HydRoStat: a quick start guide

1 Introduction

The *HydroPortailStats* R package provides functions and objects used to estimate a distribution (and related uncertainties) based on a sample of observed values.

To use *HydroPortailStats*, you should have R (https://www.r-project.org/) installed on your computer. You should also download and install the following packages:

1. evd (https://cran.r-project.org/web/packages/evd/index.html)
2. mvtnorm (https://cran.r-project.org/web/packages/mvtnorm/index.html)
3. numDeriv (https://cran.r-project.org/web/packages/numDeriv/index.html)

In practice, only the two functions described below are needed to get started. We recommend to take a look at the README at https://github.com/benRenard/HydroPortailStats to understand how these functions can be called.

```
h3 <- Hydro3_Estimation(y,dist,[Emeth],[Umeth],[options])
```

The function `Hydro3_Estimation` has 2 compulsory inputs and 3 optional inputs:

1. `y` [compulsory] : numerical vector of observations (without missing data)
2. `dist` [compulsory] : character string denoting the distribution to be estimated. See section Error: Reference source not found for the list of available distributions.
3. `Emeth` [optional] : estimation method. L-Moments (`"LMOM"`, default, recommended), moments (`"MOM"`), maximum likelihood (`"ML"`) or Bayesian (`"BAY"`).
4. `Umeth` [optional] : method for quantifying uncertainties. Parametric Bootstrap (`"PBOOT"`, default, recommended), bootstrap (`"BOOT"`), maximum likelihood (`"ML"`), Bayesian (`"BAY"`) or nothing (`"NONE"`).
5. `options` [optional] : a list containing estimation options. See section 4 for details.

The output of the function `Hydro3_Estimation` is a Hydro3 object, containing all results of the estimation procedure (parameter estimates, quantiles, uncertainties, statistical tests, etc.). See section 3 for details.

```
Hydro3_Plot(y,h3)
```

The function `Hydro3_Plot` has 2 compulsory inputs

1. `y` [compulsory] : numerical vector of observations (without missing data)
2. `h3` [compulsory] : Hydro3 object, resulting from the call of the `Hydro3_Estimation` function.

The function `Hydro3_Plot` has no output: it just produces a plot summarizing the inference (parameter estimates, data, cumulative distribution function (cdf) and quantiles).

2 Available distributions

Distribution	ID	# parameters	Typical usage*
Normal (or Gaussian)	`"Normal"`	2	QA
Log-normal	`"LogNormal"`	2	QA, QN
Gumbel	`"Gumbel"`	2	QX
Generalized extreme value	`"GEV"`	3	QX
Pearson III	`"PearsonIII"`	3	QX
Log-Pearson III	`"LogPearsonIII"`	3	QX
exponential	`"Exponential2"`	2	QS
Generalized Pareto	`"GPD3"`	3	QS
Gumbel for minima	`"Gumbel_min"`	2	QN
Generalized extreme value for minima	`"GEV_min"`	3	QN
Exponential with zero threshold	`"Exponential1"`	1	QS
Generalized Pareto with zero threshold	`"GPD2"`	2	QS
Poisson	`"Poisson"`	1	N

* QA = Annual discharge, QN = minimal discharge, QX = maximal discharge, QS = above-threshold discharge, N = count.

3 Description of a Hydro3 object

A Hydro3 object `h3` is a list containing all useful results of the estimation process. It contains the following fields (fields marked in red should be the most useful):

1. `h3$dist`: the estimated distribution.
2. `h3$empirical`: empirical estimates. A data frame with the following columns:
   a. `y`: sorted data
   b. `freq`: non-exceedance frequency
   c. `T`: return period
   d. `u`: reduced variate
3. `h3$pcdf`: estimated pdf and cdf. A data frame with the following columns:
   a. `x`: value
   b. `pdf`: associated pdf
   c. `cdf`: associated cdf
4. `h3$quantile`: estimated quantiles. A data frame with the following columns:
   a. `T`: return period
   b. `p`: non-exceedance probability
   c. `u`: reduced variate

    d. `q`: estimated quantile
    e. `IC.low`: lower bound of the uncertainty interval
    f. `IC.high`: higher bound of the uncertainty interval
5. `h3$par`: estimated parameters. A data frame with the following columns:
    a. `index`: parameter index
    b. `name`: parameter name
    c. `estimate`: estimated parameter
    d. `IC.low`: lower bound of the uncertainty interval
    e. `IC.high`: higher bound of the uncertainty interval
    f. `mean`: mean of the sampling distribution (see `h3$u`)
    g. `median`: median of the sampling distribution (see `h3$u`)
    h. `sdev`: standard deviation of the sampling distribution (see `h3$u`)
6. `h3$KS`: Result of the Kolmogorov-Smirnov goodness-of-fit test. A list with the following fields:
    a. `pval`: p-value of the test
    b. `stat`: test statistics
    c. `xtra`: not used
7. `h3$MK`: Result of the Mann-Kendall trend test. A list with the following fields:
    a. `pval`: p-value of the test
    b. `stat`: test statistics
    c. `xtra`: not used
8. `h3$Pettitt`: Result of the Pettitt step-change test. A list with the following fields:
    a. `pval`: p-value of the test
    b. `stat`: test statistics
    c. `xtra`: estimated location of the step-change
9. `h3$u`: Properties of the sampling distribution for parameters estimates, explored through simulations:
    a. `cov`: covariance matrix for parameters estimates
    b. `sim`: simulated parameter values, representing their sampling distribution
    c. `ok`: logical flag indicating whether the simulations went well
    d. `error`: integer error code (0 = no error)
    e. `message`: character string with a possible error message
10. `h3$ok`: logical flag indicating whether the estimation went well
11. `h3$error`: integer error code (0 = no error)
12. `h3$message`: character string with a possible error message

4 Estimation options

An option object `o` is a list containing all options used for estimation. In practice, the most useful ones are: (i) option `invertT`, that should be set to `TRUE` if large return periods correspond to small values of the variable (typical example: low flow variable “annual minimum”); (ii) option `splitZeros` that should be set to `TRUE` if values equal to zeros should be treated separately.

1. `o$FreqFormula`: formula used for computing non-exceedance frequencies (default: Hazen (i-0.5)/n
2. `o$pgrid`: grid of probabilities defining where estimated pdf and cdf are evaluated
3. `o$Tgrid`: grid of return periods defining where estimated quantiles are evaluated

4. `o$IClevel`: level of the confidence intervals (default: 0.9)
5. `o$p2T`: conversion factor between return period and non-exceedance probability, equal to the average number of data per year (default: 1)
6. `o$invertT`: `FALSE` if large return periods correspond to large values, `TRUE` otherwise (default: `FALSE`)
7. `o$splitZeros`: Should values smaller than or equal to zero be treated separately? (default: `FALSE`)
8. `o$lang`: language used in figure labels (default: French)
9. `o$nsim`: number of simulations used to explore the sampling distribution (default: 1000)