Statistical Hydrology

HYDROEUROPE

Copyright © Warsaw University of Technology

Motivation

Deterministic hydrology

- Based on well-known physical principles
- Ignored uncertainty of input data
- Results are highly dependent on the boundary conditions
- Simulated scenarios (boundary conditions) are based on past records or expert choice

Statistical Hydrology

- Repeatability of events as a source of information
- Extrapolation of past records as forecast for the future
- Return period expressed in years as a link between hydrology and economy/social aspects

Basic definitions

- Random experiment – the process of observing events having uncertain outcome e.g. dice roll, discharge observation

- Elementary event – event, which contains only a single outcome in the sample space e.g. getting one on a dice, observing discharge 100m3/s

- Random event – any combination of outcomes of an experiment e.g. getting odd result in a dice throw, observing discharge greater then given threshold

Probability

A measure of how likely an event will occur:

- **Classic** probability – ratio of the number of favorable per total number of possible event outcomes

e.g. throwing odd number on a dice:

$$P(A) = \frac{3}{6}$$

- **Empirical** probability – ratio based on historical record

e.g. in past 10 year, anual maximum discharge>100m3/s has been observed twice.

$$\hat{P}(Q > 100) = \frac{2}{10}$$

- **Theoretical** probability – a function which satisfies three axioms:
  1. Probability is always finite and non-negative $P(A) \geq 0$
  2. Probability of **any** elementary event $P(\Omega) = 1$
  3. For **mutually exclusive** events: $P(A \cup B) = P(A) + P(B)$

Random Variable

- A function, which assigns a number to a random event.

- Examples of **discrete** random variables:
  - dice roll result
  - number of siblings

- Examples of **continuous** random variables:
  - Annual maximum flow at Thames in London
  - Monthly sum of precipitation in Berlin
  - Any other uncertain measurement outcome

Probability distribution

- A function, which assigns probability to the value of the outcome of experiment (**empirical distribution)**
  - Empirical distribution is build using empirical probability concept and finite sample of experiment results

- A function, which assigns probability to the all values of Random Variable (**theoretical distribution)**
  - Conclusions from theoretical distributions are valid for the whole **general population**.

Probability Density Function - PDF

- A function, which specifies the probability of the random variable falling within a particular range of values.

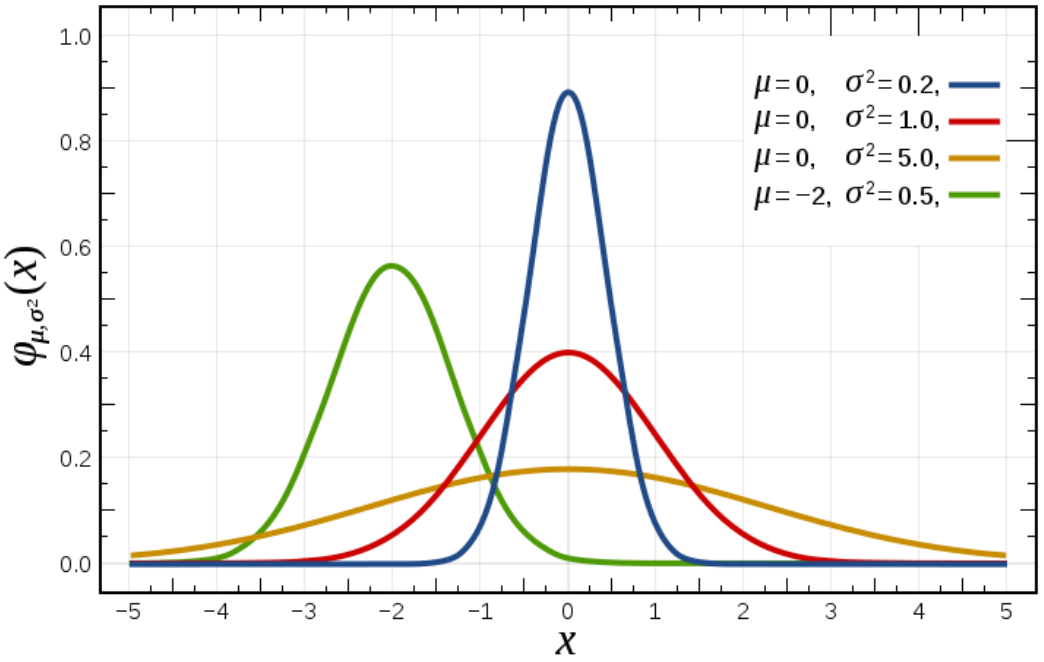
Formally: $$P(a \leq X \leq b) = \int_a^b f_X(X)\,dx$$

PDF describes analytically the way probability is distributed within the possible values of random variable.

