} = \frac{47.83 - 47.07}{14.8} = \frac{0.76}{14.8} = 0.05135$$

Marks of Section A is more Skewed. But marks of Section A is negatively Skewed.

Marks of Section B are Positively skewed.

From a moderately skewed distribution of retail prices for men's shoes, it is found that the mean price is Rs. 20 and the median price is Rs. 17. If the coefficient of variation is 20%, find the Pearsonian coefficient of skewness of the distribution.

$$C.V. = \frac{\sigma}{\bar{X}} X100$$

Given: C.V. = 20, X= 20, M = 17

$$20 = \frac{\sigma}{20} X100$$

$$\sigma = 20 \text{ X } 20 / 100 = 400/100 = 4$$

$$Sk_p = \frac{3(\bar{X} - M)}{\sigma} = \frac{3(20-17)}{4} = 9/4 = 2.25$$

# 3. Calculate Karl Pearson's coefficient of Skewness for the following data.

| X | $X^2$ |
|---|---|
| 25 | 625 |
| 15 | 225 |
| 23 | 529 |
| 40 | 1600 |
| 27 | 729 |
| 25 | 625 |
| 23 | 529 |
| 25 | 625 |
| 20 | 400 |
| $\sum X = 223$ | $\sum X^2 = 5887$ |

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \left[\frac{\sum X}{N}\right]^2}$$

$$\overline{X} = \frac{\sum X}{N}$$

$$= \frac{223}{9} = 24.78$$

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \left[\frac{\sum X}{N}\right]^2} = \sqrt{\frac{5887}{9} - (24.78)^2}$$

$$= \sqrt{654.1111 - 614.0484} = \sqrt{40.06} = 6.33$$

$$Z = 25$$

$$Sk_p = \frac{\overline{X} - Z}{\sigma} = \frac{24.78 - 25}{6.33} = \frac{-0.22}{6.33} = -0.03$$

# Calculate Karl Pearson's coefficient of Skewness for the following data.

| Wage per Item Rs.(x) | Number of items f | fx | $x^2$ | $fx^2$ |
|---|---|---|---|---|
| 12 | 10 | 120 | 144 | 1440 |
| 15 | 25 | 375 | 225 | 5625 |
| 20 | 40 | 800 | 400 | 16000 |
| 25 | 70 | 1750 | 625 | 43750 |
| 30 | 32 | 960 | 900 | 28800 |
| 40 | 13 | 520 | 1600 | 20800 |
| 50 | 10 | 500 | 2500 | 25000 |
| | $\sum f =$ | $\sum fx =$ | | $\sum fX^2 =$ |

$$\overline{X} = \frac{\sum fX}{\sum f} = \frac{5025}{200} = 25.13$$

$$\sigma = \sqrt{\frac{\sum fX^2}{\sum f} - \left[\frac{\sum fX}{\sum f}\right]^2}$$

$$= \sqrt{\frac{141415}{200} - (25.13)^2} = \sqrt{707.075 - 631.5169}$$

$$= \sqrt{75.5581} = 8.69$$

Greatest frequency = 70, $Z = 25$

$$Sk_p = \frac{\overline{X} - Z}{\sigma} = \frac{25.13 - 25}{8.69} = 0.13/8.69 = 0.0149$$

5. Calculate Karl Pearson's coefficient of Skewness for the following data.

| Profit (Rs.Lakhs) | No of Companies f | m | fm | $m^2$ | $fm^2$ |
|---|---|---|---|---|---|
| 10-20 | 18 | 15 | 270 | 225 | 4050 |
| 20-30 | $20 = f_0$ | 25 | 500 | 625 | 12500 |
| 30-40 | $30 = f_1$ | 35 | 1050 | 1225 | 36750 |
| 40-50 | $22 = f_2$ | 45 | 990 | 2025 | 44550 |
| 50-60 | 10 | 55 | 550 | 3025 | 30250 |

Bowley‘s Coefficient

This measure is based on **quartiles**.

For a symmetrical distribution, it is seen that Q1, and Q3 are equidistant from median (Q2).

Thus (Q3 − Q2) − (Q2 − Q1) can be taken as an absolute measure of skewness.

$$S_{kq} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

$$= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

# Properties of Bowley's coefficient of skewness

- $-1 \leq Skq \leq 1$.
- $Skq = 0 \Rightarrow$ distribution is symmetrical about mean.
- $Skq > 0 \Rightarrow$ distribution is skewed to the right.
- $Skq < 0 \Rightarrow$ distribution is skewed to the left.

Compare the Skewness of A and B using bowley's coefficient

| | Q1 | M | Q3 |
|---|---|---|---|
| Series A | 40 | 60 | 80 |
| Series B | 62.85 | 65.25 | 72.15 |

In series A there is no skewness, In Series B there is moderate positive skewness.

Series A

$$Sk_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{80 + 40 - 2(60)}{80 - 40} = \frac{120 - 120}{40} = 0$$

Series B

$$Sk_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{72.15 + 62.85 - 2(65.25)}{72.15 - 62.85} = \frac{135 - 130.5}{9.3} = 4.5/9.3 = 0.4839$$

# Calculate Bowley's coefficient of Skewness.

| No of child per family x | No of Families f | Cumulative Frequency cf |
|---|---|---|
| 0 | 7 | 7 |
| 1 | 10 | 17 |
| 2 | 16 | 33 |
| 3 | 25 | 58 |
| 4 | 18 | 76 |
| 5 | 11 | 87 |
| 6 | 8 | 95 |
| | $\sum f =$ | |

Position of Q1 = $\frac{\sum f+1}{4}$ th observation

= 95+1/4 = 96/4 =24 th observation

**Q1=2**

Position of Q3 = $3(\frac{\sum f+1}{4})$ th observation

= 3(24) = 72 th observation

**Q3=4**

Position M = $\frac{\sum f+1}{2}$ th observation

= = 95+1/2 = 96/2 = 48 th observation

**M(Q2)=3**

$$Sk_B = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{4+2-2(3)}{4-2} = \frac{6-6}{2} = 0$$

# Multivariate Exploration

It is the study of more than one attribute in the dataset simultaneously.

The real world example is the **weather**.

The weather at any particular place does not solely depend on the ongoing season, instead many other factors play their specific roles, like humidity, pollution, etc.

This technique is critical to understanding the relationship between the attributes, which is central to data science .

The aim is to find patterns and correlations between several variables simultaneously

allowing  a much deeper, more complex understanding of a given scenario

# Central Data Point

In any dataset, each data point can be expressed as a set of all the attributes

observation i: {Attribute 1, Attribute 2, Attribute 3….}

For example, observation one: {5.1, 3.5, 1.4, 0.2}.

This observation point can be expressed in four-dimensional Cartesian coordinates and can be plotted in a graph

In this way, if dataset has 150 observations then all 150 observations can be expressed in Cartesian coordinates.

If the objective is to find the most “typical” observation point, it would be a data point made up of the mean of each attribute in the dataset independently.

For the dataset, the central mean point is {5.006, 3.418, 1.464, 0.244}.

This data point may not be an actual observation. It will be a hypothetical data point with the most typical attribute values.

# Correlation

Correlation measures the statistical relationship between two attributes,particularly dependence of one attribute on another attribute.

When two attributes are highly correlated with each other, they **both vary at the same rate** with each other either in the same or in opposite directions.

Statistically, the two attributes that are correlated are dependent on each other and one may be used to **predict** the other.

the value of the correlation coefficient varies between +1 and -1

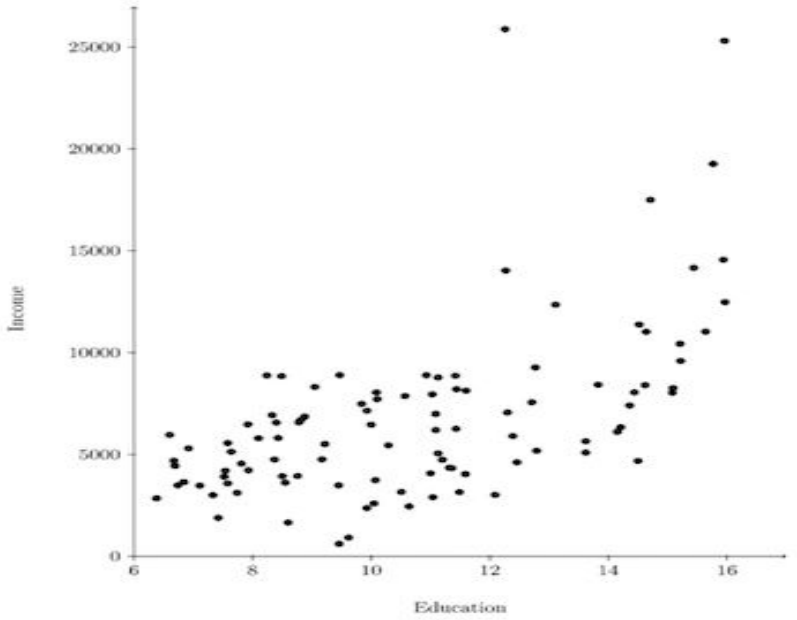
Correlation can be positive, negative, or no correlation.

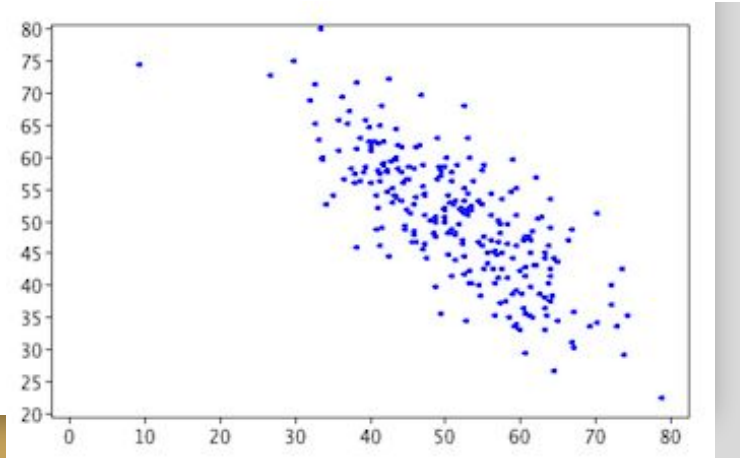
Positive correlation means that as one data set increases, the other data set increases as well

Typically, positively correlated data sets are seen as a line the goes up and to the right on a scatter plot.

Negative correlation means that as one data set increases, the other decreases.

Negatively correlated data sets are seen as a line the goes down and to the right on a scatter plot.

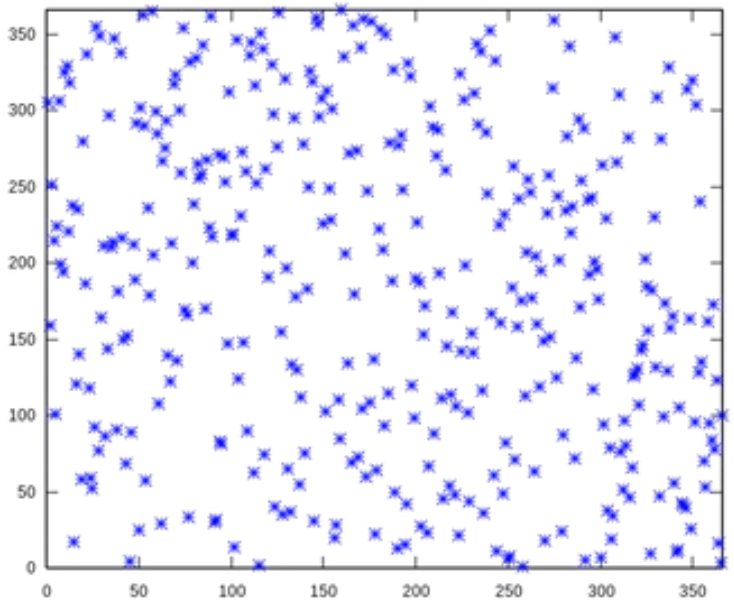


No correlation means that the two sets of data are not related at all.

In other words, this means that one set of data does not increase or decrease with the other.

No correlation is typically seen when the data points are very spread out

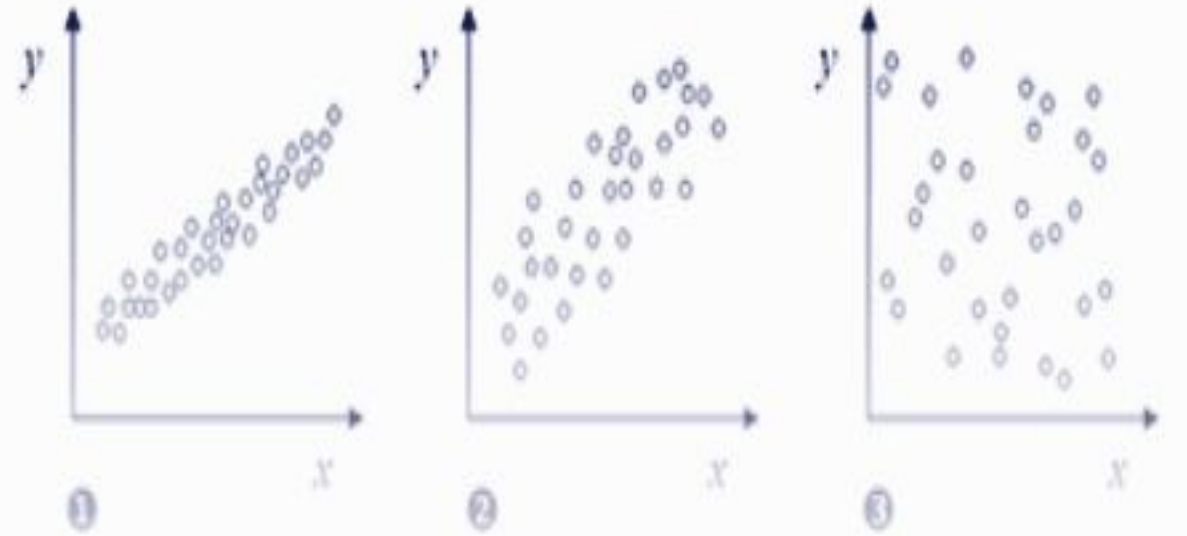


Positive and negative is not the only way to describe correlation; correlation can also be described by its strength.

Data sets can also have perfect correlation, strong correlation, or weak correlation.

The closer the data points are together and the more they form a straight line, the stronger the correlation.

The first graph has a strong positive relationship, while the second has a low or weak positive correlation. The third graph has no relationship or no correlation.



**four** types of correlations: Pearson correlation, Kendall rank correlation, Spearman correlation, and the Point-Biserial correlation

**Karl Pearson Correlation Coefficient for bivariate distribution**

measures the strength of linear dependence

1. A value closer to 1 or –1 indicates the two attributes are highly correlated, with perfect correlation at 1 or –1.

Perfect correlation also exists when the attributes are governed by **formulas and laws**.

For example, observations of the values of gravitational force and the mass of the object (Newton's second law) or the quantity of the products sold and total revenue (price*volume = revenue).

A correlation value of 0 means there is no linear relationship between two attributes.

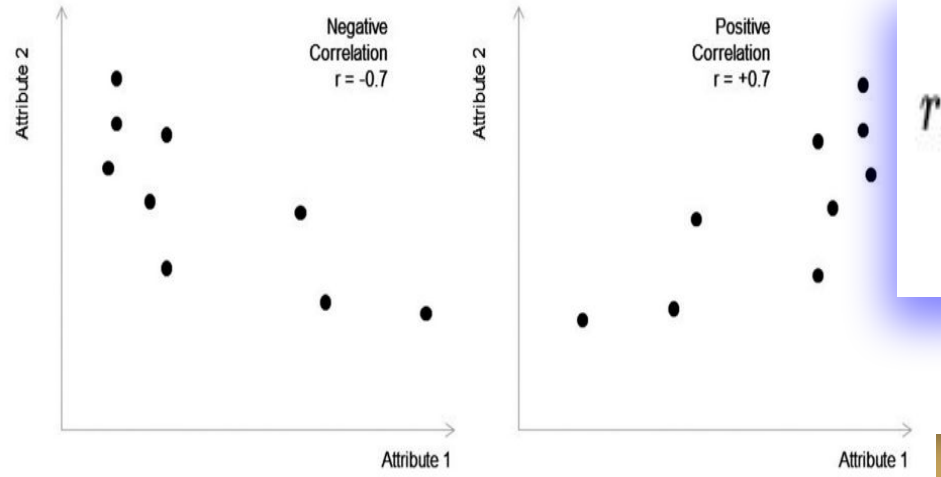
The Pearson correlation coefficient between two attributes x and y is calculated with the formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N \times s_x \times s_y}$$

where sx and sy are the standard deviations of random variables x and y, respectively

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$




$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

the value of the correlation coefficient varies between +1 and -1

| STRENGTH OF ASSOCIATION | NEGATIvE **r** | POSITIvE **r** |
|---|---|---|
| weak | -0.1 to -0.3 | 0.1 to 0.3 |
| average | -0.3 to -0.5 | 0.3 to 0.5 |
| strong | -0.5 to -1.0 | 0.5 to 1.0 |

Question: An experiment conducted on 9 different cigarette smoking subjects resulted in the following data –

$N = 9$

| Sub No | Cigarettes smoked per week (averaged over the last 5 years of their life) | Number of years lived |
|---|---|---|
| 1 | 25 | 63 |
| 2 | 35 | 68 |
| 3 | 10 | 72 |
| 4 | 40 | 62 |
| 5 | 85 | 65 |
| 6 | 75 | 46 |
| 7 | 60 | 51 |
| 8 | 45 | 60 |
| 9 | 50 | 55 |

| x | x2 | y | y2 | xy |
|---|---|---|---|---|
| 25 | 625 | 63 | 3969 | 1575 |
| 35 | 1225 | 68 | 4624 | 2380 |
| 10 | 100 | 72 | 5184 | 720 |
| 40 | 1600 | 62 | 3844 | 2480 |
| 85 | 7225 | 65 | 4225 | 5525 |
| 75 | 5625 | 46 | 2116 | 3450 |
| 60 | 3600 | 51 | 2601 | 3060 |
| 45 | 2025 | 60 | 3600 | 2700 |
| 50 | 2500 | 55 | 3136 | 2750 |
| $\Sigma xi = 425$ | $\Sigma xi^2 = 24525$ | $\Sigma yi = 542$ | $\Sigma yi^2 = 33188$ | $\Sigma xiyi = 24640$ |

$(\Sigma xi)^2 = 425^2 = 180625$

$(\Sigma yi)^2 = 542^2 = 293764$

$$r_{xy} = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{N\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{N\Sigma y_i^2 - (\Sigma y_i)^2}}$$

$$= \frac{9.24640 - 425.542}{\sqrt{9.24525 - 180625}\sqrt{9.33188 - 293764}}$$

$$= \frac{-8590}{\sqrt{40100}\sqrt{4928}}$$

$$= -0.61$$

This implies a negative correlation between the considered variables i.e. The higher the number of cigarettes smoked per week in last 5 years, the lesser the number of years lived.

**Compute Pearson's coefficient of correlation between advertisement cost and sales as per the data given below:**

| Advertisement Cost in 1000's | 39 | 65 | 62 | 90 | 82 | 75 | 25 | 98 | 36 | 78 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sales in lakhs | 47 | 53 | 58 | 86 | 62 | 68 | 60 | 91 | 51 | 84 |

**Correlation coefficient is positively correlated.**

**n = 10**

$$\sum x = 650 \quad \sum y = 660 \quad \sum xy = 45604 \quad \sum x^2 = 47648 \quad \sum y^2 = 45784$$

$$r_{xy} = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{N\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{N\Sigma y_i^2 - (\Sigma y_i)^2}}$$

**=0.7804**

Find the pearson's coefficient of correlation between price and demand from the following data.

| Price | 11 | 13 | 15 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|
| Demand | 30 | 29 | 24 | 24 | 21 | 18 | 15 |

$\sum x =$ **113** $\sum y =$ **161** $\sum xy =$ **2495** $\sum x^2 =$ **1889** $\sum y^2 =$ **3883**

Find the pearson's coefficient of correlation between the yield in grams(y) and the matured pods (x) of 10 groundnut plants

| X: | 14 | 34 | 20 | 16 | 11 | 11 | 20 | 17 | 22 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|
| Y: | 16 | 40 | 21 | 18 | 14 | 13 | 20 | 35 | 17 | 27 |

Q. Find the pearson's coefficient of correlation between price and demand from the following data.

| Price (x) | Demand (y) | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|
| 11 | 30 | 121 | 900 | 330 |
| 13 | 29 | 169 | 841 | 377 |
| 15 | 24 | 225 | 576 | 360 |
| 17 | 24 | 289 | 576 | 408 |
| 18 | 21 | 324 | 441 | 378 |
| 19 | 18 | 361 | 324 | 342 |
| 20 | 15 | 400 | 225 | 300 |
| $\Sigma$ 113 | 161 | 1889 | 3883 | 2495 |

$$r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{n(\Sigma x_i^2) - (\Sigma x_i)^2}\sqrt{n\Sigma(y_i^2) - (\Sigma y_i)^2}}$$

$$r_{xy} = \frac{7 \times 2495 - 113 \times 161}{\sqrt{9 \times 1889 - (113)^2}\sqrt{9 \times 3883 - (161)^2}}$$

$$= \frac{17465 - 18193}{65.05 \times 95} = -0.117$$

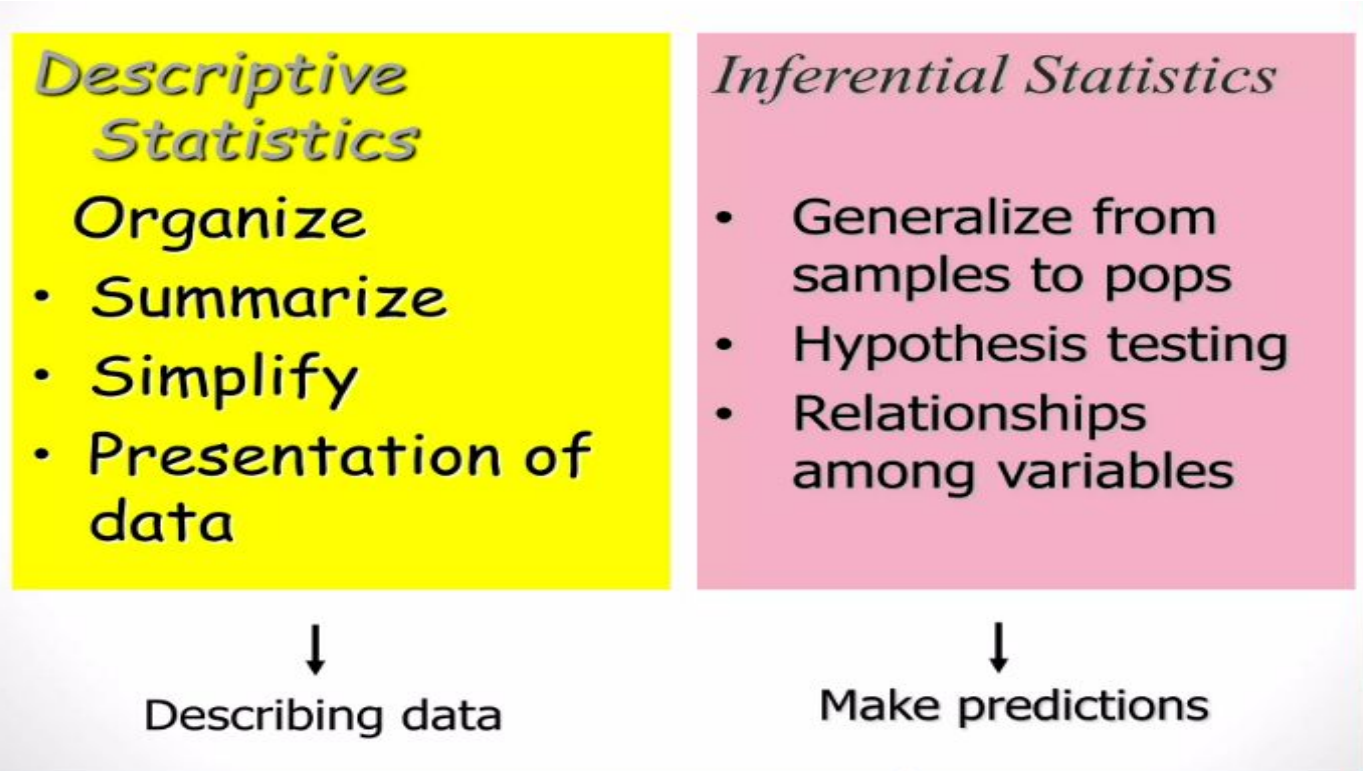
Q Find the pearson's coefficient of correlation between the yield in grams (y) and the matured pods (x) of 10 groundnut plants

| | x | y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|---|
| | 14 | 16 | 196 | 256 | 224 |
| | 34 | 40 | 1156 | 1600 | 1360 |
| | 20 | 21 | 400 | 441 | 420 |
| | 16 | 18 | 256 | 324 | 288 |
| | 11 | 14 | 121 | 196 | 154 |
| | 11 | 13 | 121 | 169 | 143 |
| | 20 | 20 | 400 | 400 | 400 |
| | 17 | 35 | 289 | 1225 | 595 |
| | 22 | 17 | 484 | 289 | 374 |
| | 17 | 27 | 289 | 729 | 459 |
| $\Sigma$ | 182 | 221 | 3712 | 5629 | 4417 |

$$r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{n(\Sigma x_i^2) - \Sigma(x_i)^2}\sqrt{n\Sigma(y_i^2) - (\Sigma y_i)^2}}$$

$$= \frac{10 \times 4417 - 182 \times 221}{\sqrt{10 \times 3712 - (182)^2}\sqrt{10 \times 5629 - (221)^2}} = 0.723$$

$$= \frac{3948}{63.214 \times 83.30} = 0723$$




https://www.youtube.com/watch?v=VHYOuWu9jQI

## Descriptive

1. Organizing and summarizing data using numbers & graphs

2. Data Summary:
Bar Graphs, Histograms, Pie Charts, etc.
Shape of graph & skewness

3. Measures of Central Tendency:
Mean, Median, & Mode

4. Measures of Variability:
Range, variance, & Standard deviation

## Inferential

1. Using sample data to make an inference or draw a conclusion of the population.

2. Uses probability to determine how confident we can be that the conclusions we make are correct. (Confidence Intervals & Margins of Error)

