

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

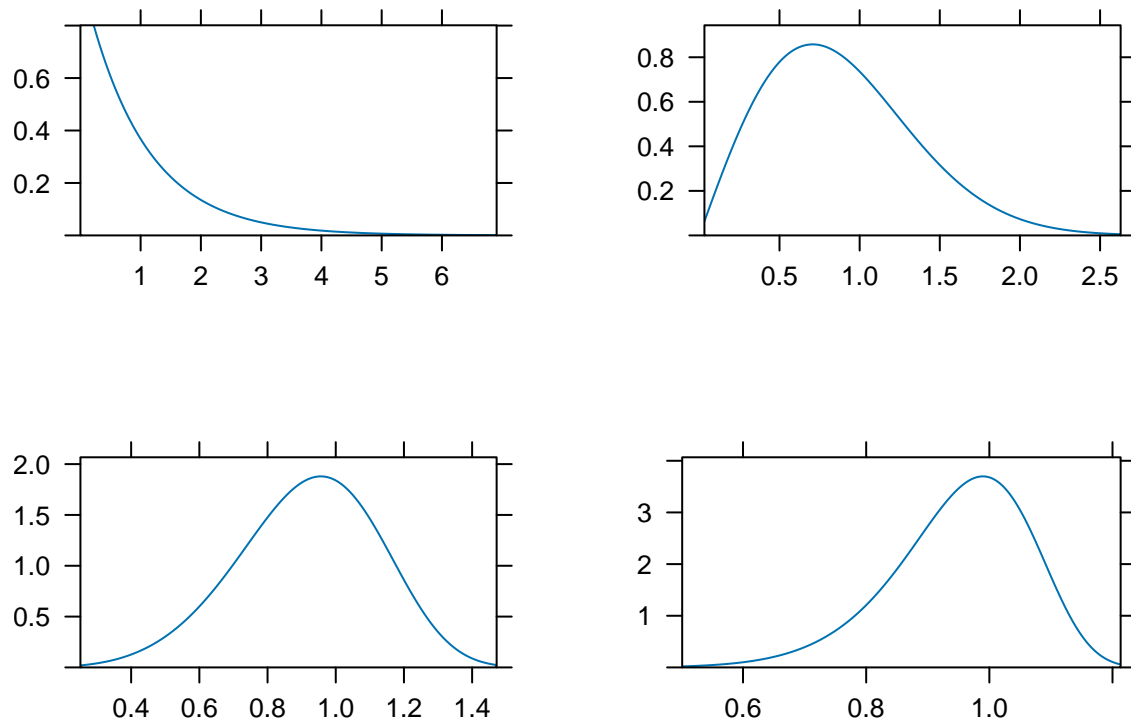
$$F(x) = \begin{cases} 1 - e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

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 = .))`) First we run the variable `k` or shape through 1 to 10 and see how the shape changes, it changes from skewed to the left to be skewed to the right.

```
lambda <- 1
k <- 1
w1 <- plotDist("weibull", params = list(shape = k, scale = lambda))
k <- 2
w2 <- plotDist("weibull", params = list(shape = k, scale = lambda))
k <- 5
w3 <- plotDist("weibull", params = list(shape = k, scale = lambda))
```

