Descriptive Statistics

Unit - I

Statistics:

The science of collecting, organising, presentin analyzing and Intepreting data to assist in making more effective decisions.

statistical analysis - Used to Manipulate, summarise and Investigate data, so that Useful decision making information results.

Why statistics Matters:

statistical procedures are used to advance our knowledge of ourselves and the World around us groups of people, gathering scores, analyzing

Statistics helps to strengthen your critical thinking skills and reasoning abilities

The sun will explode in less than 6 years (9/19/2002) - to evaluate the Validity of such assertions to see if they stand up to scentific scrunity

statistics enable you to Understand and research results in the professional journals of your area of specialization.

Knowledegable and competent professional in your area of study.

STATISTICS AND OTHER STATISTICAL TERMS:

Research, a systemmatic inquiry in search of Knowledge, Involves use of statistics.

Set of scores, referred to as data.

Before the scores have undergone any type of statistical transformation or analysis, they are called raw scores.

A population consists of all elements - individuals, items or objects - whose characteristics are being studied. The population that is being studied is also called target population

A portion of population selected for study is referred to as sample.

Usually populations are so long that a researchee cannot examine the entire group.

Therefore a sample is selected to represent the population in a research study.

The goal is to use the results obtained from the sample to help answer questions about the population.

A Sample drawn in such a way that each element of the population has a chance of being selected is called random Sample.

Samples are normally draw from only the position of the population that is accessible refferred

Discriptive Statistics Sum up and Condense a set of raw Scores so that overall trends in the data become apparent. Percentages and averages are examples of descriptive statistics.

Inferential Statistics Involves predicting characteristics of a population based on data obtained from a sample.

Scales of Measurements :

Nominal Scales:

The least specific Measurement scale only (classify the observations into different categories)

eg: colours, types of trees, Religion

These Variables have no quantitative Values.

Ordinal Scales:

In addition to classify the observations, this scale also (permits Ordering and ranking of these observations)

eg: first, second, third place winners

Interval Scales:

With Interval scales, there is equal distance between units on the scale.

eg. Temparature in degrees is Measured on the Interval scale.

In the Interval scale there is an zero point.

Ratio Scale:

In the ratio scale the real absolute zero point is known.

x	y	x^2	xy	$\Sigma(x-3)$	$y+2$	$(y+2)^2$
6	4	36	24	3	6	36
5	7	25	35	2	9	81
9	2	81	18	6	4	16
8	1	64	8	5	3	9
		206	85	16		142

Organising data Using tables & graphs...

Unorganised leadership Course of 35 Managers

54, 43, 48, 50, 52, 44, 46, 44, 51, 42, 50, 46, 51, 55, 51, 48, 53, 51, 46, 52, 50, 45, 55, 49, 48, 50, 49, 51, 48, 43 Frequency Vs cumulative frequency Vs relative frequency.

X	Tally	F
57	I	1
56	-	0
55	II	2
54	I	1
53	I	1
52	II	2
51	IIII	4
50	IIII	4
49	II	2
48	IIII	4
47	-	0
46	III	3
45	I	1
44	II	2
43	II	2
42	I	1
		30

Relative Frequency Distribution:

The frequency refers to "how many times a particular score occurs", the relative frequency refers to the "portion of the time score occurs"

$$\text{Relative frequency} = \frac{f}{N}$$

N → No. of data in population

N = 26

X	f	Relative Frequency
10	4	4/26 = 0.15 => 15%
9	2	2/26 = 0.08 => 8%
8	6	6/26 = 0.23 => 23%
7	4	4/26 = 0.15 => 15%
6	5	5/26 = 0.19 => 19%
5	3	3/26 = 0.5 => 50%
4	0	0/26 = 0 => 0%
3	2	2/26 = 0.08 => 8%
	N = 26	1.38

Cumulative Frequency Distribution :-

It indicates the frequency of scores that fall at or below a particular score value. this type of table is useful if we want to know how many people scored below certain value on a test.

X	f	Cf	PR (%)
10	4	26	100
9	2	22	84.62
8	6	20	76.92
7	4	14	53.85
6	5	10	38.46
5	3	5	19.23
4	0	2	7.69
3	2	2	7.69

Percentile Rank:

Percentile Rank gives a bit more information than frequency by indicating the percentage of score that fall at below a given score in a distribution.

$$\text{Percentile Rank} = \frac{cf}{N} \times 100$$

29	33	32	34
32	32	31	32
33	32	32	33
32	31	34	32
31	33	33	31

Complete the column with the raw score:

X	Tally	f	Relative frequency	cf	PR
29	I	1	1/20 = 0.2	20	100
31	IIII	4	4/20 = 0.2	19	95
32	~~IIII~~ III	8	8/20 = 0.4	15	75
33	~~IIII~~	5	5/20 = 0.25	7	35
34	II	2	2/20 = 0.10	2	10

N = 20

Groups frequency distribution :

This combines scores into groups refferred as class intervals thus condensing the data and making overall trends more apparent.

class Interval	f
81 - 83	1
78 - 80	3
75 - 77	6

* Interval size symbolised by I refers the no of scores in the class interval.

For eg: I = 3 for the class interval 21 to 23

* The Range symbolised by R refers to the amount of spread in the distribution of the scores.

It is determined by subtracting the lower limit from the upper limit

* Bowling scores of high school gym class

88	128	110	140	127
109	92	119	142	111
126	85	124	138	92
83	114	146	112	86
98	132	128	95	120
122	115	92	116	119
79	118	81	112	115

Upper limit
highest no + 0.5
Lower limit
lowest no - 0.5

Step : 1 Range:

Up. L = 146 , L.L = 79

R = 146 · 5 − 78 · 5

R = 68

step 2 :·

68 ÷ 2 = 34

68 ÷ 3 = 22

68 ÷ 5 = 13

Where i = 2, 3, 5

Here $\boxed{i = 5}$

Class Intervals :·

Here i = 5, so we can take multiple of 5

class Intervals	Tally	f
145 - 149	I	1
⋮		
100 - 104		
95 - 99		
90 - 94	III	3
85 - 89	II	2
80 - 84	III	3
75 - 79	I	1

1 2 1 4 3 5 6 2 2 3 3 1 35

class Intervals	Tally	Frequency f
145 - 149	I	1
140 - 144	II	2
135 - 139	I	1
130 - 134	I	1
125 - 129	IIII	4
120 - 124	III	3
115 - 119	~~IIII~~	5
110 - 114	~~IIII~~ I	6
105 - 109	I	1
100 - 104	–	0
95 - 99	II	2
90 - 94	III	3
85 - 89	II	2
80 - 84	III	3
75 - 79	I	1

N = 35

Measures Of Central tendency:.

Central tendency:.

Central tendency is a single value that describes the most typical or representative score in an entire distribution.

eg: Mode, Median, Mean

The Mode:.

The Mode (Mo) specifies the value with the highest frequency in a set of scores.

To determine the mode simply arrange scores in descending order (or create frequency distributio

If there are numerous scores

Once they are arranged so, it is easy to see at a glance which scored occured with the greatest frequency.

In the set of scores below, 73, 73, 72, 70, 68 68, 68, 68, 59, 59, 59, 55

$$Mo = 68$$

Grouped Frequency Distribution:.

The Mode would be the mid-point of the class interval with greatest frequency.

Class Intervals	Mid point	f	class Intervals	Mid point	f
36 - 38	37	8	21 - 23	22	20
33 - 35	34	11	18 - 20	19	16
30 - 32	31	18	15 - 17	16	12
27 - 29	28	26	12 - 14	13	7
24 - 26	25	32	9 - 11	10	3

Mode = 25

The Median:.

Median (Mdn): Which is the middlepoint in a distribution.

