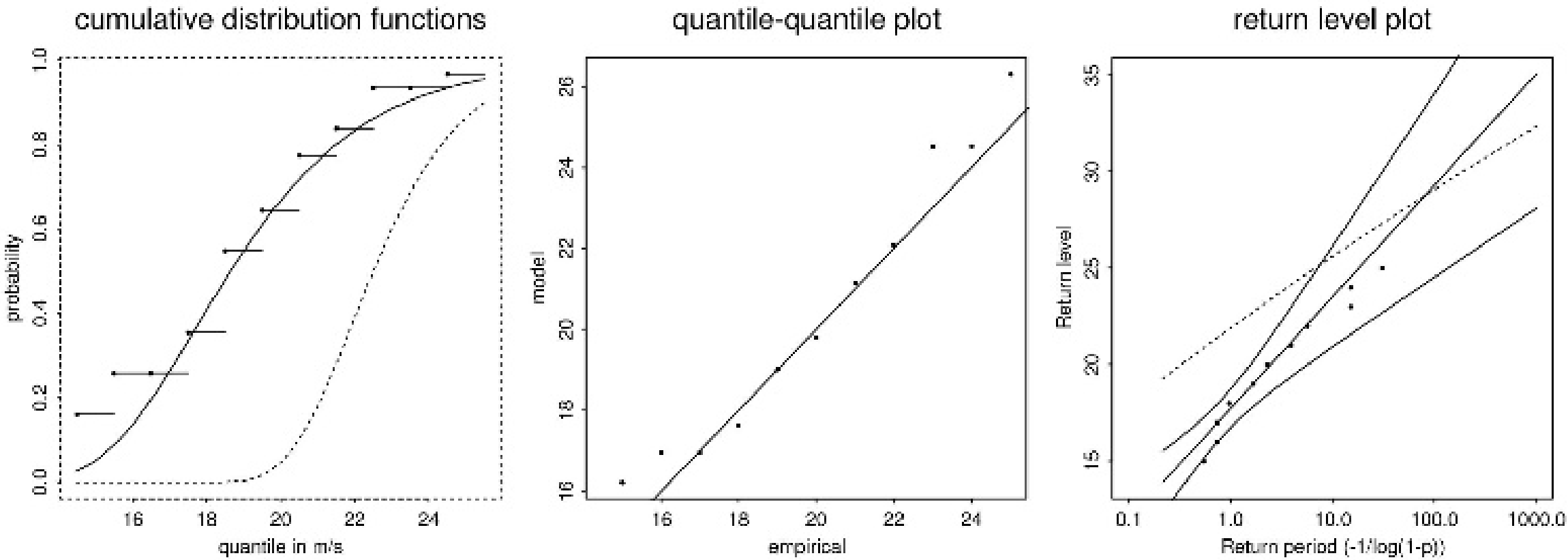
Extreme value statistics: from one dimension to many

Lecture 1: one dimension

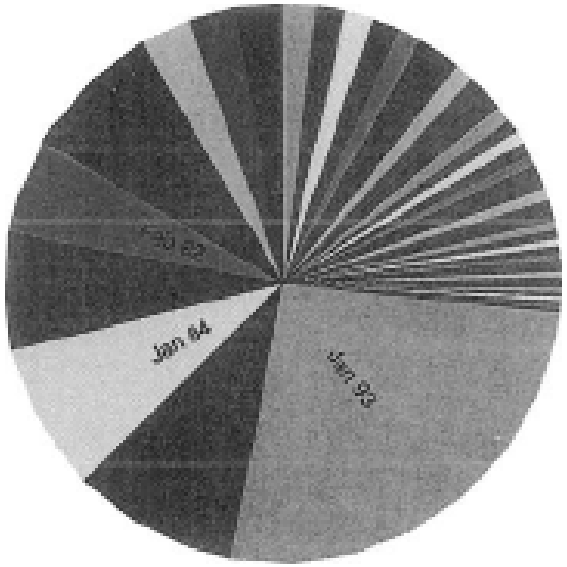
Lecture 2: many dimensions

Extremes shape much of the world around us

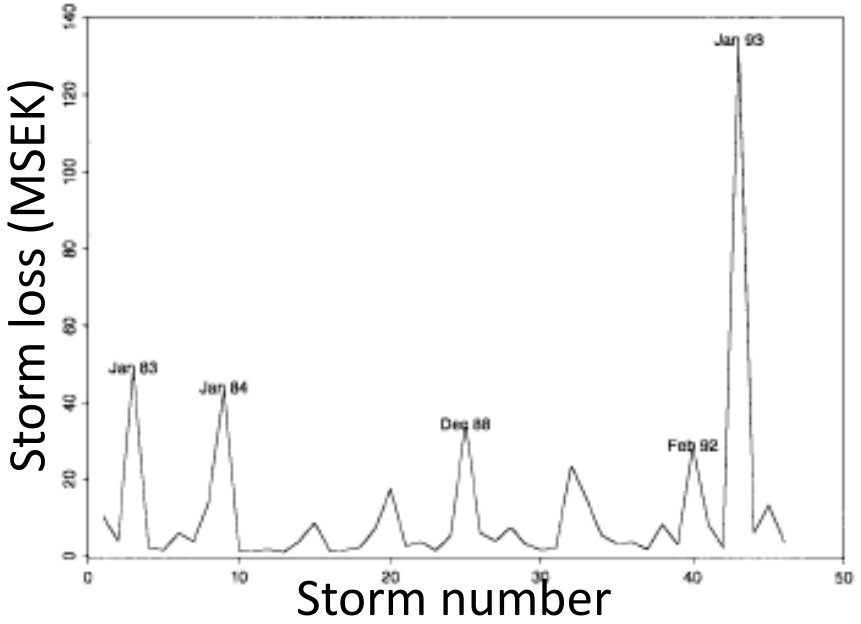
The philosophy of EVS is simple: extreme events, perhaps extreme water levels or extreme financial losses, are often quite different from ordinary everyday behavior, and ordinary behavior then has little to say about extremes, so that only other extreme events give useful information about future extreme events.



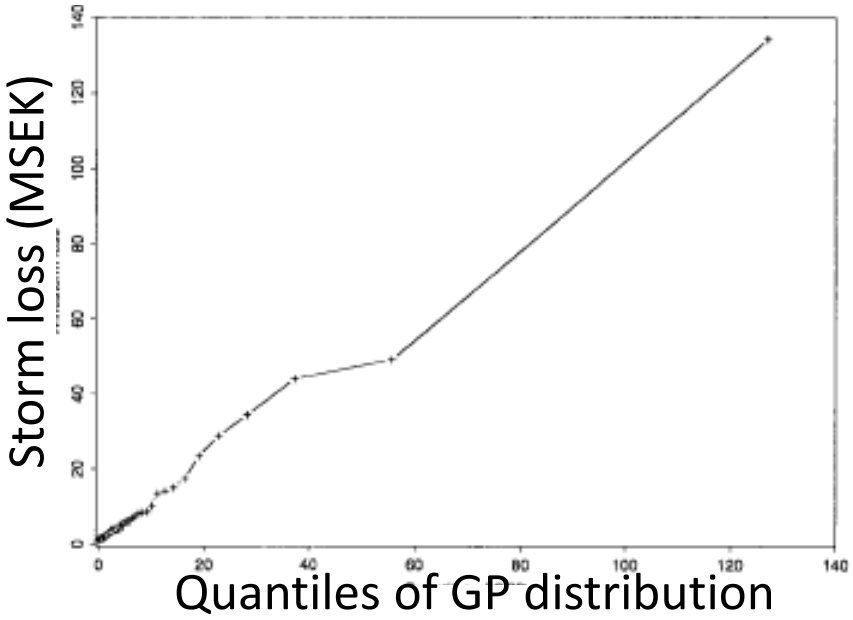
Yearly maximum 10 minute average windspeeds 1961–1990 at Barkåkra, Sweden. Solid lines: estimated GEV distribution of yearly maxima. Dotted lines: estimated Weibull distribution obtained by using all measurements (the Weibull fit was very good in the center of data)



