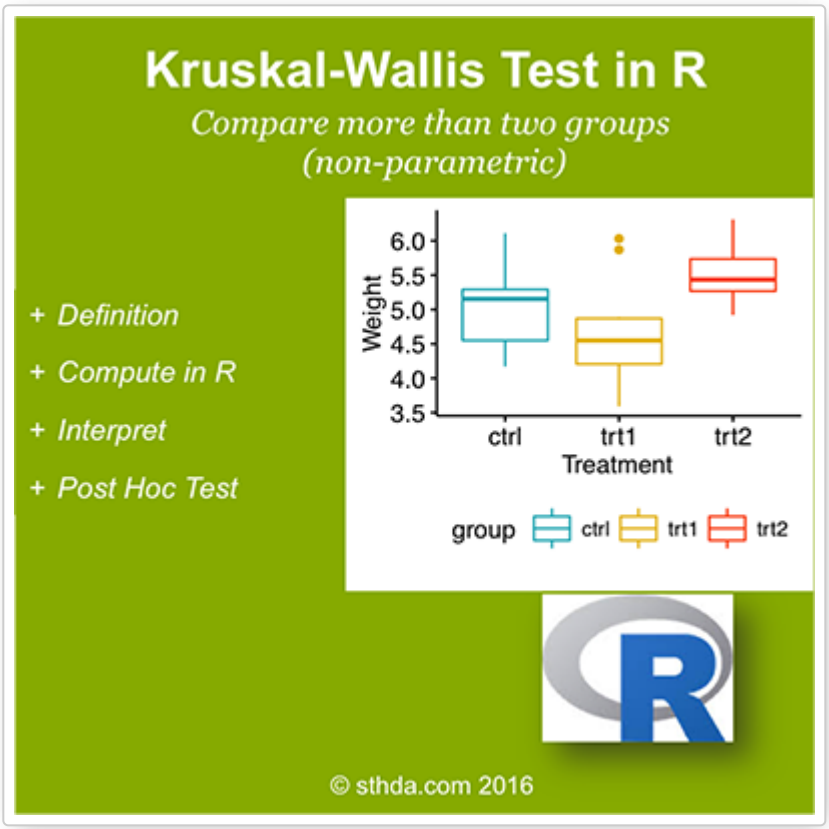
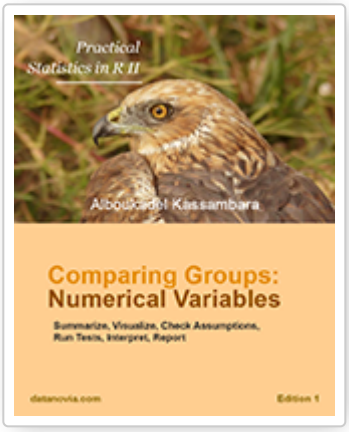
Kruskal-Wallis Test in R

What is Kruskal-Wallis test?

Kruskal-Wallis test by rank is a **non-parametric alternative** to one-way **ANOVA test**, which extends the two-samples Wilcoxon test in the situation where there are more than two groups. It's recommended when the assumptions of one-way ANOVA test are not met. This tutorial describes how to compute Kruskal-Wallis test in **R** software.



Related Book:



Practical Statistics in R for Comparing Groups: Numerical Variables

Visualize your data and compute Kruskal-Wallis test in R

Import your data into R

1. **Prepare your data** as specified here: Best practices for preparing your data set for R

2. **Save your data** in an external .txt tab or .csv files
3. **Import your data into R** as follow:

```
# If .txt tab file, use this
my_data <- read.delim(file.choose())
# Or, if .csv file, use this
my_data <- read.csv(file.choose())
```

Here, we'll use the built-in R data set named *PlantGrowth*. It contains the weight of plants obtained under a control and two different treatment conditions.

```
my_data <- PlantGrowth
```

Check your data

```
# print the head of the file
head(my_data)
```

```
  weight group
1   4.17  ctrl
2   5.58  ctrl
3   5.18  ctrl
4   6.11  ctrl
5   4.50  ctrl
6   4.61  ctrl
```

> ! In R terminology, the column "group" is called factor and the different categories ("ctr", "trt1", "trt2") are named factor levels. **The levels are ordered alphabetically**.

```
# Show the group levels
levels(my_data$group)
```

```
[1] "ctrl" "trt1" "trt2"
```

If the levels are not automatically in the correct order, re-order them as follow:

```
my_data$group <- ordered(my_data$group,
                         levels = c("ctrl", "trt1", "trt2"))
```

It's possible to compute summary statistics by groups. The dplyr package can be used.

