Unit 4: Data Analysis

M Krishna

Presentation Agenda

- Exploratory Data Analysis,
  - Statistical Estimation
    - Hypothesis Testing
      - Parametric Tests
        - Non-Parametric Tests
          - Multiple Regression
            - Factor Analysis
              - Cluster Analysis

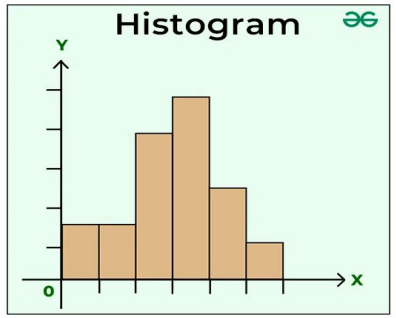
Exploratory Data Analysis

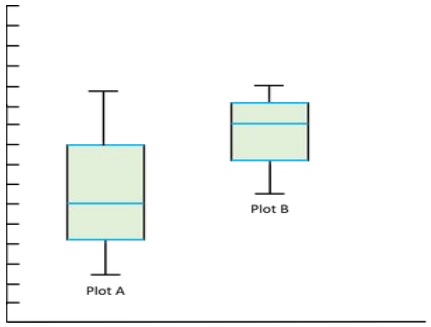
The Goal of data analysis is to gain information from data.

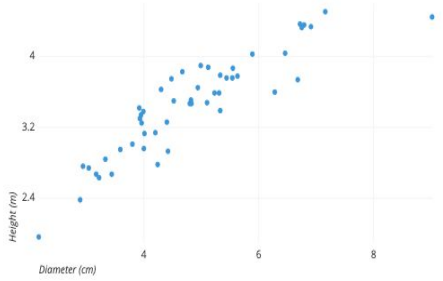
✔ Exploratory Data Analysis: is a crucial step in data analysis that involves summarizing and visualizing the key characteristics of a dataset

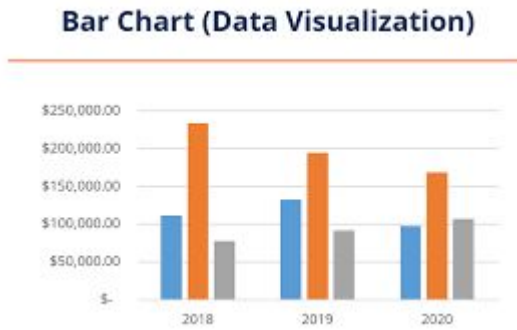
Methods of Exploratory data Analysis

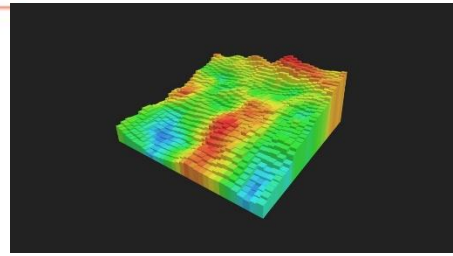
1. **Descriptive Statistics:** Summarize the central tendency, dispersion, and shape of the dataset's distribution using mean, median, mode, variance, and standard deviation.
2. **Data Visualization:** Use graphical tools to understand the data better. Common plots include (Histograms, Box Plots, Scatter Plots, Bar Charts, Heatmaps)





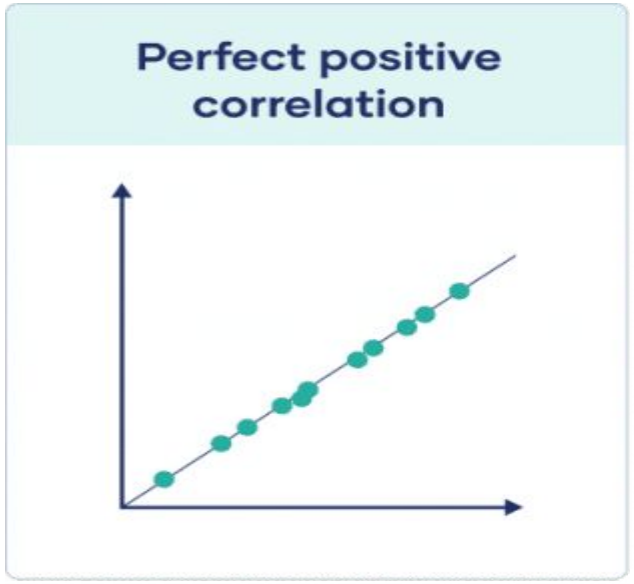


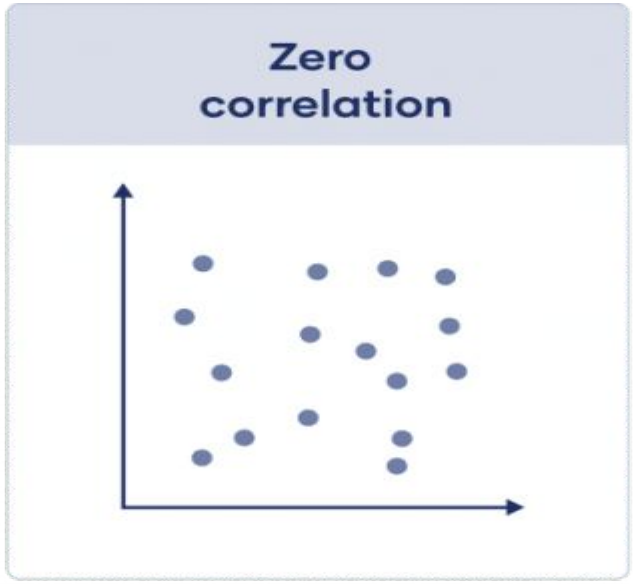


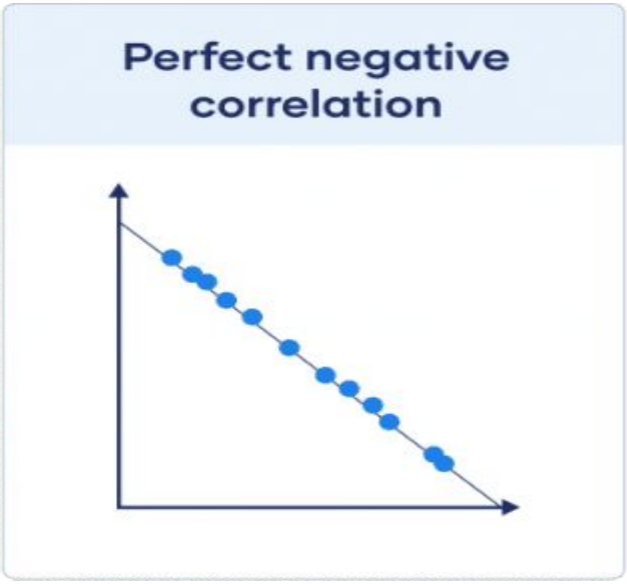


Methods of EDA

- ✔ Data Cleaning: Identify and handle missing values, outliers, and inconsistencies in the data
- ✔ **Correlation Analysis**: Examine the relationships between variables using correlation coefficients and scatter plots.

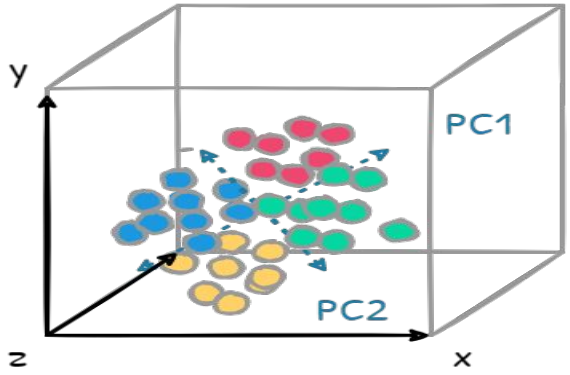


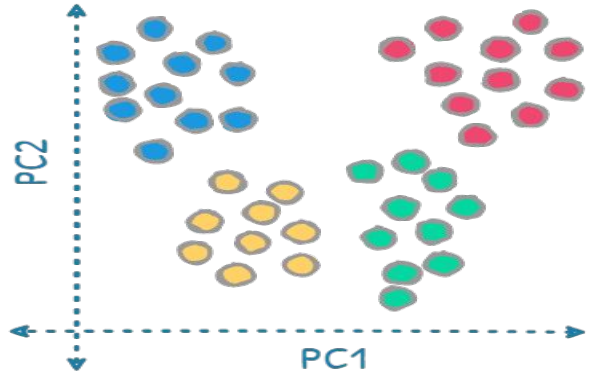


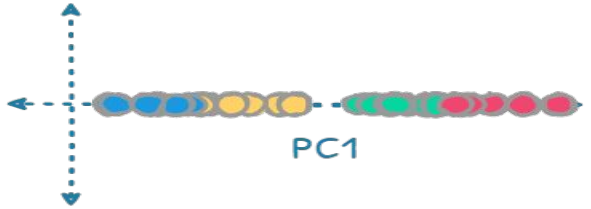


Methods of EDA

✔ **Dimensionality Reduction**: Use techniques like Principal Component Analysis (PCA) to reduce the number of variables while retaining essential information.







Methods of EDA

✔ **Data Transformation**: Apply transformations to make the data more suitable for analysis, such as normalization or scaling.

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

#	Emp	Age	Salary
1	Emp1	44	73000
2	Emp2	27	47000
3	Emp3	30	53000
4	Emp4	38	62000
5	Emp5	40	57000
6	Emp6	35	53000
7	Emp7	48	78000

Normalization

Age	Normalized Age	Salary	Normalized Salary
44	0.80952381	73000	0.838709677
27	0	47000	0
30	0.142857143	53000	0.193548387
38	0.523809524	62000	0.483870968
40	0.619047619	57000	0.322580645
35	0.380952381	53000	0.193548387
48	1	78000	1
	Range 0-1		**Range 0-1**

How to calculate Normalized value?
X = 35, min = 27, max = 48 for column Age.
Xnorm(for 35) = $\frac{35-27}{48-27} = 0.3809$

Methods of EDA

✔ **Summary Statistics:** Look at aggregate statistics to understand the overall patterns and trends in the data

	Count	Mean	Median	Max	Min	S.D.
Company Characteristics						
Later-stage dummy	12670	0.041	0	1	0	0.197
Early-stage dummy	12670	0.463	0	1	0	0.499
Seed stage dummy	12670	0.261	0	1	0	0.439
Expansion stage dummy	12670	0.199	0	1	0	0.399
Age (years)	12670	2.479	1	30	0	6.097
Round number of investors	12670	2.191	2	18	1	1.498
IPO	12670	0.090	0	1	0	0.286
Round amount ($ million)	12670	6.168	3.015	402	0.001	25.989
Lead VC Characteristics						
Fund size ($ million)	12670	219.311	110	7264	0.1	430.189
Number of prior rounds	12670	214.952	78	2821	0	340.892
Market Concentration						
HHI	2212	0.101	0.072	1	0.005	0.096
1/Total VC no	2212	0.092	0.062	1	0.002	0.107
Other Market Characteristics						
VC prior 4 Qt. inflow ($ million)	2212	30107.97	23061.2	97591.1	1803.3	25913.45
Value-weighted industry avg. book-to-market ratio	2212	0.282	0.255	1.343	0.081	0.141

✔ **Exploration of Data Subsets:** Analyze specific subsets of the data to uncover more detailed patterns or differences.

Data Distribution: Mean

The most common measures the mean or average

1. The mean or Average : To calculate the average $\hat{X}$ of the set of observation, add their value and divide by the number of observation

$$\text{Mean}, \bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n}$$

Let the mean of $x_1, x_2, x_3 \ldots\ldots x_n$ be A, then what is the mean of:

1. $(x_1+k), (x_2+k), (x_3+k), \ldots\ldots (x_n+k)$? Mean= A+K
2. $(x_1-k), (x_2-k), (x_3-k), \ldots\ldots (x_n-k)$? Mean = A-K
3. $kx_1, kx_2, kx_3 \ldots\ldots kx_n$? Mean = AK

If the heights of 5 people are 142 cm, 150 cm, 149 cm, 156 cm and 153 cm. Find the mean height.

$$\text{Mean height} = \frac{142+150+149+156+153}{5} = 150$$

Data Distribution: Mean

$$\text{Mean, } \bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots . x_n f_n}{f_+ f_2 + \ldots . f_n}$$

Find the mean of the following distribution

xi	fi	xifi
4	5	20
6	10	60
9	10	90
10	7	70
15	8	120
	$\sum f_i = 40$	$\sum x_i f_i = 360$

$$\text{Mean} = \frac{\sum x_i f_i}{\sum f_i} = \frac{360}{40} = 9$$

Data Distribution: Mean

Find the average number of patients visiting the hospital in a day.

Number of patients	Number of days visiting hospital
0-10	2
10-20	6
20-30	9
30-40	7
40-50	4
50-60	2

Class mark (xi)	frequency (fi)	xifi
5	2	10
15	6	90
25	9	225
35	7	245
45	4	180
55	2	110
Total	$\sum f_i = 30$	$\sum f_i x_i = 860$

$$\text{Mean} = \frac{\sum x_i f_i}{\sum f_i} = \frac{860}{30} = 28.97$$

Data Distribution: Median

The median M is the midpoint of a distribution, the number such that half the observations are smaller, and the other half are larger.

To Find the Median

1. Sort all the observation in order of size from smallest to largest
2. It the number of observations n is odd, the median M is the centre observation in the ordered ist e.g. M = n+1/2 the obs.
3. It the number of observations n is even, the mean of two centre observation in the ordered list

Median (odd numbers)

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th}\text{observation}$$

56, 67, 54, 34, 78, 43, 23. What is the median?

Here, n (no. of observations) = 7

Arranging in ascending order, we get: 23, 34, 43, 54, 56, 67, 78.

$$Median\ position = \frac{7+1}{2} = 4$$

Median = 54

Data Distribution: Median

Median (even)

$$\text{Median} = \frac{\frac{n}{2}^{th}\ \text{obs.} + (\frac{n}{2}+1)^{th}\text{obs.}}{2}$$

Let's consider the data: 50, 67, 24, 34, 78, 43.

What is the median?

Arranging in ascending order, we get: 24, 34, 43, 50, 67, 78.

$\frac{6}{2}$= 4

$$\text{Median} = \frac{43+50}{2} = 46.5$$

Data Distribution: Median (Grouped data)

$$\text{Median} = l + [\frac{\frac{n}{2} - c}{f}] \times h$$

where,

l=lower limit of median class

c= cumulative frequency of the class preceding the median class

f=frequency of the median class

H =class size

n is total number of observation i.e f

Find the median marks for the following distribution:

Classes	0-10	10-20	20-30	30-40	40-50
Frequency	2	12	22	8	6

Classes	Number of students	Cumulative frequency
0-10	2	2
10-20	12	2 + 12 = 14
20-30	22	14 + 22 = 36
30-40	8	36 + 8 = 44
40-50	6	44 + 6 = 50

$N = 50$

$$\frac{N}{2} = \frac{50}{2} = 25$$

Median Class $= 20 - 30$

$l = 20, f = 22, c.f = 14, h = 10$

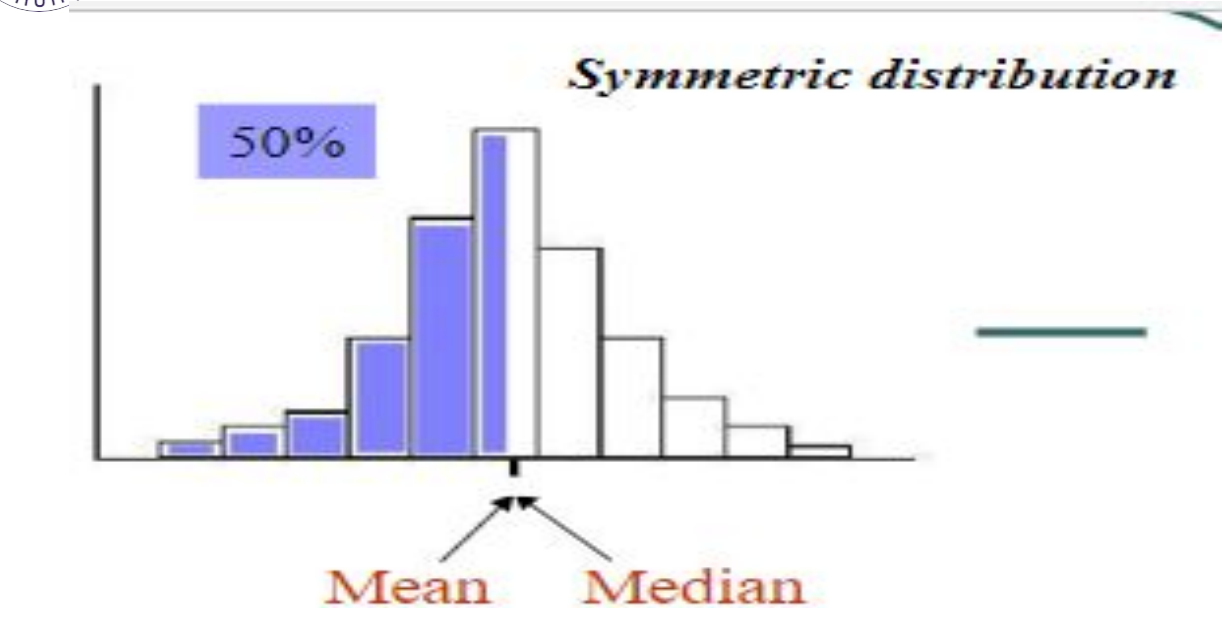
$$\text{Median} = l + [\frac{\frac{n}{2} - c}{f}] \times h$$