Pie chart of LFAB wind-storm losses 1982-1993

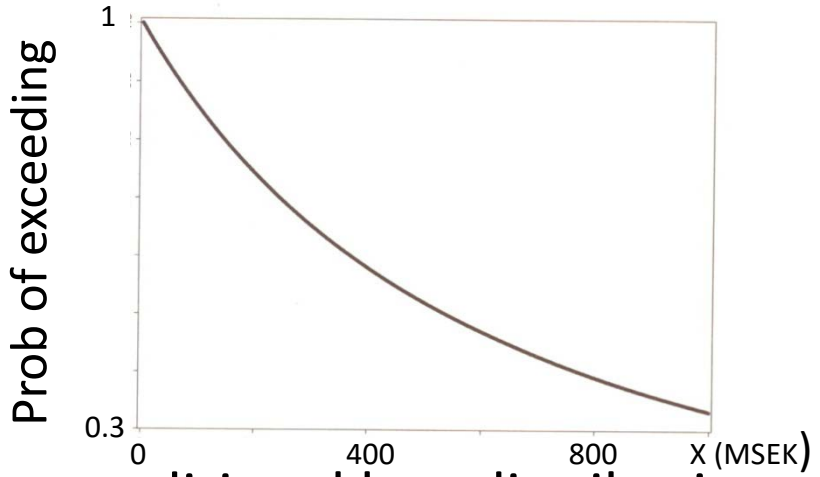


LFAB windstorm losses 1982-1993



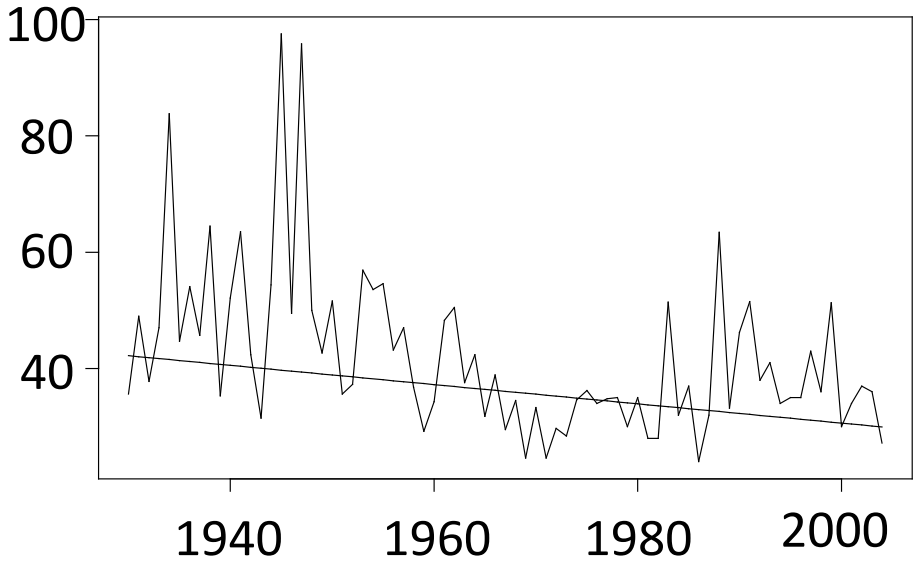
qq-plot against GP distribution

Risk (MSEK)	Next year	Next 5 years	Next 15 years
10%	66	215	473
1%	366	1149	2497

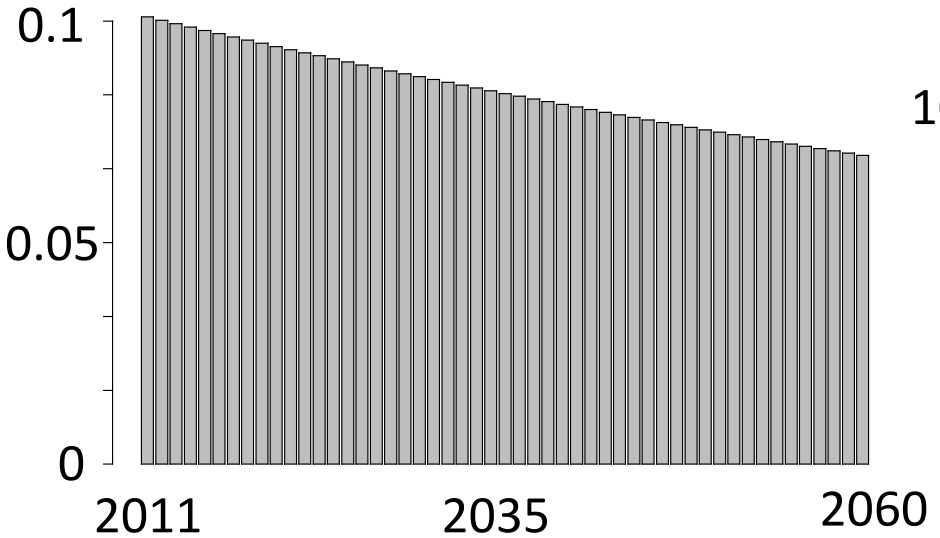


conditional loss distribution

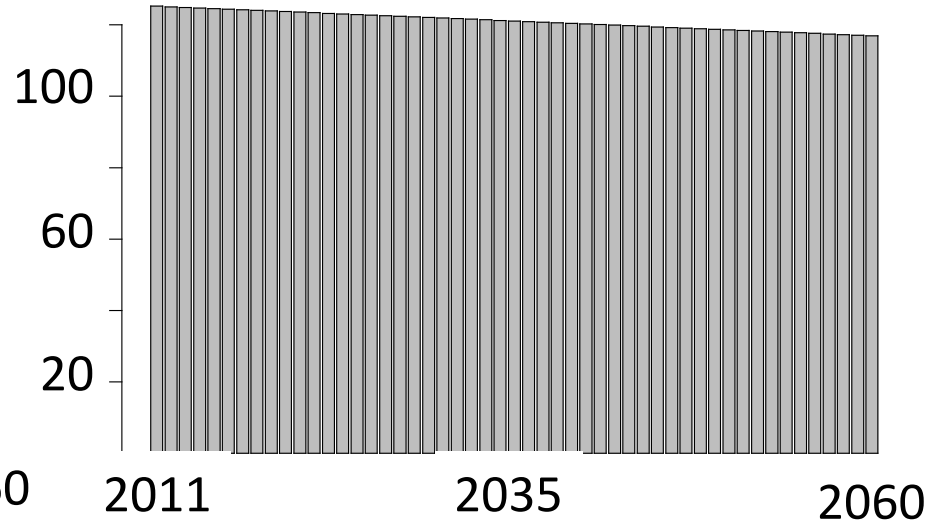
the 2011–2060 5% largest daily winter rainfall in Manjimup is 121 mm



Yearly maximum daily rainfall



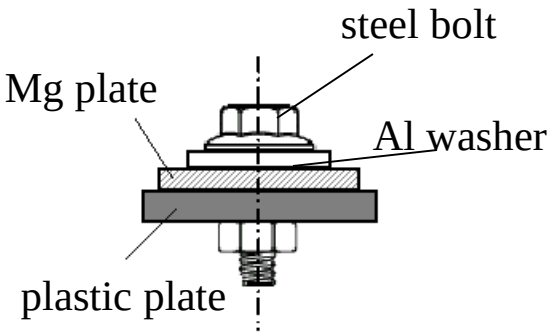
The 121 mm-s risk plot



The 0.1% constant risk plot

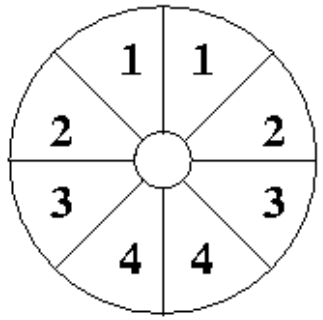
Laboratory experiment in climate chamber at VCC

- 3 different bolt coatings
- 3 time points
- 3 assemblies per coating and time point
- maximum pit depth measured in 8 sectors for each assembly



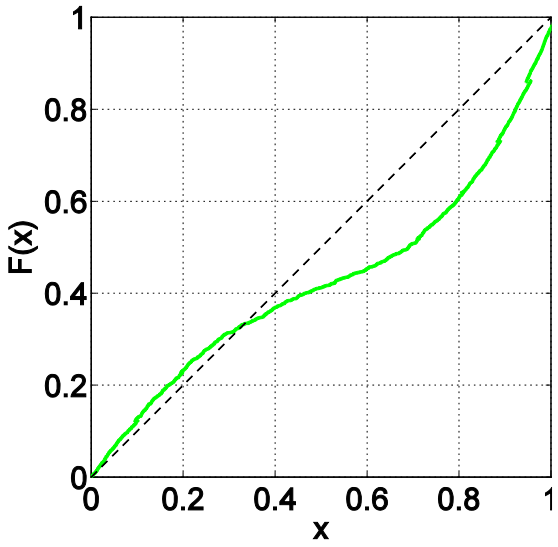
test assembly

corroded Mg plate

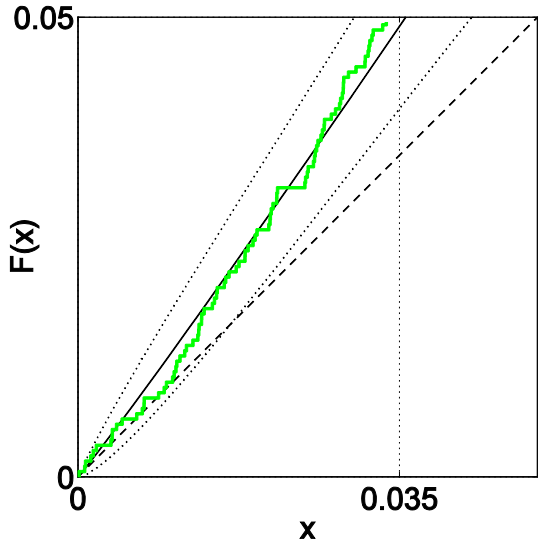


maximum pit depth mesured in sectors

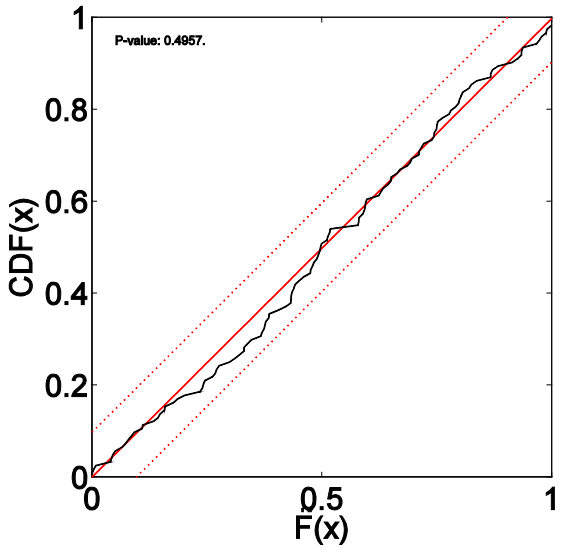
yeast: p-values for genomewide screening wildtype data set



empirical cdf of p-values (1728 p-values)

