# Statistical Inference

Non-Parametric Tests II

### Contents

- 1. Kruskal Wallis test
- 2. Chi-square test of association

#### Kruskal Wallis test

- The Kruskal Wallis test is considered as nonparametric alternative to one way analysis of variance (ANOVA).
- The Kruskal Wallis test is used to compare differences between more than two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.
- H0: K samples come from the same population H1: Not H0.

### Kruskal Wallis test procedure

- Combine all the observations from k samples into a single sample of size n and arrange them in ascending order.
- Assign ranks to them from smallest to largest as 1 to n. if there is a tie at two or more places, each observation is given the mean of the ranks for which it is tied.
- The ranks assigned to observations in each of the k groups are added separately to give k rank sums.
- The test statistic is

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

 $n_j = number of observations in j^{th} sample$ 

n = number of observations in the combined sample

 $R_j = sum \ of \ the \ ranks \ in \ the \ j^{th} \ sample.$ 

H follows Chi Square Distribution with k-1 df

## Case Study - 1

To execute Non-Parametric test in Python, we shall consider the below case as an example.

#### **Background**

Data consist of aptitude score of 3 groups of employees.

#### **Objective**

To check whether there is difference in the score among three groups.

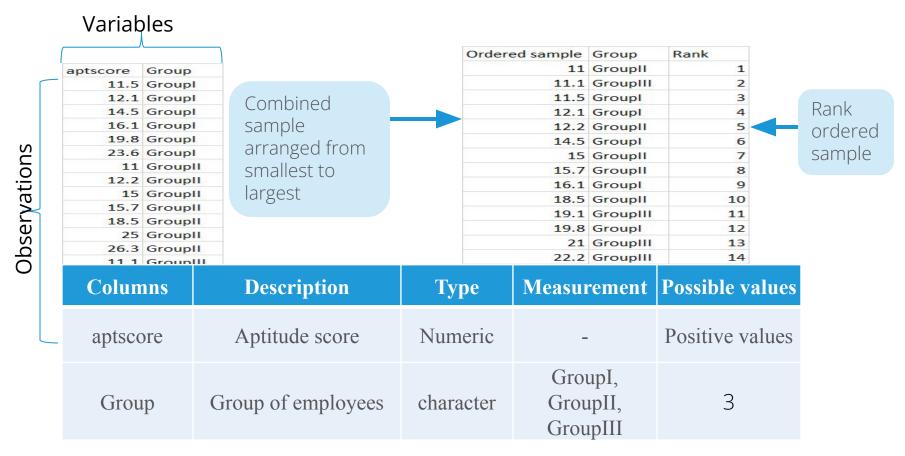
#### Sample Size

Sample size: 20

Variables: aptscore, Group

## Data Snapshot

#### Kruskal Wallis Test



### Kruskal Wallis test

Testing distribution of more than two samples

Objective

To test the **null hypothesis** that all the samples came from same population

Null Hypothesis ( $H_0$ ): The three samples are from the same population Alternate Hypothesis ( $H_1$ ): The three samples do not come from the same population

Test Statistic	$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_{j}^{2}}{n_{j}} - 3(n+1) \begin{cases} n_{j} = \text{number of observations in } j^{th} \text{ sample} \\ n = \text{number of observations in the combined sample} \end{cases}$ $R_{j} = \text{sum of the ranks in the } j^{th} \text{ sample}.$	
Decision Criteria	Reject the null hypothesis <b>if p-value &lt; 0.05</b>	

# Kruskal Wallis test example

#### Calculations:

	Value
Sample size	$n_1 = 6$ $n_2 = 7$ $n_3 = 7$
$R_1$	50
$R_2$	68
$R_3$	92
Н	2.2309
p-value	0.3278

### Kruskal Wallis test in Python

```
# Import the CSV file
import pandas as pd
data = pd.read_csv('Kruskal Wallis Test.csv')

# Kruskal wallis test

from scipy.stats import kruskal

group1 = data[data['Group'] == 'GroupI']['aptscore']
group2 = data[data['Group'] == 'GroupII']['aptscore']
group3 = data[data['Group'] == 'GroupIII']['aptscore']
kruskal(group1, group2, group3) +
```

- kruskal from scipy.stats performs the Kruskal wallis test on the data.
- aptscore is the analysis variable.
- Group is the factor variable.

