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 $X \sim \texttt{weibull}(k, \lambda)$. 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

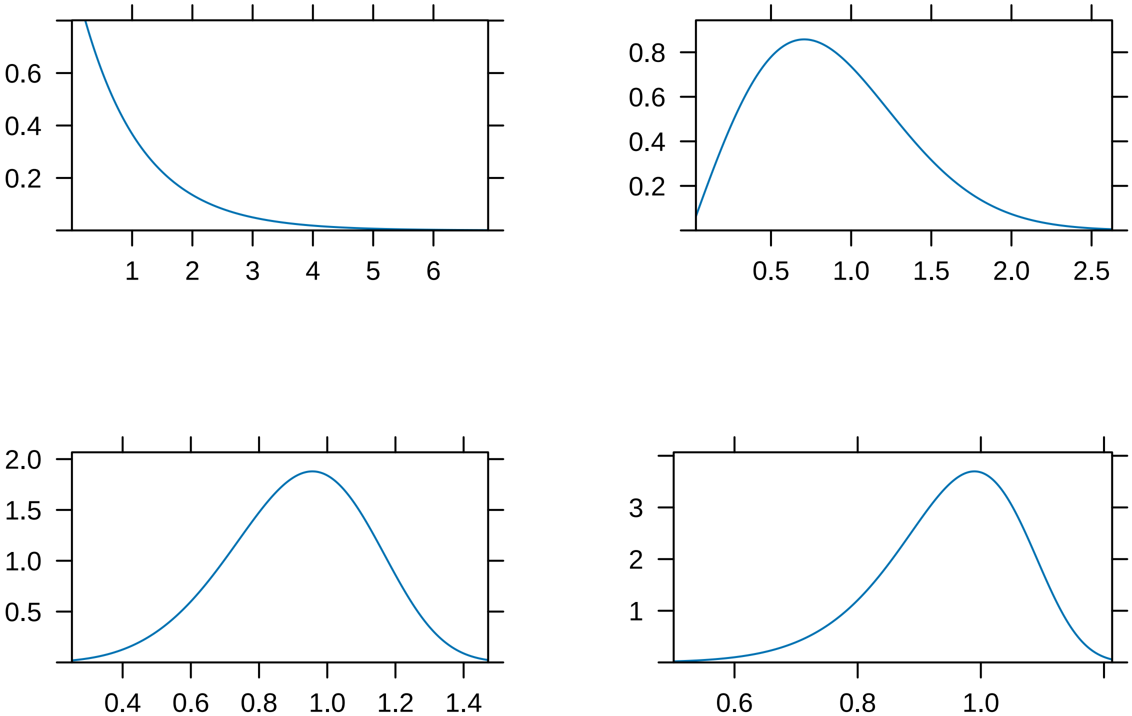
$$F(x) = \begin{cases} 1 - e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

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.

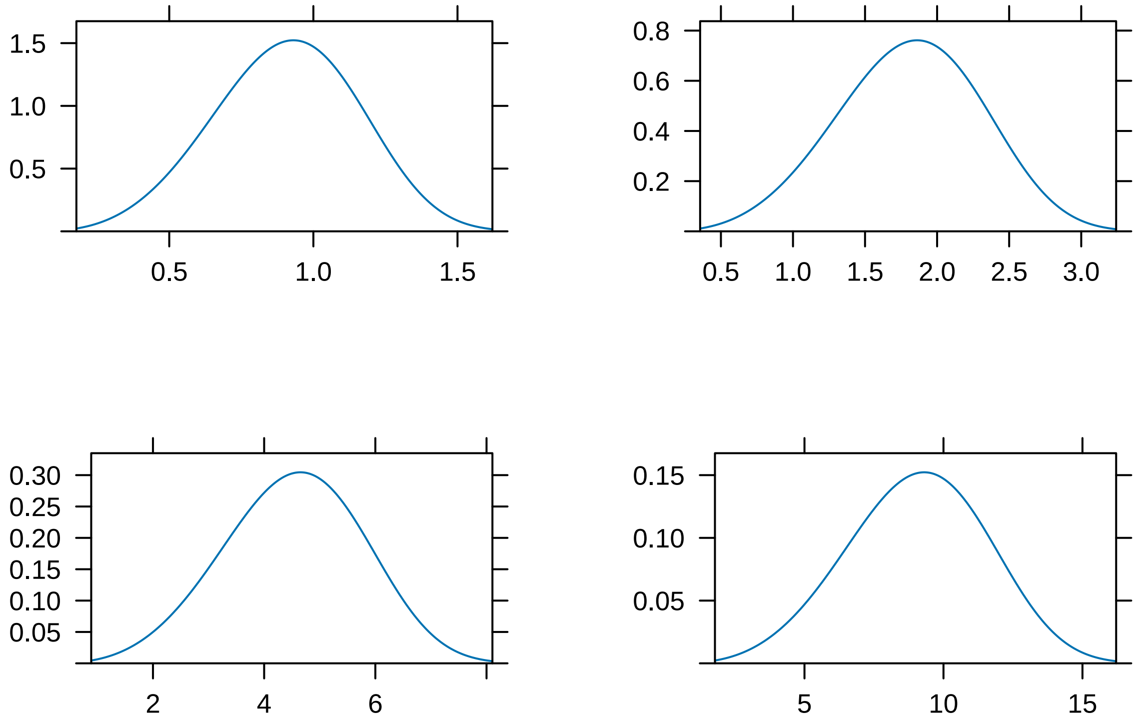
```
lambda <- 1
k <- 1
w1 <- plotDist("weibull", params = list(shape = k, scale = lambda))
k <- 2
w2 <- plotDist("weibull", params = list(shape = k, scale = lambda))
k <-5
w3 <- plotDist("weibull", params = list(shape = k, scale = lambda))
```

```r
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:

```r
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 lambdas value.