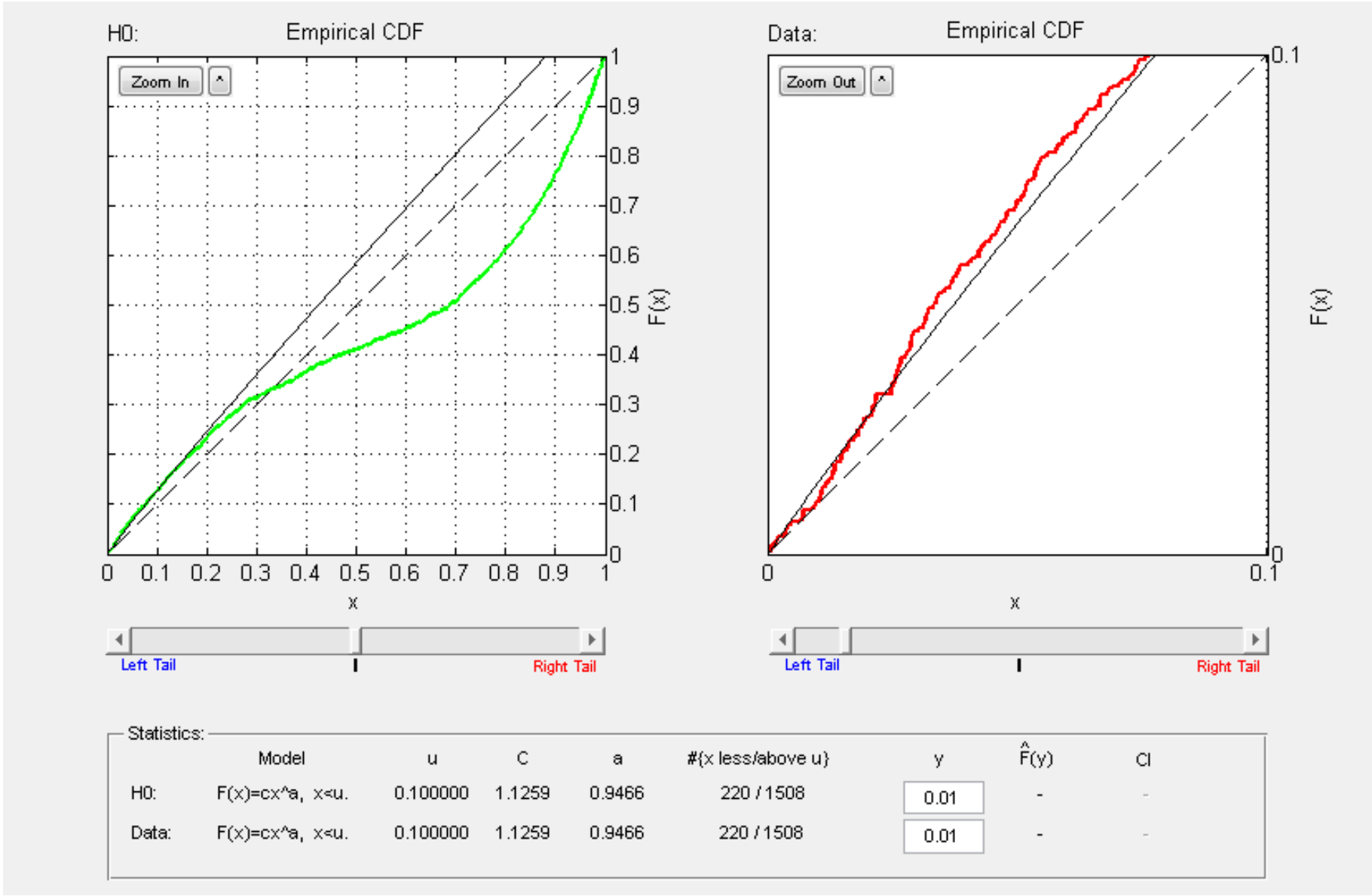


empirical cdf of p-values for p<0.05 with fitted model



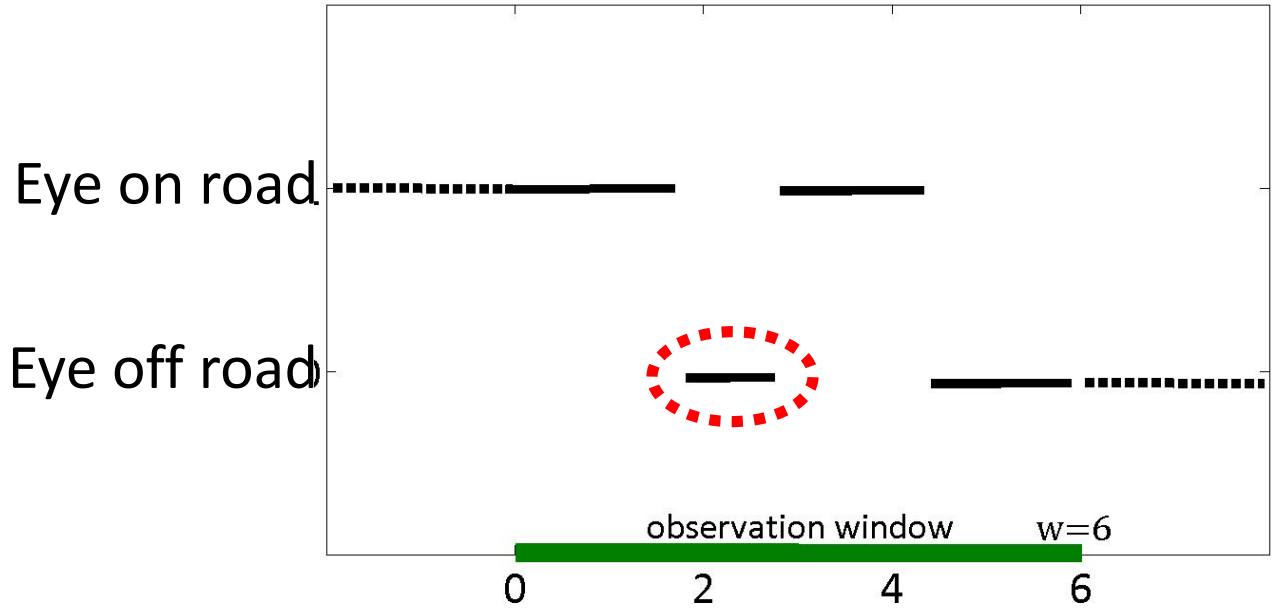
Kolmogorov - Smirnov 95% goodness of fit test (p=0.49)

SmartTail



Bioscreen testing for gene expression in yeast: wildtype data

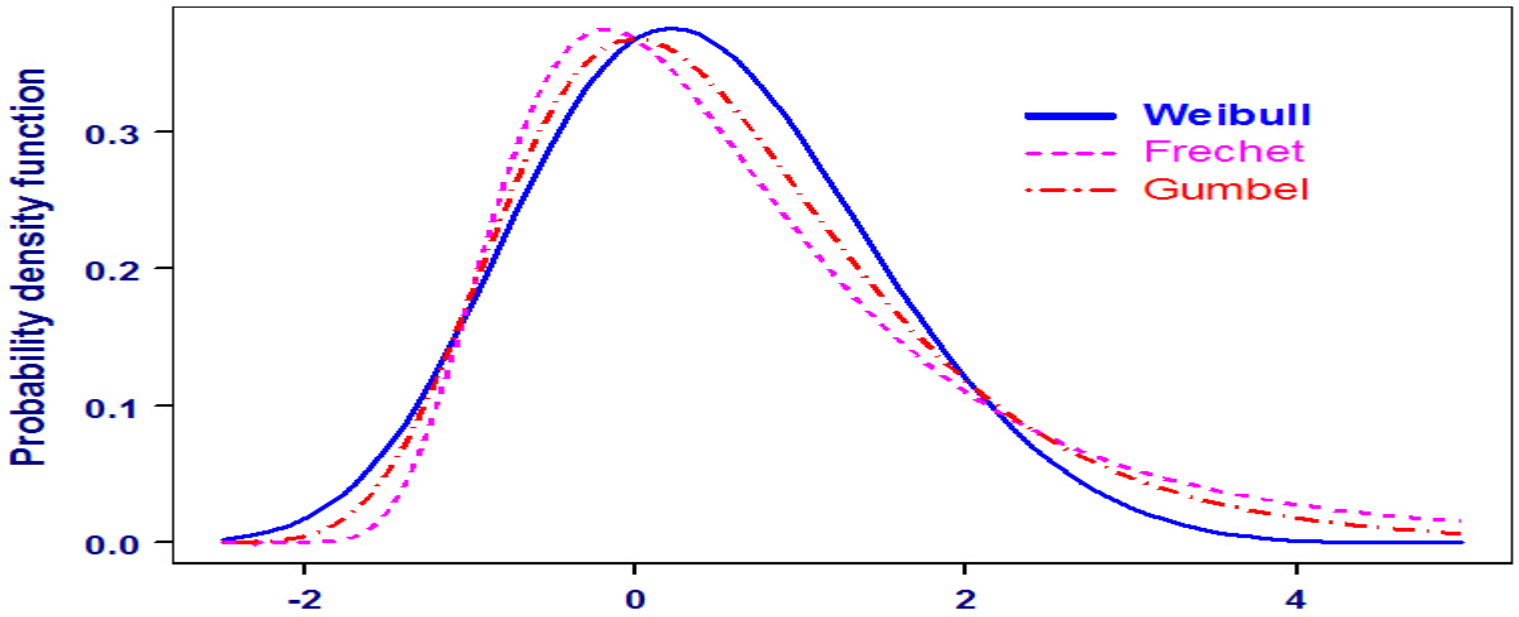
- Traffic accidents: 1.3 million deaths/year worldwide, 20-50 million severely injured
- Driver inattention major cause
- Estimate the (tail) of the distribution function of the eye-off-road intervals



Naturalistic driving study

Generalized extreme value (GEV) distributions

$$G(x) = e^{-\left(1+\frac{x-\mu}{\sigma}\right)_+^{-1/\gamma}}$$



$\gamma > 0$ Frechet distribution, finite left endpoint $x > \mu + \sigma/\gamma$

$\gamma = 0$ Gumbel distribution, $G(x) = \exp\{-e^{-(x-\gamma)/\sigma}\}$, unbounded

$\gamma < 0$ Weibull distribution, finite right endpoint $x < \mu + \sigma/|\gamma|$

The GEV cdf-s are the limit distributions of maxima: Let $X_1, X_2, \ldots$ be i.i.d. with cdf F. If for some scaling and location sequences $a_n > 0$ and b_n

$$Pr(a_n^{-1}(\vee_{i=1}^{n} X_i - b_n) \leq x) \to_d G(x), \qquad \text{as} \quad n \to \infty,$$

where G is non-degenerate, then G is a GEV distribution. Conversely all GEV distributions can be obtained in this way.