## Kruskal Wallis test in Python

#### # Output:

KruskalResult(statistic=2.230929090974231, pvalue=0.3277629827136111)

#### Interpretation:

 Since p-value is >0.05, do not reject H0. Aptitude score is same for all three groups of employees.

## Chi-square test of Association

- The chi-square test for independence, also called as Pearson's chi-square test or the chi-square test of association, is used to test if there is a relationship between two categorical variables.
- The two categorical variables can be nominal or ordinal.
- H0: Two attributes are independent (not associated)

H1: Not H0.

### Chi-square test procedure

- Assume that there are 'r' categories of attribute A and 'c' categories of attribute B. Therefore, we have a cross table of r\*c (r rows and c columns).
- Let Ri be the total of ith row and Cj be the total of jth column.
- Observed frequencies are calculated from the data. Oij: Observed frequency in ith row and jth column.
- Expected frequencies are given by Eij = (Ri \* Cj)/ n where n is total sample size. Expected frequencies are computed under null hypothesis.
- Test statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where Oij are the observed frequencies in the ith row and jth column. Eij are the expected frequencies in the ith row and jth column.

•  $\chi^2$  follows a Chi-Square Distribution with (r-1)(c-1) degrees of freedom.

## Case Study - 2

To execute Non-Parametric test in Python, we shall consider the below case as an example.

#### **Background**

Data consist of information regarding the Performance & Recruitment Source of employees.

#### **Objective**

To check whether Performance & Source of Recruitment are associated.

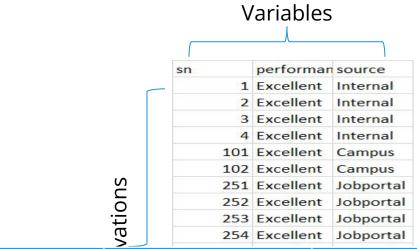
#### Sample Size

Sample size: 870

Variables: sn, performance, source

## Data Snapshot

#### chi square test of association



Columns	Description	Type	Measurement	<b>Possible values</b>
sn	Serial number	Numeric	-	-
performance	Employee performance	Character	Excellent, Good,Poor	3
source	Source of recruitment	Character	Campus, Internal, Jobportal	3

• Get the observed frequency (count) table from this data.

# Chi-square test of Association

Testing association between two categorical variables

Objective

To test the **null hypothesis** that two categorical variables are **independent** 

Null Hypothesis ( $H_0$ ): performance and source are not associated Alternate Hypothesis ( $H_1$ ): performance and source are associated

Test Statistic	$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ Oij = observed frequencies in the ith and jth column. Eij = expected frequencies in the ith and jth column.		
Decision Criteria	Reject the null hypothesis <b>if p-value &lt; 0.05</b>		

# Chi-square test example

#### Observed Frequency table

	Recruitment Source			
Performance	Campus	Internal	Jobportal	Total
Excellent	150	100	40	290
Good	100	100	100	300
Poor	80	50	150	280
Total	330	250	290	870

#### • Expected Frequency table

		Recruitment Source		
Performance	Campus	Internal	Jobportal	Total
Excellent	=(330*290)/870	83	97	290
Good	114	=(250*300)/870	100	300
Poor	106	80	=(290*280)/870	280
Total	330	250	290	870

	Value
r	3
С	3
$\chi^2$	107.3786

## Chi-Square test in Python

```
# Import the CSV file
data = pd.read_csv('chi square test of association.csv')
# create cross table of 2 categorical
cont table = pd.crosstab(data.performance, data.source)
# Chi-square test of association
from scipy.stats import chi2_contingency
chi2_contingency(cont table)
                              chi2_contingency from
                              scipy.stats function performs
                              Chi-square test of association.
                              It returns chi2(test statistic), p
                              valve, dof, expected
                              frequencies.
```

### Chi-Square test in Python

#### # Output:

```
(107.37856396477088,

2.6359873347121296e-22,

4,

array([[110. , 83.33333333, 96.66666667],

        [113.79310345, 86.20689655, 100. ],

        [106.20689655, 80.45977011, 93.33333333]]))
```

#### Interpretation:

 Since p-value is <0.05, reject H0. Recruitment source and employee performance are associated.

### Quick Recap

In this session, we continued learning non parametric tests. Here is a quick recap:

Kruskal Wallis test

Nonparametric alternative to one way ANOVA.

Chi-Square test

 Also called as Pearson's chi-square test or the chi-square test of association, is used to test if there is a relationship between two categorical variables (nominal or ordinal).