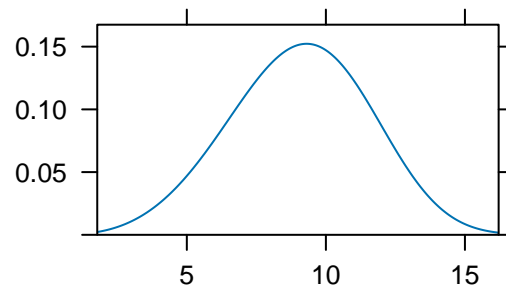
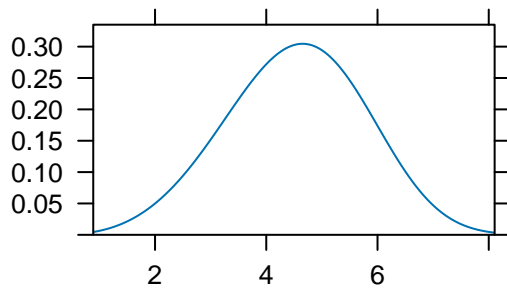
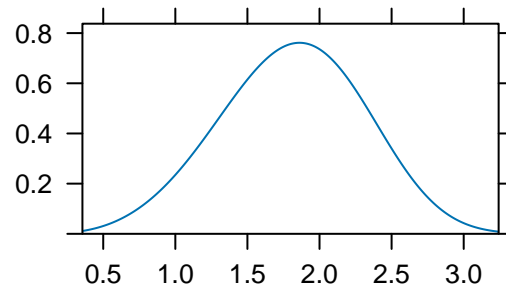
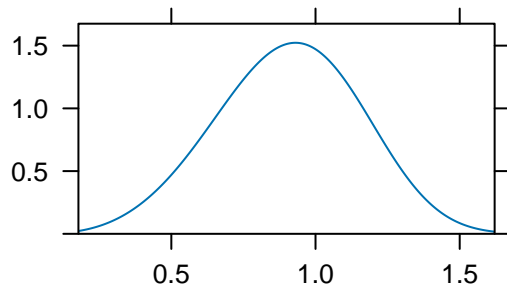
```
k <- 10
w4 <- plotDist("weibull", params = list(shape = k, scale = lambda))
grid.arrange(w1, w2, w3, w4, ncol = 2)
```



We then

the run the same changes but for lambda and see how the scale changes:

```
lambda <- 1
k <- 4
w1 <- plotDist("weibull", params = list(shape = k, scale = lambda))
lambda <- 2
w2 <- plotDist("weibull", params = list(shape = k, scale = lambda))
lambda <- 5
w3 <- plotDist("weibull", params = list(shape = k, scale = lambda))
lambda <- 10
w4 <- plotDist("weibull", params = list(shape = k, scale = lambda))
grid.arrange(w1, w2, w3, w4, ncol = 2)
```



Here we see that lambda defines how “sharp” or “pointy” the function is, larger lambda means longer tails or larger variance. It also seems that the center is around lambda's value.

2. Assume that $x \geq 0$. Show that the distribution function $F(x)$ satisfies

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

```
library("png")
pp <- readPNG("1.2.png")
plot.new()
rasterImage(pp,0,0,1,1)
```

We substitute $F(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k}$ into the equation
 $\ln\left(-\ln\left(1 - \left(1 - e^{-\left(\frac{x}{\lambda}\right)^k}\right)\right)\right) = -k \cdot \ln(\lambda) + k \cdot \ln(x)$

We are simplifying the inner (O) brackets

$$1 - \left(1 - e^{-\left(\frac{x}{\lambda}\right)^k}\right) = e^{-\left(\frac{x}{\lambda}\right)^k}$$

We substitute the simplified expression back into the equation.

$$\ln\left(-\ln\left(e^{-\left(\frac{x}{\lambda}\right)^k}\right)\right) = -k \cdot \ln(\lambda) + k \cdot \ln(x)$$

We use log rules $\ln(a^b) = b \cdot \ln(a)$ on -1 power

$$\ln\left(-1 \cdot (-1) \cdot \ln\left(e^{\left(\frac{x}{\lambda}\right)^k}\right)\right) = -k \cdot \ln(\lambda) + k \cdot \ln(x)$$

-1 cancels and we are left with

$$\ln\left(\left(\frac{x}{\lambda}\right)^k\right) = -k \cdot \ln(\lambda) + k \cdot \ln(x)$$

We use the same log rules to move k

$$k \cdot \ln\left(\frac{x}{\lambda}\right) = -k \cdot \ln(\lambda) + k \cdot \ln(x)$$