Pf: If $Pr(a_n^{-1}(\vee_{i=1}^{n} X_i - b_n) \leq x) = F(a_n x + b_n)^n \to_d G(x)$ then

$$F(a_n x + b_n)^{2n} \to_d G(x)^2 \quad \text{and} \quad F(a_{2n} x + b_{2n})^{2n} \to_d G(x)$$

By Kinchine's convergence of types theorem there hence exist $\alpha_2 > 0$ and β_2 with $G(\alpha_2 x + \beta_2)^2 = G(x)$. Generalize to conclude that to $t > 0$ there exist $\alpha_t > 0$ and β_t

$$G(\alpha_t x + \beta_t)^t = G(x).$$

This functional equation can be solved to give the GEV distributions. The converse is proved by straightforward checking.

The GEV distributions are the max-stable distributions: Let $X_1, \mathrm{X}_2, \ldots$ be i.i.d. with nondegenerate cdf G. If for some scaling and location sequences $\alpha_n > 0$ and β_n

$$\Pr(\alpha_n^{-1} \ (\mathrm{V}_{i=1}^{n} \mathrm{X}_i \ - \ \beta_n) \leq \ \mathrm{x}) \ = G(\mathrm{x}), \text{for} \ \ n = 1, 2, \ldots$$

then G is a GEV distribution. Conversely all GEV distributions can be obtained in this way.

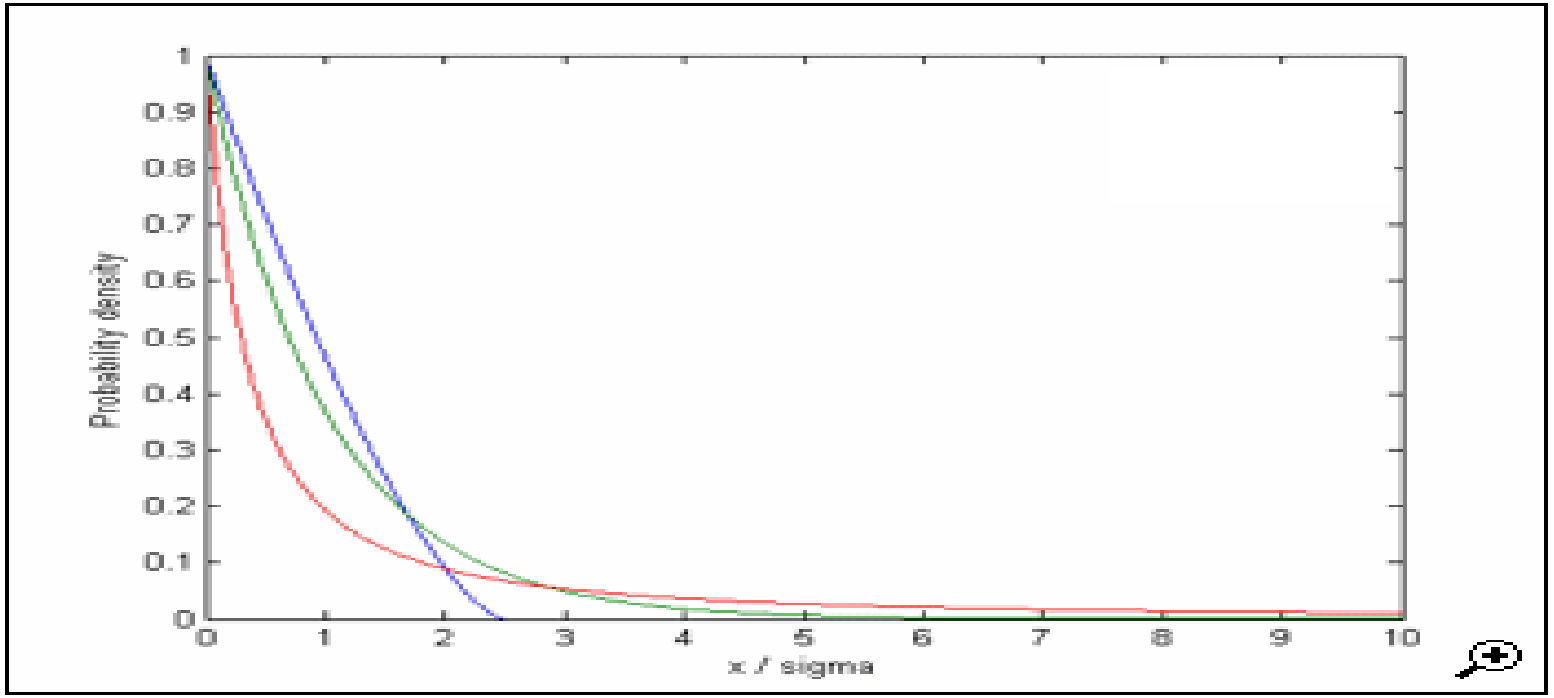
The block maxima method

- Obtain observations $X_1, \dots, X_n$ of block maxima (e.g. yearly maxima or sector maxima)
- Assume the obsevations are i.i.d and follow a GEV distribution
- Use $X_1, \dots, X_n$ to estimate the parameters of the GEV distribution
- Use the fitted GEV to compute estimates and confidence intervals for quantities of interest, e.g. quantiles of the distribution of the 50-year maximum

- Many estimation methods: ML estimators, moment estimators, probability weighted moments estimators, biascorrected estimators, …
- ML has the major advantage that it gives standardized ways for including trends in parameters and for testing of submodels
- Profile likelihood or bootstrap confidence intervals often preferable

Generalized Pareto(GP) distributions

$$\mathrm{H}(x) = 1 - \left(1 + \frac{x}{\sigma}\right)_{+}^{-1/\gamma}$$



$\gamma > 0$ left endpoint 0, right endpoint ∞

$\gamma = 0$ cdf $H(x) = 1 - e^{-x/\sigma}$

$\gamma < 0$ left endpoint 0, right endpoint $\sigma/|\gamma|$

The GP distributions are the limit distributions threshold excesses: Let X have cdf F. If there exist continuous threshold and scaling functions $u_t > 0$ and s_t with $F(u_t) < 1$ and $F(u_t) \to 1$ as $t \to \infty$, such that

$$Pr(s_t^{-1}(X - u_t) \leq \mathrm{x} \mid X > u_t) \to_d H(\mathrm{x}), \qquad \text{as} \quad n \to \infty,$$

where H is non-degenerate, then H is a GP distribution. Conversely all GP distributions can be obtained in this way.

- maxima of F are in "the domain of attraction of a GEV cdf" if and only if threshold excesses of F are in "the domain of attraction of a GP cdf H"
- if normalization is chosen appropriately,

$$H(x) = \log \frac{G(x)}{G(0)}$$

The GP distributions are the threshold-stable distribution: Let X have cdf H, with H nondegenerate. If there exist continuous threshold and scaling functions $u_t > 0$ and s_t with $F(u_t) < 1$ and $F(u_t) \to 1$ as $t \to \infty$, such that

$$Pr(s_t^{-1}(X - u_t) \leq \mathrm{x} \mid X > u_t) = H(\mathrm{x}), \qquad \text{for all } t \geq 1,$$

then H is a GP distribution. Conversely all GP distributions has this property.

Peaks over thresholds (PoT) method

- Choose (high) threshold u and from i.i.d observations $Y_1, \dots, Y_n \sim F$ obtain N threshold excesses $X_1 = Y_{t_1} - u, \dots, X_N = Y_{t_N} - u$, where $t_1, \dots, t_N$ are the times of threshold exceedance
- Assume $X_1, \dots, X_N$ are i.i.d and follow a GP distribution and that $t_1, \dots, t_N$ are the occurrence times of an independent Poisson process, so that N has a Poisson distribution
- Use $X_1, \dots, X_n$ to estimate the parameters of the GP distribution and N to estimate the mean of the Poisson distribution
- Estimate tail $\bar{F}(x) = 1 - F(x) = \bar{F}(u)\bar{F}_u(x-u)$, where $\bar{F}_u(x-u)$ is the conditional distribution of threshold excesses, by

$$\hat{\bar{F}}(x) = \frac{N}{n}\,\hat{H}(x-u)$$

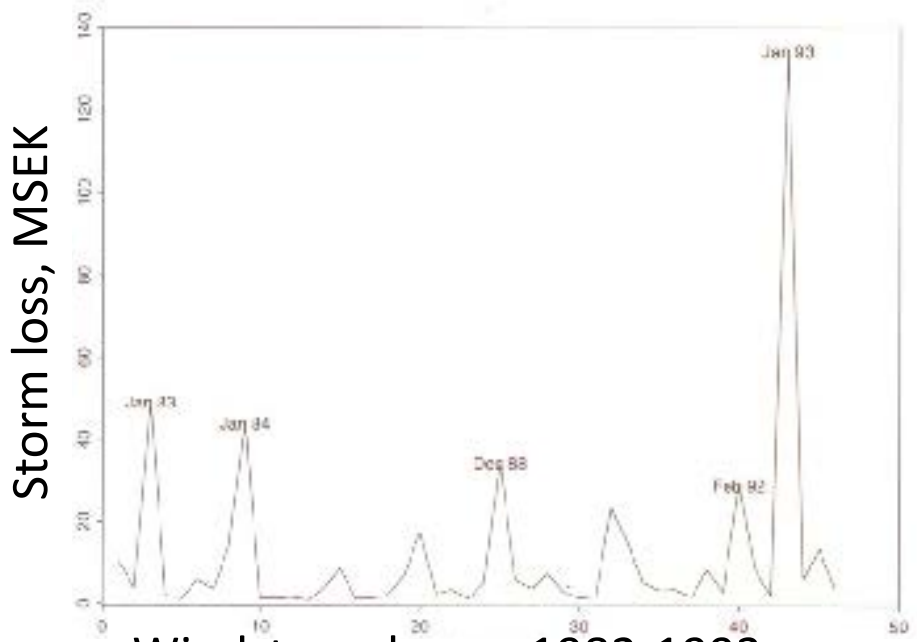
- Many estimation methods: ML estimators, moment estimators, probability weighted moments estimators, biascorrected estimators, …
- ML has the major advantage that it gives standardized ways for including trends in parameters and for testing of submodels
- Profile likelihood or bootstrap confidence intervals often preferable
- Straightforward to compute/estimate distribution of block maxima from fitted PoT model

