Input Data Analysis

DATA ANALYSIS IN TRANSPORT SYSTEMS – MASTER URBAN MOBILITY (MUM)

Course 2024-25 Term 1

Author: Lídia Montero

Introduction

1. In stochastic simulations, must decide on input probability distributions from which to generate *random variates*
   - Random *variates* = *observations* or *draws* or *realizations* of random variables from specified input distributions/processes.
   - Queueing Systems: interarrival and service times
2. Once these input probability distributions are specified, must have a way to generate random variates from them

Specifying Univariate Input Distributions

- *Univariate distribution*: a single (scalar) random input
  - May have many across simulation
- Typically have real-world observed data on a simulation input
  - Want to *fit* a probability distribution to the data
  - Then *use* this fitted distribution to generate random variates to drive the simulation
- Why not just drive simulation with actual observed data?
  - Seems more direct than fitting/generating
  - Useful for validation, but there are problems

Choosing Probability Distributions

- There are many probability distributions; some common ones:
  - Continuous: normal, exponential, logistic, continuous uniform, triangular, Weibull, gamma, Pareto, lognormal, loglogistic, Erlang k, Gumbel, etc.
  - Discrete: discrete uniform, binomial, Poisson, geometric, negative binomial
- Lists of probability distributions and their properties/relationships:
  - Many books and web sites (https://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm)
  - Beware: different sources may parameterize the same distribution differently, important to mesh with our current definition.
  - http://www.statsoft.com/Textbook/Distribution-Fitting

Choosing Probability Distributions – Qualitative

- How do you choose one distribution from among hundreds?
- Some qualitative criteria:
  - Continuous vs. discrete inputs – choose according to situation
  - Range (a.k.a. support) of the distribution – finite on both ends or infinite on right or left ends? Match with the real situation
  - Make a histogram of the observed data, match it to *shapes* of known probability density functions (continuous) or probability mass function (discrete) … Formulate hypothesis,
  - But still need to estimate distribution parameters ("fit"), test the goodness of fit. Descriptive statistical tools are a 'must'.

Fitting Distributions to Data – Preliminaries

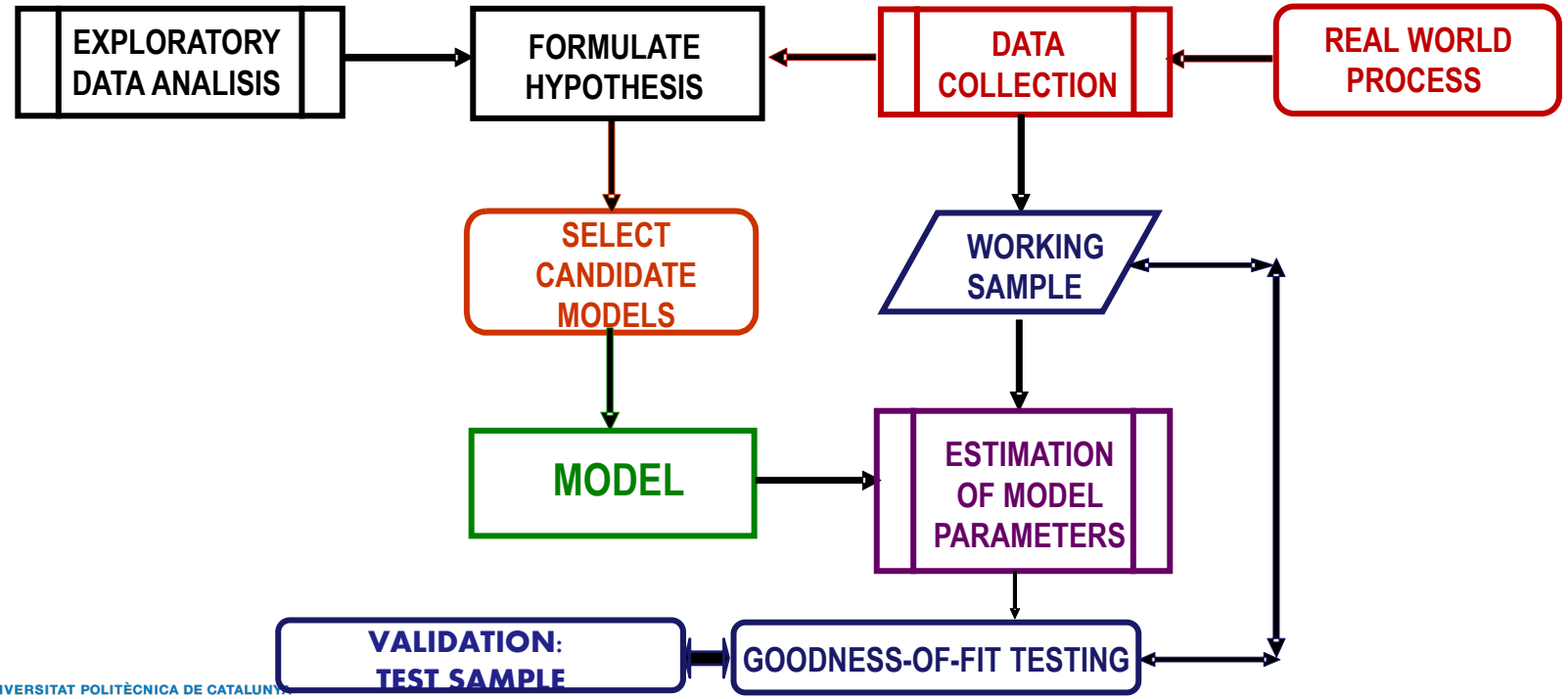
- Select a possible distribution form, estimate parameters, test goodness of fit … repeat as necessary … with your working sample.
- Final validation of selected model with test sample.
- Several distribution-fitting packages, or modules provided in MINITAB.
- Or statistical resources in R.
- Use Stat::Fit (www.geerms.com) … free student version, plus a "textbook" (Law and Kelton) and website

Fitting Distributions to Data – Preliminaries

- Or statistical resources in R (fitdistrplus, vcd, moments packages).
- Law and Kelton., "Simulation Modeling and Analysis", 1991, second Edition, published by Mc Graw Hill Inc.
- Ricci 2005 Fitting Distributions with R . http://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf
- Karian Z. A. et al., "Handbook of Fitting statistical Distributions with R", 2011 Chapman and Hall (anyway I do not like the approach)

Fitting Distributions to i.i.d univariate data

- Formulation of hypothesis once descriptive data analysis is performed: Histograms, Box Plots, etc. List of candidate distributions. Empirical moment calculation. Split data into working-test samples (75-25 or 80-20).
- Estimation of parameters with working sample: location, scale and shape parameter for each candidate distribution.
- Goodness of fit for each possible distribution (working/validation samples): Quartile and Probability Plots, X^2 Test, Kolmogorov-Smirnoff Test,… Theoretical vs Empirical moments



Fitting Distributions: Practical rules

- Coefficient of variation for random variate X: $C_X = \frac{\sigma_X}{\mu_X}$
- Intuitively is a mesure of how far the selected distribution for the sample data relies from the exponential distribution (with C_X=1)

A. Candidate distributions for nonnegative sample data shows C_X<1:

Gamma family distributions (shift parameter = 0):

$$f_X(x) = \begin{cases} \frac{e^{-x/\beta}}{\Gamma(\alpha)} \beta^{-\alpha} x^{\alpha-1} & \text{para } x \geq 0 \\ 0 & \text{para } x = 0 \end{cases} \qquad \Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-\alpha} dy$$

$$\alpha \in R_+, \beta \in R_+ \quad \text{If } \alpha = k \in Z_+ \Rightarrow \Gamma(k+1) = k!$$

α shape parameter, β scale parameter. $F_X(x)$ has an analytical expression when α is integer:

$$F_X(x) = \begin{cases} 1 - e^{-x/\beta} \sum_{j=0}^{\alpha-1} \frac{(x/\beta)^j}{j!} & x \geq 0 \\ 0 \quad \text{otherwise} \end{cases}$$

- Particular case: when $\alpha = 1$, $\Gamma(1) \equiv$ exponential distribution scale β.

$$f_X(x) \rightarrow \begin{cases} \text{Mean } \bar{x} = \alpha\beta \\ \text{Variance } \sigma^2 = \alpha\beta^2 \end{cases} \Rightarrow C_x = \frac{\sqrt{\alpha}}{\alpha} < 1 \text{ if } \alpha > 1$$

- For $\alpha = k$ integer and $\beta = 1/(k\mu)$, $\Gamma(k, 1/(k\mu))$ is Erlang distribution with shape k and mean $1/\mu$, sum k I.I.D. exponentials with mean $1/(k\mu)$,

$$\left.\begin{aligned} f_X(x) &= \frac{k\mu(k\mu)^{k-1}x^{k-1}}{(k-1)!}\left(e^{-k\mu}\right)^x \\ F_X(x) &= 1 - \left(e^{-k\mu}\right)^x \sum_{j=0}^{k-1} \frac{(k\mu)^j x^j}{j!} \end{aligned}\right\} x \geq 0;\ \text{Mean}\frac{1}{\mu};\ \text{Variance}\frac{1}{k\mu^2}$$

Fitting Distributions: Practical rules

B. Approximation of distributions with coefficient of variation is significantly greater than 1: $C_X>1$.

Example: Hiperexponential distribution (Combination of exponential distributions)

$$f_X(x) = \sum_{i=1}^{k} \pi_i \mu_i e^{-\mu_i x} \qquad \left(\pi_i \geq 0, \forall i, \sum_{i=1}^{k} \pi_i = 1 \right)$$

$$F_X(x) = \sum_{i=1}^{k} \pi_i \left(1 - e^{-\mu_i x}\right) = 1 - \sum_{i=1}^{k} \pi_i e^{-\mu_i x} \qquad x \geq 0$$

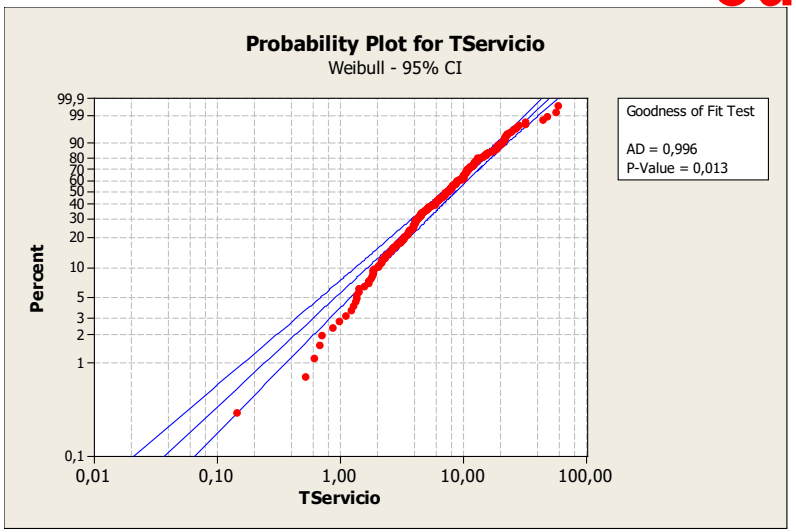
Mean **Variance :**

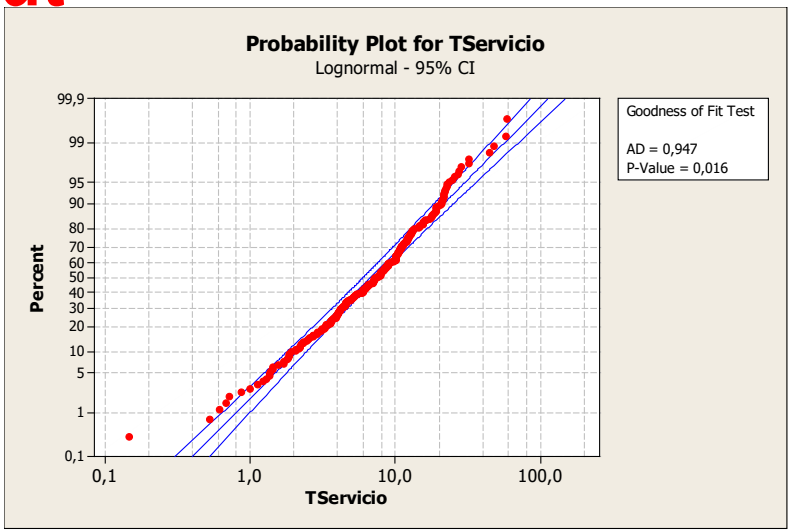
$$\bar{x} = \sum_{i=1}^{k} \frac{\pi_i}{\mu_i} = \frac{1}{\mu} \qquad \sigma_X^2 = \frac{1}{\mu^2} + \sum_{i=1}^{k} \sum_{j=1}^{k} \pi_i \pi_j \left(\frac{1}{\mu_i} - \frac{1}{\mu_j} \right)^2$$

