Principles of Statistical Data Analysis: Assignment 1

Ayala Denul, Nathan Laroy, Ole Schacht

2023-10-27

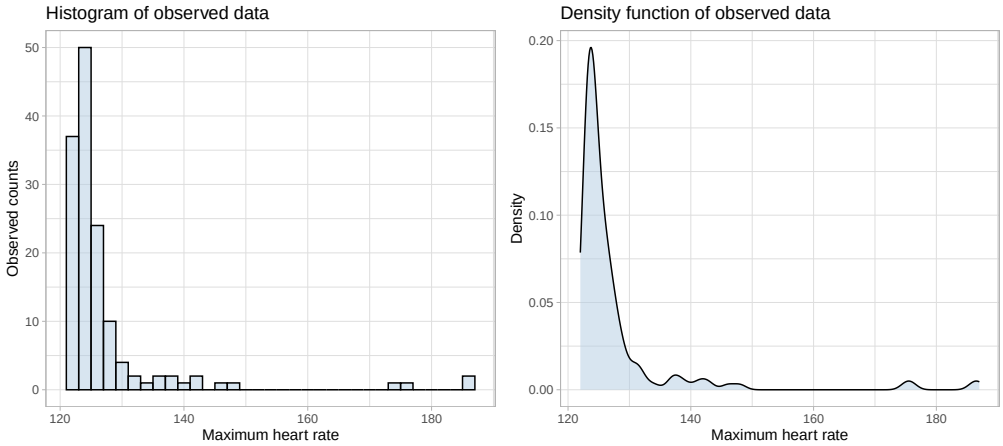
1. **Explore the data, both with numerical summary statistics and with graph(s). Include an evaluation of the shape of the distribution of the maximum heart rates. Can you find a transformation to make distribution look similar to a normal distribution?**

```r
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
```

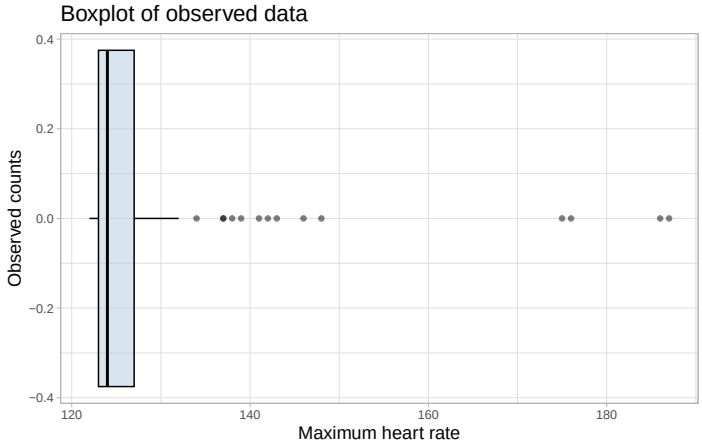
As a first step, we load in the data vector called `hr.RData`, which contains the maximal heart rate of a particular athlete during 142 sessions. Reading in the data can be done using the `load()` command.

```r
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.



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.



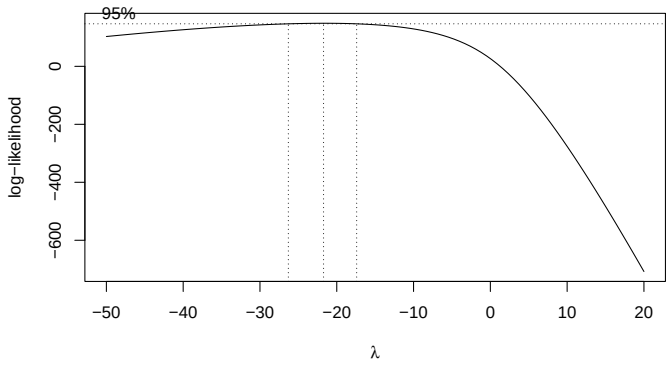
To give an estimate of the dispersion or variation in the data set, one might again naively interpret a statistic that is (highly) sensitive to outliers. One of such measures is the standard deviation, which equals 10.295 here. A better option might be to report the interquartile range (IQR). This computes $p_{75} - p_{25}$ and equals 4 in the current data set. For completeness, the minimum and maximum values equal 122 and 187, respectively.

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.

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

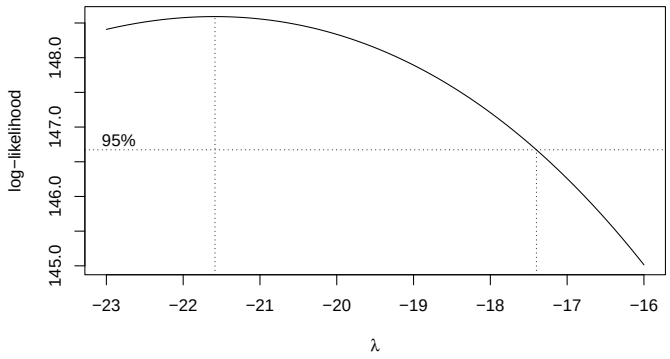



