Refresh on some basic

probability distributions

04/10/09

outline

- density, distribution, moments
- uniform distribution
- Poisson process, exponential distribution
- Pareto function
  - density and distribution
  - residual waiting time
  - tail of distribution, fitting
- hypoexponential distribution (Erlang)
  - simulation
- hyperexponential distribution
  - simulation
- fitting

moments

variance and second order moment

- variance σ^2 of a discrete random var. X with mean E[X]

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{1}{n}\sum_i x_i^2 + \frac{1}{n}\sum_i \bar{x}^2 - \frac{2\bar{x}}{n}\sum_i x_i = E[X^2] + E[X]^2 - 2E[X]^2 =$$

$$= E[X^2] - E[X]^2$$

- variance σ^2 of a continuous random var. X with mean μ

$$\sigma^2 = E[(X-\mu)^2] = E[X^2 - 2X\mu + \mu^2] = \int_{-\infty}^{+\infty} (x^2 - 2x\mu + \mu^2)\, f(x)\, dx =$$

$$= \int_{-\infty}^{+\infty} x^2 f(x)\, dx - 2\mu \int_{-\infty}^{+\infty} x\, f(x)\, dx + \mu^2 \int_{-\infty}^{+\infty} f(x)\, dx =$$

$$= E[X^2] - 2\mu\mu + \mu^2 = E[X^2] - \mu^2$$

uniform distribution

Poisson process

exponential distribution

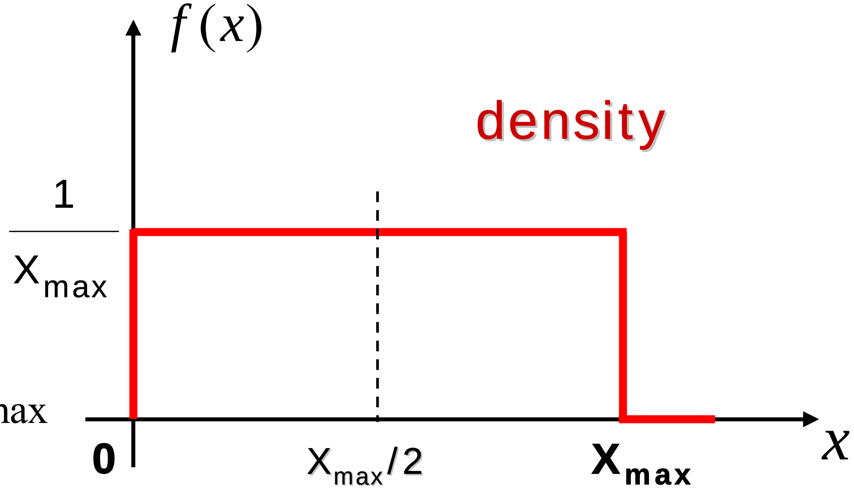
uniform distrib. over the interval (0,Xmax)

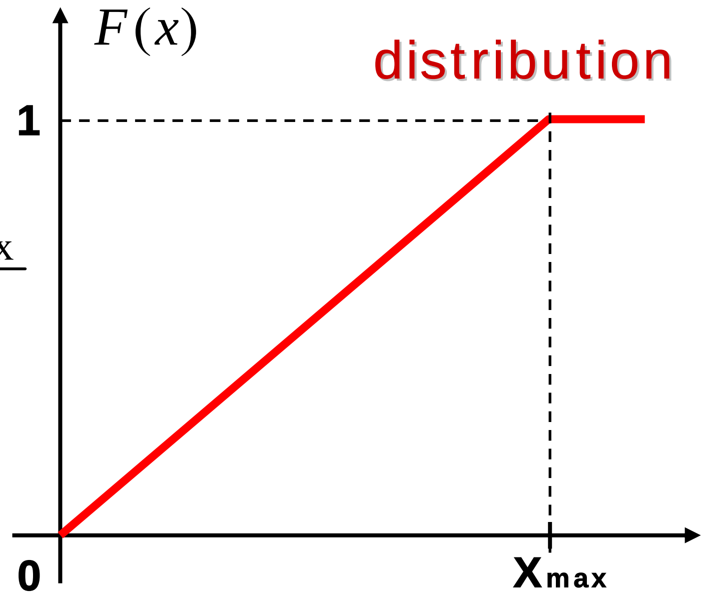
$$f(x) = \frac{1}{X_{\max} - 0} \qquad 0 < x < X_{\max}$$

$$F(x) = \int_0^x \frac{1}{X_{\max}} dt = \frac{x}{X_{\max}} \qquad 0 \leq x < X_{\max}$$

$$= 1 \qquad x \geq X_{\max}$$

$$E[X] = \int_0^{X_{\max}} x \frac{1}{X_{\max}} dx = \left[\frac{x^2}{2 X_{\max}} \right]_0^{X_{\max}} = \frac{X_{\max}}{2}$$





Poisson process

random point process whose associated counting process $N(t)$ satisfies:

- independent increments
- stationary increments
- $P[\text{one event in } (t,t+h)] = \lambda h + o(h)$
- $P[\text{two or more events in } (t,t+h)] = o(h)$

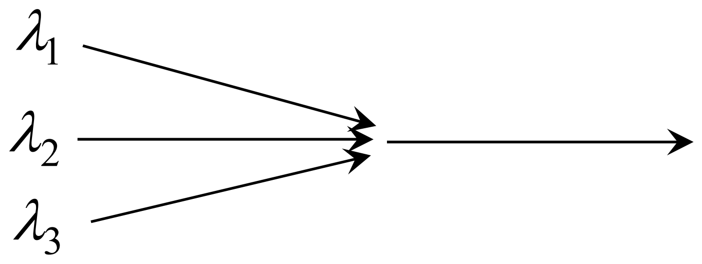
events occur singly at a rate λ uniform in time

the number of arrivals in any time interval t has a Poisson distribution with mean λt

$$P_0(t) = e^{-\lambda t} \qquad P_k(t) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

