Course of Study Bachelor Computer Science	Exercises Statistics WS 2023/24
Sheet VII - Solutions	

Statistical Inference

1. In an urn there is an unknown number N of balls numbered from 1 to N. The number of N should be estimated. A ball from the urn is used for this purpose and his number is noted. Describe the random variable X= the number of the drawn ball.

   (a) Determine the distribution of X depending on N. Calculate the expected value and variance of X.

   (b) Show that $T(X) = 2X - 1$ is an unbiased estimator for N is.

   (c) Calculate for $N = 4$ and $N = 5$ the probability for N to be exactly estimated at T.

   (d) Calculate the variance of T.

   Answer:

   (a) uniform distribution: $P(X = k) = \frac{1}{\vartheta}$ for $k = 1, ..., \theta$, i.e. $E(X) = \frac{N+1}{2}$, $\mathrm{Var}(X) = \frac{N^2-1}{12}$

   (b) $E(T(X)) = E(2X - 1) = 2E(X) - 1 = N$

   (c) $P(T(X) = N) = P(2X-1 = N) = P(X = \frac{N+1}{2}) = \begin{cases} \frac{1}{N} & \frac{N+1}{2} \in \mathbb{N} \\ 0 & \text{else} \end{cases} \Rightarrow$
   N=4: $P(T(X) = N) = 0$ and N=5: $P(T(X) = N) = 1/5$

   (d) $\mathrm{Var}(T) = \mathrm{Var}(2X - 1) = 4\mathrm{Var}(X) = \frac{N^2-1}{3}$

2. Fish are caught from a lake, until you get n ($n \geq 3$) fishes of a certain species A. The random variable X describe the number of all caught fishes to this time. The lake contained a great number of fishes, so that it can be assumed that the ratio p of the number of fishes of the species A to the total number of all fish of the lake does not change, when some fish are caught out of the lake.

   (a) Show that $P_p(X = k) = \binom{k-1}{n-1} p^n (1-p)^{k-n}$, $k = n, n+1, \ldots$

   (b) Show that $T(X) = \dfrac{n-1}{X-1}$ is an unbiased estimator for p.

Answer:

(a) $X \in \{n, n+1, n+2, ...\}$

$$\begin{aligned} P(X=k) &= P(\{\text{n-1 species A fishes among the first k-1 caught fishes}\} \cup \\ &\quad \{\text{k th caught fish is a fish from species A}\}) \\ &= \binom{k-1}{n-1} p^{n-1}(1-p)^{k-1-n+1} \cdot p \\ &= \binom{k-1}{n-1} p^{n}(1-p)^{k-n} \end{aligned}$$

(b)

$$\begin{aligned} E(T(X)) &= E(\frac{n-1}{X-1}) \\ &= \sum_{k=n}^{\infty} \frac{n-1}{k-1} \cdot \binom{k-1}{n-1} p^{n}(1-p)^{k-n} \\ &= \sum_{k=n}^{\infty} \binom{k-2}{n-2} p^{n}(1-p)^{k-n} \\ &= p \cdot \sum_{k=n}^{\infty} \binom{k-2}{n-2} p^{n-1}(1-p)^{k-n} \\ &= p \cdot \sum_{k=n-1}^{\infty} \binom{k-1}{(n-1)-1} p^{n-1}(1-p)^{k+1-n} \\ &= p \cdot \sum_{k=n-1}^{\infty} P_p(\tilde{X}=k) = p \end{aligned}$$

with $\tilde{X}$ number of all caught fishes until n-1 fishes of a certain are get.

Maximum Likelihood Estimation

1. A ticket inspector checks for Frankfurt S-Bahn lines the tickets from the passengers. He keeps checking until he sees a passenger without valid ticket. He then collects the increased fare and starts after a break with a new check of the tickets.

   For 10 such check runs, he shall have

   42 50 40 64 30 36 68 42 46 48

until he have found a non valid ticket.

Determine a maximum likelihood estimator based on the given numbers for p share of nonvalid tickets among all checked ticktes.

