Aalto University
School of Electrical Engineering

Network Traffic Measurements and Analysis

Lecture III: Probability models and measurements

Markku Liinaharja (slides originally made by Esa Hyytiä),

Department of Communications and Networking
Aalto University, School of Electrical Engineering

Version 0.2, October 4, 2017

Contents

Measurements and modeling

- Exploratory approach

  Problem $\Rightarrow$ Data $\Rightarrow$ Analysis $\Rightarrow$ (Model) $\Rightarrow$ Conclusions
- Measurement analysis is often intertwined with traffic modeling
- If the observations can be described using an idealized mathematical model their implications are often easier to understand
- E.g., input to a queue
- “All models are wrong, but some models are useful”

Why models?

- **Descriptive models** for measurements
  - Efficient summary of observed data
  - E.g., Gaussian with mean 100ms and standard deviation 10ms
- **Constructive models** for what-if scenarios
  - Model that could have produced the observed data
  - E.g., The trace could have been produced by a certain stochastic process
  - We are interested in the underlying phenomena instead of details of data
- In this lecture, we focus on descriptive probability models

Probability models

- Probability captures features that are unknown or difficult to characterize
  - Exact user behavior
  - Immense amount of functionalities in the Internet
- Probability allows us to model, reason, and proceed with inference in an uncertain environment

Modeling process

1. Model selection
   - Prefer models with a few parameters over those that have more parameters (Occam's razor)
   - Model should be parsimonious to avoid over-fitting
2. Parameter estimation
   - Choose the parameters that best describe the observed data
3. Validation
   - Descriptive: Compare the distributions
   - Constructive: Confirm that the observed data is relatively likely outcome of the model

Contents

Probability

- Consider an experiment with random outcomes
- Event is a set of outcomes
- **Probability**
  - How likely an event is?
  - Long-run proportion of an event in a series of experiments
- Mathematically defined by three axioms, for an event A
  i) $0 \leq \mathrm{P}\{A\} \leq 1$
  ii) $\mathrm{P}\{S\} = 1$, where S is the sample space
  iii) A and B disjoint $\Rightarrow \mathrm{P}\{A \text{ or } B\} = \mathrm{P}\{A\} + \mathrm{P}\{B\}$
- Everything else follows from these!

Useful notions

- **Conditional probability**:

$$\mathrm{P}\{B \mid A\} = \frac{\mathrm{P}\{A \text{ and } B\}}{\mathrm{P}\{A\}} = \frac{\mathrm{P}\{A \cap B\}}{\mathrm{P}\{A\}}.$$

- **Independence**: If the two events are independent, then

$$\mathrm{P}\{B \mid A\} = \mathrm{P}\{B\}.$$

- **Bayes' rule**: assume $S = B_1 \cup B_2 \cup \ldots \cup B_n$, where $B_i \cap B_j = \emptyset$ for $i \neq j$. Then,

$$\mathrm{P}\{B_i \mid A\} = \frac{\mathrm{P}\{A \mid B_i\} \cdot \mathrm{P}\{B_i\}}{\sum_j \mathrm{P}\{A \mid B_j\} \cdot \mathrm{P}\{B_j\}}.$$

Random variables

- Random variable X is a variable whose value is a numerical outcome of a random phenomenon
  - Coin toss: "heads" $X = 1$, "tails" $X = 0$
- **Discrete random variable** X takes discrete values $\{x_1, x_2, x_3, \ldots\}$ with probabilities $\{p_1, p_2, p_3, \ldots\}$
- **Continuous random variable** X takes continuous values $\{x\}$ according to a probability density function $f(x)$ (pdf)
  - Probability of event $x \in (a, b)$ is the area under pdf,

$$\mathrm{P}\{a < X < b\} = \int_a^b f(x)\, dx$$

Expectation and variance of a random variable

- **Expectation** (mean of a random variable) $\mu = \mathrm{E}(X)$

$$\mu = \sum_i p_i x_i, \qquad \mu = \int x \cdot f(x)\, dx.$$

- **Variance** $\mathrm{V}(X) = \mathrm{E}((X - \mu)^2)$,

$$\mathrm{V}(X) = \sum_i p_i (x_i - \mu)^2, \qquad \mathrm{V}(X) = \int (x - \mu)^2 \cdot f(x)\, dx.$$