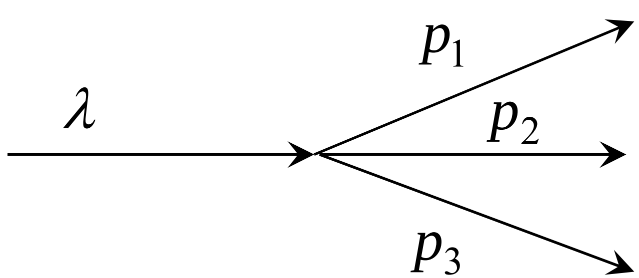
Poisson process

when multiple Poisson streams are merged the resulting stream is a Poisson stream with intensity equal to the sum of the intensities



$$\lambda = \lambda_1 + \lambda_2 + \lambda_3$$

a single Poisson stream can be split into independent Poisson streams



$$\lambda\ p_1 = \lambda_1$$

$$\lambda\ p_2 = \lambda_2$$

$$\lambda\ p_3 = \lambda_3$$

$$p_1 + p_2 + p_3 = 1$$

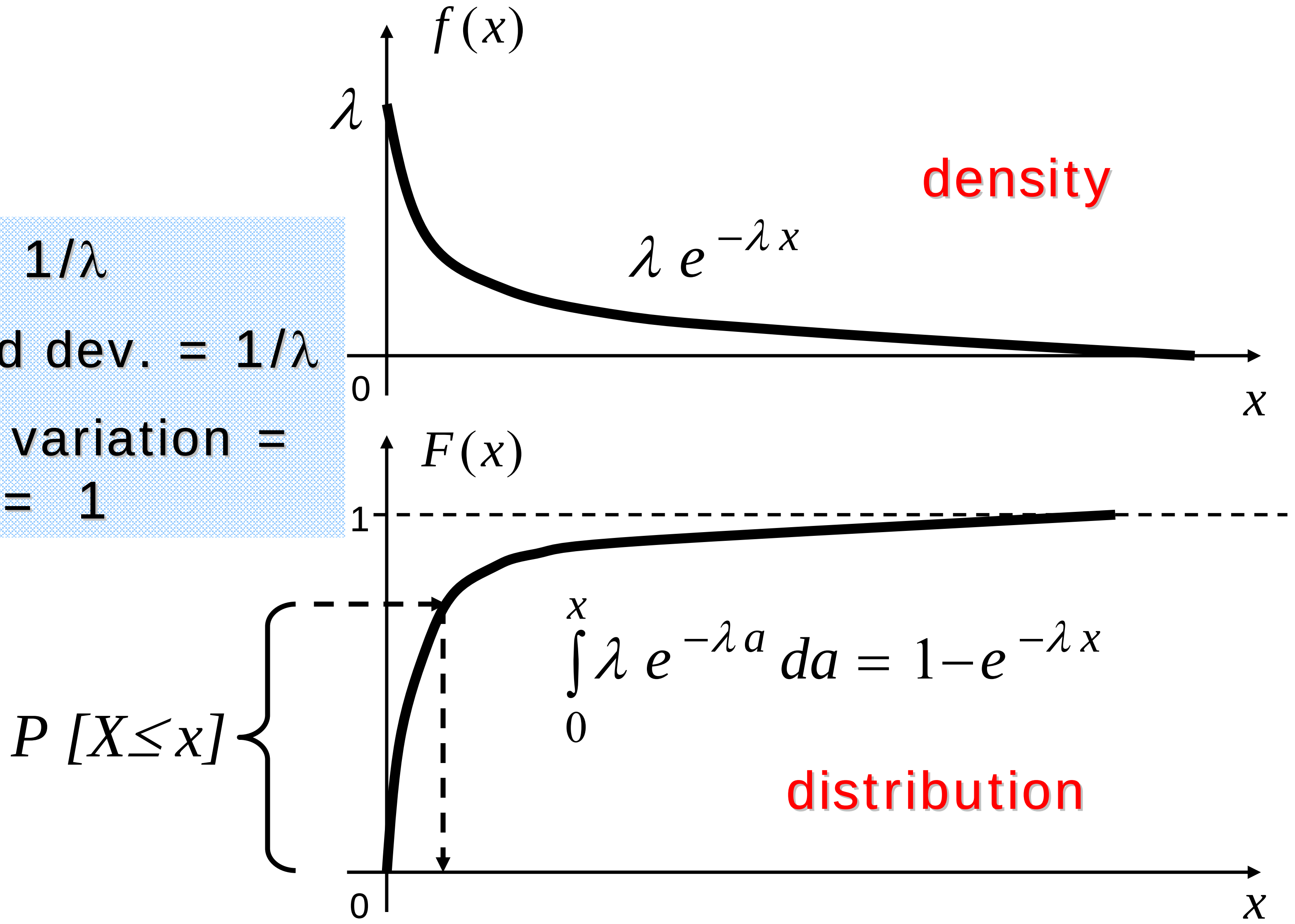
$$\lambda_1 + \lambda_2 + \lambda_3 = \lambda$$

exponential distribution

mean = $1/\lambda$

standard dev. = $1/\lambda$

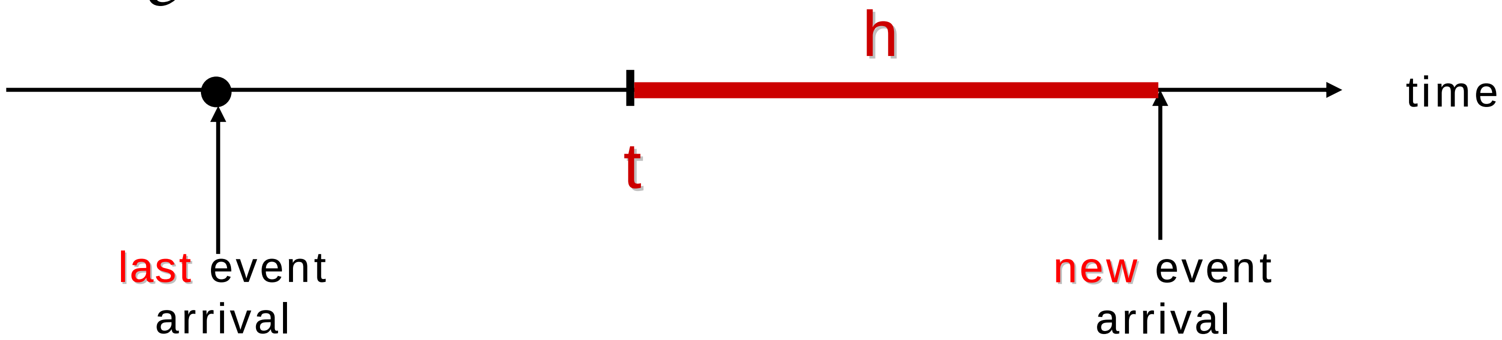
coeff.of variation =
= σ/μ = 1



exp. distr. - memoryless property (random)

$$P[X > t + h | X > t] = \frac{P[(X > t + h) \cap (X > t)]}{P[X > t]} =$$

$$\frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} = P[X > h]$$



X: interarrival time, *t* time units elapsed since last event, the distribution of the remaining waiting time *h* is independent of *t*, the system is ***memoryless***

exponential distribution: percentiles

$r-th$ percentile $\quad P[X \le \prod(r)] = r/_{100}$

90% of values are less than 90-*th* percentile

$$P[X \le \prod(90)] = 0.9 = 1 - e^{-\lambda \prod(90)}$$

$$e^{-\lambda \prod(90)} = 0.1 \qquad\qquad -\lambda \prod(90) = \log 0.1$$

$$\prod(90) = -\log 0.1 /_{\lambda} = 1/_{\lambda} \log 10 \cong \boxed{2.3\ 1/_{\lambda}}$$