Normal Distribution

Very common, continuous and symmetric distribution, with two parameters: μ-mean value, and σ^2 -variance

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

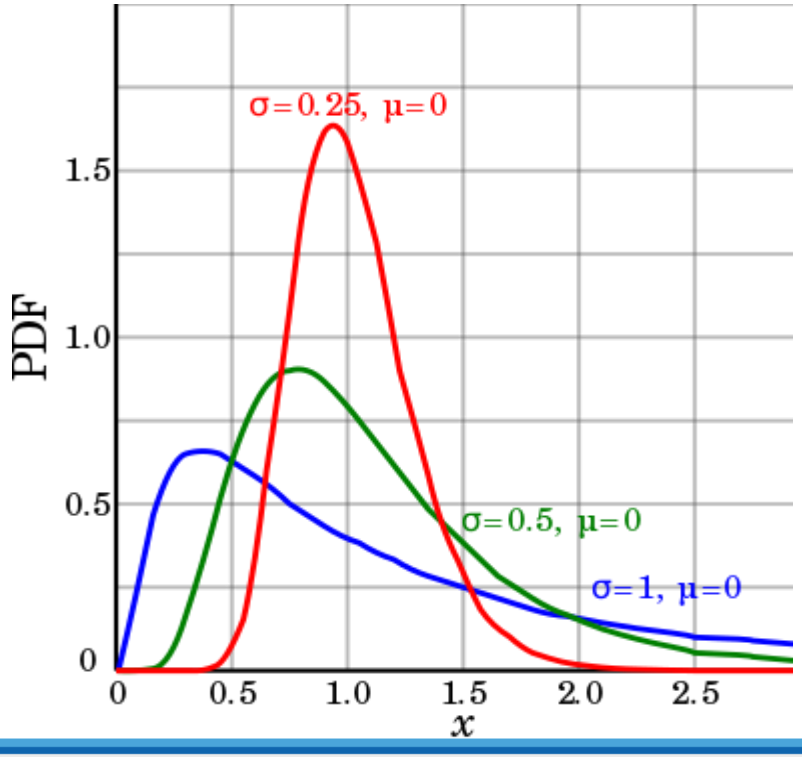


Asymmetric distributions (1)

Random variables (measurement outcomes) in hydrology are usually bounded from the left side at zero (negative flow is not possible). Therefore a special class of distributions – asymmetric distributions applies.

e.g. raindrop size follows log-normal distribution

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$



Asymmetric distributions (2)

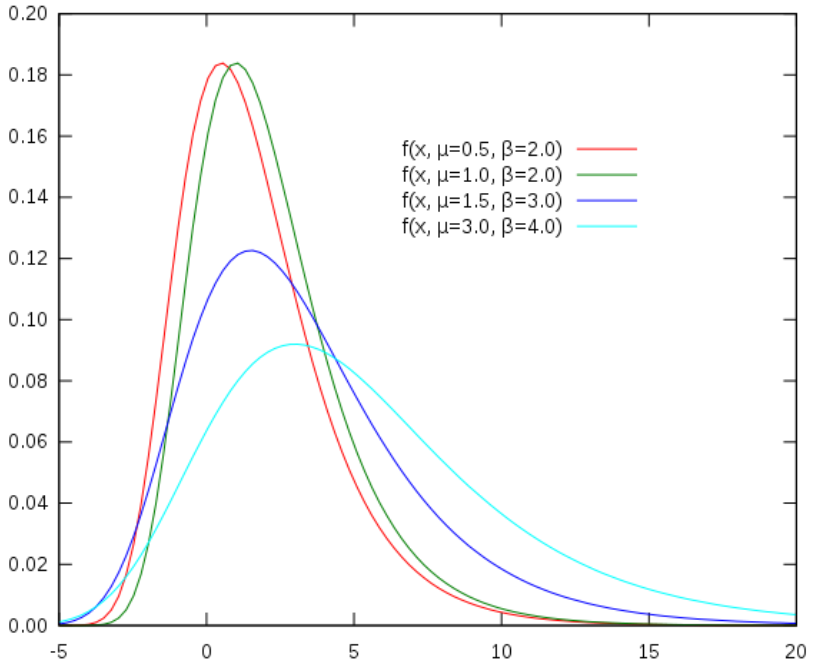
Gumbel distribution (also known as Generalized Extreme Value distribution Type-I) is used to model the distribution of the maximum (or the minimum) of a number of samples of various distributions, assuming that the number of samples is *large.*

$$f_X(x) = \frac{1}{\beta} e^{-(z+e^{-z})}$$

Where:

$$z = \frac{x - \mu}{\beta}$$

Gumbel Distribution PDF



Probability distribution – Cumulative Distribution Function (CDF)