- **Covariance** of two random variables

$$\mathrm{Cov}[X, Y] = \mathrm{E}((X - \mu_x)(Y - \mu_y)).$$

Properties of expectation and variance

- For constants a, b, and random variables X and Y

$$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y).$$

- For constants a, b, and a random variable X

$$\mathrm{V}(aX + b) = a^2\,\mathrm{V}(X).$$

- For *independent* random variables X and Y

$$\mathrm{V}(X + Y) = \mathrm{V}(X) + \mathrm{V}(Y).$$

$$\mathrm{E}(XY) = \mathrm{E}(X) \cdot \mathrm{E}(Y).$$

Contents

- Introduction
- Probability
- Distributions for network measurements
- Parameter estimation
- Model validation

Distributional models

Distributional (analytic) models for measurement data have certain benefits

- Models can be manipulated mathematically, leading to improved understanding
- Models are concise and easily communicated (only a few parameters)
- Values of model's parameters can give insight into the nature of the underlying data (distribution varies predictably, when its parameters are varied)
- Models can take into account features that have not been observed or external knowledge

Discrete distributions

- Discrete distributions are characterized by the probability mass function (pmf)

$$p_i = p(x_i) = \mathrm{P}\{X = x_i\}.$$

- Cumulative distribution $\mathrm{P}\{X \leq x\}$ is given by

$$F(x) = \sum_{i:x_i \leq x} p(x_i).$$

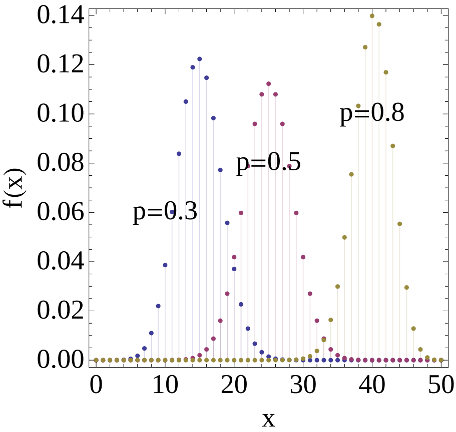
Binomial distribution

Number of successes (each with probability p) in n attempts takes values $\{0, 1, 2, \ldots, n\}$.

<u>Bin(n, p)</u>:

$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}.$$

Mean μ:	np
Variance σ^2:	$np(1-p)$



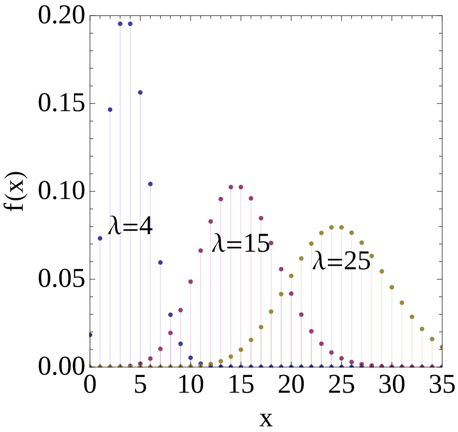
Poisson distribution

When $n \to \infty$ and $p \to 0$ in binomial distribution so that $np = \lambda$ takes values $\{0, 1, 2, \ldots\} \Rightarrow$ Poisson distribution.