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

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

```r
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 (()) 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

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

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.

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

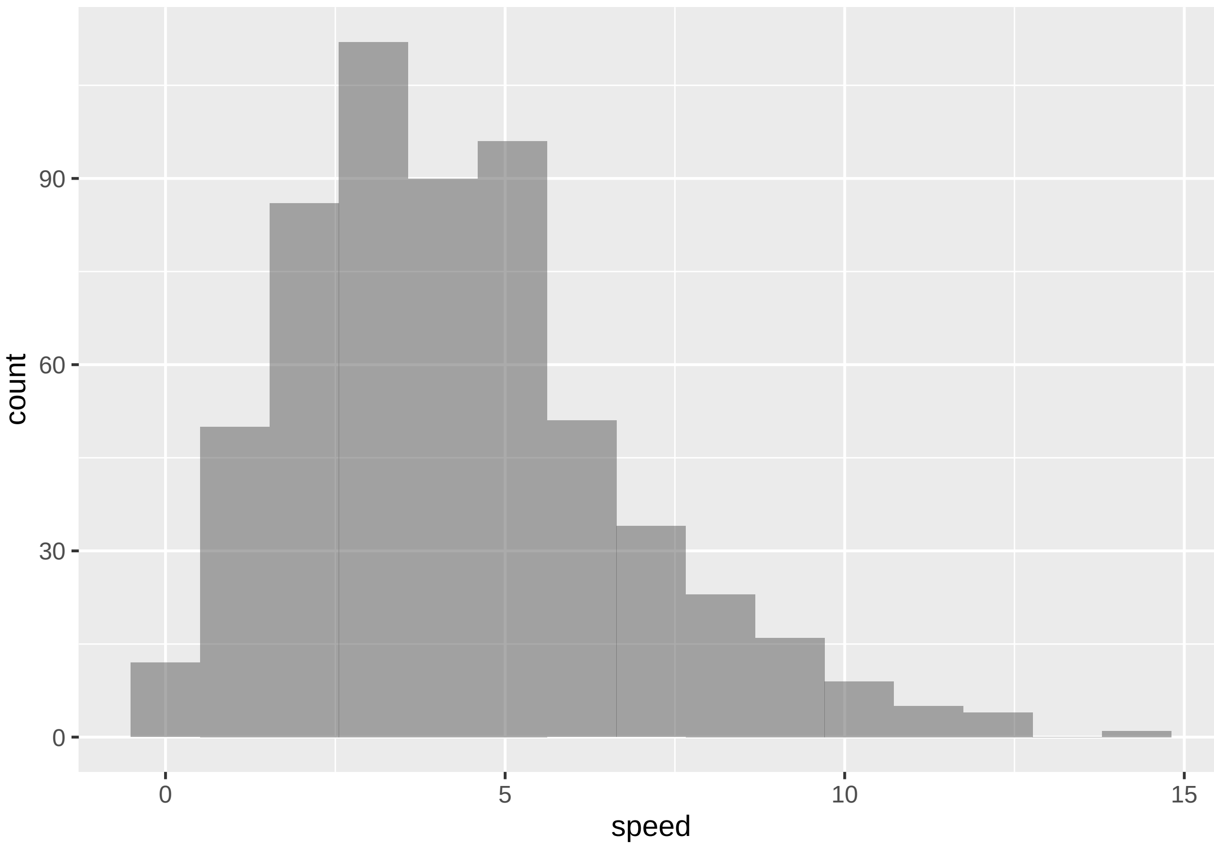
Part II: Wind speed measurements

In this part we consider a data set containing wind speed measurements from a Danish weather station located at Sjælsmark. 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:

```r
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?

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



A histogram shows how many times something ocoured in a given interval. in this instance, we want 15 bins, such that the largest bin describe 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 ocourances that fall in a given interval. A histrogram can be used to show the distribution of a measured system.

This is clearly bot 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

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

2. Argue that $F(x_{(i)}) \approx \frac{i}{n}$ for $i = 1, \ldots, n$. (Hint: How many observations are less than or equal to $x_{(i)}$?)

The distribution funciton F(x), gives the probability of a value being in the range from 0 to x. Thus if we where to compute F(10) it would return the probability of getting a value from 0 to 10. Much in the same way, if we say i =250, and compute 250/589 = 0.424 this means, that the probability of an output being in the span from 0 to the x value at index 250 would be 0.424. In other words, we know that all value below the i'th index must be smaller than the value at the i'th index and as such i/n approximates F(x).

First and formost the distribution fuction must have an area of 1 at infinity. The infitnity in our discrete case

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

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

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 $\sigma^2 = \lambda^2\left[\Gamma(1 + 2/k) - \{\Gamma(1 + 1/k)\}^2\right]$, 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 `x_bar`?

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 `x_bar`:

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

4. Calculate the mean and standard deviation of the values in `x_bar`. How do they match with what you expected?

5. Make a QQ-plot to assess the distribution of `x_bar`. Does this look like what you would expect?

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