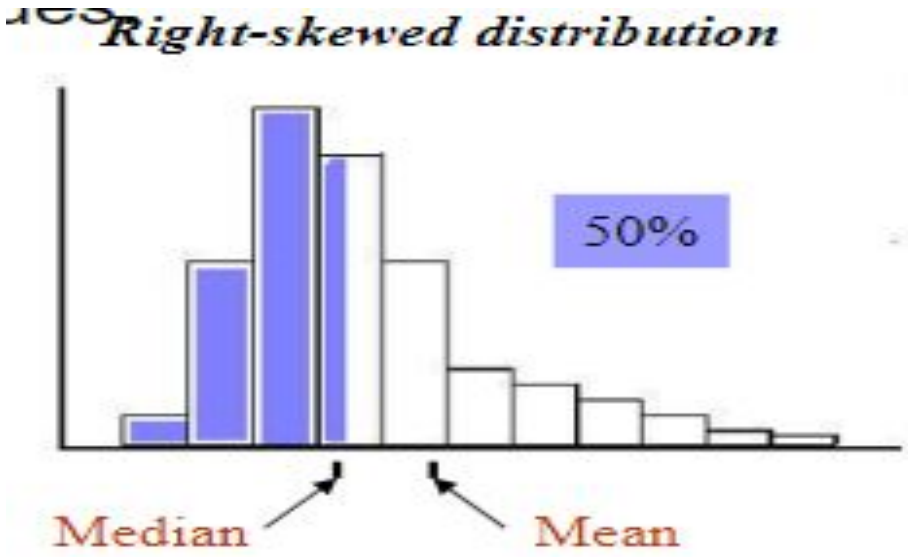
$$= 20 + \frac{25 - 14}{22} \times 10$$

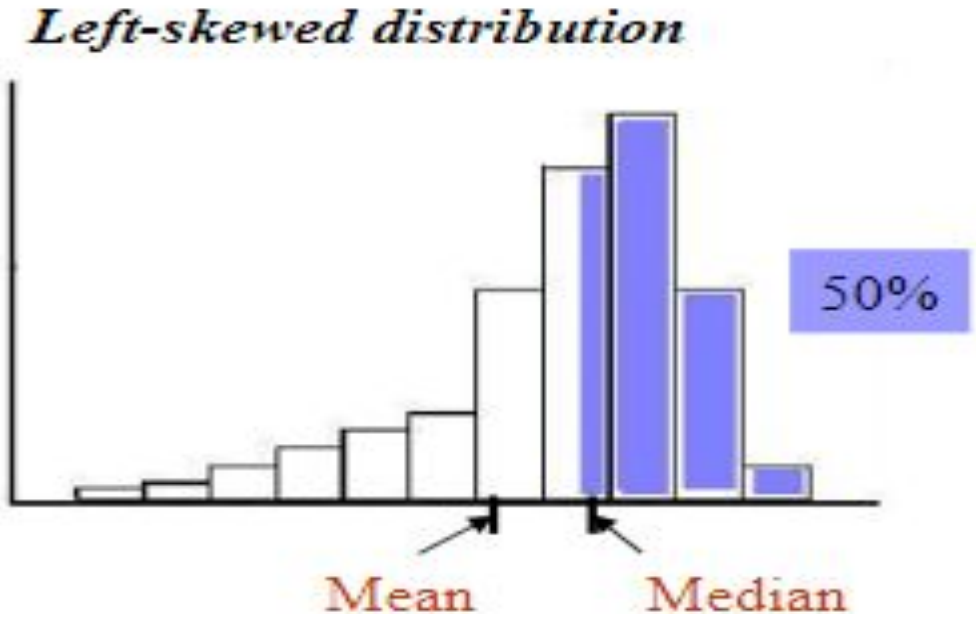
$$= 20 + \frac{11}{22} \times 10$$

$$= 20 + 5 = 25$$

Mean versus Median







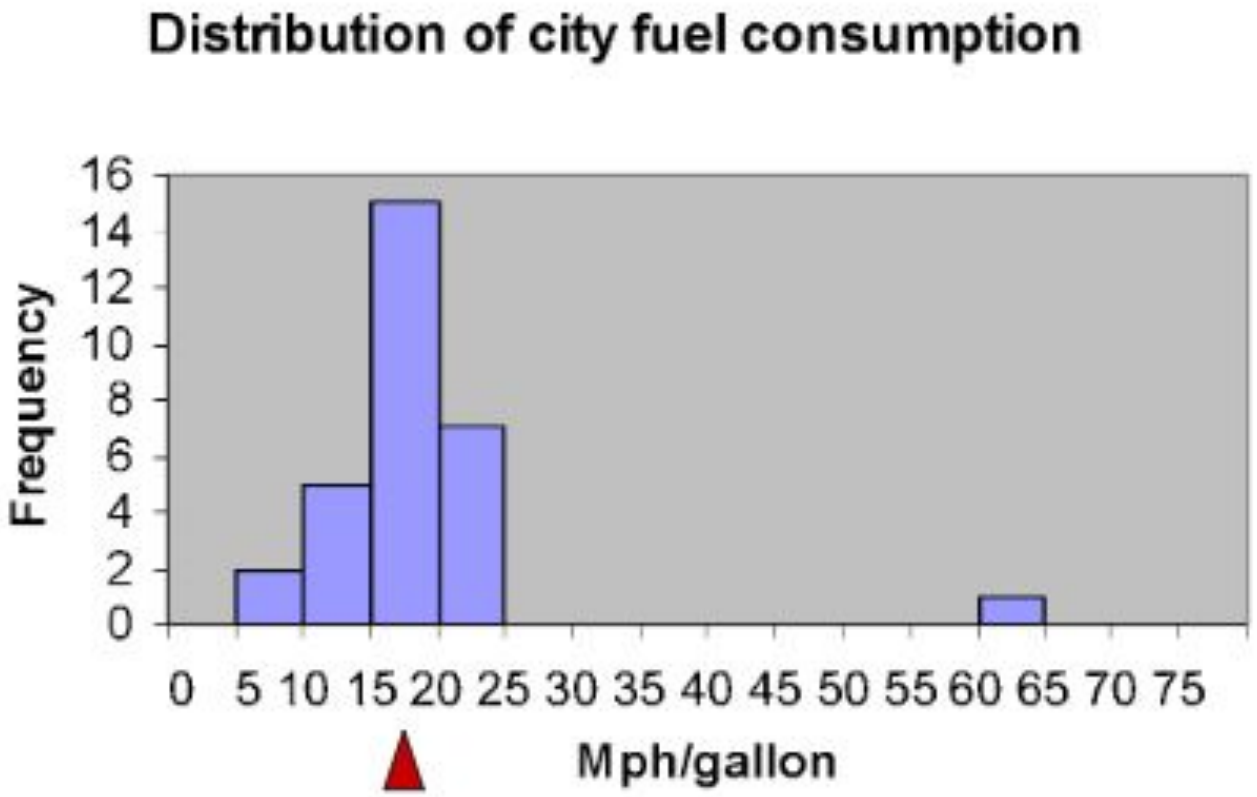
1. The Mean and median of a symmetric distribution are close together
2. In Skewed distributions, the mean is father out in the long tail than is the median. The mean is more sensitive to extreme values.

Mean or Median

- ✔ The mean is a good measure for the centre of a symmetric distribution
- ✔ The median is a resistant measure and should be used for skewed distributions.
- ✔ Its value is only slightly affected by the presence of extreme observations, no matter how large these observations are.

The Mode

- The Mode is the observation value with the highest frequency (observation with maximum frequency



An average, the cars under study drive 18.9 miles per gallon, and 50% of cars under study drive at least 18 miles per gallon

The Mode

6, 8, 9, 3, 4, 6, 7, 6, 3 the value 6 appears the most number of times.
Thus, mode = 6.

Bimodal List

List A = {1, 2, 3, 3, 4, 4, 5, 6}

Mode [A] = {3, 4}

List A has 2 modes.
Therefore, it is a bimodal list.

Trimodal List

List B = {1, 2, 3, 3, 4, 4, 5, 5, 6}

Mode [B] = {3, 4, 5}

List B has 3 modes.
Therefore, it is a trimodal list.

The Mode for grouped data

$$\text{Mode} = l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] \times h$$

where, l= lower limit of modal class,

f_m= frequency of modal class,

f_1= frequency of class preceding modal class,

f_2= frequency of class succeeding modal class,

h= class width

Marks Obtained	0-20	20-40	40-60	60-80	80-100
Number of students	5	10	12	6	3

The highest frequency = 12, so the modal class is 40-60.

l = lower limit of modal class = 40

f_m = frequency of modal class =12

f_1 =frequency of class preceding modal class = 10

f_2 =frequency of class succeeding modal class = 6

h =class width = 20

Find the mode of the given data:

$$\begin{aligned}\text{Mode} &= l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] \times h \\ &= 40 + \left[\frac{12 - 10}{2 \times 12 - 10 - 6}\right] \times 20 \\ &= 40 + \left[\frac{2}{8}\right] \times 20 \\ &= 45\end{aligned}$$

The relation between Mean, Median and Mode

2 mean + Mode = 3 Median

We have a data whose mode == 65 and median == 61.6. Find Mean

$$
\begin{aligned}
2\text{Mean} + \text{Mode} &= 3\,\text{Median} \\
\therefore 2\,\text{Mean} &= 3 \times 61.6 - 65 \\
\therefore 2\,\text{Mean} &= 119.8 \\
\Rightarrow \text{Mean} &= \frac{119.8}{2} \\
\Rightarrow \text{Mean} &= 59.9
\end{aligned}
$$

Spread of a Distribution

Quartiles are statistical measures that divide a dataset into four equal parts, each representing 25% of the data. .

- ✔ First Quartile (Q1): Also known as the lower quartile, it represents the 25^{th} percentile of the data.
- ✔ Second Quartile (Q2): This is the median of the dataset and represents the 50^{th} percentile. It divides the dataset into two equal halves.
- ✔ Third Quartile (Q3): Also known as the upper quartile, it represents the 75^{th} percentile of the data. This means that 75% of the data values fall below Q3.
- ✔ **Interquartile Range (IQR)**: This is the range between the first and third quartiles (Q3 - Q1). The IQR measures the spread of the middle 50% of the data and is used to identify outliers.

Spread of a Distribution

Calculation of Quartiles

- ✓ $Q_1=(\frac{1}{4}(n+1))^{th}$ position
- ✓ $Q_2=(\frac{1}{2}(n+1))^{th}$ position
- ✓ $Q_3=(\frac{3}{4}(n+1))^{th}$ position

n is the number of data points

Consider the dataset: 3, 7, 8, 5, 12, 7, 9, 10, 14, 21

Sort the Data: 3, 5, 7, 7, 8, 9, 10, 12, 14, 21

Find Q1: The position is $\frac{1}{4}(10+1)=2.75$ so Q_1 is between the 2nd and 3rd values, which is 5 and 7.

Interpolating, $Q_1 = 5 + 0.75 \times (7 - 5) = 6.5$.

Find Q2: The position is $\frac{1}{2}(10+1)= 5.5$, so Q_2 is between the 5^{th} and 6^{th} values, which is 8 and 9.

Interpolating, $Q_2 = 8 + 0.5 \times (9 - 8) = 8.5$.

Find Q3: The position is $\frac{3}{4}(10+1)=8.25$ so Q_3 is between the 8^{th} and 9^{th} values, which is 12 and 14.

Interpolating, $Q_3 = 12 + 0.25 \times (14 - 12) = 12.5$.

IQR: $IQR=Q3-Q1=12.5-6.0$ (**Interquartile Range)**

Spread of a Distribution

Find Interquartile Range of 5 ,6, 1, 2, 15, 12, 27, 19, 18 (Odd numbers)

- ✔ Step 1: Put the numbers in order. 1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- ✔ Step 2: Find the median. 1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- ✔ Step 3: Place parentheses around the numbers above and below the median. Not necessary statistically, but it makes Q1 and Q3 easier to spot. (1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
- ✔ Step 4: Find Q1 and Q3 (Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.)
  (1, 2, **5**, 6, 7), 9, (12, 15, **18**, 19, 27). Q1 = 5 and Q3 = 18.
- ✔ Step 5: Subtract Q1 from Q3 to find the interquartile range. 18 – 5 = 13.

Spread of a Distribution

Find Interquartile Range of 4,4,10,11,15,7,14,12,6(Odd numbers)

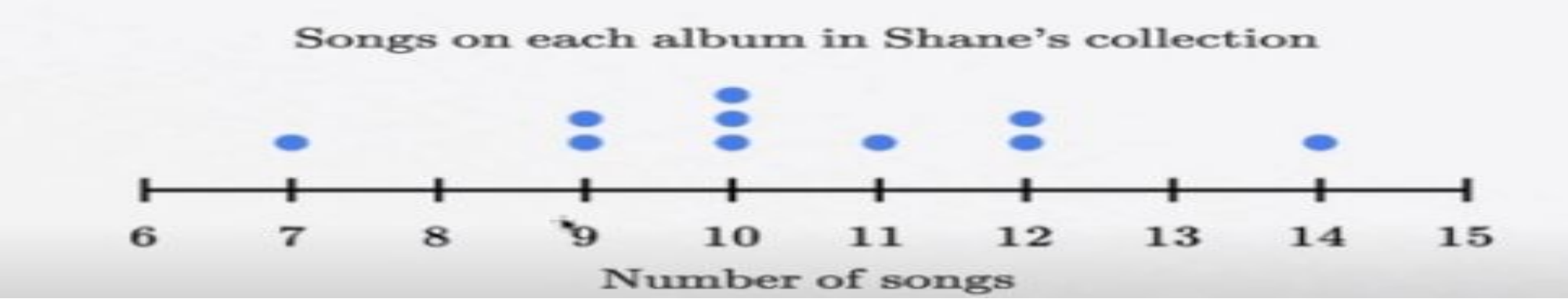
- ✔ Step 1: Put the numbers in order. 4,4,6,7,10,11,12,14,15
- ✔ Step 2: Find the median. 4,4,6,7,10,11,12,14,15
- ✔ Step 3: Place parentheses around the numbers above and below the median. Not necessary statistically, but it makes Q1 and Q3 easier to spot. (4,4,6,7), 10, (11,12,14,15).
- ✔ Step 4: Find Q1 and Q3
- ✔ (4,4,6,7), 10, (11,12,14,15). Q1 = 5 and Q3 = 13.
- ✔ Step 5: Subtract Q1 from Q3 to find the interquartile range. 13 – 5 = 8.

Spread of a Distribution

Find the IQR for the following data set: 3, 5, 7, 8, 9, 11, 15, 16, 20, 21 (Even)

Step 1: Put the numbers in order: 3, 5, 7, 8, 9, 11, 15, 16, 20, 21.
Step 2: Make a mark in the centre of the data: 3, 5, 7, 8, 9, | 11, 15, 16, 20, 21.
Step 3: Place parentheses around the numbers above and below the mark you made in Step 2–it makes Q1 and Q3 easier to spot.
(3, 5, 7, 8, 9), | (11, 15, 16, 20, 21).
Step 4: Find Q1 and Q3
Q1 is the median (the middle) of the lower half of the data, and Q3 is the median (the middle) of the upper half of the data.
(3, 5, 7, 8, 9), | (11, 15, 16, 20, 21). Q1 = 7 and Q3 = 16.
Step 5: Subtract Q1 from Q3. 16 – 7 = 9.

Spread of a Distribution



- ✔ Step 1: Put the numbers in order. 7,9,9,10,10,10,11,12,12,14
- ✔ Step 2: Make a mark in the centre of the data: 7,9,9,10,10|10,11,12,12,14

Step 3: Place parentheses around the numbers above and below the mark you made in Step 2–it makes Q1 and Q3 easier to spot.

(7,9,9,10,10), | (10,11,12,12,14).

Step 4: Find Q1 and Q3

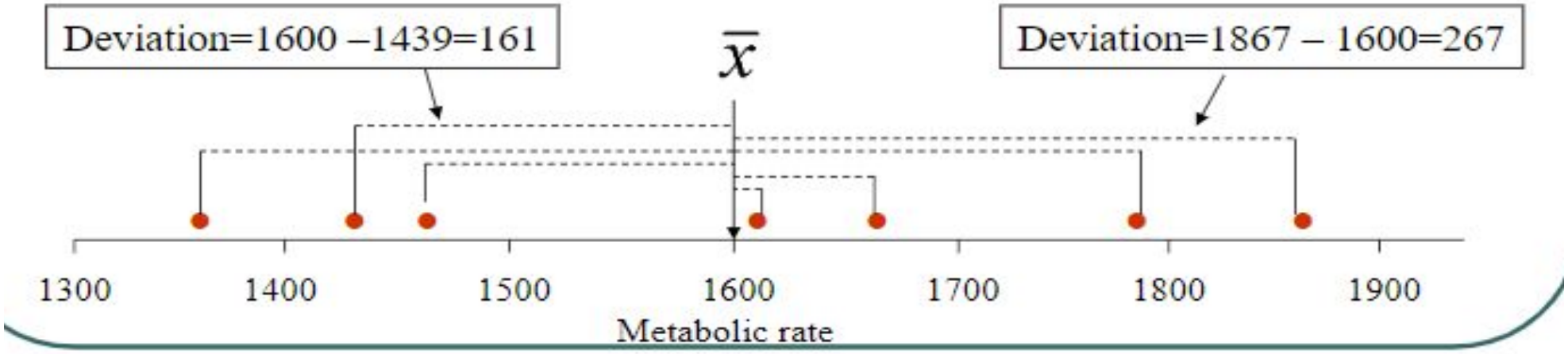
(7,9,9,10,10), | (10,11,12,12,14). Q1 = 9 and Q3 = 12.

Step 5: Subtract Q1 from Q3. 12 – 9 = 3.

Standard Deviation

- ✓ If a distribution is symmetric
  - ✓ Use the average to measure the centre and the standard deviation to measure the spread
  - ✓ The SD measures how far the observations are from the average
  - ✓ E.g. A person's Metabolic rate = rate at which the body consumes energy Rates of 7 men in a study on dieting 1792, 1666, 1614, 1460, 1867, 1439, 1362.

