EVA Tutorial #1

BLOCK MAXIMA APPROACH UNDER NONSTATIONARITY

Rick Katz

Institute for Mathematics Applied to Geosciences
National Center for Atmospheric Research
Boulder, CO USA

email: rwk@ucar.edu

Home page: www.isse.ucar.edu/staff/katz

Lecture: www.isse.ucar.edu/extremevalues/docs/eva1.pdf

Outline

(1) Traditional Methods/Rationale for Extreme Value Analysis

(2) Max Stability/Extremal Types Theorem

(3) Block Maxima Approach under Stationarity

(4) Return Levels

(5) Block Maxima Approach under Nonstationarity

(6) Trends in Extremes

(7) Other Forms of Covariates

(1) Traditional Methods/Rationale for Extreme Value Analysis

- **Fit models/distributions to all data**

-- Even if primary focus is on extremes

- **Statistical theory for averages**

-- Ubiquitous role of normal distribution

-- Central Limit Theorem for sums or averages

- **Central Limit Theorem**

-- Given time series $X_1, X_2, \ldots, X_n$

Assume independent and identically distributed (iid)

Assume common cumulative distribution function (cdf) F

Assume finite mean μ and variance σ^2

-- Denote sum by $S_n = X_1 + X_2 + \cdots + X_n$

-- Then, no matter what shape of cdf F,

$$\Pr\{(S_n - n\mu) / n^{1/2}\sigma \le x\} \to \Phi(x) \text{ as } n \to \infty$$

where Φ denotes standard normal $N(0, 1)$ cdf

- **Robustness**

-- Avoid sensitivity to extremes
(outliers / contamination)

- **Nonparametric Alternatives**

-- Kernel density estimation
Ok for center of distribution (but not for lower & upper tails)

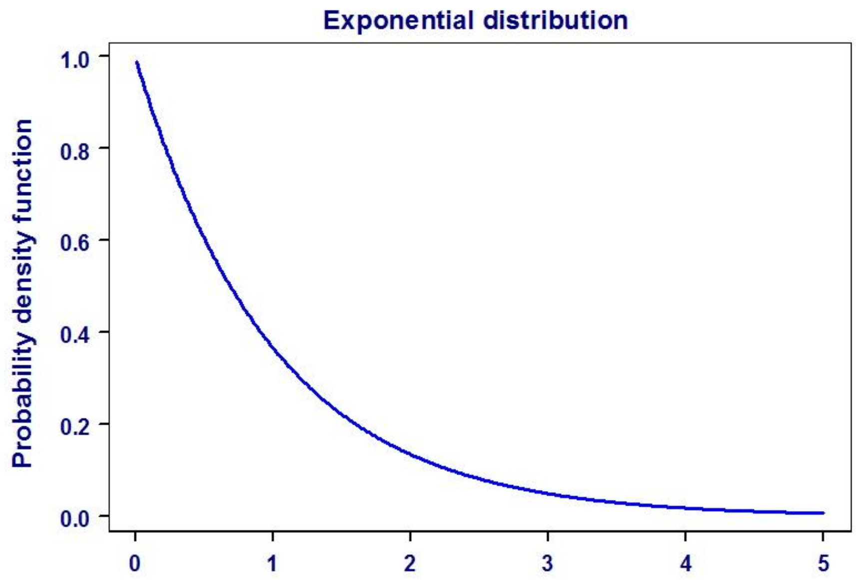
-- Resampling
Fails for maxima
Cannot extrapolate

- **Conduct sampling experiment**

-- Exponential distribution with cdf

$$F(x) = 1 - \exp[-(x/\sigma)], \quad x > 0, \sigma > 0$$

Here σ is scale parameter (also mean)



-- Draw random samples of size n = 10 from exponential distribution (with σ = 1) and calculate mean for each sample

(i) First pseudo random sample

1.678, 0.607, 0.732, 1.806, 1.388, 0.630, 0.382, 0.396, 1.324, 1.148

(Sample mean ≈ 1.009)