<u>Poisson(λ)</u>:

$$P\{X = i\} = \frac{\lambda^i}{i!} e^{-\lambda}.$$

Mean μ:	λ
Variance σ^2:	λ



Zipf's Law

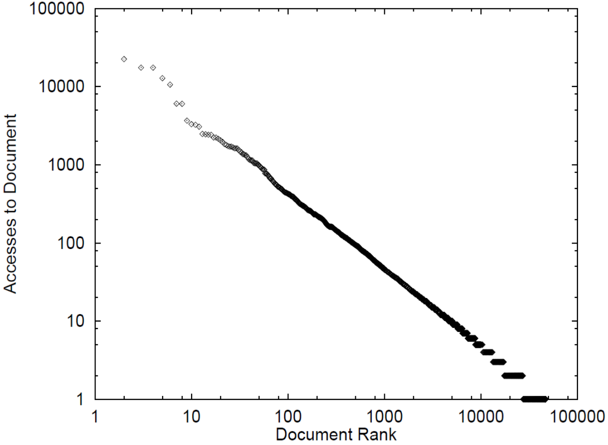
- Consider a set of categorical variables, e.g., URLs of web pages sorted in decreasing number of references made to each page
  - R number of references to a page
  - n rank of the page
- Then, for some constants c and β, Zipf's law states that

$$R = c\,n^{-\beta}.$$

- "Discrete power-law distribution"
- Linear in log-log plot

$$\log R = \log c - b \log n.$$

Example: Zipf's Law Applied To WWW Documents



Source: C.Cunha, A. Bestavros, M. Crovella, Characteristics of WWW Client-based Traces, Tech. Report BU-CS-95-010, Boston University, 1995.

Continuous distributions

- A continuous random variable has a probability density function (pdf), denoted by $f(x)$
- Cumulative distribution function (cdf) defines the probability $\mathrm{P}\{X \leq x\}$ and it is denoted by $F(x)$,

$$F(x) = \int_{-\infty}^{x} f(x)\, dx.$$

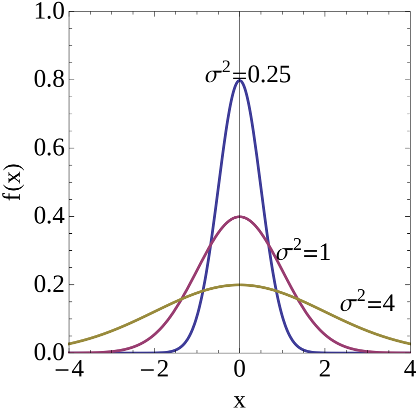
Normal distribution N(μ, σ^2)

Normal (Gaussian) distribution is denoted by N(μ, σ^2) with mean μ and variance σ^2.

Probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Mean:	μ
Variance:	σ^2



Exponential distribution Exp(λ)

Exponential distribution with intensity λ is denoted by Exp(λ).

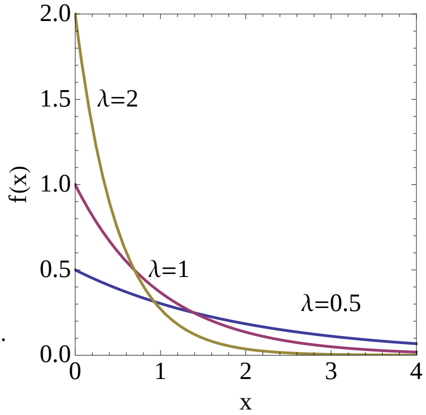
Probability density function:

$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$

Mean μ:	$1/\lambda$
Variance σ^2:	$1/\lambda^2$

Memorylessness property:

$P\{X > x+t \mid X > t\} = P\{X > x\}.$



Gamma distribution Gamma(p, λ)

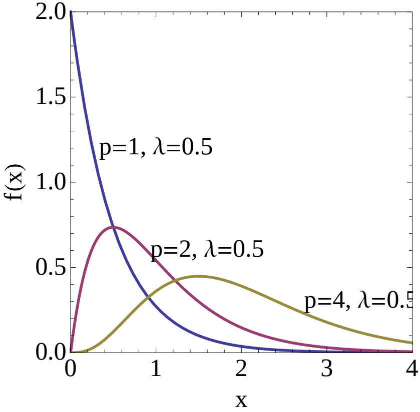
Probability density function:

$$f(x) = \frac{(\lambda x)^{p-1}}{\Gamma(p)} \lambda \, e^{-\lambda x}, \quad x \geq 0,$$

