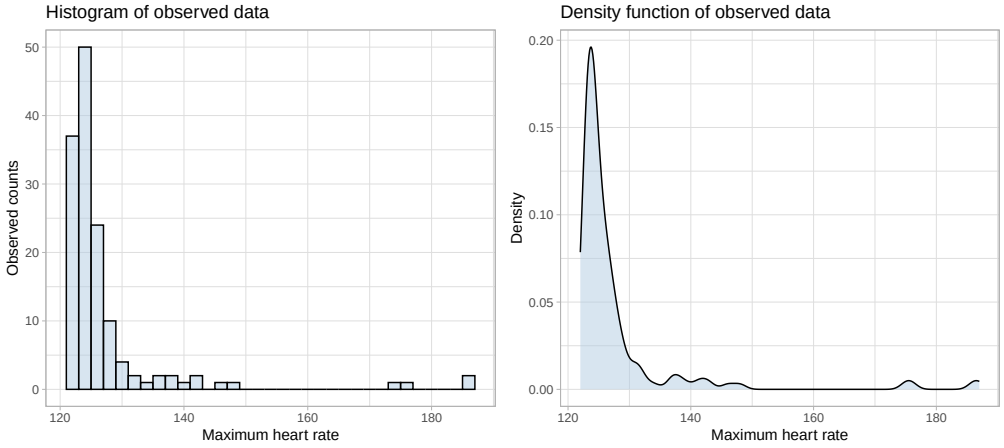
Principles of Statistical Data Analysis: Assignment 1

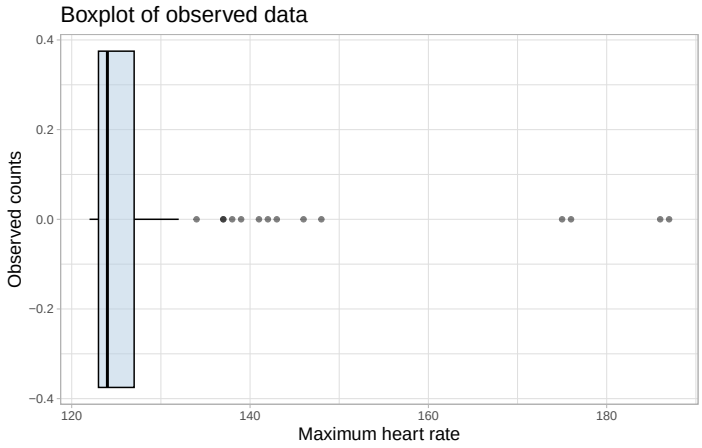
Ayala Denul, Nathan Laroy, Ole Schacht

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?**

The below two visualization show that the data is heavily skewed. The skewness is very distinct; 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. Again, it is clear that most of the data is located in said interval.



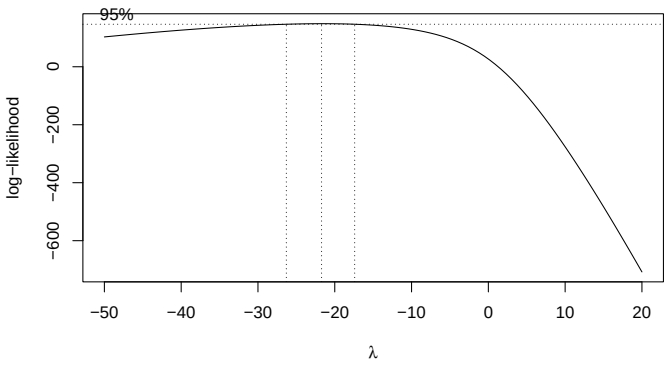
Below we present a boxplot, showcasing that the maximal heart rates are far removed to the right from the bulk of the data. For descriptive purposes, it is advised to use statistics which are relatively insensitive to outliers, e.g., the median. We find that `median(hr)` $= 124$. The minimum and maximum values equal 122 and 187, respectively.

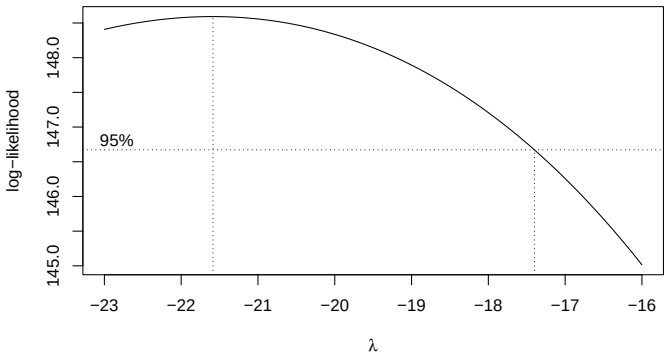


To give an estimate of data dispersion, it is also advised to employ statistics insensitive to outliers. For example, interquartile range (IQR). The IQR is defined as $p_{75} - p_{25}$ and here equals 4, which is arguably tight.