Log of a division is $\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)$

$$k \cdot (\ln(x) - \ln(\lambda)) = -k \cdot \ln(\lambda) + k \cdot \ln(x)$$

This leaves us with the proof that the distribution function $F(x)$ satisfies the equation.

$$k \cdot \ln(x) - k \cdot \ln(\lambda) = -k \cdot \ln(\lambda) + k \cdot \ln(x)$$

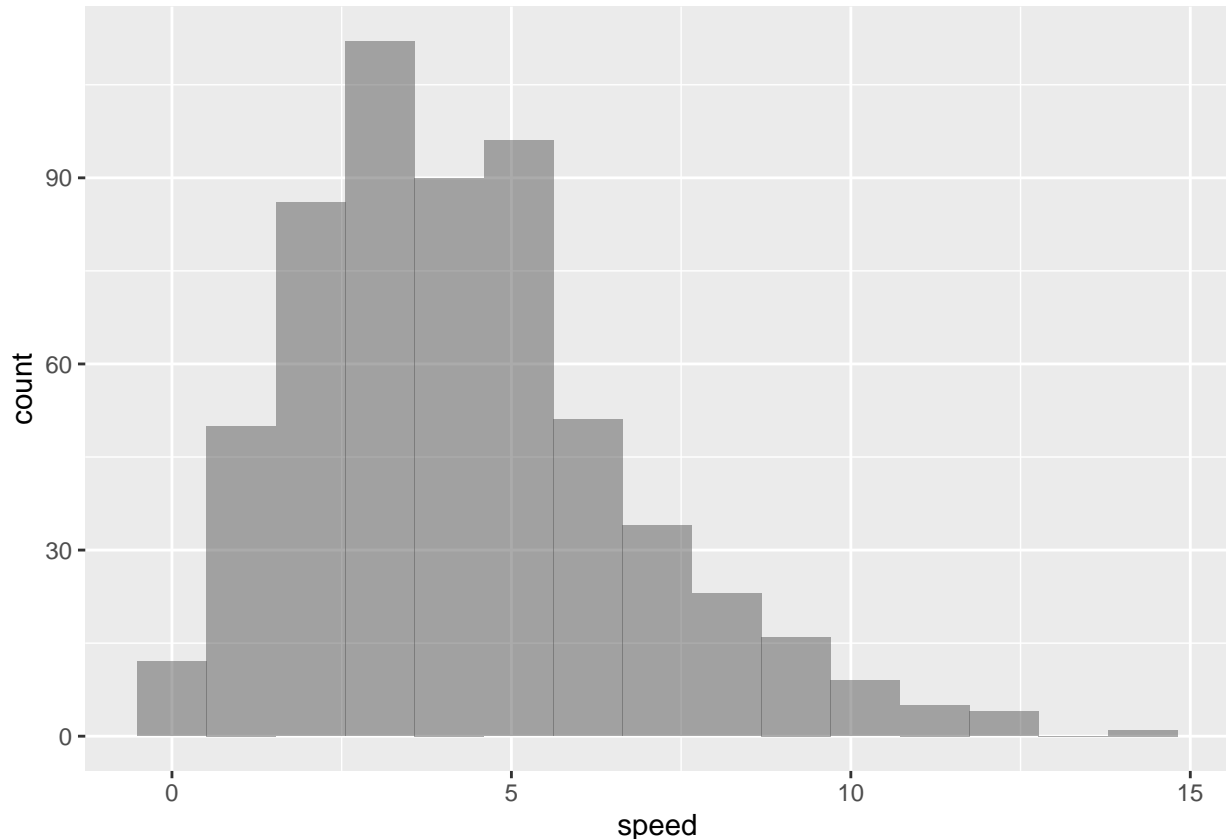
Part II: Wind speed measurements

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?

```
#hist(speed)
gf_histogram(~speed,bins=15)
```



A histogram shows how many times something occurred in a given interval. In this instance, we want 15 bins, such that the largest bin describes the maximum outcome. In this instance the max is 15, and thus we get that the bin size should be $15/15 = 1$. Thus the first bin describes all counts in the interval 0 - 0.999 m/s. bin 2 describes from 1. to 1.999 m/s etc. The y axis is the count, i.e. it is the total number of occurrences that fall in a given interval. A histogram can be used to show the distribution of a measured system.

