Kruskal Wallis Test

2025-01-27

date: “2025-01-27” Presented by Group 1 students 1. Ma Cho Cho Yee (MAS 1 - 8) 2. Mg Kaung Myat Thu (MAS 1 - 13) 3. Mg Kyaw Min Khaing (MAS 1 - 19) 4. Mg Kyaw Swar Tun (MAS 1 - 20) 5. Mg Nay Aung (MAS 1 - 30) 6. Mg Nay Myo Aung (MAS 1 - 31) 7. Mg Soe Eain Htet (MAS 1 - 34) 8. Mg Thet Paing Htet (MAS 1 - 39) 9. Mg Wai Linn Htet Thwin (MAS 1 - 43) 10 Mg Yan Naing Soe (MAS 1 - 51) —

Kruskal-Wallis H-test statistic using the formula:

Where: - is the total number of observations. - is the rank sum for each group. - is the number of observations in each group.

Assumption-

1. non-parametric statistical test
2. three or more independent groups
3. if there are statistically significant differences between them ,it is an extension of the Mann\_Whitney U test which is used for comparing two groups.

In this 1974 Motor Trend US magazine , is there a significant difference between Cylinder Count (cyl) and Horsepower (hp)? Assess the normality assumption for the data? Use the Kruskal-Wallis Test to evaluate whether there are statistically significant differences in Horsepower across the groups (4-cylinder, 6-cylinder, 8-cylinder). Present the test results using a boxplot and appropriately annotate the p-value in your visualization.

## State the Null and Research Hypotheses

### Alternative Hypothesis (H1):is HO: θL = θM = θH (All groups have the same median).

### The research hypothesis is HA: There is a tendency for self-conidence to rank systematically higher or lower for at least one level of social interaction when compared with the other levels. (at least one group differs in its median).

### Set the Level of Risk (or the Level of Signiicance) Associated with the Null Hypothesis α = 0.05, 95% chance that any observed statistical difference will be real and not due to chance

### Load necessary packages

library(pacman)  
p\_load(rmarkdown,officer,knitr,kableExtra,latex2exp,xtable,flextable,tidyverse, rstatix, ggsignif)

### Convert knit format

### check my data type

mtcars <- as.tibble(mtcars)  
glimpse(mtcars)

## Rows: 32  
## Columns: 11  
## $ mpg <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8,…  
## $ cyl <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4, 8,…  
## $ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8, 16…  
## $ hp <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 180…  
## $ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92,…  
## $ wt <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150, 3.…  
## $ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90, 18…  
## $ vs <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0,…  
## $ am <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,…  
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3,…  
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2,…

### show my data

mtcars <- as.tibble(mtcars)  
kable(mtcars, caption = "mtcars Dataset")

