REAL TIME DATA ANALYSIS

ABSTRACT

Real-time analytics is all about using data as soon as it is produced to answer questions, make predictions, understand relationships, and automate processes. The core requirements of real-time analytics is access to fresh data and fast queries. These are essentially two measures of latency, data latency and query latency. In this article, I have considered Campus Placement data to understand patterns and to draw inferences. I have tried to provide an overview of basic statistical considerations for data analysis. Selecting the appropriate statistical method, one need to know the assumption and conditions of the statistical methods, so that proper statistical method can be selected
for data analysis. Two main statistical methods are used in data analysis: descriptive statistics, which summarizes data using indexes such as mean and median and another is inferential statistics, which draw conclusions from data using statistical tests such as student's t-test. Selection of appropriate statistical method depends on the following three things: Aim and objective of the study, Type and distribution of the data used, and Nature of the observations (paired/unpaired). All type of statistical methods that are used to compare the means are called parametric while statistical methods used to compare other than means (ex-median/mean ranks/proportions) are called nonparametric methods.

Introduction

A data set is a collection of the data of individual cases or subjects. Usually,it is meaningless to present such data individually because that will not produceany important conclusions. In place of individual case presentation, we present summary statistics of our data set with or without analytical form which can be easily absorbable for the audience. Statistics which is a science of collection, analysis, presentation, and interpretation of the data,have two main branches, are descriptive statistics and inferential statistics.
The dataset used for analyzing different statistical tools is of a placement data of students in a XYZ campus. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students. A total of 215 students male and female data regarding all their educational details and placement status is analyzed to determine factors responsible for placements etc.

Selection of appropriate statistical method is very important step in analysis of biomedical data. A wrong selection of the statistical method not only creates some serious problem during the interpretation of the findings but also affects the conclusion of the study. In statistics, for each specific situation, statistical methods are available to analysis and interpretation of the data. To select the appropriate statistical method, one need to know the assumption and conditions of the statistical methods, so that proper statistical method can be selected for data analysis.[1] Other than knowledge of the statistical methods, another very important aspect is nature and type of the data collected and objective of the study
because as per objective, corresponding statistical methods are selected which are suitable on given data. Practice of wrong or inappropriate statistical method is a common phenomenon in the published articles in biomedical research. Incorrect statistical methods can be seen in many conditions like use of unpaired t-test on paired data or use of parametric test for the data which does not follow the normal distribution, etc., At present, many statistical software like SPSS, R, Stata, and SAS are available and using these softwares, one can easily perform the statistical analysis but selection of appropriate statistical test is still a difficult task for the biomedical researchers especially those with non statistical background.[2] Two main

statistical methods are used in data analysis: descriptive statistics, which summarizes data using indexes such as mean, median, standard deviation and another is inferential statistics, which draws conclusions from data using statistical tests such as student's t-test, ANOVA test, etc.

Brief summary of the data

Ssc_p		
	Statistic	Standard error
Mean	67.3034	.73841
95% Confidence Interval for Mean Lower Bound Upper Bound		65.8479 68.7589
5% Trimmed Mean	67.4619	
Median	67.0000	
Variance	117.228	
Std. Deviation	10.82721	
Minimum	40.89	
Maximum	89.40	
Range	48.51	
Interquartile Range	15.60	
Skewness	-.133	.166
Kurtosis	-.608	.330

Hsc_p		
	Statistic	Standard error
Mean	66.3332	.74320
95% Confidence Interval for Mean Lower Bound Upper Bound	64.8682 67.7981	
5% Trimmed Mean	66.2243	
Median	65.0000	
Variance	118.756	
Std. Deviation	10.89751	
Minimum	37.00	
Maximum	97.70	
Range	60.70	
Interquartile Range	12.20	
Skewness	.164	.166
Kurtosis	.451	.330

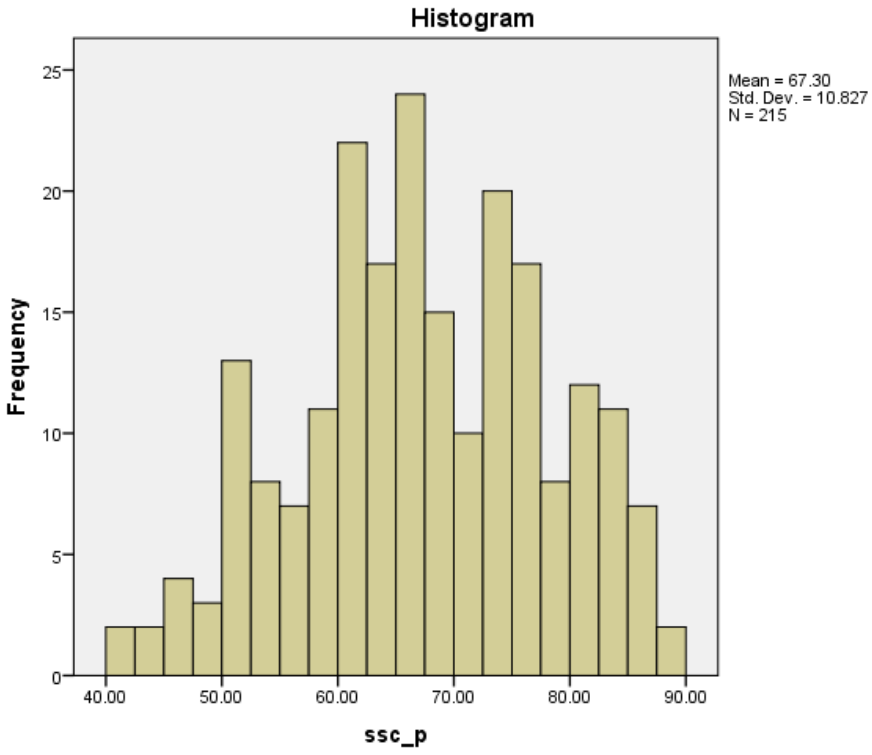
degree_p		
	Statistic	Standard error
Mean	66.3702	.50186
95% Confidence Interval for Mean Lower Bound Upper Bound	65.3810 67.3594	
5% Trimmed Mean	66.2721	
Median	66.0000	
Variance	54.151	
Std. Deviation	7.35874	
Minimum	50.00	
Maximum	91.00	
Range	41.00	
Interquartile Range	11.00	
Skewness	.245	.166
Kurtosis	.052	.330

etest_p		
	Statistic	Standard error
Mean	71.9540	.90638
95% Confidence Interval for Mean Lower Bound Upper Bound	70.1675 73.7406	
5% Trimmed Mean	71.7099	
Median	70.0000	
Variance	176.627	
Std. Deviation	13.29011	
Minimum	50.00	
Maximum	98.00	
Range	48.00	
Interquartile Range	23.00	
Skewness	.299	.166
Kurtosis	-1.078	.330