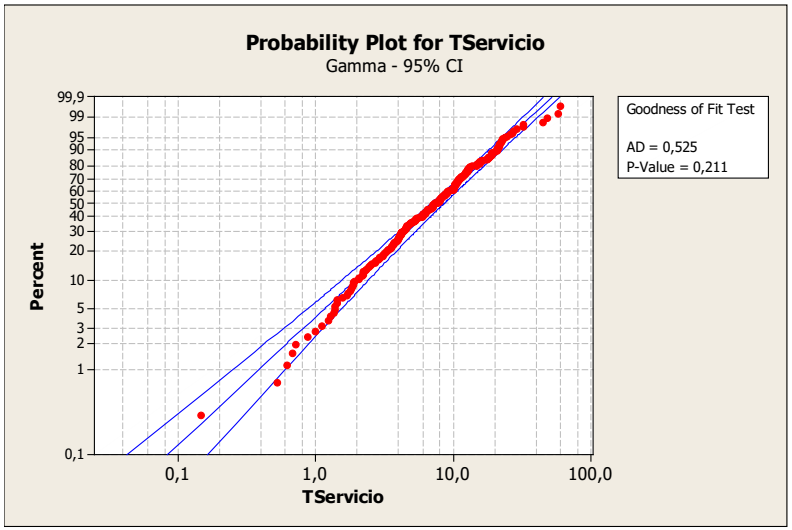
Coefficient of Variation

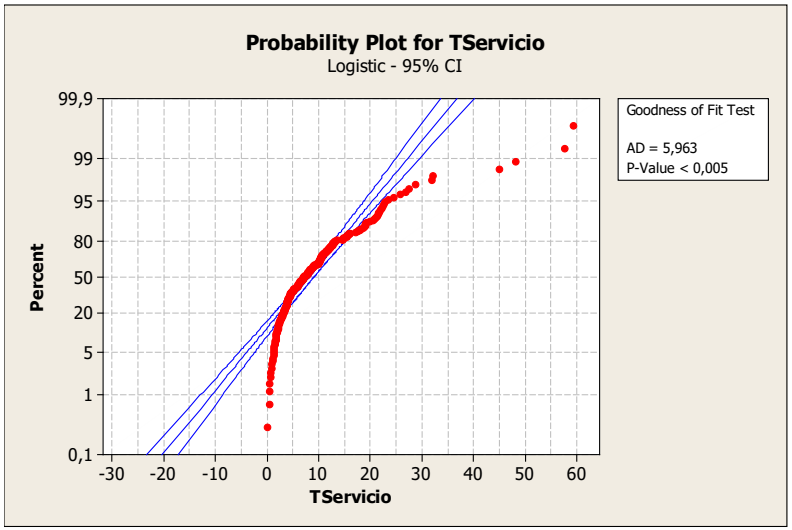
$$C_X^2 = 1 + \mu^2 \sum_{i=1}^{k} \sum_{j=1}^{k} \pi_i \pi_j \left(\frac{1}{\mu_i} - \frac{1}{\mu_j} \right)^2 \geq 1 \quad \left(= 1 \Leftrightarrow \mu_i = \mu_j \forall i, j \right)$$

Service times in a queueing system : Minitab output



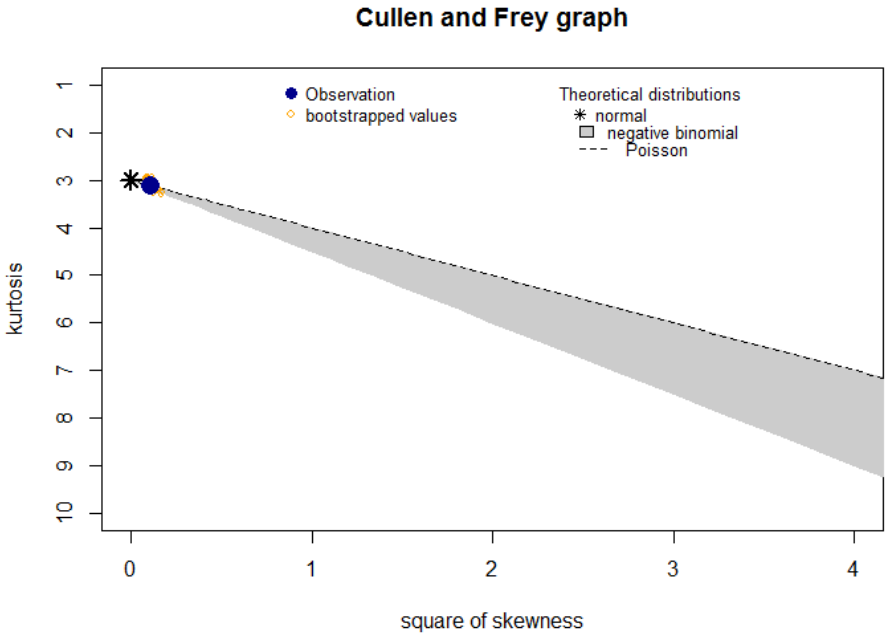




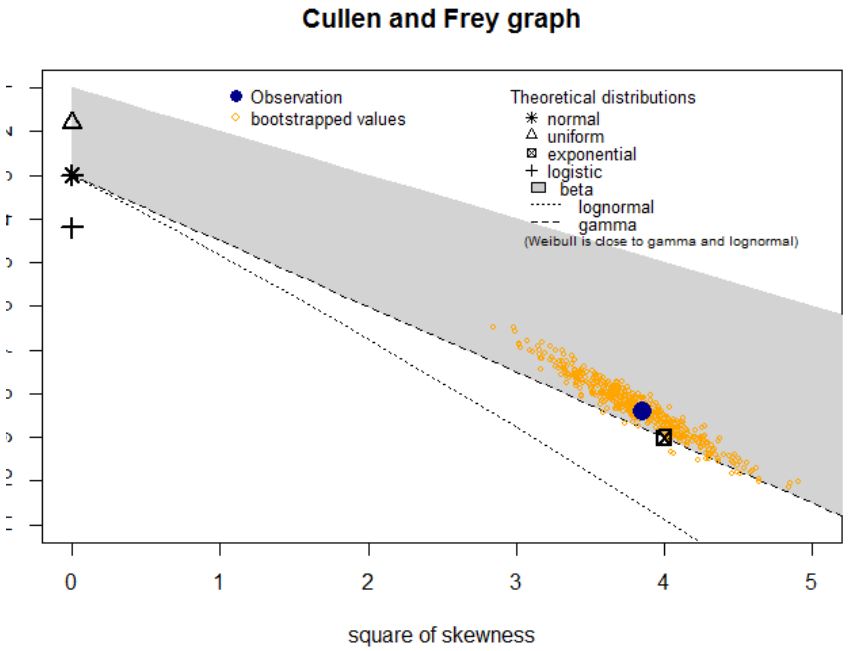


Service times in a queueing system : fitdistrplus package in R output

```
descdist(dsample,discrete=T,boot=50)
summary statistics ------
min: 2 max: 24
median: 10
mean: 9.9828
estimated sd: 3.194643
estimated skewness: 0.3344852
estimated kurtosis: 3.109237
```




```
descdist(sample,boot=500)
summary statistics ------
min: 10.00011 max: 26.70881
median: 11.40808
mean: 12.04515
estimated sd: 2.036611
estimated skewness: 1.963552
estimated kurtosis: 8.420342
```



Descriptive Statistics methods in R (I)

- Variate type: discrete variate (with few values) or continous variate (also, discrete one with many values).
- Numerical description: moments; i.e. mean, variance (standard deviation), kurtosis, etc. Classified as:
  - Centrality Trend: *Mean (not average), Median, Mode.*
  - Diversion Trend: *Variance (Standard Deviation), Quantiles, IQR (Q3-Q1), etc.*
- Rough description in R: summary(dataframe)
- Use library moments in R and compute first fourth sample moments, including skewness and kurtosis.
- Check correlation in sample (acf(), pacf())
- Remove shift for exponential, gamma, weibull type data

Descriptive Statistics methods in R (II)

- Discrete variate (with few values or cathegorical) : pie (table(*variable*)), barplot(table(*variable*)) .
- Continuous variates: Classified as:
  - Histogram: hist(variable, freq=T). *Draw theoretical distributions adding a curve()*
  - Boxplot: use Boxplot(variable) in library(car) instead that boxplot(variable).
- Default description in R: table(var), summary(var).
- Possible tools increase when 2 simultaneous variates are considered (not the case in this unit).

Auxiliary Graphic Tools: Probability Plot - PPlot

A probability plot can be thought as a graphical comparison of an estimate of the true distribution function of the available data X_1, X_2, …., X_n with the distribution function of the fitted distribution.

A *probability-probability* plot (P-P) is a graph of the model probability $\hat{F}(X_i)$ versus the sample probability

$$\tilde{F}_n(X_i) = q_i = \frac{i - 0.5}{n}, i = 1,2,....,n$$

If $\hat{F}(x)$ and $\tilde{F}_n(x)$ are close togeher then the P-P plot will also be approximately linear with an intercept 0 and a slope 1.

The lineal correlation coefficient of the fit of the P-P points is a measurement of the goodness-of-fit of the proposed distribution.

Auxiliary Graphic Tools: QQPlot

A quantile-quantile plot can be thought as a graphical comparison of the quantiles of the available data X_1, X_2, …., X_n with the quantiles given by the standardized fitted distribution.

A *quantile-quantile* plot (Q-Q) is a graph of the standard model quantiles $\hat{F}^{-1}(q_i)$ **where** $q_i = \frac{i-0.5}{n}, i = 1,2,....,n$ versus $x_{(i)}, i = 1,2,....,n$ **s.t.** $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ the ordered sample data.

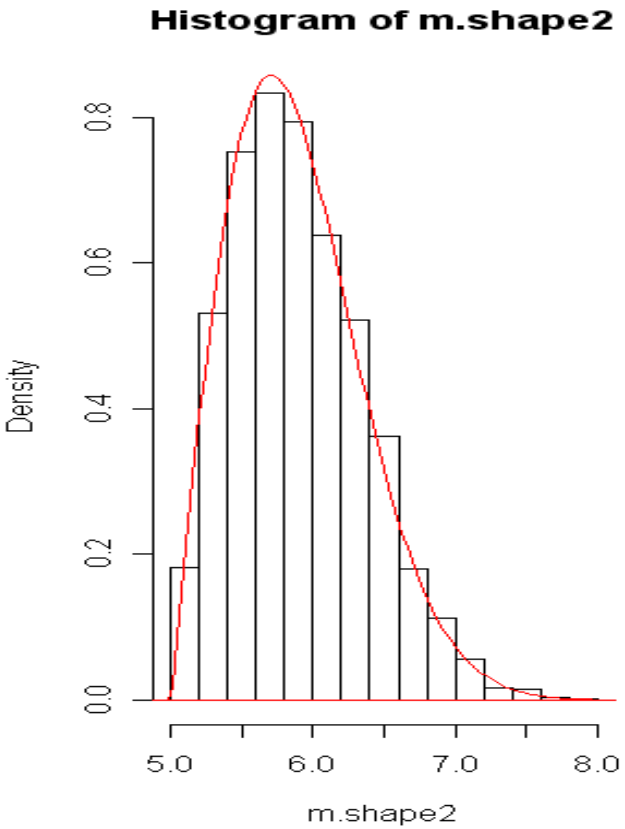
If $\hat{F}^{-1}(q_i)$ and $x_{(n)}$ are close togeher then the Q-Q plot will also be approximately linear with an **intercept meaning location** and a slope meaning scale. Also, it is a measurement of the goodness-of-fit of the proposed distribution.

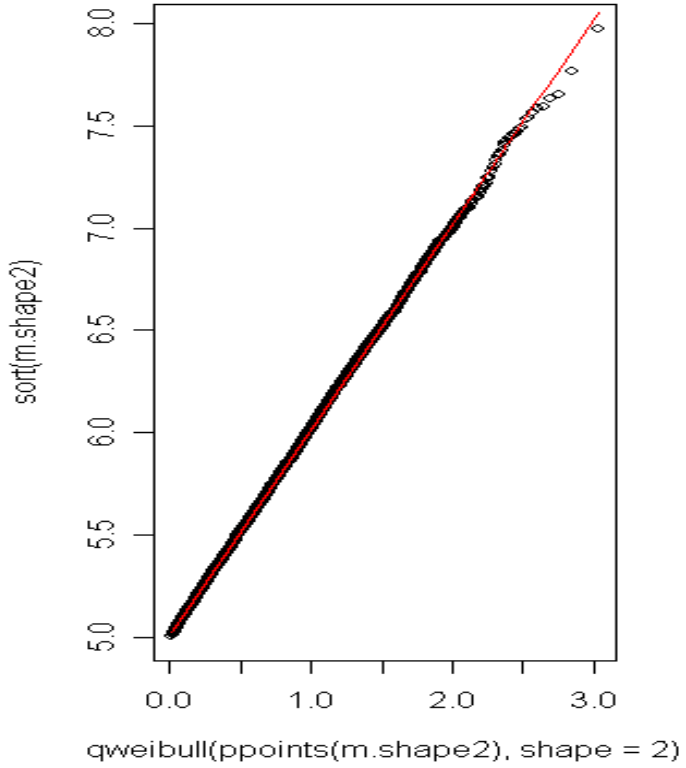
Descriptive Statistics methods in R: example

- **How to avoid the distorsion effect of non standard distributed sample data?** H0: Weibull α**=2**
- Plot theoretical standard quantiles for the theoretical distribution vs sample data quantiles. That's a QQPlot.
  - Compute the coefficient of determination R2 of the linear regression fit for QQPlot: this is a goodness of fit indicator as closer to 1 than better is the fit to H0.
  - Intercept and slope of regression are estimates of shift and scale parameters for H0.