This is clearly not gaussian (normal), it is skewed, and looks a lot like Weibull dist. with k approx. around 2, see figures from task 1.

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 $x_{(i)}$?)

We know from 2.1) that the values X_i approximately follow a Weibull distribution. We also know that $F(x_i)$ is the Weibull distribution function. Because the sample values ordered from low to high follow a Weibull distribution as shown in 2.1) it is reasonable to assume that they follow $F(x)$. Also we are asked if $F(x) \approx i/n$ for $i = 1..n$ we know that $\lim(i/n) = 1$ when $i \rightarrow n$. We also know that $F(x) = 1 - \exp(-(x/\lambda)^k)$ for any values of λ and k this will also converge at 1 when $x \rightarrow \infty$. Therefore i/n and $F(x)$ are approximately equal.

3. Using Exercise 2 in Part I, argue that if the observations come from a `weibull(k, λ)` distribution, then

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

Since we have concluded in 2.2) that $F(x) = i/n$ when X_i is a ranked list of observations from low to high, it is fair to say that the approximation i/n will also fulfill the empirical distribution function $\ln(-\ln(1-F(x))) = -k \ln(\lambda) + k \ln(x_i)$ this means that the empirical distribution function equation $\ln(-\ln(1-i/n)) \approx -k \ln(\lambda) + k \ln(x_i)$ is also valid.

Another way to look at it is that we have concluded in 2.2 that $(1-i/n) \sim 1 - \exp(-(x/\lambda)^k)$ so we can approximate the two sides of the equation as equal.

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.

If the observations come from a Weibull distribution then they will work for the discrete empirical distribution function $\ln(-\ln(1-i/n)) \approx -k \ln(\lambda) + k \ln(x_i)$. This means the equation is true. If we examine both sides of the equation then the equation is of the form $y = ax - b$ and this is the function for a straight line. They will only follow this straight line if the observations follow a weibull distribution because the empirical distribution function is only valid in this case.

We can examine the EMF further to see it is a straight line. We just need to analyze the $\ln(1-F(x))$ expression in the CDF

```
library("png")
pp <- readPNG("2.4.png")
plot.new()
rasterImage(pp,0,0,1,1)
```

$$\begin{aligned} & \ln \left(1 - \left(1 - e^{-\left(\frac{x}{\lambda} \right)^k} \right) \right) \\ & \ln \left(1 - 1 + e^{-\left(\frac{x}{\lambda} \right)^k} \right) \\ & = \ln \left(e^{-\left(\frac{x}{\lambda} \right)^k} \right) \\ & = - \left(\frac{x}{\lambda} \right)^k \end{aligned}$$

If we substitute this expression back into the CDF we get the following

```
library("png")
pp <- readPNG("2.4.2.png")
plot.new()
rasterImage(pp,0,0,1,1)
```

$$\frac{\ln\left(-\left(-\left(\frac{x}{\lambda}\right)^x\right)\right)}{\ln\left(\left(\frac{x}{\lambda}\right)^x\right)} = -k \cdot \ln(\lambda) + k \cdot \ln(x)$$

Log-log plot
y = kx - kc
Straight line!

