

Exam exercise for Module 1: Wind speed distributions



In this workshop we consider a continuous probability distribution called the Weibull distribution. Among other things, it is used to model wind speed distributions.

We recommend that you answer the exercises using Rmarkdown (you can simply use the exam Rmarkdown file as a starting point).

Part I: The Weibull distribution

The Weibull distribution depends on two parameters $k > 0$ and $\lambda > 0$. If X follows a Weibull distribution with parameters k and λ , we write $X \sim \text{weibull}(k, \lambda)$. In this case, X has the probability density function

$$f(x) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

and the distribution function

$$F(x) = \begin{cases} 1 - e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

The parameter k is called the shape parameter, since it determines the shape of the distribution, while λ is called the scale parameter, because it works by scaling the x -axis.

1. Use the `mosaic` package with the `plotDist` function to make plots of different parameter combinations to demonstrate that λ is a scale parameter and k is a shape parameter. (Hint: `plotDist("weibull", params = list(shape = ., scale = .))`)
2. Assume that $x \geq 0$. Show that the distribution function $F(x)$ satisfies

$$\ln(-\ln(1 - F(x))) = -k \ln(\lambda) + k \ln(x).$$

Part II: Wind speed measurements

In this part we consider a data set containing wind speed measurements from a Danish weather station located at Sjølsmark. The data set contains the wind speed measured at 12 noon every day of January in the years 2001-2019. We first load the data set:

```
speed<-read.delim("https://asta.math.aau.dk/datasets?file=windSpeed.txt",header=FALSE)[,1]
```

1. Draw a histogram of the wind speed observations by editing the R chunk below. Explain how a histogram is constructed. Do you think the observations come from a normal distribution?

```
# gf_histogram(~...,bins=25)
```

In the following we will convince ourselves that the data actually comes from a Weibull distribution. We order the $n = 589$ observations from smallest to largest

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

2. Argue that $F(x_{(i)}) \approx \frac{i}{n}$ for $i = 1, \dots, n$. (Hint: How many observations are less than or equal to $x_{(i)}$?)
3. Using Exercise 2 in Part I, argue that if the observations come from a `weibull(k, λ)` distribution, then

$$\ln(-\ln(1 - \frac{i}{n})) \approx -k \ln(\lambda) + k \ln(x_{(i)}).$$

The code below computes a vector containing the values $v_i = \ln(-\ln(1 - \frac{i}{n}))$ and a vector containing the values $u_i = \ln(x_{(i)})$.

```
n<-length(speed)
sortedSpeed<-sort(speed)
u<-log(sortedSpeed)
CDF<-(1:n)/n
v<-log(-log(1-CDF))
#gf_point(...~...) %>%gf_lm()
```

4. Argue that the points (u_i, v_i) should lie approximately on a straight line if the observations come from a `weibull(k, λ)` distribution. Edit the code above to check that this is the case.
5. The intercept and slope of the line can be found to be -2.82 and 1.78 , respectively. Use this to give estimates of the parameters k and λ of the model. Insert these values in the code below to plot the histogram together with the approximate density (`shape` is k and `scale` is λ).

```
#gf_dhistogram( ~ speed, bins = 25) %>%
#gf_dist("weibull", shape = ..., scale = ..., col = "red")
```

Part III: Sample mean and the central limit theorem

In this last exercise, we investigate the distribution of the sample mean when a random sample is taken from a population having a `weibull(k, λ)` distribution. We will use the values of k and λ that you found in Part II, Exercise 5 to mimic a sample of wind speed measurements.

Denote by μ the mean of the population distribution, `weibull(k, λ)`, and by σ^2 the variance of the population distribution.

The numeric values of μ and σ^2 for choices of λ and k can be calculated $\mu = \lambda \Gamma(1 + 1/k)$ and $\sigma^2 = \lambda^2 [\Gamma(1 + 2/k) - \{\Gamma(1 + 1/k)\}^2]$, where $\Gamma(x)$ denotes the gamma function.

1. Using the values of k and λ from Part II, Exercise 5, what is the mean and standard deviation? (Hint: You can use the function `gamma()` in R to compute the gamma function.)
2. Suppose that a sample consists of 30 observations from this distribution. We denote the sample mean by `x_bar`. Using the central limit theorem, answer the following questions:
 - What is the expected value of `x_bar`?
 - What is the standard deviation of `x_bar` (also called the standard error)?

- What is the approximate distribution of \bar{x} ?

The code below generates 30 independent realizations of a Weibull distribution with parameters k and λ . One may think of this of as simulated random sample of 30 independent wind speed observations.

```
# x<-rweibull(30, shape=..., scale = ... )  
# mean(x)
```

3. Insert the values of k and λ from Part II, Exercise 5 in the code. Run the command a few times. Is each sample mean close to what you expected?

Use `replicate` to repeat the sampling 500 times and save each mean value in the vector \bar{x} :

```
# x_bar <- replicate(500, mean(rweibull(30, shape=..., scale = ...) ))
```

4. Calculate the mean and standard deviation of the values in \bar{x} . How do they match with what you expected?
5. Make a QQ-plot to assess the distribution of \bar{x} . Does this look like what you would expect?

```
#qqnorm(...)  
#qqline(...)
```