Unit Test 2

Basic Data Analytic Methods

Shiva Saran
BE-B 20

Q1. Explain following terms:

- Power and Sample size
  The power of a test is the probability of correctly rejecting the null hypothesis. It is denoted by 1- β, where β is the probability of a type II error.

- ANOVA
  Analysis of Variance (ANOVA) is a generalization of the hypothesis testing of the difference of two population means. ANOVA tests if any of the population means differ from the other population means. The null hypothesis of ANOVA is that all the population means are equal.

Q2. When do we use Wilcoxon rank sum test? Write steps in the test

The Wilcoxon rank-sum test is a nonparametric hypothesis test that checks whether two populations are identically distributed. Wilcoxon test does not assume anything about the population distribution, it is generally considered more robust than the t-test. This test is often performed as a two-sided test and, thus, the research hypothesis indicates that the populations are not equal as opposed to specifying directionality.

Steps in the test are:

- Assign ranks by arranging the observations from smallest to largest.
- Summation of ranks in each group
- U test is performed

Q3. Use the above data and group them using K Means Clustering Algorithm. show calculation of centroid.

Sr. No	Height	Weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70

9	183	84
10	180	88
11	180	67
12	177	76

Solution:

Initial K1	185	72
Initial K2	170	56

Euclidean Distance for	K1	K2	New Centroid K1		New Centriod K2	
3	20.81	4.47	185	72	169	58
4	7.21	14.14	182	70	169	58
5	2.00	19.10	182	71	169	58
6	8.49	26.87	185	74	169	58
7	5.83	17.03	183	73	169	58
8	3.54	16.28	181	71	169	58
9	12.87	29.53	182	78	169	58
10	10.59	31.95	181	83	169	58
11	15.85	14.21	181	83	175	63
12	8.06	13.73	179	80	175	63

Grouping								
K1	1	4	5	6	7	8	10	12
K2	2	3	11					

Q4. Explain type 1 and type 2 errors?

Type 1 Error

A type 1 error is also known as a false positive and occurs when a researcher incorrectly rejects a true null hypothesis. This means that your report that your findings are significant when in fact they have occurred by chance.

Type 2 Error

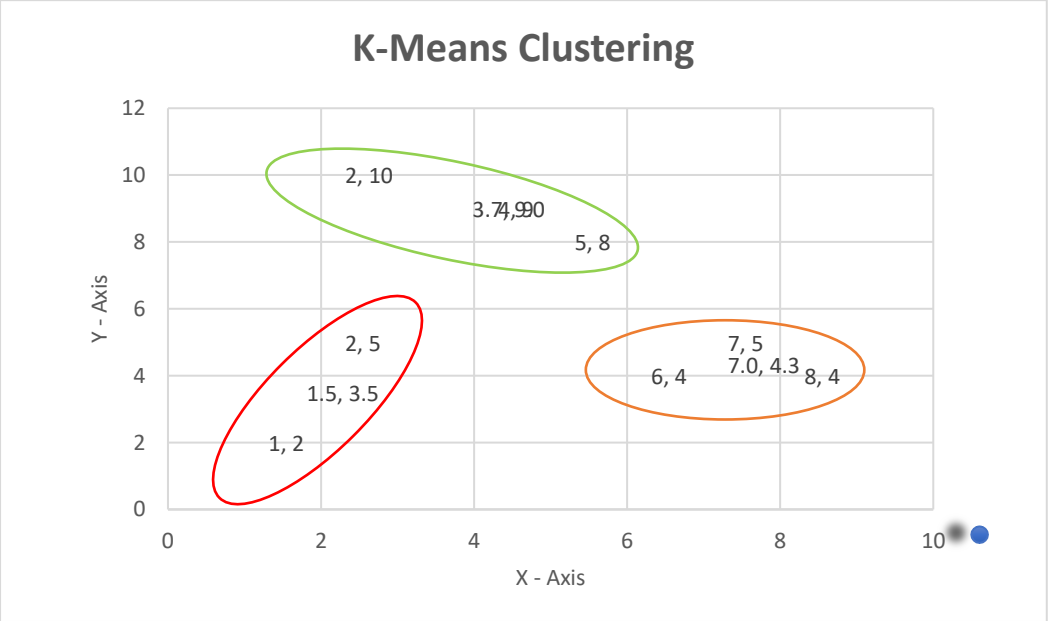
A type II error is also known as a false negative and occurs when a researcher fails to reject a null hypothesis which is really false. Here a researcher concludes there is not a significant effect, when actually there really is.

Q5. Cluster the following eight point (with(X,Y) representing locations)into three cluster:
A1(2,10),A2(2,5),A3(8,4),A4(5,8),A5(7,5),A6(6,4),A7(1,2),A8(4,9)
Note: use Manhattan distance algorithm

Solution:

Using Manhattan distance algorithm

Initial cluster centres are: A1(2,10) , A2(2,5) and A3 (8,4)



Calculations:

			2	10	2	5	8	4	
	Points		**Distance Mean 1**		**Distance Mean 2**		**Distance Mean 3**		**Cluster**
A1	2	10	0		5		12		1
A2	2	5	5		0		7		2
A3	8	4	12		7		0		3
A4	5	8	5		6		7		1
A5	7	5	10		10		2		3
A6	6	4	10		5		2		3
A7	1	2	9		4		9		2
A8	4	9	3		6		9		1

Cluster Grouping:

Cluster 1	A1	A4	A8
Cluster 2	A2	A7	
Cluster 3	A3	A5	A6

New Cluster Centres:

New C 1	3.7	9.0
New C 2	1.5	3.5
New C 3	7.0	4.3