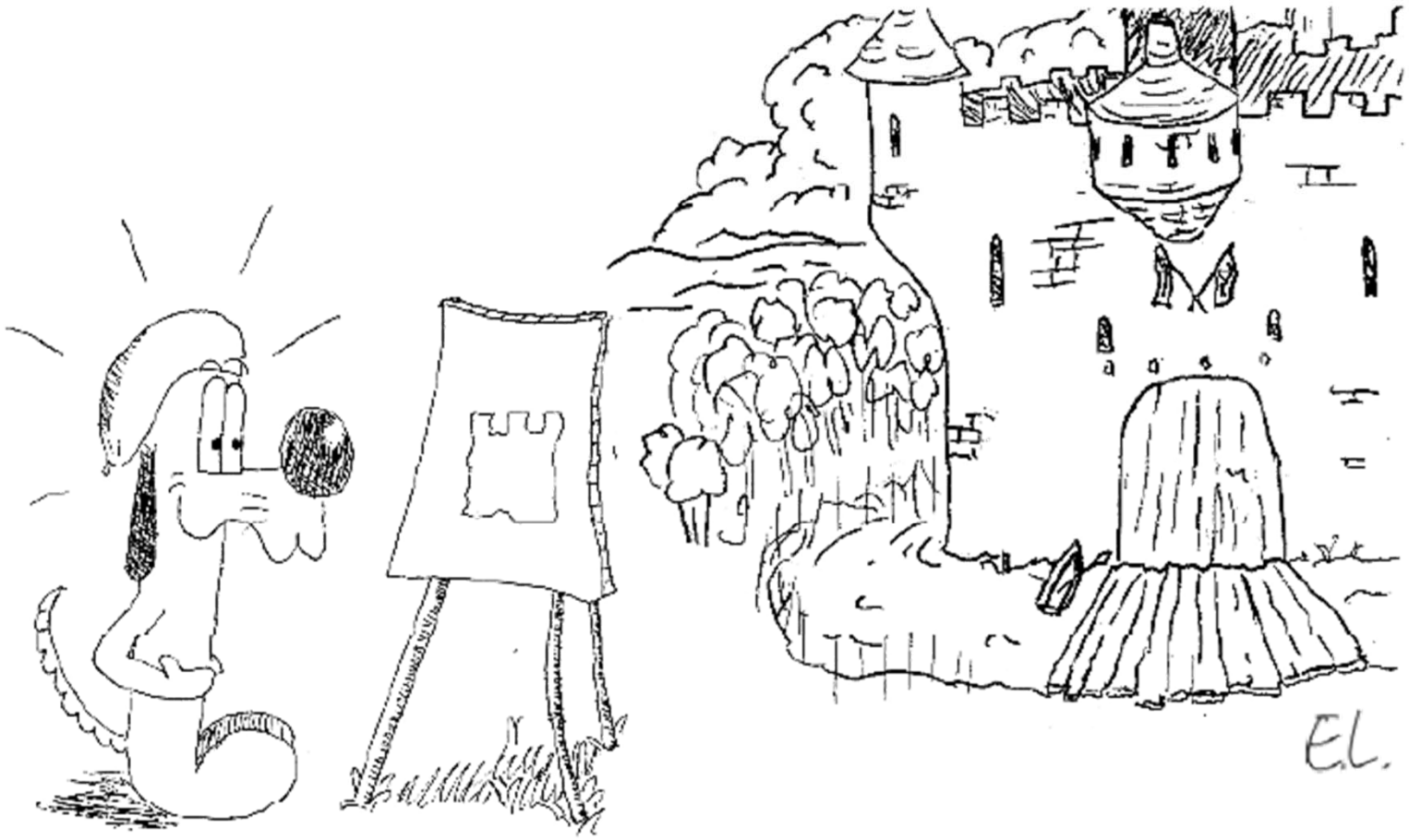
Model Fitting
Part 2



Jean-Yves Le Boudec

Contents

1. Choosing a Distribution

Know a catalog of distributions, use the following *features* to do a pre-selection

- Shape
- Kurtosis, Skewness
- Power laws
- Hazard Rate

Fit

Verify the fit visually or with a test (see later)

Feature 1: Distribution Shape

The distributions with PDFs $F(), G()$ have the same shape iff there exists some location shift m and some scaling parameter $s > 0$ s.t. $G(sx + m) = F(x) \forall x$

i.e. $F()$ is the distribution of random variable X and $G()$ is the distribution of Y with $Y = sX + m$

Which distributions have the same shape ?

A. A and B
B. A and C
C. B and C
D. All have the same shape
E. All shapes are different
F. I don't know

A: $N(0,1)$
B: $N(\mu, \sigma^2)$ with $\mu \neq 0, \sigma \neq 1, \sigma > 0$
C: $N(\mu, 0)$ with $\mu \neq 0$

Location and Scale Parameters

A distribution in a catalog (e.g. Wikipedia) usually has many parameters; it is important to know which ones are simply location and scale parameters

E.g. $N(\mu, \sigma^2)$: μ is a location parameter, σ is a scale parameter

For the exponential distribution expo(λ), $\mathbf{1/\lambda}$ is …

A. A location parameter
B. A scale parameter
C. Both
D. None
E. I don't know

$\boldsymbol{F}()$ and $\boldsymbol{G}()$ have the same shape and $F()$ has a pdf $f()$ …

A. $\Rightarrow G$ also has a pdf $g()$ and
$g(x) = f\left(\frac{x-m}{s}\right)$ for some m and $s > 0$

B. $\Rightarrow G$ also has a pdf but the formula in A does not hold, in general

C. It may be that G does not have a pdf

D. I don't know

Standard Distribution

In a good catalog of distributions, only distributions with different shapes are worth mentioning. For each shape, we pick one particular distribution, which we call standard.

Standard normal: N(0,1)

Standard exponential: Exp(1)

Standard Uniform: U(0,1)

Log-Normal Distribution

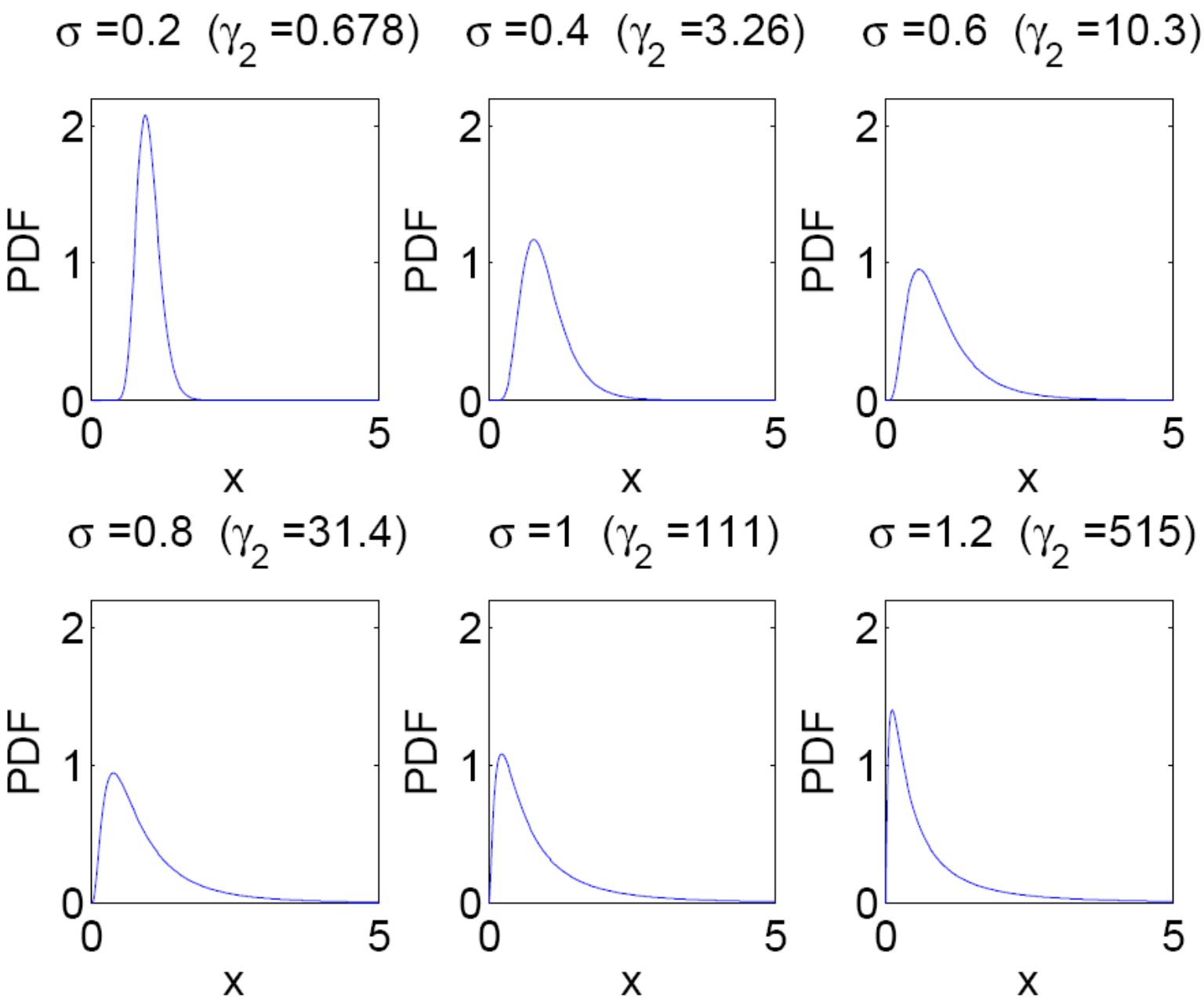
X has log-normal distribution iff $X = e^Z$ and Z has a normal (= Gaussian) distribution N_{μ,σ^2}

