Leistungsnachweis

CAS Datenanalyse 20.11 Modul A2

Stefan Schmidt

03.05.2020

1. COVID-19 Pandemie in der Schweiz

Datenbasis

Es werden COVID-19 Infektionsdaten von Wikipedia vom 14.04. und 16.04.2020 verwendet. Ausserdem enthält das verwendete Data Frame Angaben zur Einwohnerzahl der Kantone (Quelle: A.Ruckstuhl COVID-19 Arbeitsblatt *CAS-DA_ModulA2-HT3_Coronavirus.R*).

Spalten des Data Frames `df`:

- `kanton`: Kürzel des Kantons
- `inf_1404`: Anzahl COVID-19 infizierter Personen am 14.04.2020
- `inf_1604`: Anzahl COVID-19 infizierter Personen am 16.04.2020
- `einw10k`: Einwohnerzahl (in 10'000)

Infektionen per 10'000 Einwohner

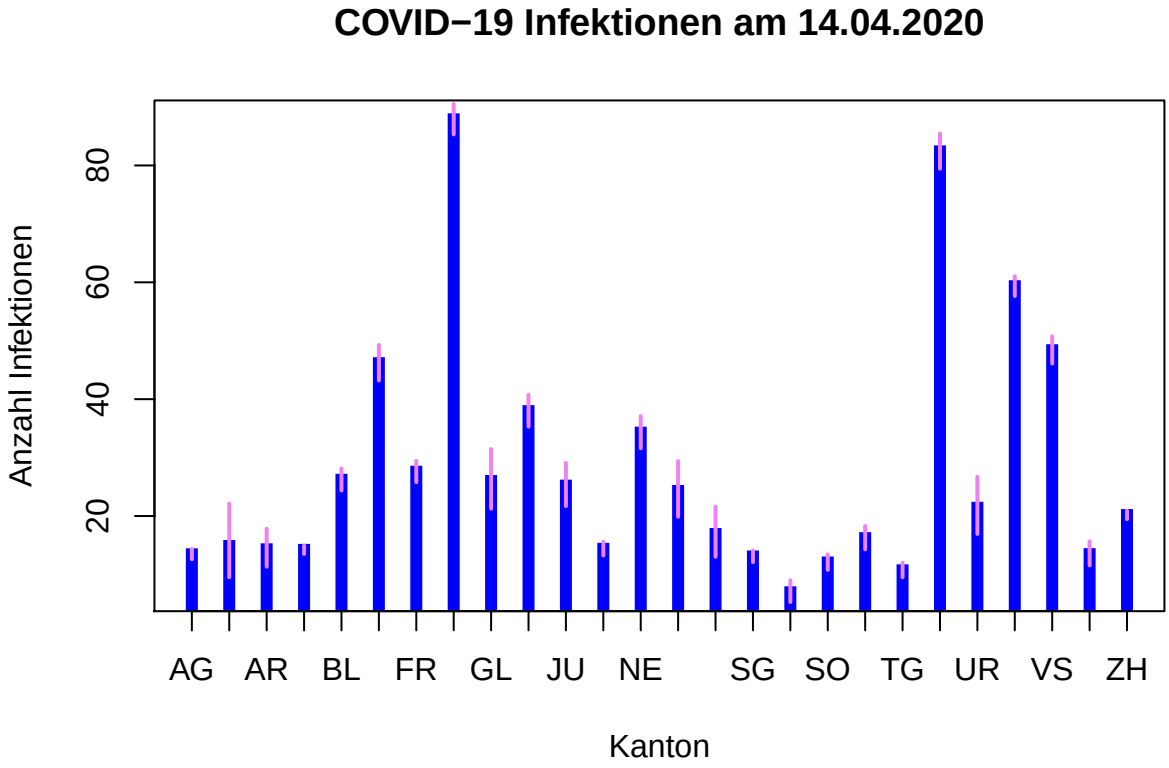
Die Kantone haben unterschiedliche Einwohnerzahlen, daher wird im Folgenden für jeden Kanton die Anzahl der COVID-19 Infektionen per 10'000 Einwohner graphisch dargestellt, die Vertrauensintervalle berechnet und eingezeichnet (lila).

```
i1404_10k <- df$inf_1404 / df$einw10k

VI <- matrix(NA, nrow = nrow(df), ncol = 2)

for(i in 1:nrow(df)){
  pt <- poisson.test(x = df$inf_1404[i], T = df$einw10k[i])
  VI[i,] <- pt$conf.int
}

plot(i1404_10k,
     main = "COVID-19 Infektionen am 14.04.2020",
     xlab = "Kanton", ylab = "Anzahl Infektionen",
     type = "h", lwd = 6, col = "blue", lend = 2, xaxt = "n")
axis(1, at = 1:length(i1404_10k), labels = df$kanton)
segments(1:nrow(df), VI[,1], 1:nrow(df), VI[,2], col = "violet", lwd = 2)
```