mtcars Dataset

| mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 21.0 | 6 | 160.0 | 110 | 3.90 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| 21.0 | 6 | 160.0 | 110 | 3.90 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |
| 22.8 | 4 | 108.0 | 93 | 3.85 | 2.320 | 18.61 | 1 | 1 | 4 | 1 |
| 21.4 | 6 | 258.0 | 110 | 3.08 | 3.215 | 19.44 | 1 | 0 | 3 | 1 |
| 18.7 | 8 | 360.0 | 175 | 3.15 | 3.440 | 17.02 | 0 | 0 | 3 | 2 |
| 18.1 | 6 | 225.0 | 105 | 2.76 | 3.460 | 20.22 | 1 | 0 | 3 | 1 |
| 14.3 | 8 | 360.0 | 245 | 3.21 | 3.570 | 15.84 | 0 | 0 | 3 | 4 |
| 24.4 | 4 | 146.7 | 62 | 3.69 | 3.190 | 20.00 | 1 | 0 | 4 | 2 |
| 22.8 | 4 | 140.8 | 95 | 3.92 | 3.150 | 22.90 | 1 | 0 | 4 | 2 |
| 19.2 | 6 | 167.6 | 123 | 3.92 | 3.440 | 18.30 | 1 | 0 | 4 | 4 |
| 17.8 | 6 | 167.6 | 123 | 3.92 | 3.440 | 18.90 | 1 | 0 | 4 | 4 |
| 16.4 | 8 | 275.8 | 180 | 3.07 | 4.070 | 17.40 | 0 | 0 | 3 | 3 |
| 17.3 | 8 | 275.8 | 180 | 3.07 | 3.730 | 17.60 | 0 | 0 | 3 | 3 |
| 15.2 | 8 | 275.8 | 180 | 3.07 | 3.780 | 18.00 | 0 | 0 | 3 | 3 |
| 10.4 | 8 | 472.0 | 205 | 2.93 | 5.250 | 17.98 | 0 | 0 | 3 | 4 |
| 10.4 | 8 | 460.0 | 215 | 3.00 | 5.424 | 17.82 | 0 | 0 | 3 | 4 |
| 14.7 | 8 | 440.0 | 230 | 3.23 | 5.345 | 17.42 | 0 | 0 | 3 | 4 |
| 32.4 | 4 | 78.7 | 66 | 4.08 | 2.200 | 19.47 | 1 | 1 | 4 | 1 |
| 30.4 | 4 | 75.7 | 52 | 4.93 | 1.615 | 18.52 | 1 | 1 | 4 | 2 |
| 33.9 | 4 | 71.1 | 65 | 4.22 | 1.835 | 19.90 | 1 | 1 | 4 | 1 |
| 21.5 | 4 | 120.1 | 97 | 3.70 | 2.465 | 20.01 | 1 | 0 | 3 | 1 |
| 15.5 | 8 | 318.0 | 150 | 2.76 | 3.520 | 16.87 | 0 | 0 | 3 | 2 |
| 15.2 | 8 | 304.0 | 150 | 3.15 | 3.435 | 17.30 | 0 | 0 | 3 | 2 |
| 13.3 | 8 | 350.0 | 245 | 3.73 | 3.840 | 15.41 | 0 | 0 | 3 | 4 |
| 19.2 | 8 | 400.0 | 175 | 3.08 | 3.845 | 17.05 | 0 | 0 | 3 | 2 |
| 27.3 | 4 | 79.0 | 66 | 4.08 | 1.935 | 18.90 | 1 | 1 | 4 | 1 |
| 26.0 | 4 | 120.3 | 91 | 4.43 | 2.140 | 16.70 | 0 | 1 | 5 | 2 |
| 30.4 | 4 | 95.1 | 113 | 3.77 | 1.513 | 16.90 | 1 | 1 | 5 | 2 |
| 15.8 | 8 | 351.0 | 264 | 4.22 | 3.170 | 14.50 | 0 | 1 | 5 | 4 |
| 19.7 | 6 | 145.0 | 175 | 3.62 | 2.770 | 15.50 | 0 | 1 | 5 | 6 |
| 15.0 | 8 | 301.0 | 335 | 3.54 | 3.570 | 14.60 | 0 | 1 | 5 | 8 |
| 21.4 | 4 | 121.0 | 109 | 4.11 | 2.780 | 18.60 | 1 | 1 | 4 | 2 |

### Convert to long format(tidy) for analysis and data manipulation

data\_long <- mtcars |>   
 select(hp, cyl) |>   
 mutate(cyl = as.factor(cyl)) |>   
 pivot\_longer(cols = hp, names\_to = "Metric", values\_to = "Value")   
kable(data\_long, caption = "data cleaning and wrangling")