often used as a result of rescaling

Note that the support of X is $[0; +\infty)$

Furthermore $Z = \mu + \sigma Z_0$ with Z_0 standard normal $N_{0,1}$ hence

$$X = e^{\mu+\sigma Z_0} = e^{\mu}(e^{Z_0})^{\sigma}$$


σ =0.2 (γ2 =0.678)
σ =0.4 (γ2 =3.26)
σ =0.6 (γ2 =10.3)
σ =0.8 (γ2 =31.4)
σ =1 (γ2 =111)
σ =1.2 (γ2 =515)
PDF
x
0
1
2
5

For the log-normal distribution…

A. μ is a location parameter

B. σ is a scale parameter

C. A and B

D. None

E. I don't know

Log-Normal Distribution

X has log-normal distribution iff $X = e^Z$ and Z has a normal (= Gaussian) distribution N_{μ,σ^2}

often used as a result of rescaling

Note that the support of X is $[0; +\infty)$

Furthermore $Z = \mu + \sigma Z_0$ with Z_0 standard normal $N_{0,1}$ hence

$$X = e^{\mu+\sigma Z_0} = e^{\mu}(e^{Z_0})^{\sigma}$$

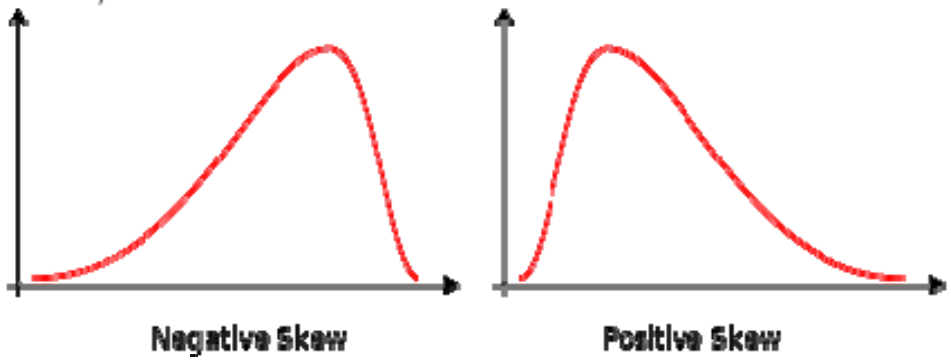
Feature 2: Skewness and Curtosis

$$\begin{cases} \kappa_1 = \mathbb{E}(X) \\ \kappa_2 = \mathbb{E}\left(X - \mathbb{E}(X)\right)^2 = \mathrm{var}(X) \\ \kappa_3 = \mathbb{E}\left(X - \mathbb{E}(X)\right)^3 \\ \kappa_4 = \mathbb{E}\left(X - \mathbb{E}(X)\right)^4 - 3\mathrm{var}(X)^2 \end{cases}$$

SKEWNESS INDEX κ_3 is called skewness. The *skewness index* is

$$\gamma_1 := \kappa_3/\kappa_2^{3/2} = \kappa_3/\sigma^3$$

Measures symmetry of distribution



By Rodolfo Hermans (Godot) at en.wikipedia. - Own work; transferred from en.wikipedia by Rodolfo Hermans (Godot)., CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=4567445

KURTOSIS INDEX κ_4 is called Kurtosis. The *Kurtosis index* is

$$\gamma_2 := \kappa_4/\kappa_2^2 = \kappa_4/\sigma^4$$

Measures departure from the Bell-shape of normal distribution
Equal to 0 for normal distribution

If $\boldsymbol{F}()$ and $\boldsymbol{G}()$ have the same shape…

A. They have the same skewness index
B. They have the same Kurtosis index
C. Both
D. None
E. I don't know

Feature 2: Skewness and Curtosis

$$\begin{cases} \kappa_1 = \mathbb{E}(X) \\ \kappa_2 = \mathbb{E}\,(X - \mathbb{E}(X))^2 = \mathrm{var}(X) \\ \kappa_3 = \mathbb{E}\,(X - \mathbb{E}(X))^3 \\ \kappa_4 = \mathbb{E}\,(X - \mathbb{E}(X))^4 - 3\mathrm{var}(X)^2 \end{cases}$$

SKEWNESS INDEX κ_3 is called skewness. The *skewness index* is

$$\gamma_1 := \kappa_3/\kappa_2^{3/2} = \kappa_3/\sigma^3$$

Measures symmetry of distribution

KURTOSIS INDEX κ_4 is called Kurtosis. The *Kurtosis index* is

$$\gamma_2 := \kappa_4/\kappa_2^2 = \kappa_4/\sigma^4$$

Measures departure from the Bell-shape of normal distribution
Equal to 0 for normal distribution

Regress+

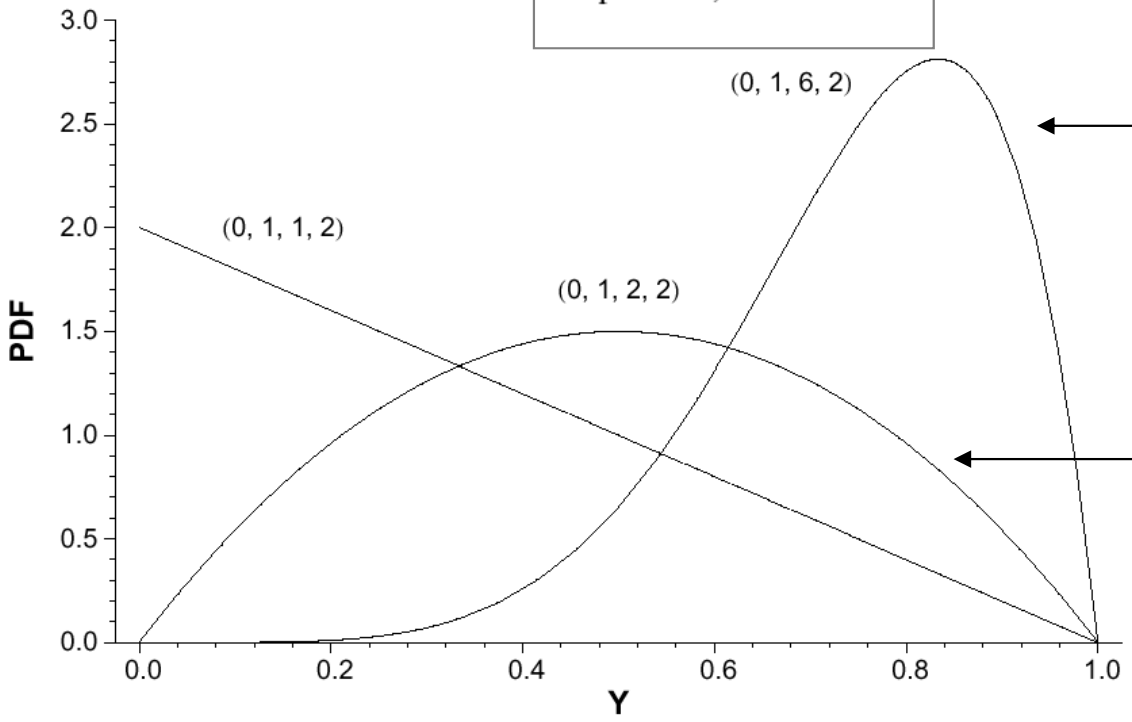
Appendix A

A Compendium of Common Probability Distributions

Michael P. McLaughlin
McLean, VA
September, 1999

[McLaughlin]