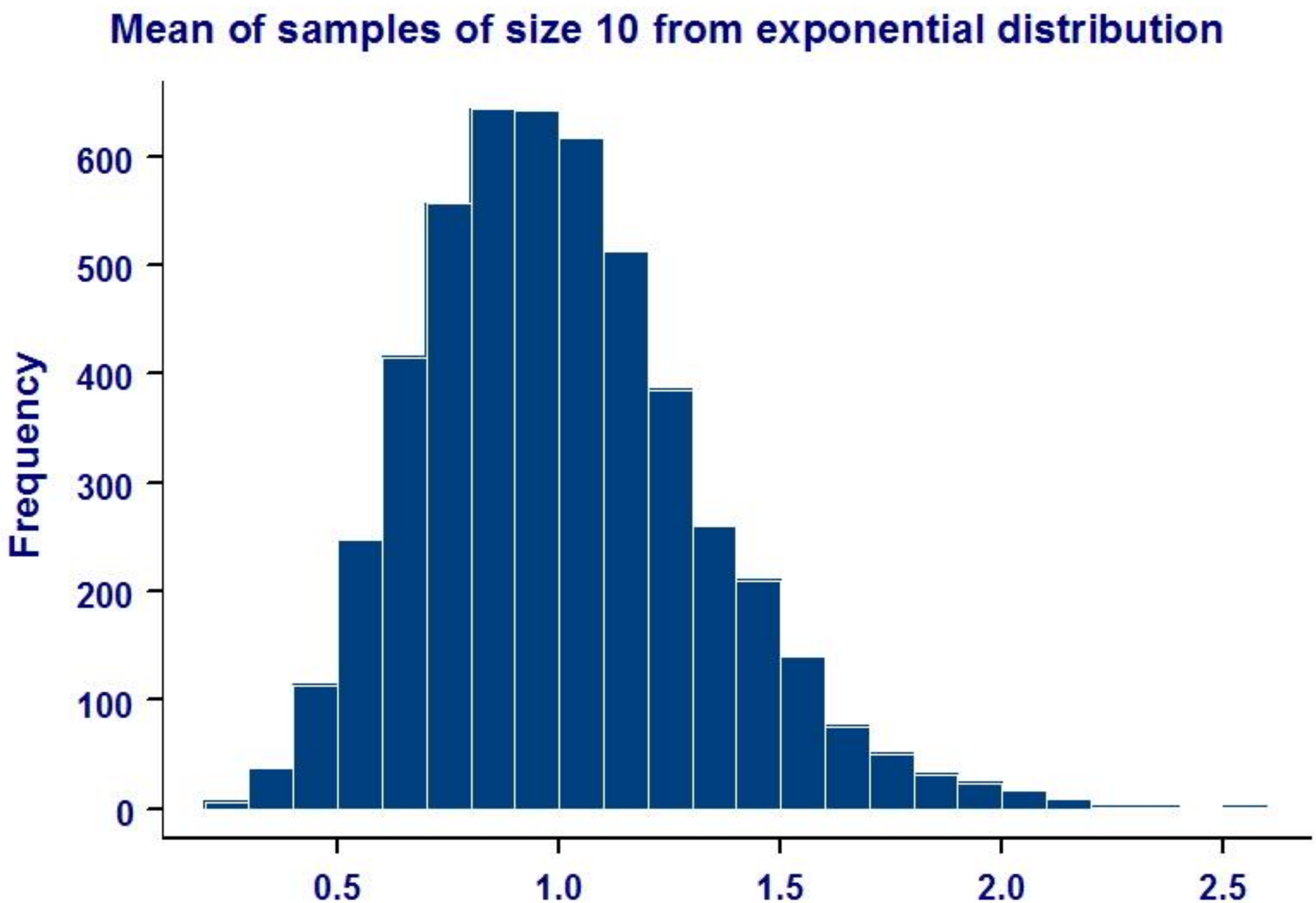
(ii) Second pseudo random sample

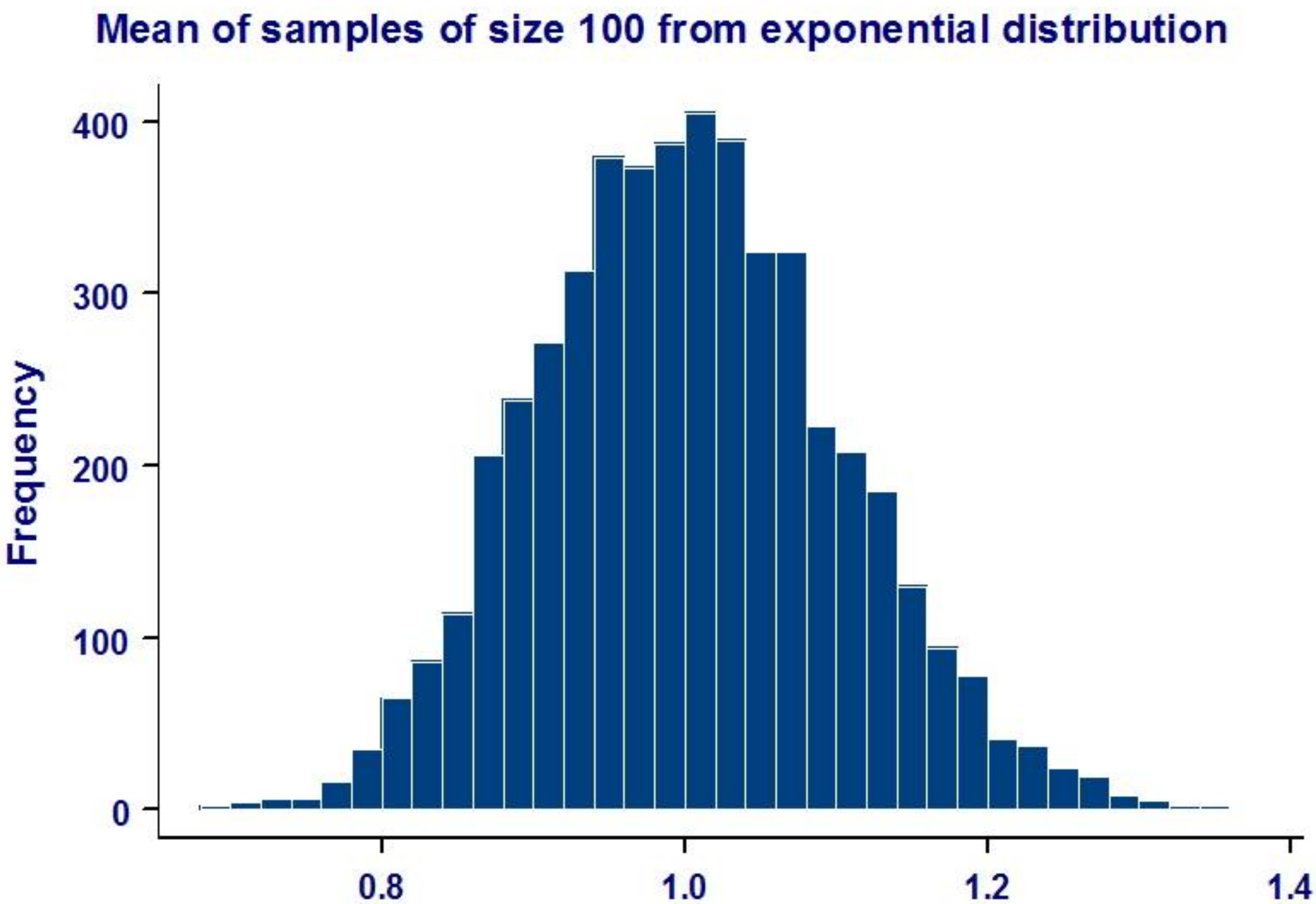
Sample mean ≈ 0.571

(iii) Third pseudo random sample

Sample mean ≈ 0.859

Repeat many more times


Mean of samples of size 10 from exponential distribution
Frequency
600
500
400
300
200
100
0
0.5
1.0
1.5
2.0
2.5


Mean of samples of size 100 from exponential distribution
Frequency
400
300
200
100
0
0.8
1.0
1.2
1.4

- **Limited information about extremes**

-- Exploit what theory is available

- **More robust/flexible approach**

-- Tail behavior of standard distributions is too restrictive

Statistical theory indicates possibility of “heavy” tails

Data suggest evidence of “heavy” tails

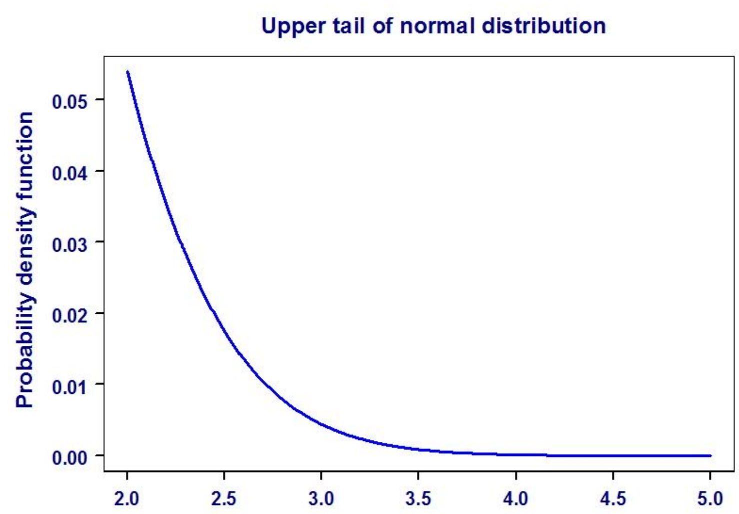
Conventional distributions have “light” tails

-- Example

Let X have standard normal distribution [i. e., $N(0, 1)$] with probability density function (pdf)

$$\varphi(x) = (2\pi)^{-1/2} \exp(-x^2 / 2)$$

Then $\Pr\{X > x\} \equiv 1 - \Phi(x) \approx \varphi(x) / x$, **for large** x



- **Statistical behavior of extremes**

-- Effectively no role for normal distribution

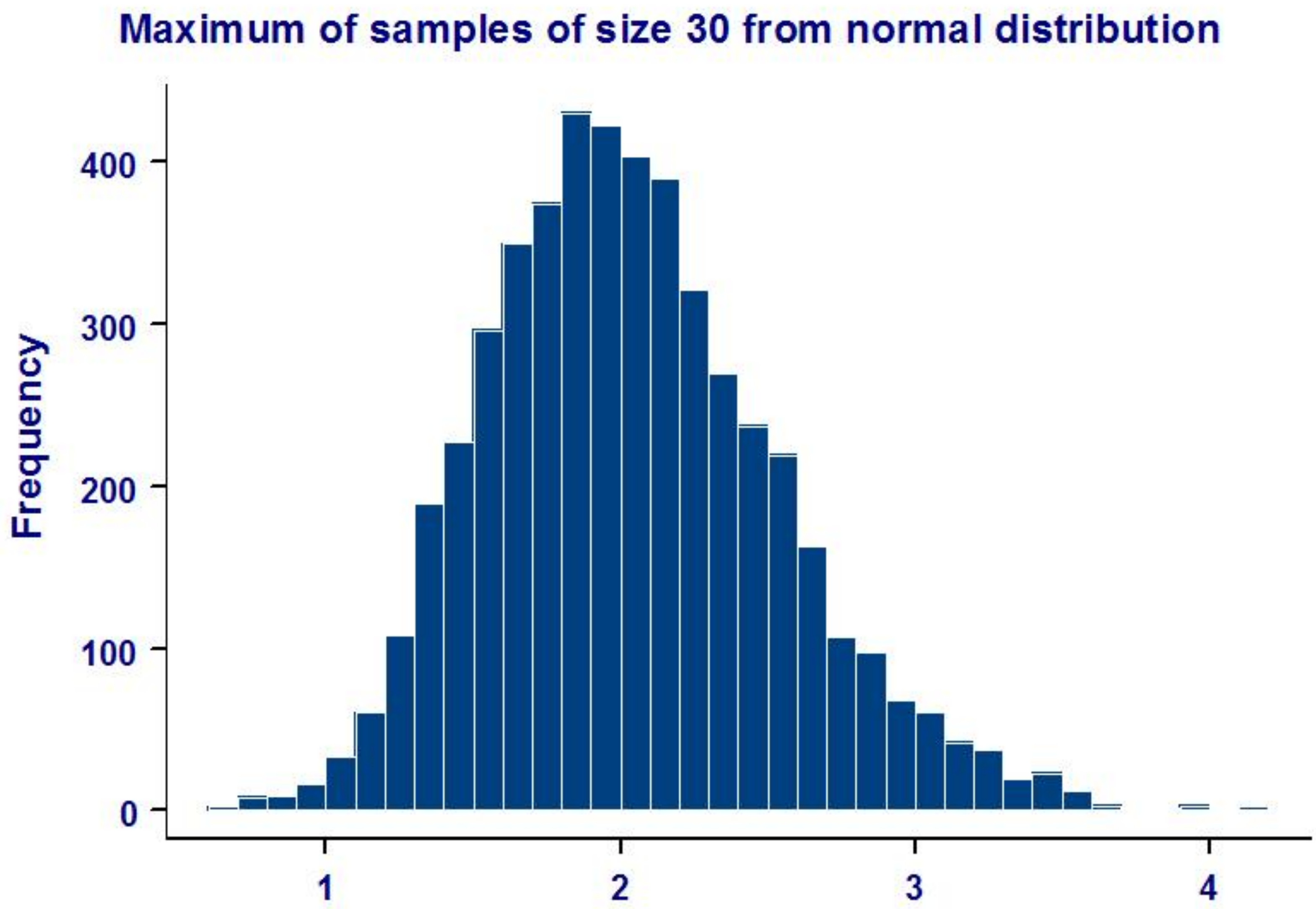
-- What form of distribution(s) instead?