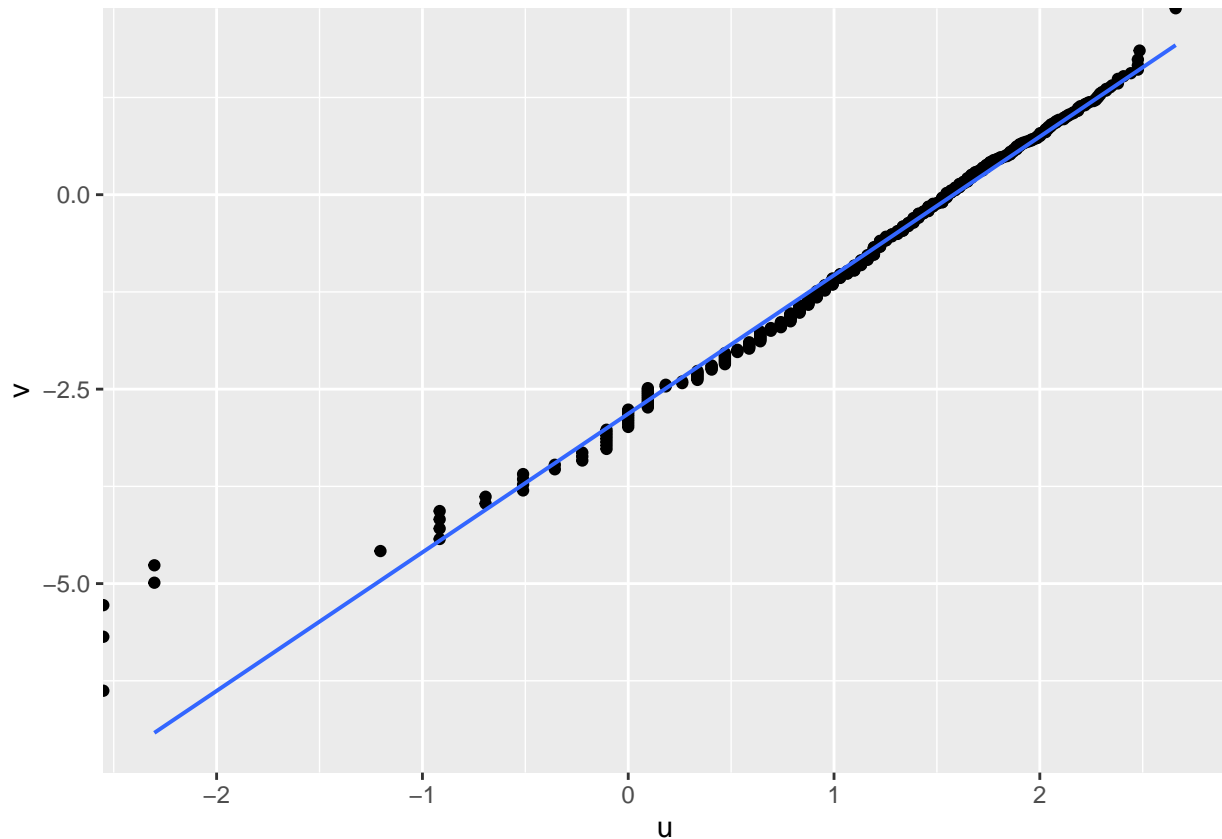
The picture shows that the log-log plot should produce a straight line if the points follow a weibull distribution. We can plot the points and see if true.

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(v~u) %>%gf_lm()
```

```
## Warning: Removed 4 rows containing non-finite outside the scale range
## (`stat_lm()`).
```

```
## Warning: Using the `size` aesthetic with geom_line was deprecated in ggplot2 3.4.0.
## i Please use the `linewidth` aesthetic instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



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).

We can find values of lambda and k from the straight line equation. We know from the picture on 2.4 that “k” is the slope of the line as this is the value that is multiplied by x. Our slope is then given as $k = 1.78$

The lambda value can be found from the intercept with the y-axis on the straight line as shown on the picture

```
library("png")
pp <- readPNG("2.4.3.png")
plot.new()
rasterImage(pp,0,0,1,1)
```