To find the median for an odd number of scores

=> Arrange the scores in descending order from highest to lowest

=> The location of the median will be the score that has an equal number of scores above and below as determined by: $N + 1/2$

26, 25, 24, 20, 18, 17, 17, 15, 12

$(9+1)/2 = 5$

We are looking for 5th score. Thus the median is 18.

Mdn = 18

5 is not median but rather the location.

To find the Median for an even number of scores :-

⇒ Arrange the scores in order from highest to lowest.

⇒ Divide the distribution in half and draw a line between the two scores that separate the distribution in to halves.

⇒ Add the two middle scores that surround the halfway point and divide by 2.

The resulting value will be the Median.

92, 91, 90, 90, 87, 82, 77, 75, 75, 70, 68, 60

Middle scores

$$Mdn = \frac{82+77}{2}$$

Mdn = 79.5

The formula method for determining the median used when working from grouped frequency distribution and cumulative frequency of scores.

$$Mdn = LL + \left(\frac{50\%. \text{ of } N - Cf \text{ below}}{f \text{ wi}} \right) i$$

Where LL = Lower limit of Class Interval that contains Median.

N = number of scores

cf_{below} = cumulative frequency below the class interval that contains Median

f_{wi} = frequency of scores in the interval that contains Medium

i = size of class interval

For example:

Class Interval	f	cf
42 - 44	4	124 → N
39 - 41	8	120
36 - 38	10	112
33 - 35	11	102
30 - 32	8	91
27 - 29	18	83
24 - 26	17	65
21 - 23	16	48
18 - 20	10	32
15 - 17	11	22
12 - 14	5	11
9 - 11	6	6

$$Mdn = 23.5 + \left[\frac{62 - 48}{17}\right] 3$$

$$= 23.5 + 2.47$$

$$= 25.97$$

$$Mdn = 25.97$$

class Intervals	f	cf
60 - 64	2	60
55 - 59	1	58
50 - 54	0	57
45 - 49	5	57
40 - 44	0	52
35 - 39	7	52
30 - 34	13	45
25 - 29	12	32
20 - 24	8	20
15 - 19	5	12
10 - 14	0	7
5 - 9	5	7
0 - 4	2	2

$$Mdn = LL + \left(\frac{50\%.N - cf}{f_{wi}} \right) i$$

$$= 24.5 + \left(\frac{30 - 20}{12} \right) 5$$

$$= 24.5 + 4.16$$

$$Mdn = 28.67$$

The Mean:

The Mean is the sum total of all of the scores in a distribution divided by the total number of scores.

For a population, $\mu = \frac{\Sigma X}{N}$

For a sample, $M = \frac{\Sigma X}{n}$

Let us calculate the mean for the following set of scores from a population

78, 63, 42, 98, 87, 52, 72, 64, 75, 89

$$\mu = \frac{\Sigma X}{N} = \frac{720}{10} = 72$$

The Mean for the following set of scores from a sample involves the same set of calculations.

3, 8, 6, 9, 10, 17, 5, 8, 1

$$M = \frac{\Sigma X}{n} = \frac{67}{9} = 7.44$$

Mean for a simple Frequency Distribution:.

To calculate the mean for scores that have been arranged into a simple frequency distribution the formula is modified as follows:.

For a population,

$$\mu = \frac{\Sigma fX}{N}$$

For a Sample

$$M = \frac{\Sigma fX}{n}$$

Where fX = frequency of score multiplied by the score itself.

Let us calculate the mean for the following scores from a sample arranged into simple frequency distribution table.

x	f	fx	x	f	fx
48	1	48	41	4	164
47	4	188	40	6	240
46	2	92	39	3	117
45	4	180	38	0	0
44	9	396	37	1	37
43	8	344	36	2	72
42	5	210	35	1	35
				n = 50	$\Sigma fx = 2123$

$$\therefore M = \frac{\Sigma fx}{n} = \frac{2123}{50}$$

$$M = 42.46$$

1) The annual incomes for the employees of a new hat company are listed below (in thousands)

65	33	52	47
72	73	68	65
58	49	18	59
31	62	61	74
42	27	54	52
80	78	66	22

Construct a grouped frequency distribution table Using the table, determine the median, mean and mode adding additional columns as needed.

Sol.

$$\text{Range} = U.L - L.L$$

$$= 80.5 - 17.5$$

$$R = 63$$

Class Interval:

$i = 2, 3, 5$

$\frac{63}{2} = 31.5$

$\frac{63}{3} = 21$

$\frac{63}{5} = 12.6$

$\boxed{i = 5}$

Class Interval	frequency		Midpoint	
	f	cf	x	fx
80 - 84	1	24	82	82
75 - 79	1	23	77	77
70 - 74	3	22	72	216
65 - 69	4	19	67	268
60 - 64	2	15	62	124
55 - 59	2	13	57	114
50 - 54	3	11	52	156
45 - 49	2	8	47	94
40 - 44	1	6	42	42
35 - 39	0	5	37	0
30 - 34	2	5	32	64
25 - 29	1	3	27	27
20 - 24	1	2	22	22
15 - 19	1	1	17	17
	24			$\Sigma fx = 1303$

(i) Mode = 67

(ii) Median $= L.L + \left[\frac{50\%.N - c.f \text{ below}}{f\, wi}\right] \times i$

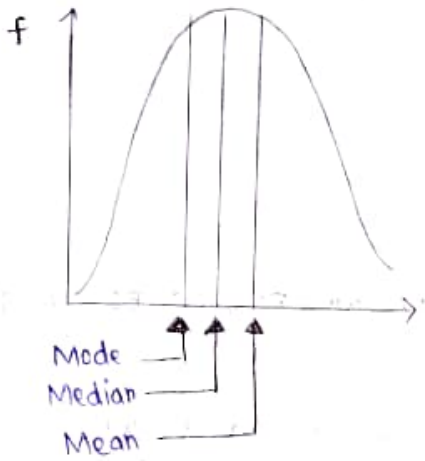
$= 54.5 + \left[\frac{12 - 11}{2}\right] \times 5 = 54.6 + (6.5) \times 5$

Median = 57

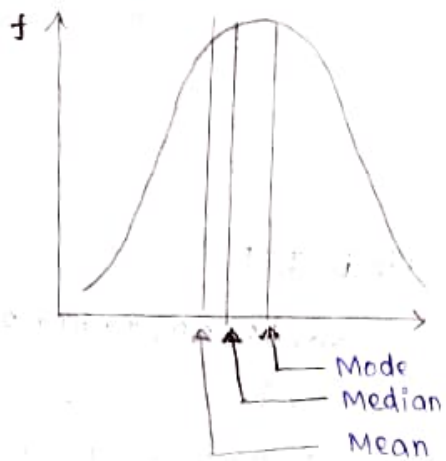
(iii) Mean = $1303 \div 24 = 54.29$

In skewed Distribution, the mode would again be at the peak; the mean would be located towards the tails in the direction of the skew (having been affected by either high or low extreme scores); and the median would be between the mode and the mean (so that half the scores lie above it and half below it).

Positive skew

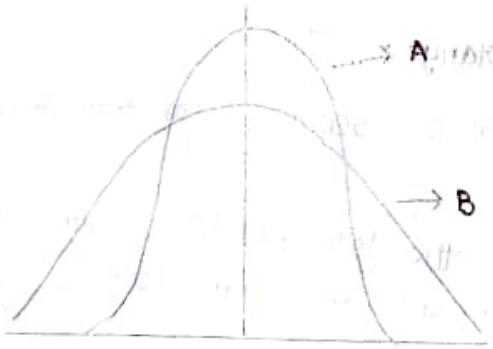


Negative Skew



Measures of Variability:

Measures of Variability provides information about how similar or different the scores are in relationship to other scores in the distribution.



Both of the distributions below have the same mean however they are quite different in the variability of their scores.

Three common measures of variability are the range, the interquartile range, and the standard deviation.