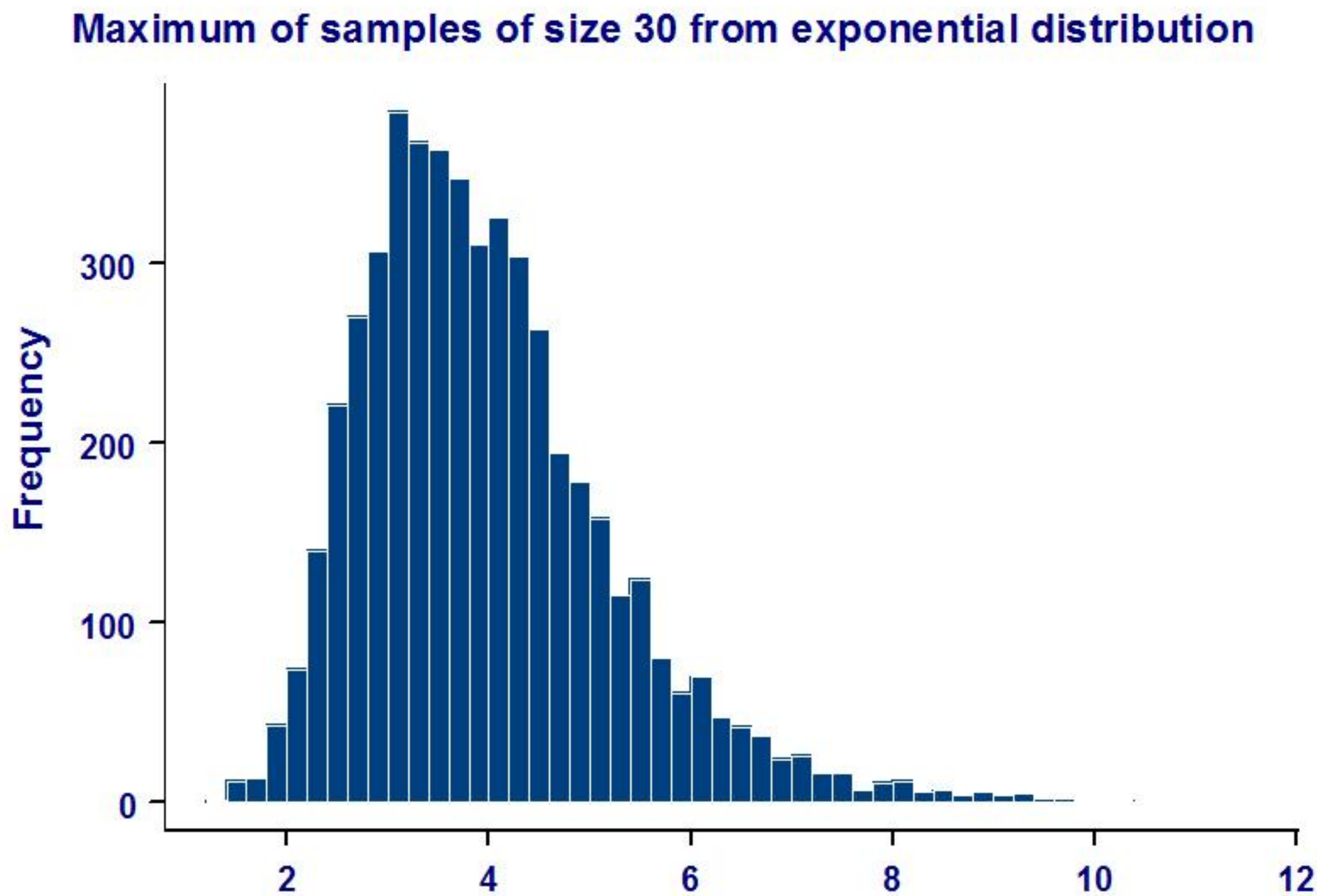
- **Conduct another sampling experiment**

-- Calculate largest value of random sample (instead of mean)

(i) Standard normal distribution $N(0, 1)$

(ii) Exponential distribution ($\sigma = 1$)


Maximum of samples of size 30 from normal distribution
Frequency
0
100
200
300
400
1
2
3
4


Maximum of samples of size 30 from exponential distribution
Frequency
0
100
200
300
2
4
6
8
10
12

(2) Max Stability/Extremal Types Theorem

- **"Sum stability"**

-- Property of normal distribution

$X_1, X_2, \ldots, X_n$ iid with common cdf $N(\mu, \sigma^2)$

Then sum $S_n = X_1 + X_2 + \cdots + X_n$

is *exactly* normally distributed

In particular, $(S_n - n\mu) / n^{1/2}\sigma$

has an exact $N(0, 1)$ distribution

- **"Max stability"**

-- Want to find distribution(s) for which maximum has same form as original sample

Note that

$$\max\{X_1, X_2, \ldots, X_{2n}\} =$$

$$\max\{\max\{X_1, X_2, \ldots, X_n\}, \max\{X_{n+1}, X_{n+2}, \ldots, X_{2n}\}\}$$

-- So cdf *G*, say, must satisfy

$$G^2(x) = G(ax + b)$$

Here $a > 0$ and b are constants

- **Extremal Types Theorem**

Time series $X_1, X_2, \ldots, X_n$ assumed iid (*for now*)

Set $M_n = \max\{X_1, X_2, \ldots, X_n\}$

Suppose that there exist constants $a_n > 0$ and b_n such that

$$\Pr\{(M_n - b_n) / a_n \le x\} \to G(x) \quad \text{as } n \to \infty$$

where G is a non-degenerate cdf

Then G must a generalized extreme value (GEV) cdf; that is,

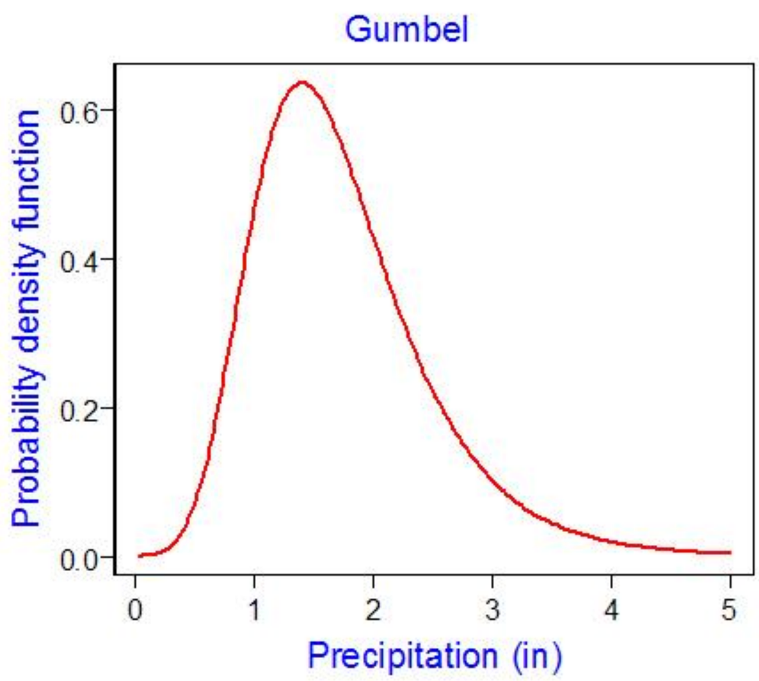
$$G(x; \mu, \sigma, \xi) = \exp\left\{-[1 + \xi\,(x - \mu)/\sigma]^{-1/\xi}\right\},\ 1 + \xi\,(x - \mu)/\sigma > 0$$

μ location parameter, $\sigma > 0$ scale parameter, ξ shape parameter

(i) $\xi = 0$ (*Gumbel type*, limit as $\xi \to 0$)

"Light" upper tail

"Domain of attraction" for many common distributions (e. g., normal, exponential, gamma)

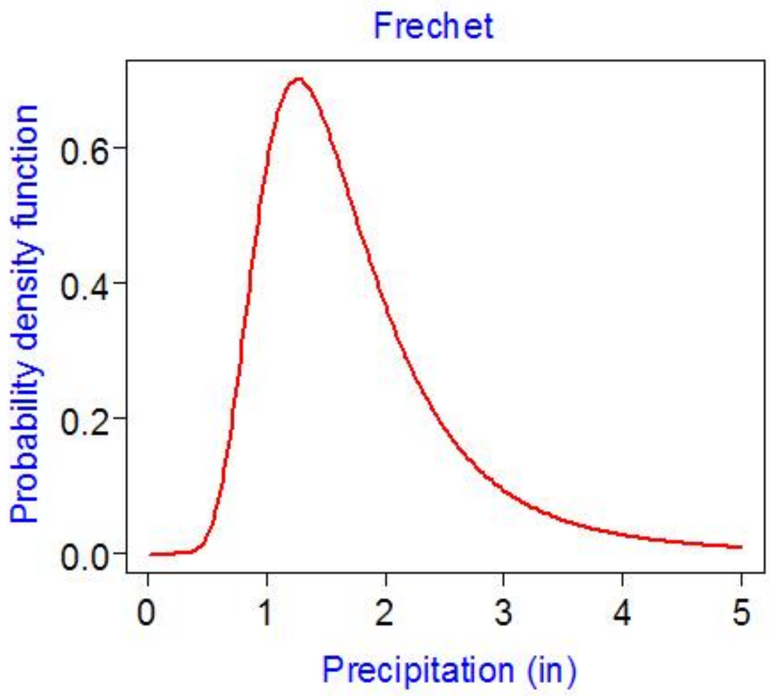


(ii) ξ > 0 (*Fréchet type*)

"Heavy" upper tail with infinite *r*th-order moment if *r* ≥ 1/ξ

(e. g., infinite variance if ξ ≥ 1/2)