# Statistics

## Descriptive

1. Organizing and summarizing data using numbers & graphs

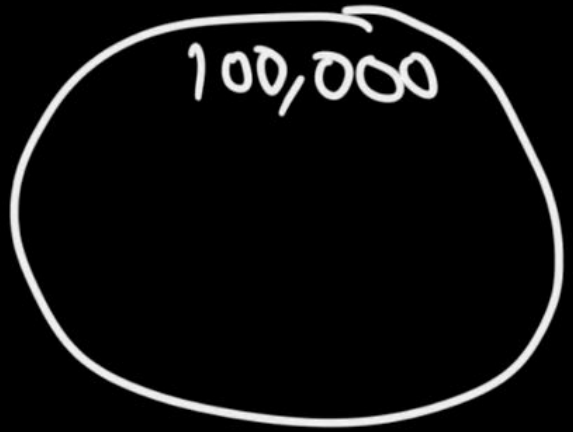
2. Data Summary:
Bar Graphs, Histograms, Pie Charts, etc.
Shape of graph & skewness

3. Measures of Central Tendency:
Mean, Median, & Mode

4. Measures of Variability:
Range, variance, & Standard deviation

blue cars?

xyz




## Inferential

1. Using sample data to make an inference or draw a conclusion of the population.

2. Uses probability to determine how confident we can be that the conclusions we make are correct. (Confidence Intervals & Margins of Error)

## Descriptive

1. Organizing and summarizing data using numbers & graphs

2. Data Summary:
Bar Graphs, Histograms, Pie Charts, etc.
Shape of graph & skewness

3. Measures of Central Tendency:
Mean, Median, & Mode

4. Measures of Variability:
Range, variance, & Standard deviation

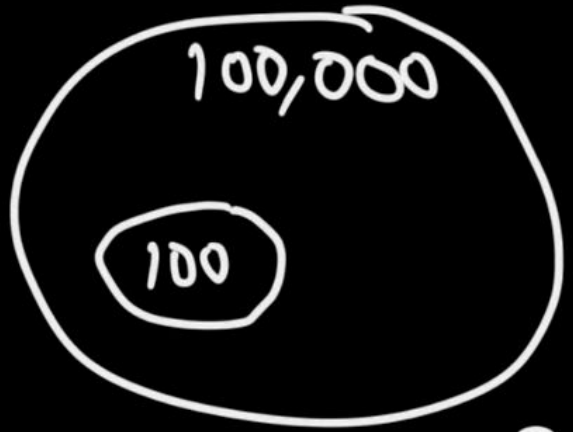
blue cars?

## Inferential

1. Using sample data to make an inference or draw a conclusion of the population.

2. Uses probability to determine how confident we can be that the conclusions we make are correct. (Confidence Intervals & Margins of Error)

xyz




20/100

20%

20% ± 2%.

# Statistics

## Descriptive

1. Organizing and summarizing data using numbers & graphs

2. Data Summary:
Bar Graphs, Histograms, Pie Charts, etc.
Shape of graph & skewness

3. Measures of Central Tendency:
Mean, Median, & Mode

4. Measures of Variability:
Range, variance, & Standard deviation

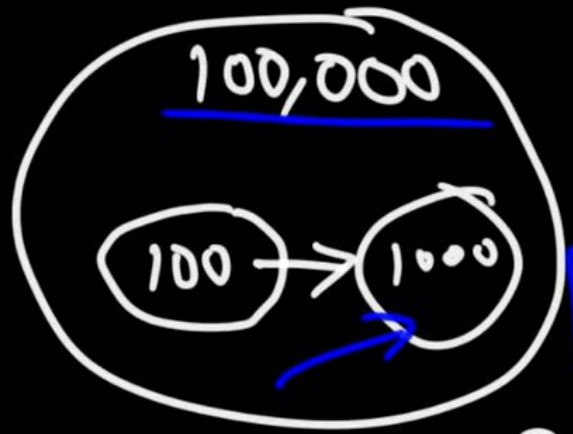
blue cars ?

95%

## Inferential

1. Using sample data to make an inference or draw a conclusion of the population.

2. Uses probability to determine how confident we can be that the conclusions we make are correct. (Confidence Intervals & Margins of Error)



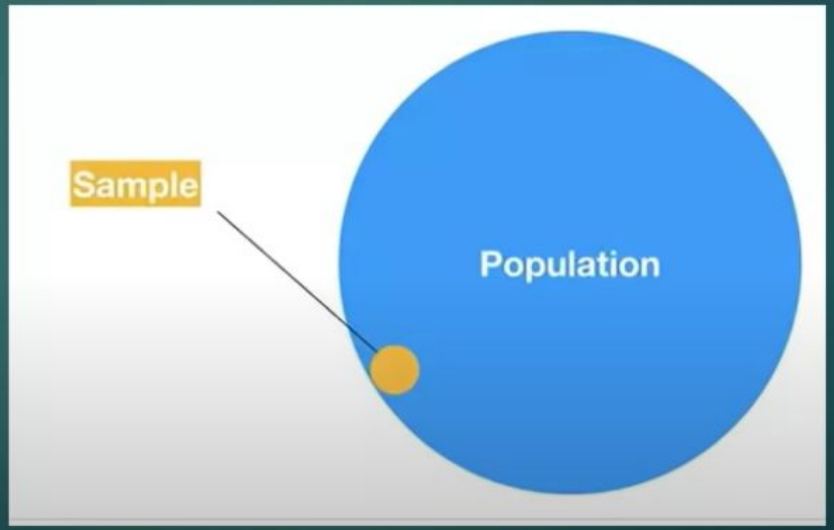

20/100

20%

21% ± 1%

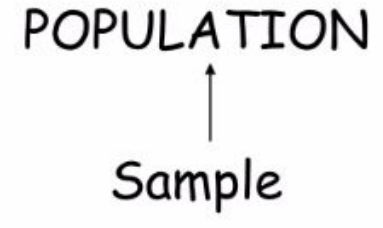
20% ± 2%

# Why do researchers use confidence intervals?



# Inferential Statistics

Used to draw conclusions about populations by examining the sample




Techniques that allow us to study samples and then make generalizations about the population.

Inferential statistics are a very crucial part of scientific research in that these techniques are used to test hypothesis

Inferential statistics is the act of generalizing from the data ("sample") to a larger phenomenon ("population") with calculated degree of certainty.

- The act of generalizing and deriving statistical judgments is the process of inference.

Inferential statistics is a statistical method that deduces from a small but representative sample the characteristics of a bigger population.

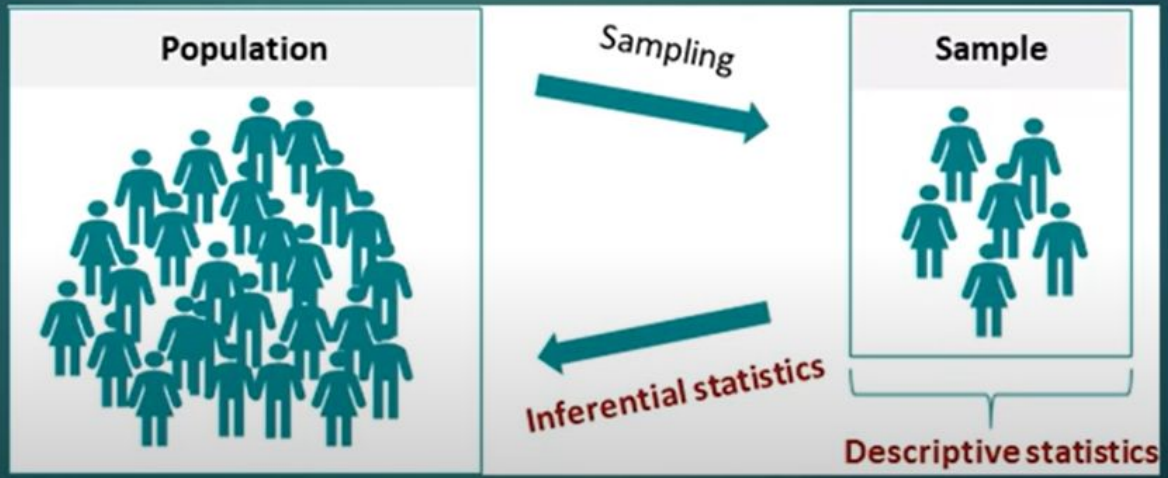
In other words, it allows the researcher to make **assumptions** about a wider group, using a smaller portion of that group as a guideline.

Inferential Statistics makes inference and prediction about population based on a sample of data taken from population.

# PURPOSE OF INFERENTIAL STATISTICS

- To determine difference between experimental and control group in experimental research.
- To enable researcher to evaluate effects of an independent variable on dependent variable.
- To determine whether the findings from the sample can generalize or be applied to the entire population
- To estimate differences in scores between groups in a research study.

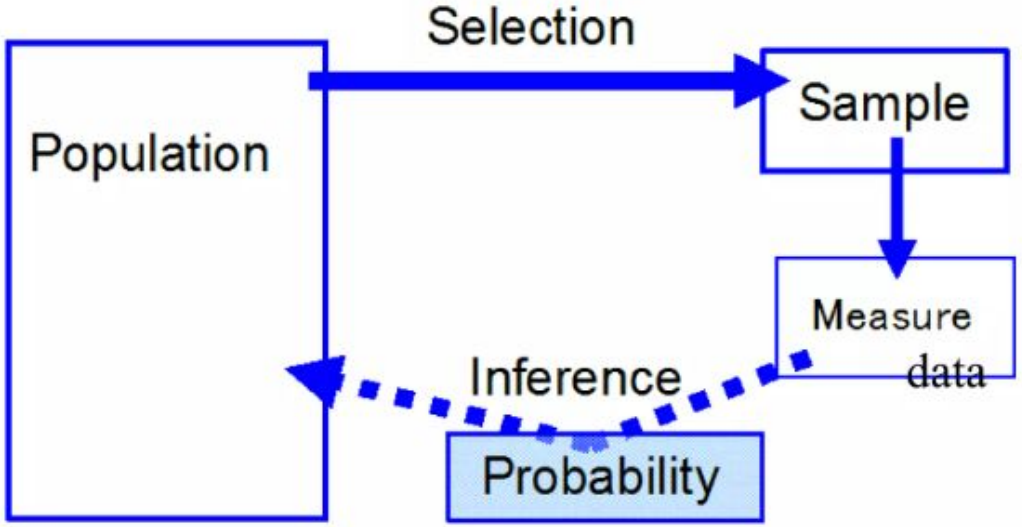
Inferential statistics are concerned with populations, and use sample data to make an inference about the population or to test hypotheses.



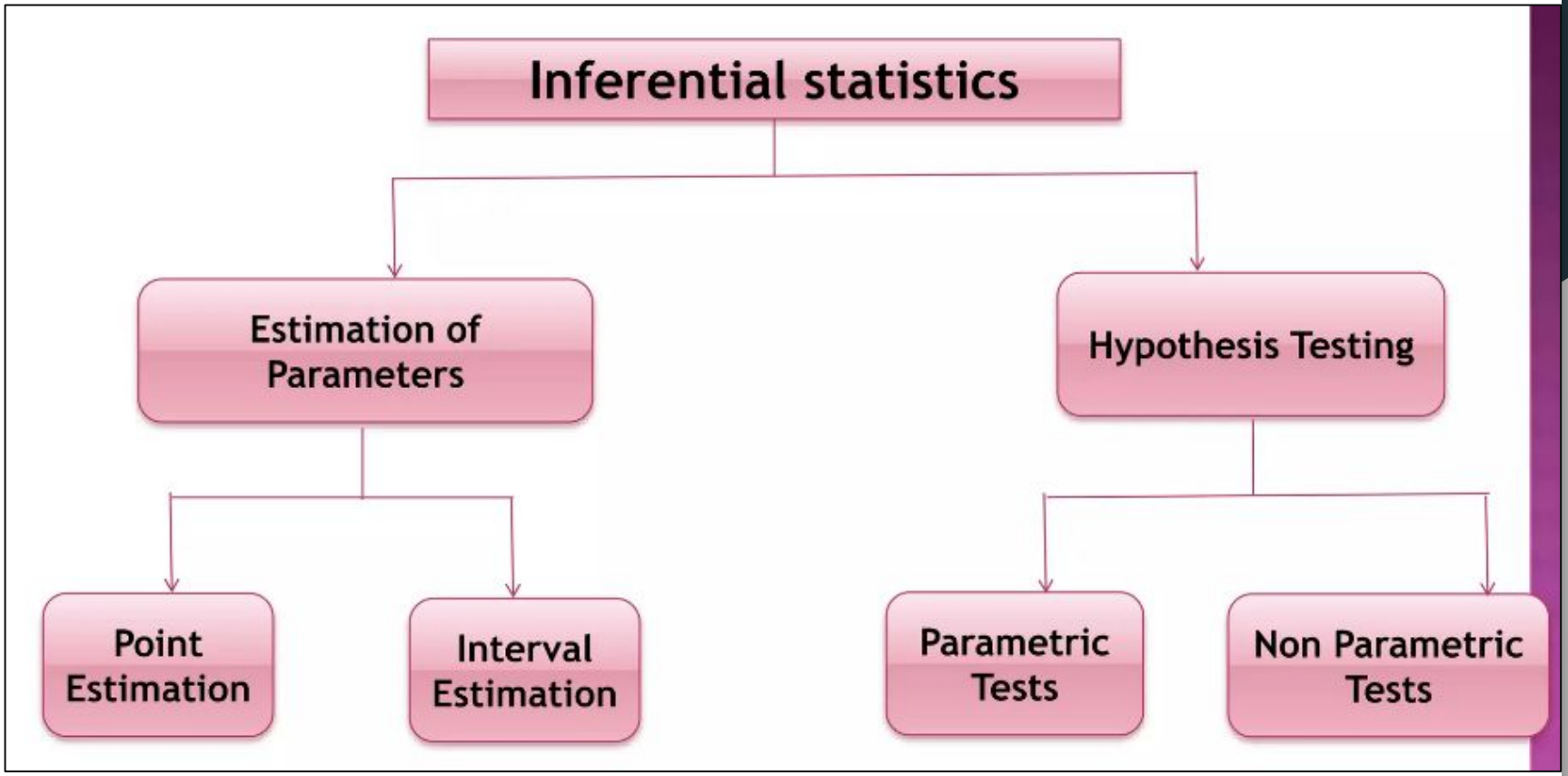
hYPOTHESIS- **A Hypothesis is a statement about the population parameter.**

**In other words,a hypothesis is a conclusion which is tentatively drawn on logical basis.**

# Chain of Reasoning for Inferential Statistics




*Are our inferences valid? ...Best we can do is to calculate probability about inferences*


Inferential statistics
Estimation of Parameters
Hypothesis Testing
Point Estimation
Interval Estimation
Parametric Tests
Non Parametric Tests

FORMS OF INFERENTIAL STATISTICS

The two forms of inferential statistics.

- Estimation
- Hypothesis testing

There are two forms of **estimation**:

- Point estimation (maximally likely value for parameter)
- Interval estimation (also called **confidence interval for parameter**)

Point estimates are single points that are used to **infer parameters** directly.

★ For Example,

**>Sample proportion** p^ ("p hat") is the point estimator of p

› **Sample mean** $\overline{X}$ ("x bar") is the point estimator of μ

› **Sample standard deviation** s is the point estimator of σ

# ESTIMATION

**Point Estimate:** A single statistic value that is the "best guess" for the parameter value.

**Interval Estimate:** An interval of numbers around the point estimate, that has a fixed "confidence level" of containing the parameter value, Called a confidence interval.

# Hypothesis testing

An objective method of **making decisions or inferences** from sample data (evidence)

Sample data used to choose between two choices i.e.  hypotheses or statements about a population

We typically do this by comparing what we have observed to what we expected

> A **statistical hypothesis** is a statement about the parameters of one or more populations.

# Hypothesis Testing

## (Procedure for testing A Hypothesis)

1. Setting up of hypothesis
2. Computation of test statistics
3. Types of errors in hypothesis testing
4. Levels of significance
5. Critical region or Rejection region
6. Two tailed test and one tailed test
7. Critical value
8. Decision

## Hypothesis Tests

used by statistician to determine whether **to reject a null hypothesis**, based on sample data. This process is called hypothesis testing and is consists of following **four steps**:

1. State the hypotheses - This step involves stating both null and alternative hypotheses. The hypotheses should be stated in such a way that they are mutually exclusive. If one is true then other must be false.

2. Formulate an analysis plan - The analysis plan is to describe how to use the sample data to evaluate the null hypothesis. The evaluation process focuses around a single test statistic. To set up level of Significance and confidence limit.

3. Analyze sample data - Find the **value of the test statistic** (using properties like mean score, proportion, t statistic, z-score, etc.) stated in the analysis plan.

4. Interpret results - Apply the decisions stated in the analysis plan. If the value of the test statistic is very unlikely based on the null hypothesis, then reject the null

Always **two hypotheses:**

- Null Hypothesis($H_0$): states that there is no exact or actual relationship between the variables.
- Alternative Hypothesis($H_1$): states that there is a statistically significant relationship between two variables.

**$H_A$: Research (Alternative) Hypothesis**

What we aim to gather evidence of

Typically that there is a difference/effect/relationship etc.

**$H_0$: Null Hypothesis**

What we assume is true to begin with

Typically that there is no difference/effect/relationship etc.

“The Court Case”

Members of a jury have to decide whether  a person is guilty or innocent based on evidence

Null: The person is innocent

Alternative: The person is not innocent (i.e. guilty)

The **null** can only be rejected if there is enough evidence to doubt it

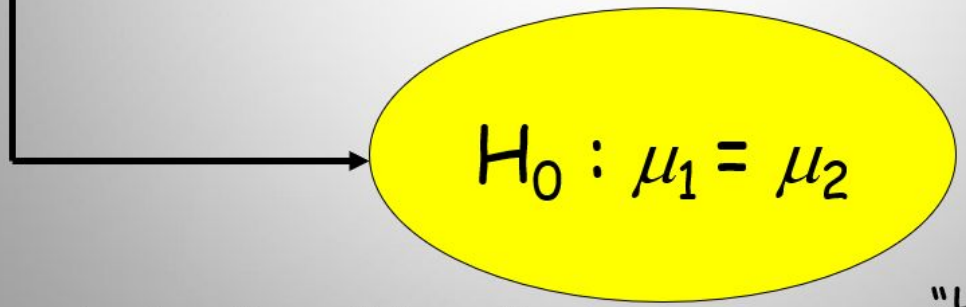
They do not know whether the person is really guilty or innocent so they may make a mistake

Inferential Statistics: uses sample data to evaluate the credibility of a hypothesis about a population

↓

NULL Hypothesis:

NULL (*nullus* - latin): "not any" → no differences between means




Always testing the null hypothesis

"H- Naught"

Inferential statistics: uses sample data to evaluate the credibility of a hypothesis about a population

↓

Hypothesis: Scientific or alternative hypothesis

Predicts that there are differences between the groups

$H_1 : \mu_1 \neq \mu_2$

# Hypothesis

A statement about what findings are expected

null hypothesis

*"the two groups will not differ"*

alternative hypothesis

*"group A will do better than group B"*
*"group A and B will not perform the same"*

**Null and Alternative hypothesis are mathematical opposites => Only one will present at at time**

## Statistical Test

These are intended to decide whether a hypothesis about distribution of one or more populations should be rejected or accepted.

**Statistical Test**

- Parametric Test
- Non Parametric Test

## Parametric Test

Gives generalizations for creating records about the mean of the original population.

These types of test includes Student's T tests and ANOVA tests, which assume data is from a normal distribution.