$\bar{x} = 1600\ and$ sd = 189.24



Standard Deviation

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

'the variance of an observed Variables is defined as the square of the standard deviation

Variance = s^2

Properties of Standard Deviation

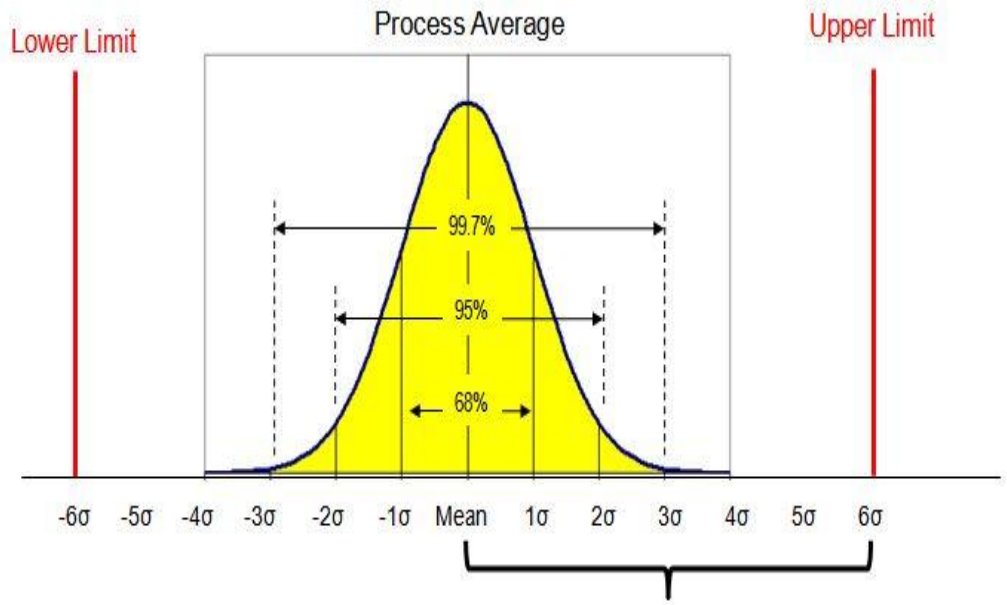
- ✔ It measures the spread about the mean
- ✔ Only used in association with the mean, Good descriptive measure for Symmetric distributions
- ✔ If s = 0 all observations have the same values
- ✔ It is NOT a resistant measure, a few extreme observations may affect its values

Interpreting the SD

For many lists of observation- especially if their histogram is bell-shaped

1. Roughly 6* % of the observation in the list lie within 1 SD of the average
2. 95% of the observation lie within 2 SD of the average

Sigma	Defects	Confidence
1S	31.8%	68%
2S	6.7%	95%
3S	2.7%	99.7%
4S	0.6%	99.99%
5S	0.02%	
6S	0.00034%	



Problems

Consider the following three data sets A, B and C and determine standard Deviation

A = {9,10,11,7,13}

B = {10,10,10,10,10}

C = {1,1,10,19,19}

Step 1: Mean

Mean of Data set A = (9+10+11+7+13)/5 = 10

Mean of Data set B = (10+10+10+10+10)/5 = 10

Mean of Data set C = (1+1+10+19+19)/5 = 10

Standard Deviation Data set A = $\sqrt{[\ ((9-10)^2+(10-10)^2+(11-10)^2+(7-10)^2+(13-10)^2\ )/5\ ]} = 2$

Standard Deviation Data set B= $\sqrt{[\ ((10-10)^2+(10-10)^2+(10-10)^2+(10-10)^2+(10-10)^2\ )/5\ ]} = 0$

Standard Deviation Data set C = $\sqrt{[\ ((1-10)^2+(1-10)^2+(10-10)^2+(19-10)^2+(19-10)^2\ )/5\ ]} = 8.05$

Variance = S^2

Variance of A = 2^2 =4

Variance of B = 0^2 = 0

Variance of C = 8^2 =64

Coefficient of Variance = CV = standard deviation / Mean

CV of A = 2/10 = 0.2

CV of B = 0/10 = 0

CV of C = 8.05/10 = 0.805

Problems

Consider the following three data sets A, B and C and determine standard Deviation

A = {9,10,11,7,13}

B = {10,10,10,10,10}

C = {1,1,10,19,19}

Step 1: Mean

Mean of Data set A = (9+10+11+7+13)/5 = 10

Mean of Data set B = (10+10+10+10+10)/5 = 10

Mean of Data set C = (1+1+10+19+19)/5 = 10

Standard Deviation Data set A = $\sqrt{[\ ((9-10)^2+(10-10)^2+(11-10)^2+(7-10)^2+(13-10)^2\ )/5\ ]} = 2$

Standard Deviation Data set B= $\sqrt{[\ ((10-10)^2+(10-10)^2+(10-10)^2+(10-10)^2+(10-10)^2\ )/5\ ]} = 0$

Standard Deviation Data set C = $\sqrt{[\ ((1-10)^2+(1-10)^2+(10-10)^2+(19-10)^2+(19-10)^2\ )/5\ ]} = 8.05$

Variance = S^2

Variance of A = 2^2 =4

Variance of B = 0^2 = 0

Variance of C = 8^2 =64

Coefficient of Variance = CV = standard deviation / Mean

CV of A = 2/10 = 0.2

CV of B = 0/10 = 0

CV of C = 8.05/10 = 0.805

Interpreting the SD

Calculate the standard deviation from the following distribution of marks by

Marks	No. of Students
1–3	40
3–5	30
5–7	20
7–9	10

$$\overline{X} = \frac{\sum fX}{\sum f} = \frac{400}{100} = 4$$

Marks	f	X	fX	$(X-\overline{X})^2$	$f(X-\overline{X})^2$
1–3	40	2	80	4	160
3–5	30	4	120	0	0
5–7	20	6	120	4	80
7–9	10	8	80	16	160
Total	100		400		400

$\sum fX$ 400

$$S = \sqrt{\frac{\sum f(X-\overline{X})^2}{\sum f}} = \sqrt{\frac{400}{100}} = \sqrt{4} = 2 \text{ marks}$$

Variance = $S^2 = 4$

CV= S/x = 2/4 = 0.5

Statistical Estimation

- ✔ **Parameter**: A numerical characteristic of a population, such as the mean (μ) or standard deviation (σ)
- ✔ **Estimator**: A statistic (a function of the sample data) used to estimate a population parameter.
- ✔ **Estimate**: The value obtained from an estimator. For instance, if the sample mean is 50, then 50 is the estimate of the population mean.
- ✔ **Point Estimate**: A single value given as an estimate of a population parameter. For example, using the sample mean to estimate the population mean.
- ✔ **Interval Estimate**: A range of values within which the population parameter is expected to lie, with a certain level of confidence. For example, a 95% confidence interval for the mean.

Factors Affecting Confidence Interval Estimates

- ✔ The factors that determine the width of a confidence interval are
  - ✔ The sample size, n
  - ✔ The variability in the population, usually SD estimated
  - ✔ The desired level of confidence

Confidence Intervals

- ✓ **Confidence intervals (CIs)** offer a range of values where a parameter likely falls, quantifying uncertainty**.**
- ✓ For example, a 95% confidence interval means that if you were to take 100 different samples and calculate a CI for each, approximately 95 of those intervals would contain the true population parameter
- ✓ Point Estimate: The point estimate is a single value from the sample used to estimate the population parameter. (Point estimate = mean)

  Point Estimation = $\bar{\bar{x}} = \frac{\sum x}{n}$
- ✓ Margin of Error: The margin of error is the range around the point estimate that accounts for variability and provides the interval

  Margin error = $z \frac{\sigma}{\sqrt{n}}$ (z-score or t-score)

Confidence Intervals

✓ **Confidence intervals (CIs)**

Confidence Interval for the Mean (known σ and z distribution)

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

Confidence level	Critical (z) value to be used in confidence interval calculation
50%	0.67449
75%	1.15035
90%	1.64485
95%	1.95996
97%	2.17009
99%	2.57583
99.9%	3.29053

Problems

1. A survey was taken of USA companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India? A radom sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years. Using this information, construct a 90% confidence internal for the mean number of years that a company has been trading in India for the population of US companies trading with firms in India.

G$\bar{x} = 10.455 \quad \sigma = 7.7 \;\; and \; n = 44 \quad z = 1.645 \; (from \; table \; for \; 90\,\% \; confidence \;)$

$$\bar{x} - Z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z\frac{\sigma}{\sqrt{n}}$$

$$10.455 - 1.645\frac{7.7}{\sqrt{44}} \leq \mu \leq 10.45 + 1.645\frac{7.7}{\sqrt{44}}$$

$$8.545 \leq \mu \leq 12.365$$

The 90% confidence interval for the mean number of years that a company has been trading with firms in India is approximately [8.545,12.365] years

Problems

2. A study is conducted in a company that employs 800 engineers. A random sample of 50 engineers reveals that the average sample age is 34.3 years. Historically, the population standard deviation of the age of the company's engineers is approximately 8 years. Construct a 98% confidence interval to estimate the average age of all the engineers in this company.

✓ Given data: N = 800 n = 50 $\bar{x}$ = 34.3 σ = 8 z = 2.33 from the table for 98% confidence)

$$\bar{x} - Z\,\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \le \mu \le \bar{x} + Z\,\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$34.3 - 2.33\,\frac{8}{\sqrt{50}}\sqrt{\frac{800-50}{800-1}} \le \mu \le 34.3 + 2.33\,\frac{8}{\sqrt{50}}\sqrt{\frac{800-50}{800-1}}$$

$$31{,}75 \le \mu \le 36.85$$

The 98% confidence interval to estimate the average age of all engineers in the company is approximately [31.67,36.93] years

Estimating the Mean of a Normal Population: Sample Size is Small and Population is unknown

- ✓ When estimating the mean of a normal population with a small sample size and an unknown population standard deviation, you use the **t-distribution** instead of the normal distribution.

  sample size (n) < 30

- ✓ Steps to Construct the Confidence Interval
  - ✓ Determine the Sample Mean ($\bar{x}$) =
  - ✓ Determine the Sample Standard Deviation (σ)
  - ✓ Find the t-Score based on degree of freedom and confidence level or $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
  - ✓ Construct the Confidence Interval (CI) $\bar{x} - t\alpha_{/2,n-1}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t\alpha_{/2,n-1}\frac{\sigma}{\sqrt{n}}$

T-Distribution Characteristics

- ✔ T distribution-symmetric, unimodal, mean =0, flatter in middle and have more area is their tails than the normal distribution
  - ✔ T distribution approach the normal curve as n becomes larger
  - ✔ T distribution is to be used when the population variance or population Std Dev is unknown, regardless of the size of the sample

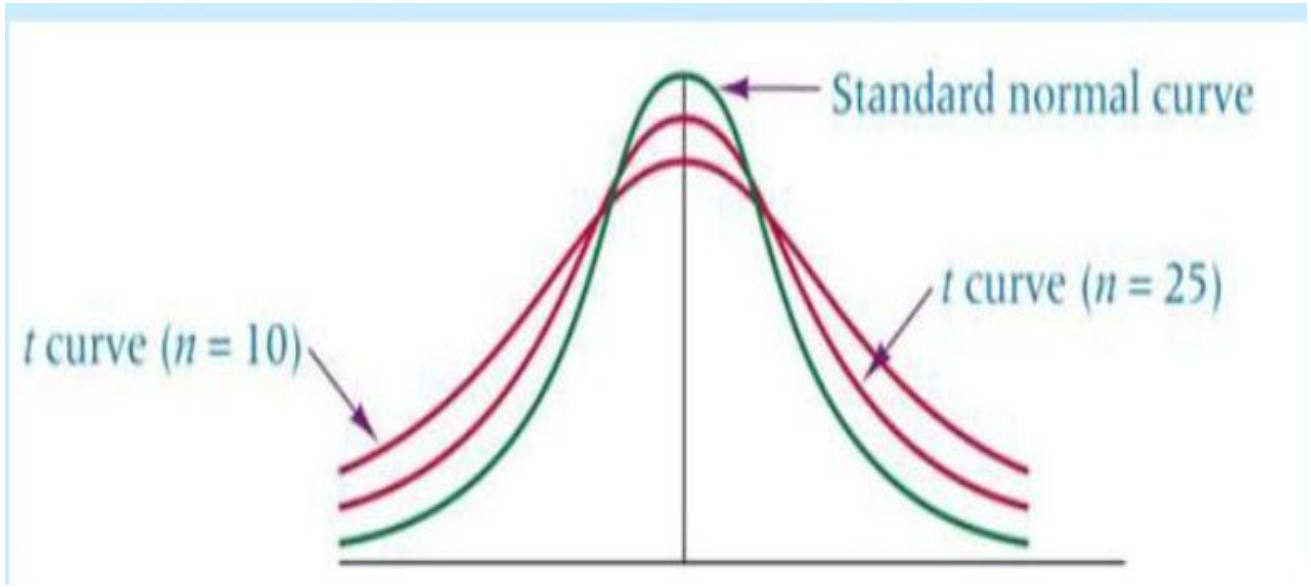
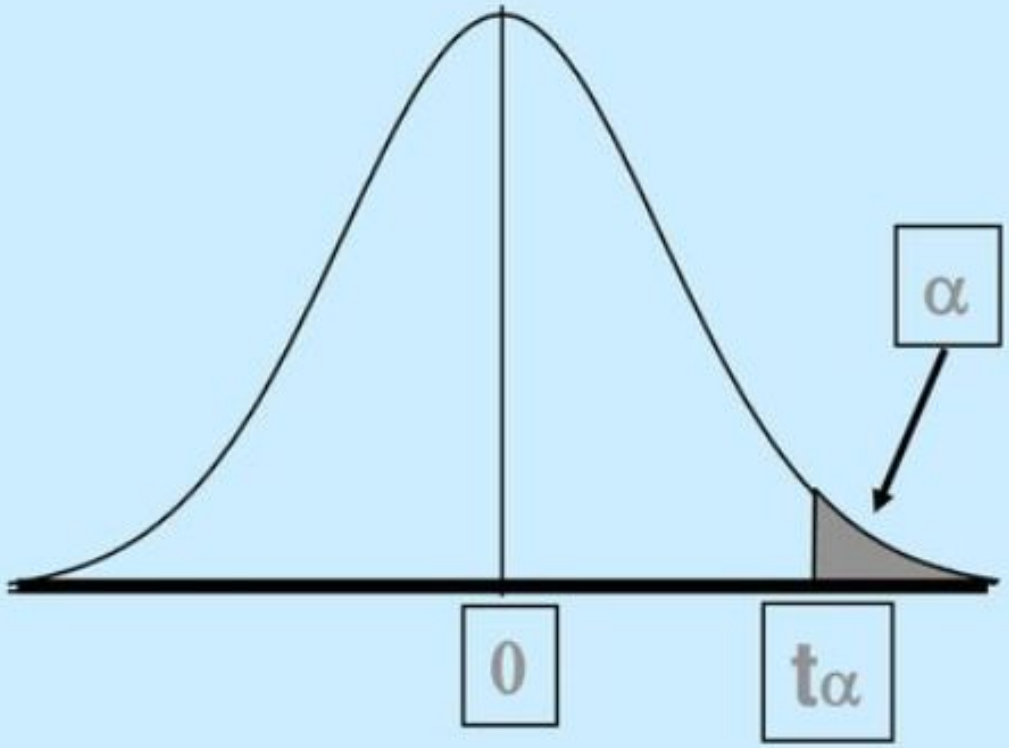


Table of Critical Values of t

df	$t_{0.100}$	$t_{0.050}$	$t_{0.025}$	$t_{0.010}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.327	2.576



With df = 24 and $\alpha = 0.05$, $t_\alpha = 1.711$.

Problem

3. The owner of a large equipment rental company wants to make rather quick estimate of the average number of days a piece of ditch digging equipment is rented out per person per time. The company has records of all rentals, but the amount of time required to conduct an audit of all accounts would be prohibitive. The owner decides to take a random sample of rental invoices. Fourteen different rentals of ditch diggers are selected randomble from the files, yielding the following data. She uses these data construct a 99% confidence interval to estimate the average number of days that a ditch digger is rented and assumes that the number of datas per rental is normally distributed in the population. The data: 3, 1, 3, 2, 5, 1, 2, 1, 4, 2, 1, 3, 1,1

Given information:

n = 14

Sample data: 3, 1, 3, 2, 5, 1, 2, 1, 4, 2, 1, 3, 1,1

Confidence level = 99%

Problem

Calculate the Sample Mean

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3+1+3+2+5+1+2+1+4+2+1+3+1+1}{14} = 1.79$$

Calculate the Sample Standard Deviation (s)

$$\sigma^2 = \frac{\sum(x_i-\bar{x})^2}{n-1} = \frac{(3-1.79)^2+(1-1.79)^2+(3-1.79)^2+(2-1.79)^2+(5-1.79)^2+\cdots+(1-1.79)^2}{14-1} = 0.52$$

$\sigma = 0.72$

Df= n-1 = 13

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073

$\alpha/_2 = \frac{1-0.99}{2} = 0.005$

t = 3.012 (from the table based on α and df)

Problem

$$\bar{x} - t\alpha_{/2,n-1}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t\alpha_{/2,n-1}\frac{\sigma}{\sqrt{n}}$$

$$1.79 - 3.012\ \frac{0.72}{\sqrt{14}} \leq \mu \leq 1.79 + \ 3.012\ \frac{0.72}{\sqrt{14}}$$

$$1.22 \leq \mu \leq 2.36$$

The 99% confidence interval for the average number of days that a ditch digger is rented is approximately [1.22,2.36] days. This interval provides a range within which the true population mean is expected to lie with 99% confidence, assuming normal distribution of rental days.

Confidence Interval to Estimate the Population Proportion

To construct a confidence interval to estimate the population proportion, you use the sample proportion and the normal approximation if the sample size is large enough. n > 30

- ✓ Calculate the Sample Proportion ($\hat{P} = \frac{x}{n}$) where x is the number of successes and n is the sample size
- ✓ Determine the z-Score for the Desired Confidence Level based on confidential level
- ✓ Construct the Confidence Interval (CI)

$$\hat{P} - z\alpha_{/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \mu \leq \hat{P} + z\alpha_{/2}\sqrt{\frac{\hat{p}(\widehat{1-p})}{n}}$$

Problem

Assume you have a sample of 200 people where 80 of them favor a new policy. Construct a 95% confidence interval for the proportion of people who favor the policy.

1. Calculate the Sample Proportion ($\hat{P} = \frac{x}{n}$) = $\frac{80}{200}$ =0.40

2. Determine the z-Score for a 95% Confidence Level z = 1.960

3. **Calculate the Interval** $\hat{P} - z\alpha/_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \mu \leq \hat{P} + z\alpha/_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$$0.40 - 1.96\sqrt{\frac{0.40(1-0.40)}{200}} \leq \mu \leq 0.40 + 1.96\sqrt{\frac{\widehat{.40}(1-0.40)}{200}}$$

$$0.332 \leq \mu \leq 0.468$$

The 95% confidence interval for the proportion of people who favor the new policy is approximately [0.332,0.468] This interval provides a range within which the true population proportion is expected to lie with 95% confidence.

Problem

A clothing company produces men's Jeans. The jeans are made and sold with either a regular cut or a boot cut. In an effort to estimate the proportion of their men's jeans market in Oklahoma city that prefers boot-cut jeans, the analyst takes a random sample of 423 jeans sales from the company two Oklahoma city retail outlets. Only 72 of the sales were for boot-cut jeans. Construct a 90% confidence interval to estimate the proportion of the population in Oklahoma city who prefer boot-cut jeans.

Given n = 423 x = 72 $\hat{p} = \frac{x}{n} = \frac{72}{423} = 0.17$ z = 1.645 for 90% confidence from the table

$$\hat{P} - z\alpha_{/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \mu \leq \hat{P} + z\alpha_{/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.17 - 1.642\sqrt{\frac{0.17(1-0.17)}{423}} \leq \mu \leq 0.17 - 1.642\sqrt{\frac{0.17(1-0.17)}{423}} \qquad \mathbf{0.14} \leq \mu \leq \mathbf{0.20}$$

The 90% confidence interval to estimate the proportion of the population in Oklahoma City who prefer boot-cut jeans is approximately [0.14,0.2]. This interval provides a range within which the true proportion of the population is expected to lie with 90% confidence.

Hypothesis Testing

- Hypothesis is considered as intelligent guess or prediction, that gives directional to the researcher to answer the research question.
- Hypothesis are defined as formal statement of the tentative or expected prediction or explanation of the relationship between two or more variables in a specified population.
- A hypothesis is formal tentative statement of the expected relationship between two or more variable under study.
- A hypothesis helps to translate the research problem and objective into a clear explaining or prediction of the expected results or outcome of the study.

Hypothesis Testing

Hypothesis testing refers to

1. Making an assumption about a population parameters
2. Collecting sample data
3. Calculating a sample statistic
4. Using the sample statistic to evaluate the hypothesis.

Null Hypothesis (H_0): State the hypothesized value of the parameter before sampling. The assumption we wish to test or the assumption we are trying to reject.

Population mean $\mu = 20$ There is no difference between coke and diet coke

Alternative Hypothesis (H_A): All possible alternatives other than the null hypothesis

e.g $\mu > 20$ and $\mu < 20$

There is a difference between coke and diet coke.