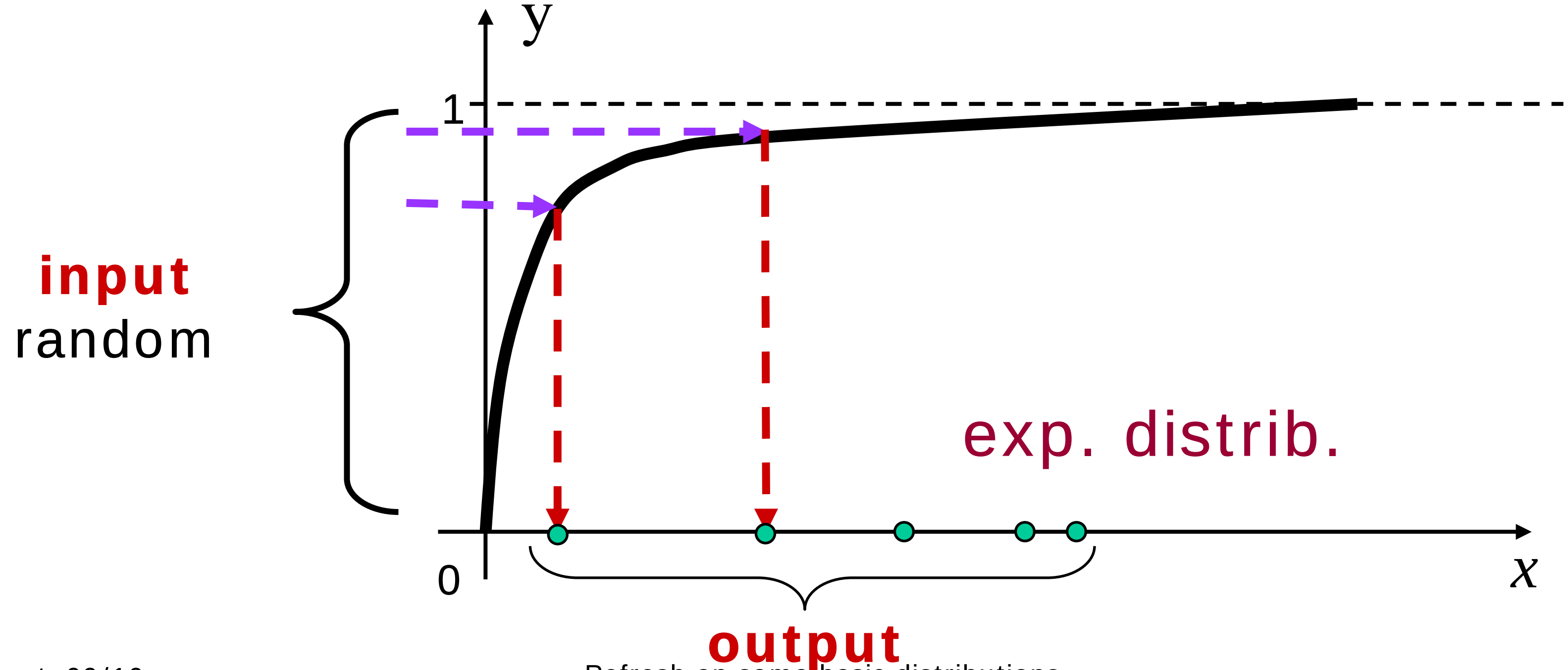
$$\prod(r) = 1/_{\lambda} \log \left(100/_{100-r}\right)$$

generation of exponentially distrib. sequence

- we can obtain a random deviate x of an EXP(λ) random variable by first generating a random number y from a uniform distribution over (0,1) and then using the relation $x = F^{-1}(y)$

$$y = 1 - e^{-\lambda x} \qquad e^{-\lambda x} = 1 - y \qquad \log(e^{-\lambda x}) = \log(1-y)$$

$$-\lambda x = \log y \qquad x = -{}^{1}\!/_{\lambda} \log y \qquad [\,0 \le y \le 1 \text{ random }]$$



Pareto function

Web performance indices

- typical phenomenon observed on the Internet for several performance indices (download times, connection times, file sizes, think time of web browser, ...)

- extreme variability of traffic characteristics (arrival times, number and sizes of downloaded files, ...)
- extreme variability of performance indices (download times, connection times, ...)
- teletraffic distributions follow an exponential decay, Web traffic distributions follow a power decay

