

Statistical Techniques for Data Science

Testing of Hypothesis

Non-Parametric Tests

Objective



After attending this session, you will be able to -

- > Perform Chi-square Test
- **Perform Fisher's Exact Test**
- **Perform Mann-Whitney U-Test**
- **Perform Wilcoxon Signed Rank Test**
- Perform Kruskal-Wallis Test

Non-Parametric Tests



- If it is assumed that the data does not follow any probability distribution (which is characterized by different parameters), then the Non-parametric test is performed
- The following are different non-parametric tests
 - **Chi-square Test**
 - Fisher's Exact Test
 - **Mann-Whitney U-Test**
 - **Wilcoxon Signed Rank Test**
 - Kruskal-Wallis Test

Chi-square test



- > Two different variables are given and we are supposed to understand that whether the two variables are dependent or not
- Chi-square test is used in 2 cases —
- (i) To test whether 2 variables are independent
- (ii) To test Goodness of Fit
- > This test is used only for frequencies- not for probability or percentage

2 x 2 contingency table – Independence Test



Categorical	Categorica	Total	
variable 1	Present	Absent	Total
Present	O ₁ E ₁	O ₂ E ₂	r ₁
Absent	O ₃ E ₃	O ₄ E ₄	r ₂
Total	C ₁	C ₂	n

Calculation of expected frequencies



$$\mathbf{E}_1 = \frac{\mathbf{r}_1 \mathbf{c}_1}{\mathbf{n}}$$

$$\mathbf{E}_2 = \frac{\mathbf{r}_1 \mathbf{c}_2}{\mathbf{n}}$$

$$\chi^{2} = \sum_{i=1}^{k} \frac{(O_{i} - E_{i})^{2}}{E_{i}} \approx \chi^{2}_{(\alpha,(r-1)*(c-1))}$$

$$\mathbf{E}_3 = \frac{\mathbf{r}_2 \mathbf{c}_1}{\mathbf{n}}$$

$$\mathbf{E}_4 = \frac{\mathbf{r}_2 \mathbf{c}_2}{\mathbf{n}}$$



A company has chosen three pension plans. Management wishes to know whether the preference for plans is independent of job classification and wants to use $\alpha = 0.05$. The opinion of a random sample of 500 employees are shown below

	Pen	T . (- 1		
Job classification	1	2	3	Total
Salaried workers	21	36	30	87
Hourly workers	48	26	19	93
Total	69	62	49	180



$$E_{1} = \frac{r_{1}c_{1}}{n} = \frac{87 \times 69}{180} = 33.35$$

$$E_{2} = \frac{r_{1}c_{2}}{n} = \frac{87 \times 62}{180} = 29.97$$

$$E_{3} = \frac{r_{1}c_{3}}{n} = \frac{87 \times 49}{180} = 23.68$$

$$E_{4} = \frac{r_{2}c_{1}}{n} = \frac{93 \times 69}{180} = 35.65$$

$$E_{5} = \frac{r_{2}c_{2}}{n} = \frac{93 \times 62}{180} = 32.03$$

$$E_{6} = \frac{r_{2}c_{3}}{n} = \frac{93 \times 49}{180} = 25.32$$



SI No	(O _i)	(E _i)	(O _i -E _i)	(O _i -E _i) ²	$(O_i-E_i)^2/E_i$
1	21	33.35	- 12.35	152.52	4.57
2	36	29.97	6.03	36.36	1.21
3	30	23.68	6.32	39.94	1.69
4	48	35.65	12.35	152.52	4.28
5	26	32.03	- 6.03	36.36	1.14
6	19	25.32	- 6.32	39.94	1.58
Total	180	180	Chi-squa	re value	14.46



- H₀: Job satisfaction and pension plan are independently distributed
- H₁: Job satisfaction and pension plan are not independently distributed (Associated)
- χ^2 = 14.46 (From table we get corresponding value at 5% level of significance, which is equal to 5.991)
- DF=2
- P<0.001
- Inference: Reject H0, which shows Job satisfaction and pension plan are associated

