DNV

WHEN TRUST MATTERS

Statistical Modelling of environmental conditions

Lecture at HIPERWIND PhD Summer School
DTU, Denmark

Erik Vanem

29 August 2023

HIPERWIND

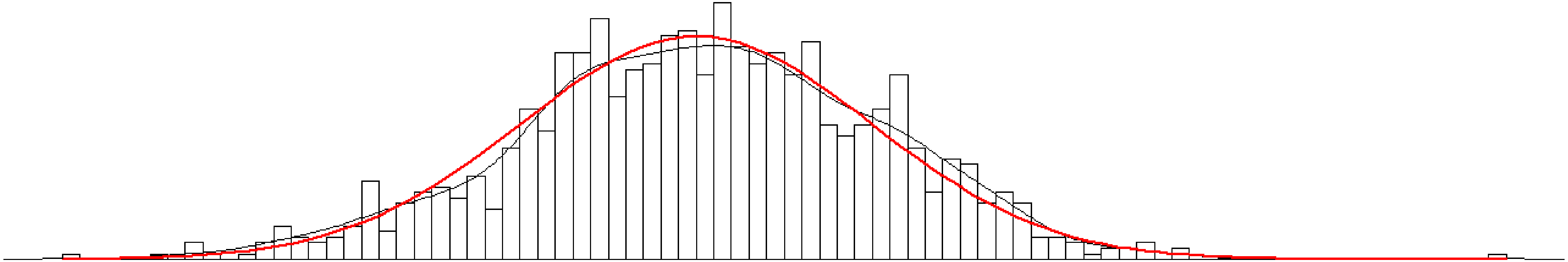
Lecture 1: Probability distributions and statistical modelling

Learning goals

- Understand basic concepts and statistical models
  - Descriptive statistics and statistical inference
  - Parametric and non-parametric models

- Be able to fit a statistical model to data
  - Choosing the right model
  - Fitting the model
  - Evaluating the model

- Understand multivariate models and be able to fit statistical models to multivariate data



Univariate analysis

Basic concepts

- Population and sample
  - A **population** is the set of “objects” of interest; e.g. all sea states at a given location
  - A **sample** is a subset of the population of which we have information
- A **variable** is any characteristic whose value may change from one object to another in the population; e.g. Significant wave height, mean wave period, …
- A **random variable** is a variable whose possible values are random and a result of some underlying random process
- **Data** typically consist of observations of one or more variables

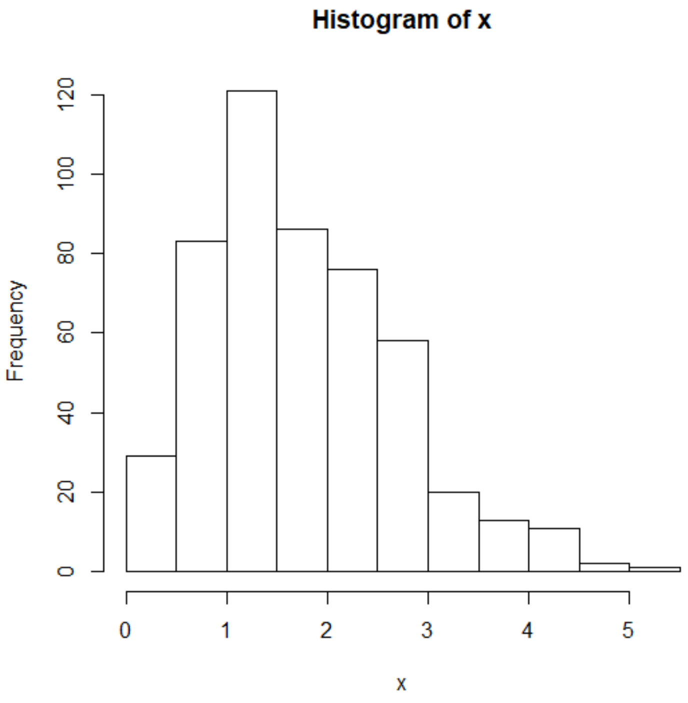
Descriptive and inferential statistics

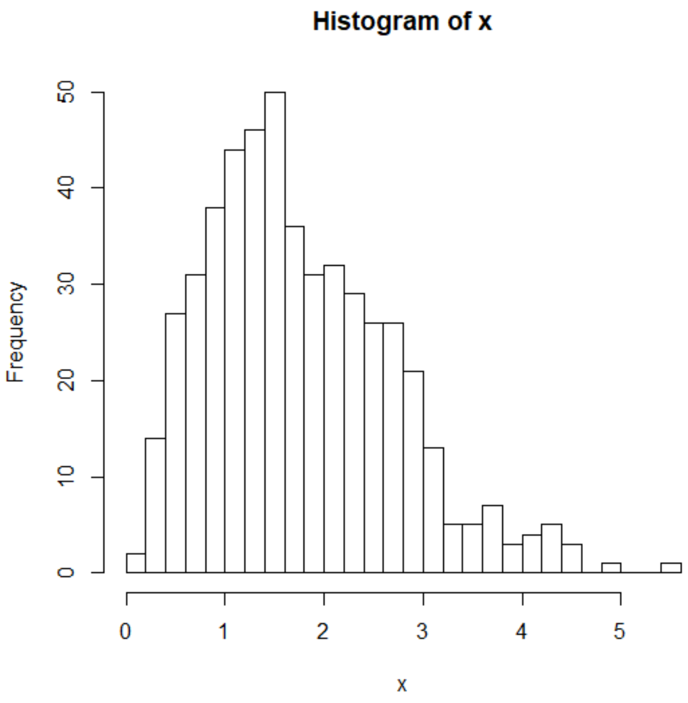
- **Descriptive statistics**: summarize and describe important features of the *data*
  - Histograms and scatterplots, summary statistics

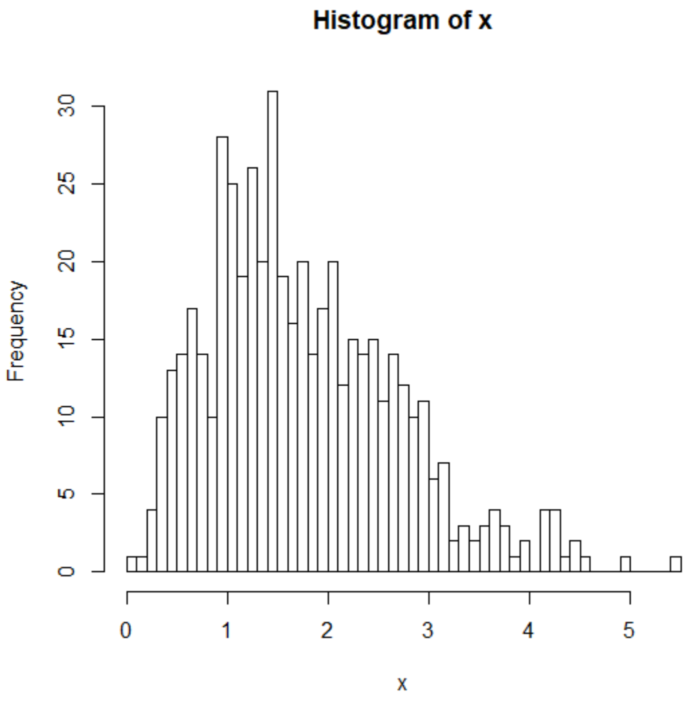
- **Statistical inference**: Use the sample to draw conclusions about the *population*, given some assumptions
  - Point estimation, hypothesis testing, confidence intervals
  - Need assumptions of the ***underlying statistical model***; If assumptions are wrong, the inference might be wrong and misleading
  - In structural offshore design one is typically interested in inference about ***extreme conditions***
    - 10-year, 50-year, 100-year extremes

Descriptive statistics - Histograms

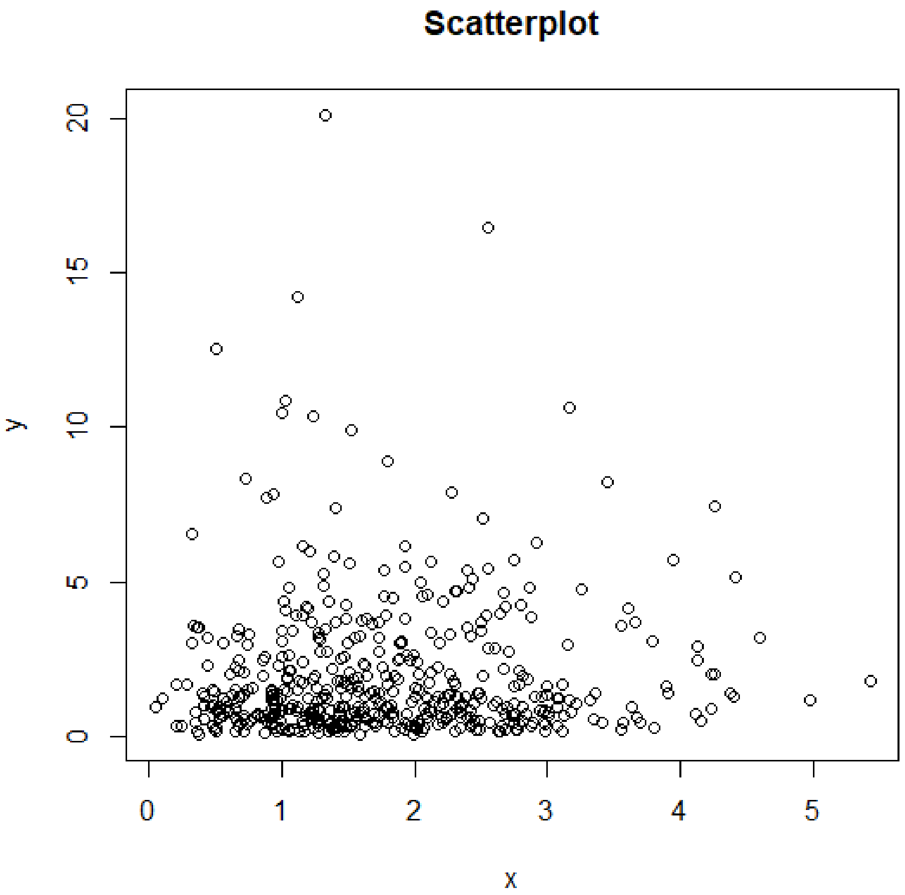
- Displays the frequency or relative frequency of observations in different class intervals from a sample

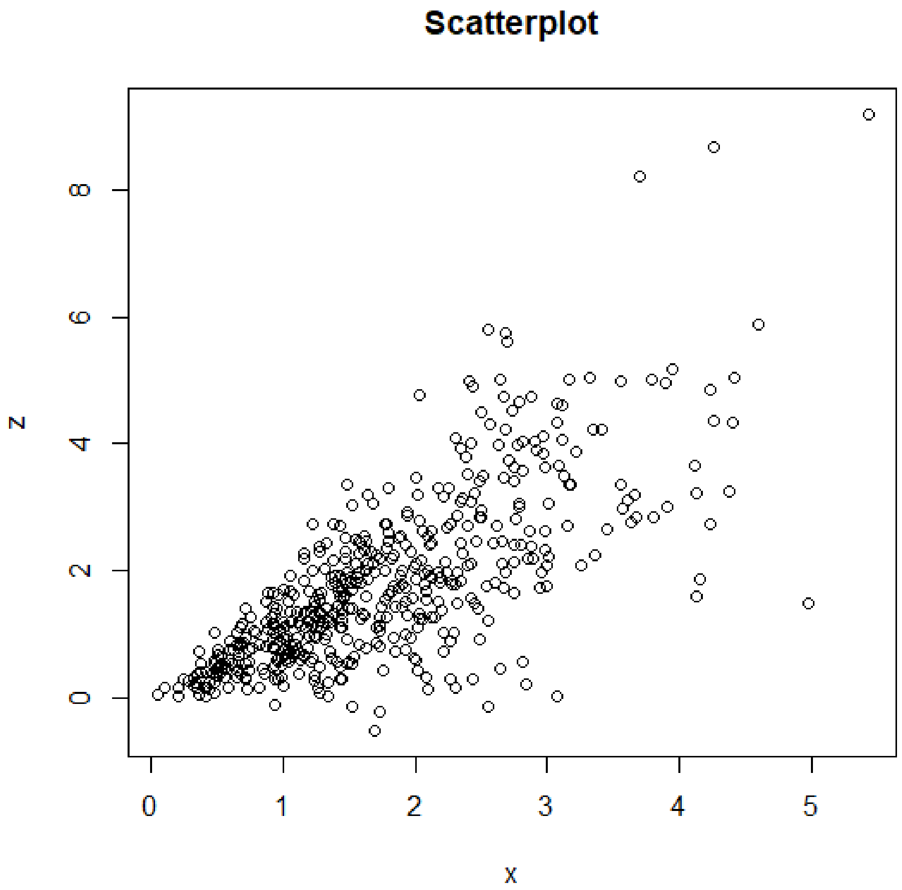




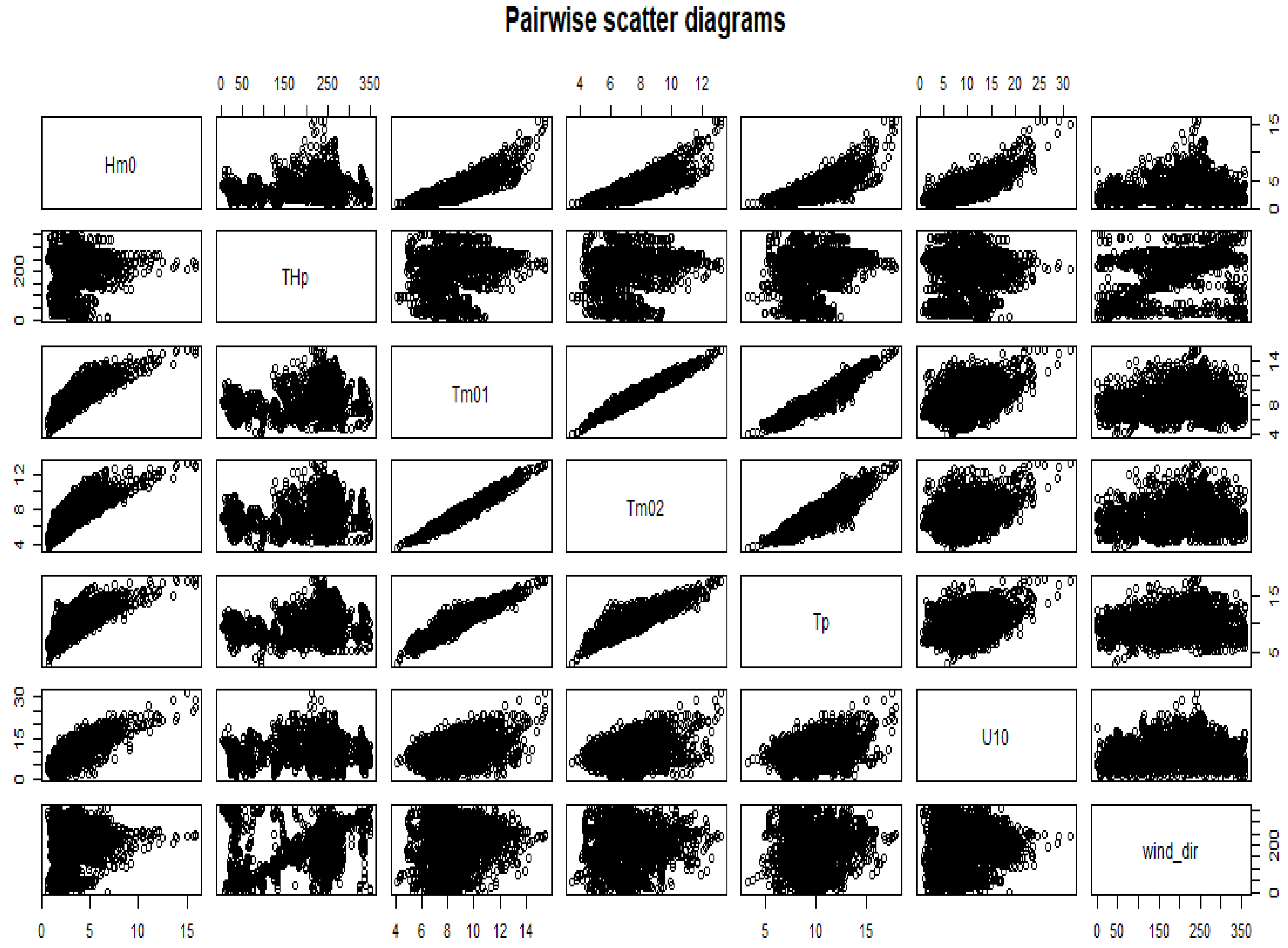


Descriptive statistics - scatterplots



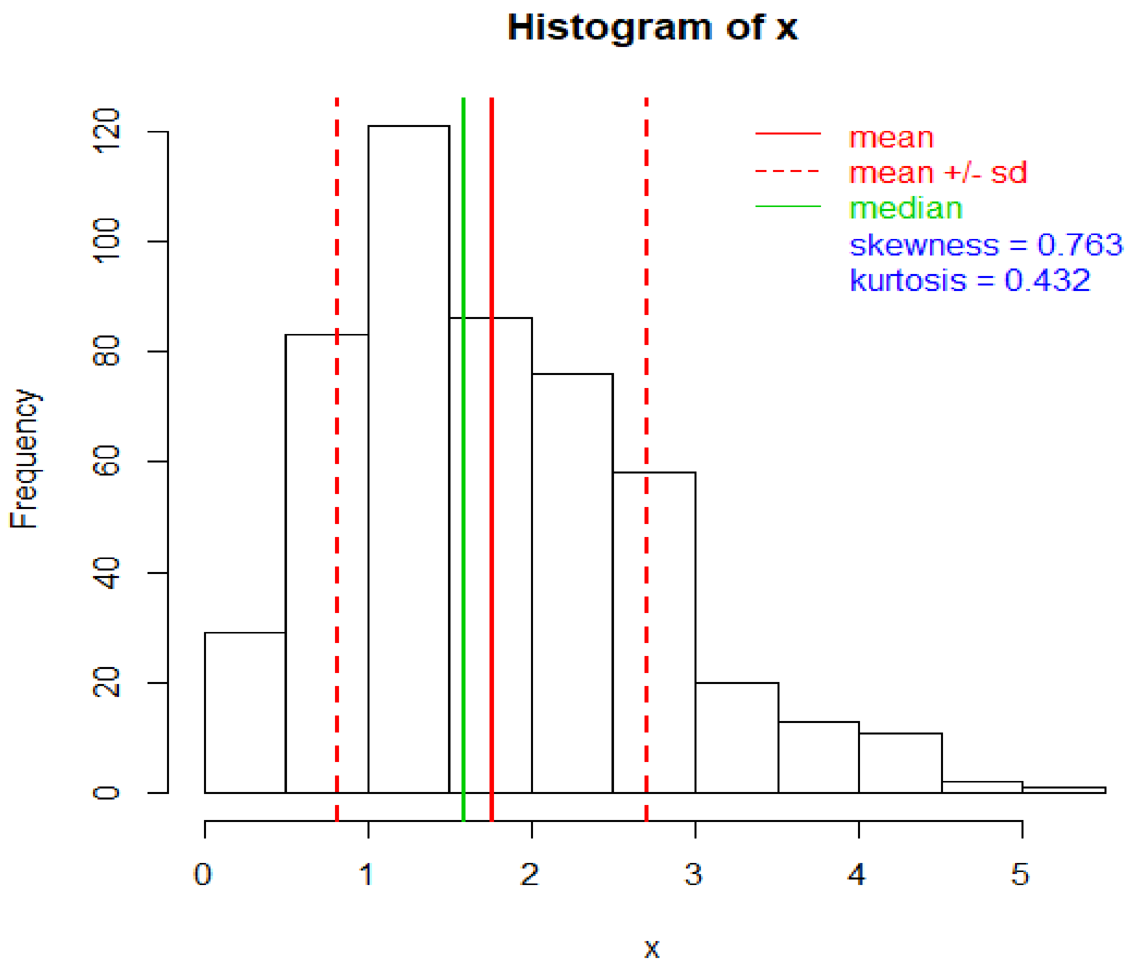


Scatterplots for data in > 2 dimensions

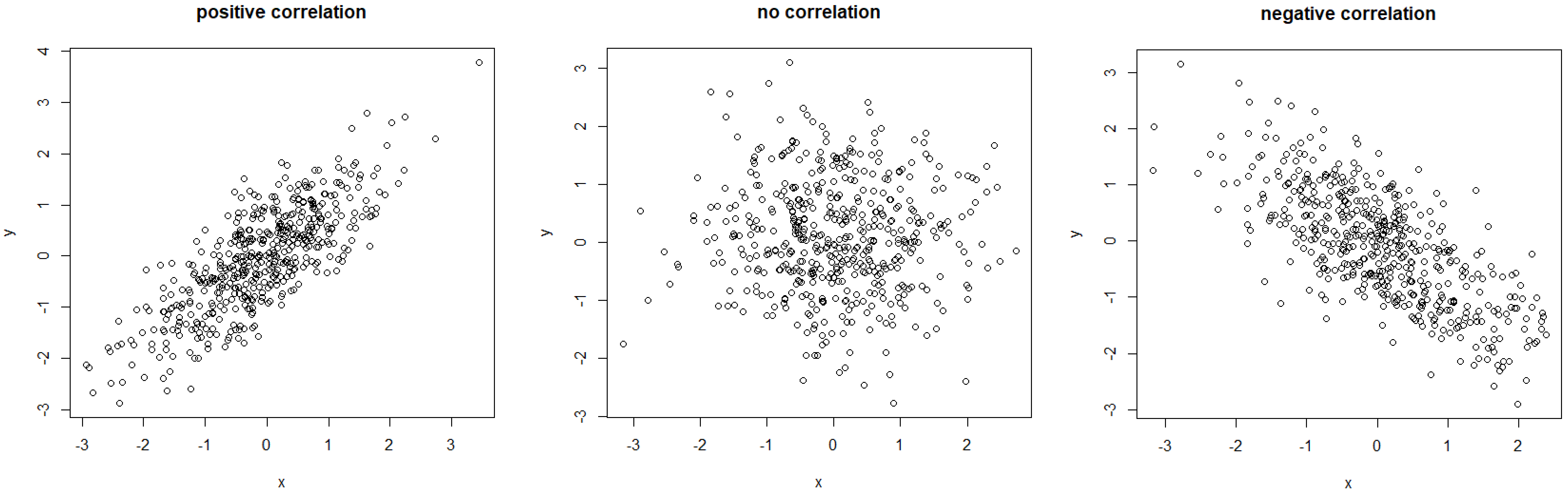


Descriptive statistics – summary statistics

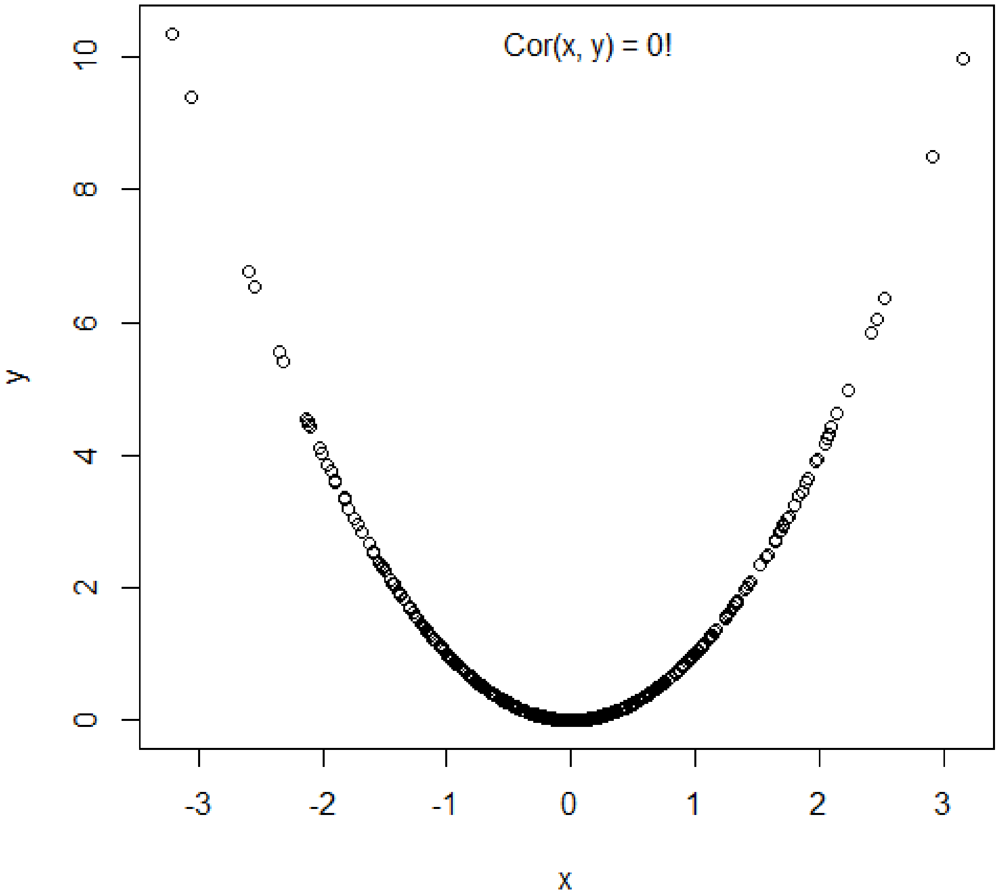
- ***Measures of location***
  - The mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - The median
  - Quantiles (quartiles, percentiles, …)
- ***Measures of variability***
  - Variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Standard deviation $s = \sqrt{s^2}$
- Skewness – ***measure of asymmetry***
- Kurtosis – ***measure of heavy-tailedness***
- Covariance and correlation
  - Linear correlation coefficient
  - Rank correlation



Covariance and correlations



Example of perfect dependence but correlation = 0 ($Y = X^2$)



Note: Independence implies correlation = 0, but correlation = 0 does not imply independence!

Probability distributions - PDF

Probability density function (pdf):

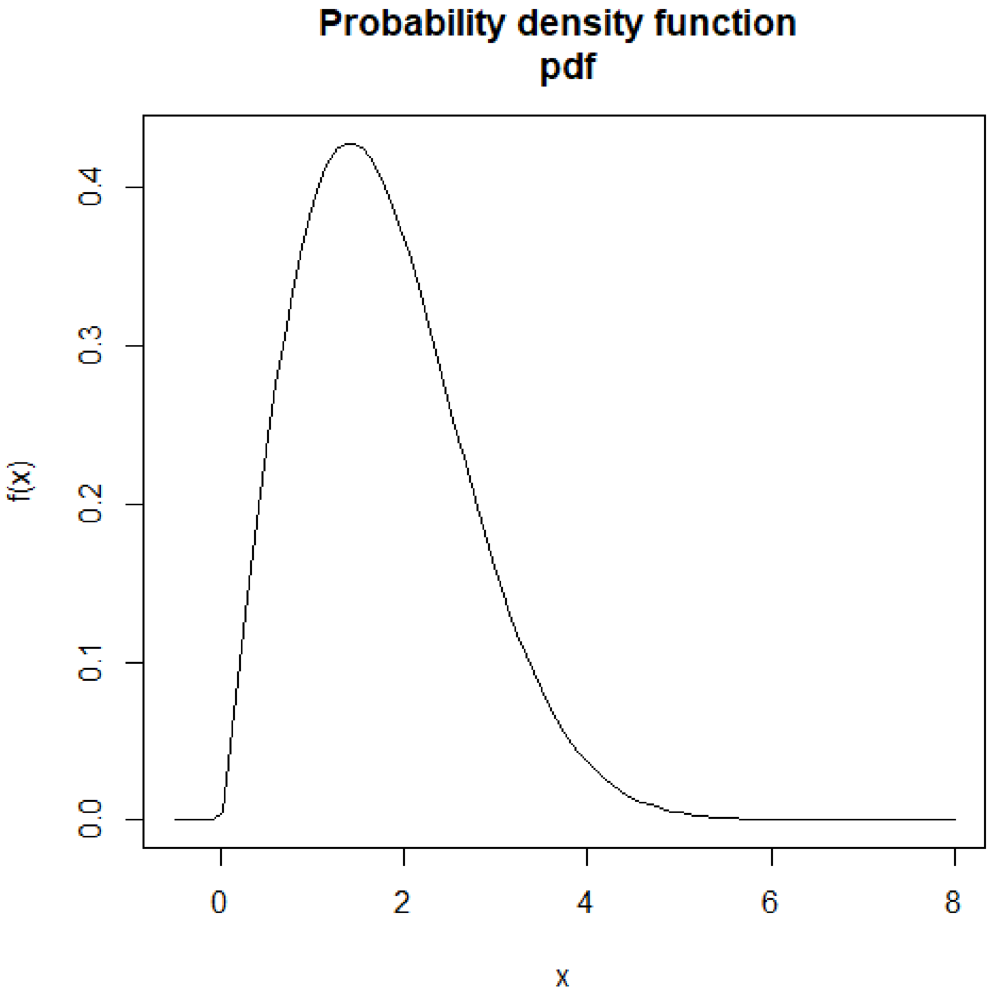
$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Must satisfy two conditions

1. $f(x) \geq 0 \; \forall \; x$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$ (area under the entire graph)

Interpretation:

$$f(x) = \frac{P(x < X < x + \Delta x)}{\Delta x}$$



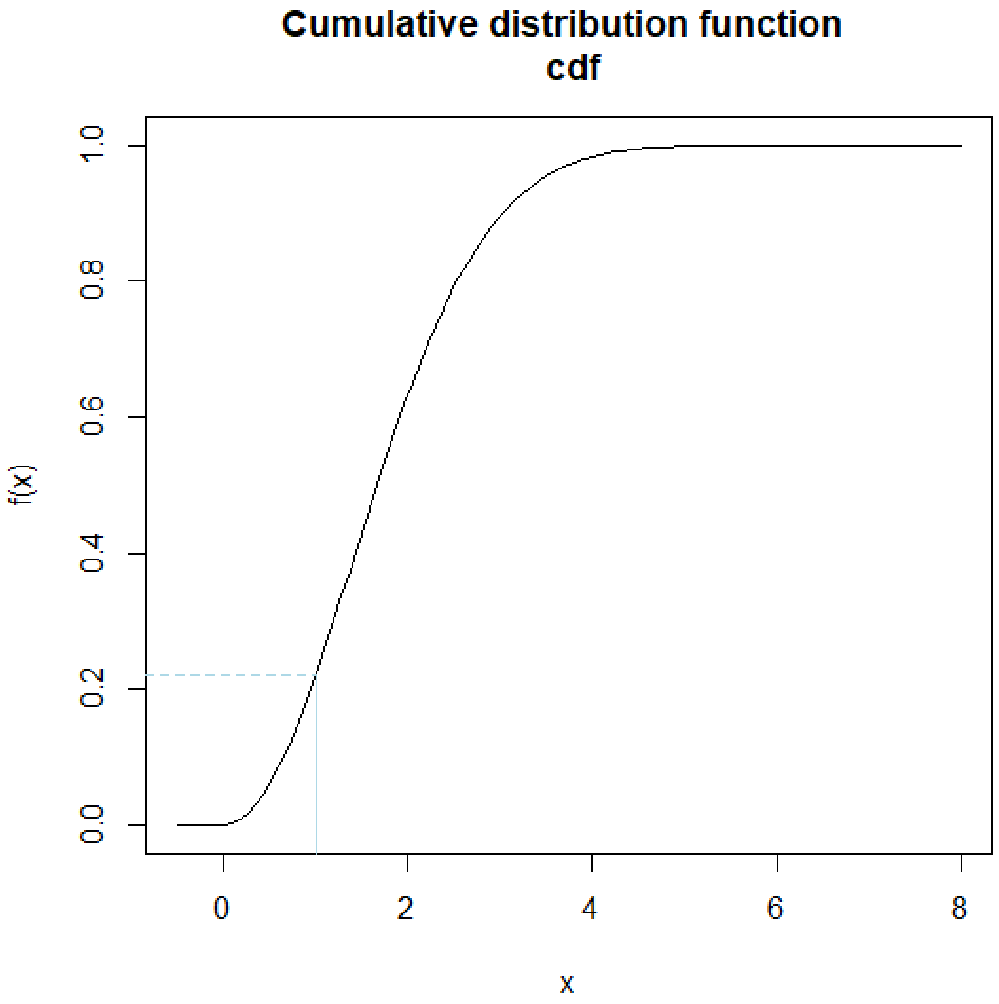
Cumulative distribution function (CDF)