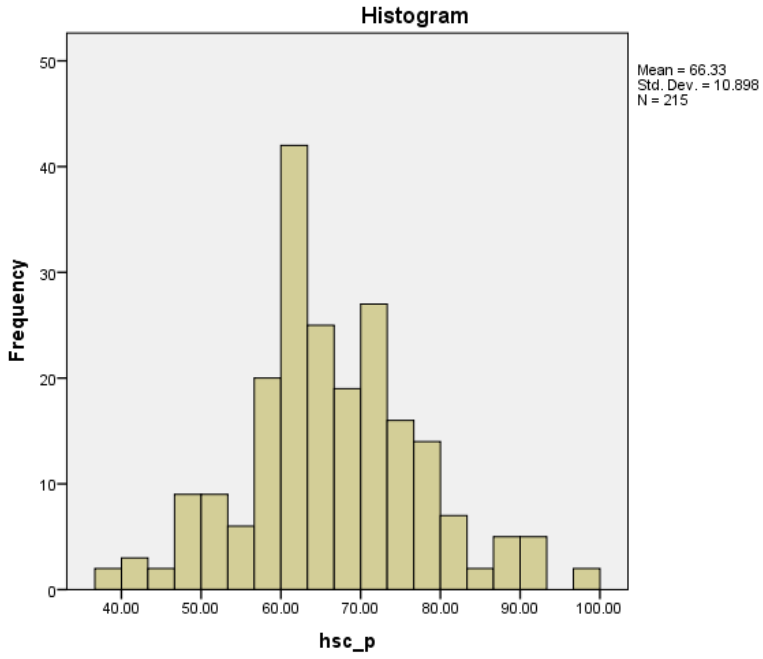
mba_p		
	Statistic	Standard error
Mean	62.2782	.39783
95% Confidence Interval for Mean Lower Bound Upper Bound	61.4940 63.0624	
5% Trimmed Mean	62.1497	
Median	62.0000	
Variance	34.028	
Std. Deviation	5.83338	
Minimum	51.21	
Maximum	77.89	
Range	26.68	
Interquartile Range	8.38	
Skewness	.314	.166
Kurtosis	-.471	.330

Histogram :

