Assignment template

James Bond, group 007

1 February 2033

Short introduction to R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. R Markdown files permit you to interweave R code with ordinary text to produce well-formatted data analysis reports that are easy to modify. The R Markdown file itself shows the readers exactly how you got the results in your report. For more details on using R Markdown see http://rmarkdown.rstudio.com.

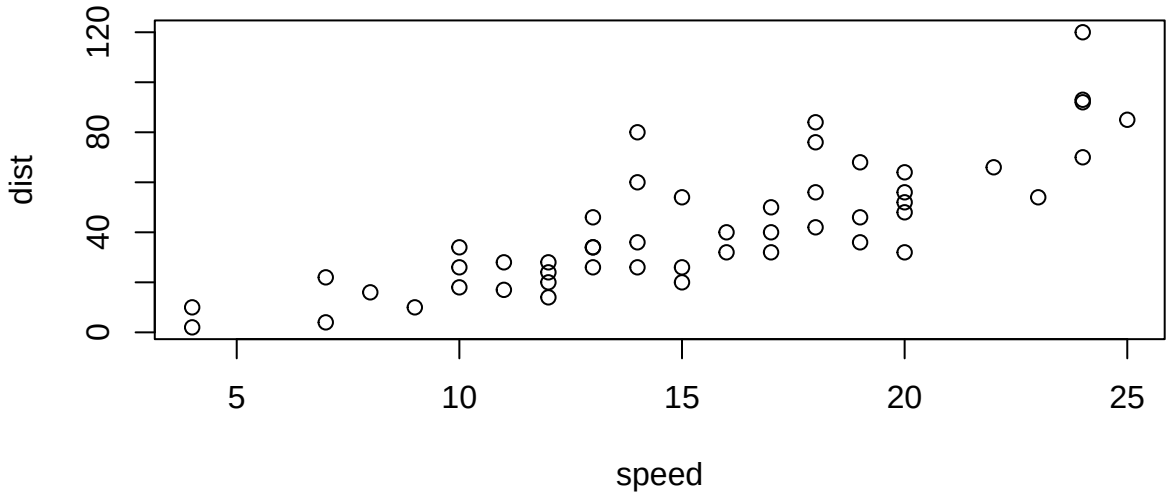
When you click the **Knit** button, a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. For inline R code, surround code with back ticks and r. R replaces inline code with its results. For example, two plus one is 3; for the build-in R dataset `cars`, there were 50 cars studied. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Figures

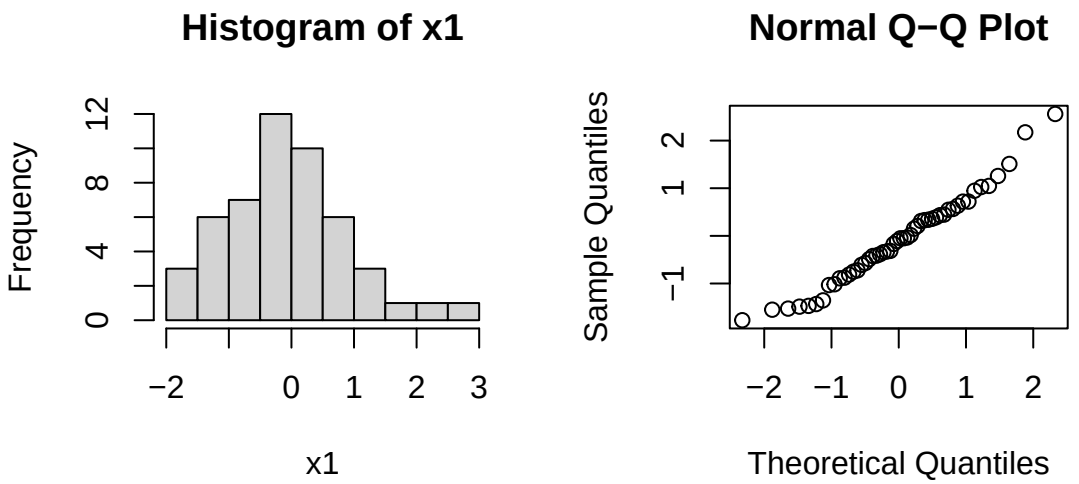
You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot. Use knitr options to style the output of a chunk. Place options in brackets above the chunk. Other options with the defaults are: the `eval=FALSE` option just displays the R code (and does not run it); `warning=TRUE` whether to display warnings; `tidy=TRUE` wraps long code so it does not run off the page.

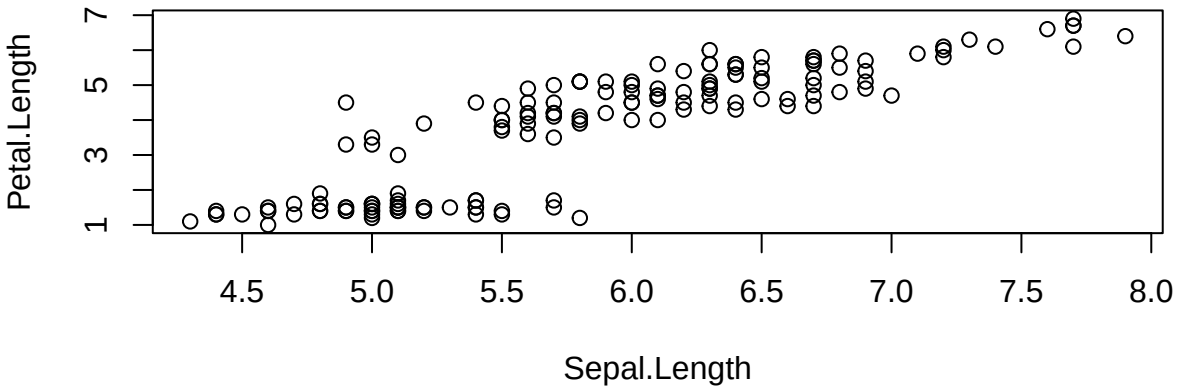
You can control the size and placement of figures. For example, you can put two figures (or more) next to each other. Use `par(mfrow=c(n,m))` to create `n` by `m` plots in one picture in R. You can adjust the proportions of figures by using the `fig.width` and `fig.height` chunk options. These are specified in inches, and will be automatically scaled down to fit within the handout margin. Chunk option `fig.align` takes values `left`, `right`, or `center` (to align figures in the output document).

```
par(mfrow=c(1,2)); x1=rnorm(50); hist(x1); qqnorm(x1)
```



You can arrange for figures to span across the entire page by using the `fig.fullwidth` chunk option.

```
plot(iris$Sepal.Length,iris$Petal.Length,xlab="Sepal.Length",ylab="Petal.Length")
```



More about chunk options can be found at https://yihui.name/knitr/options/.

