kd

Pēteris Račinskis pr20015

4/7/2021

1. uzdevums

a) Aprakstīt visus mainīgos lielumus

```
library(MASS)
library(psych)
```

- npreg: grūtniecību skaits, int
- glu: glikozes koncentrācija, int
- bp: asinsspiediens, int
- skin: ādas locījuma biezums, int
- bmi: ķermeņa masas indekss, num
- ped: diabetes pedigree function, num
- age: vecums, int
- type: tips, kategoriāls ar 2 līmeņiem

b) Aprakstošā statistika

```
combined <- rbind(MASS::Pima.te,MASS::Pima.tr)
summary(combined)
```

```
##      npreg             glu               bp              skin
##  Min.   : 0.000   Min.   : 56.00   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 98.75   1st Qu.: 64.00   1st Qu.:22.00
##  Median : 2.000   Median :115.00   Median : 72.00   Median :29.00
##  Mean   : 3.517   Mean   :121.03   Mean   : 71.51   Mean   :29.18
##  3rd Qu.: 5.000   3rd Qu.:141.25   3rd Qu.: 80.00   3rd Qu.:36.00
##  Max.   :17.000   Max.   :199.00   Max.   :110.00   Max.   :99.00
##       bmi             ped              age            type
##  Min.   :18.20   Min.   :0.0850   Min.   :21.00   No :355
##  1st Qu.:27.88   1st Qu.:0.2587   1st Qu.:23.00   Yes:177
##  Median :32.80   Median :0.4160   Median :28.00
##  Mean   :32.89   Mean   :0.5030   Mean   :31.61
##  3rd Qu.:36.90   3rd Qu.:0.6585   3rd Qu.:38.00
##  Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

```
describe(combined)
```

```
##       vars   n   mean    sd median trimmed   mad   min    max  range skew
## npreg    1 532   3.52  3.31   2.00    3.06  2.97  0.00  17.00  17.00 1.14
```

```
## glu      2 532 121.03 31.00 115.00  118.65 29.65 56.00 199.00 143.00 0.61
## bp       3 532  71.51 12.31  72.00   71.52 11.86 24.00 110.00  86.00 0.00
## skin     4 532  29.18 10.52  29.00   28.91 10.38  7.00  99.00  92.00 0.68
## bmi      5 532  32.89  6.88  32.80   32.56  6.60 18.20  67.10  48.90 0.63
## ped      6 532   0.50  0.34   0.42    0.45  0.26  0.08   2.42   2.34 1.89
## age      7 532  31.61 10.76  28.00   29.95  8.90 21.00  81.00  60.00 1.27
## type*    8 532   1.33  0.47   1.00    1.29  0.00  1.00   2.00   1.00 0.71
##       kurtosis   se
## npreg     0.74 0.14
## glu      -0.35 1.34
## bp        0.77 0.53
## skin      2.86 0.46
## bmi       1.24 0.30
## ped       5.51 0.01
## age       1.15 0.47
## type*    -1.50 0.02
```

c) Šķērstabula

```
split <- combined
age_cat <- sapply(split$age,function(i) {
  if (i >= 30) {
    "30+"
  } else {
    "30-"
  }
})
split <- cbind(split,age_cat)
tab <- table(split$type, split$age_cat); tab
```

```
##
##        30- 30+
##   No  245 110
##   Yes  62 115
```

```
c <- chisq.test(tab); c
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 54.513, df = 1, p-value = 1.544e-13
```

Neatkarīgi?

```
conf <- 0.05
(c$p.value > conf)
```

```
## [1] FALSE
```

d) Salīdzināšana:

```
comp <- function(a,b) {
  print(describe(a))
  print(describe(b))
  t.test(a,b)
```

```
}
over30 <- split[age_cat == "30+",]
under30 <- split[age_cat == "30-",]
comp(over30$ped,under30$ped)
```

```
##    vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 225 0.54 0.36   0.44     0.5 0.29 0.08 2.33  2.24 1.55     3.77 0.02
##    vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 307 0.47 0.33    0.4    0.42 0.24 0.08 2.42  2.34 2.21     7.44 0.02

## 
##  Welch Two Sample t-test
## 
## data:  a and b
## t = 2.3127, df = 454.24, p-value = 0.02119
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01062184 0.13077676
## sample estimates:
## mean of x mean of y 
## 0.5437644 0.4730651
```

```
comp(over30$bmi,under30$bmi)
```

```
##    vars   n mean sd median trimmed  mad  min max range skew kurtosis  se
## X1    1 225 33.5  6   33.6   33.25 5.78 19.3  50  30.7 0.31    -0.04 0.4
##    vars   n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 307 32.45 7.44     32   32.03 7.26 18.2 67.1  48.9  0.8     1.58 0.42

## 
##  Welch Two Sample t-test
## 
## data:  a and b
## t = 1.7966, df = 525.03, p-value = 0.07297
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09792549  2.19433664
## sample estimates:
## mean of x mean of y 
##  33.49511  32.44691
```

```
comp(over30$ped,under30$ped)
```

```
##    vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 225 0.54 0.36   0.44     0.5 0.29 0.08 2.33  2.24 1.55     3.77 0.02
##    vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 307 0.47 0.33    0.4    0.42 0.24 0.08 2.42  2.34 2.21     7.44 0.02

## 
##  Welch Two Sample t-test
## 
## data:  a and b
## t = 2.3127, df = 454.24, p-value = 0.02119
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01062184 0.13077676
## sample estimates:
```

```
## mean of x mean of y
## 0.5437644 0.4730651
```

2. uzdevums

```
library(ggplot2)
library(qqplotr)
```

a) Aprakstošās statistikas

Datu ielāde:

```
d <- read.csv("kd.csv",header = F);str(d)
```

```
## 'data.frame':    135 obs. of  1 variable:
##  $ V1: int  30 113 81 115 9 2 91 112 15 138 ...
```

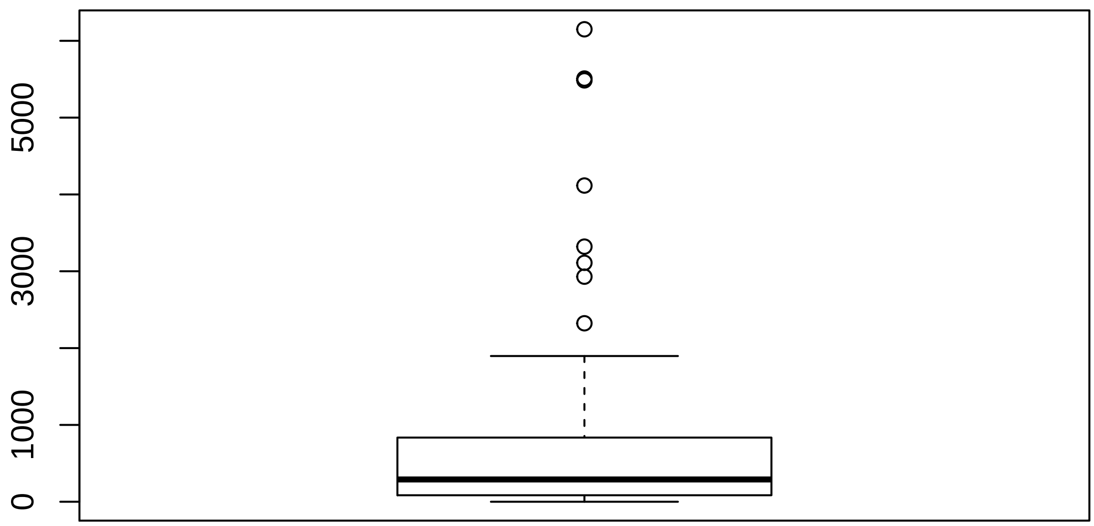
Aprakstošās statistikas:

```
describe(d)
```

```
##    vars   n   mean     sd median trimmed    mad min  max range skew kurtosis
## X1    1 135 656.88 1037.3    290  427.06 383.99   0 6150  6150  3.2    11.66
##       se
## X1 89.28
```

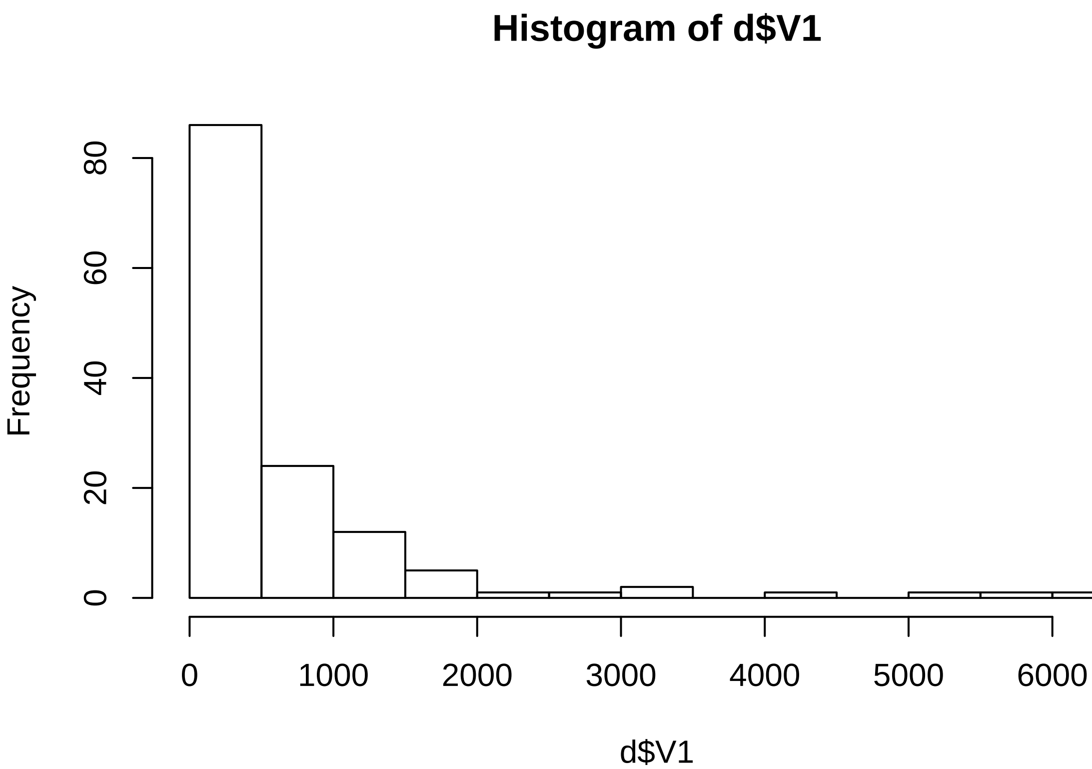
Kastu grafiks:

```
boxplot(d$V1)
```



Histogramma:

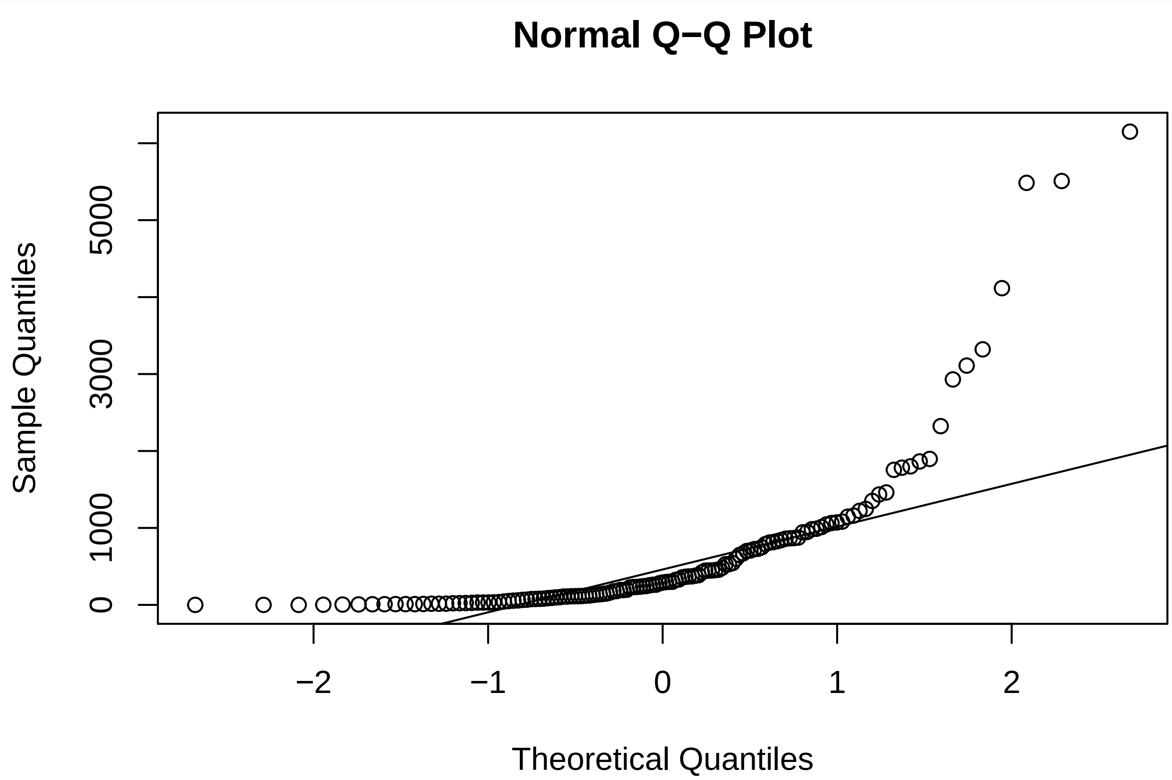
```
hist(d$V1)
```



Dati acīmredzami nav simetriski.

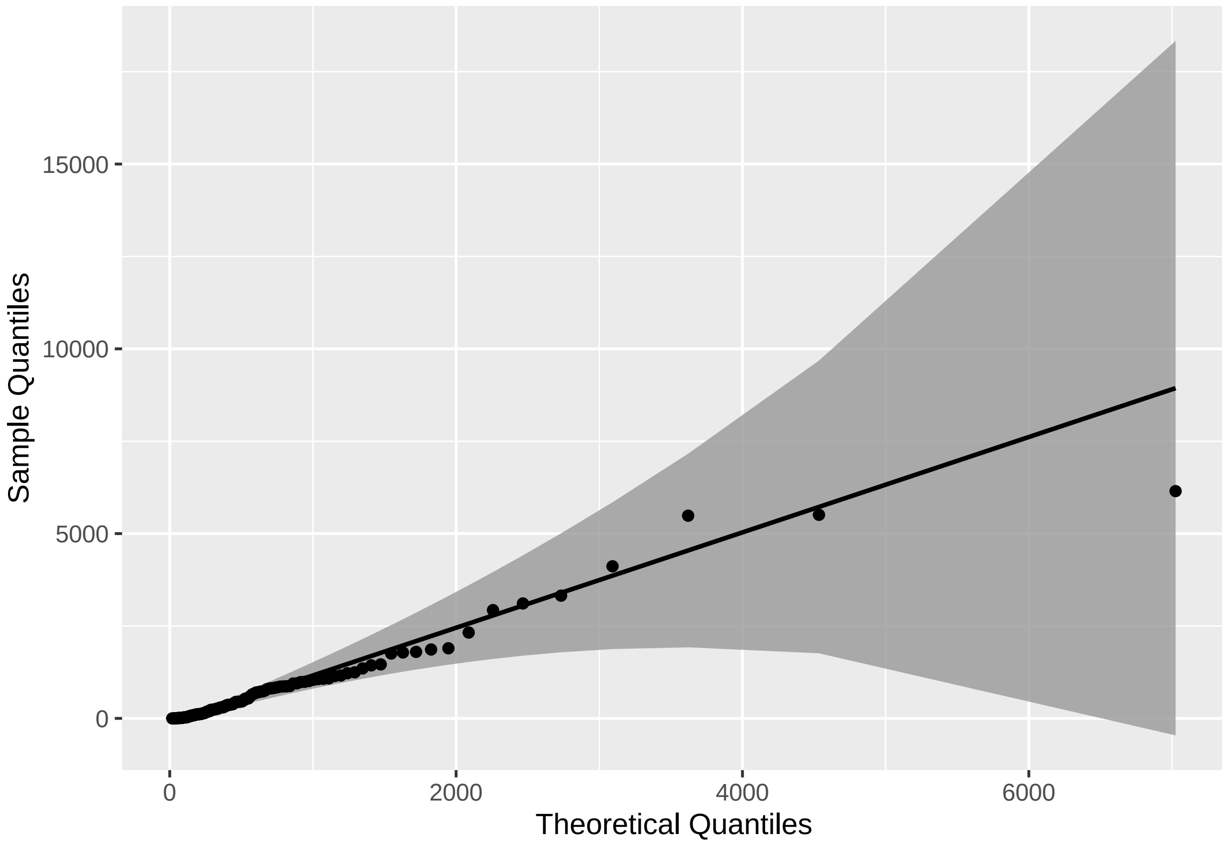
Kvantiļu-kvantiļu grafiks pret normālo sadalījumu:

```
qqnorm(d$V1)
qqline(d$V1)
```



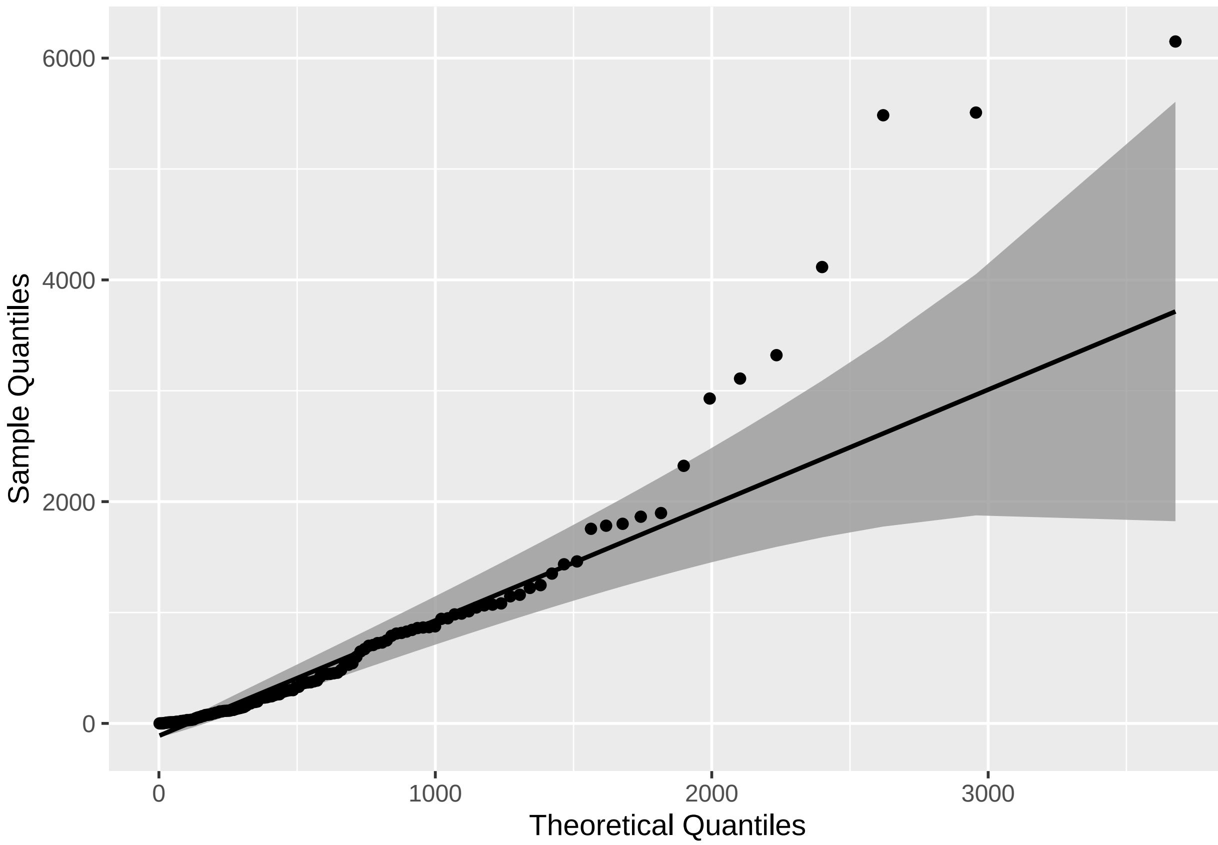
Salīdzinājums ar log-normālo sadalījumu:

```
mu <- mean(d$V1)
sigma <- sd(d$V1)
mlog <- log(mu^2 / sqrt(sigma^2 + mu^2))
slog <- sqrt(log(1 + (sigma^2 / mu^2)))
di <- "lnorm"
dp <- list(meanlog=mlog, sdlog = slog)

gg <- ggplot(data = d, mapping = aes(sample = V1)) +
    stat_qq_band(distribution = di, dparams = dp) +
    stat_qq_line(distribution = di, dparams = dp) +
    stat_qq_point(distribution = di, dparams = dp) +
    labs(x = "Theoretical Quantiles", y = "Sample Quantiles")
gg
```



Salīdzinājums ar eksponenciālo sadalījumu:

```
di <- "exp" # exponential distribution
dp <- list(rate = 1/mean(d$V1)) # exponential rate parameter

gg <- ggplot(data = d, mapping = aes(sample = V1)) +
    stat_qq_band(distribution = di, dparams = dp) +
    stat_qq_line(distribution = di, dparams = dp) +
    stat_qq_point(distribution = di, dparams = dp) +
    labs(x = "Theoretical Quantiles", y = "Sample Quantiles")
gg
```



b) Testi:

```
require(goft)
x <- sapply(d$V1,function(i){
  i + 0.01
})
exp_test(x)
```

```
##
##  Test for exponentiality based on a transformation to uniformity
##
## data:  x
## T = 5.5093, p-value < 2.2e-16
```

```
lnorm_test(x)
```

```
##
##  Test for the lognormal distribution based on a transformation to
##  normality
##
## data:  x
## p-value = 8.981e-11
```

Spriežot pēc testiem, dati neatbilst ne vienam, ne otram sadalījumam - bet pozitīvu rezultātu principā ir grūti iegūt. Datu kopā ir vairākas 0 vērtības, kas varētu bojāt visu procesu. Pēc ticamības novērtējuma it kā lnorm ir labāks, taču pēc formas exp izskatās tuvāks - it sevišķi, jo lnorm nevajadzētu būt tik daudzām vērtībām tuvām 0.

Lognormālā un eksponenciālā sadalījuma piemeklēšana:

```
library(maxLik)
llf_lnorm <- function(param) {
  mu <- param[1]
  sd <- param[2]
  llValue <- dlnorm(x, mean=mu, sd=sd, log=TRUE)
  sum(llValue)
}
llf_exp <- function(param) {
  lambda <- param[1]
  n <- length(x)
  n*log(lambda)-lambda*sum(x)
}
```

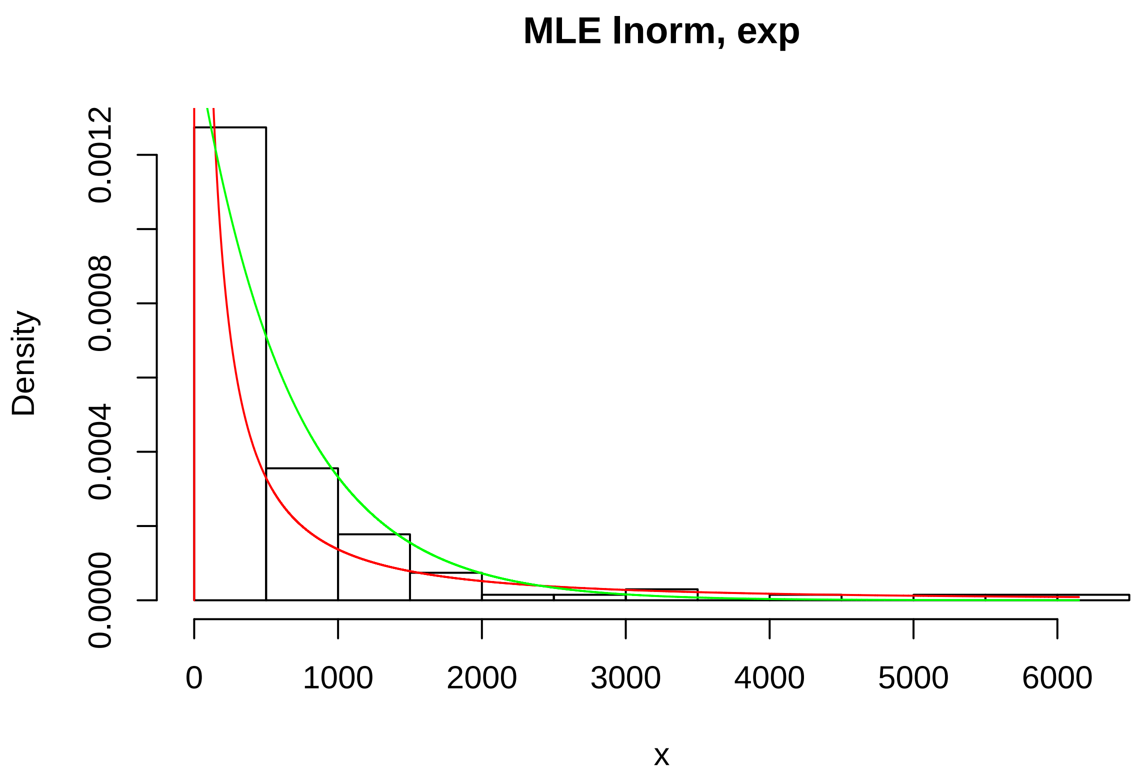
```
mu_init = 3.5
sd_init = 0.6
mllnm <- maxLik(llf_lnorm, start=c(mu_init,sd_init))
summary(mllnm)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 10 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: -1009.312
## 2  free parameters
## Estimates:
##      Estimate Std. error t value Pr(> t)
## [1,]   5.2656     0.1899   27.73  <2e-16 ***
## [2,]   2.2076     0.1344   16.43  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

```
lambda_init <- 1/mean(x)
mlexp <- maxLik(llf_exp, start=c(lambda_init))
summary(mlexp)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 1 iterations
## Return code 2: successive function values within tolerance limit (tol)
## Log-Likelihood: -1010.815
## 1  free parameters
## Estimates:
##       Estimate Std. error t value Pr(> t)
## [1,] 0.0015223  0.0001311   11.61  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

```
hist(x,prob=T,main="MLE lnorm, exp")
xax <- seq(0,max(x),0.1)
lines(xax,dlnorm(xax,mllnm$estimate[1],mllnm$estimate[2]),col="red")
lines(xax,dexp(xax,mlexp$estimate[1]),col="green")
```



c) Pārliecības intervāli

Nav īsti saprotams, kam šeit varētu veikt t-testu vai Wilkoksona testu, taču tīri mehāniski to var izdarīt. Šie testi ir paredzēti izlašu novērtēšanai attiecīgi pie normāli un simetriski sadalītiem lielumiem. Varbūt domāts kaut kā izmantot šos testus lai pārbaudītu izlases atbilstību sadalījumiem, izmantojot kādus matemātiskus pārveidojumus, taču jau pastāv iebūvēti testi šāda veida pārbaudēm.

```
t.test(x)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 7.3579, df = 134, p-value = 1.684e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  480.3181 833.4649
## sample estimates:
## mean of x
##  656.8915
```

```
wilcox.test(x)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  x
## V = 9180, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

3. uzdevums

Nosacījums, lai blīvuma funkcija būtu korekta: integrālim jābūt 1 pār visu definīcijas apgabalu; t.i., varbūtībai kopsumā jābūt = 1

```
get_c <- function(cb,lower,upper) {
  1 / integrate(cb,lower,upper)$value
}
a <- function(x){
  x^2
}
b <- function(x){
  x * exp(-x)
}
get_c(a,0,1)
```

```
## [1] 3
```

```
get_c(b,0,Inf)
```

