Statistics – Seminar 1

Group 1-2

Task 1

Measurements of the count rate from a photon detector have been taken and the data is given in the file named count_rate_data.csv, available in the File section, in the "Data files Seminar 1" folder: countrate_data.csv. (When reading the data in R, please note that the separator is ";").
You have 100 measurements of the count rate, supposedly in the very same conditions, in close but different moments in time.

a. Fit the data assuming Poisson and Gaussian distribution, respectively. You can do this in R by using the fitdistr() function that is in the MASS package.

The given data consists of 100 measurement points. Each second the count rate is given (Figure 1-1). A fit of the data is executed assuming Poisson and Gaussian distribution.

The results of the fits are:

- Gaussian distribution:
  - Mean: 19.54 (0.441)
  - Standard deviation: 4.41 (0.312)
- Poisson:
  - Lambda: 19.54 (0.442)

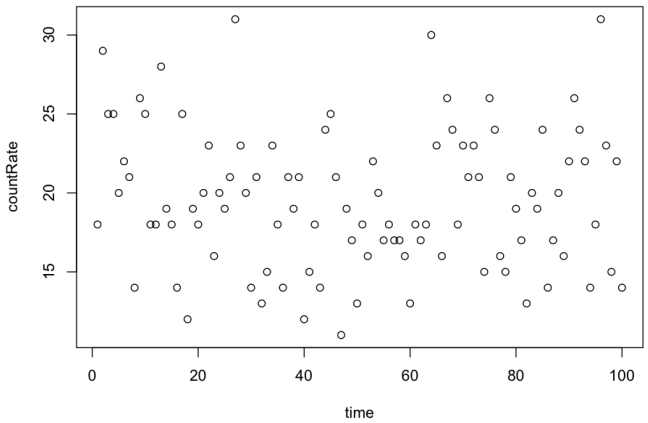

Figure 1-1: Plot of the given data countrate_data.csv

b. Calculate the mean of the data and compare it to the parameters "lambda" and "mean" obtained from the fit.

The mean of the count rate is 19.45 cps. The lambda of the Poisson distribution and the mean of the Gaussian distribution are exactly the same.

c. Plot the data as a histogram and the Gaussian and Poisson fits on the same plot

The histogram and the fits are showed in Figure 1-2.