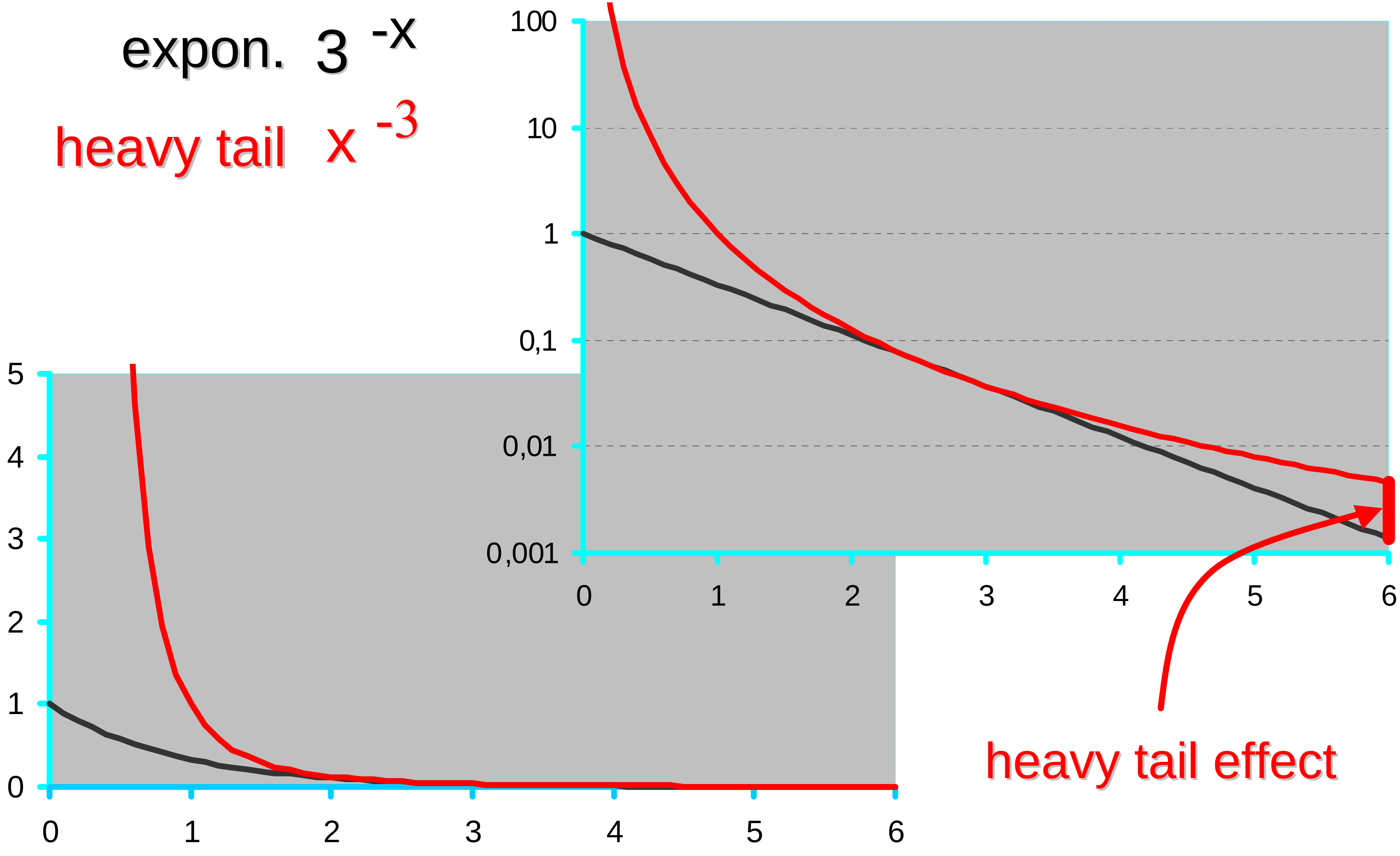
⇓

very high values of variables occur with non negligible probability (heavy tail)

exponential vs heavy tail decrease

expon. 3^{-x}

heavy tail x^{-3}



Pareto function

Pareto

exponential

$$p(x) = \alpha k^{\alpha} x^{-\alpha-1} \qquad \alpha > 0 \qquad 0 < k \leq x \qquad\qquad \lambda e^{-\lambda x}$$

$$F(x) = P[X \leq x] = \int_{k}^{x} \alpha k^{\alpha} h^{-\alpha-1} \; dh =$$

$$= \alpha k^{\alpha} \frac{1}{-\alpha} \; [h^{-\alpha}]_{k}^{x} = 1 - k^{\alpha} x^{-\alpha} \qquad\qquad 1 - e^{-\lambda x}$$

$$\alpha > 1 \qquad mean\ value = k \frac{\alpha}{\alpha - 1} \qquad\qquad \frac{1}{\lambda}$$

if $\alpha \leq 2 \Rightarrow$ infinite variance

if $\alpha \leq 1 \Rightarrow$ infinite mean

mean

$$mean \quad \mu = \int_k^\infty x\, f(x)\; dx = \alpha k^\alpha \int_k^\infty x \quad x^{-\alpha-1} \quad dx =$$

$$= \alpha k^\alpha \int_k^\infty x^{-\alpha} \quad d\,x \; = \alpha\, k^\alpha \; \frac{[x^{-\alpha+1}]_k^\infty}{-\alpha+1} \qquad\qquad 0 < k \leq x$$

$$if \;\; \alpha > 1 \qquad mean \; = \alpha\, k^\alpha \; \frac{-\,k^{-\alpha+1}}{-\alpha+1} \; = \; k\,\frac{\alpha}{\alpha-1}$$

$$if \;\; \alpha \leq 1 \qquad mean \; = \infty \qquad since \quad [x^{-\alpha+1}]_k^\infty \;\rightarrow\; \infty$$

variance

variance $\sigma^2 := E[(X - E[X])^2] = E[X^2] - (E[X])^2 =$

$$\int_k^\infty x^2 f(x)\,dx - \left(\int_k^\infty x\, f(x)\,dx\right)^2 = \alpha k^\alpha \int_k^\infty x^2\, x^{-\alpha-1}\,dx - \left(\alpha k^\alpha \int_k^\infty x\, x^{-\alpha-1}\,dx\right)^2$$

$$= \alpha k^\alpha \left[\frac{x^{-\alpha+2}}{-\alpha+2}\right]_k^\infty - \left(\alpha k^\alpha \left[\frac{x^{-\alpha+1}}{-\alpha+1}\right]_k^\infty\right)^2 \qquad \alpha > 0 \quad 0 < k \le x$$