Descriptive Statistics methods in R: example

- Sample data from Weibull location (shift) 5, scale 1 and shape 2.
- H0: Weibull distribution with shape α= 2.
- Plot QQPlot and estimate simple linear regression line.





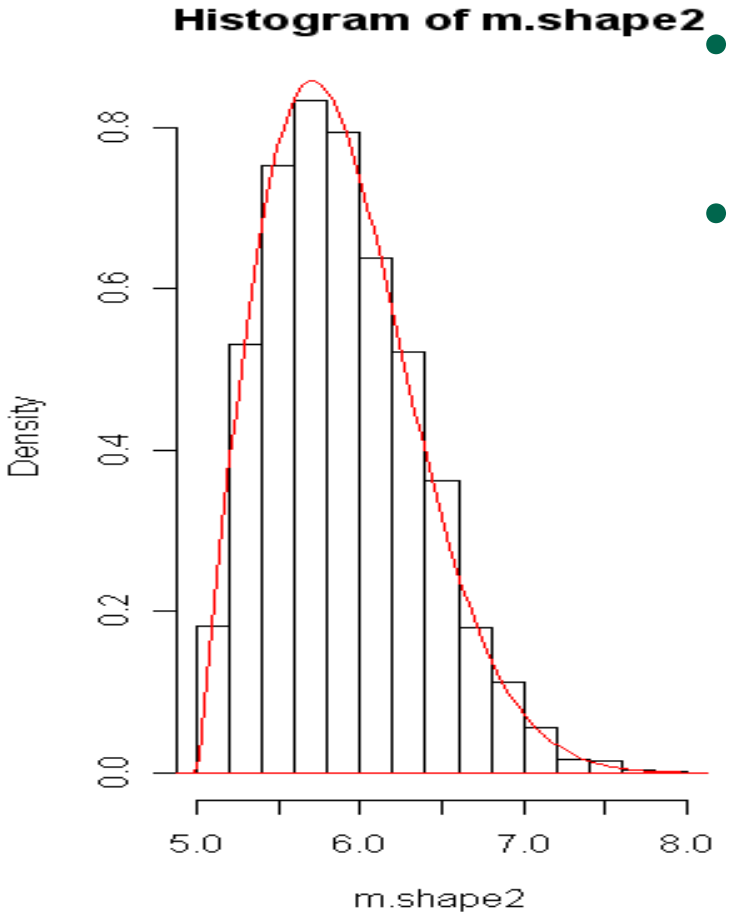
Descriptive Statistics methods in R: example

- H0: Weibull: $\alpha = 2$. Shape parameter known.

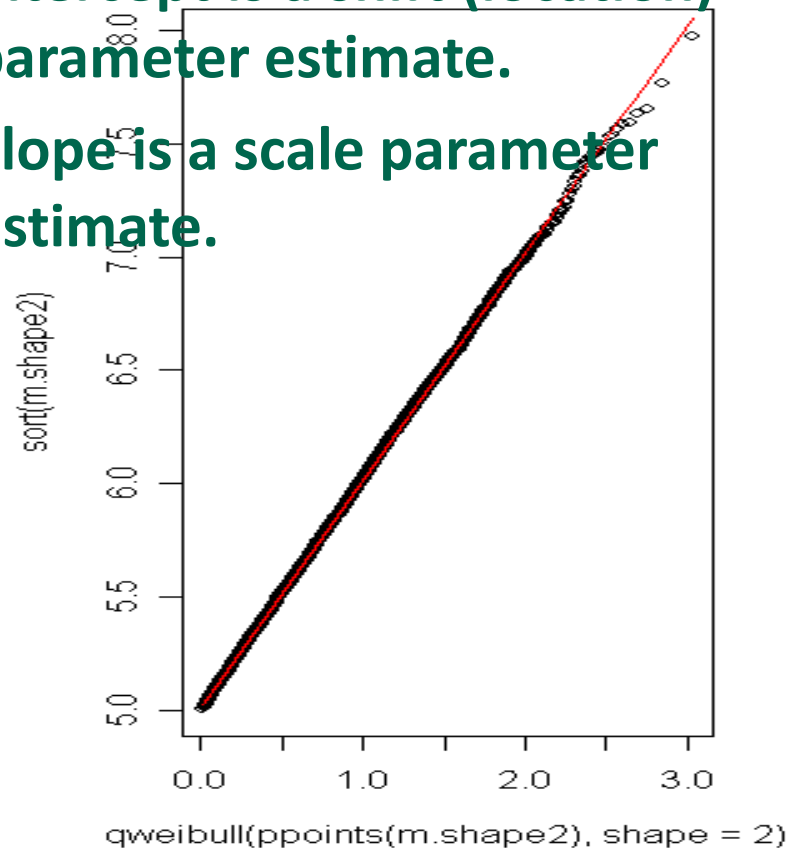
```
> # Sample m.shape2 5+rweibull(5000, shape=2, scale = 1)
> # Compute standard weibull shape 2 for probability points
in the sample
> par(mfrow=c(1,2))
> hist( m.shape2, freq=F )
> x<-seq(min(m.shape2),max(m.shape2),100)
> curve(dweibull(x-5,shape=2,scale=1),col=2,add=T)
> # QQPLOT
> xx <- qweibull( ppoints( m.shape2 ) , shape=2)
> yy<- sort( m.shape2)
> mm <- lm( yy ~ xx )
> cor(yy,xx) # Squared correlation coeff is R2
[1] 0.9998977
>plot( xx, yy, main="QQPlot") # plot std theo vs obs ptiles
>lines( xx, mm$fitted.values, col=2 )
> mm
Coefficients:
(Intercept)           xx
   5.002425     0.986965
```

Descriptive Statistics methods in R: example

- H0: Weibull $\alpha = 2$. Shape parameter known.
- Sample data **W(θ=5,α=2,λ=1).**



- **Intercept is a shift (location) parameter estimate.**
- **Slope is a scale parameter estimate.**



Fitting Distributions to i.i.d univariate data

- For each chosen distributions, Fit Distribution Parameters Moments method
  - Maximum Likelihood
    - fitdistr() in MASS
    - fitdist() in fitdistrplus
- Once distribution parameters are estimated then: Goodness of Fit with Working Sample
  - Numeric tools: moments comparison (skewness, kurtosis, mean, variance).
  - Graphics (based on overlapping theoretical distribution to empirical histogram).
  - Tests for goodness of fit

Fitting Distributions to i.i.d univariate data

- **Hipothesis**: Observations $X_1, X_2, \ldots, X_n$, are I.I.D.
- **Method**: Maximum likelihood estimation
  - Let $P_\theta(x)$ be a univariate probability function, depending on θ parameter. We infer an estimate $\hat{\theta}$ of θ parameter calculate from sample data $X_1, X_2, \ldots, X_n$.
  - Maximum Likelihood Function $L(\theta)$ is defined as:

$$L(\theta) = P_\theta(X_1)\, P_\theta(X_2) \ldots P_\theta(X_n)$$

  - The estimate $\hat{\theta}$ of θ parameter maximizes $L(\theta)$ being the one the "one that better explain sample data".

$$\hat{\theta} \Leftrightarrow \frac{dL(\theta)}{d\theta} = 0$$

Example 1: Exponential distribution

- Mean $\theta = \mu > 0, f_\mu(x) = \frac{1}{\mu} e^{-x/\mu}, x \geq 0$
- Likelihood function :

$$L(\mu) = \left(\frac{1}{\mu} e^{-x_1/\mu}\right)\left(\frac{1}{\mu} e^{-x_2/\mu}\right)\ldots\left(\frac{1}{\mu} e^{-x_n/\mu}\right) = \frac{1}{\mu^n} \exp\left(-\frac{1}{\mu}\sum_{i=1}^{n} x_i\right)$$

- Log-likelihood function:

$$l(\mu) = \ln[L(\mu)] = -n\ln\mu - \frac{1}{\mu}\sum_{i=1}^{n} x_i$$

- Determine the value of $\hat{\mu}$ maximizing $L(\mu)$ for $\mu \geq 0$:

$$\max_{\mu} L(\mu) \equiv \max_{\mu} l(\mu)$$

$$\frac{dl(\mu)}{d\mu} = -\frac{n}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{n} x_i \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- According to KKT conditions:

$$\frac{dl^2(\mu)}{d\mu^2} = \frac{n}{\mu^2} - \frac{2}{\mu^3}\sum_{i=1}^{n} x_i < 0 \text{ for } \mu = \hat{\mu}$$

Example : Weibull distribution W(θ,α,β)

- Pay attention to the functional form, that is meaning of model parameters.
- **Instead W(θ=0,α,λ=1/β), widely used in engineering.**
- Modeling of fatigue in mechanical or electronic components

$$f(t) = \lambda\alpha t^{\alpha-1}\mathrm{e}^{-\lambda t^{\alpha}} \qquad F(t) = 1 - e^{-\lambda t^{\alpha}}$$

$$E[T] = \frac{1}{\alpha}\left(\frac{1}{\lambda}\right)\Gamma\left(\frac{1}{\alpha}\right) \quad V[T] = \frac{1}{\alpha}\left(\frac{1}{\lambda}\right)^{2}\left(2\Gamma\left(\frac{2}{\alpha}\right) - \frac{1}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)^{2}\right)$$

- α shape parameter and λ is rate parameter, related to scale parameter β , $\lambda = 1/\beta$.
- Previously to the use of any statistical package, the distribution functional form must be checked.

ML parameter estimates for H0 Model

- H0: Weibull. Shift should be estimated externally.
- ML Estimation in R: fitdistr() in MASS or fitdist() in fitdistrplus package on your working sample.

```
> # Maximum likelihood estimation shape and scale (rate),
package MASS. Shift (location) estimated externally.Keep in
mind m.shape2 is 5+rweibull(5000, shape=2, scale = 1)
> fitdistr( m.shape2-5, "weibull")
      shape           scale
   2.008910258    0.990032096
  (0.022126097) (0.007340344)
Warning messages:
1: Se han producido NaNs in: dweibull(x, shape, scale, log)
2: Se han producido NaNs in: dweibull(x, shape, scale, log)
3: Se han producido NaNs in: dweibull(x, shape, scale, log)
4: Se han producido NaNs in: dweibull(x, shape, scale, log)
> fitdistr( m.shape2-5, "weibull",lower=0.001)
      shape           scale
   2.008909637    0.990030803
  (0.022126093) (0.007340332)
```

ML parameter estimates for H0 Model

- H0: Weibull. Shift should be estimated externally.
- ML Estimation in R: method fitdistr() in library MASS or fitdist() in fitdistrplus package on your working sample.

```
> ussampe <- m.shape2 - 5 + 0.001
> fitw<-fitdist(ussample,"weibull")
> plot(fitw)
> summary(fitw)
Fitting of the distribution ' weibull ' by maximum
likelihood Parameters :
       estimate Std. Error
shape 1.007415 0.01108972
scale 2.051546 0.03032513

Loglikelihood: -8577.138 AIC: 17158.28 BIC: 17171.31
Correlation matrix:
       shape         scale
shape 1.0000000 0.3131122
scale 0.3131122 1.0000000
```

Fitting Distributions to i.i.d univariate data

- For each chosen distributions, Fit Distribution Parameters Moments method
  - Maximum Likelihood
    - fitdistr() in MASS
    - fitdist() in fitdistrplus
- Once distribution parameters are estimated then: Goodness of Fit with Working Sample
  - Numeric tools: moments comparison (skewness, kurtosis, mean, variance).
  - Graphics (based on overlapping theoretical distribution to empirical histogram).
  - Tests for goodness of fit

Goodness of fit – Check theoretical vs simple moments

- Assume a sample from **W(θ=5,α=2,λ=1),**

$$F(t) = 1 - \mathrm{e}^{-\lambda(t-\theta)^{\alpha}},\ t \geq \theta$$

- Sample moments for **W(θ=0,α=2,λ=1) are not the same: pay attention.**
- Theoretical moments for Weibull; i.e., X~**W(θ=0,α=2,λ=1).**

$$\textbf{Mean } \beta\Gamma\left(\frac{\alpha+1}{\alpha}\right) \quad \textbf{Variance } \beta^2\left\{\Gamma\left(\frac{\alpha+2}{\alpha}\right)-\left[\Gamma\left(\frac{\alpha+1}{\alpha}\right)\right]^2\right\}$$

$$\Gamma(c+1)=c\Gamma(c),\quad \Gamma(0)=\Gamma(1)=1,\quad \Gamma\left(\tfrac{1}{2}\right)=\sqrt{\Pi}$$

$$E[X]=\Gamma\left((\alpha+1)/\alpha\right)\beta=\Gamma(3/2)=\sqrt{\pi}/2 \quad E[Y]=5+\sqrt{\pi}/2$$

