Principles of Statistical Data Analysis: Assignment 1

Ayala Denul, Nathan Laroy, Ole Schacht

2023-10-26

1. **Explore the data, both with numerical summary statistics and with graph(s). Include an evaluation of the shape of the distribution of the maximum heart rates.Can you find a transformation to make distribution look similar to a normal distribution?**

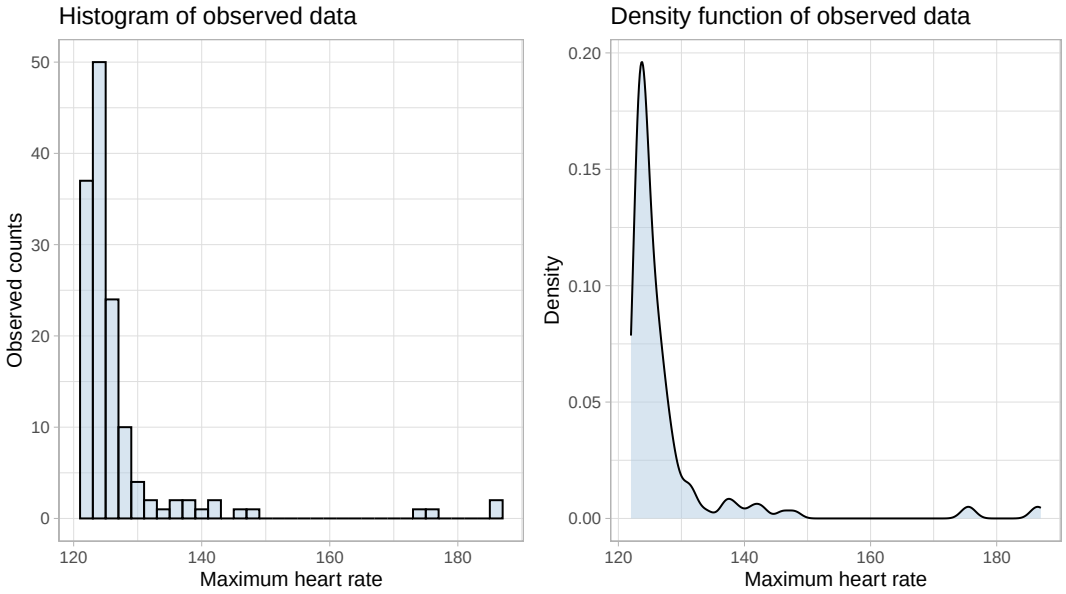
```
# load relevant packages
library(MASS) # compute lambda for boxcox transformation
library(ggplot2) # produce elegant data visualizations
library(ggpubr) # arrange multiple ggplot graphs
library(showtext) # use default LaTeX font in graphs

# je moet een package downloaden om latex lettertypes te gebruiken in je plots
# Dit zal ik op het einde dan wel terug inschakelen, anders kunnen jullie thuis
# de plots nu mss niet runnen omdat de latex font niet op jullie pc staat.
# die regeltjes code staan nu annotated om errors te vermijden.
# We moeten normaal geen rmarkdown indienen dus het installeren van packages is
# voor de beoordelaar van deze homework dus niet aan de orde en geen probleem
# (m.a.w. we kunnen de fonts dus gerust gebruiken als het indien format enkel pdf is)
```

As a first step, we load in the data vector called `hr.RData`, which contains the maximal heart rate of a particular athlete during 142 sessions. Reading in the data can be done using the `load()` command.