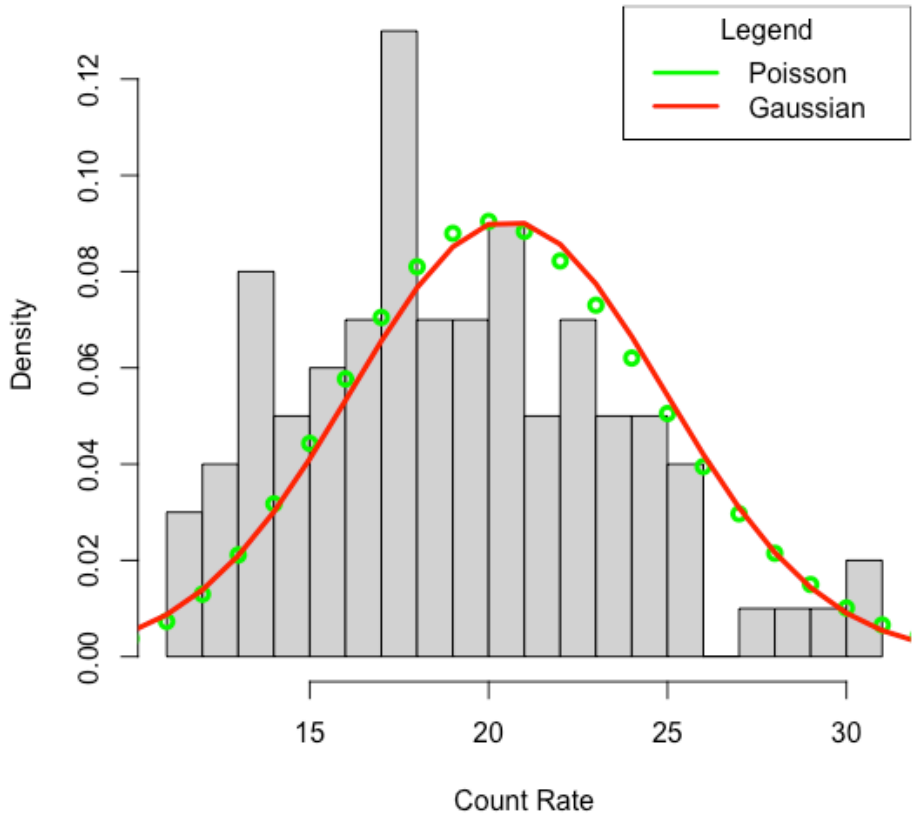

Figure 1-2: Histogram of the count rate and Gaussian and Poisson fit

d. Plot the cumulative distribution function for the fitted Gaussian and Poisson distribution

The plots of cumulative distribution functions are showed in Figure 1-3.

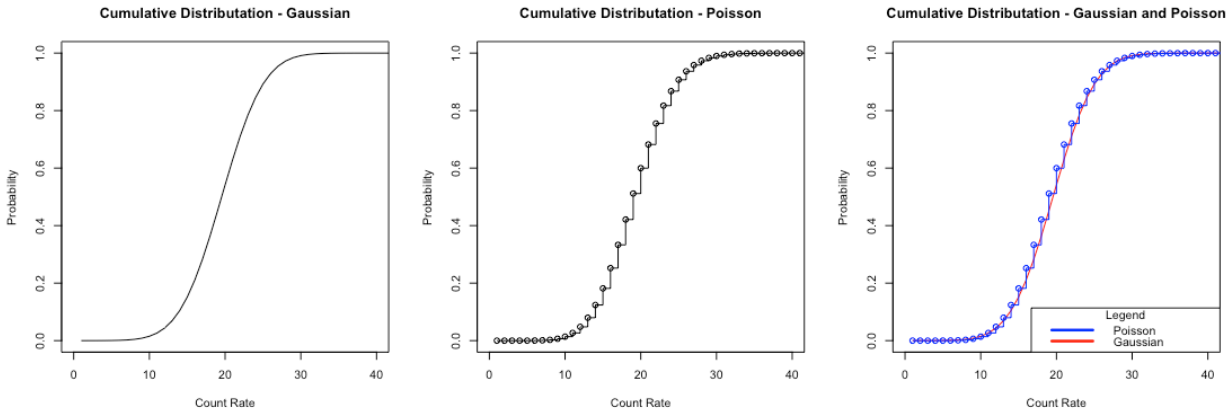

Figure 1-3: Cumulative distribution functions for the fitted Gaussian and Poisson distribution

e. Give an estimate for the count rate and give also a 95% confidence interval (a bit of help: you can use the qnorm and qpois functions for this -but you are welcome to use other methods as well)

The estimate of the count rate (50% quantile of the distributions) and the 95% confidence interval (2.5% quantile and 97.5% quantile) are:

Gaussian distribution:
- Estimate: 19.45
- 95% confidence interval: [10.9 28.18]

Poisson distribution:
- Estimate: 19
- 95% confidence interval: [11 29]

f. Estimate the probability of obtaining a measurement of the count rate of 15 cps

The estimate of obtaining a measurement of a specific count rate can be calculated with pnorm() and ppois(), respectively. For example, the estimated probability of a count rate of 15 cps, assuming a Gaussian distribution, is 15.15%. Assuming a Poisson distribution, the estimated probability of a count rate of 15 cps is 18.17%.

g. Discuss whether or not the Gaussian or the Poisson are good fit to the data (in particular discuss how the "goodness" of the fit can be assessed) and any difference in the estimates of the parameters above. Please, in your discussion try to focus as much as possible on quantitative arguments and try to be as rigorous as possible.

The "goodness" of the fit can be discussed by taking different values and tests into account.

One possibility is the use of the Chi-square test. Therefore, the observed data is compared to the expected data assuming the Poisson and Gaussian distribution, respectively. The compared data points are showed in Figure 1-4.