Chi-square test – Goodness-of-Fit



- In the previous section, we have discussed Ci-square Independence Test
- Goodness-of-Fit test is applied when one categorical variable is available from a single population and the test is used to determine whether the sample data is consistent with a hypothesized distribution
- When to Use the Chi-Square Goodness of Fit Test
 - The chi-square goodness of fit test is appropriate when the following conditions are met:
 - The sampling method is <u>simple random sampling</u>.
 - The variable under study is <u>categorical</u>.
- ➤ The expected value of the number of sample observations in each <u>level</u> of the variable is at least 5
- This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan,
 (3) analyze sample data, and (4) interpret results
- Hypotheses –

H₀: The data are consistent with a specified distribution.

H_a: The data are *not* consistent with a specified distribution

For the test statistic is a chi-square random variable (X^2) defined by the following equation - $X^2 = \Sigma \left[(O_i - E_i)^2 / E_i \right]$ (degrees of freedom k-1 where k is number of groups)

Chi-square test – Goodness-of-Fit



Example – A computer programmer has developed an algorithm for generating 5 first 5
 alphabets at random and the code has given the following result when ran for 500 times –

Alphabets	А	В	С	D	E
Frequency	104	112	102	94	88

Is there evidence that there is good-fit to show that random alphabet generator is working correctly? Use $\alpha = 0.05$.

Solution – Calculated value of Chi-square based on data is –

 $X^2 = \Sigma [(O_i - E_i)^2 / E_i] = 3.44$ (Degrees of freedom is 4 as 5 categories are there)

Alphabet	0	E	(O-E)	(O-E)^2	(O-E)^2/E
Α	104	100	4	16	0.16
В	112	100	12	144	1.44
С	102	100	2	4	0.04
D	94	100	-6	36	0.36
E	88	100	-12	144	1.44
		Chi-square	Value	Sum	3.44

Tabulated value of Chi-square with 4 degrees of freedom is 9.487

As the calculated value is less than tabulated value, we will accept the null hypothesis

Mann Whitney U test



Mann Whitney U test:

nonparametric equivalent of a t test for two independent samples





Mann Whitney U test:

$$U_1 = (n_1)(n_2) + \frac{n_1(n_1+1)}{2} - \sum R_1$$

$$U_2 = (n_1)(n_2) + \frac{n_2(n_2+1)}{2} - \sum R_2$$

Where: n_1 Size of sample one

 n_2 Size of sample two





Mann Whitney U test:

$$U_1 = (n_1)(n_2) + \frac{n_1(n_1+1)}{2} - \sum R_1$$

$$U_2 = (n_1)(n_2) + \frac{n_2(n_2+1)}{2} - \sum R_2$$

Where:

 $\sum R_1$

Sum of sample one ranks

 $\sum R_2$

Sum of sample two ranks







1) Choose the smaller of the two U values.

2) Find the critical value (Mann Whitney table)

3) When computed value is *smaller* than the critical value the outcome is significant! (i.e., the null hypothesis is to be rejected)



24 28

18 42

45 63

57 57

12 90

30 68





Step One: Rank all data across groups



group 1		group 2
24		28
18	2	42
45		63
57		57
12	1	90





24 3 28 4

18 2 42

45 63

57 57

12 1 90

30 68





24 3 28 4

18 2 42 6

45 7 63

57 57

12 1 90

30 5 68





Tied ranks:

- Find all values that are tied.
- Identify all ranks that would be assigned to those values.
- Average those ranks.
- Assign that average to all tied values.