Selecting and Interpreting Significance Level

- Selecting and interpreting the significance level (α) is crucial in hypothesis testing.
- The significance level, denoted as α, is the probability of rejecting the null hypothesis when it is actually true. It represents the threshold for determining whether the observed results are statistically significant.
- Common Choices for α
  - 0.10: Indicates a 10% risk of a Type I error
  - 0.05: Indicates a 5% risk of a Type I error (most commonly used).
  - 0.01: Indicates a 1% risk of a Type I error.
- Interpreting the Significance Level (Comparison with α)
  - If $p \leq \alpha$, reject the null hypothesis
  - If $p > \alpha$, do not reject the null hypothesis
- Implications of α Choices
  - **Higher α**: More likely to reject the null hypothesis, increasing the risk of Type I errors.
  - **Lower α**: Less likely to reject the null hypothesis, reducing the risk of Type I errors but increasing the risk of Type II errors.

Parametric Tests

1. Parametric Test
2. Non Parametric tests

Parametric test Parametric tests that make specific assumptions about the parameters (the mean and variance) of the population distribution from which the data is drawn.

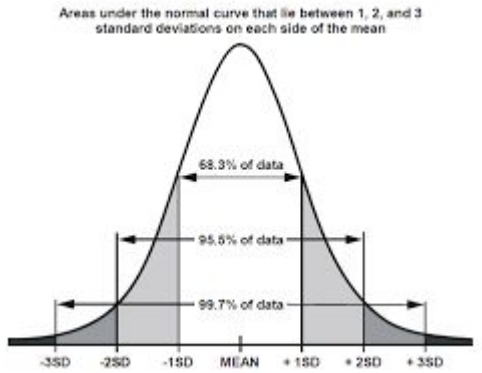
Applications

- ✔ Used for continuous variables (y = f(x))
- ✔ Used when data are measured on appropriate interval of ratio scale of measurement
- ✔ Data should follow normal distribution

Parametric Tests

Parametric tests that make specific assumptions about the parameters (the mean and variance) of the population distribution from which the data is drawn.

Normality: Many parametric tests assume that the data follows a normal distribution (bell-shaped curve)



Homogeneity of Variance: Some parametric tests assume that different groups have similar variances (σ^2)

Parametric Tests

- **t-Test**: Used to compare the means of two groups to determine if they are statistically significantly different from each other.
  - Independent t-test (comparing two separate groups)
  - Paired t-test (for comparing two related groups)
  - One-Sample t-Test (To compare the mean of a single group to a known value or population mean)
- **ANOVA (Analysis of Variance)**: Used to compare the means of three or more groups to see if there is a significant difference among them.
  - One-way ANOVA (for one factor)
  - Two-way ANOVA (for two factors)
- **Pearson Correlation Coefficient**: Measures the strength and direction of the linear relationship between two continuous variables.

Parametric Tests

- ✔ **Regression Analysis**: Used to understand the relationship between a dependent variable and one or more independent variables.
  - ✔ Linear regression
  - ✔ Multiple and logistic regressions

- ✔ **Z-Test**: Used to compare the sample mean to the population mean when the sample size is large and the population variance is known.

Parametric Tests-t test

- ✔ Developed by Prof. W S Gosseti
- ✔ A t-test compares the difference between two mean of different groups to determine whether the difference is statically significant.

One Sample t-test

Assumptions:

- ✔ Population is normally distributed
- ✔ Sample is drawn from the population, and it should be random
- ✔ Known the population mean

Condition

- ✔ Population standard deviation is not known
- ✔ Size of the sample is small (<30)

Parametric Tests-t test-one sample

Steps to Perform a One-Sample t-Test:

1. State the Hypotheses
   - **Null Hypothesis (H_0)**: $\mu=\mu 0$ (the sample mean is equal to the hypothesized population mean)
   - **Alternative Hypothesis (H_1)**: $\mu\neq\mu 0$ (the sample mean is not equal to the hypothesized population mean)
2. **Calculate the t-Statistic**: $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$
3. **Determine the Degrees of Freedom** (df = n-1)
4. **Find the p-Value**: Compare the t-statistic to the t-distribution with the appropriate degrees of freedom to find the p-value
5. **Make a Decision**
   - **Reject H_0** if the p-value is less than the significance level (α), commonly set at 0.05.
   - **Fail to Reject H_0** if the p-value is greater than α.
6. **Interpret the Results**

If you reject the null hypothesis, it means there is a significant difference between the sample mean and the hypothesized population mean. If you fail to reject it, there is not enough evidence to suggest a significant difference.

Parametric Tests-t test

1. The following Data represents haemoglobin values in gm/dl for 10 patient 10.5 , 9, 6.5, 8, 11, 7, 7.5, 8.5, 9.5, 12 Is the mean value for patents significantly differ from the mean values of general population (12 gm/dl). Evaluate the role of change

Mean

$$\bar{x} = \frac{10.5 + 9 + 6.5 + 8 + 11 + 7 + 7.5 + 8.5 + 9.5 + 12}{10}$$

Standard

$$\sigma = \frac{\sqrt{(10.5-8.95)^2+(9-8.95)^2+(6.5-8.95)^2+(8-8.95)^2+(11-8.95)^2+(7-8.95)^2+(7.5-8.95)^2+(8.5-8.95)^2+(9.5-8.95)^2+(12-8.95)^2}}{10-1}$$

$= 1.802$

Step 1: State the Hypotheses

Null Hypothesis (H0): The mean hemoglobin value of the patient sample is equal to the mean hemoglobin value of the general population.

H0:μ=12 gm/dl

Alternative Hypothesis (H1): The mean hemoglobin value of the patient sample is different from the mean hemoglobin value of the general population.

H1:μ≠12 gm/dl

Parametric Tests-t test

Step 2: **Calculate the t-Statistic**

$$t = \frac{\bar{\bar{x}} - \mu}{\frac{s}{\sqrt{n}}} = \frac{8.95-12}{1.80201/\sqrt{10}} = -5.352$$

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

Step 3 Determine the Degrees of Freedom (df = n-1)

df = 10-1 = 9

Step 4: Then compare with tabulated value for 9 df and 5% level of significance it is = -1.8331. The calculated values > tabulated value

5. Make a Decision

Reject H0 and concluded that there is a statistically significant difference between the mean of sample ad population mean and this difference is unlikely due to change

6. Interpret the Results

The significant difference in mean hemoglobin values suggests that the patient group has a lower average hemoglobin level compared to the general population.

Paired t-test

Used when measurements are taken from the same subject before and after some manipulation or treatment.

e.g. To determine the significance of a difference in blood pressure before and after administration of an experimental pressure substance.

Assumptions:

1. Populations are distributed normally
2. Samples are drawn independently and at Radom

Condition

1. Samples are related with each other
2. Sizes of the sample are small and equal

Paired t-test

1. State the Hypotheses

Null Hypothesis (H0): The mean difference between the paired measurements is zero. H0:μd=0

Alternative Hypothesis (H1): The mean difference between the paired measurements is not zero. H1:μd≠=0

2. Calculate the Differences

$$d_i = After_i - Before_i$$

3. Calculate the Mean and Standard Deviation of the Differences

Mean of the differences $\bar{d} = \frac{\sum d_i}{n}$

Standard deviation of the differences $(\sigma_d) = \frac{\sqrt{(d_i - \bar{d})^2}}{n-1}$

4. Calculate the t-Statistic

$$t = \frac{\bar{d}}{\sigma_c / \sqrt{n}}$$

Paired t-test

5. Compare t calculated and theoretical values
6. Conclusion

Blood pressure of 8 patients before and after treatment

BP before	BP after
180	140
200	145
230	150
240	155
170	120
190	130
200	140
165	130

Parametric Tests-t test
(One Sample two occasion)

1. State the Hypotheses

✔ **Null Hypothesis (H0):** The mean difference in blood pressure before and after treatment is zero, implying no effect of the treatment. H0:μd=0

✔ **Alternative Hypothesis (H1):** The mean difference in blood pressure before and after treatment is not zero, implying an effect of the treatment. H1:μd≠0

2. Calculate the Differences

BP before	BP after	Dif (x)
180	140	40
200	145	55
230	150	80
240	155	85
170	120	50
190	130	60
200	140	60
165	130	35
		∑x= 465

Parametric Tests-t test
(One Sample two occasion)

3. Calculate the Mean and Standard Deviation of the Differences

Mean of difference

$$\bar{d} = \frac{\sum d_i}{n} = \frac{40 + 55 + 80 + 85 + 50 + 60 + 60 + 35}{8} = \frac{465}{8} = 58.125$$

Standard deviation of the differences

$$(40 - 58.125)^2 = 331.64$$
$$(55 - 58.125)^2 = 9.78$$
$$(80 - 58.125)^2 = 480.64$$
$$(85 - 58.125)^2 = 727.64$$
$$(50 - 58.125)^2 = 65.14$$
$$(60 - 58.125)^2 = 3.52$$
$$(60 - 58.125)^2 = 3.52$$
$$(35 - 58.125)^2 = 532.14$$

$$331.64 + 9.78 + 480.64 + 727.64 + 65.14 + 3.52 + 3.52 + 532.14 = 2152.44$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} = \sqrt{\frac{2152.44}{7}} \approx \sqrt{307.49} \approx 17.54$$

Parametric Tests-t test
(One Sample two occasion)

4. **Calculate the t-Statistic**

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{58.125}{17.54/\sqrt{8}} = \frac{58.125}{6.20} \approx 9.37$$

Critical values of *t* for two-tailed tests

Significance level (α)

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

5. Compare t calculated and theoretical values

Tabulated (Dof = 7), with level of significance 0.05, two tails = 2.36

6. Conclusion

‘t’ calculated value is higher than table values hence reject the null hypothesis.

This indicates that there is a significant difference in blood pressure before and after treatment, suggesting that the treatment had a substantial effect.

Hypothesis Testing for ρ (Repeated Measures t-test)

Two researcher is conducting a study to evaluate the frequency of sheep collisions during a driving test. Over the course of a day driving test, the number of sheep hit by vehicles is recorded. The data collected from multiple driving tests is as follows:

Subject	Score on test A	Score on test B
1	28	25
2	26	27
3	33	28
4	30	31
5	32	29
6	30	30
7	31	32
8	18	21
9	22	25
10	24	20

Parametric Tests-t test
(One Sample two occasion)

1. State the Hypotheses

Null Hypothesis (H0): There is no significant difference in the mean number of sheep collisions between Test A and Test B. H0:$\mu_A=\mu_B$

Alternative Hypothesis (H1): There is a significant difference in the mean number of sheep collisions between Test A and Test B. H1:$\mu A \neq \mu B$

2. Calculate the Differences

Subject	Score on test A	Score on test B	Difference D
1	28	25	3
2	26	27	-1
3	33	28	5
4	30	31	-1
5	32	29	3
6	30	30	0
7	31	32	-1
8	18	21	-3
9	22	25	-3
10	24	20	4
			$\sum D= 6$

Parametric Tests-t test
(One Sample two occasion)

3. Calculate the Mean and Standard Deviation of the Differences

Mean of difference

$$\bar{d} = \frac{\sum d_i}{n} = \frac{3 + (-1) + 5 + (-1) + 3 + 0 + (-1) + (-3) + (-3) + 4}{10} = \frac{7}{10} = 0.7$$

Standard deviation of the differences

$$(3 - 0.7)^2 = 5.29$$
$$(-1 - 0.7)^2 = 2.89$$
$$(5 - 0.7)^2 = 18.49$$
$$(-1 - 0.7)^2 = 2.89$$
$$(3 - 0.7)^2 = 5.29$$
$$(0 - 0.7)^2 = 0.49$$
$$(-1 - 0.7)^2 = 2.89$$
$$(-3 - 0.7)^2 = 13.69$$
$$(-3 - 0.7)^2 = 13.69$$
$$(4 - 0.7)^2 = 11.29$$

$$5.29 + 2.89 + 18.49 + 2.89 + 5.29 + 0.49 + 2.89 + 13.69 + 13.69 + 11.29 = 75.75$$

$$s_d = \sqrt{\frac{75.75}{9}} \approx \sqrt{8.42} \approx 2.90$$

Parametric Tests-t test

(One Sample two occasion)

4. Calculate the t-Statistic

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{0.7}{2.90/\sqrt{10}} = \frac{0.7}{0.92} \approx 0.76$$

Critical values of *t* for two-tailed tests

Significance level (α)

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

5. Compare t calculated and theoretical values

Tabulated (Dof = 9), with level of significance 0.05, two tails = 2.262

6. Conclusion

‘t’ calculated value is Lower than table values hence accept the null hypothesis.

In this paired t-test, you would conclude that there is no significant difference in the number of sheep collisions between Test A and Test B

Two sample t-test

- ✔ Used when the two independent random samples come from the normal populations having unknown or same variance
- ✔ We test the null hypothesis that the two population means are same i.e $\mu_1 = \mu_2$

Assumptions

1. Population are distributed normally
2. Samples are drawn independently and at random

Conditions

1. Standard deviation in the populations are same and not known
2. Size of the sample is small

Two sample t-test-Procedure

1. State the Hypotheses

Null Hypothesis (H0): There is no significant difference between the means of the two populations. H0:μ1=μ2

Alternative Hypothesis (H1): There is a significant difference between the means of the two populations. H1:μ1≠μ2

2. Collect and Summarize Data

Sample Means: $\bar{x}_1$ for group 1 and $\bar{x}_2$ for group 2

Sample variance s_1^2 for group 1 and s_2^2 for group 2

Sample size n_1 for group 1 and n_2 group 2

Two sample t-test-Procedure

3. Calculate the Test Statistic

Equal Variances Assumed (Calculate the pooled variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Calculate the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

4. Calculate Degrees of Freedom **Equal Variances Assumed** df=n1+n2−2
5. Find t value in table
6. Conclusions

Two sample t-test

The following data represent weight in kg for 10 males and 12 females

Males 80, 75, 95, 55, 60, 70, 75, 72, 80, 65

Females 60, 70, 50, 85, 45, 60, 80. 65, 70, 62, 77, 82 Show that hypothesis

Step 1: Hypotheses:

Null Hypothesis (H0): There is no significant difference between the mean weight of males and females. H0:$\mu_{males}=\mu_{females}$

Alternative Hypothesis (H1): There is a significant difference between the mean weight of males and females.H1:$\mu_{males}\neq\mu_{females}$

Step 3: Calculate Sample Statistics:

Males:

$$\bar{x}_1 = \frac{80+75+95+55+60+70+75+72+80+65}{10} = \frac{630}{10} = 63$$

Deviations

$$\text{Deviations} = [(80-63)^2, (75-63)^2, (95-63)^2, (55-63)^2, (60-63)^2, (70-63)^2,$$

$$\text{Deviations} = [289, 144, 1024, 64, 9, 49, 144, 81, 289, 4]$$

Two sample t-test

$$s_1^2 = \frac{289 + 144 + 1024 + 64 + 9 + 49 + 144 + 81 + 289 + 4}{10 - 1} = \frac{2097}{9} \approx 233$$

Sample Size (n1)=10

For Females

$$\bar{x}_2 = \frac{60 + 70 + 50 + 85 + 45 + 60 + 80 + 65 + 70 + 62 + 77 + 82}{12} = \frac{734}{12} \approx 61.17$$

$$\text{Deviations} = [(60 - 61.17)^2, (70 - 61.17)^2, (50 - 61.17)^2, (85 - 61.17)^2, (45 - 61.1$$

$$s_2^2 - \frac{1.37 + 77.87 + 124.63 + 573.39 + 260.59 + 1.37 + 353.79 + 14.82 + 77.87 + (}{12 - 1}$$

S_2^2 = 192.23

Sample size = 12

Two sample t-test

4. Calculate t values

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(10-1)\cdot 233 + (12-1)\cdot 192.23}{10+12-2}$$

= 210.56

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{63-61.17}{\sqrt{210.56\left(\frac{1}{10}+\frac{1}{12}\right)}}$$

4. Calculate 5.Degrees of freedom df = n_1 +n_2-2 =20

5. Find the p-value

Use a t-distribution table or calculator to find the p-value for t=0.29 df=20.

For a two-tailed test, you compare the p-value to your significance level (α typically 0.05)

6. Make a Decision

If the p-value is large (e.g., greater than 0.05), fail to reject the null hypothesis. There is no significant difference in average weights between males and females.

Perform the above calculations to finalize the results. The t-test result indicates whether there is a significant difference in average weights between the two groups.

Z-test

- ✔ A Z-test is a statistical test used to determine whether there is a significant difference between sample and population means or between the means of two independent samples. It is based on the Z-distribution, which is a special case of the normal distribution with a mean of 0 and a standard deviation of 1. The Z-test is appropriate when the sample size is large (typically n>30) or when the population variance is known

Assumption

- Population is normally distributed
- The sample is drwon at random

Condition

- ✔ Population standard deviation σ is known
- ✔ Size of the sample is large (n>30)

One-Sample Z-Test

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Two-Sample Z-Test

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Z-test-steps

1. State the Hypotheses

Null Hypothesis (H0): There is no effect or difference.

Alternative Hypothesis (H1): There is a significant effect or difference.

2. Calculate the Z-Statistic:
3. Determine the p-value:
4. Make a Decision
5. Report the Results

Hypothesis Testing Z- *test*

1. One of Real Estate advertises that the mean selling time of a residential home is 40 days or less after it is listed in their company. A sample of 50 recently sold homes shows a sample mean selling time of 45 days and a standard deviation of 20 days. Using a 0.02 level of significance, test the validity of the company's claim.

Given: n = 50, mean = 45 days, standard devotion = 20 days, significant level = 0.02

Step 1: Hypotheses

a) **Null Hypothesis (H0)**: The mean selling time is 40 days or less. H0:$\mu \leq 40$

b) **Alternative Hypothesis (H1)**: The mean selling time is greater than 40 days. H1:$\mu > 40$

This is a one –tailed test with $\alpha=0.02$ then critical value is $Z_{0.02} = 2.05$

Step 2: Calculate the Test Statistic

$$Z = \frac{\bar{\bar{x}} - \mu_0}{S_{\bar{x}}} = \frac{45-40}{\frac{20}{\sqrt{50}}} = 1.77$$

Hypothesis Testing Z- *test*

d) Since z= 1.77 is less than 2.05, we do not reject the null hypothesis. A sample selling time 45 days is not significantly greater than population mean of 40 days at 0.02 level of confidence.

The sample does not provide sufficient evidence to suggest that the mean selling time is greater than 40 days.