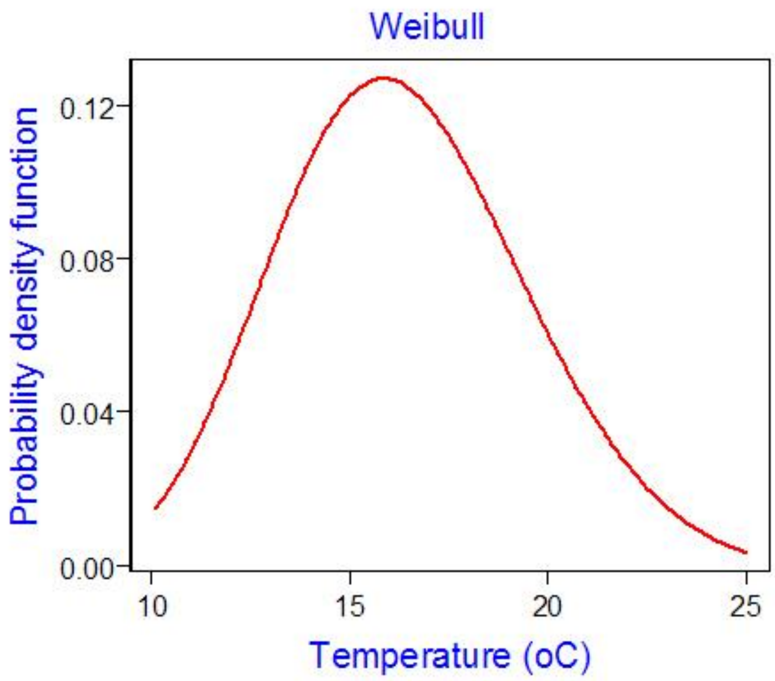
Fits precipitation, streamflow, economic damage

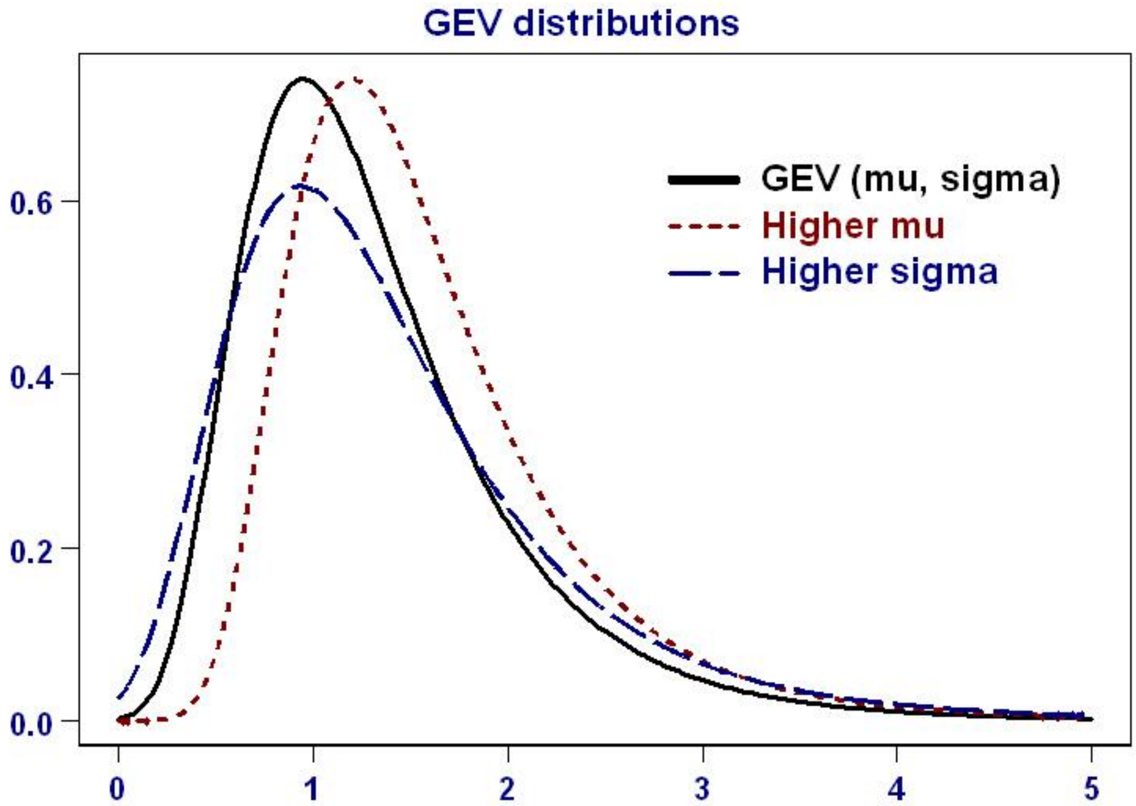


(iii) $\xi < 0$ (*Weibull type*)

Bounded upper tail [$x < \mu + \sigma / (-\xi)$]

Fits temperature, wind speed, sea level





Location parameter of GEV is *not* equivalent to mean

Scale parameter of GEV is *not* equivalent to standard deviation

(3) Block Maxima Approach under Stationarity

- **GEV distribution**

-- Fit directly to maxima (say with block size *n*)

e. g., annual maximum of daily precipitation amount or highest temperature over given year or annual peak stream flow

-- Advantages

Do not necessarily need to explicitly model annual and diurnal cycles

Do not necessarily need to explicitly model temporal dependence

- **Maximum likelihood estimation (mle)**

-- Given observed block maxima $X_1 = x_1, X_2 = x_2, \ldots, X_T = x_T$

-- Assume exact GEV dist. with pdf

$$g(x; \mu, \sigma, \xi) = G'(x; \mu, \sigma, \xi)$$

-- Likelihood function

$$L(x_1, x_2, \ldots, x_T; \mu, \sigma, \xi) = g(x_1; \mu, \sigma, \xi)\, g(x_2; \mu, \sigma, \xi) \cdots g(x_T; \mu, \sigma, \xi)$$

Minimize

$$-\ln L(x_1, x_2, \ldots, x_T; \mu, \sigma, \xi)$$

with respect to μ, σ, ξ

-- Likelihood ratio test (LRT)

For example, to test whether ξ = 0 fit two models:

(i) −ln $L(x_1, x_2, \ldots, x_T; \mu, \sigma, \xi)$ minimized with respect to μ, σ, ξ

(ii) −ln $L(x_1, x_2, \ldots, x_T; \mu, \sigma, \xi = 0)$ minimized with respect to μ, σ

If ξ = 0, then 2 [(ii) – (i)] has approximate chi square distribution with 1 degree of freedom (df) for large *T*

-- Confidence interval (e. g., for ξ) based on "profile likelihood"

Minimize −ln $L(x_1, x_2, \ldots, x_T; \mu, \sigma, \xi)$ with respect to μ, σ as function of ξ

Use chi square dist. with 1 df

- **Fort Collins daily precipitation amount**

-- Fort Collins, CO, USA

Time series of daily precipitation amount (in), 1900-1999

Semi-arid region

Marked annual cycle in precipitation

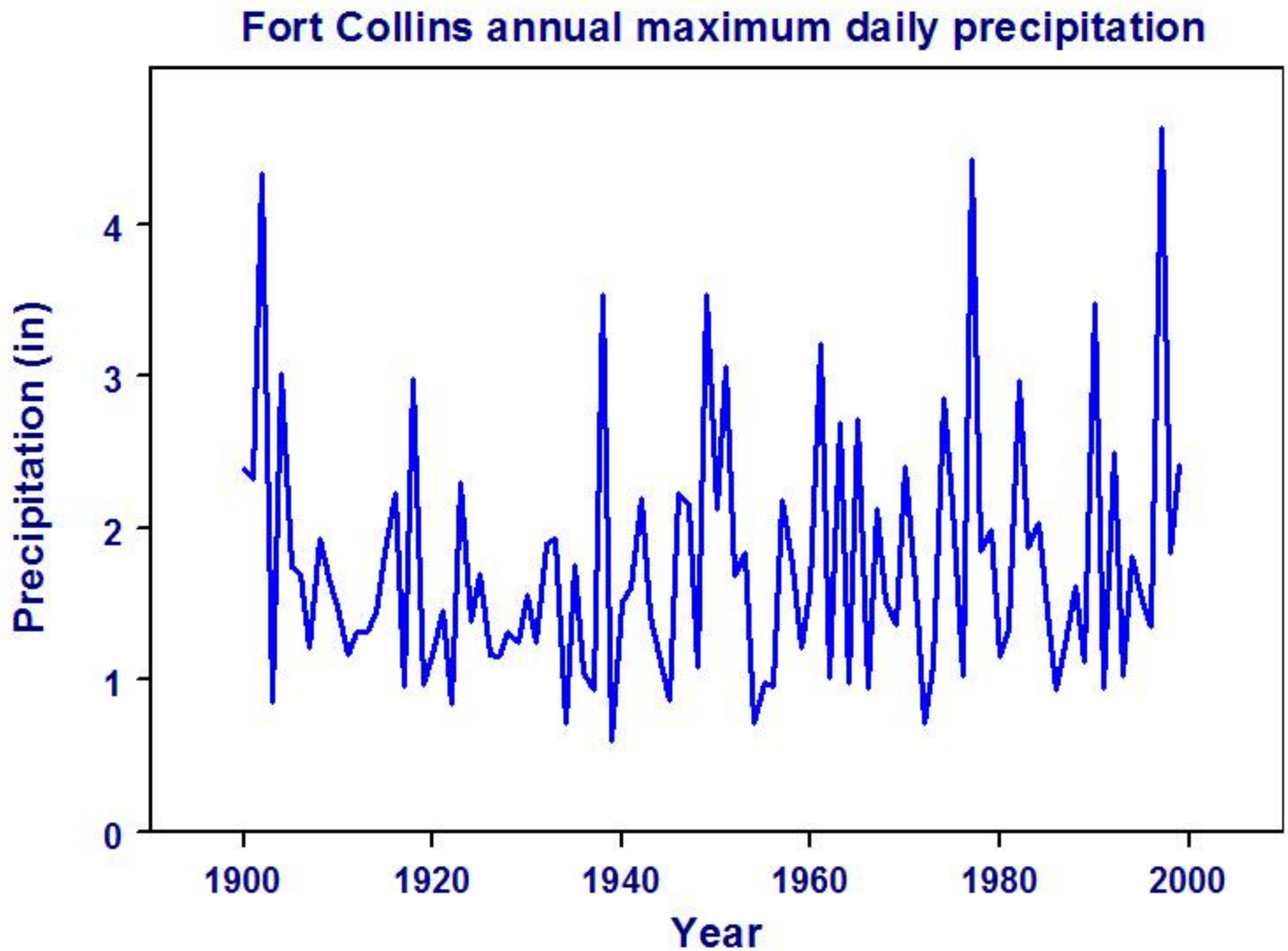
(peak in late spring/early summer, driest in winter)

