Week 1 **Problems 1**

To run the following exercises, first install and then load the `R` packages:

`evd, mev, scales, lubridate, gridExtra, ggplot2, dplyr, tidyr, ggdist, ggpubr`

using `library(evd)` etc.

1. To get some intuition for the information in a probability plot, here is a function to generate samples, standardize them and make a normal probability plot of them:

```
tp <- function(n=c(10,20,50), line=F, ran.gen=rnorm, lims=c(-4,4), ...)
{ m <- length(n)
for (i in 1:m)  for (j in 1:m)
{ y <- ran.gen(n[i],...)
y <- (y-mean(y))/sqrt(var(y))
qqnorm(y,xlim=lims,ylim=lims,main=paste("n=",n[i]))
if (line) abline(0,1,lty=2) }
invisible()  }
```

(a) To produce plots for normal samples of sizes 10, 20 and 50:

```
par(mfrow=c(3,3),pty="s")
tp() # without line
tp(line=TRUE) # with line
```

Repeat the last two commands a few times. Is the line useful? What effect has sample size on the variability of the plot?

(b) To assess what happens for non-normal data, here are samples from the gamma distribution with shape parameter 4 and from the t distribution with 5 degrees of freedom:

```
tp(ran.gen=rgamma,shape=4,line=TRUE)
tp(ran.gen=rt,df=5,line=TRUE)
```

Try each several times, with and without lines and with various values of `shape` and `df`.

Write a short summary of your findings.

(c) The function below generates data that are either (1) normal, (2) heavy-tailed, (3) skewed, (4) light-tailed, (5) have outliers, or (6) rounded.

```
ran.gen <- function(i,n,m=5)
switch(i,rnorm(n),rt(n,df=m),rgamma(n,shape=m)/m,
(rbeta(n,m,m)-0.5)*m,c(rnorm(n-m),rcauchy(m)),
round(m*rnorm(n)))
par(mfrow=c(3,3),pty="s")
gen <- sample(x=1:6,size=9,replace=TRUE)  # make data type
for (i in 1:9) qqnorm(ran.gen(gen[i],500))  # make data and plot them
```

Which normal scores plot(s) correspond to which types of data? Type `gen` to see if you're right. Try the last two lines again, with 50 replaced by 25, 100, or 500.

(d) The Abisko dataset from the `mev` package contains precipitation data for all rainy days within the time frame 01.01.1913–31.12.2014 for Abisko. The following loads the data and sets precipitation to zero for the dry days:

```
data(abisko)
prec <- seq(from=min(abisko$date),to=max(abisko$date),by="day")
prec.y <- rep(0,length(prec))
prec.y[prec %in% abisko$date] <- abisko$precip
abisko <- data.frame(date=as.Date(prec),precip=prec.y)
```

To plot precipitation as a function of time:

```
plot_abisko1 <- function () {
dates <- as.Date(c("1940-01-01","1979-12-31"))
plot <- ggplot(abisko, mapping=aes(x=date, y=precip))+
geom_point(pch=16,cex=0.7)+
labs(x="", y = "Precipitation (mm)")+
scale_y_continuous(limits = c(0,79), expand = c(0, 0))+
scale_x_date(limits = as.Date(c('1913-01-01','2015-01-01')), expand = c(0, 0))+
theme_classic(base_size=11)+
theme(axis.text = element_text(size = 10),
panel.background = element_rect(fill = "white",
colour = "white",
size = 0.5, linetype = "blank"))
ggsave(filename = "figures/abisko1.png", plot = plot, bg = "white", width = 2000,
height = 1000, unit = 'px', dpi=250)
}
```

Apply `plot_abisko1()` and comment on the resulting plot (which is stored in a directory). Alternatively, you may try `plot(abisko$precip[abisko$precip>0],pch=".")`, which drops the zeros.