- To install **dplyr** package, type this:

```
install.packages("dplyr")
```

- Compute summary statistics by groups:

```
library(dplyr)
group_by(my_data, group) %>%
  summarise(
    count = n(),
    mean = mean(weight, na.rm = TRUE),
    sd = sd(weight, na.rm = TRUE),
    median = median(weight, na.rm = TRUE),
    IQR = IQR(weight, na.rm = TRUE)
  )
```

```
Source: local data frame [3 x 6]
   group count  mean        sd median    IQR
  (fctr) (int) (dbl)     (dbl)  (dbl)  (dbl)
1   ctrl    10 5.032 0.5830914  5.155 0.7425
2   trt1    10 4.661 0.7936757  4.550 0.6625
3   trt2    10 5.526 0.4425733  5.435 0.4675
```

Visualize the data using box plots

- To use R base graphs read this: R base graphs. Here, we'll use the **ggpubr** R package for an easy ggplot2-based data visualization.
- Install the latest version of ggpubr from GitHub as follow (recommended):

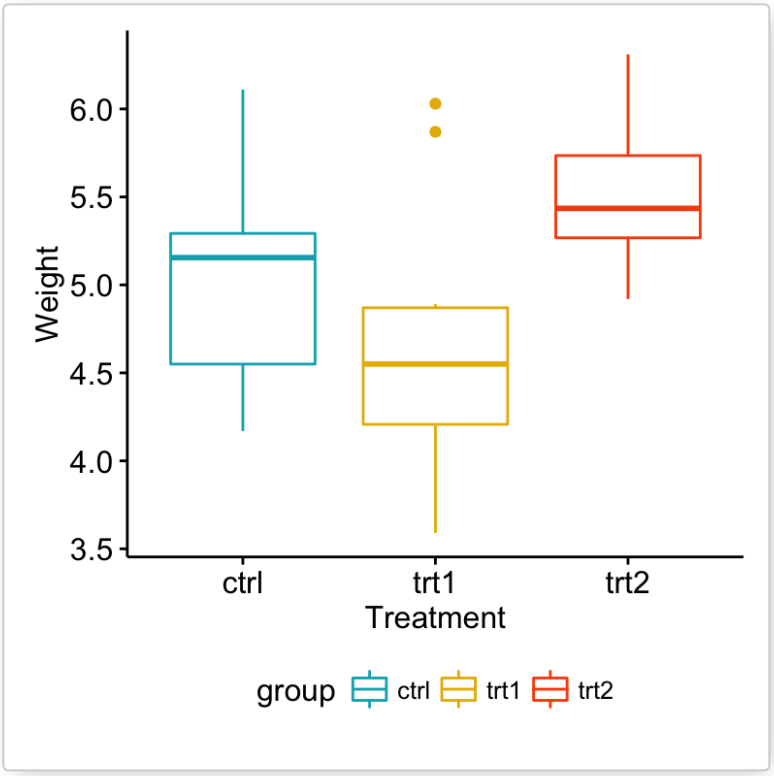
```
# Install
if(!require(devtools)) install.packages("devtools")
devtools::install_github("kassambara/ggpubr")
```

- Or, install from CRAN as follow:

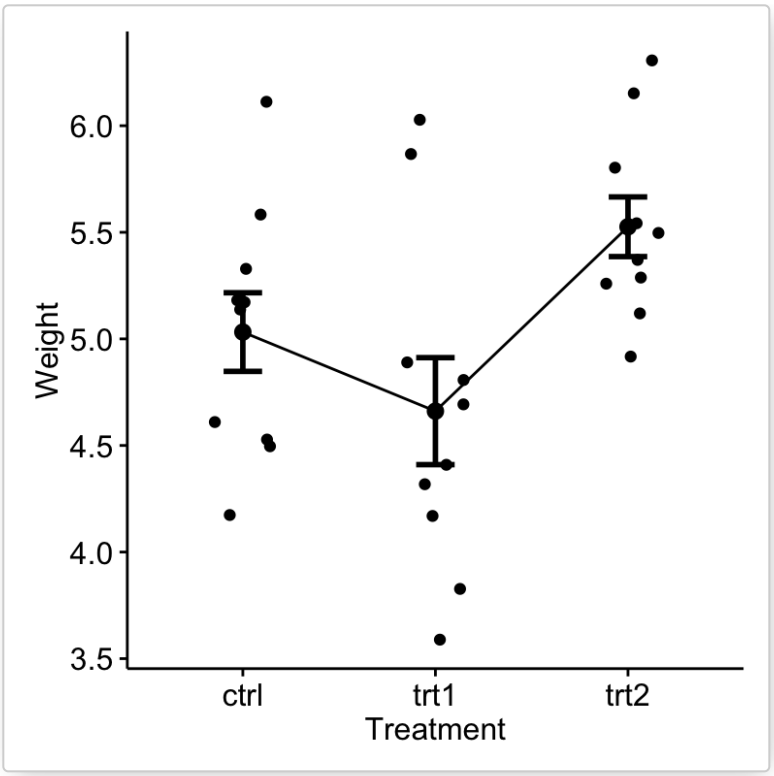
```
install.packages("ggpubr")
```

- Visualize your data with ggpubr:

```
# Box plots
# ++++++++++++++++++++
# Plot weight by group and color by group
library("ggpubr")
ggboxplot(my_data, x = "group", y = "weight",
          color = "group", palette = c("#00AFBB", "#E7B800", "#FC4E07"),
          order = c("ctrl", "trt1", "trt2"),
          ylab = "Weight", xlab = "Treatment")
```




```
# Mean plots
# ++++++++++++++++++++
# Plot weight by group
# Add error bars: mean_se
# (other values include: mean_sd, mean_ci, median_iqr, ....)
library("ggpubr")
ggline(my_data, x = "group", y = "weight",
       add = c("mean_se", "jitter"),
       order = c("ctrl", "trt1", "trt2"),
       ylab = "Weight", xlab = "Treatment")
```



Compute Kruskal-Wallis test

? We want to know if there is any significant difference between the average weights of plants in the 3 experimental conditions.

The test can be performed using the function **kruskal.test**() as follow:

```
kruskal.test(weight ~ group, data = my_data)
```

```
    Kruskal-Wallis rank sum test
data:  weight by group
Kruskal-Wallis chi-squared = 7.9882, df = 2, p-value = 0.01842
```

Interpret

As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the treatment groups.

Multiple pairwise-comparison between groups

From the output of the Kruskal-Wallis test, we know that there is a significant difference between groups, but we don't know which pairs of groups are different.

It's possible to use the function **pairwise.wilcox.test**() to calculate pairwise comparisons between group levels with corrections for multiple testing.

```
pairwise.wilcox.test(PlantGrowth$weight, PlantGrowth$group,
                 p.adjust.method = "BH")
```

```
    Pairwise comparisons using Wilcoxon rank sum test
data:  PlantGrowth$weight and PlantGrowth$group
     ctrl  trt1
trt1 0.199 -
trt2 0.095 0.027
P value adjustment method: BH
```

The pairwise comparison shows that, only trt1 and trt2 are significantly different ($p < 0.05$).

See also

- Analysis of variance (ANOVA, parametric):
  - One-Way ANOVA Test in R
  - Two-Way ANOVA Test in R
  - MANOVA Test in R: Multivariate Analysis of Variance

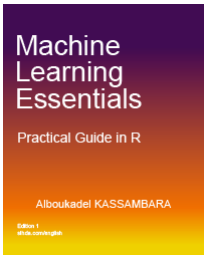
Infos

This analysis has been performed using **R software** (ver. 3.2.4).

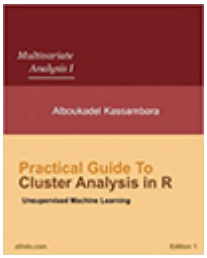
Enjoyed this article? I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

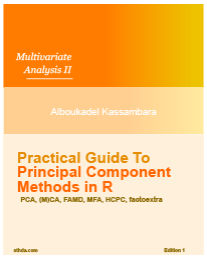
Recommended for You!



Machine Learning Essentials: Practical Guide in R

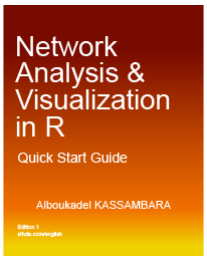


Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R

R Graphics Essentials for Great Data Visualization



Network Analysis and Visualization in R

More books on R and data science

Recommended for you

This section contains best data science and self-development resources to help you on your path.

Coursera - Online Courses and Specialization

Data science

- Course: Machine Learning: Master the Fundamentals by Standford
- Specialization: Data Science by Johns Hopkins University
- Specialization: Python for Everybody by University of Michigan
- Courses: Build Skills for a Top Job in any Industry by Coursera
- Specialization: Master Machine Learning Fundamentals by University of Washington
- Specialization: Statistics with R by Duke University
- Specialization: Software Development in R by Johns Hopkins University
- Specialization: Genomic Data Science by Johns Hopkins University

Popular Courses Launched in 2020

- Google IT Automation with Python by Google
- AI for Medicine by deeplearning.ai
- Epidemiology in Public Health Practice by Johns Hopkins University
- AWS Fundamentals by Amazon Web Services

Trending Courses

- The Science of Well-Being by Yale University
- Google IT Support Professional by Google
- Python for Everybody by University of Michigan
- IBM Data Science Professional Certificate by IBM
- Business Foundations by University of Pennsylvania
- Introduction to Psychology by Yale University
- Excel Skills for Business by Macquarie University
- Psychological First Aid by Johns Hopkins University
- Graphic Design by Cal Arts