Answer: $\vartheta \in (0,1) =$ ratio non valid tickets
The random variable X = "number of tickets until the first non valid ticket" is geometricaly distributed with parameter ϑ, i.e. $P(X = k) = (1-\vartheta)^{k-1}\vartheta$, $k = 1, 2, ...$
Likelihoodfunction

$$L(x_1, ..., x_n; \vartheta) = \prod_{i=1}^{n}(1-\vartheta)^{x_i-1}\vartheta = \vartheta^n(1-\vartheta)^{(\sum_{i=1}^{n} x_i)-n}$$

Easier to consider is
$f(\vartheta) = \ln L(x_1, ..., x_n; \vartheta) = n \ln \vartheta + (\sum_{i=1} n x_i - n)\ln(1-\vartheta)$
From $f'(\vartheta) = \frac{n}{\vartheta} - \frac{\sum_{i=1}^{n} x_i - n}{1-\vartheta} = 0$ we get, $\hat{\vartheta} = \frac{n}{\sum_{i=1}^{n} x_i}$. f' has a sign change from + to -. Thus there is local maximum.
Here: $\hat{\vartheta} = 0.0215$

2. A device consists of the components K_1, K_2 and K_3. The device becomes defective as soon as one or more of the components are defective. The lifetimes L_1, L_2 and L_3 (in h) of the three components are independent random variables.

   The distribution function of L_1 is $F_1(x) = \begin{cases} 1 - e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$

   The distribution functions of L_2 and L_3 are $F_2(x) = \begin{cases} 1 - e^{-\lambda \sqrt[3]{x}} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$.

   λ is an unknown parameter > 0.

   (a) Calculate the distribution function and density for the lifetime S of the device.

   (b) When measuring the lifetime of randomly from production of the devices removed resulted in following values in hours:

   82.2 94.0 122.5 95.8 106.4

   Use a maximum likelihood estimator to determine the an estimate for λ.

   Answer:

(a)

$$\begin{aligned} P(S \le s) &= 1 - P(S > s) = 1 - P(S_1 > s) \cdot P(S_2 > s) \cdot P(S_3 > s) \\ &= \begin{cases} 1 - e^{-\lambda(s+2\sqrt[3]{s})} & \text{für } s > 0 \\ 0 & \text{sonst} \end{cases} \end{aligned}$$

density function: $f(\lambda, s) = \lambda(1 + \frac{2}{3\sqrt[3]{s^2}})e^{-\lambda(s+2\sqrt[3]{s})}$

(b) Likelihoodfunktion

$$L(s_1, ..., s_5; \lambda) = \lambda^5 \prod_{i=5}^{5}(1 + \frac{2}{3\sqrt[3]{s_i^2}})e^{-\lambda(s_i+2\sqrt[3]{s_i})}$$

Taking the logarithm of the likelihood we get

$$f(\lambda) = \ln(L(s_1, ..., s_5; \lambda) = 5\ln\lambda + \sum_{i=1}^{4}\left(\ln(1 + \frac{2}{3\sqrt[3]{s_i^2}}) - \lambda(s_i + 2\sqrt[3]{s_i})\right)$$

Taking the first derivative of $f(\lambda)$ and set it zero

$$f'(\lambda) = \frac{5}{\lambda} - \sum_{i=1}^{5}(s_i + \sqrt[3]{s_i}) = 0$$

we get that we have a local maximum at

$$\hat{\lambda} = \frac{5}{\sum_{i=1}^{5}(s_i + 2\sqrt[3]{s_i})} = 0.00914$$

3. To determine the number of N of red deers living in a precinct region 7 red deer were caught and marked in a trapping action. Afterwards the animals were again released. After a certain time, another trapping action was started. Thereby 3 red deer were caught, whereby 2 already were marked. It is assumed that between is no influx or outflow of red deer in the region and that the animals were able to pass the region within a short period of time.

   (a) Determine a maximum likelihood estimator for the total number N of the red deer living in the region.

   (b) A third trapping action started, where 8 red deers were caught. 4 of them were marked. What is no the maximum likelihood estimation of N?

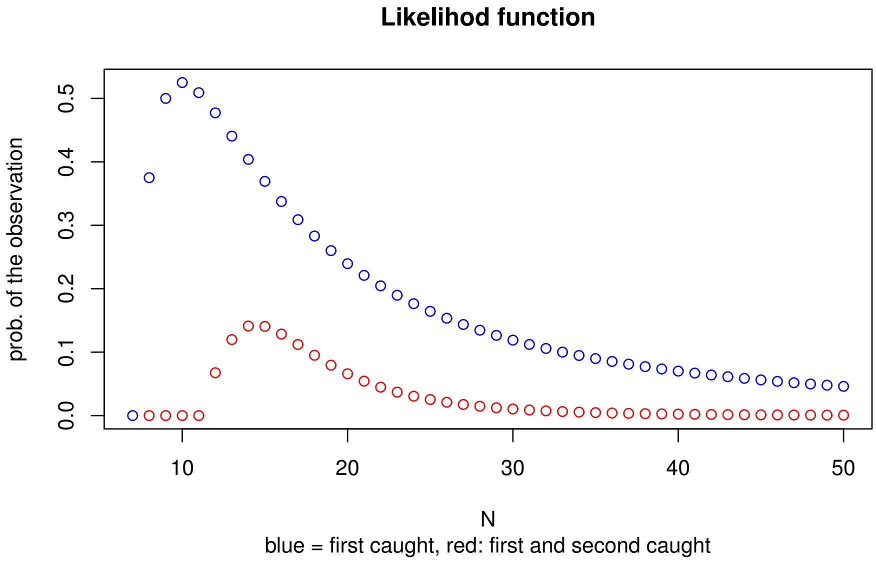
   Answer:

(a) If N denotes the unknown number of red deers and X denotes the random variables which counts the number of caught marked red deers in the second trapping action we have

$$P_N(X = 2) = \frac{\binom{7}{2}\binom{N-7}{1}}{\binom{N}{3}}$$

The Likelihoodfunktion $L(2; N)$ is nothing else then this probability.

N	p
7	0.00
8	0.38
9	0.50
10	0.52
11	0.51
12	0.48
13	0.44
14	0.40
15	0.37
16	0.34



$\Rightarrow$ maximum likelihood estimation of N is 10.

(b) Let Y denotes the number of caught marked red deers in the third trapping action

$$P_N(Y = 4) = \frac{\binom{7+1}{4}\binom{N-7-1}{4}}{\binom{N}{8}}$$

The probability of both observation is $P_N(X = 2) \cdot P_N(Y = 4)$, which is the likelihood function $L(2, 4; N)$

N	p
9	0
10	0
11	0
12	0.0675
13	0.120
14	0.141
15	0.141
16	0.128
17	0.112
18	0.0951
19	0.0795
20	0.0659

$\Rightarrow$ maximum likelihood estimation of N is 14.

```
##########################################################
# To determine the number of N of red deers living in a
# precinct region 7 red deer were caught and marked in a
# trapping action. Afterwards the animals were again
# released. After a certain time, another trapping action
# was started. Thereby 3 red deer were caught, whereby 2
# already were marked. It is assumed that between the no
# influx or outflow of red deer in the region and that the
# animals were not able to pass the region within a short
# period of time.
# Determine a maximum likelihood estimator for the total
# number N of the red deer living in the region.
#
# file: max_likelihood_deers_sol.R
##########################################################

library(tidyverse)
library(xtable)

n.marked <- 7
n.caught.1 <- 3
n.rd.1 <- 2
n.caught.2 <- 8
n.rd.2 <- 4

# create a tibble with prob. of the observation dep. on N
ml.est.N <- tibble(
  N = n.marked:50,
  est.1 = dhyper(x=n.rd.1,m=n.marked,n=N-n.marked,k=n.caught.1),
  est.2 = est.1 *
    dhyper(x=n.rd.2,m=n.marked+(n.caught.1-n.rd.1),
           n=N-n.marked-(n.caught.1-n.rd.1),k=n.caught.2)
)
head(ml.est.N,20)

# diagramm of the likelihood functions
plot(x = ml.est.N$N, y = ml.est.N$est.1, col = "blue",
     xlab = "N", ylab = "prob. of the observation",
     main = "Likelihod function",
     sub = "blue = first caught, red: first and second caught")
points(x = ml.est.N$N, y = ml.est.N$est.2, col = "red")

# find the maxima
ML.EST.1 <- ml.est.N %>%
  select(N, est.1) %>%
  filter(est.1 == max(est.1))
ML.EST.2 <- ml.est.N %>%
  select(N, est.2) %>%
  filter(est.2 == max(est.2))
ML.EST.1; ML.EST.2
```

Confidence Intervals

1. Strictly speaking, what is the correctinterpretation of a 95% confidence interval for the mean?

   ◯ If repeated samples were taken and the 95% confidence interval was computed for each sample, 95% of the intervals would contain the population mean.

   ◯ A 95% confidence interval has a 0.95 probability of containing the population mean.

   ◯ 95% of the population distribution is contained in the confidence interval.

   Answer: The first is the most accurate interpretation of a 95% confidence interval.

2. A population is known to be normally distributed with a standard deviation of 2.8.

   (a) Compute the 95% confidence interval on the mean based on the following sample of nine: 8, 9, 10, 13, 14, 16, 17, 20, 21.

   (b) Now compute the 99% confidence interval using the same data.