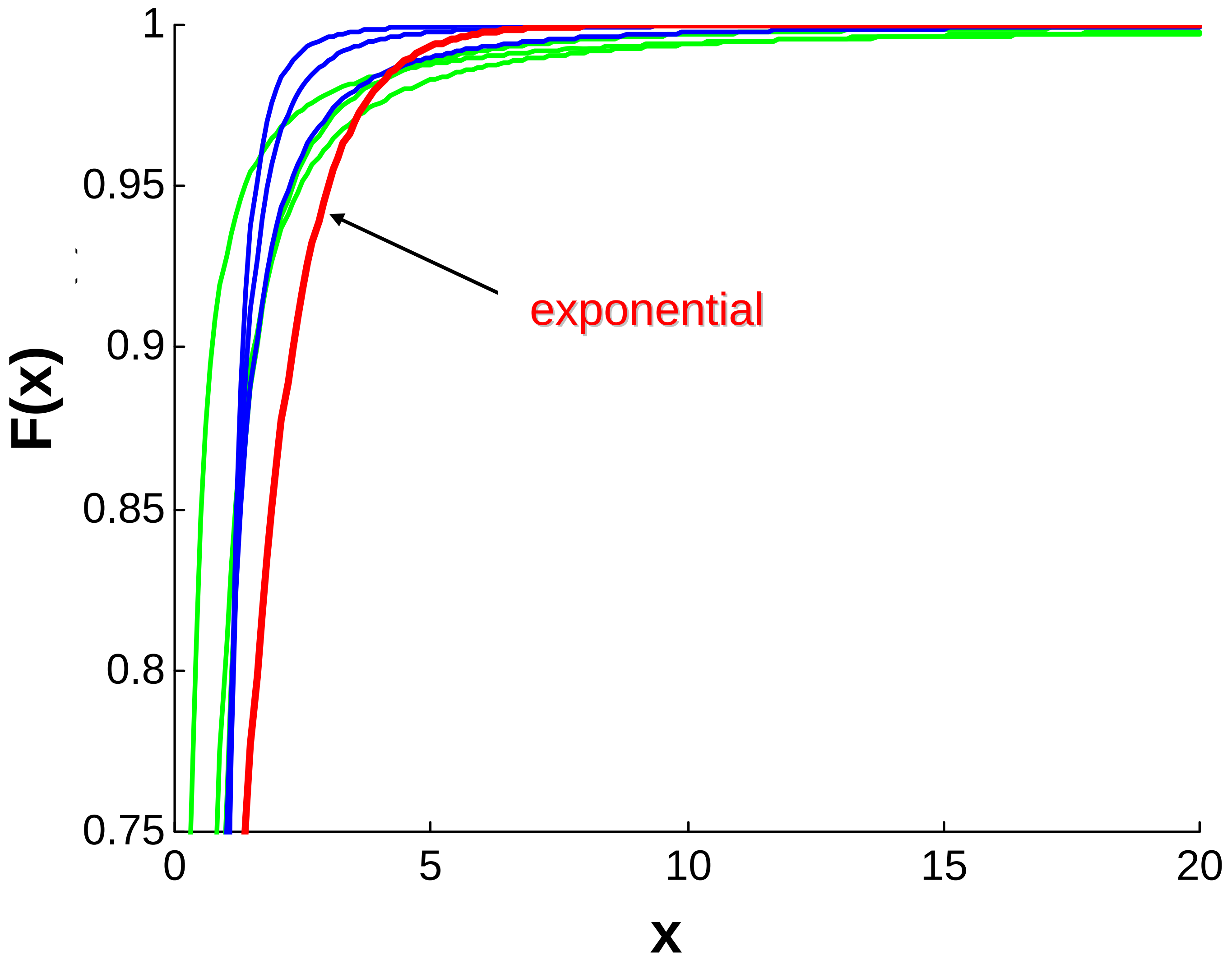
if $\alpha > 2$ $\quad \sigma^2 = \alpha k^\alpha \dfrac{-k^{-\alpha+2}}{-\alpha+2} - \left(k\dfrac{\alpha}{\alpha-1}\right)^2 = k^2 \dfrac{\alpha}{\alpha-2} - \left(k\dfrac{\alpha}{\alpha-1}\right)^2$

if $\alpha \le 2$ $\quad \sigma^2 = \infty$ because $[x^{-\alpha+2}]_k^\infty \to \infty$

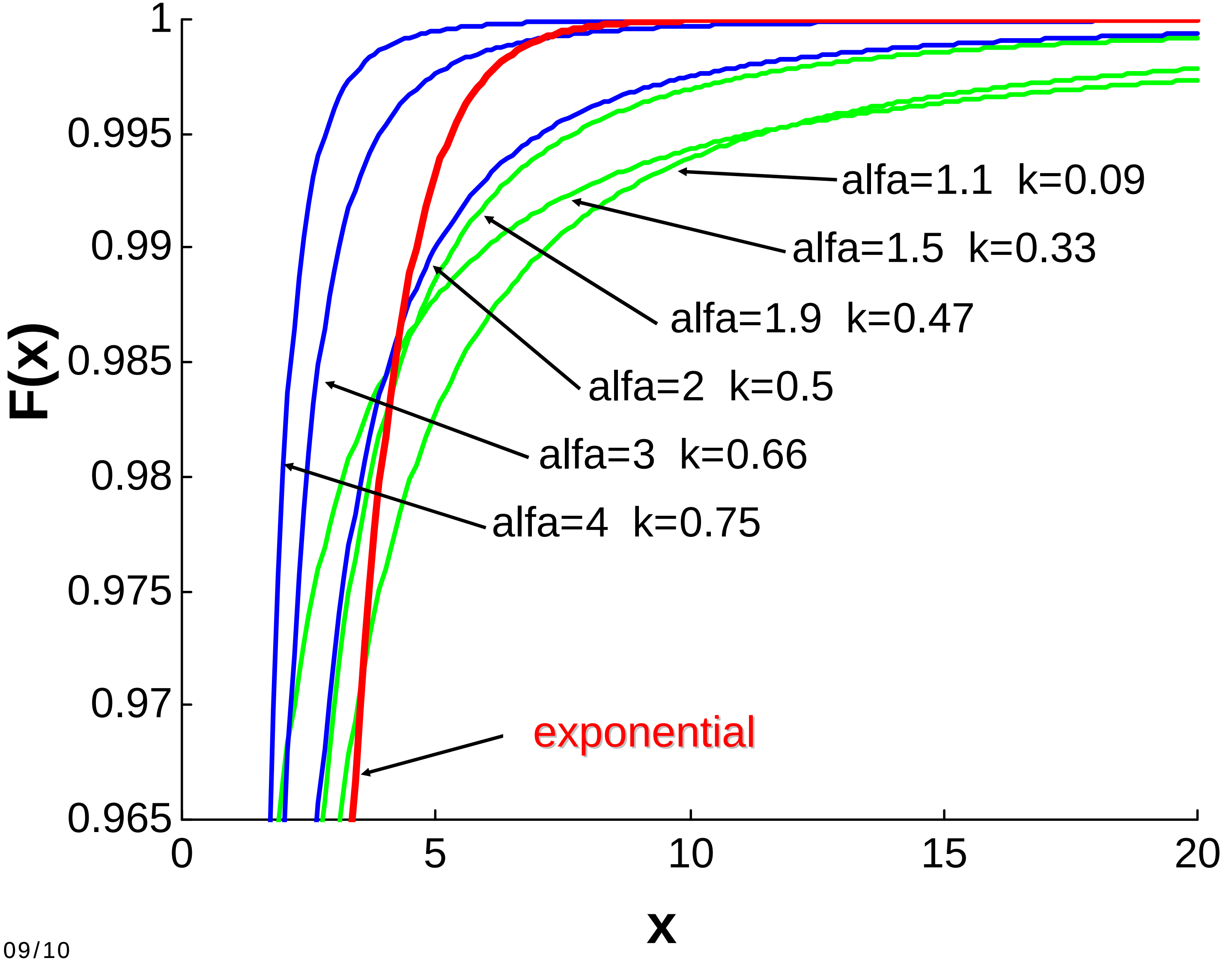
shape parameter α

- decreasing the value of α increases the portion of probability mass that is present in the tail of the distribution
- a random variable distributed according to a Pareto function with $\alpha \le 2$ can give rise to very large values with a non negligible probability