where

$$\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} \, dt.$$

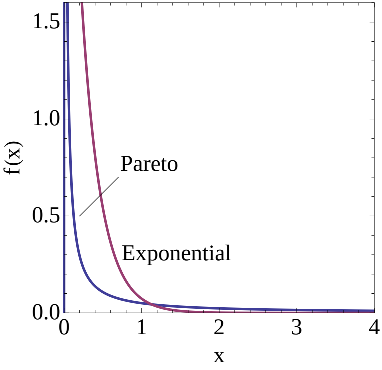
Mean:	p/λ
Variance:	p/λ^2



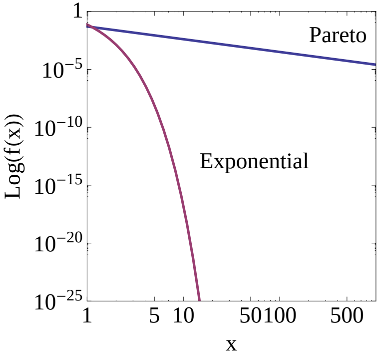
Heavy-tailed distributions

- Heavy-tailed distributions are distributions with a right tail that decays slower than exponentially
- Evidence found, e.g., in sizes of
  - Files stored on Web servers
  - Data transferred through the Internet
  - Files stored in general-purpose Unix file systems
  - I/O traces of file system, disk, and tape activity

Visual comparison



(a) Linear scale



(b) Log scale

Definitions

- Distribution has a **heavy tail** if for all $\gamma > 0$

$$\lim_{x \to \infty} e^{\gamma x} G(x) \to \infty,$$

where $G(x) = 1 - F(x)$, i.e., the ccdf.

- Distribution has a long tail if

$$G(x + t) \sim G(x), \quad \text{as } x \to \infty.$$

- Distribution has a **power tail** if for some α and $\beta > 0$

$$G(x) \sim \alpha x^{-\beta}, \quad \text{as } x \to \infty.$$

- In a nutshell:
  - Large values likely
  - High variability

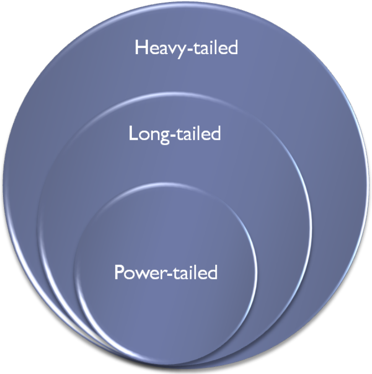
Relations

$$\text{Power-tailed} \subset \text{Long-tailed} \subset \text{Heavy-tailed.}$$

Distribution with a short tail has

$$\lim_{x \to \infty} e^{\gamma x} G(x) \to 0,$$

for some $\gamma > 0$.



Effects of “heavy tails”

- Expectation paradox:
  - The longer we have waited for an event the longer we have to wait
- Aggregate size of small variables is negligible compared to the largest one

$$\lim_{x \to \infty} \frac{P\{X_1 + X_2 + \ldots + X_n > x\}}{P\{\max\{X_1, X_2, \ldots, X_n\} > x\}} = 1, \quad \forall\, n \geq 2.$$

  - Typical flow is small, but typical transferred byte belongs to a large flow

Examples of utilizing heavy tails

- Load balancing in distributed systems
  - Only a few flows are redirected with a significant effect on load distribution
- Scheduling in web servers
  - Shortest-remaining-processing-time scheduling lets the small tasks interrupt larger ones and with heavy tails the benefit becomes large
- Routing and switching in the Internet
  - Shortcuts established only for large flows (cf. data centers)

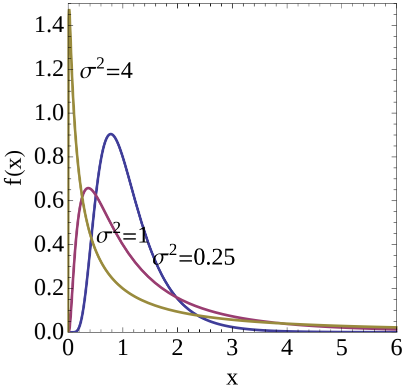
Log-normal distribution, LogNormal(μ, σ^2)