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

4. uzdevums

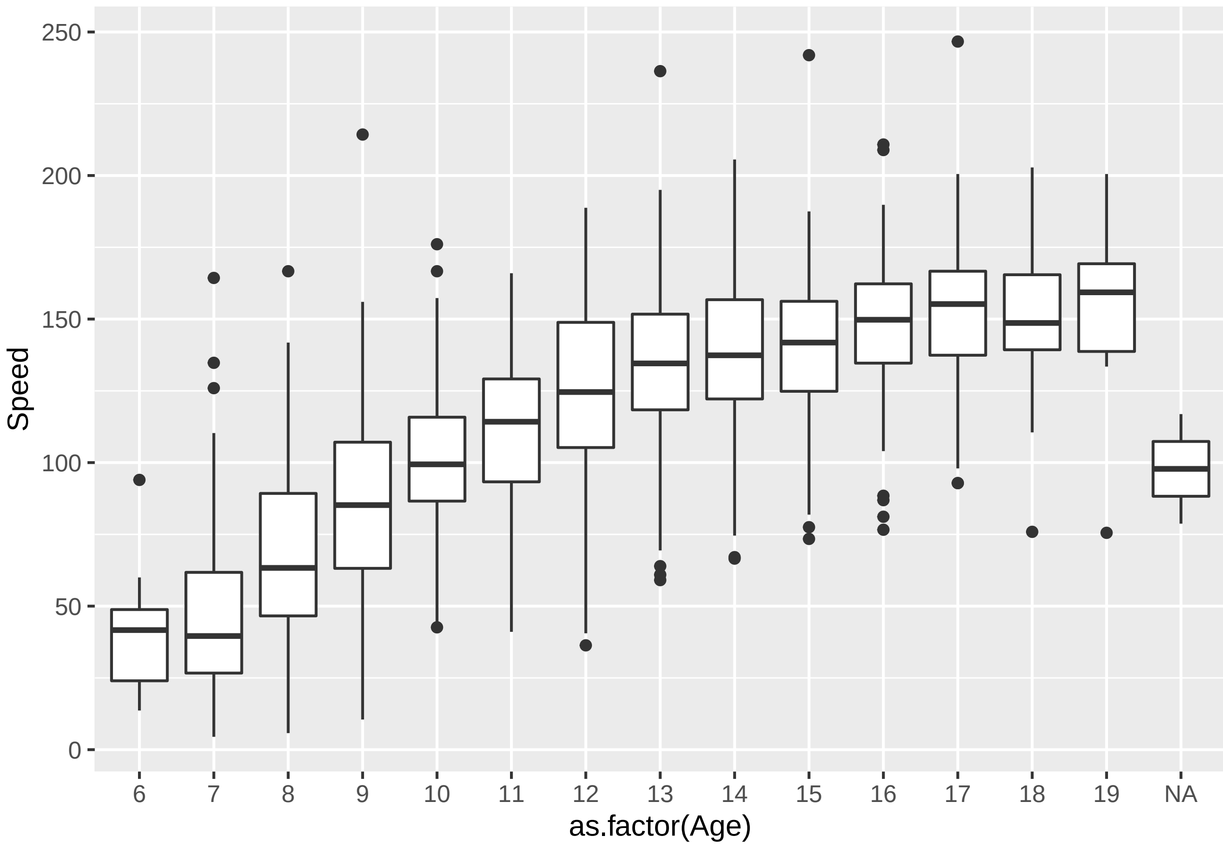
```
library(readxl)
d <- read_excel('kd_data.xls')
summary(d)
```

```
##      Sex                 Speed          L(sek) P, A     L(sek) P, B
##  Length:2166        Min.   :  4.482   Min.   : 14.00   Length:2166
##  Class :character   1st Qu.: 81.073   1st Qu.: 33.00   Class :character
##  Mode  :character   Median :115.865   Median : 43.00   Mode  :character
##                     Mean   :110.542   Mean   : 48.04
##                     3rd Qu.:142.470   3rd Qu.: 57.00
##                     Max.   :246.679   Max.   :447.00
##                     NA's   :219       NA's   :1137
##       Age
##  Min.   : 6.00
##  1st Qu.: 9.00
##  Median :11.00
##  Mean   :11.53
##  3rd Qu.:14.00
##  Max.   :19.00
##  NA's   :2
```

a) Kastu grafiks:

```
ggplot(data=d,aes(x=as.factor(Age),y=Speed)) + geom_boxplot()
```

```
## Warning: Removed 219 rows containing non-finite values (stat_boxplot).
```



Izskatās, ka lasītprasme īpaši nemainās pēc 14-15 gadu vecuma.