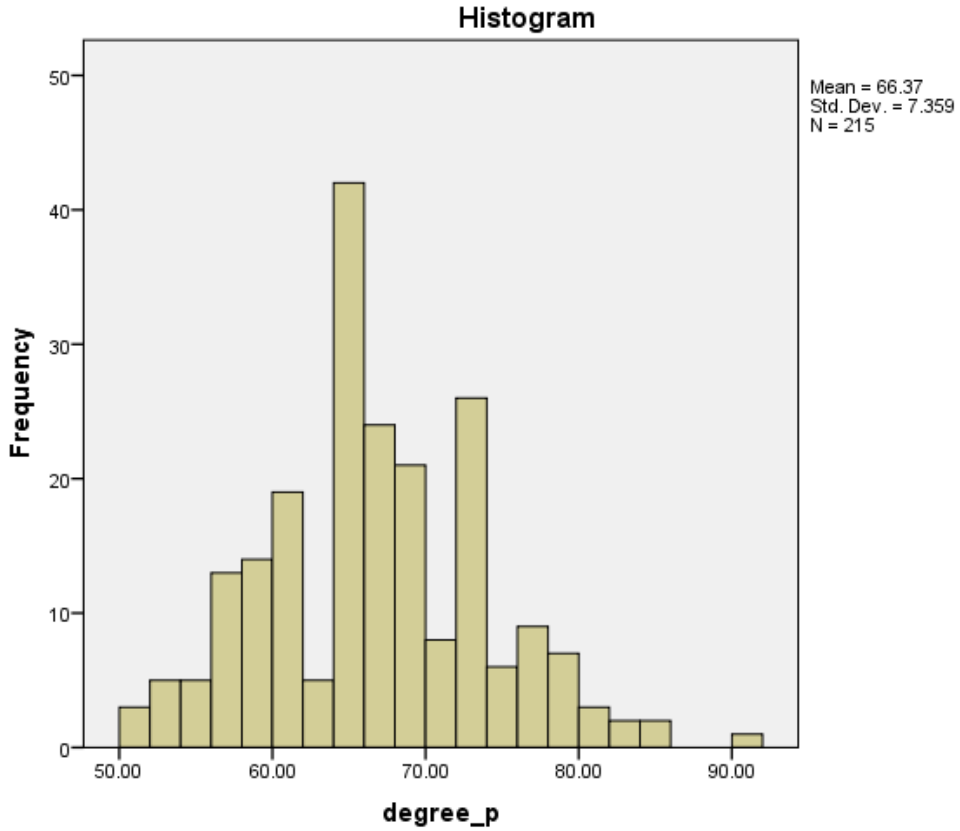
1. Ssc_p



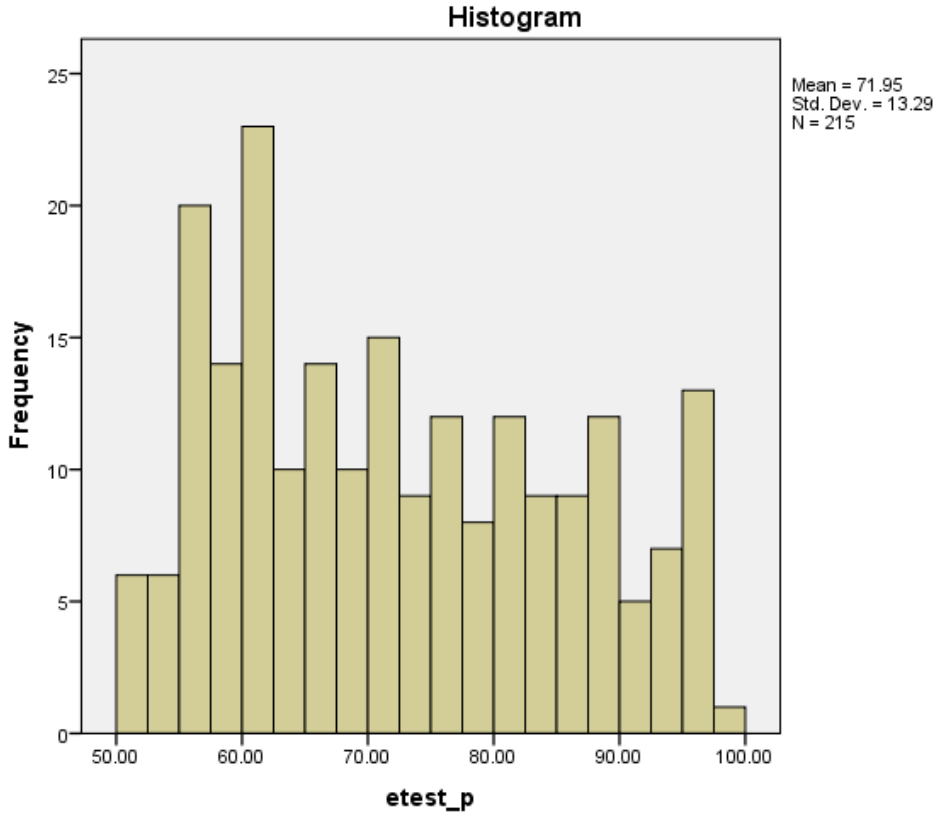
2. Hsc_p ：



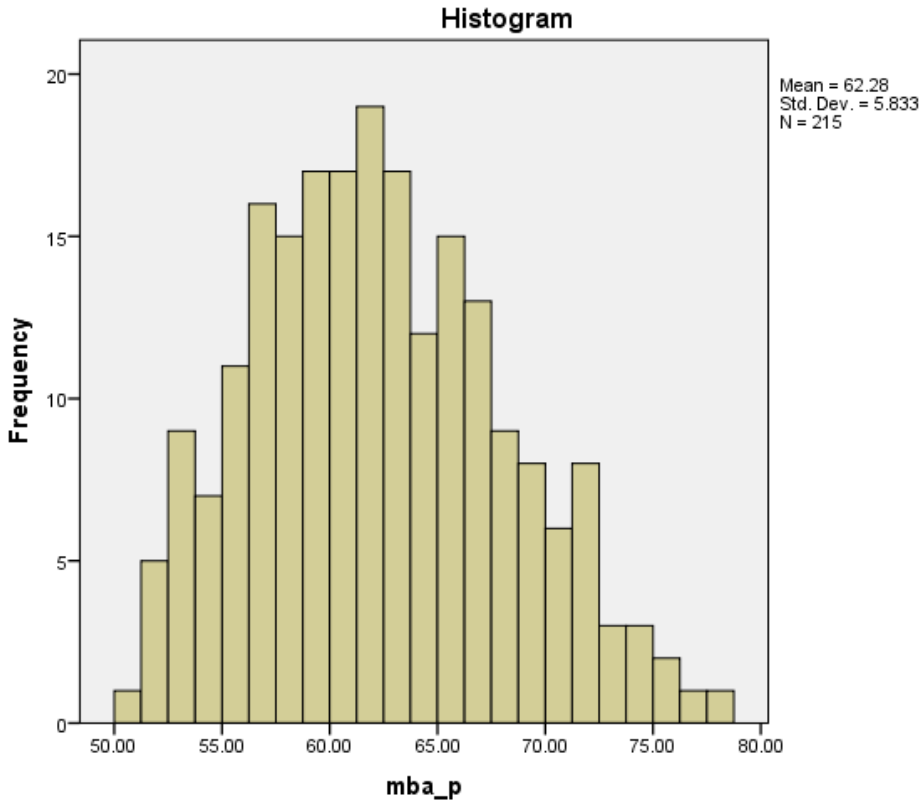
3. degree_p



4. Etest_p :



5. mba_p :



- **Understanding relation between Arithmetic mean, Geometric mean and Harmonic mean** :

	AM	GM	HM
ssc_p	67.3034	66.40355	65.45964
hsc_p	66.33316	65.42462	64.47488
degree_p	66.37019	65.96693	65.56107
etest_p	71.95405	70.74636	69.57394
mba_p	62.27819	62.00862	61.74342

General relation between AM, GM and HM = > **1. AM > GM > HM**
2. $(GM)^2$ = AM * HM

From data above calculated from dataset, AM > GM > HM is well observed.

	GM	=	AM * HM
ssc_p	4409.431	=	4405.656
hsc_p	4280.381	=	4276.823
degree_p	4351.636	=	4351.301
etest_p	5005.047	=	5006.126
mba_p	3845.069	=	3845.268

From table above, with few deviation in ssc_p and hsc_p data, it is found our database value follow the relation $(GM)^2$ = AM * HM.

- **Relation between mean , median and mode and understanding based on it skewness of distribution** :

In statistics, for a moderately skewed distribution, there exists a relation between mean, median and mode. This mean median and mode relationship is known as the **"empirical relationship"** which is defined as Thrice of **Median is equal to the sum of Mode and 2 times the mean**.

- Mean is the average of the data set which is calculated by adding all the data values together and dividing it by the total number of data sets.
- Median is the middle value among the observed set of values and is calculated by arranging the values in ascending order or in descending order and then choosing the middle value.
- Mode is the number from a data set which has the highest frequency and is calculated by counting the number of times each data value occur.

Following conclusions can be drawn for mean , median and mode from table below :

	MEAN	MEDIAN	MODE
ssc_p	67.3034	67	62
hsc_p	66.33316	65	63
degree_p	66.37019	66	65
etest_p	71.95405	70	60
mba_p	62.27819	62	56.7

3 * MEDIAN	=	2*MEAN + MODE	Results
201	=	196.6067907	Significant diff
195	=	195.6663256	Moderately skewed
198	=	197.7403721	Moderately skewed
210	=	203.908093	Significant diff
186	=	181.2563721	Significant diff

Also as observed in above table of mean, median and mode for every data MEAN > MEDIAN > MODE, Thus it can concluded that data are positively skewed distribution.

Normality of data :

Various statistical methods used for data analysis makeassumptions about normality, including correlation,regression, t-tests, and analysis of variance. Central limittheorem states that when sample size has 100 or more observations, violation of the normality is not a major issue. Although for meaningful conclusions, assumption ofthe normality should be followed irrespective of the sample size. If a continuous data follow normal distribution, then we present this data in mean value. Further, this mean value is used to compare between/among the groups to calculate the significance level (P value). If our data are not normally distributed, resultant mean is not a representative value of our data. A wrong selection of the representative value of a data set and further calculated significance level using this representative value might give wrong interpretation.] That is why, first we test the normality of the data, then we decide whether mean is applicable as representative value of the data or not. If applicable, then means are compared using parametric test otherwise medians are used to compare the groups, using nonparametric methods.

Checking normality of our database :

The two well-known tests of normality, namely, the Kolmogorov–Smirnov test and the Shapiro–Wilk test are most widely used methods to test the normality of the data. Normality tests can be conducted in the statistical software “SPSS”. The Shapiro–Wilk test is more appropriate method for small sample sizes (<50 samples) although it can also be handling on larger sample size while Kolmogorov– Smirnov test is used for n ≥50. For both of the above tests, null hypothesis states that data are taken from normal distributed population. When P > 0.05, null hypothesis accepted and data are called as normally distributed.

Hypothesis testing :
Let us assume :
H_0 = Data taken follows normal distribution.
H_a = Data taken does not follows normal distribution.

From SPSS,

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ssc_p	.059	215	.069	.986	215	.032
hsc_p	.085	215	.001	.985	215	.021
degree_p	.076	215	.004	.989	215	.100
etest_p	.095	215	.000	.949	215	.000
mba_p	.052	215	.200*	.985	215	.019

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

As n > 50 , thus By referring to Kolmogorov– Smirnov test, only **ssc_p and mba_p** data accept null hypothesis as p >0.05 and thus data here follow normal distribution , whereas hsc_p, degree_p and etest_p have p <0.05 thus alternate hypothesis is accepted.

Skewness is a measure of symmetry, or more precisely, the lack of symmetry of the normal distribution. Kurtosis is a measure of the peakedness of a distribution. The original kurtosis value is sometimes called kurtosis (proper).
A distribution is called approximate normal if skewness or kurtosis (excess) of the data are between – 1 and + 1. Although this is a less reliable method in the small-to-moderate sample size (i.e., n <300) because it can not adjust the standard error (as the sample size increases, the standard error decreases). To overcome this problem, a z-test is applied for normality test using skewness and kurtosis. A Z score could be obtained by dividing the skewness values or excess kurtosis value by their standard errors. For small sample size (n <50), z value ± 1.96 are sufficient to establish normality of the data. However, medium-sized samples (50≤ n <300), at absolute z-value ± 3.29, conclude the distribution of the sample is normal. For sample size >300, normality of the data is depend on the histograms and the absolute values of skewness and kurtosis. Either an absolute skewness value ≤2 or an absolute kurtosis (excess) ≤4 may be used as reference values for determining considerable normality.