   Answer: Assumption: Normal distribution with known standard deviation $\sigma = 2.8$

   (a) Wanted: 95% confidence interval for μ

   Data: $$\begin{aligned} n &= 9 \\ \bar{x} &= \sum_{i=1}^{9} \frac{x_i}{9} = \frac{128}{9} = 14.22 \end{aligned}$$

   Confidence interval:

   $\bar{x} \pm \frac{\sigma}{\sqrt{n}} \cdot u_{0.975} = 14.22 \pm \frac{2.8}{\sqrt{9}} \cdot 1.96 = 14.22 \pm 1.829 \Rightarrow [12.39,\ 16.05]$

   (b) Wanted: 99% confidence interval for μ

   Data: $$\begin{aligned} n &= 9 \\ \bar{x} &= \sum_{i=1}^{9} \frac{x_i}{9} = \frac{128}{9} = 14.22 \end{aligned}$$

   Confidence interval:

   $\bar{x} \pm \frac{\sigma}{\sqrt{n}} \cdot u_{0.995} = 14.22 \pm \frac{2.8}{\sqrt{9}} \cdot 2.5758 = 14.22 \pm 7.2122/\sqrt{9} \Rightarrow [11.82,\ 16.62]$

```
####################################################
# A population is known to be normally distributed
# with a standard deviation of 2.8.
#
# file: infstat_conf_interval_normal_mean.R
####################################################

# a) Compute the 95% confidence interval on the mean
```

```
sample <- c(8, 9, 10, 13, 14, 16, 17, 20, 21)
alpha <- 0.05
m <- mean(sample)
m
s <- 2.8
q_a <- qnorm(1-alpha/2,0,1)
q_a
u <- m-q_a*s/sqrt(length(sample))
o <- m+q_a*s/sqrt(length(sample))
u;o

# b) Now compute the 99% confidence interval using the same data.
alpha <- 0.01
q_a <- qnorm(1-alpha/2,0,1)
q_a
u <- m-q_a*s/sqrt(length(sample))
o <- m+q_a*s/sqrt(length(sample))
u;o

# Solution applying z.test() from the TeachingDemos package
library(TeachingDemos)
z.test(x= sample, sd = 2.8, alternative = "two.sided", conf.level = 0.95)$conf.int
# a)
z.test(x = sample, sd = 2.8, alternative = "two.sided", conf.level = 0.99)$conf.int # b)
```

3. You take a sample of 22 from a population of test scores, and the mean of your sample is 60.

   (a) You know the standard deviation of the population is 10. What is the 99% confidence interval on the population mean?

   (b) Now assume that you do not know the population standard deviation, but the standard deviation in your sample is 10. What is the 99% confidence interval on the mean now?

   Hint: Assume that the test scores follow a normal distribution.

   Answer: Assumption: Normal distribution, Data: $\begin{array}{rcl} n &=& 22 \\ \bar{x} &=& 60 \end{array}$

   (a) Wanted: 99% confidence interval for μ
   Assumption: Known standard deviation $\sigma = 10$
   Confidence interval:
   $\bar{x} \pm \frac{\sigma}{\sqrt{n}} \cdot u_{0.995} = 60 \pm \frac{10}{\sqrt{22}} \cdot 2.5758 = 60 \pm 5.492 \Rightarrow [54.508,\ 65.492]$

   (b) Wanted: 99% confidence interval for μ
   Assumption: Unknown standard deviation, but already estimated $s = 10$ (i.e. t_{n-1}−distribution is used)
   Confidence interval:
   $\bar{x} \pm \frac{\sigma}{\sqrt{n}} \cdot t_{21,0.995} = 60 \pm \frac{10}{\sqrt{22}} \cdot 2.8314 = 60 \pm 6.036 \Rightarrow [53.963,\ 66.036]$

```
#####################################################
# You take a sample of 22 from a population of test
# scores, and the mean of your sample is 60.
#
# file: infstat_conf_intervall_normal_mean_sd_unknown.R
######################################################
n <- 22
m <- 60
```

```
# a) You know the standard deviation of the population is 10. What
#    is the 99\% confidence interval on the population mean.
alpha <- 0.01
s <- 10
q_a <- qnorm(1-alpha/2,0,1)
q_a
u <- m-q_a*s/sqrt(n)
o <- m+q_a*s/sqrt(n)
u;o


# Solution applying z.test() from the TeachingDemos package
library(TeachingDemos)
z.test(x = m, sd = 10, alternative = "two.sided", n = 22, conf.level = 0.99)$conf.int

# b) Now assume that you do not know the population standard
#    deviation, but the standard deviation in your sample is 10. What
#    is the 99\% confidence interval on the mean now?
s_sample <- 10
t_a <- qt(1-alpha/2,n-1)
t_a
u <- m-t_a*s/sqrt(n)
o <- m+t_a*s/sqrt(n)
u;o
```