THE RANGE

Subtracting the lower limit of the lowest score from the upper limit of the highest score.

$$R = X_{UL\text{-}High} - X_{LL\text{-}Low}$$

For example:

For the following set of scores

3, 3, 5, 7, 7, 7, 8, 9

$$R = 9.5 - 2.5$$

$$R = 7$$

Find IQR

36, 42, 30, 7, 51, 29, 45, 35, 44, 53, 32, 50, 28, 43 33, 29.

7 28 29 29 | 30 32 33 35 | 36 42 43 44 | 45 50 51 53

$$Q_1 = \frac{29+30}{2}$$

$$Q_1 = 29.5$$

$$Q_3 = \frac{44+45}{2}$$

$$Q_3 = 44.5$$

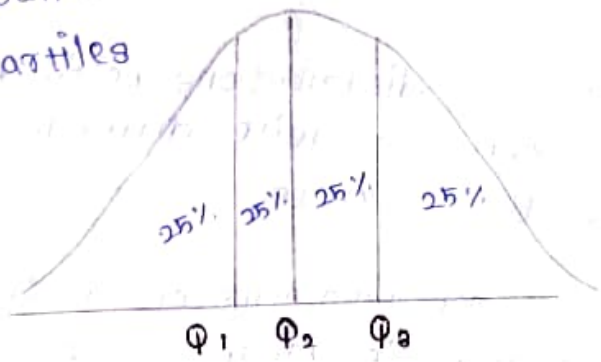
$$IQR = Q_3 - Q_1 = 44.5 - 29.5$$

$$\boxed{IQR = 15}$$

Interquartile range:

The range of scores from the middle 50 % of a distribution.

To determine the interquartile range, we first divide the distribution into four equal parts, which produces three quartiles



Quartiles

Q_1 = the point at or below which 25 % of the scores lie.

Q_2 = the point at or below which 50 % of scores lie

Q_3 = the point at or below which 75 % of scores lie

Standard Deviation:.

Standard Deviation is the measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of set, while the high standard deviation indicates that the values are spread out over wider range.

(i) **Definitional formula** - they are written the way that statistics are defined, these formulas usually involves more computations.

(ii) **Computational formula** - these are easier to use with a hand held calculator

Definitional formula:

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

σ = population standard deviation

X = each value

μ = population mean

N = number of values in population

1) 17, 24, 22, 26, 18, find S.D

X	μ	$X - \mu$	$(X - \mu)^2$
17	21.4	-4.4	19.36
24	21.4	2.6	6.76
22	21.4	0.6	0.36
26	21.4	4.6	21.16
18	21.4	-3.4	11.56

$\Sigma (X - \mu)^2 = 59.26$

Mean (μ) $= \frac{17 + 24 + 22 + 26 + 18}{5}$

$$\sigma = \sqrt{\frac{59.26}{5}}$$

$$\sigma = 3.44$$

2) 3, 3, 0, 1, 1, 2, 2, 2, 6, 0

X	μ	$(X - \mu)$	$(X - \mu)^2$
3	2	1	1
3	2	1	1
0	2	-2	4
1	2	-1	1
1	2	-1	1
2	2	0	0
2	2	0	0
2	2	0	0
6	2	4	16
0	2	-2	4

$$\Sigma (X - \mu)^2 = 28$$

$$\sigma = \sqrt{\frac{28}{10}}$$

$$\sigma = 1.67$$

Sum of squares $= \Sigma (X - \mu)^2$

$$\sigma^2 = \frac{SS}{N} \quad \text{(Variance)}$$

(ii) Computational formula:

$$\sigma = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}}$$

$$\frac{\Sigma (X - \mu)^2}{N}$$

$$\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}$$

(1) 17, 24, 22, 26, 18

X	X^2
17	289
24	576
22	484
26	676
18	324

$\Sigma X = 107$ $\Sigma X^2 = 2349$

$$\frac{(\Sigma X)^2}{N} = \frac{(107)^2}{5} = \frac{11449}{5} = 2,289.8$$

$$\sigma = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}}$$

$$= \sqrt{\frac{2349 - 2289.8}{5}}$$

$$= \sqrt{11.84}$$

$$\boxed{\sigma = 3.44}$$

Unit - II

Inferential Statistics - I

Sampling Distribution:

A distribution of statistics obtained by selecting all the possible samples of specific size from a population.

Distribution of Sample Means:

The collection of sample means for all the possible random samples of a particular size (n) that can be obtained from a population.

The central limit Theorem:

The central limit theorem states that

(i) as the size of the sample (n) increases, the shape of the sampling distribution of means approximates the shape of the normal distribution;

(ii) the mean of the sampling distribution of means will equal the population mean (μ); and

(iii) the standard deviation will equal $\sigma/\sqrt{n}$.

Shape: Even if the shape of the population distribution from which a sample is drawn is not normal, if the sample size is large (i.e $n = 60$ or more), the sampling distribution itself will be normal in shape.

Mean: The mean of the distribution of sample means is the mean of the population. Sample size does not affect the center of the distribution.

Standard Deviation: The standard deviation of the sampling distribution of means is called the standard error of the mean (σ_M) or simply standard error. The standard error (σ_M)

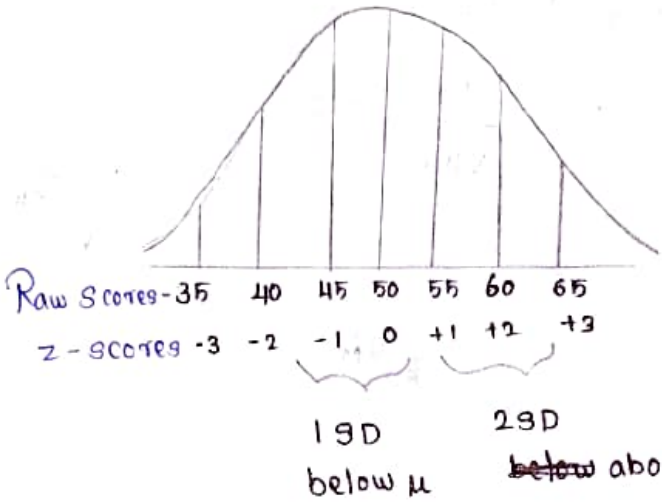
represents the amount that a sample mean (M) is expected to vary from the population mean (μ).

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

Probabilities, Proportions And Percentages of Sample Means:

z-score describes how far a particular raw score deviates from the mean in standard units.

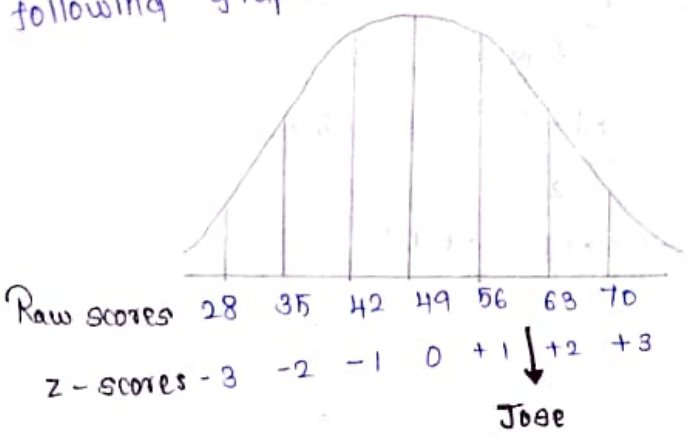
The relationship between raw scores and z-scores is illustrated below:



$$z = \frac{X - \mu}{\sigma}$$

Jose scored 60 on his history test. The class mean was 49, and the standard deviation was 7. What is Jose's equivalent z-score? $z = \frac{60 - 49}{7} = +1.57$

