Module-2
Data Analysis-1

Prepared By
Arya Kumar
Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Comparison Chart

Basis for Comparison	Parametric Test	Nonparametric Test
Meaning	A statistical test, in which the population parameter is known as parametric test.	A statistical test, in which the population parameter is not known as parametric test.
Basis of test statistic	Distribution	Arbitrary
Measurement level	Interval or ratio	Nominal or ordinal
Measure of central tendency	Mean	Median
Information about population	Completely known	Unavailable
Applicability	Variables	Variables and Attributes
Correlation test	Pearson	Spearman

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

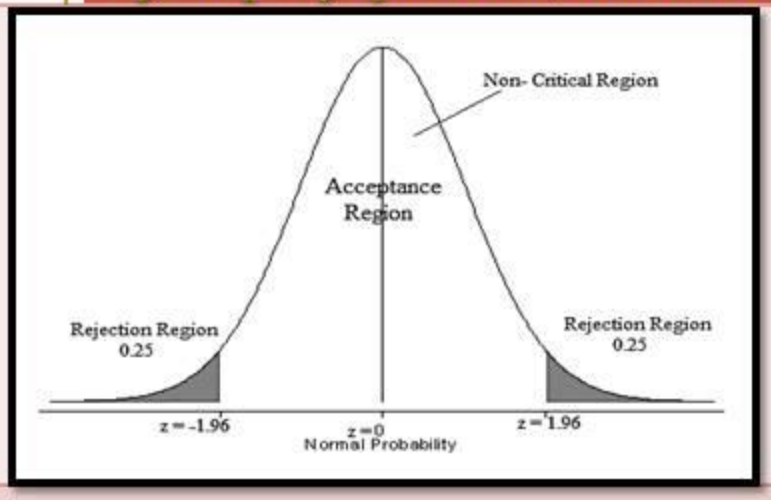
Process of Hypothesis

Hypothesis Formulation	• Null Hypothesis (H_0) • Alternative Hypothesis (H_1) or (H_a)
Hypothesis testing	• one tailed hypothesis • two tailed hypothesis
Level of Significance	1%, 5% or 10%.
Test Statistics or Deciding the distribution	The statistics test can be a mean difference, proportion difference, mean score, proportion, z-score, t-score, chi-score, f-score and etc. if: • The size of sample (n) is more than 30 i.e. (n>30) then z-score is used. • The size of sample (n) is less than 30 i.e. (n<30) then t-score is used.
P-Value	value of probability
Result Interpretation	• CV<TV, H_0 is accepted • CV>TV, H_0 is rejected calculated value (CV) and tabulated value (TV)

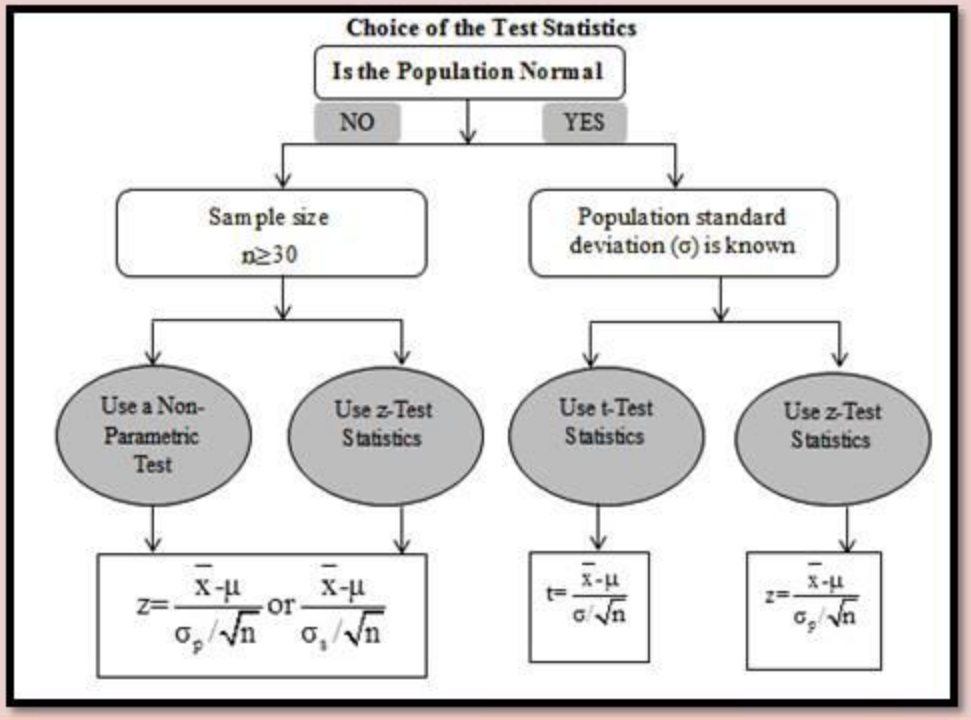
Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Acceptance and Rejection Region



Test Statistics



Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

HYPOTHESIS TESTING WITH LARGE SAMPLE (n > 30)

	Single Population	Two Population
If standard deviation σ of the population is known, population is normal & infinite, sample size may be large or small	$z=\frac{\bar{x}-\mu}{\sigma_p/\sqrt{n}}$	$z=\frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{\sigma_{s_1}^2}{n_1}+\frac{\sigma_{s_2}^2}{n_2}}}$ $\sigma_{s_1}=\sqrt{\frac{\Sigma(x-\bar{x}_1)^2}{n-1}}$ and $\sigma_{s_2}=\sqrt{\frac{\Sigma(x-\bar{x}_2)^2}{n-1}}$
If standard deviation σ of the population is not known, population is normal & infinite, sample size may be large or small	$z=\frac{\bar{x}-\mu}{\sigma_s/\sqrt{n}}$	$z=\frac{(\bar{x}_1-\bar{x}_2)}{\sigma_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$
If standard deviation σ of the population is known, population is normal & finite sample size is small Prepared By Arya Kumar	$z=\frac{\bar{x}-\mu}{\sigma_p/\sqrt{n}}\times\frac{1}{\sqrt{(N-n)/(N-1)}}$ $\sigma_s=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$	$z=\frac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\frac{\sigma_{s_1}^2}{n_1}+\frac{\sigma_{s_2}^2}{n_2}}}$ $\sigma_{s_1}=\sqrt{\frac{\Sigma(x-\bar{x}_1)^2}{n-1}}$ and $\sigma_{s_2}=\sqrt{\frac{\Sigma(x-\bar{x}_2)^2}{n-1}}$