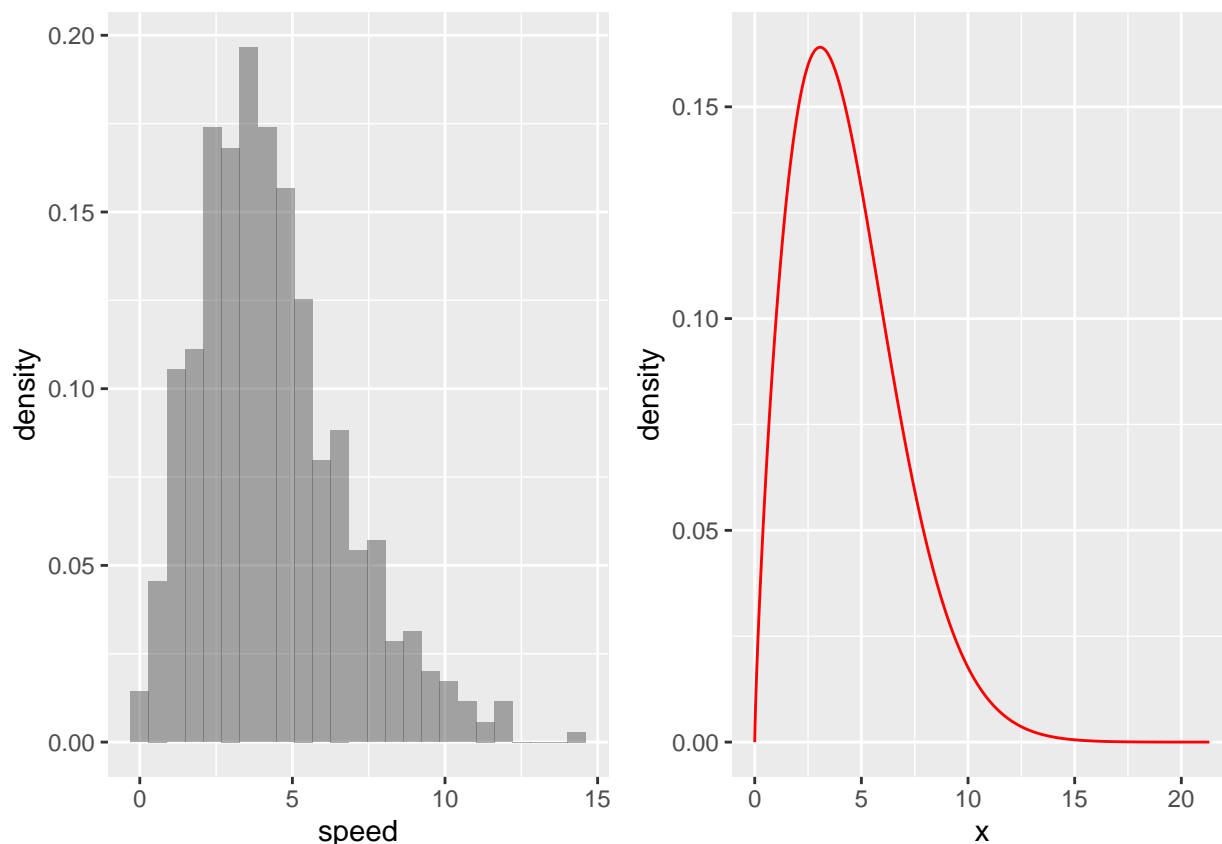
$$c = -k \cdot \ln(\lambda)$$

$$\lambda = e^{-\frac{c}{k}}$$

$$\lambda = e^{-\frac{-2.82}{1.78}} = \lambda = 4.875729152$$

Using $k = 1.78$ and $\lambda = 4.876$ we can plot the approximate distribution along with the histogram

```
k1 <- 1.78
lambda1 <- 4.876
w5 <- gf_dhistogram( ~ speed, bins = 25) #>%
w6 <- gf_dist("weibull", shape = k1, scale = lambda1, col = "red")
grid.arrange(w5,w6, ncol = 2)
```



We can see that the approximated distribution function looks very similar to the histogram.

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 $\text{weibull}(k, \lambda)$ 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, $\text{weibull}(k, \lambda)$, 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 \bar{x} . Using the central limit theorem, answer the following questions:
 - What is the expected value of \bar{x} ?
 - What is the standard deviation of \bar{x} (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(...)
```