Def.: *Random variable X follows the LogNormal(μ, σ^2) distribution if* $\log(X)$ *is distributed as N(μ, σ^2).*

Probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma \cdot x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}.$$

Mean:	$e^{\mu+\sigma^2/2}$
Variance:	$\left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2}.$



Pareto distribution Pareto(α, k)

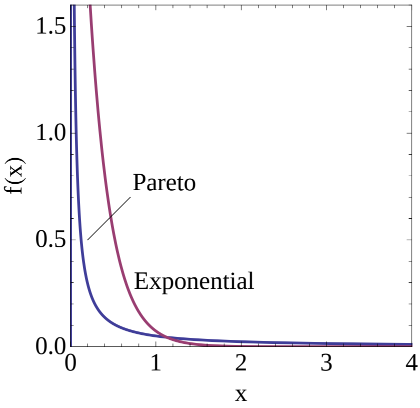
Probability density function:

$f(x) = \alpha k^{\alpha} x^{-\alpha-1}$.

- α is the **shape parameter**
- k is the **scale parameter**

Pareto has a **power tail**,

$$G(x) = \left(\frac{k}{x}\right)^{\alpha}, \quad x \geq k.$$



Mean:	$\frac{\alpha k}{\alpha - 1}$
Variance:	$\left(\frac{k}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2}$

Note: mean is infinite for $\alpha \leq 1$, and variance for $\alpha \leq 2$.

Weibull distribution, Weibull(α, β)

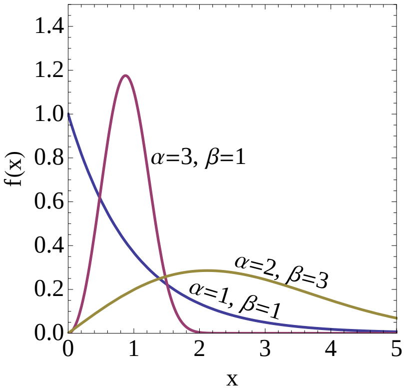
In Weibull(α, β) distribution α is the shape parameter and β the scale parameter.

- $\alpha < 1$ "failure rate decreases in time"
- $\alpha > 1$ "failure rate increases in time"

The pdf has the form

$f(x) = \alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^{\alpha}}$.

Mean:	$\beta\,\Gamma\,(1 + 1/\alpha)$
Variance:	$\beta^2\,\Gamma\,(1 + 2/\alpha)$ $-\beta^2\,\Gamma\,(1 + 1/\alpha)^2$



Choosing the distribution

- Distribution can be chosen according to “best fit”
- Distribution function can be adopted from a similar situation
  - Often assumed that the distribution function remains valid in other “similar conditions”, only parameters vary
- Distribution functions can emerge from generative processes
  - E.g., CLT and Gaussian distribution
  - Allows taking into account external knowledge on the variable
- As a result probabilistic modeling of network measurements is seldom purely objective process ...

Contents

Parameter estimation

- Given a distribution (or distribution family) and data, the next step is to fit the parameters of the distribution to match the data
- **Estimator** is a function of (sample) data that attempts to estimate an unknown (population) parameter
- We try to **optimize** the parameters of a density with respect to some **measure of fit**

Estimation

- We briefly outline two practical methods of parameter estimation
  1. Method of moments
  2. Maximum likelihood (ML)
- Note that a lot of literature exists on parameter estimation
  - Quality of the estimators ignored here ...
  - E.g., sample mean and sample variance of data are "best" estimates for Normal distribution

Method of moments

- k:th moment of a random variable: $\mathrm{E}(X^k)$
- k:th sample moment of data: $m_k = \frac{1}{n}\sum_i x_i^k$
- **Method of moments**:
  1. Derive as many moments of the distribution as there are parameters
  2. Compute the corresponding sample moments from data
  3. Solve the parameters so that the moments and sample moments are equal
- Simple, but not always available
- The estimates are not necessarily “optimal”

Maximum likelihood

- Idea is to maximize the probability/likelihood that the selected distribution has produced the observations
- Likelihood function for independent observations