## Parametric Tests

t test (n<30)

t test

t test for one sample

t test for two samples

Unpaired two samples

Paired two samples

Pearson's Correlation

Z test for large samples (n>30)

ANOVA (Analysis of Variance)

ANOVA

ONE WAY

TWO WAY

# Key Difference Between Parametric & Non-parametric

| Properties | Parametric | Non-parametric |
|---|---|---|
| Assumptions | Yes | No |
| Value for central tendency | Mean value | Median value |
| Correlation | Pearson | Spearman |
| Probabilistic distribution | Normal | Arbitrary |
| Population knowledge | Requires | Does not require |
| Used for | Interval & Ratio data | Nominal & Ordinal data |
| Applicability | Variables | Attributes & Variables |
| Examples | t-test, z-test, ANOVA etc. | Kruskal-Wallis, Mann-Whitney |

## Type I and Type II Errors

- In the context of testing of hypotheses, there are basically two types of errors we can make

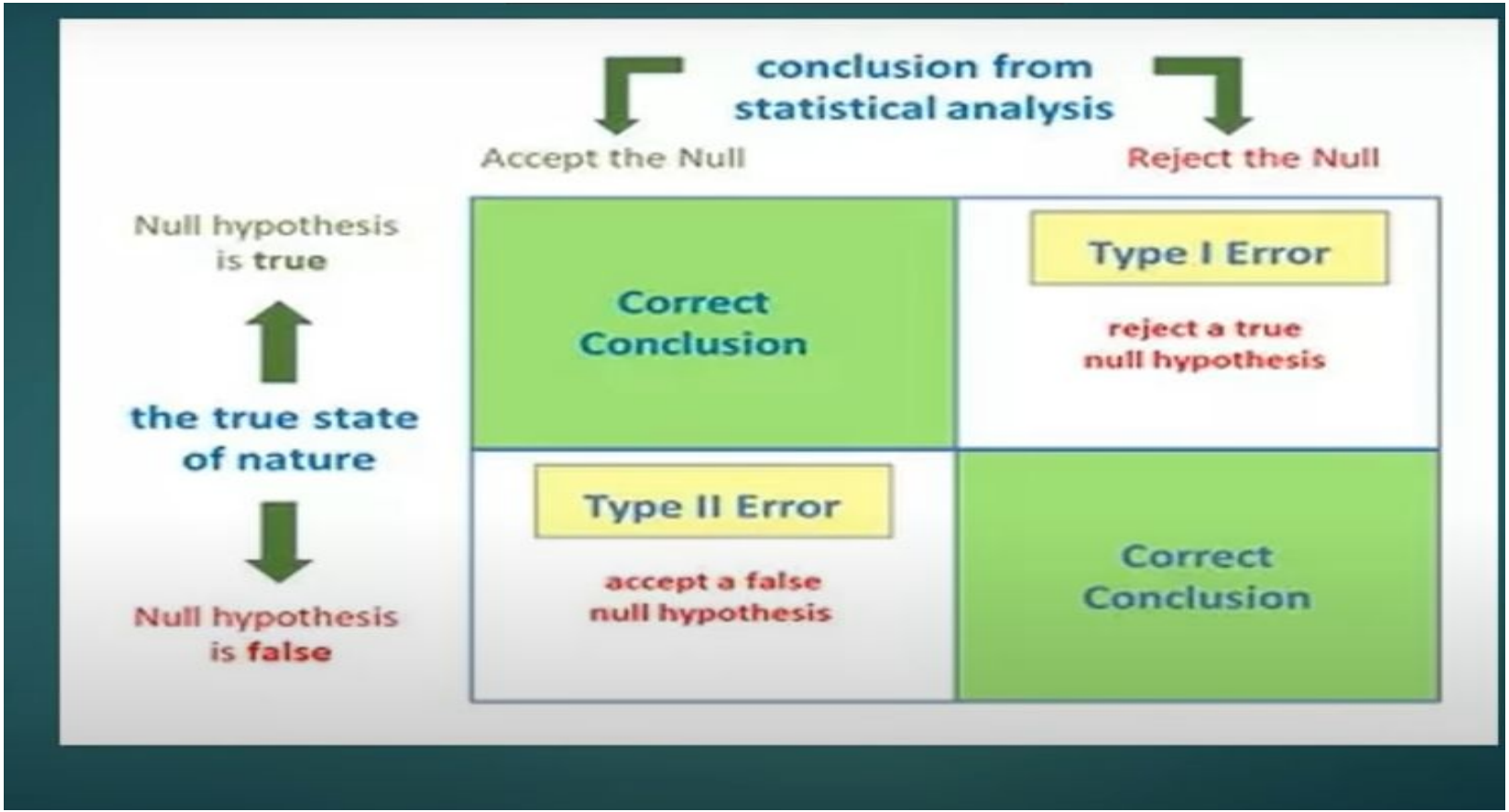
***Type I and Type II Errors***

## Type I error

- Known as an error of the first kind or alpha error(a)
- Occurs when the null hypothesis (H0) is true, but is rejected.
- It usually equals the significance level of a test.
- If type I error is fixed at 5 %, it means that there are about 5 chances in 100 that we will reject H0 when H0 is true

## Type II error

- Known as an error of the second kind, or β (beta)error
- Occurs when the null hypothesis is false, but erroneously fails to be rejected.


conclusion from statistical analysis
Accept the Null
Reject the Null
Null hypothesis is true
the true state of nature
Null hypothesis is false
Correct Conclusion
Type I Error
reject a true null hypothesis
Type II Error
accept a false null hypothesis
Correct Conclusion

# Level of Significance or p-value

- The level of statistical significance is often expressed as a *p*-value between 0 and 1.
- The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

## Level of significance:

After setting up the Null hypothesis, we set up the limits within we expect the null hypothesis to lie

It is a fixed probability of wrongly rejecting a True Null Hypothesis.

The probability that a random value of a statistic will lie in the critical region is called the level of significance & is expressed in percentage as **α=5% or 1%** level of significance.

α=5%=0.05 (the probability of rejecting a true hypothesis is 0.0.5)

Note that: Hypothesis is rejected it does mean that hypothesis is disproved.

For example, if α=5%, that means we are okay to take a 5% risk and conclude there exists a difference when there is no actual difference.

Usually we will have two levels of significance. 5% and 1% level of significance.

5% level of significance means that, if this experiment is repeated under identical conditions 100 times,then the chance for this conclusion to go wrong is five out of 100 .

1% level of significance means that, if this experiment is repeated under identical conditions 100 times,then the chance for this conclusion to go wrong is one out of 100 .

## Confidence limits :-

The limits within which an hypothesis should lie with specified probability

are called confidence limits .

Generally, the confidence limits are set up with 5% or 1% level of significance.

If level of Significance is 5% then confidence limits = 1- α ie. 0.95=95%

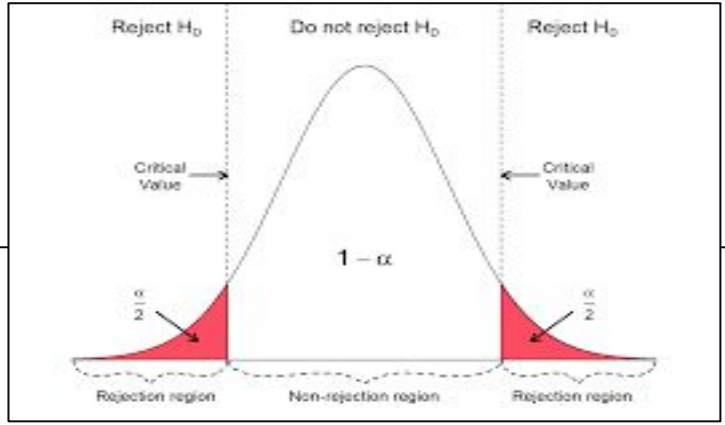
** If after testing result and null hypothesis has 5% or less that 5% difference then Null hypothesis is accepted else Null hypothesis is rejected.

# Hypothesis Testing

## (Procedure for testing A Hypothesis)

### Critical region or Rejection region

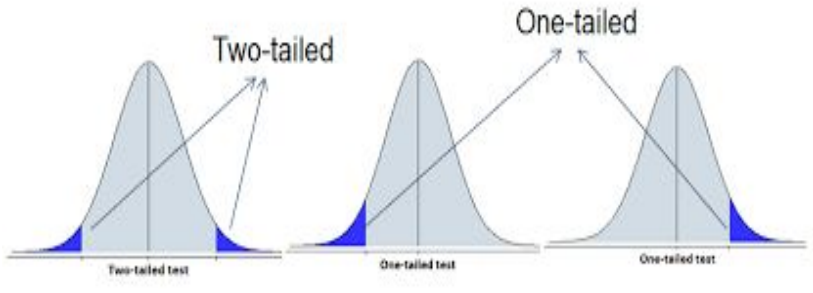
A critical region, also known as the rejection region, is a set of values for the test statistic for which the null hypothesis is rejected. i.e. if the observed test statistic is in the critical region then we reject the null hypothesis and accept the alternative hypothesis.

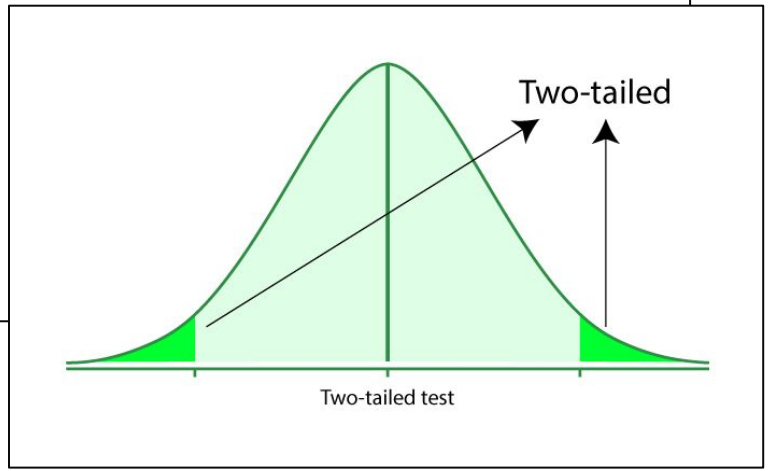


# Hypothesis Testing

(Procedure for testing A Hypothesis)

## Two tailed test and one tailed test



# Hypothesis Testing

## (Procedure for testing A Hypothesis)

**Critical value**

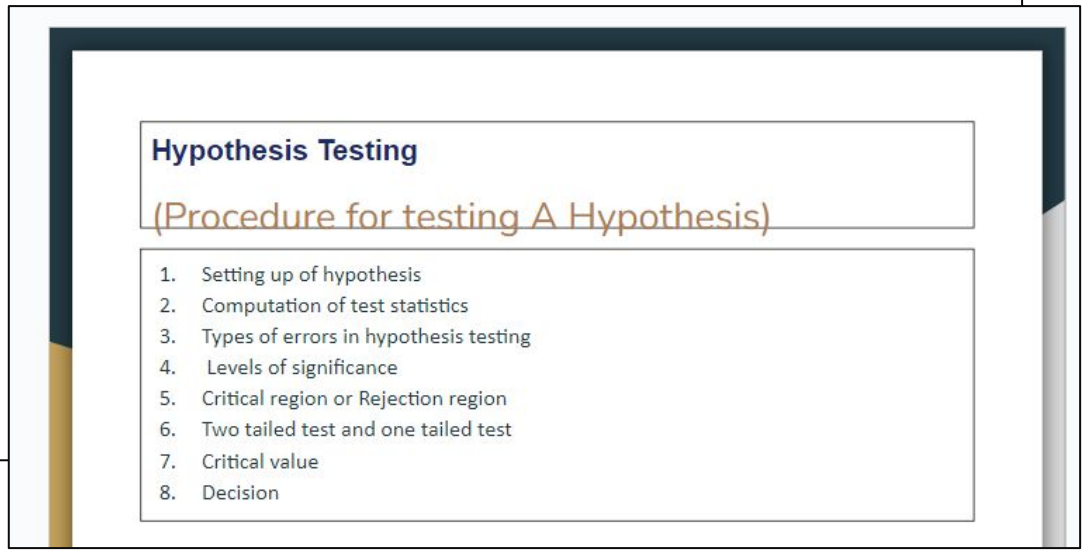
A critical value is a point on the distribution of the test statistic under the null hypothesis that defines a set of values that call for rejecting the null hypothesis. This set is called critical or rejection region. Usually, one-sided tests have one critical value and two-sided test have two critical values.

| DF | A = 0.1 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | 0.0005 |
|---|---|---|---|---|---|---|---|
| ∞ | ta = 1.282 | 1.645 | 1.96 | 2.326 | 2.576 | 3.091 | 3.291 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.656 | 318.289 | 636.578 |
| 2 | 1.886 | 2.92 | 4.303 | 6.965 | 9.925 | 22.328 | 31.6 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.214 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.61 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.894 | 6.869 |
| 6 | 1.44 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 1.397 | 1.86 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.25 | 4.297 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.93 | 4.318 |
| 13 | 1.35 | 1.771 | 2.16 | 2.65 | 3.012 | 3.852 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.14 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 1.337 | 1.746 | 2.12 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 1.333 | 1.74 | 2.11 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 1.33 | 1.734 | 2.101 | 2.552 | 2.878 | 3.61 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.85 |
| 21 | 1.323 | 1.721 | 2.08 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 1.319 | [illegible] | 2.069 | 2.5 | 2.807 | 3.485 | 3.768 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |

# Hypothesis Testing

(Procedure for testing A Hypothesis)

**Decision**

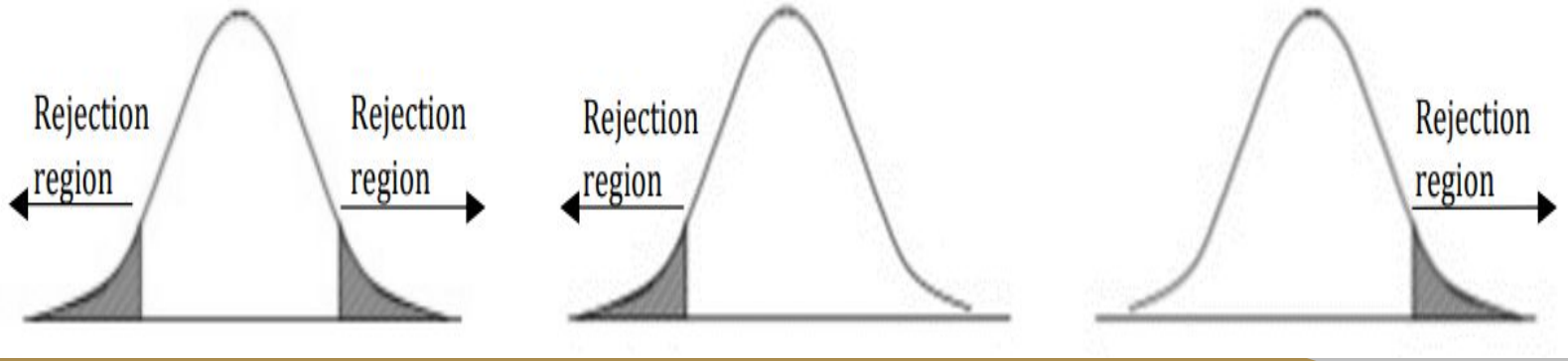


## Conclusion

**Note: H0 must always contain equality(=).**
**Ha always contains difference(≠, >, <).**

For example, if we were to test the equality of average means (μ) of two groups:for a two-tailed test, we define H0: μ1 = μ2 and Ha: μ1≠μ2

for a one-tailed test, we define H0: μ1 = μ2 and Ha: μ1 > μ2(right tailed) or Ha: μ1 < μ2(Left Tailed)



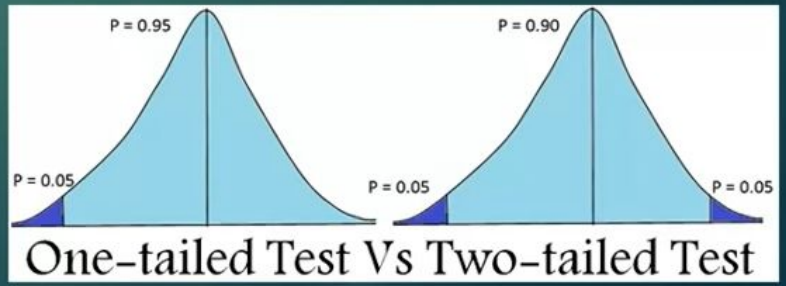
**Assumptions made in the use of 't' test**

1. Samples are randomly selected
2. Data utilised is Quantitative
3. Variable follow normal distribution
4. Sample variances are mostly same in both the groups under the study
5. Samples are small, mostly lower than 30

A t-test compares the difference between two means of different groups to determine whether that difference is statistically significant.

Student's 't' test for different purposes

- 't' test for one sample
- 't' test for unpaired two samples
- 't' test for paired two samples



## ONE SAMPLE T-TEST

" When compare the mean of a single group of observations with a specified value "

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

**Here:**

- **The sample mean**($\bar{x}$).
- **The population mean**($\mu$).
- **The sample standard deviation**(s)
- **Number of observations**(n)

If absolute value of 't' obtained is greater than table value then reject the null hypothesis and if it is less than table value, the null hypothesis may be accepted.

**Exercise:**

Your company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work? Test your hypothesis at a 5% alpha level.

h

**Exercise:**

Your company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work? Test your hypothesis at a 5% alpha level.

1. Null hypothesis is that there is no difference in sales, so: $H_0$: $\mu = \$100$.

2. Alternate hypothesis- There *is* a difference (that the mean sales increased),

so: $H_1$: $\mu > \$100$.

3.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The sample mean($\bar{x}$). = $130.

The population mean($\mu$). = $100

The sample standard deviation(s) = $15.

Number of observations(n) = 25.

**Exercise:**

Your company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work? Test your hypothesis at a 5% alpha level.

1. Null hypothesis is that there is no difference in sales, so: $H_0$: $\mu = \$100$.

2. Alternate hypothesis- There *is* a difference (that the mean sales increased),

so: $H_1$: $\mu > \$100$.

3.

The sample mean($\bar{x}$). = $130.

The population mean($\mu$). = $100

The sample standard deviation(s) = $15.

Number of observations(n) = 25.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{(130 - 100)}{(15 / \sqrt{(25)}} = \frac{30}{3} = 10$$

Find the t-table value.

The alpha level: given as 5% in the question.

The degrees of freedom, which is the number of items in the sample (n) minus 1

: 25 – 1 = 24.

Look up 24 degrees of freedom in the table at 0.05

| DF | A = 0.1 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | 0.0005 |
|---|---|---|---|---|---|---|---|
| ∞ | ta = 1.282 | 1.645 | 1.96 | 2.326 | 2.576 | 3.091 | 3.291 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.656 | 318.289 | 636.578 |
| 2 | 1.886 | 2.92 | 4.303 | 6.965 | 9.925 | 22.328 | 31.6 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.214 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.61 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.894 | 6.869 |
| 6 | 1.44 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 1.397 | 1.86 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.25 | 4.297 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.93 | 4.318 |
| 13 | 1.35 | 1.771 | 2.16 | 2.65 | 3.012 | 3.852 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.14 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 1.337 | 1.746 | 2.12 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 1.333 | 1.74 | 2.11 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 1.33 | 1.734 | 2.101 | 2.552 | 2.878 | 3.61 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.85 |
| 21 | 1.323 | 1.721 | 2.08 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 1.319 | [illegible] | 2.069 | 2.5 | 2.807 | 3.485 | 3.768 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |

The intersection is 1.711.This is your one-tailed critical t-value.

The calculated t value (10) is greater than table value(1.711).

Thus null hypothesis rejected and accepting the alternate hypothesis.

i.e,

***There is a difference (that the mean sales increased), so: $H_1$: $\mu > \$100$.***

# TWO SAMPLE 'T' TEST

## Unpaired Two sample 't'- test

Used when the two independent random samples come from the normal populations having unknown or same variance

**Null Hypothesis vs Alternative Hypothesis**

## Assumptions

- The samples are random & independent of each other
- The distribution of dependent variable is normal.
- The variances are equal in both the groups

## PAIRED TWO-SAMPLES T-TEST

Used when we have paired data of observations from one sample only, when each individual gives a pair of observations.

Same individuals are studied more than once in different circumstances- measurements made on the same people before and after interventions

## Assumptions

- The outcome variable should be continuous
- The difference between pre-post measurements should be normally distributed

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

***Where:***

X1 is the mean of sample 1

S1 is the standard deviation of sample 1

n**1** is the sample size of sample 1

X2 is the mean of sample 2

S2 is the standard deviation of sample 2

n**2** is the sample size in sample 2

# Steps for interpreting t test results

Step 1: State Null hypotheses & Alternate hypotheses

Step 2: Identify the given values

Step 3: put values in the equation and find t value

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

## Step 4: Compute the degree of freedom (df)

Unpaired Sample:

$$df = n1 + n2 - 2$$

Paired Sample:

$$df = n - 1$$

Step 5: Identify table value of t distribution using df.

Step 6: Compare and interpret using the calculated t value with table value…

t value > table value = null hypothesis rejected

t value < table value = null hypothesis accepted

**Exercise 1**

Percentage of cartoons labeled as "funny".

Women:84,97,58,90
Men: 88,90,52,97,86

Find Significance either men or women are more likely than the opposite gender, on average, to find cartoons funny.

$H_o$ = there is no significant difference between men and women to find cartoons as funny

$H_1$ = there is a significant difference between men and women to find cartoons as funny

| | Men | Women |
|---|---|---|
| | 88 | 84 |
| | 90 | 97 |
| | 52 | 58 |
| | 97 | 90 |
| | 86 | |
| Mean | 82.6 | 82.25 |
| SD | 17.601 | 17.0171 |
| N | 5 | 4 |

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

| | Men | Women |
|---|---|---|
| | 88 | 84 |
| | 90 | 97 |
| | 52 | 58 |
| | 97 | 90 |
| | 86 | |
| Mean | 82.6 | 82.25 |
| SD | 17.601 | 17.0171 |
| N | 5 | 4 |

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

$$= \frac{82.6-82.25}{\sqrt{\frac{(17.6)^2}{5}+\frac{(17.01)^2}{4}}} = \frac{82.6-82.25}{\sqrt{\frac{309.76}{5}+\frac{289.34}{4}}} = \frac{0.35}{\sqrt{61.95+72.33}}$$

$$= \frac{0.35}{\sqrt{134.28}} = \frac{0.35}{11.58} = \underline{\mathbf{0.03}}$$

Calculate degree of freedom (df)

Unpaired Sample: df= n1+n2 – 2

Therefore df=5+4-2= **7**

Identify table value for df 7 at p value 0.05.

Calculate degree of freedom (df)

Unpaired Sample: df= n1+n2 – 2

Therefore df=5+4-2= **7**

Identify table value for df 7 at p value 0.05.

Table value is calculated to be 2.37 which is higher than the calculated t value (0.03)and **Null hypothesis accepted**

Calculate degree of freedom (df)

Unpaired Sample: df= n1+n2 – 2

Therefore df=5+4-2= **7**

Identify table value for df 7 at p value 0.05.

Table value is calculated to be 2.37 which is higher than the calculated t value (0.03)and **Null hypothesis accepted**

**i.e** there is not enough (or significant) evidence to conclude that either men or women are more likely than the opposite gender, on average, to find cartoons funny.

Exercise 2

Find is there any significant change in mood after an intervention among 9 students

| Mood (Pre) | 3 | 0 | 6 | 7 | 4 | 3 | 2 | 1 | 4 |
|---|---|---|---|---|---|---|---|---|---|
| Mood (Post) | 5 | 1 | 5 | 7 | 10 | 9 | 7 | 11 | 8 |

Ho = there is no significant change in mood after intervention

$H_1$ = there is a significant change in mood after intervention

| | Mood Pre | Mood Post |
|---|---|---|
| | 3 | 5 |
| | 0 | 1 |
| | 6 | 5 |
| | 7 | 7 |
| | 4 | 10 |
| | 3 | 9 |
| | 2 | 7 |
| | 1 | 11 |
| | 4 | 8 |
| Mean | 3.33 | 7 |
| SD | 2.23 | 3.04 |

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

| | Mood Pre | Mood Post |
|---|---|---|
| | 3 | 5 |
| | 0 | 1 |
| | 6 | 5 |
| | 7 | 7 |
| | 4 | 10 |
| | 3 | 9 |
| | 2 | 7 |
| | 1 | 11 |
| | 4 | 8 |
| Mean | 3.33 | 7 |
| SD | 2.23 | 3.04 |

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

$$= \frac{3.33-7}{\sqrt{\frac{(2.23)^2}{9}+\frac{(3.04)^2}{9}}} = \frac{-3.67}{\sqrt{\frac{4.97}{9}+\frac{9.24}{9}}}$$

$$= \frac{-3.67}{\sqrt{0.55+1.02}} = \frac{-3.67}{\sqrt{1.57}}$$

$$= \frac{-3.67}{1.25} = \underline{\mathbf{-2.94}}$$

Calculate degree of freedom (df)

Paired Sample: df= n – 1

Therefore df=9-1= **8**

Identify table value for df 8 at p value 0.05.

Table value is calculated to be 2.31 which is less than the calculated t value (2.94)and **Null hypothesis rejected**

Calculate degree of freedom (df)

Paired Sample: df= n – 1

Therefore df=9-1= **8**

Identify table value for df 8 at p value 0.05.

Table value is calculated to be 2.31 which is less than the calculated t value (2.94)and **Null hypothesis rejected**

So it stated that there is a significant change in mood after intervention

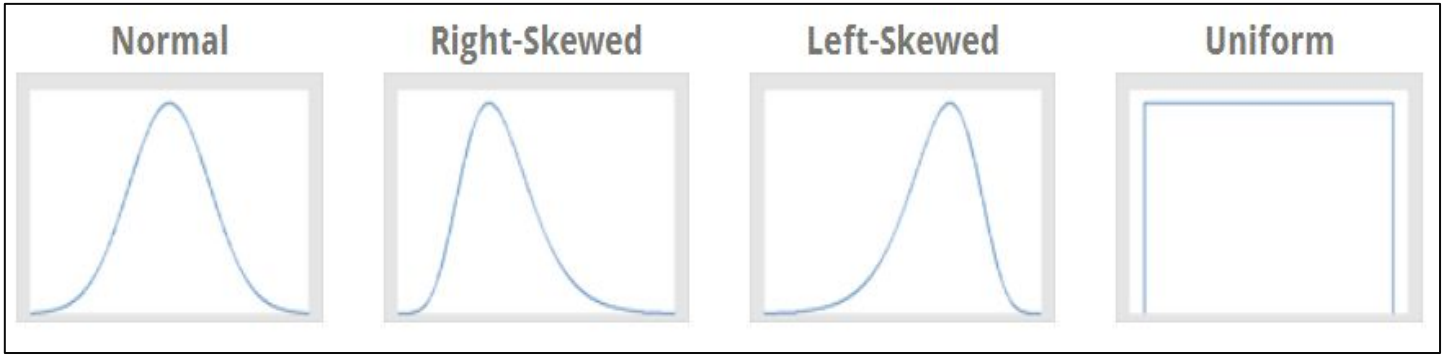
# Central Limit Theorem

The central limit theorem in statistics states that, given a sufficiently large sample size,the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population.