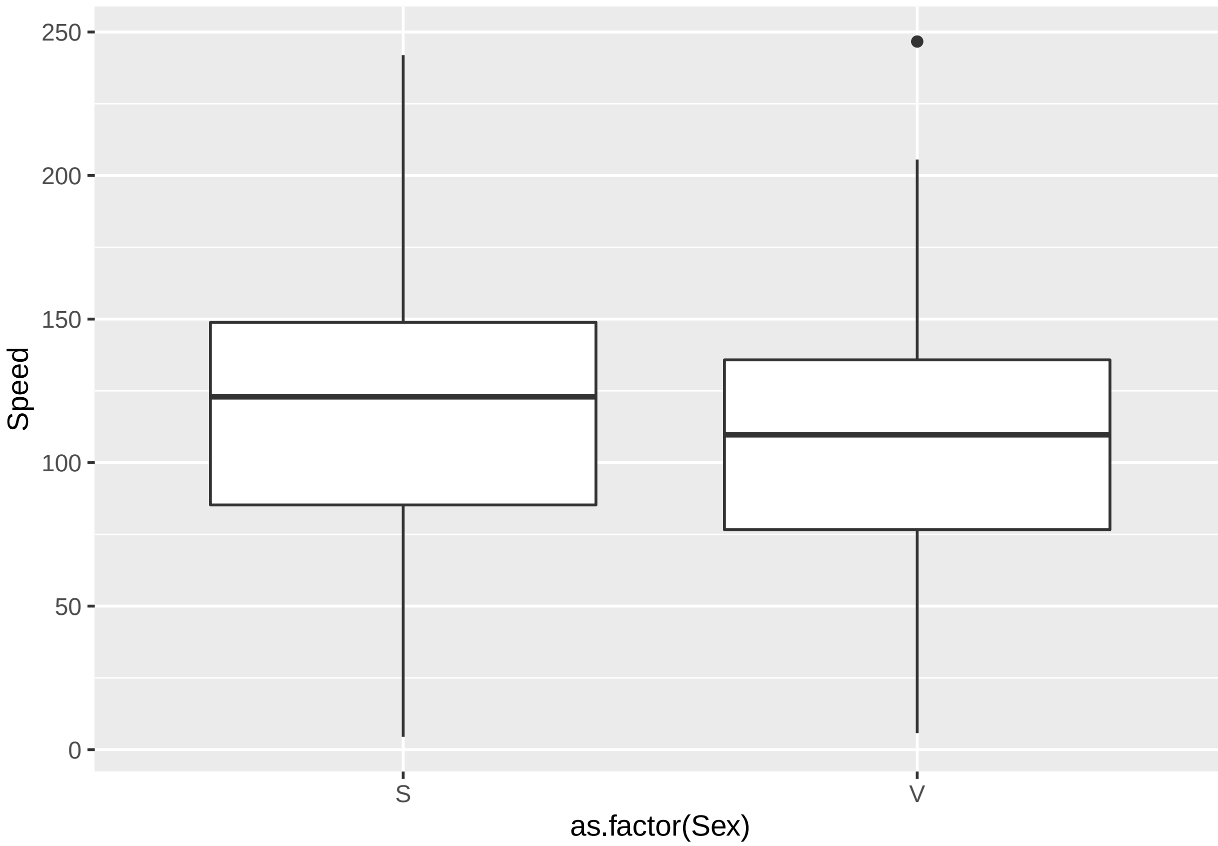
b) Salīdzinājums pa dzimumiem

```
attach(d)
df <- subset(d, Sex == "V" | Sex == "S"); df
```

```
## # A tibble: 2,095 x 5
##    Sex   Speed `L(sek) P, A` `L(sek) P, B`   Age
##    <chr> <dbl>         <dbl> <chr>         <dbl>
##  1 V     17.7             NA <NA>              7
##  2 V     13.0             NA <NA>              7
##  3 V     58.4             80 <NA>              7
##  4 S     47.6            159 <NA>              7
##  5 V     13.1            149 <NA>              7
##  6 S     43.0            128 <NA>              7
##  7 S     29.2            108 <NA>              7
##  8 S     26.3             94 <NA>              7
##  9 V     15.3             NA <NA>              7
## 10 V      5.76            NA <NA>              8
## # ... with 2,085 more rows
```

```
ggplot(data=df,aes(x=as.factor(Sex),y=Speed)) + geom_boxplot()
```

```
## Warning: Removed 151 rows containing non-finite values (stat_boxplot).
```



Salīdzinājums 7 gadu vecumā:

```
age7 <- subset(df, Age == 7)
n <- length(age7)
p1 <- age7[age7$Sex == "V",]$Speed
p2 <- age7[age7$Sex == "S",]$Speed
tres <- t.test(p1,p2,conf=0.95); tres
```

```
##
##  Welch Two Sample t-test
##
## data:  p1 and p2
## t = 0.36408, df = 162.38, p-value = 0.7163
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.438025  9.348677
## sample estimates:
## mean of x mean of y
##  46.98225  45.52693
```

```
alpha <- 0.025
cutoff <- qt(1-alpha,n-1); cutoff
```

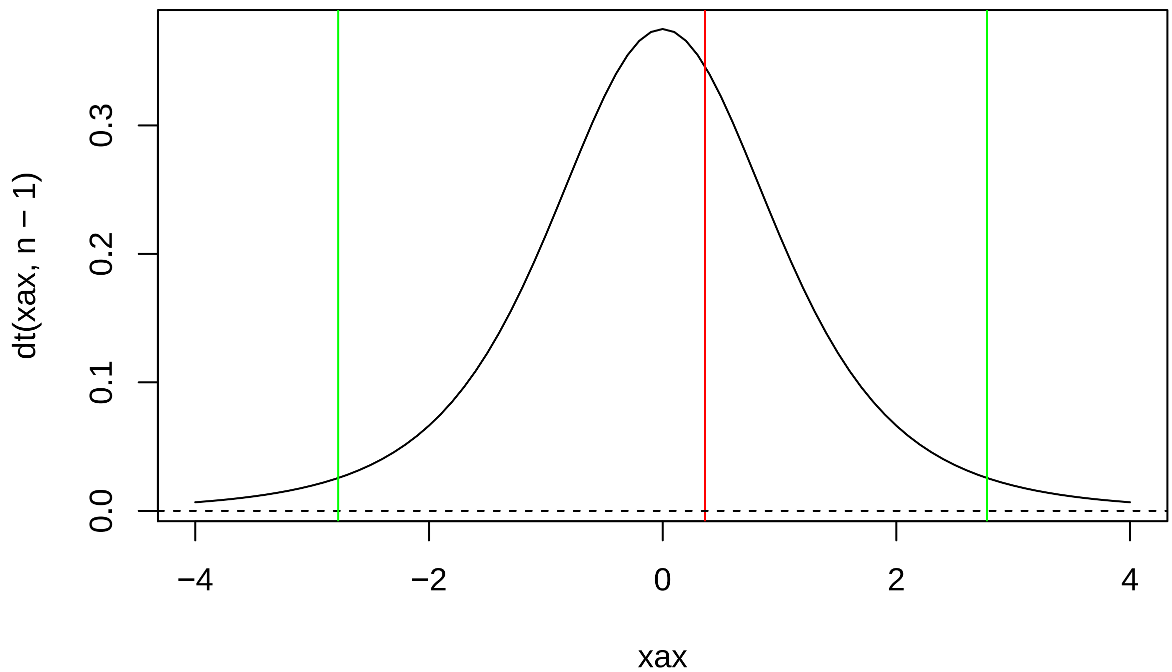
```
## [1] 2.776445
```

```
tres$conf.int
```

```
## [1] -6.438025  9.348677
## attr(,"conf.level")
```

```
## [1] 0.95
```

```
xax <- seq(-4,4,0.1)
plot(xax,dt(xax,n-1),type="l")
abline(v=tres$statistic,col="red")
abline(h=0,lty=2)
abline(v=cutoff,col="green")
abline(v=-cutoff,col="green")
```