			Standard error	z score
ssc_p	skewness	-0.133	0.166	-0.8012
	kurtosis	-0.608	0.33	-1.84242
hsc_p	skewness	0.164	0.166	0.987952
	kurtosis	0.451	0.33	1.366667
degree_p	skewness	0.245	0.166	1.475904

	kurtosis	0.052	0.33	0.157576
etest_p	skewness	0.299	0.166	1.801205
	kurtosis	-1.078	0.33	-3.26667
mba_p	skewness	0.314	0.166	1.891566
	kurtosis	-0.471	0.33	-1.42727

As we have n = 215 , thus observed z score for ssc_p, hsc_p , degree_p , etest_p (almost) mba_p are within z score value ±3.29 , thus these data can be called as approximately normal distribution.

Outlier detection :

An outlier is **an observation that lies an abnormal distance from other values in a random sample from a population**. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Outliers exists due to variability in data and an experimental measurement error.
Criteria to identify an outlier :
a. Data point that falls outside of 1.5 times of an interquartile range above the 3rd quartile aand below the 1st quartile
b. Data point that falls outside of 3 deviations , we use a z score and if the z score falls outside of 2 standard deviations.

As data here is approximately normal distribution, to get more accurate results we will use Quartile method to detect outliers.
This method of Outliers' detection is done with the help of GOOGLE COLABORATORY.

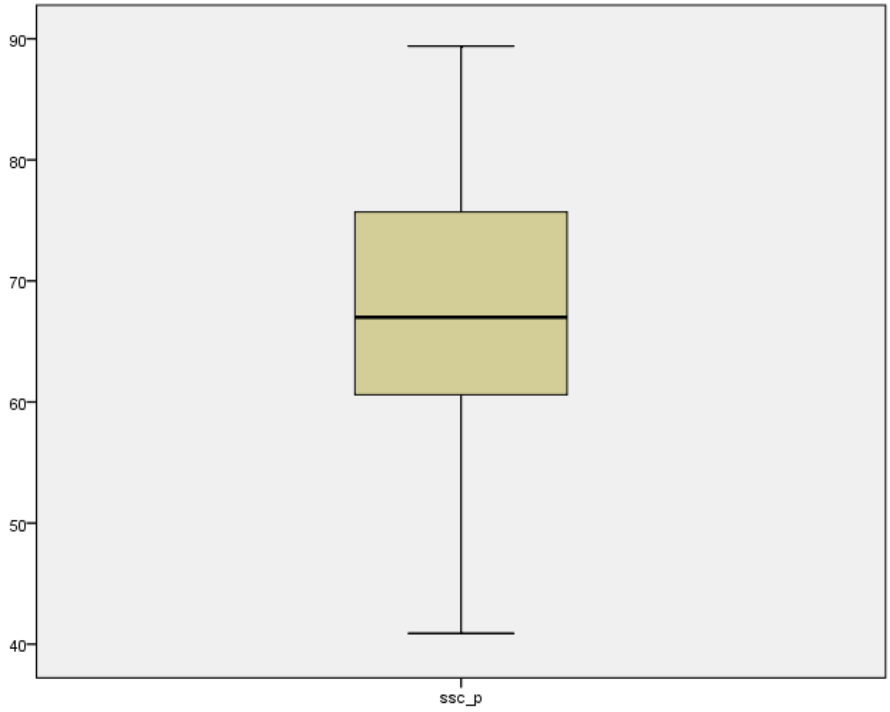
(Please refer google colaboratory program pdf attached at end)

- RESULTS SUMMARY :

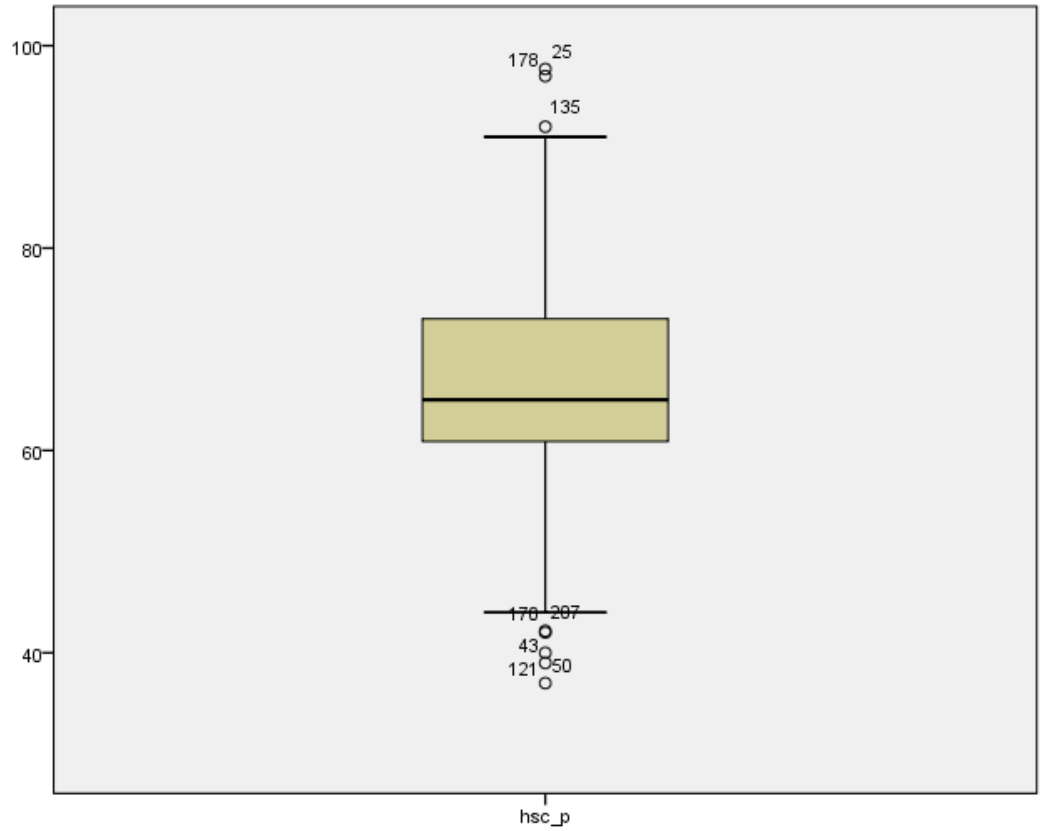
Sr. No.	Data	Outliers' detection
1.	ssc_p	No outliers
2.	hsc_p	97.7, 39, 37, 40, 92, 42.16, 97, 42
3.	degree_p	91
4.	etest_p	No outliers
5.	mba_p	No outliers

- **Box and Whisker's plot** :