Choice of threshold/number of order statistics in PoT + model diagnostics

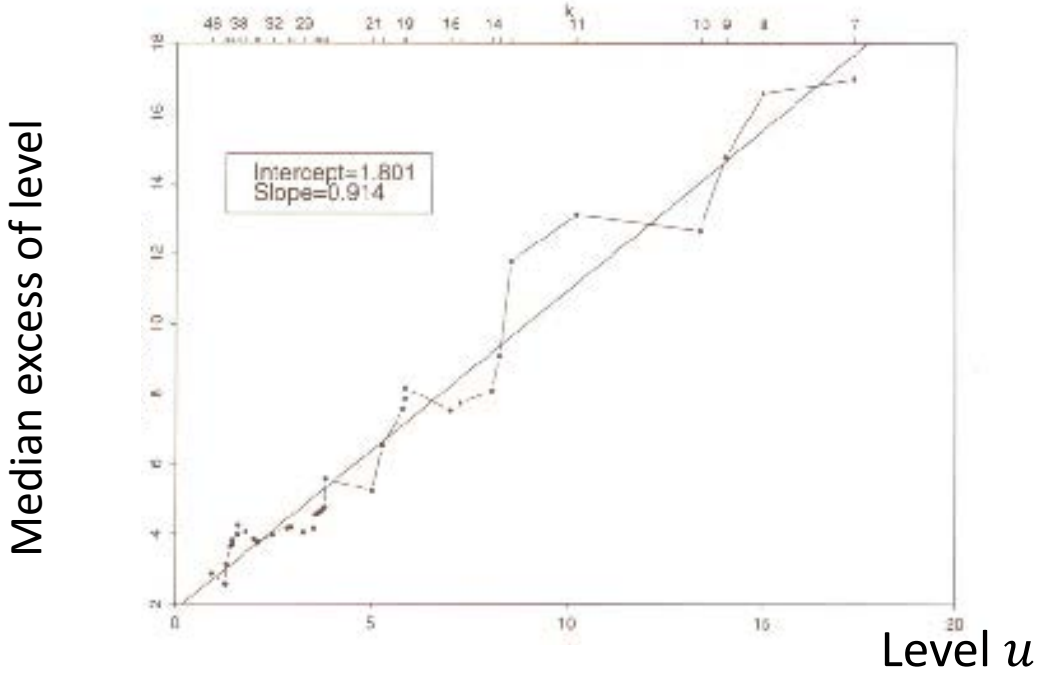
Threshold choice compromise between low bias (= good fit of model), which requires high threshold/few order statistics, and low variance, which requires low threshold/many order statistics. Tools aiding choice include

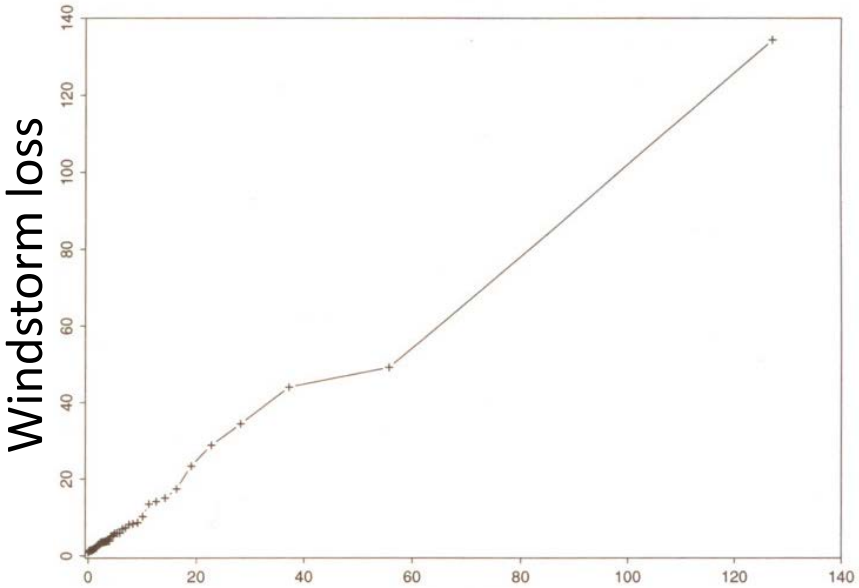
- mean excess plots (high variability for heavy tails)
- median excess plots
- plots of parameter estimates as function of threshold/number of order statistics
- qq- and pp-plots

automatic threshold selection procedures exist, but perhaps not always reliable ("optimal" threshold depends on the underlying unknown)