Thus a raw score of 60 converts to a z-score of +1.57, meaning that Jose scored, 1.57 standard deviations above the mean. This situation, is illustrated in following graph:



Probabilities, Proportions And Percentages of Sample Means:

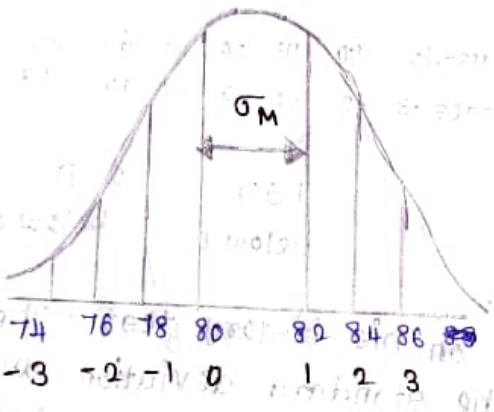
The Sampling distribution of means can be used to determine the probabilities, proportions and percentages associated with particular sample means.

$$z = \frac{M - \mu}{\sigma_M} \quad \text{and} \quad M = \mu + (z)(\sigma_M)$$

$\mu = 80$, $\sigma = 14$ and $n = 49$. Using the appropriate formula we find the standard error (σ_M) to be 2. Thus

$$\sigma_M = \frac{14}{\sqrt{49}} = 2$$

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



1) Given normally shaped distribution with $\mu = 80$ and $\sigma_m = 2$, A. What is the probability that an obtained sample mean will below 81?

$$z = \frac{M - \mu}{\sigma_M}$$

$$= \frac{81 - 80}{2} = +.50$$

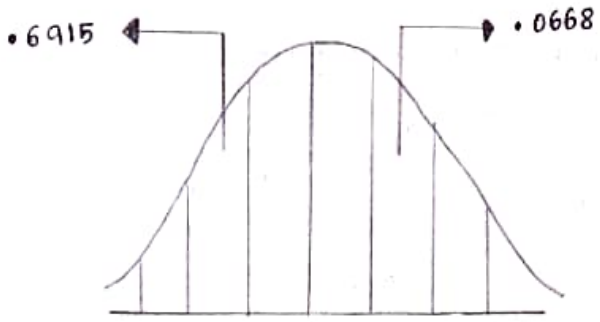
$$P(M < 81) = .6915$$

2) Greater than 83 :

$$z = \frac{M - \mu}{\sigma_M}$$

$$= \frac{83 - 80}{2} = +1.50$$

$P(M < 83) = .0668$



3) Given a normally shaped population distribution with $\mu = 95$ and $\sigma = 5$, a sample size of $n = 25$ is drawn at random. The probability is 0.05 that mean of sample will above what value?

To find standard error :

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

$$\sigma_M = \frac{5}{\sqrt{25}} = 1$$

To find z-score associated with closest proportion above 0.0500, which is 1.65.

Finally convert the z-score to a sample mean

$$M = \mu + (z)(\sigma_M)$$

$$= 95 + (1.65)(1)$$

$$M = 96.65$$

$$z = \frac{M - \mu}{\sigma_M}$$

$$(z)(\sigma_M) = M - \mu$$

$$M = \mu + z(\sigma_M)$$

To find z-score

above (+) → C
above (-) → B
below (-) → C
below (+) → B

What range of sample mean to fall in 70%. of the middle of the distribution?

$\mu = 95, \quad \sigma = 5, \quad \sigma_M = 1$

$M = \mu + (z)(\sigma_M)$

$\Rightarrow M = 95 + (-1.04)(1)$

$M = 93.96$

$\Rightarrow M = 95 + (1.04)(1)$

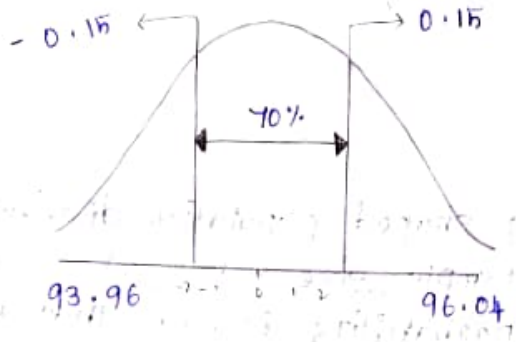
$M = 96.04$

$100\% - 70\% = 30\%$

$= \frac{30}{100} = 0.3$ ↓ middle tall

$\alpha = \frac{0.3}{2}$

$\alpha = 0.15$



1) Finding Probability or proportion of Given Sample Means. Given a normal distribution with $\mu = 100$ and $\sigma = 12$, a sample of $n = 36$ is drawn at random.

A) What is the probability that sample mean will fall above 99?

B) What proportion of sample mean will have value less than 95?

$z = \frac{M - \mu}{\sigma_M} \qquad \sigma_M = \frac{12}{\sqrt{36}} = 2$

$M = z(\sigma_M) + \mu$

$99 = z(\sigma_M) + 100$

$-1 = 2z$

$z = -\frac{1}{2} = -0.5$

Hypothesis Testing :.

=> Hypothesis testing is a procedure used in inferential statistics to estimate population parameters based on sample data.

=> Sample research question → test of reading comprehension for 4th graders is normally distributed with mean $\mu = 70$ and S.D $\sigma = 10$

=> Random sample of N = 25, 4th graders are thaught and then tested for reading comprehension.

→ Sample mean M = 75 is obtained

=> Thus the sample mean M differ enough from population mean (μ) to conclude that reading technique made a difference in level of comprehension.

Testing steps :.

-> Formulate the hypothesis

=> Indicate the α-level and determine the critical values

=> calculate relevant statistics

=> Make decisions and report the result

Formulating the hypothesis :-

(i) Formulate the hypothesis. there will be to mutually exclusive hypothesis

(ii) Either the new reading technique does not have an effect on comprehension or it does

(iii) The null hypothesis (H_0) attributes any difference between obtained sample mean and population mean to chance.

* chance :. (due to chance or random sampling error)

* equality :

* Ineffective treatment

(iv) The alternative hypothesis H_1 called the research hypothesis is the opposite of the null. H_1 describes true differences

(v) The effectiveness of treatment.

Step 1:.

=> $H_0 = \mu = 70$

=> Ho states that the new reading technique will not change the mean level of reading comprehension.

=> The population mean will be 70.

Step 2:.

=> $H_1 = \mu \neq 70$

=> It states that the new reading techniques does have an effect on comprehension

Step 3:.

=> Indicate alpha level and determine the critical values

=> Most of an alpha level is set at 0.05 (or) 0.01 and occassionally at 0.001

Sample means that fall in this areas have a 5% or less likelyhood of occuring by chance alone if Ho is true. $Z_{\text{Critical Values}} = \pm 1.96$

Step 3:.

Calculate the relevant statistics

$$Z_{obt} = \frac{M - \mu}{\sigma_M}$$

Step 4:.

Make a decision and report the results Finally we must either we must reject the null hypothesis or accept the null hypothesis.

SRQ:

Given $\mu = 70$, $\sigma = 10$, $n = 25$, did the new reading technique have effect on Comprehension?

Step 1: Formulate hypothesis

$H_0 : \mu = 70$

$H_1 : \mu \neq 70$

Step 2:

Indicate the alpha level and determine critical values

$\alpha = 0.05$

$z_{crit} = \pm 1.96$

Step 3:

Calculate Relevant Statistics

Given: $n = 25$, $\mu = 70$, $\sigma = 10$, $M = 75$

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

$$z_{obt} = \frac{M - \mu}{\sigma_M}$$

$$= \frac{75 - 70}{2}$$

$$= +2.50$$

Step 4: Make a decision and report the results:

The new reading program had a significant effect on reading comprehension.

Reject H_0, $z = +2.50$ then $p < 0.05$.