data cleaning and wrangling

| cyl | Metric | Value |
| --- | --- | --- |
| 6 | hp | 110 |
| 6 | hp | 110 |
| 4 | hp | 93 |
| 6 | hp | 110 |
| 8 | hp | 175 |
| 6 | hp | 105 |
| 8 | hp | 245 |
| 4 | hp | 62 |
| 4 | hp | 95 |
| 6 | hp | 123 |
| 6 | hp | 123 |
| 8 | hp | 180 |
| 8 | hp | 180 |
| 8 | hp | 180 |
| 8 | hp | 205 |
| 8 | hp | 215 |
| 8 | hp | 230 |
| 4 | hp | 66 |
| 4 | hp | 52 |
| 4 | hp | 65 |
| 4 | hp | 97 |
| 8 | hp | 150 |
| 8 | hp | 150 |
| 8 | hp | 245 |
| 8 | hp | 175 |
| 4 | hp | 66 |
| 4 | hp | 91 |
| 4 | hp | 113 |
| 8 | hp | 264 |
| 6 | hp | 175 |
| 8 | hp | 335 |
| 4 | hp | 109 |

### calculate rank and handling Ties for the Test Statistic doing Ties Correction (C\_H)

ranked\_data <- data\_long |>   
 mutate(Rank = rank(Value),   
 Tie\_Case = duplicated(Value) | duplicated(Value, fromLast = TRUE)) |>   
 arrange(Rank) # Rank အလိုက် စဉ်သည်  
  
kable(ranked\_data, caption = "sorting,Ranking and Tie")

sorting,Ranking and Tie

| cyl | Metric | Value | Rank | Tie\_Case |
| --- | --- | --- | --- | --- |
| 4 | hp | 52 | 1.0 | FALSE |
| 4 | hp | 62 | 2.0 | FALSE |
| 4 | hp | 65 | 3.0 | FALSE |
| 4 | hp | 66 | 4.5 | TRUE |
| 4 | hp | 66 | 4.5 | TRUE |
| 4 | hp | 91 | 6.0 | FALSE |
| 4 | hp | 93 | 7.0 | FALSE |
| 4 | hp | 95 | 8.0 | FALSE |
| 4 | hp | 97 | 9.0 | FALSE |
| 6 | hp | 105 | 10.0 | FALSE |
| 4 | hp | 109 | 11.0 | FALSE |
| 6 | hp | 110 | 13.0 | TRUE |
| 6 | hp | 110 | 13.0 | TRUE |
| 6 | hp | 110 | 13.0 | TRUE |
| 4 | hp | 113 | 15.0 | FALSE |
| 6 | hp | 123 | 16.5 | TRUE |
| 6 | hp | 123 | 16.5 | TRUE |
| 8 | hp | 150 | 18.5 | TRUE |
| 8 | hp | 150 | 18.5 | TRUE |
| 8 | hp | 175 | 21.0 | TRUE |
| 8 | hp | 175 | 21.0 | TRUE |
| 6 | hp | 175 | 21.0 | TRUE |
| 8 | hp | 180 | 24.0 | TRUE |
| 8 | hp | 180 | 24.0 | TRUE |
| 8 | hp | 180 | 24.0 | TRUE |
| 8 | hp | 205 | 26.0 | FALSE |
| 8 | hp | 215 | 27.0 | FALSE |
| 8 | hp | 230 | 28.0 | FALSE |
| 8 | hp | 245 | 29.5 | TRUE |
| 8 | hp | 245 | 29.5 | TRUE |
| 8 | hp | 264 | 31.0 | FALSE |
| 8 | hp | 335 | 32.0 | FALSE |

### Total Rank Variability for hp

rank\_sums <- ranked\_data |>   
 group\_by(cyl) |> # Ensure cyl exists  
 summarise(  
 n = n(),  
 Rank\_Sum = sum(Rank)  
 )  
  
# Print rank sums  
kable(rank\_sums, caption = "total Rank")

total Rank

| cyl | n | Rank\_Sum |
| --- | --- | --- |
| 4 | 11 | 71 |
| 6 | 7 | 103 |
| 8 | 14 | 354 |

### total number of observation

N <- nrow(ranked\_data)  
  
cat("Total number of observation (N):", N, "\n")

## Total number of observation (N): 32

### number of category

k <- n\_distinct(ranked\_data$cyl)  
cat(" Number of category (N):", k, "\n")