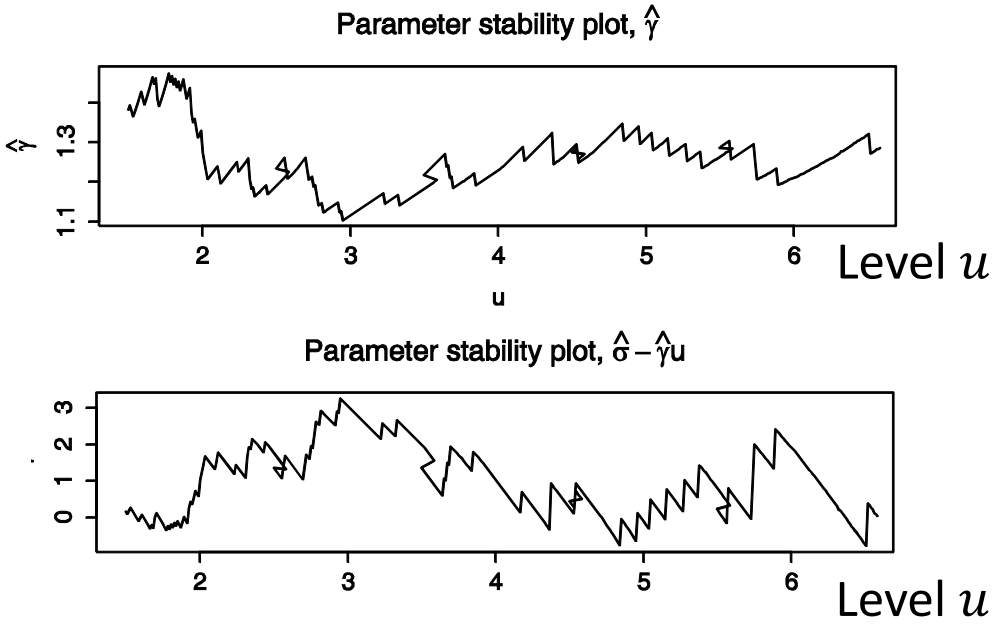


Windstorm losses 1982-1993
3-day excesses of 0.9 MSEK



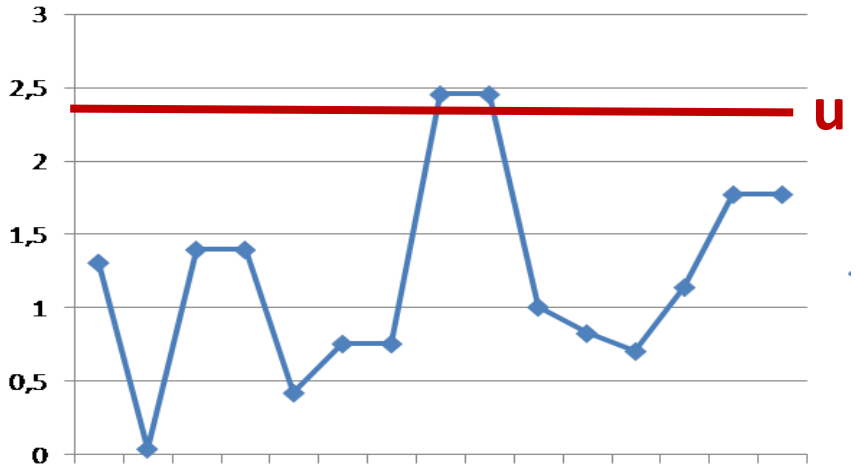


Quantiles of GP distribution

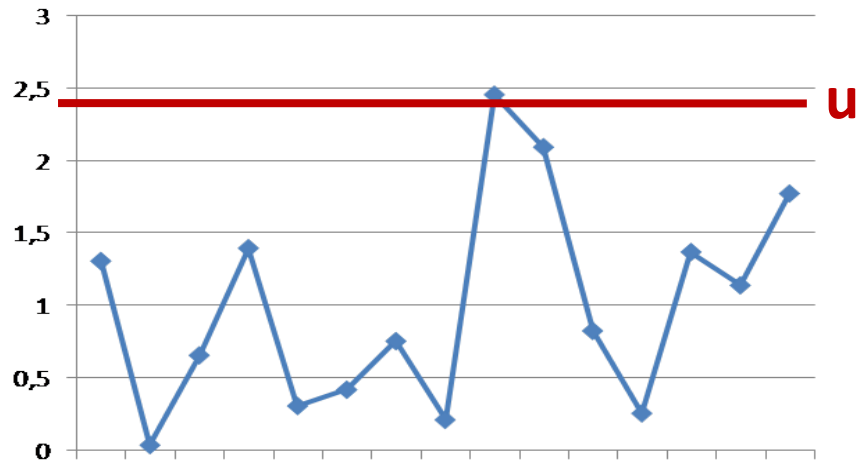


parameter stability plots

Dependence→ extremes typically come in clusters



stationary, cdf $F(x)$

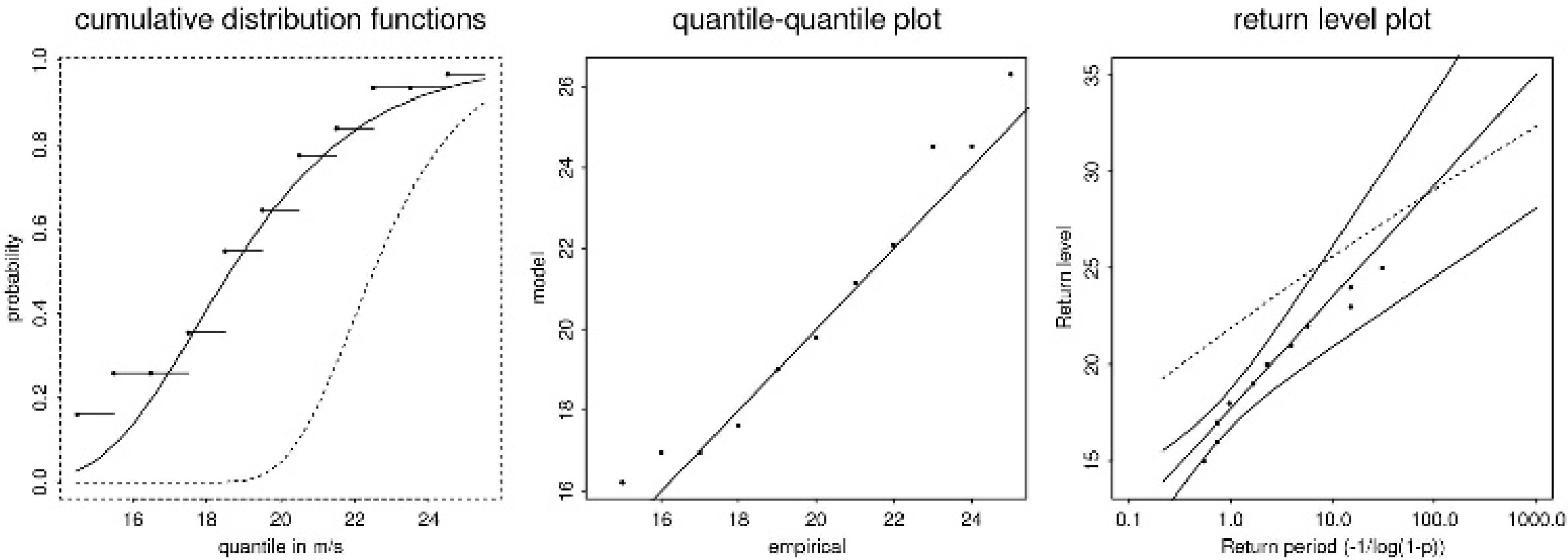


i.i.d, cdf $F(x)$

- θ = "Extremal index" = 1/asymptotic mean cluster length
- typically $Pr(\vee_{i=1}^{n} X_i \leq x) \approx F(x)^{n\theta}$
- typically clusters asymptotically i.i.d., dependence within clusters
- typically tail of cluster maxima asymptotically same as tail of F**!!**
- typically the EV distributions the only possible limit distributions

Dependence, cont'd

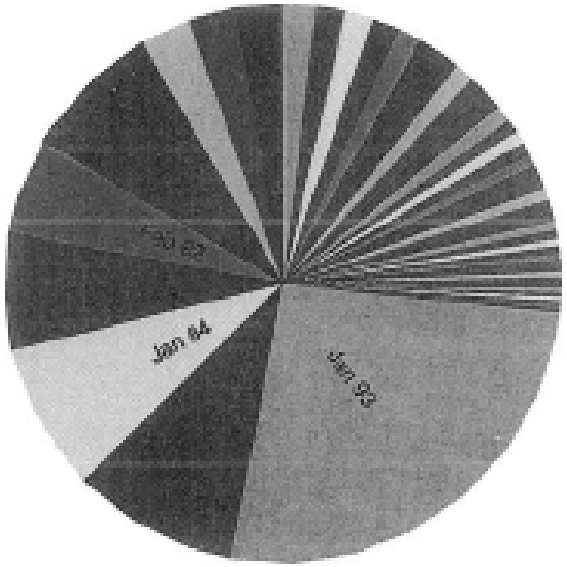
- *Block maxima method*: If blocks are long, then block maxima (typically) are approximately independent, and one can proceed as for i.i.d. observations
- *PoT:* Standard method is to use i.i.d. PoT method with excesses replaced by cluster maxima, and exceedance times replaced by the times when cluster maxima occur. (Though sometimes it may be better to use all excesses: work in progress with H. Drees, A. Janssen)
- *PoT:* Estimate extremal index by $\hat{\theta} = \frac{\#\text{clusters}}{\#\text{exceedances}}$
- PoT: Use $Pr(\vee_{i=1}^{n} X_i \leq x) \approx F(x)^{n\theta}$ to switch between PoT and block maxima



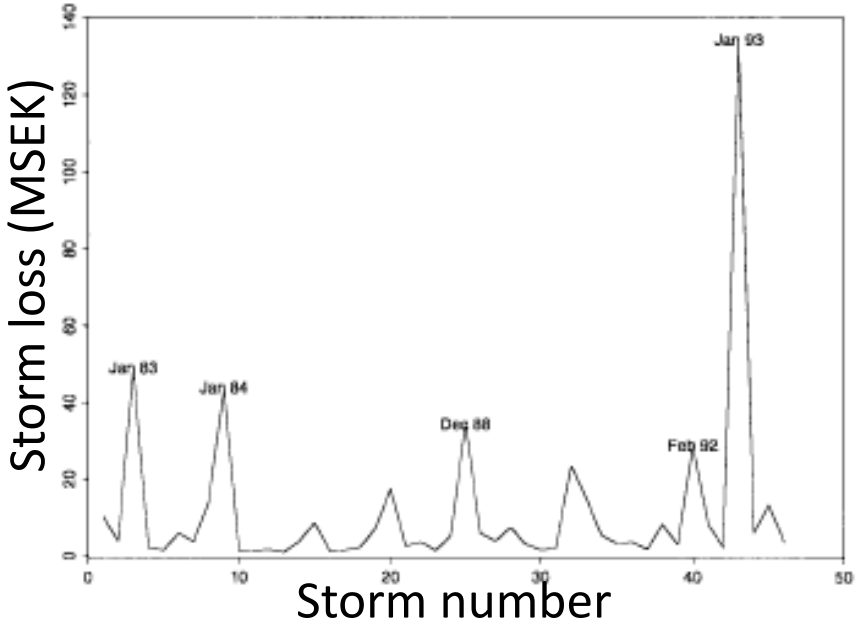
Yearly maximum 10 minute average windspeeds 1961–1990 at Barkåkra, Sweden. Solid lines: estimated GEV distribution of yearly maxima. Dotted lines: estimated Weibull distribution obtained by using all measurements (the Weibull fit was very good in the center of data)

- *Data:* 10 minute average wind speed at the start of each hour during 1961–1990, from 12 synoptic meteorological stations in Sweden
- *Aim:* Investigate methods to estimate the $1/50$ or $1/100$ upper quantiles of yearly maximum windspeed, as contribution to development of Swedish wind standards
- *Analysis:* Block maxima method with block $=$ year. Maximum likelihood estimation which took rounding of wind speeds to whole knots into account. Separate analysis for each station
- *Conclusions: U*se Block maxima method, not Weibull fitting; if rounding is neglected then estimation uncertainty is underestimated

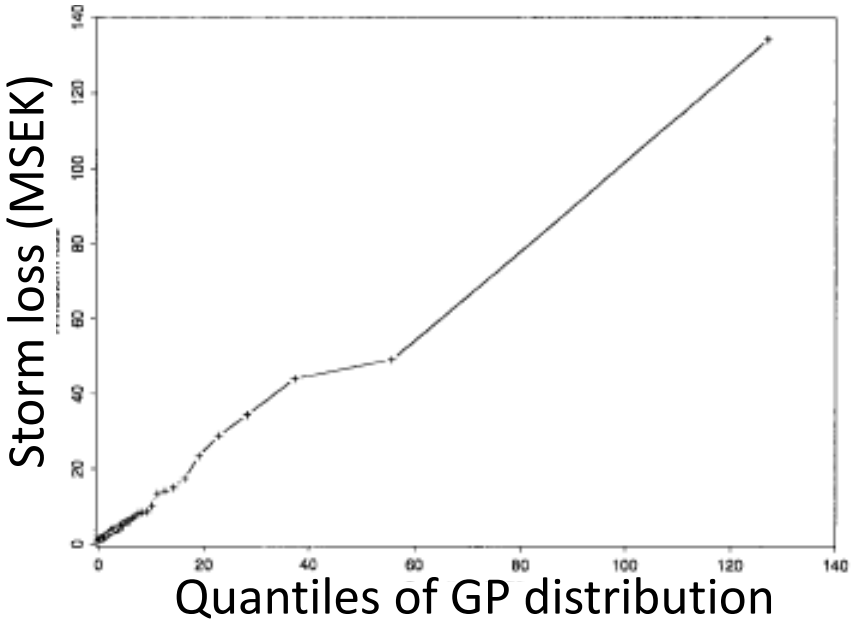
Perrin et al, Theor. Appl. Climatol. 2006