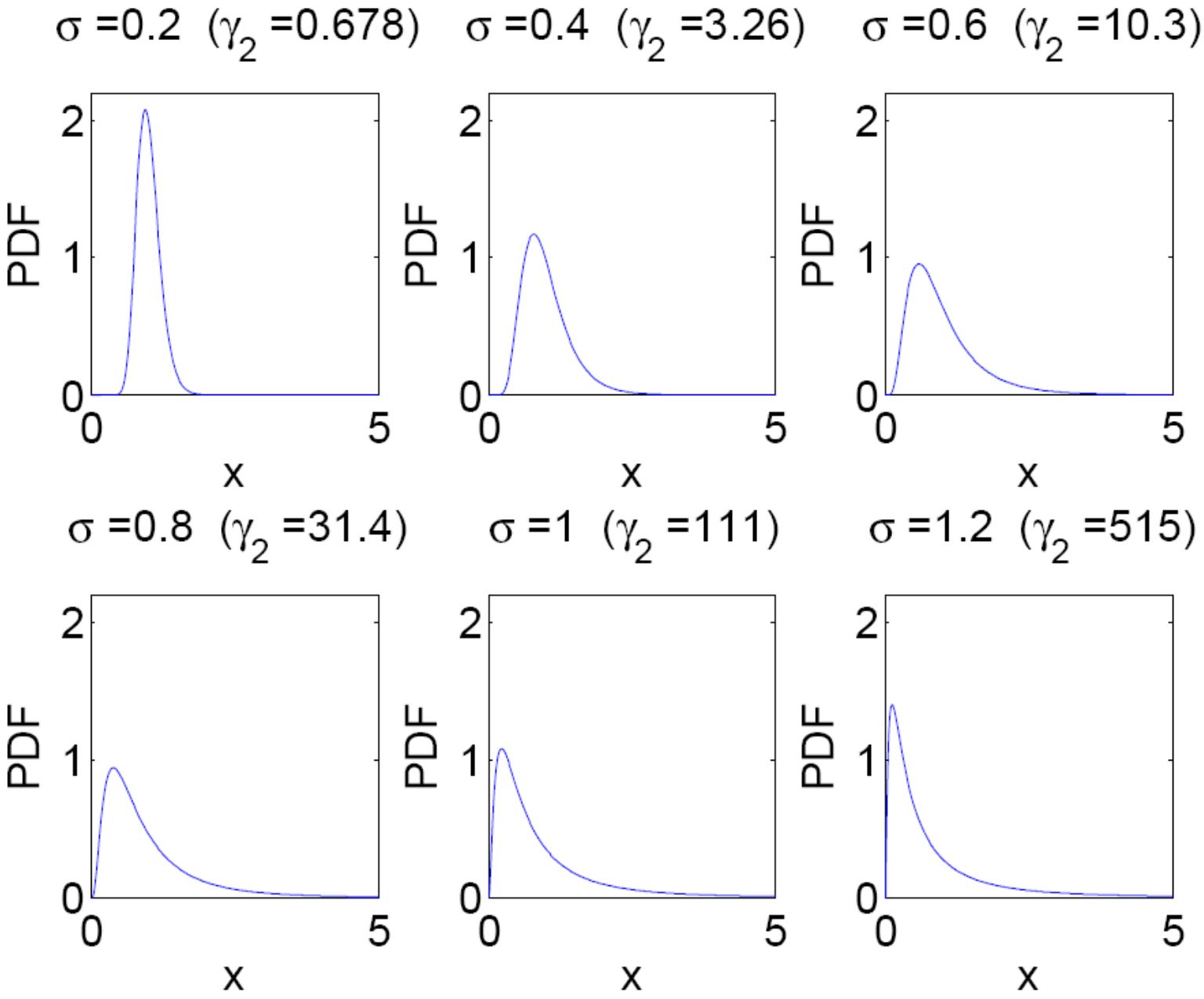
Beta(A,B,C,D) $y < B, \quad C, D > 0$



skewness < 0
kurtosis >0

skewness = 0
kurtosis < 0

Log-normal distributions



shape is independent of μ - μ chosen such that mean is 1

Jarque Bera test of normality (Chapter 4)

Based on Kurtosis and Skewness

Should be 0 for normal distribution

JARQUE-BERA. The *Jarque-Bera* statistic is used to test whether an iid sample comes from a normal distribution. It is equal to $\frac{n}{6}\left(\hat{\gamma}_1^2 + \frac{\hat{\gamma}_2^2}{4}\right)$, the distribution of which is asymptotically χ_2^2 for large sample size n. In the formula, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are the sample indices of skewness and kurtosis, obtained by replacing expectations by sample averages in Equation (6.3).

$$\begin{cases} \kappa_1 = \mathbb{E}(X) \\ \kappa_2 = \mathbb{E}\left(X - \mathbb{E}(X)\right)^2 = \mathrm{var}(X) \\ \kappa_3 = \mathbb{E}\left(X - \mathbb{E}(X)\right)^3 \\ \kappa_4 = \mathbb{E}\left(X - \mathbb{E}(X)\right)^4 - 3\mathrm{var}(X)^2 \end{cases}$$

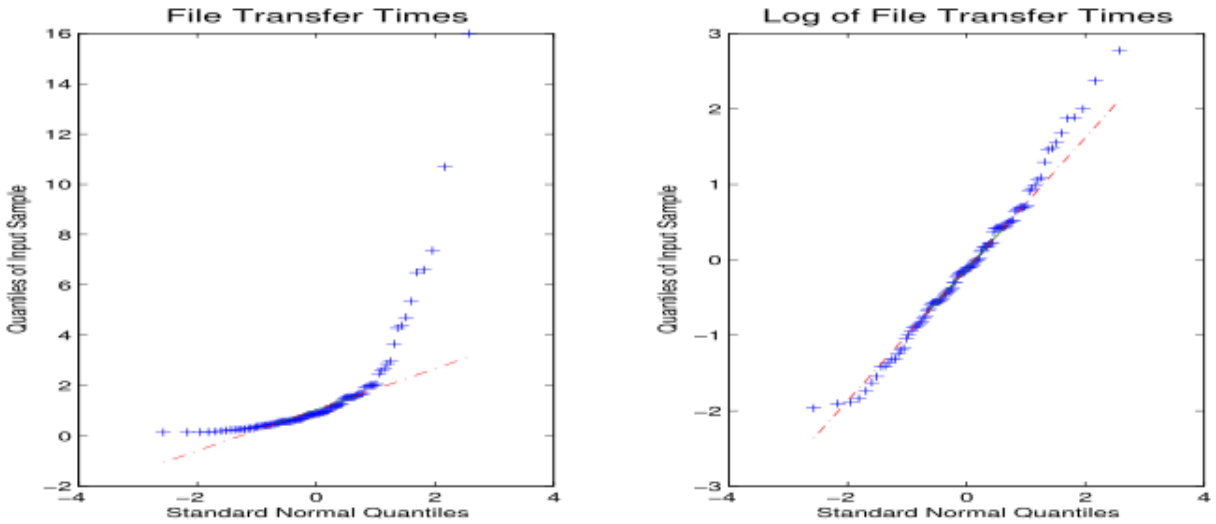


Figure 4.4: Normal qqplots of file transfer data and its logarithm.

EXAMPLE 4.18: APPLICATION TO EXAMPLE 4.17. We would like to test whether the data in Example 4.17 and its transform are normal.

```
Original Data          h = 1          p =    0.0010
Transformed Data       h = 0          p =    0.1913
```

The conclusions are the same as in Example 4.17, but for the original data the normality assumption is clearly rejected, whereas it was borderline in Example 4.17.

Feature 3: Power Laws

Zipf's "law": probability that the jth most popular object in a catalog (e.g. movies on Netflix): is chosen is proportional to $1/j^{p+1}$ for some index p.

Empirically found to for recommendation systems, for distribution of file sizes, of cluster groups in facebook, of incomes in a population etc…

Model 1 (Zipfian distribution): user picks one out of N objects; proba that object $j \in \{1, \dots N\}$ is selected is $\theta_j = \frac{\eta}{j^{p+1}}$ where η is some constant and $p \geq 0$.