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

HYPOTHESIS TESTING WITH SMALL SAMPLE (n=30)

	Single Population	Two Population
If standard deviation σ of the population is not known, population is normal & infinite, sample size may be small	$t=\frac{(\bar{x}-\mu)}{\sigma_s/\sqrt{n}}$ with df =(n − 1)	
If standard deviation σ of the population is not known, population is normal & finite sample size is small	$t=\frac{(\bar{x}-\mu)}{\sigma_s/\sqrt{n}} \times \frac{1}{\sqrt{(N-n)/(N-1)}}$ $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$	$t=\frac{\bar{x}_1-\bar{x}_2}{\frac{\Sigma(x_1-\bar{x}_1)^2+\Sigma(x_2-\bar{x}_2)^2}{n_1+n_2-2}} \times \frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ with df = n_1+n_2-1

Prepared By

Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Module-2
ANOVA

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Meaning of ANOVA

- ANOVA provides a statistical test of whether or not.
- The means of several groups are all equal, and therefore generalize t-test to more than two groups.
- It is helpful for testing as it considers the two-sample t-test. While by multiplying two sample t-tests there is chance of committing a type I error.
- For this reason, ANOVAS are useful in comparing three or more means.

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

ANOVA will test:

- The null hypothesis **H_0: $\mu_1, \mu_2 \ldots\ldots \mu_k$** against
- The alternative hypothesis **H_1: there exists difference in mean at least from one variable**

ANOVA Approach

The primary activity is to segregate the total variation in the sample data into two components.

1. The amount of variation **among/between the sample means** or the variation attributable to the difference among samples means. The difference is denoted by **SSC or SSTR**.
2. The amount of variation **within the sample observations**. The difference in the values of various elements in a sample due to chance is called an estimate and denoted as **SSE**.

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

The **sample data** observed can be classified in to two forms i.e.

- **one factor (criterion)**
- **two factors (criterion)**

The calculations for **total variation and its components may** be carried out in each of the two-types of classifications by

A. **direct method,**
B. **short-cut method, and**
C. **Coding method.**

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

ANOVA TECHNIQUE

One-way ANOVA

In case of one way ANOVA, only single factor is considered and measured it is done for the reason that several possible types of sample can occur within that factor

A. Direct Method

Process

Step-1: Mention null and alternative hypothesis:

Null hypothesis H_0: $\mu_{1=}\ \mu_{2=}\ \ldots\ldots\ldots = \mu_k$

Alternative Hypothesis H_1: $\mu_{1=}\ \mu_{2=}\ldots\ldots\ldots\ \text{“}\mu_k$

Step-2: Calculate variation between sample mean:

a. Calculate the mean of each sample *i.e.*

$\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots\ldots\ldots \bar{x}_k$ of all k samples

***b.* Calculate the mean of the sample means**

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \ldots\ldots\ldots\bar{x}_k}{No.ofsample(k)} = \frac{T}{n}$$

Where, T = Grand total of all observations
n = Number of observations in k samples

c . Calculate the sum of squares for variance between the samples (or SS between)

$$SS\ between = \sum_{j=1}^{r} n_j (\bar{x}_j - \bar{\bar{x}})^2$$

$$SS\ between = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \ldots\ldots\ldots n_k(\bar{x}_k - \bar{\bar{x}})^2$$

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step 3: Calculate variation within sample:

a. Calculate the mean of each sample i.e.,

$\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots\ldots\ldots \bar{x}_k$ *of all k' samples.*

b. Calculate the sum of squares for variance within the samples (or SS within):.

SS within= $\sum_{i=1}^{k} \sum_{j=1}^{n} \left(\bar{x}_{ij} - \bar{\bar{x}} \right)^2$

SS within = $\sum_{i} \left(x_{1i} - \bar{x}_1 \right)^2 + \sum_{i} \left(x_{2i} - \bar{x}_2 \right)^2 + \ldots\ldots\ldots + \sum_{i} \left(x_{ki} - \bar{x}_k \right)^2$

i = 1,2,3,....

This sum is also called the sum of squares for error i.e., SSE = SS Total – SS between.

Step-4: Calculate the total variance:

Total variation is represented by the sum of squares total (SS total) and equal to the sum of the squared difference between each sample value from the grand mean

SS total = $\Sigma \left(x_{ij} - \bar{\bar{x}} \right)^2$

i = 1, 2,....and j= 1,2,...

This total should be equal to: SS for total variance = SS between + SS within and (n-1)=(k – 1)+ (n - k)

Step-5: Calculate average variation between and within samples and Mean Square:

When the sum of squares is divided by their associated degrees of freedom, we get the following variances of mean square:

$$MSbetween = \frac{SSbetween}{k-1}, MSwithin = \frac{SSwithin}{n-k}$$

and

$$MSE = \frac{SSE}{k-r}$$

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step-6: Calculate F-ratio:

$$Fratio = \frac{SSbetween/(k-1)}{SSwithin/(n-k)} = \frac{MSbetween}{MSwithin}$$

This ratio is useful to confirm whether the difference among several sample means is significant or is just a matter of sampling fluctuations.

For this let us look into the table; giving the value of F for given degrees of freedom at different levels of significance.