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

- Likelihood is a function of the parameter(s) of the distribution, for which the estimate is

$$\arg\max_{\theta} L(x_1, x_2, \ldots, x_n; \theta)$$

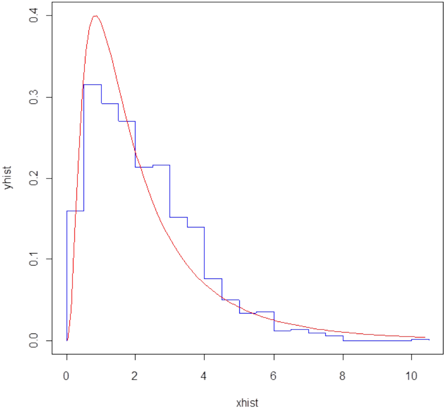
- Maximum likelihood estimates have many favorable properties, but require often complex non-linear optimization

Other estimates

- Non-linear least-squares optimization of the density
  - Non-linear optimization methods more generally available in mathematical software
- Statistical software offer many direct ways of fitting parameters for given densities
  - Often based on maximum likelihood
  - General optimization methods can be sensitive to selected starting values
  - Results are affected by outliers

Example

```
> require(MASS)
> fitdistr(logftp,"lognormal")
   meanlog         sdlog
 0.52062324     0.84232334
(0.02663660)   (0.01883492)
```



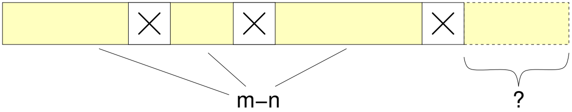

```
> h<-hist(logftp,n=20)
> xhist<-c(min(h$breaks),h$breaks)
> yhist<-c(0,h$density,0)
> xfit<-seq(min(logftp),max(logftp),length=100)
> yfit<-dlnorm(xfit,meanlog=0.52062,sdlog=0.84232)
> plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)),col=4)
> lines(xfit,yfit,col="red")
```

Example: German tank problem

Suppose that a measurement has given us

- A set of serial numbers of some device
  - How many devices has been sold?
- A list of user-id's of people using some service
  - How many users the given service has (say globally)?

These questions are examples of the *German tank problem*.



Example: German tank problem (2)

Assumptions:

- The existing serial numbers are $1, \ldots, N$ (all equally likely)
- Sampling without replacement, $\{X_i\}$, $i = 1, \ldots, n$
- Task is to estimate N based on the given n samples

Clearly, $n \leq m \leq N$, where $m = \max\{X_i\}$, but . . . ?

1. Bayesian estimate,

$$\hat{N} = \frac{(m-1)(n-1)}{n-2}.$$

2. Minimum-variance unbiased estimator,

$$\hat{N} = m + \frac{m}{n} - 1.$$

Off-line vs. On-line estimation

Off-line estimation

- Collect samples and store them
- Estimate the quantities of interest
- No (strict) memory or time constraints

On-line estimation

- Collect samples in real-time (streaming data)
  - Update estimates at the same time
- Both memory and time constraints (typically)

The latter is important, e.g., for real-time monitoring systems.

Example: On-line estimation

Mean:

Init:

1. $S \leftarrow 0$
2. $n \leftarrow 0$

Per sample:

1. $S \leftarrow S + x_i$
2. $n \leftarrow n + 1$
3. $\hat{m} \leftarrow S/n$

- Two state variables
- Fast constant computation time
- Unfortunately, e.g., median or mode are more difficult! (Why?)

Sometimes(!) sufficient:

$$\begin{cases} \text{mean} & \leftarrow & \text{mean} & + & \eta \times (x_i - \text{mean}) \\ \text{median} & \leftarrow & \text{median} & + & \eta \times \text{sgn}(x_i - \text{median}) \end{cases}$$

(cf. EWMA)

Example: On-line estimation (2)

- Moving average (MA):

$$y \leftarrow \frac{x_{i-k} + \ldots + x_i}{k+1}$$

- Exponentially Weighted Moving Average (EWMA):

$$y \leftarrow \alpha x_i + (1-\alpha)y$$

In general, on-line estimation of the streaming data is an interesting topic itself, and there are advanced algorithms for different scenarios. For example,

R. Jain and I. Chlamtac, “*The P-Square Algorithm for Dynamic Calculation of Percentiles and Histograms without Storing Observations*”, Communications of the ACM, October 1985.

