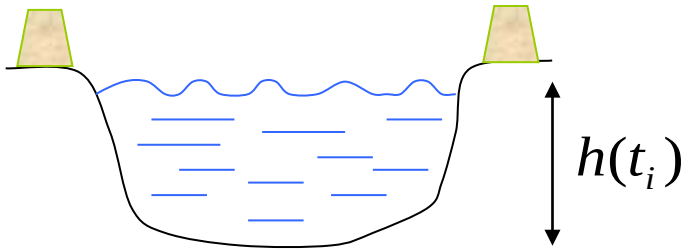
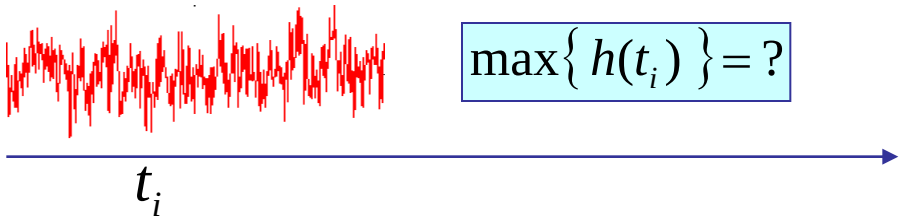
Extreme value-, order- and record statistics

Zoltán Rácz

Institute for Theoretical Physics
Eötvös University
E-mail: racz@general.elte.hu
Homepage: cgl.elte.hu/~racz

Motivation: Do witches exist if there are 2 extreme hurricanes in a century?





~100 years (data) ~200 years (design)

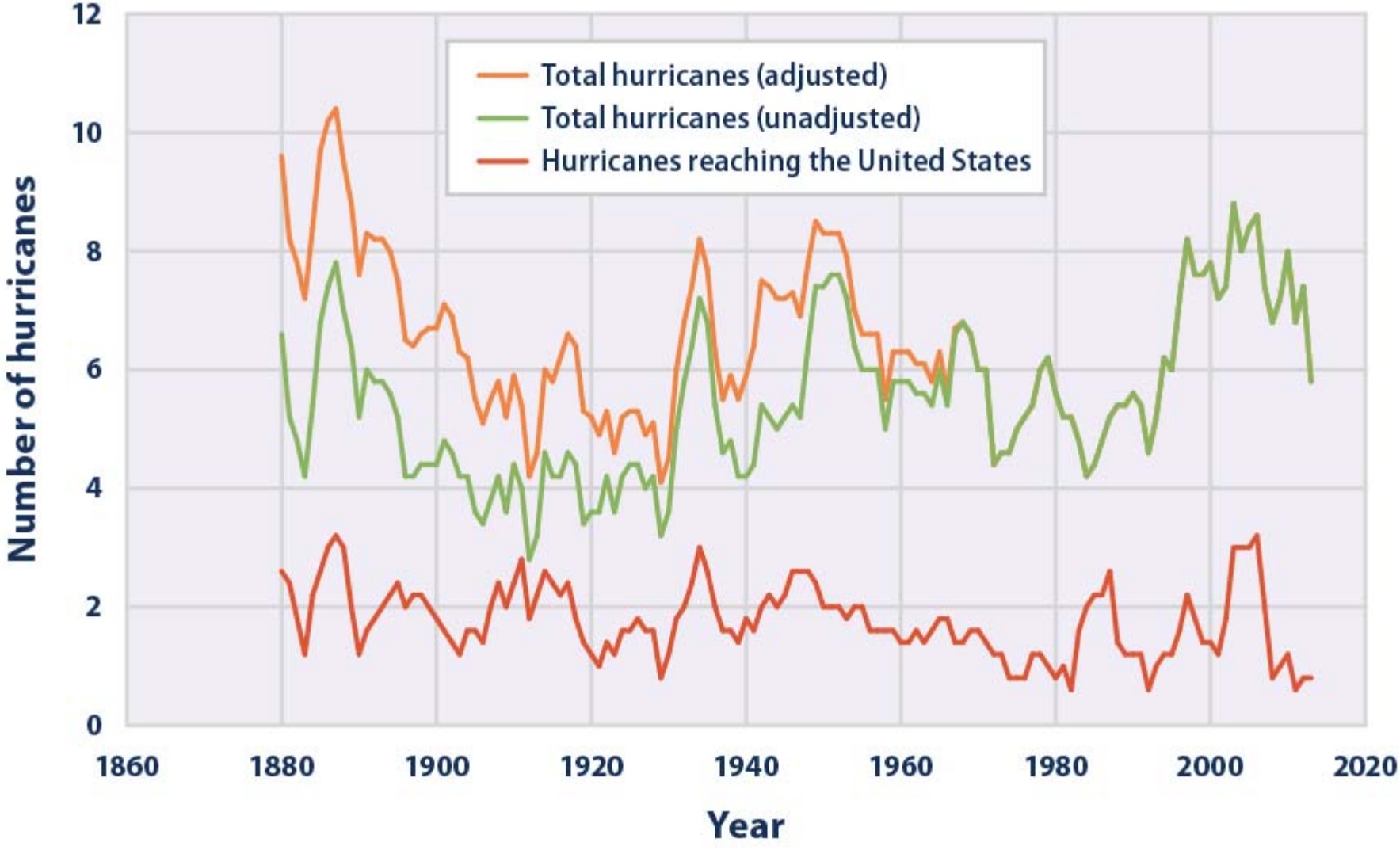
Problems of extrapolating to values where no data exist.

↑

unusually large or small

Question:
Can this be done at all?

Number of Hurricanes in the North Atlantic, 1878–2015



Data sources:

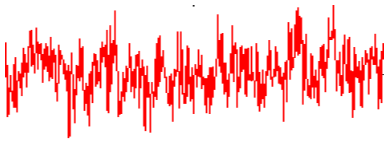
- NOAA (National Oceanic and Atmospheric Administration). 2016. The Atlantic Hurricane Database Re-analysis Project. www.aoml.noaa.gov/hrd/hurdat/comparison_table.html.
- Vecchi, G.A., and T.R. Knutson. 2011. Estimating annual numbers of Atlantic hurricanes missing from the HURDAT database (1878–1965) using ship track density. J. Climate 24(6):1736–1746. www.gfdl.noaa.gov/bibliography/related_files/gav_2010JCLI3810.pdf.

For more information, visit U.S. EPA's "Climate Change Indicators in the United States" at www.epa.gov/climate-indicators.

Examples of extreme value problems I.

winds

$v(t_i)$



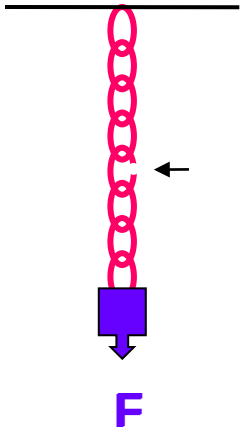
$\max\{ v(t_i) \} = ?$

t_i

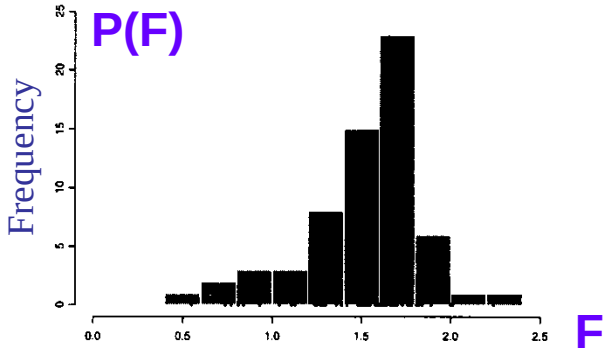
~5 years (data) ~100 years (design)

Q: How long will they stand?

Distribution of the breaking strength of the weakest link



F

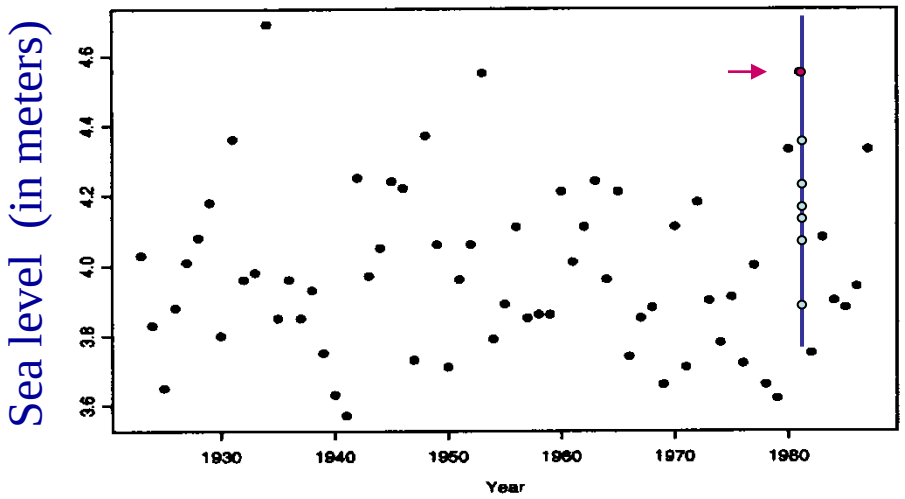


Breaking strength of glass fibres

Q: Can we calculate **P(F)** ?

Examples of extreme value problems II.

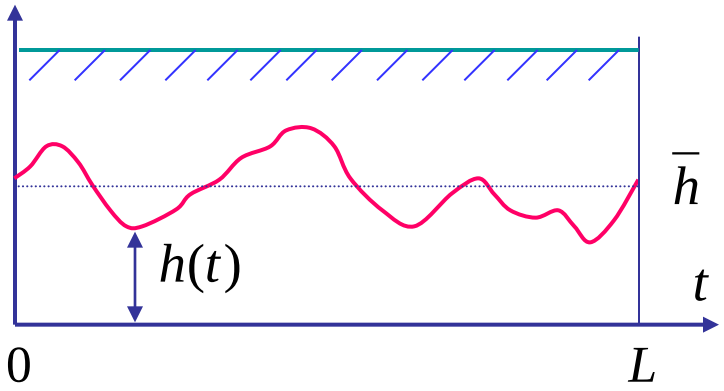
Stuart Coles:
An Introduction to Statistical Modeling of Extreme Values



The 1841 sea level benchmark (centre) on the `Isle of the Dead', Tasmania. According to Antarctic explorer, Capt. Sir James Clark Ross, it marked mean sea level in 1841.

Annual maximum sea levels at Port Pirie, South Australia

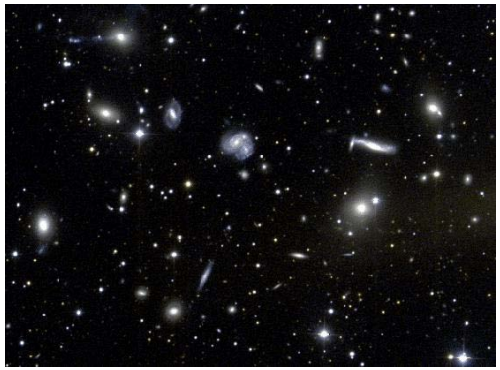
Rusting through



Extremes on a moving boundary

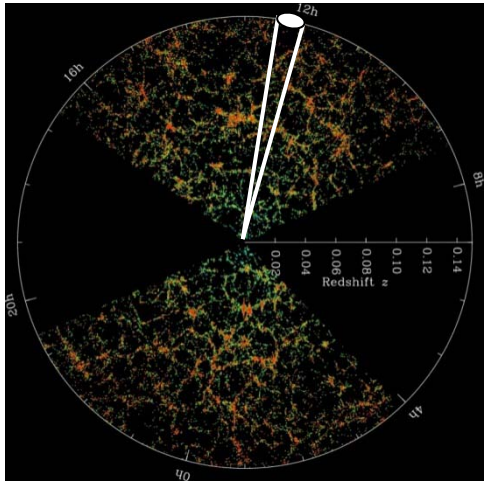
Examples of extreme value problems III.