Hypothesis Testing With z-Test Research Question:

A standardized productivity scale has a $\mu = 2$ and a $\sigma = 5$. The CEO of company A wants to know if employee participation in company decisions has productivity. A sample of $n = 75$ employees

participated in the decision making process was administrated the productivity scale with a result of $M = 27$. Does participation in company decisions have an effect on Productivity? Use $\alpha = 0.01$ and a two tailed test.

Step 1:-

$H_0 : \mu = 25$

H_0 states that the employee participation in company decisions will not change the Productivity.

$H_1 : \mu \neq 25$

It states that employee participation in company decision will have an effect on productivity

Step 2:-

Indicate alpha level and determine the critical values.

$\alpha = 0.01$

$Z_{crt} = \pm 2.58$

Step 3:-

Given $n = 75$, $\mu = 25$, $\sigma = 5$, $M = 27$

Calculate the relevant statistics

$$Z_{obt} = \frac{M - \mu}{\sigma_M}$$

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{75}} = 0.58$$

$$= \frac{27 - 25}{0.58}$$

$$= \frac{2}{0.58}$$

$$Z = 3.45$$

Step 4:-

Make a decision and report the results The ~~mean~~ ~~reading~~ ~~technique~~. The employee participation has a significant effect on productivity. Reject H_0, $Z = +3.45$, $P < 0.01$.

Effect Size for a z-test:

Effect size indicates magnitude of a treatment effect:

General formula,

$$d = \frac{|M - \mu|}{\sigma}$$

For our research Problem,

$$d = \frac{|75 - 70|}{10}$$

$$d = 0.5$$

$\Rightarrow$ d = 0.20 to 0.49 - Small effect

$\Rightarrow$ d = 0.50 to 0.79 - Moderate effect

$\Rightarrow$ d = 0.80 and above - Large effect

One Sample t-Test:

The One-Sample t test is a test of hypothesis about a population mean (μ) when the population standard deviation (σ) is not known.

This test is used when researchers want to know (1) if a sample is representative of a population and/or (2) if a particular treatment or condition has a significant effect. df = n-1

1) The population mean of a standardized test of critical thinking is $\mu = 53$. A group of faculty members at a small community college underwent a 10-week training program to learn techniques designed to help students develop their critical thinking skills. After the training, the new techniques were implemented in the classrooms. The mean critical thinking score for a sample of $n = 87$ students exposed to the new techniques was $M = 55$ with $ss = 6013$. Do the results suggest that the training program had a significant effect? Use a two-tailed test and $\alpha = 0.5$

Step 1:. Formulate hypothesis

$H_0 : \mu = 53$

$H_1 : \mu \neq 53$

The null hypothesis

Step 2: Indicate the alpha level and determine critical values

$\alpha = 0.05$

$df = 86$

$t_{crit} = \pm 2.000$

Step 3:.

calculate relevant statistics

$$s = \sqrt{\frac{SS}{n-1}}$$

$$= \sqrt{\frac{6013}{87-1}}$$

$$s = 8.36$$

Before we can determine our obtained t-value, we first need to calculate standard error using estimated standard deviation above:

$$s_M = \frac{s}{\sqrt{n}} = \frac{8.36}{\sqrt{87}} = 0.90$$

Finally, We can calculate the t-statistic:

$$t_{obt} = \frac{M - \mu}{s_M}$$

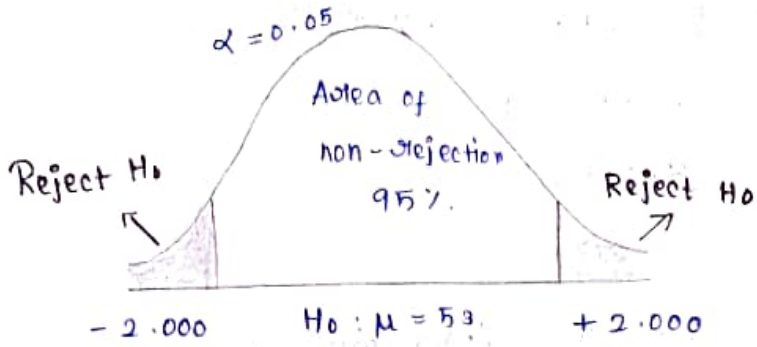
$$= \frac{55 - 53}{0.90}$$

$$= +2.22$$

Step 4:.

Make a decision and report the results

The t-distribution for this example is below:



The t_{crit} values are ± 2.000. To reject H_0, we need a sample mean with an obtained t-value beyond 2.000 in either direction. Our obtained t-value was +2.22 which falls in the direction of the region. Thus we reject the null-hypothesis

Test	Hypothesis	t_{crt}
Two-tailed	H_0 : μ = Value specified in pblm H_1 : μ = Value specified in pblm	$t_{crt} = \pm$ Value in chart
one-tailed, left	H_0 : $\mu \geq$ Value spec in pblm H_1 : $\mu <$ Value spec in pblm	$t_{crt} = -$ Value in chart
one-tailed, right	H_0 : $\mu \leq$ Value spec H_a : $\mu >$ Value spec	$t_{crt} = +$ Value in chart

Step 1: Formulate the hypothesis

$H_0 : \mu \leq 9$

$H_1 : \mu > 9$

Step 2: Indicating alpha level and determine the critical values

$\alpha = 0.05$

$df = n - 1 = 25 - 1$

$df = 24$

$t_{crit} = +1.711$

Step 3:

$$S_M = \frac{s}{\sqrt{n}} = \frac{6}{\sqrt{25}}$$

$$= \frac{6}{5} = 1.2$$

$$t_{obt} = \frac{11 - 9}{1.2}$$

$$= 1.66666$$

$$t_{obt} = +1.67$$

Step 4:

The study did not show that extroverts have significantly more friends than national average. Fail to reject H_0, $t(24) = +1.67$, $P > 0.05$

Research Question:

The mean scores

Step 1 :

Formulate hypothesis

Lower scores reflect better performance in golf

$H_0 : \mu \geq 83$

$H_1 : \mu < 83$

Step 2 : Indicating alpha level and determine the critical values.

$\alpha = 0.05$

$df = n - 1$

$df = 24$

$t_{crit} = -1.711$

Step 3 :

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{788}{24}}$$

$$s = 5.73$$

$$s_M = \frac{s}{\sqrt{n}} = 1.15$$

$$t_{obt} = \frac{81 - 83}{1.15}$$

$$t_{obt} = -1.74$$

Step 4 :

The flags has significantly improved golf scores.

Reject H_0, $t_{crit} > t_{obt}$, $P > 0.05$

$t(24) = -1.74$

Two - Sample t-test:

Standard Error of Difference :-

$$S_{M_1 - M_2} = \sqrt{\left(\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The standard deviation of sampling distribution of difference between means is called the standard error of difference between means.

Formula for T-test, Independent sample Designs:

$$t_{obt} = \frac{M_1 - M_2}{S_{M_1 - M_2}}$$

Degrees of freedom :-

$$df = n_1 + n_2 - 2$$

political

Step 1 : Formulate hypothesis

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

step 2 : Indicate the alpha level and determine critical values

$\alpha = 0.05$.

$df = n_1 + n_2 - 2$

$= 11 + 10 - 2$

$= 19$

$t_{crit} = \pm 2.093$

Step 3 : Calculate relevant statistics

X_1	X_1^2	X_2	X_2^2
59	3481	36	1296
48	2308	42	1764
45	2025	50	2500
39	1521	37	1369
52	2704	51	2601
56	3136	32	1024
50	2500	47	2209
41	1681	38	1444
46	2116	40	1600
45	2025	31	961
48	2304		
$\Sigma X_1 = 529$	$\Sigma X_1^2 = 25,797$	$\Sigma X_2 = 404$	$\Sigma X_2^2 = 16,768$

$n_1 = 11$ $\quad n_2 = 10$

$M_1 = 48.09$ $\quad M_2 = 40.40$

Calculate the sum of squares for each sample

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$SS_1 = 25,797 - \frac{(529)^2}{11}$

$= 356.91$

$SS_2 = 16,768 - \frac{(404)^2}{10}$

$= 446.40$

Calculate the standard error of difference.