# Regardless of that variable's distribution in the population

In a population, the values of a variable can follow different probability distributions.

These distributions can range from normal, left-skewed, right-skewed, and uniform among others.

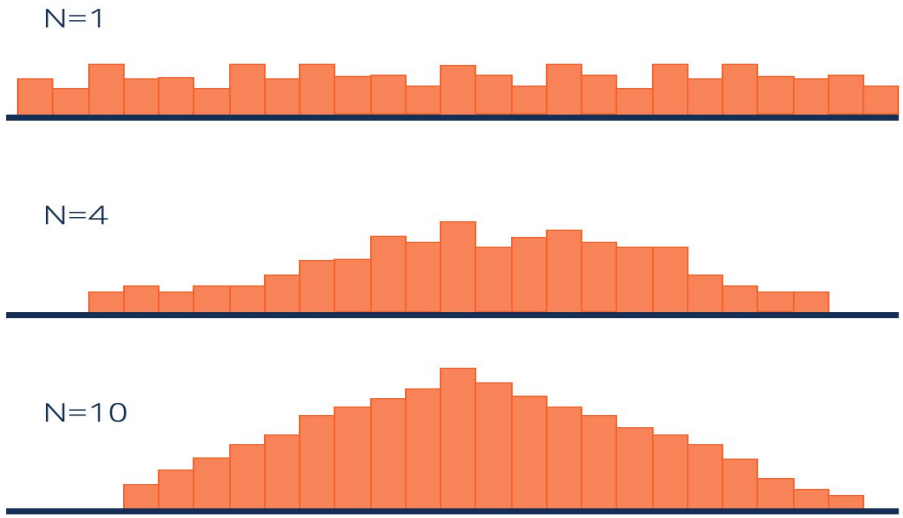


The shape of the sampling distribution depends on the sample size.

If you perform the study using the same procedure and change only the sample size, the shape of the sampling distribution will differ for each sample size.

**the shape of the sampling distribution changes with the sample size**

And, that brings us to the next part of the CLT definition!



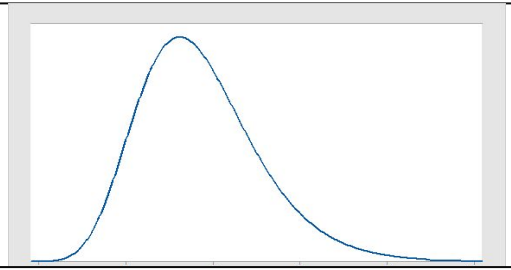
The central limit theorem links the following two distributions:

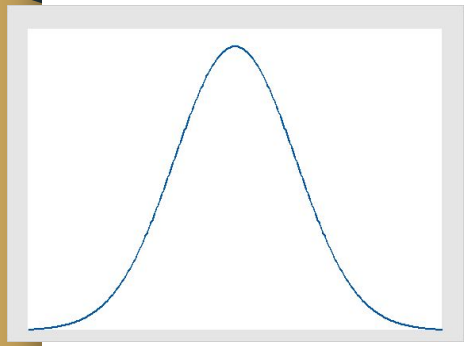
1. The distribution of the variable in the population.
2. The sampling distribution of the mean.

the CLT states that regardless of the variable's distribution in the population, the sampling distribution of the mean will tend to approximate the normal distribution.

the population distribution can look like→:

←the sampling distribution can appear like





It's not surprising that a normally distributed variable produces a sampling distribution that also follows the normal distribution.

But, surprisingly, non normal population distributions can also create normal sampling distributions.

Properties of the Central Limit Theorem

Normal distributions have two parameters, the mean and standard deviation.

As the sample size increases, the sampling distribution converges on a normal distribution where the **mean equals the population mean, and the standard deviation equals $\sigma/\sqrt{n}$.** Where:

$\sigma$ = the population standard deviation  n = the sample size

As the sample size (n) increases, the standard deviation of the sampling distribution becomes **smaller** because the square root of the sample size is in the denominator.

In other words, the sampling distribution clusters more tightly around the **mean** as sample size increases.

# Central Limit Theorem Formula

Sample mean = Population mean = $\mu$

$$\text{Sample standard deviation} = \frac{\text{(Standard deviation)}}{\sqrt{n}}$$

$$= \frac{\sigma}{\sqrt{n}}$$

## Applications of the central limit theorem are listed below:

In data science, the central limit theorem is used to make accurate assumptions of the population in order to build a robust statistical model.

In applied machine learning, the CLT helps to make inferences about the model performance.

In statistical hypothesis testing the central limit theorem is used to check if the given sample belongs to a designated population.

What Is a Z-Test?

A z-test is a statistical test used to determine whether two population means are different when t**he variances are known and the sample size is large.**

The test statistic is assumed to have a normal distribution, and standard deviation should be known in order for an accurate z-test to be performed.

A z-test is a hypothesis test in which the z-statistic follows a normal distribution.

A z-statistic, or z-score, is **a number** representing the result from the z-test.

Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size.

Z-tests assume the standard deviation is known, while t-tests assume it is unknown.

The z-test is best used for **greater-than-30 samples** because, under the central limit theorem, as the number of samples gets larger, the samples are considered to be approximately normally distributed.

When conducting a z-test, the null and alternative hypotheses, alpha and z-score should be stated.

Next, the test statistic should be calculated, and the results and conclusion stated.

## Table for Critical Values $Z_\alpha$ of Z.

| Critical value ($Z_\alpha$) | Level of Significance α. | | |
|---|---|---|---|
| | 1 % | 5 % | 10 % |
| Two Tailed Test | $\|Z_\alpha\| = 2.58$ | $\|Z_\alpha\| = 1.96$ | $\|Z_\alpha\| = 1.645$ |
| Right Tailed Test | $Z_\alpha = 2.33$ | $Z_\alpha = 1.645$ | $Z_\alpha = 1.28$ |
| Left Tailed Test | $Z_\alpha = -2.33$ | $Z_\alpha = -1.645$ | $Z_\alpha = -1.28$ |

... the decision about the null hypothesis *i.e.*, whether to accept it or ... with the critical

★ Z-distribution :- Sample is large ($n \geq 30$)

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

where $\bar{x}$ = sample Mean

$\mu$ = population mean

$\sigma$ = standard deviation of the population

$n$ = size of the sample.

Steps of Performing Z-test

(i) First identify the null and alternate hypothesis.

(ii) Determine the level of significance (α).

(iii) Find the critical value of Z in Z-test.

(iv) Calculate the Z-test statistics, using the formula for calculating the Z-test statistics:

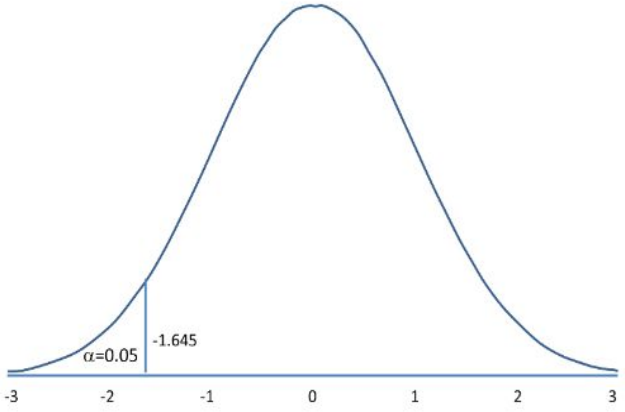
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

(v) Compare with the hypothesis and decide whether to reject or not to reject the null hypothesis.

## Type of Z-test

Left-tailed Test

In this test, the region of rejection is located to the extreme left of the distribution. Here, the population mean at least as small as some specified value of the mean .



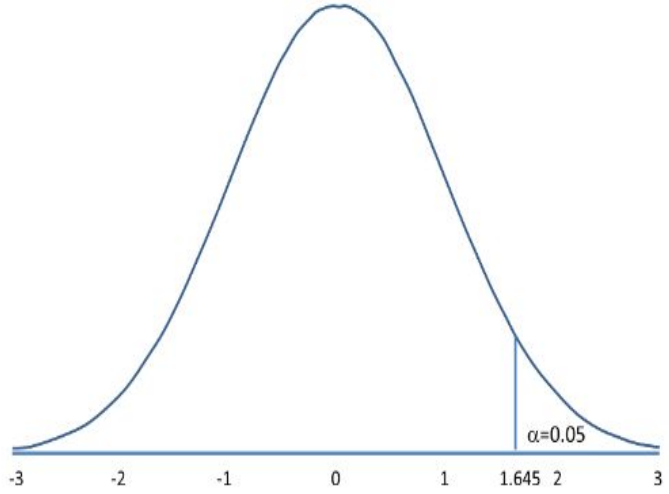

Rejection Region for Lower-Tailed Z Test ($H_1$: $\mu < \mu_0$ ) with $\alpha$ =0.05

The decision rule is: Reject $H_0$ if $Z \leq 1.645$.

| Lower-Tailed Test | |
|---|---|
| a | Z |
| 0.10 | -1.282 |
| 0.05 | -1.645 |
| 0.025 | -1.960 |
| 0.010 | -2.326 |
| 0.005 | -2.576 |
| 0.001 | -3.090 |
| 0.0001 | -3.719 |

## Right-tailed Test

In this test, the region of rejection is located to the extreme right of the distribution. Here the population mean is at least as large as some specified value of the mean.



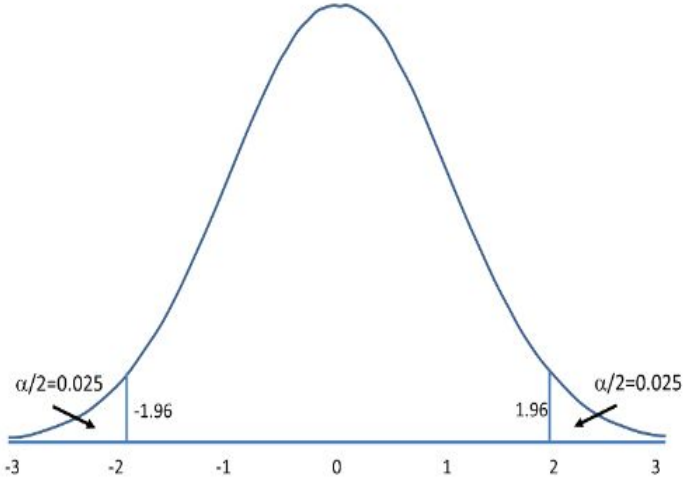

Rejection Region for Upper-Tailed Z Test ($H_1$: $\mu > \mu_0$ ) with $\alpha$=0.05

The decision rule is: Reject $H_0$ if $Z \geq 1.645$.

| Upper-Tailed Test | |
|---|---|
| α | Z |
| 0.10 | 1.282 |
| 0.05 | 1.645 |
| 0.025 | 1.960 |
| 0.010 | 2.326 |
| 0.005 | 2.576 |
| 0.001 | 3.090 |
| 0.0001 | 3.719 |

# Two-tailed Test

Here, the region of rejection is located to both extremes of the distribution. Here our NUll hypothesis is that the claimed value is equal to the mean population value.




Rejection Region for Two-Tailed Z Test ($H_1: \mu \neq \mu_0$) with $\alpha = 0.05$

The decision rule is: Reject $H_0$ if $Z \leq -1.960$ or if $Z \geq 1.960$.

| Two-Tailed Test | |
|---|---|
| α | Z |
| 0.20 | 1.282 |
| 0.10 | 1.645 |
| 0.05 | 1.960 |
| 0.010 | 2.576 |
| 0.001 | 3.291 |
| 0.0001 | 3.819 |

A gym trainer claimed that all the new boys in the gym are above average weight. A random sample of thirty boys weight have a mean score of 112.5 kg and the population mean weight is 100 kg and the standard deviation is 15. Is there a sufficient evidence to support the claim of gym trainer.

A gym trainer claimed that all the new boys in the gym are above average weight. A random sample of thirty boys weight have a mean score of 112.5 kg and the population mean weight is 100 kg and the standard deviation is 15. Is there a sufficient evidence to support the claim of gym trainer.

**Step-1: State Null and Alternate Hypothesis**

**Null Hypothesis:**

$H_0: \mu = 100$

**Alternate Hypothesis:**

$H_a: \mu > 100$

**Step-2: Set the significance level (alpha-value)**

Let alpha-value is 0.05, so corresponding z-score is 1.645

**Step-3: Find the z-value**

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{112.5 - 100}{\frac{15}{\sqrt{30}}} = 4.56$$

**Step-4: Comparing with the significance level:**

From step-3, we have

$4.56 > 1.645$

So, we have to reject the null hypothesis.

i.e. average weight of new boys are greater than 100 kg

## Two Sample Z- Test

A two sample z test is used to check if there is a difference between the means of two samples.

The z test statistic formula is given as follows:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$. $\overline{x_1}$, $\mu_1$, $\sigma_1^2$ are the sample mean, population mean and population variance respectively for the first sample. $\overline{x_2}$, $\mu_2$, $\sigma_2^2$ are the sample mean, population mean and population variance respectively for the second sample.

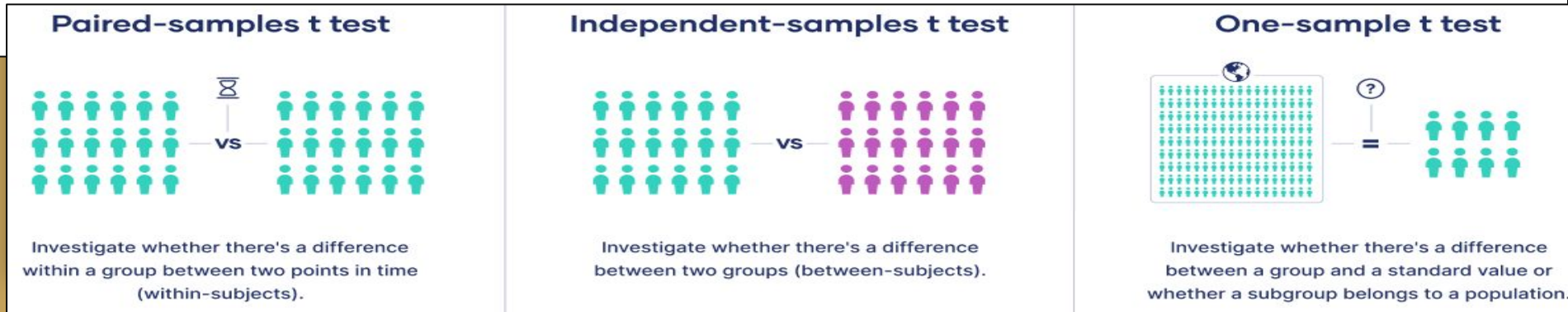
The two-sample z test can be set up in the same way as the one-sample test. However, this test will be used to compare the means of the two samples. For example, the null hypothesis is given as H0 :μ1=μ2.

The t-test is a test that is mainly used to compare the **mean** of two groups of samples. It is meant for evaluating whether the means of the two sets of data are statistically significantly different from each other.

Types of t-test:

1. The **one-sample t-test**, which is used to compare the mean of a population with a theoretical value.
2. The **unpaired two-sample t-test**, which is used to compare the mean of two **independent** given samples.
3. The **paired t-test**, which is used to compare the means between two groups of samples that are **related**

t-test compares the average values of two data sets and determines **if they came from the same population.**



Calculating a t-test requires three fundamental data values.

They include

1. the difference between the mean values from each data set, or the mean difference,
2. the standard deviation of each group, and
3. the number of data values of each group.

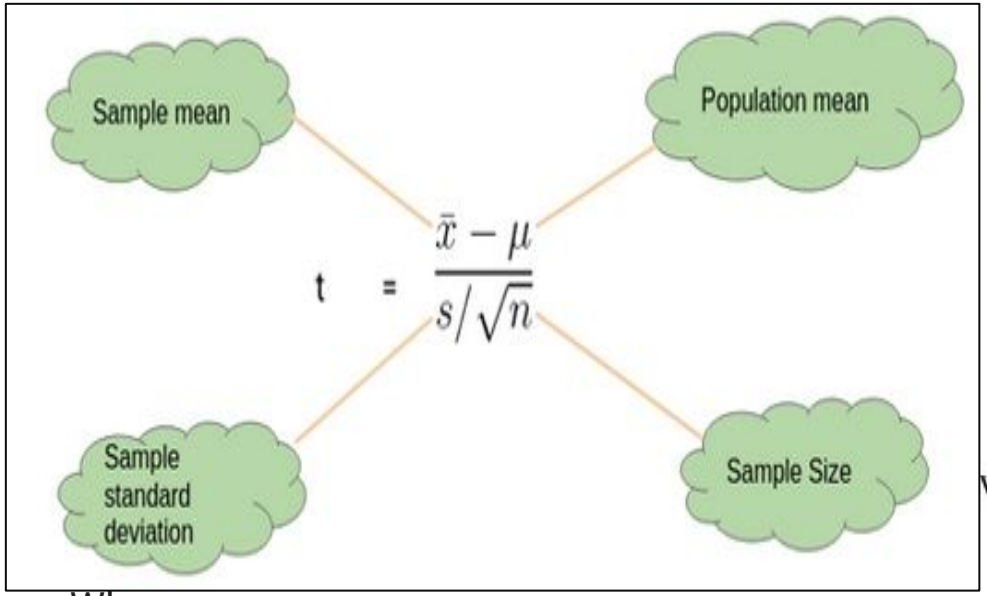
The t-test produces **two values** as its output: **t-value and degrees of freedom**.

This calculated t-value is then compared against a value obtained from **a critical value table** called the T-distribution table.

Higher values of the t-score indicate that a large difference exists between the two sample sets.

The smaller the t-value, the more similarity exists between the two sample sets.

# One-Sample T-Test




where $\bar{x} = \frac{\sum x_i}{n}$ and $s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$

Where,

- T = t-statistic
- $\bar{x}$ = mean of the group ( It is also sample mean)
- μ= theoretical mean value of the population
- s = standard deviation of the group
- n = sample size

Based on field experiments, a new variety green gram is expected to given an yield of 12.0quintals per hectare.
The variety was tested on 10 randomly selected farmers fields. The yield ( quintals/hectare) were recorded as 14.3,12.6,13.7,10.9,13.7,12.0,11.4,12.0,12.6,13.1. Do the results conform the expectation?

Null hypothesis H0: $\mu=12.0$ (i.e) the average yield of the new variety of green gram is 12.0 quintals/hectare.
Alternative Hypothesis: H1:$\mu \neq 12.0$ (i.e) the average yield is not 12.0 quintals/hectare
Level of significance: 5 %

$$\sum x = 126.3 \qquad \sum x^2 = 1605.77$$

$$\bar{x} = \frac{\sum x}{n} = \frac{126.3}{10} = 12.63$$

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{1605.77 - 1595.169}{9}} = \sqrt{\frac{10.601}{9}}$$

$$= 1.0853$$

$$\frac{s}{\sqrt{n}} = \frac{1.0853}{\sqrt{10}} = 0.3432$$

Now $t = \left| \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \right| \qquad t = \frac{12.63 - 12}{0.3432} = 1.836$

t < t table

We accept the null hypothesis H0

We conclude that the new variety of green gram will give an average yield of 12 quintals/hectare

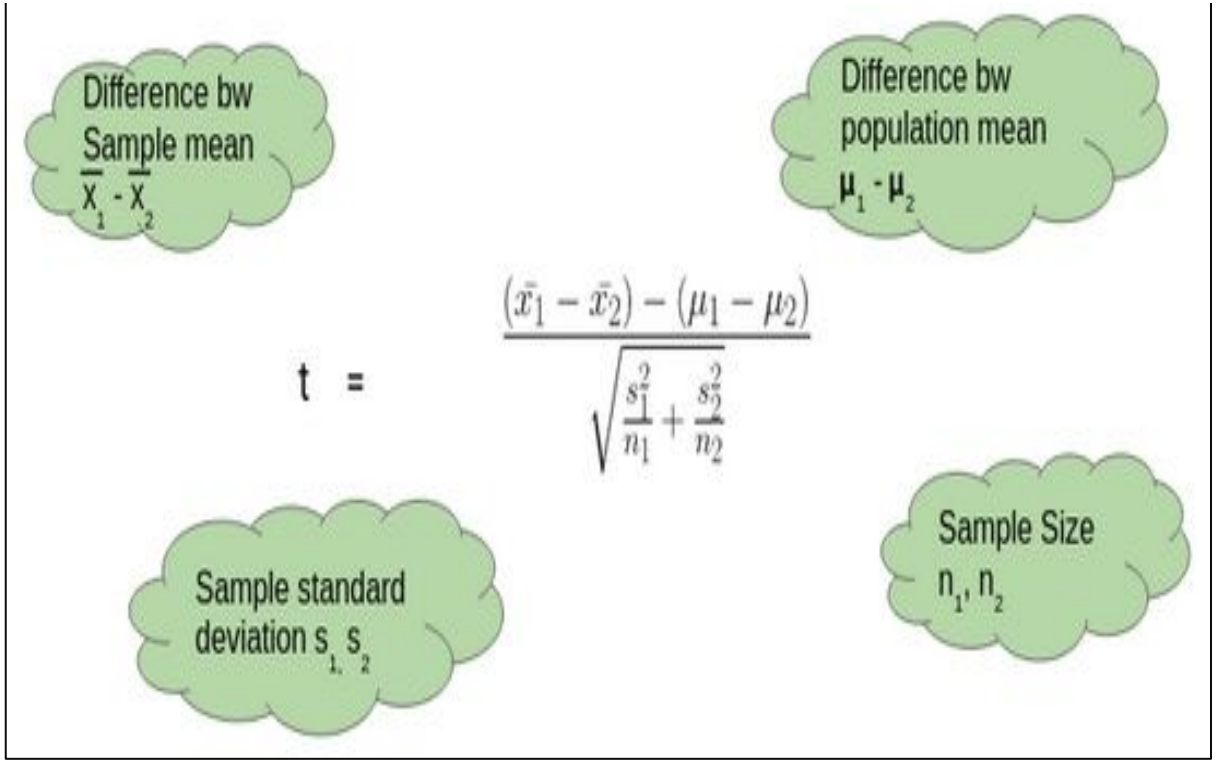
Table value for t corresponding to 5% level of significance and 9 d.f. is 2.262 (two tailed test)

t < ttab
We accept the null hypothesis H0
We conclude that the new variety of green gram will give an average yield of 12 quintals/hectare

| Degrees of freedom (df) | .2 | .15 | .1 | .05 | .025 | .01 | .005 | .001 |
|---|---|---|---|---|---|---|---|---|
| 1 | 3.078 | 4.165 | 6.314 | 12.706 | 25.452 | 63.657 | 127.321 | 636.619 |
| 2 | 1.886 | 2.282 | 2.920 | 4.303 | 6.205 | 9.925 | 14.089 | 31.599 |
| 3 | 1.638 | 1.924 | 2.353 | 3.182 | 4.177 | 5.841 | 7.453 | 12.924 |
| 4 | 1.533 | 1.778 | 2.132 | 2.776 | 3.495 | 4.604 | 5.598 | 8.610 |
| 5 | 1.476 | 1.699 | 2.015 | 2.571 | 3.163 | 4.032 | 4.773 | 6.869 |
| 6 | 1.440 | 1.650 | 1.943 | 2.447 | 2.969 | 3.707 | 4.317 | 5.959 |
| 7 | 1.415 | 1.617 | 1.895 | 2.365 | 2.841 | 3.499 | 4.029 | 5.408 |
| 8 | 1.397 | 1.592 | 1.860 | 2.306 | 2.752 | 3.355 | 3.833 | 5.041 |
| 9 | 1.383 | 1.574 | 1.833 | 2.262 | 2.685 | 3.250 | 3.690 | 4.781 |
| 10 | 1.372 | 1.559 | 1.812 | 2.228 | 2.634 | 3.169 | 3.581 | 4.587 |
| 11 | 1.363 | 1.548 | 1.796 | 2.201 | 2.593 | 3.106 | 3.497 | 4.437 |
| 12 | 1.356 | 1.538 | 1.782 | 2.179 | 2.560 | 3.055 | 3.428 | 4.318 |
| 13 | 1.350 | 1.530 | 1.771 | 2.160 | 2.533 | 3.012 | 3.372 | 4.221 |
| 14 | 1.345 | 1.523 | 1.761 | 2.145 | 2.510 | 2.977 | 3.326 | 4.140 |
| 15 | 1.341 | 1.517 | 1.753 | 2.131 | 2.490 | 2.947 | 3.286 | 4.073 |
| 16 | 1.337 | 1.512 | 1.746 | 2.120 | 2.473 | 2.921 | 3.252 | 4.015 |
| 17 | 1.333 | 1.508 | 1.740 | 2.110 | 2.458 | 2.898 | 3.222 | 3.965 |
| 18 | 1.330 | 1.504 | 1.734 | 2.101 | 2.445 | 2.878 | 3.197 | 3.922 |
| 19 | 1.328 | 1.500 | 1.729 | 2.093 | 2.433 | 2.861 | 3.174 | 3.883 |
| 20 | 1.325 | 1.497 | 1.725 | 2.086 | 2.423 | 2.845 | 3.153 | 3.850 |
| 21 | 1.323 | 1.494 | 1.721 | 2.080 | 2.414 | 2.831 | 3.135 | 3.819 |
| 22 | 1.321 | 1.492 | 1.717 | 2.074 | 2.405 | 2.819 | 3.119 | 3.792 |
| 23 | 1.319 | 1.489 | 1.714 | 2.069 | 2.398 | 2.807 | 3.104 | 3.768 |
| 24 | 1.318 | 1.487 | 1.711 | 2.064 | 2.391 | 2.797 | 3.091 | 3.745 |
| 25 | 1.316 | 1.485 | 1.708 | 2.060 | 2.385 | 2.787 | 3.078 | 3.725 |
| 26 | 1.315 | 1.483 | 1.706 | 2.056 | 2.379 | 2.779 | 3.067 | 3.707 |
| 27 | 1.314 | 1.482 | 1.703 | 2.052 | 2.373 | 2.771 | 3.057 | 3.690 |
| 28 | 1.313 | 1.480 | 1.701 | 2.048 | 2.368 | 2.763 | 3.047 | 3.674 |
| 29 | 1.311 | 1.479 | 1.699 | 2.045 | 2.364 | 2.756 | 3.038 | 3.659 |
| 30 | 1.310 | 1.477 | 1.697 | 2.042 | 2.360 | 2.750 | 3.030 | 3.646 |
| 40 | 1.303 | 1.468 | 1.684 | 2.021 | 2.329 | 2.704 | 2.971 | 3.551 |
| 50 | 1.299 | 1.462 | 1.676 | 2.009 | 2.311 | 2.678 | 2.937 | 3.496 |
| 60 | 1.296 | 1.458 | 1.671 | 2.000 | 2.299 | 2.660 | 2.915 | 3.460 |
| 70 | 1.294 | 1.456 | 1.667 | 1.994 | 2.291 | 2.648 | 2.899 | 3.435 |
| 80 | 1.292 | 1.453 | 1.664 | 1.990 | 2.284 | 2.639 | 2.887 | 3.416 |
| 100 | 1.290 | 1.451 | 1.660 | 1.984 | 2.276 | 2.626 | 2.871 | 3.390 |
| 1000 | 1.282 | 1.441 | 1.646 | 1.962 | 2.245 | 2.581 | 2.813 | 3.300 |
| Infinite | 1.282 | 1.440 | 1.645 | 1.960 | 2.241 | 2.576 | 2.807 | 3.291 |

Two-Sample T-Test

We perform a Two-Sample t-test when we want to compare the mean of two samples.