24 3 28 4

18 2 42 6

45 7 63

57 57

12 1 90

30 5 68







$$8+9 = 17$$
 Averaging $17/2 = 8.5$ ranks





24 3 28 4

18 2 42 6

45 7 63

57 8.5 57 8.5

12 1 90

30 5 68





24 3 28 4

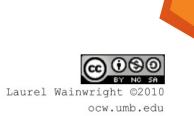
18 2 42 6

45 7 63 10

57 8.5 57 8.5

12 1 90 12

30 5 68 11





Step Two: Sum the ranks for each group

group 1

group 2

8.5

8.5

26.5

51.5





Check the rankings:

$$\sum R = \frac{n(n+1)}{2}$$





$$\sum R = \frac{(12)(13)}{2}$$

$$\sum R = \frac{156}{2}$$

$$\sum R = 78$$





24 3 28 4

18 2 42 6

45 7 63 10

57 8.5 57 8.5

12 1 90 12

30 5 68 11

26.5 51.5









Laurel Wainwright @2010

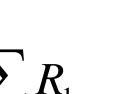
ocw.umb.edu



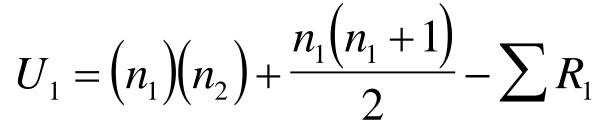
Step Three: Compute U₁

$$U_1 = (n_1)(n_2) + \frac{n_1(n_1+1)}{2} - \sum R_1$$









$$U_1 = (6)(6) + \frac{6(7)}{2} - 26.5$$

$$U_1 = 36 + 21 - 26.5$$

$$U_1 = 30.5$$





Step Four: Compute U₂

$$U_2 = (n_1)(n_2) + \frac{n_2(n_2+1)}{2} - \sum R_2$$





$$U_2 = (n_1)(n_2) + \frac{n_2(n_2+1)}{2} - \sum R_2$$

$$U_2 = (6)(6) + \frac{6(7)}{2} - 51.5$$

$$U_2 = 36 + 21 - 51.5$$

$$U_2 = 5.5$$









$$U_1 = 30.5$$

$$U_2 = 5.5$$

$$U = 5.5$$







Critical Value = 5
This is a nonsignificant outcome



Wilcoxon Signed rank test



- Denote the before and after observation by X and Y
- Find the difference between X and Y
- Ignore the sign of the difference and rank the difference with rank 1 for smaller difference, rank 2 for next smaller difference, so on and rank n for the larger difference. Assign average rank for the tied values in the difference
- Attach the original sign to the ranks assigned to the difference
- Find the sum of the positive ranks and negative ranks separately. Choose the minimum of sum of positive and negative ranks.
- If the calculated value is more than the critical value the null hypothesis in not rejected, otherwise
 it is rejected

Example on Wilcoxon signed rank test

Serum fibronogen degradation product values (µgm/ml) of a group of 12 persons

SI No	Before	After	Difference	Ranks for magnitude	Ranks
1	5.0	7.8	- 2.8	2	- 2
2	10.0	180.0	- 170.0	11	- 11
3	18.0	10.0	8.0	5	+ 5
4	5.0	80.0	- 75.0	10	- 10
5	10.0	15.0	- 5.0	3.5	- 3. <mark>5</mark>
6	20.0	10.0	10.0	6	+ 6
7	5.0	180.0	- 175.0	12	- 12
8	2.5	40.0	- 37.5	8	- 8
9	15.0	10.0	5.0	3.5	+ 3.5
10	10.0	7.5	2.5	1	+ 1
11	80.0	10.0	70.0	9	+ 9
12	5.0	20.0	- 15.0	7	-7

manipalglobal



Sum of (+) ranks = 24.5

Sum of (-) ranks = 53.5

For 12 pairs, a minimum rank sum of less than or equal to 14 is required for rejection of the null hypothesis at 5% level.

Since the calculated rank sum 24.5 is more than 14, the null hypothesis, that the pre and post operative values of F.D.P. is not significantly different is accepted at P>0.05.