A function F(x), that answer the question:

What is the probability, that the value of random variable X will be lower than the given argument x?

$$F(x) = P(X < x)$$

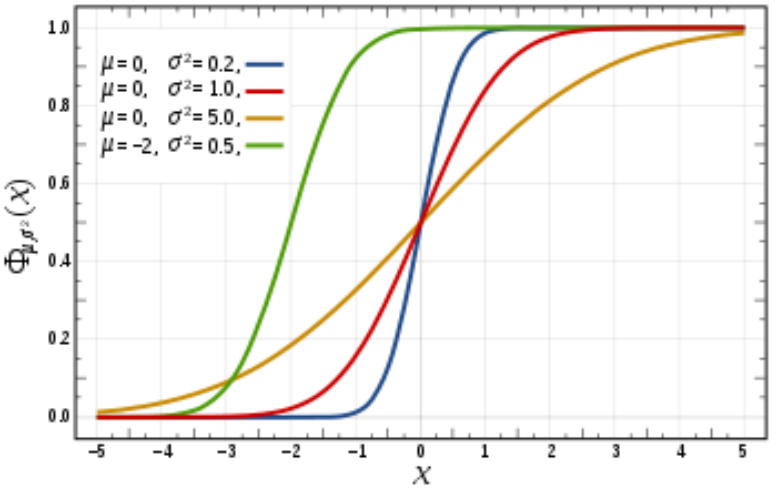
General properties of CDF:

$$F(x) = \int_{-\infty}^{x} f(\tau)\, d\tau$$

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$

Normal Distribution CDF



Probability distribution – excedence function

- A function $p_{ex}(x)$ that answers the question:

What is the probability, that the value of random variable X is greater than the argument x?

- Important example: The flow value $Q_{0.01}$ is a flow, which has the exceedance probability equal to 0.01;

$p_{ex}(Q_{0.01}) = 0.01$

A sum of CDF and exceedance function of the same argument is always equal to one:

Quantile

- Quantiles are values of random variable, which cut the probability distribution into subgroups of given size.

e.g. median is a quantile of rank 0.5, since it cuts the distribution into halves

In Flood Frequency Analysis quintile refers to **discharge** with a given probability of **exceedance**

e.g. $Q_{0.5}$ is the discharge with 0.5 probability of exceedance (biannual flood)

$Q_{0.01}$ is the discharge with 0.01 probability of exceedance (hundred years flood)

Return period

- In hydrology, together with probability of exceedance, we use the return period, assuming, that a one year returns one realisation of an experiment

Probability of excedence	Return period	Symbol
0.5	2 years	$Q_{0.5}$
0.1	10 years	$Q_{0.1}$
0.01	100 years	$Q_{0.01}$
0.002	500 years	$Q_{0.002}$

Flood Frequency Analysis (FFA)

- Magnitude of extreme flows is related to their frequency of occurrence:

$$Magnitude \propto \frac{1}{frequency\_of\_occurence}$$

- The **objective of FFA** is to relate the magnitude of events to their frequency of occurrence, through probability distribution
- It is assumed, that the events are independent and come from the same probability distribution.

Flood Frequency Analysis

- In order to estimate quantiles we need to know the theoretical distribution (type and parameters)
- We need to perform **sampling** from general population i.e. collect past observations of flood discharges
- Based on historical sample an empirical distribution can be build
- From the shape of empirical distribution we can predict theoretical distribution, assuming that the sample is representative for general popultaion

Types of sampling

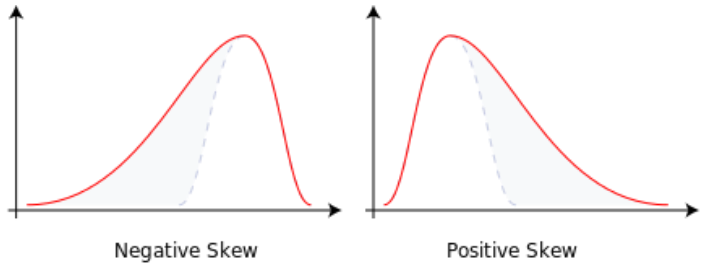
- Random sampling – equal likehood of selection of each member of population
  - pick any streamflow value from a population
- Stratified sampling – population divided into groups, and then random sampling applies
  - pick a flow value from annual maximum series
- Uniform sampling – data are selected uniformly in time or space
  - pick a streamflow value at 6am
- Convenience sampling – data collected according to the convenience of experimenter
  - pick a streamflow in the summer

Summary statistics

Statistics, which are used to summarize a set of observations (communicate the largest amount of information as simply as possible).