Equations

To produce mathematical symbols, you can also include LaTeX expessions/equations in your report: inline $\frac{d}{dx}\left(\int_0^x f(u)\,du\right) = f(x)$ and in the display mode: To be able to use this functionality, LaTeX has to be installed.

Footnotes

Here is the use of a footnote[1].

Images

Want an image? This will do it. To depict an image (say, `my_image.jpg` which should be in your current working directory), use this command

Tables

Want a table? This will create one (note that the separators *do not* have to be aligned).

Table Header	Second Header
Table Cell	Cell 2
Cell 3	Cell 4

You can also make table by using knit's `kable` function:

[1]

Exercise 1. Waiting time

A researcher measured (in minutes) how long patients have to wait in the waiting room of a doctor's office: 15.4, 17.9, 19.0, 0.5, 15.9, 2.7, 6.2, 2.5, 4.7, 6.9, 10.8, 24.3, 5.6, 23.0, 10.7. Denote the mean waiting time by /mu.

Table 2: A knit kable.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

Block quote

> This will create a block quote, if you want one.

Verbatim

```
This text is displayed verbatim/preformatted.
```

Links

Links: http://example.com, in-text link to Google.

This is a hyperlink.

This

is where the hyperlink jumps to.

Itimization, italicized and embolded text

- Single asterisks italicize text *like this*.
- Double asterisks embolden text **like this**.

One more way to italicize and embold: *italic* and **bold**.

Exercise 1

Below is a template for reporting the exercises from the assignments.

a) Here are some consequitive R-commands.

```
x=rep(c("A","B"),each=5); x
```

```
##  [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B"
```

```
sample(x)
```

```
##  [1] "A" "A" "A" "B" "A" "B" "B" "B" "A" "B"
```

```
x=rnorm(100)
```

Now the same code chunk but with all the output collapsed into signle block.

```
x=rep(c("A","B"),each=5); x
##  [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B"
sample(x)
##  [1] "A" "B" "B" "B" "A" "A" "B" "B" "A" "A"
x=rnorm(100)
```

b) Below we perform a one sample t-test for the artificial data (that we generate ourselves).

```
mu=0.2
x=rnorm(100,mu,1) # creating artificial data
t.test(x,mean=0)   # t.test(x,alternative=c("two.sided"),conf.level=0.95,mu=10)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 1.8699, df = 99, p-value = 0.06445
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.01114777  0.37591306
## sample estimates:
## mean of x
## 0.1823826
```

c) We often do not need to report the whole output of R-commands, only certain values of the output. For example, below we perform a two-sample t-test and report only the (appropriately rounded) values of t-statistics and the p-pavue.

```
mu=0;nu=0.5
x=rnorm(50,mu,1); y=rnorm(50,nu,1) # creating artificial data
ttest=t.test(x,y)
```

The value of t-statistics in the above evaluation is -1.62 and the p-value is 0.1096.

Exercise 1. Waiting time.

A researcher measured (in minutes) how long patients have to wait in the waiting room of a doctor's office: 15.4, 17.9, 19.0, 0.5, 15.9, 2.7, 6.2, 2.5, 4.7, 6.9, 10.8, 24.3, 5.6, 23.0, 10.7. Denote the mean waiting time by μ.

```
x <- as.numeric(list(15.4, 17.9, 19.0, 0.5, 15.9, 2.7, 6.2, 2.5, 4.7, 6.9, 10.8, 24.3, 5.6, 23
```

a) Check normality of the data. Assuming normality (irrespective of your conclusion about normality of the data), construct a 97%-CI for μ. Evaluate the sample size needed to provide that the length of the 97%-CI is at most 2. Compute a bootstrap 97%-CI for μ and compare it to the above CI.

Let's check the normality using Shapiro-Wilk test. H_0 is that sample x came from normally distributed population.

```
shapiro.test(x)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.93473, p-value = 0.3207
```

From the output, the p-value > 0.05 implying that the distribution of the data are not significantly different from normal distribution, i.e. the null hypothesis can not be rejected. In other words, we can assume the normality.

Estimated mean value:

```
mu = mean(x)
mu
```

```
## [1] 11.07333
```

Next, we are going to construct a 97%-CI for μ. The standard deviation σ is unknown, therefore, we estimate it by s.

```
s = sd(x)
s
```

```
## [1] 7.727545
```

The confidence interval in such a case is based on a t-distribution and the upper t-quantile.

```
alpha <- 1 - 0.97
n <- length(x)
ta <- qt(1-alpha/2, df=n-1)
ta
```

```
## [1] 2.414898
```

t-confidence interval of level 97% for μ:

```
CI_97 <- c(mu - ta*s/sqrt(n), mu + ta*s/sqrt(n))
CI_97
```

```
## [1]  6.255024 15.891642
```

Next, we evaluate the sample size needed to provide that the length of the 97%-CI is at most 2. For this, we have to solve $t_{\alpha/2}\frac{s}{\sqrt{n}} \leq E$ for n.

```
E <- 2
n_min <- (ta*s/E)^2
n_min
```

```
## [1] 87.06039
```

To provide the length of the 97%-CI less than 2, we have to collect the sample of at lest 88 objects.