Step-7: Set up a ANOVA table: **ANOVA Table: One-way**

Sources of Variation	Sum of Square (SS)	Degree of Freedom (d.f.)	Mean Squares (MS)	F-value
Between Samples	SS between	k-1	MS between = SS between / k-1	F= MS between / MS within
Within Samples	SS within	n-k	MS between = SS between / n-k	
Total	SS total	n - 1		

Decision Rules:

- If $F_{cal} < F_{critical\ value}$, accept null hypothesis H_0.
- Otherwise Reject H_0

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

B. Short-Cut Method:

There is a shortcut method for calculating SS between and SS within (SSE)

a. Calculate the grand total of all observations in samples, T is given by: ..

$$T = \sum x_1 + \sum x_2 + ... + \sum x_k$$

b. Calculate the correction factor, CF is given by:

$$CF = \frac{T^2}{n} \quad ; \quad n = n_1 + n_2 + ... + n_k$$

c. Find the sum of the squares of all observations in samples from each of k samples and subtract CF from this sum to obtain the total sum of the squares of squares of deviation SS total:

$$SStotal = \left(\Sigma x_1^2 + \Sigma x_2^2 + ...\Sigma x_k^2\right) - CF = \Sigma_{i-1}^{k}\Sigma_{j=1}^{n} x_{ij}^2 - CF$$

$$\text{SS between} = \frac{\left(\Sigma x_j\right)^2}{n_j} - CF$$

and SS within = SS total - SS between

d. Prepare Anova table

e. Degree of freedom (n-k) & (k-1)

C. Coding Method:

It is method that depends on F- test Statistics. It is used in analysing of the ratio of variances without unit of measurement. Even any other constant value is either multiplied, subtracted or added to each observations in the sample data t values does not change. This adjustment reduces the magnitude of numerical values in the sample data and reduces computational time to calculate f value without any change.

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Illustration-1 (ANOVA One Way)

The following data table is the annual production data in metric ton per acre of one agricultural industry producing wheat floor.

The industry uses 3 types of seeds for producing what floor keeping $\alpha = 5\%$ test the hypothesis that there is no significant difference between the annual production of wheat cultivated through this 3 types of seeds.

Seeds / Production	A	B	C
a	6	5	5
b	7	5	4
c	3	3	3
d	8	7	4

Prepared By
Arya Kumar

Note: solving the equation with short-cut method

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Solution-

Step-1 – Establish the hypothesis

H0: there is no significant difference in annual production of wheat with respect to 3 types of seeds

H1:there is significant difference in annual production of wheat with respect to 3 types of seeds

Step-2- Add all observation and give a symbol T and correction error T^2

Seeds / Production	A	A^2	B	B^2	C	C^2
a	6	36	5	25	5	25
b	7	49	5	25	4	16
c	3	9	3	9	3	9
d	8	64	7	49	4	16
Total (T)	24	x	20	x	16	x
T^2	x	158	x	108	x	66

Total (T)=24+20+16=60 &

Correction Error T^2= 158+108+66= 332

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step-3- Allocate data

$n=12, \quad k=3, \quad T=60, \quad T^2=332$

Step-4 Calculate the correction factors (CF)= $\frac{(T)^2}{n}=\frac{(60)^2}{12}=\frac{3600}{12}=300$

Step-5 Calculate $SStotal=\left(\Sigma x_1^2+\Sigma x_2^2+\ldots.\Sigma x_k^2\right)-CF=\Sigma_{i-1}^{k}\Sigma_{j=1}^{n}x_{ij}^2-CF$

or Sstotal= sum of squared variance for total number of observation

$=T^2-CF=332-300=32$

Step-6 Calculate SS between = $\frac{\left(\Sigma x_j\right)^2}{n_j}-CF$

$\sum\frac{(Tj)^2}{nj}$ - CF= $\left[\frac{(24)^2}{4}+\frac{(20)^2}{4}+\frac{(16)^2}{4}\right]$ -300

=308-300=8

Step- 7 Calculate SS within= SS total- SS between= 32-8 =24

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step- 8 ANOVA TABLE

Sources of Variable	SS	df	MS= SS / df	f= MS between / MS within
SS Between	8	k-1=3-1=2	8/2=4	4/2.66 =1.50
SS Within	24	n-k=12-3=9	24/9=2.66	
SS Total	32			

Step- 9 Result

f calculated = 1.50 **f critical value= f(df)= f(2,9)= 4.26 (check f table)**

As we get fc<ft, so we accept null hypothesis i.e. there is no significant difference in annual production of wheat with respect to 3 types of seeds perhaps the differences is due to some other factors.

Prepared By
Arya Kumar

F-Distribution (p=0.01) Table

Table A.6 (continued) Critical Values of the F-Distribution

	$f_{0.01}(v_1, v_2)$								
v_2	1	2	3	4	5	6	7	8	9
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
27	7.68	5.49	4.60	4.11	3.78	3,56	3.39	3.26	3.15
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2,56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

F Values for $\alpha = 0.05$

	d_1								
d_2	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.3	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
inf	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Two-way ANOVA

As per on-way ANOVA the partition of the total variation in the sample data is done into two components:

- Variation among the samples due to different samples and
- Variation within the samples due to random error

Process

Step-1: Calculation the Correction Factor:

$$\text{Correction factor} = \frac{T^2}{n}$$

T= Total of the values of individual items in the sample.

Step-2: Calculate the sum of squares of deviations for total variance:

$$\text{SS total} = \sum x_{ij}^2 - \frac{T^2}{n}$$

Step-3: Calculate the sum of squares between the columns treatment:

SS between treatments

$$\Sigma \frac{T_j^2}{n_j} - \frac{T^2}{n}, \text{ j=column}$$

Step-4: Calculate the sum of squares between the rows treatment:

SS between rows treatment

$$\Sigma \frac{T_i^2}{n_i} - \frac{T^2}{n}, \text{ i=row}$$

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step-5: Calculate the sum of squares residual or error

SS residual or error = SS total – (SS between columns + SS between rows)

Step-6: Calculate the degree of freedom:

Degree of freedom between columns = C-1

Degree of freedom between rows = r - 1

Degree of freedom for residual or error = (c – 1) (r – 1)

Degree of freedom total variance = cr - 1

Where c=number of columns and r= number of rows

Step-7: Set up the ANOVA Table:

ANOVA Table: Two-way

Sources of Variation	Sum of Square (SS)	Degree of Freedom (d.f.)	Mean Squares (MS)	F-value
Between columns	SS between columns	c-1	$\frac{SS\ between\ columns}{c-1}$	$\frac{MS\ between\ columns}{MS\ residual\ or\ error}$
Between Rows	SS Between Rows	r-1	$\frac{SS\ between\ rows}{r-1}$	$\frac{MS\ between\ rows}{MS\ residual\ or\ error}$
Residual or error	SS total – (between columns Or error + SS between rows)	(c-1)(r-1)	$\frac{SS\ residuals\ or\ errors}{(c-1)(r-1)}$	
Total	SS total	cr - 1		

Decisions Rule:

- If $F_{cal} < F_{critical\ value}$, accept null hypothesis H_0.
- Otherwise Reject H_0.