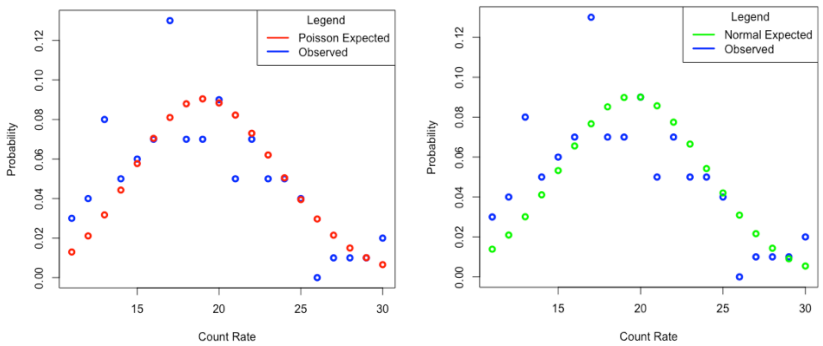

Figure 1-4: Cumulative distribution functions for the fitted Gaussian and Poisson distribution

The result of the Chi-square tests are:

Gaussian distribution:
- P-value: 0.1509

Poisson distribution:
- P-value: 0.2754

Based on these tests, we can say that both functions fit the data good.

h. ***This point is not mandatory but it could be a good exercise****. You will find in the same directory as the one for the count rate data another file with 100 data points with the number of detector counts registered every 10 s (countrate_data_10s_avg.csv). Read this data set and plot the data as an histogram, do a fit assuming Normal and Poisson distributions and plot the two fits on the same graph as the histogram. Compare this graph with the one obtained for the count rate. Can you explain why now the Normal and Poisson fits are closer to each other?*

Task 2

The function of the lungs is described by the vital capacity (VC), measured in litres. In a study of two groups of subjects, patients with pulmonary disease and normal controls, the VC was measured, and the results are found in vcdata.txt. Use statistical analysis to decide whether there is a difference between the patients and the controls.

(2.1) Research question

The research study deals with pulmonary diseases. Therefore, the vital capacity (VC) of patients with pulmonary diseases and a control group is compared. The central hypotheses of the test are:

H_0: "There is no difference between the patient group and the control group."
H_1: "There is a difference between the patient group and the control group."

$\alpha = 0.05$ (Significance level)

(2.2) Analysis of the given data

The results of the test are given in table vcdata.txt and consists of three parts:

1. Patient number (V1)
2. Group affiliation (V2)
3. VC in liters (V3)

In total, we have 28 subjects, 10 subjects in the group "patient" and 18 subjects in the group "control". It has to be noted that it is a very small sample size.

Furthermore, VC is given in liters. This is a form of continuous data. A first comparison of both groups is shown in Figure 2-1. The boxplot for the control group is very symmetrical, which suggests a normal distribution of the data. The patient group, however, look a little bit more asymmetrical.

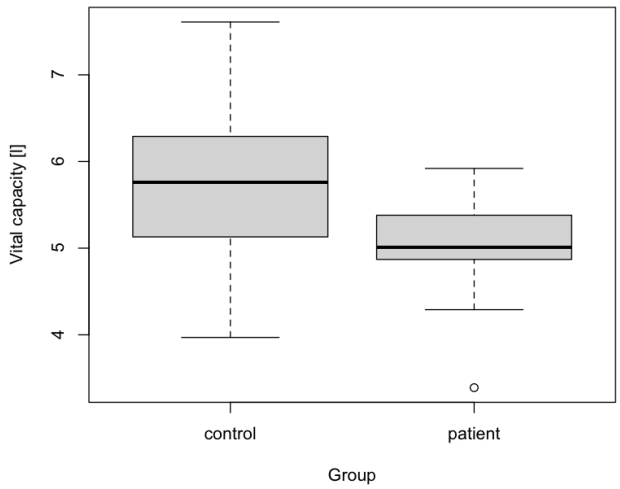

Figure 2-1: Boxplot of the control and patient group

However, a histogram and a Q-Q-plot for normal distribution is used to get a better understanding of the data (Figure 2-2). Almost all data points lay close to the fitted line, so we can assume in the following that the data is normal distributed.

This observation is confirmed by the Shapiro-Wilk test of normality, which return a p-value of 0.8858 for the control group and 0.3441 for the patient group. The used null-hypotheses, that the data is not normal distributed, can thus not be rejected.

To conclude, we have two unpaired groups of parametric continuous data.