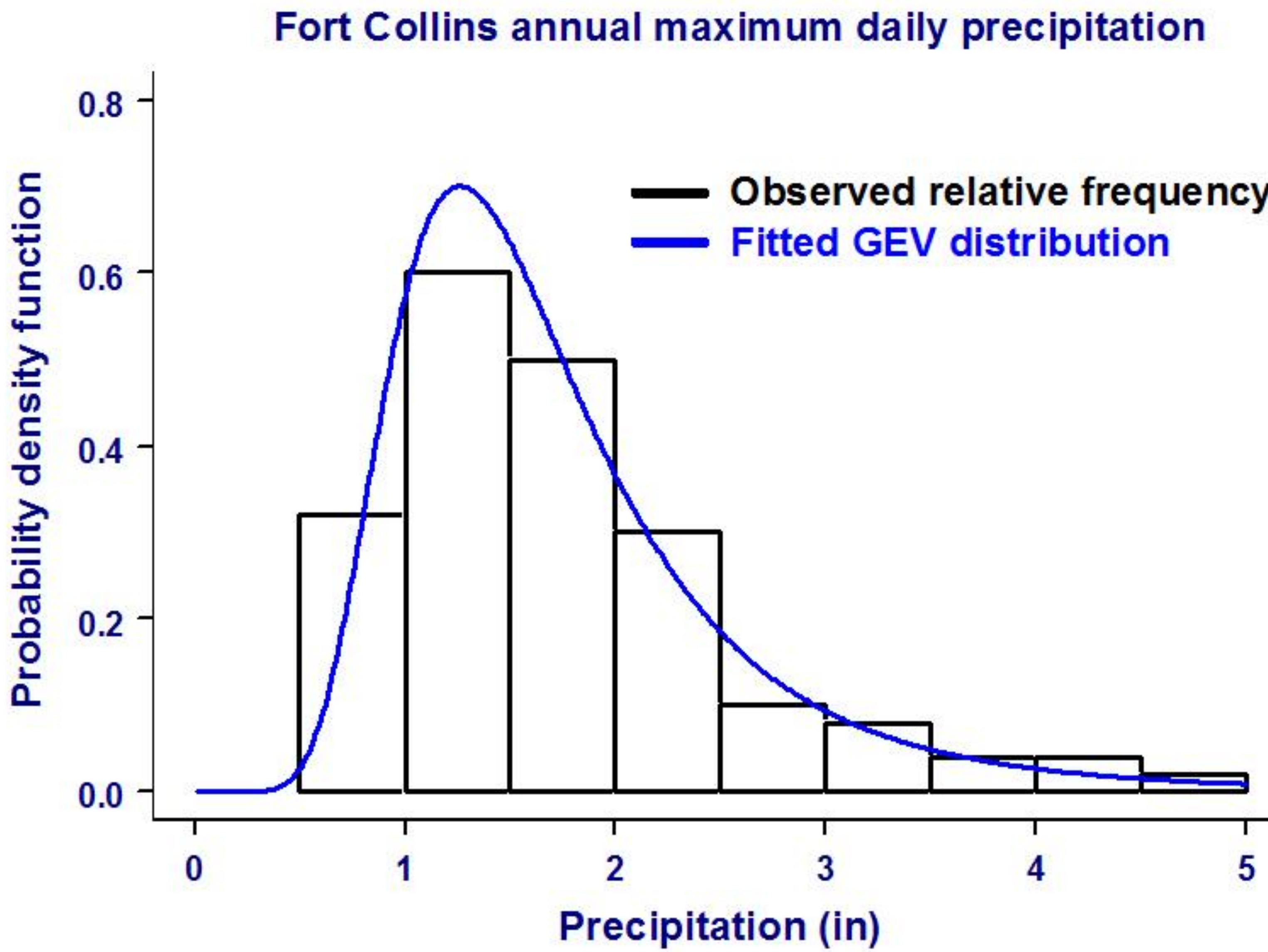
Consider annual maxima (block size $n \approx 365$)

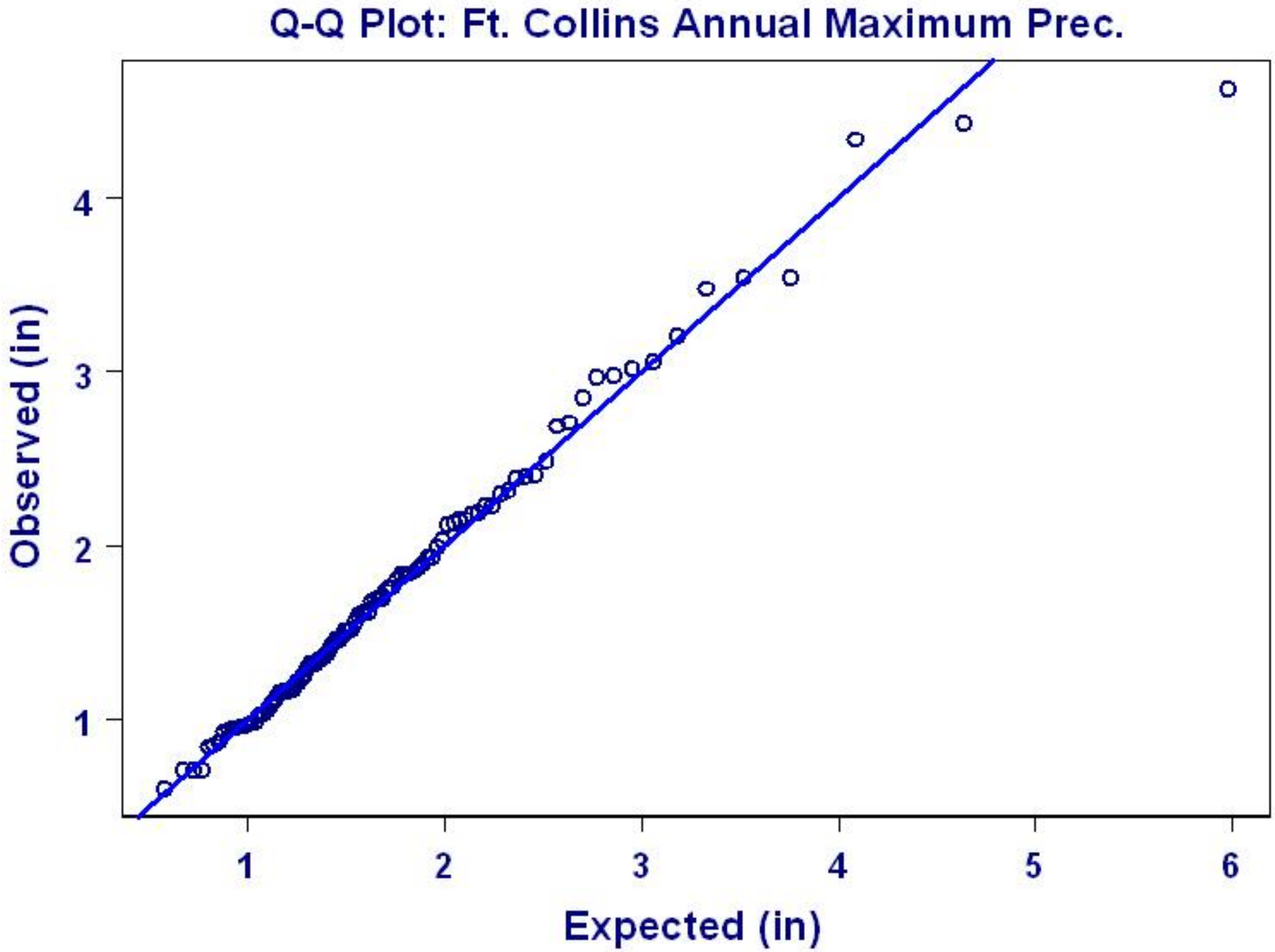
No obvious long-term trend in annual maxima ($T = 100$)

Flood on 28 July 1997

(Damaged campus of Colorado State Univ.)