```
lambda <- boxcox1$x[which.max(boxcox1$y)]
lambda
```

```
## [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.

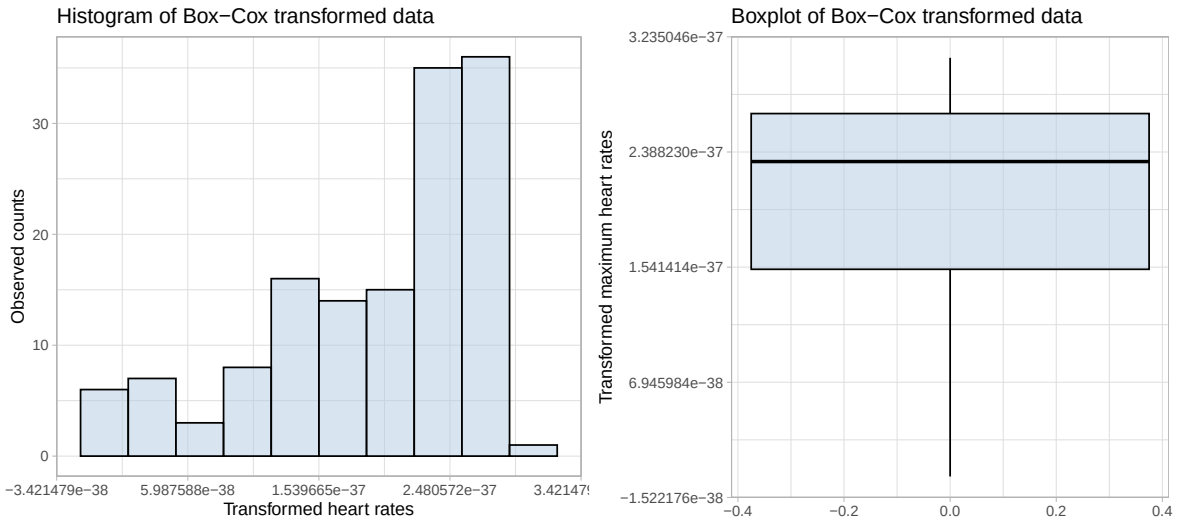
However, it is often not desirable to take the optimum with respect to λ, but rather to take the value of the sequence in λ (in steps of 0.5) that lies closest to -1, while still lying within the 95% confidence interval. For the current data set, this value corresponds to -17.5. The graph below shows a closer inspection of this log-likelihood function, which is specified between the interval $[-23, 16]$. The desired λ values equals -17.5 (indeed, 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 first construct the **likelihood function** we assume that the data originate from a Fréchet probability distribution. We therefore proceed as follows: recall from question 1 that the probability density function is given by

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

and that we are able to rely on mathematical simplifications because of the i.i.d. assumption (the observations are mutually independent, and 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. We define:

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

which 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 $\zeta = 1.2$.**

Now that we have derived the log-likelihood function of the Fréchet probability distribution, it is possible to plot these log-likelihood values for a range of different $\hat{\sigma}$ values, assuming the other two parameters are known. Plotting this relationship typically reveals a function where its vertex corresponds to the $\hat{\sigma}$ value that is most likely given the observed data. This value corresponds to the maximum likelihood estimate.

In calculating the log-likelihood function, we ran into a statistical problem because the 34-th element in the data set corresponds exactly to the given μ value. Because the deviation of this element with μ equals zero ($x_{34}-\mu=0$), the logarithm of this deviation will equal $-\inf$, which is not instructive to construct the log-likelihood function.

```
hr[34]
```

```
## [1] 122
```

We therefore exclude the 34-th observation from further analysis such that we avoid the problem of having zero-valued deviations in $x_i-\mu$. This concern was also raised in the Ufora discussion forum. Note that $n_{new} = 140$. We do acknowledge that simply removing data that doesn't fit the underlying model is not what constitutes a good research practice in real-world problems, but for the purpose of the current exercise we choose to do so nonetheless.