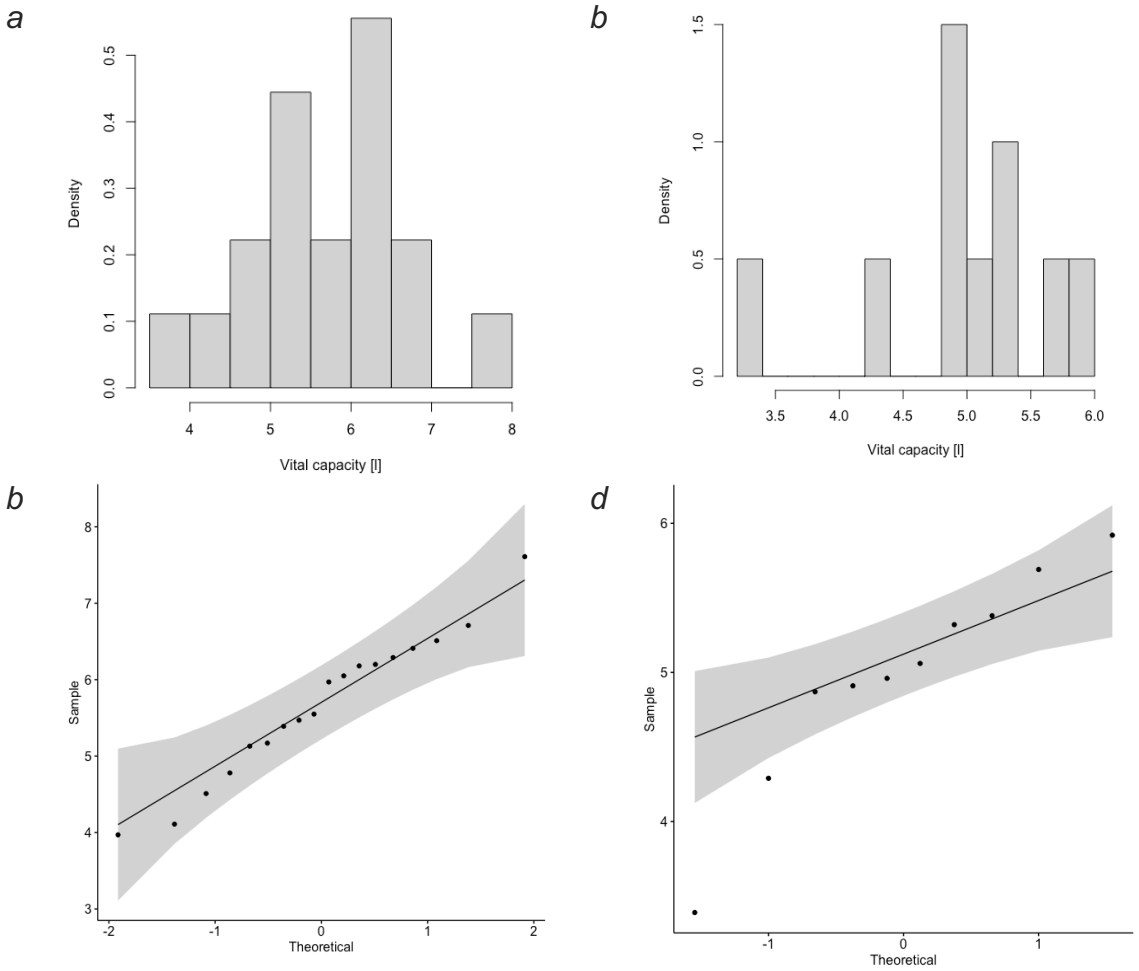

Figure 2-2: Histograms of control (a) and patient group (b) and Q-Q-plots for control (c) and patient group (d)

(2.3) Suggest at least two different statistical approaches, based on the type of data available, considering the structure of the dataset as well as how the numbers appear.

After our first analysis of the data, three statistical approaches are suitable:

1. Unpaired t test (Welch test)
   (the Welch test is used instead of the student t-test due to the uneven sample size and the variation of the groups)
2. Linear regression
3. Mann-Whitney test (Wilcoxon rank sum exact test)

(2.4) Compare the results.