Calculate the t-statistic.

$$S_{M_1 - M_2} = \sqrt{\left(\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\right)\left(1/n_1 + 1/n_2\right)}$$

$$S_{M_1 - M_2} = \sqrt{\left(\frac{356.91 + 446.4}{11 + 10 - 2}\right)\left(\frac{1}{11} + \frac{1}{10}\right)}$$

$$= \sqrt{(42.28)(.19)}$$

$$= \pm 2.83$$

$$t_{obt} = \frac{(M_1 - M_2)}{S_{M_1 - M_2}} = \frac{(48.09 - 40.40)}{2.83}$$

$$t_{obt} = \pm 2.72$$

Step 4:

Make decision and report the results

Students who read the newspaper scored significantly higher on knowledge of current events than students who watched the evening news. Reject H_0, $t(19) = \pm 2.72$, $P < 0.05$.

Research Ques:

An Industrial.

Step 1 : Formulate hypothesis

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

step 2 : Indicate alpha level and determine critical values

$\alpha = 0.05$

$df = n_1 + n_2 - 2$

$= 7 + 6 - 2$

$df = 11$

$t_{crit} = 2.201$

step 3 :

X_1	X_1^2	X_2	X_2^2
8	64	7	49
10	100	6	36
9	81	8	64
7	49	10	100
8	64	7	49
11	121	9	81
12	144		
$\Sigma X_1 = 65$	$\Sigma X_1^2 = 623$	$\Sigma X_2 = 47$	$\Sigma X_2^2 = 379$

$n_1 = 7$

$n_2 = 6$

$M_1 = \frac{65}{7} = 9.285$

$M_2 = 7.8333$

calculate sum of squares

$SS_1 = 623 - \frac{(65)^2}{7}$

$SS_1 = 19.428$

$SS_2 = 379 - \frac{(47)^2}{6}$

$SS_2 = 10.8333$

$$S_{M_1 - M_2} = \sqrt{\left(\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\right)\left(1/n_1 + 1/n_2\right)}$$

$$S_{M_1 - M_2} = \sqrt{\left(\frac{19.428 + 10.833}{7 + 6 - 2}\right)\left(\frac{1}{7} + 1/6\right)}$$

$$= \sqrt{2.751 \times 0.309}$$

$$= 0.9219$$

$$t_{obt} = \frac{(M_1 - M_2)}{S_{M_1 - M_2}} = \frac{9.285 - 7.8333}{0.9219}$$

$$\pm t_{obt} = 1.57446$$

Step 4:

Make decision and report the results

Listening to smooth jazz music did not significantly affect widget productivity. Failed to reject H_0, $\pm t(11) = +1.59$, $P < 0.05$

Two Sample t-Test related Sample Designs:

Standard Deviation of the difference:

$$S_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{n-1}}$$

$$M_D = \frac{\Sigma D}{n}$$

M_D → Mean difference

Standard error:

$$S_{M_D} = \frac{S_D}{\sqrt{n}}$$

$$t_{obt} = \frac{M_D}{S_{M_D}}$$

Sample Research Question:

A dietician

Step 1:

Formulate the hypothesis

$H_0 : \mu_D = 0$

$H_1 : \mu_D \neq 0$

Step 2:

Indicate the alpha level and determine the critical values.

$\alpha = 0.05$

$df = n - 1 = 7$

$t_{crit} = \pm 2.365$

Step 3:

Calculate relevant statistics

Subject	Weight before	Weight after	D	D^2
1	156	142	-14	196
2	192	173	-19	361
3	138	140	2	4
4	167	151	-16	256
5	110	109	-1	1
6	159	151	-8	64
7	171	154	-17	289
8	129	133	4	16
			$\Sigma D = -69$	$\Sigma D^2 = 1187$

$$M_D = \frac{\Sigma D}{n} = \frac{-69}{8} = -8.63$$

$$s_{M_D} = \frac{s_D}{\sqrt{n}} = \frac{9.20}{\sqrt{8}} = 3.25$$

$$s_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{n-1}} = \sqrt{\frac{1187 - \frac{(-69)^2}{8}}{8-1}} = 9.$$

$$t_{obt} = \frac{M_D}{s_{M_D}} = \frac{-8.63}{3.25} = -2.66$$

Step 4:

Make a decision and report the results subjects on the "no nightime snack" diet showed a significant weight loss. Reject H_0, $t(7) = -2.66$, $P < 0.05$.

Research Ques:

A psychology Professor

Step 1:

Formulate hypothesis

$H_0: \mu_D \geq 0$

$H_1: \mu_D \not\geq 0$

Step 2:

Indicate alpha level and determine the critical values

$\alpha = 0.05$

$df = n - 1 = 18 - 1 = 17$

$t_{crt} = -1.740$

Step 3:

$$S_D = \sqrt{\frac{SSD}{n-1}} = \sqrt{\frac{286}{17}} = 4.1016$$

$$S_{MD} = \frac{SD}{\sqrt{n}} = \frac{4.1016}{\sqrt{18}} = 0.9667$$

$$t_{obt} = \frac{MD}{S_{MD}} = \frac{-3}{0.9667} = -3.10334$$

Unit – 3

One - Way Analysis of Variance

ANOVA is used to abbreviate analysis of Variance

Variance in ANOVA:

⇒ Total Variance is separated into two kinds. Within – treatment Variance and between – treatments Variance.

⇒ Within – treatment Variance refers to be the Variability within a particular sample

⇒ Between treatments Variance refers to the variability between the treatment groups.

⇒ Between - treatments Variance = Individual differences + Experimental error + Treatment

⇒ Within – treatment Variance = Individual differences + Experimental error.

Sum of Squares:

$$SS_{wi} = \Sigma \left[\Sigma X^2_t - \frac{(\Sigma x_t)^2}{n_t} \right]$$

$$SS_{bet} = \Sigma \left[\frac{(\Sigma x_t)^2}{n_t} \right] - \frac{(\Sigma x_{tot})^2}{N}$$

$$SS_{total} = \Sigma X^2_{tot} - \frac{(\Sigma x_{tot})^2}{N}$$

Degrees of freedom:

$$df_{wi} = N - K$$

$$df_{bet} = K - 1$$

$$df_{total} = N - 1$$

$$df_{wi} + df_{bet} = df_{total}$$

Mean Square:

$$MS_{bet} = \frac{SS_{bet}}{df_{bet}}$$

$$MS_{wi} = \frac{SS_{wi}}{df_{wi}}$$

F - statistics:

$$F_{obt} = \frac{MS_{bet}}{MS_{wi}}$$

Research Question:

A counselling psychologist

Step 1: Formulate hypothesis

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

H_1 : Some μ's are not equal.

Step 2: Indicate alpha level and determine critical values.