- Mean value (1st raw moment) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Variance (2nd central moment) $s^2 = \frac{1}{n-1}\sum (x - \bar{x})^2$
- Skewness coefficient(3 rd moment)

$$b = \frac{\frac{1}{n}\sum (x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum (x_i - \bar{x})^2\right]^{\frac{3}{2}}}$$

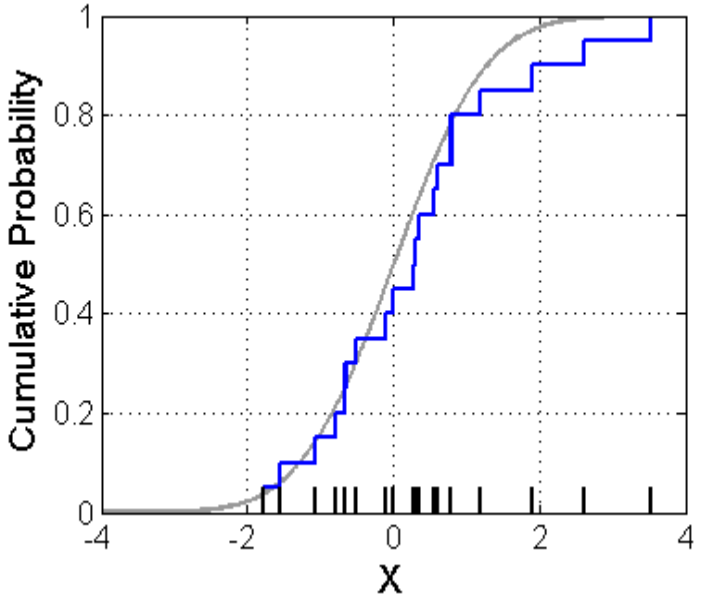


Empirical CDF

In order to build an empirical CDF, two steps are needed:

- Sort the sample in descending order
- For each value assign the empirical probability of exceedance:

$$p(m, N) = \frac{m}{N + 1}$$



where: *m* is the rank in the ordered sample

N is the sample size

Methods of distribution fitting (1)

- The conclusions about flows with given return periods a drawn from a distribution, which describes general population.
- However we only have a finite sample set. Thus, we need to fit theoretical distribution to available sampling outcomes.
- By *fitting* we understand choosing a proper theoretical distribution and estimating its parameters.
- Typical distribution fitting methods, used in FFA are:
  - Method of moments (MOM)
  - Graphical fitting
  - Maximum likehood estimation (MLE)

Methods of distribution fitting (2)

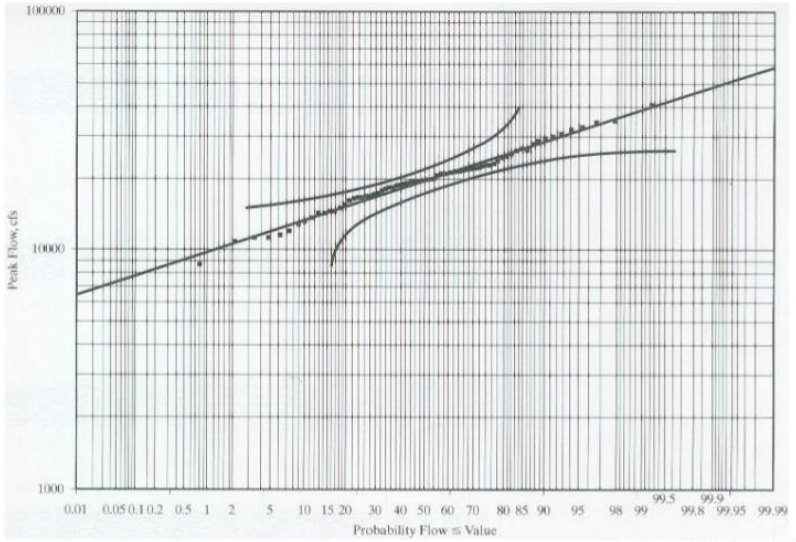
- **Method of moments** assumes, that moments of theoretical distribution are equal to empirical moments of sample set
- If a distribution has two parameters, two moments (usually mean and variance) are enough to estimate them.