Fort Collins annual maximum daily precipitation
Precipitation (in)
0
1
2
3
4
1900
1920
1940
1960
1980
2000
Year


Fort Collins annual maximum daily precipitation
Observed relative frequency
Fitted GEV distribution
Probability density function
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
Precipitation (in)


Q-Q Plot: Ft. Collins Annual Maximum Prec.
Observed (in)
Expected (in)
1
2
3
4
5
6

- **Parameter estimates and standard errors**

Parameter	Estimate	(Std. Error)
Location μ	**1.347**	**(0.062)**
Scale σ	**0.533**	**(0.049)**
Shape ξ	**0.174**	**(0.092)**

-- LRT for ξ = 0 (*P*-value ≈ 0.038)

-- 95% confidence interval for shape parameter ξ (based on profile likelihood)

$$0.009 < \xi < 0.369$$

(4) Return Levels

- **Assume stationarity**

-- i. e., unchanging climate

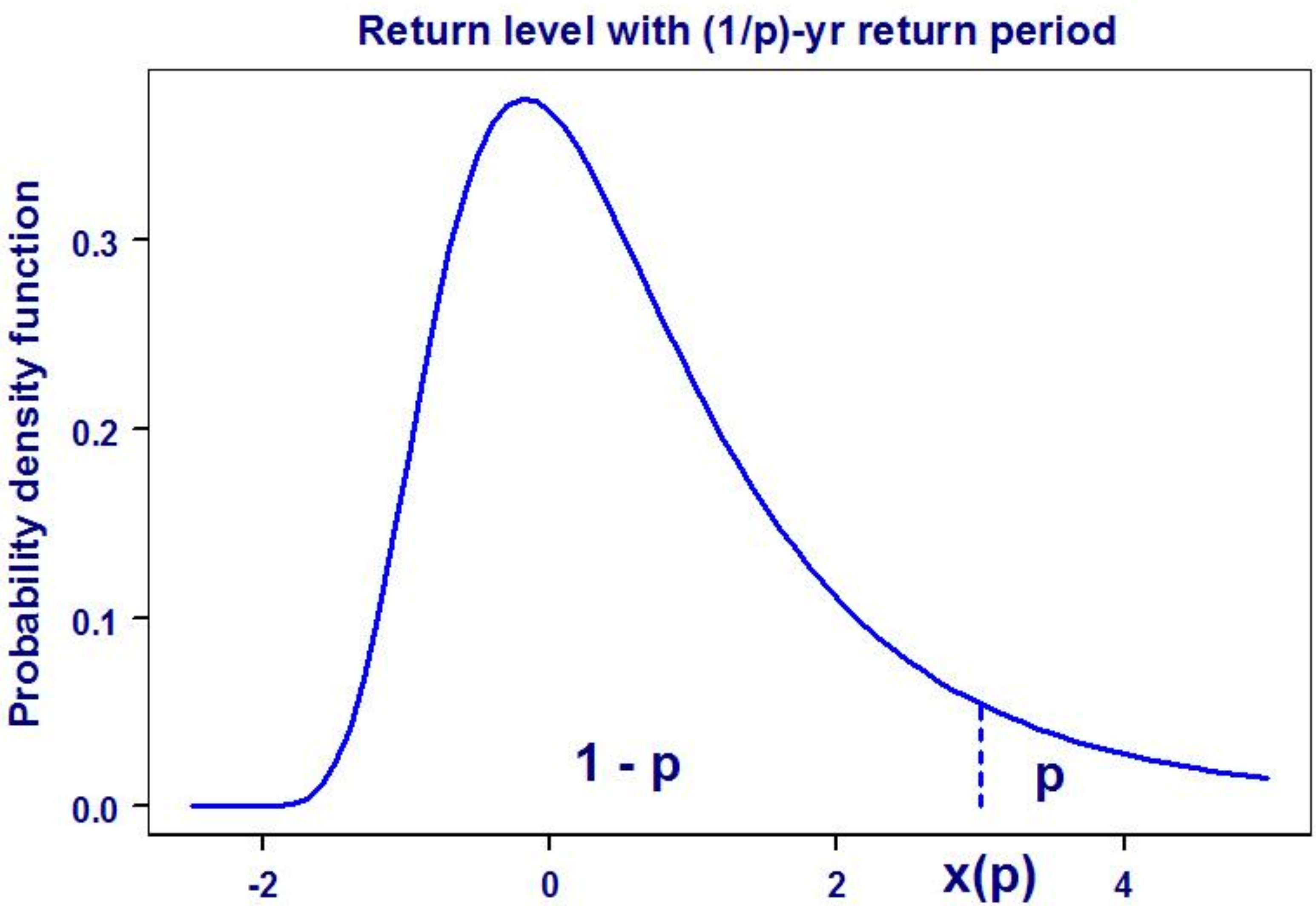
- **Return period / Return level**

-- "Return level" with (1/*p*)-yr "return period"

$$x(p) = G^{-1}(1 - p;\ \mu, \sigma, \xi), \quad 0 < p < 1$$

Quantile of GEV cdf *G*

(e. g., *p* = 0.01 corresponds to 100-yr return period)


Return level with (1/p)-yr return period
Probability density function
0.3
0.2
0.1
0.0
1 - p
p
-2
0
2
x(p)
4

- **GEV distribution**

$$x(p) = \mu - (\sigma/\xi)\,\{1 - [-\ln(1 - p)]\}^{-\xi}$$

Confidence interval: Re-parameterize replacing location parameter μ with *x*(*p*) & use profile likelihood method

-- Fort Collins precipitation example (annual maxima)

Estimated 100-yr return level: 5.10 in

95% confidence interval (based on profile likelihood):

$$3.93 \text{ in} < x(0.01) < 8.00 \text{ in}$$

(5) Block Maxima Approach under Nonstationarity

- **Sources**

-- Trends

Associated with global climate change (e. g.)

-- Cycles

Annual & diurnal cycles (e. g.)

-- Physically-based

Use in statistical downscaling (e. g.)

- **Theory**

-- No general extreme value theory under nonstationarity

Only limited results under restrictive conditions

- **Methods**

-- Introduction of covariates resembles "generalized linear models"

-- Straightforward to extend maximum likelihood estimation

- **Issues**

-- Nature of relationship between extremes & covariates

Resembles that for overall / center of data?

(6) Trends in Extremes

- **Trends**

-- Example (Urban heat island)

Trend in summer minimum temperature at Phoenix, AZ (i. e., block minima)

$$\min\{X_1, X_2, \ldots, X_n\} = -\max\{-X_1, -X_2, \ldots, -X_n\}$$

Assume negated summer minimum temperature in year *t* has GEV distribution with location and scale parameters:

$$\mu(t) = \mu_0 + \mu_1 t, \quad \ln \sigma(t) = \sigma_0 + \sigma_1 t, \quad \xi(t) = \xi, \quad t = 1, 2, \ldots$$

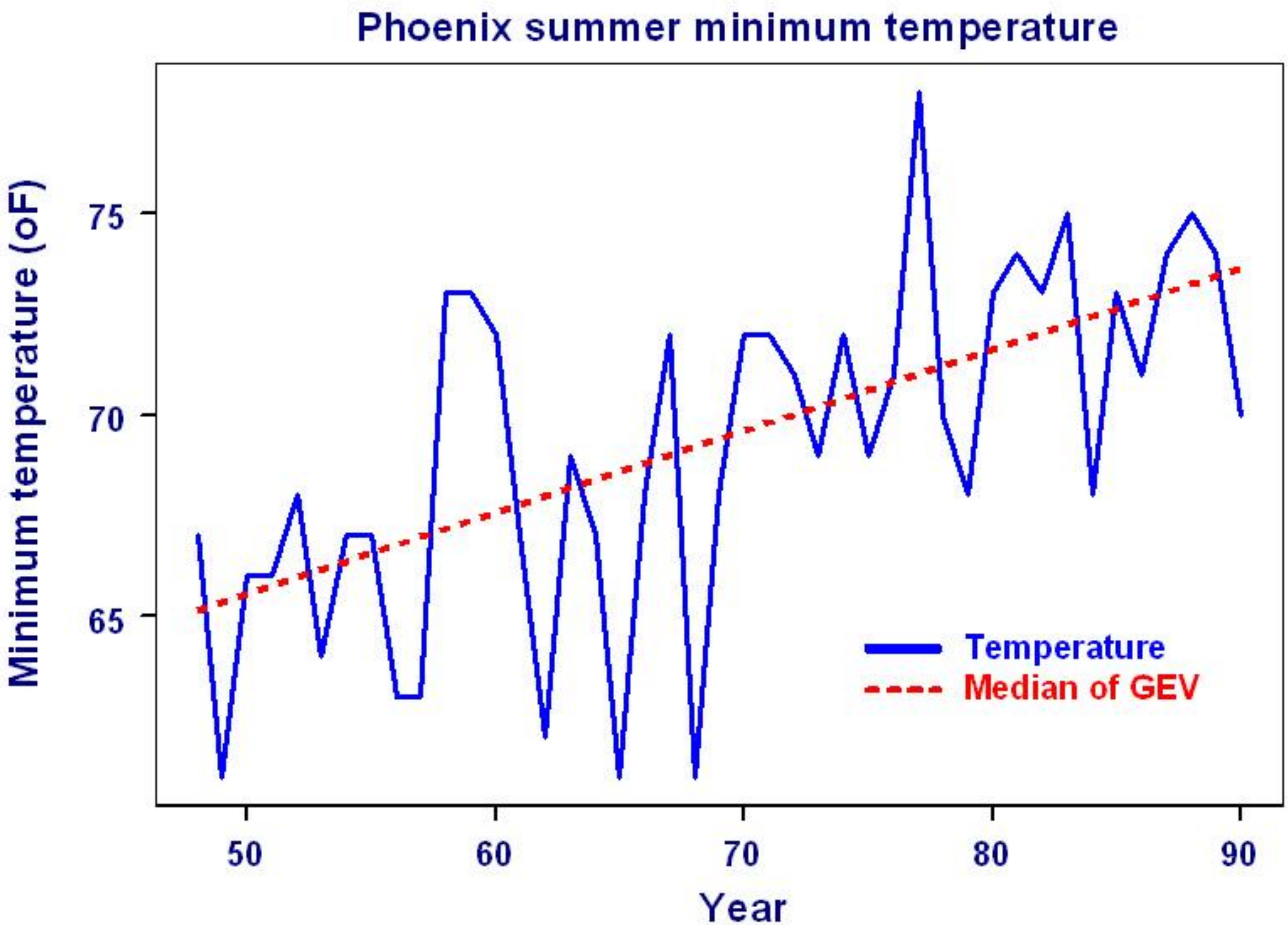
- **Parameter estimates and standard errors**