4. Calculate for the below given sample from a normally distributed population the 95% confidence intervals

   (a) for the mean, if the standard deviation is 2
   (b) for the mean, if the standard deviation is unknown
   (c) for the variance, if the mean is 250
   (d) for the variance, if the mean is unknown

   x_i : 247.4, 249.0, 248.5, 247.5, 250.6, 252.2, 253.4, 248.3, 251.4, 246.9, 249.8, 250.6, 252.7, 250.6, 250.6, 252.5, 249.4, 250.6, 247.0, 249.4

   Answer: sample size n=20, $\bar{x} = 249.92$, $s = 1.9479$, $\alpha = 0.05$

   (a) $\left[\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = [249.04, 250.80]$

   (b) $\left[\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right] = [229.01, 250.83$

   (c) $\left[\frac{Q_n}{\chi^2_{n,1-\alpha/2}}, \frac{Q_n}{\chi^2_{n,\alpha/2}}\right] = [2.11, 7.53]$ with $Q_n = \sum_{i=1}^n (x_i - \mu)^2 = 72.22$

   (d) $\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right] = [2.19, 8.09]$

```
#######################################################
# Calculate for the given sample from normally
# distributed population the 95% confidence intervals
# a) for the mean, if the standard deviation is 2
# b) for the mean, if the standard deviation is unknown
# a) for the variance, if the mean is 250
# a) for the variance, if the mean is unknown
#
# file: infstat_conf_intervall_normal_mu_sigma.R
```

```
##########################################################
# create sample values
# s.values <- round(rnorm(n=20, mean = 251, sd = 2),1)
s.values <- c(247.4,249.0,248.5,247.5,250.6,252.2,253.4,248.3,251.4,246.9,
              249.8,250.6,252.7,250.6,250.6,252.5,249.4,250.6,247.0,249.4)
# characteristics of the sample
n <- length(s.values)
xbar <- mean(s.values)
s <- sd(s.values)
# level 1-alpha
alpha <- 0.05

# confidence intervalls for mu
# a) assumption: sigma = 2
sigma <- 2
l.a <- xbar - qnorm(1-alpha/2)*sigma/sqrt(n)
u.a <- xbar + qnorm(1-alpha/2)*sigma/sqrt(n)
l.a; u.a
# b) assumption: sigma = unknown
l.b <- xbar - qt(1-alpha/2, df = n-1)*s/sqrt(n)
u.b <- xbar + qt(1-alpha/2, df = n-1)*s/sqrt(n)
l.b; u.b

# confidence intervalls for sigma^2
# c) assumption: mu = 250
mu <- 250
Qn <- sum((s.values - mu)^2)
l.c <- Qn/qchisq(1-alpha/2, df = n)
u.c <- Qn/qchisq(alpha/2, df = n)
l.c; u.c
# d) assumption: mu unknown
l.d <- (n-1)*s^2/qchisq(1-alpha/2, df = n-1)
u.d <- (n-1)*s^2/qchisq(alpha/2, df = n-1)
l.d; u.d

# solutions applying z.test(), sigma.test() from TeachingDemos and t.test()
library(TeachingDemos)
z.test(x = s.values, sd = 2, alternative = "two.sided", conf.level = 0.95)$conf.int # a)
t.test(x = s.values, alternative = "two.sided", conf.level = 0.95)$conf.int
# b)
sigma.test(x = s.values, alternative = "two.sided", conf.level = 0.95)
# d)
```

5. At a telemarketing firm, the length of a telephone solicitation (in seconds) is a normally distributed random variable with mean μ and standard deviation σ, both unknown. A sample of 51 calls has mean length 300 and standard deviation 60.

   (a) Construct the 95% confidence upper bound for μ.

   (b) Construct the 95% confidence lower bound for σ.

   Answer: Sample size $n = 51$ and sample mean $\bar{x} = 300$ and sample standard deviation $s = 60$

   (a) Wanted: Confidence interval for μ at level $1 - \alpha = 95\%$
   In general we have the two-sided confidence interval.
   $\left[\bar{x} - t_{n-1,\,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1,\,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]$
   A one-sided confidence interval (upper boundary):
   $\left(-\infty,\ \bar{x} + t_{n-1,\,1-\alpha} \cdot \frac{s}{\sqrt{n}}\right] = \left[-\infty,\ 300 + 1.6759 \cdot \frac{60}{\sqrt{51}}\right] = (-\infty, 314,23]$
   with $t_{50,0.95} = 1.6759$

(b) Wanted: Confidence interval for σ at level $1-\alpha = 95\%$

In general we have the two sided confidence interval for σ^2: $\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right]$