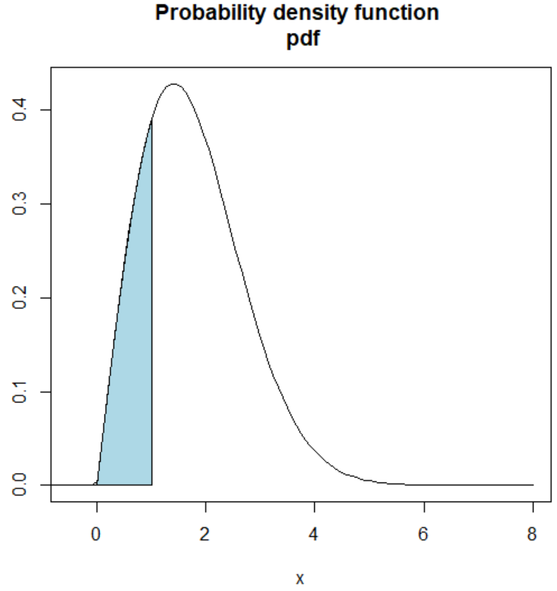
Probability of X being at most x:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)dy$$

Some properties:

- $F(\infty) = 1 = P(X \le \infty)$
- $P(X > a) = 1 - F(a)$
- $P(a \le X \le b) = F(b) - F(a)$
- $F'(x) = f(x)$ (if it exist)
- $p = F[\eta(p)], \quad \eta(p) = (100p)\text{th percentile:}$





Parametric probability distributions

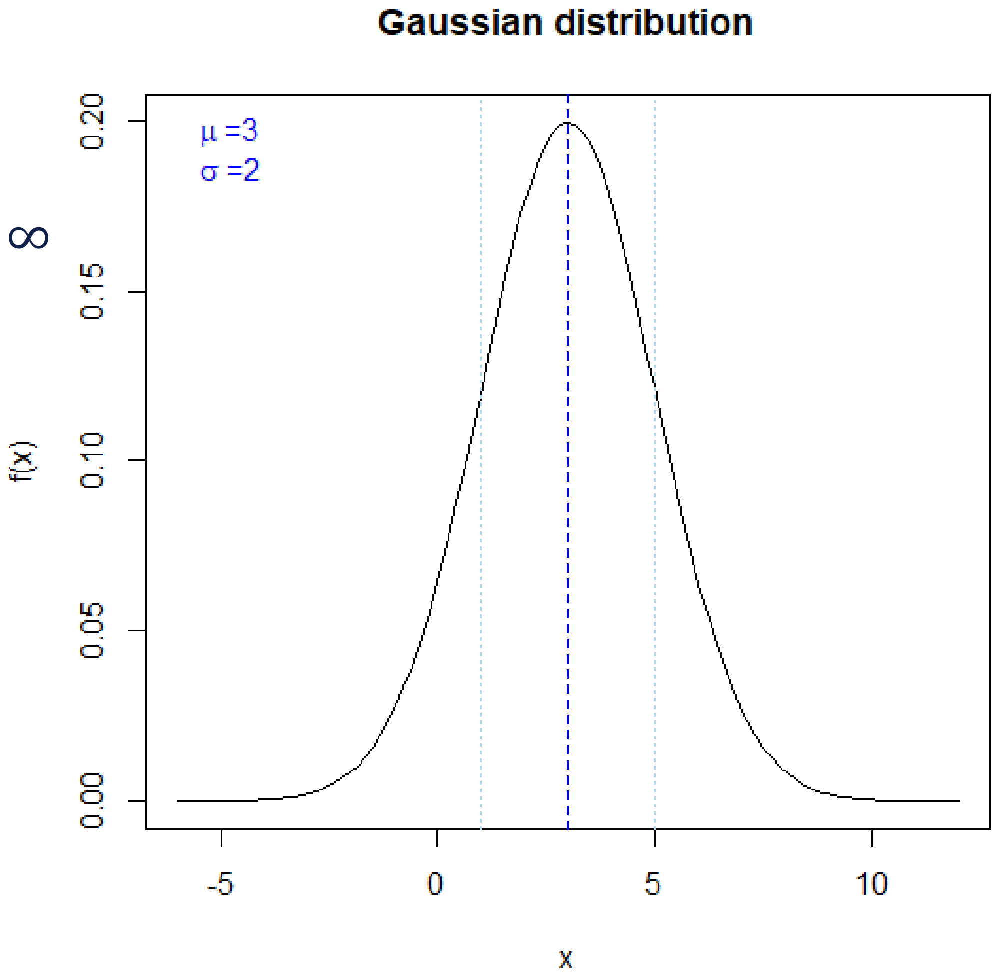
- **Parametric models**
  - Probability distribution is a function of a **fixed set of parameters**, $\boldsymbol{\theta}$, $f(x;\ \theta)$
  - Various ***families of distributions*** exist, e.g. Normal distribution, log-normal distribution, Weibull distribution, gamma distribution, …
  - Discrete and continuous distributions
  - Inference about the population involve choosing the right family and ***estimation*** of the population parameters
  - **Predictions** about future observations can be made
  - **Extrapolation** beyond the support of the data is possible
  - **Simulated data** may be generated

Example 1: Normal distribution (Gaussian)

- Fully specified by its ***mean μ*** and ***standard deviation σ***

$$f(x) = f(x;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \qquad \text{for } -\infty \le x \le \infty$$

- Symmetric around its mean
- Absolutely continuous
- CLT – (normalized) sum of independent random variables tends towards a normal distribution
- Log-normal distribution:
  - If Y is normally distributed, then $X = e^Y$ has a log-normal distribution

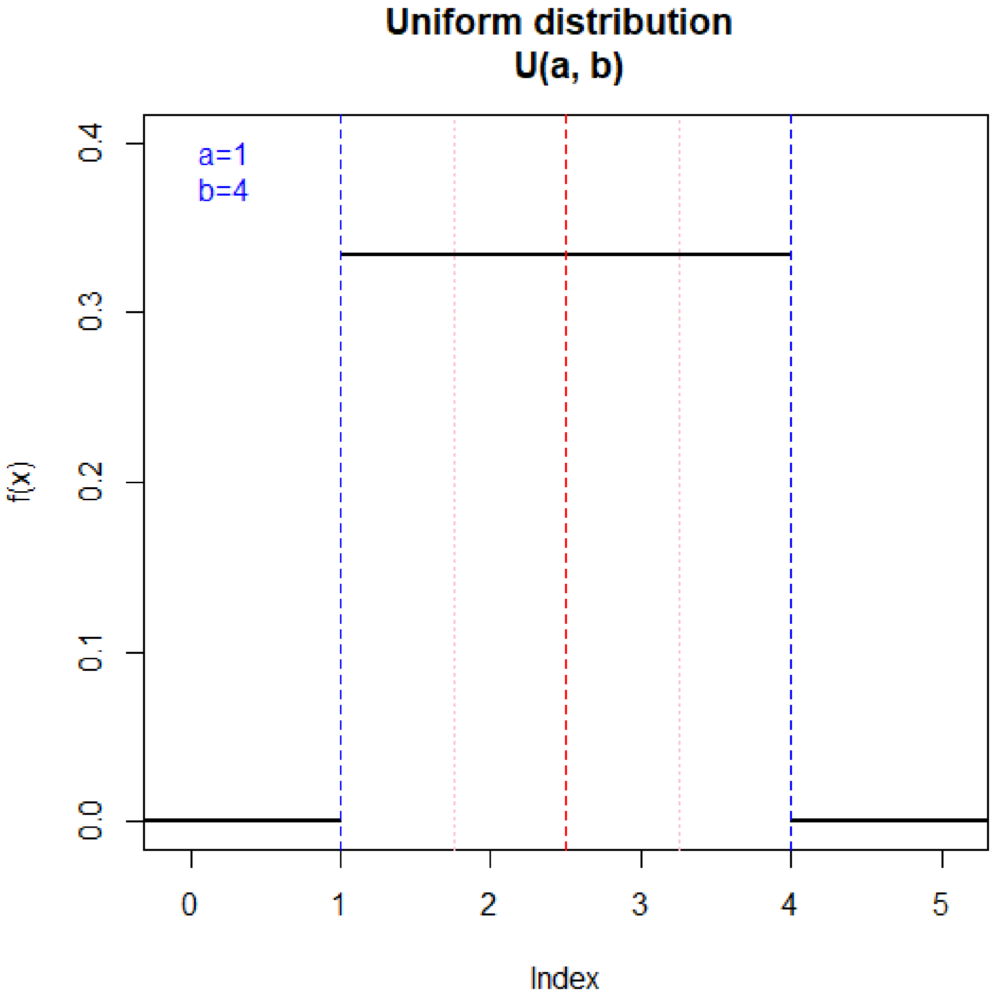


Example 2: Uniform distribution

- Fully specified by its ***range*** defined by parameters a and b

$$f(x) = f(x; a, b) = \frac{1}{b-a}, \qquad \text{for } a \le x \le b$$

- All intervals within the support equally likely
- Exist for continuous and discrete variables
- ***Probability integral transform:***
  - If X has cdf $F_X(x)$, then $Y = F_X(X)$ has a standard uniform distribution, $Y \sim U_{[0,1]}$
  - If $Y \sim U_{[0,1]}$ and X has cdf $F_X(x)$, then $F_X^{-1}(Y)$ has the same distribution as X.
  - Can be used to generate *random samples* from a probability distribution given its cdf and samples from $U_{[0,1]}$

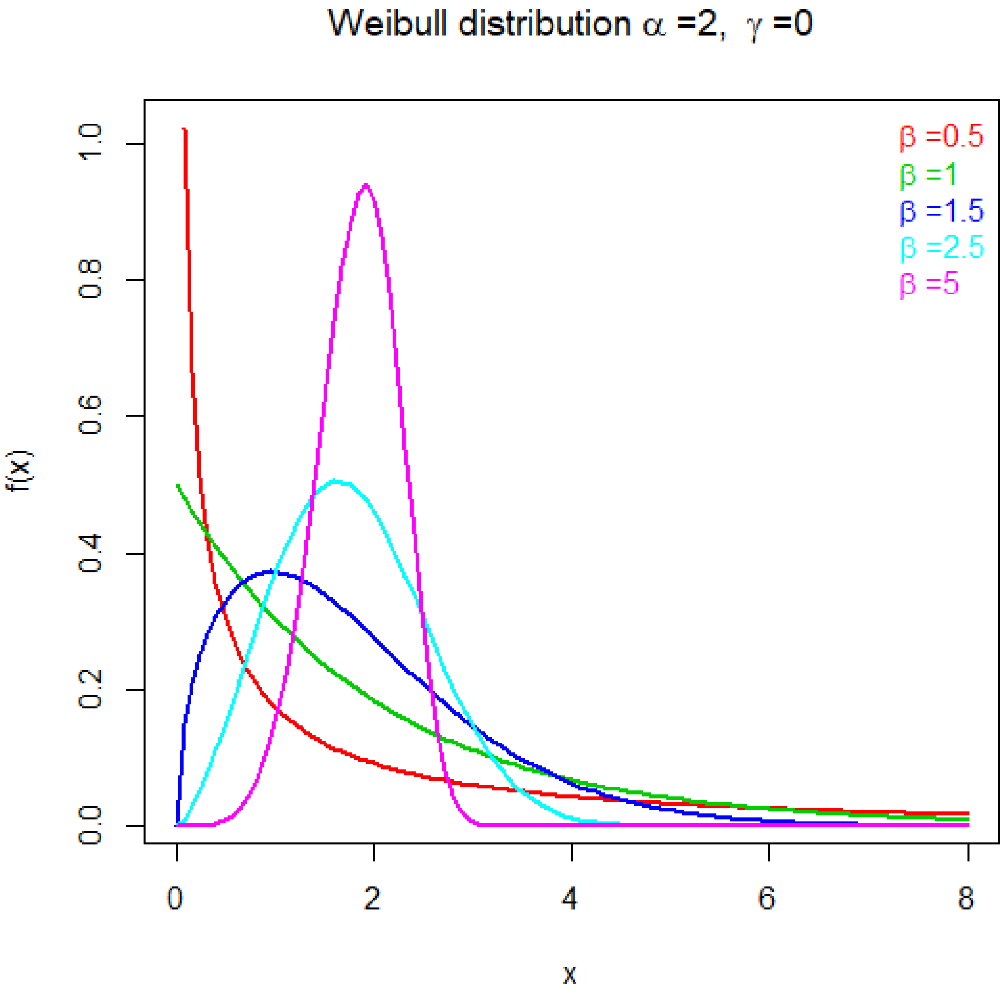


Example 3: 3-parameter Weibull distribution

- Specified by its *scale* ($\alpha > 0$), *shape* ($\beta > 0$) and *location* (γ) parameters

$$f(x;\ \alpha,\beta,\gamma) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}}, \qquad \text{for } x \geq \gamma$$

  - Location specifies a translation of the 2-parameter Weibull distribution
  - Much used in reliability analysis; Often used to model ocean waves
  - Different values of the shape parameter gives different shapes of the distribution



Non-parametric probability distributions

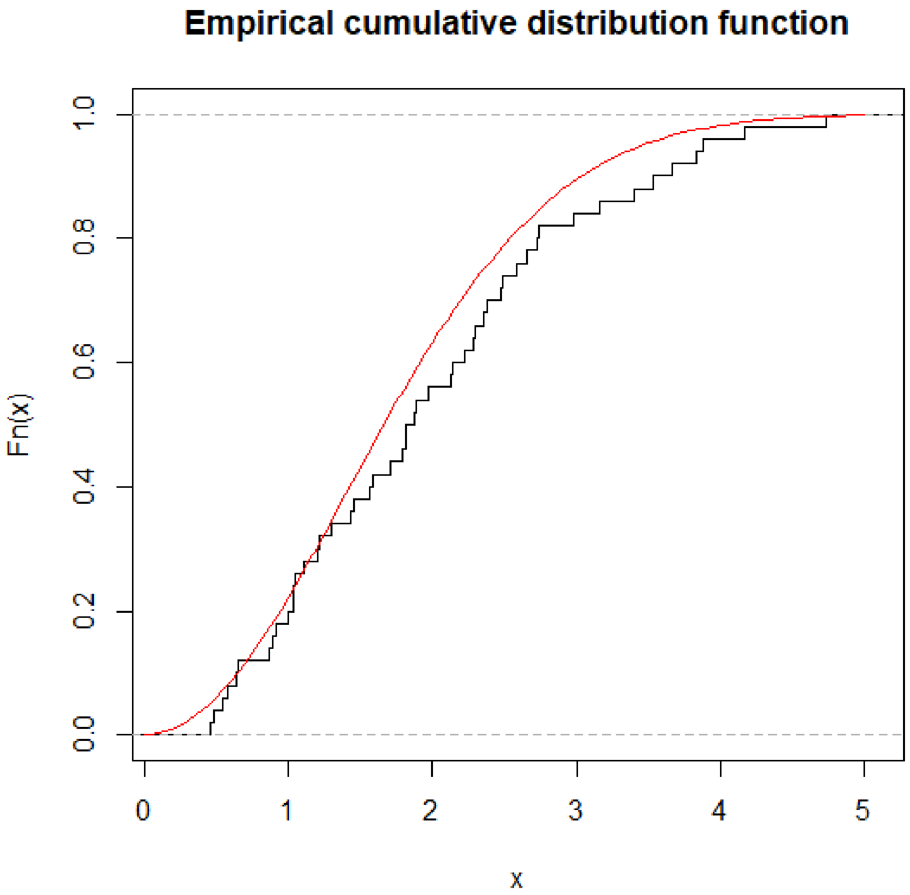
- ***Distribution-free*** - less assumptions about the underlying distribution
  - A function of a data sample – no dependency on parameters
  - May give very good fit to the data
  - But extrapolation is questionable
- Examples:
  - Histograms – provide an estimate of the probability distribution
  - Empirical distribution function – unbiased estimator for the CDF (step function)

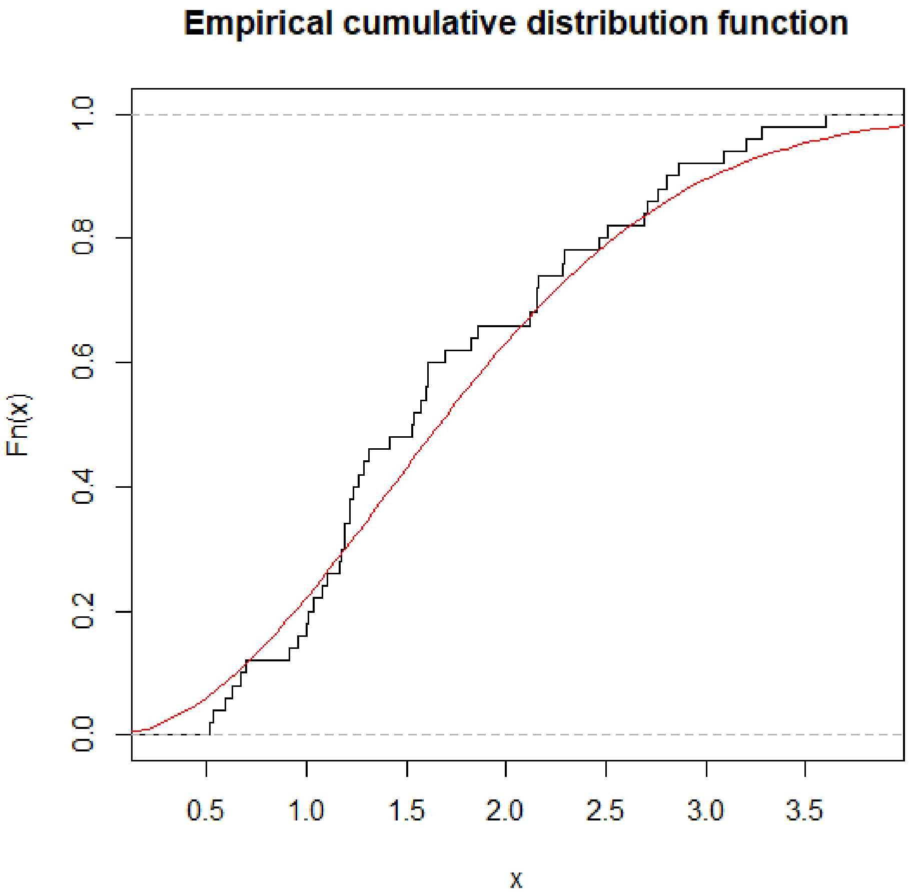
$$\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$$

  - Kernel density estimation

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Empirical distribution function

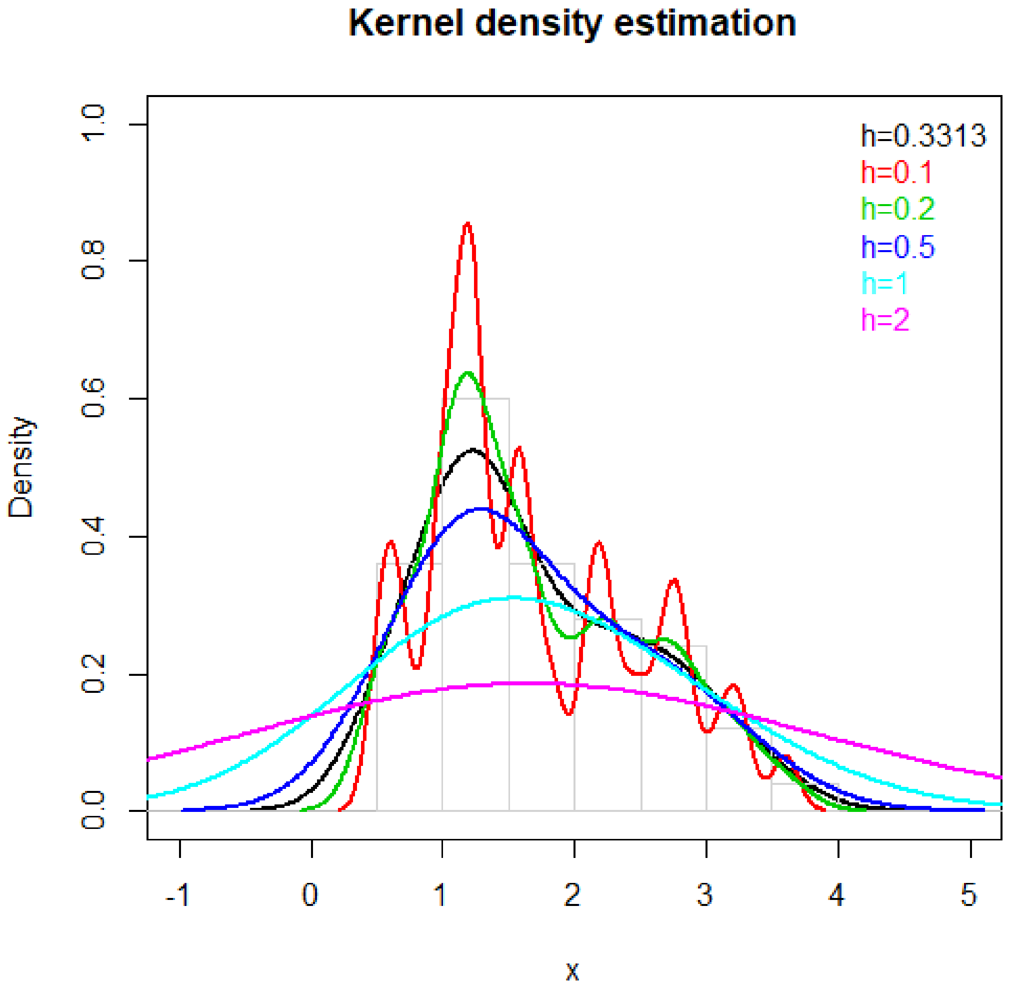

Empirical cumulative distribution function


Empirical cumulative distribution function

Kernel density estimation

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- Density estimated locally around each data point
- Need to specify a ***kernel function*** and the ***bandwidth*** parameter
  - Large bandwidth –> smooth function; small bandwidth –> very good data fit
  - Bias-variance trade-off
- Should be used with care for extrapolation, e.g. extremes
- Useful tools for descriptive data exploration



Population parameters and moments

- The **mean value** (expected value) of a random variable X with pdf f(x) is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x\, f(x) dx$$

- The **variance** of a random variable X with pdf f(x) and mean value μ is

$$\sigma_X^2 = Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x) dx = E[(X-\mu)^2]$$

- **Moments** are expected values of integer powers of X
  - The n-th moment about 0 (raw moment): $E[X^n] = \int_{-\infty}^{\infty} x^n\, f(x) dx$
  - The n-th moment about the mean (central moment): $E[(X-\mu)^n] = \int_{-\infty}^{\infty} (x-\mu)^n\, f(x) dx$
- The mean is the first moment; The variance is the second central moment; skewness is the normalized third central moment; kurtosis is the normalized fourth central moment, …

Likelihood function

- The **likelihood function** for a parametric model is a function of the model parameters conditioned on the observed data
  - Describes the *probability of observing the observed data*
- Assuming a parametrized model $f(\boldsymbol{x};\ \theta)$ and data $\boldsymbol{x}$

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x} \mid \boldsymbol{\theta})$$

  - $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ as a function of $\boldsymbol{x}$ with $\boldsymbol{\theta}$ fixed: ***probability density function***
  - $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ with $\boldsymbol{x}$ fixed: ***likelihood function***
- For $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$ assumed iid

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x} \,|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \,|\boldsymbol{\theta})$$

Building a statistical model from data

- Parametric families of distributions are specified by ***parameters***
- Building a statistical model from data typically involves
  1. ***Selecting*** the appropriate family
  2. ***Estimating*** the population parameters
  3. ***Checking*** the model

Implicit assumption:

- Data are **IID** (*independent and identically distributed*) from the same distribution
  - Under this assumption, it is possible to ***fit a parametric distribution*** model to the data
  - Fitted models may still be useful even if this assumption is not entirely fulfilled, but model improvements do exist

IID assumption may be violated due to different reasons

Identically distributed

- Different mechanisms responsible for data generation
  - Storm vs. calm weather
  - Summer vs. Winter
  - Data from different locations
  - Directional effects

Possible solutions:

- Include ***covariate effects*** in the models, e.g.

$$\alpha \rightarrow \alpha(t, x, \theta, \dots);\ \beta \rightarrow \beta(t, x, \theta, \dots); \dots$$

- Mixture models, e.g.

$$f(x) = \sum_{k=1}^{K} \pi_k \phi(x;\ \mu_k, \sigma_k)$$

- Pre-processing

Independence

- Dependence may occur if
  - Data are autocorrelated (memory)
  - Correlations in space
  - …

Possible solutions

- Time-series models, e.g. AR(p)

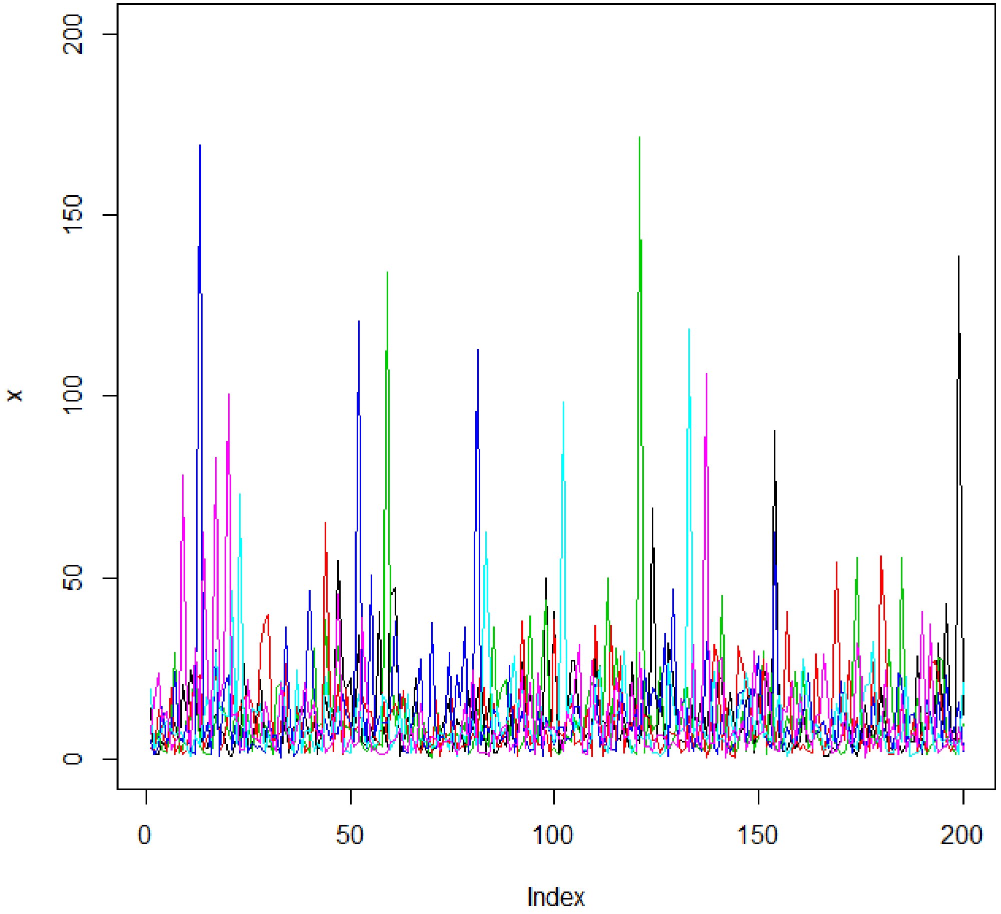
$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

- Spatial models, e.g. Kriging or Markov Random Fields
- De-clustering
- Pre-processing/sub-sampling

Sampling variability

- A sample is just ***one possible realization*** of the process
- This introduces ***sampling variability*** – different observations of the same process might give different result
- Extremes will be particularly sensitive to sampling variability
- Sampling variability may be reduced by **increasing the sample size**
  - But will never be completely removed
- Cross-validation can be used to avoid over-fitting
  - Out-of-sample validation

6 simulations from the same model (log-normal)



Choosing the “correct” probability distribution

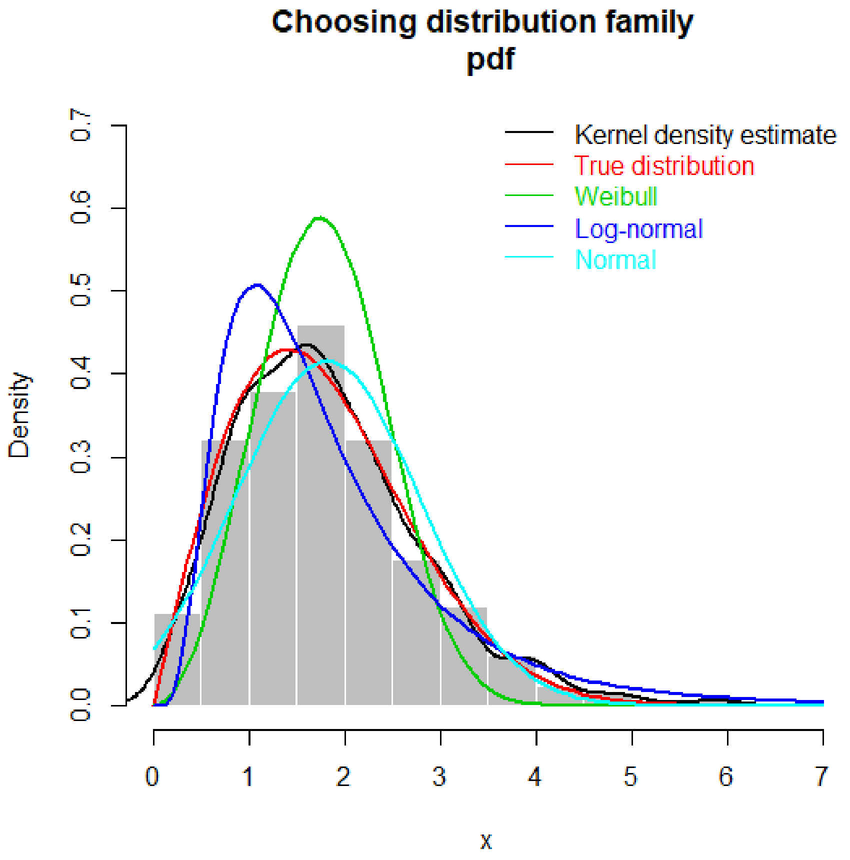
Choosing the best family of distribution model may be challenging and can be guided by different techniques

- Range of support may exclude some models
- ***Visual inspection*** of data
  - Compare histogram with pdf
  - Compare empirical cumulative distribution function with cdf
- QQ-plot
- Likelihood-based methods; AIC, likelihood ratio tests
- Statistical goodness-of-fit tests