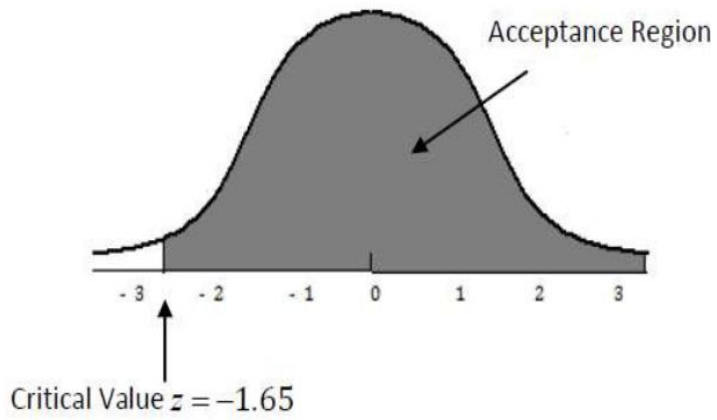
Hypothesis Testing Z- *test*

2. The height of adults in a certain town is found to have a mean of 166.17 cm with standard deviation of 5.89 cm. A random sample of 144 adults in the slum section of the town is discovered to have a mean height of 164.65 cm. Does this height indicate theat the residents of the slum area are significantly shorter in height at 0.05 level of significance?.

a) Set the null and alternative hypothesis

$H_0: \mu \geq 166.17$ $H_a: \mu < 166.17$ cm

b) Population standard deviation σ is 5.89 and sample size is large at n = 144 use the Z-static. This is also a one tailed test with $\alpha=0.05$. the critical point is $Z_{0.05}=1.65$ We shall reject H_0 if Z<-1.65



c) Computation

$$Z = \frac{\bar{\bar{x}} - \mu_0}{S_{\bar{x}}} = \frac{164.65-166.7}{\frac{5.89}{\sqrt{114}}} = -3.93$$

d) Since z= -3.93 is less than -1.65, we reject the null hypothesis. This means the height of adults in the slum section of the town significantly lower than 166.7 cm at 0.05 level of signicance.

Pearson's 'r' Correlation

Pearson's correlation coefficient, often denoted by r, measures the strength and direction of the linear relationship between two continuous variables

1. Formula

The formula for Pearson's correlation coefficient r is:

$$r = \frac{n\sum(XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

2. Interpretation

Type of correlation	Correlation coefficient
Perfect positive correlation	r = +1
Partial positive correlation	0 < r < +1
No correlation	r = 0
Partial negative correlation	0 > r > -1
Perfect negative correlation	r = -1

Pearson's 'r' Correlation

3. Significance Testing

To determine if the correlation is statistically significant, you can use a t-test:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The t-value can then be compared to the critical value from the t-distribution table with n−2n - 2n−2 degrees of freedom to determine significance.

Pearson's 'r' Correlation-Problem

We have collected data on the length of hands (in centimeters) and height (in centimeters) for 5 individuals. The data is as follows:

Calculate and Interpret the Correlation coefficient of the two variable		
Person	Hand	Height
A	17	150
B	15	154
C	19	169
D	17	172
E	21	175

	x	y	xy	x^2	y^2
A	17	150	2550	289	22500
B	15	154	2310	225	23716
C	19	169	3211	361	28561
D	17	172	2924	289	29584
E	21	175	3675	441	30625
Total	89	820	14670	1605	134986

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$$

r = 0.72

The formula gives us a correlation coefficient of **0.72**, (Critical value is 832) which is a high, positive correlation. Meaning that in this data set, as height increases, so does hand height.

Pearson's 'r' Correlation-Problem

Calculate and Interpret the Correlation coefficient of the two variable		
Person	Weight	Blood Pressure
A	**150**	**125**
B	169	130
C	175	160
D	180	169
E	200	150

	x	y	xy	x^2	y^2
A	**150**	**125**	18750	22500	15625
B	169	130	21970	28561	16900
C	175	160	28000	30625	25600
D	180	169	30420	32400	28561
E	200	150	30000	40000	22500
Total	874	734	129140	154086	109186

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$$

r = 0.61

Weight and blood pressure have a moderate, positive correlation.

Pearson's 'r' Correlation-Problem

1. Suppose you computed r = 0.801 using n = 10 data. Explain significant of the data

From the Table for n =10 the critical values ± 0.685 then
r is not significant between + 0.685 and – 0.685 but it is significant > 0.685

2. Suppose you computed r = -0.624 using n = 14 data. Explain significant of the data

From the Table for n =14 the critical values ± 0.592 then
r is not significant between + 0.592 and – 0.592 but -0.624 it is significant s

3. Suppose you computed r = 0.776 using n = 6 data. Explain significant of the data

From the Table for n =6 the critical values ± 0.832 then
r is not significant between + 0.832 and – 0.832 but it is significant > 0.811

ANOVA (analysis of Variance)

- ✔ ANOVA is a collection of statistical models used to analyse the difference between group means or variance
- ✔ Compares Multiple groups at on time
- ✔ Developed by R A Fischer

There are two ANOVA

1. One Way ANOVA
2. Two Way ANOVA

One Way ANOVA

- ✔ Compares two or more unmatched groups when data are categorised in on Factor
- ✔ E.g Comparing a control group with three different does aspirin
- ✔ Comparing the proceptivity of three or more employees based on working hours in a company

ANOVA-Problems- One Way

1. Set up an analysis of variance table for the following per acre production data for three Varieties of wheat, each grown on four plots. Consider variety differences to be significant

	Per Area Production data variety of Wheat		
Plot of Land	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

ANOVA-Problems- One Way

Solutions

A	B	C		A^2	B^2	C^2
6	5	5		36	25	25
7	5	4		49	25	16
3	3	3		9	9	9
8	7	4		64	49	16
$\sum A = 24$	$\sum B = 20$	$\sum C = 16$		$\sum A^2 = 24$	$\sum B^2 = 20$	$\sum C^2 = 16$
$\sum X = \sum A + \sum B + \sum C = 60$				$\sum X^2 = \sum A^2 + \sum B^2 + \sum C^2 = 332$		

Step 5

DoF for sum of squares = N-1 = 12-1 = 11

DoF for sum of squares between SSC = K-1 = 3-1 =2

DoF for squares with the groups = N-K = 9

Step 1: Null Hypothesis (H0): let us take the hypothesis that there is no significant difference in production of three varieties

Step 2: Correction Factor (CF)

$CF = (\sum X)^2/N = 60^2/12 = 300$

Step 3: Sum of Squares Total = $\sum X^2 - CF = 332-300$

SST = 30

Step 4: Sum of squares between SSB

$SSB = (\sum A)^2/4 + (\sum B)^2/4 + (\sum C)^2/4 - CF = 8$

Step 5: SSW = SST-SSB = 30 – 8 = 24

ANOVA-Problems- One Way

Solutions

Sum of variance	Sum of Squares	DoF	Mean of Squares	F Ratio	F critical for 5% level
Between Sample	SSB = 8	K-1 = 2	8/2 = 4	4 / 2.67 =1.5	F(2,9) =19.45 H0 accepted
With in sample	SSW = 24	N-K = 9	24/9 = 2.67		
Total	32	N-1 =11			

ANOVA-Problems- One Way

2. Is there a statistically significant difference in the mean weight loss among the four diets? We will run the ANOVA using the five-step approach

Per Area Production data variety of Wheat			
Low Calorie(A)	Low Fat(B)	Low carbohydrate(C)	Control (D)
8	2	3	2
9	4	5	2
6	3	4	-1
7	5	2	0
3	1	3	3

ANOVA-Problems- One Way

Solutions

A	B	C	D		A^2	B^2	C^2	D^2
8	2	3	2		64	4	9	4
9	4	5	2		81	16	25	4
6	3	4	-1		36	9	16	1
7	5	2	0		49	25	4	0
3	1	3	3		9	1	9	9
$\sum A$ =33	$\sum B$ =15	$\sum C$ =17	$\sum C$ =6		$\sum A^2$ =239	$\sum B^2$ =55	$\sum C^2$ =63	$\sum D^2$ =18
$\sum A+\sum B+\sum C+\sum D=71$					$\sum X^2 =\sum A^2 +\sum B^2+\sum C^2 +\sum D^2 =375$			

Step 5

DoF for sum of squares = N-1 = 20-1 = 19

DoF for sum of squares between SSB = K-1 = 4-1 =3

DoF for squares with the groups = N-K = 20-4 = 16

Step 1: Set up hypotheses and determine level of significance. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ H_1: Means are not all equal α=0.05

Step 2: Correction Factor (CF)

$CF = (\sum X)^2/N = 71^2/20 = 252.5$

Step 3: Sum of Squares Total = $\sum X^2$ –CF = 375-252.5 SST = 119.95

Step 4: Sum of squares between SSB

$SSB= (\sum A)^2/5 + (\sum B)^2/5 + (\sum C)^2/5 + (\sum D)^2/5$ –CV = 75.75

Step 5: Sum of squares with the groups

SSW = SST-SSB = 119.95 – 75.75 = 47.2

ANOVA-Problems- One Way

Solutions

Sum of variance	Sum of Squares	DoF	Mean of Squares	F Ratio	F critical for 5% level
Between Sample	SSB = 75.75	K-1 = 3	75.75/3 = 25.25	25.25 / 2.95 =8.56	F(3,16) =3.25 H0 Rejected
With in sample	SSW = 47.2	N-K = 16	47.2/16 = 2.95		
Total	122.95	N-1 =19			

We reject H_0 because $8.43 \geq 3.24$. We have statistically significant evidence at $\alpha=0.05$ to show that there is a difference in mean weight loss among the four diets.

ANOVA-Problems- One Way

3. Is there a statistically significant difference in mean calcium intake in patients with normal bone density as compared to patients with osteopenia and osteoporosis? We will run the ANOVA using the five-step approach.

Normal Bone Density (A)	**Osteopenia (B)**	**Osteoporosis (C)**
1200	1000	890
1000	1100	650
980	700	1100
900	700	1100
780	500	400
800	700	350

ANOVA-Problems- One Way

Solutions

A	B	C		A^2	B^2	C^2
1200	1000	890		1440000	1000000	792100
1000	1100	650		1000000	1210000	422500
980	700	1100		960400	490000	1210000
900	700	1100		810000	490000	1210000
780	500	400		608400	250000	160000
800	700	350		640000	490000	122500
5660	4700	4490		5458800	3930000	3917100
$\sum X = \sum A + \sum B + \sum C = 14850$				$\sum X^2 = \sum A^2 + \sum B^2 + \sum C^2 = 13305900$		

Step 5

DoF for sum of squares = N-1 = 18-1 = 17

DoF for sum of squares between SSC = K-1 = 6-1 =5

DoF for squares with the groups = N-K = 13

Step 1: Set up hypotheses and determine level of significance H_0: $\mu_1 = \mu_2 = \mu_3$ H_1: Means are not all equal $\alpha=0.05$

Step 2: Correction Factor (CF)

$CF = (\sum X)^2/N = 12251250$

Step 3: Sum of Squares Total $SST = \sum X^2 - CF = 1054650$

Step 4: Sum of squares between SSB

$SSB = (\sum A)^2/4 + (\sum B)^2/4 + (\sum C)^2/4 - CV = 129700$

Step 5: SSW = SST-SSB = 924950

ANOVA-Problems- One Way

Solutions

Sum of variance	Sum of Squares	DoF	Mean of Squares	F Ratio	F critical for 5% level
Between Sample	SSB = 129700	K-1 = 5	129700/5 = 25940	4 / 2.67 =1.5	F(5,13) =3.68 H0 accepted
With in sample	SSW = 924950	N-K = 13	924950/13 =71150	71150/259 40 = 2.74	
Total	1054650	N-1 =17			

we do not reject H_0 because 1.395 < 3.68. We do not have statistically significant evidence at a =0.05 to show that there is a difference in mean calcium intake in patients with normal bone density as compared to osteopenia and osterporosis.

ANOVA-Problems- Two Way

Used to determine the effect of two nominal predictor variables on a continuous outcome variable

It analyses the effect of the independent variables on the expected outcome along with their relationship to the outcome itselt.

Ex. Comparing the employee productivity based on the working hours and working conditions.

Assumptions of ANOVA

- ✔ The sample are independent and selected randomly
- ✔ Parent population from which samples are taken is of normal distribution
- ✔ Various treatment and environmental effects are additive in nature
- ✔ The experimental errors are distributed normally with mean zero and variance σ^2

ANOVA-Problems- Two Way

ANOVA compares variance by means of F ratio

$$F = \frac{Variance\ between\ sampels}{Variance\ within\ sample}$$

It again depends on experimental design

Null hypothesis

H_0 All population means are same

- ✓ If the computed F_c is greater than F-critical values we are likely to reject the null hypothesis
- ✓ If the computed F_c lesser than the F critical value, then the nuyll hypothesis is accepted.

ANOVA-Problems- Two Way

ANOVA compares variance by means of F ratio

$$F = \frac{Variance\ between\ sampels}{Variance\ within\ sample}$$

It again depends on experimental design

Null hypothesis

H_0 All population means are same

- ✓ If the computed F_c is greater than F-critical values we are likely to reject the null hypothesis
- ✓ If the computed F_c lesser than the F critical value, then the nuyll hypothesis is accepted.

ANOVA-Problems- Two Way

ANOVA compares variance by means of F ratio

$$F = \frac{Variance\ between\ sampels}{Variance\ within\ sample}$$

It again depends on experimental design

Null hypothesis

H_0 All population means are same

- ✓ If the computed F_c is greater than F-critical values we are likely to reject the null hypothesis
- ✓ If the computed F_c lesser than the F critical value, then the nuyll hypothesis is accepted.

ANOVA-Problems- Two Way

1. A farmer applied three types of fertilizers on Four separate plots. Two Figure on yield per acre are tabulated bellow

	Yield			
Plots Fertilizers	A	B	C	D
Nitrogen	**6**	**4**	**8**	**6**
Potash	7	6	6	9
Phosphates	8	5	10	9

Find out if the plots are materially different in fertility as also, if three fertilizers any material difference in yield

ANOVA-Problems- Two Way

Step 1: Null hypothesis: Let us take the hypothesis that

i. All plots are not significantly differ in fertility (Column wise Analysis)

ii. All the fertilizers are not significantly differ in the yields (Row wise analysis)

Step 2:

A	B	C	D	Total	A^2	B^2	C^2	D^2
6	**4**	**8**	**6**	24	36	16	64	36
7	6	6	9	28	49	36	36	81
8	5	10	9	32	64	25	100	81
21	15	24	24	T=84	149	77	200	198

Step 3: Correlation Factor (CF) = $(\sum x)^2/N = 88^2/12 = 588$

Step 4: Sum of Square Total (SST) = $\sum x^2 - C = 149 + 77 + 200 + 198 - 588 = 36$

Step 5: Sum of square Between Columns = SSC= $(\sum A)^2/n + (\sum B)^2/n + (\sum C)^2/n + (\sum D)^2/n - C =$

$21^2/3 + 15^2/3 + 24^2/3 + 24^2/4 - 588 = 18$

ANOVA-Problems- Two Way

Step 6: Sum of square between Rows =SSR = $24^2/4+28^2/4+32^2/4 -588 = 8$

Step 7: Sum of square with the Group = SSW = SST –SSC –SSR = 36 -18 – 8 =10

Step 8:

DoF for total sum of square = N-1 =12-1 = 11

DoF for sum of squares between the column = C -1 = 4-1 = 3

Dof for sum of squares between the Rows = R-1 =3-1 = 2

DoF for sum of square with the Group = (C-1)(R-1) = 6

Step 9: Anova construction

Source of variance	Sum of Squares	Degree of Freedom	Mean of sum of squares	F Ratio	Table value
Between Column	SSC =18	C-1 = 4-1 = 3	18/3 =6	6/1.667 =3.6	F(3,6) = 8.94 H0 accepted
Between Rows	SSR =8	R-1 = 3-1 =2	8/2 =4	4/1.667 = 2.4	F(2,6) = 19.33 H) Accepted
With the Group	SSW =10	(R-1)(c-1) = 6	10/6 =1.667		
Total	SST = 36	N-1 = 11			

ANOVA-Problems- Two Way

Step 10: Interpretation

Columns wise analysis: The computed values of F =3.6 is less than the Table value 8.94, hence he null hypothesis is accepted, it means the plots are not significantly differ I fertility

Row wise Analysis: The calculated value of F =2.4 is less than the Table Value 19.33, hence the null hypothesis is accepted, it means the fertilizers are alike sof for as productivity concern

ANOVA-Problems- Two Way

2. A Manger applied three five type of operator and four type of machines. Two Figure on production are tabulated bellow

Operataor (j)	Machne (i)			
	1	2	3	4
1	46	56	35	47
2	54	55	51	56
3	48	56	50	58
4	46	60	51	59
5	51	53	53	55

Find out if the operator are materially different in Machine as also, if four machines any material difference in yield

ANOVA-Problems- Two Way

Step 1: Null hypothesis: Let us take the hypothesis that

i. All operator are not significantly differ in productivity (Column wise Analysis)

ii. All the Machines are not significantly differ in the productivity (Row wise analysis)

Step 2:

0	1	2	3	4	R-sum						
1	46	56	35	47	184		2116	3136	1225	2209	
2	54	55	51	56	216		2916	3025	2601	3136	
3	48	56	50	58	212		2304	3136	2500	3364	
4	46	60	51	59	216		2116	3600	2601	3481	
5	51	53	53	55	212		2601	2809	2809	3025	
C-Sum	245	280	240	275	T=1040		12053	15706	11736	15215	$\sum X^2$ = 54710

Step 3: Correlation Factor (CF) = $(\sum x)^2/N = 1040^2/12 = 54080$

Step 4: Sum of Square Total (SST) = $\sum x^2 - C = 57710-54080 = 630$

Step 5: Sum of square Between Columns = SSC= $(\sum 1)^2/n + (\sum 2)^2/n + (\sum 3)^2/n + (\sum 4)^2/n - C = 250$

ANOVA-Problems- Two Way

Step 6: Sum of square between Rows =SSR = 54264–54080 = 184

Step 7: Sum of square with the Group = SSW = SST –SSC –SSR = 196

Step 8:

DoF for total sum of square = N-1 =20-1 = 19

DoF for sum of squares between the column = 4 -1 = 4-1 = 3

Dof for sum of squares between the Rows = R-1 =5-1 = 4

DoF for sum of square with the Group = (C-1)(R-1) = 12

Step 9: Anova construction

Source of variance	Sum of Squares	Degree of Freedom	Mean of sum of squares	F Ratio	Table value
Between Column	SSC =250	C-1 = 4-1 = 3	250/3 =83.33	83.33/16.33 =5.1	F(3,12) = 3.49 H0 rejected
Between Rows	SSR =184	R-1 = 5-1 =4	184/4 =46	46/16.33 = 2.81	F(4,12) = 3.26 H) Accepted
With the Group	SSW =196	(R-1)(c-1) = 12	196/12 =16.33		
Total	SST = 36	N-1 = 19			

ANOVA-Problems- Two Way

Step 10: Interpretation

From the F-tables F(4,12) = 3.26 and F(3,12) = 3.49.

- Since 6.81 > 3.49 we conclude that we do have sufficient evidence at the 5% level of significance to reject the null hypothesis that there in difference between the machines.

- Since 2.81 < 3.26 we conclude that we do not have sufficient evidence to reject the null hypothesis that there is no difference between the operators.