(1) Unpaired t test (Welch)

The results of the Welch t-test are:

- p-value: 0.04214
- 95% confidence interval: [0.02658413, 1.34986031]

(2) Linear regression

The results of the linear regression are:

- p-value: 0.05775
- 95% Confidence Interval: [-1.400774, 0.02432928]
- R-squared: 0.1316

(3) Mann-Whitney test (Wilcoxon rank sum exact test)

The result of the Mann-Whitney test is:

- p-value: 0.04522

(2.5) Discuss what conclusions can be drawn.

All tests return similar p-values. However, for the t-test and the Mann-Whitney test the p-value is smaller than alpha and for the linear regression the p-value is higher than alpha. So, we cannot simply reject the null hypotheses.

Instead, the suitability of the models needs to be discussed. First, we assumed in the beginning that the data is parametric. To assure that we draw the right assumption here, data of more subjects should be included in the study. The used non-parametric test shows a p-value smaller than alpha, which would show a significant difference between both groups. However, the non-parametric test is not as accurate than the parametric tests.

Furthermore, the R-squared value of the linear regression is very low, which indicates that the linear regression may not be the best approach for the given data. A possible reason could be the small sample size and its sensitivity to outliers.

To conclude, a major drawback of the study is the small sample size which makes it difficult to draw a final conclusion. The Welch t-test, which seems to be the more suitable approach, shows a significant difference between two groups. More precisely, according to the t-test the vital capacity is significantly lower for patients with lung diseases than for the control group. This result was also given by the Mann-Whitney test. In the end, however, this conclusion must be confirmed in further studies.

Task 3

Nurses working with patients in a surgical unit may be exposed to anaesthetic gas, e.g. nitrous oxide, which has been shown in epidemiological studies to cause reproductive toxicity. The hospital has implemented a new ventilation system, and they would like to assess the impact that this system has had on the nurses' workplace exposure to the gas. They measured the concentration of nitrous oxide (in ppm) at the workplace of each nurse before and after installing the new system. The results are found in concdata.txt. (The exposure limit for nitrous oxide established by the US National Institute of Occupational Safety and Health is 25 ppm.)

(3.1) Research question

The task is to compare the values before and after the change: has the new system affected the concentration of nitrous oxide at the workplace?

The hypotheses are:

H_0: "There is no significant difference in concentration before and after "
H_1: "There is a significant difference in concentration before and after "

Significance level: 5%

(3.2) Analysis of the given data

The data consists of measurements of one group (consisting of 80 individuals) in two occasions (before and after) of concentrations of Nitrous Oxide. As the box plot in Figure 3-1 suggests, there is a decrease in concentration of Nitrous Dioxide after installing the ventilation system.

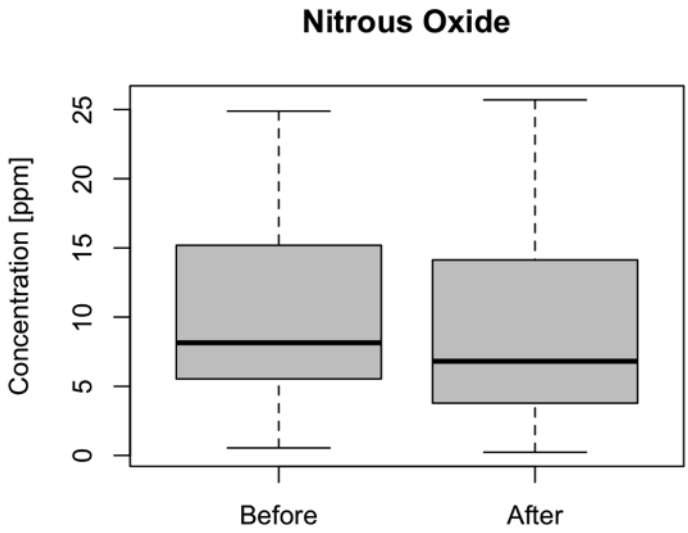

Figure 3-1. Box plot of sample data

The sample data is not normally distributed, which can be observed both in the box plot in Figure 3-1 as well as when plotting each of the samples in a density function (Figure 3-2).