Two-tailed tests are used when t**he alternative hypothesis is non-directional.**

A non-directional hypothesis states that a population parameter (such as a mean or regression coefficient) is not equal to a certain value (such as 0). Two-tailed tests are appropriate for most studies.

If you're calculating a confidence interval, choose two-tailed.

One-tailed tests are used when the **alternative hypothesis is directiona**l.

A directional hypothesis states that a population parameter is greater than or less than a certain value.

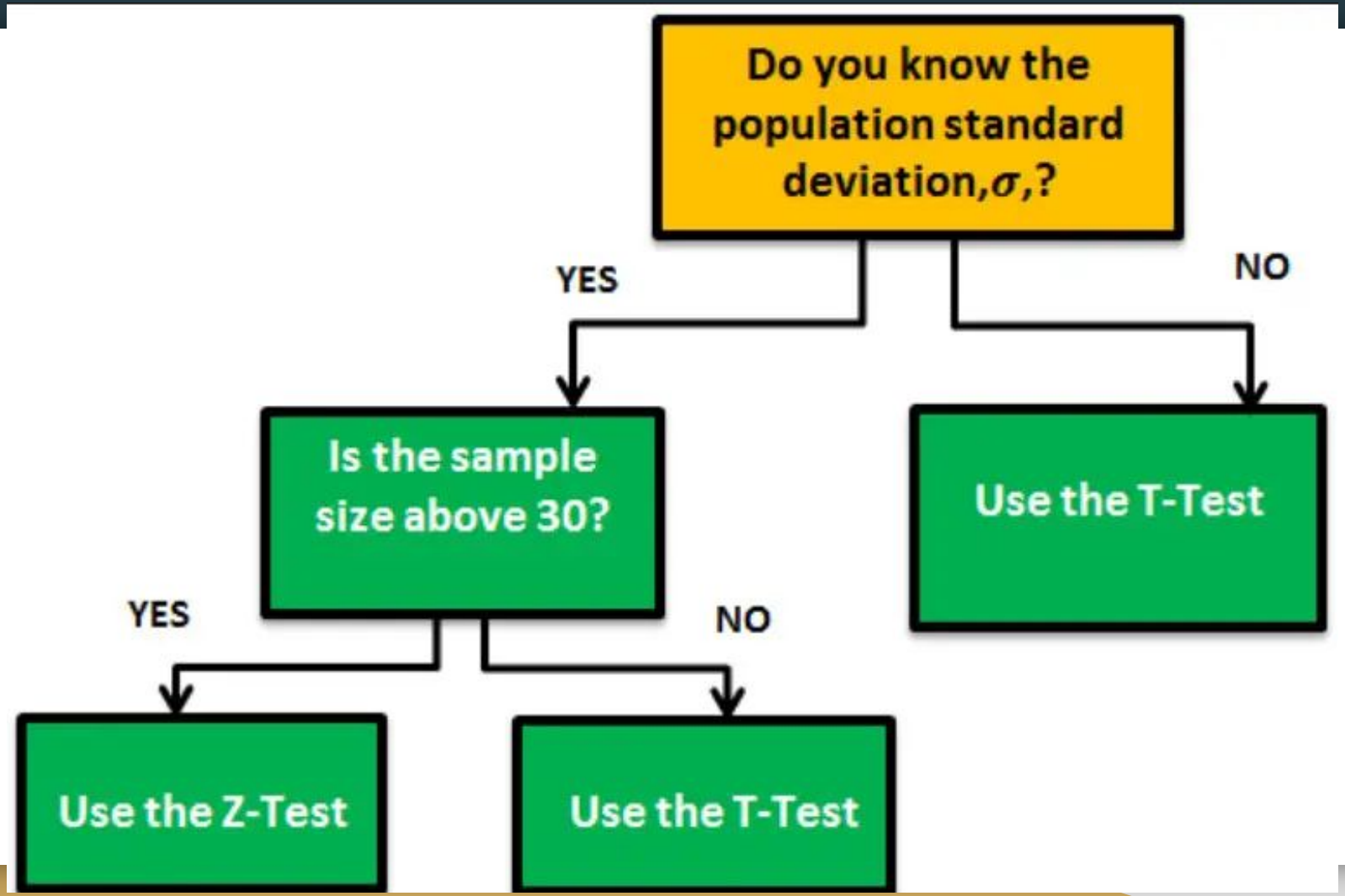
Your alternative hypothesis is directional if it includes words such as "greater than," "less than," "increases," "decreases," or the "<" or ">" sign. If it doesn't include these (or similar), it is probably non-directional.

## Critical values of *t* for one-tailed tests

Significance level (α)

| Degrees of freedom *(df)* | .2 | .15 | .1 | .05 | .025 | .01 | .005 | .001 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.376 | 1.963 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.309 |
| 2 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 |
| 3 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 |
| 4 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 |
| 5 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 |
| 6 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 |
| 7 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 |
| 8 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 |
| 9 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 |
| 10 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 |
| 11 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 |
| 12 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 |
| 13 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 |
| 14 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 |
| 15 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 |
| 16 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 |
| 17 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 |
| 18 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 |
| 19 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 |
| 20 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 |
| 21 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 |
| 22 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 |
| 23 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 |
| 24 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 |
| 25 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 |
| 26 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 |
| 27 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 |
| 28 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 |
| 29 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 |
| 30 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 |
| 40 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 |
| 50 | 0.849 | 1.047 | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 | 3.261 |
| 60 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 |
| 70 | 0.847 | 1.044 | 1.294 | 1.667 | 1.994 | 2.381 | 2.648 | 3.211 |
| 80 | 0.846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 |
| 100 | 0.845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 |
| 1000 | 0.842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 | 3.098 |
| Infinite | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 |

Scribbr


Do you know the population standard deviation, $\sigma$,?
YES
NO
Is the sample size above 30?
Use the T-Test
YES
NO
Use the Z-Test
Use the T-Test

# F Distribution

The F-Distribution is a probability distribution that is commonly used in statistical analysis. It arises when comparing the variances of two normal populations.
The shape of the F-Distribution depends on the degrees of freedom. As the degrees of freedom increase, the distribution becomes more symmetrical and approaches a normal distribution.

**Applications of the F-Distribution in Statistics**

- The F-Distribution is commonly used in statistical analysis to compare the variances of two populations.
- For example, it can be used in analysis of variance (ANOVA) to test for differences in means of three or more groups.
- It can also be used in regression analysis to test the overall significance of a regression model, or to compare the variances of the residuals for two or more models.

# F Distribution

- Suppose we want to compare the effectiveness of three different treatments for a medical condition. We randomly assign 20 patients to each treatment and measure their recovery time. We can use ANOVA with the F-distribution to test whether there is a significant difference in the means of the recovery times for the three treatments.

- The null hypothesis is that the means of the recovery times for the three treatments are equal, and the alternative hypothesis is that they are not equal. We can use the F-test to determine whether we have sufficient evidence to reject the null hypothesis.

# F Distribution

## Real-World Examples of the F-Distribution:

- The F-distribution has numerous real-world applications. For example, it is used in finance to test whether the variances of stock returns are equal across two or more portfolios.
- It is also used in engineering to test the effectiveness of different manufacturing processes by comparing the variances of the outcomes.
- Additionally, the F-distribution is used in biostatistics to compare the variances of health outcomes across different treatments or interventions.

# What is Analysis Of Variance?

Analysis of variance (ANOVA) is a statistical test for detecting differences in the group means when there is one parametric dependent variable & one or more independent variable

Specifically, we are interested in determining whether differences exist between the population means

An ANOVA Analysis of Variance

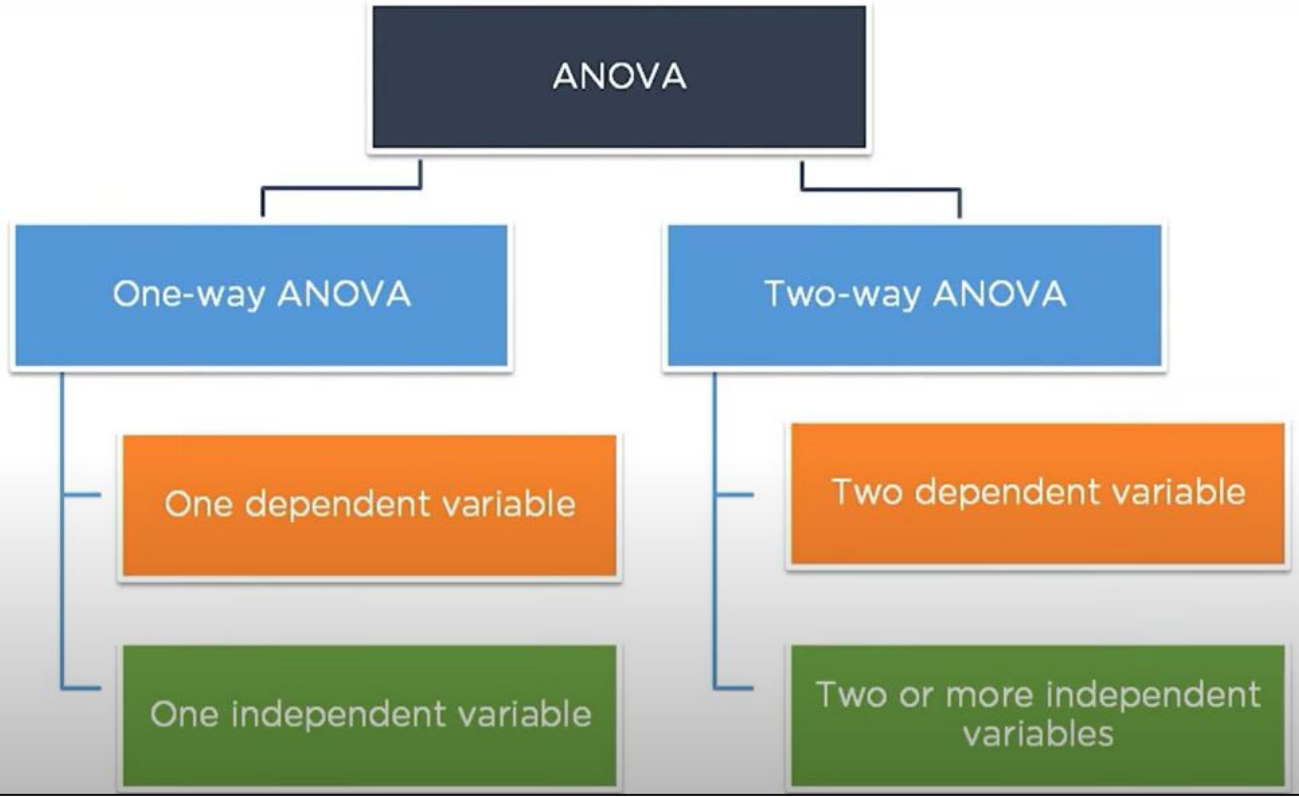
ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests

This test is a way to find out if **survey or experiment results are significant.**

In other words, they help you to figure out if you n**eed to reject the null hypothesis or accept the alternate hypothesis.**

Basically, **you're testing groups to see if there's a difference between them.** Examples of when you might want to test different groups:

1. A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
2. A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
3. Students from different colleges take the same exam. You want to see if one college outperforms the other.


ANOVA
One-way ANOVA
Two-way ANOVA
One dependent variable
One independent variable
Two dependent variable
Two or more independent variables

# Important Terminologies

**Let's assume that a new drug is developed with the goal of lowering the blood pressure more than the existing drug**

**Null Hypothesis**

The new drug doesn't lower the blood pressure more than the existing drug

**Alternative Hypothesis**

The new drug does significantly lower the blood pressure more than the existing drug

# Important Terminologies

**Let's assume that a new drug is developed with the goal of lowering the blood pressure more than the existing drug**

**F-Statistics**

The extent of difference between the means of different medical trials

Important Terms Related to ANOVA

**Means (Grand and Sample)**

A sample mean is the average value for a group, whereas the grand mean is the average of sample means from various groups or the mean of all observations combined.

**F-Statistics**

F-statistic or F-ratio is a statistical measure that tells us about the extent of difference between the means of different samples. Lower the F-ratio, closer are the sample means.

**Sum of Squares**

The sum of squares is a technique used in regression analysis to determine the dispersion of data points. It is used in the ANOVA test to compute the value of F.

**Mean Squared Error (MSE)**

The Mean Squared Error gives us the average error in the data set.

**Hypothesis**

In ANOVA, we have Null Hypothesis and an Alternative Hypothesis. The Null hypothesis is valid when all the sample means are equal, or they don't have any major difference.

The Alternate Hypothesis is valid when at least one of the sample means is different from the other.

**Group Variability**

In ANOVA, a group is a set of samples within the independent variable.

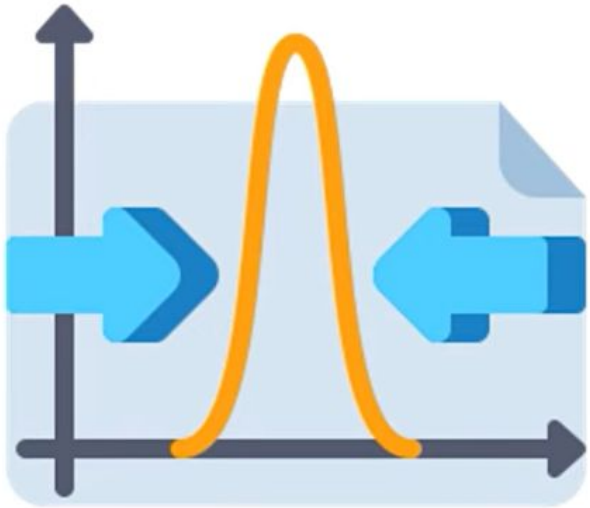
Between-group variability occurs when there is a significant variation in the sample distributions of individual groups.

Within-group variability occurs when there are variations in the sample distribution within a single group.

# How does ANOVA work?

ANOVA determines whether the groups created by the levels of the independent variable are statistically different by calculating whether the means of the different samples are different from the overall mean of the dependent variable

If any of the group means is significantly different from the overall mean, then the null hypothesis is rejected



One Way ANOVA

A one way ANOVA is used to compare **two means** from **two independent (unrelated) groups** using the **F-distribution**.

The **null hypothesis** for the test is that the **two means are equal.**

Therefore, a significant result means that the two means are unequal.

**Examples of when to use a one way ANOVA**

Situation 1: You have a group of individuals randomly split into smaller groups and completing different tasks. For example, you might be studying the effects of tea on weight loss and form three groups: green tea, black tea, and no tea.

## One Way ANOVA

The one way ANOVA test is used to determine whether there is any difference between the means of three or more groups. A one way ANOVA will have only one independent variable. The hypothesis for a one way ANOVA test can be set up as follows:

**Null Hypothesis,** $H_0$**:** $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

**Alternative Hypothesis,** $H_1$**:** The means are not equal

**Decision Rule:** If test statistic > critical value then reject the null hypothesis and conclude that the means of at least two groups are statistically significant.

The steps to perform the one way ANOVA test are given below:

Step 1: Calculate the mean for each group.

Step 2: Calculate the total mean. This is done by adding all the means and dividing it by the total number of means.

Step 3: Calculate the SSB.

Step 4: Calculate the between groups degrees of freedom.

Step 5: Calculate the SSE.

Step 6: Calculate the degrees of freedom of errors.

Step 7: Determine the MSB and the MSE.

Step 8: Find the f test statistic.

Step 9: Using the f table for the specified level of significance, $\alpha$, find the critical value. This is given by $F(\alpha, df1. df2)$.

Step 10: If $f > F$ then reject the null hypothesis.

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares | F Value |
|---|---|---|---|---|
| Between Groups | $SSB = \Sigma n_j(\bar{X}_j - \bar{X})^2$ | $df_1 = k - 1$ | $MSB = SSB / (k - 1)$ | $f = MSB / MSE$ |
| Error | $SSE = \Sigma\Sigma(X - \bar{X}_j)^2$ | $df_2 = N - k$ | $MSE = SSE / (N - k)$ | |
| Total | $SST = SSB + SSE$ | $df_3 = N - 1$ | | |

Sum of squares between groups, $SSB = \sum n_j(\bar{\bar{X}}_j - \bar{\bar{X}})^2$. Here, $\bar{\bar{X}}_j$ is the mean of the $j^{th}$ group, $\bar{\bar{X}}$ is the overall mean and $n_j$ is the sample size of the $j^{th}$ group.

Sum of squares of errors, $SSE = \sum\sum(X - \bar{\bar{X}}_j)^2$. Here, X refers to each data point in the $j^{th}$ group.

Total sum of squares, $SST = SSB + SSE$

Degrees of freedom between groups, $df_1 = k - 1$. Here, k denotes the number of groups.

Degrees of freedom of errors, $df_2 = N - k$, where N denotes the total number of observations across k groups.

Total degrees of freedom, $df_3 = N - 1$.

Mean squares between groups, $MSB = SSB / (k - 1)$

Mean squares of errors, $MSE = SSE / (N - k)$

ANOVA test statistic, $f = MSB / MSE$

Critical Value at $\alpha = F(\alpha, k - 1, N - k)$

# Real World Example

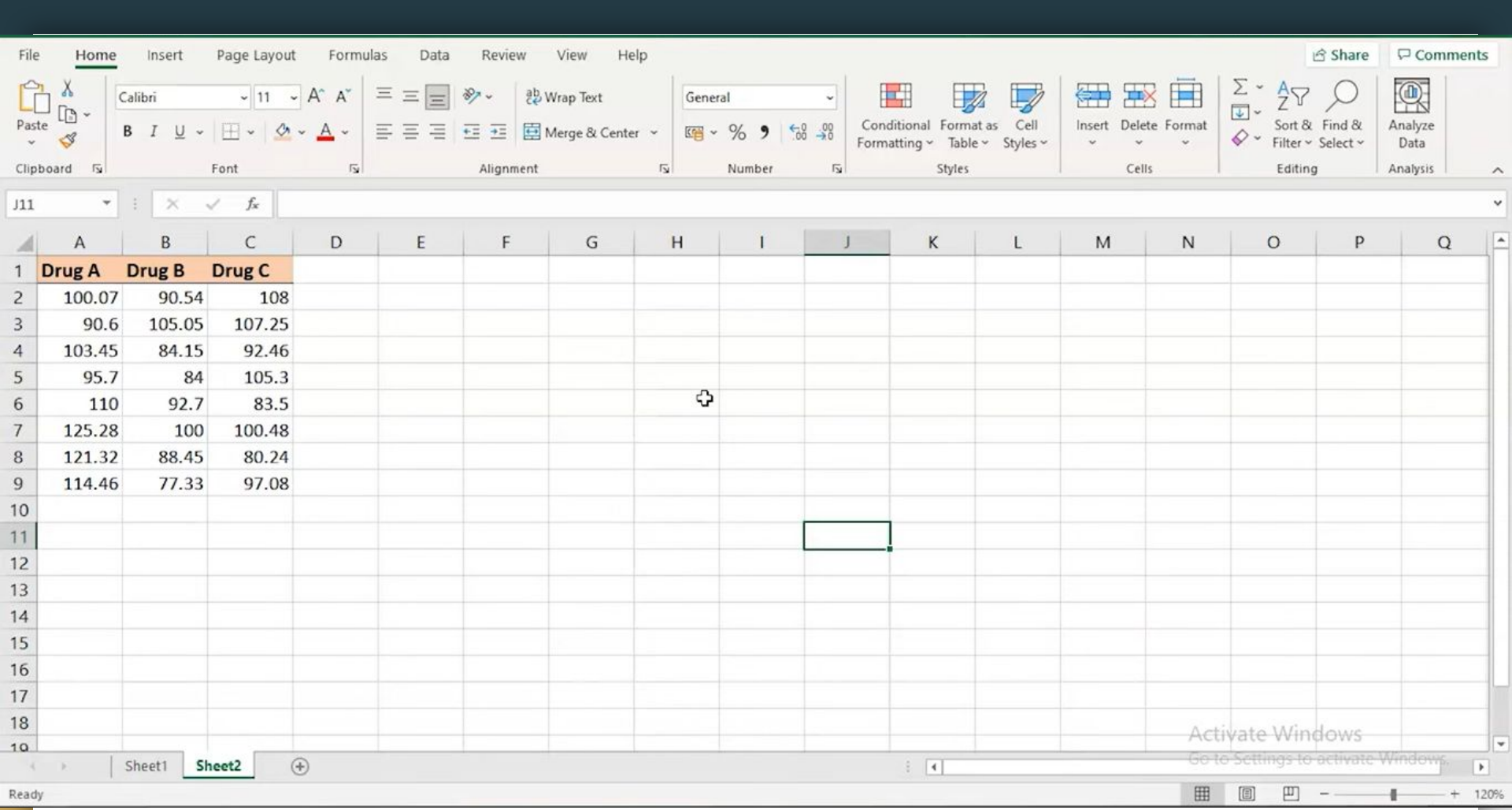
Suppose you are a marketing manager of a product company, and you want to know if the three different types of advertisement effect mean sales differently