Schätzen des Parameters λ

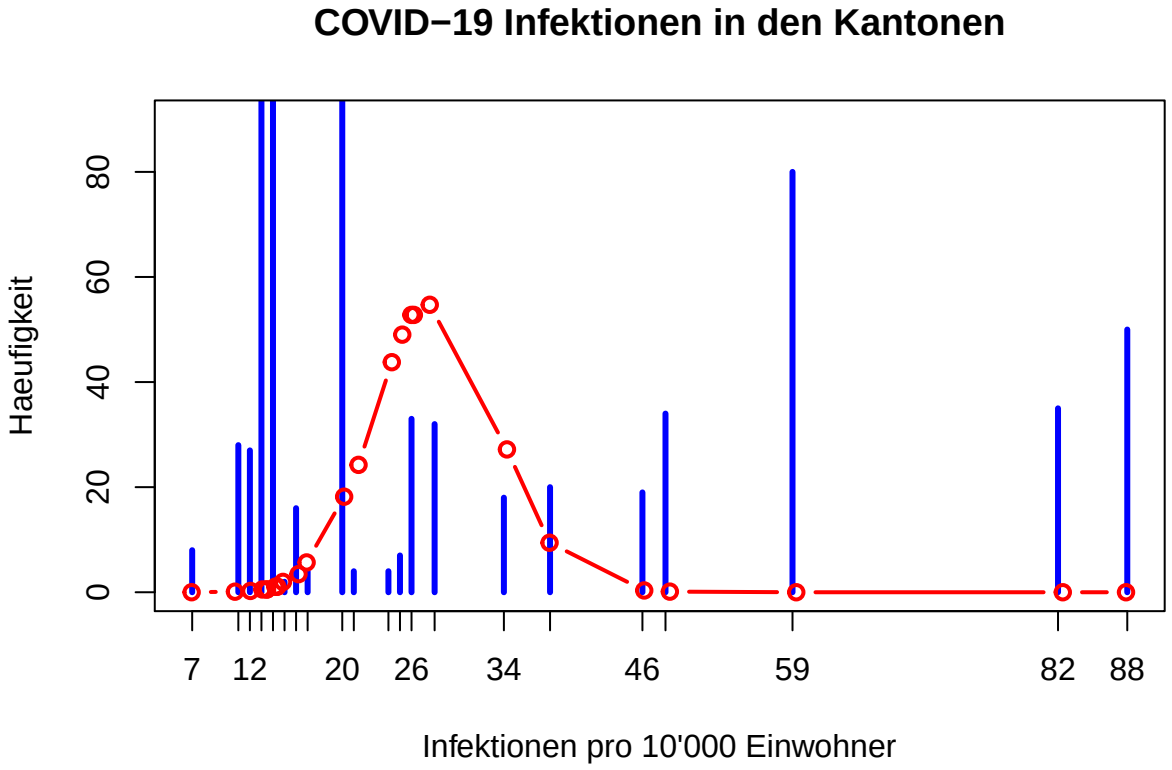
Der Mittelwert schätzt den Parameter λ.

```
mu <- mean(i1404_10k) # lambda: mittlere Anzahl per 10'000
```

Es ergibt sich für λ also ein Wert von 27.99.

