Hydrological statistics and extremes

Marc F.P. Bierkens
Professor of Hydrology
Faculty of Geosciences

Hydrological statistics

Mostly concerns with the statistical analysis of hydrological time series in relation to extremes, i.e. floods and droughts.
Example: Rhine at Lobith

Daily averaged discharge (m³/s) 1902-2002

Flow duration curve of Rhine discharge
Extreme events

Want to know:
What is the probability distribution that a flood of given size occurs?
What is the size of a flood that belongs to a given design frequency?

First question that must be answered is:

What constitutes a flood?

Two methods:
1. The largest discharge per year (Maximum values)
2. All discharge values above a certain threshold (Peak over Threshold (POT) data or partial duration data)

Maximum values of Rhine at Lobith

Maximum daily averaged discharge (m³/s) for each year in 1902-2002
Assumptions about maximum values

1. The maximum values are realisations of independent random variables.
2. There is no trend in time.
3. The maximum values are identically distributed.

In this case a single probability distribution can be assumed for the maximum values.

Probability and recurrence times

Cumulative probability distribution: \( F(y) = \Pr(Y \leq y) \)

Exceedence probability: \( P(y) = \Pr(Y \geq y) \)

Return period or recurrence time: \( T(y) = \frac{1}{P(y)} = \frac{1}{1 - F(y)} \)

Recurrence time: the average number years between two consecutive flood events of a given size

Note: the actual number of years between flood events of a given size is itself random.
Recurrence times from data

Analysis of maximum values of Rhine discharge at Lobith for recurrence time

<table>
<thead>
<tr>
<th>$T$</th>
<th>Rank</th>
<th>$F(y)$</th>
<th>$P(y)$</th>
<th>$T(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2790</td>
<td>1</td>
<td>0.00971</td>
<td>0.99029</td>
<td>1.0098</td>
</tr>
<tr>
<td>2800</td>
<td>2</td>
<td>0.01942</td>
<td>0.98058</td>
<td>1.0198</td>
</tr>
<tr>
<td>2905</td>
<td>3</td>
<td>0.02913</td>
<td>0.97087</td>
<td>1.0300</td>
</tr>
<tr>
<td>3061</td>
<td>4</td>
<td>0.03883</td>
<td>0.96117</td>
<td>1.0404</td>
</tr>
<tr>
<td>3220</td>
<td>5</td>
<td>0.04854</td>
<td>0.95146</td>
<td>1.0510</td>
</tr>
<tr>
<td>3444</td>
<td>6</td>
<td>0.05825</td>
<td>0.94175</td>
<td>1.0619</td>
</tr>
<tr>
<td>3459</td>
<td>7</td>
<td>0.06796</td>
<td>0.93204</td>
<td>1.0729</td>
</tr>
</tbody>
</table>

* ...*

9140 90 0.87379 0.12621 7.9231
9300 91 0.88350 0.11650 8.5833
9372 92 0.89320 0.10680 9.3636
9413 93 0.90291 0.09709 10.3000
9510 94 0.91262 0.08738 11.4444
9707 95 0.92233 0.07767 12.8750
9785 96 0.93204 0.06796 14.7143
9850 97 0.94175 0.05825 17.1667
10274 98 0.95146 0.04854 20.6000
11000 99 0.96117 0.03883 25.7500
11365 100 0.97087 0.02913 34.3333
11931 101 0.98058 0.01942 51.5000
12280 102 0.99029 0.00971 103.0000

Large recurrence times

From a series of $n$ maximum values: largest recurrence time to be assessed: $n+1$ years!

For design purposes: $y_T$ for large $T$ is necessary!

To this end:

1. Fit a probability distribution (which one?)
2. Use it to extrapolate to large values of $y_T$
   (maximum values $y$ with $F(y)$ close to 1).
The Gumbel distribution

It can be proven (see section 4.2.2. Syllabus) that maximum values of the following distributions follow a Gumbel distribution:

- Exponential
- Gaussian
- logGaussian
- Gamma
- Logistic
- Gumbel

\[ f_y(y) = be^{-b(y-a)} \exp(-e^{-b(y-a)}) \]
\[ F_y(z) = \exp(-e^{-bt(y-a)}) \]

Fitting the Gumbel distribution

Types of methods:

- graphical using Gumbel paper and linear regression
- method of moments
- maximum likelihood estimation
Taking the double logarithm of the cpdf:

\[
\ln[F_Y(y)] = \ln[(\exp(-e^{-b(y-a)})] = -e^{-b(y-a)}
\]

\[
\ln\{-\ln[F_Y(y)]\} = -b(y - a)
\]

\[
y_T = a - \frac{1}{b} \ln\{-\ln[F_Y(y)]\}
\]

\[
y_T = a - \frac{1}{b} \ln\{-\ln[\frac{T(y)-1}{T(y)}]\}
\]

Plot maximum \(y_i\) vs \(-\ln\{-\ln\left[\frac{i}{n+1}\right]\}\)

Plot maximum \(y_i\) vs \(T(y_i)\) on special Gumbel (double logarithmic) paper

1. Plot maximum \(y_i\) vs \(-\ln\{-\ln\left[\frac{i}{n+1}\right]\}\)

or, plot maximum \(y_i\) vs \(T(y_i)\) on special Gumbel (double logarithmic) paper