Model 2 (Zeta distribution): user picks one out of an infinite collection of objects; proba that object $j \in \mathbb{N}^+$ is selected is $\theta_j = \frac{\eta}{j^{p+1}}$ where η is some constant and $p > 0$

Pareto Distribution

Model 3 (Pareto distribution): user picks one object with feature $x \in [1, \infty)$; PDF of feature x is $f(x) = \frac{p}{x^{p+1}}$ with $p > 0$

Pareto distribution is the continuous approximation of Zeta and Zipf. Is much easier to use:

Standard Pareto with index $p > 0$:

PDF: $f(x) = \frac{p}{x^{p+1}} 1_{\{x \geq 1\}}$; CDF: $F(x) = \left(1 - \frac{1}{x^p}\right) 1_{\{x \geq 1\}}$

Complementary CDF (CCDF, Survival function):

$\mathbb{P}(X > x) = F(x) = \frac{1}{x^p}$ for $x \geq 1$

PDF and CCDF of Pareto follow a power law
i.e. a relation of the form $y = ax^b$ for some a, b.

In log-log scale, $\log y = \log a + b \log x$: a linear relation

For the Pareto distribution with index p…

Standard Pareto with index $p > 0$:

PDF: $f(x) = \frac{p}{x^{p+1}} 1_{\{x \geq 1\}}$; CDF: $F(x) = \left(1 - \frac{1}{x^p}\right) 1_{\{x \geq 1\}}$

Complementary CDF: $\mathbb{P}(X > x) = F(x) = \frac{1}{x^p}$ for $x \geq 1$

A. p is a location parameter

B. p is a scale parameter

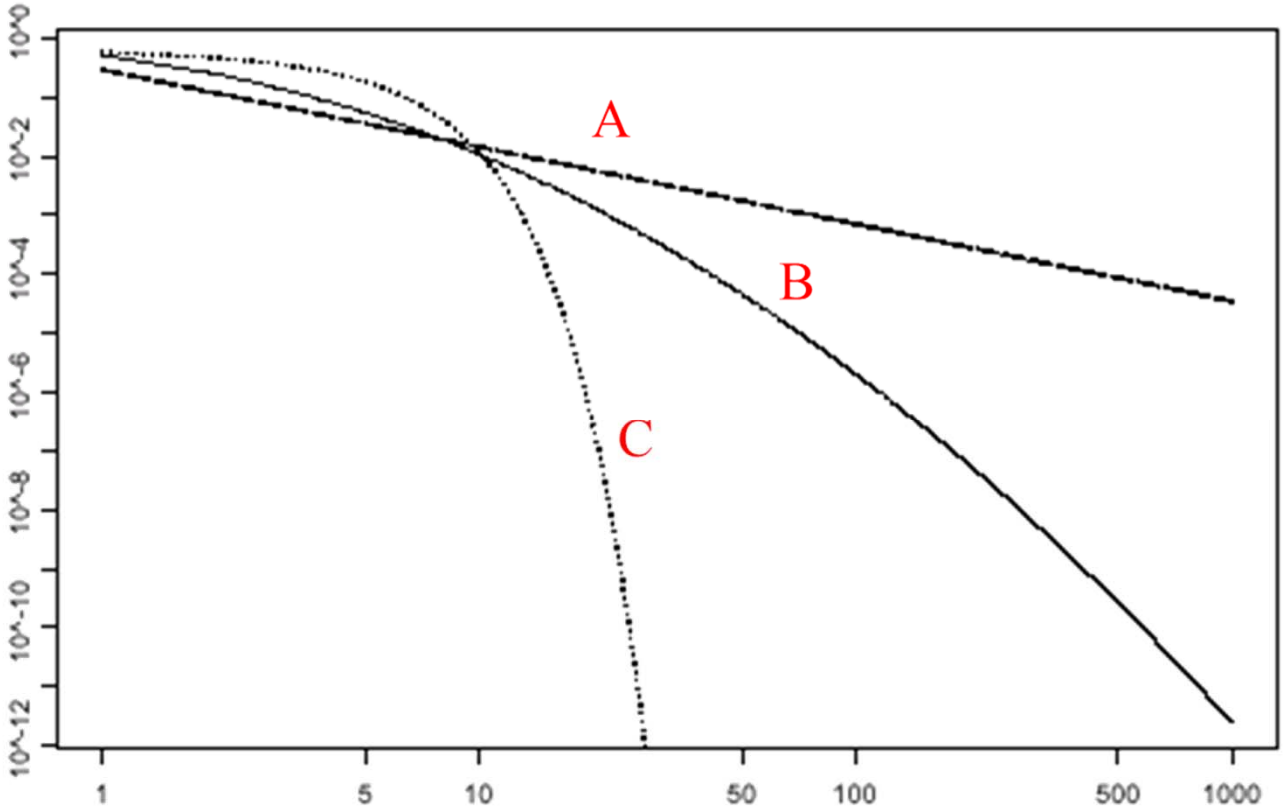
C. A and B

D. None

E. I don't know

Complementary Distribution Functions in Log-log Scales
Say which is what

A. A-lognormal, B- pareto, C-normal
B. A-p, B-l, C-n
C. A-l, B-n, C-p
D. A-n, B-l, C-p
E. A-p, B-n, C-l
F. A-n, B-p, C-l
G. I don't know



Feature 4: Hazard Rate

A property of the tail of the distribution

Definition: $\lambda(x) = \lim_{dx \to 0} \frac{1}{dx} P(X \leq x + dx \mid X > x)$

$$= \lim_{dx \to 0} \frac{1}{dx} P(x < X \leq x + dx \mid X > x)$$

interpret X as a lifetime; $\lambda(x)$ is the rate of death given reached age x

$\lambda(x) = \frac{f(x)}{1-F(x)}$ where $f =$ pdf and $F =$ CDF

Used to classify distributions

Aging: $\lim_{x \to \infty} \lambda(x) = \infty$

Memoryless: $\lim_{x \to \infty} \lambda(x) = c > 0$

Fat tail (Vanishing Hazard Rate): $\lim_{x \to \infty} \lambda(x) = 0$

Which is what ?

A. Exponential: aging; Pareto: memoryless, Normal: fat-tailed
B. E-M; P-A; N-F
C. E-A; P-F; N-M
D. E-F; P-A; N-M
E. E-F; P-M; N-A
F. E-M; P-F; N-A
G. I don’t know

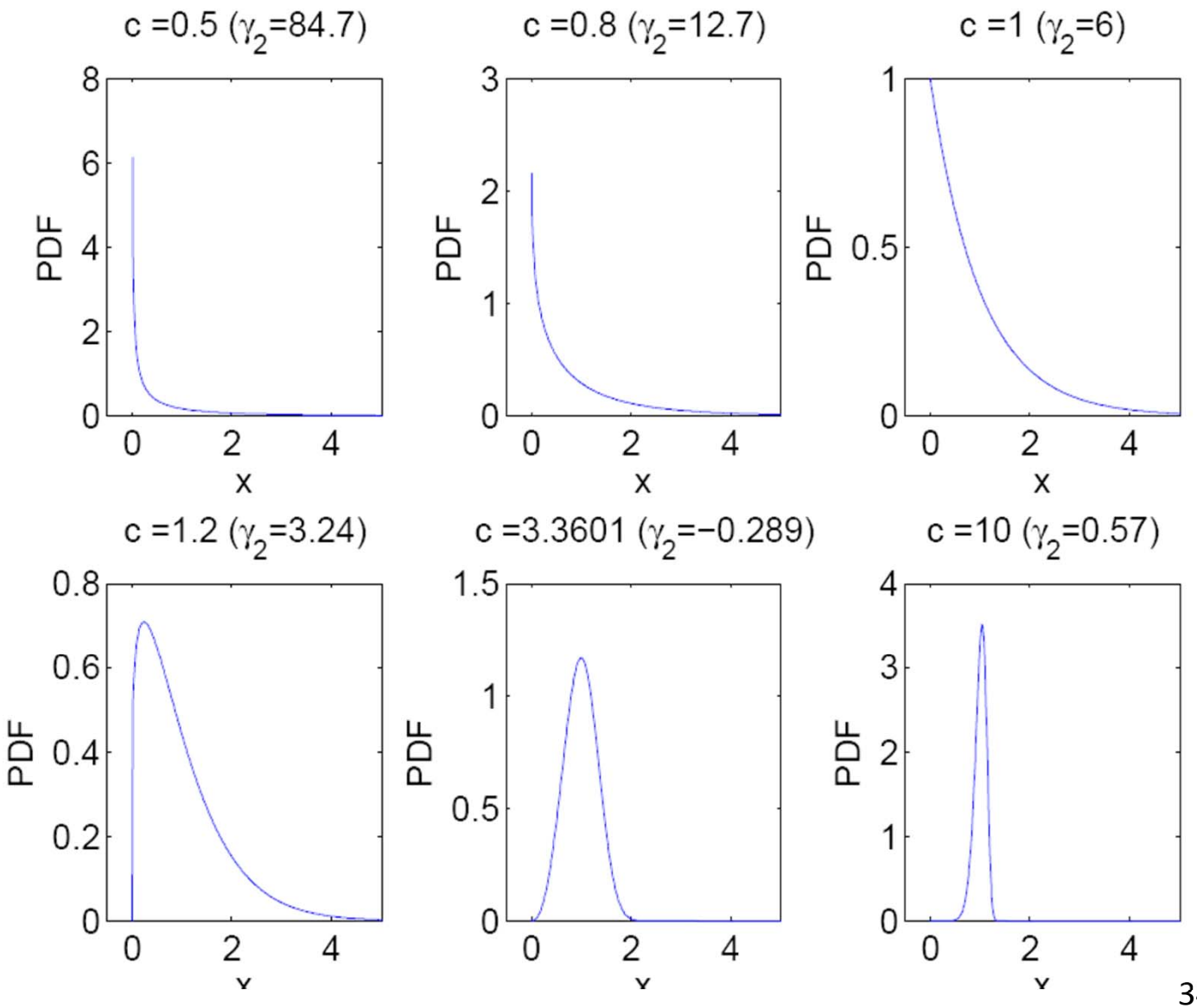
The Weibull Distribution with shape parameter *c*