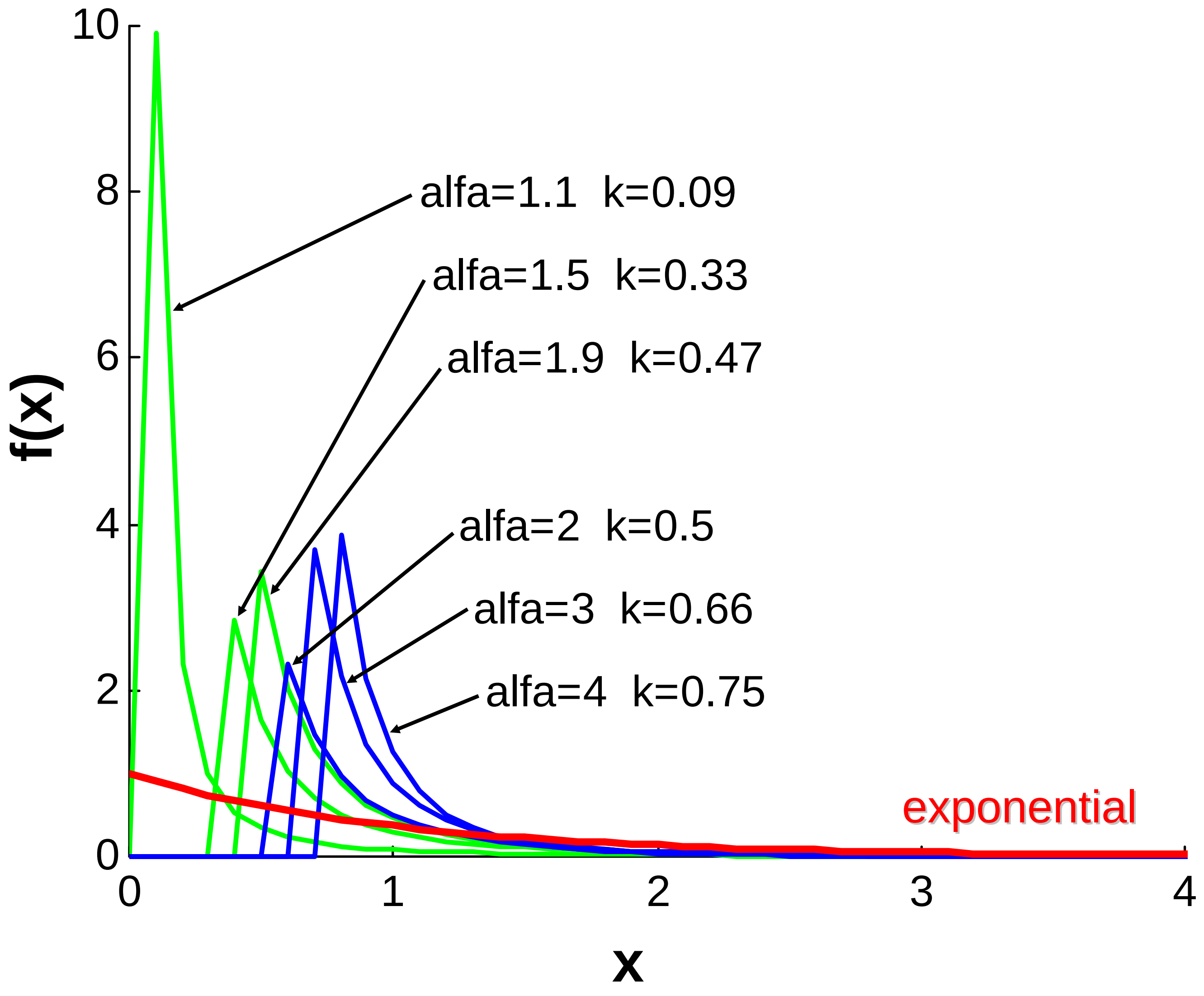
Pareto - distrib. mean = 1



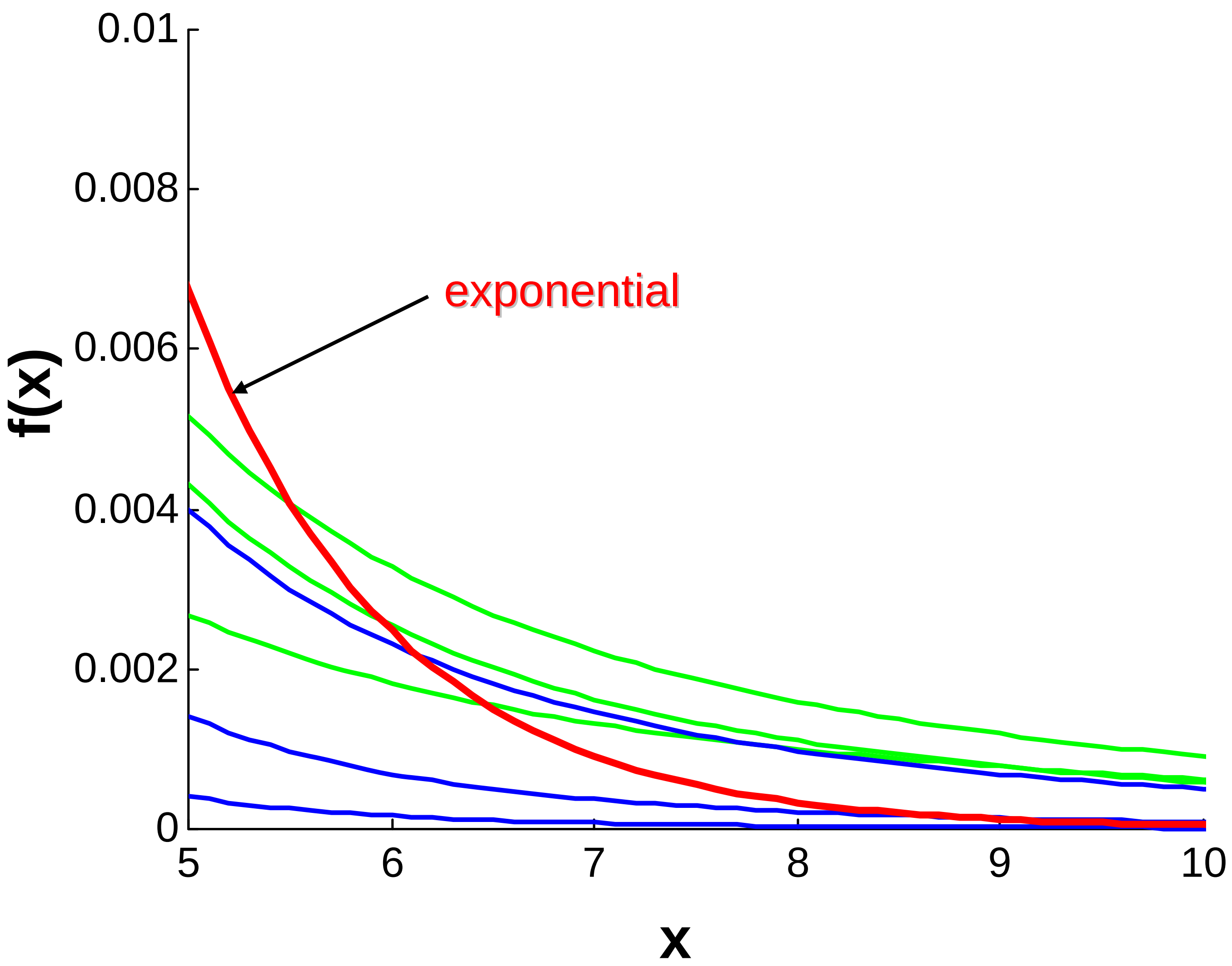
Pareto distrib.   mean = 1



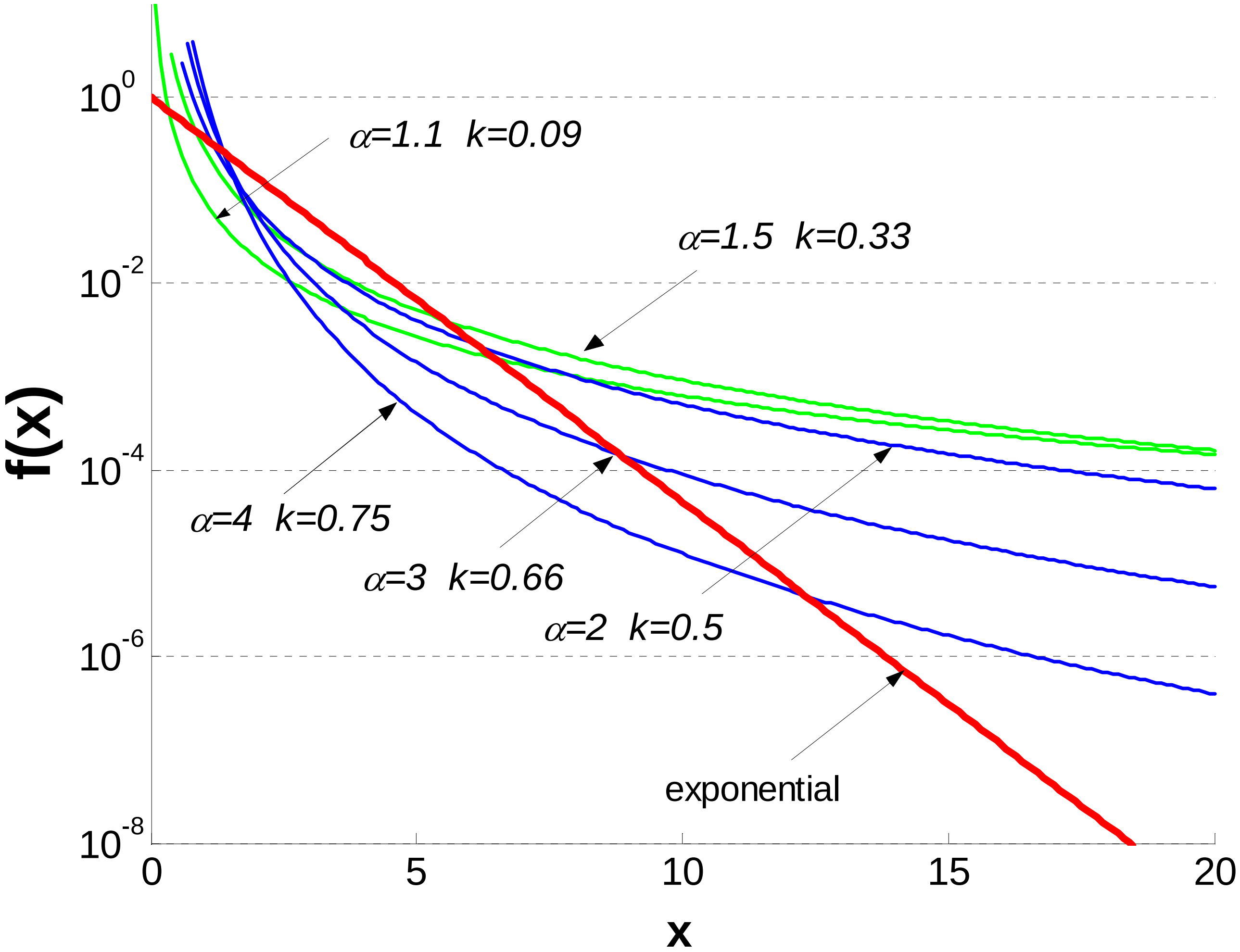
Pareto dens.   mean = 1



Pareto dens.   mean = 1



log dens. of Pareto funct. mean 1



distribution tail

- reliability *R(t)* of a system: probability that the system survives until time *t* (*F(t)* unreliability function)

