# STK121 – Chapter 18

Learning outcomes:

- Be able to understand, explain and execute the hypothesis test about the difference of more than two populations based on independent samples.
- Know what the requirements for the sample size is in order to do a Chi-square approximation.
- Be able to calculate a p-value by hand and by using EXCEL's CHI.DIST.RT function

## 2.4.3 - Kruskal-Wallis test

### Previously:

- We used ANOVA parametric test for more than two independent samples
- **H0:** μ1 = μ2 = μ3 … μk **Ha:** At least one pair of means is not equal to each other
- Parameters were known

### Currently:

- Three or more random samples and parameters unknown
- Kruskal-Wallis test is based on the analysis of independent random samples from each of K populations.
- Can be ordinal or quantitative data and does not require the assumption that the populations have normal distribution
- This is a Hypothesis test about the difference between more than two populations based on k-independent samples where k = the number of populations
-
- The ranks of the data and not the observations itself is used to compare the populations.
- The process is done as follows:
  - All the data are combined into one group, sorted and ranked.
  - Determine the sum of ranks for the separate groups.
- Kruskal-Wallis test is always a UPPER TAIL TEST

## Steps:

1) Hypothesis test
2) Use ranks of data to compare K populations
3) Combine the samples and award ranks to data points. Deal with the ties by awarding average ranks
4) Determine separate sample RANKS
5) Determine Test the statistic H
6) Determine P-value
7) Rejection rule
8) Conclusion

## Hypothesis:

- H0: All population means are identical
- Ha: Not all population means are identical

## Test statistic:

- When the populations are identical, the sampling distribution of the test statistic H can be approximated by a chi-square distribution with k – 1 degrees of freedom.
- This approximation is acceptable if each of the sample sizes $n_i \geq 5$
- $H = \left[\frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} \frac{Ri^2}{n_i}\right] - 3(n_T + 1)$
- $k = number\ of\ populations$
- $n_i = number\ of\ observations\ in\ smaple\ i$
- $n_T = total\ number\ of\ observations\ in\ all\ samples$
- $Ri = sum\ of\ ranks\ for\ sample\ i$

## Rejection rule:

- Reject H0 if p-value ≤ alpha

## Example 1:

A company employ people from three different colleges A, B and C. The employees get an annual performance evaluation rating and management want to see if the people from the three colleges perform equally or not. A sample of 20 employees, 7 from College A, 6 from college B and 7 from college C is drawn at random and the results are below.

| College A | College B | College C |
|---|---|---|
| 25 | 60 | 50 |
| 70 | 20 | 70 |
| 60 | 30 | 60 |
| 85 | 15 | 80 |
| 95 | 40 | 90 |
| 90 | 35 | 70 |
| 80 | | 75 |

- $n_1; n_2; n_3 \geq 5$ , $H \sim \chi^2$ (2)

## Step 1: Hypothesis test

- H0: All population means are identical
- Ha: Not all population means are identical

## Step 2 and 3: Ranking

| Data | College | Rank |
|---|---|---|
| 15 | B | 1 |
| 20 | B | 2 |
| 25 | A | 3 |
| 30 | B | 4 |
| 35 | B | 5 |
| 40 | B | 6 |
| 50 | C | 7 |
| 60 | C | 9 |
| 60 | B | 9 |
| 60 | A | 9 |
| 70 | C | 12 |
| 70 | A | 12 |
| 70 | C | 12 |
| 75 | C | 14 |
| 80 | C | 15.5 |
| 80 | A | 15.5 |
| 85 | A | 17 |
| 90 | A | 18.5 |
| 90 | C | 18.50 |
| 95 | A | 20 |

**8 + 9 + 10 / 3 = 9**

**11+ 12 + 13 /3 = 12**

**18 + 19 / 2 = 18.50**

## Step 4: Determine separate sample RANKS

| ni | College A | Rank | College B | Rank | College C | Rank |
|---|---|---|---|---|---|---|
| **1** | **25** | **3** | **60** | **9** | **50** | **7** |
| **2** | **70** | **12** | **20** | **2** | **70** | **12** |
| **3** | **60** | **9** | **30** | **4** | **60** | **9** |
| **4** | **85** | **17** | **15** | **1** | **80** | **15.5** |
| **5** | **95** | **20** | **40** | **6** | **90** | **18.5** |
| **6** | **90** | **18.5** | **35** | **5** | **70** | **12** |
| **7** | **80** | **15.5** | | | **75** | **14** |
| | **Sum** | **95** | **Sum** | **27** | **Sum** | **88** |

## Step 5: Determine test stat

- $H = \left[\frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} \frac{Ri^2}{n_i}\right] - 3(n_T + 1)$
- k = 3 populations
- n1 = 7, n2 = 6, n3 = 7, nT = 20
- $H = \left[\frac{12}{20(20+1)} \sum_{i=1}^{3} \frac{(95)^2}{7} + \frac{(27)^2}{6} + \frac{(88)^2}{7}\right] - 3(20 + 1) = 8.9163$