Ein mit dem geschätzten Wert für λ angepasstes Modell wird in die tatsächliche Häufigkeit der Infektionsrate eingezeichnet (rot).

```
n <- sum(i1404_10k)   # n: totale Anzahl Infektionen / pro 10'000
yModel <- n * dpois(sort(round(i1404_10k)), lambda = mu)
plot(table(rep(round(i1404_10k), round(df$einw10k))),
     main = "COVID-19 Infektionen in den Kantonen",
     xlab = "Infektionen pro 10'000 Einwohner",
     ylab = "Haeufigkeit", ylim = c(0, 90),
     col = "blue", type = "h", lwd = 3)
lines(sort(i1404_10k), yModel, type="b", lwd=2, col="red")
```



Testen des Modells

Mit dem geschätzten Wert für λ prüfen wir nun die Plausibilität der Infektionsraten für den Kanton Waadt.

```
# Infektionen / 10'000 Einwohnern
# Waadt am 14.04.2020:
as.character(df$kanton[23])
```

```
## [1] "VD"
```

```
i1404_10k[23]
```

```
## [1] 59.3259
```

Frage: Ist diese Infektionsrate bei angenommener Poisson-Verteilung und Signifikanzniveau von 95% plausibel?

```
poisson.test(x = round(i1404_10k[23]), r = mu, conf.level = 0.95)
```

```
##
##  Exact Poisson test
##
## data:  round(i1404_10k[23]) time base: 1
## number of events = 59, time base = 1, p-value = 3.414e-07
## alternative hypothesis: true event rate is not equal to 27.99269
## 95 percent confidence interval:
##  44.91353 76.10570
## sample estimates:
## event rate
##         59
```

Antwort: Nein, aufgrund des niedrigen p-Werts (3.414e-07) muss die Nullhypothese verworfen werden, dass die Infektionsrate Poisson-verteilt ist.

Bootstrap-Verteilung der Dispersion

Prüfen wir das Modell durch Bootstrap-Simulation der Dispersion:

```
library(boot)
## Bootstrap-Vertrauensintervall für die Dispersion
f.disp <- function(x, ind){
  ## x   = ursprünglicher Beobachtungsvektor
  ## ind = Beobachtungsnummer für die Bootstrap-Stichprobe
  xx <-x[ind]       # erzeugen der Bootstrap-Stichprobe
  var(xx)/ mean(xx) # Berechnet die Dispersion für die Bootstrap-Stichprobe
}

set.seed(seed=123)
inf.boot2 <- boot(i1404_10k, f.disp, R=999, stype="i")
boot.ci(inf.boot2, conf=0.95, type="perc")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 999 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = inf.boot2, conf = 0.95, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   ( 5.83, 22.56 )
## Calculations and Intervals on Original Scale
```

Da die Nullhypothese $\sigma^2/xq = 1$ NICHT im 95%-Vertrauensintervall liegt kann die Nullhypothese auf dem 2.5% Niveau verworfen werden.

Folgerung: Die Poisson-Verteilung ist nicht geeignet um die COVID-19 Infektionen in der Schweiz zu beschreiben.

χ^2-Test

Zuletzt prüfen wir das Modell noch mit dem χ^2-Test:

```
chisq.test(i1404_10k)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  i1404_10k
## X-squared = 399.61, df = 25, p-value < 2.2e-16
```

Auch aufgrund des P-Werts des χ^2-Tests wird die Null-Hypothese "Daten können durch eine Poisson-Verteilung beschrieben werden" verworfen.