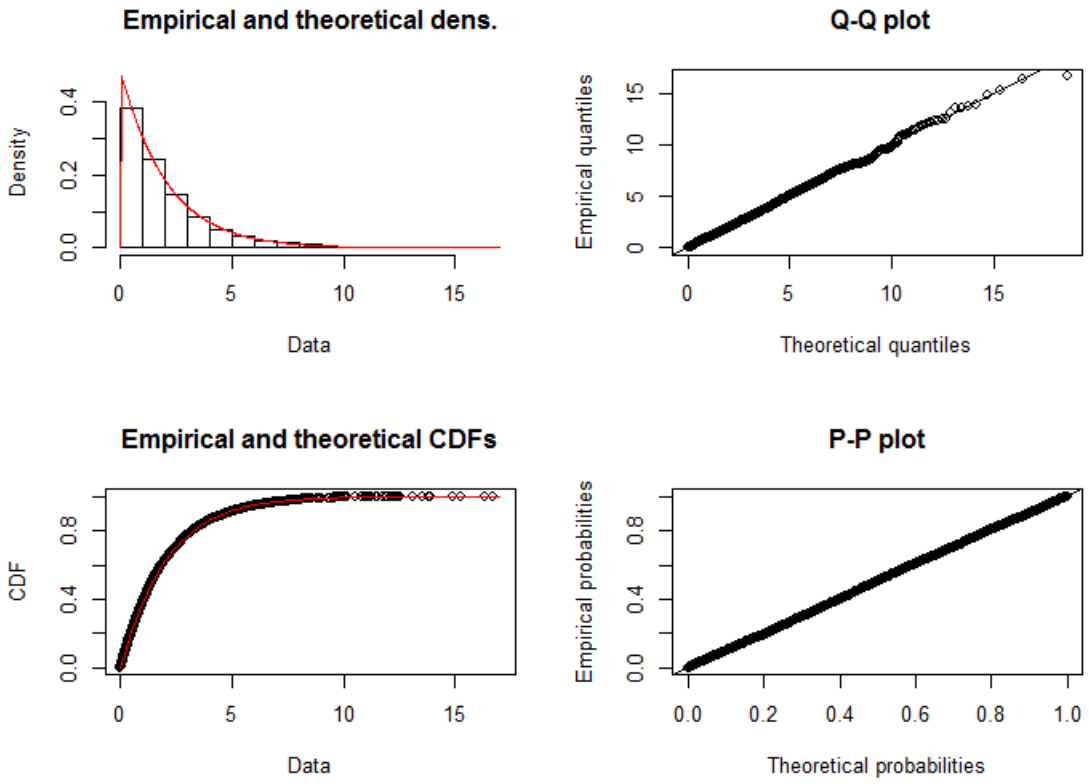
- **Compute sample statistics: check moment consistency.**
- **Pay attention if non-null shift/location applies**

Fitting Distributions to i.i.d univariate data

- For each chosen distributions, Fit Distribution Parameters Moments method
  - Maximum Likelihood
    - fitdistr() in MASS
    - fitdist() in fitdistrplus
- Once distribution parameters are estimated then: Goodness of Fit with Working Sample
    - Numeric tools: moments comparison (skewness, kurtosis, mean, variance).
    - Graphics (based on overlapping theoretical distribution to empirical histogram).
    - Tests for goodness of fit

ML parameter estimates for H0 Model

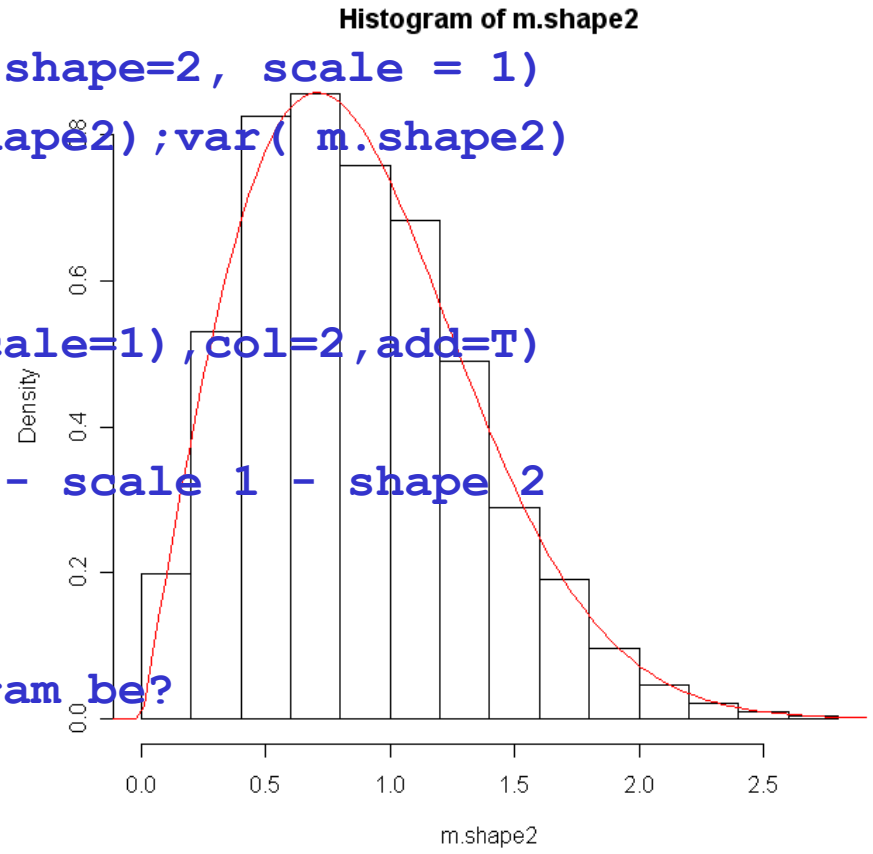
- H0: Weibull. Shift should be estimated externally.
- ML Estimation in R: fitdist() in fitdistrplus package on your working sample, graphical output.



Goodness of fit – Graphical Assessment: Overlap theoretical distribution

- Weibull: $\alpha = 2$ and $\lambda = 1$

```
# Exemple Weibull: shift 0 - scale 1 - shape 2
par(mfrow=c(1,1))
m.shape2<-  rweibull(5000, shape=2, scale = 1)
mean( m.shape2);stdev( m.shape2);var( m.shape2)
hist( m.shape2, freq=F )
x<-seq(0,max(m.shape2),100)
curve(dweibull(x,shape=2,scale=1),col=2,add=T)

# Exemple Weibull: shift 5 - scale 1 - shape 2
m.shape2<- 5 + m.shape2

# How would the new histogram be?
```



Fitting Distributions to i.i.d univariate data

- For each chosen distributions, Fit Distribution Parameters Moments method
  - Maximum Likelihood
    - fitdistr() in MASS
    - fitdist() in fitdistrplus
- Once distribution parameters are estimated then: Goodness of Fit with Working Sample
    - Numeric tools: moments comparison (skewness, kurtosis, mean, variance).
    - Graphics (based on overlapping theoretical distribution to empirical histogram).
    - Tests for goodness of fit

Fitting Distributions to i.i.d univariate data

- For each chosen distributions, Fit Distribution Parameters → Goodness of Fit with Working Sample
  - Several Goodness of Fit tests: Chi Squared (most popular test, df must be provided), K-S (Kolmogorov-Smirnoff), A-D (Anderson-Darling).
  - Minitab: Table with all distributions, all goodness-of-fit tests' test statistics
    - Smaller test statistics $\Rightarrow$ better fit
  - Detail report on each chosen distribution, and each goodness-of-fit test applied to it
  - Validation of selected distribution with test sample

Fitting Distributions: Inferential Issues

- **H0: "Sample Data distributed as ..."**
- Each test provides a p value that means "probability of H0 to be true"

How to interpret p value:

- If p value is less than the selected confidence level (usually 5 or 10%) then "REJECT H0", meaning that the proposed distribution does not fit available data.
- If p value is greater than the selected confidence level (usually 5 or 10%) then there is no evidence to reject H0", thus "DO not REJECT H0" (→"ACCEPT H0"), meaning that there is no evidence against the proposed distribution fitting available data.

Goodness of fit - Inferential Issues: Chisq Test, KS Test, Anderson-Darling Test, etc.

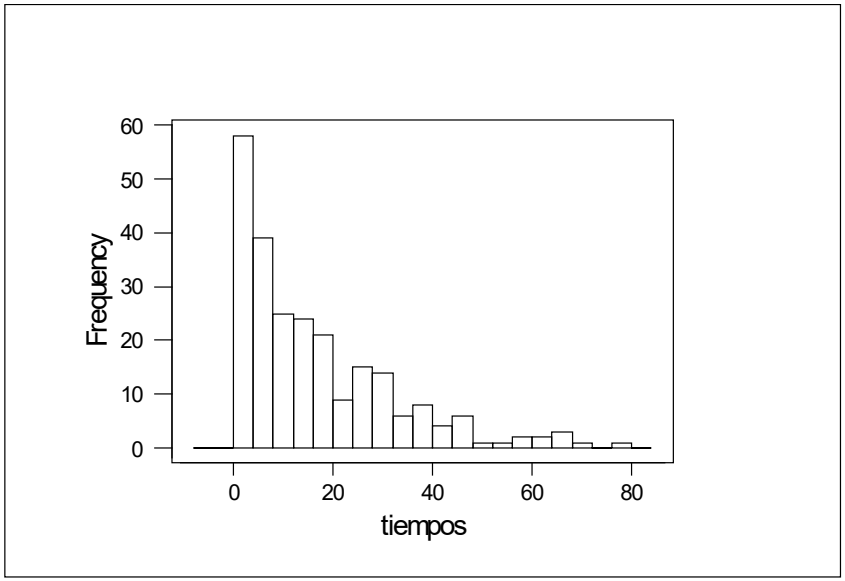
- H0: Weibull θ=5, $\alpha = 2$ and $\lambda = 1$.
- General goodness of fit Kolmogorov-Smirnoff (no shift parameter). Not available: Chi Squared Test.

```
> chisq.test( m.shape2-5, "pweibull", shape= 2, scale=1 )
# Non existent: Not authomatically provided by R
# Count samples in interval defined as percentiles of
theoretical distribution:  Weibull(shape= 2, scale=1,
shift=5): how to do this? Practical lab session.

> ks.test( m.shape2-5, "pweibull", shape= 2, scale=1  )
   One-sample Kolmogorov-Smirnov test data: m.shape2
                D = 0.015, p-value = 0.2131
           alternative hypothesis: two.sided

> # library(nortest): several normality tests
> shapiro.test( m.shape2)
        Shapiro-Wilk normality test data:  m.shape2
               W = 0.9717, p-value < 2.2e-16
```

Interarrival times to a queueing system (I)

0.0056	0.0967	0.1807	0.2262	0.2389	0.2628	0.4112	0.5298	0.5525
0.5861	0.6353	0.6510	0.6731	0.7571	0.7587	1.0471	1.0651	1.1033
1.1745	1.1891	1.2248	1.2468	1.3303	1.3524	1.3836	1.4123	1.4896
1.5370	1.5647	1.6214	1.6926	1.7026	1.8342	1.9083	1.9799	2.0623
2.1642	2.1831	2.2502	2.3509	2.3802	2.4070	2.4423	2.5339	2.7035
2.7615	2.7816	2.8736	2.9131	2.9503	2.9543	3.3028	3.4130	3.4334
3.7070	3.7101	3.7975	3.9954	4.0307	4.2456	4.3496	4.3519	4.4318
4.4611	4.5225	4.6649	4.8233	4.8930	4.9570	5.0024	5.0737	5.1245
5.4059	5.4331	5.4764	5.5202	5.5676	5.7514	5.9360	5.9792	6.1983
6.3889	6.4247	6.4959	6.6339	6.6549	6.7329	6.9757	7.1417	7.1698
7.1863	7.3967	7.6357	7.6863	7.7195	7.8727	7.9692	8.1695	8.3292
8.5473	8.5664	8.5788	8.7657	8.7732	8.7898	8.9490	9.0254	9.2393
9.5786	9.6571	9.7378	9.9193	10.0864	10.4750	10.5871	10.8425	10.8702
11.0024	11.1306	11.1442	11.5322	11.5343	12.1626	12.2130	12.5622	13.0473
13.0725	13.2578	13.6691	13.8986	14.0160	14.0924	14.1088	14.2914	14.3412
14.3476	14.3792	14.3882	14.5239	14.5886	15.1880	15.2705	15.4756	15.5005
15.6594	15.9621	16.0468	16.0755	16.3010	16.3375	16.5762	16.7091	16.7427
16.8580	16.9638	17.2893	17.3069	17.3494	17.6276	18.0241	18.1331	18.5128
19.3180	19.3221	19.7837	19.8183	19.9215	20.4094	21.2502	21.3287	21.3893
21.9020	22.5750	22.6202	23.0887	23.4108	24.2013	24.2341	24.3140	24.8287
25.5066	26.0540	26.1539	26.3157	26.4727	26.8948	26.9533	26.9969	27.0315
27.7501	27.9995	28.8724	29.1316	29.2133	29.5204	29.8822	30.1434	30.2184
30.2375	30.3884	30.4502	31.3845	31.6547	31.7689	31.9442	32.2406	33.2581
33.4450	34.4730	35.3383	35.5309	36.2030	37.0102	37.3794	37.3898	38.2246
38.5539	38.6309	38.6955	40.2452	40.7250	41.0193	41.6327	44.2501	45.0125
45.0766	46.2803	46.6233	47.8474	48.5374	53.3607	56.1566	58.3313	60.6484
61.9682	66.3411	66.7154	67.0041	70.4577	79.3826			

