Guide to statistical methods in R

Pauline Busch

2024-08-16

Table of Contents

Note: For demonstration purposes I am using the a package with 19 medical datasets for teaching reproducible medical research with R. The package is called “medicaldata” and needs to be installed if you would like to reproduce the results shown in this document in your R-script.

## Install ggstatsplot

If not done already install the package called ggstatsplot. For this copy the code down below into the console and press enter. You might need to install other packages as well. Do this in the same fashion using the install.packages() function.

## install.packages("ggstatsplot")

## Load the required R libraries

There are many nice libraries that support statistics in R. I recommend the following:

library(ggstatsplot)

In addition to that, we want to load, manipulate and visualize our data from excel:

library(readxl)  
library(openxlsx)  
library(dplyr)  
library(car)  
library(ggplot2)  
library(ggthemes)  
library(ggpubr)  
library(ggsci)  
library(patchwork)

## Choosing the right statistical method

### Data preparation

The first step is to identify whether your data is categorical or continuous. This will influence your choice of statistical methods.

* Categorical data: Data that represent categories (e.g., gender, color).
* Continuous data: Data that represent measurements on a continuous scale (e.g., height, weight).

Furthermore, you should check for outliers and missing data and decide how you want to handle these.

### Testing assumptions

Two common assumptions are normality and homogeneity of variances. The following section demonstrates how to do this in R.

#### Normality

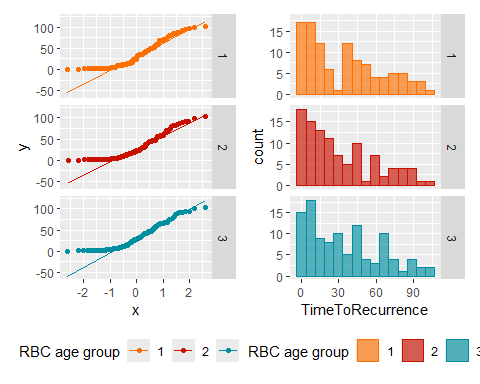
Parametric statistical tests assume that the data are normally distributed.Use the following tests to assess normality:

* **Shapiro-Wilk Test:** A formal test for normality.  
  Let’s test whether our data is normally distributed. The following code shows an example on how to conduct a Shapiro-Wilk Test on the previously loaded “medicaldata”. For comprehension: We want to understand whether RBC storage duration and biochemical prostate cancer recurrence after radical prostatectomy are associated. RBC age is categorized in three groups (1-3) from “younger” to “older” depending on storage time (see: RBC.Age.Group). Cancer recurrence can be found in the last column (TimeToRecurrence).
  + Normality is assumed when p > 0.05.
  + Note: If the Shapiro-Wilk test is performed on individual groups separately, it’s possible that some groups show normality (p > 0.05), while others do not (p < 0.05). This suggests that the assumption of normality for ANOVA may be violated.

grouped\_data <- split(blood$TimeToRecurrence, blood$RBC.Age.Group) #data preparation  
shapiro\_results <- lapply(grouped\_data, shapiro.test) #lapply() applies the shapiro-function to all groups  
  
# without lappy() its more complicated: shapiro\_results <- shapiro.test(grouped\_data[["1"]])

* **Q-Q Plot and Histogram:** A visual methods to check if your data follows a normal distribution. The following code shows how you can visually check the distribution of your data. This is not necessary but might help to comprehend the results of the Shapiro-Wilk test and get a feeling for your data.

# How to make a Q-Q Plot  
QQ\_plot <- ggplot(data = blood,   
 aes(sample = TimeToRecurrence,   
 color = factor(RBC.Age.Group))) +   
 stat\_qq() +  
 stat\_qq\_line() +  
 facet\_grid(vars(RBC.Age.Group)) +  
 scale\_color\_futurama(name = "RBC age group")  
  
# How to make a Histogram  
Histogram <- ggplot(data = blood,   
 aes(x = TimeToRecurrence,   
 fill = factor(RBC.Age.Group),   
 color = factor(RBC.Age.Group))) +  
   
 geom\_histogram(bins = 15) +  
 facet\_grid(vars(RBC.Age.Group)) +  
 scale\_fill\_futurama(alpha = 0.65, name = "RBC age group") +  
 scale\_color\_futurama(name = "RBC age group")   
   
# Combine Q-Q Plot and Histogram  
combined\_plot <- QQ\_plot + Histogram +   
 plot\_layout(ncol = 2) & theme(legend.position = "bottom")  
  
print(combined\_plot)



#### Homogeneity of variances

Some tests require that the variances between groups are equal. Use Levene’s Test to check for equality of variances.