Let's compute a bootstrap 97%-CI for μ using 1000 samples.

```
B = 1000
Tstar = numeric(B)

for(i in 1:B) {
```

```
  Xstar = sample(x, replace=TRUE)
  Tstar[i] = mean(Xstar)
}

TstarLower = quantile(Tstar, alpha/2)
TstarUpper = quantile(Tstar, 1-alpha/2)

bootstrap_CI_97 <- c(2*mu - TstarUpper, 2*mu - TstarLower)
bootstrap_CI_97
```

```
##     98.5%      1.5%
##  6.766167 15.067067
```

The confidence intervals look very close to each other. The one, calculated with a bootstrapping, is stochastic and therefore differs from launch to launch.

b) The doctor claims that the mean waiting time is less than 15 minutes. Under an assumption, verify this claim by a relevant t-test, explain the meaning of the CI in the R-output for this test. Propose and perform a suitable sign tests for this problem. Can we use yet another test based on ranks?

One-sided t-test with H_0: mean waiting time ≥ 15; H_1: mean waiting time < 15:

```
t.test(x, mu=15, alt='l')
```

```
##
##  One Sample t-test
##
## data:  x
## t = -1.968, df = 14, p-value = 0.0346
## alternative hypothesis: true mean is less than 15
## 95 percent confidence interval:
##      -Inf 14.58758
## sample estimates:
## mean of x
##  11.07333
```

H_0 is rejected. The doctor's claim (alternative hypothesis) is accepted. The confidence interval is also one-sided (left-sided). The given value of 15 is outside CI and this also tells about rejecting H_0.

A sign test for median of a single sample may be applied if we state the claim as "the median waiting time is less than 15 minutes":

```
binom.test(sum(x<15), length(x), p = 0.5, alternative = "less", conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  sum(x < 15) and length(x)
## number of successes = 9, number of trials = 15, p-value = 0.8491
## alternative hypothesis: true probability of success is less than 0.5
```

```
## 95 percent confidence interval:
##  0.0000000 0.8091353
## sample estimates:
## probability of success
##                    0.6
```

The calculated p-value is 0.85. Since this is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that median waiting time is less than 15 minutes.

In the same manner one-sample Wilcoxon signed rank test may be applied.

c) Propose a way to compute the powers of the t-test and sign test from b) at $\mu = 14$ and $\mu = 13$, comment.

The powers may be computed during a simulation as a probability of rejecting H_0 when H_1 is true. For this, we have to generate samples from H_1. For both tests we can generate from normal distribution with the mean of 15, 14, 13.

```
B <- 1000

for(m in 13:15){
  ttest <- numeric(B)
  sign <- numeric(B)
  for(i in 1L:B){
    h1_sample = rnorm(n, mean=m, sd=s)

    ttest[i] <- t.test(h1_sample, mu=mu, alt='l')[[3]]
    sign[i] <- binom.test(sum(h1_sample<mu), length(h1_sample), p = 0.5,
                          alternative = "less", conf.level = 0.95)[[3]]
  }
  print(paste0("H1 mu=", m))
  print(paste0("t-test power ", sum(ttest < 0.05)/B))
  print(paste0("sign test power ", sum(sign < 0.05)/B))
}
```

```
## [1] "H1 mu=13"
## [1] "t-test power 0.005"
## [1] "sign test power 0.09"
## [1] "H1 mu=14"
## [1] "t-test power 0.001"
## [1] "sign test power 0.168"
## [1] "H1 mu=15"
## [1] "t-test power 0"
## [1] "sign test power 0.302"
```

d) Let p be the probability that a patient has to wait longer than 15.5 minutes. Using asymptotic normality, the researcher computed the right end $\hat{p}_r = 0.53$ of the confidence interval $[\hat{p}_l, \hat{p}_r]$ for p. Recover the whole confidence interval and its confidence level.

Let's estimate a proportion of patients to wait longer than 15.5 minutes. *p_hat* is a point estimate for p.

```r
p_hat = mean(x > 15.5)
p_hat
```

```
## [1] 0.3333333
```

(1-α)-confidence interval for p is $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

```r
p_hat_r <- 0.53
margin_error = p_hat_r - p_hat
p_hat_l <- p_hat - margin_error
p_hat_l
```

```
## [1] 0.1366667
```

Let's calculate Z alpha/2 quantile:

```r
se <- sqrt((p_hat * (1 - p_hat)) / n)
z_alpha_by_2 <- margin_error / se
z_alpha_by_2
```

```
## [1] 1.615782
```

```r
alpha = (1 - pnorm(z_alpha_by_2))*2
1-alpha
```

```
## [1] 0.8938584
```

It was a 0.89-confidence interval for p

e) The researcher also reported that there were 3 men and 2 women among 5 patients who had to wait more than 15.5 minutes, 4 men and 6 women among the remaining 10 patients. The researcher claims that the waiting time is different for men and women. Verify this claim by an appropriate test.

Here we test whether the proportions of men and women in two groups waiting more and less than 15.5 minutes are significantly different. We apply the approximate proportion test:

```r
prop.test(c(2, 6), c(5, 10))
```

```
## Warning in prop.test(c(2, 6), c(5, 10)): Chi-squared approximation may be
## incorrect
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(2, 6) out of c(5, 10)
## X-squared = 0.033482, df = 1, p-value = 0.8548
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.8759135  0.4759135
## sample estimates:
## prop 1 prop 2
##    0.4    0.6
```