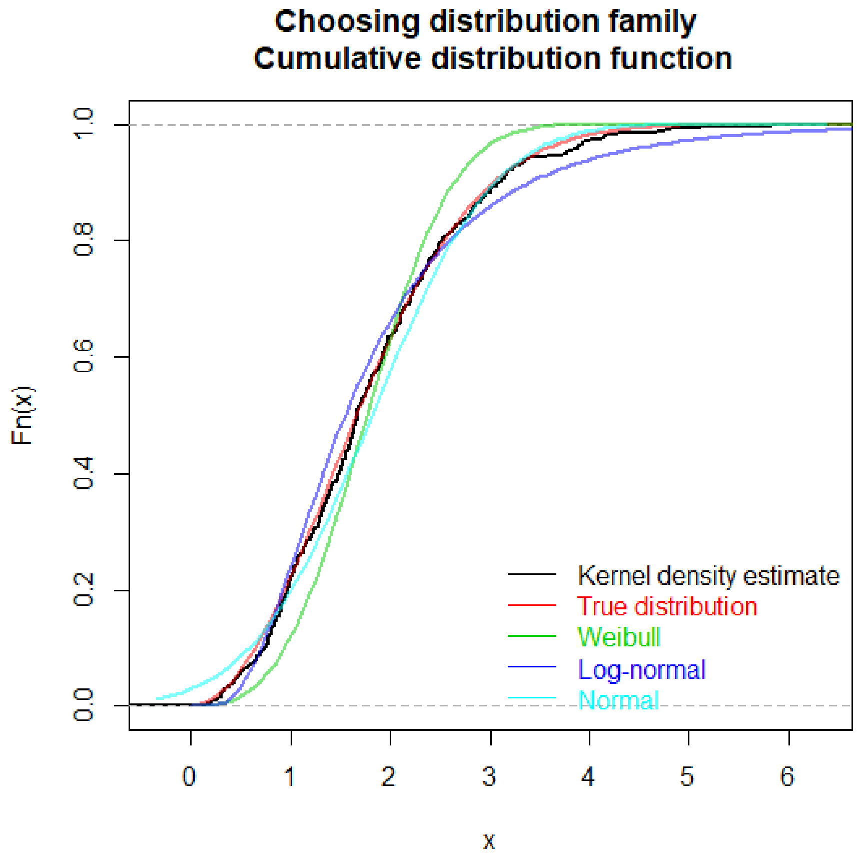
- ***Prior knowledge*** – e.g. Recommended practice, DNV-RP-C205

Visual inspection to choose the right probability distribution

Compare histogram or kernel density estimate to pdf



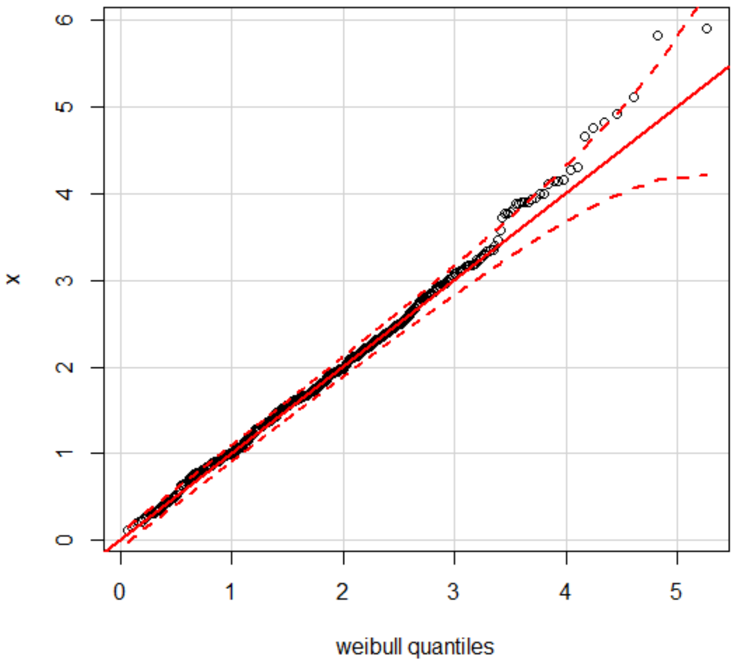
Compare ecdf with cdf



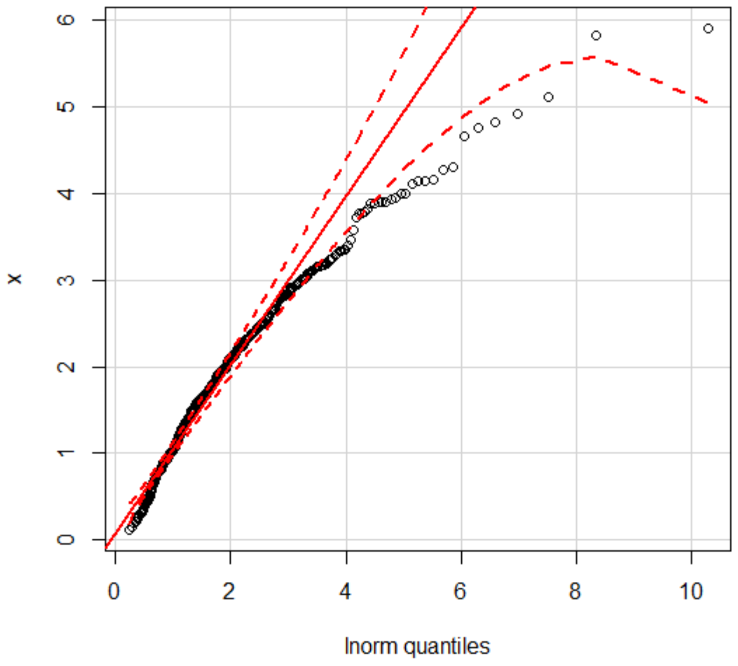
Probability-plots

- May prepare ***quantile-quantile plots*** for candidate distributions
  - Plot empirical quantile vs. theoretical quantile
  - Straight line indicate a good fit

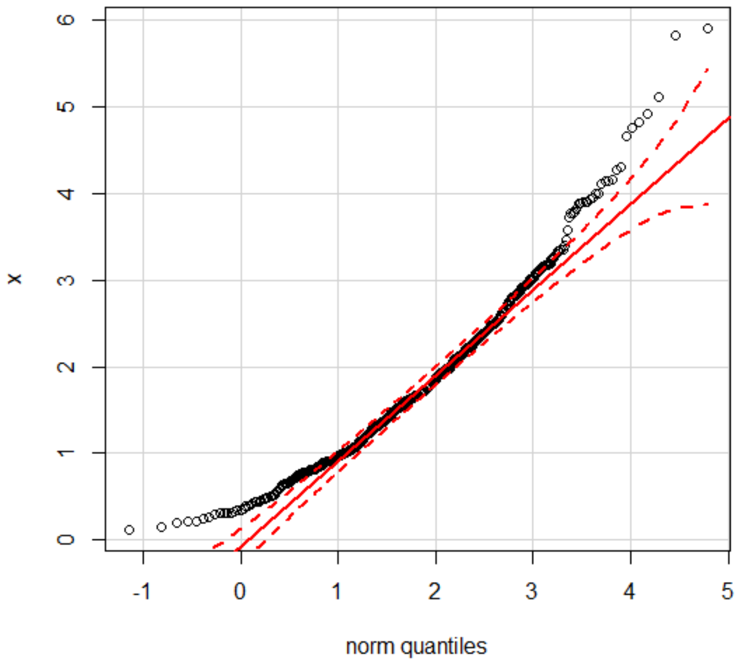
QQ plot true distribution



QQ plot log-normal distribution



QQ plot normal distribution



Likelihood-based methods - AIC

- May choose model with highest **likelihood**

$$\widehat{\boldsymbol{L}} = \mathcal{L}\left(\widehat{\boldsymbol{\theta}} \,\middle|\, \boldsymbol{x}\right)$$

  - where $\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$

- Tend to favour more complicated models – may penalize models with high number of parameters

AIC = Akaike Information Criterion:

$$AIC = 2k - 2\ln\widehat{\boldsymbol{L}}$$

- Choose the model with minimum AIC value
  - Only useful in relative comparison of different candidate models
- Other variants exist which gives different penalties for model complexity, e.g. BIC, DIC, FIC, …

Likelihood ratio tests

- May be used to compare ***nested*** models by taking the ***likelihood ratio***

$$\boldsymbol{\lambda}(\boldsymbol{x}) = \frac{\mathcal{L}_{\mathbf{0}}(\boldsymbol{\theta}_{\mathbf{0}}|\boldsymbol{x})}{\mathcal{L}_{\mathbf{1}}(\boldsymbol{\theta}_{\mathbf{1}}|\boldsymbol{x})}$$

  - Where subscript indicate the null model or the alternative model

- This ratio expresses how likely the data are under one model compared to the other

- The likelihood ratio statistic $-\mathbf{2}\,\mathbf{log}\,\boldsymbol{\lambda}$ has asymptotically a $\boldsymbol{\chi}_{\boldsymbol{q}}^{\mathbf{2}}$-distribution
  - with degrees of freedom equal to the difference in model parameters, q, between the models

- This may be used to compute ***p-values*** under the null model and test whether this should be rejected

Goodness-of-fit tests

- Numerous ***goodness-of-fit*** statistics measures how good a data sample fits to a theoretical distribution
- May be used for statistical hypothesis testing and rejection of distribution models
- Examples:
  - Anderson-Darling statistic
  - Cramer-von Mises statistic
  - Kolmogorov–Smirnov statistic
  - …
- P-values represent the probability of the data belonging to the assumed distribution
  - May reject the model for small p-values

Prior knowledge

- In some situations, ***theory*** or ***prior knowledge*** can be used to assume a particular distribution family
- For example, DNV-RP-C205 recommends certain families of distribution for certain environmental variables
  - 10-minute wind speed may be modelled by a Weibull distribution
  - Individual wave heights in a sea state may be modelled as a Rayleigh distribution
  - Crest above still water level can be modelled as a Forristall crest distribution
  - Significant wave height (H_S) can be modelled as a 3-parameter Weibull distribution
  - Annual maximum H_S can be modelled as a Gumbel distribution
  - Conditional zero up-crossing wave period can be modelled as a log-normal distribution
  - …

Notwithstanding – finding the correct distribution is challenging, and will influence the inference and any subsequent analysis

Fitting a distribution to data – parameter estimation

- Having decided on the probability distribution one needs to ***estimate the parameters***
- Several techniques exist and will give different results
  - Maximum Likelihood (ML)
  - Method of Moments (MoM)
  - Least squares (LS)
  - Quantile matching
  - Bayesian inference
  - Method of L-moments
  - Probability paper
  - Maximum Goodness-of-fit
  - …

Maximum Likelihood estimation

- Choose the parameter values that ***maximizes the likelihood*** of the data

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \arg\max_{\theta} ln\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \arg\max_{\theta} l(\boldsymbol{\theta} \mid \boldsymbol{x})$$

  - If data are iid from $f(x\,;\,\theta)$, we have

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i \mid \theta) \;\Rightarrow\; l(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n} \ln f(x_i \mid \theta)$$

  - Then, MLE is found by

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \ln f(x_i \mid \theta)$$

  - Theoretically sound approach with good asymptotic properties
  - Will also give estimates of the uncertainty of the parameter estimates (standard error)
  - However – might give wrong results if the assumed underlying distribution is not correct

Example:
MLE of a 2-parameter Weibull distribution

$$f(x;\ \alpha,\beta\ ) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}}, \qquad \text{for } x\ \geq 0$$

- Assume data x = x_1, x_2, …, x_n, iid, then

$$\mathcal{L}(\boldsymbol{\theta}\mid\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i\mid\theta) = \prod_{i=1}^{n}\frac{\beta}{\alpha}\left(\frac{x_i}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x_i}{\alpha}\right)^{\beta}}$$

$$l(\boldsymbol{\theta}|\ \boldsymbol{x}\ ) = \sum_{i=1}^{n}\ln f(x_i\mid\theta) = \sum_{i=1}^{n}\left(\ln\beta - \beta\ln\alpha + (\beta-1)x_i - \left(\frac{x_i}{\alpha}\right)^{\beta}\right) = n\ln\beta\ - n\beta\ln\alpha + (\beta-1)\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\left(\frac{x_i}{\alpha}\right)^{\beta}$$

- MLE are α,β that maximizes this expression (in practice – minimizes negative log-likelihood)
- May be found by taking the partial derivatives and setting them to 0
- In practice often found by optimization functions in standard software
  - For example, *nlm*() in *R*.

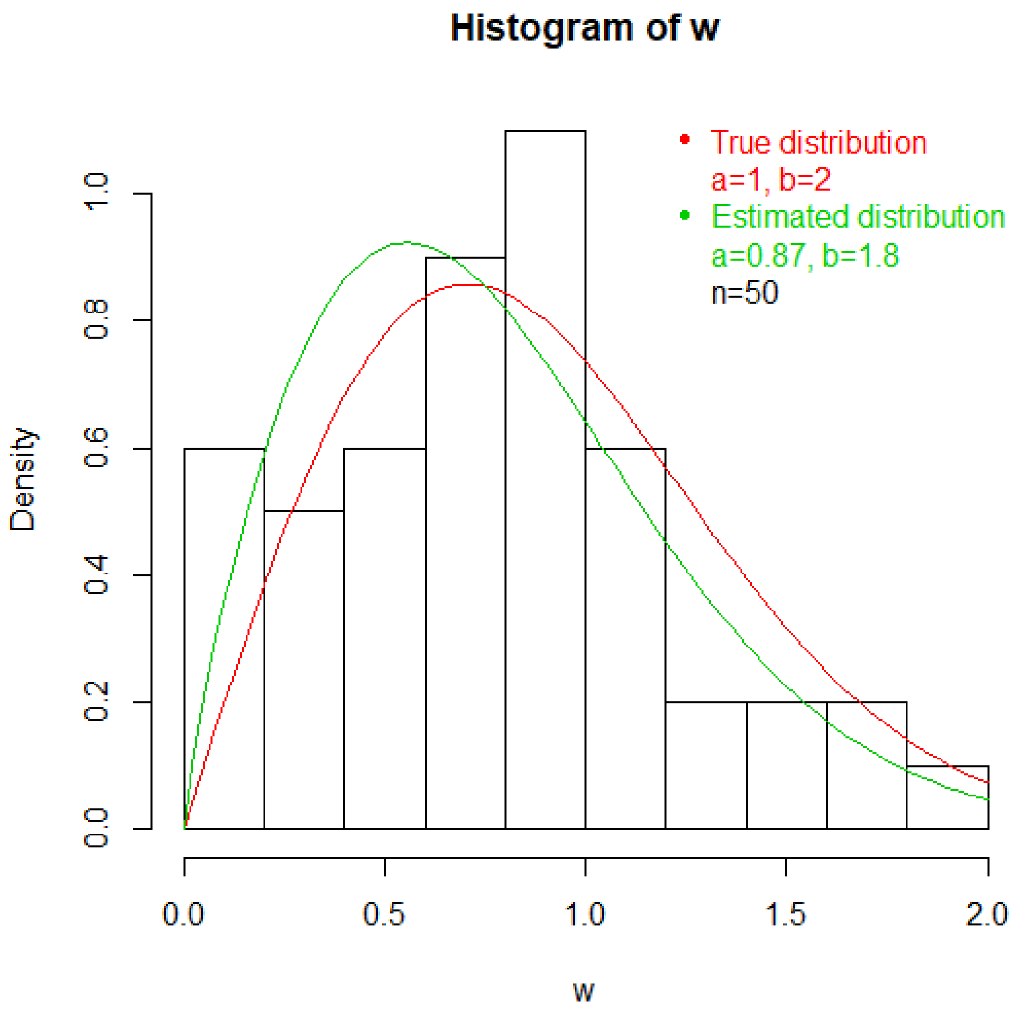
Example: MLE of a Weibull distribution in R

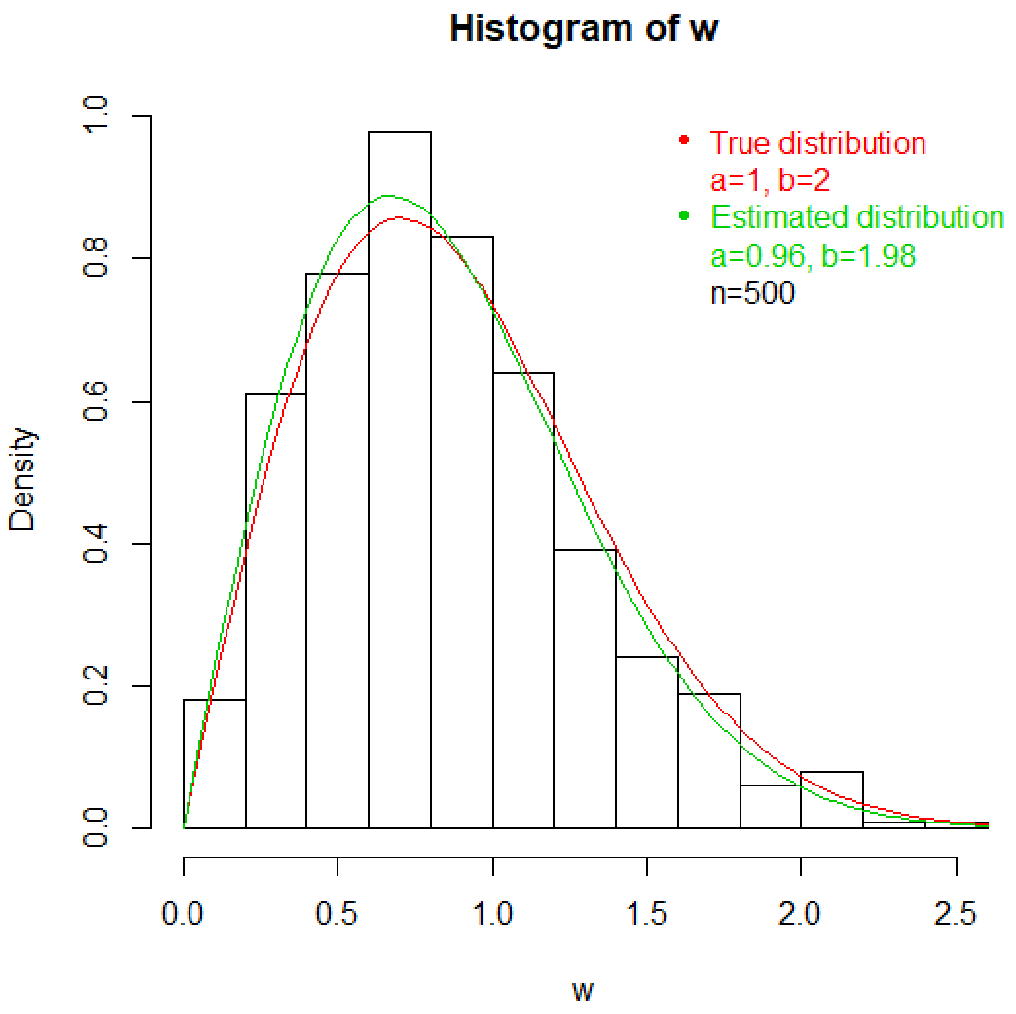
```
### Simulate from a Weibull distribution to get some data

n = 50; scale = 1; shape = 2

w = rweibull(n, scale = scale, shape = shape)


#### Define negative log-likelihood function a = scale; b = shape

logL = function(data, par){

        a = par[1]; b = par[2]

        l = log(b) -b*log(a) -(data/a)^b + (b-1)*log(data)

        sum(l)

}

minuslogL = function(data, par){

        (-logL(data, par))

}
```

```
#### Find minimum of negative likelihood

theta_hat = nlm(function(par) minuslogL(w, par), c(1, 1), hessian = T)


#### Plot results

hist(w, freq=F)

curve(dweibull(x, scale=scale, shape=shape), col=2, add=T)

curve(dweibull(x, scale=theta_hat$estimate[1], shape= theta_hat$estimate[2]),
col=3, add=T)
```

- The MLE can be found from **theta_hat$estimate**

Weibull MLE - results





Parameter uncertainty with MLE

- Theoretical result: $\widehat{\boldsymbol{\theta}} \approx N_p\left(\boldsymbol{\theta}, \hat{J}_{obs}^{-1}\right)$,

where $\hat{J}_{obs} = -\frac{\partial^2 \hat{l}_n(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2}{\partial \theta_2^2} \end{pmatrix} \hat{l}_n(\theta)$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ is the observed information matrix

- This is returned as the Hessian matrix with *nlm*() in *R* for example
- The ***standard errors*** of the parameter estimates are the square root of the diagonal elements of the inversed information matrix
- In the example above, the standard error estimates can be obtained by

se = *sqrt(diag(solve(theta_hat$hessian)))*

- A 95% confidence interval for the parameters can then be found by

$$\alpha = \hat{\alpha} \pm 1.96\, \widehat{se}_\alpha$$
$$\beta = \hat{\beta} \pm 1.96\, \widehat{se}_\beta$$

Example above; Weibull with $\boldsymbol{\alpha} = \mathbf{1}$ and $\boldsymbol{\beta} = \mathbf{2}$

n = 50

- $\hat{\alpha} = 1.120, \hat{\beta} = 1.988$
- $\widehat{se}_\alpha = 0.0846,\ \widehat{se}_\beta = 0.2076$

n = 100

- $\hat{\alpha} = 1.0198, \hat{\beta} = 2.0325$
- $\widehat{se}_\alpha = 0.0528,\ \widehat{se}_\beta = 0.1533$

n = 250

- $\hat{\alpha} = 0.944, \hat{\beta} = 1.8840$
- $\widehat{se}_\alpha = 0.0334,\ \widehat{se}_\beta = 0.0922$

n = 500

- $\hat{\alpha} = 0.994, \hat{\beta} = 2.0663$
- $\widehat{se}_\alpha = 0.02267,\ \widehat{se}_\beta = 0.0720$

Method of Moment estimation

- Main idea: put ***empirical moments equal to*** the corresponding ***population moments***
  1. Express the population moments as functions of the model parameters
  2. Calculate the sample moments from the data
  3. Set the population and sample moments equal and solve for the parameters
  4. The solutions to these equations are the ***method of moment estimates*** of the parameters
- Need same number of equations as the number of parameters

Population moments as a function of the parameters: $\mu_j = E[X^j] = g_j(\theta_1, \theta_2, \dots, \theta_p),\ \text{for } j = 1, \dots p$

Sample moments from a sample of size n: $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$

The method of moments estimators for $\theta_1, \theta_2, \dots, \theta_p$ are then the solutions to the set of equations

$$\begin{gathered}\hat{\mu}_1 = g_1(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p)\\ \hat{\mu}_2 = g_2(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p)\\ \vdots\\ \hat{\mu}_p = g_p(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p)\end{gathered}$$

Example: MoM for Normal distribution

$$f(x) = f(x;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \qquad \text{for } -\infty \le x \le \infty$$

- First population moment (mean): $\mathrm{E}(\mathrm{X}) = \mu$
- Second population moment: $\mathrm{E}(X^2) = \sigma^2 + \mu^2 \Rightarrow \sigma^2 = \mathrm{E}(X^2) - \mu^2$
- First sample moment: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Second sample moment: $\frac{1}{n}\sum_{i=1}^{n} X_i^2$

Equating population and sample moments give the MoM-estimators

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Method of Moments estimators

- May be biased, and if the distribution is known, then MLE will be better
- However, MoM might be OK, even if the assumed underlying distribution is wrong
  - It will always get the moments right!
  - May be simple to calculate by hand

- Related techniques that are more robust estimate parameters by equating
  - Population and sample probability weighted moments (PWM)
  - Population and sample L-moments

- Estimation by quantile matching equates theoretical quantiles (as a function of population parameters) with sample quantiles
  - May give more weight to particular parts of the distributions

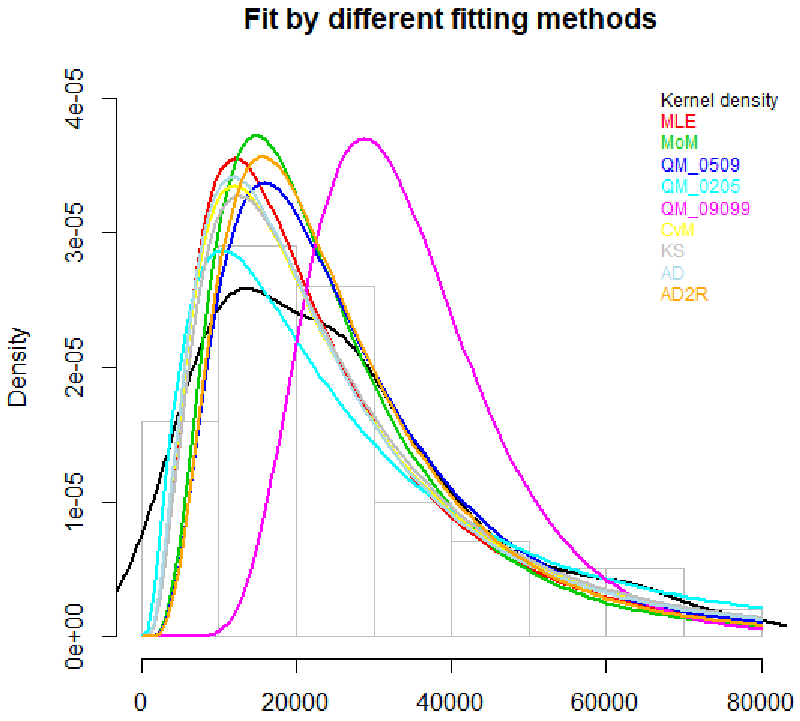
Parameter estimation in R – some libraries

- Many techniques for parameter estimation exist
  - ML and MoM two of the most commonly used
- Different techniques are typically implemented in standard software and may give different results
- Examples of libraries in R:
  - ***MASS::fitdistr*** – ML estimation with support for many in-build distributions
  - ***Fitdistrplus::fitdist*** – Supports estimation for many distributions using different estimation techniques
    - mle: Maximum likelihood
    - mme: moment matching estimation
    - qme: quantile matching estimation
    - mge: maximum goodness-of-fit estimation; different GoF-statistics
  - …

Example: Fitting lognormal distribution using different methods

- Using R and *fitdistrplus* package
- Fitting a Weibull distribution to simulated data (n = 100) by
  - MLE
  - MoM
  - Quantile fitting (different quantiles)
  - Goodness-of-fit (different measures)

True parameters: $\mu = 10, \sigma = 0.8$



	True	MLE	MoM	QM1	QM2	QM3	CvM	KS	AD	AD2R
μ	10	9.915	9.966	10.052	10.052	10.391	9.954	9.983	9.938	10.016
σ	0.8	0.720	0.604	0.617	0.893	0.353	0.751	0.738	0.746	0.598
$\eta_{0.95}$	82 116	66 121	57 547	64 045	100 744	58 177	72 281	72 932	70 555	59 841

Exercise 1 – fitting distributions to wave data

1. Fit a log-normal distribution to SWH-data using method of moments (MoM)
   a. Fit a log-normal distribution using ML
2. Fit a Weibull distribution to the data using Maximum Likelihood
   a. Also calculate the 95% confidence bands for the parameter estimates
   b. Fit a Weibull distribution to data by maximum goodness-of-fit.
   c. Fit a Weibull by quantile matching, using e.g. the 75% and 95% quantiles
3. Fit an exponential and a gamma distribution to the data by ML
4. Estimate the median, 95%- quantile and 99%-quantile from the models and compare (only MLE). Comment
5. Do visual model checking by plotting the empirical and fitted density curves
   a. Show QQ-plots for the various distributions
6. Compare by AIC to find the best alternative (Only for MLE)

Solution – Exercise 1: MoM estimators for the lognormal distribution

Population moments for log-normal distribution: $E(X^n) = e^{n\mu+\frac{1}{2}n^2\sigma^2}$

Two parameters – need two moments: $m_1 = E(X) = e^{\mu+\frac{1}{2}\sigma^2}$ and $m_2 = E(X^2) = e^{2\mu+2\sigma^2}$

$$\Longrightarrow \begin{array}{l} \mu = \ln\left(\frac{m_1^2}{\sqrt{m_2}}\right) \\ \sigma = \sqrt{\ln\left(\frac{m_2}{m_1^2}\right)} \end{array}$$

Equate with the corresponding sample moments: $$\begin{array}{l} \widehat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \widehat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{array}$$

MoM estimators for the log-normal distribution

$$\hat{\mu} = \ln\left(\frac{\widehat{m}_1^2}{\sqrt{\widehat{m}_2}}\right) \quad \text{and} \quad \hat{\sigma} = \sqrt{\ln\left(\frac{\widehat{m}_2}{\widehat{m}_1^2}\right)}$$

Solution – Exercise 1:
ML estimators for the log-normal distribution

$$f(x) = \frac{1}{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\left(\frac{(\ln x - \mu)^2}{2\sigma^2}\right)} \;\Rightarrow\; \mathcal{L}(\mu,\sigma|\boldsymbol{x}) = \prod_{i=1}^{n}\frac{1}{x_i}\frac{1}{\sigma\sqrt{2\pi}}e^{-\left(\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right)}$$

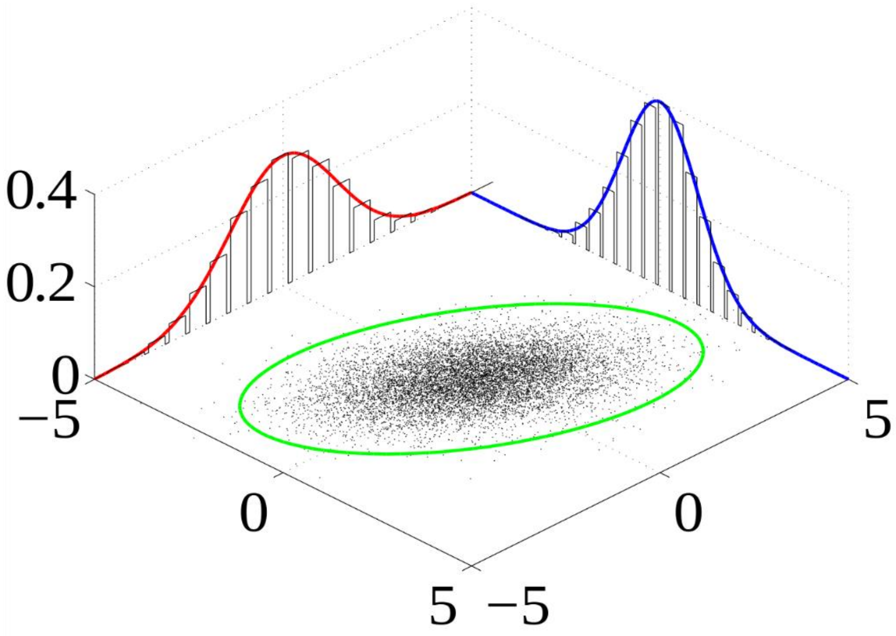
$$\Rightarrow l(\mu,\sigma|x) = \sum_{i=1}^{n}\left(-\ln x_i - \frac{1}{2}\ln\sigma^2 - \frac{1}{2}\ln 2\pi - \frac{1}{2\sigma^2}(\ln x_i - \mu)^2\right)$$