Ar Vilkoksona testu:

```
wilcox.test(p1,p2,paired=F,exact=F)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  p1 and p2
## W = 4417.5, p-value = 0.8658
## alternative hypothesis: true location shift is not equal to 0
```

Abos gadījumos H0 netiek noraidīta, p-vērtības līdzīgas.

14 gadu vecumā:

```
age14 <- subset(df, Age == 14)
n <- length(age14)
p1 <- age14[age14$Sex == "V",]$Speed
p2 <- age14[age14$Sex == "S",]$Speed
tres <- t.test(p1,p2,conf=0.95); tres
```

```
##
##  Welch Two Sample t-test
##
## data:  p1 and p2
## t = -3.7833, df = 171.36, p-value = 0.0002136
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -21.294383  -6.692488
## sample estimates:
## mean of x mean of y
##  131.6360  145.6294
```

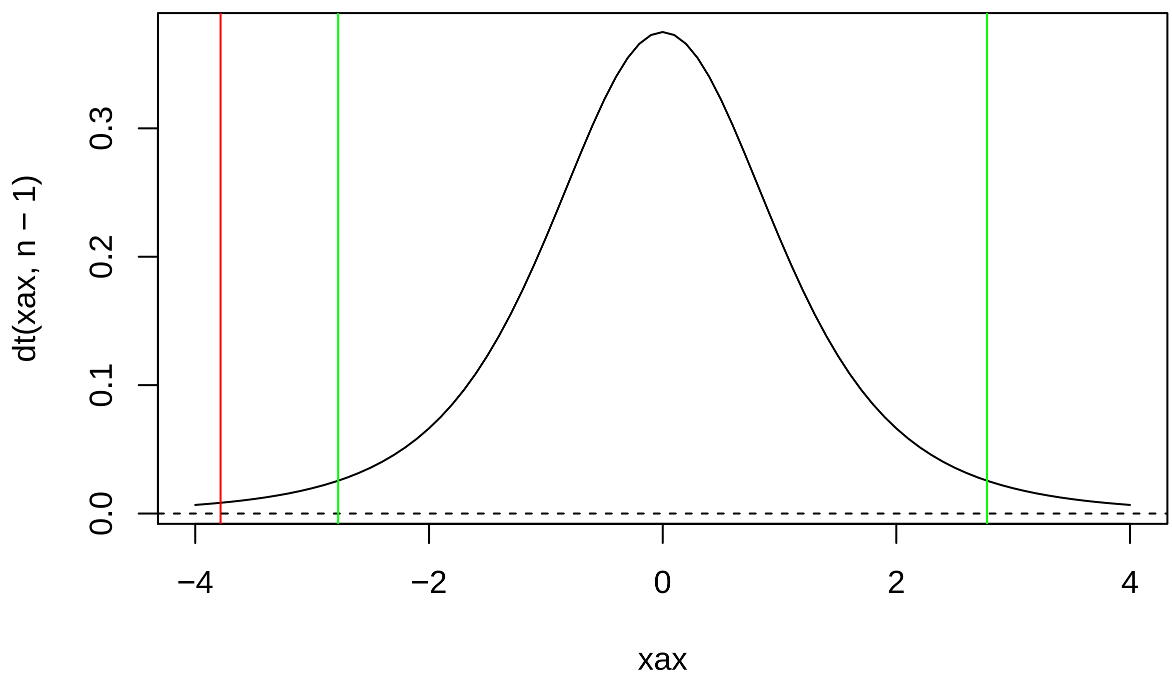
```
alpha <- 0.025
cutoff <- qt(1-alpha,n-1); cutoff
```

```
## [1] 2.776445
```

```
tres$conf.int
```

```
## [1] -21.294383  -6.692488
## attr(,"conf.level")
## [1] 0.95
```

```
xax <- seq(-4,4,0.1)
plot(xax,dt(xax,n-1),type="l")
abline(v=tres$statistic,col="red")
abline(h=0,lty=2)
abline(v=cutoff,col="green")
abline(v=-cutoff,col="green")
```



Ar Vilkoksona testu:

```
wilcox.test(p1,p2,paired=F,exact=F)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  p1 and p2
## W = 3370.5, p-value = 0.0003148
## alternative hypothesis: true location shift is not equal to 0
```

Abos gadījumos H0 tiek noraidīta, p-vērtības līdzīgas.

c) Normalitātes pārbaude vecumiem 7-14 (ne visur ir vismaz 4 lielumi) (uzdevumu nepaspēju pabeigt)

```
require(nortest)
norms <- function(d) {
  s <- d[d$Sex == "S",]$Speed
  v <- d[d$Sex == "V",]$Speed
  list(S=lillie.test(s)$p.value,V=lillie.test(v)$p.value)
}
sapply(7:14, function(i) norms(subset(df, Age == i)))
```

```
##   [,1]         [,2]        [,3]       [,4]       [,5]       [,6]      [,7]
## S 0.0003100828 0.009791243 0.8489264  0.01819721 0.03925419 0.1990035 0.259164
## V 0.0002109343 0.01320186  0.03708332 0.07400062 0.6879285  0.6705764 0.9847893
##   [,8]
## S 0.03646765
## V 0.1725062
```