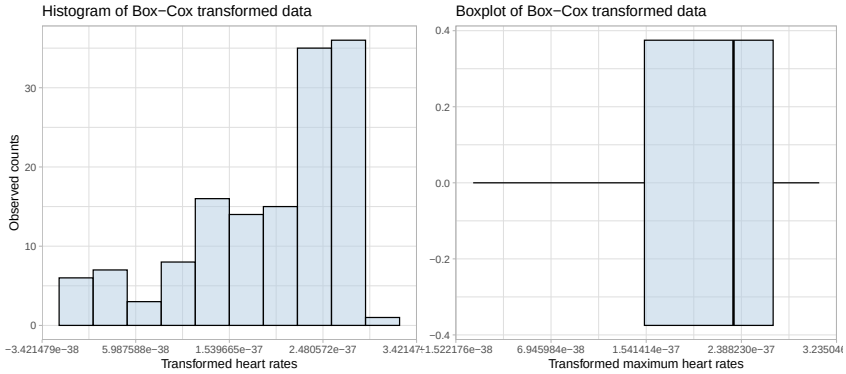
transformation

Sometimes, it may be advised to transform data so as to better approximate a normal distribution. We apply the popular Box-Cox power transformation procedure, which consists of finding an exponent λ which optimizes the transformation in yielding a close to normal distribution. We employ the `boxcox()` function from `R`'s `MASS` package. Inspection of the resulting log-likelihood function over a broad interval and after zooming in on its vertex shows a maximum at $\lambda = -21.717$. We consider an estimate λ closest to -1 within the 95% CI around the maximum as optimal, or here $\lambda = -17.5$.





We transform the data as follows: Y^{λ} if $\lambda \neq 0$. Where Y equals the `hr` data set. This yields a moderately left-skewed distribution which approximates a normal distribution more closely, but admittedly not satisfactorily. 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.**

The following Fréchet probability density function was given:

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

For constructing an MLE, we may rely on the data being i.i.d. Using the multiplication rule (and retaining the constraint on x), we may formulate the following **likelihood function**:

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

A log-transform of this function will allow for easier calculations when deriving $\hat{\sigma}_{MLE}$. Hence, we define

$l(\mu, \sigma, \zeta) \equiv \log\left[L(\mu, \sigma, \zeta)\right]$, which may be shown, using standard rules of algebra, to equal:

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

which in turn, by further applying the ratio rule of logarithms, and after distributing $(-1-\zeta)$ over terms, cancelling $\log(\sigma)$ and extracting of x_i-independent terms from the summation, resolves to:

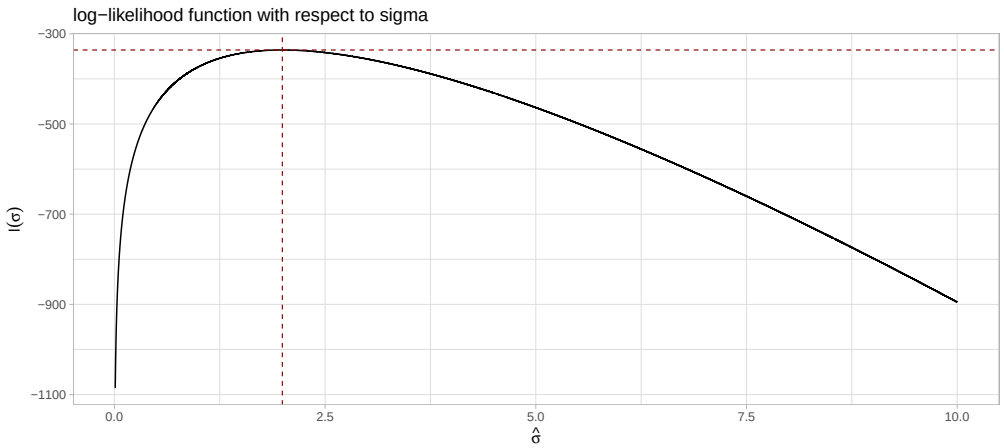
$$\begin{aligned} &= \sum_{i=1}^{n} \log(\zeta) - \log(\sigma) + (-1-\zeta)\left[\log(x_i - \mu) - \log(\sigma)\right] - \left(\frac{x_i - \mu}{\sigma}\right)^{-\zeta} \\ &= \sum_{i=1}^{n} \log(\zeta) - \log(\sigma) + (-1-\zeta)\log(x_i - \mu) + \log(\sigma) + \zeta\log(\sigma) - \left(\frac{x_i - \mu}{\sigma}\right)^{-\zeta}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

Note that we put $-\sum$ to get only $+$ inside the summation.

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 $\zeta = 1.2$.

Given that $\mu = 122$, we may exclude the 34-th observation from further analysis on account of the constraint that $f(x)$ is defined for $x > \mu$ (see also Ufora forum discussion). Note that $n_{new} = 140$.