Standard Weibull $F(x) = 1 - e^{-(x^c)}$

Aging for c > 1

Memoryless for c = 1

Fat tailed for c <1



2. Heavy Tail

Recall what fat tail / vanishing hazard rate is -- Heavier than fat tail is *heavy tail*

A property found in many high resolution data sets: financial data, network measurements at second time scale, power grid measurements at 50 Hz, etc where rare but very large values exist

A distribution defined on $[a, \infty)$ with CDF F is heavy tailed with index $p, 0 < p < 2$ if

$$1 - F(x) \sim \frac{k}{x^p} \text{ for } x \to \infty \text{ for some constant } k$$

$\Rightarrow$ Variance is infinite

$\Rightarrow$ for $0 < p < 1$ mean is also infinite

NB: we use the terminology used e.g. by Taqqu and Crovella. Other (confusing) definitions exist. Our definition of heavy tail always implies infinite variance.

Examples

Pareto distribution: $1 - F(x) = \frac{p+1}{x^p}$ is heavy tailed for $0 < p < 2$

Log-normal distribution is not heavy tailed (its variance is finite)

Weibull distribution: $1 - F(x) = e^{-(x^c)}$ is not heavy tailed

One-sided Cauchy distribution $f(x) = \frac{2}{\pi(1+x^2)} 1_{\{x \geq 0\}}$ is heavy tailed with $p = 1$

Heavy Tail means Central Limit does not hold

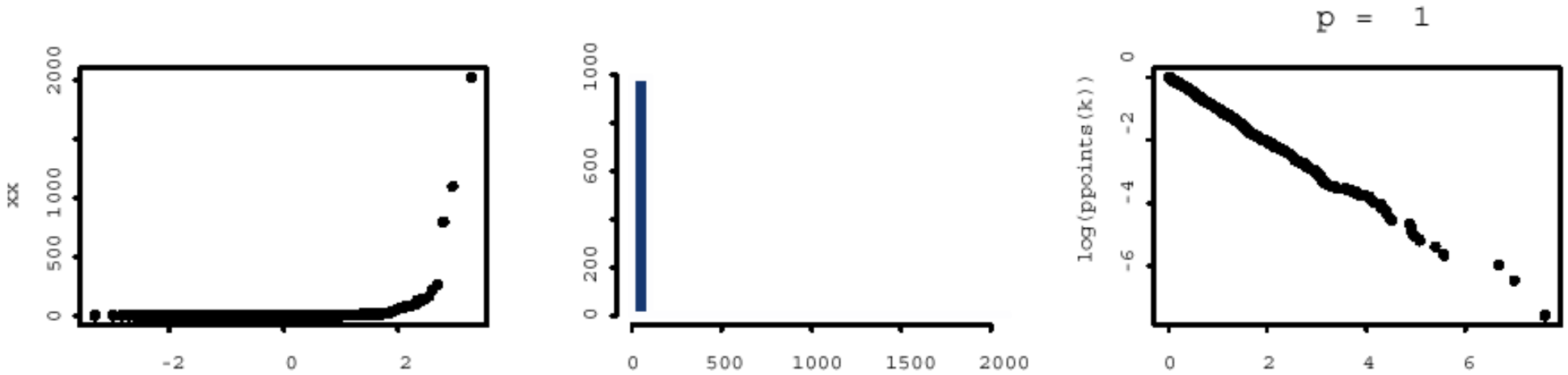
Central limit theorem:

a sum of *n* independent random variables with finite second moment tends to have a normal distribution, when *n* is large

this explains why we can often use normal assumption

But it does not always hold. It does not hold if random variables have infinite second moment.

Central Limit Theorem for Heavy Tails

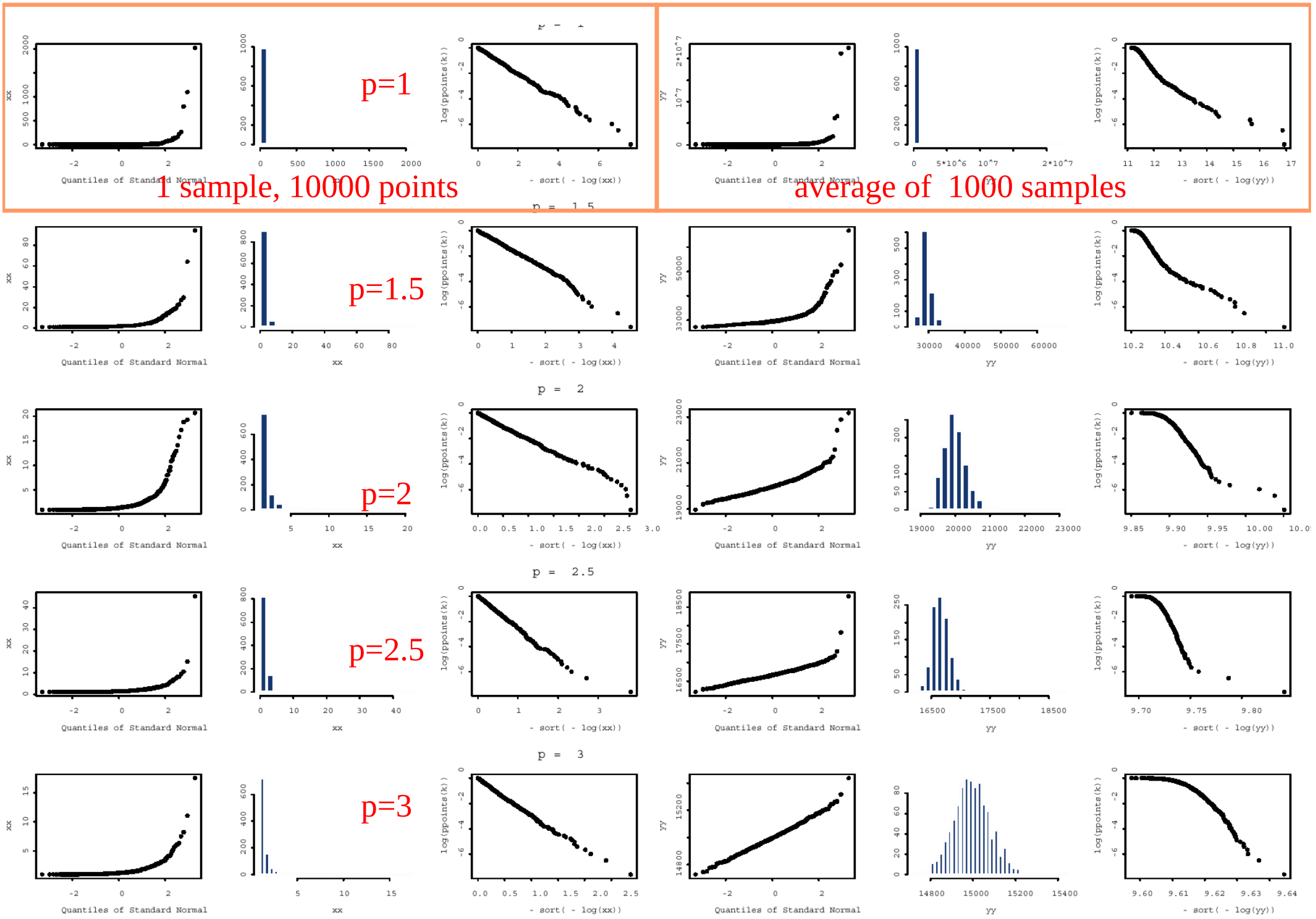


normal qqplot

histogram

complementary d.f.
log-log

One Sample of 10000 points
Pareto p = 1


p=1
1 sample, 10000 points
average of 1000 samples
p = 1.5
p=1.5
p = 2
p=2
p = 2.5
p=2.5
p = 3
p=3
Quantiles of Standard Normal
- sort(- log(xx))
- sort(- log(yy))
xx
yy

