TP : An Introduction to Extreme Value Theory

by

Jonathan El Methni

1. Download and install the following packages :
   - *evd*
   - *evir*
   - *ismev*
   - *fExtremes*
2. For each dataset in the following list :
   - *marseille*
   - *portpirie*
   - *dowjones*
   - *oldage*
   - *lisbon*
   - *oxford*
   - *temps*100*m*
3. Give a complete study :
   - Descriptives statistics.
   - Choice of the modelisation : GEV and/or GPD.
   - Choice of the blocks or choice of the threshold.
   - Estimation of the parameters using the MLE and the PWM.
   - Maximum domain of attraction.
   - Return level plot and interpretation.
   - Estimation of a return level corresponding to a return period of 100 years and 1000 years or estimation of the endpoint.
   - A small conclusion.

- Each data set depicts a different situation.

1 *marseille*

The file "*Marseille.txt*" contains daily rainfall accumulations data in 10^{-1} mm in Marseille during 127 years of observations (1864–1991). The first date is August 1, 1864. All February 29 was removed. For this dataset you have to use both approaches (GEV and GPD) and compare them.

```
> pluies = read.table('marseille.txt')[,1]
> pluies.ts = ts(pluies,start=c(1864,213),frequency=365) # 213: 1 aout 1864
> plot(pluies.ts,main='Pluies maximales journalieres Marseille')
> matPluies = matrix(pluies,365,127) # decoupage en annees aout-juillet
> maxPluies = apply(matPluies,2,max,na.rm = TRUE)
> vecAn = 1864:1990
```

2 *portpirie*

This dataset shows the annual maximum sea-levels recorded at Port Pirie, a location just north of Adelaide, South Australia, over the period 1923–1987. From such data it may be necessary to obtain an estimate of the maximum sea-level that is likely to occur in the region over the next 100 or 1000 years. It seems reasonable to assume that, the pattern of variation has stayed constant over the observation period, so we model the data as independent observations from the GEV distribution.

3 *dowjones*

This dataset shows the daily closing prices of the Dow Jones Index over a 5-year period. Evidently, the level of the process has changed dramatically over the observed period, and issues about extremes of daily behavior are swamped by long-term time variation in the series. Because of the strong non-stationnary observed on the original series $X_1, \ldots, X_n$ the data are transformed as

$$\tilde{X}_i = \log(X_i) - \log(X_{i-1}).$$

The transformed series is reasonably close to stationarity. For convenience of presentation we re-scale the data as

$$\tilde{X} \to 100\tilde{X}$$

4 *oldage*

The oldage data frame has 66 rows and 2 columns. The columns contain the oldest ages at death for men and women in Sweden, for the period 1905–1970. The row names give the years of observation.

5 *lisbon*

A numeric vector containing annual maximum wind speeds, in kilometers per hour, from 1941 to 1970 at Lisbon, Portugal.

6 *oxford*

A numeric vector containing annual maximum temperatures, in degrees Fahrenheit, from 1901 to 1980 at Oxford, England.

7 *temps*100*m*

We have collected the fastest personal best times set between January 1991 and April 2017. We would like to know : how fast can we run ? In other words, we are interested in the ultimate world record. Most research concerning ultimate world records considers the development of the world record over time and extrapolates the trend to the future. That approach necessarily uses only a limited number of past world records and therefore gives rather unstable estimates.

Our approach, however, is based on as many as possible personal best times of top athletes. As a consequence the estimated ultimate world record tells us what could be achieved "tomorrow", not what could happen in 500 years from now. Also, we base our estimates on a data set which is more up-to-date and, in addition, excludes performances from before 1991 (when modern doping control was introduced), in order to avoid as much as possible doping related times. For this purpose we collect the fastest personal best times that are set in a certain period. So each athlete only appears once on our list. Including multiple times of one athlete would not be in line with the assumption of independent data. To exclude, as much as possible, doping related times, our observation period therefore starts on 01-01-1991. The observation period ends at 01-04-2017. Note that we only consider officially recognized times ; so times with a wind speed of more than 2m/s are not taken into account.

Times for the 100m are available in hundredths of seconds. Since many athletes have equal personal bests the data show ties. To avoid estimation problems the data are smoothed. When m athletes have a personal best of 9.85 then these m results are smoothed equally over the interval $(9.845, 9.855)$ as follows :

$$t_j = 9.845 + 0.01\frac{2j-1}{2m} \quad j = 1, \ldots, m$$

```
> time=temps100m[,1]
> Time=sort(time,decreasing=TRUE) # pour ordonner les données.
> plot(Time, ylab='Temps (secondes)', main="Records au 100M",type='l')

# As a last step all smoothed times are converted into speeds.
# Higher speeds correspond to "better" times.

> Speed=360/Time
```

For this dataset you have to use the GPD approach.

If you want you can update the data. For the period before 2001 the data were taken from the Swedish website : `http://hem.bredband.net/athletics/athleticsall-timebest.htm`. This website provides all best times up to 2001. For the time period 01-01-2001 to 01-04-2017 we use the official seasonal time lists of the IAAF website : `www.iaaf.org`.