For the sake of parsimony, we only print the code for the first option but will also give the resulting $\hat{\sigma}_{MLE}$ for the second option. For both possible solutions we choose to plot the log-likelihood function over the following interval for σ in increments of `0.0001`: $[0.001, 10]$.

In the code below we create a new data vector without element 34. We then define the function `loglik` which computes the log-likelihood function for a given input of arguments. The following formal arguments are needed: `sigma` taking the value of σ, `x` specifying the data vector together with the sample size `n`, and `zeta` and `mu` taking their respective given values. We then construct a vector of possible σ values (taking

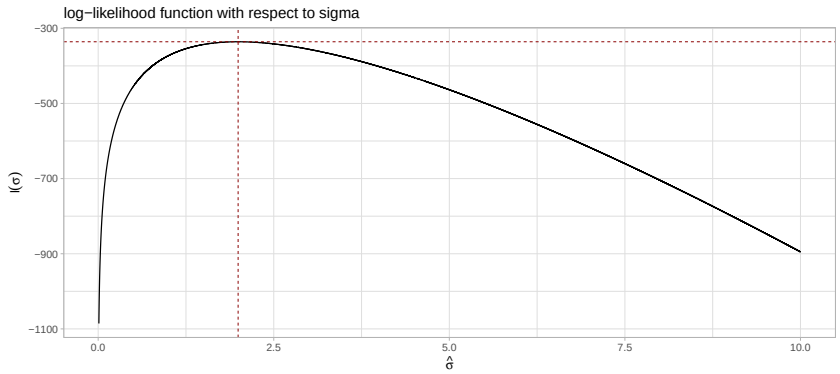
increments of `0.0001`: $[0.001, 15]$) across which the function will be computed. Finally, we use the meta-function `sapply()` to compute the function across the range of `$\sigma$` values, and store this information in a vector called `loglik_value`.

```
hr_new <- hr[-34]

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

sigma <- seq(0.01, 10, by = 0.0001)
loglik_values <- sapply(sigma, loglik, x = hr_new)
```

The graph below shows the log-likelihood function for σ. The maximum likelihood estimate ($\hat{\sigma}_{MLE}$) corresponds to the σ value for which the derivative of this function equals zero (the vertex point). This maximum is shown as the intersection of the dotted red lines. With this data, $\hat{\sigma}_{MLE} = 1.993$.




```
sigma[which.max(loglik_values)]
```

```
## [1] 1.9925
```

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 $\hat{\sigma}_{MLE}$, we will calculate the partial derivative of $l(\mu, \sigma, \zeta)$ with respect to σ. Starting from the log-likelihood function derived earlier:

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

we derive the partial derivative as follows:

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

The following mathematical rules were applied per step:

1) The derivative of a sum is the sum of the derivatives.
2) All terms without σ resolve to zero. The derivative of $a \cdot \log(x)$ is equal to $\frac{a}{x}$.
3) Chain rule for derivatives: first we derive with respect to $a = \frac{x_i-\mu}{\sigma}$, multiplied by the derivative of a with respect to σ.
4) Cancel out the negative signs within the summation.

From this result, we then 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 the partial derivative of $l(\mu, \sigma, \zeta)$ with respect to σ to zero.
2) Isolate the first term containing σ. Distribute the exponent $-\zeta - 1$ among both the numerator and denominator of the fraction.

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

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

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

6) Apply the 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.

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:

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

With R code this is calculated as:

```
zeta <- 1.2 ; mu <- 122 ; n <- length(hr_new)
(n/sum((hr_new-mu)^(-zeta)))^(1/zeta)
```

```
## [1] 1.992539
```

(2) We can also use a **numerical optimization algorithm** in `R` to solve for σ. Note that we now seek to minimize the function (hence the minus sign). Both `R` functions `nlm()` and `optimize()` give identical results to the analytical calculation.

```
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']
optim(par = 0.001, fn=loglik_optimization, x=hr_new)['par']
optimize(f=loglik_optimization, x=hr_new, interval = c(0.01,10))['minimum']
```

```
## $estimate
## [1] 1.992538
##
## $par
## [1] 1.9929
##
## $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 CDF, this estimated probability equals 0.069.

```
round(1 - exp(-((140-mu)/sigma_hat)^-zeta),3) # probability
```

```
## [1] 0.069
```

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:

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

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

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

which may be resolved into

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

and this fully expands the new log-likelihood function $l(\cdot)$.