$N = 18, \quad K = 3$

$df_{wi} = N - K = 18 - 3 = 15$

$df_{bet} = K - 1 = 3 - 1 = 2$

$df_{total} = N - 1 = 17$

$t_{crit} = 3.68$

step 3:

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
2	4	3	9	8	64
4	16	11	16	7	49
5	25	2	4	10	100
3	9	4	16	6	36
4	16	3	9	8	64
6	36	2	4	9	81

$\Sigma X_1 = 24$ $\Sigma X_1^2 = 106$ $\Sigma X_2 = 18$ $\Sigma X_2^2 = 58$ $\Sigma X_3 = 48$ $\Sigma X_3^2 = 394$

$$SS_{wi} = \Sigma \left[\Sigma X_t^2 - \frac{(\Sigma X_t)^2}{n_t} \right]$$

$$= \left[106 - \frac{(24)^2}{6} \right] + \left[58 - \frac{(18)^2}{6} \right] + \left[394 - \frac{(48)^2}{6} \right]$$

$$= [10 + 4 + 10]$$

$$SS_{wi} = 24$$

$$SS_{bet} = \Sigma \left[\frac{(\Sigma X_t)^2}{n_t} \right] - \frac{(\Sigma X_{tot})^2}{N} \Rightarrow \frac{(90)^2}{18}$$

$$= \left[\frac{(24)^2}{6} + \frac{(18)^2}{6} + \frac{(48)^2}{6} \right] - 450$$

$$= [96 + 54 + 384] - 450$$

$$SS_{bet} = 84$$

$$SS_{total} = \Sigma X^2_{tot} - \frac{(\Sigma X_{tot})^2}{N}$$

$$= 558 - 450$$

$$SS_{tot} = 108$$

$$MS_{bet} = \frac{SS_{bet}}{df_{bet}} = \frac{84}{2} = 42$$

$$MS_{wi} = \frac{SS_{wi}}{df_{wi}} = \frac{24}{15} = 1.6$$

$$F = \frac{MS_{bet}}{MS_{wi}} = 2.625$$

$$F_{obt} = 2.625$$

Summary Data:

Questions:

Three different techniques meditation

step 1: Formulate the hypothesis

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

H_1: Some μ's are not equal

step 2: Indicate alpha level and determine critical values

$N = 15, \quad K = 3$

$df\ wi = N - K = 15 - 3 = 12$

$df\ bet = K - 1 = 3 - 1 = 2$

$df\ tot = N - 1 = 15 - 1 = 14$

$t\,crit = 3.88$

step 3 :

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
10	100	6	36	5	25
12	144	8	64	9	81
9	81	3	9	12	144
15	225	0	0	8	64
13	169	2	4	4	16
$\Sigma X_1 = 59$	$\Sigma X_1^2 = 719$	$\Sigma X_2 = 19$	$\Sigma X_2^2 = 113$	$\Sigma X_3 = 38$	$\Sigma X_3^2 = 330$

$$SS_{wi} = \sum \left[\sum X^2_t - \frac{(\sum X_t)^2}{n_t} \right]$$

$$= \left[719 - \frac{(59)^2}{5} \right] + \left[113 - \frac{(19)^2}{5} \right]$$

$$+ \left[330 - \frac{(38)^2}{5} \right]$$

$$= 22.8 + 240.8 + 41.2$$

$$SS_{wi} = 104.8$$

$$SS_{bet} = \sum \left[\frac{(\sum X_t)^2}{n_t} \right] - \frac{(\sum X_{tot})^2}{N}$$

$$= \frac{(59)^2}{5} + \frac{(19)^2}{5} + \frac{(38)^2}{5} - \frac{(116)^2}{15}$$

$$= 696.2 + 72.2 + 288.8 - 897.066$$

$$SS_{bet} = 160.134$$

$$SS_{tot} = \sum X^2_{tot} - \frac{(\sum X_{tot})^2}{N}$$

$$= 1162 - \frac{(116)^2}{15}$$

$$= 1162 - 897.066$$

$$SS_{tot} = 264.934$$

$$Ms\ bet = \frac{SS\ bet}{df\ bet} = \frac{160.134}{2} = 80.067$$

$$Ms\ wi = \frac{SS\ wi}{df\ wi} = \frac{104.8}{12} = 8.733$$

$$F = \frac{Ms\ bet}{Ms\ wi} = \frac{80.067}{8.733} = 9.1683$$

$$Fobt = 9.1683$$

Step 4:

Summary Data

source	SS	df	MS	F	P
Medication	104.8	02	80.067	9.168	<0.05
Exercise	160.134	12	8.733		
Diet	264.934	14			

Step 4: Conclusion:

There is significant difference in mean reduction of blood pressure among three techniques.

Reject Ho. $t(15) = 9.1683$

Chi - Square test

THE CHI-SQUARE STATISTIC:

The chi-square statistic measures the amount of discrepancy between observed frequencies and the frequencies that would be expected due to chance, or random sampling error. The formula for chi-square is

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

Where: f_o = observed frequencies
f_e = expected frequencies

⇒ Observed frequencies are actual frequencies from our sample that fall into each category

⇒ Expected frequencies are the frequency values that would be expected if the null hypothesis is true.

⇒ If we find no differences between our observed and expected frequencies, then our obtained χ^2 value will equal 0.

⇒ But if our observed frequencies differ from those that would be expected, then χ^2 will be greater than 0.

(1) GOODNESS OF FIT FOR KNOWN PROPORTIONS:

Sample Research Question:

A new professor at a midsize college wanted to see if her grade distribution after her first year of teaching was comparable to the overall grade distribution, which has the following percentages: A - 10%, B - 22%, C - 40%, D - 21% and F - 7%. The distribution of the new professor's grade for 323 students at the end of

her first year as follows: 38 students recieved As, 78 recieved Bs, 139 recieved Cs, 55 recieved Ds and 13 recieved Fs. Does the new professor's grade distribution fit the overall college's distribution? Test using $\alpha = 0.05$.

chi-square is Ho-null hypothesis always fits
H1 - not fit

Sol:

Step 1: Formulate the hypothesis

Ho: The distribution of grades for the new professor fits overall grade distribution of the college.

H1: The distribution of grades for the new professor does not fit the overall grade distribution of the college.

step 2: Indicate alpha level and determine the critical values.

$\alpha = 0.05$

$df = K - 1$

$= 5 - 1 = 4$

$df = 4$

$\chi^2_{crit} = 9.488$

Step 3: calculate Relevant statistics

fe = (Known proportion) (n)

$A = 0.10 \times 323 = 32.30$

$B = 0.22 \times 323 = 71.06$

$C = 0.40 \times 323 = 129.20$

$D = 0.21 \times 323 = 67.83$

$F = 0.07 \times 323 = 22.61$

$$\chi^2_{obt} = \sum \frac{(fo - fe)^2}{fe}$$

$$= \left(\frac{(38 - 32.3)^2}{32.3}\right) + \frac{(18 - 71.06)^2}{71.06} + \frac{(139 - 129.2)^2}{129.2} + \frac{(55 - 67.83)^2}{67.83} + \left(\frac{(13 - 22.61)^2}{22.61}\right)$$

$= 1.01 + 0.68 + 0.74 + 2.43 + 4.08$

$= 8.94$

Step 4 :

(ii) TEST OF INDEPENDENCE :

* The other type of problem, where chi-square has its greatest usefulness is in testing for independence of variables.

* This test examines the frequencies for two variables at different levels in order to determine whether one variable is independent of the other or if the variables are related.

Sample Research Question :

A professor at a veterinary school of medicine is curious about whether or not a relationship exists between gender and type of pets owned in childhood. She randomly asks a sample of n = 260 students (132 female and 128 male) about their pet ownership, the frequency of which is recorded in the table below. Is there a relationship between gender and type of pet ownership? Test at $\alpha = 0.$

	Type of pet					
Gender	Dogs	cats	Birds	Reptiles	Rodents	Row total
Female	58	36	22	4	12	132
Male	62	22	14	10	20	128
column total	120	58	36	14	32	n = 260

Expected frequency :-

$$fe = \frac{(fc)(fr)}{n}$$

f_c = column total

f_r = row total

n = sample size

$$fe = \frac{(fc)(fr)}{n} = \frac{(120)(132)}{260} = 60.92$$

Type of pets

Gender	Dogs	Cats	Birds	Reptiles	Rodents	Row
Female	58 (60.92)	36 (29.45)	22 (18.28)	4 (7.11)	12 (16.24)	132
Male	62 (59.08)	22 (28.55)	14 (17.72)	10 (6.89)	20 (15.75)	128
column	120	58	36	14	32	n = 260

Step 1: Formulate hypothesis

H_0 : Gender and pet ownership are independent and Unrelated.