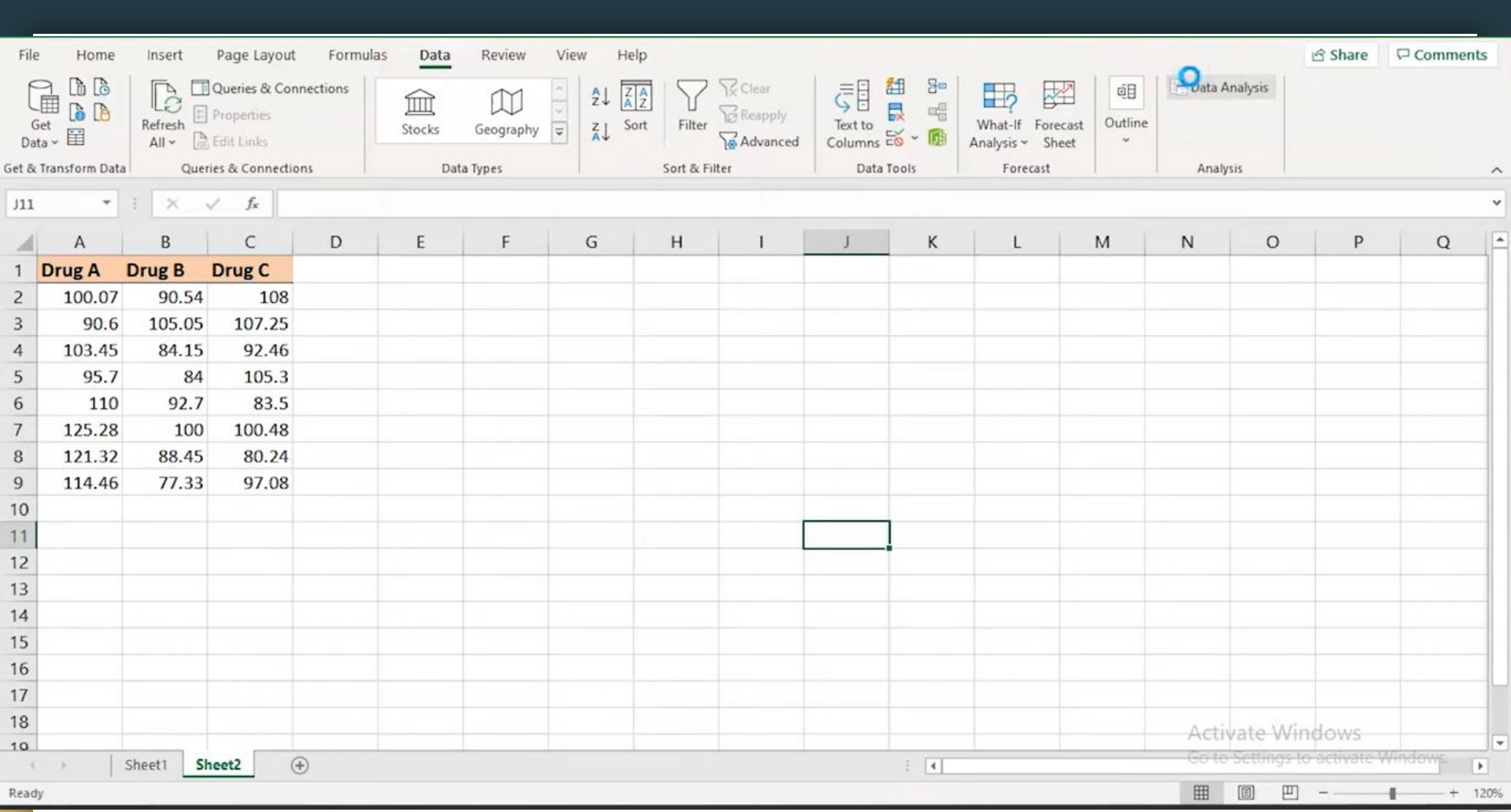
You use each type of advertisement at 20 different stores for one month and measure the total sales of each store for a month

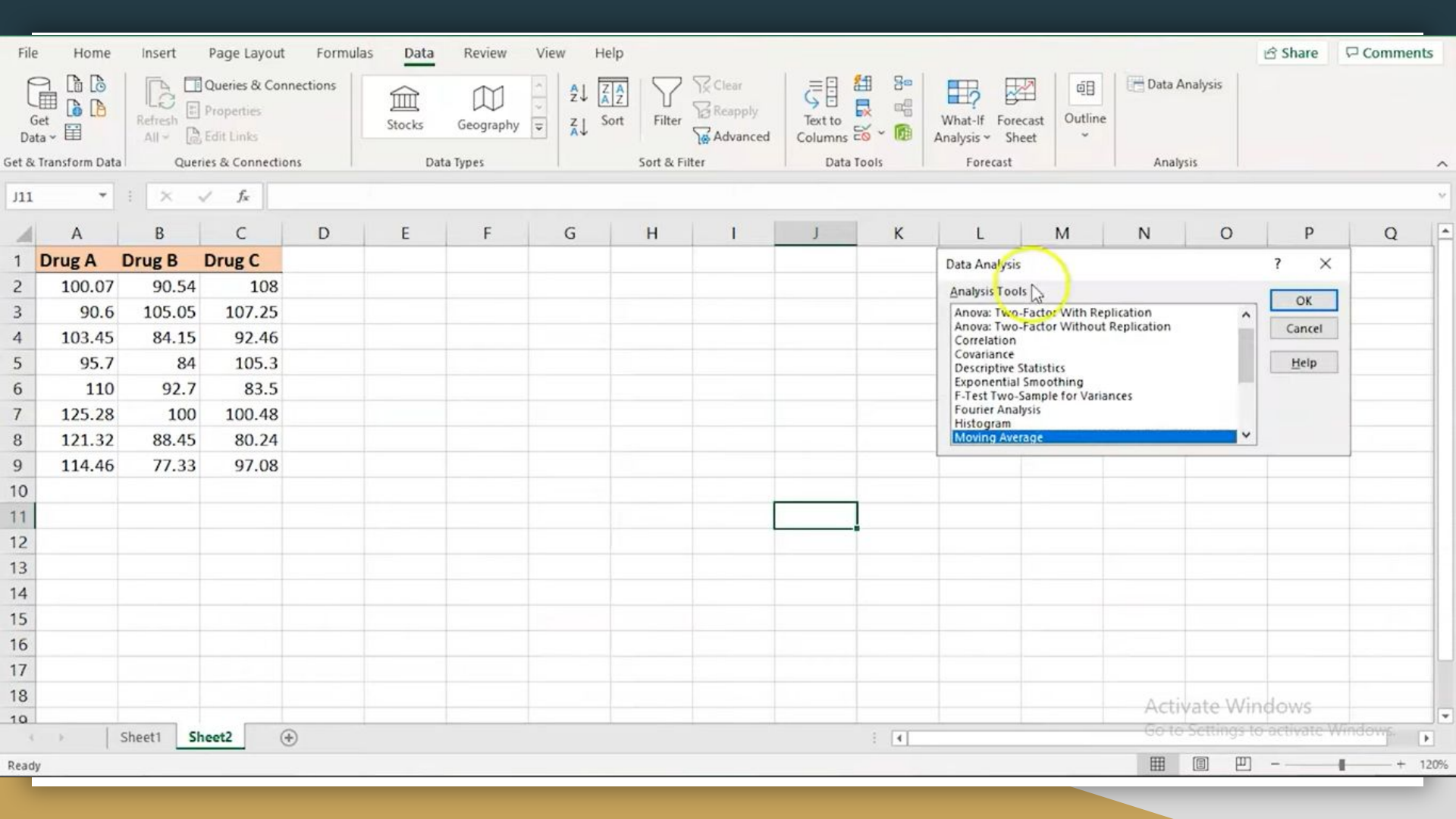
# Real World Example

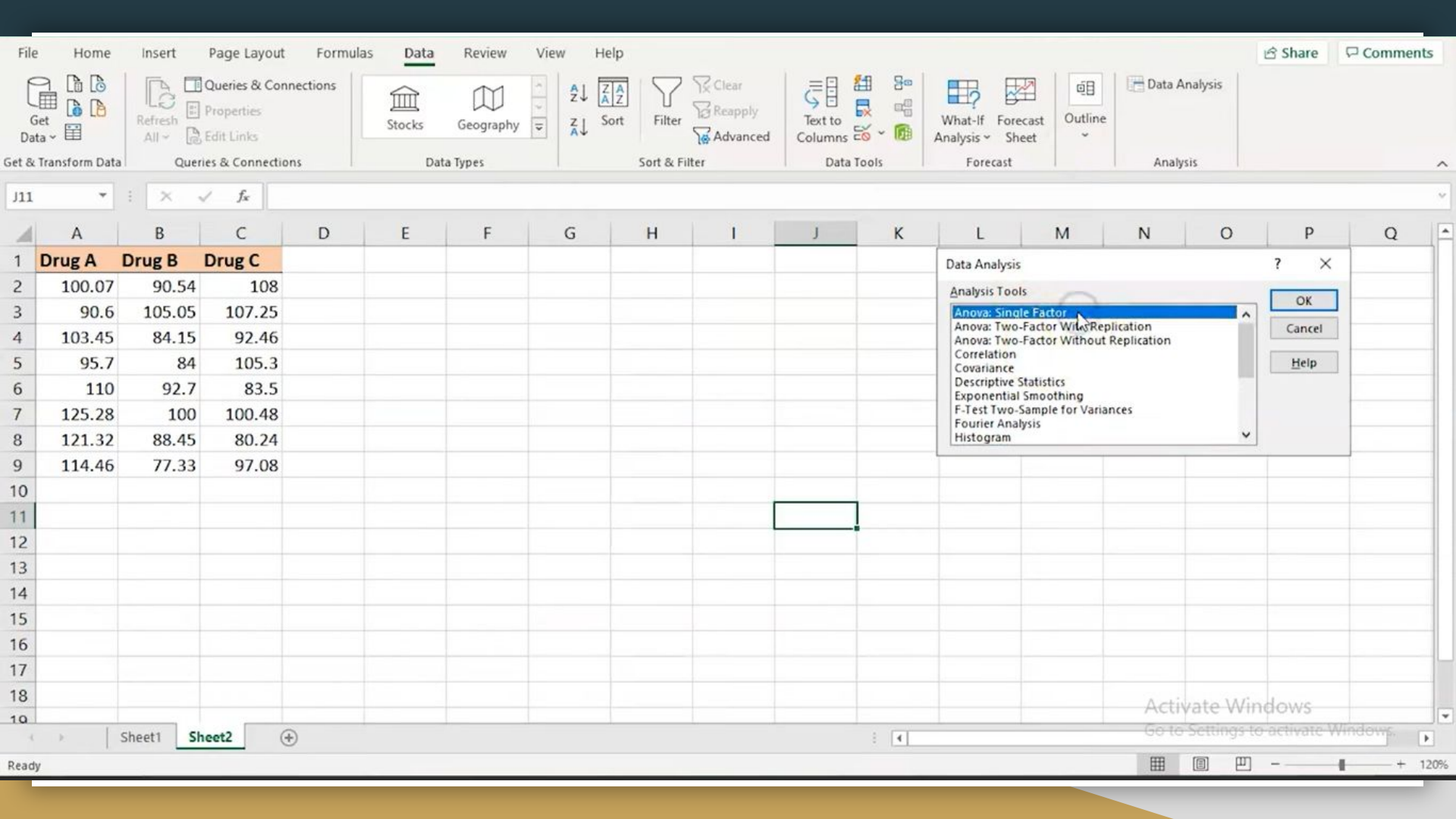
To observe if there is statistically significant difference in the mean sales between these three types of advertisements, you can conduct a one-way ANOVA

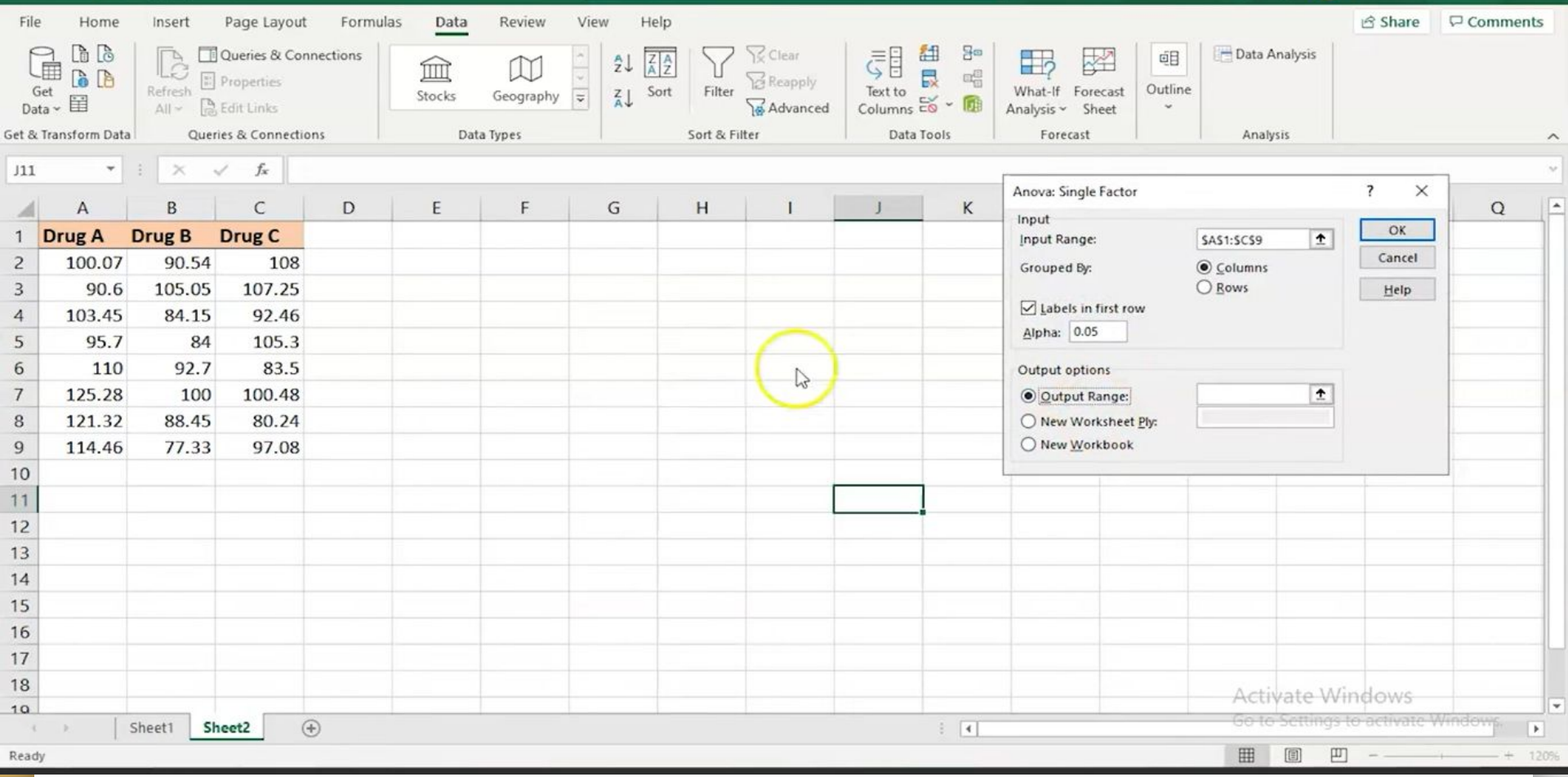
You will use types of advertisement as the factor and the sales as responsive variable

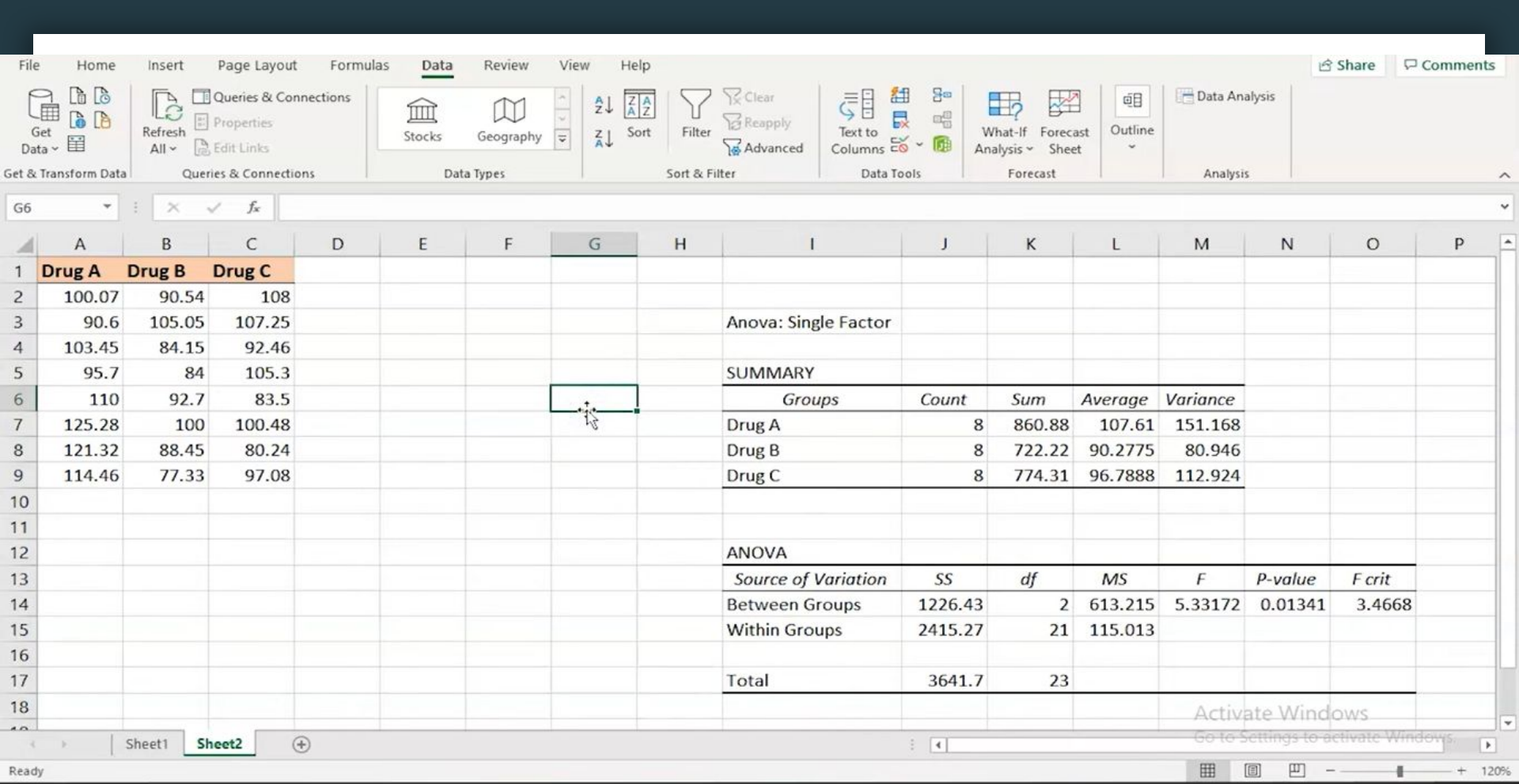

File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Share
Comments
Paste
Calibri
11
Wrap Text
General
Merge & Center
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Analyze Data
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
Analysis
J11
Drug A
Drug B
Drug C
100.07 90.54 108
90.6 105.05 107.25
103.45 84.15 92.46
95.7 84 105.3
110 92.7 83.5
125.28 100 100.48
121.32 88.45 80.24
114.46 77.33 97.08
Sheet1
Sheet2
Activate Windows
Go to Settings to activate Windows.
Ready
120%


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Share
Comments
Get Data
Refresh All
Queries & Connections
Properties
Edit Links
Stocks
Geography
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
What-If Analysis
Forecast Sheet
Outline
Data Analysis
Get & Transform Data
Queries & Connections
Data Types
Sort & Filter
Data Tools
Forecast
Analysis
J11
Drug A
Drug B
Drug C
100.07 90.54 108
90.6 105.05 107.25
103.45 84.15 92.46
95.7 84 105.3
110 92.7 83.5
125.28 100 100.48
121.32 88.45 80.24
114.46 77.33 97.08
Sheet1
Sheet2
Activate Windows
Go to Settings to activate Windows.
Ready
120%


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Share
Comments
Get Data
Refresh All
Queries & Connections
Properties
Edit Links
Stocks
Geography
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
What-If Analysis
Forecast Sheet
Outline
Data Analysis
Get & Transform Data
Queries & Connections
Data Types
Sort & Filter
Data Tools
Forecast
Analysis
J11
Drug A
Drug B
Drug C
100.07 90.54 108
90.6 105.05 107.25
103.45 84.15 92.46
95.7 84 105.3
110 92.7 83.5
125.28 100 100.48
121.32 88.45 80.24
114.46 77.33 97.08
Data Analysis
Analysis Tools
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
OK
Cancel
Help
Activate Windows
Go to Settings to activate Windows.
Sheet1
Sheet2
Ready
120%


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Share
Comments
Get Data
Refresh All
Queries & Connections
Properties
Edit Links
Stocks
Geography
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
What-If Analysis
Forecast Sheet
Outline
Data Analysis
Get & Transform Data
Queries & Connections
Data Types
Sort & Filter
Data Tools
Forecast
Analysis
J11
Drug A
Drug B
Drug C
100.07 90.54 108
90.6 105.05 107.25
103.45 84.15 92.46
95.7 84 105.3
110 92.7 83.5
125.28 100 100.48
121.32 88.45 80.24
114.46 77.33 97.08
Data Analysis
Analysis Tools
Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
OK
Cancel
Help
Activate Windows
Go to Settings to activate Windows.
Sheet1
Sheet2
Ready
120%


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Share
Comments
Get Data
Refresh All
Queries & Connections
Properties
Edit Links
Stocks
Geography
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
What-If Analysis
Forecast Sheet
Outline
Data Analysis
Get & Transform Data
Queries & Connections
Data Types
Sort & Filter
Data Tools
Forecast
Analysis
J11
Drug A
Drug B
Drug C
100.07 90.54 108
90.6 105.05 107.25
103.45 84.15 92.46
95.7 84 105.3
110 92.7 83.5
125.28 100 100.48
121.32 88.45 80.24
114.46 77.33 97.08
Anova: Single Factor
Input
Input Range: $A$1:$C$9
Grouped By: Columns Rows
Labels in first row
Alpha: 0.05
Output options
Output Range:
New Worksheet Ply:
New Workbook
OK
Cancel
Help
Sheet1
Sheet2
Ready
Activate Windows
Go to Settings to activate Windows.



File Home Insert Page Layout Formulas Data Review View Help

Share Comments

Get Data | Refresh All | Queries & Connections | Properties | Edit Links | Stocks | Geography | Sort | Filter | Clear | Reapply | Advanced | Text to Columns | What-If Analysis | Forecast Sheet | Outline | Data Analysis

Get & Transform Data | Queries & Connections | Data Types | Sort & Filter | Data Tools | Forecast | Analysis

G6

| | A | B | C |
|---|---|---|---|
| 1 | **Drug A** | **Drug B** | **Drug C** |
| 2 | 100.07 | 90.54 | 108 |
| 3 | 90.6 | 105.05 | 107.25 |
| 4 | 103.45 | 84.15 | 92.46 |
| 5 | 95.7 | 84 | 105.3 |
| 6 | 110 | 92.7 | 83.5 |
| 7 | 125.28 | 100 | 100.48 |
| 8 | 121.32 | 88.45 | 80.24 |
| 9 | 114.46 | 77.33 | 97.08 |

Anova: Single Factor

SUMMARY

| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
|---|---|---|---|---|
| Drug A | 8 | 860.88 | 107.61 | 151.168 |
| Drug B | 8 | 722.22 | 90.2775 | 80.946 |
| Drug C | 8 | 774.31 | 96.7888 | 112.924 |

ANOVA

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
|---|---|---|---|---|---|---|
| Between Groups | 1226.43 | 2 | 613.215 | 5.33172 | 0.01341 | 3.4668 |
| Within Groups | 2415.27 | 21 | 115.013 | | | |
| Total | 3641.7 | 23 | | | | |

Activate Windows
Go to Settings to activate Windows.