A one-sided confidence interval (lower boundary) for σ:
$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}}}, \infty\right) = [51,57\,; \infty)$ with $\chi^2_{n-1,1-\alpha} = \chi^2_{50,0.95} = 67.505$

```
#####################################################
# At a telemarketing firm, the length of a telephone
# solicitation (in seconds) is a normally distributed
# random variable with mean mu and standard deviation
# sigma, both unknown. A sample of 50 calls has mean
# length 300 and standard deviation 60.
#
# file: infstat_conf_interval_telefirm.R
######################################################
n <- 51; m <- 300; s_sample <- 60; alpha <- 0.05

#  a) Construct the 95% confidence upper bound for mu.
t_a <- qt(1-alpha,n-1)
t_a
o <- m+t_a*s_sample/sqrt(n)
o

#  b) Construct the 95% confidence lower bound for sigma.
chi <- qchisq(1-alpha,n-1)
chi
u <- (n-1)*s_sample^2/chi
sqrt(u)
```

6. At a certain farm the weight of a peach (in ounces) at harvest time is a normally distributed random variable with standard deviation 0.5. How many peaches must be sampled to estimate the mean weight with a margin of error ± 0.2 and with 95% confidence.

Answer: Standard deviation $\sigma = 0.5$ known.
Confidence interval: $\bar{x} \pm \frac{\sigma}{\sqrt{n}} \cdot u_{0.975}$

Wanted: n with $\frac{\sigma}{\sqrt{n}} \cdot u_{0.975} = 0.2$ i.e. $\frac{0.5}{\sqrt{n}} \cdot 1.96 = 0.2$ i.e. $\sqrt{n} = \frac{0.5\cdot 1.96}{0.2}$ i.e. $n = \left(\frac{0.5\cdot 1.96}{0.2}\right)^2 = 4.9^2 = 24.01 \Rightarrow n = 25$ since we must round upwards.

```
#########################################################
# At a certain farm the weight of a peach (in ounces)
# at harvest time is a normally distributed random
# variable with standard deviation 0.5. How many peaches
# must be sampled to estimate the mean weight with a
# margin of error pm 0.2 and with 95% confidence.
#
# file: infstat_conf_interval_peach.R
#########################################################

alpha <- 0.05; s <- 0.5; margin <- 0.2
q_a <- qnorm(1-alpha/2,0,1); q_a
# margin >= q*s/n^0.5
n <- ceiling((q_a*s/margin)^2)
n
```

7. You read about a survey in a newspaper and find that 70% of the 250 people sampled prefer candidate A.

(a) Compute the 95% confidence interval.

(b) You are surprised by this survey because you thought that more like 50% of the population preferred this candidate. Based on this sample, is 50% a possible population proportion?

Answer: Data: $\begin{array}{rcl} n & = & 250 \\ \hat{x} & = & 0.70 \end{array}$

(a) Wanted: Confidence interval for p at level $1-\alpha = 0.95$
$\hat{p} \pm u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
We have: $0.70 \pm 1.96 \cdot \sqrt{\frac{0.70(1-0.70)}{250}} = 0.70 \pm 0.057$
i.e. $[0.6432, \quad 0.7568]$

(b) Possible, but with a very low probability since 50% is not in the confidence interval.

```
###########################################################
# You read about a survey in a newspaper and find
# that 70% of the 250 people sampled prefer Candidate A.
# a) Compute the 95% confidence interval.
# b) You are surprised by this survey because you thought
#    that more like 50% of the population preferred this
#    candidate. Based on this sample, is 50% a possible
#    population proportion?
#
# file: infstat_conf_interval_prop_survey.R
###########################################################

n <- 250; p <- 0.7; alpha <- 0.05

# normal approximation
l.appr <- p - qnorm(1-alpha/2)*sqrt(p*(1-p)/n)
u.appr <- p + qnorm(1-alpha/2)*sqrt(p*(1-p)/n)
l.appr; u.appr
# Rule of thumb: n*p and n*(1-p) should be > 10
n*p; n*(1-p)

# exact
xp <- seq(0,1,length=1+10^4)
l.ex <- xp[min(which(qbinom(1-alpha/2,n,xp) == p*n))]
u.ex <- xp[max(which(qbinom(alpha/2,n,xp) == p*n))]
l.ex; u.ex

# exact confidence interval with R-function
binom.test(x=0.7*250,n=250,conf.level=1-alpha)$conf.int

# If p = 0.5, the probability to observe 0.7*250 or more voters
# in a sample of size 250 is
1-pbinom(n*p-1, size = n, prob = 0.5) # approx 0
```

8. A researcher was interested in knowing how many people in the city supported a new tax. He sampled 100 people from the city and found that 40% of these people supported the tax. What is the upper limit of the 95% (one-side) confidence interval on the population proportion?