Central Limit for heavy tailed distributions

What we saw on previous example is an application of a general theorem that say: the aggregation $X_1 + \cdots + X_n$ of n iid random variables has approximately, for large n,

a normal distribution if not heavy tailed

a "*stable*" distribution if heavy tailed ; the limit is heavy tailed with same index p

The stable distribution is a family of distributions with two shape parameters p and β. It is closed by aggregation.

Heavy tail is conserved by aggregation

Standard Stable Distributions with Index p

For $p \in]0,2[$, there is one standard stable distribution for each p and for each $\beta \in [-1,1]$, a shape parameter similar to skewness; they are all heavy-tailed (or constant).

No closed form for the pdf or CDF, no easy computation of CDF or inverse CDF.

Closed form for Fourier transform of CDF.

Hard to use in practice -- you can replace it by a mixture with Pareto tail

For $p = 2$, stable = normal

$\bar{X}$ is the mean of n iid random variables $X_1, \ldots X_n$, n is large ($\bar{X}$ is not constant) Say what is true.

A. If X_i is heavy tailed with index p (<2), then $\bar{X}$ has approximately a stable distribution with index p
B. If X_i is Pareto with index p, then $\bar{X}$ is also Pareto with index p
C. If X_i is Pareto with index p, then $\bar{X}$ is heavy tailed with index p
D. A and B
E. A and C
F. B and C
G. All
H. None
I. I don't know

$\bar{X}$ is the mean of n iid random variables $X_1, \ldots X_n$ n is not large. Say what is true.

A. If X_i is stable with index p, then $\bar{X}$ is also stable with index p
B. If X_i is normal then $\bar{X}$ is normal
C. A and B
D. None
E. I don't know

Application to Confidence Intervals

$X_i \sim$ iid Standard Pareto $p = 1.25$

i.e. $f_X(x) = \frac{1.25}{x^{2.25}}$ for $x \geq 1$

True mean of X_i is $\mu = 5$ and true median is 1.74

Assume we don't know the model and have received a sample of n values, where n is large; we want to compute the mean and the median of X_i

---

EXAMPLE 2.7: PARETO DISTRIBUTION. This is a toy example where we generate artificial data, iid, from a Pareto distribution on $[1, +\infty)$. It is defined by its cdf equal to $F(c) := \mathbb{P}(X > c) = \frac{1}{c^p}$ with $p = 1.25$; its mean is $= 5$, its variance is infinite (i.e. it is heavy tailed) and its median is 1.74.

Assume we would not know that it comes from a heavy tailed distribution and would like to use the asymptotic result in Theorem 2.2 to compute a confidence interval for the mean.

Which formula is correct (confidence level =0.95)?

A. An approximate confidence interval for the mean is $x \pm 1.96\frac{s}{\sqrt{n}}$ with $\bar{x} = \frac{1}{n}\sum_i x_i$ and $s = \frac{1}{n}\sum_i (x_i - \bar{x})^2$

B. An approximate confidence interval for the median is $[x_{(i)}, x_{(j)}]$ with $i = \lfloor 0.5\, n - 0.980\sqrt{n} \rfloor$ and $j = \lceil 0.50\, n + 1 + 0.980\sqrt{n} \rceil$

C. A and B

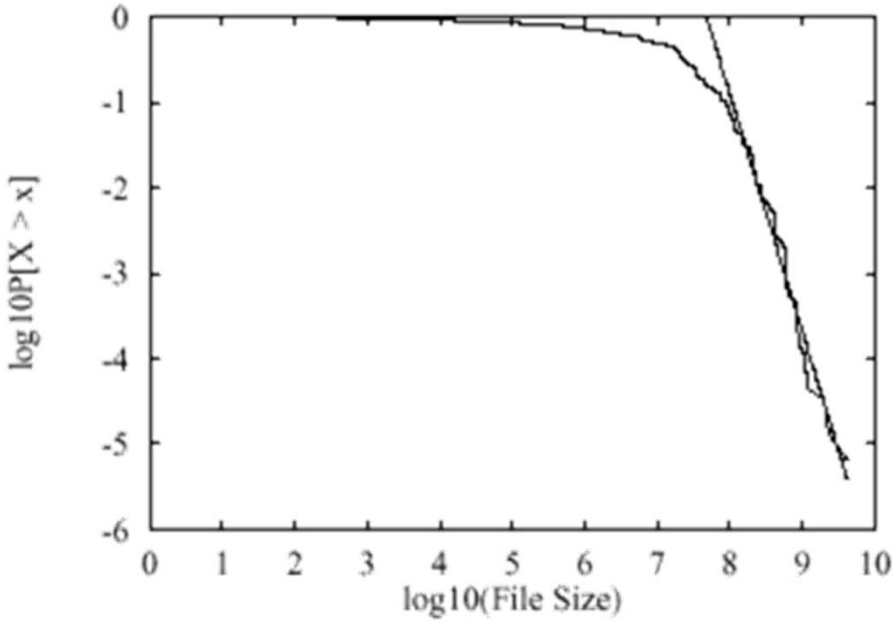
D. None

E. I don't know

Deciding for Heavy Tail

Assume you have very large data set and you suspect your distribution has infinite variance --- hard to test because everything we do is finite

An alternative is to look for heavy tail by plotting CCDF in log-scale



Estimating the index p can be done with Taqqu's method (see lecture notes)

3. Fitting A Distribution

Assume iid

Use maximum likelihood

We know how to do this (model fitting)

-> maximum likelihood estimation

Frequent issues

- Censoring
- Combinations

Censored Data

Example: We want to propose a distribution for file sizes transferred over a network; we think a lognormal distribution is adequate but we can never observe very large values, by the nature of the experiment

Lognormal is fat tailed so we cannot ignore the tail

Idea: we assume that what we observed is produced by the following simulator

sample $X \sim F_0()$ (a log-normal distribution)
if $X \leq a$ deliver X else drop X

The samples produced by this model have the following pdf... (f_0 = pdf of F_0)

A. $f(x) = f_0(x) \times 1_{\{x \leq a\}} \times \text{constant}$

B. $f(x) = f_0(x - a)$

C. $f(x) = \frac{1}{a} f\left(\frac{x}{a}\right)$

D. None of the above

E. I don't know

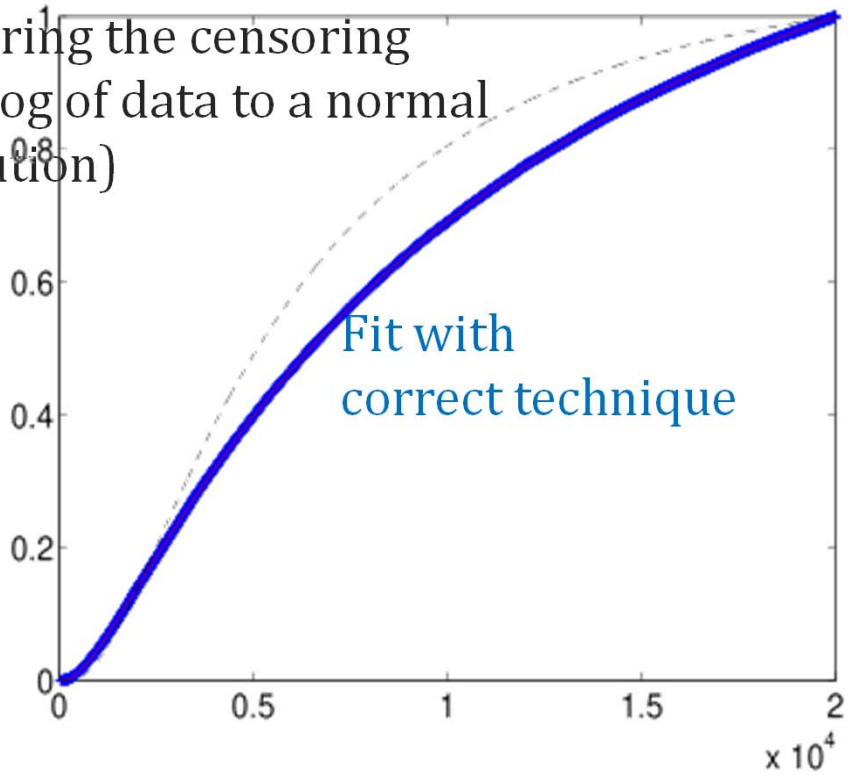
Idea: we assume that what we observed is produced by the following simulator

sample $X \sim F_0()$ (a log-normal distribution)