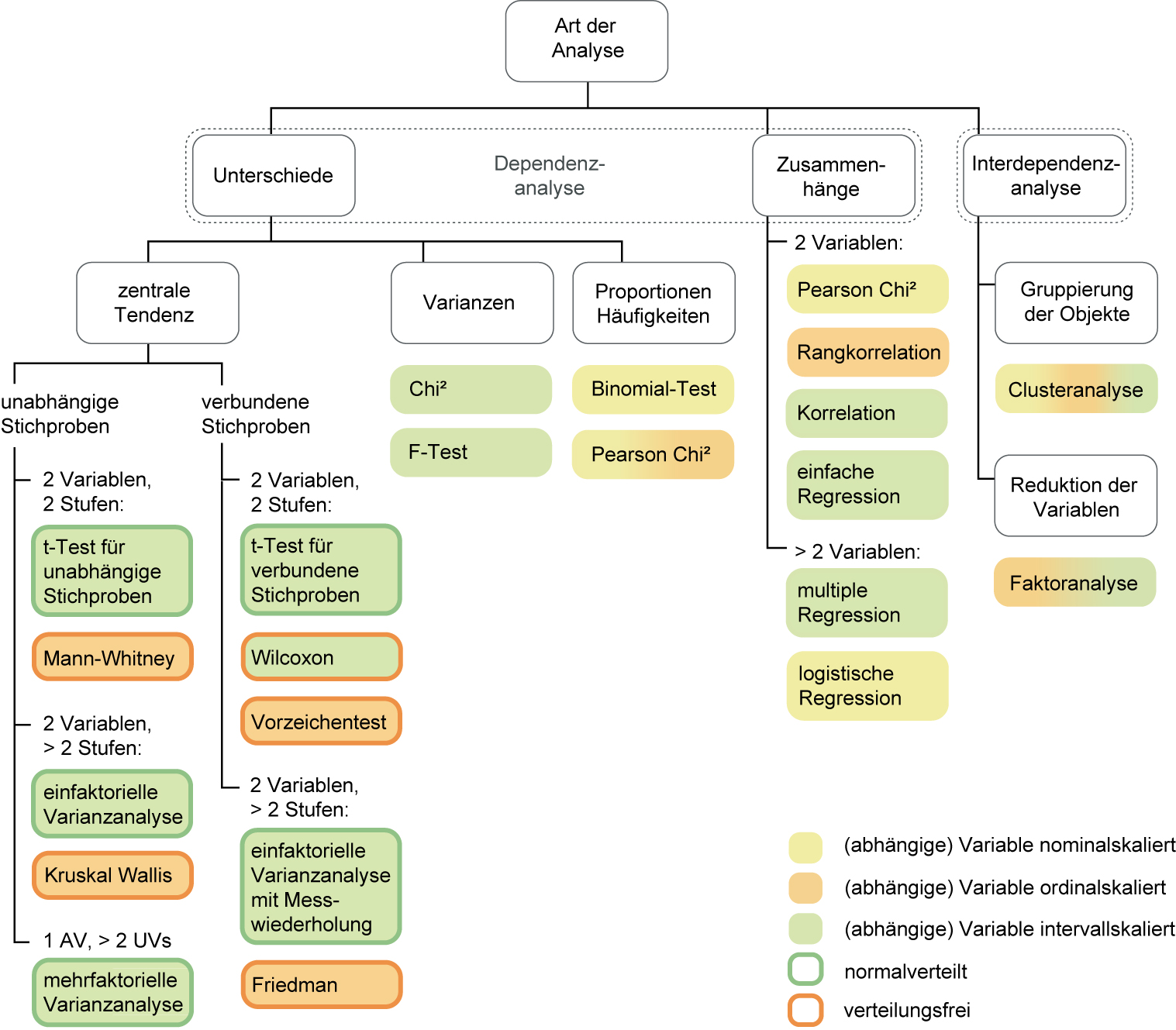
* **Levene’s Test:** A tests that compares the variances between groups.  
  The following code demonstrates how to conduct the Levene’s test in R.
  + Homogeneity of variances is assumed when ‘Pr(>F)’ (p-value) > 0.05.

levene\_results <- leveneTest(TimeToRecurrence ~ factor(RBC.Age.Group), data = blood)

### The approporiate statistical test for your research question

Once you’ve checked the assumptions, the next step is to choose the appropriate statistical test based on your data type and research question. The decision tree in the next chapter can help with that. If you are not sure or need help you can ask ChatGPT to assist you with this (e.g., first prompt: “Can you find me a suitable statistical method to analyse my experiment?” - then fill out all the information ChatGPT asks from you).

### Decision tree



## Statistical methods

### Kruskal-Wallis Test

## Post-hoc analysis

### Dunn’s Test