Sheet1 Sheet2

Ready

120%

Example 1: Three types of fertilizers are used on three groups of plants for 5 weeks. We want to check if there is a difference in the mean growth of each group. Using the data given below apply a one way ANOVA test at 0.05 significant level.

| Fertilizer 1 | Fertilizer 2 | Fertilizer 3 |
|---|---|---|
| 6 | 8 | 13 |
| 8 | 12 | 9 |
| 4 | 9 | 11 |
| 5 | 11 | 8 |
| 3 | 6 | 7 |
| 4 | 8 | 12 |

$H_0$: $\mu_1 = \mu_2 = \mu_3$

$H_1$: The means are not equal

| Fertilizer 1 | Fertilizer 2 | Fertilizer 3 |
|---|---|---|
| 6 | 8 | 13 |
| 8 | 12 | 9 |
| 4 | 9 | 11 |
| 5 | 11 | 8 |
| 3 | 6 | 7 |
| 4 | 8 | 12 |
| $\overline{X}_1 = 5$ | $\overline{X}_1 = 9$ | $\overline{X}_1 = 10$ |

Total mean, $\overline{\overline{X}} = 8$

$n_1 = n_2 = n_3 = 6, k = 3$

$SSB = 6(5 - 8)^2 + 6(9 - 8)^2 + 6(10 - 8)^2$

$= 84$

$df_1 = k - 1 = 2$

| Fertilizer 1 | $(X - 5)^2$ | Fertilizer 2 | $(X - 9)^2$ | Fertilizer 3 | $(X - 10)^2$ |
|---|---|---|---|---|---|
| 6 | 1 | 8 | 1 | 13 | 9 |
| 8 | 9 | 12 | 9 | 9 | 1 |
| 4 | 1 | 9 | 0 | 11 | 1 |
| 5 | 0 | 11 | 4 | 8 | 4 |
| 3 | 4 | 6 | 9 | 7 | 9 |
| 4 | 1 | 8 | 1 | 12 | 4 |
| $\overline{X}_1 = 5$ | Total = 16 | $\overline{X}_1 = 9$ | Total = 24 | $\overline{X}_1 = 10$ | Total = 28 |

SSE = 16 + 24 + 28 = 68

N = 18

$df_2 = N - k = 18 - 3 = 15$

$MSB = SSB / df_1 = 84 / 2 = 42$

$MSE = SSE / df_2 = 68 / 15 = 4.53$

ANOVA test statistic, f = MSB / MSE = 42 / 4.53 = 9.33

Using the f table at $\alpha = 0.05$ the critical value is given as F(0.05, 2, 15) = 3.68

As f > F, thus, the null hypothesis is rejected and it can be concluded that there is a difference in the mean growth of the plants.

**Answer:** Reject the null hypothesis

## F-table of Critical Values of $\alpha = 0.05$ for F(df1, df2)

| DF2=1 | DF1=1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 233.99 | 236.77 | 238.88 | 240.54 | 241.88 | 243.91 | 245.95 | 248.01 | 249.05 | 250.10 | 251.14 | 252.20 | 253.25 | 254.31 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 | 19.40 | 19.41 | 19.43 | 19.45 | 19.45 | 19.46 | 19.47 | 19.48 | 19.49 | 19.50 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 | 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.57 | 8.55 | 8.53 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 5.63 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.43 | 4.40 | 4.37 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.74 | 3.70 | 3.67 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.30 | 3.27 | 3.23 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.01 | 2.97 | 2.93 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.79 | 2.75 | 2.71 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.62 | 2.58 | 2.54 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.79 | 2.72 | 2.65 | 2.61 | 2.57 | 2.53 | 2.49 | 2.45 | 2.40 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.38 | 2.34 | 2.30 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 | 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.30 | 2.25 | 2.21 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 | 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.22 | 2.18 | 2.13 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.48 | 2.40 | 2.33 | 2.29 | 2.25 | 2.20 | 2.16 | 2.11 | 2.07 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 | 2.42 | 2.35 | 2.28 | 2.24 | 2.19 | 2.15 | 2.11 | 2.06 | 2.01 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 | 2.38 | 2.31 | 2.23 | 2.19 | 2.15 | 2.10 | 2.06 | 2.01 | 1.96 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.34 | 2.27 | 2.19 | 2.15 | 2.11 | 2.06 | 2.02 | 1.97 | 1.92 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 | 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 | 1.98 | 1.93 | 1.88 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 | 2.28 | 2.20 | 2.12 | 2.08 | 2.04 | 1.99 | 1.95 | 1.90 | 1.84 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 | 2.25 | 2.18 | 2.10 | 2.05 | 2.01 | 1.96 | 1.92 | 1.87 | 1.81 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 | 2.23 | 2.15 | 2.07 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.78 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 | 2.20 | 2.13 | 2.05 | 2.01 | 1.96 | 1.91 | 1.86 | 1.81 | 1.76 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 | 2.18 | 2.11 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.79 | 1.73 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 | 2.16 | 2.09 | 2.01 | 1.96 | 1.92 | 1.87 | 1.82 | 1.77 | 1.71 |
| 26 | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 | 2.15 | 2.07 | 1.99 | 1.95 | 1.90 | 1.85 | 1.80 | 1.75 | 1.69 |
| 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 | 2.20 | 2.13 | 2.06 | 1.97 | 1.93 | 1.88 | 1.84 | 1.79 | 1.73 | 1.67 |
| 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 | 2.19 | 2.12 | 2.04 | 1.96 | 1.91 | 1.87 | 1.82 | 1.77 | 1.71 | 1.65 |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 | 2.10 | 2.03 | 1.94 | 1.90 | 1.85 | 1.81 | 1.75 | 1.70 | 1.64 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 | 2.09 | 2.01 | 1.93 | 1.89 | 1.84 | 1.79 | 1.74 | 1.68 | 1.62 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 | 2.00 | 1.92 | 1.84 | 1.79 | 1.74 | 1.69 | 1.64 | 1.58 | 1.51 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 | 1.92 | 1.84 | 1.75 | 1.70 | 1.65 | 1.59 | 1.53 | 1.47 | 1.39 |
| 120 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 | 1.91 | 1.83 | 1.75 | 1.66 | 1.61 | 1.55 | 1.50 | 1.43 | 1.35 | 1.25 |
| ∞ | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.88 | 1.83 | 1.75 | 1.67 | 1.57 | 1.52 | 1.46 | 1.39 | 1.32 | 1.22 | 1.00 |

| | F-table of Critical Values of α = 0.10 for F(df1, df2) | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DF1=1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
| DF2=1 | 39.86 | 49.50 | 53.59 | 55.83 | 57.24 | 58.20 | 58.91 | 59.44 | 59.86 | 60.19 | 60.71 | 61.22 | 61.74 | 62.00 | 62.26 | 62.53 | 62.79 | 63.06 | 63.33 |
| 2 | 8.53 | 9.00 | 9.16 | 9.24 | 9.29 | 9.33 | 9.35 | 9.37 | 9.38 | 9.39 | 9.41 | 9.42 | 9.44 | 9.45 | 9.46 | 9.47 | 9.47 | 9.48 | 9.49 |
| 3 | 5.54 | 5.46 | 5.39 | 5.34 | 5.31 | 5.28 | 5.27 | 5.25 | 5.24 | 5.23 | 5.22 | 5.20 | 5.18 | 5.18 | 5.17 | 5.16 | 5.15 | 5.14 | 5.13 |
| 4 | 4.54 | 4.32 | 4.19 | 4.11 | 4.05 | 4.01 | 3.98 | 3.95 | 3.94 | 3.92 | 3.90 | 3.87 | 3.84 | 3.83 | 3.82 | 3.80 | 3.79 | 3.78 | 3.76 |
| 5 | 4.06 | 3.78 | 3.62 | 3.52 | 3.45 | 3.40 | 3.37 | 3.34 | 3.32 | 3.30 | 3.27 | 3.24 | 3.21 | 3.19 | 3.17 | 3.16 | 3.14 | 3.12 | 3.11 |
| 6 | 3.78 | 3.46 | 3.29 | 3.18 | 3.11 | 3.05 | 3.01 | 2.98 | 2.96 | 2.94 | 2.90 | 2.87 | 2.84 | 2.82 | 2.80 | 2.78 | 2.76 | 2.74 | 2.72 |
| 7 | 3.59 | 3.26 | 3.07 | 2.96 | 2.88 | 2.83 | 2.78 | 2.75 | 2.72 | 2.70 | 2.67 | 2.63 | 2.59 | 2.58 | 2.56 | 2.54 | 2.51 | 2.49 | 2.47 |
| 8 | 3.46 | 3.11 | 2.92 | 2.81 | 2.73 | 2.67 | 2.62 | 2.59 | 2.56 | 2.54 | 2.50 | 2.46 | 2.42 | 2.40 | 2.38 | 2.36 | 2.34 | 2.32 | 2.29 |
| 9 | 3.36 | 3.01 | 2.81 | 2.69 | 2.61 | 2.55 | 2.51 | 2.47 | 2.44 | 2.42 | 2.38 | 2.34 | 2.30 | 2.28 | 2.25 | 2.23 | 2.21 | 2.18 | 2.16 |
| 10 | 3.29 | 2.92 | 2.73 | 2.61 | 2.52 | 2.46 | 2.41 | 2.38 | 2.35 | 2.32 | 2.28 | 2.24 | 2.20 | 2.18 | 2.16 | 2.13 | 2.11 | 2.08 | 2.06 |
| 11 | 3.23 | 2.86 | 2.66 | 2.54 | 2.45 | 2.39 | 2.34 | 2.30 | 2.27 | 2.25 | 2.21 | 2.17 | 2.12 | 2.10 | 2.08 | 2.05 | 2.03 | 2.00 | 1.97 |
| 12 | 3.18 | 2.81 | 2.61 | 2.48 | 2.39 | 2.33 | 2.28 | 2.24 | 2.21 | 2.19 | 2.15 | 2.10 | 2.06 | 2.04 | 2.01 | 1.99 | 1.96 | 1.93 | 1.90 |
| 13 | 3.14 | 2.76 | 2.56 | 2.43 | 2.35 | 2.28 | 2.23 | 2.20 | 2.16 | 2.14 | 2.10 | 2.05 | 2.01 | 1.98 | 1.96 | 1.93 | 1.90 | 1.88 | 1.85 |
| 14 | 3.10 | 2.73 | 2.52 | 2.39 | 2.31 | 2.24 | 2.19 | 2.15 | 2.12 | 2.10 | 2.05 | 2.01 | 1.96 | 1.94 | 1.91 | 1.89 | 1.86 | 1.83 | 1.80 |
| 15 | 3.07 | 2.70 | 2.49 | 2.36 | 2.27 | 2.21 | 2.16 | 2.12 | 2.09 | 2.06 | 2.02 | 1.97 | 1.92 | 1.90 | 1.87 | 1.85 | 1.82 | 1.79 | 1.76 |
| 16 | 3.05 | 2.67 | 2.46 | 2.33 | 2.24 | 2.18 | 2.13 | 2.09 | 2.06 | 2.03 | 1.99 | 1.94 | 1.89 | 1.87 | 1.84 | 1.81 | 1.78 | 1.75 | 1.72 |
| 17 | 3.03 | 2.64 | 2.44 | 2.31 | 2.22 | 2.15 | 2.10 | 2.06 | 2.03 | 2.00 | 1.96 | 1.91 | 1.86 | 1.84 | 1.81 | 1.78 | 1.75 | 1.72 | 1.69 |
| 18 | 3.01 | 2.62 | 2.42 | 2.29 | 2.20 | 2.13 | 2.08 | 2.04 | 2.00 | 1.98 | 1.93 | 1.89 | 1.84 | 1.81 | 1.78 | 1.75 | 1.72 | 1.69 | 1.66 |
| 19 | 2.99 | 2.61 | 2.40 | 2.27 | 2.18 | 2.11 | 2.06 | 2.02 | 1.98 | 1.96 | 1.91 | 1.86 | 1.81 | 1.79 | 1.76 | 1.73 | 1.70 | 1.67 | 1.63 |
| 20 | 2.97 | 2.59 | 2.38 | 2.25 | 2.16 | 2.09 | 2.04 | 2.00 | 1.96 | 1.94 | 1.89 | 1.84 | 1.79 | 1.77 | 1.74 | 1.71 | 1.68 | 1.64 | 1.61 |
| 21 | 2.96 | 2.57 | 2.36 | 2.23 | 2.14 | 2.08 | 2.02 | 1.98 | 1.95 | 1.92 | 1.87 | 1.83 | 1.78 | 1.75 | 1.72 | 1.69 | 1.66 | 1.62 | 1.59 |
| 22 | 2.95 | 2.56 | 2.35 | 2.22 | 2.13 | 2.06 | 2.01 | 1.97 | 1.93 | 1.90 | 1.86 | 1.81 | 1.76 | 1.73 | 1.70 | 1.67 | 1.64 | 1.60 | 1.57 |
| 23 | 2.94 | 2.55 | 2.34 | 2.21 | 2.11 | 2.05 | 1.99 | 1.95 | 1.92 | 1.89 | 1.84 | 1.80 | 1.74 | 1.72 | 1.69 | 1.66 | 1.62 | 1.59 | 1.55 |
| 24 | 2.93 | 2.54 | 2.33 | 2.19 | 2.10 | 2.04 | 1.98 | 1.94 | 1.91 | 1.88 | 1.83 | 1.78 | 1.73 | 1.70 | 1.67 | 1.64 | 1.61 | 1.57 | 1.53 |
| 25 | 2.92 | 2.53 | 2.32 | 2.18 | 2.09 | 2.02 | 1.97 | 1.93 | 1.89 | 1.87 | 1.82 | 1.77 | 1.72 | 1.69 | 1.66 | 1.63 | 1.59 | 1.56 | 1.52 |
| 26 | 2.91 | 2.52 | 2.31 | 2.17 | 2.08 | 2.01 | 1.96 | 1.92 | 1.88 | 1.86 | 1.81 | 1.76 | 1.71 | 1.68 | 1.65 | 1.61 | 1.58 | 1.54 | 1.50 |
| 27 | 2.90 | 2.51 | 2.30 | 2.17 | 2.07 | 2.00 | 1.95 | 1.91 | 1.87 | 1.85 | 1.80 | 1.75 | 1.70 | 1.67 | 1.64 | 1.60 | 1.57 | 1.53 | 1.49 |
| 28 | 2.89 | 2.50 | 2.29 | 2.16 | 2.06 | 2.00 | 1.94 | 1.90 | 1.87 | 1.84 | 1.79 | 1.74 | 1.69 | 1.66 | 1.63 | 1.59 | 1.56 | 1.52 | 1.48 |
| 29 | 2.89 | 2.50 | 2.28 | 2.15 | 2.06 | 1.99 | 1.93 | 1.89 | 1.86 | 1.83 | 1.78 | 1.73 | 1.68 | 1.65 | 1.62 | 1.58 | 1.55 | 1.51 | 1.47 |
| 30 | 2.88 | 2.49 | 2.28 | 2.14 | 2.05 | 1.98 | 1.93 | 1.88 | 1.85 | 1.82 | 1.77 | 1.72 | 1.67 | 1.64 | 1.61 | 1.57 | 1.54 | 1.50 | 1.46 |
| 40 | 2.84 | 2.44 | 2.23 | 2.09 | 2.00 | 1.93 | 1.87 | 1.83 | 1.79 | 1.76 | 1.71 | 1.66 | 1.61 | 1.57 | 1.54 | 1.51 | 1.47 | 1.42 | 1.38 |
| 60 | 2.79 | 2.39 | 2.18 | 2.04 | 1.95 | 1.87 | 1.82 | 1.77 | 1.74 | 1.71 | 1.66 | 1.60 | 1.54 | 1.51 | 1.48 | 1.44 | 1.40 | 1.35 | 1.29 |
| 120 | 2.75 | 2.35 | 2.13 | 1.99 | 1.90 | 1.82 | 1.77 | 1.72 | 1.68 | 1.65 | 1.60 | 1.55 | 1.48 | 1.45 | 1.41 | 1.37 | 1.32 | 1.26 | 1.19 |
| ∞ | 2.71 | 2.30 | 2.08 | 1.94 | 1.85 | 1.77 | 1.72 | 1.67 | 1.63 | 1.60 | 1.55 | 1.49 | 1.42 | 1.38 | 1.34 | 1.30 | 1.24 | 1.17 | 1.00 |

# Confidence Intervals

How much uncertainty is associated with a point estimate of a population parameter?

An interval estimate provides **more information about a population** characteristic than does a point estimate

Such interval estimates are called **confidence intervals**

## Confidence Interval Estimate

An interval gives a **range of values**:

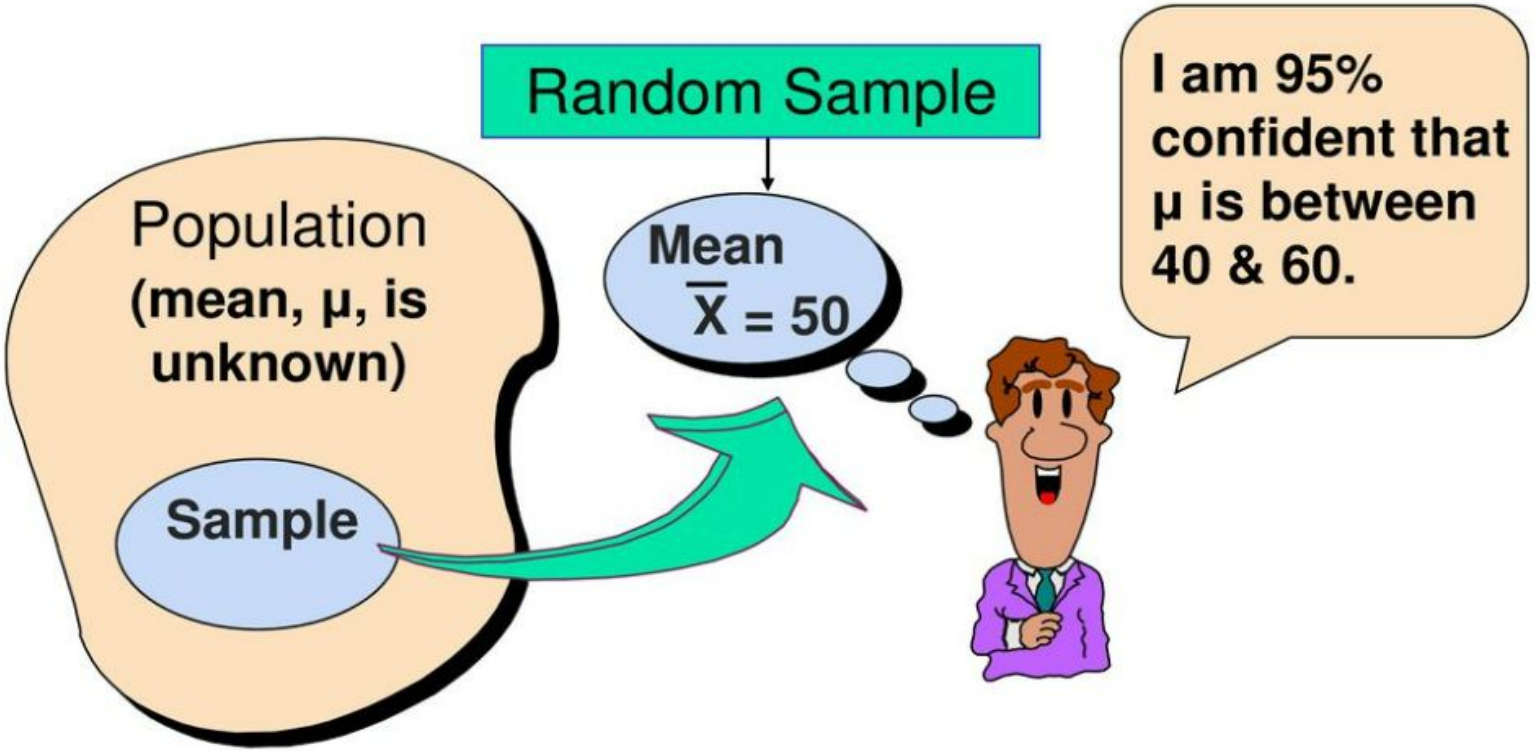
- Takes into consideration variation in sample statistics from sample to sample
- Based on observation from 1 sample
- Gives information about closeness to unknown population parameters
- Stated in terms of level of confidence
- Can never be 100% confident

General Formula

The general formula for all confidence intervals is:

Point Estimate ± (Critical Value) Standard Error

# Estimation Process



Confidence Level, $(1-\propto)$ Suppose confidence level = 95%

Suppose confidence level = 95%

Also written $(1 - \propto) = .95$

A relative frequency interpretation:

In the long run, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter

A specific interval either will contain or will not contain the true parameter

No probability involved in a specific interval

# Confidence Interval for μ (σ Known)

- Assumptions
  - Population standard deviation σ is known
  - Population is normally distributed
  - If population is not normal, use large sample
- Confidence interval estimate:

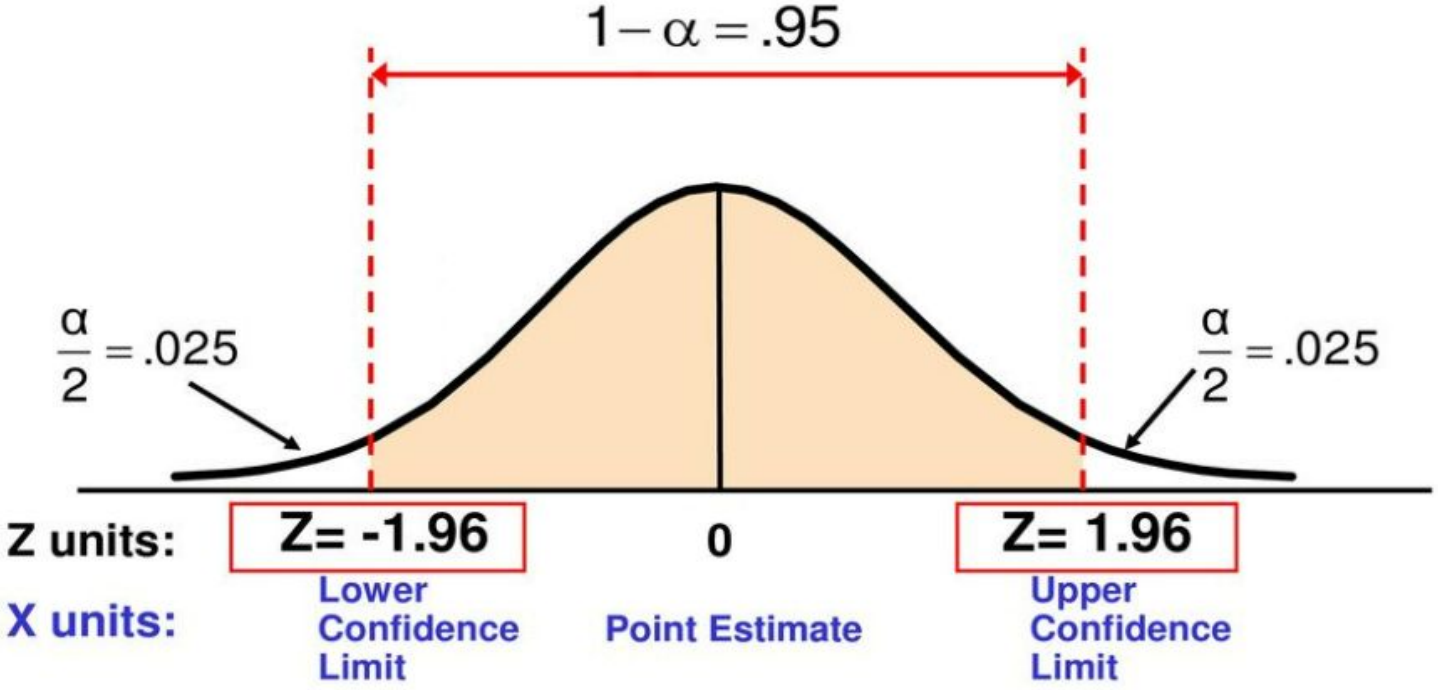
$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

(where Z is the normal distribution critical value for a probability of α/2 in each tail)

# Finding the Critical Value, Z

$$Z = \pm 1.96$$

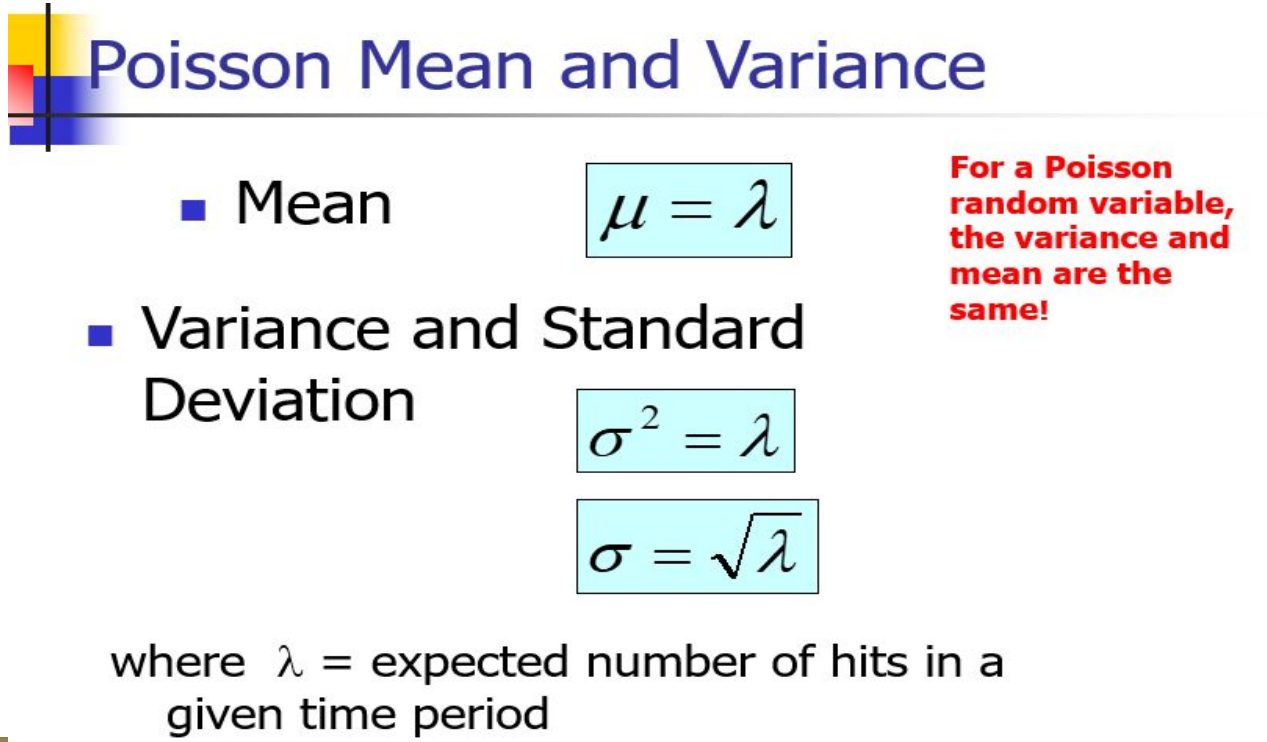
- Consider a 95% confidence interval:



**Critical Value:** Denoted by C and it is a value in the distribution beyond which leads to the rejection of the Null Hypothesis.