Standard candles: Using the brightest, 2nd brightest, …



Hercules Cluster

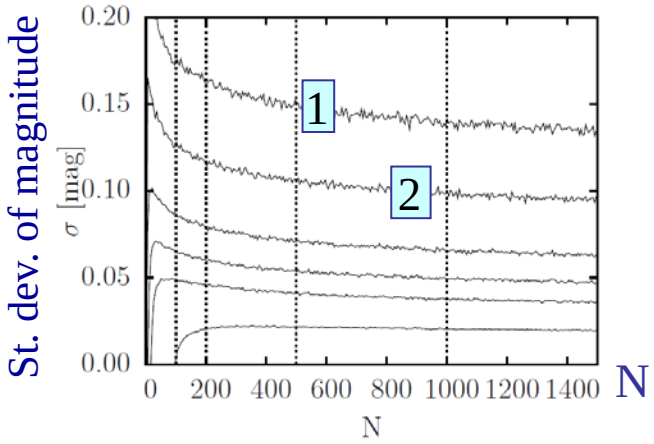
N ~ 100 galaxies



Sloan Digital Sky Survey

Brightest varies slowly with N

Sandage et al, 1956, 1973

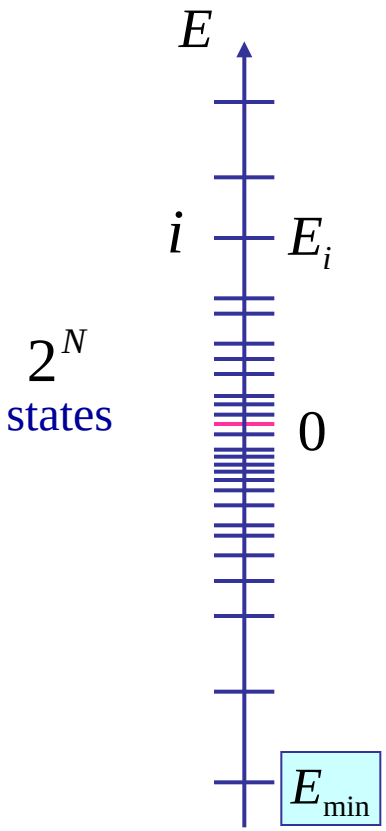


Dobos-Csabai, 2011

(No of galaxies in the cluster)

Random energy model of spin-glasses

B. Derrida (1980)
J.-P. Bouchaud and M. Mézard (1997)



S-K model

$$\mathrm{H} \sim -\sum_{ij} J_{ij}\sigma_i\sigma_j \qquad P(J_{ij}) \sim \exp\left[-\frac{J_{ij}^2 N}{J^2}\right]$$

replaced by

$$\mathrm{P}_0(E_i) \sim \exp\left[-\frac{E_i^2}{J^2 N}\right]$$

Low temperature behavior:

Q: What is the distribution of the ground state energy?

$$u = \frac{\sqrt{\ln 2}}{J} E_{\min} - N \ln 2$$

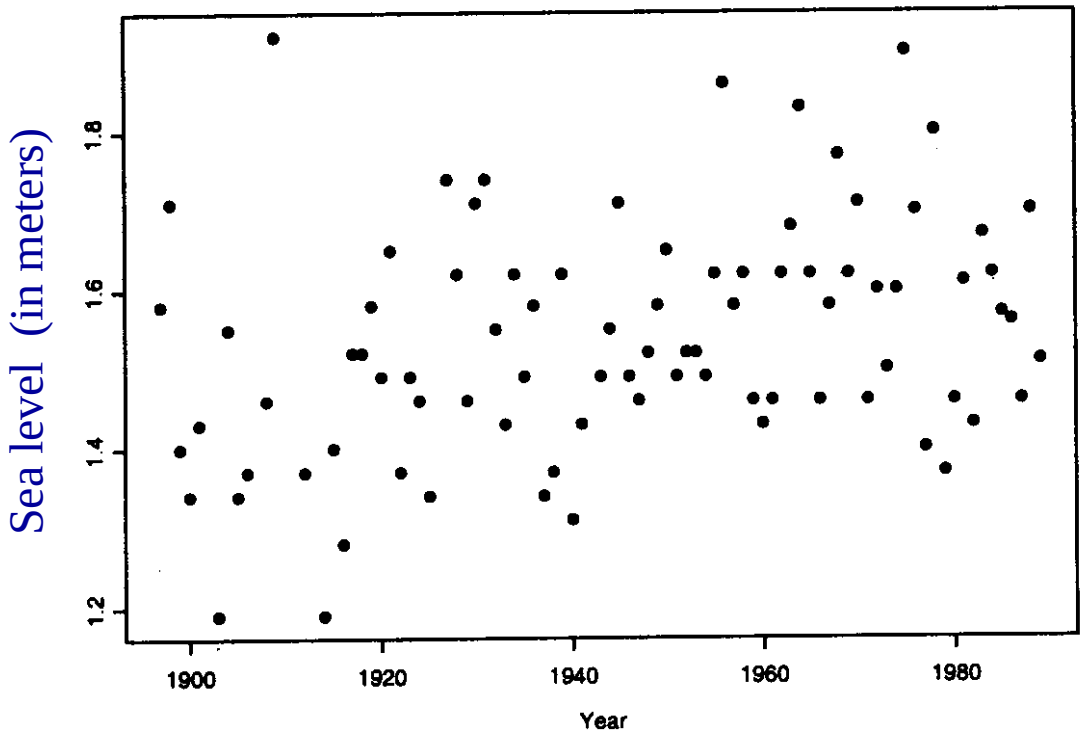
$$P(u) = e^{u+e^u} \sim e^u$$

exponential distribution

correlations?

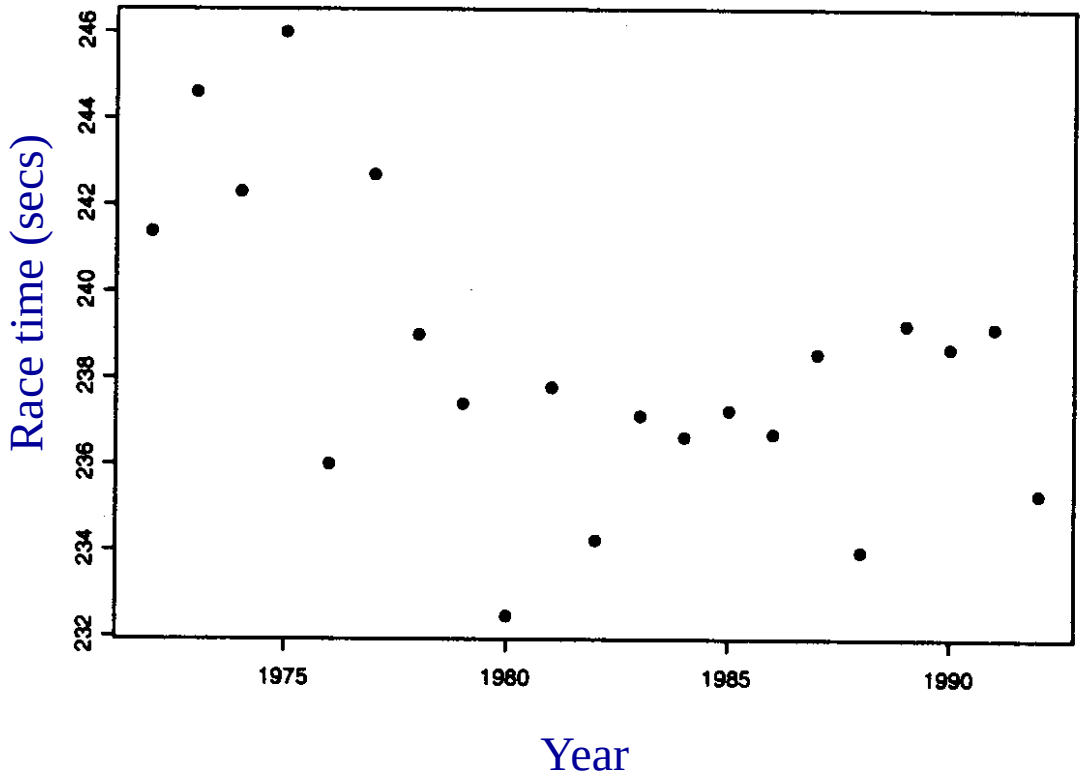
Problems with EVS: Trends I.

Figures are from **Stuart Coles: An Introduction to Statistical Modeling of Extreme Values**



Annual maximum sea levels at Fremantle, Western Australia

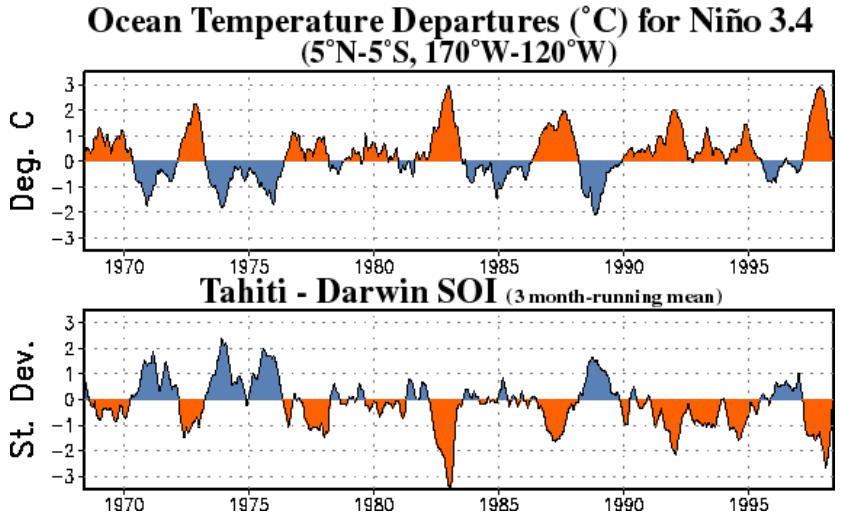
Problems with EVS: Trends II.



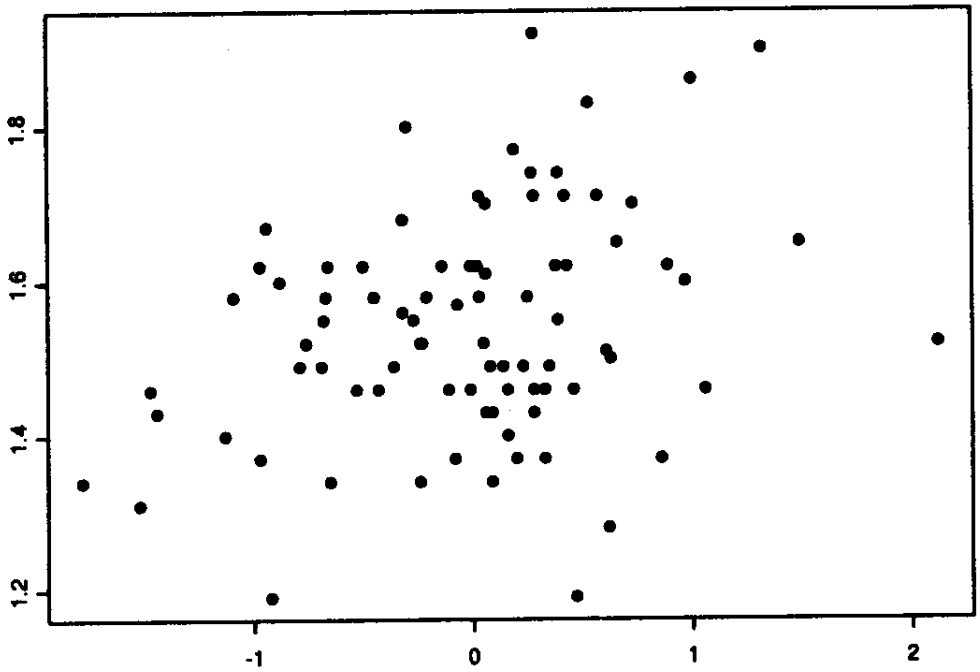
Fastest annual women's 1500 meters race times