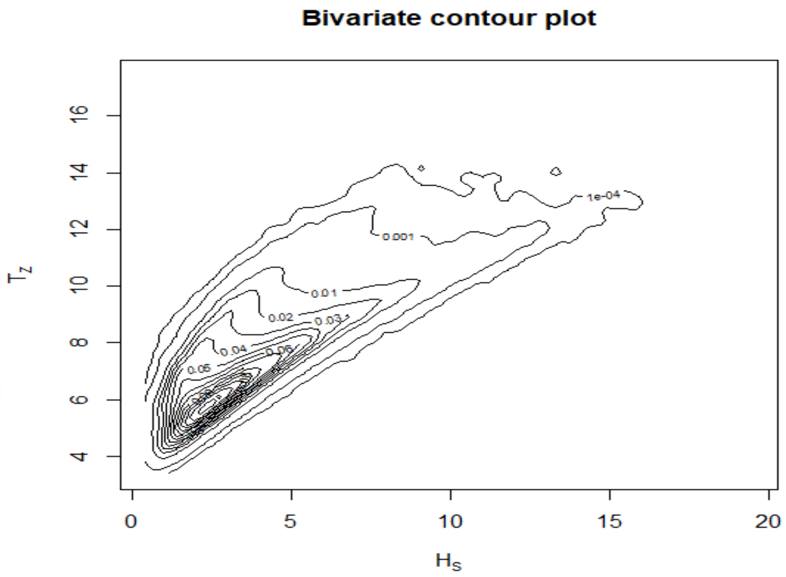
Find maximum by $\frac{\partial l}{\partial\mu} = 0$ and $\frac{\partial l}{\partial\sigma} = 0$

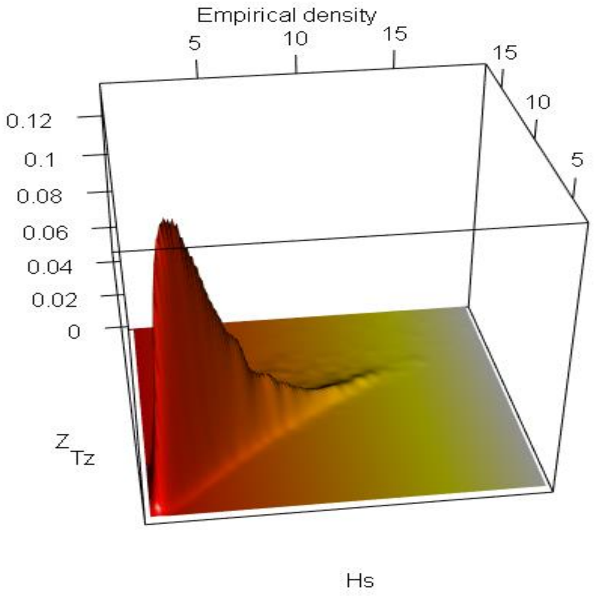
$$\frac{\partial l(\mu,\sigma\,|\boldsymbol{x})}{\partial\mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(\ln x_i - \mu) = 0 \Rightarrow \boldsymbol{\hat{\mu}} = \boldsymbol{\frac{1}{n}\sum_{i=1}^{n}\ln x_i} \qquad \text{for } \sigma^2 \neq 0$$

$$\frac{\partial l(\mu,\sigma\,|\boldsymbol{x})}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(\ln x_i - \mu)^2}{2\sigma^4} \Rightarrow \boldsymbol{\hat{\sigma}^2} = \boldsymbol{\frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \hat{\mu})^2} \qquad \text{for } \sigma^2 \neq 0$$

Typically found by numerical optimization using computer programs such as R




Bivariate contour plot


Empirical density

Multivariate analysis

Multivariate analysis: Joint probability distributions

Definitions (2d):

$f(x,y)$ is the ***joint probability density function*** for X and Y if, for any two-dimensional set A,

$$P[(X,Y) \in A] = \iint_A f(x,y)dx\,dy$$

The ***joint cumulative distribution*** function is given by

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$$

The ***marginal probability density functions*** of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Some properties:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dx\,dy = 1$$

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du$$

$$F_{X,Y}(\infty,\infty) = 1$$

$$F_X(x) = F_{X,Y}(x,\infty)$$

$$F_Y(y) = F_{X,Y}(\infty,y)$$

Independence

- Variables X and Y are said to be ***independent*** if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

- Variables $X_1, X_2, \dots, X_n$ are said to be ***independent*** if

$$f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n)$$

- For independent variables, finding a joint distribution boils down to finding the marginal distribution of each variable and multiplying them together.
- In most cases, however, this is not the case, and such models may give wrong results
  - For example: extreme winds and extreme waves tend to appear together – strongly positive correlation
  - Failure to model such dependencies may seriously under- or overestimate the probability of extreme conditions

Conditional distribution

The **conditional probability density function** of Y given that $X = x$ is

$$f_{Y|X}(y\,|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

- Independence implies that the conditional distribution of Y given $X = x$ is the same as the unconditional distribution
  - Knowing the value of X gives no additional information about the value of Y

$$f_{Y|X}(y\,|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \xrightarrow{X,Y\ \text{independent}} \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y)$$

Combining long-term and short-term statistics

- Sometimes there is a need to combine ***long-term*** and ***short term*** statistics
  - E.g. combining statistical description of ***sea states*** with short-term description of ***wave heights*** within a sea state for long-term description of wave heights
- Typically comprises ***three steps***
  - Model for long-term sea state parameters such as H_S and T_Z
  - Short-term modelling of individual wave heights conditioned on the sea state, $h \mid H_S = h_s, T_Z = t_z$
  - Combining the two distributions by ***integrating out*** the sea state parameters

$$F_H(h) = \int_{h_s} \int_{t_z} F_H(h|h_s, t_z)\, f_{H_S,T_Z}(h_s, t_z)\; dh_s dt_z$$

Joint modelling of depended variables

- Non-parametric
  - E.g. kernel density estimation

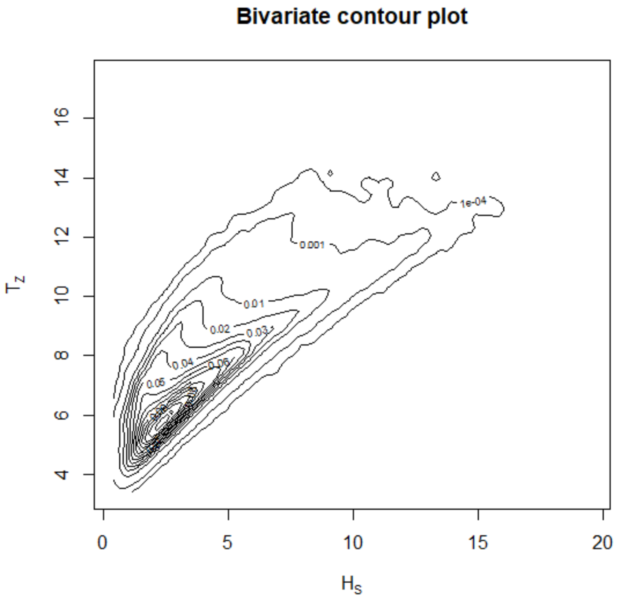
- Parametric multivariate distributions

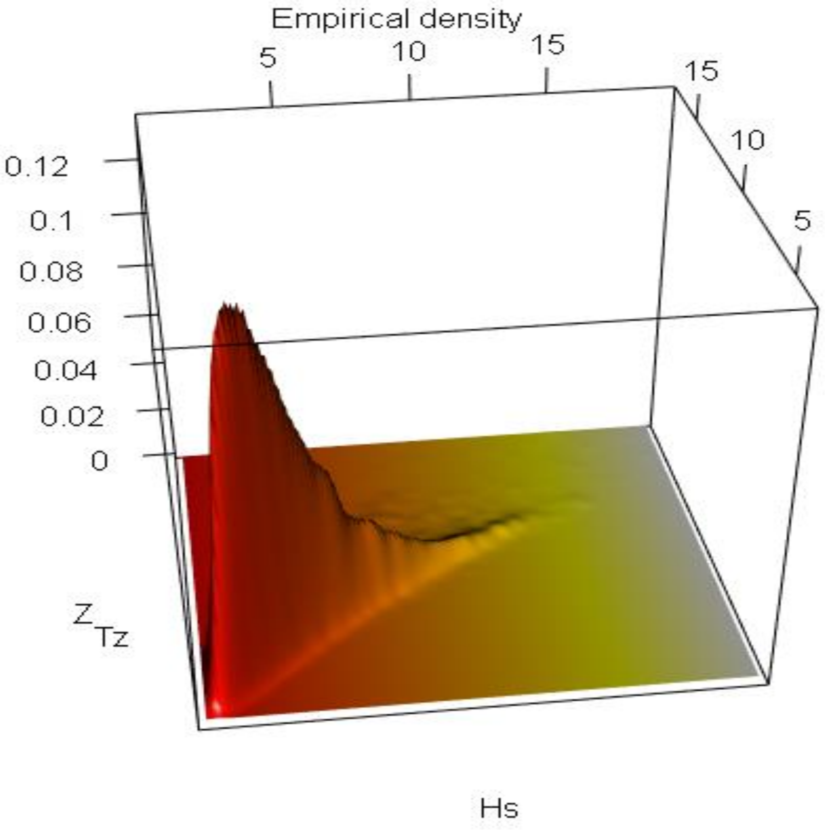
- Conditional modelling

- Copula models

Non-parametric models

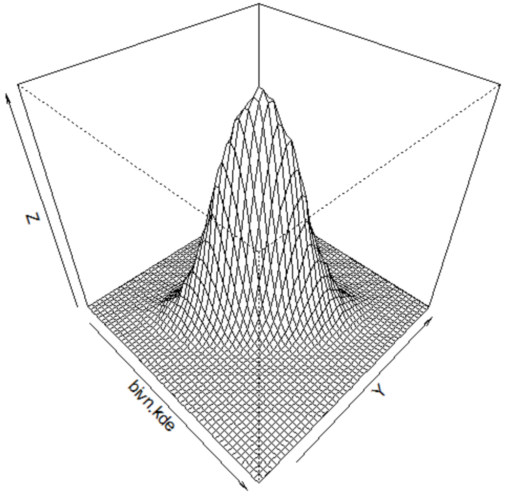
- Empirical density may be estimated by kernel density estimation, as in the univariate case
- Useful for visualizing the data
  - Perspective plots
  - Contour plots
- Not good for extrapolation





Parametric multivariate distributions

- There are some multivariate parametric distribution functions that may be fitted to the data
- Multivariate normal distribution
  - Fully specified by its mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$



$$\boldsymbol{f_X(x_1, x_2, \dots, x_n) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}}$$

- Multivariate log-normal distribution
  - If $X \sim N(\mu, \Sigma)$ has a multivariate normal distribution, then $Y = e^X$ has a multivariate log-normal distribution

- Model parameters may be fitted in a similar way, e.g. by ML, MoM, goodness-of-fit, etc…

Conditional modelling approach

- A somewhat more flexible modelling approach is to build a joint distribution model with a ***conditional modelling approach***, e.g.

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y\,|x)$$

$$\begin{aligned} & f_{X_1,\,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) \\ & = f_{X_1}(x_1) f_{X_2|X_1}(x_2\,|x_1) f_{X_3|X_2,X_1}(x_3\,|x_2,x_1) \cdots f_{X_n|X_{n-1},X_{n-2},\ldots X_1}(x_n|x_{n-1},x_{n-2},\ldots,x_1) \end{aligned}$$

  - Need then to estimate the marginal distribution of one variable and the conditional distribution of the others
  - For the conditional model – may model the conditional parameters as a function of the conditioning variables
  - Much used in offshore engineering, and e.g. recommended in DNV-RP-C205

Example: Conditional model for significant wave height and wave period

- DNV-RP-C205 (sec. 3.6.3) recommends the following joint distribution for H_S and T_Z

$$f_{H,T}(h,t) = f_H(h) f_{T|H}(t\,|h)$$

with a 3-parameter marginal for H_S and a conditional log-normal for T_Z as follows

$$h_H(h) = \frac{\beta}{\alpha}\left(\frac{h-\gamma}{\alpha}\right)^{\beta-1} e^{-\left(\frac{h-\gamma}{\alpha}\right)^{\beta}}$$

$$f_{T|H}(t|h) = \frac{1}{t\,\sigma(h)\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\ln t - \mu(h))^2}$$

and with

$$\mu(h) = E[\ln T_Z\,|H_S] = a_0 + a_1 h^{a_2}$$
$$\sigma(h) = sd[\ln T_Z\,|H_S] = b_0 + b_1 e^{b_2 h}$$

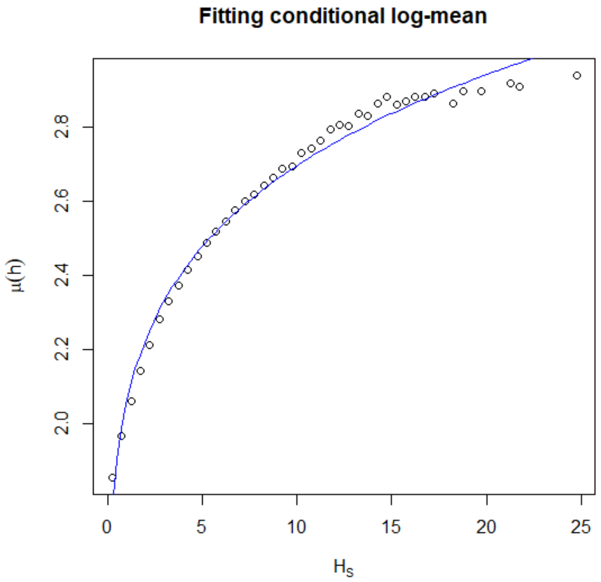
- Model parameters are $\alpha, \beta, \gamma, a_0, a_1, a_2, b_0, b_1, b_2$ which may be fitted by e.g. maximum likelihood, least squares, etc…

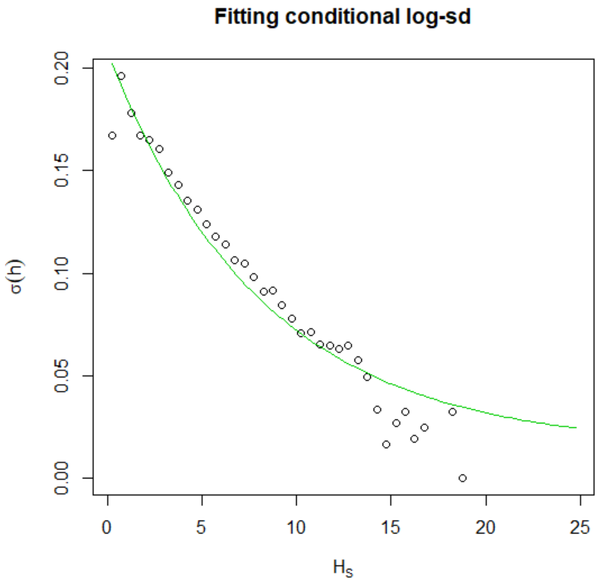
Fitting conditional model by least squares

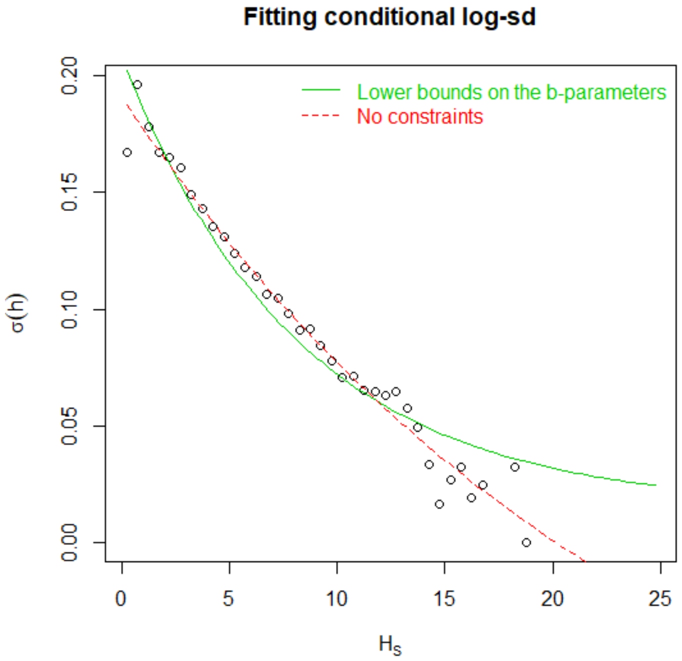
Alternative way of fitting the conditional log-normal distribution

1. Bin data in intervals of the value of H_S; e.g. 1-m bins
2. Calculate the mean and standard deviation of $\log(T_Z)$ in each bin
3. Fit the parametric models for $\mu(h)$ and $\sigma(h)$ to the binned data by minimizing least squares

- The binning will effectively give more emphasize of the tails







Multivariate modelling with copula

- Any joint probability density can be expressed as the product of the marginal probability densities and a copula density

$$h(x,y) = f(x)g(y)c(F(x),G(y))$$

- The copula is the joint cumulative distribution function of $(F_X(x), F_Y(y))$

$$C(u,v) = P[F_X(x) \leq u, F_Y(y) \leq v]$$

- Estimating such a model consist of estimating the marginal probability distributions and the copula describing the dependence structure
  - De-couples the marginal modelling and the dependence modelling
- Several parametric copulas exist
  - May also be combined by extra-parametrization/combining several copulas
  - May be fitted by e.g. ML, GoF, pseudo-ML …
- Higher-dimensional models may be constructed by pair-wise copula constructions

Some parametric copula models (2d only)

1-parameter models

- Independent copula: $c_I(u,v) = uv$
- Gaussian copula: $C_{\boldsymbol{\Sigma}}(u,v) = \boldsymbol{\Phi}_{\boldsymbol{\Sigma}}(\Phi^{-1}(u), \Phi^{-1}(v))$
- Gumbel copula: $C_\theta(u,v) = \exp\left(-\left((-\ln u)^\theta + \ (-\ln v)^\theta\right)^{\frac{1}{\theta}}\right)$
- Clayton copula: $C_\theta(u,v) = \ \left(\max\{u^{-\theta} + v^{-\theta} - 1; 0\}\right)^{-\frac{1}{\theta}}$
- Frank copula: $C_\theta(u,v) = -\frac{1}{\theta}\log\left(1 + \frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)}{\exp(-\theta)\ -1}\right)$

2-parameter model

- Marshall-Olkin copula: $C_{\alpha,\beta}(u,v) = \min\left(u^{1-\alpha}v, uv^{1-\beta}\right)$

Building asymmetric models by extra-parametrization

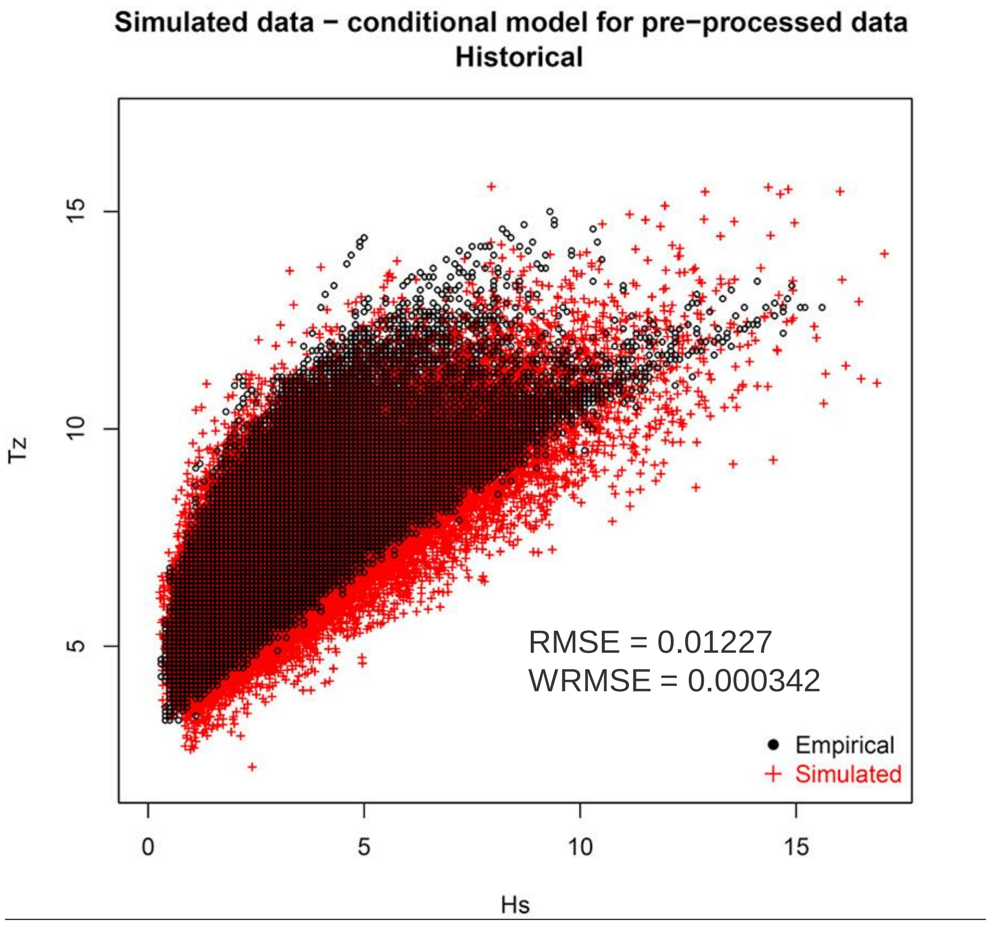
If $A(u,v)$ and $B(u,v)$ are copulas, then also

$$C(u,v) = A_{\theta_1}\left(u^{\alpha}, v^{\beta}\right)B_{\theta_2}\left(u^{\alpha-1}, v^{\beta-1}\right)$$

Is a copula, with $0 \leq \alpha, \beta \leq 1$ additional model parameters.

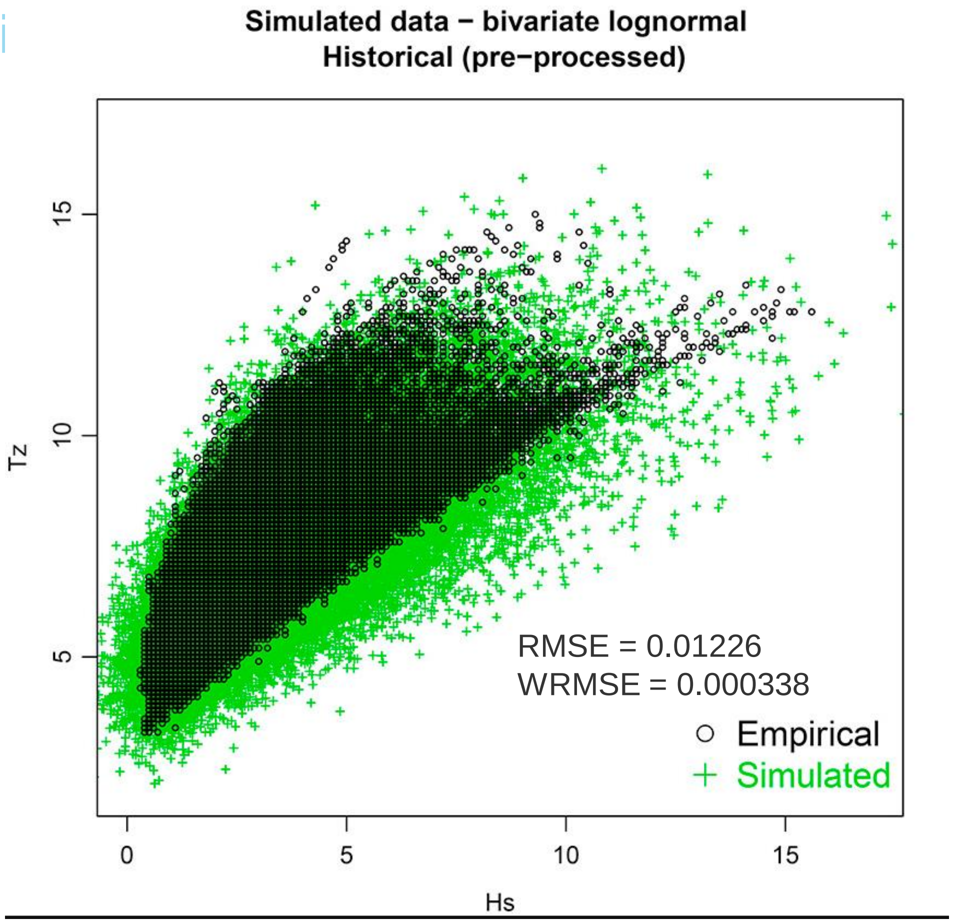
- This may be used to construct asymmetric copulas from a set of simple symmetric copulas.

Example: Fitting bivariate distribution to (H_S, T_Z) data



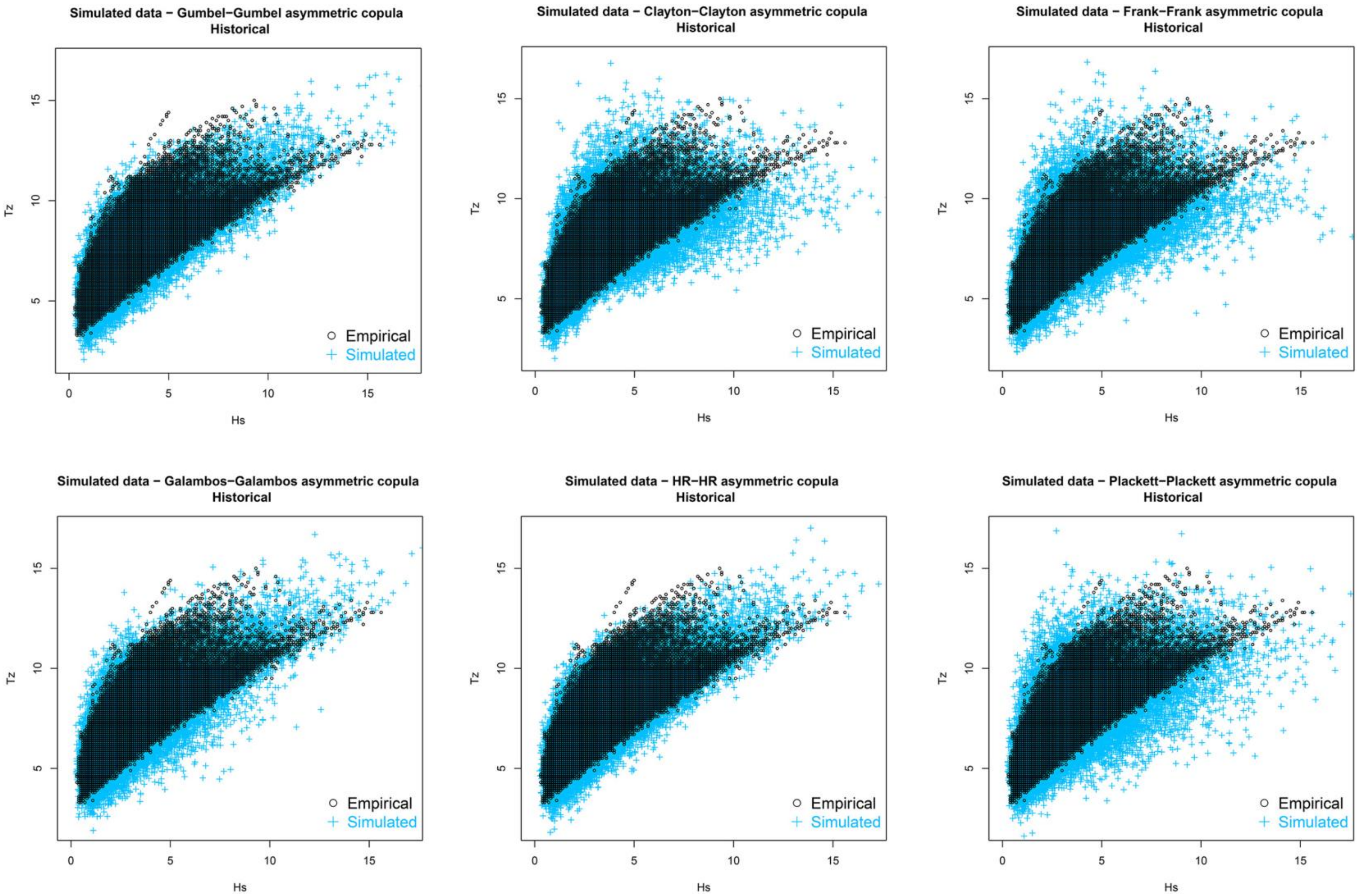
Simulated data − conditional model for pre−processed data
Historical

Bi



Simulated data − bivariate lognormal
Historical (pre−processed)

Example: Fitting bivariate distribution to (H_S, T_Z) data with copulas



Simulated data – Gumbel–Gumbel asymmetric copula
Historical

Simulated data – Clayton–Clayton asymmetric copula
Historical

Simulated data – Frank–Frank asymmetric copula
Historical

Simulated data – Galambos–Galambos asymmetric copula
Historical

Simulated data – HR–HR asymmetric copula
Historical

Simulated data – Plackett–Plackett asymmetric copula
Historical

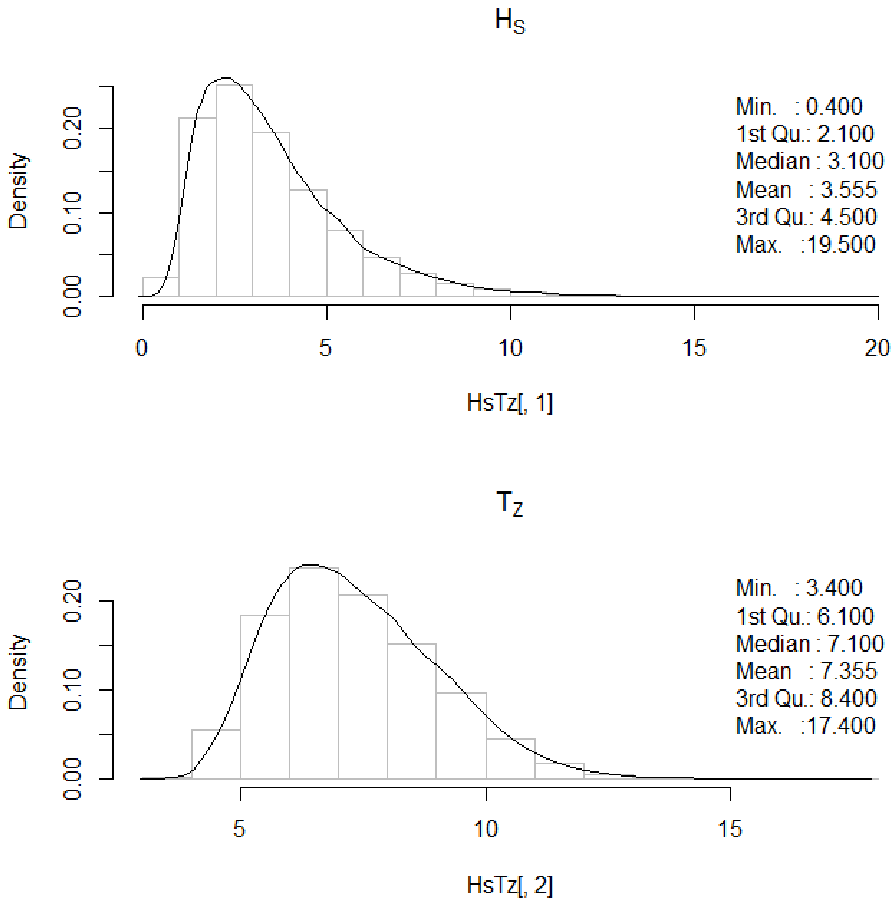
Copula model	AIC
Gumbel-Ind	-764.1
Clayton-Ind	-804.5
Frank-Ind	-850.9
Normal-Ind	-889.8
Gumbel-Gumbel	-817.5
Clayton-Clayton	-929.4
Frank-Frank	*-953.1*
Normal-Normal	-882.6
Galambos-Galambos	-848.7
HR-HR	-759.3
Plackett-Plackett	-945.8