2. Internetnutzung in der Schweiz

Datenbasis

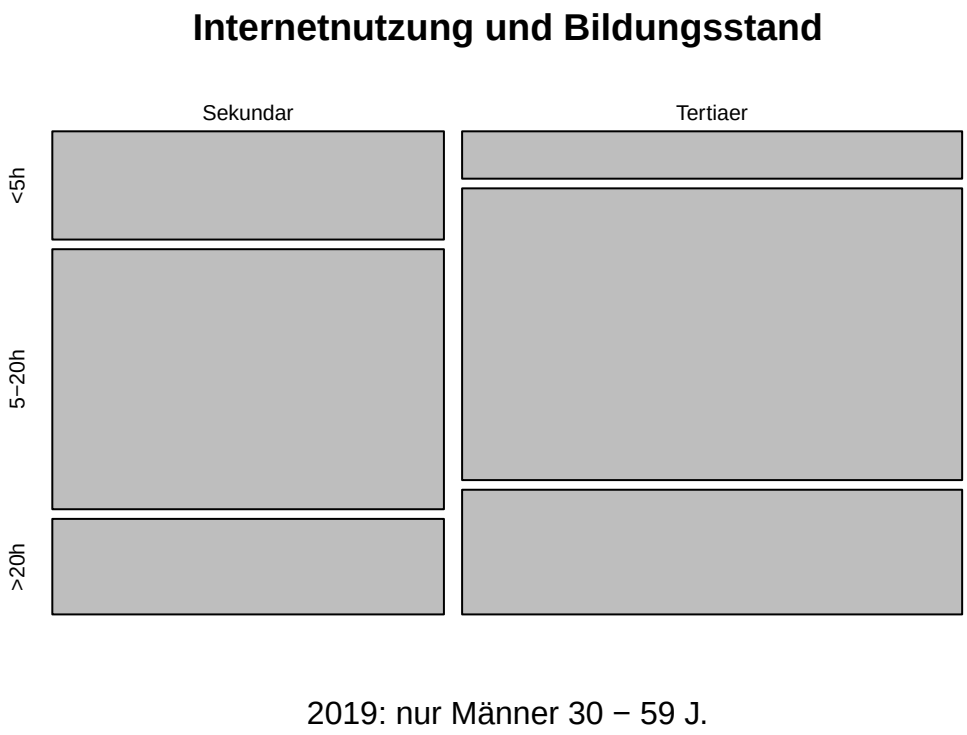
Der bereitgestellten Datei *Internetnutzung_korr.xlsx* wurde für das Jahr 2019 folgende Kontingenztabelle für Ausbildungsstufe und Internetnutzung von Männer zwischen 30 und 59 Jahren entnommen:

```
kt <- rbind(c(102, 245, 90), c(57, 351, 150))
dimnames(kt) <- list(c("Sekundar", "Tertiaer"), c("<5h", "5-20h", ">20h"))
kt
```

```
##          <5h 5-20h >20h
## Sekundar 102   245   90
## Tertiaer  57   351  150
```

Diese lässt sich im Mosaicplot darstellen.

```
mosaicplot(kt, main = "Internetnutzung und Bildungsstand", sub = "2019: nur Männer 30 - 59 J.")
```



Schätzung von Erfolgswahrscheinlichkeit π

Die Erfolgswahrscheinlichkeit π wird durch pi = X / m geschätzt.

Dafür, dass ein Mann (30 bis 59 J.) mit hohem Bildungsstand (Tertiärstufe) mehr als 20h das Internet nutzt ist die Erfolgswahrscheinlichkeit π hier:

```
X <- kt[2, 3]
m <- sum(kt)

X / m
```

```
## [1] 0.1507538
```

Testen von π

Frage: Ist $\pi = 0.15$ plausibel, wenn man von einer Erfolgswahrscheinlichkeit von 11% (3 Bildungsstufen und 3 Stufen der Internetnutzung: 1/3 * 1/3) ausgeht?

```
binom.test(x = X, n=m, p=1/3 * 1/3)
```

```
##
##  Exact binomial test
##
## data:  X and m
## number of successes = 150, number of trials = 995, p-value = 0.0001484
## alternative hypothesis: true probability of success is not equal to 0.1111111
## 95 percent confidence interval:
##  0.1290780 0.1745185
## sample estimates:
## probability of success
##              0.1507538
```

Antwort: Nein, aufgrund des P-Wertes von 0.0001484 wird die Nullhypothese verworfen.