## Number of category (N): 3

## Compute the Kruskal-Wallis H-test statistic using the formula:

Where: - is the total number of observations. - is the rank sum for each group. - is the number of observations in each group.

df <- k - 1  
  
H <- 12 / (N \* (N + 1)) \* sum((rank\_sums$Rank\_Sum^2) / rank\_sums$n) - 3 \* (N + 1)  
p\_value <- pchisq(H, df, lower.tail = FALSE)  
  
cat("Kruskal-Wallis H statistic:", round(H, 4), "\n")

## Kruskal-Wallis H statistic: 25.1476

cat("Degrees of Freedom:", df, "\n")

## Degrees of Freedom: 2

cat("p-value:", round(p\_value, 6), "\n")

## p-value: 3e-06

if (p\_value < 0.05) {  
 cat("The result is significant (p < 0.05). At least one group differs.\n")  
} else {  
 cat("The result is not significant (p >= 0.05). Groups are similar.\n")  
}

## The result is significant (p < 0.05). At least one group differs.

## Compute ties table based on Tie\_Case

ties\_table <- ranked\_data |>   
 filter(Tie\_Case) |>   
 count(Rank)  
  
# Print the updated ties table  
kable(ties\_table, caption = "Number of Tie")

Number of Tie

| Rank | n |
| --- | --- |
| 4.5 | 2 |
| 13.0 | 3 |
| 16.5 | 2 |
| 18.5 | 2 |
| 21.0 | 3 |
| 24.0 | 3 |
| 29.5 | 2 |

# Compute C\_H correction factor

## To Compute the Test Statistic doing Ties Correction (C\_H)

The ties correction is applied to adjust for tied ranks. The formula for the ties correction is:

Where: - is the number of tied ranks in each group of ties. - is the total number of observations.

# Compute C\_H correction factor  
C\_H <- 1 - sum((ties\_table$n^3 - ties\_table$n) / ((nrow(ranked\_data)^3) - nrow(ranked\_data)))  
  
# Print C\_H  
cat("Ties Correction Factor (C\_H):", round(C\_H, 4), "\n")

## Ties Correction Factor (C\_H): 0.9971

Calculate H statistic

H\_corrected <- H / C\_H   
cat("Corrected H Statistic:", round(H\_corrected, 4), "\n")

## Corrected H Statistic: 25.2215

# **Pairwise Wilcoxon Rank Sum Test with Bonferroni Correction**

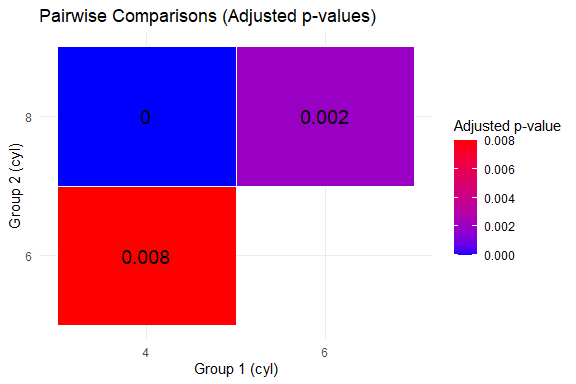
pairwise\_results <- ranked\_data |>   
 pairwise\_wilcox\_test(Value ~ cyl, p.adjust.method = "bonferroni") |>   
 mutate(p = round(p, 6), p.adj = round(p.adj, 3))  
  
kable(pairwise\_results, caption = "pairwise results")

pairwise results

| .y. | group1 | group2 | n1 | n2 | statistic | p | p.adj | p.adj.signif |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Value | 4 | 6 | 11 | 7 | 5 | 0.003000 | 0.008 | \*\* |
| Value | 4 | 8 | 11 | 14 | 0 | 0.000027 | 0.000 | \*\*\*\* |
| Value | 6 | 8 | 7 | 14 | 3 | 0.000647 | 0.002 | \*\* |