Parameter		**Estimate**	**(Std. Error)**
Location:	μ_0	**66.17***	
	μ_1	**0.196***	**(0.041)**
Scale:	σ_0	**1.338**	
	σ_1	**−0.009**	**(0.010)**
Shape:	ξ	**−0.211**	

***Sign of location parameters reversed to convert back to minima**

-- LRT for $\mu_1 = 0$ (*P*-value $< 10^{-5}$)

-- LRT for $\sigma_1 = 0$ (*P*-value ≈ 0.366)


Phoenix summer minimum temperature
Minimum temperature (oF)
75
70
65
50
60
70
80
90
Year
Temperature
Median of GEV

- **Q-Q plots under non-stationarity**

-- Transform to common distribution

Non-stationary GEV [$\mu(t)$, $\sigma(t)$, $\xi(t)$]

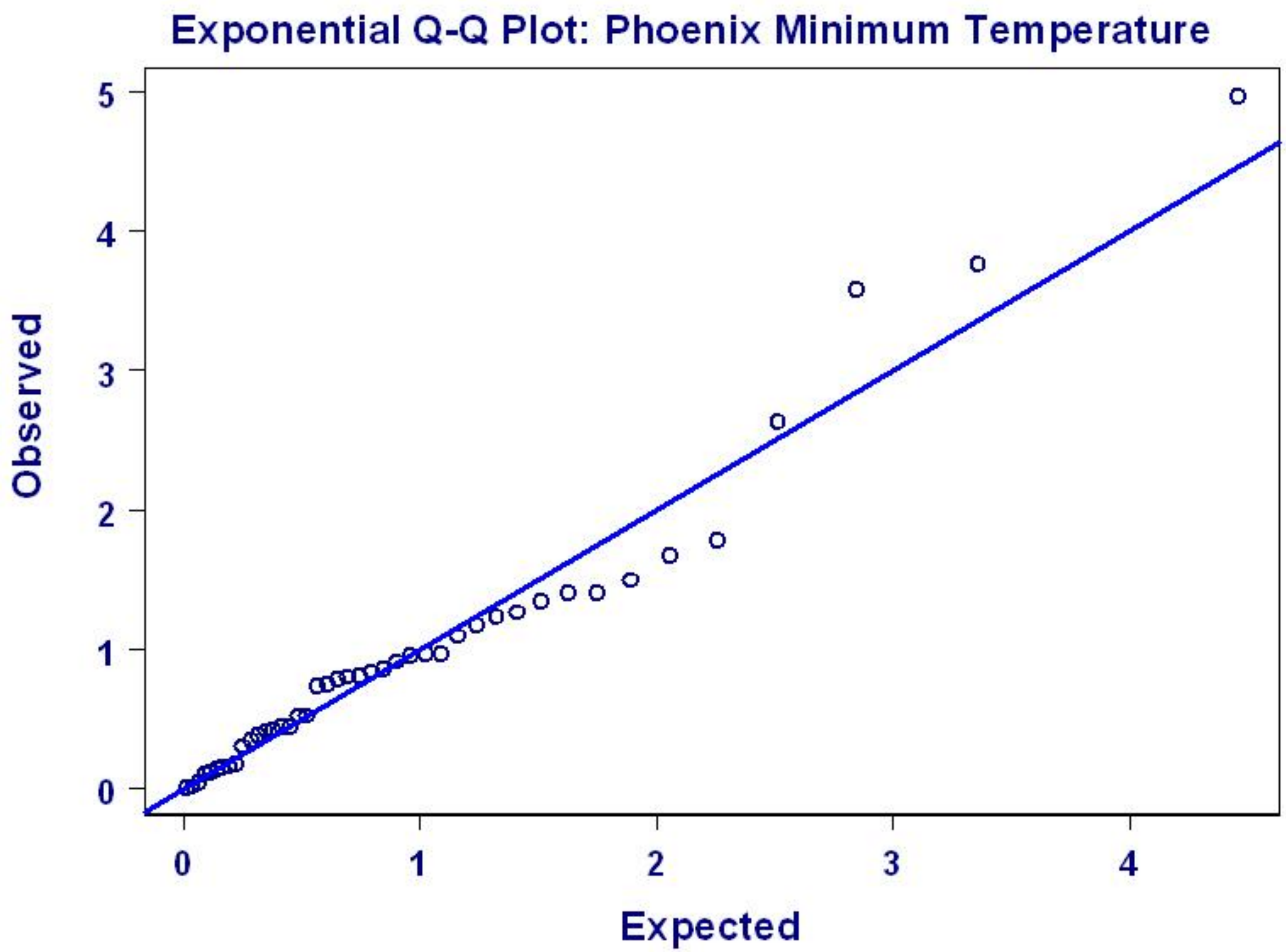
***Not* invariant to choice of transformation**

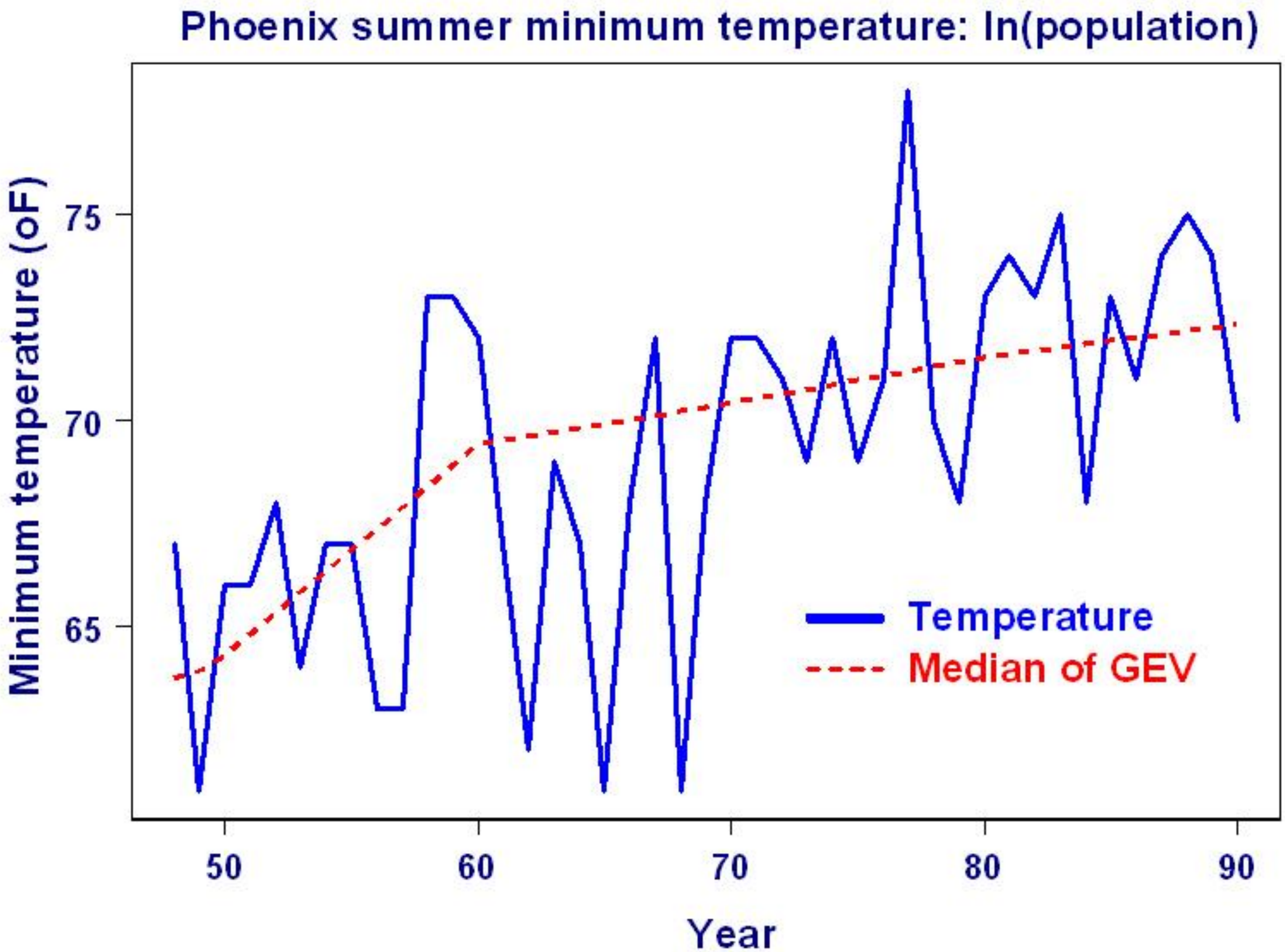
(i) Non-stationary GEV to standard exponential

$$\varepsilon_t = \{1 + \xi(t)\,[X_t - \mu(t)] \,/\, \sigma(t)\}^{-1/\xi(t)}$$

(ii) Non-stationary GEV to standard Gumbel (used by extRemes)

$$\varepsilon_t = [1/\xi(t)] \log \{1 + \xi(t)\,[X_t - \mu(t)] \,/\, \sigma(t)\}$$