Problem of time-correlations

Most of the EVS theory
is about i.i.d. variables but …

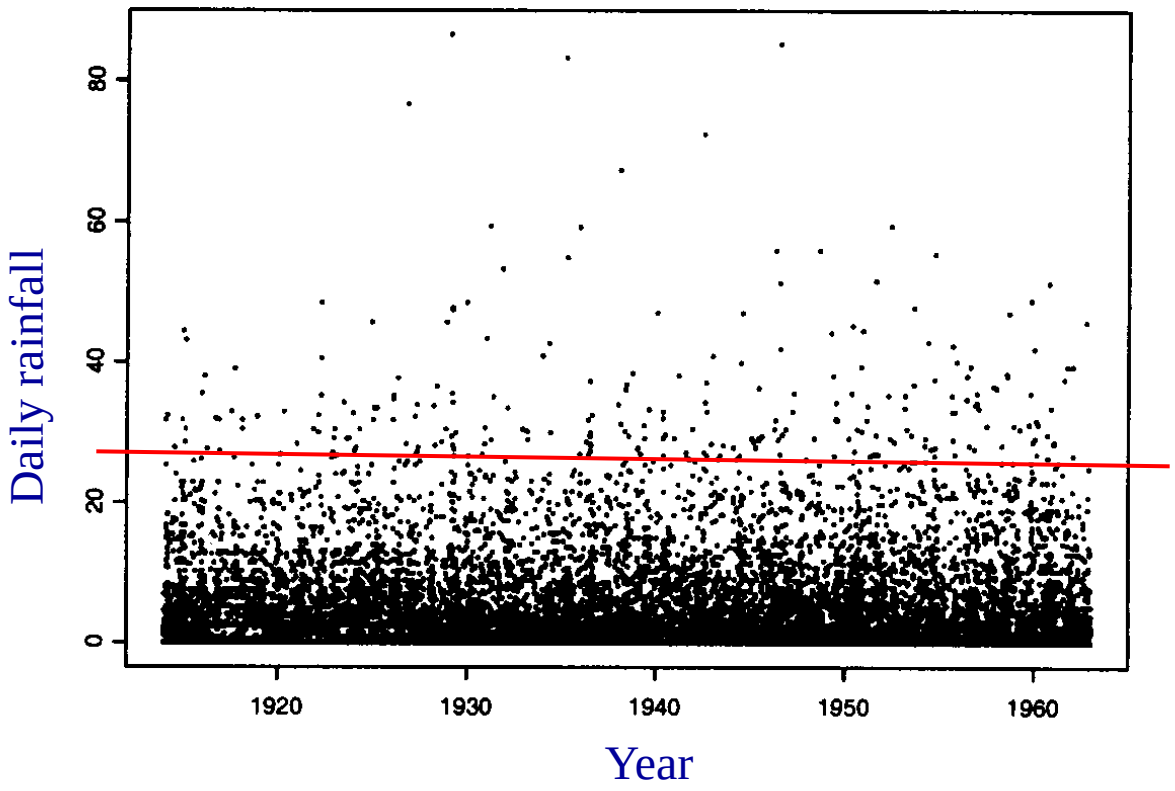


Annual
maximum
sea levels at
Fremantle,
W. Australia



Southern Oscillation Index

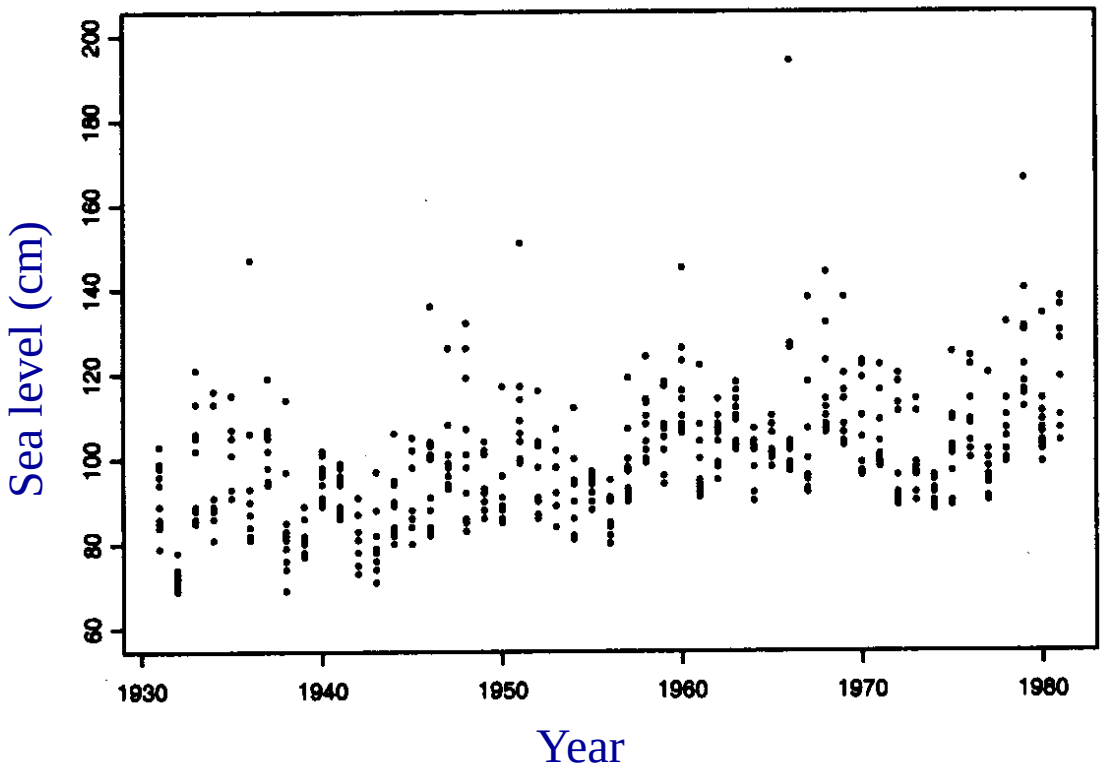
Problem of information loss



This part of the distribution gets lost in EVS

Daily rainfall accumulations

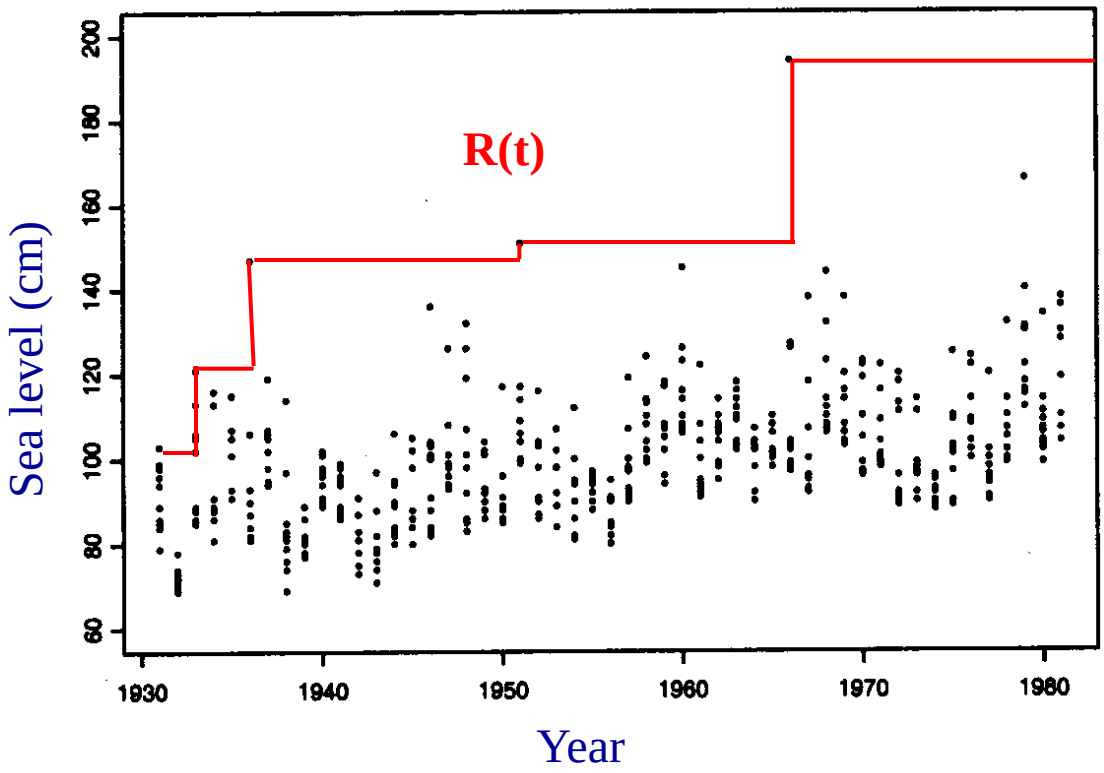
Order statistics: Second-, third-, …, largest values



Largest 10 annual sea-levels in Venice

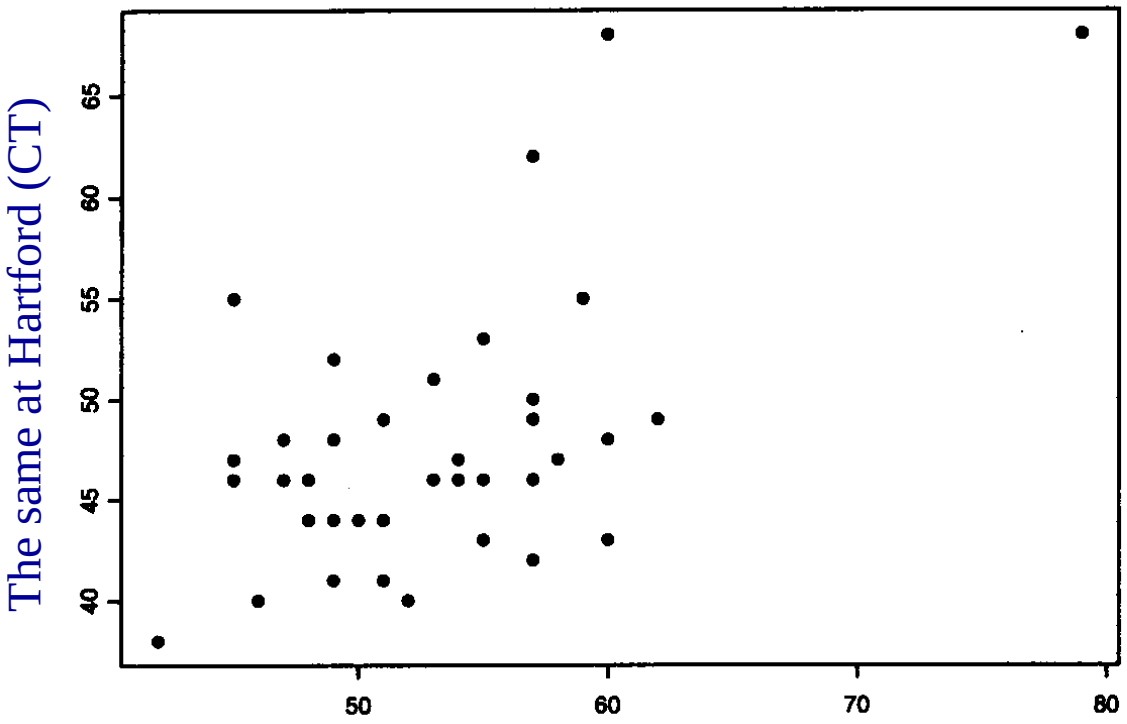
Trying to correct for the loss of interesting, discarded data.

Record statistics



Largest 10 annual sea-levels in Venice

Problem of spatial correlations

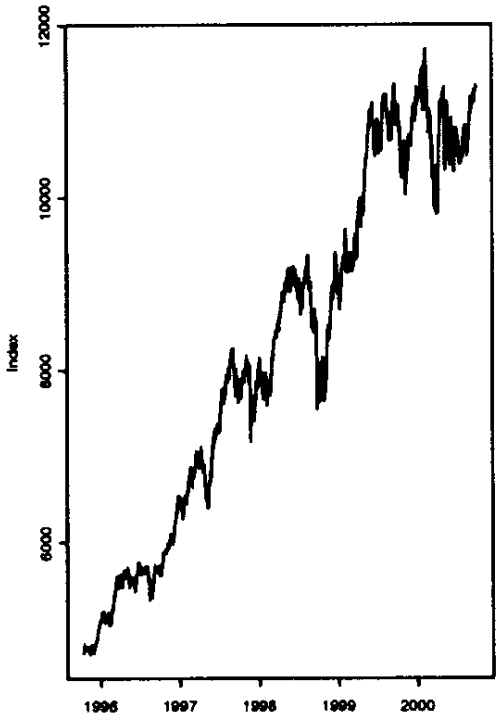


Annual maximal Wind Speed (in knots) at Albany (NY)

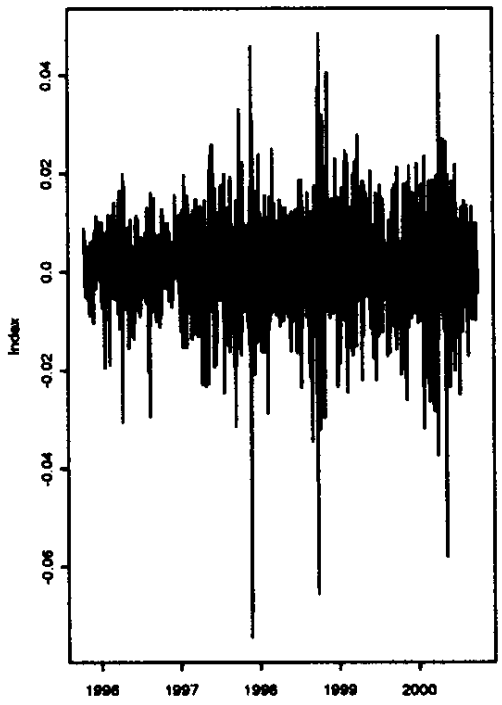
Q: At what distances are the correlations negligible?

Problem of trends and the choice of variables

Searching for independent variables.



Year

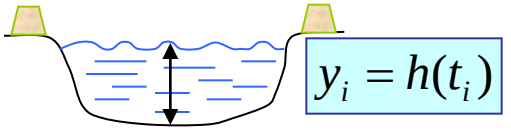


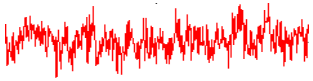
Year

$$\Delta = \ln \frac{M_i}{M_{i-1}}$$

Left panel: daily closing prices of the Dow Jones Index.
Right panel: log daily returns of the same.

Extreme value paradigm

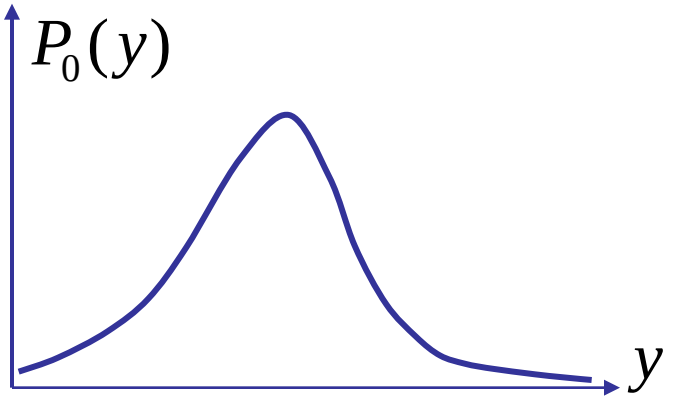




Y is measured: $y_1, y_2, \ldots, y_N$

$$z_N = \max\{y_1, y_2, \ldots, y_N\}$$

Question: What is the distribution of the largest number?



Logics:

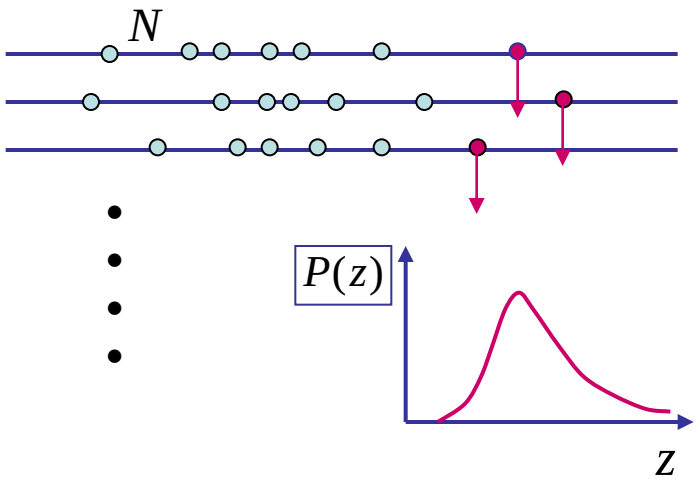
Assume something about y_i

E.g. independent, identically distributed

Use limit arguments: $(N \to \infty)$

A family of limit distributions (models) emerges

Calibrate the family of models by the measured values of z_N .