Vertrauensintervall für π

Bestimmen wir das Vertrauensintervall für π:

```
binom.test(x = X, n = m, conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  X and m
## number of successes = 150, number of trials = 995, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1290780 0.1745185
## sample estimates:
## probability of success
##              0.1507538
```

Für π sind also Werte zwischen 0.129 und 0.174 plausibel.

Test auf Homogenität

```
chisq.test(kt)
```

```
##
##  Pearson's Chi-squared test
##
## data:  kt
## X-squared = 32.352, df = 2, p-value = 9.437e-08
```

Da der P-Wert von 9.437e-08 kleiner als das Niveau von 5% ist, wird die Nullhypothese "die Verteilung der Internetnutzung ist gleich für jeden Bildungsstand" auf dem 5% Niveau verworfen.

3. Wasserverbrauch Zürich

Datenbasis

Die bereitgestellte Datei *WasserverbrauchKtZH.csv* wird geladen, in Tabellenform gebracht (jedes Jahr eine Spalte) und die Spaltennamen bereinigt. Für weitere Berechnungen wird eine Spalte der Verbrauchsdifferenz 2018 - 2017 angehängt.

```
library(reshape2)
library(janitor)
path <- "/Users/schmis12/wrk/studio/ZHAW_CAS_Data_Analysis/Leistungsnachweis_A2/data/"
wv <- read.csv(paste0(path, "WasserverbrauchKtZH.csv"), na.strings = c("null"))
wv <- dcast(wv, BFS_NR + GEBIET ~ JAHR, value.var = "WvpTE")
wv <- clean_names(wv)
wv$d2018_2017 <- wv$x2018 - wv$x2017
head(wv)
```

```
##   bfs_nr             gebiet x2006 x2007 x2008 x2009 x2010 x2011 x2012 x2013
## 1      0   Bezirk Affoltern   276   252   244   233   232   233   237   237
## 2      0 Bezirk Andelfingen   334   317   303   282   281   276   248   282
## 3      0      Bezirk Bülach   288   277   268   260   254   247   249   250
## 4      0   Bezirk Dielsdorf   311   282   278   283   262   266   256   262
## 5      0    Bezirk Dietikon   319   303   296   291   289   281   278   289
## 6      0      Bezirk Hinwil   290   273   264   261   241   245   236   233
##   x2014 x2015 x2016 x2017 x2018 d2018_2017
## 1   228   234   223   226   236         10
## 2   282   300   286   298   297         -1
## 3   237   257   245   259   256         -3
## 4   249   265   245   247   261         14
## 5   273   272   266   266   264         -2
## 6   224   233   227   219   234         15
```