Interarrival times to a queueing system (II): sample moments



Histogram 240 observed inter arrival times

Mean and standard deviation of the sample :

$$\hat{\mu} = 16.398179$$

$$\hat{\sigma} = 15.855114$$

Coefficient of Variation $cv = 0.966882 \approx 1.0 \rightarrow$

Hipothesis : exponential inter arrival times : $f(x) = 0.060982e^{-0.60982x}$

Interarrival times to a queueing system (III): chisquared test of the theoretical distribution

- Group observations into intervals: plot them in a histogram
- N_j is the number of observations in the sample belonging to j-th class, whose lower and upper values are: $[a_{j-1}, a_j)$
- According to the theoretical distribution considered f(x) (depending on the hypothesis), then the expected number of observation in j-th class Np_j, being N the total sample size, and p_j the expected probability for j-th interval, that according to f(x) is:

$$p_j = \int_{a_{j-1}}^{a_j} f(x)dx$$

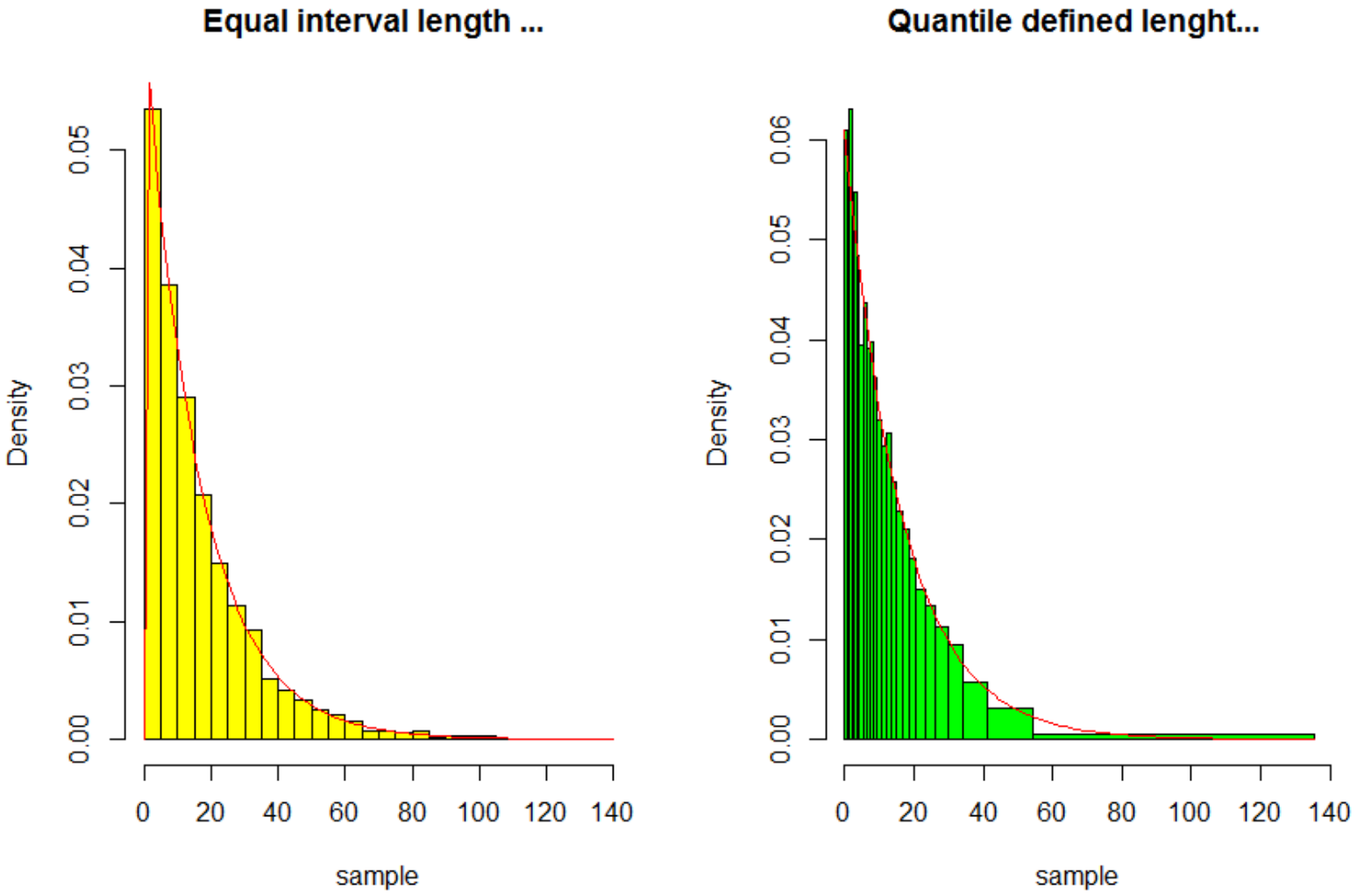
- Compute the χ^2 statistic for distribution matching, distributed as χ^2_{k-1}:

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - Np_j)^2}{Np_j}$$

If the fit is good enough then χ^2 value should be less than the theoretical distribution of the statistic $\chi^2_{k-1,1-\alpha}$, for a confidence level of α.

- In the example taking equal probability for each interval p_j=0.04, fiven k = 25 classes, then the theoretical number of observations should be Np_j = 240 x 0.04 = 9.6. Lower and upper interval values are given by $a_j = -16.398179 \ln(1 - \frac{j}{25})$, leading to the following table .

Ensure equal number of expected obs. according to hypothetical distribution



Ensure equal number of expected obs. according to hypothetical distribution

```
par(mfrow=c(1,2))
hist(sample,freq=F,breaks=25,col="yellow",main="Equal
   interval length ... ")
curve(dgamma(x,shape=1,scale=16.398179),col=2,add=T)


sequence<-seq(0,1,by=0.04)
# qualist<-quantile(sample,sequence) # Restricted to
   when non-candidate parametric distribution is stated
qualist<-qgamma(sequence,shape=1,scale=16.398179)
qualist[1]<-0
sequence;qualist
hist(sample,freq=F,breaks=qualist,col="green",main="Qu
   antile defined lenght... ")
curve(dgamma(x,shape=1,scale=16.398179),col=2,add=T)
```

Interarrival times to a queueing system (III): validation of theoretical distribution

k	$[a_{j-1},a_j)$	N_j	Np_j	$(N_j - Np_j)^2/Np_j$
1	[0,0.6694)	12	9.6	0.6
2	[0.6694,1.3673)	12	9.6	0.6
3	[1.3673,2.0962)	12	9.6	0.6
4	[2.0962,2.8591)	11	9.6	0.2041
5	[2.8591,3.6591)	7	9.6	0.7041
6	[3.6591,4.5002)	10	9.6	0.0166
7	[4.5002,5.3868)	8	9.6	0.2666
8	[5.3868,6.3242)	9	9.6	0.0375
9	[6.3242,7.3183)	10	9.6	0.0166
10	[7.3183,8.3766)	8	9.6	0.2666
11	[8.3766,9.5079)	9	9.6	0.0375
12	[9.5079,10.7232)	7	9.6	0.7041
13	[10.7332,12.0357)	7	9.6	0.7041
14	[12.0357,13.4626)	6	9.6	1.35
15	[13.4626,15.0255)	12	9.6	0.6
16	[15.0255,16.7532)	13	9.6	1.2041
17	[16.7532,18.6846)	9	9.6	0.0375
18	[18.6846,20.8743)	6	9.6	1.35
19	[20.8743,23.4021)	7	9.6	0.7041
20	[23.4021,26.3918)	9	9.6	0.0375
21	[26.3918,30.051)	12	9.6	0.6
22	[30.051,34.7865)	13	9.6	1.2041
23	[34.7865,41.4173)	13	9.6	1.2041
24	[41.4173,52.7837)	8	9.6	0.2666
25	[52.7837,∞)	10	9.6	0.0166
				χ^2=13.3324

Threshold for chi-square distribution $\chi^2_{24,0.90} = 33.196 \geq \chi^2$=13.3324 **statistic** $\Rightarrow$ **H0 accepted**

Goodness of fit - Inferential Issues: Chisq Test, KS Test, Anderson-Darling Test, etc.

- H0: Weibull θ=5, $\alpha = 2$ and $\lambda = 1$.
- Chi Squared Test.

```
> sequence<-seq(0,1,by=0.02)
> qualist<-
      qweibull(shape=res$estimate[1],scale=res$estimate[2],seq
uence) # Percentiles defined by theoretical distribution
> qualist[51]<-max(ussample) # Tune upper/lower thresholds
> dsample<-cut(ussample,breaks=qualist)# Split your sample into
groups defined by theoretical percentiles
> iobs<-as.vector(table(dsample))  # Obs in the groups defined
by nb intervals-percentiles
> pexp<-as.vector(rep(1/50,50))    # Expected probability for
each groups: sample size/nb intervals
    # iexp<-length(ussample)*pexp
    # iobs;pexp;iexp
    # X2<-sum(((iobs-iexp)^2)/iexp);X2
> chisq.test( iobs,p=pexp ) # Works
```

Goodness of fit - Inferential Issues: Chisq Test, KS Test, Anderson-Darling Test, etc.

```
> ks.test(mostra72,"pgamma", shape= 1.5, scale=6.47899  )

        One-sample Kolmogorov-Smirnov test data:  mostra72
D = 0.0159, p-value = 0.1591    alternative hypothesis: two.sided

> ks.test(mostra72,"pgamma", shape= 1.525, scale=6.30  )

        One-sample Kolmogorov-Smirnov test data:  mostra72
D = 0.0126, p-value = 0.4018  alternative hypothesis: two.sided
```

```
# Max LK estimate: package MASS in R:
> fitdistr( mostra72, "gamma",lower=0.001)
      shape         rate
  1.487304711   0.154110460
 (0.027060660) (0.003324827)
> beta=1/rate
> 6.488852
```

Goodness of fit: discrete distributions

- Use library(vcd) in R
- Method: goodfit() to compute p values for discrete distributions – binomial, poisson, neg binomial.
  - Estimated parameters are attributes $par.
  - A summary of goodfit object gives p value of H0.

```
> sample<- rpois(3000, 10)
> gf <- goodfit(table(sample), type = "poisson",
method = "ML")
> summary(gf)
         Goodness-of-fit test for poisson
distribution   X^2 df  P(> X^2)
       Likelihood Ratio 23.36815 19 0.2215152
> gf$par
$lambda
[1] 10.045
```

Some useful random variates for modeling purposes

Weibull distribution W(θ,α,β)

- Probability Density function $f_X(x)$ and distribution function $F_X(x)$ when no shift is considered; i.e. $\theta = 0$:

$$f_X(x) = \begin{cases} \alpha\beta^{-\alpha}x^{\alpha-1}e^{-\left(x/\beta\right)^{\alpha}} & x \geq 0 \\ 0 \quad \text{otherwise} \end{cases} \qquad F_X(x) = \begin{cases} 1-e^{-\left(x/\beta\right)^{\alpha}} & x \geq 0 \\ 0 \quad \text{otherwise} \end{cases}$$

- Shape parameter $\alpha > 0$, scale parameter $\beta > 0$

$$\text{Mean } \beta\Gamma\left(\frac{\alpha+1}{\alpha}\right) \qquad \text{Variance } \beta^2\left\{\Gamma\left(\frac{\alpha+2}{\alpha}\right)-\left[\Gamma\left(\frac{\alpha+1}{\alpha}\right)\right]^2\right\}$$

$$\Gamma(c+1) = c\Gamma(c), \quad \Gamma(0) = \Gamma(1) = 1, \quad \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\Pi}$$

Weibull distribution W(θ,α,β)

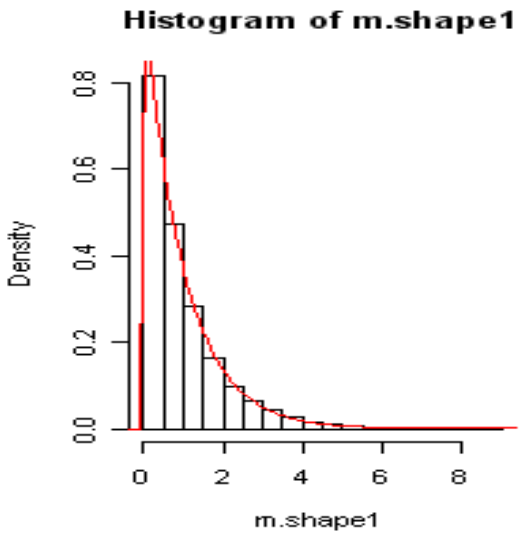
- Pay attention to the functional form, that is meaning of model parameters.
- **Instead W(θ=0,α,λ=1/β), widely used in engineering.**
- Modeling of fatigue in mechanical or electronic components