Slightly suspicious but no alternatives exist at present.

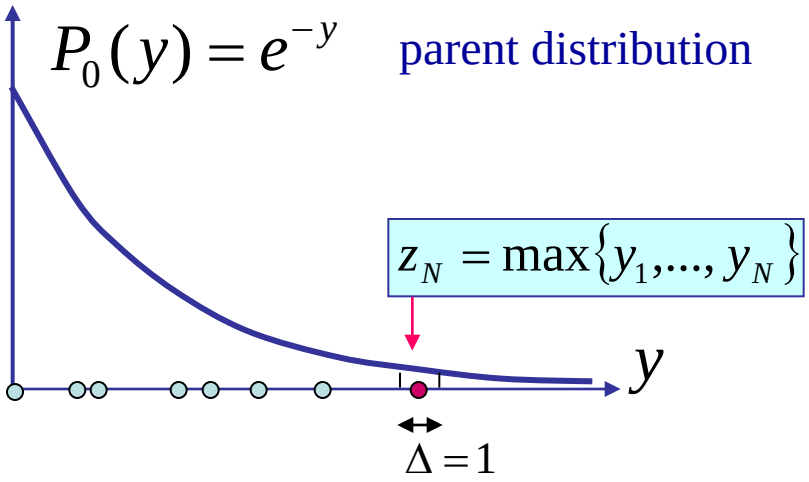
General considerations through an example

Y is measured: $y_1, y_2, \ldots, y_N$

$P_0(y) = e^{-y}$ parent distribution

Assumption: Independent, identically distributed random variables with

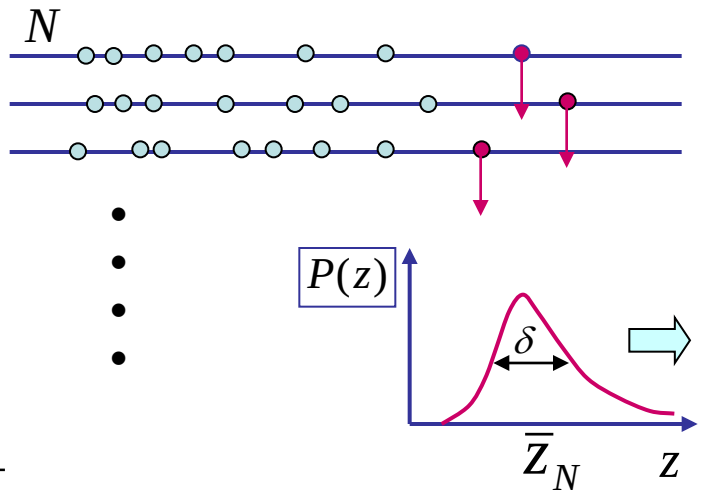
$z_N = \max\{y_1, \ldots, y_N\}$



1st question: Can we estimate $\bar{z}_N$?

$$P_0(\bar{z}_N)\Delta \cdot N \approx e^{-\bar{z}_N} N \approx 1$$

$$\bar{z}_N \approx \ln N$$



Note: $\bar{z}_N \xrightarrow[N\to\infty]{} \infty$

2nd question: Can we estimate $\delta^2 = \overline{(z - \bar{z}_N)^2}$?

$$P_0(\bar{z}_N + \delta)\Delta \cdot N \approx 1/e \quad \rightarrow \quad \delta \approx 1$$

Homework: Carry out the above estimates for a Gaussian parent distribution $P_0(y) \approx e^{-y^2}$!

Fisher-Tippett-Gumbel distribution

Fisher & Tippett (1928)

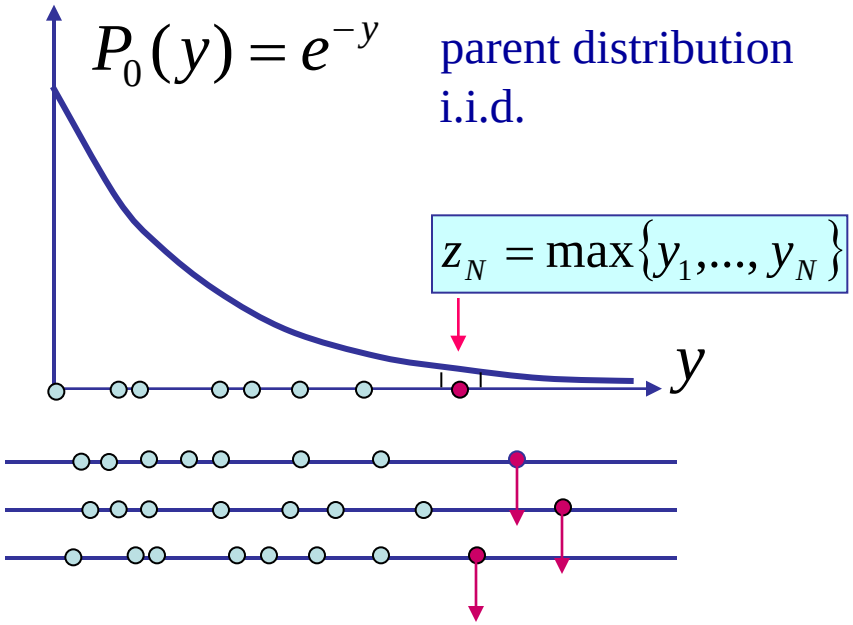
Q1: Can we calculate $P_N(z_N)$?

Q2: Is there a limit distribution?

$$\lim_{N\to\infty} P_N(a_N x + b_N) = P(x)$$

$P_0(y) = e^{-y}$ parent distribution i.i.d.

$z_N = \max\{y_1, \ldots, y_N\}$

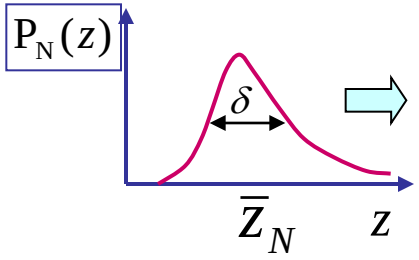


Probability of $z_N < z$:

$$M_N(z) = \int_{-\infty}^{z} P_N(z_N)\,dz_N = \left[\int_0^z P_0(y)\,dy\right]^N$$

$$M_N(z) = (1-e^{-z})^N = \left[1 - \frac{e^{-(z-\ln N)}}{N}\right]^N = \left[1 - \frac{e^{-x}}{N}\right]^N = e^{-e^{-x}}$$

$x = z - \ln N$



$$P(x) = \lim_{N\to\infty} \frac{d}{dx} M_N(z = x + \ln N) = e^{-x-e^{-x}}$$

FTG distribution

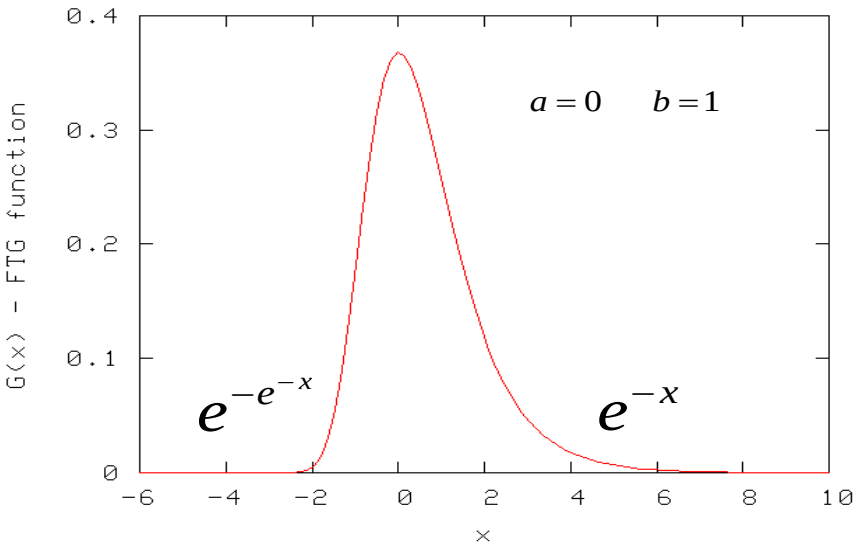
Shape of the FTG function

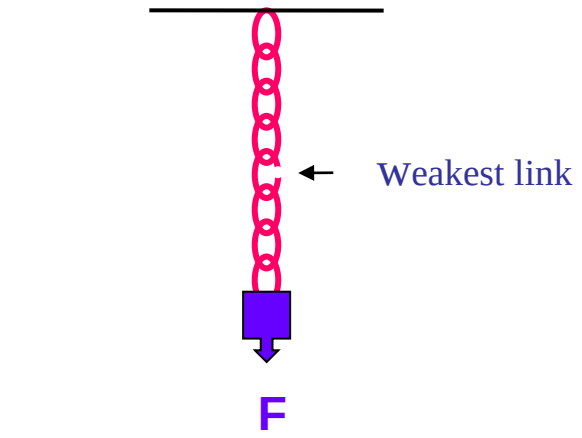
Fisher & Tippett (1928)
Gumbel (1950)

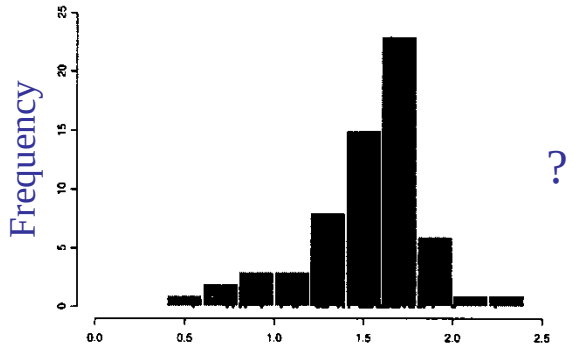
$$P(x) = \frac{1}{b} e^{-\frac{x-a}{b} - e^{-\frac{x-a}{b}}}$$

location parameter

scale parameter







?

Breaking strength of glass fibres

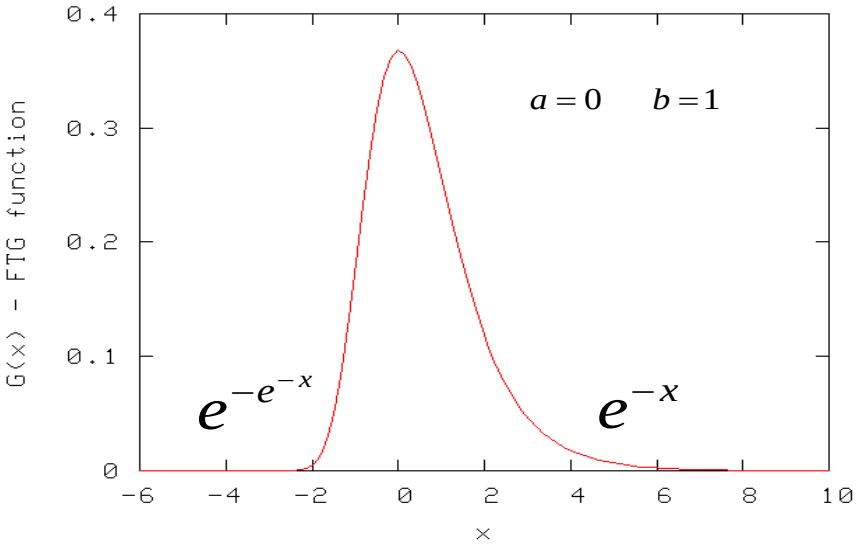
Shape of the FTG function

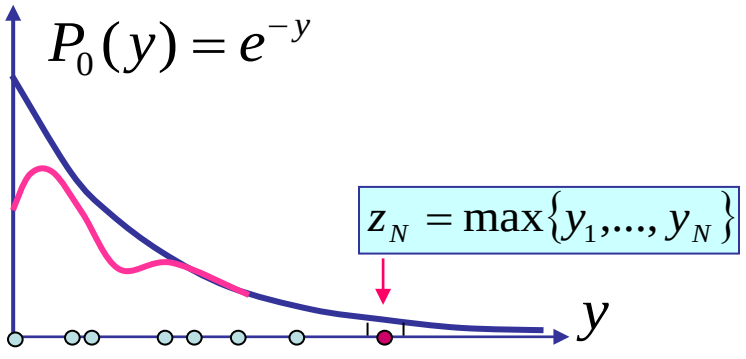
Fisher & Tippett (1928)

$$P(x) = \frac{1}{b} e^{-\frac{x-a}{b} - e^{-\frac{x-a}{b}}}$$

location parameter

scale parameter





The limit distribution should not depend on small y details of $P_0(y)$ but there is more generality to this result.