Exercise 2: Fit a bivariate model to data for H_S and T_Z

- Read in the data. Explore the data and make some illustrative plots
  - Marginal kernel density plots; scatterplot; summary statistics, …
- Fit marginal models to each variable and construct a joint model assuming independence
  - 3-parameter Weibull for Hs and log-normal for T_Z
- Fit a conditional model to the data according to the recommended model from DNV-RP-C205
  - Marginal 3-parameter Weibull for H_S and conditional log-normal model for T_Z
- Fit a joint model using copula and the marginal models fitted above
  - Explore different parametric copulas; Normal, Gumbel, Frank, Clayton, ….
  - You may use the R library *copula* for this
- Visualise the fit and compare results. Comment
  - AIC for model comparison

Solution – Exercise 2: Exploratory plots

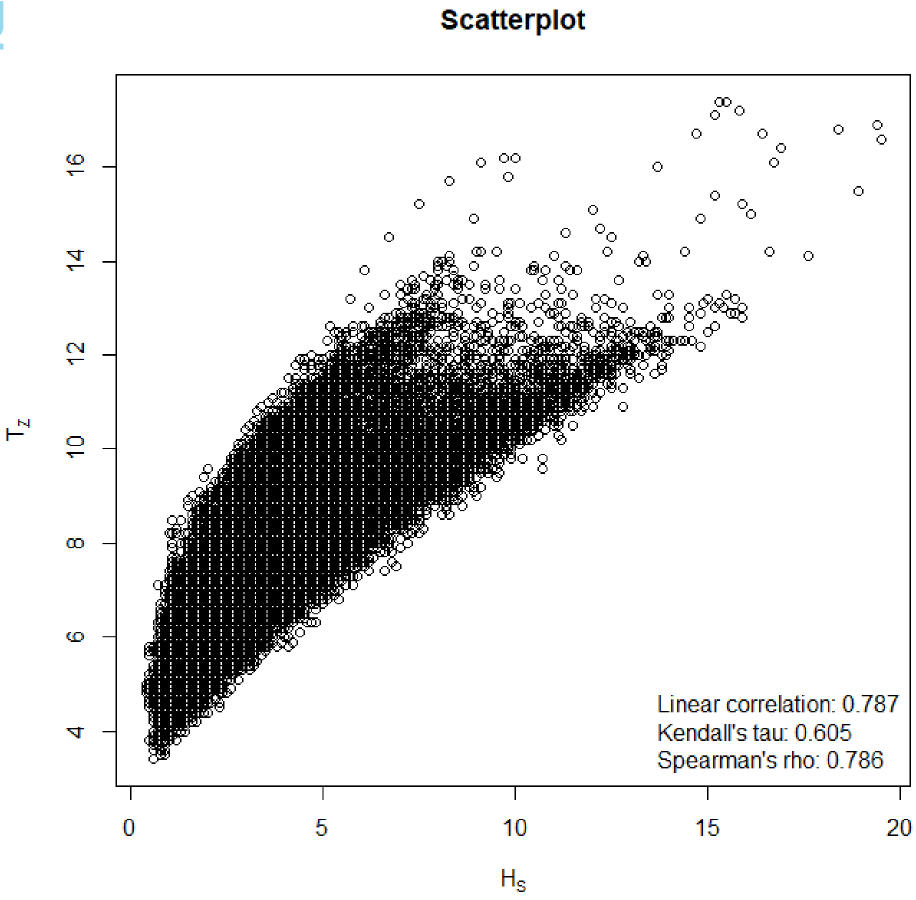
M



```
Min.   : 0.400
1st Qu.: 2.100
Median : 3.100
Mean   : 3.555
3rd Qu.: 4.500
Max.   :19.500
```

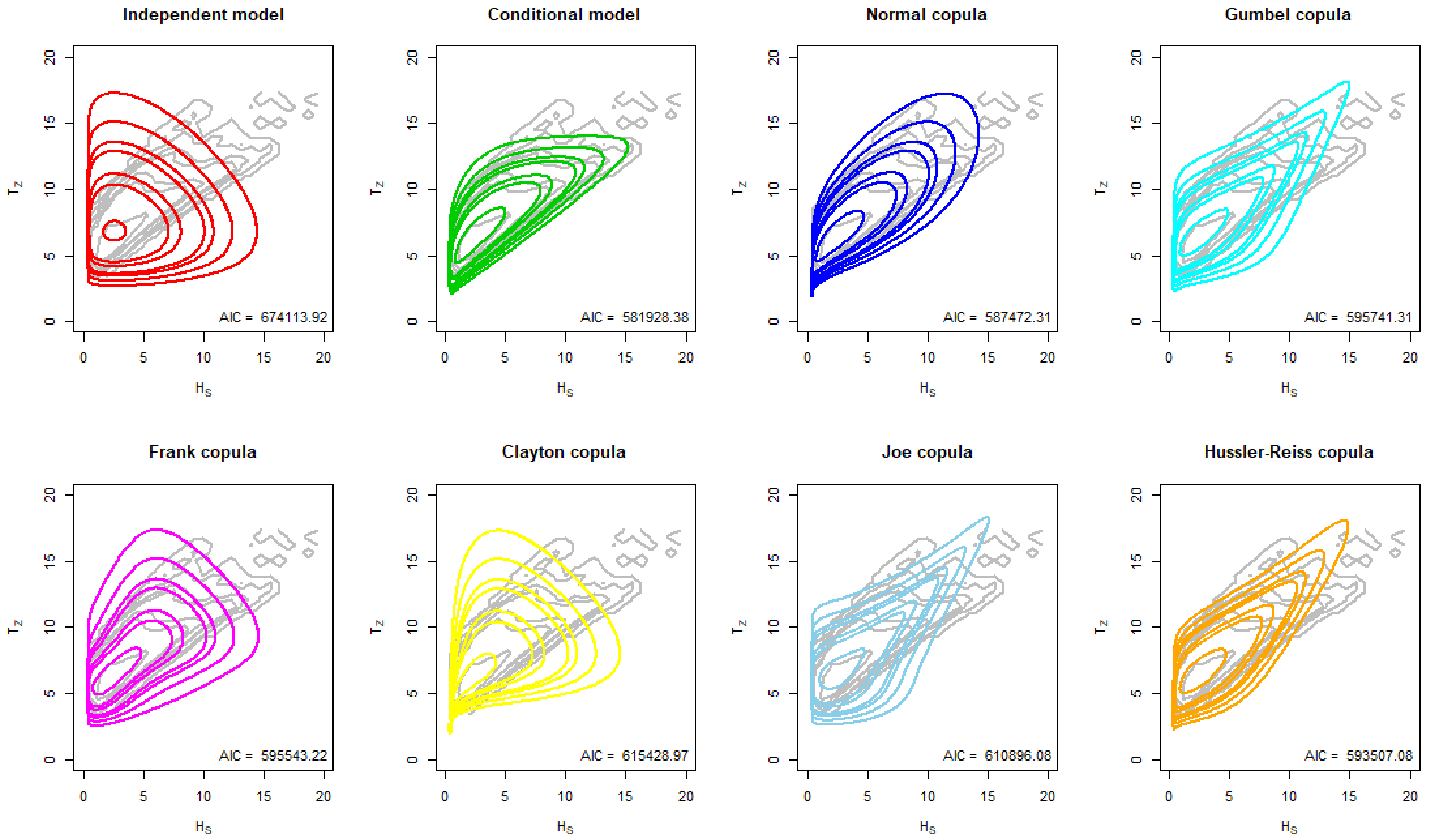
```
Min.   : 3.400
1st Qu.: 6.100
Median : 7.100
Mean   : 7.355
3rd Qu.: 8.400
Max.   :17.400
```

J



```
Linear correlation: 0.787
Kendall's tau: 0.605
Spearman's rho: 0.786
```

Solution – Exercise 2: Joint models



Summary: Univariate probability distributions

- ***Statistical inference*** draws conclusions about a ***population*** based on a random sample
- Typically, this involves fitting a ***probability distribution*** to data
- This may then be used for ***prediction*** of future values and estimation of ***extreme values***

- Different ***parametric*** probability distributions exist
- These may be fitted to data using different ***fitting techniques***

- Results are ***sensitive*** to the chosen distribution and the method used to fit it to data
- Model ***checking***, ***validation*** and ***selection*** is important to have a useful model

Summary: multivariate joint distributions

- Needed in may situations for marine design
  - Not accounting for the dependence between environmental variables may give wrong results
- ***Parametric models*** may be needed for extrapolation
  - Several alternative modelling techniques exist, including conditional modelling approach and using copula
- Finding a good model is increasingly difficult in higher dimensions
  - Challenging even in the 2-dimensional case

- Recommended practices and guidelines exist to suggest which types of models can be used
  - E.g. DNV-RP-C205
  - But possibly other models could be even better

Lecture 2: Extreme value analysis

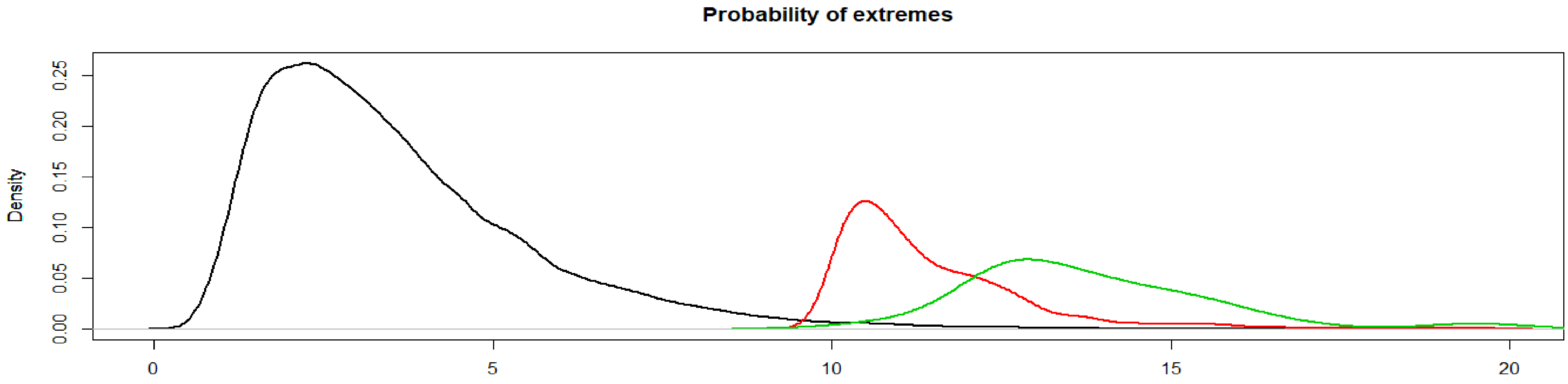
Learning goals

- Be able to perform extreme value analysis on a set of data

- Be familiar with the different approaches to extreme value analysis
  - Use all data, peaks and block maxima
  - Understand the pros and cons of the different approaches

- Be able to estimate return values associated with long return periods from a finite dataset

- Be familiar with some methods for multivariate extreme value analysis

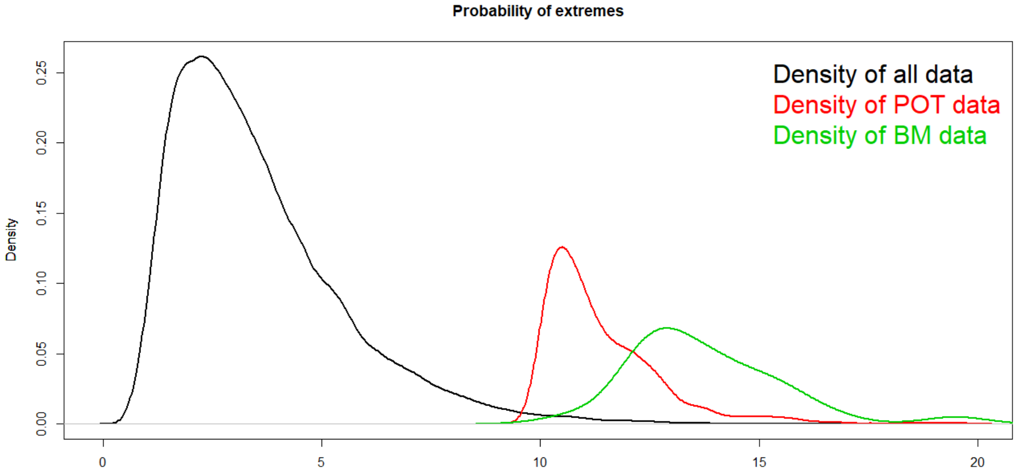


Univariate extreme value analysis

Introduction and motivation

- ***Extreme environmental conditions*** typically leads to extreme structural responses
- These are of most importance for structural design and reliability analysis
- Hence, analysis of the extremes is important → Extreme value analysis

- Extreme value analysis focuses on the far tails of the distribution and makes inference of return values corresponding to long return periods
  - High quantiles of the distribution
  - Often ***beyond the support of data***
  - The quantiles have their own distributions
    - ***Extreme value distributions***



Basic concepts: Return period and return values

- ***Return period*** T is the recurrence interval between events
  - Average time between events
  - Reciprocal of event frequency (or probability)

$$\frac{1}{T} = P(X \geq x_T)$$

- The associated ***Return value*** X_T is the value that is exceeded on average once per return period
  - Corresponds to a particular ***quantile*** of the distribution function, the $\left(1 - \frac{1}{T}\right)$-quantile

$$P(X \geq x_T) = \int_{x_T}^{\infty} f(x)dx \; = 1 - \int_{-\infty}^{x_T} f(x)dx = \frac{1}{T} \Rightarrow x_T = \eta_{1-\frac{1}{T}}$$

<u>**10-year extreme**</u>

3-hourly data: $x_T = \eta_{0.9999658}$

Annual max data: $x_T = \eta_{0.9}$

  - We typically want to estimate the return value associated with certain return periods
    - 10-year extremes, 100-year extremes, …

Different modelling approaches

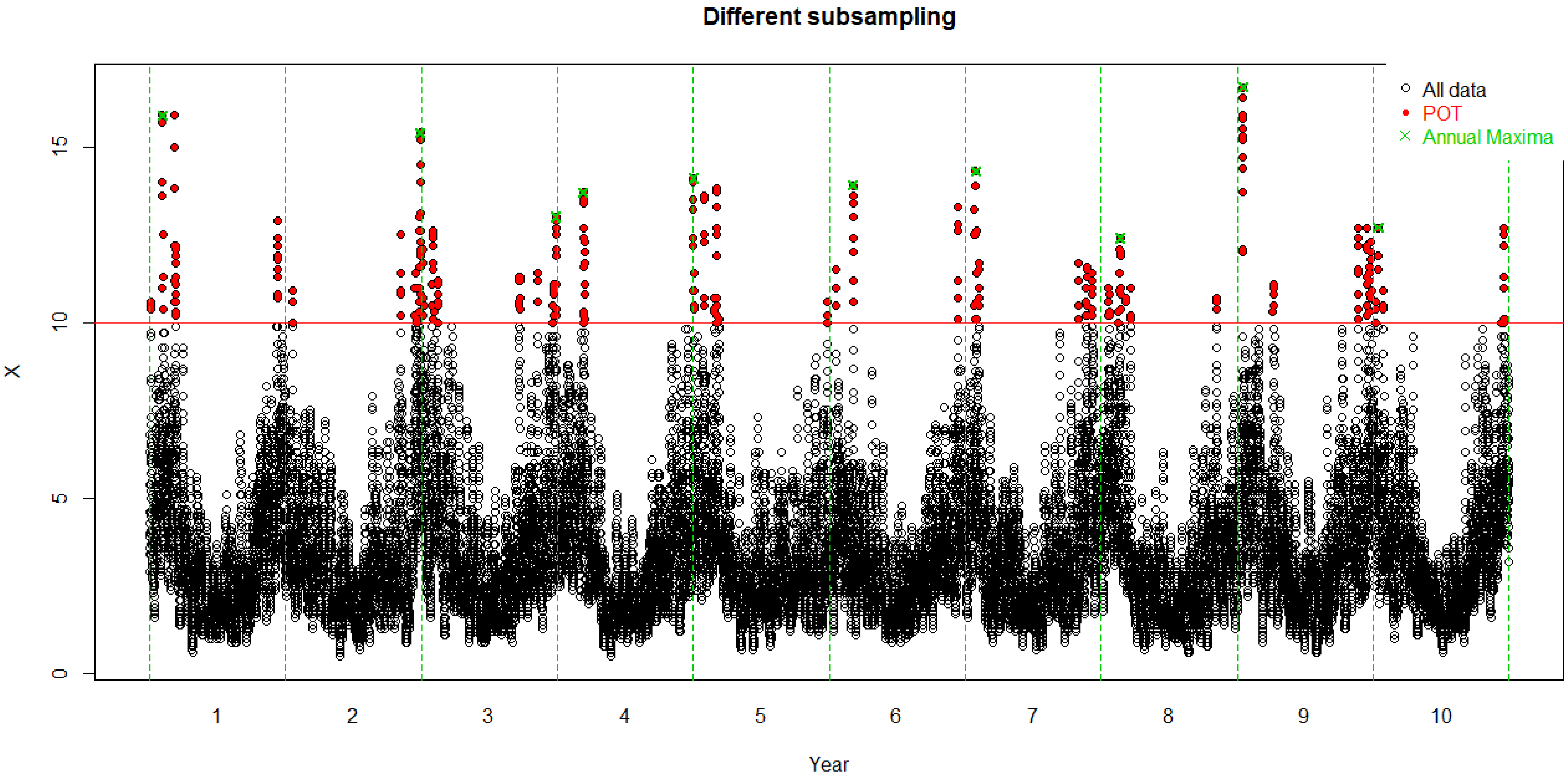
- Initial distribution approach
  - Uses ***all*** data

- Peaks over threshold (POT)
  - Uses peaks over a defined threshold – ***storm peaks*** approach

- Block maxima
  - Uses block maxima, e.g. ***annual maxima***

- Conditional exceedances approach
  - Relaxes the iid assumption

Different subsampling



Extreme value theory:
Distribution of the maximum value in a random sample

- Let $M_n = \max\{X_1, X_2, \dots, X_N\}$ where $X \sim^{iid} F(x)$
- Then, the distribution of M_n can be found from order statistics:

$$F_M(z) = P(M_n \le z) = P(X_1 \le z, \dots, X_n \le z) = P(X_1 \le z) \cdots P(X_n \le z) = \ (F_X(z)^n)$$

  - This is the distribution of the ***highest-order statistic*** of the sample

- However, when F_X is unknown, small estimation errors in F_X may give large errors in $F_M = F_X^n$
- May want to establish probabilistic models for F_M estimated from extreme data only

Extreme value theory: Generalized Extreme Value (GEV) distribution

- Let $M_n = \max\{X_1, X_2, \dots, X_N\}$ where $X \sim^{iid} F(x)$

Theorem:

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \rightarrow G(z) \text{ as } n \rightarrow \infty$$

for a non-degenerate distribution function G, then G is a member of the ***GEV*** family

$$G(z) = \exp\left\{-\left[1 - \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

for $z > \mu - {}^{\sigma}/_{\xi}$, with $\sigma > 0 \text{ and } -\infty < \mu, \xi < \infty$

Note that, if$Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \approx G(z)$, for n large enough, then $Pr\{M_n \leq z\} \approx G\left(\frac{z-b_n}{a_n}\right) = G^*(z)$ where $G^*(z)$ is another member of the GEV family.

Three types of extreme value models

- The Generalized Extreme value distribution has three special cases, depending on the value of the ***shape parameter, ξ***

Type I: Gumbel distribution ($\xi = 0$)

$$\boldsymbol{G(z)} = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\} \quad -\infty < z < \infty$$

- ***Type II: Fréchet distribution ($\xi > 0$)***

$$\boldsymbol{G(z)} = \exp\left\{-\left(\frac{z-\mu}{\sigma}\right)^{-\xi}\right\} \quad z > \mu$$

Type III: (reversed) Weibull distribution ($\xi < 0$)

$$\boldsymbol{G(z)} = \exp\left\{-\left[-\left(\frac{z-\mu}{\sigma}\right)^{-\xi}\right]\right\} \quad z < \mu$$

Extreme value analysis: GEV distribution in practice

- Interpreting the limit above as an approximation for large values of n, then the ***GEV*** family can be used for modelling the distribution of maxima of long sequences
- ***GEV*** model may be reasonable if the ***block maxima*** are ***iid***, even if the initial X_i‘s are not
- For data of independent observations, block data into sequences of length n for some large value of n to generate block maxima, $M_{n,1}, M_{n,2}, \dots M_{n,m}$ and fit the ***GEV*** model to these
- May assume any of the special cases if this can be validated by the data (or prior knowledge)
  - Often, assume $\xi = 0$ and use the 2-parameter ***Gumbel*** model
  - Less parameters give more robust estimation, but may introduce bias if the assumption is wrong
  - Statistical tests for $\xi\ =\ 0$ may be performed to check the assumption

Extreme value theory: Threshold models

- If $X_1, X_2, \ldots$ are ***iid*** with common distribution function F and that the extreme value theorem is satisfied, then, for large enough u, the distribution function of $(X - u)$ conditional on $X > u$ is approximately the ***Generalized Pareto distribution*** (GDP) with cdf

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma + \xi(u - \mu)}\right)^{-\frac{1}{\xi}}$$

- If block maxima have approximating distribution ***GEV***, then threshold excesses have corresponding approximating distribution ***GPD***
- Moreover, the parameters of the GPD are ***uniquely determined*** by those of the associated GEV distribution
  - Parameter ξ is the same in both distributions

Extreme value analysis:
Threshold models in practice

- ***GPD*** model may be reasonable if the ***threshold excesses*** are ***iid***, even if the initial X_i's are not
  - But may need to perform de-clustering

- GPD model must be accompanied by a model for the ***probability of exceeding*** the threshold
  - May often assume a Poisson model for this

- Return value estimation by combining model for threshold exceedance rate with the conditional model for the threshold exceedances
  - Both models depend on the threshold value

- Special case: $\xi = 0$ gives the ***exponential distribution***

Initial distribution approach

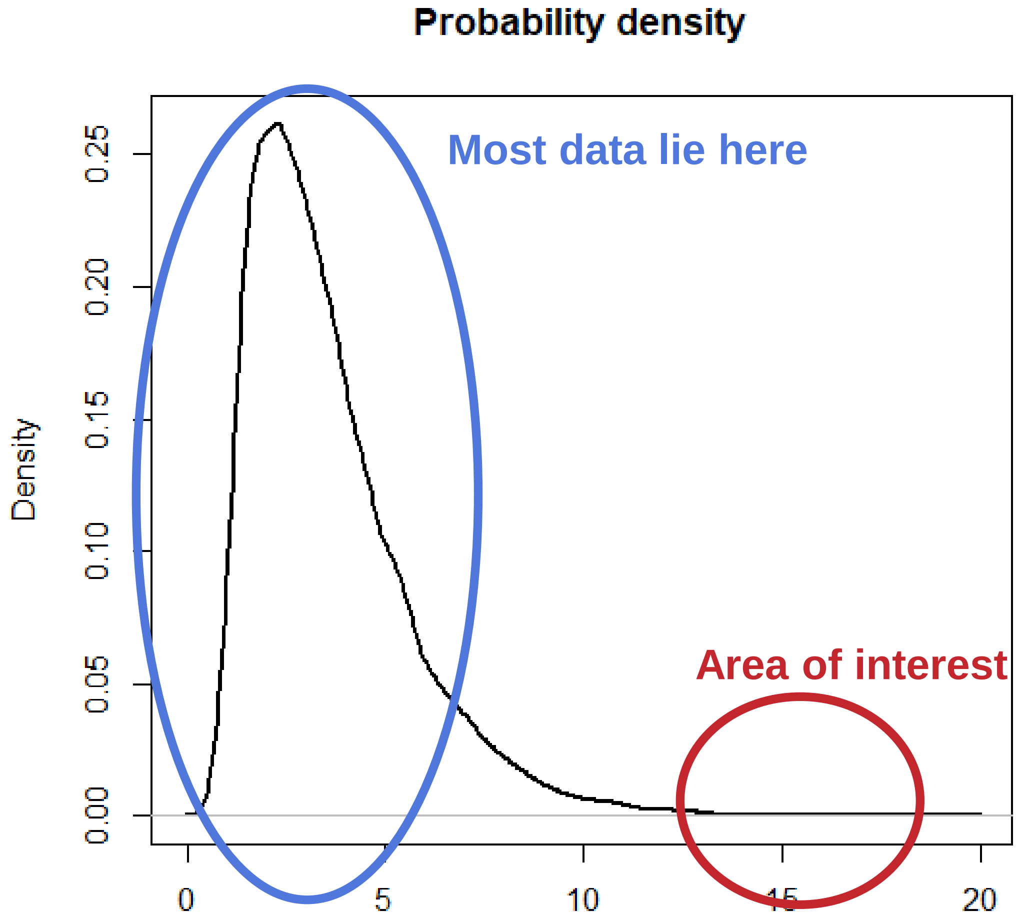
- Use ***all data*** to estimate extreme quantiles of the distribution

Pros:

- More data -> less variance
- More robust estimation and less affected by sampling variability

Cons:

- Data not iid –> may introduce bias
- Dependence in the data effectively reduces information content
- More data near the mode of the distribution and less weight on data from the tails
- Prone to mis-specification of the underlying probability distribution



EVA by initial distribution approach

- Fit the data to a parametric distribution function as before and estimate the desired return values
  1. Choose a parametric distribution function
  2. Estimate the parameters
  3. Check model fit
  4. Calculate the desired quantiles of the distribution

Block maxima

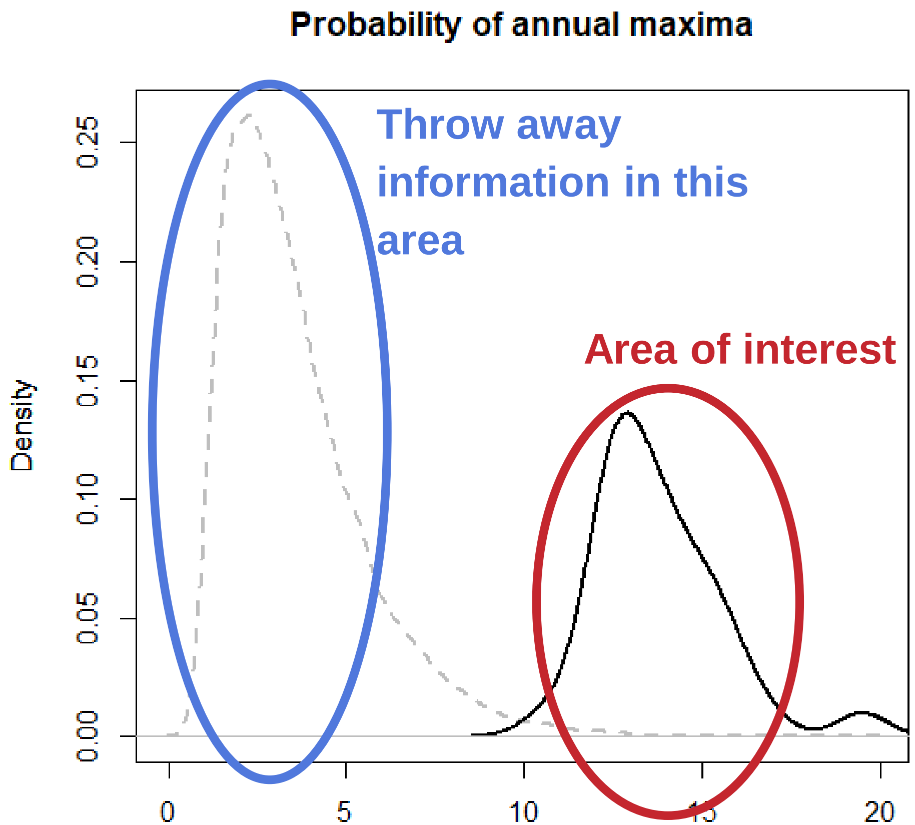
- Uses only ***block maxima***
  - Need to define ***block size*** and calculate the maximum value within each block
  - Typically, block size = 1 year → annual maxima

Pros:

- All data are extremes and carry information about the area of interest
- Asymptotic theory about the distribution of block maxima
- IID assumption more likely to be satisfied

Cons:

- Less data gives higher variance
- More sensitive to sampling variability
- Disregard information in non-extreme data



EVA using block maxima

- Fit the block maximum data to an ***extreme value distribution*** function and estimate the desired return values

1. Extract the block maxima from the raw time series
2. Choose a extreme value distribution function
    - GEV distribution or one of it’s variants
3. Estimate the parameters
4. Check model fit
5. Calculate the desired quantiles of the distribution

Peaks over threshold

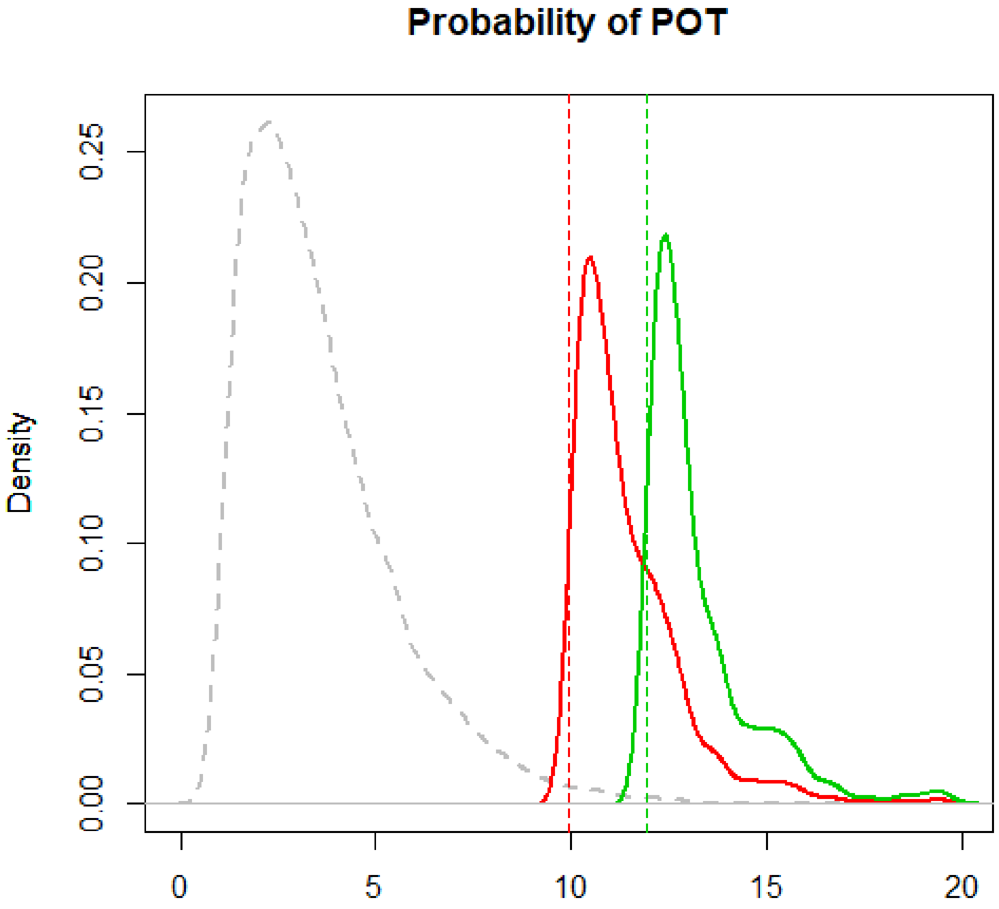
- Uses ***peaks over threshold data***
  - A compromise between all data and block maxima
  - Need to define ***threshold*** value and minimum separation distance