Non-Parametric Tests

- While most common statistical analyses (e.g., t-tests, ANOVA) are parametric, they need to fulfil a number of criteria before we use them
- These criteria include satisfying the assumptions of outliers, linearity, normality, homoscedasticity, to name a few
- If the data do not fulfil the criteria to conduct the parametric tests, we can opt for non-parametric tests, which do not require those assumptions
- Do note that non-parametric tests make *less* assumptions, not *no* assumptions!
- The trade-off is that non-parametric tests are generally lower in power

Non-Parametric Tests

- Techniques that do not rely on data belonging to any particular distribution
- Nonparametric statistics do not assume any underlying distribution of parameter
- Nonparametric does not meant that model lack parameters but that the number and nature of the parameters are flexible
- There are some situations when it is clear that the outcome does not follow a normal distribution. These include:
  - When the outcome is an ordinal variable or a rank
  - When there are definite ouliers
  - When the outcome has clear limits of detection

Difference between Parametric and Non-Parametric Tests

Parametric test	Non –parametric test
It is use when the information about the population parameters is completely known .	It is use when there is no or few information available about the population parameters
It assumes that the data is normally distributed.	It makes no assumptions about the distribution of data.
Interval scale or ratio scale	Nominal and ordinal scales
It uses mean	It uses median
More powerful than non parametric	Less powerful
Eg Independent sample T test, paired sample T test, one way ANOVA can be use.	Eg Mann-whitney test, Wilcoxon signed rank test, Kruskal-wallis test can be used.

Difference between Parametric and Non-Parametric

Description	Parametric	Non-Parametric
Assumed Distribution	Normal	Any
Typical data	Ratio or interval	Nominal or ordinal
Usual Central Measures	Mean	Median
Benefits	Can draw Many conclusion	Simplicity less affecte Outliers
TESTS		
One sample test		Chi-square test One Sample sign test
Independent Measures, 2 groups	Independent Measure t-test	Mann-Whitney test
Independent Measures, >2 groups	One way independent measure ANOVA	Kruskal Walles Test
Repeated Measures, 2 Condition	Matched pair t-test	Wilcoxon text

Chi-Square test

1. ***Chi-square test (x^2)*** The Chi-Square test is used to assess whether observed frequencies in categorical data differ significantly from expected frequencies. It is commonly applied in two main types:

   Chi-Square Test of Independence: Evaluates if two categorical variables are independent or associated.

   Chi-Square Test of Goodness of Fit: Determines if a sample data matches the expected distribution.

Chi-Square test

Chi-Square Test of Independence: Procedure

1.Set Up the Hypotheses:

Null Hypothesis (H_0): The variables are independent (no association).

Alternative Hypothesis (H_1): The variables are dependent (there is an association).

2. Create a Contingency Table:Tabulate the observed frequencies for each combination of the variables.

3. Calculate the Expected Frequencies: $$E_{ij} = \frac{(\text{row total} \times \text{column total})}{\text{grand total}}$$

4. Compute the Chi-Square Statistic: (O_{ij} observe frequency)

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

5. Determine the Degrees of Freedom (df) = (number of rows-1) (number of column-1)

6. Compare to the Critical Value:

If $\chi 2_{calculated} > \chi 2_{critical}$, reject the null hypothesis.

Chi-Square test

Suppose we want to test if there is an association between gender and voting preference in an election. We collect the following data:

	Vote for A	Vote for B	Total
Male	30	20	50
Female	25	25	50
Total	55	45	100

Step 1: Hypothesis

H0:Gender and Preference are independent

H1:Gender and Preference are dependent

Create a Contingency Table Chi-Square Calculation

$$E_{\text{Male, A}} = \frac{50 \times 55}{100} = 27.5$$

$$E_{\text{Male, B}} = \frac{50 \times 45}{100} = 22.5$$

$$E_{\text{Female, A}} = \frac{50 \times 55}{100} = 27.5$$

$$E_{\text{Female, B}} = \frac{50 \times 45}{100} = 22.5$$

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(30-27.5)^2}{27.5} + \frac{(20-22.5)^2}{22.5} + \frac{(25-27.5)^2}{27.5} + \frac{(25-22.5)^2}{22.5}$$

=1.01

Degree of freedom = (2-1)(2-1) =1

Critical value α = 0.05 and df = 1 is 3.841

ecision: Since χ2=1.01<3.841 we fail to reject the null hypothesis,

suggesting no significant association between gender and voting preference.

Chi-Square test

Example 1 A die is thrown with following results. Is the die unbiased?

Number of turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Solution: Let us take the hypothesis that the die is unbiased. If that is so, the probability of obtaining any one of the six number is 1/6 and as such the expected frequency of any one number coming upward is E_f = 132 *1/6 = 22. Now we can write the observed frequencies along with expected frequencies and work out the values of λ^2 as follows

Chi-Square test

No. of turned up	Observed Frequency, O_f	Expected Frequency, E_f	$(O_f - E_f)$	$(O_f - E_f)^2$	$(O_f - E_f)^2/E_f$
1	16	22	-6	36	36/22
2	20	22	-2	4	4/22
3	25	22	3	9	9/22
4	14	22	-8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22
Σ	132				9

$\Sigma[(O_f-E_f)^2/E] = 9$

Hence the calculated value of $\lambda^2 = 9$ then Degree of freedom in the given problem is (n-1) = (6-1) =5

The Table value of λ^2 for 5 DoF at 5% level of significance is 11.071. Comparing computed values and table values of λ^2, we find that computed values is less than the table values and as such could have arisen due to fluctuations of sampling. The result, thus, support the hypothesis and it can be concluded that the die is unbiased.

Chi-Square test

2. One sample sign test: It is based on the direction or the plus or minus signs of observation in a sample and not on their numerical magnitudes

Problem: The compressive strength of insulating blocks used in the construction of new houses is tested by a civil engineer. The engineer needs to be certain at the 5% level of significance that the median compressive strength is at least 1000 psi. Twenty randomly selected blocks give the following results:

Observation	Compressive Strength	Observation	Compressive Strength	Observation	Compressive Strength	Observation	Compressive Strength
1	1128.7	6	718.4	11	1167.1	16	1153.6
2	679.1	7	787.4	12	1387.5	17	1423.3
3	1317.2	8	1562.3	13	679.9	18	1122.6
4	1001.3	9	1356.9	14	1323.2	19	1644.3
5	1107.6	10	1153.2	15	788.4	20	737.4

Test (at the 5% level of significance) the null hypothesis that the median compressive strength of the insulting blocks is 1000 psi against the alternative that it is greater

Chi-Square test

Solution The hypotheses are H0 : θ = 1000 and H1 : θ > 1000

Comp. Strength	Sign	Comp. Strength	Sign	Comp. Strength	Sign	Comp. Strength	Sign
1128.7	+	718.4	−	1167.1	+	1153.6	+
679.1	−	787.4	−	1387.5	+	1423.3	+
1317.2	+	1562.3	+	679.9	−	1122.6	+
1001.3	+	1356.9	+	1323.2	+	1644.3	+
1107.6	+	1153.2	+	788.4	−	737.4	−

We have 14 plus signs and the required probability value is calculated directly from the binomial formula as (for binomial distribution p =1/2

$$P(X = r) = \binom{n}{r} q^{n-r} p^r = \binom{n}{r} (1-p)^{n-r} p^r$$

$$P(X \geq 14) = \sum_{r=14}^{20} \binom{20}{r} \left(\frac{1}{2}\right)^{20-r} \left(\frac{1}{2}\right)^r$$

Chi-Square test

$$= \frac{20.19.18.17.16.15}{1.2.3.4.5.6}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18.17.16}{1.2.3.4.5}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18.17}{1.2.3.4}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18}{1.2.3}\left(\frac{1}{2}\right)^{20} + \frac{20.19}{1.2}\left(\frac{1}{2}\right)^{20} + \frac{20}{1}\left(\frac{1}{2}\right)^{20} + \left(\frac{1}{20}\right)^{20}$$

$$= \left(\frac{1}{2}\right)^{20}(38760 + 15504 + 4845 + 1140 + 190 + 20 + 1)$$

$$= 0.05766$$

$= 0.05766$

Since we are performing a one-tailed test, we must compare the calculated value with the value 0.05. Since 0.05 < 0.05766 we conclude that we cannot reject the null hypothesis and that on the basis of the available evidence, we cannot conclude that the median compressive strength of the insulating blocks is greater than 1000 psi.

Chi-Square test

2. A certain type of solid rocket fuel is manufactured by bonding an igniter with a propellant. In order that the fuel burns smoothly and does not suffer either “flame-out” or become unstable it is essential that the material bonding the two components of the fuel has a shear strength of 2000 psi. The results arising from tests performed on 20 randomly selected samples of fuel are as follows:

Observation	Shear Strength	Observation	Shear Strength	Observation	Shear Strength	Observation	Shear Strength
1	2128.7	6	1718.4	11	2167.1	16	2153.6
2	1679.1	7	1787.4	12	2387.5	17	2423.3
3	2317.2	8	2562.3	13	1679.9	18	2122.6
4	2001.3	9	2356.9	14	2323.2	19	2644.3
5	2107.6	10	2153.2	15	1788.4	20	1737.4

Using the 5% level of significance, test the null hypothesis that the median shear strength is 2000 psi.

Chi-Square test

The hypotheses are H0 : θ = 2000 H1 : θ ≠ 2000 We determine the signs associated with each observation as shown below and perform a two-tailed test.

Shear Strength	Sign	Shear Strength	Sign	Shear Strength	Sign	Shear Strength	Sign
2128.7	+	1718.4	−	2167.1	+	2153.6	+
1679.1	−	1787.4	−	2387.5	+	2423.3	+
2317.2	+	2562.3	+	1679.9	−	2122.6	+
2001.3	+	2356.9	+	2323.2	+	2644.3	+
2107.6	+	2153.2	+	1788.4	−	1737.4	−

We have 14 plus signs and the required probability value is calculated directly from the binomial formula:

$$P(X \geq 14) = \sum_{r=14}^{20} \binom{20}{r} \left(\frac{1}{2}\right)^{20-r} \left(\frac{1}{2}\right)^{r}$$

$$= \frac{20.19.18.17.16.15}{1.2.3.4.5.6}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18.17.16}{1.2.3.4.5}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18.17}{1.2.3.4}\left(\frac{1}{2}\right)^{20} + \frac{20.19.18}{1.2.3}\left(\frac{1}{2}\right)^{20} + \frac{20.19}{1.2}\left(\frac{1}{2}\right)^{20} + \frac{20}{1}\left(\frac{1}{2}\right)^{20} + \left(\frac{1}{2}\right)^{20}$$

$$= \left(\frac{1}{2}\right)^{20}(38760 + 15504 + 4845 + 1140 + 190 + 20 + 1) = 0.05766$$

Since we are performing a two-tailed test, we must compare the calculated value with 0.025. Since 0.025 < 0.05766 we cannot reject the null hypothesis on the basis of the evidence and conclude that the median shear strength is not significantly different from 2000 psi.

Chi-Square test

3. A certain type of solid rocket fuel is manufactured by binding an igniter with a propellant. In order that the fuel burns smoothly and does not suffer either “flame-out” or become unstable it is essential that the material bonding the two components of the fuel has a shear strength of 2000 psi. The results arising from tests performed on 10 randomly selected samples of fuel are as follows

Observation	Shear Strength	Observation	Shear Strength
1	2128.7	6	1718.4
2	1679.1	7	1787.4
3	2317.2	8	2562.3
4	2001.3	9	2356.9
5	2107.6	10	2153.2

Using the 5% level of significance, test the null hypothesis that the median shear strength is 2000 psi.

Chi-Square test

Answer The hypotheses are H0 : θ = 2000 H1 : θ ≠ 2000

Shear Strength	Sign	Shear Strength	Sign
2128.7	+	1718.4	−
1679.1	−	1787.4	−
2317.2	+	2562.3	+
2001.3	+	2356.9	+
2107.6	+	2153.2	+

We have 7 plus signs and the required probability value is calculated directly from the binomial formula as

$$P(X \geq 7) = \sum_{r=7}^{10} \binom{10}{r} \left(\frac{1}{2}\right)^{10-r} \left(\frac{1}{2}\right)^{r} = \frac{10.9.8}{1.2.3}\left(\frac{1}{2}\right)^{10} + \frac{10.9}{1.2}\left(\frac{1}{2}\right)^{10} + \frac{10}{1}\left(\frac{1}{2}\right)^{10} + \left(\frac{1}{2}\right)^{10}$$

$$= \frac{10.9.8}{1.2.3}\left(\frac{1}{2}\right)^{10} + \frac{10.9}{1.2}\left(\frac{1}{2}\right)^{10} + \frac{10}{1}\left(\frac{1}{2}\right)^{10} + \left(\frac{1}{2}\right)^{10} = \left(\frac{1}{2}\right)^{10}(120 + 45 + 10 + 1) \simeq 0.172$$

Since we are performing a two-tailed test, we must compare the calculate value with the value 0.025. Since 0.025 < 0.172 we cannot reject the null hypothesis on the basis of the available evidence and we cannot conclude that the median shear strength is different to 2000 psi.

Chi-Square test

The sign test for paired data

Automotive development engineers are testing the properties of two anti-lock braking systems in order to determine whether they exhibit any significant difference in the stopping distance achieved by different cars. The systems are fitted to 10 cars and a test is run ensuring that each system is used on each car under conditions which are as uniform as possible. The stopping distances (in yards) obtained are given in the table below

	Anti-lock Braking System	
Car	1	2
1	27.7	26.3
2	32.1	31.0
3	29.6	28.1
4	29.2	28.1
5	27.8	27.9
6	26.9	25.8
7	29.7	28.2
8	28.9	27.6
9	27.3	26.5
10	29.9	28.3

Use a sign test to test the null hypothesis that the mean difference between the hardnesses produced by the two methods is zero against the alternative that it is not zero. Use the 1% level of significance.

Chi-Square test

Solution We are testing to find any differences in the median stopping distance figures for each braking system. The null and alternative hypotheses are:

H0 : θ1 = θ2 or H0 : θ differences = 0

H1 : θ1 ≠ θ2 or H1 : θ differences 6 ≠ 0

e perform a two-tailed test. The signed differences shown by the two systems are shown in the table below:

	Anti-lock Braking System		
Car	1	2	Sign
1	27.7	26.3	+
2	32.1	31.0	+
3	29.6	28.1	+
4	29.2	28.1	+
5	27.8	27.9	−
6	26.9	25.8	+
7	29.7	28.2	+
8	28.9	27.6	+
9	27.3	26.5	+
10	29.9	28.3	+

We have 9 plus signs and the required probability value is calculated directly from the binomial formula as

$$P(X \geq 9) = \sum_{r=9}^{10} \binom{10}{r} \left(\frac{1}{2}\right)^{10-r} \left(\frac{1}{2}\right)^{r} = \frac{10}{1}\left(\frac{1}{2}\right)^{10} + \left(\frac{1}{2}\right)^{10} = 11 \times \left(\frac{1}{2}\right)^{10} \simeq 0.011$$

Since we are performing a two-tailed test, we must compare the calculated value with the value 0.025. Since 0.011 < 0.025 we reject the null hypothesis on the basis of the available evidence and conclude the the differences in the median stopping distances recorded is significant at the 5% level

Mann-Whitney test

3. Mann-Whitney Test (U test): It is a nonparametric counterpart of the t test used to compare the means of two independent populations.

The following assumptions underlie the use of the Mann-Whitney test

1. The sample are independent
2. The level of data is at least ordinal

The two-tailed hypothesis being tested

H_0: The two population identical

H_a: The two populations are not identical

The problem can be solved two category

Small sample case: When both n_1 and $n_2 \leq 10$

Larger sample case: When both n_1 and $n_2 > 10$

Small sample case: When both n_1 and $n_2 \leq 1$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1 \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2$$

Mann-Whitney test

Is there a difference between health service workers and police service workers in the amount of compensation employers pay them per hour during Corona effect time. (randomly selected).

Health Service Worker (Rupees)	Police Service workers (Rupees)
20.10	26.19
19.80	23.88
22.36	25.50
18.75	2164
21.90	24.85
22.96	25.30
20.75	24.12
	23.45

Hypothesis

H_0: The health service population is identical to the police service population on employee compensation

H_a: The health service population is not identical to the police service population on employee compensation

Mann-Whitney test

Total Employee Compensation	Rank	Group
18.75	1	H
19.80	2	H
20.10	3	H
20.75	4	H
21.64	5	P
21.90	6	H
22.36	7	H
22.96	8	H
23.45	9	P
23.88	10	P
24.12	11	P
24.85	12	P
25.30	13	P
25.50	14	P
26.19	15	P

$W_1 = 1+2+3+4+6+7+8 = 31$

$W_2 = 5+9+10+11+12+13+14+15 = 89$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1 \qquad U_1 = 7 * 8 + \frac{7(7+1)}{2} - 31 = 53$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2 \qquad U_2 = 7 * 8 + \frac{8(8+1)}{2} - 89 = 3$$

Because of U_2 is the smallest values of U, we use U =3 as the test statistic. Because it is the smallest size let n_1 = 7 and n_2 =8

From table P values of 0.0011 x2 = 0.0022 (because two tailed)

The statistical conclusion is that the populations are not identical

Mann-Whitney test

Larger sample case: When both n_1 and $n_2 > 10$

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1 \qquad \mu_U = \frac{n_1 n_2}{2} \qquad \sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Do construction workers who purchase lunch from street vendors sped less per mean than construction workers who go to restaurants for lunch?