Kruskal-Wallis test



- Denote the k samples as G₁, G₂, G₃, ..., G_k
- Denote the size of each of the samples as n₁, n₂, n₃, ...,n_k
- = $n=n_1+n_2+n_3+...+n_k$
- Combine the data, keeping track of the sample from which each datum arose
- Rank the data in such a way the lowest value with rank 1, next value with rank 2 so on and the highest value with rank n. If there are tied value assign the average rank
- Separate the groups along with their respective ranks, add the ranks of each sample separately, naming the sums R₁, R₂, R₃,...,R_k

Kruskal-Wallis test



Calculate the test-statistic given by

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

will be distributed as Chi-square with k-1 degrees of freedom, where R_i is the sum of the rank of the ith group and n_i is the number of observation in the ith group

 If the calculated value is more than the critical value the null hypothesis in rejected, otherwise it is not rejected

Example on Kruskal-Wallis test

_	manipa	lgloba
---	--------	--------

Acad	emy	O†	Data	Scienc	

врн	Positive	Negative
	biopsy	biopsy
5.3	7.1	11.4
7.9	6.6	0.5
8.7	6.5	1.6
4.3	14.8	2.3
6.6	17.3	3.1
6.4	3.4	1.4
	13.4	4.4
	7.6	5.1



Example on Kruskal-Wallis test



Sc	ie	n	C	e
-			_	

SI No	Group	PSA (ng/ml)	Ranks
16	Negative biopsy	0.5	1
20	Negative biopsy	1.4	2
17	Negative biopsy	1.6	3
18	Negative biopsy	2.3	4
19	Negative biopsy	3.1	5
12	Positive biopsy	3.4	6
4	BPH	4.3	7
21	Negative biopsy	4.4	8
22	Negative biopsy	5.1	9



Science

SI No	Group	PSA (ng/ml)	Ranks
1	BPH	5.3	10
6	BPH	6.4	11
9	Positive biopsy	6.5	12
5	BPH	6.6	13.5
8	Positive biopsy	6.6	13.5
7	Positive biopsy	7.1	15
14	Positive biopsy	7.6	16
2	BPH	7.9	17
3	BPH	8.7	18



SI No	Group	PSA (ng/ml)	Ranks
15	Negative biopsy	11.4	19
13	Positive biopsy	13.4	20
10	Positive biopsy	14.8	21
11	Positive biopsy	17.3	22

					-	manipalglo	
	BPH		Positive k	piopsy	Negative	biopsy Academy of Data Sc	cience
PSA (n	g/ml)	Rank	PSA (ng/ml)	Rank	PSA (ng/ml)	Rank	
5.3	3	10	7.1	15	11.4	19	
7.9		17	6.6	13.5	0.5		
8.7	7	18	6.5	12	1.6	3	
4.3		7	14.8	21	2.3	4	
6.6		13.5	17.3	22	3.1	5	
6.4		11	3.4	6	1.4	2	
			13.4	20	4.4	8	
			7.6	16	5.1	9	

125.5

76.5

Total

51



n = total no. of observations = 6+8+8=22

$$H = \frac{12}{22(22+1)} \left[\frac{76.5^2}{6} + \frac{125.5^2}{8} + \frac{51^2}{8} \right] - 3(22+1)$$

$$= 8.53$$

H has Chi-square distribution with 2 degrees of freedom. The critical values for 2 degrees of freedom is 5.99 and the null hypothesis of equality of medians is rejected and alternative hypothesis is accepted.







Copyright Manipal Global Education Services Pvt. Ltd. All Rights Reserved.

All product and company names used or referred to in this work are trademarks or registered trademarks of their respective holders.

Use of them in this work does not imply any affiliation with or endorsement by them.

This work contains a variety of copyrighted material. Some of this is the intellectual property of Manipal Global Education, some material is owned by others which is clearly indicated, and other material may be in the public domain. Except for material which is unambiguously and unarguably in the public domain, permission is not given for any commercial use or sale of this work or any portion or component hereof. No part of this work (except as legally allowed for private use and study) may be reproduced, adapted, or further disseminated without the express and written permission of Manipal Global Education or the legal holder of copyright, as the case may be.