Books - Data Science

Our Books

- Practical Guide to Cluster Analysis in R by A. Kassambara (Datanovia)
- Practical Guide To Principal Component Methods in R by A. Kassambara (Datanovia)
- Machine Learning Essentials: Practical Guide in R by A. Kassambara (Datanovia)
- R Graphics Essentials for Great Data Visualization by A. Kassambara (Datanovia)
- GGPlot2 Essentials for Great Data Visualization in R by A. Kassambara (Datanovia)
- Network Analysis and Visualization in R by A. Kassambara (Datanovia)
- Practical Statistics in R for Comparing Groups: Numerical Variables by A. Kassambara (Datanovia)
- Inter-Rater Reliability Essentials: Practical Guide in R by A. Kassambara (Datanovia)

Others

- R for Data Science: Import, Tidy, Transform, Visualize, and Model Data by Hadley Wickham & Garrett Grolemund
- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems by Aurelien Géron
- Practical Statistics for Data Scientists: 50 Essential Concepts by Peter Bruce & Andrew Bruce
- Hands-On Programming with R: Write Your Own Functions And Simulations by Garrett Grolemund & Hadley Wickham
- An Introduction to Statistical Learning: with Applications in R by Gareth James et al.
- Deep Learning with R by François Chollet & J.J. Allaire
- Deep Learning with Python by François Chollet

Get involved :

Click to **follow us** on Facebook and Google+ :

Comment this article by clicking on "Discussion" button (top-right position of this page)

Sign in

Login

Login

Password

Password

Auto connect

Sign in

Register

Forgotten password

Welcome!

Want to Learn More on R Programming and Data Science?

Follow us by Email

Subscribe

by FeedBurner

on Social Networks

Click to see our collection of resources to help you on your path...

Course & Specialization

Recommended for You (on Coursera):

- Course: Machine Learning: Master the Fundamentals
- Specialization: Data Science
- Specialization: Python for Everybody
- Course: Build Skills for a Top Job in any Industry
- Specialization: Master Machine Learning Fundamentals
- Specialization: Statistics with R
- Specialization: Software Development in R
- Specialization: Genomic Data Science

See More Resources

---

factoextra

survminer

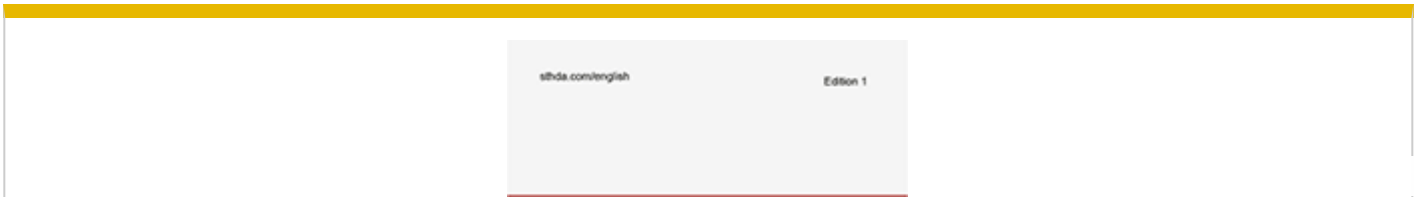
ggpubr

ggcorrplot

fastqcr

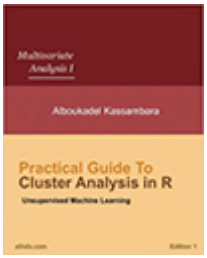
Our Books

3D Plots in R

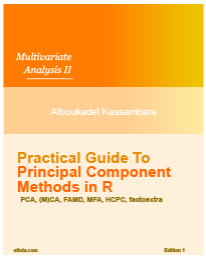


R Graphics Essentials for Great Data Visualization: 200 Practical Examples You Want to Know for Data Science

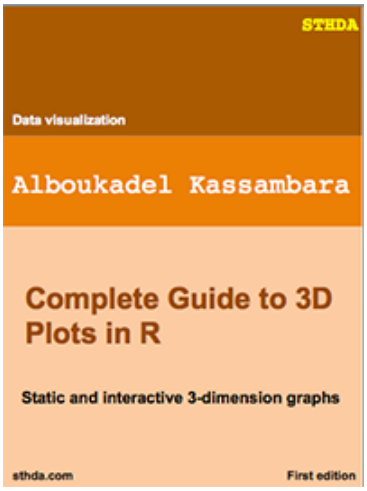
★ NEW!!



Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R



Datanovia: Online Data Science Courses

R-Bloggers

Recommended for you

Reading Data From TXT|CSV Files: R Base Functions - Easy Guid...

www.sthda.com

ggplot2 histogram plot : Quick start guide - R software and data vis...

www.sthda.com

ggplot2 - Easy Way to Mix Multiple Graphs on The Same Page - Articl...

www.sthda.com