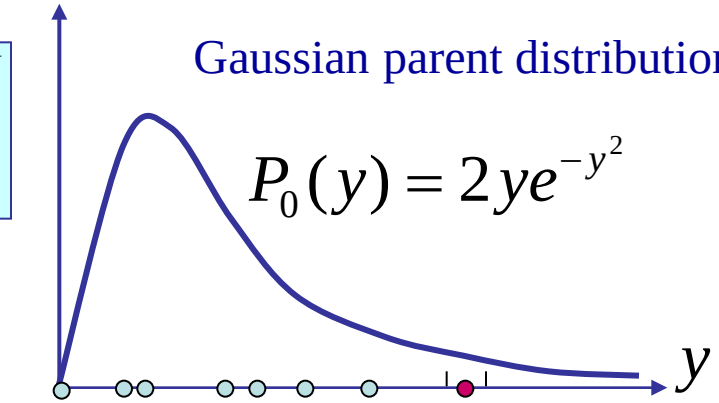
FTG distribution – from Gaussian parent

Probability of $z_N < z$:

$$M_N(z) = \int_{-\infty}^{z} P_N(z_N)\,dz_N = \left[\int_{0}^{z} P_0(y)\,dy\right]^N$$

Gaussian parent distribution

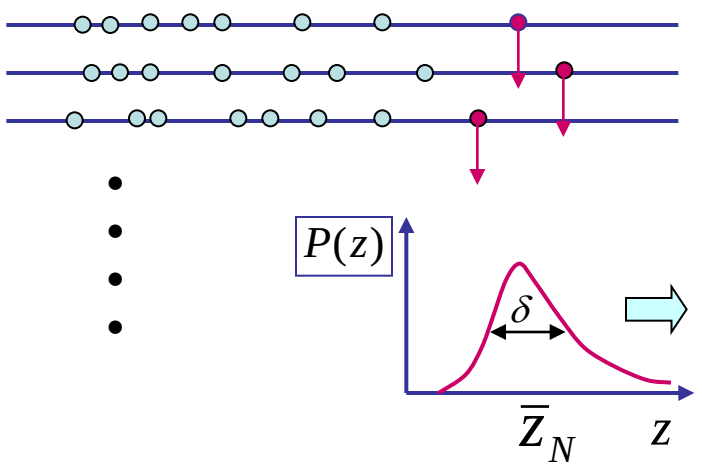
$$P_0(y) = 2ye^{-y^2}$$



$$M_N(z) = (1-e^{-z^2})^N$$

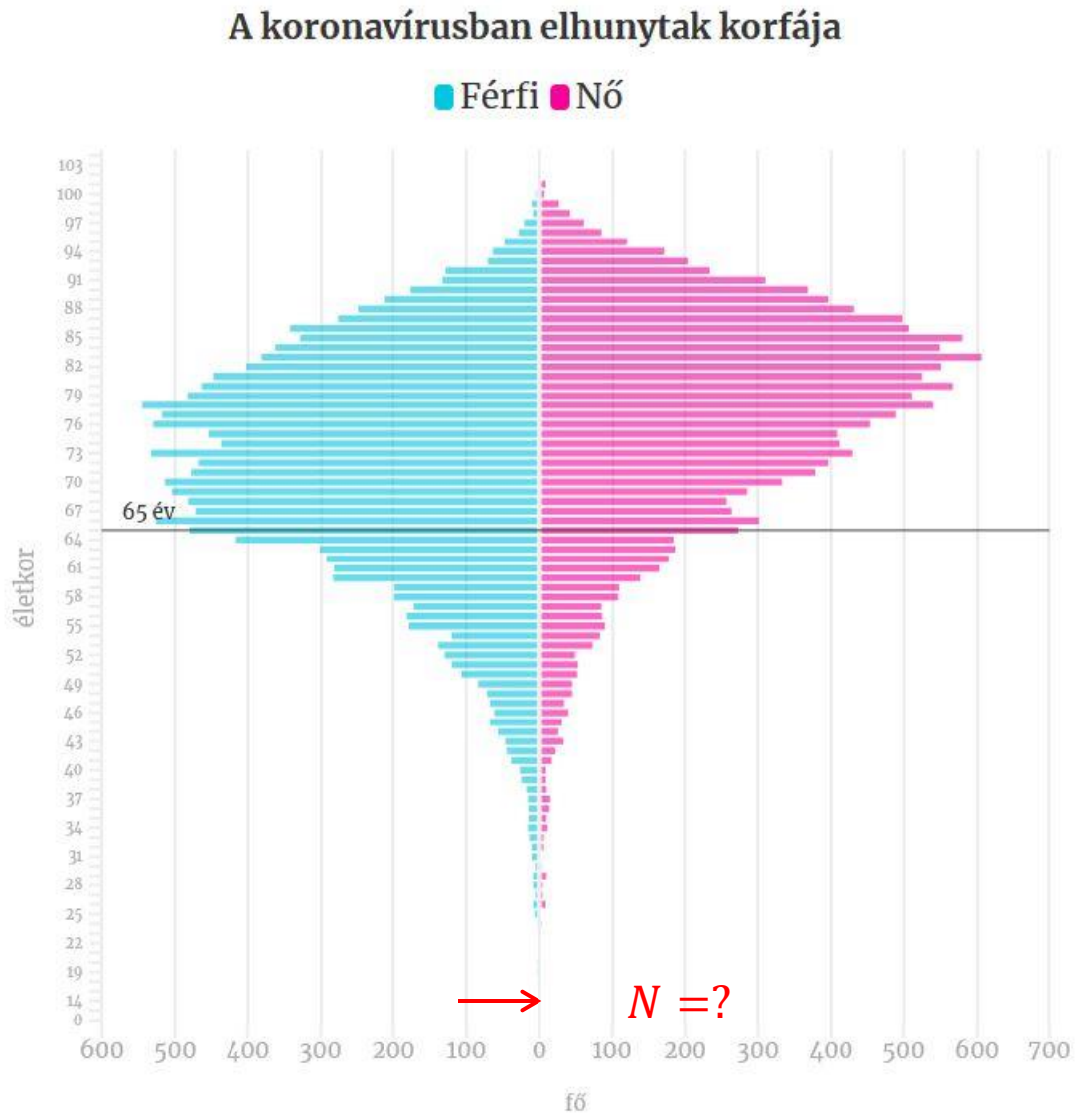
$$= \left[1-\exp\left[-\left[\frac{x}{2\sqrt{\ln N}} + \sqrt{\ln N}\right]^2\right]\right]^N$$

$$= \left[1-\frac{\exp\left[-x-x^2/4\ln N\right]}{N}\right]^N = e^{-e^{-x}}$$



$$P(x) = \lim_{N\to\infty} \frac{d}{dx} M_N\left(z = x/2\sqrt{\ln N} + \sqrt{\ln N}\right) = e^{-x-e^{-x}}$$

FTG distribution

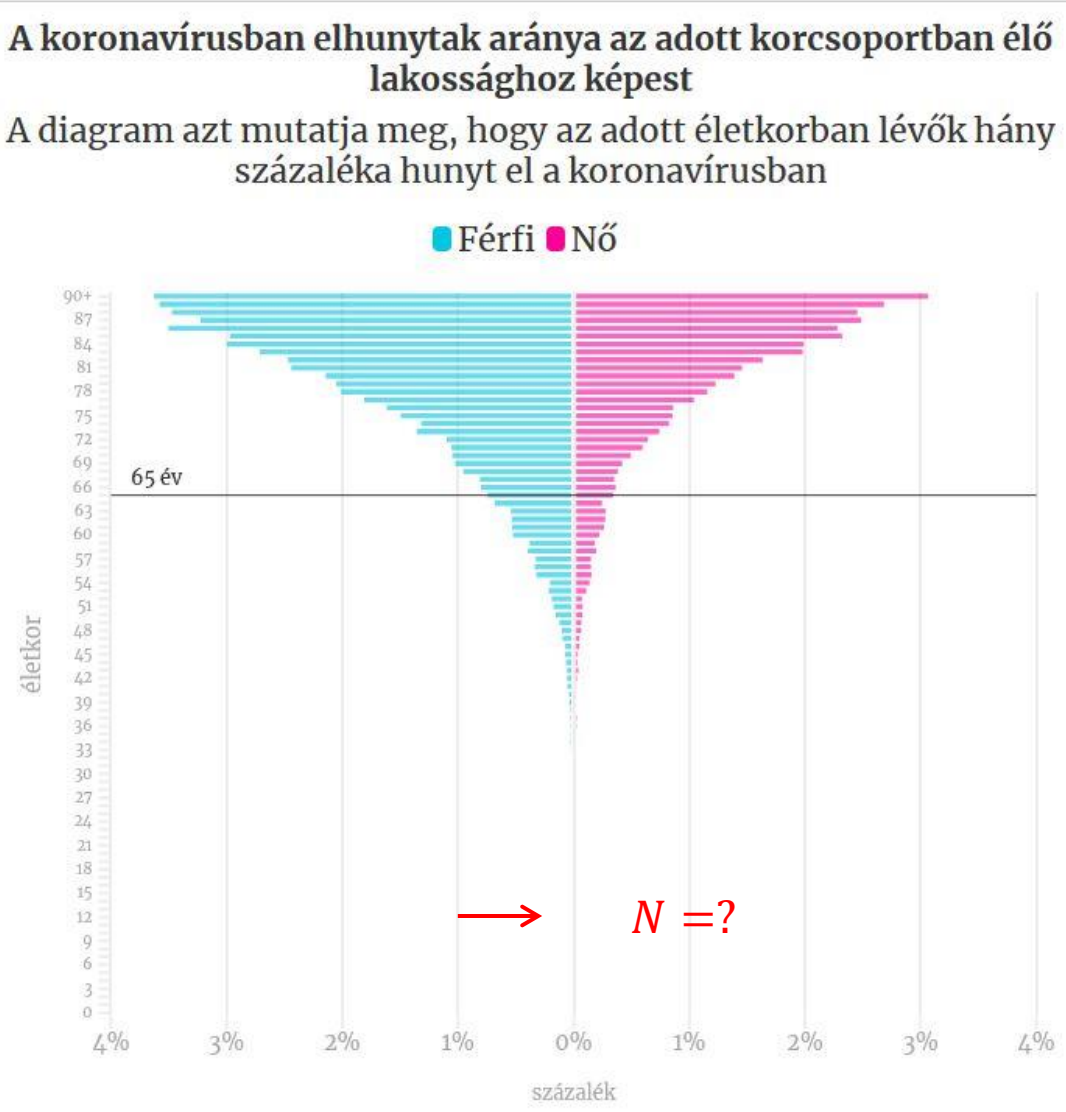


Adat forrása: koronavirus.gov.hu, Adatok letöltése

Halottak száma:

$N \approx 15000$ nő
≈ 15000 férfi

A háttérben van egy eloszlás – ez az, amit megfigyelünk, s fittelünk. Ezután feltételezve az események i.i.d. karakterét, meghatározhatjuk egy extrém események (pl. a legfiatalabb halott korának) időbeli (N-függő) fejlődését.



Halottak száma:

$N \approx 15000$ nő

≈ 15000 férfi

A háttérben van egy eloszlás – ez az, amit megfigyelünk, s fittelünk. Ezután feltételezve az események i.i.d. karakterét, meghatározhatjuk egy extrém események (pl. a legfiatalabb halott korának) időbeli (N-függő) fejlődését.

Extreme value statistics: i.i.d. variables

Y is measured: $y_1, y_2, \ldots, y_N$

$$z_N = \max\{y_1, y_2, \ldots, y_N\}$$

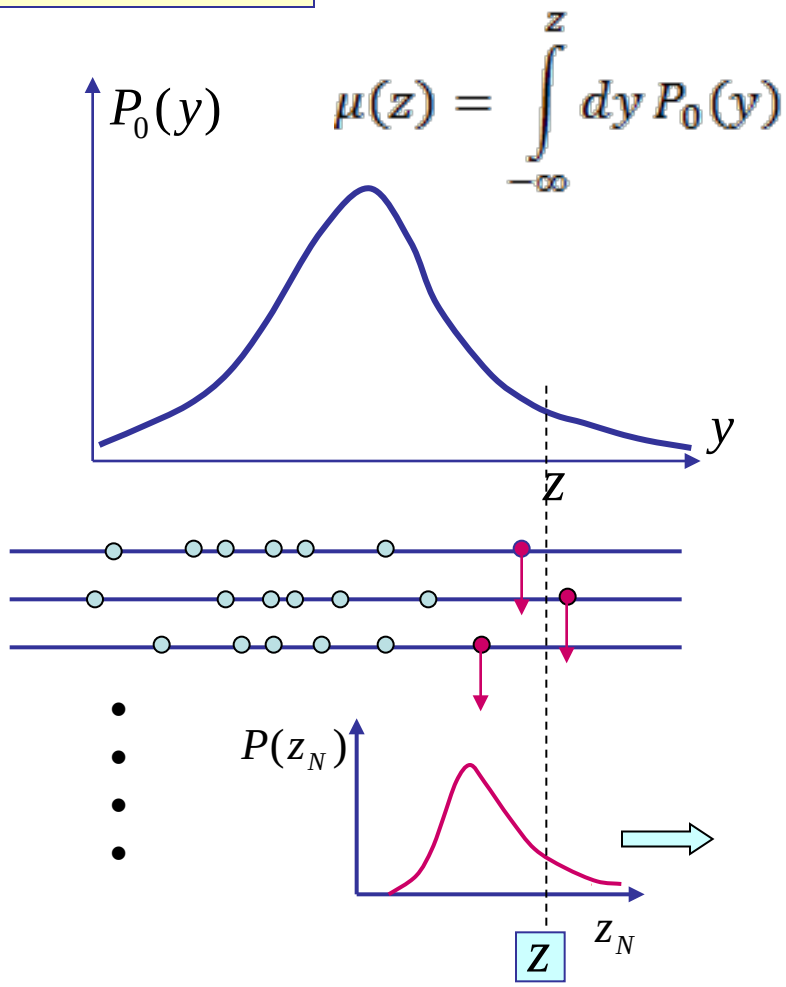
$M_N(z)$ probability of $z_N < z$

$$M_N(z) = [\mu(z)]^N$$

Question: Is there a limit distribution for $N \to \infty$?