Wenders (n_1= 14)	**2.75**	**3.29**	**4.63**	**3.61**	**3.10**	**4.29**	**2.25**	**2.97**	**4.01**	**3.68**	**3.15**	**2.97**	**4.05**	**3.60**		
Restaurant (n_2= 16)	4.10	4.75	3.95	3.50	4.25	4.98	5.75	4.10	2.70	3.65	5.11	4.80	6.25	3.89	4.80	5.50

Mann-Whitney test

Hypothesis

H_0 =The populations of construction worker spending for lunch at vendors and restaurants are the same

H_a = The populations of construction worker spending at vendors is shifted to the left of the population of

Value	Rank	Group	Value	Rank	Group
2.25	1	V	4.01	16	V
2.70	2	R	4.05	17	V
2.75	3	V	4.10	18.5	R
2.97	4.5	V	4.10	18.5	R
2.97	4.5	v	4.25	20	R
3.10	6	V	4.29	21	V
3.15	7	V	4.53	22	V
3.29	8	V	4.75	23	R
3.50	9	R	4.80	24.5	R
3.60	10	V	4.80	24.5	R
3.61	11	V	4.98	26	R
3.65	12	R	5.11	27	R
3.68	13	V	5.50	28	R
3.89	14	R	5.75	29	R

W_1 = 1+3+4.5+4.5+6+7+8+10+11+13+16+17+21+22 = 144

Solving for U, μ_u and σ_u using formula

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1 = 185$$

$$\mu_U = \frac{n_1 n_2}{2} = 112$$

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = 24.1$$

$$Z = \frac{U - \mu_u}{\sigma_U} = 3.03$$

The p-value associated with Z = 3.03 is 0.0012 <0.0022. the null hypothesis is rejected

Kruskal Wallies Test

4. Kruskal Wallies Test: *This test determines whether all of the groups come from the same or equal populations or whether at least one group comes from a different population.*

The process of computing a Kruskal-Wallis K Statistic begins with ranking the data in the groups together, as though they were from one group.

$$K = \frac{12}{n(n+1)}\left(\sum_{i=1}^{c}\frac{T_i^2}{n_i}\right) - 3(n+1)$$

Where

C= number of groups

n = total number of items

T_i = Total of ranks in a group

n_i = number of items in a group

$K = X^2$ with Dof = c-1

Kruskal Wallies Test

Problem: Agribusiness researcher are interested in determining the conditions under which Christmas trees grow fastest. A random sample of equivalent size seedling is divided into four groups. Use the Kruskal-Wallies test to determine whether there is a significant difference in the growth of trees in the groups use a =0.01

Group 1 (Native)	Group 2 (+ water)	Group 3 (+ Fertilizer)	Group 4 (+ water and Fertilizer)
8	1	11	18
5	12	14	20
7	11	10	16
11	9	16	15
9	13	17	14
6	12	12	22

Kruskal Wallies Test

Hypothesis

H_0 : Group 1 = Group 2 = Group 3 = Group 4

H_a :At least one group is different

DoF = C -1 = 4-1 = 3 and α =0.01 (the critical values of chi-square is $X^2_{0.1 \text{ and } 3}$ =11.3449

if K > 11.3449 the decision is to reject the null hypothesis

Assigning Raking																					
5	6	7	8	9	9	10	10	11	11	11	12	12	12	13	14	14	15	16	16	17	18	20	22
1	2	3	4	5.5	5.5	7.5	7.5	10	10	10	13	13	13	15	16.5	16.5	18	19.5	19.5	21	22	23	24

Group 1 (Native)	Group 2 (+ water)	Group 3 (+ Fertilizer)	Group 4 (+ water and Fertilizer)
4	7.5	10	22
1	13	16.5	23
3	10	7.5	19.5
10	5.5	19.5	18
5.5	15	21	16.5
2	13	13	24
T_1 = 25.5 (n_1=6)	T_2 =64.0(n_2=6)	T_3 = 87.5(n_3=6)	T_4 =123.0(n_3=6)

Kruskal Wallies Test

$$\sum_{i=1}^{c} \frac{T_i^2}{n_1} = \frac{25.5^2}{6} + \frac{64^2}{6} + \frac{87.5^2}{6} + \frac{123^2}{6} = 4588.6$$

$$K = \frac{12}{n(n+1)} \left(\sum_{i=1}^{c} \frac{T_i^2}{n_i} \right) - 3(n+1)$$

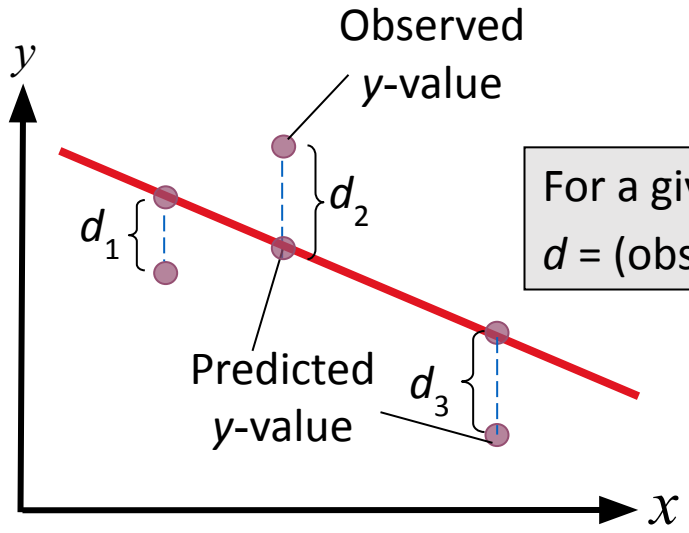
$$K = \frac{12}{24(24+1)} (4588.6) - 3(24+1) = 16.77$$

The observed K values is 16.77 and the critical value $X^2_{0.01,3}$ =11.3449. because the observed values is grater than the table value, the null hypothesis is rejected. There is a significant difference in the way the trees grow.

Linear Regression

Residuals

After verifying that the linear correlation between two variables is significant, next we determine the equation of the line that can be used to predict the value of *y* for a given value of *x*.



For a given *x*-value,
d = (observed *y*-value) – (predicted *y*-value)

Each data point d_i represents the difference between the observed *y*-value and the predicted *y*-value for a given *x*-value on the line. These differences are called **residuals**.

Linear Regression

Example:

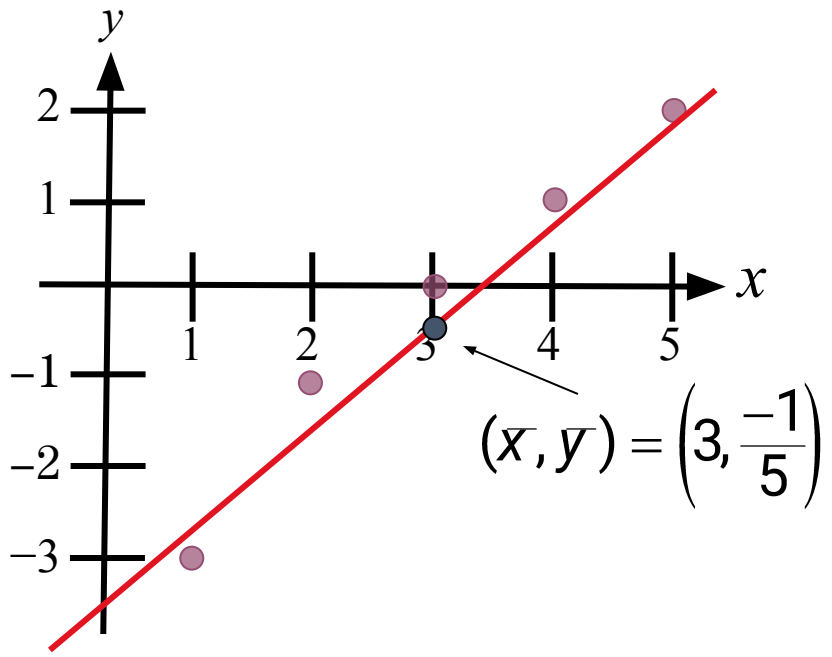
Find the equation of the regression line.

x	y	xy	x^2	y^2
1	-3	-3	1	9
2	-1	-2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{5(9) - (15)(-1)}{5(55) - (15)^2} = \frac{60}{50} = 1.2$$

$$b = \bar{y} - m\bar{x} = \frac{-1}{5} - (1.2)\frac{15}{5} = -3.8$$

The equation of the regression line is $\hat{y} = 1.2x - 3.8$.



Linear Regression

The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Find the equation of the regression line.

b.) Use the equation to find the expected test score for a student who watches 9 hours of TV.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

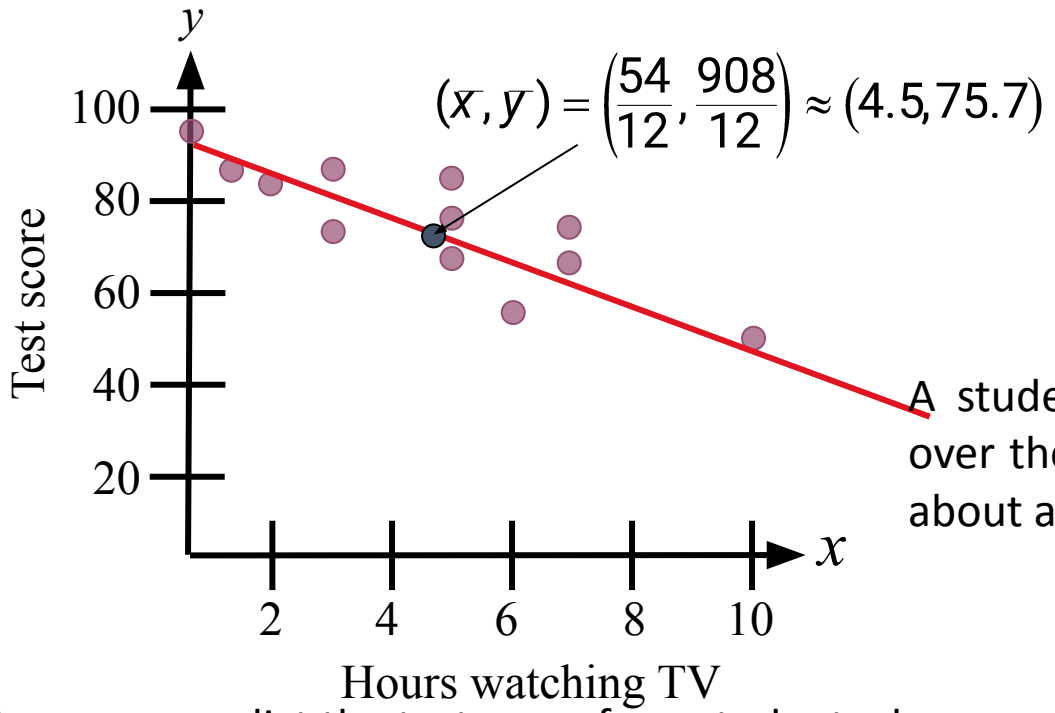
$\sum x = 54$ $\sum y = 908$ $\sum xy = 3724$ $\sum x^2 = 332$ $\sum y^2 = 70836$

Linear Regression

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{12(3724) - (54)(908)}{12(332) - (54)^2} \approx -4.067$$

$$b = \bar{y} - m\bar{x}$$
$$= \frac{908}{12} - (-4.067)\frac{54}{12}$$
$$\approx 93.97$$

$\hat{y} = -4.07x + 93.97$



A student who watches 9 hours of TV over the weekend can expect to receive about a 57.34 on Monday's test.

Using the equation $\hat{y} = -4.07x + 93.97$, we can predict the test score for a student who watches 9 hours of TV.

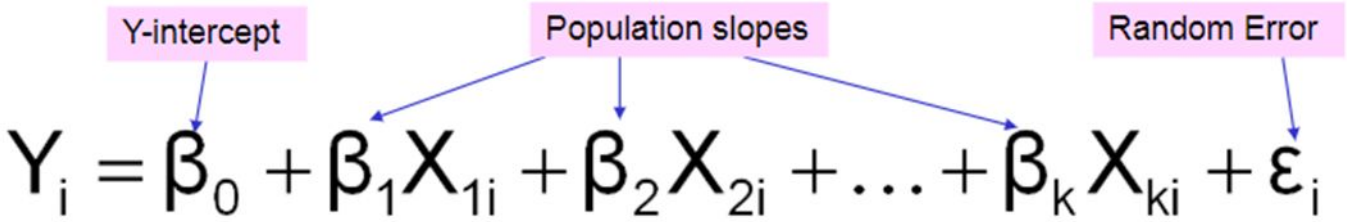
$\hat{y} = -4.07x + 93.97 \quad = -4.07(9) + 93.97 \quad = 57.34$

Multiple Regression

In statistical modeling, **regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable ('outcome variable') and one or more independent variables ('predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which a researcher finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.

Examine the linear relationship between 1 dependent (Y) and 2 or more independent variables (X_j)

Multiple Regression Model with 'k' independent variables



Multiple Regression

Problem: A Distributor of frozen desert pies want to evaluate factors thought to influence demand

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Depend Variable: Pie Sales (units per week)

Independent Variables: Price

Advertising

Data are collected for 15 weeks

Solution

$$\text{Sales} = b_o + b_1 (\text{price}) + b_2 (\text{Advertising})$$

Multiple Regression

	y	x_1	x_2	x_1y	x_2y	x_1x_2	x_1^2	x_2^2
1	350	5.5	3.3	1925	1155	18.15	30.25	10.89
2	460	7.5	3.3	3450	1518	24.75	56.25	10.89
3	350	8	3	2800	1050	24	64	9
4	430	8	4.5	3440	1935	36	64	20.25
5	350	6.8	3	2380	1050	20.4	46.24	9
6	380	7.5	4	2850	1520	30	56.25	16
7	430	4.5	3	1935	1290	13.5	20.25	9
8	470	6.4	3.7	3008	1739	23.68	40.96	13.69
9	450	7	3.5	3150	1575	24.5	49	12.25
10	490	5	4	2450	1960	20	25	16
11	340	7.2	3.5	2448	1190	25.2	51.84	12.25
12	300	7.9	3.2	2370	960	25.28	62.41	10.24
13	440	5.9	4	2596	1760	23.6	34.81	16
14	450	5	3.5	2250	1575	17.5	25	12.25
15	300	7	2.7	2100	810	18.9	49	7.29
	Σ	99.2	52.2	39152	21087	345.46	675.26	185

$$b_1 = \frac{\sum x_1 y \sum x_2^2 - \sum x_1 x_2 \sum x_2 y}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{\sum x_2 y \sum x_1^2 - \sum x_1 x_2 \sum x_1 y}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

$(\Sigma x_1 x_2)^2 = 26814169$

$b_1 = -24.97509$

$b_2 = 74.131$

$b_0 = \frac{\sum y}{n} - \frac{\sum x_1}{n} - \frac{\sum x_2}{n}$

$B_0 = 306.526$

$$\text{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

Multiple Regression

	y	x_1	x_2	x_1y	x_2y	x_1x_2	x_1^2	x_2^2
1	350	5.5	3.3	1925	1155	18.15	30.25	10.89
2	460	7.5	3.3	3450	1518	24.75	56.25	10.89
3	350	8	3	2800	1050	24	64	9
4	430	8	4.5	3440	1935	36	64	20.25
5	350	6.8	3	2380	1050	20.4	46.24	9
6	380	7.5	4	2850	1520	30	56.25	16
7	430	4.5	3	1935	1290	13.5	20.25	9
8	470	6.4	3.7	3008	1739	23.68	40.96	13.69
9	450	7	3.5	3150	1575	24.5	49	12.25
10	490	5	4	2450	1960	20	25	16
11	340	7.2	3.5	2448	1190	25.2	51.84	12.25
12	300	7.9	3.2	2370	960	25.28	62.41	10.24
13	440	5.9	4	2596	1760	23.6	34.81	16
14	450	5	3.5	2250	1575	17.5	25	12.25
15	300	7	2.7	2100	810	18.9	49	7.29
	Σ	99.2	52.2	39152	21087	345.46	675.26	185

$$b_1 = \frac{\sum x_1 y \sum x_2^2 - \sum x_1 x_2 \sum x_2 y}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{\sum x_2 y \sum x_1^2 - \sum x_1 x_2 \sum x_1 y}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

$(\Sigma x_1 x_2)^2 = 26814169$

$b_1 = -24.97509$

$b_2 = 74.131$

$b_0 = \frac{\sum y}{n} - \frac{\sum x_1}{n} - \frac{\sum x_2}{n}$

$B_0 = 306.526$

$$\text{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

Factor Analysis

- FA refers to a method that reduces a large variable into a smaller variable factor. Furthermore, this technique takes out maximum ordinary variance from all the variables and put them in common score.

- Type of Factor Analysis
  - Principal component analysis
  - Common Factor Analysis
  - Image Factoring
  - Maximum likelihood method
  - Other methods of factor analysis

Factor Analysis

Factor analysis is used in the following circumstances

- ✔ To identify underlying dimensions or factors, that explain the correlations among a set of variable
- ✔ To identify a new, smaller, set of uncorrelated variables to replace the original set of correlated variables in subsequently multivariate analyses
- ✔ To identify a smaller set of salient variables from a larger set for use in subsequent multivariate analysis

Mathematically, each variable is expressed as a linear combination of underlying factors.

$X_i = A_{j1}F_1 + A_{j2}F_2 + A_{j3}F3 + \quad + A_{jm}F_m + V_j U_j$

X_j = j^{th} standardized variable

A_j = Standardized multiple regression coefficient of variable

F = Common factor

V_j = Standarized regression coefficient of variable

U_i = The unique factor variable

m= number of common factor

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \ldots + W_{ik}X_k$$

F_i = estimate of *i* th factor

W_i = weight or factor score coefficient

k = number of variables

Factor Analysis

Factor analysis is used in the following circumstances

- ✔ To identify underlying dimensions or factors, that explain the correlations among a set of variable
- ✔ To identify a new, smaller, set of uncorrelated variables to replace the original set of correlated variables in subsequently multivariate analyses
- ✔ To identify a smaller set of salient variables from a larger set for use in subsequent multivariate analysis

Mathematically, each variable is expressed as a linear combination of underlying factors.

$X_i = A_{j1}F_1 + A_{j2}F_2 + A_{j3}F3 + \quad + A_{jm}F_m + V_j U_j$

X_j = j^{th} standardized variable

A_j = Standardized multiple regression coefficient of variable

F = Common factor

V_j = Standarized regression coefficient of variable

U_i = The unique factor variable

m= number of common factor

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \ldots + W_{ik}X_k$$

F_i = estimate of *i* th factor

W_i = weight or factor score coefficient

k = number of variables

Factor Analysis Model

- It is possible to select weights or factor score coefficient so that the first factor explains the largest portion of the total variance
- Then a second set of weights can be selected, so that the second factor accounts for most of the residual variance, subject to being uncorrelated with the first factor
- This same principle could be applied to selecting additional weights for the additional factors

Statistics Associated with Factor Analysis

- **Bartlett's test of Sphericity**: Bartlett's test of spericity is a test statistic used to examine the hypothesis that variables are uncorrelated in the population. In other words, the population correlation matrix is an identity matrix; each variable correlates perfectly with itself(r=1) but has no correlation with the other variables (r=0)
- The following correlation matrix shows the correlation coefficients between different variables for professional basketball teams.

NBA Correlation Matrix				
	Age	Wins	True Shooting %	Turnover %
Age	1.00	0.48	0.56	-0.29
Wins		1.00	0.77	-0.15
True Shooting %			1.00	0.10
Turnover %				1.00

Statistics Associated with Factor Analysis

- **Correlation matrix:** A Correlation matrix is a lower triangle matrix showing the simple correlations, r_1 between all possible pairs of varible included in the analysis. The diagonal element, which are all 1 are usually omitted.

	Variable 1	*Variable 2*	*Variable 3*	*Variable 4*	*Variable 5*	*Variable 6*	*Variable 7*	*Variable 8*	*Variable 9*	*Variable 10*
Variable 1	1									
Variable 2	0.272724	1								
Variable 3	-0.03205	0.267882	1							
Variable 4	0.108167	0.282454	0.344223	1						
Variable 5	-0.03914	-0.36504	-0.01383	-0.37275	1					
Variable 6	0.013927	-0.36842	-0.03052	-0.14839	-0.20021	1				
Variable 7	-0.08183	0.307122	0.596709	0.539091	-0.23986	0.001919	1			
Variable 8	0.270934	-0.17073	0.057761	-0.05328	-0.18467	0.311216	-0.02321	1		
Variable 9	0.213589	0.161635	0.156619	0.159255	-0.26331	0.059653	0.066852	0.18052	1	
Variable 10	-0.03829	-0.07516	-0.31795	-0.12114	-0.21915	-0.28395	-0.34985	-0.07299	-0.31734	1

Statistics Associated with Factor Analysis

- **Correlation matrix:** A Correlation matrix is a lower triangle matrix showing the simple correlations, r_1 between all possible pairs of varible included in the analysis. The diagonal element, which are all 1 are usually omitted.