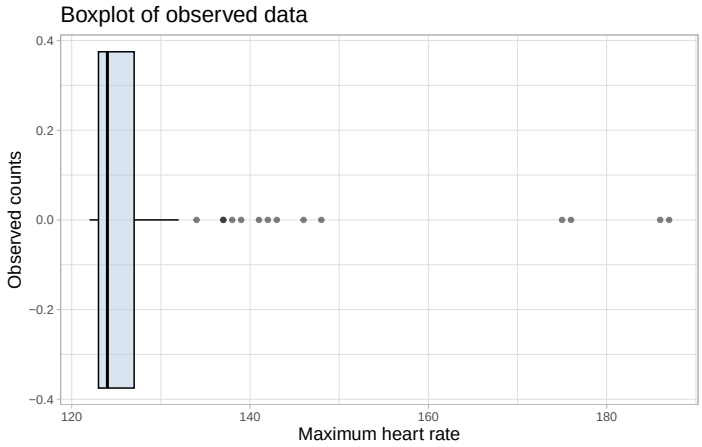
```
load("hr.RData")
```

We start by visually inspecting the data. It is immediately clear that the maximal heart rate across sessions is not normally distributed. Below are two visualizations. The left-hand figure shows a histogram of the observed heart rate data. It is clear to see that the observed distribution is heavily right-skewed; most of the data is located in the interval $[120, 140]$, with just a few observations located in the far-right tail. The right-hand graph shows a kernel density estimate of the same data, which is a smoothed version of the histogram on the left. Here, too, it is clear that most of the data is located in the interval $[120, 140]$. In fact, as was given in the assignment, maxima are often well approximated by a Fréchet distribution, whose density function is presented below:

$$\frac{\zeta}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{-1-\zeta}\exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{-\zeta}\right]$$



A third alternative visualization is presented below in the format of a boxplot. We observe here again that the maximal heart rates are far removed to the right from the bulk of the data. Because of this heavy skew, interpreting a location summary statistic that is sensitive to outliers is not advised. Indeed, a naive estimation of the mean statistic gives us the following value: `mean(hr)` = 127.652. A more robust measure may be, e.g., the median, because this value is less sensitive to outliers. The corresponding value of the median is: 124 in the current data set.




```
mean(hr)
```

```
## [1] 127.6525
```

```r
median(hr)
```

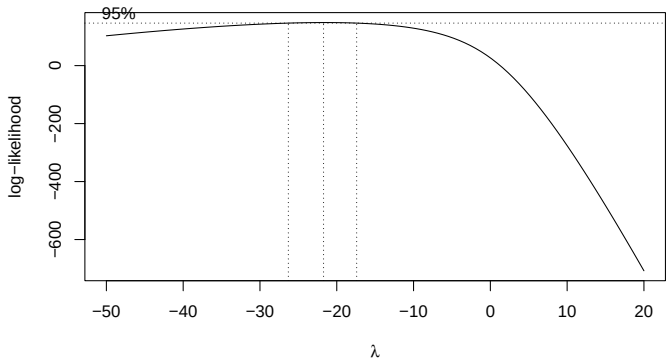
```
## [1] 124
```

transformation

In particular contexts, one often aims to transform the data in such a way so as to approximate a normal distribution. Multiple options to this end are at our disposal. A naive option might be to use a simple log-transformation. Log-transforming data makes has the ability to make it more normally distributed, because taking the logarithm of a set of right-skewed values makes the most extreme observations less extreme.

However, one may easily observe that the resulting log-transformed distribution is still heavily skewed. We therefore try a different strategy, namely, a Box-Cox power transformation, which consists of transforming the data with a given exponential value λ. In order to find this optimal solution for λ we rely on a numerical optimization algorithm from the `MASS` package in `R`. The function `boxcox()` takes as input arguments the formula to be maximized (which can be specified in generic form as `lm(hr ~ 1)`), and a sequence of possible `lambda` values for which the log-likelihood profiles will be calculated. Because we are dealing with heavily right-skewed data, we expect to find a negative value for λ. We therefore inspect the log-likelihood profile function over the interval $[-50, 20]$ in steps of 1. The graph below shows the log-likelihood function for the specified interval of `lambda` values.

```r
boxcox1 <- boxcox(lm(hr ~ 1),
       lambda = seq(-50, 20, 1), plotit = TRUE, eps = 1/50,
       xlab = expression(lambda), ylab = "log-likelihood")
```

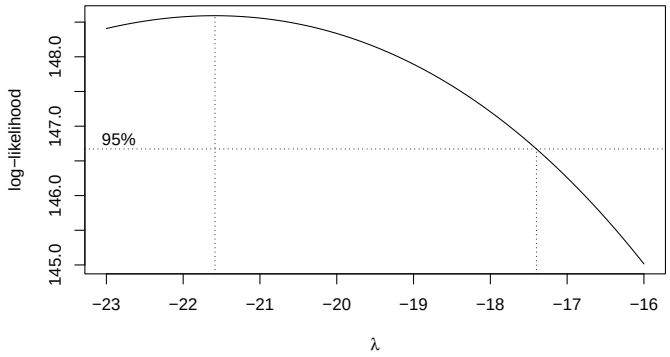



```r
lambda <- boxcox1$x[which.max(boxcox1$y)]
lambda
```

```
## [1] -21.71717
```

We observe that an optimum is found at the value of $\lambda = -21.71717$. The left (right) vertical dotted line corresponds to the lower (upper) limit of the 95% confidence interval for λ. The middle vertical dotted line corresponds with the maximum value for λ. Finally, the horizontal dotted line corresponds to the value of the log-likelihood of this optimum.