It is compared to the test statistic.

Poisson distribution is for counts—if **events happen at a constant rate over time**, the Poisson distribution gives the probability of X number of events occurring in time T.



# Poisson distribution

If events happen independently of each other, with **average number of events in some fixed interval $\lambda$**, then the distribution of the number of events $k$ in that interval is Poisson.

A random variable $X$ has the Poisson distribution with parameter $\lambda(> 0)$ if

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!} \qquad (k = 0,1,2, \dots)$$

Example: On average lightning kills three people each year in the UK, $\lambda=3$. What is the probability that only one person is killed this year?

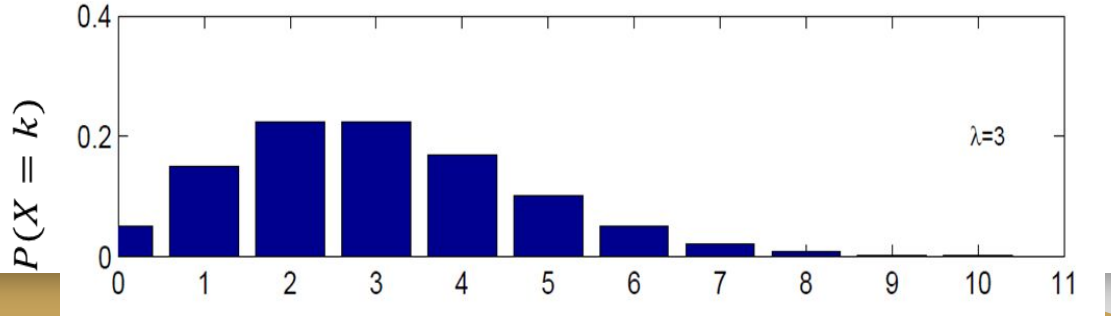
*Answer*:

Assuming these are independent random events, the number of people killed in a given year therefore has a Poisson distribution:

Let the random variable x be the number of people killed in a year.

Poisson distribution $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$ with $\lambda = 3$

$$\Rightarrow P(X = 1) = \frac{e^{-3}3^1}{1!} \approx 0.15$$



# “Poisson Process” (rates)

Note that the Poisson parameter can be given as the mean number of **events that occur in a defined time period(fixed interval)** OR,

equivalently, **can be given as a rate**, such as v=2/month (2 events per 1 month) that must be multiplied by t=time (called a “Poisson Process”)--> X ~ Poisson ()

$\lambda = vt$ v: rate, t: time

***Example:* Telecommunications**

Messages arrive at a switching centre at random and at an average rate of 1.2 per second.

(a) Find the probability of **5 messages arriving in a 2-sec interval.**

(b) For how long can the operation of the centre be interrupted, if the probability of losing one or more messages is to be no more than 0.05?

***Answer:***

Times of arrivals form a Poisson process, rate $v=1.2$/sec.

(a) Let $Y=5$ number of messages arriving in a $t=2$-sec interval.

Then $Y \sim$ Poisson, mean number $\lambda=vt=1.2\times2=2.4$

$$P(Y=k=5)=\frac{e^{-\lambda}\lambda^{k}}{k!}=\frac{e^{-2.4}2.4^{5}}{5!}=0.060$$

(b) For **how long** can the operation of the centre be interrupted, if the probability of **losing one or more messages** is to be **no more than 0.05**?

*Answer:*

(b) Let the required time = ***t* seconds**. Average rate of arrival is **1.2/second**.

Let k = number of messages in *t* seconds, so that

k~ Poisson, with $\lambda = 1.2 \times t = 1.2t$

Want P(At least one message) = $P(k \geq 1) = 1 - P(k = 0) \leq 0.05$

$$P(k = 0) = \frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-1.2t}(1.2t)^0}{0!} = e^{-1.2t}$$

$$\Rightarrow 1 - e^{-1.2t} \leq 0.05$$

$$\Rightarrow \quad -e^{-1.2t} \leq 0.05 - 1$$

$$\Rightarrow e^{-1.2t} \geq 0.95$$

$$\Rightarrow -1.2t \geq \ln(0.95) = -0.05129$$

$$\boxed{\Rightarrow t \leq 0.043 \text{ seconds}}$$

It's the logarithm function, to the base e.

So if e^x = a, then **ln(a) = x.**

If calls to your cell phone are a Poisson process with a constant rate λ=2 calls per hour, what's the probability that, if you forget to turn your phone off in a 1.5 hour movie, your phone rings during that time?

1b. How many phone calls do you expect to get during the movie?

X ~ Poisson (v or λ=2 calls/hour), t=1.5 hrs λ=λt=2*1.5

**P(X≥1)=1 – P(X=0)**

$$P(X=0)=\frac{(2*1.5)^0 e^{-2(1.5)}}{0!}\frac{(3)^0 e^{-3}}{0!}=e^{-3}=.05$$

P(X≥1)=1 – .05 = 95% chance

1b. How many phone calls do you expect to get during the movie?

$$E(X) = \lambda t = 2(1.5) = 3$$

if new cases of West Nile in New England are occurring at a rate of about **2 per month**, then what's the probability that exactly **4 cases will occur in the next 3 months**?

X ~ Poisson (v or λ =2/month)  t=3 months  k=4

$$P(X = 4 \text{ in } 3 \text{ months}) = \frac{(2*3)^4 e^{-(2*3)}}{4!} = \frac{6^4 e^{-(6)}}{4!} = 13.4\%$$

Exactly 6 cases?  k=6

$$P(X = 6 \text{ in } 3 \text{ months}) = \frac{(2*3)^6 e^{-(2*3)}}{6!} = \frac{6^6 e^{-(6)}}{6!} = 16\%$$

# 1)

Patients arrive at a hospital accident and emergency department at random at a rate of 6 per hour.

(a) Find the probability that, during any 90 minute period, the number of patients arriving at the hospital accident and emergency department is

(i) exactly 7

(ii) at least 10

**(5)**

A patient arrives at 11.30 a.m.

(b) Find the probability that the next patient arrives before 11.45 a.m.

**(3)**

Given ,
rate=6/hr
t=90 min=1.5 hr

λ=6*1.5=9

$$\frac{e^{-\lambda}\lambda^{k}}{k!}$$ :

A i) x=k=7
P(x=k=7)=0.117

A ii) P(x>=10)=1-P(x<=9)
= 1-0.587 =o.413

b) t=15 min=0.4 hr
rate=6/hr
λ=6*0.4=1.5

P(x>=1)=1-P(x=0)
=0.777

An online shop sells a computer game at an average rate of 1 per day.

(a) Find the probability that the shop sells more than 10 games in a 7 day period.

**(3)**

Once every 7 days the shop has games delivered before it opens.

(b) Find the least number of games the shop should have in stock immediately after a delivery so that the probability of running out of the game before the next delivery is less than 0.05

**(3)**

In an attempt to increase sales of the computer game, the price is reduced for six months. A random sample of 28 days is taken from these six months. In the sample of 28 days, 36 computer games are sold.

(c) Using a suitable approximation and a 5% level of significance, test whether or not the average rate of sales per day has increased during these six months. State your hypotheses clearly.

**(7)**

https://youtu.be/YolZTZDSu_I
https://youtu.be/BJPcDV0rx88
https://youtu.be/f8XueW9LKgo
https://youtu.be/oB0l6JgW4kI

# What Is a Chi-Square Test?**(not in syllabus)**

This test was introduced by **Karl Pearson in 1900 for categorical data analysis and distribution.** So it was mentioned as Pearson's chi-squared test.

The Chi-Square test is a statistical procedure for determining the difference **between observed and expected data.**

This test can also be used to determine whether **it correlates to the categorical variables** in our data.

A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable.

Nonparametric tests are used for data that don't follow the assumptions of parametric tests, especially the assumption of a normal distribution.

Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal.

They cannot have a normal distribution since they can only have a few particular values.

# When to use a chi-square test

A Pearson's chi-square test may be an appropriate option for your data if all of the following are true:

1. You want to test a hypothesis about one or more categorical variables. If one or more of your variables is quantitative, you should use a different statistical test. Alternatively, you could convert the quantitative variable into a categorical variable by separating the observations into intervals.
2. The sample was randomly selected from the population.
3. There are a minimum of five observations expected in each group or combination of groups

It is used to determine whether your data are significantly different from what you expected.

There are two types of Pearson's chi-square tests:

**The chi-square goodness of fit test** is used to test whether the frequency distribution of a categorical variable is different from your expectations.

- when you have one categorical variable

**The chi-square test of independence** is used to test whether two categorical variables are related to each other.

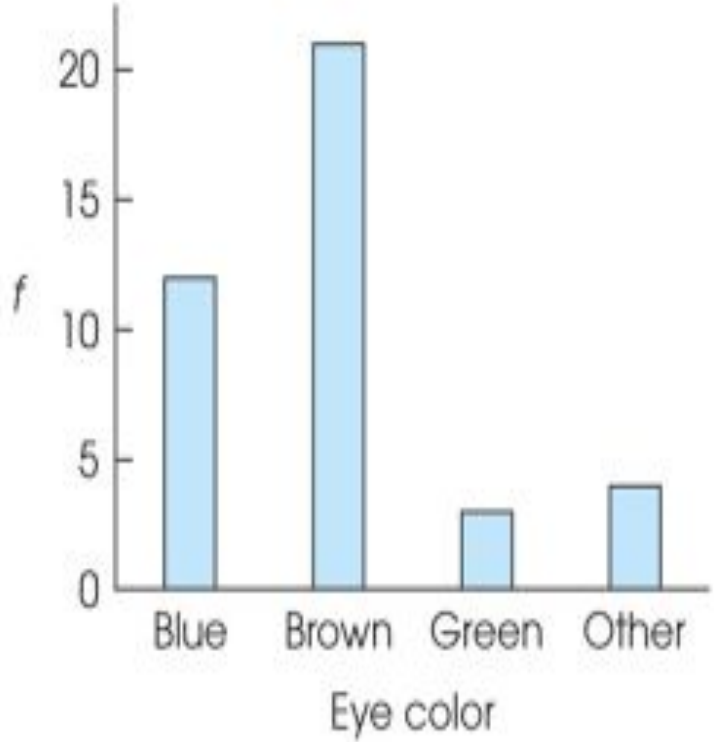
# The chi-square test for goodness-of-fit

The chi-square test for goodness-of-fit uses frequency data from a sample to test hypotheses about the shape or proportions of a population.

Each individual in the sample is classified into one of the category on the scale of measurement.

The data, called observed frequencies, simply count how many individuals from the sample are in each category.

The null hypothesis specifies the **proportion** of the population that should be in each category.




| Eye color (X) | f |
|---|---|
| Blue | 12 |
| Brown | 21 |
| Green | 3 |
| Other | 4 |

| Blue | Brown | Green | Other |
|---|---|---|---|
| 12 | 21 | 3 | 4 |

| Bird species | Frequency |
|---|---|
| House sparrow | 15 |
| House finch | 12 |
| Black-capped chickadee | 9 |
| Common grackle | 8 |
| European starling | 8 |
| Mourning dove | 6 |

Both the types(**goodness of fit test, test of Independence)** test whether

the observed frequency distribution of a categorical variable

is significantly different from

its expected frequency distribution.

A frequency distribution describes how observations are distributed between different groups.

Frequency distributions are often displayed using frequency distribution tables.

A frequency distribution table shows the number of observations in each group.

When there are **two categorical variables**, you can use a specific type of frequency distribution table called a **contingency table** to show the number of observations in each combination of groups.

**Frequency of visits by bird species at a bird feeder during a 24-hour period**

| Bird species | Frequency |
|---|---|
| House sparrow | 15 |
| House finch | 12 |
| Black-capped chickadee | 9 |
| Common grackle | 8 |
| European starling | 8 |
| Mourning dove | 6 |

**Contingency table of the handedness of a sample of Americans and Canadians**

| | Right-handed | Left-handed |
|---|---|---|
| **American** | 236 | 19 |
| **Canadian** | 157 | 16 |

## Contingency Table

Let A and B be the attributes of the given data.

Let the data be classified into S classes A1, A2,AS according to attribute A and into t classes B1, B2, ...Bt, according to the attribute B.

Let Oij be the observed frequency of the cell belonging to the classes Ai (i = 1, 2, ..., s) and Bj (j = 1, 2, t).

The data can be set into a s x t contingency table of s rows and t columns as follows:

| Classes | $B_1$ | $B_2$ | ... | $B_j$ | ... | $B_t$ | Total |
|---|---|---|---|---|---|---|---|
| $A_i$ | $O_{1i}$ | $O_{12}$ | ... | $O_{1j}$ | ... | $O_{1t}$ | $A_1$ |
| $A_2$ | $O_{21}$ | $O_{22}$ | ... | $O_{2j}$ | ... | $O_{2t}$ | $A_2$ |
| : | | : | | | | | |
| $A_i$ | | | | | | | |
| $A_s$ | $O_{s1}$ | $O_{s2}$ | ... | $O_{5j}$ | ... | $O_{st}$ | $A_s$ |
| Total | $B_1$ | $B_2$ | ... | $B_j$ | ... | $B_t$ | N |

Suppose I ask you to pick five integers that sum to 100.

The requirement of summing to 100 is a **restriction on your number choices.**

For the first number, you can choose any integer you want. Whatever your choice, the sum of the five numbers can still be 100. This is also true of the second, third, and fourth numbers.

You have no choice for the final number; it has only one possible value and it isn't free to vary. For example, imagine you chose 15, 27, 42, and 3 as your first four numbers. For the numbers to sum to 100, the final number needs to be 13.

Due to the restriction, you could only choose four of the five numbers. The first four numbers were free to vary. In contrast, the fifth number wasn't free to vary; it depended on the other four numbers.

Consider a data sample consisting of five positive integers. The values of the five integers must have an average of six. If four of the items within the data set are {3, 8, 5, and 4}, the fifth number must be 10. Because the first four numbers can be chosen at random, the degrees of freedom is four.

The number of **independent pieces of information** used to calculate the statistic is called the degrees of freedom.

The degrees of freedom of a statistic depend on the sample size:

When the sample size is **small**, there are only a few independent pieces of information, and therefore only a few degrees of freedom.

When the sample size is **large**, there are many independent pieces of information, and therefore many degrees of freedom.

## Degrees of Freedom

The term degrees of Freedom refers to the number of **"independent constraints"** in a set of data.

(1) If the data are given in a contingency table, then the degree of freedom is calculated by the formula **Y=(c-1) (r-1).**

Where Y stands for degree of freedom, c for number of columns and r for the number of rows. Thus in a 2x 2 table, degrees of freedom are (2-1). (2-1) = 1 and so on.

The degrees of freedom can be calculated to ensure that chi-square tests are statistically valid.

(2) If data is given in frequencies the degree of freedom**=n-1** where n=no of frequencies

# Formula For Chi-Square Test

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

Where

c = Degrees of freedom

O = Observed Value

E = Expected Value

NOTE: This formula is used for both one-way and two way chi-square tests

## How to perform a chi-square test

$$\frac{\Sigma (O_i - E_i)^2}{E_i}$$

1. Create a table of the observed and expected frequencies. This can sometimes be the most difficult step because you will need to carefully consider which expected values are most appropriate for your null hypothesis.
2. Calculate the chi-square value from your observed and expected frequencies using the chi-square formula.
3. Find the critical chi-square value in **a chi-square critical value table** or using statistical software.
4. Compare the chi-square value to the critical value to determine which is larger.
5. Decide whether to reject the null hypothesis. You should reject the null hypothesis if the chi-square value is **greater** than the critical value. If you reject the null hypothesis, you can conclude that your data are s**ignificantly different from what you expected.**

| month | observed | expected | O-E | $(O-E)^2$ | $(O-E)^2/E$ |
|---|---|---|---|---|---|
| 1 | 12 | 10 | | | |
| 2 | 8 | 10 | | | |
| 3 | 20 | 10 | | | |
| 4 | 2 | 10 | | | |
| 5 | 14 | 10 | | | |
| 6 | 10 | 10 | | | |
| 7 | 15 | 10 | | | |
| 8 | 6 | 10 | | | |
| 9 | 9 | 10 | | | |
| 10 | 4 | 10 | | | |
| ToT | 100 | 10 | 0 | - | 26.6 |

The number of scooter accidents per month in a certain town were as follows: 12, 8, 20, 2, 14, 10, 15, 6, 9,4 Are these frequencies in agreement with the belief that **accidently conditions were the same during this 10 month period**

Null Hypothesis: Ho: The given frequencies (i.e. number of accidents per month) are consistent with the belief that the accident conditions were same during the 10-month period.

total number of accidents = 12+8+20+2+14+10+15+6+9+4=100,

Under null hypothesis,

Expected number of accidents for each of the 10 months= 100/10=10, I.e. these accidents are uniformly distributed

d.f=10-1=9

$X^2_{0.05}$ for 9 d.f=16.191

## Critical values of the Chi-square distribution with *d* degrees of freedom

| | Probability of exceeding the critical value | | | | | | |
|---|---|---|---|---|---|---|---|
| *d* | 0.05 | 0.01 | 0.001 | *d* | 0.05 | 0.01 | 0.001 |
| 1 | 3.841 | 6.635 | 10.828 | 11 | 19.675 | 24.725 | 31.264 |
| 2 | 5.991 | 9.210 | 13.816 | 12 | 21.026 | 26.217 | 32.910 |
| 3 | 7.815 | 11.345 | 16.266 | 13 | 22.362 | 27.688 | 34.528 |
| 4 | 9.488 | 13.277 | 18.467 | 14 | 23.685 | 29.141 | 36.123 |
| 5 | 11.070 | 15.086 | 20.515 | 15 | 24.996 | 30.578 | 37.697 |
| 6 | 12.592 | 16.812 | 22.458 | 16 | 26.296 | 32.000 | 39.252 |
| 7 | 14.067 | 18.475 | 24.322 | 17 | 27.587 | 33.409 | 40.790 |
| 8 | 15.507 | 20.090 | 26.125 | 18 | 28.869 | 34.805 | 42.312 |
| 9 | 16.919 | 21.666 | 27.877 | 19 | 30.144 | 36.191 | 43.820 |
| 10 | 18.307 | 23.209 | 29.588 | 20 | 31.410 | 37.566 | 45.315 |

***INTRODUCTION TO POPULATION GENETICS*, Table D.1**

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

=26.6

Since calculated value of x = 26.6 is greater than tabulated value from (i) = 16.919, it is significant and hence the null hypothesis is rejected at 5% level of significance.

Hence, we conclude that the accident conditions are certainly not uniform over the 10-month period.

A poker-dealing machine is supposed to deal cards at random, as if from an infinite deck.

In a test, you counted 1600 cards, and observed the following:

| | |
|---|---|
| Spades | 404 |
| Hearts | 420 |
| Diamonds | 400 |
| Clubs | 376 |

Could it be that the suits are equally likely? Or are these discrepancies too much to be random?

Null Hypothesis(H0) : Suits are equally distributed

Under null hypothesis
Expected no of Cards in each suit=(404+420+400+376)/4=400

The number of degrees of freedom is 3 (number of categories minus 1=4-1=3).
The critical value is from a table you'll have on the exam (using  = 0.05).

| Suits | observed | expected | O-E | $(O-E)^2$ | $(O-E)^2/E$ |
|---|---|---|---|---|---|
| Spades | 404 | 400 | | | |
| Hearts | 420 | 400 | | | |
| Diamonds | 400 | 400 | | | |
| Clubs | 376 | 400 | | | |
| ToT | 1600 | 1600 | | | 2.480 |

The critical value is from a table(using α = 0.05).= 7.815

Conclusion: Null Hypothesis is accepted as 2.48<7.815

## Critical values of the Chi-square distribution with *d* degrees of freedom

| | Probability of exceeding the critical value | | | | | | |
|---|---|---|---|---|---|---|---|
| *d* | 0.05 | 0.01 | 0.001 | *d* | 0.05 | 0.01 | 0.001 |
| 1 | 3.841 | 6.635 | 10.828 | 11 | 19.675 | 24.725 | 31.264 |
| 2 | 5.991 | 9.210 | 13.816 | 12 | 21.026 | 26.217 | 32.910 |
| 3 | 7.815 | 11.345 | 16.266 | 13 | 22.362 | 27.688 | 34.528 |
| 4 | 9.488 | 13.277 | 18.467 | 14 | 23.685 | 29.141 | 36.123 |
| 5 | 11.070 | 15.086 | 20.515 | 15 | 24.996 | 30.578 | 37.697 |
| 6 | 12.592 | 16.812 | 22.458 | 16 | 26.296 | 32.000 | 39.252 |
| 7 | 14.067 | 18.475 | 24.322 | 17 | 27.587 | 33.409 | 40.790 |
| 8 | 15.507 | 20.090 | 26.125 | 18 | 28.869 | 34.805 | 42.312 |
| 9 | 16.919 | 21.666 | 27.877 | 19 | 30.144 | 36.191 | 43.820 |
| 10 | 18.307 | 23.209 | 29.588 | 20 | 31.410 | 37.566 | 45.315 |

*INTRODUCTION TO POPULATION GENETICS*, Table D.1