Pie chart of LFAB wind-storm losses 1982-1993

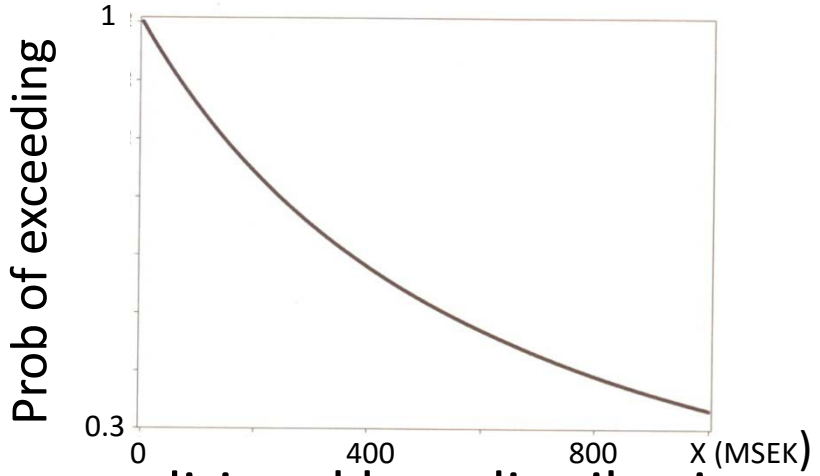


LFAB windstorm losses 1982-1993



qq-plot against GP distribution

Risk (MSEK)	Next year	Next 5 years	Next 15 years
10%	66	215	473
1%	366	1149	2497

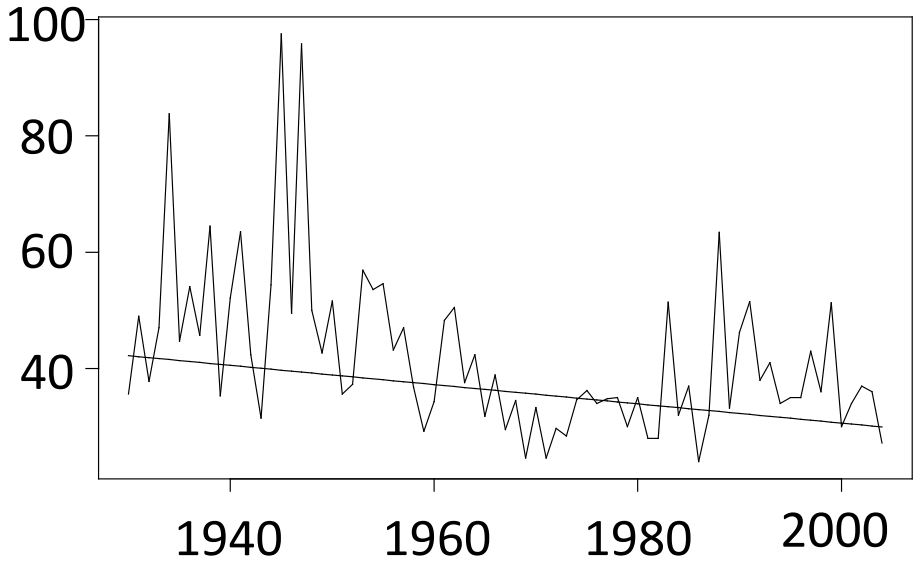


conditional loss distribution

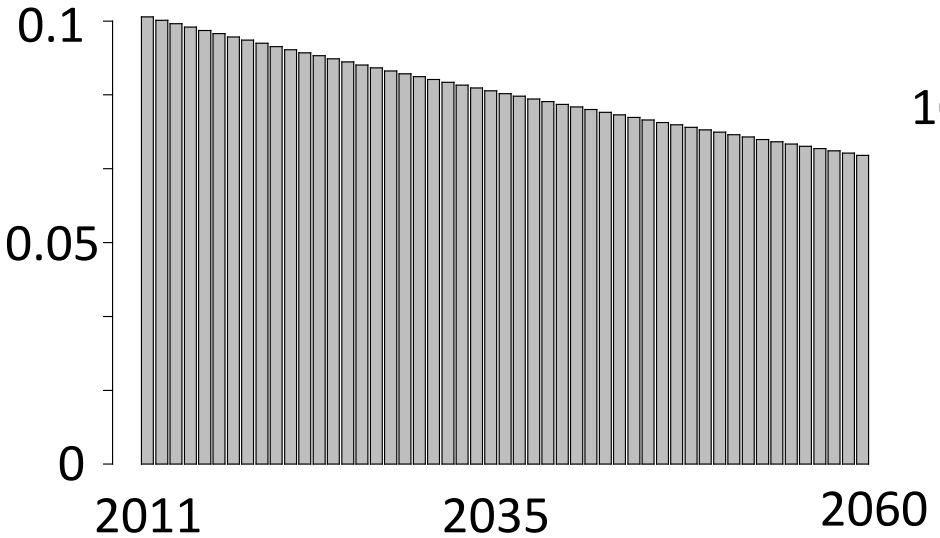
- *Data:* all 3-day periods during 1982-1993 with more than 0.9 MSEK windstorm loss to LFAB
- *Aim:* Method development, advice LFAB on how much wind storm reinsurance to buy
- *Analysis:* Standard ML PoT with time trend in scale parameter and Poisson process
- *Results:* No significant time trend; often used lognormal analysis gave bad predictions; LFAB increased it's reinsurance coverage substantially
- *Story to be continued tomorrow!*

Rootzén & Tajvidi, Scand. Actuarial. J. 1997

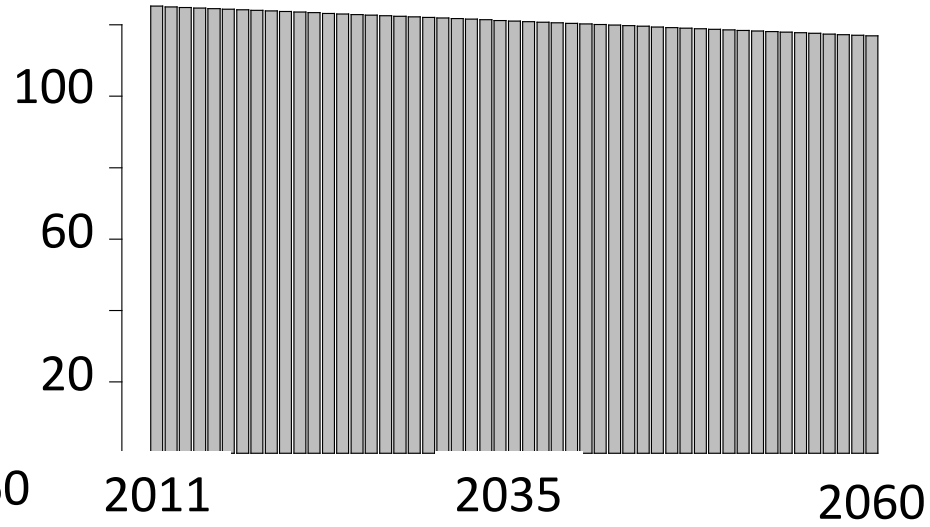
the 2011–2060 5% largest daily winter rainfall in Manjimup is 121 mm



Yearly maximum daily rainfall



The 121 mm-s risk plot



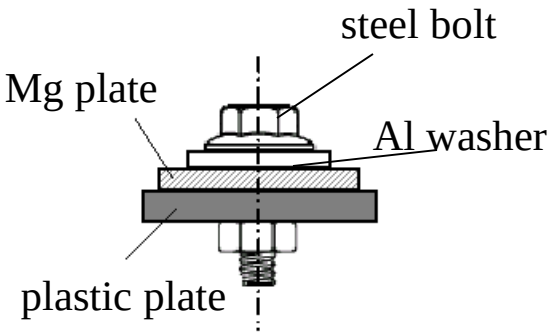
The 0.1% constant risk plot

- *Data:* Yearly maximum daily winter rainfall 1929-2004 in Manjimup, Western Australia
- *Aim:* Illustrate a new concept “Design Life Level” aimed at handling design in a changing climate
- *Analysis:* Standard ML block maxima method with time trends in parameters
- *Results:* Significant trend in location parameter

- Rootzén & Katz, Water Resources Research 2013

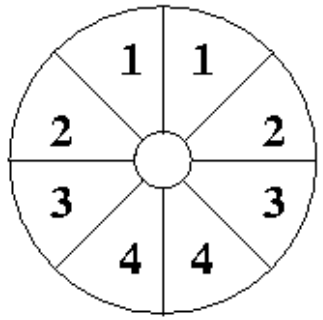
Laboratory experiment in climate chamber at VCC

- 3 different bolt coatings
- 3 time points
- 3 assemblies per coating and time point
- maximum pit depth measured in 8 sectors for each assembly



test assembly

corroded Mg plate

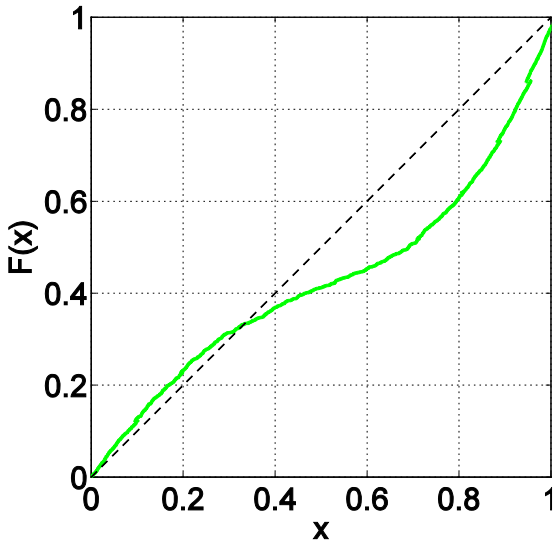


maximum pit depth mesured in sectors

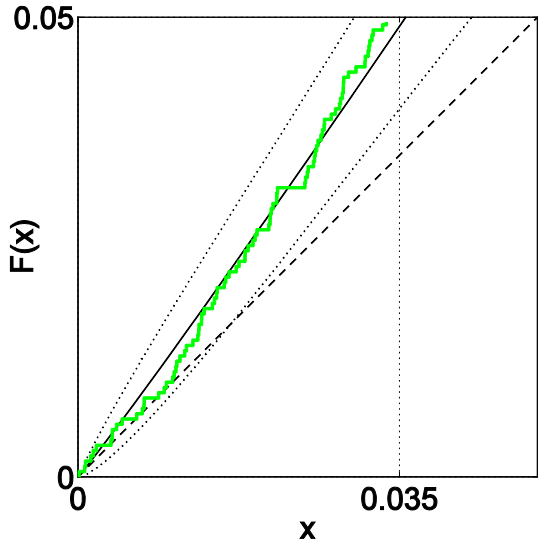
- *Data:* See previous slide
- *Aim:* Analysis of corrosion experiments at Ford/Volvo
- *Analysis:* New GEV random effects model $X_{i,j} = \mu + M_i + G_{i,j}$ where the M_i are exp-stable with parameters $\sigma,\ 0, \alpha < 1$ and the $G_{i,j}$ are Gumbel with parameters $0, \alpha$ (building on results by Hougaard (1986), Crowder (1989), Tawn (1990), Coles & Tawn (1991), Stephenson (2003))
- *Results:* Estimate of risk of penetration 50% higher than estimate from i.i.d. analysis