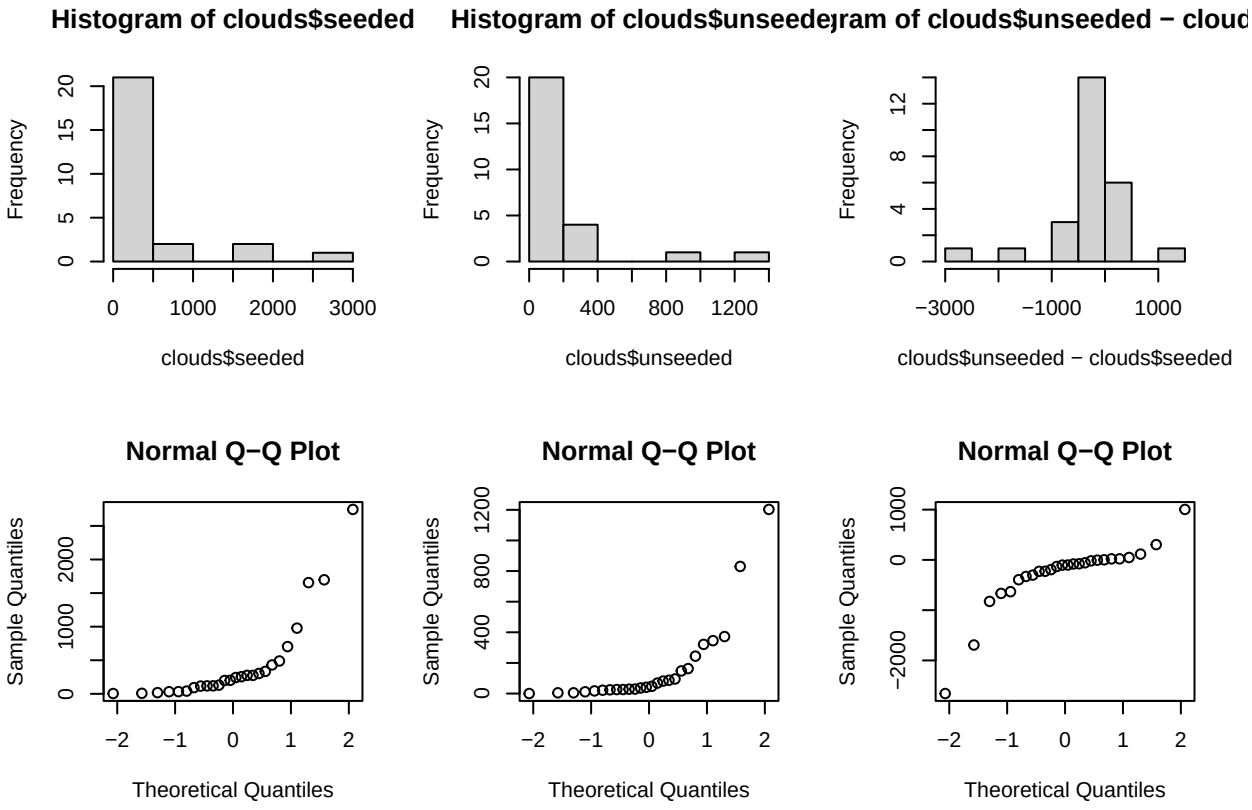
There is no significant evidence that the waiting time is different for men and women.

Exercise 2. Seeded clouds.

To improve rain fall in dry areas, an experiment was carried out with 52 clouds. Scientists investigated whether the addition of silver nitrate leads to more rainfall. They chose 26 out of a sample of 52 clouds and seeded it with silver nitrate. The file clouds.txt contains the precipitation values (records the rainfall in feet per acre) of seeded and unseeded clouds.

```
clouds <- read.table("data/clouds.txt", header=TRUE)

par(mfrow=c(2, 3))
hist(clouds$seeded)
hist(clouds$unseeded)
hist(clouds$unseeded - clouds$seeded)
qqnorm(clouds$seeded)
qqnorm(clouds$unseeded)
qqnorm(clouds$unseeded - clouds$seeded)
```




```
shapiro.test(clouds$unseeded - clouds$seeded)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  clouds$unseeded - clouds$seeded
```

```
## W = 0.75882, p-value = 3.791e-05
```

From the histograms we can easily notice that data is distributed not normally. The distributions look closer to exponential. The difference also is not distributed normally according to Shapiro-Wilk test.

a) Test whether silver nitrate has an effect by performing three tests: the two samples t-test (argue whether the data are paired or not), the Mann-Whitney test and the Kolmogorov-Smirnov test. Indicate whether these tests are actually applicable for our research question. Comment on your findings.

The data might be counted as paired if the data was collected in the following way: two more or less similar clouds are found not far from each other and only one of them is seeded. In the target experiment, the half of clouds was selected without any requirements so we are not assuming that the samples are paired.

```
t.test(clouds$unseeded, clouds$seeded, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  clouds$unseeded and clouds$seeded
## t = -1.9984, df = 33.856, p-value = 0.05375
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -559.585876    4.740491
## sample estimates:
## mean of x mean of y
##  164.5619  441.9846
```

According to two not paired samples t-test, the H_0 states that the means are equal is not rejected. T-test actually may not be performed on our data as the columns are even approximately not distributed normally as well as their difference.

Mann-Whitney test doesn't assume normality and, therefore may be applied. The data is continuous and we can limit the alternative to a shift in location.

```
wilcox.test(clouds$unseeded, clouds$seeded)
```

```
## Warning in wilcox.test.default(clouds$unseeded, clouds$seeded): cannot compute
## exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  clouds$unseeded and clouds$seeded
## W = 203, p-value = 0.01383
## alternative hypothesis: true location shift is not equal to 0
```

```
median(clouds$unseeded); median(clouds$seeded)
```

```
## [1] 44.2
```

```
## [1] 221.6
```

According to Mann-Whitney test, H_0 of equal means is rejected. The underlying distribution of precipitation for seeded clouds is shifted to the right from that of unseeded ones.