$$\lim_{N\to\infty} M_N(a_N x + b_N) =$$

$$\lim_{N\to\infty} [\mu(a_N x + b_N)]^N = M(x)$$

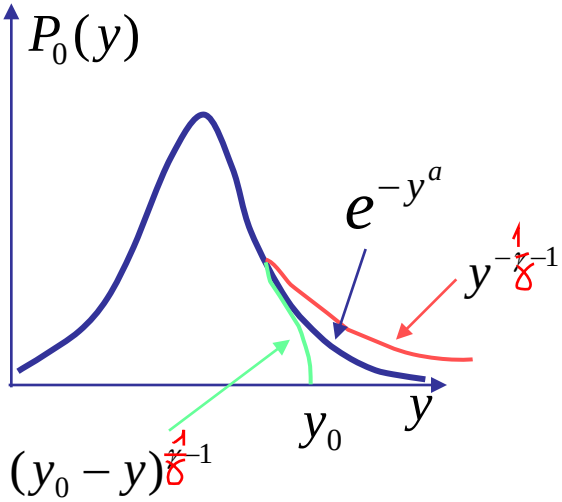


Result: Three possible limit distributions depending on the tail of the parent distribution, $P_0(y)$.

$$z = a_N x + b_N$$

Extreme value limit distributions: i.i.d. variables

Fisher & Tippett (1928)
Gnedenko (1941)
J. Galambos (1978)



- Fisher-Tippett-Gumbel (exponential tail)

$$M_{FTG}(x) = \exp(-\exp(-x))$$

- Fisher-Tippett-Frechet (power law tail)

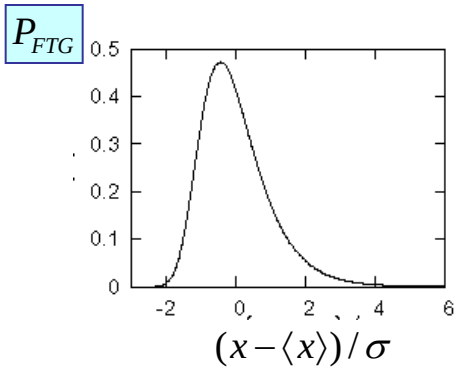
$$M_{FTF}(x) = \exp(-x^{-1/\gamma}) \quad x > 0$$
$$= 0 \quad x < 0$$

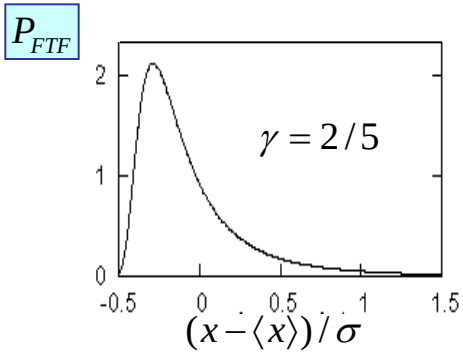
- Weibull (finite cutoff)

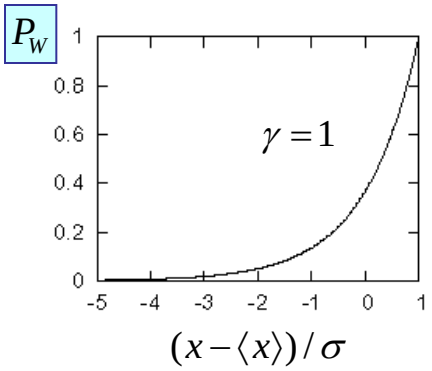
$$M_W(x) = \exp(-(-x)^{1/\gamma}) \quad x < 0$$
$$= 1 \quad x > 0$$

Characteristic shapes of probability densities:

$$P_I(x) = dM_I(x)/dx$$







HF2: Picture gallery as a function of γ. Do we understand the trends?

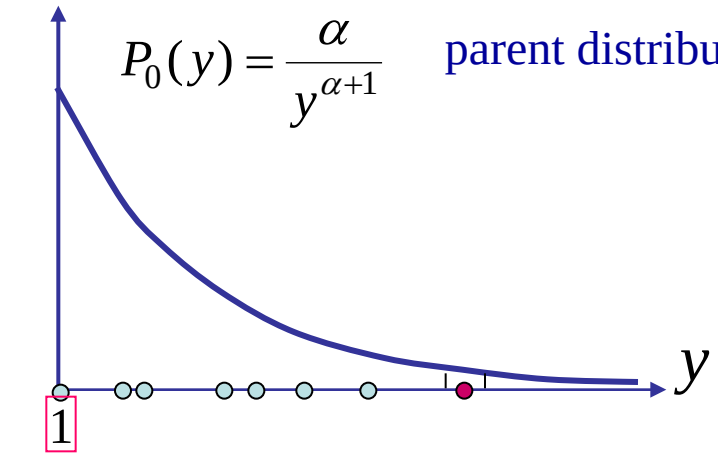
Fisher-Tippett-Frechet distribution

Y is measured: $y_1, y_2, \ldots, y_N$

$P_0(y) = \dfrac{\alpha}{y^{\alpha+1}}$ parent distribution

Assumption: Independent, identically distributed random variables with

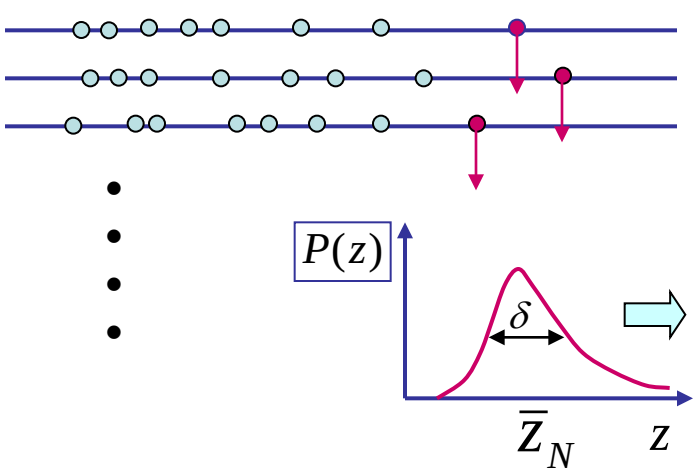
$$z_N = \max\{y_1, y_2, \ldots, y_N\}$$



Question: Can we calculate $P_N(z_N)$?

Probability of $z_N < z$:

$$M_N(z) = \int_{-\infty}^{z} P_N(z_N)\,dz_N = \left[\int_0^z P_0(y)\,dy\right]^N$$



$$M(z) = \left(1 - \frac{1}{z^\alpha}\right)^N = \left(1 - \frac{1}{N(N^{-1/\alpha}z)^\alpha}\right)^N$$

$$x = N^{-1/\alpha} z$$

$$= \left(1 - \frac{1}{Nx^\alpha}\right)^N = \exp(-x^{-\alpha})$$

$$P(x) = \lim_{N\to\infty} \frac{d}{dx} M_N(z = xN^{+1/\alpha}) = \alpha x^{-\alpha-1} \exp(-x^{-\alpha})$$

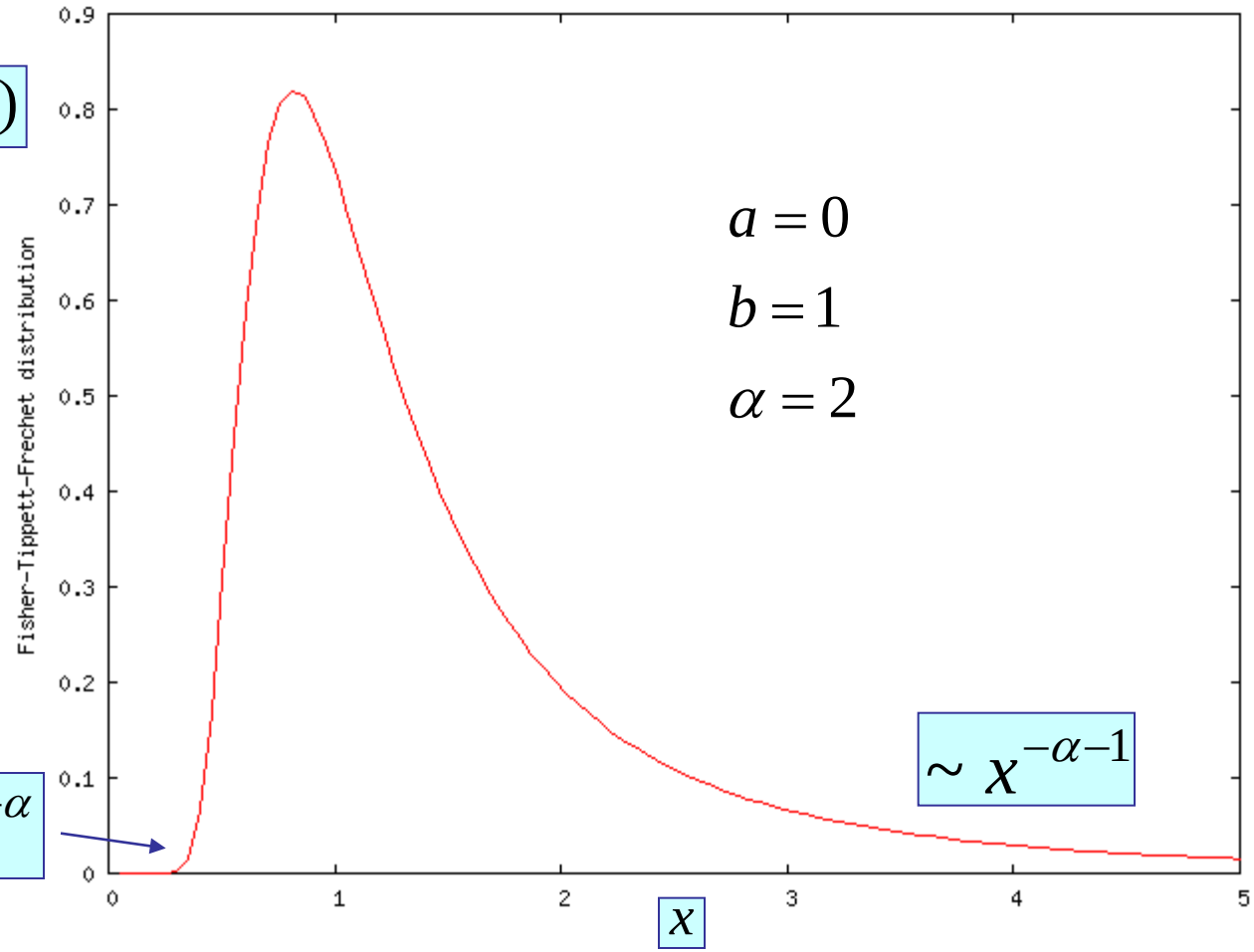
FTF distribution

$x \geq 0$

Shape of the FTF distribution

$$P(x) = \alpha \left[\frac{x-a}{b} \right]^{-\alpha-1} e^{-\left[\frac{x-a}{b} \right]^{-\alpha}}$$

$x \geq a$



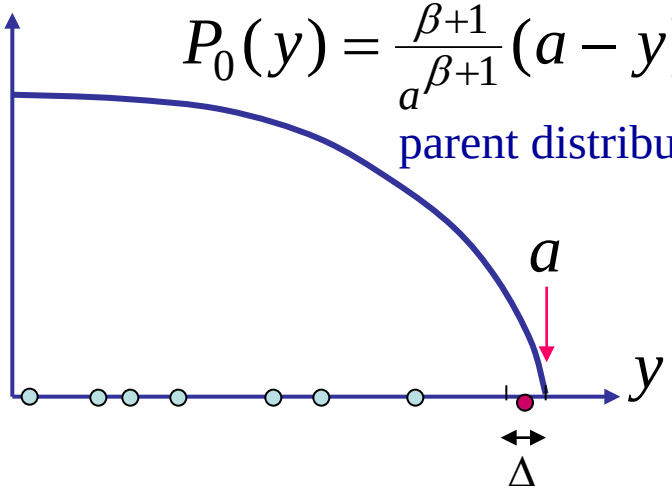
Finite cutoff: Weibull distribution

Y is measured: $y_1, y_2, \ldots, y_N$

Assumption: Independent, identically distributed random variables with

$$P_0(y) = \frac{\beta+1}{a^{\beta+1}}(a-y)^{\beta}$$

parent distribution

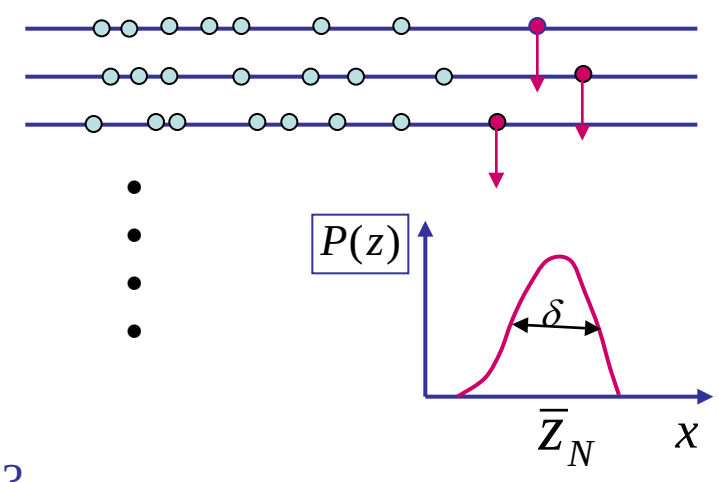


$$z_N = \max\{y_1, y_2, \ldots, y_N\} \xrightarrow[N\to\infty]{} a$$

1st question: Can we estimate $a - \bar{z}_N$?

$$P_0(\bar{z}_N)\Delta \cdot N \approx 1 \qquad \Delta \approx a - \bar{z}_N$$

$$a - \bar{z}_N \approx N^{-1/(\beta+1)}$$



2nd question: Can we estimate $\delta^2 = \overline{(z-\bar{z}_N)^2}$?

$$P_0(a) = 0 \quad \rightarrow \quad \delta \approx a - \bar{z}_N \quad \rightarrow \quad \delta \approx N^{-1/(\beta+1)}$$