Answer: Survey with $n = 100$ and 40% approve the taxes
Wanted: Upper-boundary confidence interval for a proportion $p =$ at level $1-\alpha = 0.95$: $\hat{p} + u_{1-\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
We have $n = 100$, $\hat{p} = 0.4$ and $1-\alpha = 0.95 \Rightarrow u_{1-\alpha} = 1.645$, i.e. we become $0.4 + 1.645 \cdot \sqrt{\frac{0.4 \cdot 0.6}{100}} = 0.48$

```
##################################################
# A researcher was interested in knowing how many
# people in the city supported a new tax. She sampled
# 100 people from the city and found that 40% of
# these people supported the tax. What is the upper
# limit of the 95% (one-side) confidence interval
# on the population proportion?
#
# file: infstat_conf_intervall_prop_one_sided.R
##################################################

n <- 100; p <- 0.4; alpha <- 0.05

# normal approximation
u.appr <- p + qnorm(1-alpha)*sqrt(p*(1-p)/n)
u.appr
# Rule of thumb: n*p and n*(1-p) should be greater than 10
n*p; n*(1-p)

# exact
xp <- seq(0,1,length=1+10^4)
u.ex <- xp[max(which(qbinom(alpha,n,xp) == p*n))]
u.ex

# exact confidence interval with R-function
binom.test(x=40, n=100, alternative = "less",
           conf.level=1-alpha)$conf.int
```

9. An advertising agency wants to construct a 99% confidence lower bound for the proportion of dentists who recommend a certain brand of toothpaste. The margin of error is to be 0.02. How large should the sample be?

   Answer: The lower boundary at level $1-\alpha = 0.99$ for the proportion p is denoted z. Thus, we have
   $z = \hat{p} - u_{1-\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ with $u_{1-\alpha} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.02$
   $\alpha = 0.01 \Rightarrow u_{0.99} = 2.326$
   n is unknown, thus $2.326 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0.02$ i.e.
   $2.326^2 \cdot \hat{p}(1-\hat{p}) \leq 0.02^2 \cdot n$ i.e. $n \geq \frac{2.326^2}{0.02^2}\hat{p}(1-\hat{p})$
   For which $\hat{p}$ has the function $y = \hat{p}(1-\hat{p})$ a maximum? We take the derivative: $y = \hat{p} - \hat{p}^2 \Rightarrow y' = 1 - 2\hat{p}$ and then $y' = 1 - 2\hat{p} = 0 \Rightarrow \hat{p} = \frac{1}{2}$.
   Thus, we have $y = \hat{p}(1-\hat{p}) \leq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

   This gives $n \geq \frac{2.326^2}{0.02^2} \cdot \frac{1}{4} = 3381$
   If we suppose that $p \leq 0.25$, i.e. $n \geq \frac{2.326^2}{0.02^2} \cdot \frac{1}{4} \cdot \frac{3}{4} = 2537$

```
##################################################
# An advertising agency wants to construct a 99%
# confidence lower bound for the proportion of
# dentists who recommend a certain brand of toothpaste.
# The margin of error is to be 0.02. How large should
# the sample be?
#
# file: infstat_conf_interval_prop_sample_size.R
##################################################

alpha <- 0.01; margin <- 0.02
c <- qnorm(1-alpha,0,1)
f <- seq(0,1,length=101)
n <- max(ceiling(c^2 * f*(1-f)/(margin^2)))
n

# If f <= 0.2, we get
f <- seq(0,0.2,length=21)
```

```
n <- max(ceiling(c^2 * f*(1-f)/(margin^2)))
n
```

10. The interval [45.6, 47.8] is a symmetric 99% confidence interval for the unknown parameter μ based on a sample $x_1, \ldots, x_{10}$ from a normal distribution $N(\mu, \sigma^2)$ with unknown σ. Calculate the sample mean $\bar{x}$ and the sample standard deviation s.

    Answer: Mean: $\bar{x} = \frac{45.6+47.8}{2} = 46.7$ and using the lower limit 45.6 we get $45.6 = \bar{x} - t_{9,\,0.995} \cdot \frac{s}{\sqrt{n}}$ i.e. $s = \frac{\bar{x}-45.6}{t_{9,\,0.995}} \cdot \sqrt{n} = \frac{46.7-45.6}{3.25} \cdot \sqrt{10} = 1.07$

11. The waiting time at the pay desk of a certain supermarket is normally distributed with mean waiting time μ and known standard deviation $\sigma = 1,8$ minutes. A confidence interval for the mean waiting time (in minutes) for this supermarket is [5.12; 8.32]. If the sample size is $n = 10$, what is then the confidence level?
    Answer: The length of the interval is $8.32 - 5.12$ and
    $8.32 - 5.12 = 2 \cdot u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 2 \cdot u_{1-\frac{\alpha}{2}} \cdot \frac{1.8}{\sqrt{10}}$ i.e. $u_{1-\frac{\alpha}{2}} = 2.81$ and the normal distribution table gives $1 - \frac{\alpha}{2} = 0.9975$ i.e. $\alpha \approx 0.005$. So the confidence level is $1 - \alpha = 99.5\%$.