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Illustration-2 (two way ANOVA- without repetition)

The following data table is the annual production data in metric ton per acre of one agricultural industry producing wheat floor.

The industry uses 3 types of seeds and 4 types of fertilizers for producing what floor keeping $\alpha = 5\%$ test the hypothesis that there is no significant difference between the annual production of wheat cultivated through this 3 types of seeds or through fertilizers .

Prepared By
Arya Kumar

		Seeds		
Fertilizers		A	B	C
	w	6	5	5
	x	7	5	4
	y	3	3	3
	z	8	7	4

Note: solving the equation with same problem, only fertilizer is considered in row wise

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Solution-

Step-1 – Establish the hypothesis (there will be 2 hypothesis)

1.H0: there is no significant difference in annual production of wheat with respect to 3 types of seeds

H1:there is significant difference in annual production of wheat with respect to 3 types of seeds

2.H0: there is no significant difference in annual production of wheat with respect to 3 types of seeds

H1:there is significant difference in annual production of wheat with respect to 3 types of seeds

Prepared By
Arya Kumar

Step-2- Add all observation and give a symbol T and correction error T^2

Seeds / Production	A	A^2	B	B^2	C	C^2	Total A+B+C
w	6	36	5	25	5	25	16
x	7	49	5	25	4	16	16
y	3	9	3	9	3	9	9
z	8	64	7	49	4	16	19
Total (T) (w+x+y+z)	24	x	20	x	16	x	60
T^2	x	158	x	108	x	66	

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step-3- Allocate data

n=12, k=3, T=60, T^2=332

Step-4 Calculate the correction factors (CF)= $\frac{(T)^2}{n} = \frac{(60)^2}{12} = \frac{3600}{12} = 300$

Step-5 Calculate $SStotal = \left(\Sigma x_1^2 + \Sigma x_2^2 + \ldots \Sigma x_k^2\right) - CF = \Sigma_{i-1}^{k} \Sigma_{j=1}^{n} x_{ij}^2 - CF$

or Sstotal= sum of squared variance for total number of observation

$= T^2 - CF = 332 - 300 = 32$

Step-6 (a) Calculate SS between column (seeds) =

$$\sum \frac{(Tj)^2}{nj} - CF = \left[\frac{(24)^2}{4} + \frac{(20)^2}{4} + \frac{(16)^2}{4}\right] - 300$$

$$\frac{\left(\Sigma x_j\right)^2}{n_j} - CF$$

$= 308 - 300 = 8$

Prepared By

Arya Kumar

Step-6 (b) Calculate SS between row (fertilizers)

$$= \left[\frac{(16)^2}{3} + \frac{(16)^2}{3} + \frac{(9)^2}{3} + \frac{(19)^2}{3}\right] - 300$$

$= 318 - 300 = 18$

Step- 7 Calculate SS residual = SS total- (SS between column+ SS between row)=

$= 32 - (8+18) = 6$

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step- 8 ANOVA TABLE

Sources of Variable	SS	df	MS= SS / df	f ratio f1 & f2
SS Between column	8	c-1=3-1=2	8/2=4	MS between column / MS Residual 4/1 =4
SS Between row	18	r-1=4-1=3	18/3=6	MS between row / MS Residual 6/1 =6
residual	32-8-18= 6	(c-1)(r-1)= 2x3=6	6/6=1	Prepared By
SS Total	32	(cxr)-1= (3x4)-1=11		Arya Kumar

Step- 9 Result

f1calculated column (seeds)= 4 **f1 critical value= f(df)= f(2,6)= 5.14**

f2 calculated row (fertilizers)= 6 **f2 critical value= f(df)= f(3,6)= 4.76**

As we get, f1c>f1t, so we reject null hypothesis and i.e. f2c<f2t, so we accept null hypothesis it means there is no significant difference in annual production of wheat with respect to 3 types of seeds but due to fertilizers.

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Illustration-3 (two way ANOVA- with repetition)

The following data table gives the information relating to 3 drugs testing, to judge the effectiveness in treating the COVD-19 for three different group of people

		Drugs		
Group of People		X	Y	Z
	A	14	10	11
		15	9	11
	B	12	7	10
		11	8	11
	C	10	11	8
		11	11	7

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Solution-

Step-1 – Establish the hypothesis (there will be 3 hypothesis)

1. **H0**: Drugs has no significant effect on treating COVID-19
2. **H0**: People has no significant effect
3. **H0**: Interact has no significant effect

Prepared By
Arya Kumar

Step-2- Add all observation and give a symbol T and correction error T^2

		Drugs						Total
Group of People		X	X^2	Y	Y^2	Z	Z^2	
	A	14	196	10	100	11*	121*	70
		15	225	9	81	11*	121*	
	B	12	144	7	49	10	100	59
		11	121	8	64	11	121	
	C	10	100	11*	121*	8	64	58
		11	121	11*	121*	7	49	
Total (T)		73	x	56	x	58	x	187
T^2		x	907	x	536	x	576	2019

*** Refers to repetition (do mention and write in exam)**

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step-3- Allocate data

n=18(count the numbers in the box) , k= 9(count the boxes), T=187, T^2=2019

Step-4 Calculate the correction factors (CF)= $\frac{(T)^2}{n} = \frac{(187)^2}{18} = \frac{34969}{18} = 1942.72$

Step-5 Calculate $SStotal = \left(\Sigma x_1^2 + \Sigma x_2^2 + \ldots \Sigma x_k^2\right) - CF = \Sigma_{i-1}^{k} \Sigma_{j=1}^{n} x_{ij}^2 - CF$

or Sstotal= sum of squared variance for total number of observation

$= T^2 - CF = 2019 - 1942.72 = 76.28$

Step-6 (a) Calculate SS between column = $\frac{\left(\Sigma x_j\right)^2}{n_j} - CF$

$$\sum \frac{(Tj)^2}{nj} - CF = \left[\frac{(73)^2}{6} + \frac{(56)^2}{6} + \frac{(58)^2}{6}\right] - 1942.72$$