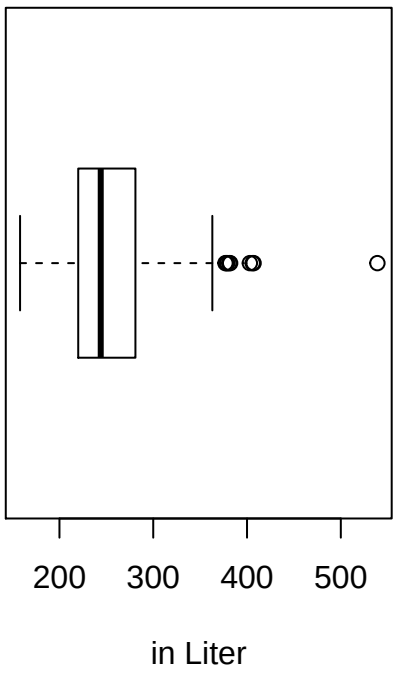
Wasserverbrauch Zürich 2018

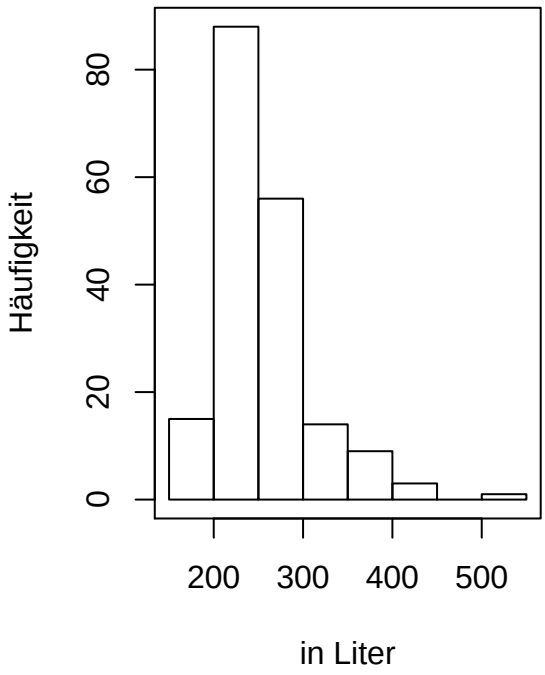
Betrachten wir zunächst die vollständigen Daten (ohne NA) für das Jahr 2018.

```
wv.cc2018 <- wv$x2018[complete.cases(wv$x2018)]
```

Die Verteilung des durchschnittlichen Jahres-pro-Kopf-Verbrauchs 2018 aus 186 Zürcher Gebieten stellt sich in Boxplot und Histogramm unimodal rechtsschief dar:

```
par(mfrow = c(1, 2))
boxplot(wv.cc2018, horizontal = TRUE, xlab = "in Liter")
hist(wv.cc2018, ylab = "Häufigkeit", xlab = "in Liter", main = "")
box()
```





Schätzung der Parameter μ und σ

Schätzwerte für μ und σ erhalten wir folgendermassen:

```
(xq <- mean(wv$x2018, na.rm = T))
```

```
## [1] 255.172
```

```
(s <- sd(wv$x2018, na.rm = T))
```