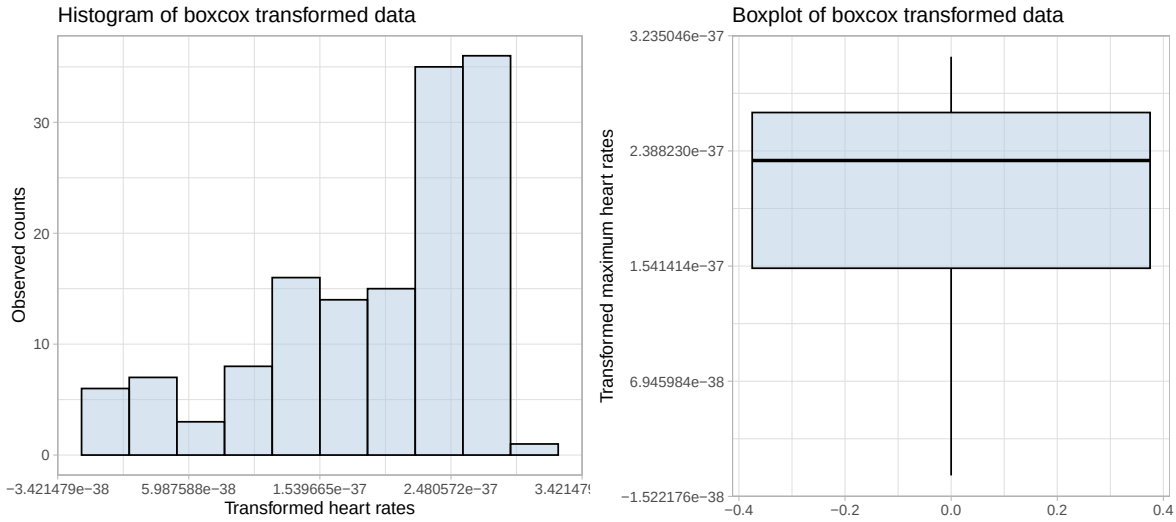
However, it is often not desirable to take the optimum with respect to λ, but rather to take the value of the sequence in λ (in steps of 0.5) that lies closest to -1, while still lying within the 95% confidence interval. For the current data set, this value corresponds to the -17.5. The graph below shows a closer inspection of this log-likelihood function, which is specified between the interval $[-23, 16]$. The desired λ values equals -17.5 (the value within the CI that is closest to 1 in steps of 0.5).



Now that we have obtained the desired value for λ, we may transform the data. This can be achieved using one the following transformations, depending on the value for λ: Y^{λ} if $\lambda \neq 0$, or $\log(\lambda)$ if $\lambda = 0$. Because we obtain $\lambda = -17.5$ we apply the former conversion formula to the data. In R this can be done as follows:

```
hr_transformed <- hr^(-17.5)
```

After transforming the data, we observe that the transformation approximates more a normal distribution, although a visual inspection brings more nuance to this claim. We now in fact observe a moderately left-skewed distribution. It is important to note that performing statistical analyses on this transformed data is not straightforward, because the interpretation of estimated model parameters now changes considerably. In the remainder of this report we use the non-transformed data.



2. **Construct the likelihood and the log-likelihood functions. You may assume that all observations are independently distributed.**

To construct the **likelihood function** we assume that the data originates from a Fréchet probability distribution. We proceed as follows: recall from question 1 that the probability density function is given by

$$f(x) = \frac{\zeta}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{-1-\zeta} \exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{-\zeta}\right]$$

and that we are allowed to rely on mathematical simplifications because of the i.i.d. assumption (the observations are mutually independent, but identically distributed). Thus, using the multiplication rule we find:

$$L(\mu,\sigma,\zeta) = \prod_{i=1}^{n} \frac{\zeta}{\sigma}\left(\frac{x_i-\mu}{\sigma}\right)^{-1-\zeta} \exp\left[-\left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta}\right]$$