$= 28.78$

Prepared By

Arya Kumar

Step-6 (b) Calculate SS between row

$$\sum \frac{(Ti)^2}{ni} - CF = \left[\frac{(70)^2}{6} + \frac{(59)^2}{6} + \frac{(58)^2}{6}\right] - 1942.72$$

$= 14.78$

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step- 7 Calculate SS within

$= (14-14.5)^2 + (15-14.5)^2 + (10-9.5)^2 + (9-9.5)^2 + (11-11)^2 + (12-11.5)^2 + (11-11.5)^2 + (7-7.5)^2 + (8-7.5)^2 + (10-10.5)^2 + (11-10.5)^2 + (10-10.5)^2 + (11-10.5)^2 + (11-11)^2 + (8-7.5)^2 + (7-7.5)^2$

=3.5

How to get mean value

In each box, there are two values add the values and divide it with 2

Ex- check 1st box i.e. (14+15)/2= 14.5

Now deduct mean i.e. 14.5 from two values i.e. 14 and 15

Similarly,

2nd box i.e. (10+9)/2= 9.5

Now deduct mean i.e. 9.5 from two values i.e. 10 and 9

And so on till last box

i.e. (8+7)/2= 7.5

Now deduct mean i.e. 7.5 from two values i.e. 8 and 7

Step- 8 Calculate SS interaction = SS total- (SS between column+ SS between row+ SS within)

=76.28-(28.78+14.78+3.5)= 29.2

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step- 9 ANOVA TABLE

Sources of Variable	SS	df	MS= SS / df	f ratio f1 & f2
SS Between column	28.78	c-1=3-1=2	14.38	MS between column / MS within =14.38/0.389= 36.1
SS Between row	14.78	r-1=3-1=2	17.39	MS between row / MS within =17.3/0.389= 19
SS Interaction	76.28-(28.78+14.78+3.5)= 29.22	?= 17-(2+2+9)= 4	7.30	MS between interaction / MS within =7.30/0.389= 18.76
SS within	3.5	n-k=18-9= 9	0.389	
SS Total	76.28	n-1=18-1= 17		

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step- 10 Result

f1 calculated column = 36.1

f1 critical value= f(df)= f(2,9)= 4.26

f2 calculated row = 19

f2 critical value= f(df)= f(2,9)= 4.26

f2 calculated interaction = 18.76

f2 critical value= f(df)= f(4,9)= 3.63

(df= check the previous table)

So it is clear that the calculated values are more than the table value so the null hypothesis is rejected.

It means, there exist some significance between drugs, group of people and interaction.

Prepared By
Arya Kumar

Prepared By

Arya Kumar

Faculty of MBA, CIME

aryantripathy@yahoo.com, 09853422575

Module-3 Non-Parametric Test

RUN- TEST

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Run test- Test of Randomness

Random means not occurring similarly, it means one cannot predict

- Run test means, whether a string of data is occurring randomly from a specific distribution
- Let us suppose,
  - there are n_1 observations of random variable x and
  - n_2 observations random variable y
- Suppose we combine two sets of observations into one large collection n_1+n_2 observations, and then arrange the observations in increasing order of magnitude.
- We may get a sequence like this:

(x)(y y)(x x x)(y)(x x)(y y)

Here, the run is 6

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Illustration-1

During Lockdown it is observed that people are roaming around, the observation states that both male and female are on the street. The police commissioner asked the police men about the frequency.

MMFMFFFMMMMFFFMMFFFMMMMMFFMMMFFFMFFFFFMMFFFFF

State from the frequency of male and female movements, that it maintain a random approach.

FORMULA FOR RUN TEST

$$Z = \frac{R - \mu_R}{\sigma_R}$$

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Find R i.e. RUN So R= 16	**MM F M FFF MMMM FFF MM FFF MMMMM FF MMM FFF M FFFFF MM FFFFF** check the underlines i.e. MM =1, F =2, M=3, FFF=4 ..... FFFFF=16
N=45	Check the total number of people on the street.
n1= 20	Check the total number of Male
n2= 25	Check the total number of female
Calculate μ_r	$1+\left[\frac{2*n1*n2}{n1+n2}\right]$
σ_r	$\sqrt{\frac{2.n1.n2(2.n1.n2-n1-n2)}{(n1+n2)^2(n1+n2+1)}}$
Find Zc	formula already given
Value of Zt	Z table value, as nothing is mentioned consider it as 5% level of significance. so our acceptance area is .95 It's a two tailed test, as our H0 is there is no randomness H1 is there is randomness So, we need to confirm either random or not so it is 2 tail. As 2 tail so we should check .95/2 i.e. 0.457 (check Z-table)

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

SOLUTION TO THE PROBLEM

$$\mu_r = 1 + \left[\frac{2n_1 \cdot n_2}{n_1 + n_2}\right] = 1 + \left[2\frac{(20 \cdot 25)}{20+25}\right]$$

$$\Rightarrow 1 + \frac{1000}{45} = \frac{1045}{45} = 23.22.$$

$$\sigma_r = \sqrt{\frac{2n_1 \cdot n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1+n_2)^2 (n_1+n_2-1)}}$$

$$= \sqrt{\frac{2 \times 20 \times 25 \ (2 \cdot 20 \times 25 - 20 - 25}{(20+25)^2 \ (20+25-1)}}$$

$$\Rightarrow \sqrt{\frac{1000 \ (955)}{2025 \ (44)}} = \sqrt{\frac{955000}{89100}}$$

$$= 3.27.$$

$$Z_c = \frac{r - \mu_r}{\sigma_r} = \frac{16 - 23.22}{3.27}$$

$$= |-2.21| = 2.21$$

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

As our Zc is 2.21

Zt is 1.96 (check Z table- 0.475)

As Zc> Zt so we have to reject null hypothesis and confirm that the frequency of male and female are random in nature. In simple term it can be said that the both male and female are not following a predictable patter while leaving home and roaming on the street during lockdown

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Illustration-2

Data below are the life time of battery in hours produce by a company in a particular order @ α=5%. Determine the sample life time of battery is random.