(e) The following function first takes the maximum precipitation for each month, and makes a QQ-plot of them against the theoretical quantiles of a Gumbel distribution. Extend the function so that it takes also yearly maxima and adds the corresponding points to the QQ-plot.

```
plot_abisko2 <- function (){
abisko.max <- matrix(NA, 102, 12)
year <- c(1913:2014)
for (i in 1:102) for (j in 1:12)
{ k <- (year(abisko$date)-1912==i & month(abisko$date)==j)
abisko.max[i,j] <- max(abisko$precip[k]) }
mon.max <- c(abisko.max)
mon.n <- length(mon.max)

plot1 <- ggplot()+
geom_point(aes(x=qgumbel(c(1:mon.n)/(mon.n+1)),
```

```
 y=sort(mon.max)), pch=16,cex=0.7)+
labs(y="Ordered maxima (mm)", x="Gumbel plotting positions")+
theme_classic(base_size=11)+
theme(axis.text = element_text(size = 10),
panel.background = element_rect(fill = "white",
colour = "white",
size = 0.5, linetype = "blank"))

pl <- cowplot::plot_grid(plotlist = list(plot1),
labels = c(""),
ncol = 1)
ggsave(filename = "figures/abisko2.png", plot = pl,
bg = "white", width = 900, height = 900, unit = 'px', dpi=250)
}
```

Apply `plot_abisko2()` and look at the resulting QQ-plots. Do you notice differences? In which case does the Gumbel distribution provide a better fit to the maxima?

2. (a) An exponential random variable X with mean $1/\lambda$, with $\lambda > 0$, has density function $f(x) = \lambda \exp(-\lambda x)$ for $x > 0$. Using the `rexp()` function, generate samples of standard exponential random variables of sizes $n = 50, 100$. Use a QQ-plot to compare the empirical and theoretical quantiles.

   (b) A homogeneous Poisson process $N(t)$ is a random process that takes values in $\mathbb{N} \cup \{0\}$ and is indexed by $t \in \mathbb{R}_+$. At time $t > 0$ its probability mass function is $\mathrm{P}(N(t) = k) = (\lambda t)^k \exp(-\lambda t)/k!$, $k \in \mathbb{N} \cup \{0\}$, where $\lambda > 0$ is called the rate or intensity.

   For independent and non-negative random variables X_i consider the process $S_k = \sum_{i=1}^{k} X_i$. Each of the X_i's is the interval between events $i-1$ and i (we set $S_0 = 0$). Let $N(t)$ denote the number of events up to time t, and assume that this is a homogeneous Poisson process with intensity λ. Show that the intervals X_i are exponentially distributed. Find the mean parameter.

   Hint: The event $N(t) = k$ occurs only when $S_k \leq t$ and $S_{k+1} > t$.

   (c) For the variable $Y = 1/X$ find the cumulative distribution function, and compare the empirical and theoretical quantiles via a QQ-plot.

3. The Lomax distribution is given by

$$\mathrm{P}(Y \leq y) = 1 - \frac{\beta^\alpha}{(\beta + y)^\alpha}, \quad y > 0, \alpha, \beta > 0.$$

   (a) Create a function in `R` for the negative log-likelihood $-\ell(\alpha, \beta)$ of a sample of size n.

   (b) Compute the maximum likelihood estimates $(\widehat{\alpha}_{\mathrm{MLE}}, \widehat{\beta}_{\mathrm{MLE}})$ of (α, β) and their standard errors for the (positive values in the) Abisko dataset from Problem 1.

   Hint: Use the function `optim(..., hessian = T)`.

   (c) Show that the log-likelihood for a sample of size n can be expressed as

$$\ell(\alpha, \beta) = n \log(\alpha/\beta) - (\alpha + 1) S(\beta),$$

where $S(\beta) = \sum_{i=1}^{n} \log(1+y_j/\beta)$. Show that, apart from additive constants (i.e., terms that do not depend on y, α, or β) the profile log-likelihood for the parameter β can be written as

$$\ell_{\mathrm{P}}(\beta) = \max_{\alpha} \ell(\alpha, \beta) = -n \log S(\beta) - n \log \beta - S(\beta), \quad \beta > 0.$$

Hence plot this function for the Abisko data. What do you conclude?