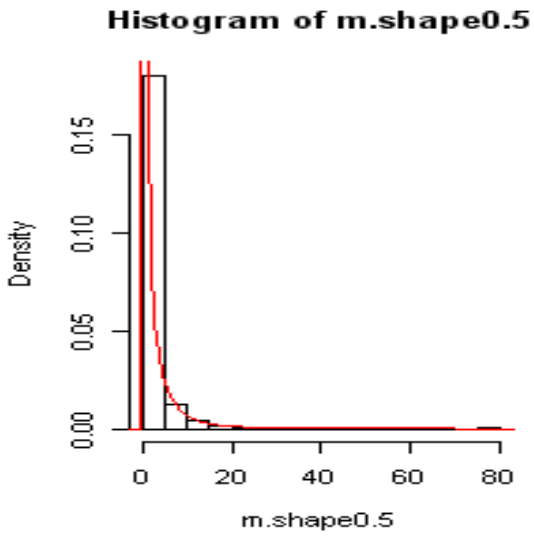
$$f(t) = \lambda\alpha t^{\alpha-1}\mathrm{e}^{-\lambda t^{\alpha}} \qquad F(t) = 1 - e^{-\lambda t^{\alpha}}$$

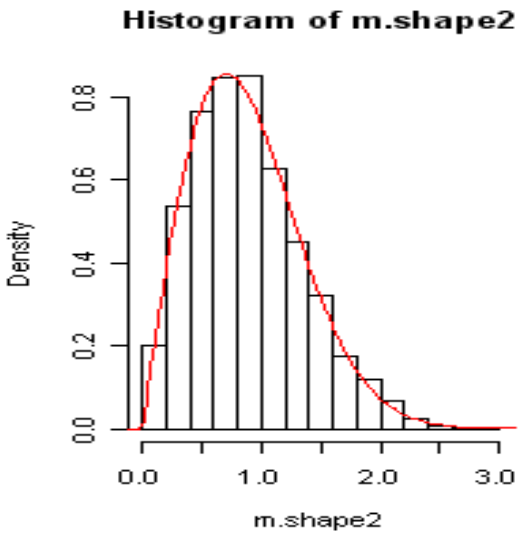
$$E[T] = \frac{1}{\alpha}\left(\frac{1}{\lambda}\right)\Gamma\left(\frac{1}{\alpha}\right) \quad V[T] = \frac{1}{\alpha}\left(\frac{1}{\lambda}\right)^{2}\left(2\Gamma\left(\frac{2}{\alpha}\right) - \frac{1}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)^{2}\right)$$

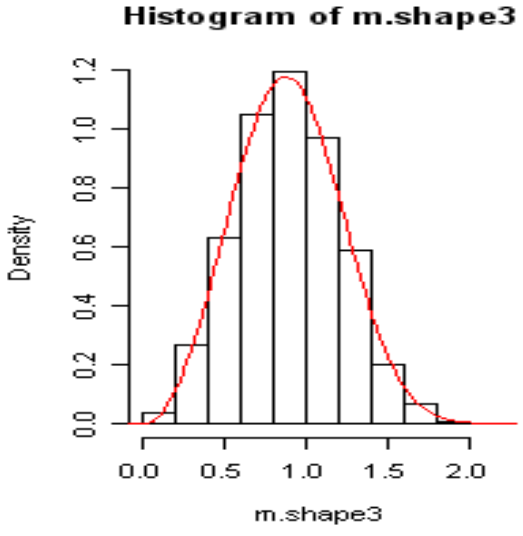
- α shape parameter and λ is rate parameter, related to scale parameter β , $\lambda = 1/\beta$.
- Previously to the use of any statistical package, the distribution functional form must be checked.