Pros:

- More data compared to block maxima
- IID assumption more likely to be satisfied
- Asymptotic theory about the distribution of POT
- All data carry information about the extremes

Cons:

- Results sensitive to threshold value
- Results sensitive to de-clustering
- Need also a model for the exceedance rate



EVA using POT

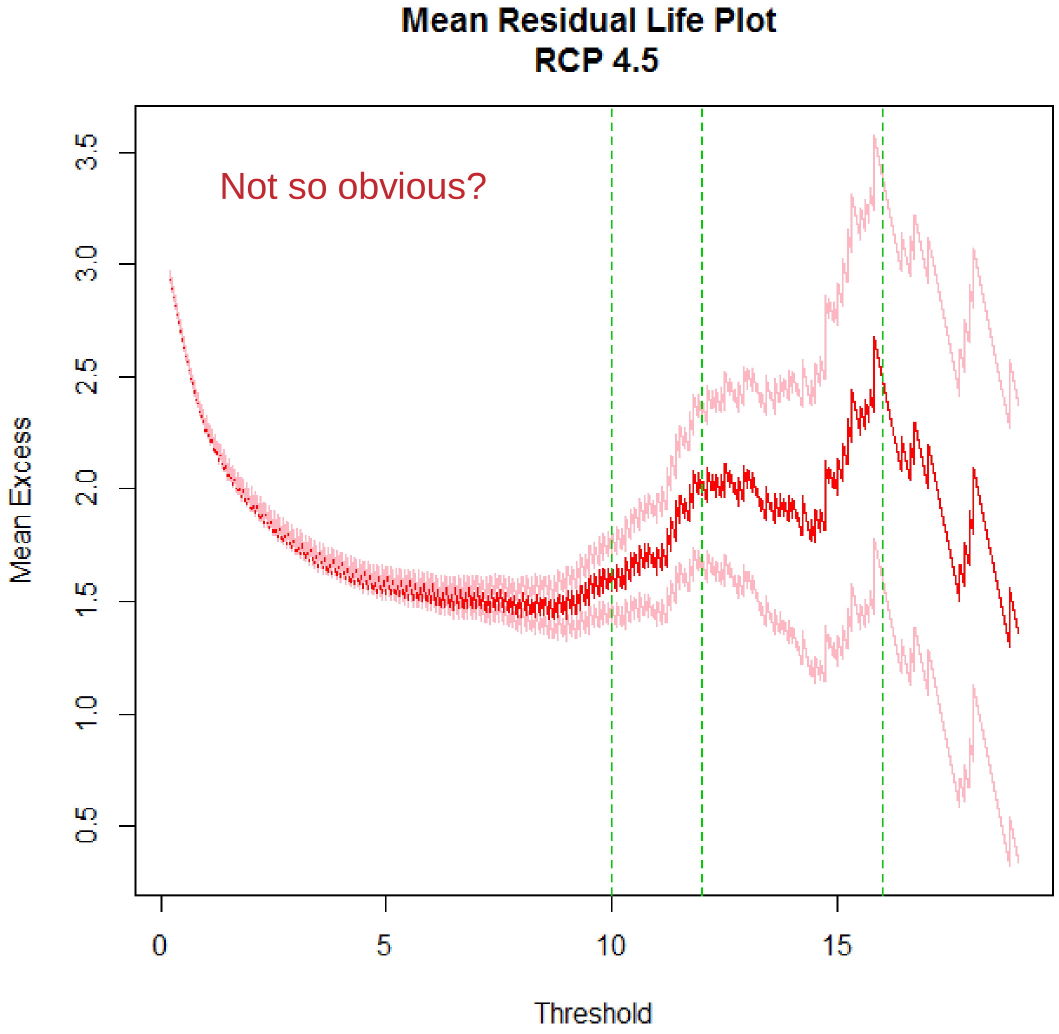
- Fit the POT data to a POT-distribution and estimate the desired return values
  1. Peak-picking
     a. Select threshold value
     b. Collect all data above the threshold
     c. Separate data into clusters and take the maximum value (peak) within each cluster
  2. Choose a POT model
     a. GPD distribution or one of it’s variants
  3. Estimate the parameters
  4. Establish a model for threshold exceedance rate
     a. Poisson model, or simply average number of exceedances per year
  5. Check model fit
  6. Calculate the desired quantiles of the distribution

Threshold selection with the POT approach

- ***Selection of threshold*** perhaps the most critical step of the POT approach
  - Results will be sensitive to this choice
  - Not straightforward
- Several tools exist in guiding the threshold selection
  - Simple approach is to use a specific empirical quantile
  - Mean residual life plot
  - Threshold choice plots/parameter stability plots
  - Dispersion index plots
- Note that different methods may suggest different thresholds
  - Not straightforward to pinpoint an exact value even for one method
- Good threshold selection involves subjective judgement and some trial and error

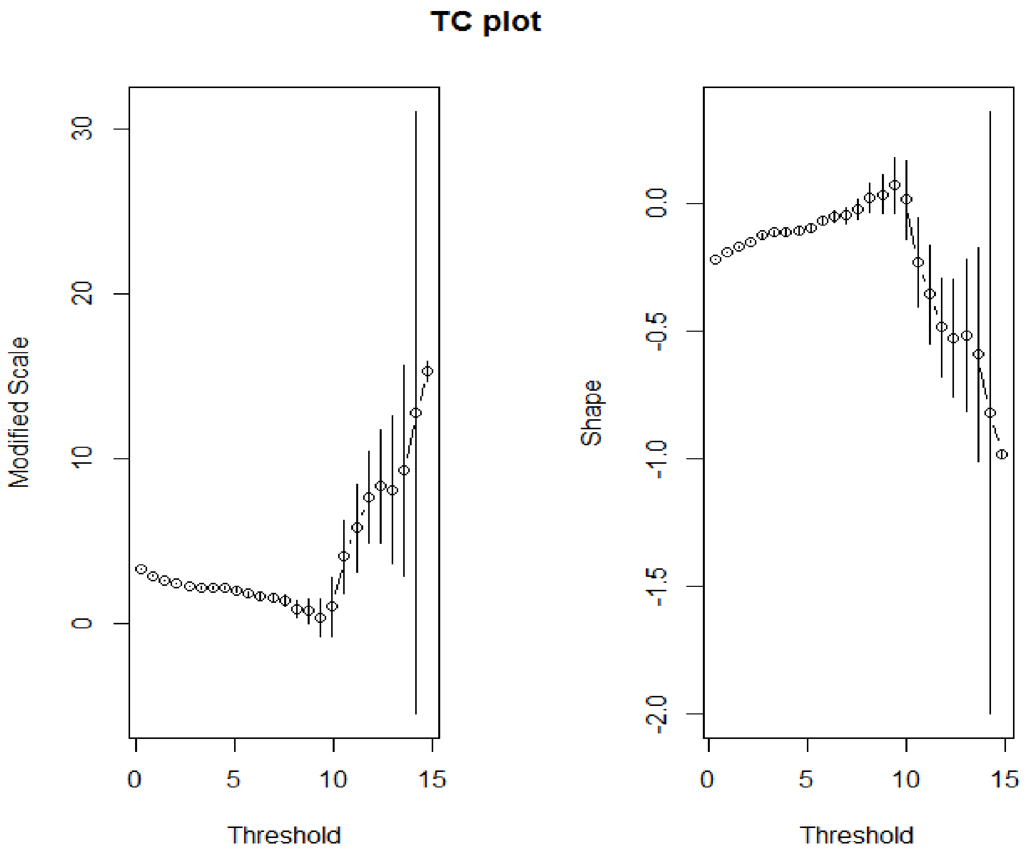
Threshold selection; Mean residual life plot

- Plot the threshold against the mean excess above that threshold
- Plots should be linear above the appropriate threshold



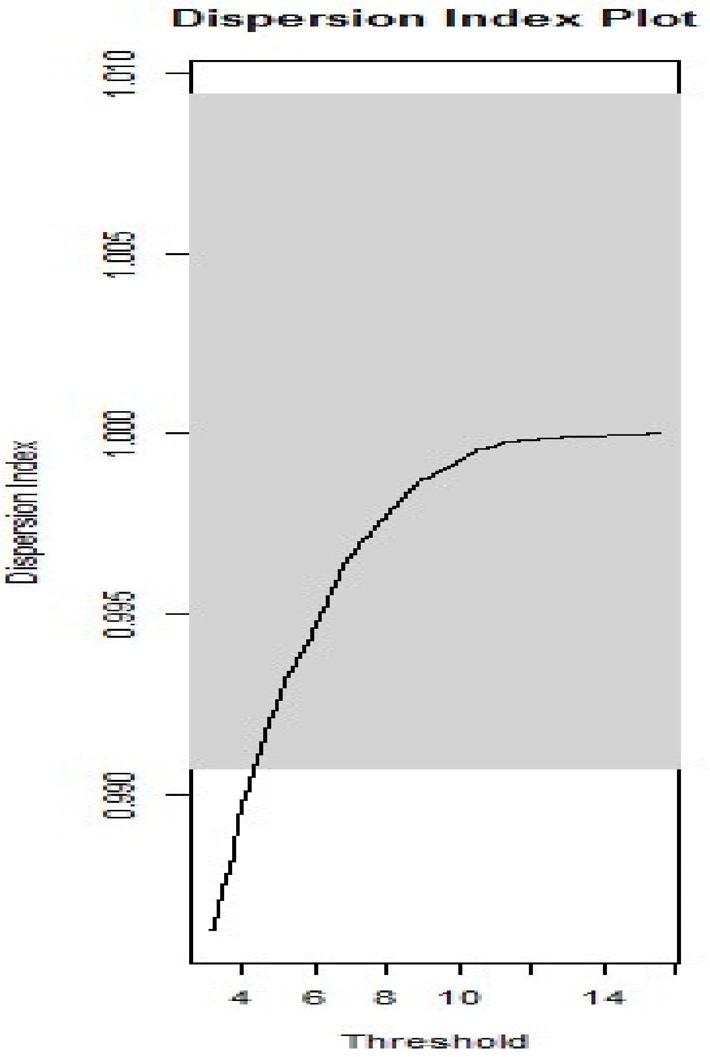
Threshold selection:
TC-plots aka Parameter stability plots

- Threshold choice plots plot modified scale and shape parameters for different thresholds
- These should be constant for any value above the correct threshold



Threshold selection: Dispersion index plot

- Dispersion index: ration of variance to the mean: $D = \frac{\sigma^2}{\mu}$
- Dispersion index plots plot the dispersion index for number of exceedance probabilities within a fixed period for different thresholds
- Poisson distribution has $D = 1$ and is often used to model threshold exceedance
  - Poisson has $\mu = \sigma^2 = \lambda$
- An appropriate threshold should correspond to a dispersion index “not too far from” 1



Modelling number of threshold exceedances

- POT model needs to be accompanied by a model for the number of threshold exceedances
- May typically assume that event (threshold exceedances) times are independent
- May then be modelled as a ***Poisson process***
    - For a Poisson process, the number of events within a time interval follows a ***Poisson distribution***
- The Poisson probability distribution, with parameter λ, for number of events, N, is

$$P(N = n) = \frac{\lambda^n}{n!} e^{-\lambda}$$

- The Poisson distribution has $E(N) = \lambda$ and $Var(N) = \lambda$
- May estimate the parameter λ from data and assume this model for annual threshold exceedances

- In practice, often use the empirical average number of exceedances per year, n_y

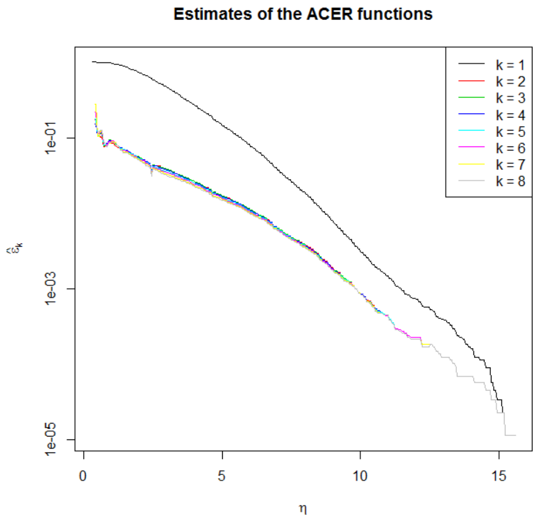
The ACER method

- Average conditional exceedance rate method
  - Relaxes the independent data assumptions
  - Conditions on previous datapoints in the time-series
- K-parameter reflects the ***(k-1)-step memory*** in the data
  - $K = 1$ corresponds to independent data
  - $K = 2$ corresponds to 1-step memory (condition on preceding value only)
  - …
- Assumes a non-linear functional form of the ACER function above a ***tail marker***

- Not as commonly used as the other methods
  - Results may be sensitive to choice of k and value of the tail marker
- May use all data, even if not independent

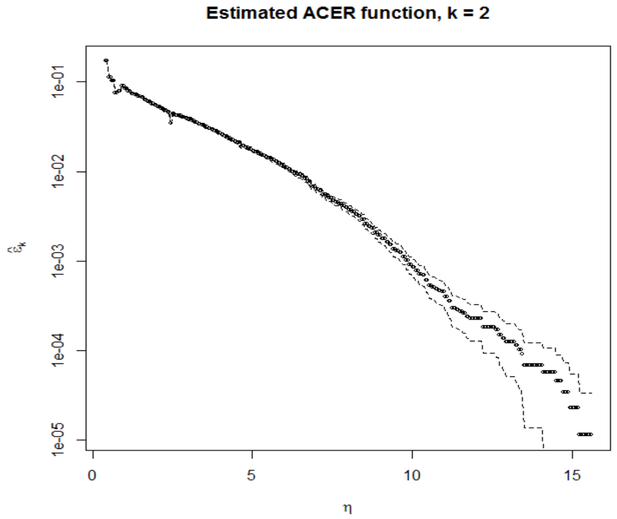
EVA using ACER

- Estimate an ACER function and estimate the desired return values
  1. Determine the value of k
  2. Calculate the empirical ACER function from the data
  3. Specify a tail marker
     a. Indication of when the tail of the distribution begins
  4. Fit a non-linear parametric function to the empirical ACER functions above the tail marker
  5. Estimate the return value from the parametric function by extrapolation

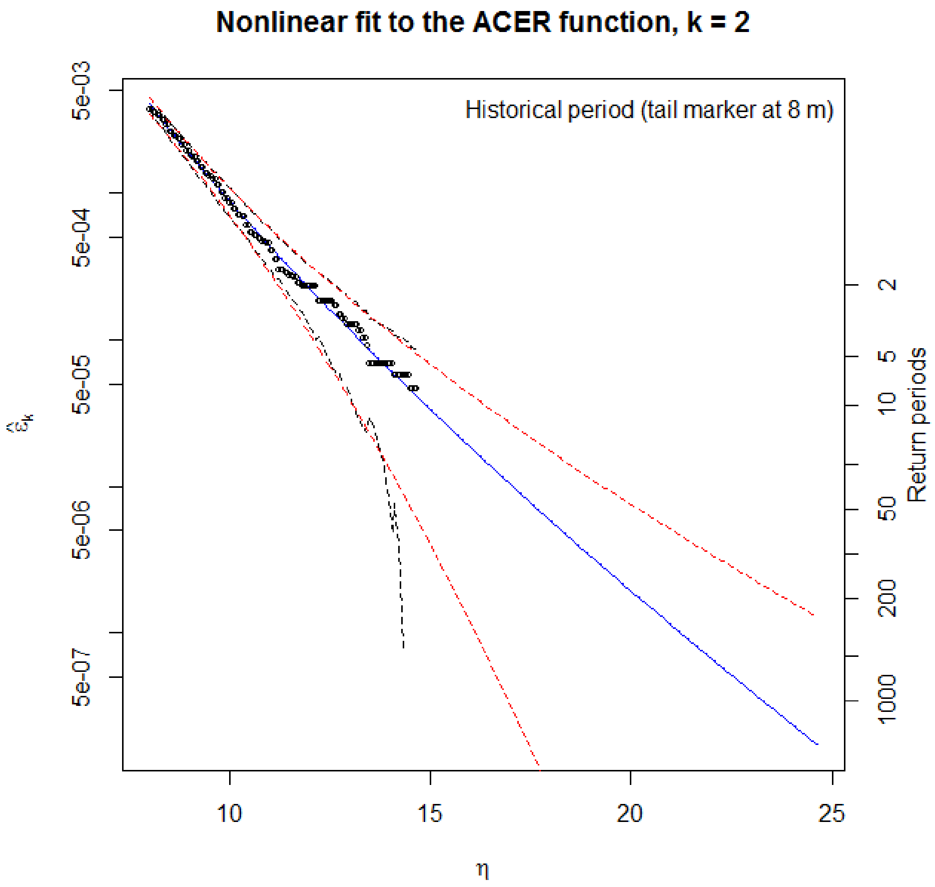


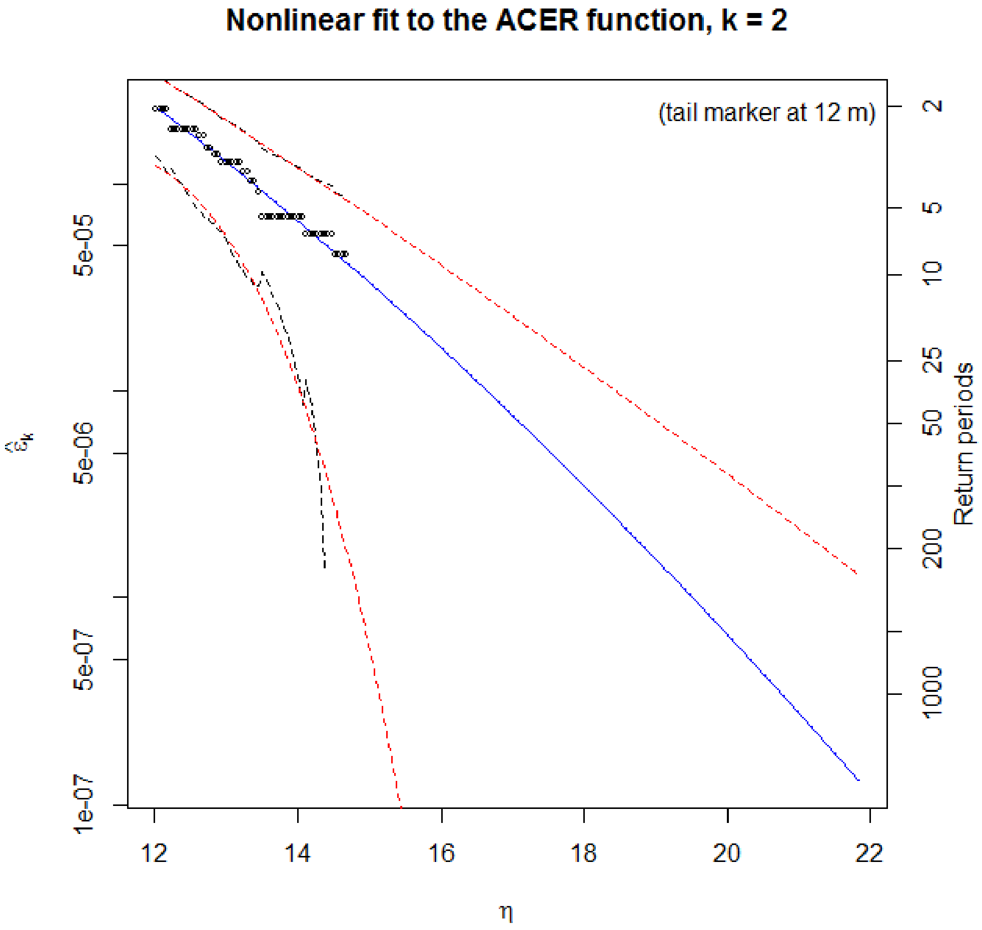
ACER functions for different values of k.

Suggests $k = 2$ is OK



Fitting non-linear ACER functions for different values of the tail marker





Exercise 3: Extreme value analysis of H_S data

1. Fit a 3-parameter Weibull distribution to the data and estimate the 10-, 20- and 100 year return values

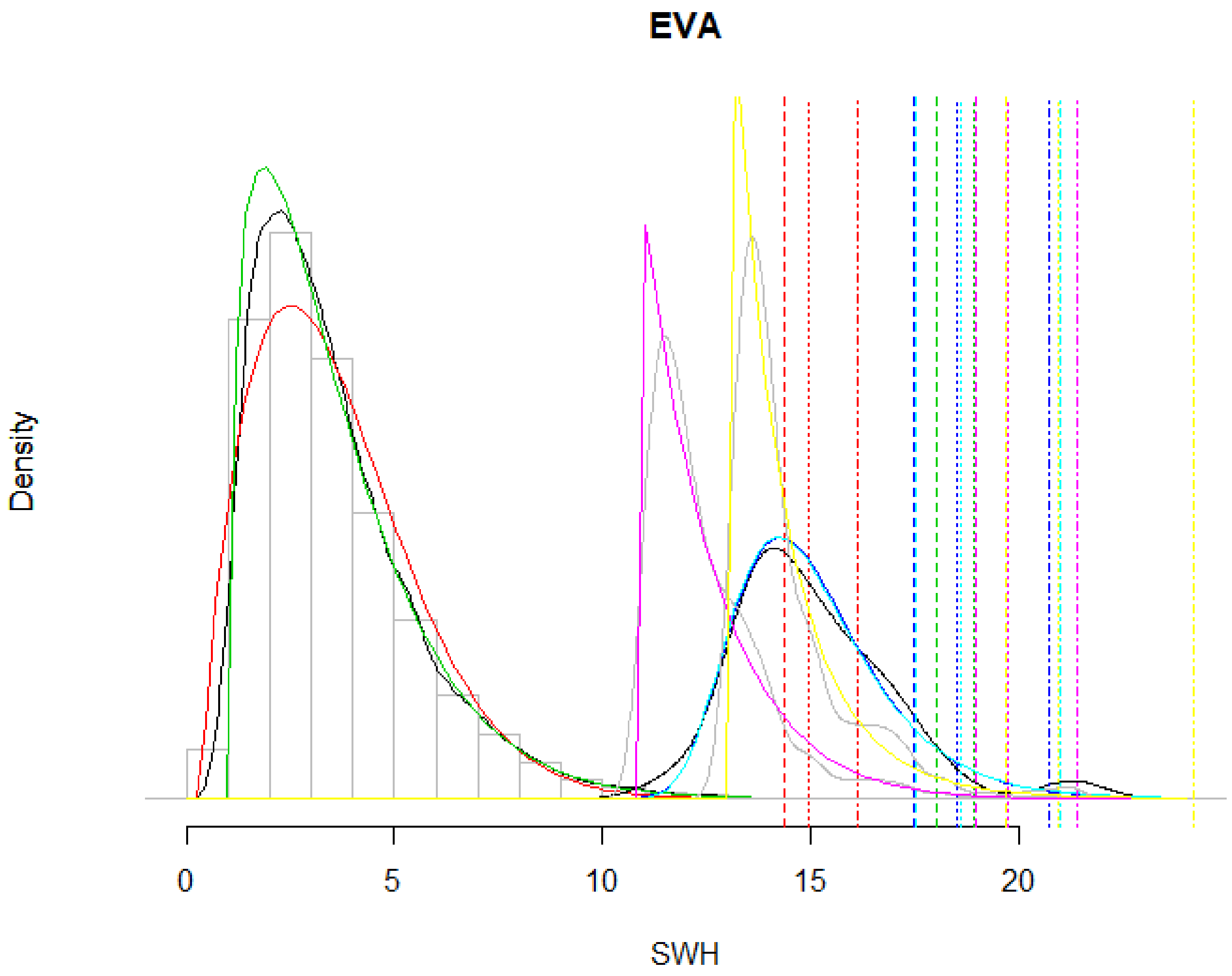
2. Extract the annual maxima from the time series and do block maximum analysis
   a. Fit a GEV model to the annual maxima and estimate the 10-, 20- and 100 year return values
   b. Fit a Gumbel model to the annual maxima and estimate the 10-, 20- and 100 year return values

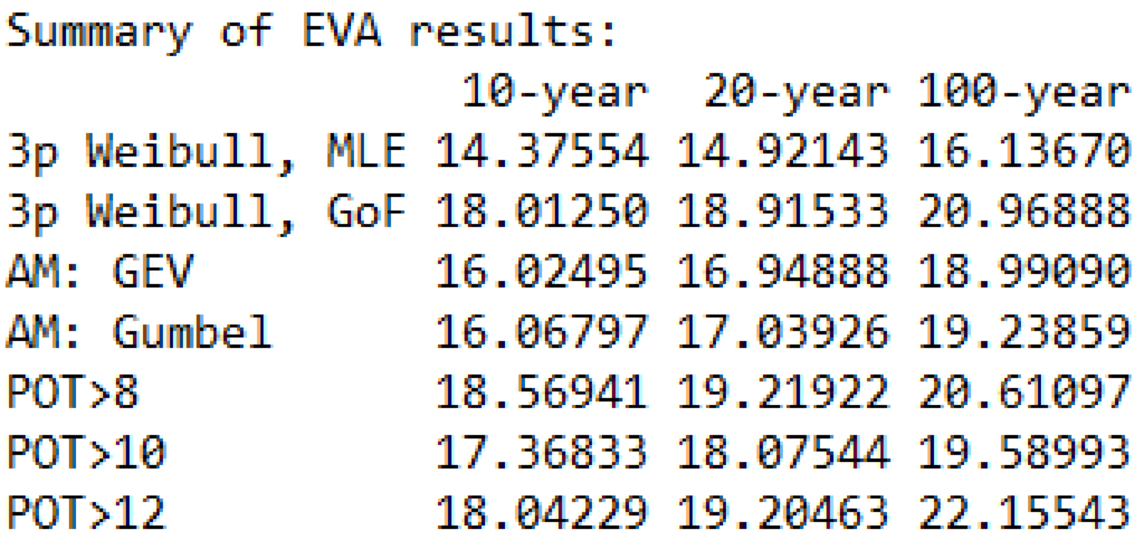
3. Extract the peaks above 8, 10 and 12 m from the time series (disregard de-clustering)
   a. Estimate the annual threshold exceedance rates from the data
   b. Fit GPD models to the POT data and estimate 10-, 20- and 100-year return values

   Hint: You may use a library in R for this, e.g. POT-package

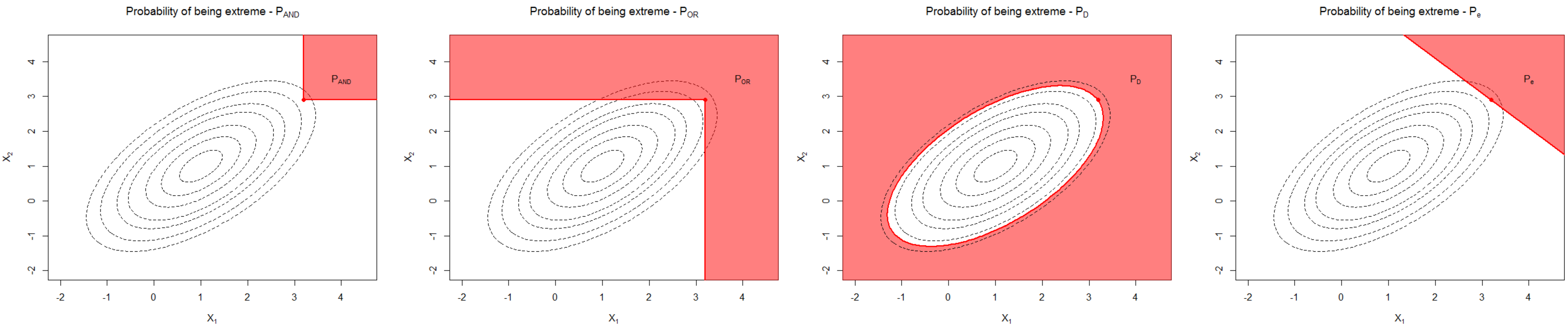
4. Compare results from the various methods and comment
   a. Which estimate would you choose?

Exercise 3: Summary of results





```
Summary of EVA results:
                 10-year  20-year 100-year
3p Weibull, MLE 14.37554 14.92143 16.13670
3p Weibull, GoF 18.01250 18.91533 20.96888
AM: GEV         16.02495 16.94888 18.99090
AM: Gumbel      16.06797 17.03926 19.23859
POT>8           18.56941 19.21922 20.61097
POT>10          17.36833 18.07544 19.58993
POT>12          18.04229 19.20463 22.15543
```



Probability of being extreme - P_{AND}

Probability of being extreme - P_{OR}

Probability of being extreme - P_D

Probability of being extreme - P_e

Multivariate extreme value analysis

Multivariate extreme value analysis

- In many situations, the ***joint extremal behaviour*** of several environmental variables is of interest
- Need methods to analyse and describe ***multivariate extremes***

- ***Environmental contour method*** is one way of doing this that is often used in structural reliability and design of offshore structures
  - Method is covered by DNV-RP-C205

- Another approach is the ***conditional extremes model***

- Examples and illustrations in 2-dimensions only, but in principle extendable to higher dimensions

What are extremes in the multivariate case?