Kolmogorov-Smirnov test also doesn't assume normality. H_0: equality of continuous distributions.

```
ks.test(clouds$unseeded, clouds$seeded)
```

```
## Warning in ks.test(clouds$unseeded, clouds$seeded): cannot compute exact p-value
## with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  clouds$unseeded and clouds$seeded
## D = 0.42308, p-value = 0.01905
## alternative hypothesis: two-sided
```

```
mean(clouds$unseeded); mean(clouds$seeded)
```

```
## [1] 164.5619
```

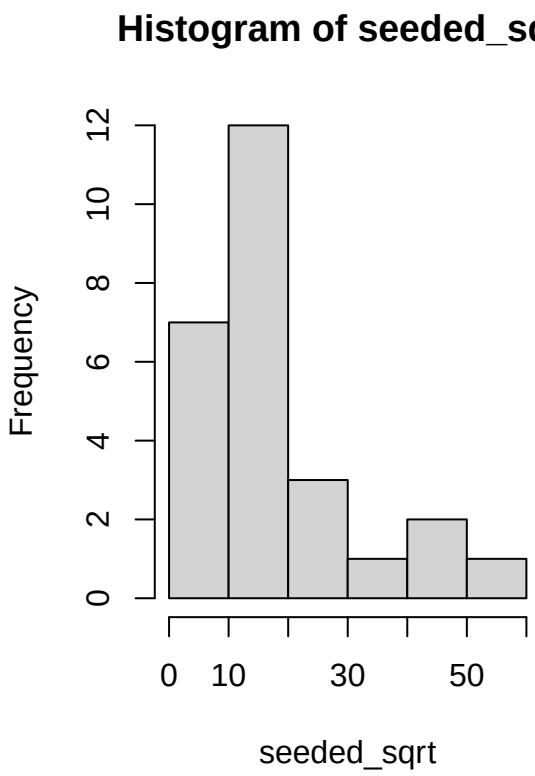
```
## [1] 441.9846
```

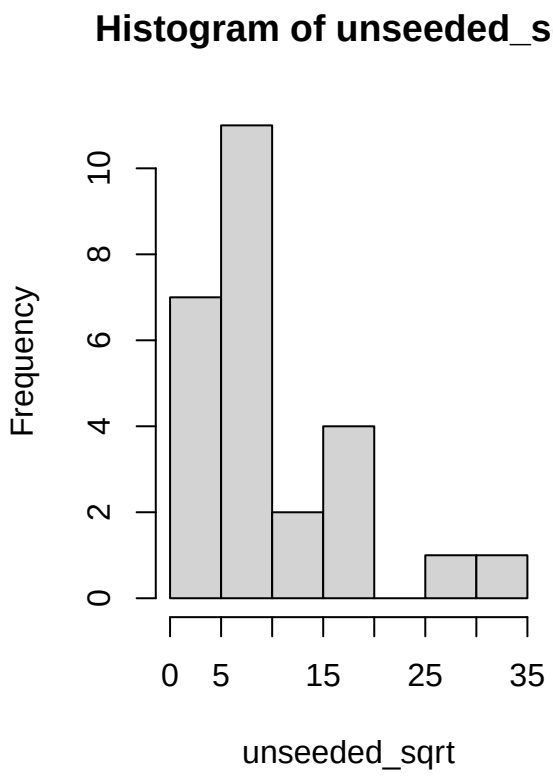
Kolmogorov-Smirnov test also rejects H_0. The mean amount of precipitation is larger for seeded clouds than for unseeded.

b) Repeat the procedures from a) first on the square root of the values in *clouds.txt*, then on the square root of the square root of the values in *clouds.txt*. Comment on your findings.

```
unseeded_sqrt <- sqrt(clouds$unseeded)
seeded_sqrt <- sqrt(clouds$seeded)

par(mfrow=c(1, 2))
hist(seeded_sqrt)
hist(unseeded_sqrt)
```





Not the data looks more normal. Let's check it for normality once again.

```
shapiro.test(unseeded_sqrt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  unseeded_sqrt
## W = 0.83744, p-value = 0.0008196
```

```
shapiro.test(seeded_sqrt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  seeded_sqrt
## W = 0.87394, p-value = 0.004298
```

The p-value < 0.05 for both columns. This implies that the distributions of the data are significantly different from normal distribution. This means that t-test may not be performed on our data and applied just for interest.

```
t.test(unseeded_sqrt, seeded_sqrt, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  unseeded_sqrt and seeded_sqrt
## t = -2.4246, df = 43.363, p-value = 0.01956
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.071300  -1.202087
## sample estimates:
## mean of x mean of y
##  9.931321 17.068014
```

```
wilcox.test(unseeded_sqrt, seeded_sqrt)
```

```
## Warning in wilcox.test.default(unseeded_sqrt, seeded_sqrt): cannot compute exact
## p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  unseeded_sqrt and seeded_sqrt
## W = 203, p-value = 0.01383
## alternative hypothesis: true location shift is not equal to 0
```

```
ks.test(unseeded_sqrt, seeded_sqrt)
```