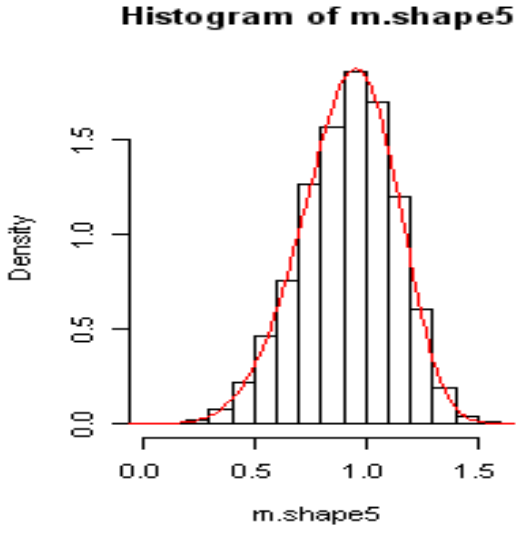
Take a look to standard Weibull pdf's

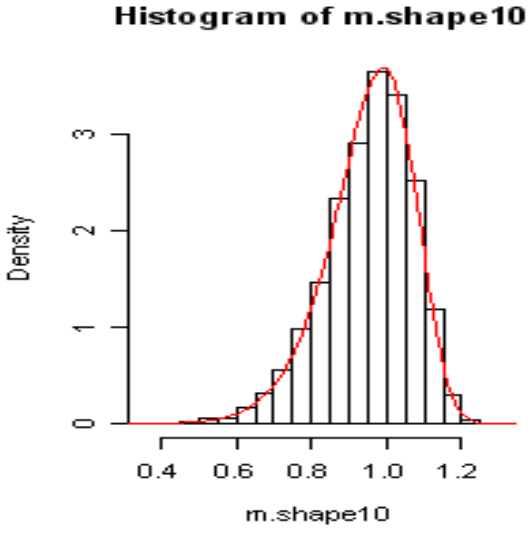




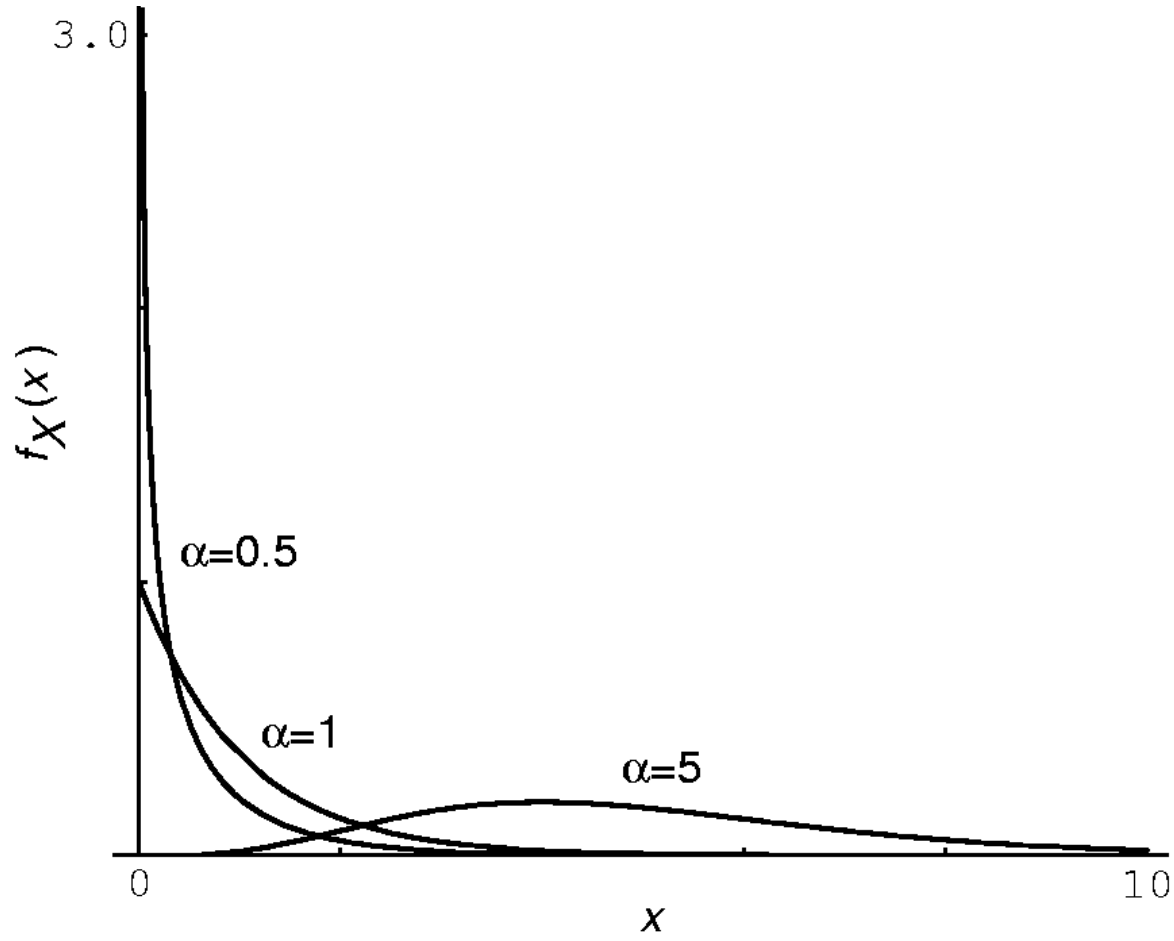




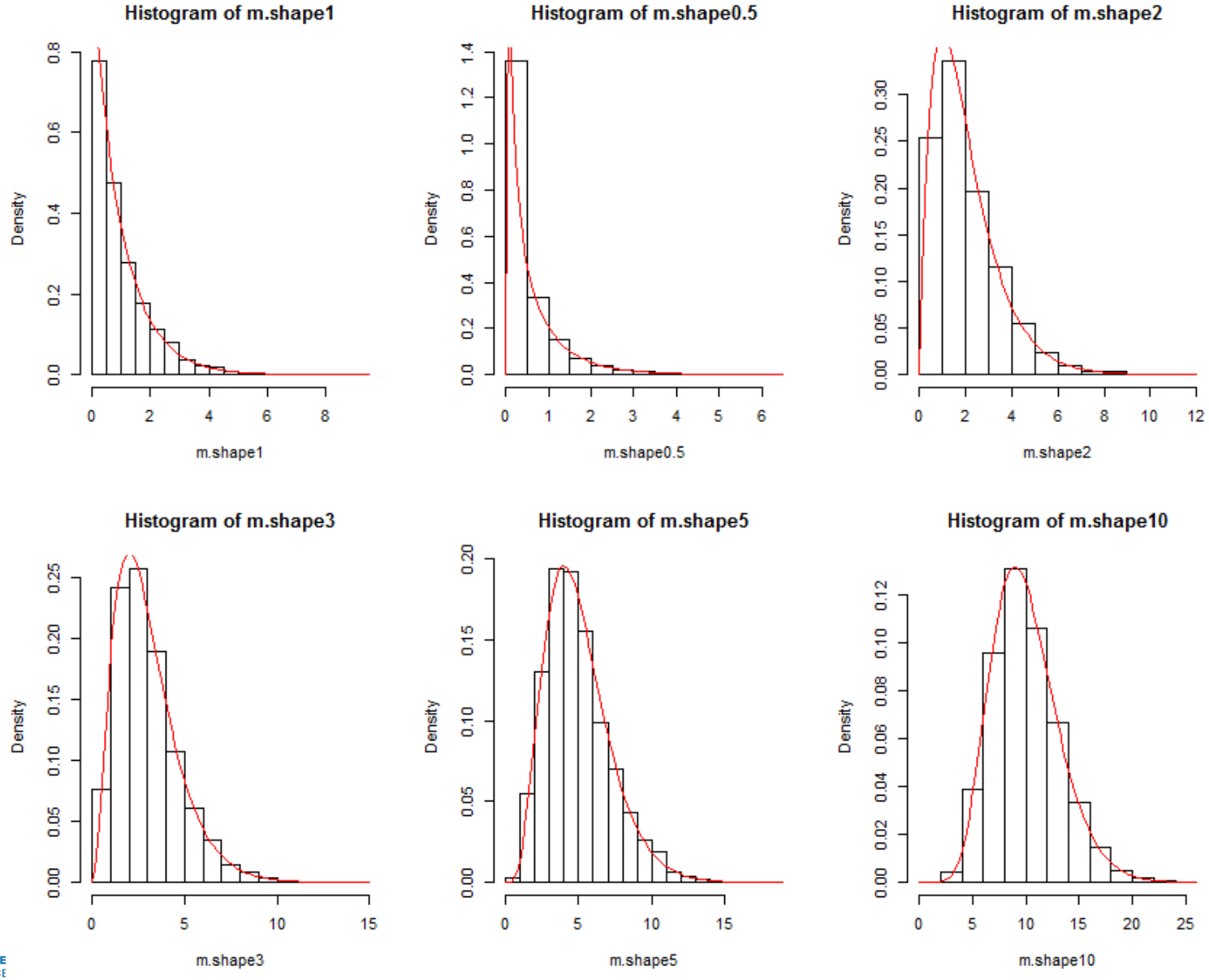




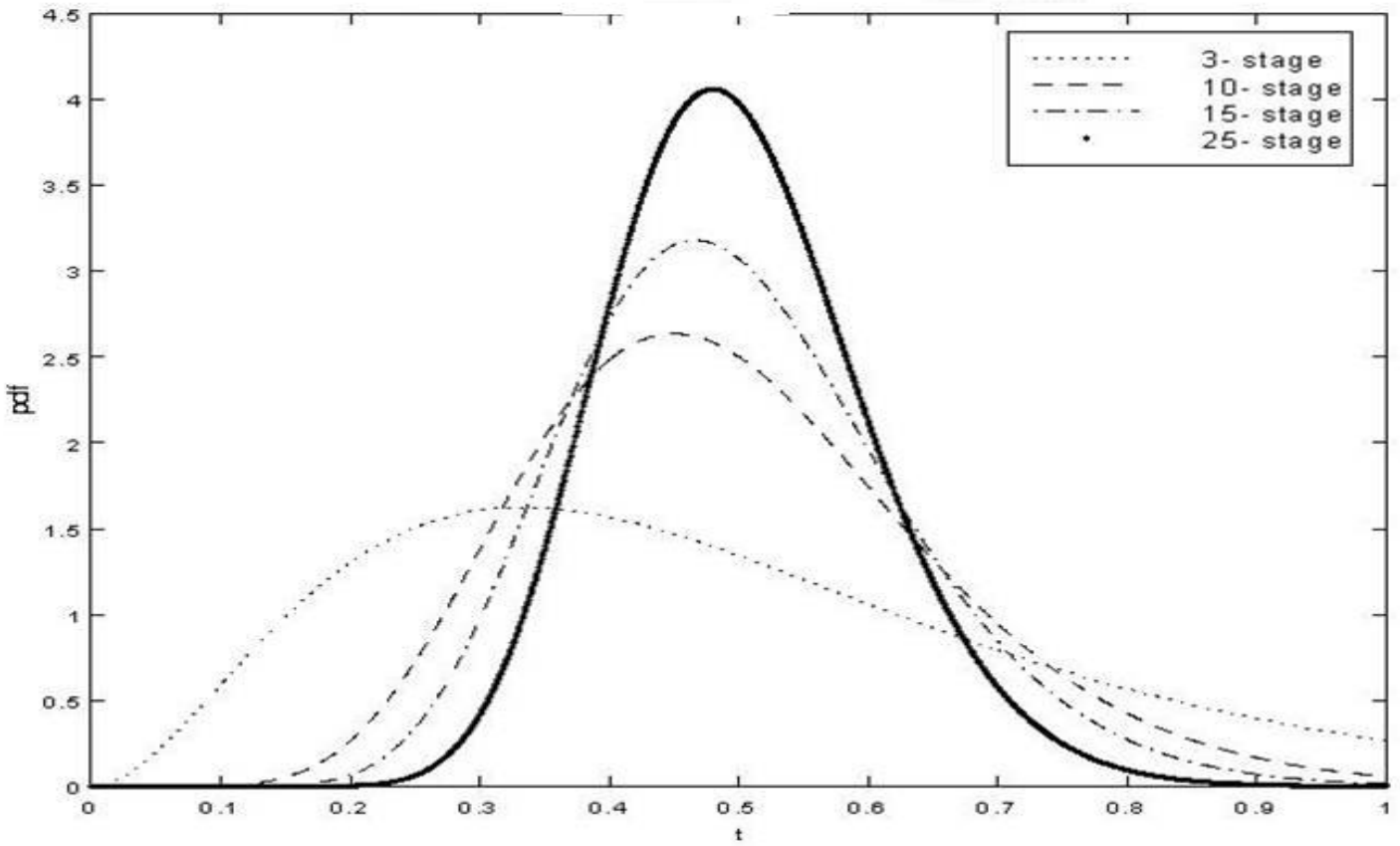
Gamma $\Gamma(\alpha,\beta=1)$, no location parameter considered. Scale $\beta= 1$.



Take a look to standard Gamma pdf's



Erlang Distribution:for statisticians a particular gamma when shape parameter is integer



In computer science a case of hypoexponential distribution

HipoExponential distributions (HIPOEXP)

- HipoExp: Serial connection of Exponentials phases
- 2 phase HipoExp notated as *HiPO*(λ_1, λ_2).
- Density, distribution and hazard functions are:

$$f(t) = \frac{\lambda_1\lambda_2}{\lambda_2-\lambda_1}(\mathrm{e}^{-\lambda_1 t} - \mathrm{e}^{-\lambda_2 t}),\ t > 0$$

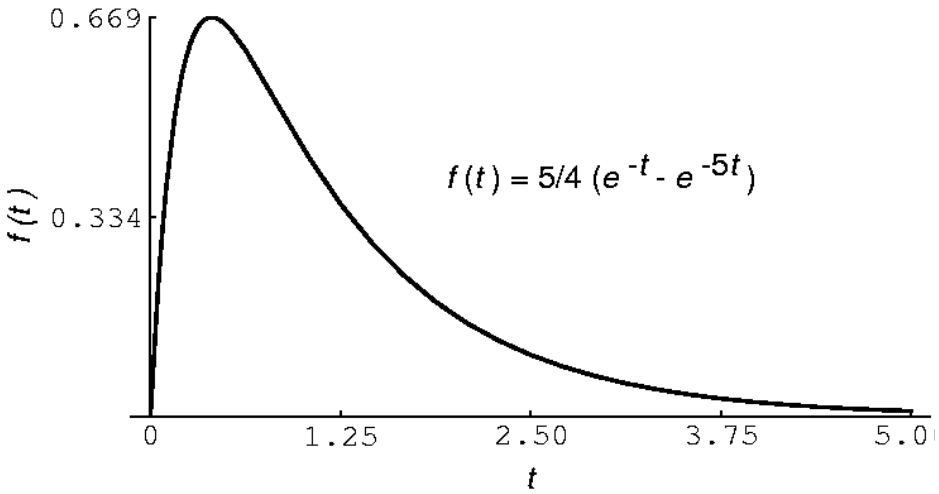
$$F(t) = 1 - \frac{\lambda_2}{\lambda_2-\lambda_1}\mathrm{e}^{-\lambda_1 t} + \frac{\lambda_1}{\lambda_2-\lambda_1}\mathrm{e}^{-\lambda_2 t},\ t \geq 0$$

$$h(t) = \frac{\lambda_1\lambda_2(\mathrm{e}^{-\lambda_1 t} - \mathrm{e}^{-\lambda_2 t})}{\lambda_2\mathrm{e}^{-\lambda_1 t} - \lambda_1\mathrm{e}^{-\lambda_2 t}}.$$

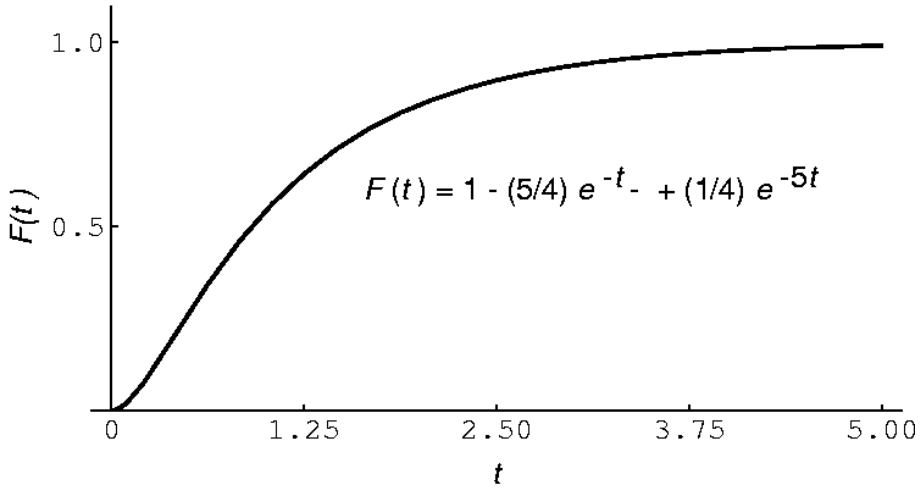
- Service time for a disk, usually modeled as 3 phase HipoExp since total access time is the sum of searching time plus latency plus transfer timeés.

HipoExponential distributions (HIPOEXP)

- HipoExp(1,5)



HipoExp(1,5) density function



HipoExp(1,5) Distribution function

Erlang distribution

- Particular case of HIPOEXP process: all phases are i.i.d exponentially distributed.

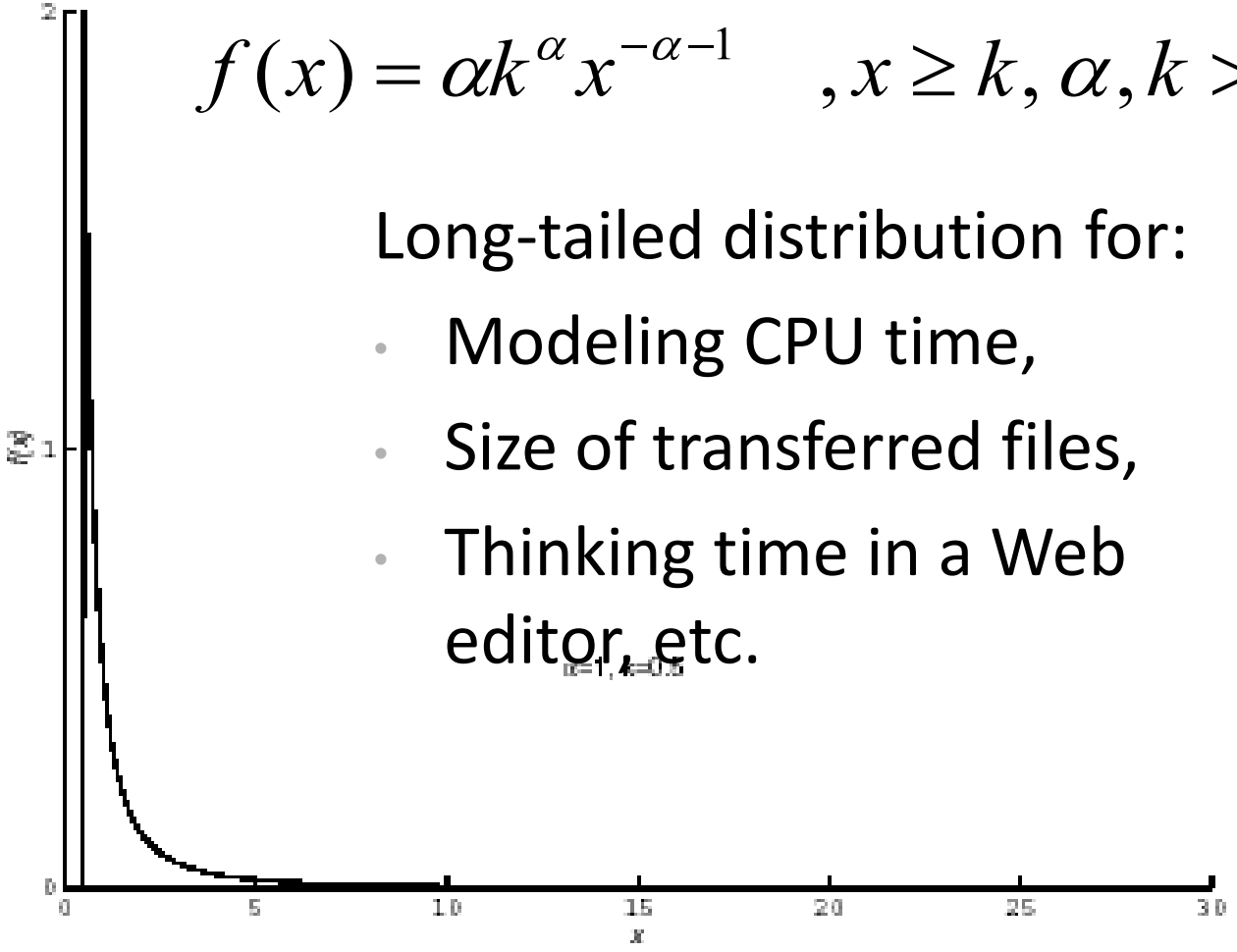
$$f(t) = \frac{\lambda^r t^{r-1} \mathrm{e}^{-\lambda t}}{(r-1)!},\ t > 0,\ \lambda > 0,\ r = 1, 2, \ldots$$

$$F(t) = 1 - \sum_{k=0}^{r-1} \frac{(\lambda t)^k}{k!} \mathrm{e}^{-\lambda t},\ t \geq 0,\ \lambda > 0,\ r = 1, 2, \ldots$$

- Also a particular case of gamma distribution when shape parameter is integer (meaning, number of modeling phases).
- No specific distribution in R for Erlang variates.