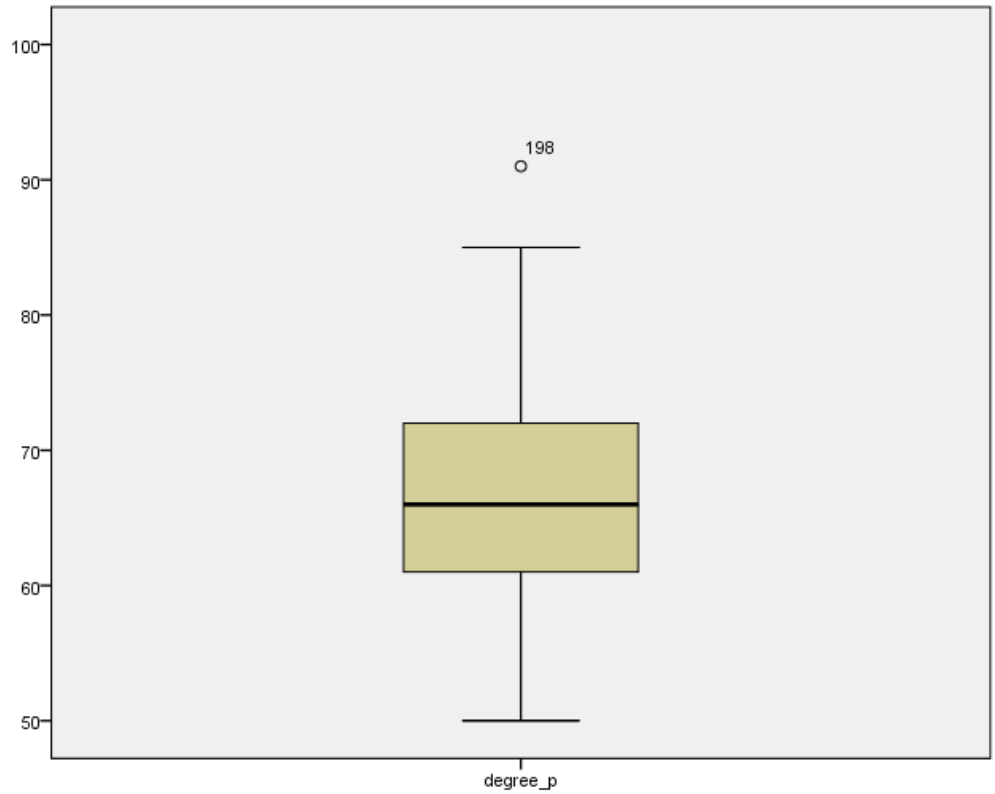
- 1. Ssc_p



2. Hsc_p

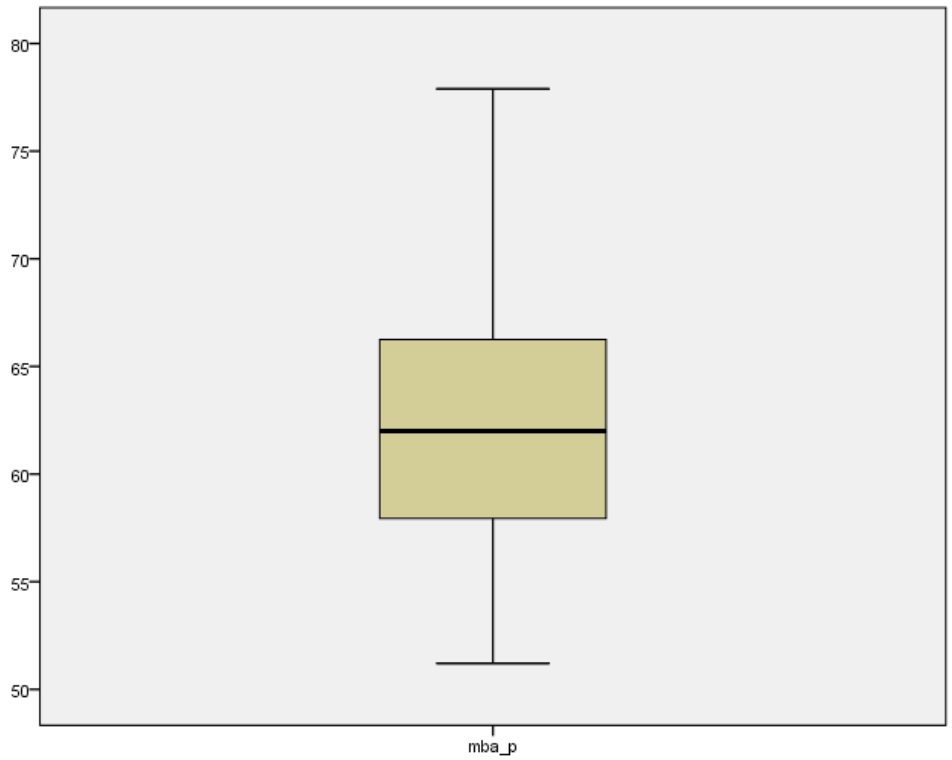


3. Degree_p



4. etest_p

5. mba_p



Pearson's Correlation Coefficient :

In statistics, the **Pearson correlation coefficient** the **bivariate correlation** is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

Correlations

		ssc_p	hsc_p	degree_p	etest_p	mba_p
ssc_p	Pearson Correlation	1	.511**	.538**	.249**	.388**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	215	215	215	215	215
hsc_p	Pearson Correlation	.511**	1	.434**	.233**	.355**
	Sig. (2-tailed)	.000		.000	.001	.000
	N	215	215	215	215	215
degree_p	Pearson Correlation	.538**	.434**	1	.208**	.402**
	Sig. (2-tailed)	.000	.000		.002	.000
	N	215	215	215	215	215
etest_p	Pearson Correlation	.249**	.233**	.208**	1	.210**
	Sig. (2-tailed)	.000	.001	.002		.002
	N	215	215	215	215	215
mba_p	Pearson Correlation	.388**	.355**	.402**	.210**	1
	Sig. (2-tailed)	.000	.000	.000	.002	
	N	215	215	215	215	215

**. Correlation is significant at the 0.01 level (2-tailed).

Hypothesis testing :

Let us assume :

H_0 = Data taken has no significant correlation.

H_a = Data taken has significant correlation.

Thus, for every data Sig. < 0.01 , rejecting null hypothesis , we conclude that correlation values are significant and by considering values : ssc_p, hsc_p and degree_p have better correlation than values and rest all others have comparatively low correlation.

F test :

As data were found approximately normal, thus we can apply parametric test. An F-test is **any statistical test in which the test statistic has an F-distribution under the null hypothesis**.

The F-test is used by a researcher in order **to carry out the test for the equality of the two population variances**. If a researcher wants to test whether or not two independent samples have been drawn from a normal population with the same variability, then he generally employs the F-test.

Determining equal variance in population of male and female for ssc_p, hsc_p, degree_p, etest_p and mba_p for condition (so as to limit size of n < 30) that HSC stream is to be Science, degree in Science & Technology and MBA specialization in Marketing &Finance.