H_1 : There is a relationship between gender and pet ownership.

Step 2: Indicate alpha level and determine the critical values.

$$df = (R-1)(C-1)$$

R = no. of rows

C = no. of columns

$$\alpha = 0.05$$

$$df = (2-1)(5-1) = 4$$

$$\chi^2_{crit} = 9.488$$

Step 3: Calculate Relevant Statistics

$$X^2_{obt} = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(58 - 60.92)^2}{60.92} + \frac{(36 - 29.45)^2}{29.45} + \frac{(22 - 18.28)^2}{18.28} +$$

$$+ \frac{(20 - 15.75)^2}{15.75}$$

$$= 0.14 + 1.46 + 0.76 + 1.36 + 1.11 + 0.14 + 1.50 + 0.78 + 1.40 + 1.15$$

$$= 9.80$$

Step 4: Make a decision and Report the results

There is a relationship between gender and type of pet owned. Reject H_0, χ^2 (4df, n = 260) 9.80, $p < 0.05$

A Researcher is Interested

Penmanship Quality

Handedness	Low	Medium	High	Row total
left-handed	8 (7.31)	29 (28.38)	6 (7.31)	43
Right-handed	9 (9.69)	37 (37.62)	11 (9.69)	57
column total	17	66	17	

Formulate hypothesis

Step 1:

H_0 : Handedness and penmanship quality are Unrelated

H_1 : there is a relationship between handedness and penmanship quality.

Step 2: Indicate alpha level and determine critical values

$$df = (R-1)(C-1)$$
$$= (2-1)(3-1)$$
$$df = 2$$
$$\alpha = 0.05$$
$$\chi^2 crit = 5.991$$

Step 3: calculate Relevant statistics

$$X^2 obt = \sum \frac{(f_o - fe)^2}{fe}$$

$$= \frac{(8-7.31)^2}{7.31} + \frac{(29-28.38)^2}{28.38} + \frac{(6-7.31)^2}{7.31} +$$

$$\frac{(9-9.69)^2}{9.69} + \frac{(37-37.62)^2}{37.62} + \frac{(11-9.69)^2}{9.69}$$

$= 0.065 + 0.01 + 0.23 + 0.04 + 0.01 + 0.17$

$= 0.52$

Non-Parametric Test

Mann Whitney U Test

Two Sample T-test:

Procedure:

1) Rank data (1 being the lowest)
2) Find the sum of the ranks for each group
3) Find U for both groups

$$U_1 = N_1 N_2 + \frac{N_1 (N_1+1)}{2} - T_1 \quad \text{(or)}$$

$$U_2 = N_1 N_2 + \frac{N_2 (N_2+1)}{2} - T_2$$

4) Using a table for Mann Whitney U, find the critical value
5) Compare the critical value with the computed value (smaller value between U_1 and U_2)

Mann Whitney U Test

FLUSHEM	3	4	2	6	2	5
KLEEROUT	9	7	5	10	6	8

							Total
FLUSHEM	3	4	2	6	2	5	
RANK	3	4	1.5	7.5	1.5	5.5	23
KLEEROUT	9	7	5	10	6	8	
RANK	11	9	5.5	12	7.5	10	55

PROCEDURE:

Find U for both groups

$$U_1 = (6)(6) + \frac{6(7)}{2} - 23$$

$$U_2 = (6)(6) + \frac{6(7)}{2} - 55$$

Thus,

$U_1 = 34$ and $U_2 = 2$

Hence the computed value $U = 2$

Find critical value

$n_1 = 6$, $n_2 = 6$

The critical value is 5

* => critical value: 5

computed value: 2

critical value - opposite to other test

since the computed value is less than the critical value we reject the null hypothesis.

Therefore, there is a significant difference in the rated effectiveness of two laxatives.

KRUSKAL - WALLIS TEST PROCEDURE:

1) Rank all the scores (1 being the lowest)

2) Find T_c (the total of all the ranks of each group)

3) Solve for $H = \frac{12}{N(N+1)} \sum \frac{T_c^2}{n_c} - 3(N+1)$

4) Using the Appropriate Table, find the critical value

5) Compare the critical value with the computed value

Note: Reject null if computed value is greater than or equal to the critical value.

a	Rank	b	Rank	c	Rank	d	Rank
68		78		94		54	
63		69		82		51	
58		58		73			
51		51		67			
41		53		66			
				61			

Unit - IV

Simple Linear Regression

The two main objectives:

(i) Establish if there is a relationship between two variables. More specifically establish, if there is a statistically significant relationship between the two.

Ex: Income and spending, Wage & gender, student height and exam score.

(ii) Forecasting new observations, We can use what we know about the relationship to forecast. Unobserved values, there are two variables

* Dependent Variable - This is the variable whose values we want to explain our forecast, denoted by y.

* Independent Variable - This is the variable that explains the other one, we denote it by x

Linear Equation:

$$y = mx + c$$

Simple Linear Regression:

$$Y = \beta_0 + \beta_1 x$$

$\beta_0 \rightarrow$ Intercept

$\beta_1 \rightarrow$ slope

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (\text{epsilon})$$

$$\beta_1 = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

$\bar{x} \rightarrow$ Mean of x

$\bar{y} \rightarrow$ Mean of y

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
15	20			$\Sigma (x - \bar{x})^2$	$\Sigma (x - \bar{x})(y - \bar{y})$
$\bar{x} = 3$	$\bar{y} = 4$			10	6

$$B_1 = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

$$B_1 = \frac{6}{10}$$

$$B_1 = 0.6$$

$$y = B_0 + B_1 x$$

$$4 = B_0 + (0.6)(3)$$

$$B_0 = 4 - 0.6(3)$$

$$= 4 - 1.8$$

$$B_0 = 2.2$$

$$\boxed{y = 2.2 + 0.6x}$$

$\hat{y} = 2.2 + 0.6(1)$

$= 2.8$

Standard Error:

$$\text{Standard error} = \sqrt{\frac{\Sigma (y^1 - y)^2}{(n-2)}}$$

y^1	$y^1 - y$	$(y^1 - y)^2$
2.8	0.8	0.64
3.4	-0.6	0.36
4	-1	1
4.6	0.6	0.36
5.2	0.2	0.04
		$\Sigma (y^1 - y)^2 = 2.4$

Standard error $= \sqrt{\frac{\Sigma (y^1 - y)^2}{n-2}}$

$= \sqrt{\frac{2.4}{3}}$

$= 0.89$

It should be less than 1

2) The values of x and corresponding values of y are shown below.

x	0	1	2	3	4
y	2	3	5	4	6

(i) Find the least square regression line $y = ax + b$

(ii) Estimate the value of y, when $x = 10$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$	y^1	$y^1 - y$	$(y^1 - y$
0	2	-2	-2	4	4			
1	3	-1	-1	1	1			
2	5	0	1	0	0			
3	4	1	0	1	0			
4	6	2	2	4	4			
				$\Sigma (x - \bar{x})^2$	$\Sigma (x - \bar{x})(y - \bar{y})$			
				10	9			

Σx: 10 Σy: 20

$\bar{x}$: 2 $\bar{y}$: 4

$$B_1 = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

$$B_1 = \frac{9}{10}$$

$$B_1 = 0.9$$

$$y = B_0 + B_1 x$$

$$4 = B_0 + (0.9) \times 2$$

$$4 = B_0 + 1.8$$

$$B_0 = 2.2$$

$$\boxed{y = 2.2 + 0.9x}$$

(ii)

$$y = 2.2 + 0.9 (10)$$

$$\boxed{y = 11.2}$$