Exponential Q-Q Plot: Phoenix Minimum Temperature
Observed
Expected
0
1
2
3
4
5


Phoenix summer minimum temperature: ln(population)
Minimum temperature (oF)
75
70
65
50
60
70
80
90
Year
Temperature
Median of GEV

(7) Other Forms of Covariates

- **Physically-based covariates**

-- Example [Arctic Oscillation (AO)]

Winter maximum temperature at Port Jervis, NY, USA
(i. e., block maxima)

***Z* denotes winter index of AO**

Given *Z* = *z*, assume conditional distribution of winter maximum temperature is GEV distribution with parameters:

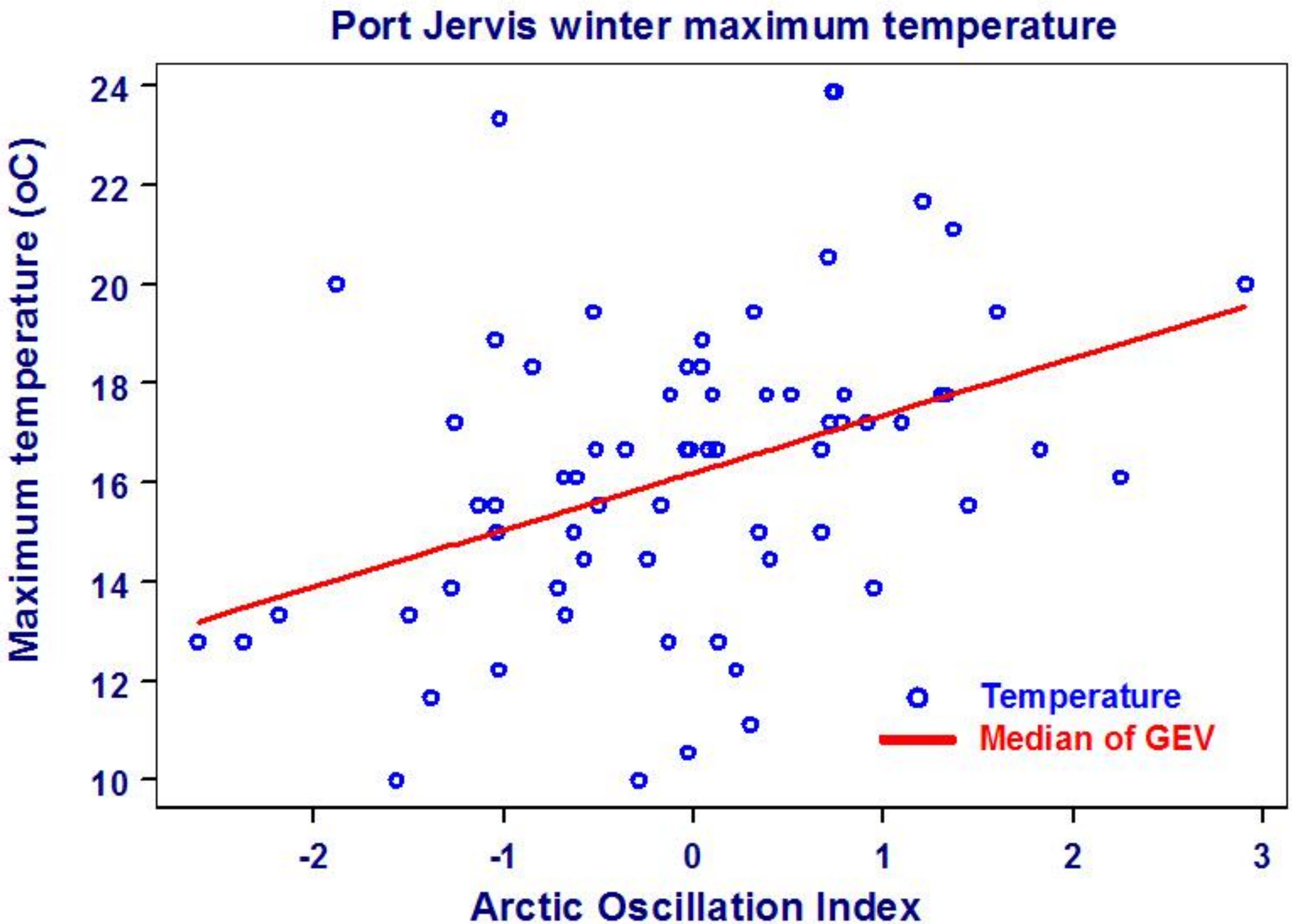
$$\mu(z) = \mu_0 + \mu_1 z, \quad \ln \sigma(z) = \sigma_0 + \sigma_1 z, \qquad \xi(z) = \xi$$

- **Parameter estimates and standard errors**

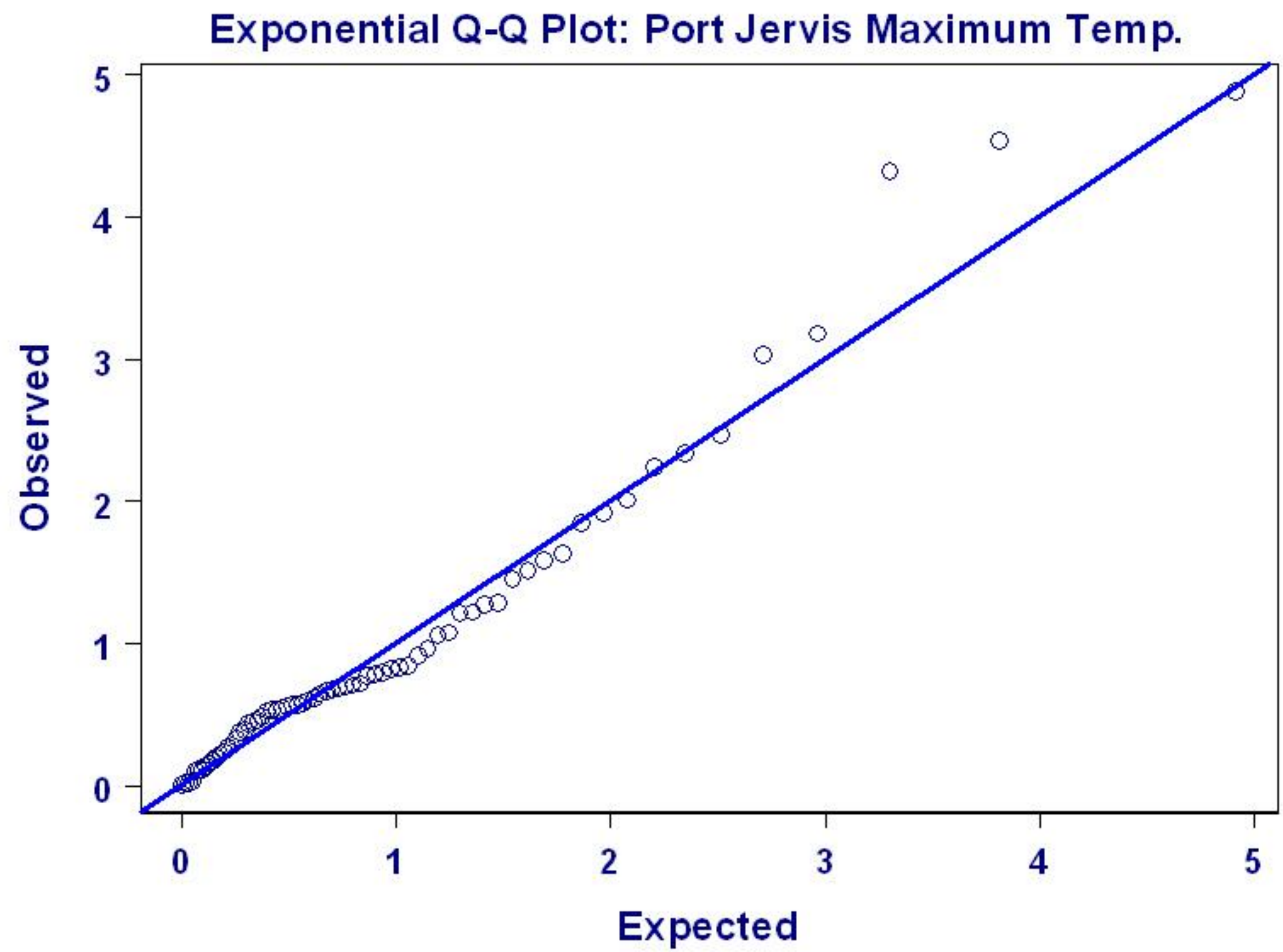
Parameter		Estimate	(Std. Error)
Location:	μ_0	15.26	
	μ_1	1.175	(0.319)
Scale:	σ_0	0.984	
	σ_1	−0.044	(0.092)
Shape:	ξ	−0.186	

-- LRT for $\mu_1 = 0$ (*P*-value < 0.001)

-- LRT for $\sigma_1 = 0$ (*P*-value ≈ 0.635)



Port Jervis winter maximum temperature
Maximum temperature (oC)
24
22
20
18
16
14
12
10
-2
-1
0
1
2
3
Arctic Oscillation Index
Temperature
Median of GEV


Exponential Q-Q Plot: Port Jervis Maximum Temp.
Observed
Expected
0
1
2
3
4
5

Homework

A random variable *X* has a *lognormal distribution* if the log-transformed variable

$$Y = \ln X$$

has a normal distribution. Then *Y* is in the domain of attraction of the Gumbel type.

What is the domain of attraction of *X*?
(i. e., Gumbel, Fréchet, or Weibull type?)

***Answer*: *X* is in the domain of attraction of the Gumbel type.**