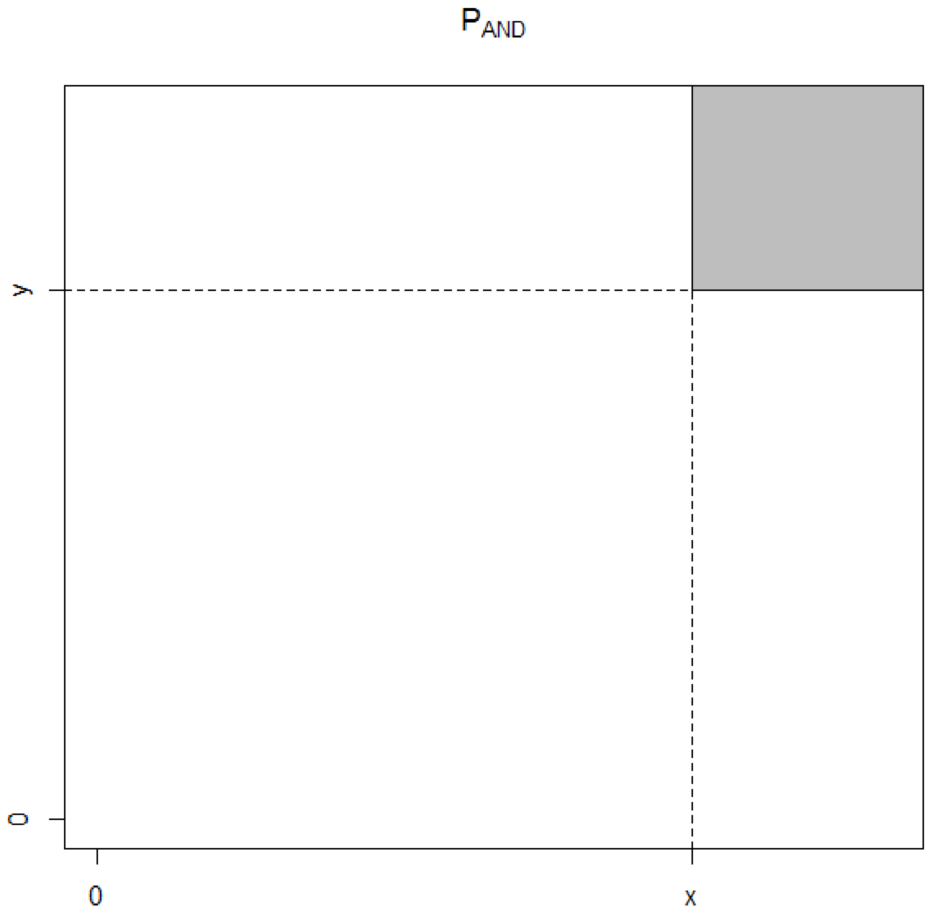
- No single definition of return value
- Even with an agreed definition – no unique solution
  - The possible solutions will typically define a ***contour*** in the parameter space

- The extremeness of an observation, will depend on the definition of return value

- Some possible definitions in the bivariate case will be presented in the following; $P_{AND}, P_{OR}, P_{Cond1}, P_{Cond2}, P_e, P_D$ ...
  - For one and the same observation from a distribution, these probabilities will be different

- Which definition is most appropriate may depend on the structural problem and can be related to the form of the limit state function

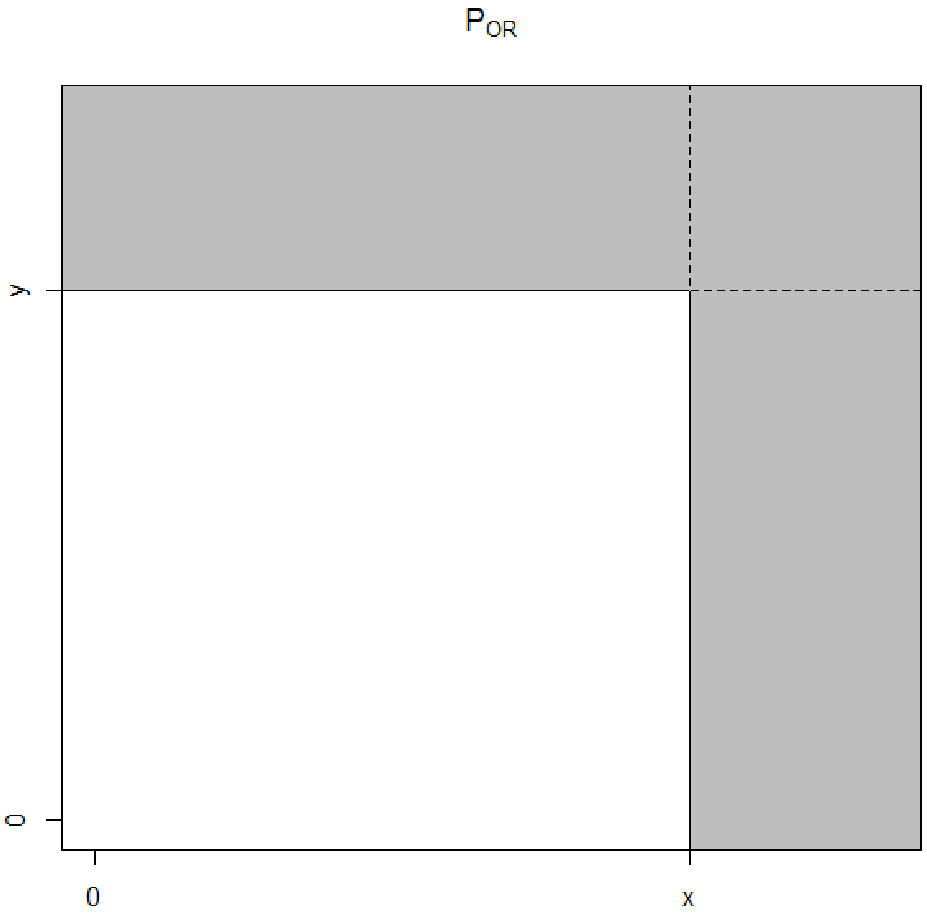
Both X and Y are extreme, $\boldsymbol{P_{AND}}$



$$X \geq x \bigcap Y \geq y$$

$$P_{AND} = P(X \geq x, Y \geq y)$$

Either X or Y are extreme, $\boldsymbol{P_{OR}}$



$$X \geq x \bigcup Y \geq y$$

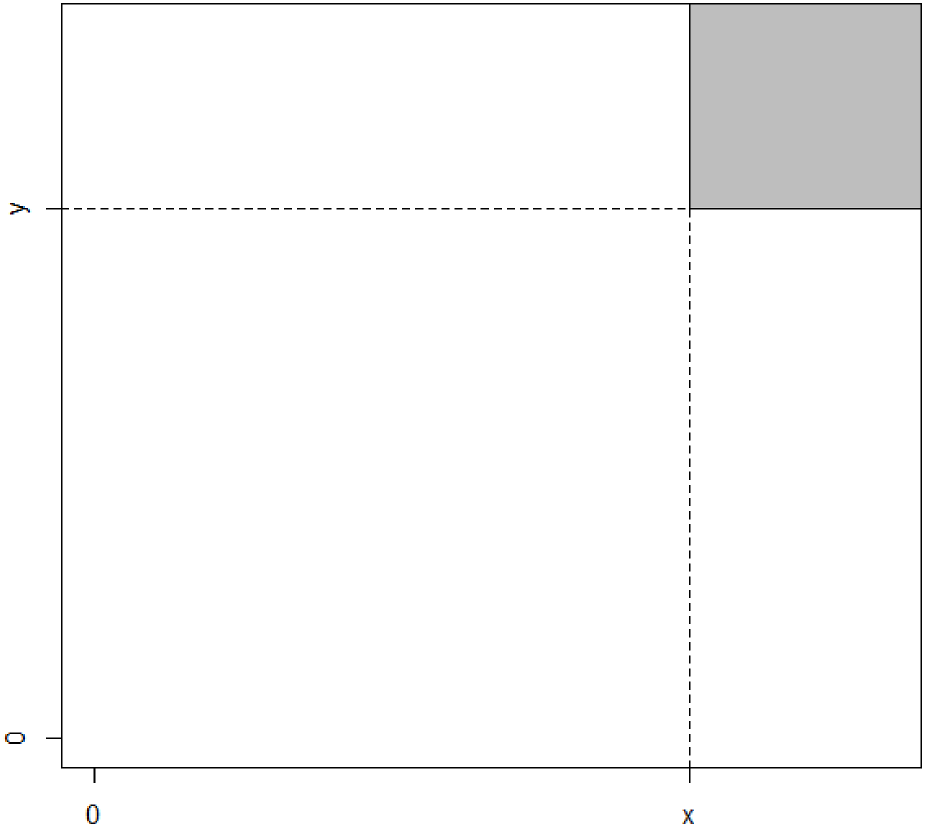
$$P_{OR} = P(X \geq x) + P(Y \geq y) - P(X \geq z, Y \geq y)$$

Y are extreme conditional on X being extreme, $\boldsymbol{P_{Cond1}}$

$$Y \geq y \,|X \geq x$$

$$P_{Cond1} = P(X \geq x)P(Y \geq y \,|X \geq \; x)$$

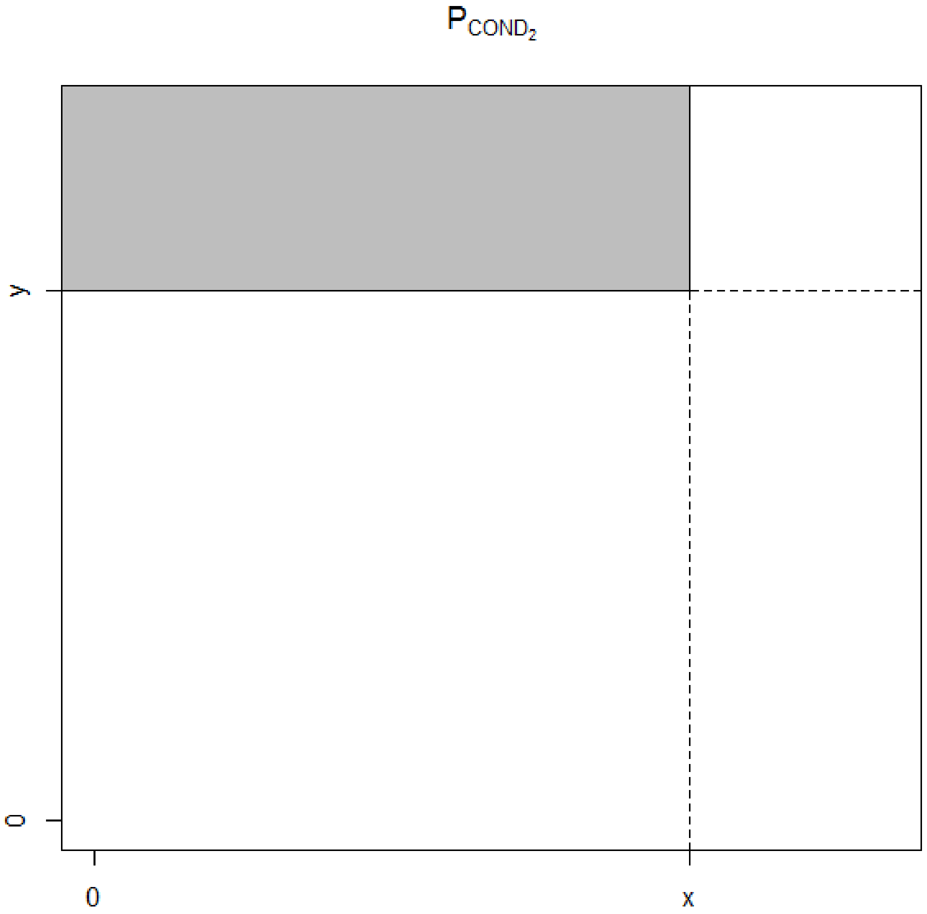
P_{COND_1}



Y are extreme conditional on X *not* being extreme, $\boldsymbol{P_{Cond2}}$

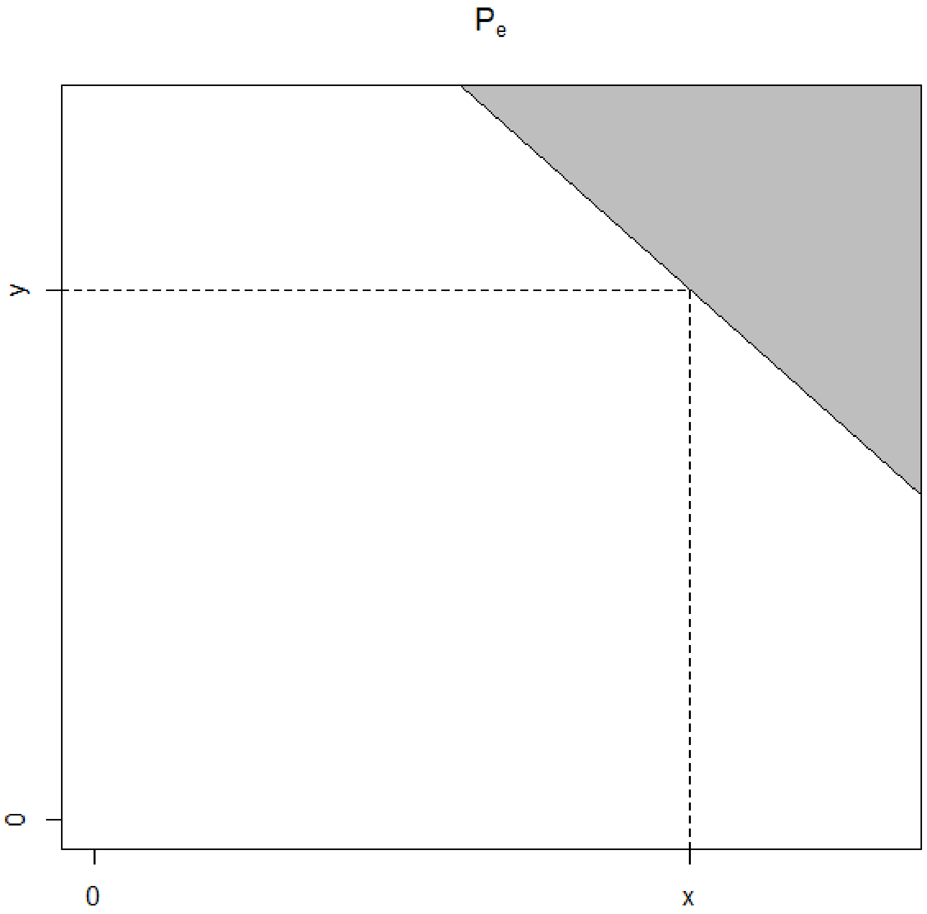
$$Y \geq y \,|X < x$$

$$P_{Cond2} = P(X \leq x)P(Y \geq y\,|X \leq \ x)$$

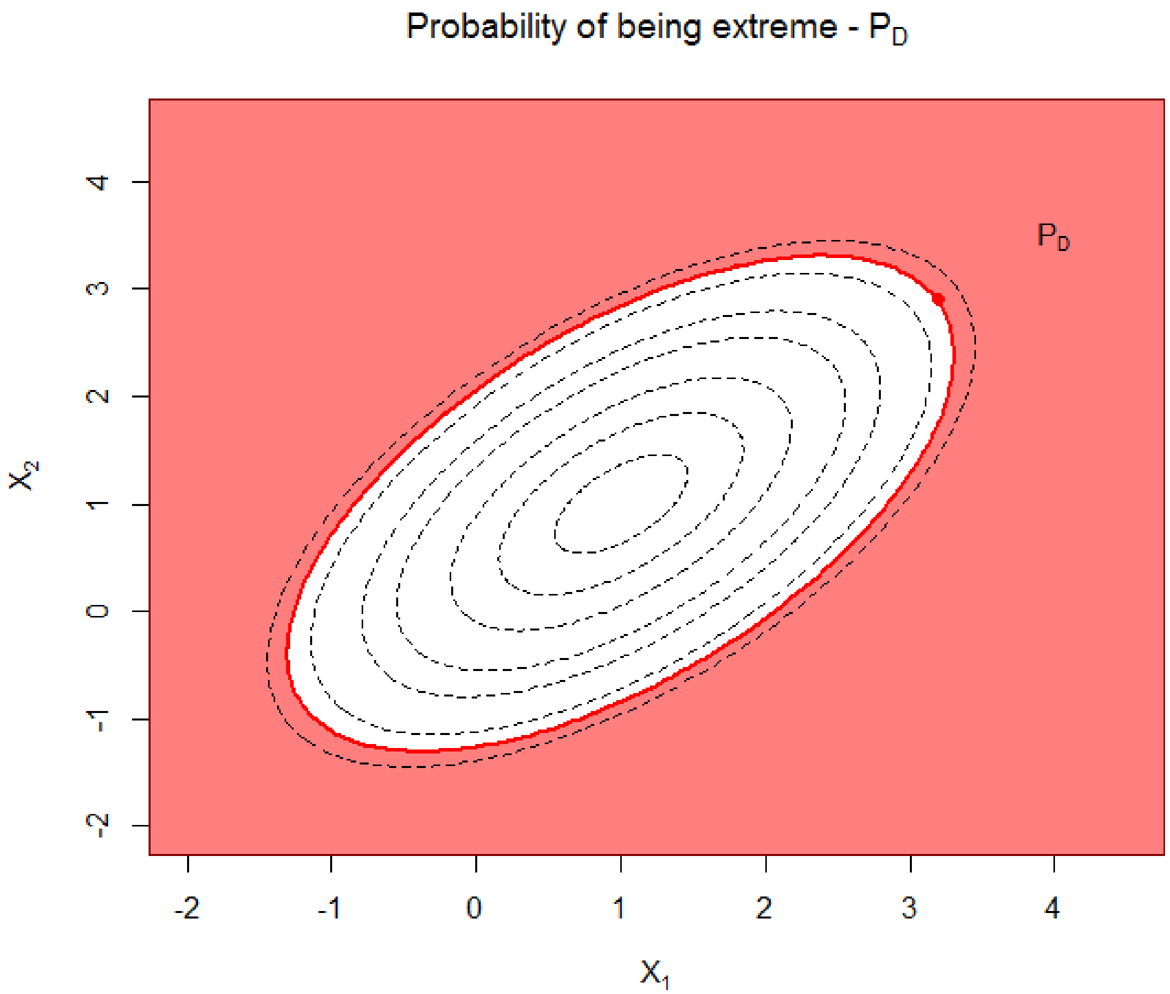


(X, Y) outside an exceedance hyperplane, $\boldsymbol{P_e}$

$$(X, Y) \in \Pi^+(\theta)$$
$$P_e = P\big((X, Y) \in \Pi^+(\theta)\big)$$



(X, Y) has extreme Mahalanobis distance, $\boldsymbol{P_D}$



$$D(\boldsymbol{x} = (x, y)) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \geq D$$

Where

- $\boldsymbol{\mu}$ is the mean vector and
- $\boldsymbol{\Sigma}$ is the sample covariance matrix

$$P_D = P(D(x, y) \geq D)$$

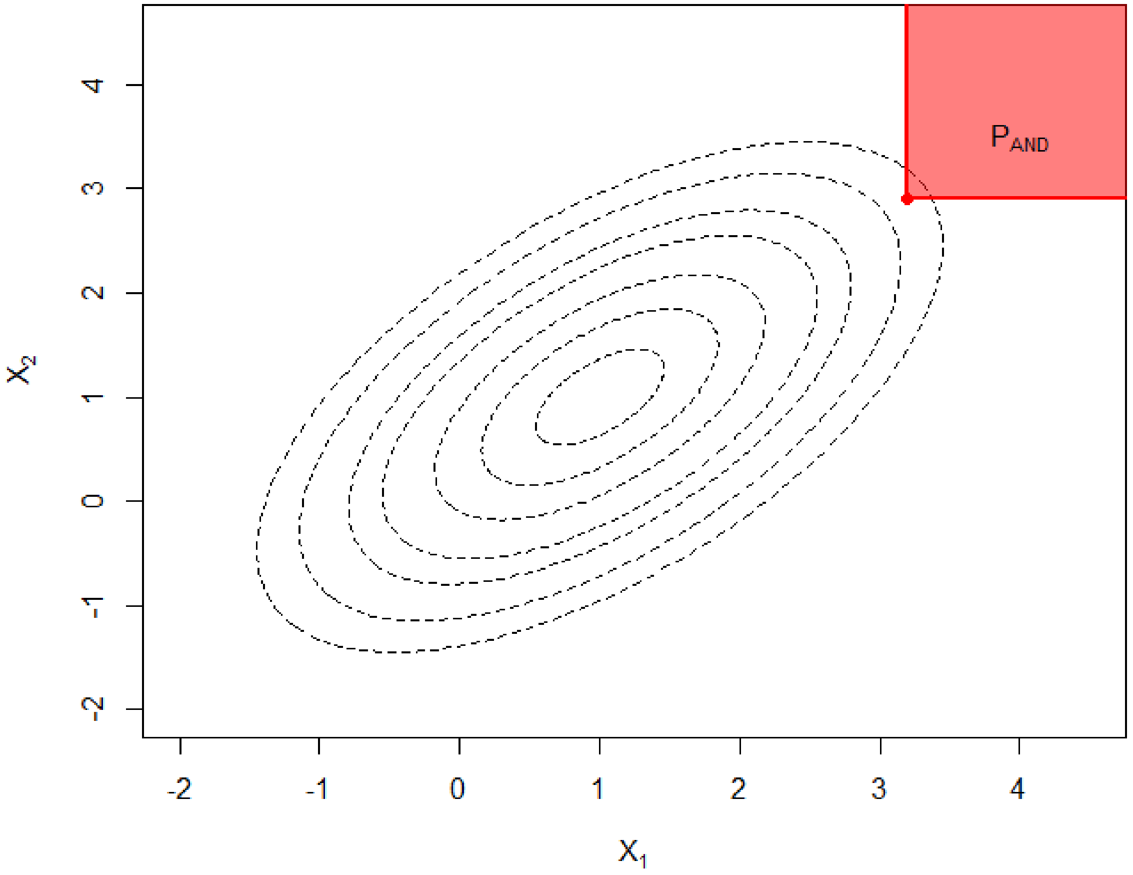
No unique solution, even with a single definition

- Assuming the P_{AND} definition one may define the T-year return value, (x_T, y_T) as follows, from annual maximum data, M_X and M_Y

$$P(M_X \geq x_T, M_Y \geq y_T) = 1 - F_{M_X,M_Y}(x_T, y_T) = \frac{1}{T}$$

- This has ***no unique solution***:
  - Given any pair (x_T, y_T) that satisfies this, one can find another pair satisfying this by e.g. increasing x_T and reducing y_T
  - A continuum of solutions satisfies this equation, and will form a contour in (x, y)-space
- The same applies to the other definitions of extremeness

Probability of being extreme - P_{AND}



Environmental contour method for describing extreme environmental conditions

- Environmental contours are used to account for extreme environmental conditions in marine design
- Inverse structural reliability problem:
  - Rather than computing the failure probability of a given design, one starts with a minimum required reliability and investigates what restrictions this impose on possible designs
- Commonly used; Included in our recommended practices and other standards, e.g. NORSOK

Probabilistic structural reliability

- Often, one is interested in the ***probability of failure***, P_f of a structure subject to stochastic environmental loads
- Define a ***performance function*** g as a function of a number of stochastic input variables, $\boldsymbol{X} = (X_1, X_2, \dots, X_n)^T$
  - Typically, g = capacity (strength) – demand (stress)
- Then,

$$\begin{aligned} g(\boldsymbol{X}) > 0&: \text{ The structure survives} \\ g(\boldsymbol{X}) < 0&: \text{ The structure fails} \end{aligned}$$

- The ***limit state function*** is defined as the boundary between safe and unsafe regions, $g(\boldsymbol{X}) = 0$
- The reliability, $\boldsymbol{R}$ of the structure is then defined as the probability of surviving

$$R = 1 - P_f = P(g(\boldsymbol{X}) > 0)$$

Probabilistic structural reliability

- Assuming that g is a function of the environmental variables $\boldsymbol{X}$ only, then

$$R = 1 - P_f = P(g(\boldsymbol{X}) > 0) = \int_{g(\boldsymbol{X})>0} f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}$$

  - Probability of failure corresponds to the probability of the environmental variables being in the unsafe region

- In practice, this integral can be difficult to evaluate
  - Both $g(X)$ and $f_X(x)$ may be complicated functions

- May transform problem to ***standard normal space*** and solve

$$R = 1 - P_f = P(\tilde{g}(\boldsymbol{U}) > 0) = \int_{\tilde{g}(\boldsymbol{U})>0} \phi(\boldsymbol{u}) d\boldsymbol{u}$$

  - $\tilde{g}(\boldsymbol{U})$: Transformed performance function

Structural reliability: FORM

Approximate solution to the reliability integral can be found by the ***First Order Reliability Method*** (FORM)

- Transform $\boldsymbol{X}$ to independent standard normal variables, $\boldsymbol{U} = (U_1, U_2, \dots, U_n)^T$
  - Using the Rosenblatt transform
- Approximate the limit state function by a first-order Taylor expansion at the ***design point***
  - Design point: Point on the limit state function closest to origin (in transformed space),

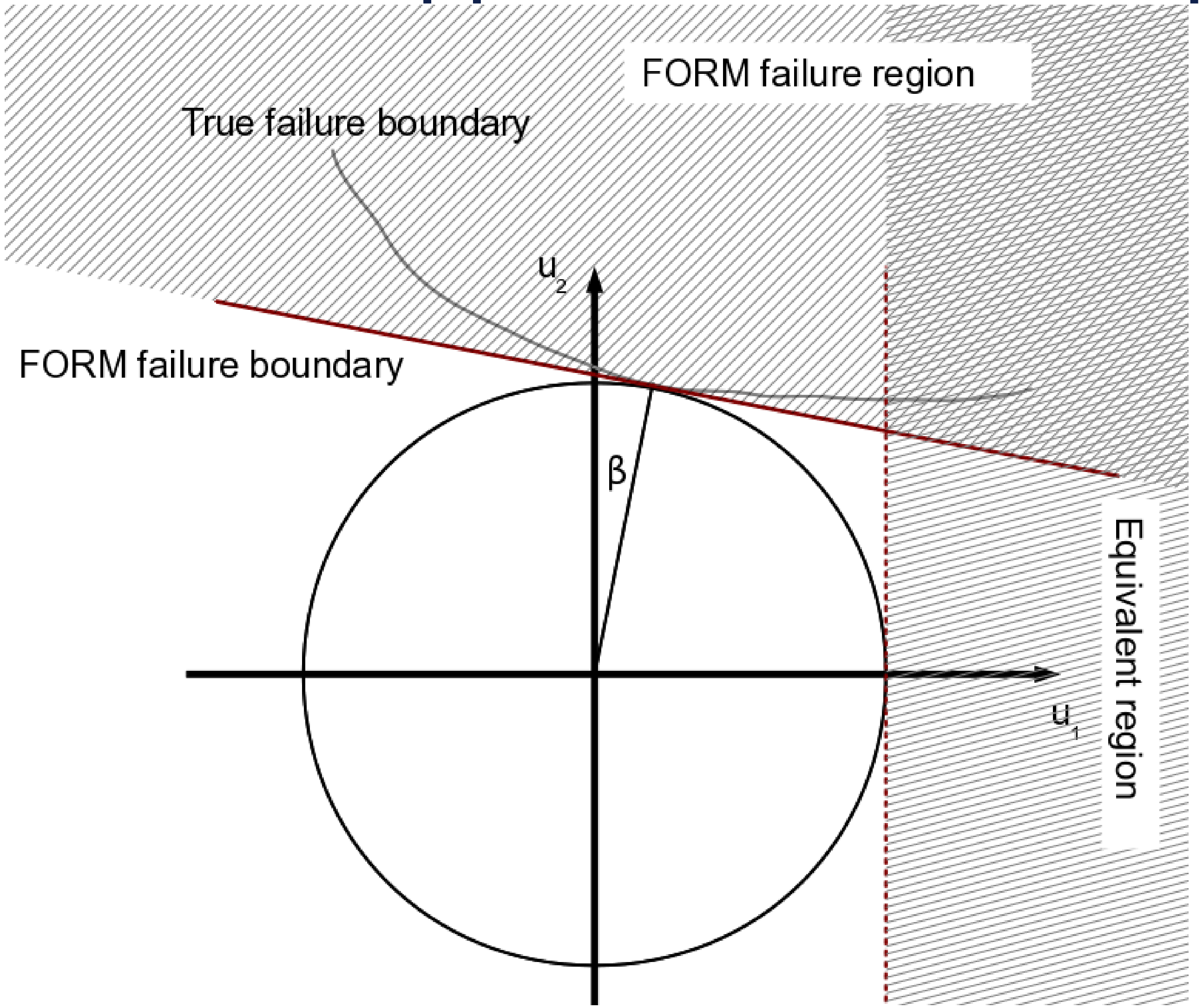
$$\boldsymbol{u_d} = \left(u_{d,1}, u_{d,2}, \dots, u_{d,n}\right)^T$$

- Calculate the ***reliability index***, β_R**,** as the distance between the design point and the origin

$$\beta_R = \sqrt{\sum_{i=1}^{n} u_{d,i}^2}$$

- Then, $P_f \approx \Phi(-\beta_R)$

The FORM approximation in U-space



Rosenblatt transform from X-space to U-space:

$$U_1 = \Phi^{-1}\left(F_{X_1}(X_1)\right)$$

$$U_2 = \Phi^{-1}\left(F_{X_2|X_1}(X_2)\right)$$

$$\vdots$$

$$U_n = \Phi^{-1}\left(F_{X_n|X_1,\cdots,X_n}(X_n)\right)$$

The inverse reliability problem – Environmental contours

- Rather than calculating the failure probability of a given design, start with a given ***target failure probability*** and explore what restrictions this imposes on possible design
  - Inverse First-Order Reliability Method: ***I-FORM***

- May construct ***contours*** corresponding to a given target reliability level in the space of the environmental input variables: ***Environmental contours***

- Environmental contours is a way of describing ***multivariate extremes*** of various environmental variables
  - De-couples the environmental description from the structural problem
  - Describes ***extreme combinations*** of the environmental variables

- ***Design sea states approach:*** Assume that the n-year extreme response can be estimated from the n-year sea-state condition

Environmental contours

- Several different definitions exist of environmental contours. DNV-RP-C205 mentions 3:
  1. I-FORM contours
  2. Iso-density contours
  3. Direct sampling contours

- Different definitions give different contours
  - All aiming at solving the same problem – which designs are safe given the target reliability level and assuming a joint probabilistic model for the environmental input variables, $f_{\boldsymbol{X}}(\boldsymbol{x})$

I-FORM contours

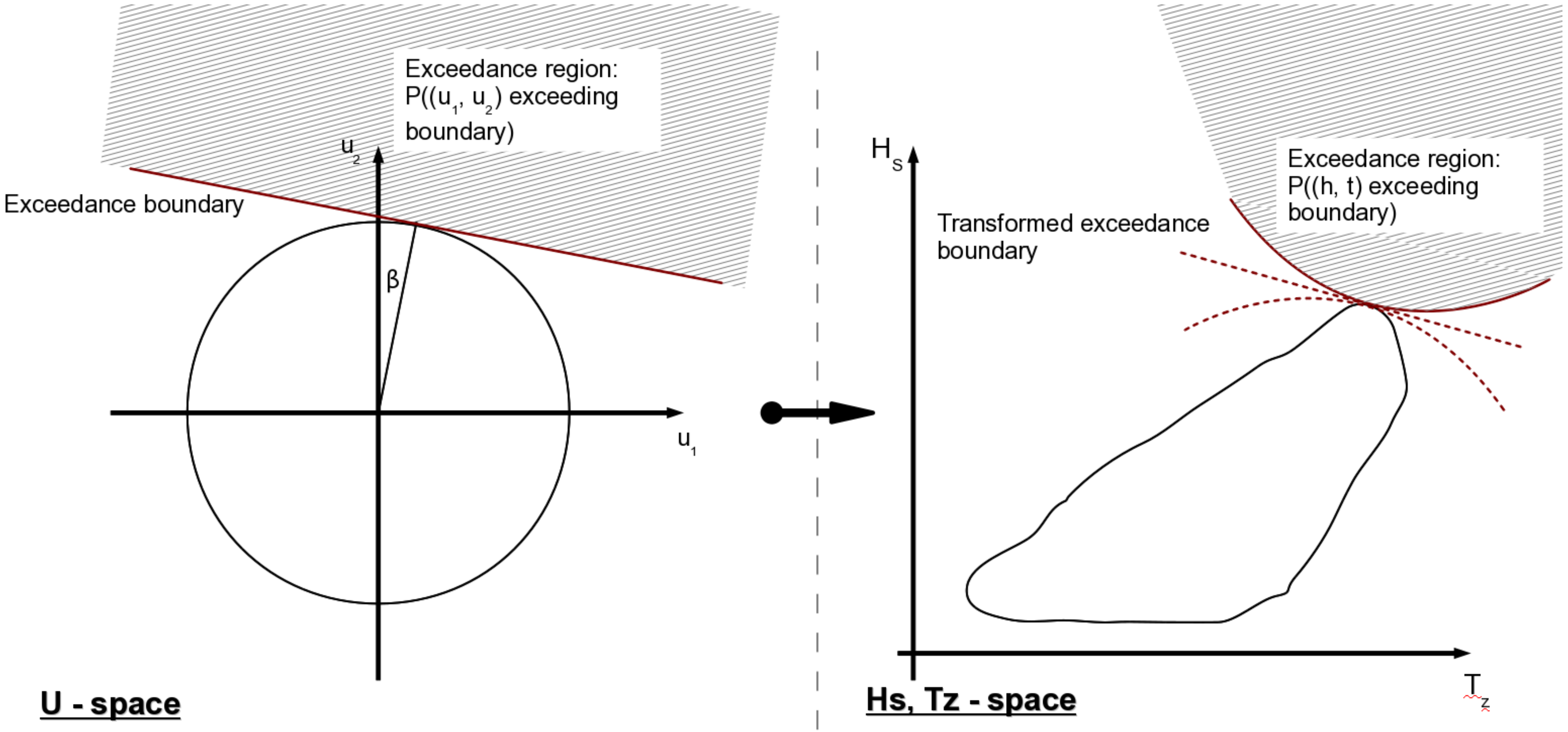
- Start with a hypersphere in standard normal space with radius $\beta_R = -\Phi^{-1}\left(P_f\right) = \Phi^{-1}\left(1 - P_f\right)$
  - Circle in 2D
- Transform the hypersphere to the physical parameter space using the ***inverse Rosenblatt transform***

$$\begin{gathered} X_1 = F_{X_1}^{-1}(\Phi(U_1)) \\ X_2 = F_{x_2|X_1}^{-1}(\Phi(U_2)) \\ \vdots \\ X_n = F_{x_n|X_1,X_2,\dots,X_n}^{-1}(\Phi(U_n)) \end{gathered}$$