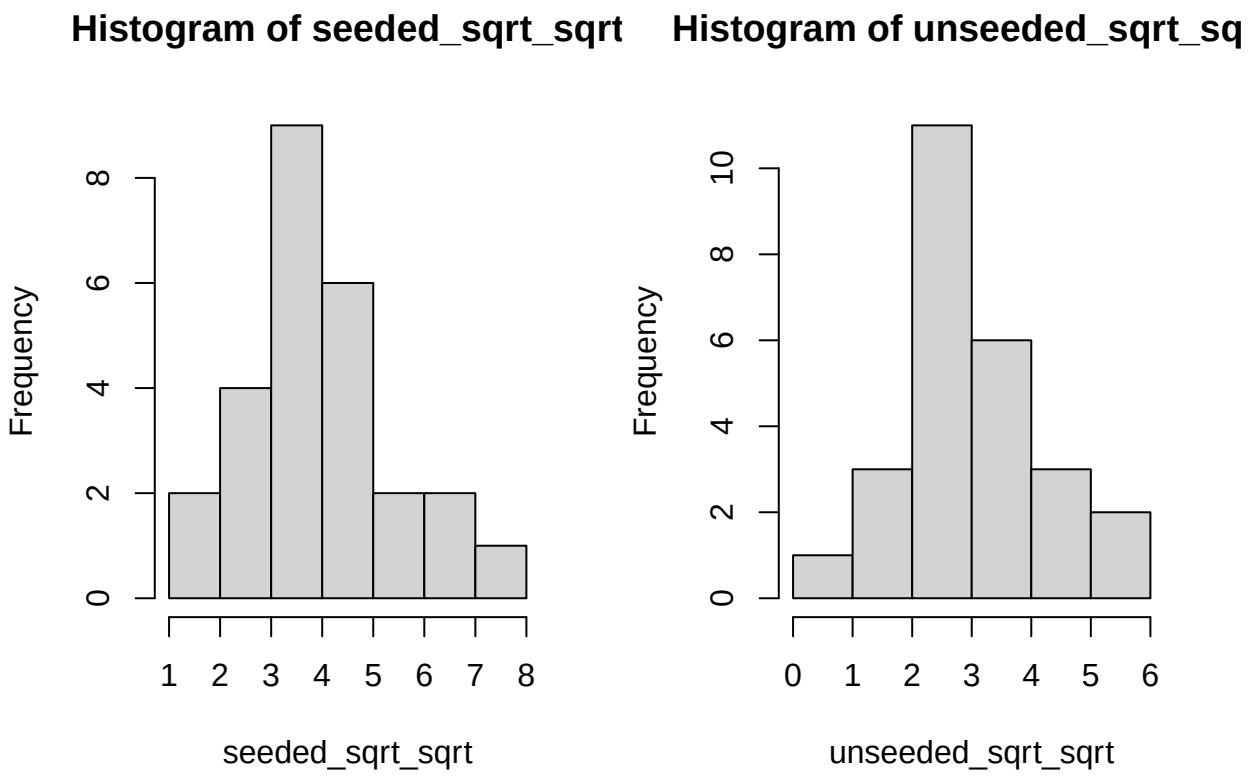
```
## Warning in ks.test(unseeded_sqrt, seeded_sqrt): cannot compute exact p-value
## with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  unseeded_sqrt and seeded_sqrt
## D = 0.42308, p-value = 0.01905
## alternative hypothesis: two-sided
```

In this case t-test rejects the H_0, so means of squared values a significantly different. Interestingly, Wilcoxon and Kolmogorov-Smirnov tests remained completely the same as they are both based on ranks. The ranks remain the same because square root function increases monotonically.

```
unseeded_sqrt_sqrt <- sqrt(unseeded_sqrt)
seeded_sqrt_sqrt <- sqrt(seeded_sqrt)

par(mfrow=c(1, 2))
hist(seeded_sqrt_sqrt)
hist(unseeded_sqrt_sqrt)
```



Not the data looks normal. Let's check it for normality once again.

```
shapiro.test(unseeded_sqrt_sqrt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  unseeded_sqrt_sqrt
## W = 0.95778, p-value = 0.3497
```

```
shapiro.test(seeded_sqrt_sqrt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  seeded_sqrt_sqrt
## W = 0.96504, p-value = 0.5004
```

From the output, the p-value > 0.05 for both columns implying that the distributions of the data are not significantly different from normal distribution. Only now, for 4th roots of columns we can apply t-test.

```
t.test(unseeded_sqrt_sqrt, seeded_sqrt_sqrt, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  unseeded_sqrt_sqrt and seeded_sqrt_sqrt
## t = -2.5968, df = 48.826, p-value = 0.0124
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7236468 -0.2196477
## sample estimates:
## mean of x mean of y
##  2.907340  3.878988
```

```
wilcox.test(unseeded_sqrt_sqrt, seeded_sqrt_sqrt)
```

```
## Warning in wilcox.test.default(unseeded_sqrt_sqrt, seeded_sqrt_sqrt): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  unseeded_sqrt_sqrt and seeded_sqrt_sqrt
## W = 203, p-value = 0.01383
## alternative hypothesis: true location shift is not equal to 0
```

```
ks.test(unseeded_sqrt_sqrt, seeded_sqrt_sqrt)
```

```
## Warning in ks.test(unseeded_sqrt_sqrt, seeded_sqrt_sqrt): cannot compute exact
## p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  unseeded_sqrt_sqrt and seeded_sqrt_sqrt
## D = 0.42308, p-value = 0.01905
## alternative hypothesis: two-sided
```

Wilcoxon and Kolmogorov-Smirnov tests didn't change for the same reason as before. But now all three tests reject H_0 and we can conclude that for 4th roots of measurements, the columns are distributed differently.