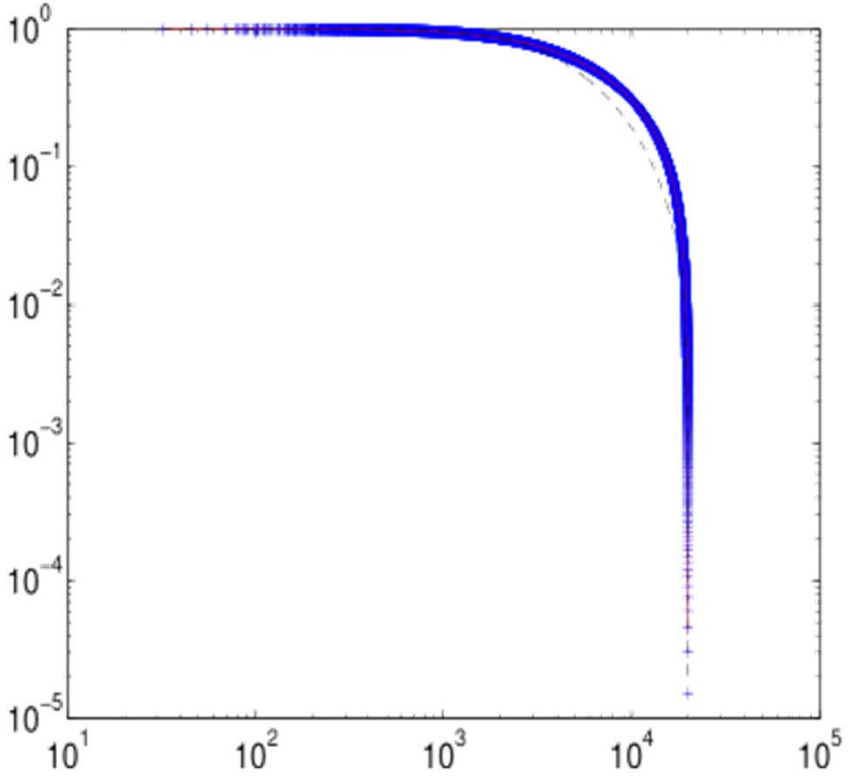
if $X \leq a$ deliver X else drop X

EXAMPLE 3.10: CENSORED LOG-NORMAL DISTRIBUTION. Figure 3.7(a) shows an artificial data set, obtained by sampling a log-normal distribution with parameters $\mu = 9.357$ and $\sigma = 1.318$, truncated to 20000 (i.e. all data points larger than this value are removed from the data set).

Fit ignoring the censoring (i.e. fit log of data to a normal distribution)



(a) CDF



(b) CCDF in log-log scales

Combinations

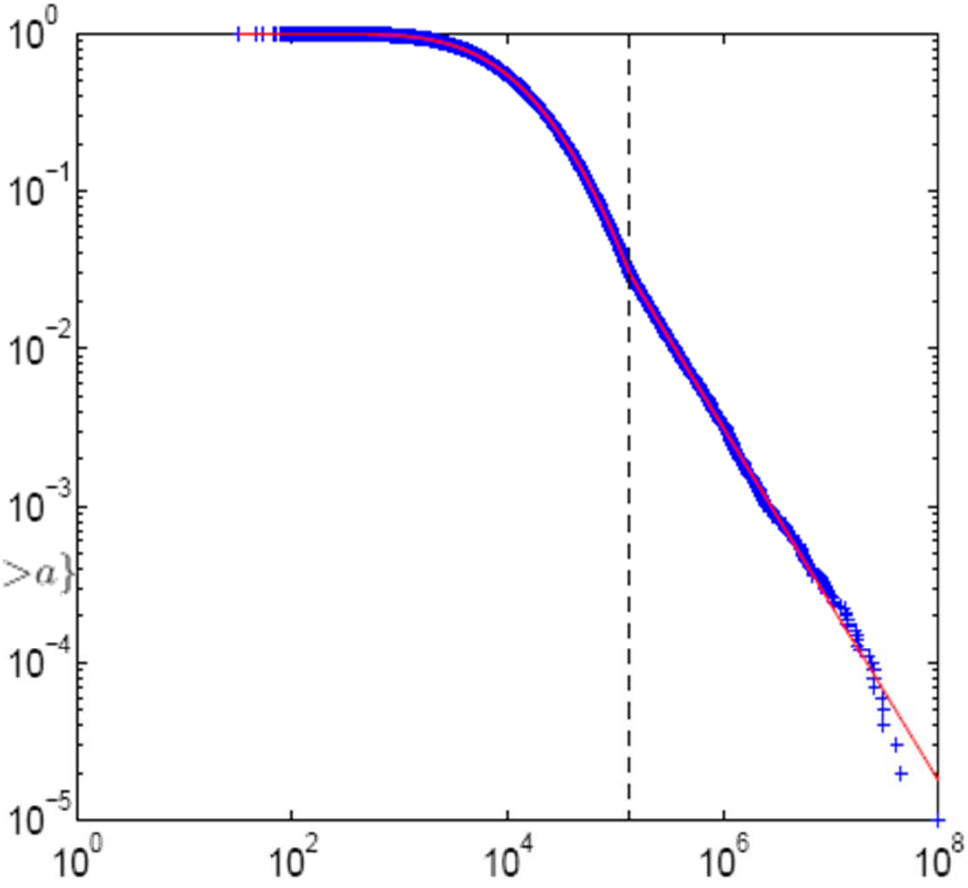
We want to fit a log normal distrib to the body and pareto to the tail

Often used when the tail is well identified

This corresponds to the pdf

$$f_X(x) = q\frac{f_1(x)}{F_1(a)}\mathbf{1}_{\{x\leq a\}} + (1-q)\frac{f_2(x)}{1-F_2(a)}\mathbf{1}_{\{x>a\}}$$