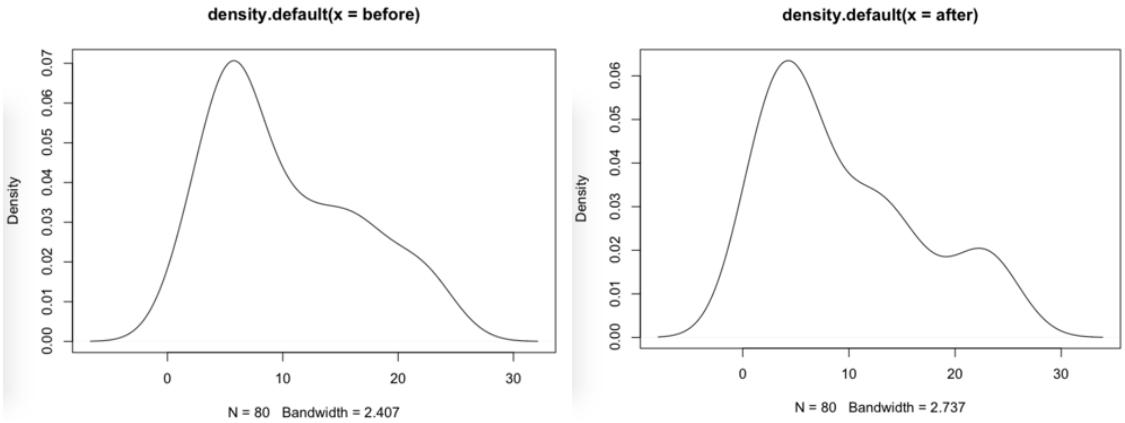

Figure 3-2. Density functions of sample data

However, the density function for the differences in concentration before and after the installation looks more normally distributed than the sample data themselves (Figure 3-3).

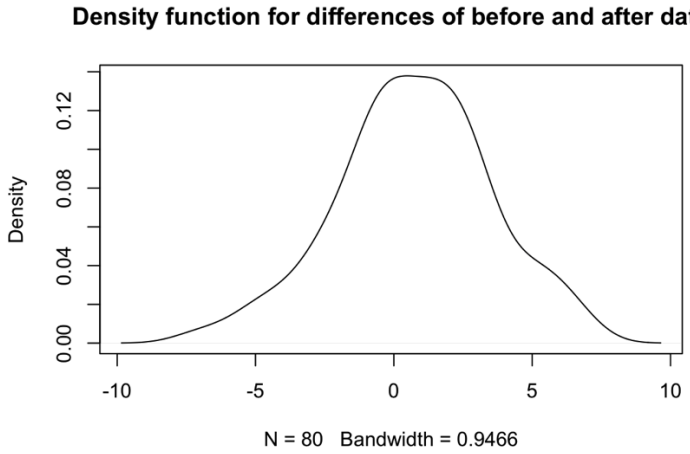

Figure 3-3. Density function of differences between measurements before and after installation

The assumption of normal distribution is confirmed with the Shapiro-Wilk normality test, which generated a p-value of 0.887, indicating that we cannot reject the Shapiro-Wilk test null hypothesis.

(3.3) Suggest two different statistical approaches to do this comparison.

The data for before and after measurements is continuous interval/ratio, and dependent on each other since it is measurements of the same sample in two different cases. Two appropriate tests are:

- The paired t-test
- The paired Wilcoxon rank sum test.

(3.4) Compare the results

Paired t-test
- p-value: 0.02595
- 95% confidence interval: [0.08799374, 1.34196127]

Paired Wilcoxon rank sum test:
- p-value: 0.02201

(3.5) Discuss what conclusions can be drawn.

Both the t-test and Wilcoxon test shows that there is a significant difference in the concentration measurements before and after installing the ventilation system, since both have a p-value < 0.05. The null hypothesis, that there is no difference, can therefore be rejected.

Task 4

In a clinical study of rheumatoid arthritis, four different drugs were compared in 40 patients, and the clinical response was classified as "improved", "unchanged" or "worsened". Analyze the numeric results to determine whether there is a difference in response between the four treatments.

(4.1) Research question

The task is to verify that there is no difference in response between the four treatments. The hypotheses are:

H_0: "There is no difference in response between the four treatments"
H_1: "There is a difference in response between the four treatments"

$\alpha = 0.05$ (Significance level)

(4.2) Create a suitable presentation of the data, which are available in *treatmentdata.txt*.