The graph below shows the log-likelihood function. Note that $\hat{\sigma}_{MLE}$ corresponds to the σ value for which the partial derivative of this function equals zero. Here, $\hat{\sigma}_{MLE} = 1.993$ (i.e., the intersection between red dotted lines).



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

An analytic expression for $\hat{\sigma}_{MLE}$ may be achieved via the the partial derivative of $l(\mu, \sigma, \zeta)$ with respect to σ. By applying rules of derivation (sum of derivatives, chain rule), one may show that:

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

From this result, by reordering terms and applying standard rules of algebra (e.g., exponentiation with equal bases), we isolate σ and obtain the following $\hat{\sigma}_{MLE}$:

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

5. **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).**

(1) Calculating $\hat{\sigma}_{MLE}$ can be done **analytically** as follows:

$$\sqrt[1.2]{\frac{140}{\sum_{i=1}^{140}(x_i-122)^{-1.2}}} \approx 1.993$$

(2) We can also use a **numerical optimization algorithm** in `R` to solve for σ. Note that we now seek to minimize the function. The `R` functions `nlm()` , `optim()` and `optimize()` return identical rounded results.

```
loglik_optimization <- function(sigma, x, n=length(x), zeta=1.2, mu=122) {
  log_lik_optimization <- -(n*log(zeta) + n*zeta*log(sigma)
                            - sum((1+zeta)*log(x-mu) + ((x - mu)/sigma)^(-zeta)))}

nlm(f=loglik_optimization, x=hr_new, p=0.001)['estimate'] # 1.993
```

```
## $estimate
## [1] 1.992538
```

```r
optim(par = 0.001, fn=loglik_optimization, x=hr_new)['par'] # 1.993
```

```
## $par
## [1] 1.9929
```

```r
optimize(f=loglik_optimization, x=hr_new, interval = c(0.01,10))['minimum'] # 1.993
```

```
## $minimum
## [1] 1.992532
```

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.**

Using the law of total probability and the Fréchet cumulative distribution function, this probability can be estimated as:

$$\begin{aligned} P(X > 140) = 1 - P(X \leq 140) &= 1 - \exp\left(-\left(\frac{x-\mu}{\hat{\sigma}}\right)^{-\zeta}\right) \\ &= 1 - \exp\left(-\left(\frac{140-122}{1.993}\right)^{-1.2}\right) \\ &\approx 0.069 \end{aligned}$$

.

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.**

Given that the second batch of data ($n_2 = 81$) is similar to the first ($n_1 = 141 - 1$), we may continue to assume i.i.d. It is given that the distribution remains formally equivalent, albeit with given parameters known to differ between the two batches (i.e., location and scale). Hence, the **likelihood function** $L(\mu, \sigma, \zeta)$ which was previously derived, may be extended to now describe a composite of two formally equivalent batches:

$$\begin{aligned} L(\mu_1, \mu_2, \sigma_1, \sigma_2, \zeta) \equiv \prod_{i=1}^{n_1} \frac{\zeta}{\sigma_1}\left(\frac{x_i - \mu_1}{\sigma_1}\right)^{-\zeta-1} \exp\left(-\left[\frac{x_i - \mu_1}{\sigma_1}\right]^{-\zeta}\right) \\ \cdot \prod_{j=1}^{n_2} \frac{\zeta}{\sigma_2}\left(\frac{x_j - \mu_2}{\sigma_2}\right)^{-\zeta-1} \exp\left(-\left[\frac{x_j - \mu_2}{\sigma_2}\right]^{-\zeta}\right) \end{aligned}$$

Consequentially, the **log-likelihood function** may simply be extended in a similar fashion, as follows:

$$\begin{aligned} l(\mu_1, \mu_2, \sigma_1, \sigma_2, \zeta) = \sum_{i=1}^{n_1} \log\left(\frac{\zeta}{\sigma_1}\right) + \log\left(\left[\frac{x_i - \mu_1}{\sigma_1}\right]^{-\zeta-1}\right) - \left(\frac{x_i - \mu_1}{\sigma_1}\right)^{-\zeta} \\ + \sum_{j=1}^{n_2} \log\left(\frac{\zeta}{\sigma_2}\right) + \log\left(\left[\frac{x_j - \mu_2}{\sigma_2}\right]^{-\zeta-1}\right) - \left(\frac{x_j - \mu_2}{\sigma_2}\right)^{-\zeta}, \end{aligned}$$

which may be resolved into

$$l(\mu_1, \mu_2, \sigma_1, \sigma_2, \zeta) = n_1 \log(\zeta) + n_1 \zeta \log(\sigma_1) - \sum_{i=1}^{n_1} (1+\zeta) \log(x_i - \mu_1) + \left(\frac{x_i - \mu_1}{\sigma_1} \right)^{-\zeta}$$
$$+ n_2 \log(\zeta) + n_2 \zeta \log(\sigma_2) - \sum_{j=1}^{n_2} (1+\zeta) \log(x_j - \mu_2) + \left(\frac{x_j - \mu_2}{\sigma_2} \right)^{-\zeta} \quad .$$