with $q \in [0,1]$



(a) CCDFs

This pdf corresponds to the following simulator output

A. A
B. B
C. None
D. Both
E. I don't know

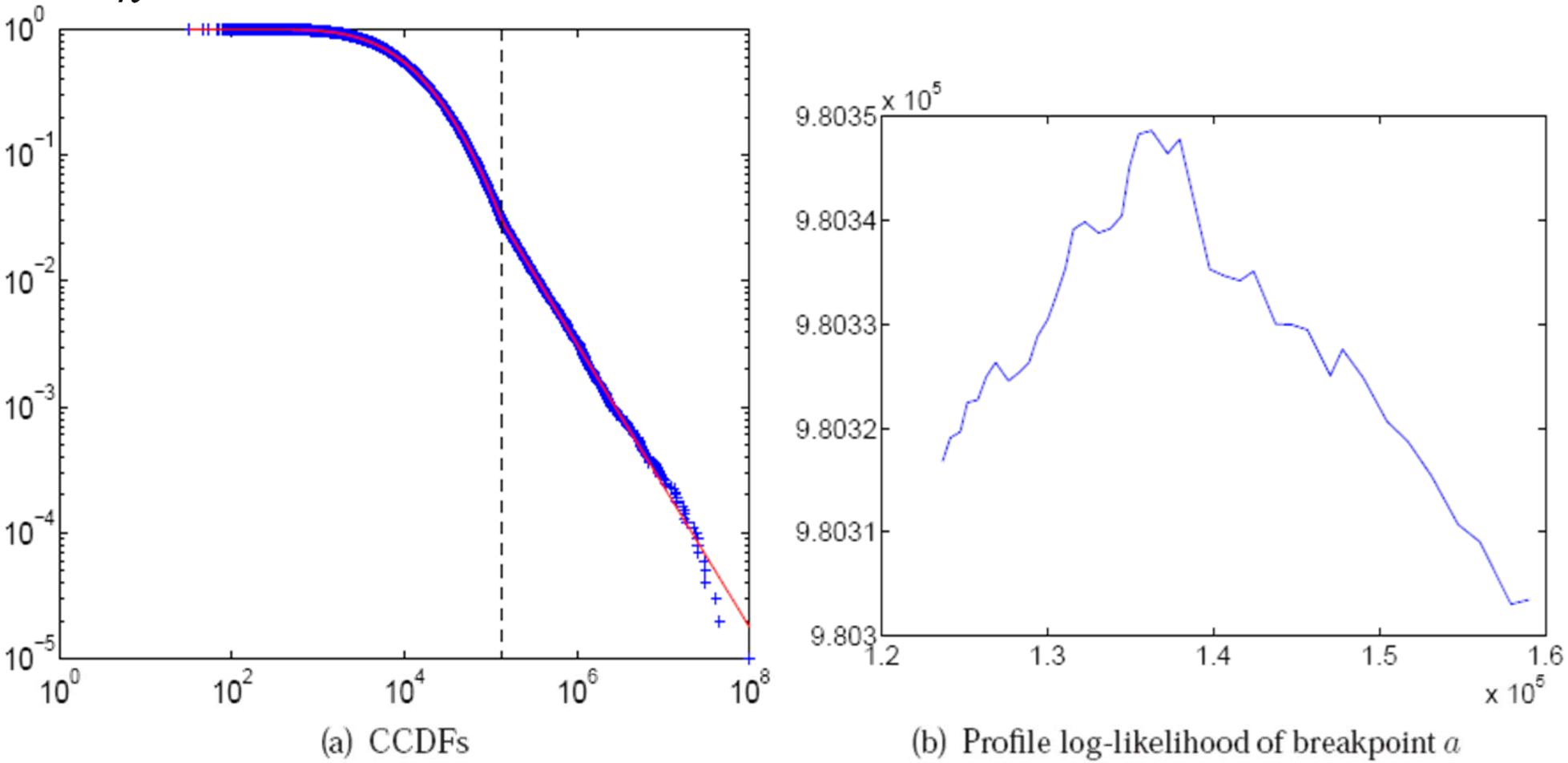
A:
Draw U ~U(0,1)
if $U < q$ draw X from F_1 until $X \leq a$
  else draw X from F_2 until $X > a$
Deliver X

A:
Draw U ~U(0,1)
if $U < q$ draw X from F_1 until $X \leq a$
  else draw X from F_2 until $X > a$
Deliver X

B:
Draw X from F_1 until $X \leq a$
Draw Y from F_2 until Y $> a$
With proba q deliver X else deliver Y

Maximum Likelihood estimation of combination

Solved by brute force, after observing that the MLE satisfies $\hat{q} = \frac{n_1(a)}{n}$ where $n_1(a) = \sum_{i=1:n} 1_{x_i \leq a}$



(a) CCDFs

(b) Profile log-likelihood of breakpoint a

Figure 3.8: Fitting a combination of Log-Normal for the body and Pareto for the tail. Dashed vertical line: breakpoint.

4 Illustration A Load Generator: Surge

Designed to create load for a web server

Sophisticated load model by Crovella and Barford

It is an example of a well constructed benchmark.

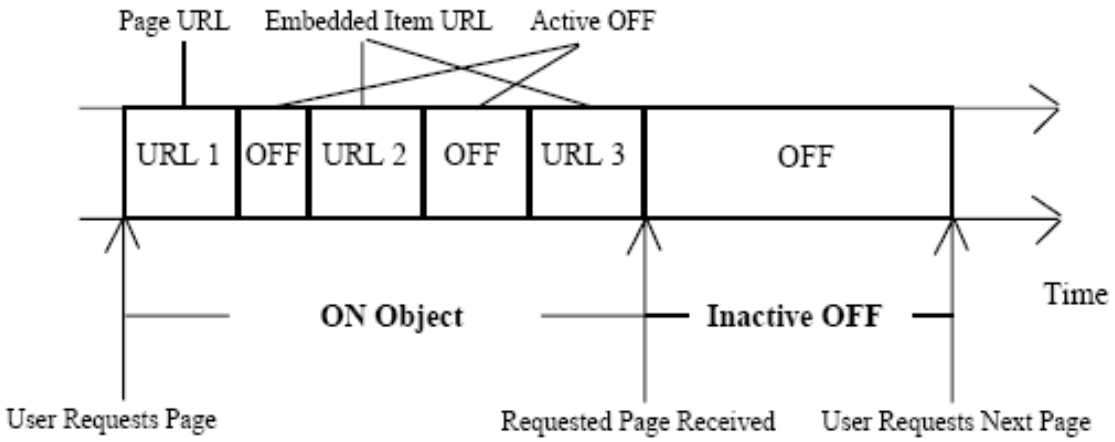
User Equivalent Model

Idea: find a stochastic model that represents user well

User modelled as sequence of downloads, followed by “think time”

Tool can implement several “user equivalents”

Used to generate real work over TCP connections



Characterization of UE

1. One UE alternates between ON-object periods and "Inactive OFF periods". Inactive OFF periods are iid with a Pareto distribution (Table 6.1).
2. During an ON-object period, a UE sends a request with embedded references. Once the first reference (URL1) is received, there is an "Active OFF period", then the request for the second reference is sent, and so on, until all embedded references are received. There is only one TCP connection at a time per UE, and one TCP connection for each reference (an assumption that made sense with early versions of HTTP).
3. The active OFF times are modelled as iid random variables with Weibull dsitributions
4. The number of embedded references is modelled as a set of iid random variables, with a Pareto distribution.

Fitting the distributions

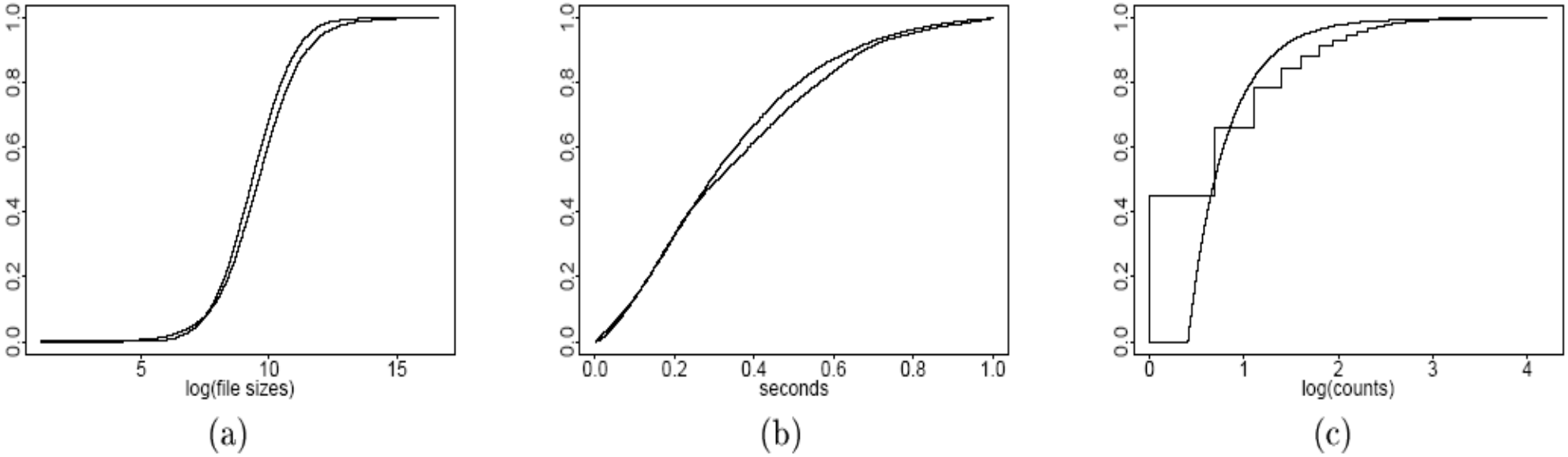


Figure 2: CDF of (a) Log-transformed File Sizes vs. Fitted Normal Distribution (b) Active OFF Times vs. Fitted Weibull Distribution (c) Embedded Reference Count vs. Fitted Pareto Distribution