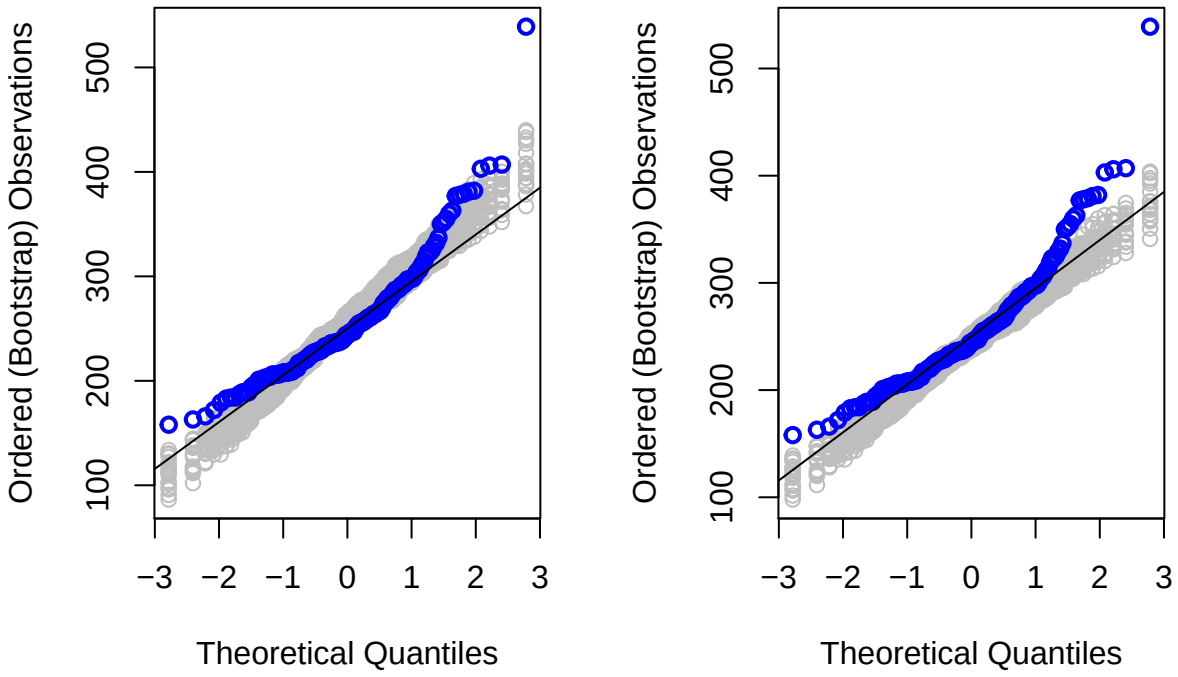
```
## [1] 53.95997
```

Überprüfung des Modells

Frage: Kann für den Wasserverbrauch 2018 eine **Standardnormalverteilung** angenommen werden?

Wir prüfen dies mit mit dem QQ-Plot (links mit, rechts ohne Aussreisser).

```
par(mfrow = c(1, 2))
source(paste0(path, "RFn-qqnormSim.R"))
qqnormSim(wv.cc2018, SEED = 123); qqline(wv.cc2018)
qqnormSim(wv.cc2018, rob = TRUE, SEED = 123); qqline(wv.cc2018)
```