2. Fit a straight line by eye or by regression to determine \(a\) and \(b\).
Method of moments

Mean and variance of a Gumbel variate $Y$:

$$\mu_Y = a + \frac{0.5772}{b} \quad \sigma_Y^2 = \frac{\pi^2}{6b^2}$$

1) Estimate mean $m_Y$ and variance $s_Y^2$
2) Equate these to the above expressions
3) Solve for $a$ and $b$:

$$\hat{a} = m_Y - \frac{0.5772}{b}$$

$$\frac{1}{b} = \sqrt{\frac{6s_Y^2}{\pi^2}}$$
Method of moments

Rhine data set

\[ \hat{g} = 5621, \hat{b} = 0.00061674 \]

Estimation of the \( T \)-year event

\[ \hat{y}_T = \hat{a} - \frac{1}{\hat{b}} \ln(-\ln\left(\frac{T - 1}{T}\right)) \]

In case of the Rhine data set:

MoM parameters:

\[ y_{1250} = 5621 - 1621 \cdot \ln(-\ln(0.9992)) = 17182 \text{ m}^3/\text{s}. \]

with parameters from regression: 17736 m\(^3\)/s
Estimation of confidence limits

In case of regression:

\[ \hat{y}_T - t_{0.05}s_{\hat{y}}(T) \leq y_T \leq \hat{y}_T + t_{0.05}s_{\hat{y}}(T) \]

\( t_{0.05} \) : the 95-point of the student’s t-distribution

and the standard error of prediction \( s_{\hat{y}}(T) \) estimated as:

\[ s_{\hat{y}}^2(T) = \left( 1 + \frac{1}{N} \sum_{i=1}^{N} (x(T_i) - \bar{x})^2 \right) \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \]

with \( x(T_i) = -\ln[-\ln((T_i - 1)/T_i)] \)

\[ \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x(T_i) \]

Estimation of confidence limits

In case of method of moments:

\[ \text{Vár}(\hat{y}_T) \approx \frac{(1.11 + 0.52x + 0.61x^2)}{\hat{b}^2 N} \]

with \( x(T) = -\ln[-\ln((T - 1)/T)] \)

Assuming a Gaussian estimation error:

\[ \hat{y}_T - 1.96\sqrt{\text{Vár}(y_T)} \leq y_T \leq \hat{y}_T + 1.96\sqrt{\text{Vár}(y_T)} \]
The number $T$-year events in a given period

The *expected* number of $T$-year events in $N$ years:

$$Np = N / T$$

The *actual* number $n$ of $T$-year events in $N$ years is a random variable obeying a binomial distribution:

$$\Pr(n \text{ events } y_T \text{ in } N \text{ years}) = \binom{N}{n} p^n (1 - p)^{N-n}$$

Examples:  
$$\Pr(1 \text{ events } y_{100} \text{ in 10 years}) = \binom{10}{1} 0.01 \cdot (1 - 0.01)^9 = 0.0914$$

$$\Pr(\text{one or more flood events occur in 10 years}) = 1 - \Pr(\text{no events}) = 1 - 0.99^{10} = 0.0956.$$ 

The time until the next $T$-year event

The *expected* number of years $n$ until the next $T$-year event: $T$

The *actual* number of years $n$ until the next $T$-year event is a random variable obeying a geometric distribution (with $p=1/T$):

$$\Pr(m \text{ years until event } y_T) = (1 - p)^{m-1} p$$
Other extreme value distributions

Generalized extreme value distributions

- Type I (Gumbel)
- Type II
- Type III (Weibull)

Other extreme value distributions

Other maximum value distributions used for maximum values: log-normal, log-Pearson-type III.
Minimum values (e.g. low flows)

- Take $-Z$ or $1/Z$
- or use Weibull distribution on $Z_{\text{min}}$

Testing the assumptions

Independence: Von Neuman’s $Q$

$$Q = \frac{\sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Lower critical area.

If larger than critical value: no evidence that data are dependent!

Rhine data set: $Q=2.071$; upper critical value at $\alpha=0.05$: 1.618 -> no evidence for dependence
Testing the assumptions

Trends: Mann-Kendall test

\[ T = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \text{sgn}(Y_i - Y_j) \]

\[ T' = 18T / [n(n-1)(2n+5)] \]

For \( n > 40 \) \( T' \) has standard Gaussian distribution with two sided critical area (i.e. significant at 95% (\( \alpha = 0.05 \)) accuracy trend is \( T' < -1.96 \) or \( T' > 1.96 \)); otherwise no evidence of a trend.

Rhine data set: \( T' = 1.992 \rightarrow \) significant trend at 95% accuracy.

Yearly maxima of average daily runoff of the Rhine at Lobith

\[ y = 12.682x - 18205 \]
\[ R^2 = 0.0326 \]
Testing the assumptions

Testing for some distribution: $\chi^2$ test

1. Define $m$ classes (as in a histogram) and assign the $n$ data values
2. Count the number of data falling in each each class $i$: $n_i$
3. Fit the proposed distribution function $F(Y)$ to the data
4. Calculate the expected number of data falling into each class $i$ as:
   \[ n^*_i = n(F_Y(y_{up}) - F_Y(y_{low})) \]
5. The following test statistic is calculated:
   \[ X^2 = \sum_{i=1}^{m} \frac{(n_i - n^*_i)^2}{n^*_i} \]

Testing the assumptions

Testing for some distribution: $\chi^2$ test

$X^2$ follows a chi-squared distribution with $m-1$ degrees of freedom: $\chi^2_{m-1}$

There is an upper critical area for the 0-hypothesis that the data follow the proposed distribution.

Rhine data and proposed Gumbel distribution and 20 classes: $X^2 = 23.70$
Rhine data and proposed lognormal distribution and 20 classes: $X^2 = 14.36$

Lower boundary critical area for $m-1 = 19$ degrees of freedom and $\alpha=0.05$: 30.144 -> both distributions cannot be discarded!