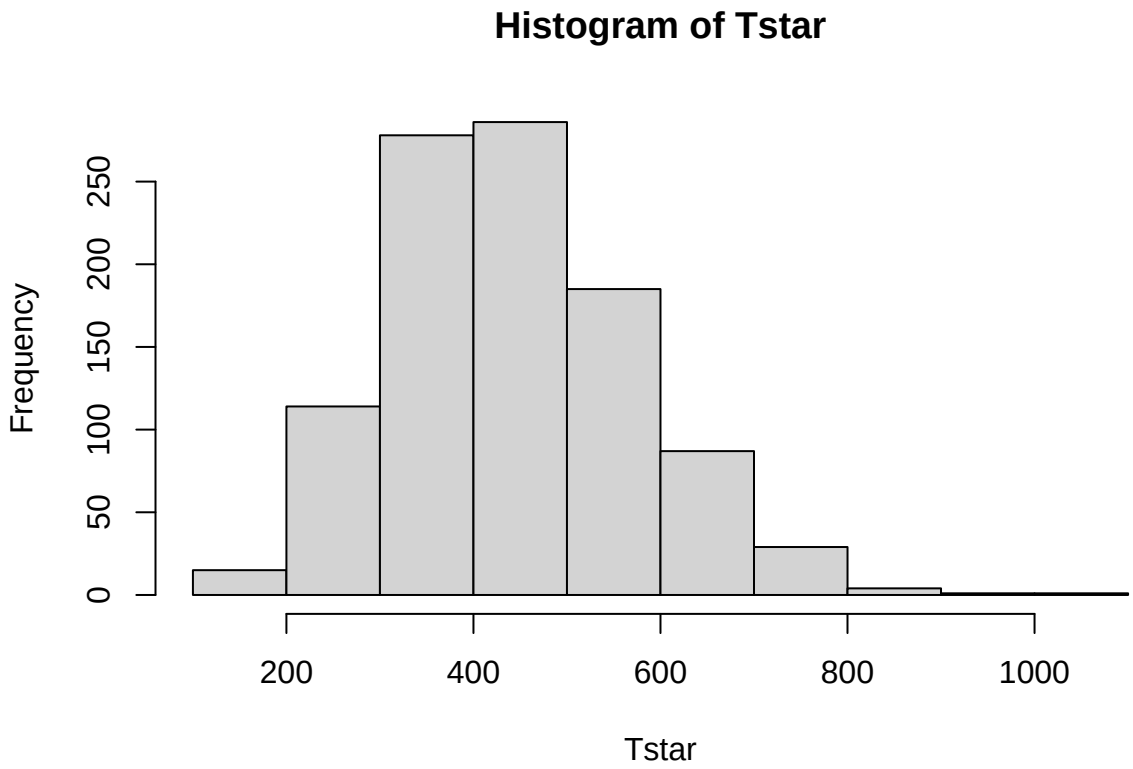
c) Let X1,...,X26 be the sample for seeded clouds (column *seeded*). Assuming X1,...,X26$\sim$Exp(λ) and using the central limit theorem, find an estimate $\hat{\lambda}$ of λ and construct a 95%-CI for λ. By using a bootstrap test with the test statistic T=median(X1,...,X26), test the hypothesis H_0:X1,...,X26$\sim$Exp(λ_0) with the parameter $\lambda_0=\hat{\lambda}$. Test this also by the Kolmogorov-Smirnov test.

```
seeded <- clouds$seeded

B <- 1000
Tstar <- numeric(B)
for(i in 1:B){
  Xstar <- sample(seeded, replace=TRUE)
  Tstar[i] <- mean(Xstar)
}
lambda_hat <- 1/mean(Tstar)
lambda_hat
```

```
## [1] 0.002252054
```

```
hist(Tstar)
```




```
alpha <- 1 - 0.95
deltastar <- 1/Tstar - lambda_hat
d <- quantile(deltastar, c(alpha/2, 1-alpha/2))
CI95 = lambda_hat - c(d[2], d[1])
lambda_hat; CI95
```

```
## [1] 0.002252054
```

```
##        97.5%         2.5%
## 0.0001372394 0.0031029465
```

Next, we check H_0: X1,...,X26$\sim$Exp(λ_0) with the parameter λ_0=$\hat{\lambda}$ using a bootstrap test.

```
B <- 1000
t <- median(seeded)
tstar <- numeric(B)
n <- length(seeded)
for(i in 1:B){
  xstar <- rexp(n, lambda_hat)
  tstar[i] <- median(xstar)
}
pl <- sum(tstar<t)/B
pr <- sum(tstar>t)/B
p <- 2*min(pl, pr)
pl;pr;p
```

```
## [1] 0.121
```

```
## [1] 0.879
```

```
## [1] 0.242
```

There is no evidence against H_0. Let's test the same hypothesis with Kolmogorov-Smirnov test:

```
ks.test(seeded, rexp(n, lambda_hat))
```

```
## Warning in ks.test(seeded, rexp(n, lambda_hat)): cannot compute exact p-value
## with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  seeded and rexp(n, lambda_hat)
## D = 0.26923, p-value = 0.3027
## alternative hypothesis: two-sided
```

This test also doesn't reject the null hypothesis.

d) Using an appropriate test, verify whether the median precipitation for seeded clouds is less than 300. Next, design and perform a test to check whether the fraction of the seeded clouds with the precipitation less than 30 is at most 25%.

To check whether the median precipitation for seeded clouds is less than 300 (H_1), we will use binomial test for a proportion. The test is non–parametric, so we do not assume that the data is normally distributed. As the theoretical probabilities are equal, the binomial test becomes its special case - sign test.

```
binom.test(sum(seeded<300), length(seeded), p = 0.5, alternative = "less", conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  sum(seeded < 300) and length(seeded)
## number of successes = 17, number of trials = 26, p-value = 0.9622
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.8060396
## sample estimates:
## probability of success
##              0.6538462
```

Since this is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that median precipitation for seeded clouds is less than 300.

Similarly, we check whether the fraction of the seeded clouds with the precipitation less than 30 is at most 25%.

```
binom.test(sum(seeded<30), length(seeded), p = 0.25, alternative = "less", conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  sum(seeded < 30) and length(seeded)
## number of successes = 3, number of trials = 26, p-value = 0.08019
## alternative hypothesis: true probability of success is less than 0.25
## 95 percent confidence interval:
##  0.000000 0.271902
## sample estimates:
## probability of success
##              0.1153846
```

Again, we do not have sufficient evidence to say that the fraction of the seeded clouds with the precipitation less than 30 is at most 25%.