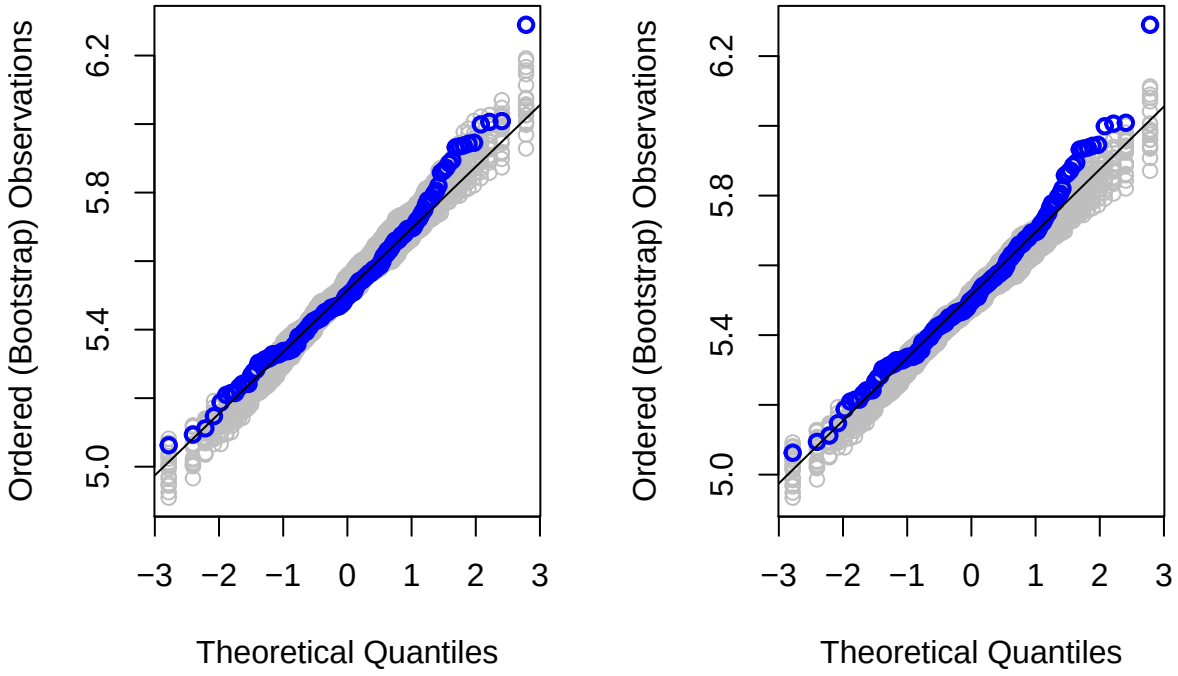


Antwort: Nein. Beide Simulationen zeigen, dass die Verteilung an beiden Enden etwas langschwänziger als die Gauss'sche Normalverteilungs-Kurve ist.

Frage: Kann für den Wasserverbrauch 2018 eine **Lognormalverteilung** angenommen werden?

Wir prüfen erneut mit dem QQ-Plot (links mit, rechts ohne Aussreisser).

```
par(mfrow = c(1, 2))
qqnormSim(log(wv.cc2018), SEED = 123); qqline(log(wv.cc2018))
qqnormSim(log(wv.cc2018), rob = TRUE, SEED = 123); qqline(log(wv.cc2018))
```



Antwort: Ja, vom Ausreisser abgesehen passen die Daten zu einem lognormal verteilten Modell.

Vertrauensintervalle

Frage: Wo lag 2018 der reale durchschnittliche Wasserverbrauch μ für die gesamte Region Zürich?

```
t.test(wv$x2018, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  wv$x2018
## t = 64.494, df = 185, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  247.3663 262.9778
## sample estimates:
## mean of x
##   255.172
```

Beruecksichtigt man **alle Daten**, kommt man bei einem Signifikanz-Niveau von 95% zu einem Mittelwert μ von 255.17 Litern pro Kopf und Jahr.

Wobei das Vertrauensintervall für μ von 247.37 bis 262.98 Litern reicht.

```
t.test(wv$x2018[wv$x2018 < 500], alternative="two.sided", conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  wv$x2018[wv$x2018 < 500]
## t = 69.173, df = 184, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  246.4036 260.8721
## sample estimates:
## mean of x
##  253.6378
```

Lässt man den **Aussreisser unberücksichtigt** (539 l für Berg a.I.), kommt man zu einem Mittelwert μ von 253.64 l und einem Vertrauensintervall zwischen 246.40 und 260.87 l.

Vorzeichentests: Binomial- und Wilcoxon-Test

Im Vergleich zu 2017 hat der Wasserverbrauch 2018 etwas abgenommen.

```
sum(wv$d2018_2017, na.rm = T)
```

```
## [1] 1244
```

Frage: Ist diese Abnahme signifikant oder rein zufaellig?

Hierzu untersuchen wir alle für 2017 und 2018 vollständigen Datensätze.

Zunächst mit dem Binomial-Test:

```
wv.cc.d2018_2017 <- wv$d2018_2017[complete.cases(wv$d2018_2017)]
binom.test(
  sum(wv.cc.d2018_2017 > 0),
  n = length(wv.cc.d2018_2017),
  p = 0.5,
  alternative = "two.sided",
```

```
  conf.level = 0.95
)
```

```
##
##  Exact binomial test
##
## data:  sum(wv.cc.d2018_2017 > 0) and length(wv.cc.d2018_2017)
## number of successes = 119, number of trials = 184, p-value = 8.381e-05
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5730232 0.7156417
## sample estimates:
## probability of success
##              0.6467391
```

Dann mit dem Wilcoxon-Test:

```
wilcox.test(wv$d2018_2017, alternative="two.sided", mu=0, conf.level=0.95)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  wv$d2018_2017
## V = 11692, p-value = 9.73e-07
## alternative hypothesis: true location is not equal to 0
```

Antwort: Da der P-Wert beider Tests kleiner dem Signifikanz-Niveau von 0.05 ist, kann die Nullhypothese (Differenz = 0) verworfen werden. Die Verbrauchsabnahme ist also nicht rein zufällig.