## Step 6 and 7: Determine P-value and state rejection rule

- Reject H0 if p-value < 0.05
- 0.0116 < P-value < 0.025

OR

- Degrees of freedom (3-1) = 2
- Alpha = 0.05
- $\chi^2_{(2;0.05)} = 5.991$
- H > $\chi^2_{(2;0.05)}$ – Reject H0
- 8.9163 > 5.991 - Reject H0

## Step 8: Conclusion

- Conclusion Reject H0. Using excel will show that P-value for H = 8.92 is 0.0116.
- As P-value is < α = .05, we reject H0.
- Not all populations are identical.
- College B perform worse than A and C since RANK sum total is only 27.
- There is insufficient evidence to conclude that the populations are identical.

## Example 2:

A well-know car magazine has decided to compare the average fuel consumption of midsize cars between four manufacturers. A random selection of 20 cars (new0 were selected and the were subjected to identical road tests. The fuel consumption in $liter\ /\ 100km$ is given below.

| | | | | | |
|---|---|---|---|---|---|
| Manufacturer 1 | 8.8 | 8.9 | 9.3 | 8.9 | 8.7 |
| Manufacturer 2 | 9.6 | 9.9 | 9.1 | 9.3 | 9.7 |
| Manufacturer 3 | 9.7 | 8.3 | 8.9 | 9.1 | 8.7 |
| Manufacturer 4 | 9.2 | 9.5 | 9.7 | 9.5 | 9.8 |

What can we conclude regarding the fuel consumption of the manufacturers, are they the same or not?

- $n_1 = n_2 = n_3 = n_4 = 5$ , $H \sim \chi^2(3)$

## Step 1: Hypothesis test

- H0: All four population **are identical** with regard to fuel consumption
- Ha: All four populations are **not identical** with regards to fuel consumption

## Step 2 and 3: Ranking

| manufacturer | Consumption | Ranks |
|---|---|---|
| 3 | 8.3 | 1 |
| 3 | 8.7 | 2.5 |
| 1 | 8.7 | 2.5 |
| 1 | 8.8 | 4 |
| 1 | 8.9 | 6 |
| 1 | 8.9 | 6 |
| 3 | 8.9 | 6 |
| 2 | 9.1 | 8.5 |
| 3 | 9.1 | 8.5 |
| 4 | 9.2 | 10 |
| 1 | 9.3 | 11.5 |
| 2 | 9.3 | 11.5 |
| 4 | 9.5 | 13.5 |
| 4 | 9.5 | 13.5 |
| 2 | 9.6 | 15 |
| 2 | 9.7 | 17 |
| 3 | 9.7 | 17 |
| 4 | 9.7 | 17 |
| 4 | 9.8 | 19 |
| 2 | 9.9 | 20 |
| | | |

## Step 4: Determine separate sample RANKS

Sum of the ranks:

| | | |
|---|---|---|
| Manufacturer 1 | $R_1$ | 8.7 + 8.8 + 8.9 + 8.9 + 9.3 = 30 |
| Manufacturer 2 | $R_2$ | 9.1 + 9.3 + 9.6 + 9.7 + 9.9 = 72 |
| Manufacturer 3 | $R_3$ | 8.3 + 8.7 + 8.9 + 9.1 + 9.7 = 35 |
| Manufacturer 4 | $R_4$ | 9.2 + 9.5 + 9.5 + 9.7 +9.7 = 73 |

## Step 5: Determine test stat

- $H = \left[\frac{12}{n_T(n_T+1)} \sum_{i=1}^{k} \frac{Ri^2}{n_i}\right] - 3(n_T + 1)$
- $H = \left[\frac{12}{(20)(20+1)} \sum_{i=1}^{k} \frac{(30)^2}{5} + \frac{(72)^2}{5} + \frac{(35)^2}{5} + \frac{(73)^2}{5}\right] - 3(20+1) = 9.2171$

1 2 3

Do the test stat in three parts - see highlighted parts

## Step 6 and 7: Determine P-value and state rejection rule

- Reject H0 if p-value < alpha = 0.05
- 0.025 < P-value < 0.05

OR

- Degrees of freedom (4-1) = 3
- Alpha = 0.05
- $\chi^2_{(3;0.05)} = 7.8147$
- H > $\chi^2_{(2;0.05)}$ – Reject H0
- 9.2171 > 7.8147 - Reject H0

## Step 8: Conclusion

- All four populations are NOT IDENTICAL with regard to fuel consumption

## YouTube videos:

1) https://www.youtube.com/watch?v=agtKm_SBIPw
2) https://www.youtube.com/watch?v=JJ6RKqRUK10
3) https://www.youtube.com/watch?v=q1D4Di1KWLc
4) https://www.youtube.com/watch?v=At0NMw7HWMw