Pareto variate: Density function plot

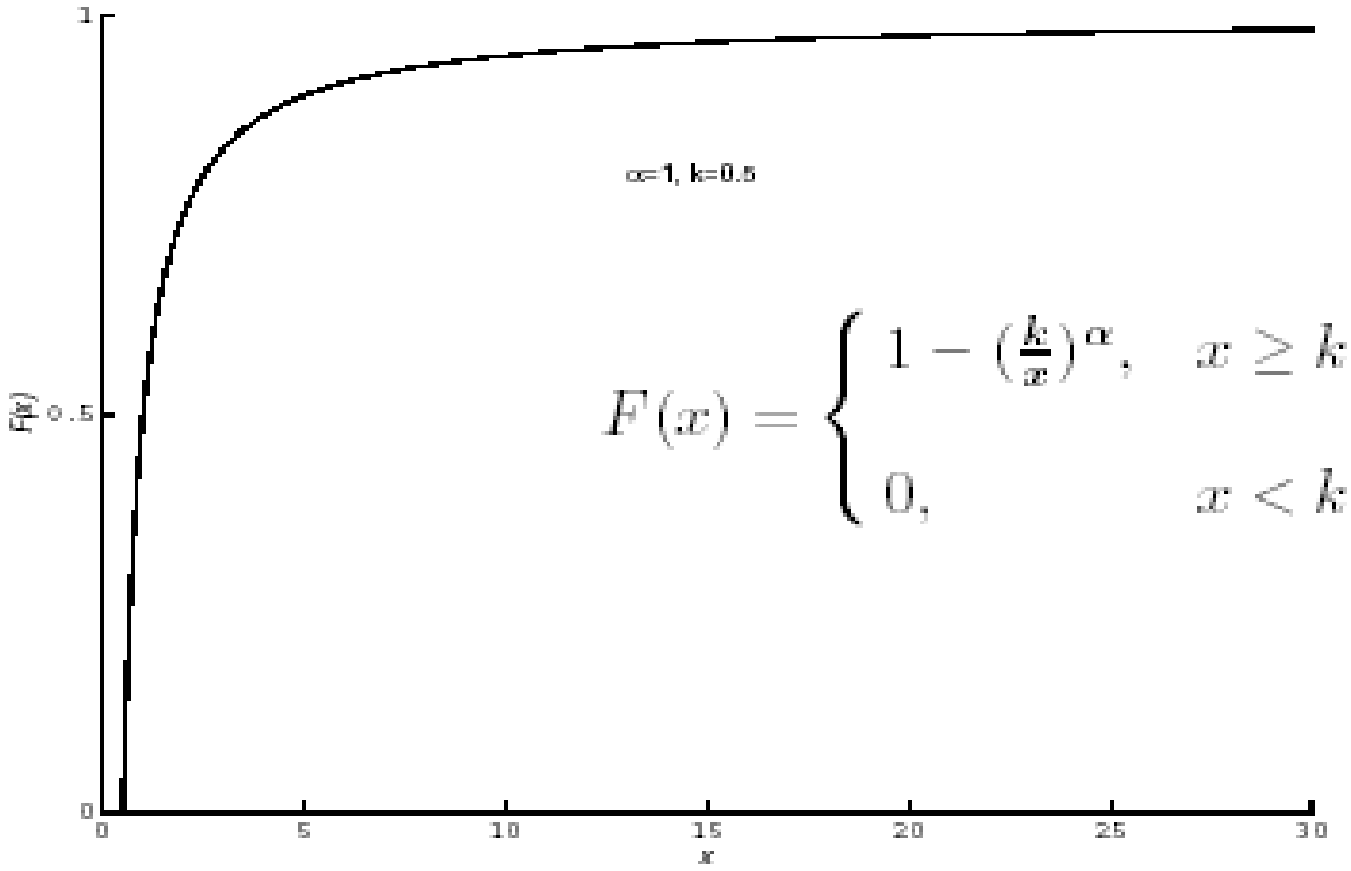
$$f(x) = \alpha k^{\alpha} x^{-\alpha-1} \quad , x \geq k, \alpha, k > 0$$



Long-tailed distribution for:

- Modeling CPU time,
- Size of transferred files,
- Thinking time in a Web editor, etc.

Pareto variate: Distribution function plot



$$F(x) = \begin{cases} 1 - \left(\frac{k}{x}\right)^{\alpha}, & x \geq k \\ 0, & x < k \end{cases}$$

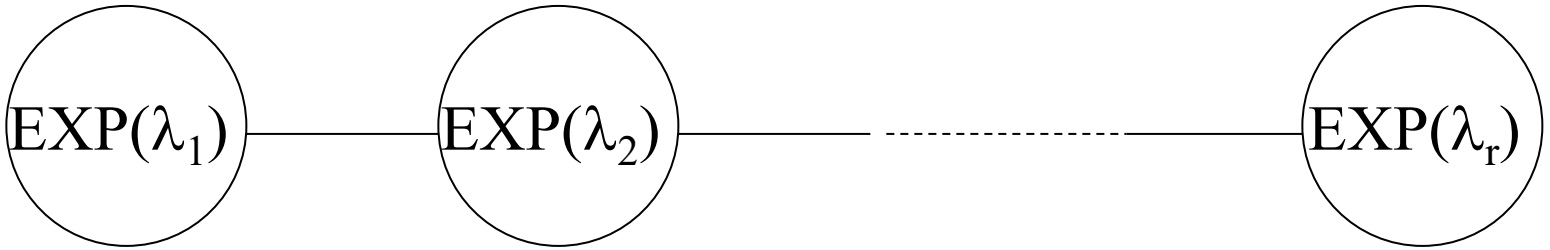
Hipoexponential: General case

- $Z = \sum_{i=1}^{r} X_i$, where $X_1, X_2, \ldots, X_r$ are independent variates with exponential distribution of rate λ_i

$$\lambda_i \neq \lambda_j \quad for\ i \neq j$$

- Then Z is an r-phase hipoexponential variate.

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{1}{\lambda_i} \qquad \mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[X_i] = \sum_{i=1}^{n} \frac{1}{\lambda_i^2},$$



HiperExponential variate (HIPEREXP)

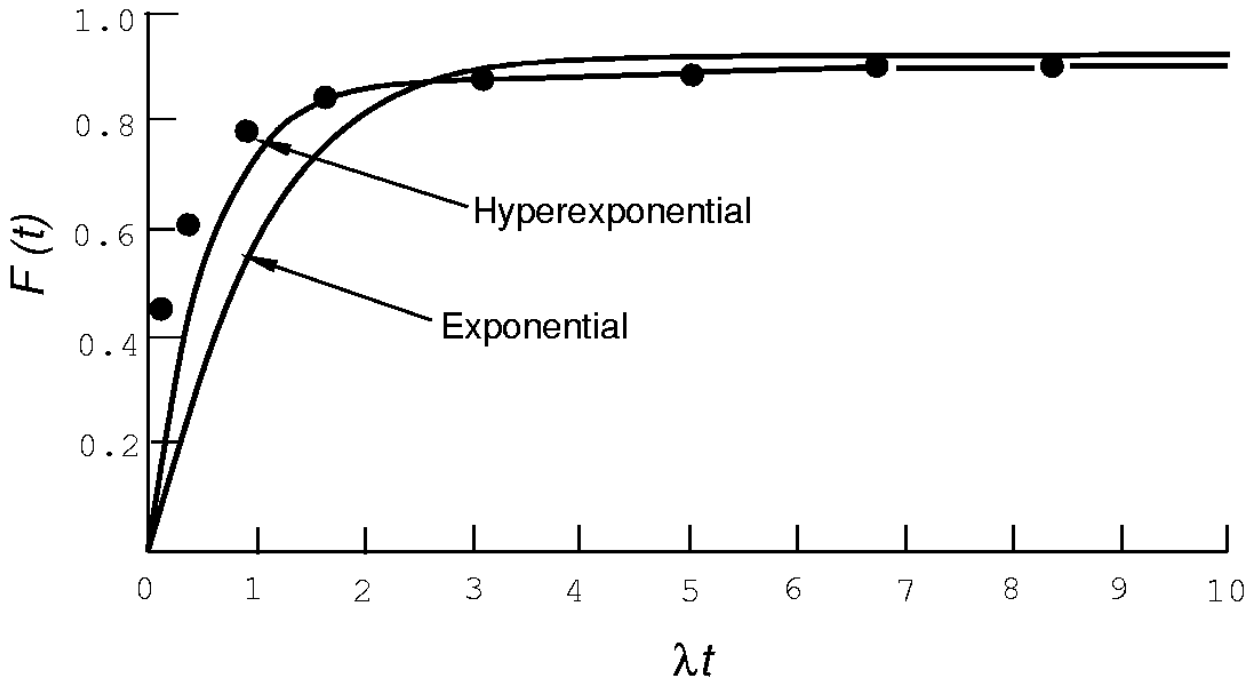
- Composed Distribution for alternance of K Exponential variates (non identical)
- Ellapsed CPU Time sometimes modeled as a HiperExp.

$$f(t) = \sum_{i=1}^{k} \alpha_i \lambda_i \mathrm{e}^{-\lambda_i t}, \ t > 0, \ \lambda_i > 0, \ \alpha_i > 0, \ \sum_{i=1}^{k} \alpha_i = 1$$

$$F(t) = \sum \alpha_i (1 - \mathrm{e}^{-\lambda_i t}), \ t \geq 0$$

$$\begin{aligned} E[X] &= \int_0^\infty (\sum_{i=1}^{n} x \alpha_i \lambda_i e^{-\lambda_i x}) dx \\ &= \sum_{i=1}^{n} \alpha_i \int_0^\infty x \lambda_i e^{-\lambda_i x} dx \\ &= \sum_{i=1}^{n} \frac{\alpha_i}{\lambda_i}. \end{aligned}$$

Hiper Exponencial Vs. Exponencial CDF



Transformed variates: Y pdf from X pdf since $Y = \Phi(X)$

- Let X be a normal variate $X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- *and let us define* $Y = \Phi(X) = e^X$, thus

$$\Phi^{-1}(y) = \ln(y) \;\Rightarrow\; [\Phi^{-1}]'(y) = \frac{1}{y}$$

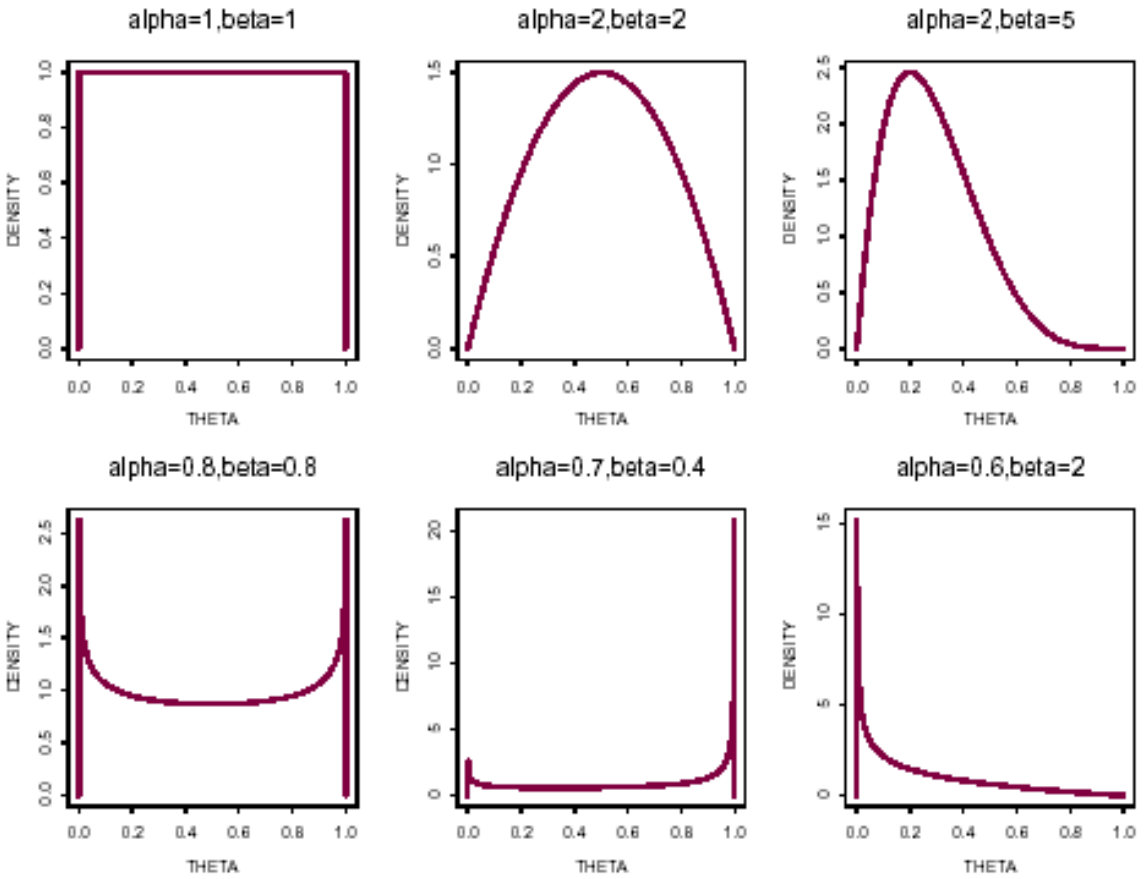
and

$$\begin{aligned} f_Y(y) &= \frac{f(\ln y)}{y} \\ &= \frac{1}{\sigma y\sqrt{2\pi}} \exp\left[-\frac{(\ln y - \mu)^2}{2\sigma^2}\right], \qquad y > 0. \end{aligned}$$

- Y is a *log-normal* variate-

Beta distribution… 2 shape parameters

- Stàndard defined (0,1).
- Let α , β, be shape parameters, thus unimodal, J-shape, U-shape, … can be obtained.



Distribution-Fitting Issues

- What if nothing fits?
  - Empirical distributions
- Too much data?
  - “False” rejection
  - Watch for bad data
- Sensitivity analysis
  - Test both means and variation
- Types of inputs:
  - Deterministic vs. Stochastic
  - Scalar vs. Multivariate vs. Stochastic
  - Time-Varying Arrival Rates