			MALE							
ssc_p	**ssc_b**	**hsc_p**	**hsc_b**	**hsc_s**	**degree_p**	**degree_t**	**workex**	**etest_p**	**specialisation**	**mba_p**
79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	55	Mkt&Fin	66.28
55	Central	49.8	Central	Science	67.25	Sci&Tech	Yes	55	Mkt&Fin	51.58
82	Others	64	Others	Science	66	Sci&Tech	No	67	Mkt&Fin	62.14
76.5	Others	97.7	Others	Science	78.86	Sci&Tech	No	97.4	Mkt&Fin	74.01
81	Central	68	Others	Science	64	Sci&Tech	No	93	Mkt&Fin	62.56
52.6	Central	65.58	Central	Science	72.11	Comm&Mgmt	Yes	57.6	Mkt&Fin	56.66
73	Central	73	Central	Science	66	Sci&Tech	Yes	70	Mkt&Fin	68.07
82	Others	61	Others	Science	62	Sci&Tech	No	89	Mkt&Fin	65.45
64	Others	80	Others	Science	65	Comm&Mgmt	No	69	Mkt&Fin	57.65
84	Others	90.9	Others	Science	64.5	Sci&Tech	Yes	86.04	Mkt&Fin	59.42
84	Others	79	Others	Science	68	Sci&Tech	No	84	Mkt&Fin	66.69
70	Others	63	Others	Science	70	Comm&Mgmt	Yes	55	Mkt&Fin	62
61.08	Others	50	Others	Science	54	Sci&Tech	Yes	71	Mkt&Fin	65.69
77	Central	75	Others	Science	73	Sci&Tech	Yes	80	Mkt&Fin	67.05
85	Others	60	Others	Science	73.43	Sci&Tech	No	60	Mkt&Fin	61.29
71	Central	58.66	Others	Science	58	Sci&Tech	Yes	56	Mkt&Fin	61.3

49	Central	59	Central	Science	65	Sci&Tech	Yes	86	Mkt&Fin	62.48
67	Others	63	Others	Science	64	Sci&Tech	Yes	60	Mkt&Fin	61.87
63	Central	67	Others	Science	64	Comm& Mgmt	Yes	75	Mkt&Fin	66.46
67.9	Others	62	Others	Science	67	Sci&Tech	No	58.1	Mkt&Fin	75.71
73.24	Others	50.83	Others	Science	64.27	Sci&Tech	Yes	64	Mkt&Fin	66.23
78.5	Central	65.5	Central	Science	67	Sci&Tech	Yes	95	Mkt&Fin	64.86
72	Others	63	Others	Science	77.5	Sci&Tech	Yes	78	Mkt&Fin	54.48
58	Others	60	Others	Science	72	Sci&Tech	No	74	Mkt&Fin	53.62

FEMALE

ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p
77.4	Others	60	Others	Science	64.74	Sci&Tech	Yes	92	Mkt&Fin	63.62
86.5	Others	64.2	Others	Science	67.4	Sci&Tech	No	59	Mkt&Fin	59.69
79	Others	61	Others	Science	75.5	Sci&Tech	Yes	70	Mkt&Fin	68.2
82	Others	64	Others	Science	73	Sci&Tech	Yes	96	Mkt&Fin	71.77
75.4	Others	60.5	Central	Science	84	Sci&Tech	No	98	Mkt&Fin	65.25

Hypothesis testing :

Let us assume :

H_0 = No difference in variance between male and female population

H_a = Significant difference in variance between male and female population

By using excel functions, observed F test results for alpha = 0.05 :

Sr. No.	Data	F test value	Results
1.	ssc_p	0.095586	P > 0.05 : H_0 accepted : Equal variance population
2.	hsc_p	0.003352	P < 0.05 : Ha accepted : Unequal variance population
3.	degree_p	0.419716	P > 0.05 : H_0 accepted : Equal variance population
4.	etest_p	0.431483	P > 0.05 : H_0 accepted : Equal variance population
5.	mba_p	0.701483332	P > 0.05 : H_0 accepted : Equal variance population

Student's T test :

Student's t-test, in statistics, **a method of testing hypothesis about the mean of a small sample drawn from a normally distributed population when the population standard deviation is unknown**.

As data of ssc_p and mba_p are normally distributed , thus determining difference in values of male and female ssc_p and mba_p by using Student t test of paired test equal variance(as determined by f test above)

Hypothesis testing :

Let us assume :

H_0 = No difference in values between ssc_p and mba_p

H_a = Significant difference in values between ssc_p and mba_p

By using excel functions ; (alpha = 0.05)

Sr. No.	Description	F test value	Results
1.	Male	0.000694	P <0.05 : Ha accepted ; Significant difference in values between ssc_p and mba_p
2.	Female	0.010758	P <0.05 : Ha accepted ; Significant difference in values between ssc_p and mba_p

Analysis of Variance (ANOVA test) :

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

t-test is statistical hypothesis test used to compare the means of two population groups. ANOVA is an observable technique used to compare the means of more than two population groups.

Determining relation between results of t test and ANOVA :

For this we are referring data of mba_p and getting placed or not placed :

Group Statistics

	status	N	Mean	Std. Deviation	Std. Error Mean
mba_p	Placed	25	63.4892	4.93359	.98672
	Not Placed	4	63.7125	9.97401	4.98701

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
mba_p	Equal variances assumed	2.891	.101	-.073	27
	Equal variances not assumed			-.044	3.239

Independent Samples Test

		t-test for Equality of Means			
		Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
					Lower
mba_p	Equal variances assumed	.943	-.22330	3.07894	-6.54077
	Equal variances not assumed	.968	-.22330	5.08369	-15.74737

Independent Samples Test

		t-test for Equality of Means
		95% Confidence Interval of the Difference
		Upper
mba_p	Equal variances assumed	6.09417
	Equal variances not assumed	15.30077

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
status	.0	Not Placed	4
	1.0	Placed	25

Tests of Between-Subjects Effects

Dependent Variable: mba_p

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.172[a]	1	.172	.005	.943
Intercept	55794.043	1	55794.043	1706.798	.000
status	.172	1	.172	.005	.943
Error	882.611	27	32.689		
Total	117891.705	29			
Corrected Total	882.783	28			

From this it can be observed (t test value) ^{2} = (ANOVA test value)
as ; t test value = -0.073 ANOVA test value (F value) = 0.005

= $(-0.073)^2$ = 0.005329 which approximately equal to ANOVA f test value = 0.005

Considering one more such test , to verify this relation;
For this we are referring data of salary and getting placed or not placed :

Group Statistics

	status	N	Mean	Std. Deviation	Std. Error Mean
salary	Placed	25	322720.000	115811.4560	23162.2912
	Not Placed	4	.000	.0000	.0000

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
salary	Equal variances assumed	5.121	.032	5.488	27
	Equal variances not assumed			13.933	24.000

Independent Samples Test

		t-test for Equality of Means			
		Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
					Lower
salary	Equal variances assumed	.000	322720.0000	58799.5848	202073.2176
	Equal variances not assumed	.000	322720.0000	23162.2912	274915.3805

Independent Samples Test

		t-test for Equality of Means
		95% Confidence Interval of the Difference
		Upper
salary	Equal variances assumed	443366.7824
	Equal variances not assumed	370524.6195

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
status	.0	Not Placed	4
	1.0	Placed	25

Tests of Between-Subjects Effects

Dependent Variable: salary

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	35913171862 0.690[a]	1	35913171862 0.690	30.123	.000
Intercept	35913171862 0.690	1	35913171862 0.690	30.123	.000
status	35913171862 0.690	1	35913171862 0.690	30.123	.000
Error	32189504000 0.000	27	11922038518 .519		
Total	29256000000 00.000	29			
Corrected Total	68102675862 0.690	28			

a. R Squared = .527 (Adjusted R Squared = .510)

From this it can be observed (t test value) ^{2} = (ANOVA test value)
as ; t test value = 5.488 ANOVA test value (F value) = 30.123

= $(5.488)^2$ = 30.118 which approximately equal to ANOVA f test value = 30.123

Hence, this relation is verified.