Because using the **log-likelihood** function results in less tedious calculations for deriving the MLE for a given unknown parameter, we construct this very function in the following section. The log-likelihood function is generically defined as the logarithm of the likelihood function. In the current case, we define:

$$l(\mu,\sigma,\zeta) \equiv \log\left[L(\mu,\sigma,\zeta)\right]$$

The log-likelihood function can now be calculated as follows:

$$
\begin{aligned}
l(\mu,\sigma,\zeta) &= \sum_{i=1}^{n} \log\left[\frac{\zeta}{\sigma}\left(\frac{x_i-\mu}{\sigma}\right)^{-1-\zeta}\right] + \log\left[\exp\left(-\left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta}\right)\right] \\
&= \sum_{i=1}^{n} \log\left(\frac{\zeta}{\sigma}\right) + \log\left(\left[\frac{x_i-\mu}{\sigma}\right]^{-1-\zeta}\right) - \left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta} \\
&= \sum_{i=1}^{n} \log(\zeta) - \log(\sigma) + (-1-\zeta)\log\left(\frac{x_i-\mu}{\sigma}\right) - \left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta} \\
&= \sum_{i=1}^{n} \log(\zeta) - \log(\sigma) + (-1-\zeta)\left[\log(x_i-\mu) - \log(\sigma)\right] - \left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta} \\
&= \sum_{i=1}^{n} \log(\zeta) - \log(\sigma) + (-1-\zeta)\log(x_i-\mu) - (-1-\zeta)\log(\sigma) - \left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta} \\
&= \sum_{i=1}^{n} \log(\zeta) - \log(\sigma) + (-1-\zeta)\log(x_i-\mu) + \log(\sigma) + \zeta\log(\sigma) - \left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta} \\
&= \sum_{i=1}^{n} \log(\zeta) + \zeta\log(\sigma) + (-1-\zeta)\log(x_i-\mu) - \left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta} \\
&= n\log(\zeta) + n\zeta\log(\sigma) + \sum_{i=1}^{n}(-1-\zeta)\log(x_i-\mu) - \left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta} \\
&= n\log(\zeta) + n\zeta\log(\sigma) - \sum_{n=1}^{n}(1+\zeta)\log(x_i-\mu) + \left(\frac{x_i-\mu}{\sigma}\right)^{-\zeta}
\end{aligned}
$$

The following mathematical rules were applied per step:

$1 > 4)$ Properties of logarithms (log of a division, log of an exponentiated term, log of ln()).

5) Distribute $(-1-\zeta)$ over $[\log(x_i-\mu) - \log(\sigma)]$.

6) Distribute $\log(\sigma)$ over $(-1-\zeta)$, and resolve with preceding $-$ sign.

7) Cancel $-\log(\sigma) + \log(\sigma)$ and reorder terms (in)dependent of x_i within summation for visual ease.

8) Extract x_i-independent terms from summation (multiplying by n).

9) Put minus before summation sign, so as to remove all minus signs from within summation (easier for later partial derivation).

---

3. **Make a plot of the log-likelihood function as a function of the scale parameter. You may assume that the other two parameters are known:** $\mu = 122$ and $\zeta = 1.2$.

```
hr[34] <- 122 # dit is de waarde die voor problemen zorgt omdat dit gelijk is aan mu
# dus log(122-122) = log(0) = -inf

loglik <- function(sigma, x) {
```

```
  zeta <- 1.2
  mu <- 122
  n <- length(hr)
  n*log(zeta) + n*zeta*log(sigma) - sum( (1+zeta)*log(x-mu) + ((x - mu)/sigma)^(-zeta))
}

sigma <- seq(1, 14000, by = 10)
loglik_values <- unlist(lapply(sigma, loglik, x = hr))
head(loglik_values) # produceert enkel NaN

data <- data.frame(sigma = sigma, loglik = loglik_values)
ggplot(data, aes(x = sigma, y = loglik)) +
  geom_line(color = "black") +
  labs(x = expression(sigma), y = expression(L(sigma), col = "black")) +
  ggtitle("1 (b)") + theme_light()
# normaal kan dit mooi geplot worden maar nu zitten we met errors
```

4. **Derive an explicit formula for the maximum likelihood estimate of the scale parameter (assuming that the other parameters are known).**

To arrive at an analytic expression for the MLE of the scale parameter σ, we proceed as follows, starting from the log-likelihood function derived earlier:

$$l(\mu, \sigma, \zeta) = n\log(\zeta) + n\zeta\log(\sigma) - \sum_{n=1}^{n}(1+\zeta)\log(x_i - \mu) + \left(\frac{x_i - \mu}{\sigma}\right)^{-\zeta},$$

we calculate the partial derivative of $l(\mu, \sigma, \zeta)$ with respect to σ.