Weibull distribution II

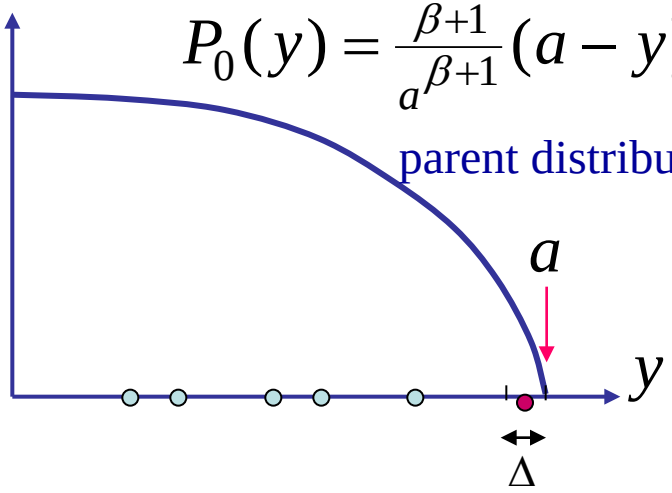
Y is measured: $y_1, y_2, \ldots, y_N$

Assumption: Independent, identically distributed random variables with

$$z_N = \max\{y_1, y_2, \ldots, y_N\}$$

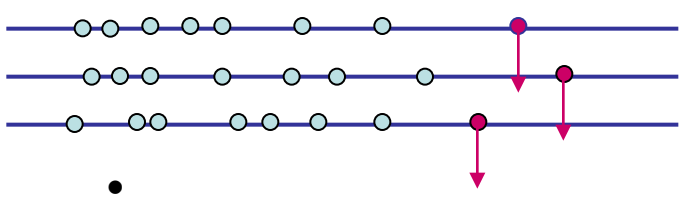
$$P_0(y) = \frac{\beta+1}{a^{\beta+1}}(a-y)^{\beta}$$

parent distribution

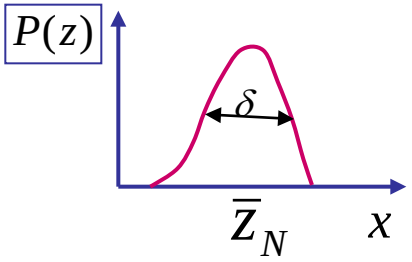


Probability of $z_N < z$:

$$M_N(z) = \int_{-\infty}^{z} P_N(z_N)\,dz_N = \left[\int_0^z P_0(y)\,dy\right]^N$$



$$M(z) = \left[1-(1-\tfrac{z}{a})^{\beta+1}\right]^N = (1-(-x)^{\beta+1}/N)^N = e^{-(-x)^{\beta+1}}$$



$$a - \bar{z}_N \approx N^{-1/(\beta+1)}$$

$$z = a + \alpha x N^{-1/(\beta+1)}$$

$$P(x) = \lim_{N\to\infty} \frac{d}{dx} M_N(z = a + \alpha x N^{-1/(\beta+1)}) = (\beta+1)(-x)^{\beta}\exp[-(-x)^{\beta+1}] \qquad x \le 0$$

Weibull distribution

$$P(x) = 0 \qquad x \ge 0$$

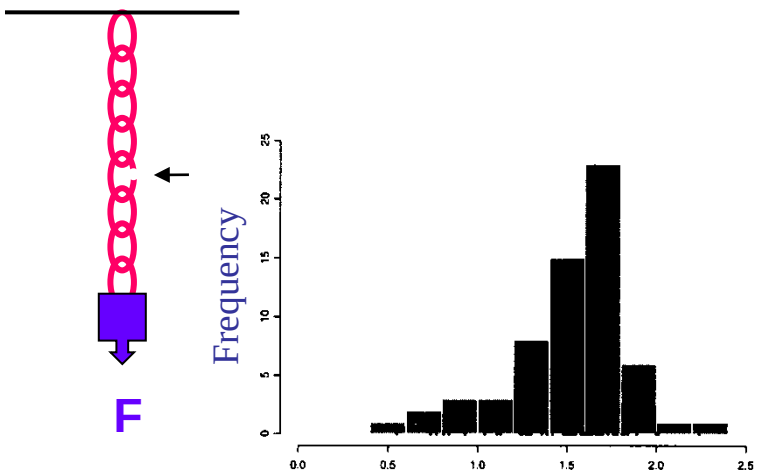
Weibull function and fitting

$$P(x) = \begin{cases} \frac{\beta+1}{b}\left(\frac{a-x}{b}\right)^{\beta} e^{-\left(\frac{a-x}{b}\right)^{\beta+1}} & x \le a \\ 0 & x > a \end{cases}$$

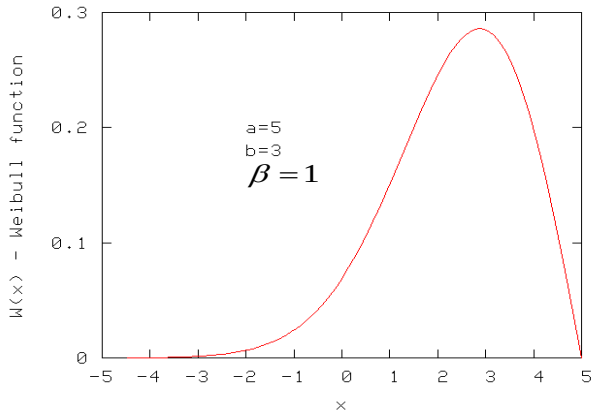
$z_N = \max\{y_1, y_2, \ldots, y_N\}$ is measured.

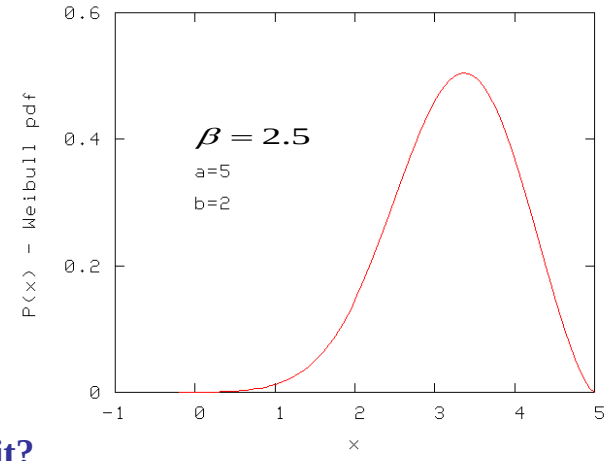
N is not known!

Question:
What is the fitting procedure?



Breaking strength of glass fibres

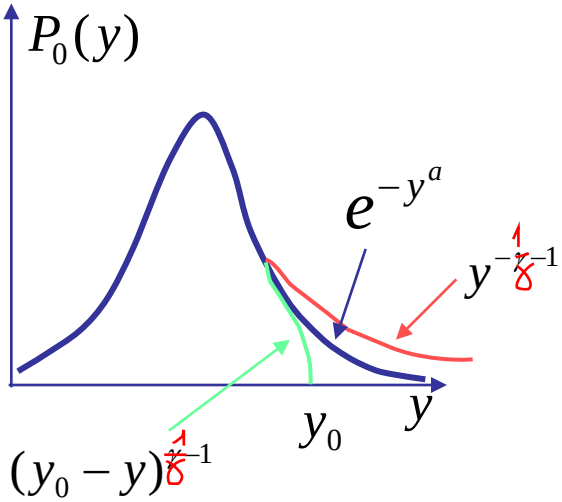




HF.3: Is there a better fit?

Extreme value limit distributions: i.i.d. variables

Fisher & Tippett (1928)
Gnedenko (1941)
J. Galambos (1978)



- Fisher-Tippett-Gumbel (exponential tail)

$$M_{FTG}(x) = \exp(-\exp(-x))$$

- Fisher-Tippett-Frechet (power law tail)

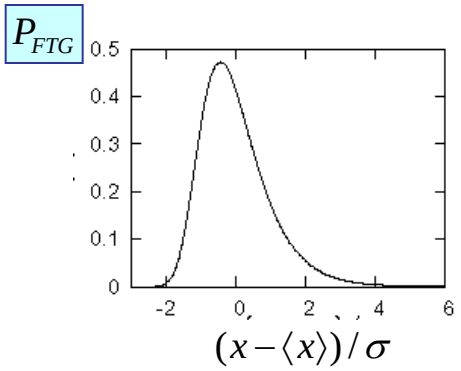
$$M_{FTF}(x) = \exp(-x^{-1/\gamma}) \quad x > 0$$
$$= 0 \quad x < 0$$

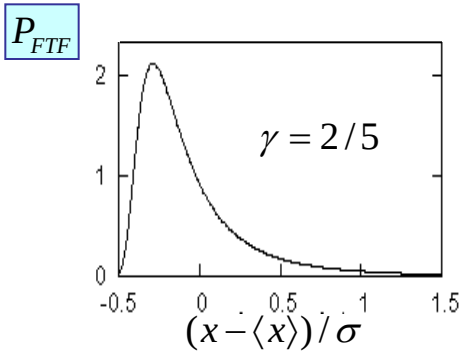
- Weibull (finite cutoff)

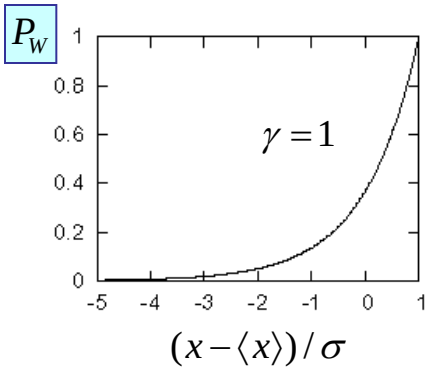
$$M_W(x) = \exp(-(-x)^{1/\gamma}) \quad x < 0$$
$$= 1 \quad x > 0$$

Characteristic shapes of probability densities:

$$P_I(x) = dM_I(x)/dx$$





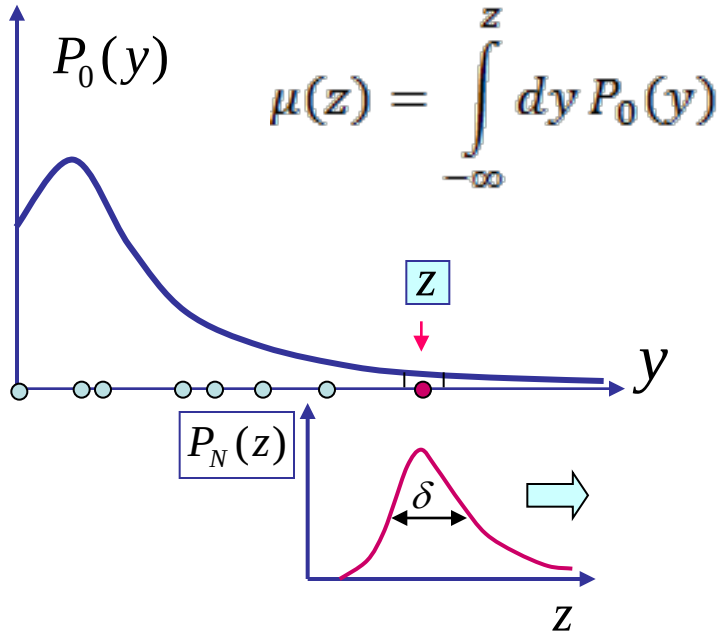


HF2: Picture gallery as a function of γ. Do we understand the trends?

Extreme value statistics and renormalization group

Q: Limit distribution for $N \to \infty$?

$$\lim_{N\to\infty} [\mu(z)]^N \Big|_{z=a_N x + b_N} = M(x)$$



Fix point condition: $N \to N' = pN$

$$[M(a_p x + b_p)]^p = M(x)$$

Interpretation:

$$\Re_p[M(z)] = [M(z)]^p = [M(a_p x + b_p)]^p = M(x)$$

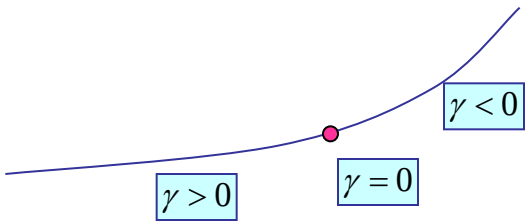
transformation — scale change — shift

$$M(x) = \exp\left[-(1+\gamma x)^{-1/\gamma}\right]$$

How to get γ?

$$g(z) = -\ln[-\ln F_0(z)]$$

$$g''/(g')^2\Big|_{z\to z^*} = -1/\gamma$$



Solving the RG fix point equation I.

For arbitrary p, we have $M(x) = \left[M(a_p x + b_p)\right]^p$

Let $M(x) = \exp[-\exp(-f(x))]$

Standardization (shift and scaling) in the x ~ O(1) variable:

$$M(0) = M'(0) = 1/e \quad \rightarrow$$

$$f(x) = f(a_p x + b_p) + \ln p$$

$$f(0) = 0 \qquad f'(0) = 1$$

Solution:

$$f'(x) = a_p f'(a_p x + b_p)$$

The right hand side should not depend on p.

$$f'(x) = \frac{1}{1+\gamma x}$$

where $\gamma = \dfrac{a_p - 1}{b_p}$

Solving the RG fixed point equation II.

$$f(x) = f(a_p x + b_p) + \ln p$$

$$f(0) = 0 \qquad f'(0) = 1$$

$$\rightarrow \quad f'(x) = \frac{1}{1+\gamma x}$$

$$f(x) = \frac{1}{\gamma}\ln(1+\gamma x) + C$$

$C = 0$ from $f(0) = 0$

Generalized **E**xtreme **V**alue statistics

$$M(x) = \exp[-\exp(-f(x))] = \exp[-(1+\gamma x)^{-1/\gamma}]$$

Probability density:

$$P(x) = (1+\gamma x)^{-(1+\gamma)/\gamma} \exp[-(1+\gamma x)^{-1/\gamma}]$$

$\gamma \to 0$ limit

$$M(x) = \exp[-\exp(-x)] \qquad P(x) = \exp[-x - \exp(-x)]$$

Solving the RG fix point equation III.