Chi- Square test :

A **chi-squared test** (also **chi-square** or χ^2 **test**) is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof. Pearson's chi-between the expected frequencies and the observed frequencies in one or more categories of a contingency table. In the standard applications of this test, the observations are classified into mutually exclusive classes. If the null hypothesis that there are no differences between the classes in the population is true, the test statistic computed from the observations follows a χ^2 frequency distribution. The purpose of the test is to evaluate how likely the observed frequencies would be assuming the null hypothesis is true.

Hypothesis testing :

Let us assume :

H_0 = Data are independent that is no association.

H_a = Data are dependent that is association between values.

Determining by using SPSS ;

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
degree_t * status	215	100.0%	0	0.0%	215	100.0%
specialisation * status	215	100.0%	0	0.0%	215	100.0%
ssc_b * status	215	100.0%	0	0.0%	215	100.0%
hsc_s * status	215	100.0%	0	0.0%	215	100.0%
gender * status	215	100.0%	0	0.0%	215	100.0%

degree_t * status

Crosstab

			status		
			Not Placed	Placed	Total
degree_t	Comm&Mgmt	Count	43	102	145
		Expected Count	45.2	99.8	145.0
	Others	Count	6	5	11
		Expected Count	3.4	7.6	11.0
	Sci&Tech	Count	18	41	59

	Expected Count	18.4	40.6	59.0
Total	Count	67	148	215
	Expected Count	67.0	148.0	215.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2.969[a]	2	.227
Likelihood Ratio	2.734	2	.255
N of Valid Cases	215		

a. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 3.43.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.118	.227
	Cramer's V	.118	.227
N of Valid Cases		215	

<u>Result</u> : As Pearson Chi-Square value > 0.05; thus H_0 null hypothesis is accepted. So , it can be concluded that type of degree and getting placed or not are independent (not associated with each other)

specialisation * status

Crosstab

			status		
			Not Placed	Placed	Total
specialisation	Mkt&Fin	Count	25	95	120
		Expected Count	37.4	82.6	120.0
	Mkt&H	Count	42	53	95

	R	Expected Count	29.6	65.4	95.0
Total		Count	67	148	215
		Expected Count	67.0	148.0	215.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13.508[a]	1	.000		
Continuity Correction[b]	12.440	1	.000		
Likelihood Ratio	13.532	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	215				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 29.60.
b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-.251	.000
	Cramer's V	.251	.000
N of Valid Cases		215	

<u>Result</u> : As Pearson Chi-Square value(000) < 0.05; thus H_a alternate hypothesis is accepted. So , it can be concluded that type of specialisation and getting placed or not are dependent (associated with each other) and Cramer's V tell how much associated , so here we can say moderately associated.

ssc_b * status

Crosstab

			status		
			Not Placed	Placed	Total
ssc_b	Central	Count	38	78	116
		Expected Count	36.1	79.9	116.0

	Others	Count	29	70	99
		Expected Count	30.9	68.1	99.0
Total		Count	67	148	215
		Expected Count	67.0	148.0	215.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.299[a]	1	.584		
Continuity Correction[b]	.159	1	.690		
Likelihood Ratio	.300	1	.584		
Fisher's Exact Test				.658	.345
N of Valid Cases	215				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 30.85.
b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.037	.584
	Cramer's V	.037	.584
N of Valid Cases		215	

<u>Result</u> : As Pearson Chi-Square value > 0.05; thus H_0 null hypothesis is accepted. So , it can be concluded that ssc_b and getting placed or not are independent (not associated with each other)

hsc_s * status

Crosstab

			status		
			Not Placed	Placed	Total
hsc_s	Arts	Count	5	6	11

		Expected Count	3.4	7.6	11.0
	Commerce	Count	34	79	113
		Expected Count	35.2	77.8	113.0
	Science	Count	28	63	91
		Expected Count	28.4	62.6	91.0
Total		Count	67	148	215
		Expected Count	67.0	148.0	215.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.115[a]	2	.573
Likelihood Ratio	1.050	2	.592
N of Valid Cases	215		

a. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 3.43.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.072	.573
	Cramer's V	.072	.573
N of Valid Cases		215	

<u>Result</u> : As Pearson Chi-Square value > 0.05; thus H_0 null hypothesis is accepted. So , it can be concluded that hsc_b and getting placed or not are independent (not associated with each other)

gender * status

Crosstab

			status		
			Not Placed	Placed	Total
gender	F	Count	28	48	76
		Expected Count	23.7	52.3	76.0
	M	Count	39	100	139
		Expected Count	43.3	95.7	139.0
Total		Count	67	148	215
		Expected Count	67.0	148.0	215.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.768[a]	1	.184		
Continuity Correction[b]	1.382	1	.240		
Likelihood Ratio	1.746	1	.186		
Fisher's Exact Test				.218	.120
N of Valid Cases	215				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 23.68.
b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.091	.184
	Cramer's V	.091	.184
N of Valid Cases		215	

Result : As Pearson Chi-Square value > 0.05; thus H_0 null hypothesis is accepted. So , it can be concluded that gender and getting placed or not are independent (not associated with each other)

Summary : Only type of specialization associated or dependent on getting placement, rest all are independent factors.

Cochran's Q test :

In statistics, in the analysis of two-way randomized block designs where the response variable can take only two possible outcomes (coded as 0 and 1), **Cochran's Q test** is a non-parametric statistical test to verify whether *k* treatments have identical effects.[1][2][3] It is named after William Gemmell Cochran. Cochran's Q test should not be confused with Cochran's C test, which is a variance outlier test. Put in simple technical terms, Cochran's Q test requires that there only be a binary response (e.g. success/failure or 1/0) and that there be more than 2 groups of the same size. The test assesses whether the proportion of successes is the same between groups. Often it is used to assess if different observers of the same phenomenon have consistent results (interobserver variability)

Determining relation between type of degree and type of specialization to placement status.

Hypothesis testing : For alpha = 0.05
Let us assume :
H_0 = Data values are statistically insignificant
H_a = Data values are statistically significant.

Cochran Test

Frequencies

	Value	
	0	1
degree_t	156	59
specialisation	120	95
status	67	148

Test Statistics

N	215
Cochran's Q	67.184[a]
df	2
Asymp. Sig.	.000

Result : As Sig < 0.05 , so accepting alternalte hypothesis that type of degree and specialization have some relation with placement status.

To understand this difference, we use McNemar test, McNemar's test is **a statistical test used on paired nominal data**. It is applied to 2 × 2 contingency tables with a dichotomous trait, with matched pairs of subjects, to determine whether the row and column marginal frequencies are equal (that is, whether there is "marginal homogeneity").

Hypothesis testing : For alpha = 0.05
Let us assume :
H_0 = Data values are statistically insignificant
H_a = Data values are statistically significant.