Z	$\frac{R-\mu_R}{\sigma_R}$
Calculate μ_r	$1+\left[\frac{2*n1*n2}{n1+n2}\right]$
σ_r	$\sqrt{\frac{2.n1.n2(2.n1.n2-n1-n2)}{(n1+n2)^2(n1+n2+1)}}$
Find Zc	formula already given
Value of Zt	Z table value, as nothing is mentioned consider it as 5% level of significance. so our acceptance area is .95 It's a two tailed test, as our H0 is there is no randomness H1 is there is randomness So, we need to confirm either random or not so it is 2 tail. As 2 tail so we should check .95/2 i.e. 0.457 (check Z-table)

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Steps in solving the problem

1. Check the battery life and present in column wise name it as Days
2. Make a column and mention it as grade

   (Grade A= more than median value, Grade B= less than median value)*
3. Assign all the figures in ascending orders

   *Here, the median value is 28th value i.e. 266 selected after ascending order , As the total number is 55. (check the process of finding median if any doubts)
4. Find the run from grade column i.e. assign

   - AA as 1 then
   - BBB as 2
   - AA0A0A as 3 (because 0 is coming in same run, 0 is assign as the values are 266) and so on.

Prepared By
Arya Kumar

Days	Grading	ascending order
270	A A 1	201
280		212
248	B B B 2	215
260		217
220		220
285	A A O A O A 3	222
270		225
266		228
269		236
266		238
272		242
225	B B 4	244
228		247
290	A A A A A 5	248
284		248
282		249
276		249
269		249
250	B B B 6	250
249		256
262		258
273	A A 7	260
277		262
258	B B 8	262
264		262
269	A A A 9	264
276		264
278	B- 10	266
249	A A 11	266
286		267
282	B B B B 12	269
264		270
201		271
215		272
222		273
238		276
212		277
242		278
236		280
247		282
249		282
248		282
256		284
271	A A A 13	285
282		286
305		290
217	B - 14	303
303	A A A A 15	305
305		305
309		309
320		320
262	B B B 16	
244		
262		
267	A - 17	

H0: There exist no randomness

H1: There exist randomness

Median = 28th no; ie 266

as there exist 2; 266 so

55 - 2 = 53 number.

Compare all given data

A = 26 = n_1

B = 27 = n_2

$r = 17$

Let A = above 266

B = below 266

Prepared By

Arya Kumar

Faculty of MBA, CIME

antripathy@yahoo.com,

09853422575

Find R i.e. RUN So R= 17	check the repetition of A and B
N=53	As there exist two 266 in the list so we consider 55-2=53
n1= 26	Check the total number of A
n2= 27	Check the total number of B

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

$\mu_r = 1 + \left(\frac{2n_1 \cdot n_2}{n_1 + n_2}\right) \Rightarrow 1 + \frac{1404}{53} = \frac{1457}{53}$

$= 27.49$

$\sigma = \sqrt{\frac{2n_1 \cdot n_2 \,(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2 \,(n_1+n_2-1)}}$

$= \sqrt{\frac{1404\,(1404-53)}{2809\,(52)}}$

$= \sqrt{12.99} = 3.604$

Prepared By
Arya Kumar

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

$$Z_c = \frac{\bar{x} - \mu_r}{\sigma_{\bar{x}}} = \frac{17 - 27.49}{3.604}$$

$$= 2.91$$

Prepared By
Arya Kumar

Z_t @ $\alpha = 5\%$ $= 1.96$

so, $Z_c > Z_t$ which means null hypothesis is rejected. so, we can conclude that there is randomness in life of battery.

Prepared By

Arya Kumar

Faculty of MBA, CIME

aryantripathy@yahoo.com, 09853422575

Module-3 Non-Parametric Test

SIGN- TEST

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

SIGN TEST

- The **sign test** is a statistical method to test for consistent differences between pairs of observations
- Example: Effect of medicine- before and after the treatment

- Let X is a continuous random variable with median

  That is, $P(X < \bar{\mu}) = P(X > \bar{\mu}) = 0.5$

  Thus, the null hypothesis about the median or mean can be set up accordingly.

$H_0: \mu = \mu_0$

$H_1: \bar{\mu}_1 \neq \bar{\mu}_0$ (Two Sided)

$\bar{\mu} > \bar{\mu}_0$ (One Sided)

$\bar{\mu} < \bar{\mu}_0$ (One Sided)

- As you can see we can have tests on one tail or two tail so the probability of proportion (p) will always be considered as 50% chance or ½

  i.e. $H_0: p = \frac{1}{2}$

  against $H_1: p > \frac{1}{2}$ or $p < \frac{1}{2}$ or $p \neq \frac{1}{2}$

- In general, null hypothesis will always be written as p is equal to 50% and alternative may be more/ less/ not equal to 50%

Prepared By
Arya Kumar

Illustration-1
One sample test

Suppose during the semester ,11 students study for hours in a day i.e.

6, 7, 7, 5, 4, 8, 6, 10, 5, 6, 4

By using 5% level of significance can it be said that the average hours study of the students is 8? i.e. ($\mu_{H0}=8$ and against alternative $\mu_{H0}<8$) * this is not the hypothesis

Prepared by
Arya Kumar

Basic information for Solution

1. Assign the values in row or column
2. Compare each values with average value provided in this case its 8
3. Mention -/+ sign if the values found less than 8 or more than 8 respectively

Hours	6	7	7	5	4	8	6	10	5	6	4
Average Hours	8	8	8	8	8	8	8	8	8	8	8
Sign (hrs. - Avg. hrs.)	-	-	-	-	-	0	-	+	-	-	-

Here the signs are (-,-,-,-,-,-,+,-,-,-); 0 is not considered as it is neither + nor -

4. Put the formula

$$Zc = \frac{\bar{p} - p_{H0}}{\sigma_{prop}} \qquad \sigma_{prop} = \sqrt{\frac{p_{H0} x q_{H0}}{n}}$$