- The resulting contours in $\boldsymbol{X}$-space will be the environmental contours

- This approach is based on the ***linear approximation*** of the failure boundary in ***standard normal space***

I-FORM contours: transformed limit state function will depend on the joint environmental model


Exceedance region:
P((u1, u2) exceeding boundary)
u2
Exceedance boundary
β
u1
U - space
Hs
Transformed exceedance boundary
Exceedance region:
P((h, t) exceeding boundary)
Tz
Hs, Tz - space

Iso-density contours

- Establish contours in the parameter space with ***equal density***
  - Need to define an anchor point for this to be uniquely defined

- Estimate the return value for the governing (primary) variable for the prescribed return period and associated values for the other variables
  - For example, 100-year return value for H_S and conditional median for T_Z

- The contour line is estimated form the joint model as the contour of ***constant probability density*** going through this point

Direct sampling contours

- The main idea is to perform the linearization of the failure probability in $\boldsymbol{X}$-space rather than in $\boldsymbol{U}$-space

- Simulate ***sufficient*** number of ***Monte Carlo simulations*** from the joint environmental model, $f_{\boldsymbol{X}}(x)$
- Establish the environmental contours by estimating supporting hyperplanes according to the specified failure probability P_f
  - For any given angle $\theta \in [0, 360)$, identify a hyperplane $\Pi(\theta)$, defined by an equation on the form $t\cos(\ ) + h\sin(\ ) = C(\ )$
  - The hyperplanes partition the sample space into two halfspaces $\Pi(\theta)^+$ and $\Pi(\theta)^-$ such that the fraction of sample points in $\Pi(\theta)^+$ is approximately equal to P_f
  - The resulting environmental contours are defined by the circumference of the set obtained by intersecting the sets $\Pi(\theta)^-$ for all $\theta \in [0, 360)$

Direct sampling contours – algorithm (2D)

1. Simulate n points $(T_1, H_1), \dots, (T_n, H_n)$
2. Calculate the projections at angle θ, $X_i = T_i \cos\theta + H_i \sin\theta\,,\ i = 1, \dots, n$
3. Sort in ascending order: $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ with corresponding samples $\left(T_{(1)}, H_{(1)}\right), \dots, \left(T_{(n)}, H_{(n)}\right)$
4. Calculate number of samples to be kept within the desired failure boundary: $k = n(1 - P_f)$
5. Estimate $\hat{C}(\theta) = X_{(k)}$ corresponding to the halfspace $\mathcal{B}(\theta) = \left\{(t, h)\colon\ t\cos\theta + h\sin\theta \ \le\ X_{(k)}\right\}$
6. Environmental contours: $\mathcal{B} = \bigcap_{\theta \in [0,360)} \mathcal{B}(\theta)$

Direct sampling contours – estimating the boundary (2D)

- Boundary may be identified by calculating ***intersection points*** (t, h) of neighbouring hyperplanes, corresponding to angles $\theta \text{ and } \theta + \delta$, by solving

$$t\cos\theta + h\sin\theta = C(\theta)$$

$$t\cos(\theta+\delta) + h\sin(\theta+\delta) = C(\theta+\delta)$$

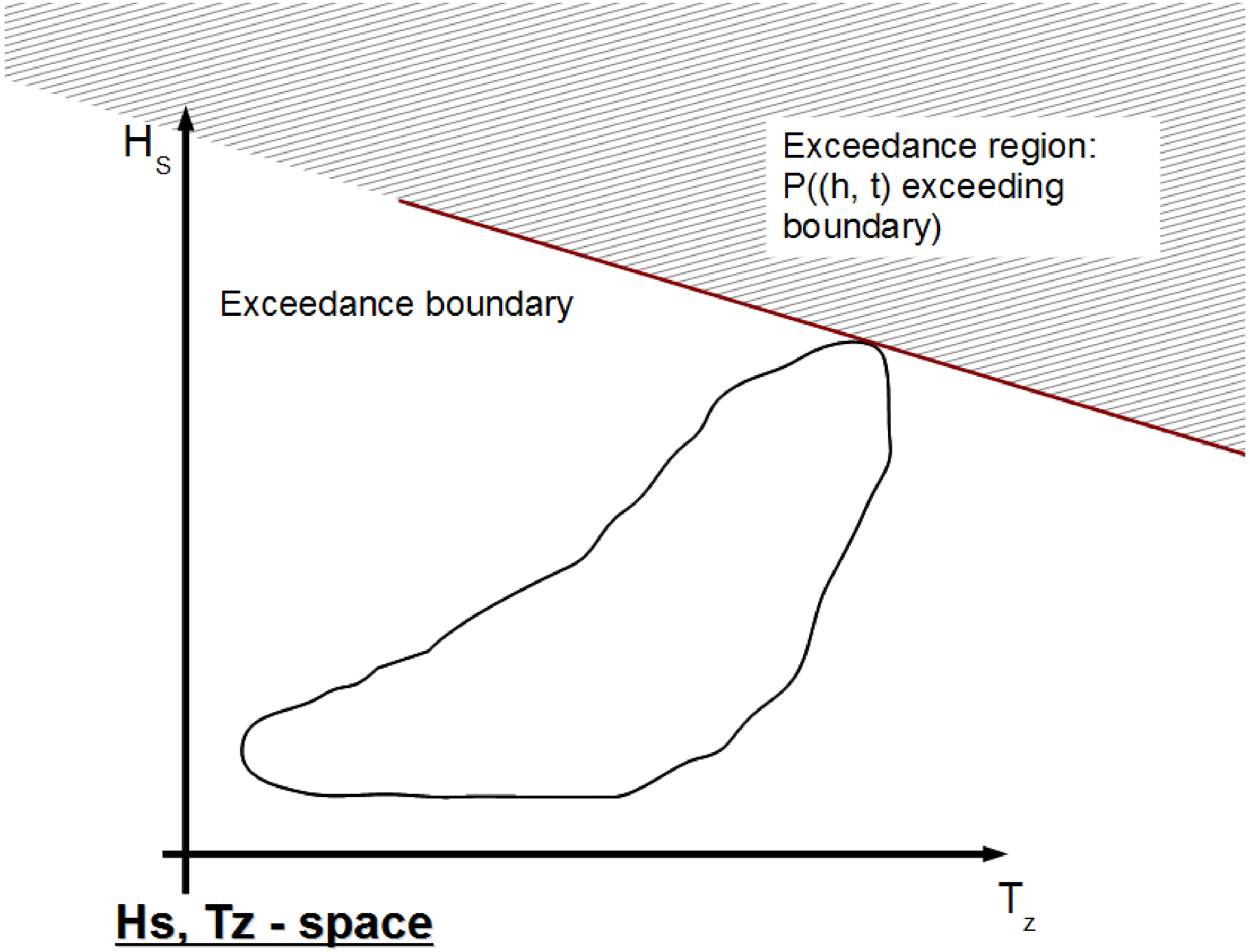
- The solutions are the points (t, h), for each θ

$$t = \frac{\sin(\theta+\delta)C(\theta) - \sin\theta\, C(\theta+\delta)}{\sin(\theta+\delta)\cos\theta - \sin\theta\cos(\theta+\delta)}$$

$$h = \frac{-\cos(\theta+\delta)C(\theta) + \cos\theta\, C(\theta+\delta)}{\sin(\theta+\delta)\cos\theta - \sin\theta\cos(\theta+\delta)}$$

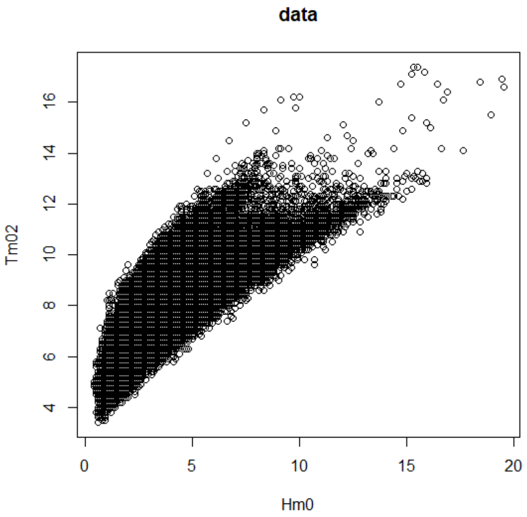
- Extensions to higher dimensional problems are possible (but somewhat more cumbersome)

Failure region with direct sampling contours

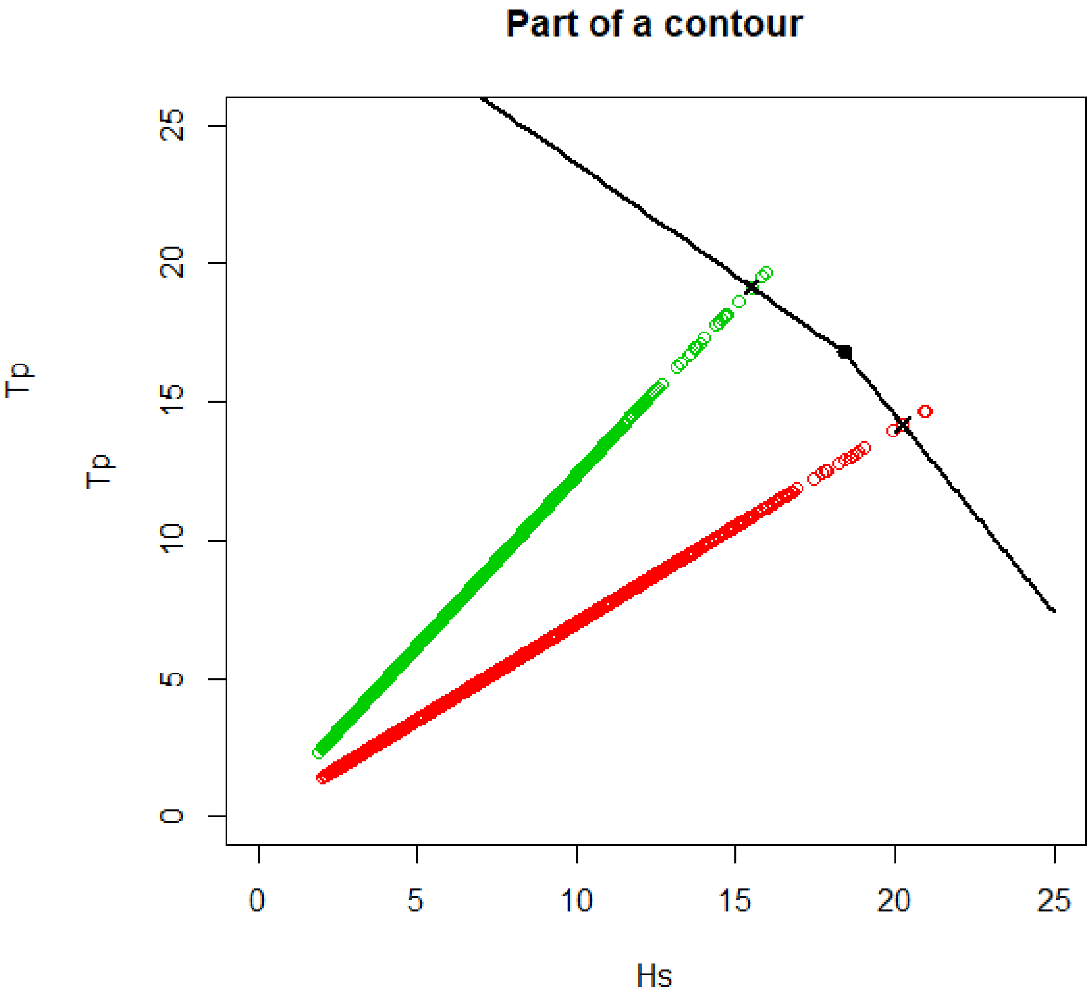


- Straightforward interpretation: Each point along the contour corresponds to a ***tangent line*** with a specified ***exceedance probability***

Direct sampling contours – calculating an intersection |

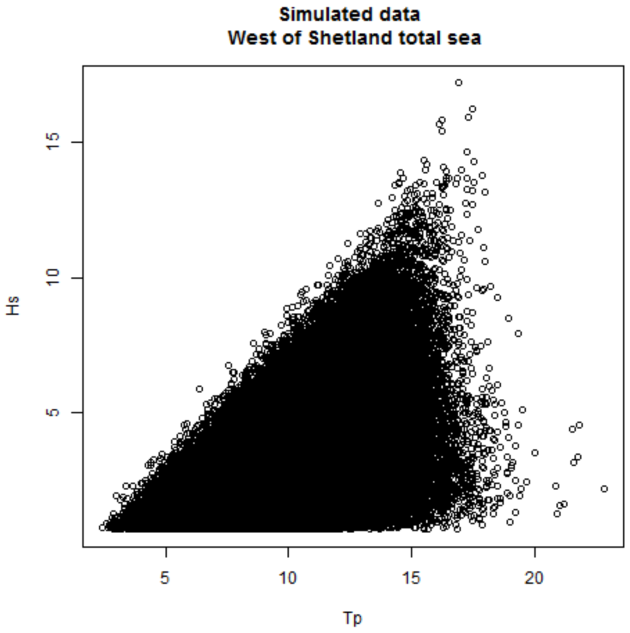


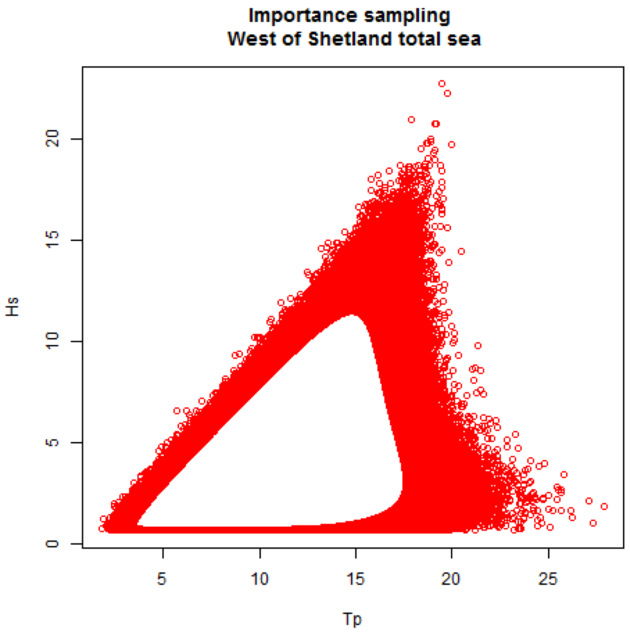
Data projected along two angles



Monte Carlo sampling using importance sampling

- When sampling from the joint distribution – only interested in the tails
- Importance sampling or tail sampling may be used to collect extremes and disregard samples near the mode
- Example shows N = 250 000 samples with and without importance sampling
  - Effective number of samples with importance sampling ≈ 5 000 000

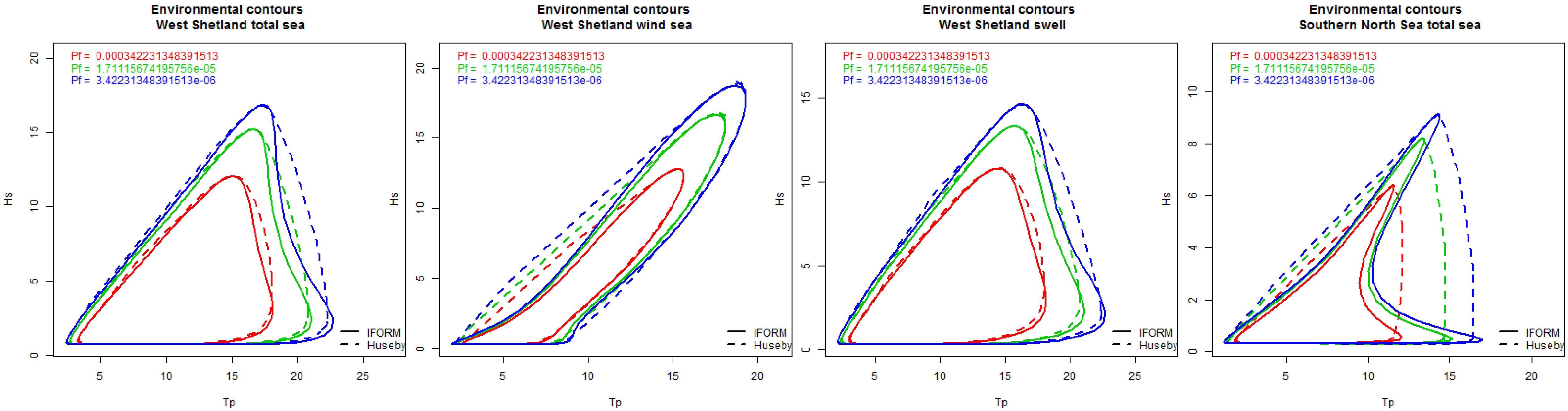




$P_d = 0.995$

Examples: Environmental contours for sea state variables

- Joint model for (HS, TZ) fitted to data from various locations
  - Total sea, wind sea or swell
  - 10-, 20- and 100-year contours
- IFORM contours and direct sampling contours



Environmental contours - Software

- Recent software from the ***ECSADES*** project are available for calculating contours from a given set of data in *R*
  - Assuming the conditional model for H_S and T_Z suggested in DNV-RP-C205
  - For other models, quite straightforward to implement from scratch
    - For I-FORM approach: Need to be able to transform from U-space to X-space
    - For direct sampling approach: need to be able to sample from the joint distribution

- Software may be downloaded from:

https://github.com/ECSADES/ecsades-r

Conditional extremes model

- Extreme value model for a random vector $\boldsymbol{X} = (X_1, \ldots X_d)$
- Estimate the distribution of $\boldsymbol{X}$ when $\boldsymbol{X}$ is extreme in at least one component
- Estimate the ***marginals*** of all X_i and the ***extremal dependence structure***, $\boldsymbol{X}_{-i} | \, X_i > v$
  - Estimate the marginals above thresholds, φ_i, by a Generalized Pareto distribution
  - Transform the marginals to ***standard Gumbel*** form by the probability integral transform, $Y_i = T_i(X_i)$
  - Model the dependence structure above some dependence threshold $\phi_{\bar{\tau}}$ as d semi-parametric regression models on the form (for $d = 2$):

$$(Y_2 | Y_1 = y_1) = \alpha_{\bar{\tau}} y_1 + y_1^{\beta_{\bar{\tau}}} W_{\bar{\tau}} \quad \text{for} \quad y_1 > \phi_{\bar{\tau}}$$

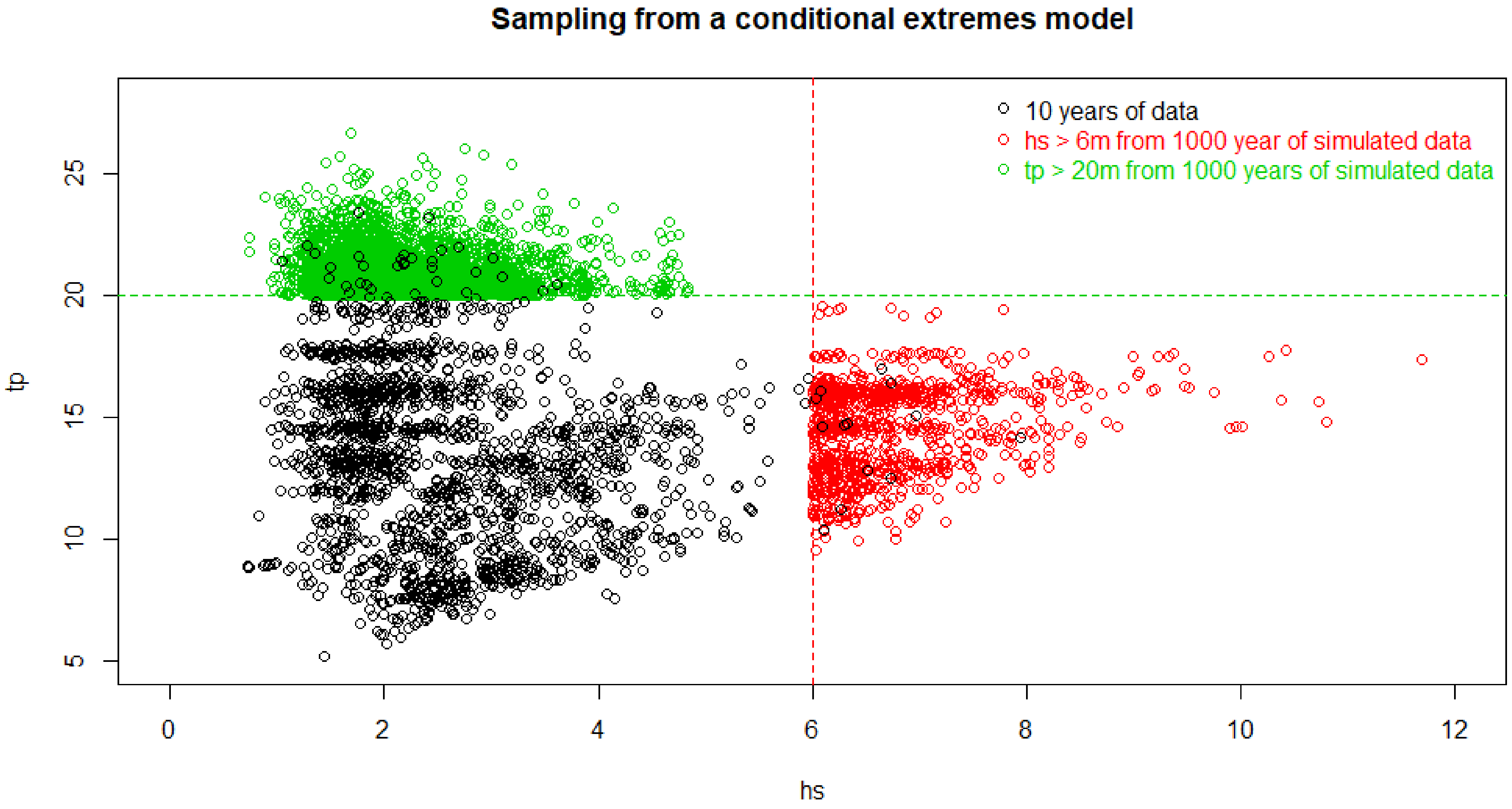
  - Where the threshold $\phi_{\bar{\tau}}$ is defined as the quantile of the standard Gumbel distribution with non-exceeding probability $\bar{\tau}$.
  - For estimation, assume $W_{\bar{\tau}} \sim N\left(\mu_{\bar{\tau}}, \sigma_{\bar{\tau}}^2\right)$

Conditional extremes model – extrapolation

- Once fitted, the conditional extremes model may be used to sample from the extremes of the multivariate distribution

1. Simulate Y_1 from a Gumbel distribution conditional on exceeding the marginal threshold $T_i(\varphi_i)$
2. Sample $W_{\bar{\tau}}$ from the empirical distribution of the residuals
3. Obtain $Y_2 = \widehat{\alpha_{\bar{\tau}}}\, Y_1 + Y_1^{\widehat{\beta_{\bar{\tau}}}} W_{\bar{\tau}}$
4. Transform $\boldsymbol{Y} = (Y_1, Y_2)$ to the original scale by using the inverse transformation, $\boldsymbol{X} = T^{-1}(\boldsymbol{Y})$
5. The resulting vector X constitutes a simulated value from the conditional distribution of $\boldsymbol{X} \mid X_1 > \ \varphi_i$

Sampling from a conditional extremes model



Exercise 4: Environmental contours

1. Download and install the ECSADES software in *R*
2. Calculate Environmental contours for the (H_S, T_Z) dataset used in previous exercises
   - Assume the conditional model (supported by the software) and estimate 10-, 20- and 100-year contours

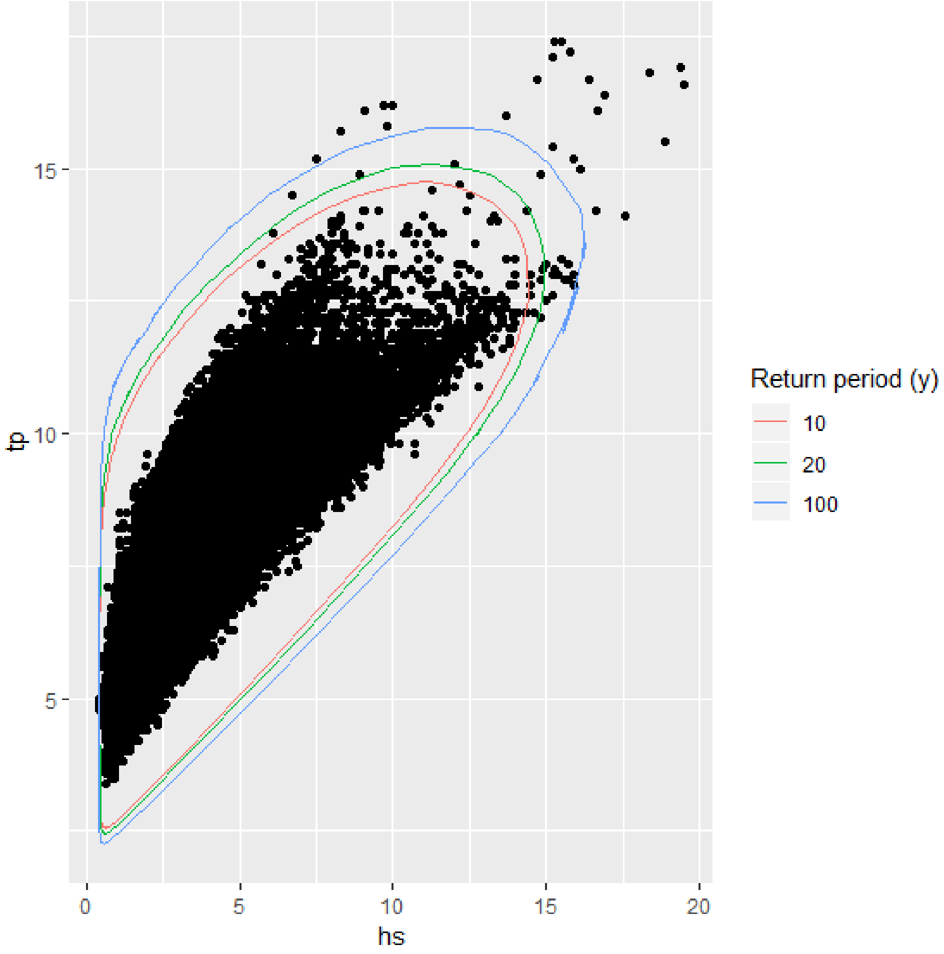
   a. Direct sampling contours

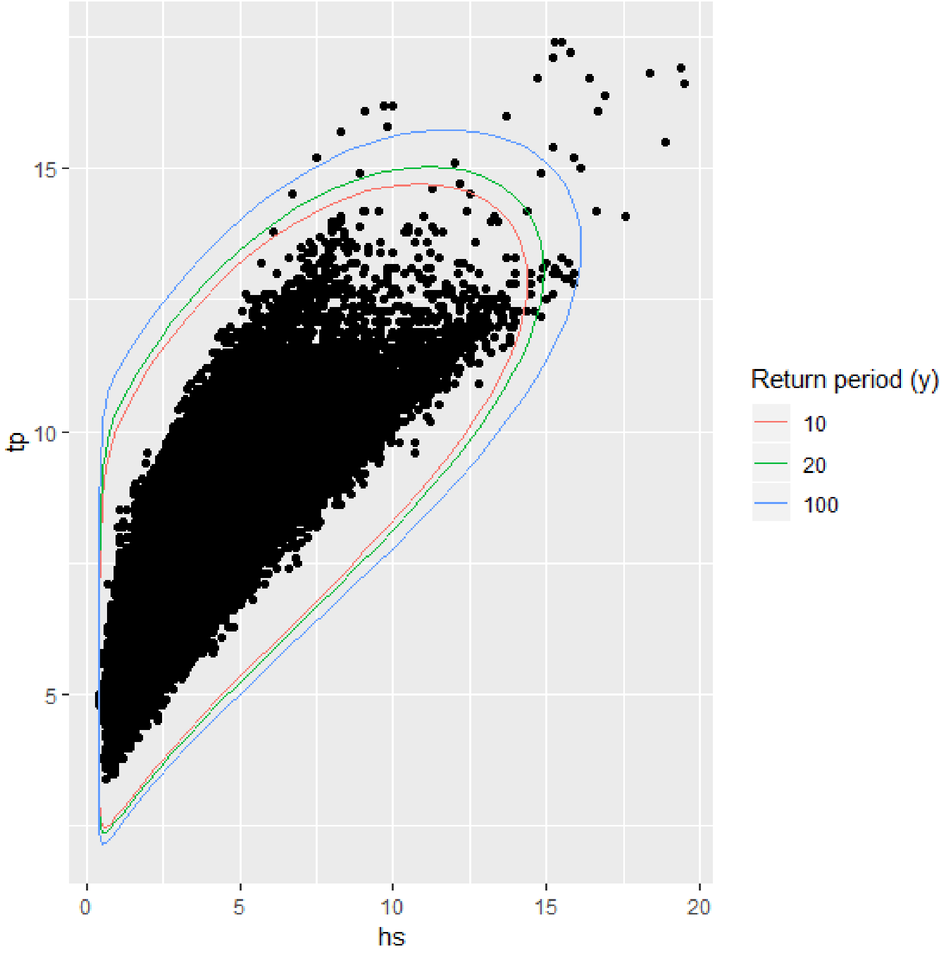
   b. I-FORM contours

3. Plot results

4. Compare and comment on the results

Exercise 4: Summary of results





Summary: Extreme value analysis

- Extreme value analysis aims at estimating ***return values*** for long return periods from data
- Parametric models are needed for ***extrapolation***
- Different modelling techniques exist utilizing different subsets of data
  - ***All data***
  - ***Subsampled*** data
  - Only ***peaks over threshold***
  - Only ***block maxima***
- Different statistical models can be used to fit such data and estimate extreme values

- Extreme value analysis is challenging even in the univariate case and becomes increasingly challenging in higher dimensions
- ***Environmental contours*** is one method for analysing joint extremes of environmental variables

WHEN TRUST MATTERS

Erik.Vanem@dnv.com
+47 6757 9900

www.dnv.com

DNV