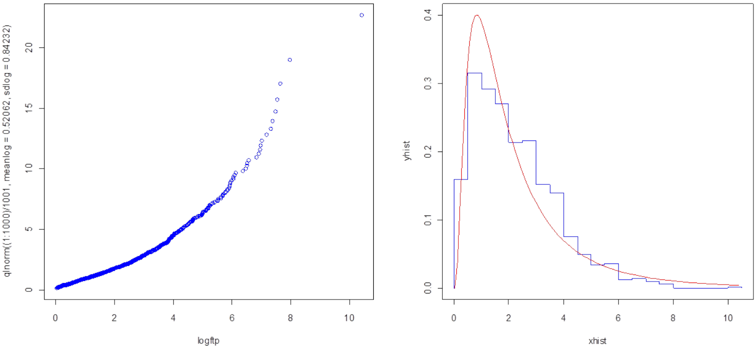
Contents

Checking for model fit

- Visual tools
  - Plot density and histogram in the same figure
  - Plot cdf and ecdf in the same figure
  - Compare data with the distribution in a QQ-plot
  - For highly variable data, log-log complementary distribution could be considered
  - Plot $\log(1 - F(x))$ against $\log x$ for CDF and ECDF
- Statistical tests
  - χ^2-test
  - Kolmogorov-Smirnov -test

Example

- Our log-normal fit
  - QQ-plot does not support the fit!

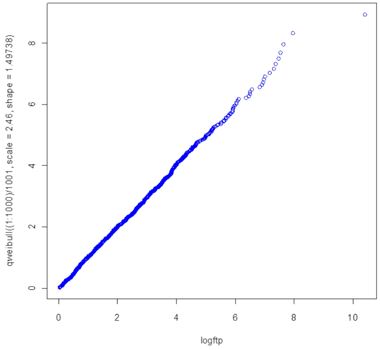


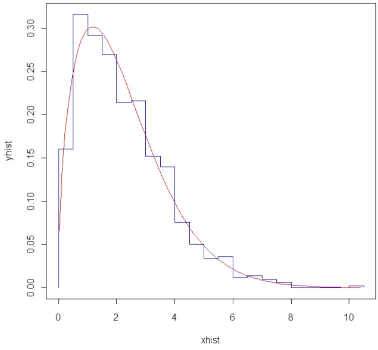

```
>qqplot(logftp,qlnorm((1:1000)/1001,meanlog=0.52062,sdlog=0.84232),col=4)x))
```

Example Continued

Let's try with Weibull distribution . . .

```
> fitdistr(logftp,"weibull")
     shape          scale
1.49738316    2.47066018
(0.03691000) (0.05496849)
```





Statistical testing for goodness of fit

- As the model selection is usually subjective, visual tools are generally sufficient for validating the model
- However, there are also statistical tests available for the **goodness of fit**
  - A *p*-value is computed for null hypothesis

    "*Sample comes from a population with a given distribution*"
  - *p*-value is roughly the probability that given the null hypothesis, we actually observe the data
  - If *p*-value is small, there is only a small probability that the data is from the distribution and null hypothesis is rejected
  - A large *p*-value does not automatically mean that the distribution is correct

Literature

- Mark Crovella, Balachander Krishnamurthy, **Internet Measurement: Infrastructure, Traffic & Applications**, John Wiley and Sons Ltd, 2006.
- David J. Marchette, **Computer Intrusion Detection and Network Monitoring: A Statistical Viewpoint**, Springer, 2001.
- Pierre Baldi, Paolo Frasconi, Padhraic Smyth, **Modeling the Internet and the Web, John Wiley and Sons Ltd**, 2003.
- NIST/SEMATECH, Engineering Statistics Handbook, `http://www.itl.nist.gov/div898/handbook/index.htm`