	Variable 1	*Variable 2*	*Variable 3*	*Variable 4*	*Variable 5*	*Variable 6*	*Variable 7*	*Variable 8*	*Variable 9*	*Variable 10*
Variable 1	1									
Variable 2	0.272724	1								
Variable 3	-0.03205	0.267882	1							
Variable 4	0.108167	0.282454	0.344223	1						
Variable 5	-0.03914	-0.36504	-0.01383	-0.37275	1					
Variable 6	0.013927	-0.36842	-0.03052	-0.14839	-0.20021	1				
Variable 7	-0.08183	0.307122	0.596709	0.539091	-0.23986	0.001919	1			
Variable 8	0.270934	-0.17073	0.057761	-0.05328	-0.18467	0.311216	-0.02321	1		
Variable 9	0.213589	0.161635	0.156619	0.159255	-0.26331	0.059653	0.066852	0.18052	1	
Variable 10	-0.03829	-0.07516	-0.31795	-0.12114	-0.21915	-0.28395	-0.34985	-0.07299	-0.31734	1

Statistics Associated with Factor Analysis

- **Communality:** Communality is the amount of variance a variable shares with all the other variables being considered. This is also the proportion of variance explained by the common factors.

Communalities

	Initial	Extraction
A1 – Environmentalist	1.000	.758
A2 – Go Green	1.000	.568
A3 – Support Green	1.000	.622
A4 – Responsibility	1.000	.708
A6 – Reduce CO2	1.000	.594
A7 – Together	1.000	.499
A8 – Community	1.000	.575
A9 – Collaboration	1.000	.724

Extraction Method: Principal Component Analysis.

Statistics Associated with Factor Analysis

- **Eigenvalue:** The eigenvalue represents the total variance explained by each factor.

Component	Total	Variance (%)	Cumulative (%)
Initial eigenvalues			
1	2.358	29.481	29.481
2	1.172	14.649	44.129
3	1.116	13.950	58.079
4	0.891	11.133	69.212
5	0.831	10.386	79.598
6	0.713	8.916	88.513
7	0.634	7.920	96.433
8	0.285	3.567	100.00
Extraction sums of squared loadings			
1	2.358	29.481	29.481
2	1.172	14.649	44.129
3	1.116	13.950	58.079

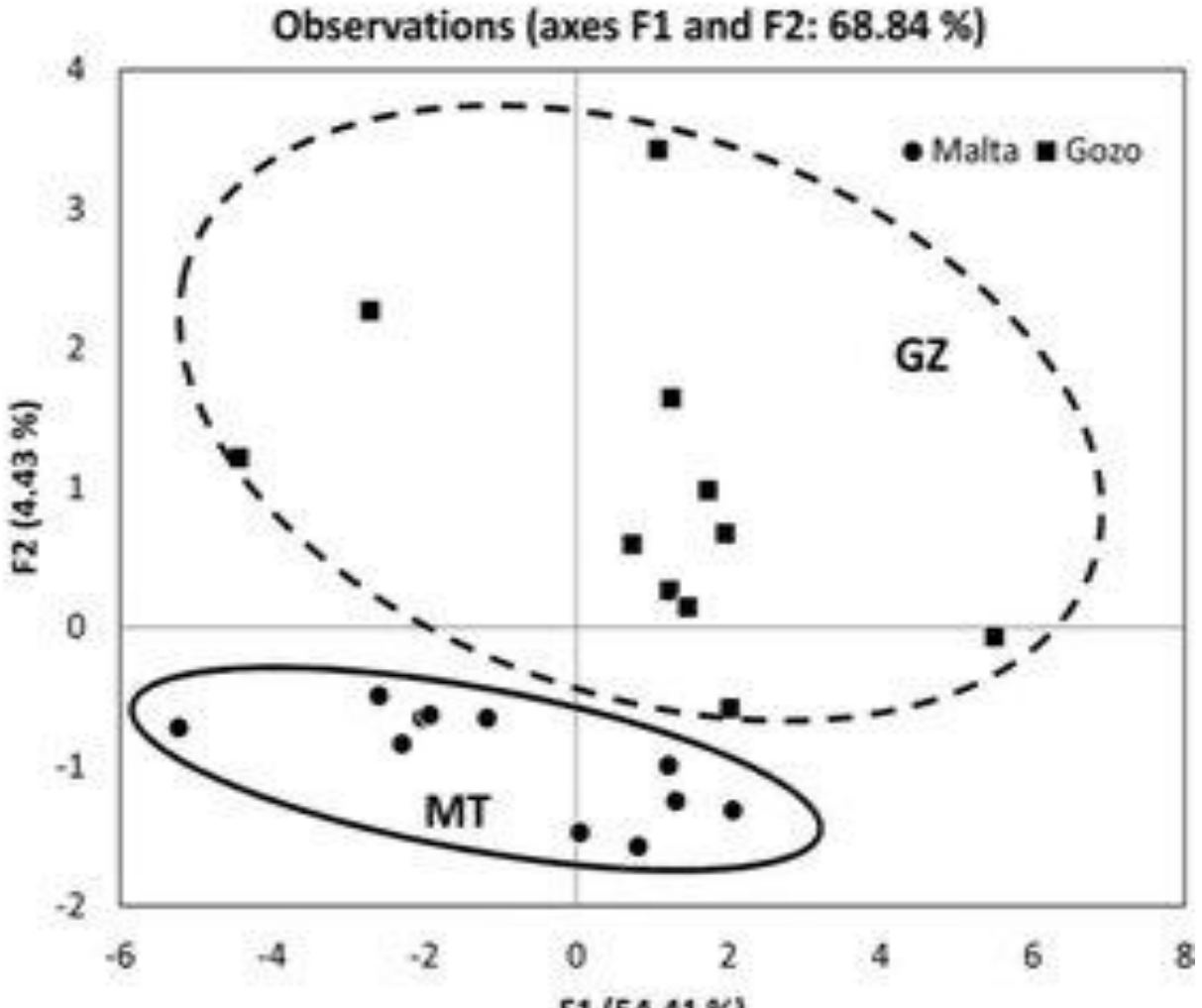
Statistics Associated with Factor Analysis

- **Factor Loadings:** Factor loadings are simple correlation between the variable and factors .

- Factor loading is basically the correlation coefficient for the variable and factor. Factor loading shows the variance explained by the variable on that particular factor. In the SEM approach, as a rule of thumb, 0.7 or higher factor loading represents that the factor extracts sufficient variance from that variable.

Statistics Associated with Factor Analysis

- **Factor Loading plot:** A factor loading plot is a plot of the original variables using the factor loadings as coordinates.



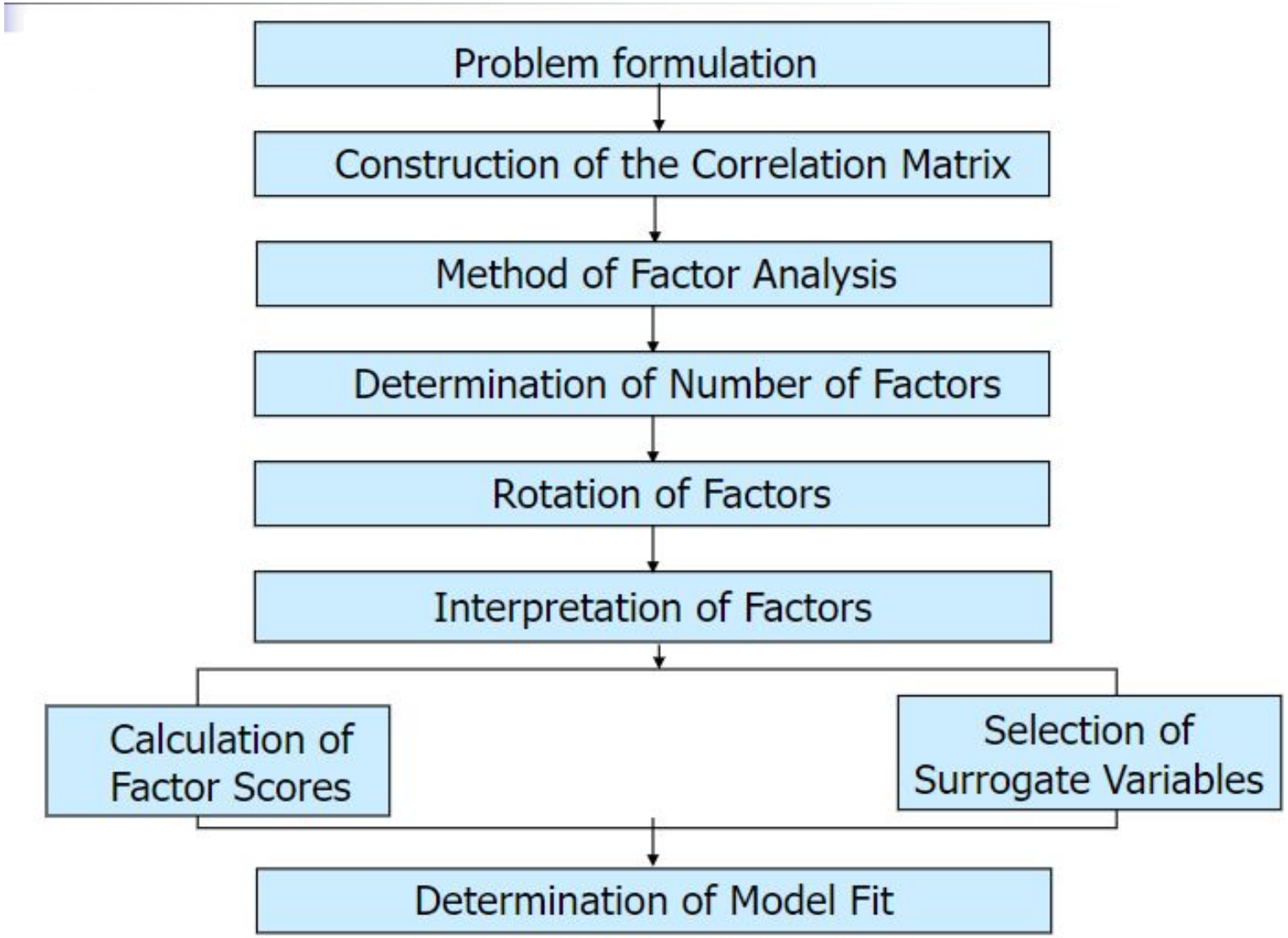
Statistics Associated with Factor Analysis

- **Factor Matrix:** A factor matrix contains the factor loading of all the variables on all the factors extracted. .

Participants	Factor 1	Factor 2	Factor 3
A1	0.7154**X**	0.2708	0.2869
A2	0.4672	0.1668	0.8683**X**
A3	0.4273	0.6806**X**	0.2166
A4	0.7588**X**	0.3648	0.2371
A5	0.5998**X**	0.4460	0.2584
A6	0.6380	0.6633**X**	0.1458
A7	0.8402**X**	0.2095	0.3065
A8	0.0586	0.8470**X**	0.2309
A9	0.7463**X**	0.4589	0.3640
A10	0.7919**X**	0.2217	0.3084
A11	0.6134**X**	0.2642	0.5338
A12	0.6233**X**	0.3902	0.4938
A13	0.5928	0.6131**X**	0.3983
A14	0.3109	0.7106**X**	0.0101
%Explanation Variance	38	25	15

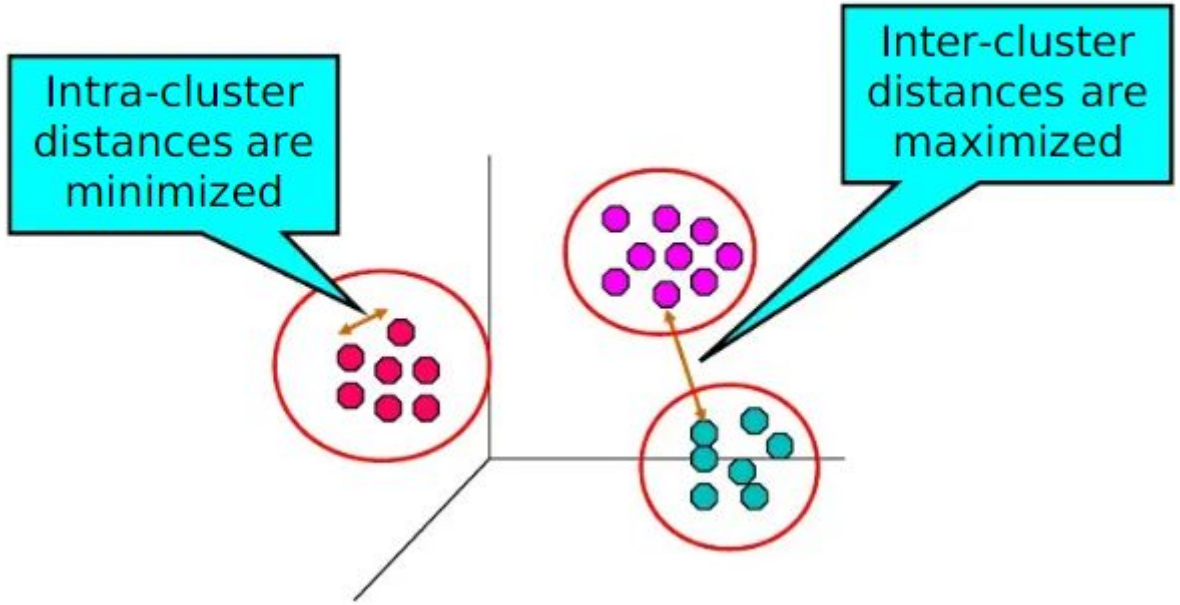
Mean: 0.00; Standard Deviation: 2.130

Conducting Factor Analysis



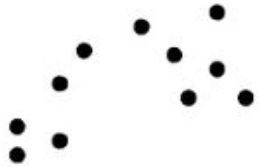
Cluster Analysis

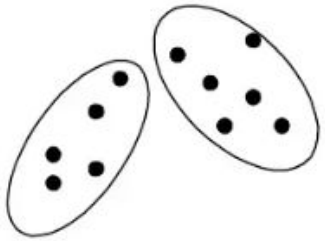
Finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups

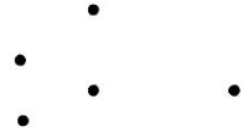


Type of cluster

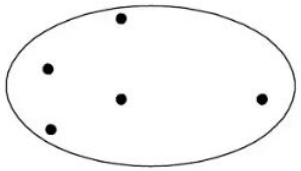
- ✔ A Clustering is a set of clusters
- ✔ Important distinction between hierarchical and partitional sets of cluster
- ✔ Partitional Clustering (A division data objects into non-overlapping subsets (Clusters) such that each data object is in exactly one subset)





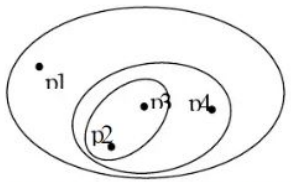


Original Points

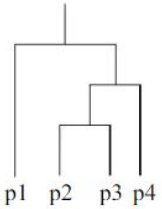


A Partitional Clustering

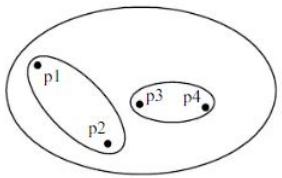
- ✔ Hierarchical Clustering (A set of nested clusters organised as a hierarchical tree)



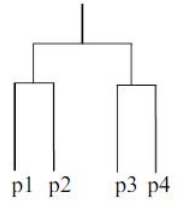
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



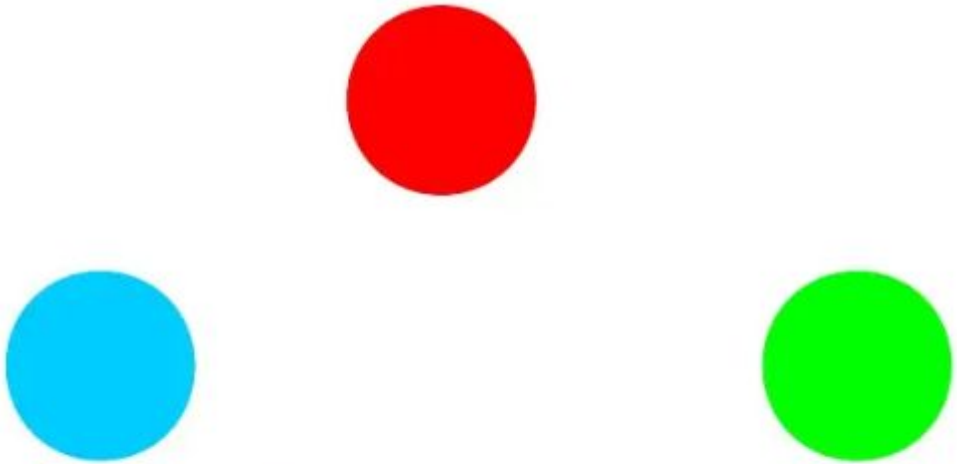
Non-traditional Dendrogram

Type of cluster

- ✔ Well-separated Cluster
- ✔ Centre-based A Clustering is a set of clusters
- ✔ Contiguous Clustering
- ✔ Density based Cluster
- ✔ Property or conceptual
- ✔ Described by Objective Function

Well-Separated Cluster

- ✔ A Cluster is a set of points such that any point in a cluster is closer (more similar) to every other point in the cluster than to any point not in the cluster



3 well-separated clusters

Center-based

- ✔ A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “Cetner” of a cluster, than to the centre of any other cluster.
- ✔ The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, th emost representative point of a cluster

4 center-based clusters

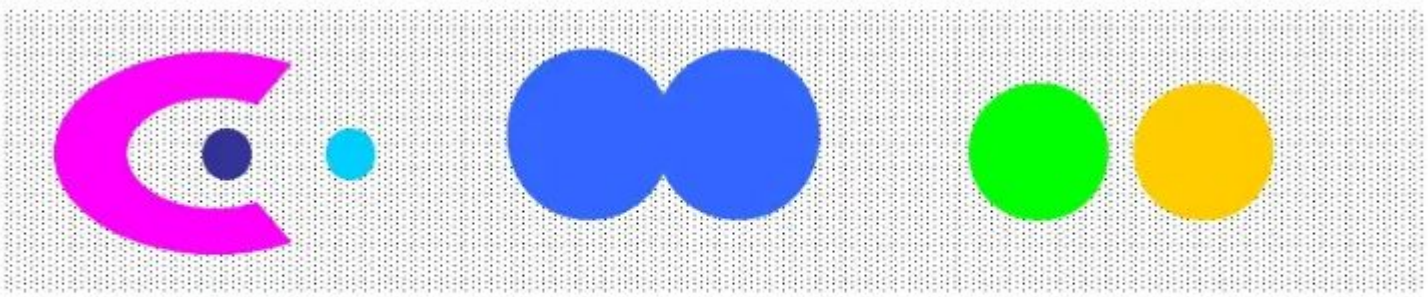
Contiguous Clustering

- ✔ A cluster is a set of points such that a point in a cluster is closer (more similar) to one or more other points in the couster than to any point not in the cluster.

8 contiguous clusters

Density based Cluster

- ✔ A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density
- ✔ Used when the cluster are irregular or intertwined and when noise and outlier are present



6 density-based clusters

Property or Conceptual

- ✔ Finds clusters that share some common property or represent a particular concept