Faculty of MBA, CIME
aryantripathy@yahoo.com, 9853422575

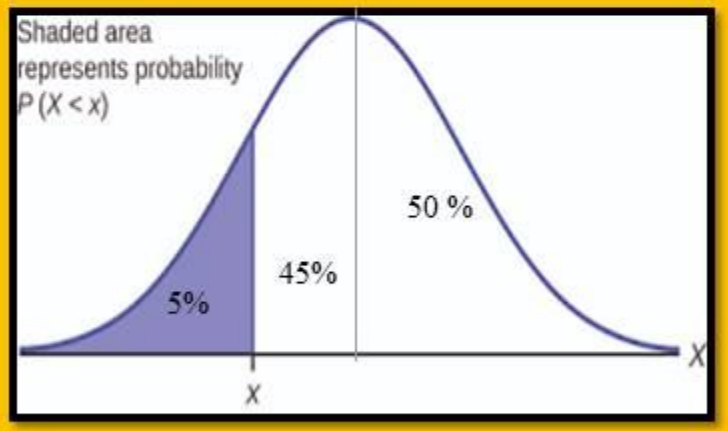
Step-1- Mention the data available:

Prepared by
Arya Kumar

- **If, $p_{H0} = ½$ then $qH0 = ½$,**
- **n = 10** (as out of total observation given in question is 11 nos., but in those 11 nos. one of them is eight so when we compare with all the figures we will get +/- sign but 8-8 is 0 which will not be considered. (like run test)
- **$\bar{p} = .9$** (refers to number of – sign which is 9 so the **average of p** is 9/n =9/10)
- **Zt=1.64 (**check the value of Z from the table for 0.45, it shows 1.64.) *

*This question is a one tailed test so we have to concentrate on one side with 5% level of significance

- it means in normal curve one side i.e. right hand side 50% no need to worry as it is good
- But in left hand side out of 50%, 5% is the rejection side i.e. the shaded part remaining is 45% or 0.45.



Step-2. Identify the null and alternative hypothesis

H0: $p_{H0} = ½$ and

H1: $p_{H0} < ½$ (as the question asked that p is equal to ½ or less than ½).

Faculty of MBA, CIME
aryantripathy@yahoo.com, 9853422575

Step.3 mathematical:

a) Compute the Zc= $\frac{\bar{p}-p_{H0}}{\sigma_{prop}}$, prop refers to proportion (don't get confuse with null hypothesis as you see H0)

Prepared by
Arya Kumar

- $\sigma prop=\sqrt{\frac{p_{H0}xq_{H0}}{n}}=\sqrt{\frac{.5x.5}{10}}=0.1581$
- Therefore , $Zc=\frac{\bar{p}-p_{H0}}{\sigma_{prop}}=\frac{0.9-0.5}{0.1581}=2.5300$

b) Now, we get the value Zc=2.5300.

c) To confirm whether the hypothesis is accepted or rejected. (Follow the method)

- ✓ check the Zt i.e Z table value for one tailed test
- ✓ For one tail test link the row and column for 0.45 (as it is one tail so out of 50%, 5% is level of significance) we get the value as 1.64.
- ✓ Now compare with the Zc with Zt, i.e. Zc> Zt as the calculated value is more than table value so the null hypothesis is rejected.

d) **It means in an average the students are studying less than 8 hours which was our alternative hypothesis**

Faculty of MBA, CIME
aryantripathy@yahoo.com, 9853422575

Illustration-2
Two sample test

Vice Chancellor asked the principal to give report of the performance of senior faculty an junior faculty, that is there any significant difference between two?

Sr.	2	1	4	4	3	3	4	2	4	1	3	3	4	4	4	1	1	2	2	4	4	4	4	3	3	2	3	4	3	4	3	1	4	3	2	2	2	1	3	3
Jr.	3	2	2	3	4	2	2	1	3	1	2	3	4	4	3	2	3	2	3	3	1	4	3	3	2	2	1	1	1	3	2	2	4	4	3	3	1	1	4	2

Here, A means Senior Faculty, B means Junior faculty, The numbers refers to the ratings given to the faculties

Solution

Prepared by
Arya Kumar

Assign +/- sign

Sr.	2	1	4	4	3	3	4	2	4	1	3	3	4	4	4	1	1	2	2	4	4	4	4	3	3	2	3	4	3	4	3	1	4	3	2	2	2	1	3	3
Jr.	3	2	2	3	4	2	2	1	3	1	2	3	4	4	3	2	3	2	3	3	1	4	3	3	2	2	1	1	1	3	2	2	4	4	3	3	1	1	4	2
+/-	-	-	+	+	-	+	+	+	+	0	+	0	0	0	+	-	-	0	-	+	+	0	+	0	+	0	+	+	+	+	+	-	0	-	-	-	+	0	-	+

While comparing A and B, if the A value is more than B then mention +, if the A value is less than B mention – sign.

Faculty of MBA, CIME
aryantripathy@yahoo.com, 9853422575

So we get the + or lets say X = 19 , - or lets say Y=11 and 0 has no relevance so no need to consider 0

Now n= 30 (as from total 40 observations, 10 is deducted due to 0's)

Mathematical solution,

Prepared by
Arya Kumar

H0: PH0=50%

H1: PH0≠ 50%, its two sample and tow tailed test. Check the question it has asked both are same or not same. There is no question for higher or lower side.

Formula:

$$\mathbf{Zc=}\ \frac{\bar{X}-p_{H0}}{\sigma_{prop}},$$

$$\mathbf{\sigma prop=}\ \sqrt{\frac{p_{H0}xq_{H0}}{n}},$$

$$\bar{X}=\frac{X}{n}$$

(similarly you can use for Y, and implement y instead of X the answer will be same)

Faculty of MBA, CIME
aryantripathy@yahoo.com, 9853422575

Prepared by
Arya Kumar

Compute the $Zc = \frac{\bar{X} - p_{H0}}{\sigma_{prop}}$, prop refers to proportion (don't get confuse with null hypothesis as you see H0)

Now, $\bar{X}$= proportion mean, check the number of (+ sign), it is 19 so mean is 19/n=19/30=0.63

$$\sigma prop = \sqrt{\frac{p_{H0} \times q_{H0}}{n}} = \sqrt{\frac{.5x.5}{30}} = 0.091$$

Therefore , $Zc = \frac{\bar{X} - p_{H0}}{\sigma_{prop}} = \frac{0.63 - 0.5}{0.091} = 1.429$

Now, we get the value Zc=1.429.