McNemar Test

Crosstabs

degree_t & specialisation

	specialisation	
degree_t	Mkt.&Fin	Mkt&HR
Others	90	66
Sci&Tech	30	29

status & degree_t

	degree_t	
status	Others	Sci&Tech
Not Placed	49	18
Placed	107	41

specialisation & status

	status	
specialisation	Not Placed	Placed
Mkt.&Fin	25	95
Mkt&HR	42	53

Test Statistics[a]

	degree_t & specialisation	status & degree_t	specialisation & status
N	215	215	215
Chi-Square[b]	12.760	61.952	19.737
Asymp. Sig.	.000	.000	.000

a. McNemar Test
b. Continuity Corrected

Result : As McNemar are run, we need to make Beltourni adjustment, by dividing alpha 0.05 to 3, then we get 0.01667. As all three combination as below 0.01667, thus accepting alternative hypothesis that above three combinations have some statistics significance.

Multiple Regression :

Multiple regression is **a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables**. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.

Here we are determining that mba_p, etest_p and work experience determine the salary .

By using SPSS;

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	mba_p, workex, etest_p[b]	.	Enter

a. Dependent Variable: salary
b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.105[a]	.011	-.019	69651.949

a. Predictors: (Constant), mba_p, workex, etest_p

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5368419130.818	3	1789473043.606	.369	.776[b]
	Residual	480288008053.648	99	4851394020.744		
	Total	485656427184.466	102			

a. Dependent Variable: salary
b. Predictors: (Constant), mba_p, workex, etest_p

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	275020.050	79613.099		3.454	.001
	workex	10431.336	13890.585	.076	.751	.454
	etest_p	376.481	515.774	.077	.730	.467
	mba_p	-448.981	1354.999	-.035	-.331	.741

a. Dependent Variable: salary

Results :

1. As R^2 = 0.011, takenas a set, the mba_p, etest_p and work experience accounts for about only 1.1% of variance in salary after getting placed.

2 . By referring to ANOVA table, alpha = 0.05; the overall regression model is insignificant as F(3,99) = 0.369 , p > 0.05, R^2 = 0.011.

Conclusion :

Thus various statistical tools were applied and we tried to understand factors in different conditions affecting placement and salary and also understand distribution of database considered.

References :

1. Research paper on “Scales of Measurement and Presentation of Statistical Data”
2. Reseach paper on “Descriptive Statistics and Normality Tests for Statistical Data”
3. Reseach paper on “Statistical data analysis of cancer incidences in insurgency affected states in Nigeria”
4. Research paper on “Statistical data presentation: a primer for rheumatology researchers”
5. Database from kaggle website.

```
from google.colab import drive
drive.mount ('/content/gdrive')
```

```
Mounted at /content/gdrive
```

```
import pandas as pd
```

```
data = pd.read_excel (r'/content/gdrive/My Drive/Colab Notebooks/placement2.xlsx')
```

```
display(pd.DataFrame(data))
```

	sl_no	gender	ssc_p	hsc_p	degree_p	etest_p	mba_p
0	1	M	67.00	91.00	58.00	55.0	58.80
1	2	M	79.33	78.33	77.48	55.0	66.28
2	3	M	65.00	68.00	64.00	75.0	57.80
3	4	M	56.00	52.00	52.00	66.0	59.43
4	5	M	85.80	73.60	73.30	96.8	55.50
...	...	...	...	...	...	...	...
210	211	M	80.60	82.00	77.60	91.0	74.49
211	212	M	58.00	60.00	72.00	74.0	53.62
212	213	M	67.00	67.00	73.00	59.0	69.72
213	214	F	74.00	66.00	58.00	70.0	60.23
214	215	M	62.00	58.00	53.00	89.0	60.22

215 rows × 7 columns

Double-click (or enter) to edit

```
data.describe()
```

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p
count	215.000000	215.000000	215.000000	215.000000	215.000000	215.000000
mean	108.000000	67.303395	66.333163	66.370186	71.954047	62.278186
std	62.209324	10.827205	10.897509	7.358743	13.290111	5.833385

25%	54.500000	60.600000	60.900000	61.000000	60.000000	57.945000
50%	108.000000	67.000000	65.000000	66.000000	70.000000	62.000000
75%	161.500000	75.700000	73.000000	72.000000	83.000000	66.255000
max	215.000000	89.400000	97.700000	91.000000	98.000000	77.890000

```
Q1 = data.ssc_p.quantile (0.25)
Q3 = data.ssc_p.quantile(0.75)
Q1,Q3
```

```
(60.599999999999994, 75.7)
```

```
IQR = Q3 - Q1
IQR
```

```
15.100000000000009
```

```
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
lower_limit, upper_limit
```

```
(37.94999999999998, 98.35000000000002)
```

```
data[(data.ssc_p<lower_limit)|(data.ssc_p>upper_limit)]
```

	sl_no	gender	ssc_p

```
Q1 = data.hsc_p.quantile (0.25)
Q3 = data.hsc_p.quantile(0.75)
Q1,Q3
```

```
(60.9, 73.0)
```

```
IQR = Q3 - Q1
IQR
```

```
12.100000000000001
```

```
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
lower_limit, upper_limit
```

```
(42.75, 91.15)
```

```
data[(data.hsc_p<lower_limit)|(data.hsc_p>upper_limit)]
```

	sl_no	gender	ssc_p	hsc_p	degree_p	etest_p	mba_p
24	25	M	76.50	97.70	78.86	97.40	74.01
42	43	M	49.00	39.00	65.00	63.00	51.21
49	50	F	50.00	37.00	52.00	65.00	56.11
120	121	M	58.00	40.00	59.00	73.00	58.81
134	135	F	77.44	92.00	72.00	94.00	67.13
169	170	M	59.96	42.16	61.26	54.48	65.48
177	178	F	73.00	97.00	79.00	89.00	70.81
206	207	M	41.00	42.00	60.00	97.00	53.39

```
Q1 = data.degree_p.quantile (0.25)
Q3 = data.degree_p.quantile(0.75)
Q1,Q3
```

```
(61.0, 72.0)
```

```
IQR = Q3 - Q1
IQR
```

```
11.0
```

```
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
lower_limit, upper_limit
```

```
(44.5, 88.5)
```

```
data[(data.degree_p<lower_limit)|(data.degree_p>upper_limit)]
```

	sl_no	gender	ssc_p	hsc_p	degree_p	etest_p	mba_p
197	198	F	83.96	53.0	91.0	59.32	69.71

```
Q1 = data.etest_p.quantile (0.25)
Q3 = data.etest_p.quantile(0.75)
```

```
Q1,Q3
```

```
(60.0, 83.0)
```

```
IQR = Q3 - Q1
IQR
```

```
23.0
```

```
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
lower_limit, upper_limit
```

```
(25.5, 117.5)
```

```
data[(data.etest_p<lower_limit)|(data.etest_p>upper_limit)]
```

	sl_no	gender	ssc_p	hsc_p	degree_p	etest_p	mba_p

```
Q1 = data.mba_p.quantile (0.25)
Q3 = data.mba_p.quantile(0.75)
Q1,Q3
```

```
(57.945, 66.255)
```

```
IQR = Q3 - Q1
IQR
```

```
8.309999999999995
```

```
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
lower_limit, upper_limit
```

```
(45.480000000000004, 78.71999999999998)
```

```
data[(data.mba_p<lower_limit)|(data.mba_p>upper_limit)]
```

	sl_no	gender	ssc_p	hsc_p	degree_p	etest_p	mba_p