Generalized **E**xtreme **V**alue statistics $M(x) = \exp[-(1+\gamma x)^{-1/\gamma}]$

$0 < \gamma$

$-1/\gamma < x$

$$P(x) = (1+\gamma x)^{-(1+\gamma)/\gamma} \exp[-(1+\gamma x)^{-1/\gamma}]$$ Frechet

$\gamma < 0$

$x < -1/|\gamma|$

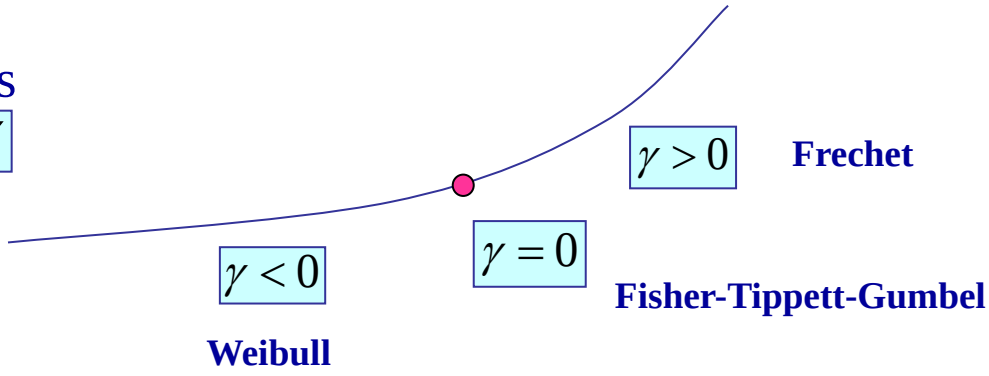
$$P(x) = (1-|\gamma|\, x)^{(1-|\gamma|)/|\gamma|} \exp[-(1-|\gamma|\, x)^{1/|\gamma|}]$$ Weibull

$\gamma = 0$

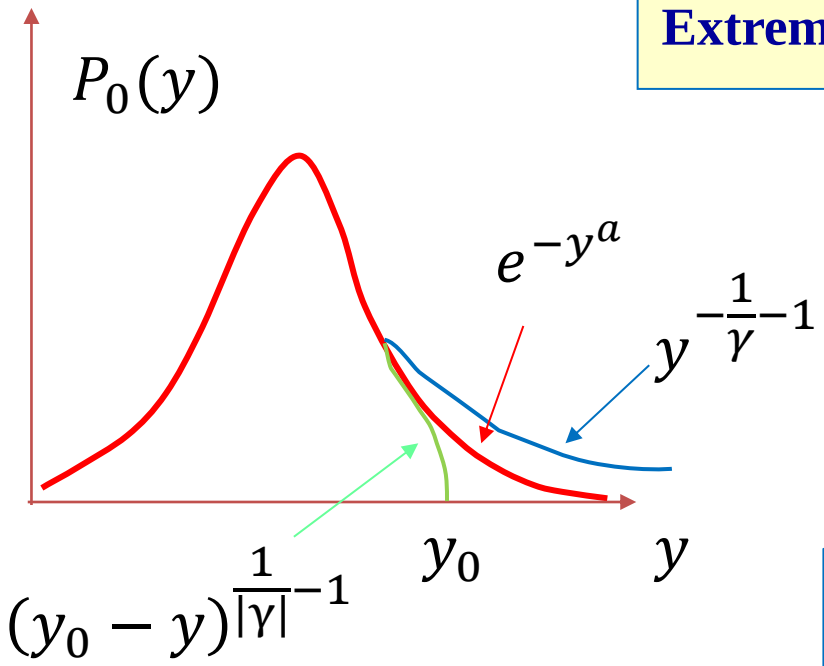
$-\infty < x < \infty$

$$P(x) = \exp[-x - \exp(-x)]$$ Fisher-Tippett-Gumbel

A line of fix points parametrized by γ



Extreme value limit distributions: i.i.d. variables



- Fisher-Tippet-Gumbel (exponential tail)

$$P_{FTG}(x) = e^{-x-e^{-x}}$$

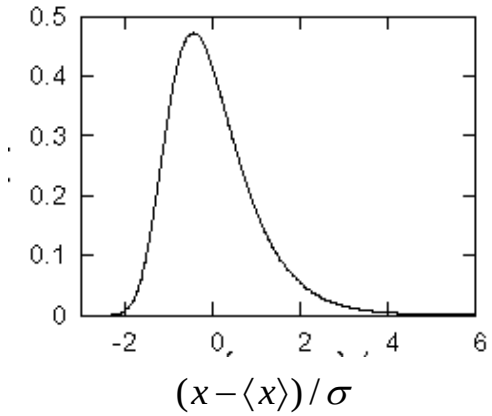
$$-\infty < x < \infty$$

- Fisher-Tippet-Frechet (power-law tail)

$$P_{FTF}(x) = (1+\gamma x)^{-\frac{1}{\gamma}-1} e^{-(1+\gamma x)^{-\frac{1}{\gamma}}}$$

$$0 < \gamma \qquad -1/\gamma < x$$

Gumbel

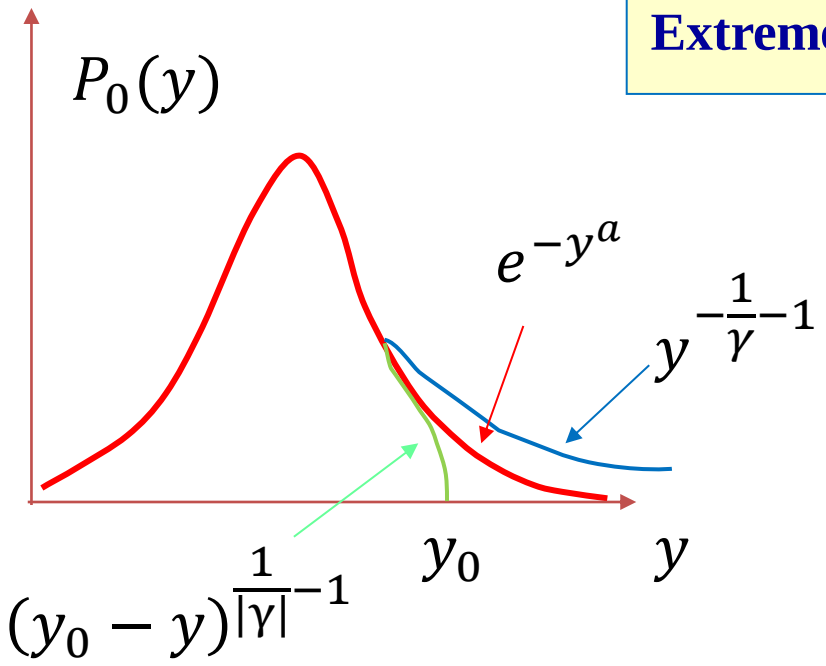


- Weibull (finite cutoff)

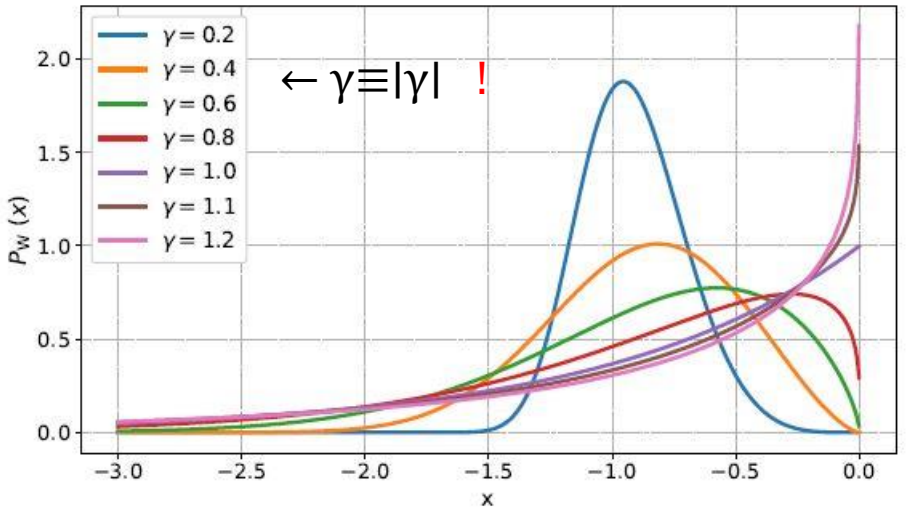
$$P_W(x) = (1-|\gamma|x)^{\frac{1}{|\gamma|}-1} e^{-(1-|\gamma|x)^{\frac{1}{|\gamma|}}}$$

$$\gamma < 0 \qquad x < 1/|\gamma|$$

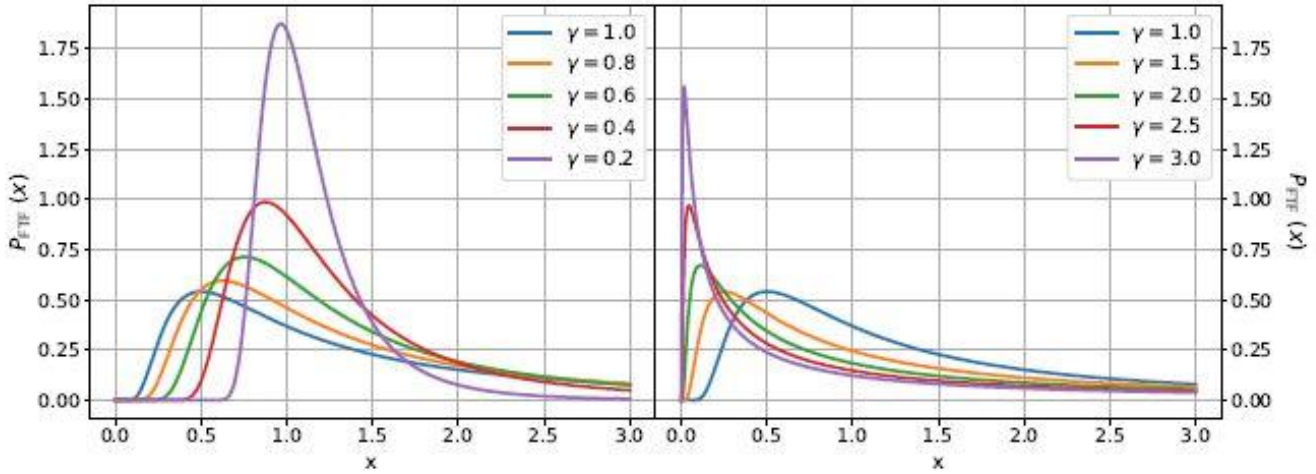
Extreme value limit distributions: i.i.d. variables



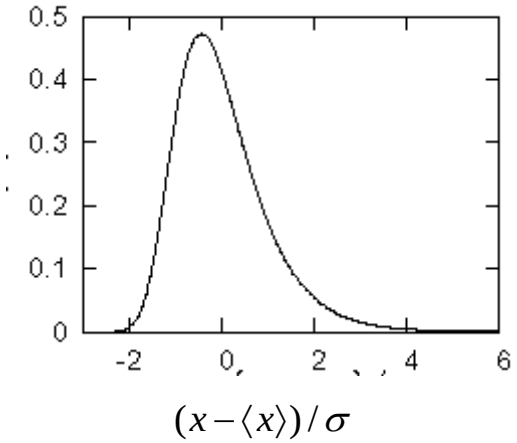
Weibull



Fisher-Tippett-Frechet



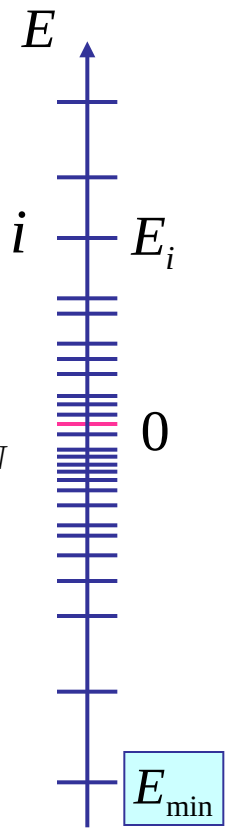
Gumbel



Return to the random energy model: What can be deduced from EVS?

B. Derrida (1980)
J.-P. Bouchaud and M. Mézard (1997)

S-K model $\mathrm{H} \sim -\sum_{ij} J_{ij}\sigma_i\sigma_j$ $P(J_{ij}) \sim \exp\left[-\frac{J_{ij}^2 N}{J^2}\right]$



Number of states: $N_{EVS} = 2^N$

replaced by $\mathrm{P}_0(E_i) \sim \exp\left[-\frac{E_i^2}{J^2 N}\right]$ $y^2 = \frac{E^2}{J^2 N}$

Number of spins in the system

Distribution of the ground state energy:

Change of variables for Gaussian: $y_{\min} = \frac{u}{\sqrt{\ln N_{EVS}}} + \sqrt{\ln N_{EVS}}$

$u = \frac{\sqrt{\ln 2}}{J} E_{\min} - N \ln 2$

$P(u) = e^{u+e^u} \sim e^u \sim e^{E_{\min}}$ exponential distribution

Low temperature behavior: Second lowest energy and crowding at E_{min} is needed.