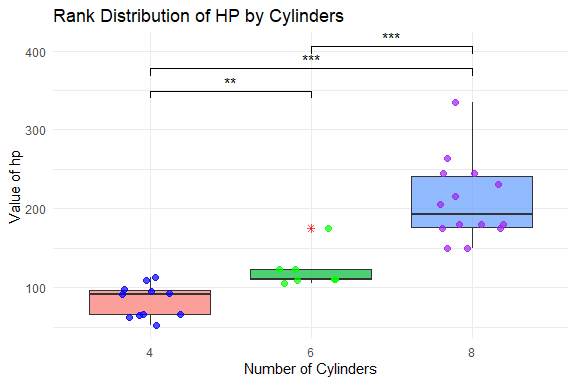
# Visualizing pairwise comparisons with exact p-values with heatmap

filtered\_results <- pairwise\_results |>   
 filter((group1 == "4" & group2 == "6") |   
 (group1 == "4" & group2 == "8") |   
 (group1 == "6" & group2 == "8"))  
ggplot(filtered\_results, aes(x = group1, y = group2, fill = p.adj)) +  
 geom\_tile(color = "white") + # Heatmap style  
 geom\_text(aes(label = round(p.adj, 3)), color = "black", size = 5) + # Adjusted p-values (3 digits)  
 scale\_fill\_gradient(low = "blue", high = "red") + # Color scale for p-values  
 theme\_minimal() +  
 labs(title = "Pairwise Comparisons (Adjusted p-values)",   
 x = "Group 1 (cyl)",   
 y = "Group 2 (cyl)",   
 fill = "Adjusted p-value") +  
 theme\_minimal()



visturalization boxplot

rank\_plot <- ggplot(data\_long , aes(x = cyl, y = Value, fill = cyl)) +  
 geom\_boxplot(outlier.color = "red", outlier.shape = 8, alpha = 0.7) + # Transparent boxplot  
 geom\_jitter(aes(color = cyl), width = 0.2, size = 2, alpha = 0.7) + # Jitter points  
 labs(  
 title = "Rank Distribution of HP by Cylinders",  
 x = "Number of Cylinders",  
 y = "Value of hp"  
 ) +  
 theme\_minimal() +  
 geom\_signif(  
 comparisons = list(  
 c("4", "6"),  
 c("4", "8"),  
 c("6", "8")  
 ),  
 map\_signif\_level = TRUE,  
 test = "wilcox.test",  
 step\_increase = 0.1  
 ) +  
 scale\_color\_manual(values = c("4" = "blue", "6" = "green", "8" = "purple")) + # Custom colors  
 theme(legend.position = "none") # Hide legend  
  
print(rank\_plot)



# Perform Kruskal-Wallis Test

rstatix\_in\_kruskal\_result <- data\_long |>   
 kruskal\_test(Value ~ cyl)  
kable(rstatix\_in\_kruskal\_result, caption = "Rstatix in kruskal result")

Rstatix in kruskal result

| .y. | n | statistic | df | p | method |
| --- | --- | --- | --- | --- | --- |
| Value | 32 | 25.22154 | 2 | 3.3e-06 | Kruskal-Wallis |

# Perform pairwise Wilcoxon comparisons with Bonferroni correction

rstatix\_in\_pairwise\_results <- data\_long |>   
 pairwise\_wilcox\_test(Value ~ cyl, p.adjust.method = "bonferroni") |>   
 mutate(p = round(p, 6), p.adj = round(p.adj, 3))  
kable(rstatix\_in\_pairwise\_results, caption = "rstatix\_in\_pairwise\_results")

rstatix\_in\_pairwise\_results

| .y. | group1 | group2 | n1 | n2 | statistic | p | p.adj | p.adj.signif |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Value | 4 | 6 | 11 | 7 | 5 | 0.003000 | 0.008 | \*\* |
| Value | 4 | 8 | 11 | 14 | 0 | 0.000027 | 0.000 | \*\*\*\* |
| Value | 6 | 8 | 7 | 14 | 3 | 0.000647 | 0.002 | \*\* |