- Pros: fast, simple, intuitive
- Cons: sensitive to outliers, especialy when sample size is small

Methods of distribution fitting (3)

- Graphical fitting- fit the distribution.

by plotting it on a probability paper together with empirical CDF



(Bedient & Huber 2002)

- Pros: old method, no computer needed, robust to outliers
- Cons: partially subjective, highly relies on user experience

Methods of distribution fitting (4)

- Maximum likehood estimation defines likehood function as product of probabilities of sampling outcomes. Then, partial derivatives are calculated, so that the parameter set provides the maximum likehood for given sample set.

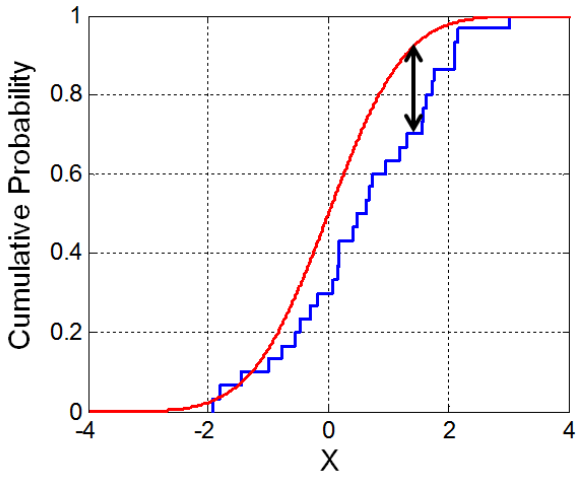
- Pros: consistent for large sample size
- Cons: computationally intensive, complex implementation

How good is the fit of the distribution?

- Graphic check: visual check of the plotted graph (How accurate does the CDF/PDF reproduce empirical observations?)
- Statistical tests:
  - E.g. Kolmogorov-Smirnov test, which checks if the sample comes from the theoretical distribution which has just been fitted, by quantifying the distance:

$$D_n = \frac{1}{n} \max \left| F_X(x_i) - S_n(x_i) \right|$$

The value of Dn is compared with tabulated critical Values for given significance level (in this case it's The allowed probability of fitting wrong distribution)



Confidence limits

- From the variability of the sample we can not only predict the value of discharge with requested return period, but also quantify the uncertainty of this prediction
- This estimation is important in cases when the designer is capable of additional investment, just to be on a *safe side*.

$$CL(Q_T) = Q_T \pm t_{1-\alpha/2} SE_q$$

Where:

Q_T is the discharge with return period T

$t_{1-\alpha/2}$ is the value of t-Student's distribution for $1 - \alpha/2$ confidence level

SE_q is the standard error of sample with size *n* and standard deviation σ

$SE_q = \frac{\sigma}{\sqrt{n}}$

General steps of FFA

1. Collect historical data (usually annual maximum discharge)
2. Check for possible outliers
3. Sort the sample in non-descending order
4. Assign empirical probability of exceedance and draw empirical CDF
5. Estimate the parameters of theoretical distribution
6. Check if the goodness of fit is satisfactory
7. Calculate the sought quantiles
8. Estimate their confidence limits

Assumptions of frequency analysis

- All data points are correct and precisely measured (any measurement error may bias the results)

- All points represent independent events (beware of floods, which start in December and end in January)
- Random sample – every value in the population has equal chance of being included in the sample
- All extreme flows originate from the same statistical distribution (sample homogeneity)
- Hydrologic regime has remained constant during sampling (climate change is usually neglected)

Sources of uncertainty in FFA

- Measurement uncertainty (for high flows, extrapolation of rating curve is usually used)
- Poor representativeness of short sample (i.e. It may include a series of relatively „dry years” by conincidence)
- Changes in runoff conditions (i.e. recent urbanization was not „included” in the long data sample)
- Subjective choice of distribution

Take home message

- Flood Frequency Analysis is an example branch of Statistical Hydrology
- It provides you with estimations, which are not possible to be neither observed directly nor modelled deterministically
- If the input data are biased by large uncertainty, also the results will be biased.
- There is no universal approach, usually local (national) standard defines details of FFA procedurę (allowed distribution type, fitting methods, minimum sample size, etc.)