Exercise 3. Concentrations of epinephrine.

a) Is it reasonable to assume that the three columns of dogs.txt were taken from normal populations?

```
dogs <- read.table("data/dogs.txt", header=TRUE)
shapiro.test(dogs$isofluorane)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dogs$isofluorane
## W = 0.83093, p-value = 0.03434
```

```
shapiro.test(dogs$cyclopropane)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dogs$cyclopropane
## W = 0.93334, p-value = 0.4815
```

```
shapiro.test(dogs$halothane)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dogs$halothane
## W = 0.9234, p-value = 0.3862
```

Only the data from isofluorane show a normal distribution p-value = 0.03434. However the other values cyclopropane, and halothane clearly are not normal distribution. We conclude that these dogs are not from a normal population.

b) Investigate whether the columns isofluorane and halothane are correlated. Apply relevant tests to verify whether the distributions of these columns are different. Is a permutation test applicable?

```r
with(dogs, cor.test(isofluorane, halothane , method="spearman"))
```

```
## Warning in cor.test.default(isofluorane, halothane, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  isofluorane and halothane
## S = 128.89, p-value = 0.5436
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.218846
```

```r
if (!require("coin")) install.packages("coin")
```

```
## Loading required package: coin
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'coin'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     alpha
```

```r
library(coin)
```

```r
independence_test(dogs$isofluorane ~ dogs$halothane)
```

```
##
##  Asymptotic General Independence Test
##
## data:  dogs$isofluorane by dogs$halothane
## Z = 0.47487, p-value = 0.6349
## alternative hypothesis: two.sided
```

To test whether isofluorane and halothane is correlated it is used a spearman correlation test. The result show small correlation according to Cohen rho = 0.218846. Therefore we conclude that the correlation is small however a permuation test is applicable because of a small sample size. However the permutation test also does not show a significant correlation p-value = 0.6349.

c) Conduct a one-way ANOVA to determine whether the type of drug has an effect on the concentration of plasma epinephrine. Give the estimated concentrations for each of the three anesthesia drugs.

```r
data2 <- read.table("data/dogs.txt", header=TRUE)
attach(data2)

combined<-data.frame(cbind(isofluorane, halothane, cyclopropane))
stacked <- stack(combined)
```

```
anova_results <-aov(values ~ ind, data = stacked )
summary(anova_results)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## ind          2  1.081  0.5404   5.355  0.011 *
## Residuals   27  2.725  0.1009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First of all the data has been conducted to one column of data set and then a one way anova has been conducted. The result of this one way anova is significant ; p = 0,011. From this it has to be conluded that the type of the drug (isofluorane/halothane/cyclopropane) makes.

D)Does the Kruskal-Wallis test arrive at the same conclusion about the effect of drug as the test in c)? Explain possible differences between conclusions of the Kruskal-Wallis and ANOVA tests.

```
kruskal.test(values ~ ind, data = stacked)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  values by ind
## Kruskal-Wallis chi-squared = 5.6442, df = 2, p-value = 0.05948
```

The Kruskal-Wallis test did not arrive at the same conclusion as the one way anova. The result was insignificant p-value = 0.05948 which showed a significant result p = 0.011. compared to the anova The Kruskal-Wallis test is a non-parametric counterpart of ANOVA which does not rely on normality but on ranks thereby a bit less powerful results than 1-way ANOVA.

Exercise 4. Hemoglobin in trout.

a) Present an R-code for the randomization process to distribute 80 fishes over all combinations of levels of factors rate and method.

```
BLOOD <- read.table("data/hemoglobin.txt", header=TRUE)
summary(BLOOD)
```

```
##    hemoglobin          rate         method
##  Min.   : 5.500   Min.   :1.00   Length:80
##  1st Qu.: 7.350   1st Qu.:1.75   Class :character
##  Median : 8.700   Median :2.50   Mode  :character
##  Mean   : 8.736   Mean   :2.50
##  3rd Qu.:10.200   3rd Qu.:3.25
##  Max.   :11.900   Max.   :4.00
```

```
set.seed(42)
rows <- sample(nrow(BLOOD))
randomized <- BLOOD[rows, ]
```

b) Perform the two-way ANOVA to test for effects of factors rate, method and their interaction on the response variable hemoglobin. Comment on your findings.

```
res.aov3 <- aov(hemoglobin ~ rate * method, data = randomized)
summary(res.aov3)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## rate          1  27.93  27.931  11.933 0.000905 ***
## method        1   2.42   2.415   1.032 0.312963
## rate:method   1   1.24   1.243   0.531 0.468373
## Residuals    76 177.90   2.341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the Two Way Anova analysis show that Rate has a significant effect F =11.933, p =0.000905, but method has no significant effect F = 1.032, p= 0.312963. Furthermore, there is no interaction effect between rate and method F = 0.531, p = 0.468373. The results show that only rate is a significant factor in influencing the Hemoglobin levels. The method that is used is not important because it does not influence the result, accordingly, the method does neither decrease nor increase the influence of the rate

c) Which of the two factors has the greatest influence? Is this a good question? Consider the additive model. Which combination of rate and method yield the highest hemoglobin? Estimate the mean hemoglobin value for rate 3 by using method A. What rate leads to the highest mean hemoglobin?

d) Test the null hypothesis that the hemoglobin is the same for all rates by a one-way ANOVA test, ignoring the variable method. Is it right/wrong or useful/not useful to perform this test on this dataset?

```
res.aov <- aov(hemoglobin ~ rate, data = randomized)
summary(res.aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## rate          1  27.93  27.931      12 0.000867 ***
## Residuals    78 181.55   2.328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes it is usefull because we already have shown that method has no influence at all (HOWEVER NOT SURE ABOUT THAT)

Exercise 5. Sour cream.

a) Analyze the data in a three-way experiment without interactions with acidity as response and starter, batch and position as factors.

By using summary command, can you tell whether there is a significant difference between the effects of starter 1 and starter 2 on acidity?

```
cream <- read.table("data/cream.txt", header=TRUE)
attach(cream)

model <- lm(acidity ~ batch + position + starter, data = cream)
```

```
summary(model)
```

```
##
## Call:
## lm(formula = acidity ~ batch + position + starter, data = cream)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9442 -1.2164 -0.0976  0.8296  3.7508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.3330     1.3373   5.484 1.93e-05 ***
## batch         0.3116     0.2483   1.255    0.223
## position     -0.0674     0.2483  -0.271    0.789
## starter       0.1992     0.2483   0.802    0.431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.756 on 21 degrees of freedom
## Multiple R-squared:  0.09839,    Adjusted R-squared:  -0.03041
## F-statistic: 0.7639 on 3 and 21 DF,  p-value: 0.527
```