Fougères et al., Scand. J. Statist 2009

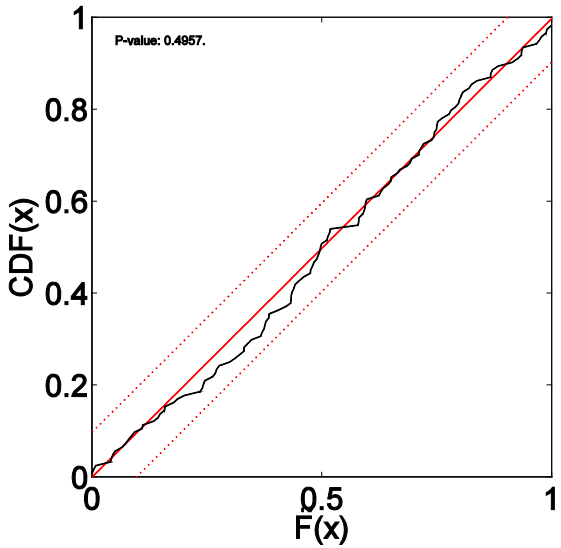
yeast: p-values for genomewide screening wildtype data set



empirical cdf of p-values (1728 p-values)



empirical cdf of p-values for p<0.05 with fitted model



Kolmogorov - Smirnov 95% goodness of fit test (p=0.49)

The standard approach to multiple testing: Benjamini-Hochberg

Smart method to adjust significance level of the tests to guarantee that the false discovery rate

$$FDR = \frac{E\#\text{false positives}}{E\#\text{positives}} 1_{\{\#\text{positives}>0\}}$$

is below some user-defined constant θ

But

- Typically one wants to know “how many positives are false”, not just that FDR is less than θ
- Very often $\frac{\Pr(\text{ a test is a false positive})}{\Pr(\text{a test is a positive})} \rightarrow \text{a limit } c$, as one goes out into the tails of the distribution of the test statistic. If c is bigger than θ, then Benjamini-Hochberg makes $FDR < \theta$ by making $Pr(\#\text{rejections} > 0)$ smaller than θ/c. This says little about whether rejections are false or not
- Benjamini-Hochberg assumes that the theoretical null distribution is the true null distribution. Often this is substantially wrong

New answers

- Conditional distribution of $\#\text{false positives}$ given the total number of positives is approximately binomial
- Methods to estimate the success probability ($pFDR$) of this binomial distribution
- Techniques to handle cases when theoretical and true null distributions differ

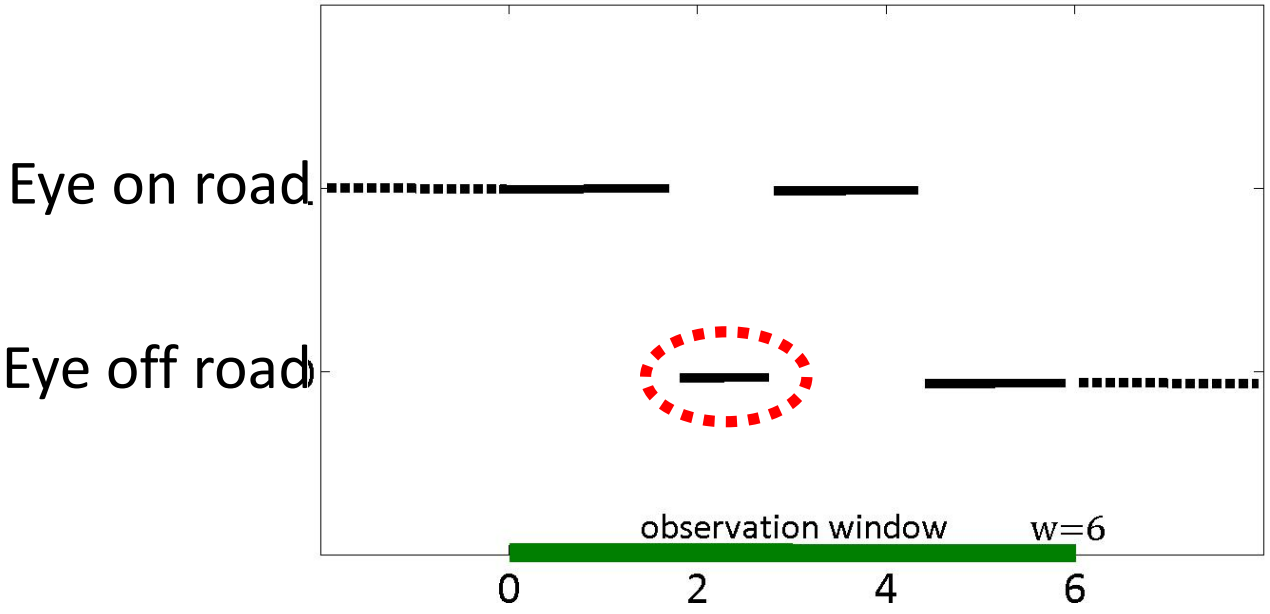
Uses PoT methods and extreme value theory motivated mixture model for P-values in a test:

$$Pr(P \leq x) = \pi_0 c_0 x^{1/\gamma_0} + \pi_1 c1_0 x^{1/\gamma_1} \qquad 0 \leq x \leq u,$$

for some suitable low threshold u.

Rootzén & Zholud, Biometrika 2015

- Traffic accidents: 1.3 million deaths/year worldwide, 20-50 million severely injured
- Driver inattention major cause
- Estimate the (tail) of the distribution function of the eye off road intervals



- Fit GP distribution to long eye off road intervals
- Only use first interval
- Handle selection bias

Naturalistic driving study

Visual behavior

How much do you look off road while driving?

- ❑ 5% of the time
- ❑ 10% of the time
- ❑ 15% of the time
- ☒ 20% of the time

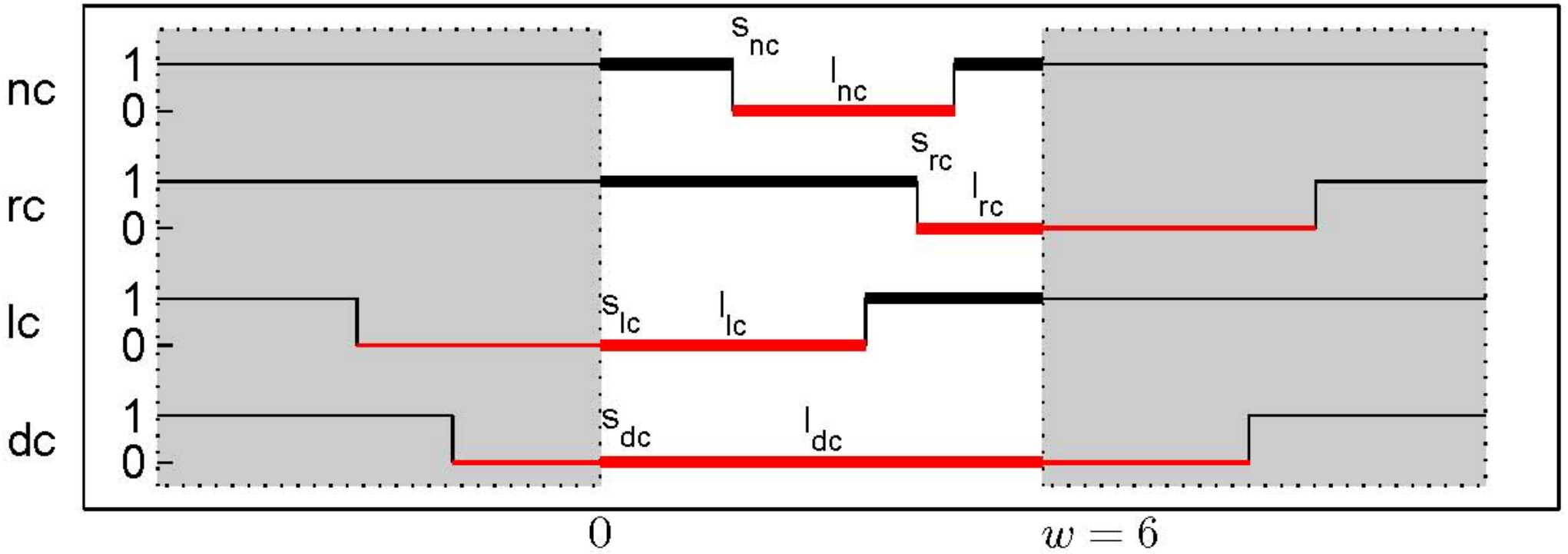
What is the .999 quantile of the lengths of off road glances?

- ❑ 1 second
- ❑ 2 seconds
- ❑ 3 seconds
- ❑ 4 seconds
- ☒ 5 seconds
- ❑ 10 seconds

Is glance behavior different in different situations?

Not well understood

- *Data:* 4803 randomly chosen 6 sec long observation windows obtained during normal driving, human annotators used web camera recordings of the driver's face to construct a 6 sec 0–1 process, where 0 means that the driver looks away from the road, and 1 that she looks at the road



White is observation window, S start of eyes-off-road interval, l observed length of eyes-off-road interval, nc is non-censored, rc is right censored, lc is left censored, dc is doubly censored

Crash Acc.mpg

- *Model as* alternating ”zero-one” renewal process (valid more generally)
- *Analysis:* Only use first glance in observation interval; ML estimation of parametric or PoT model (threshold $u = 2$) for length of eyes-off-road intervals; likelihood consisting of 4 kinds of terms corresponding to the different censorings in figure on previous slide
- *Results:* GP distribution estimates $\hat{\sigma} = 1.09, \hat{\gamma} = 0.13$. Same analysis for “task durations” gave $\hat{\sigma} = 1.89, \hat{\gamma} = 0.03$
- *Outlook:* Predict survival times beyond follow up time in medical experiments?

Rootzén & Zholud, Technometrics 2016

- E. Gilleland, M.Ribatet, A.G: Stephenson “A software review of extreme value analysis”, *Extremes* 2013
- S. Coles “An introduction to statistical modelling of extreme values”, Springer 2001
- J. Beirlant, Y. Goegebeur, J. Segers, J. Teugles “Statistics of extremes: theory and application” Wiley 2004

- *Not covered includes*: Regular variation, Hill estimator, point process approach, and much more

Tomorrow: wind storm insurance, part 2; landslides; portfolio risk; flu epidemics; - and more theory!