12. **R programming task:** Consider an urn with M white balls and N-M black. n balls are drawn without replacement and X denotes the number of white balls in the sample. N=500 and n=50 are known but M the number of white balls is unknown. Construct an two sided $1-\alpha = 0.95$ confidence intervall for M based on the H(N,M,n)-distribution of X. Compare it with a binomial and a normal approximation.

```
##########################################################
# Consider an urn with M white balls and N-M black. n
# balls are drawn without replacement and X denotes the
# number of white balls in the sample. N=500 and n=50
# are known but M the number of white balls is unknown.
# Construct an two sided 1-alpha=0.95 confidence intervall
# for M based on the H(N,M,n)-distribution of X. Compare
# it with a binomial and a normal approximation.
#
# file: infstat_conf_interval_hypergeo_M.R
##########################################################

library(tidyverse)

# urn modell: N total number of balls, M = number of white
# balls, n = number of drawn balls
# X = number of white balls ~ H(N,M,n)
N <- 500
n <- 50
alpha <- 0.05

# symmetric intervals [lb,ub] for X with probability 1-alpha
# for different values of M
sy.intervals <- tibble(
  M = 0:N,
  # quantils of H(N,M,n)
  lb = qhyper(alpha/2,M,N-M,n),
  ub = qhyper(1-alpha/2,M,N-M,n)
)
# plot of the the intervals
plot(x=sy.intervals$M, y=sy.intervals$lb, col="blue",
```

```
      type = "p",
      xlab = "M", ylab = "lower and upper bounds",
      main = "symmetric 95% intervals for X")
points(x=sy.intervals$M, y=sy.intervals$ub, col="red")

# Mention the lb- and ub-functions are not strictly monotonously
# increasing: use for given value of X the min of the
# corresponding ub values and the max of the corresponding lb
# values of M as an inverse of the two function. These values
# are the bounds of the confidence intervals.
ex.conf.intervall <- function(x) {
  return(c(
    sy.intervals %>%
      filter(ub == x) %>%
      mutate(l = min(M)) %>%
      select(l) %>%
      unique() %>%
      as.numeric(),
    sy.intervals %>%
      filter(lb == x) %>%
      mutate(u = max(M)) %>%
      select(u) %>%
      unique() %>%
      as.numeric()
  ))
}

# The binom.test(x,n) function returns in the variable
# conf.int the confidence interval for p=M/N if they are X
# white balls in a sample of n balls drawn from the urn
# with replacement
binom.appr.conf.intervall <- function(x) {
  return(
    c(
      binom.test(x, n, conf.level = 1-alpha)$conf.int[1]*N,
      binom.test(x, n, conf.level = 1-alpha)$conf.int[2]*N
    )
  )
}

# normal approximation of the confidence interval for an
# unknown proportion if x white balls are in a sample of
# n balls drwan with replacement
normal.appr.conf.intervall <- function(x) {
  return(
    c(
      N*(x/n -qnorm(1-alpha/2)*sqrt(x*(1-x/n)/n^2)),
      N*(x/n +qnorm(1-alpha/2)*sqrt(x*(1-x/n)/n^2))
    )
  )
}

# tibble of the bounds of the confidence intervalls for M
# for all possibloe values of X
tab <- tibble(
  X = 0:n) %>%
  group_by(X) %>%
  mutate(ex.lb=ex.conf.intervall(X)[1],
         ex.ub=ex.conf.intervall(X)[2],
         binom.lb=binom.appr.conf.intervall(X)[1],
         binom.ub=binom.appr.conf.intervall(X)[2],
         norm.lb=normal.appr.conf.intervall(X)[1],
         norm.ub=normal.appr.conf.intervall(X)[2]
         )

# plot of all bounds
plot(x=tab$X, y=tab$ex.lb, col="red",
      xlab = "x", ylab = "M",
      main = "95% confidence intervall for M in H(N=500,M,n=50)",
      sub = "red = exact, blue = binomial approx, black = normal approx.")
points(x=tab$X, y=tab$ex.ub, col="red")
points(x=tab$X, y=tab$binom.lb, col="blue")
points(x=tab$X, y=tab$binom.ub, col="blue")
points(x=tab$X, y=tab$norm.lb, col="black")
points(x=tab$X, y=tab$norm.ub, col="black")
```