The results of the clinical study are presented in Table 4-1 and in Figure 4-1.

Treatment	improved	unchanged	worsened
A	5	3	2
B	8	2	0
C	4	2	4
D	2	1	7

Table 4-1: Results of the clinical study of rheumatoid arthritis

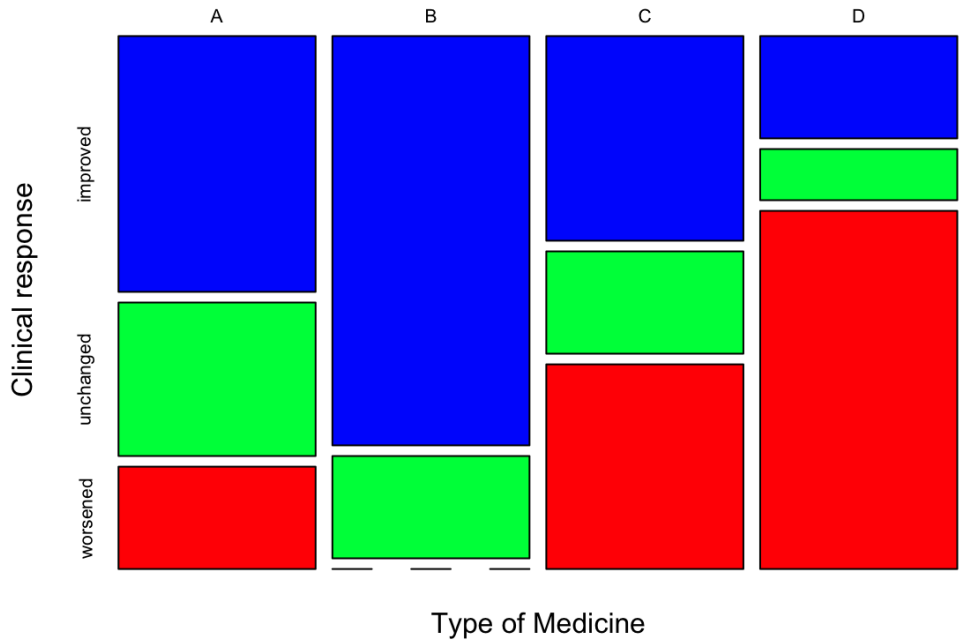

Figure 4-1: Results of the clinical study of rheumatoid arthritis; the clinical response can either improved the patients' health (blue), left it unchanged (green) are worsened it (red)

(4.3) Consider and compare at least two different approaches and make a recommendation on which one to use.

First, it must be checked if the data is parametric. Therefore, the Shapiro-Wilk normality test is used. The null hypothesis is, that the data is normal distributed. The p-value of the Shapiro-Wilk normality test is 4.7e-07. Therefore, we can reject the null hypothesis and need to assume that the data is non-parametric.

We can analyze those data as ordinal and add one more column 'rank' that scored -1, 0, 1 for worsened, unchanged, improved respectively. So,

- One-way ANOVA and
- Kruskal-Willis tests

are chosen, because there are four groups that are compared. Furthermore, the Chi-square test is ruled out, because the expected numbers in each field < 5 and the sample size < 60 as well. Even though the data is assumed not to be non-parametric, the ANOVA is used for comparison reasons.

Results

Two suitable approaches are:
- One-way ANOVA test
- Kruskal-Willis test

The summary of the One-way ANOVA test is:

	Df	Sum (sq)	Mean (sq)	F value	Pr(>F)
Treatment	3	8.9	2.9667	4.811	0.00643
Residuals	36	22.2	0.6167		

Table 4-2: Summary of the One-way ANOVA test

The result of the Kruskal-Willis test are:
- chi-squared: 10.775
- df: 3
- p-value: 0.01301

Discussion

Both One-way ANOVA test and Kruskal-Wallis rank sum test shows that there is significant difference: p-value < 0.05, observed. However, concerning about the results of Shapiro-Wilk normality test, p-value < 0.5, so these data are not normally distributed. Since the precondition of One-way ANOVA test is that the data has normality, in this case, **Kruskal-Wallis rank sum test result is more reliable**.