$$\begin{aligned}
\frac{\partial l}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left(n\log(\zeta) + n\zeta\log(\sigma)\right) - \frac{\partial}{\partial \sigma}\sum_{i=1}^{n}(1+\zeta)\log(x_i - \mu) + \left(\frac{x_i - \mu}{\sigma}\right)^{-\zeta} = 0 \\
&= \frac{n\zeta}{\sigma} - \sum_{i=1}^{n}\frac{\partial}{\partial \sigma}\left(\frac{x_i - \mu}{\sigma}\right)^{-\zeta} = 0 \\
&= \frac{n\zeta}{\sigma} - \sum_{i=1}^{n}(-\zeta)\left(\frac{x_i - \mu}{\sigma}\right)^{-\zeta-1}\left(-\frac{x_i - \mu}{\sigma^2}\right) = 0 \\
&= \frac{n\zeta}{\sigma} - \sum_{i=1}^{n}\zeta\left(\frac{x_i - \mu}{\sigma}\right)^{-\zeta-1}\left(\frac{x_i - \mu}{\sigma^2}\right) = 0
\end{aligned}$$

The following mathematical rules were applied per step:

1) Derivative of a sum is sum of derivatives.

2) All terms without σ resolve to zero. Derivative of $a \cdot \log(x)$ is $\frac{a}{x}$.

3) Chain rule: first derive with respect to $a = \frac{x_i - \mu}{\sigma}$, multiplied by derivative of a with respect to σ.

4) Cancel out the negative signs within the summation.

From this result, we isolate σ:

$$\begin{aligned}
0 &= \frac{n\zeta}{\sigma} - \sum_{i=1}^{n} \zeta \left(\frac{x_i - \mu}{\sigma}\right)^{-\zeta-1} \left(\frac{x_i - \mu}{\sigma^2}\right) \\
\frac{n\zeta}{\sigma} &= \sum_{i=1}^{n} \zeta \frac{(x_i - \mu)^{-\zeta-1}}{\sigma^{-\zeta-1}} \left(\frac{x_i - \mu}{\sigma^2}\right) \\
\frac{n\zeta}{\sigma} &= \sum_{i=1}^{n} \left[\frac{\zeta}{\sigma^{-\zeta-1}\sigma^2}(x_i - \mu)^{-\zeta-1}(x_i - \mu)\right] \\
\frac{n\zeta}{\sigma} &= \frac{\zeta}{\sigma^{-\zeta+1}} \sum_{i=1}^{n} (x_i - \mu)^{-\zeta} \\
\frac{\sigma^{-\zeta+1}}{\sigma} &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^{-\zeta} \\
\sigma^{\zeta} &= \frac{n}{\sum_{i=1}^{n} (x_i - \mu)^{-\zeta}} \\
\sigma &= \sqrt[\zeta]{\frac{n}{\sum_{i=1}^{n} (x_i - \mu)^{-\zeta}}}
\end{aligned}$$

The following mathematical rules were applied per step:

1) Set partial derivative of $l(\mu, \sigma, \zeta)$ with respect to σ to zero.

2) Isolate first term containing σ. Distribute exponent $-\zeta - 1$ among numinator and denominator of fraction.

3) Reassemble factors such that all σ and all $(x_i - \mu)$ are together.

4) Apply product rule on exponents with shared bases for both all σ and all $(x_i - \mu)$, and put x_i-independent factor in front of summation sign.

5) Isolate σ to left of equation, and ζ to right, and apply product rule of exponents with shared bases.

6) Apply product rule of exponents with shared bases on σ, and raise both sides of the equation to the power of -1.

7) Take ζ-degree root of both sides of the equation to cancel left-side exponent.

4. **Calculate the maximum likelihood estimate of scale parameter. Use two methods: (1) via the formula derived in the previous question; (2) via numerical optimization in R (e.g. with the `optim` function).**

Calculating the MLE of the scale parameter σ can be calculated **(1) analytically** as follows:

$$

$$ with R code this is calculated as:

```
sqrt(length(hr)*sum((hr-122)^1.2))
```

```
## [1] 438.14
```

We can also use a **(2) numerical optimization algorithm** in R to solve for σ.

6. **With the estimated parameter (and the given location and shape parameters), give an estimate of the probability that the maximum heart rate is larger than 140 beats per minute.**

---

7. **Suppose that there is similar data from another athlete (81 observations). Under the assumption that the shape parameter is the same for the two athletes, but their location and scale parameters may be different, construct the log-likelihood function.**