To confirm whether the hypothesis is accepted or rejected. (Follow the method)

- ✓ check the Zt i.e Z table value for two tailed test
- ✓ For **two tail test** link the row and column for **0.475** (as it is two tail so at 5% level of significance, we get 100%-5%= 95%. Due to two tail 95%/2 =47.5) we get the value as **1.96**
- ✓ Now compare with the Zc with Zt, i.e. **Zc< Zt** as the calculated value is less than table value so the null hypothesis is accepted.

It means there is no significant difference between Senior Faculty and Junior Faculty. So the principal is right that there is no difference between the faculties.

Faculty of MBA, CIME
aryantripathy@yahoo.com, 9853422575

Prepared By Arya Kumar

Faculty of MBA, CIME

aryantripathy@yahoo.com,
09853422575

Non-Parametric Test

Krushkal Wallis Test

More than 2 sample

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

- A Mann-Whitney U test (also called a Mann-Whitney-Wilcoxon test or the Wilcoxon rank-sum test) puts everything in terms of rank rather than in terms of raw values.

- By this some power is lost, If your two samples have very different distributions or very unequal variances, these tests will cause an inflated Type I error rate (check type 1 error)
- The major difference between the Mann-Whitney U and the Kruskal-Wallis H is simply that the KW-H consider more than 2 sample while MW-U consider only 2 samples.

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Formula:

Prepared By ARYA KUMAR

Kruskal-Wallis Formula

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

Compare the result of H value with Chi square table value to find the acceptance or rejection of hypothesis.

- Check the significance level
- Find degree of freedom i.e. (df)= K-1, k refers to number of samples not items selected from each samples.

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Illustration-1

Three machines are used in the packaging of 16kg of wheat floor. Each machine is design, so as to pack on an average of 16kg of wheat floor per bag. **Prepared By ARYA KUMAR**

Sample of bags are selected from each machine and the amount of wheat floor packaging in each bag is shown below:

M1	**15.8**	**15.9**	**16.2**	**15.7**	**16.3**	**15.8**
M2	**16.5**	**16**	**15.4**	**15.9**	**16.2**	**16.1**
M3	**15.7**	**16.4**	**16.2**	**15.9**	**15.7**	**16.3**

At 5% level of significance test the hypothesis that the amount of wheat floor package of machine is same.

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Weight	Machine	Rank
15.4	M2	1
15.7	M1	3
15.7	M3	3
15.7	M3	3
15.8	M1	5.5
15.8	M1	5.5
15.9	M1	8
15.9	M2	8
15.9	M3	8
16	M2	10
16.1	M2	11
16.2	M1	13
16.2	M2	13
16.2	M3	13
16.3	M1	15.5
16.3	M3	15.5
16.4	M3	17
16.5	M2	18

Solution: Steps are as similar like that of Mann Whitney U test

Prepared By ARYA KUMAR

Step 1- Represent all data in ascending order and name it as weight

Step-2 Check the table in the question again and assign Bank name i.e. M1, M2 or M3 as per their respective weight like:
15.4 belongs to M2, then 15.7 belongs to M1 and so on.

Step-3-then rank them all as per the weight . But in case you find any similar figure like 15.7,15.7 and 15.7 then see what is the actual position of it here it belongs to rank 2,3 & 4 so we add (2+3+4)and divide 3 as there are 3 similar case which gives 3 so assign for each. Again 15.8 is repeated two times i.e. for rank 5 and 6 so add both and divide it with 2 as two cases so we get 5.5 each and so on

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Solution: similar like that of Mann Whitney U test

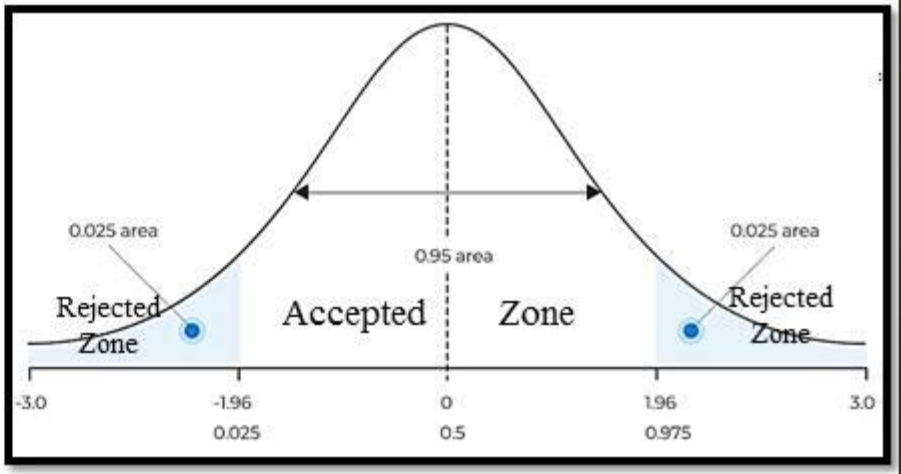
Step-4- Formulate hypothesis

H0= There is no significant difference between the three machines

H1= There is significant difference between the three machines

Step 5- Calculate sum of ranks for M1, M2 and M3(not the weight of the wheat)

- R1= sum of Rank M1= 50.5 (Check Mann Whiteny, if unable to understand)
- R2= sum of Rank-M2=61
- R3= sum of Rank-M3=59.5
- n1=6
- n2=6
- n3=6
- n=18 (6+6+6)
- Two tailed test because the question asked either significant difference or not, it means that both the banks has same median or different median



Note: Draw the diagram in all such kind of questions, show whether 1 tailed test or two tailed test and show the level of significance, to fetch more marks

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Step-6- Put the formula and calculate

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{18(18+1)} \left(\frac{50.5^2}{6} + \frac{61^2}{6} + \frac{59.5^2}{6}\right) - 3(18+1)$$

$$= 0.035(425.04 + 620.16 + 590.04) - 57$$

$$= 0.2334$$

Step-7- Check chi square table

- ✓ 5% level of significance
- ✓ Df= K-1= 3-1=2

So, $\chi^2_{0.05} = 5.99$

Prepared By ARYA KUMAR

So, we can confirm that null hypothesis is accepted as H value is less than chi-square value. It means the machines are significant and all the machines pack same amount of floor, there must be some other reason for the differences.

Faculty of MBA, CIME
aryantripathy@yahoo.com, 09853422575

Prepared By ARYA KUMAR

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38