$$R(t)=P(X>t)=1-F(t)=\overline{F(t)}$$

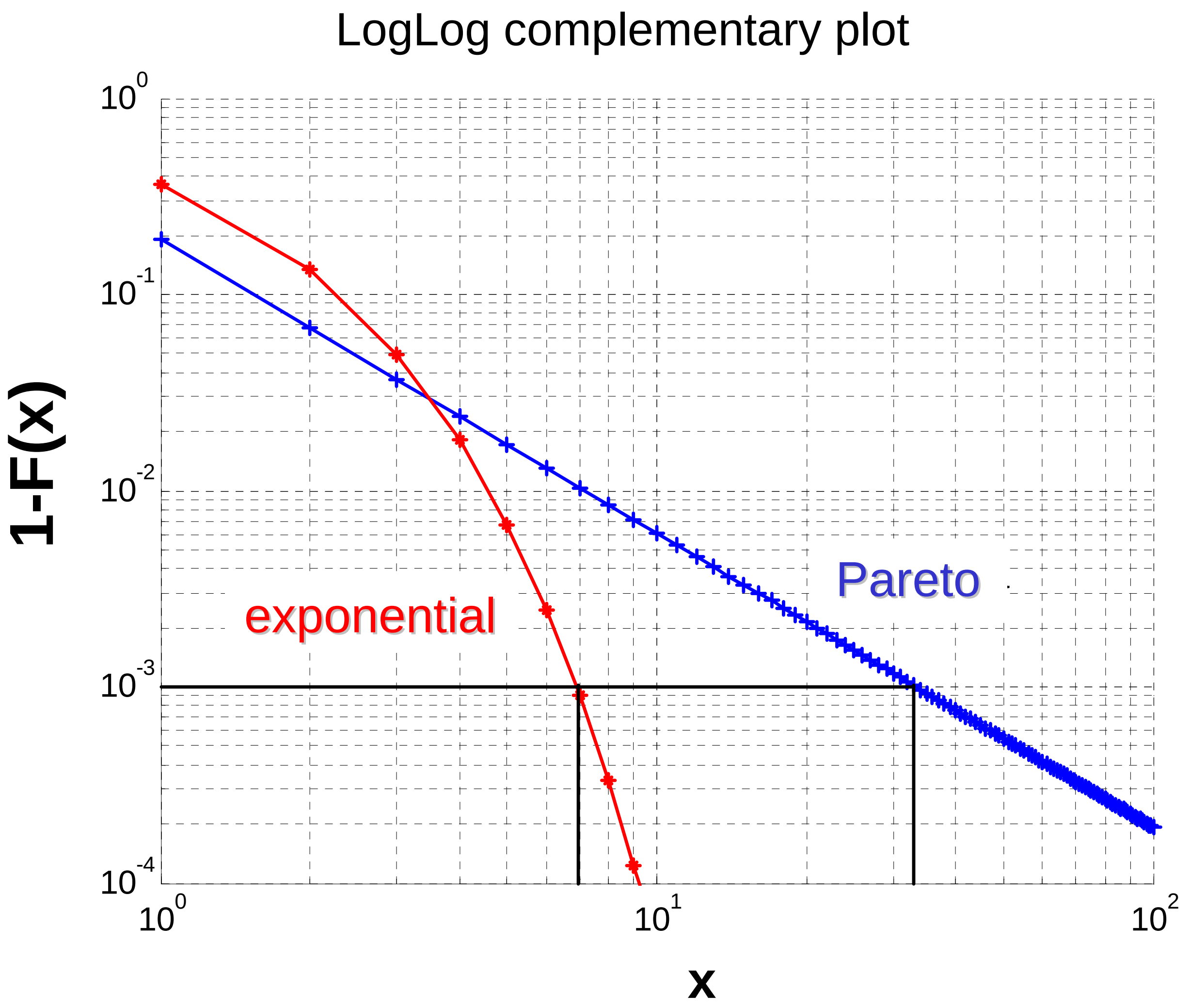
- high values are the most critical for performance (tail distribution is very important)

exp. distrib. $\overline{F(x)}=e^{-\lambda x}$

Pareto distrib. $\overline{F(x)}=\left(\frac{k}{x}\right)^{\alpha} \quad 0<\alpha\leq 2 \quad 0<k\leq x$

tail of the distributions, mean=1

esp. $\overline{F(x)} = e^{-\lambda x} = e^{-x}$ Pareto $\overline{F(x)} = \left(\frac{k}{x}\right)^{\alpha} = \left(\frac{1}{3x}\right)^{1.5}$



generation of a seq.of r.n. with Pareto distribution

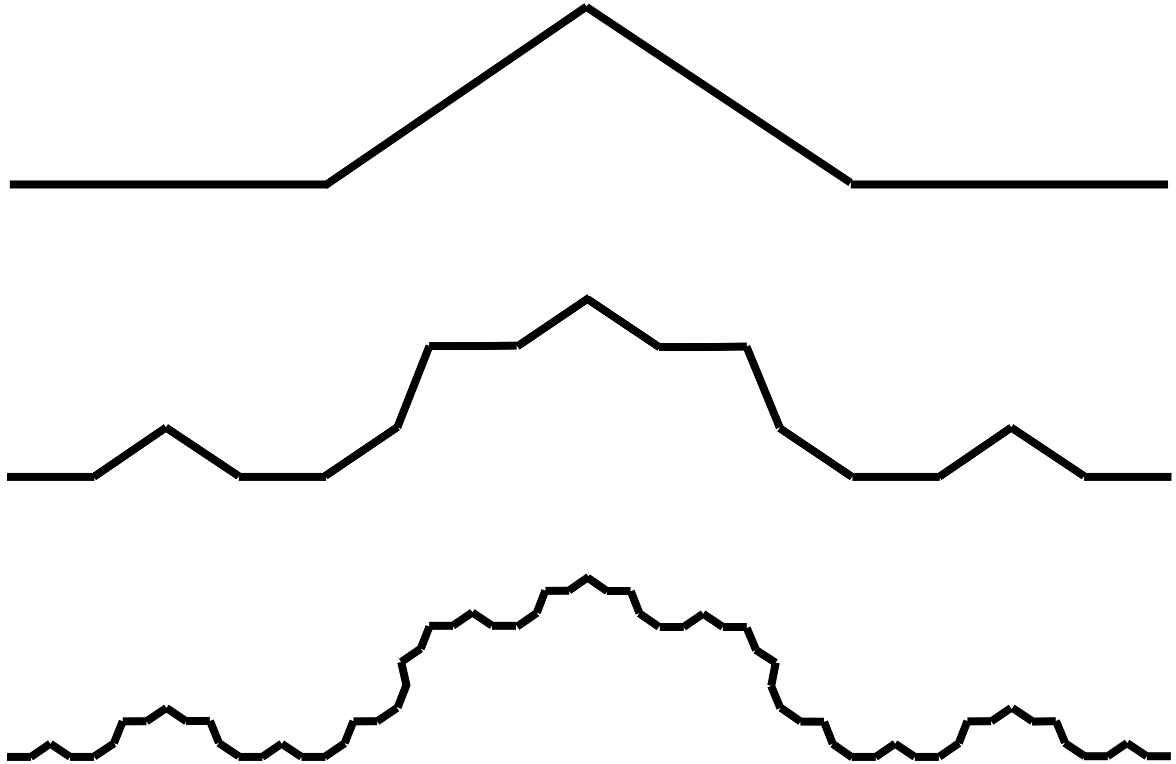
$$\mu = \begin{cases} k\dfrac{\alpha}{\alpha - 1} & \alpha > 1 \\ \infty & \alpha \leq 1 \end{cases} \qquad \sigma^2 = \begin{cases} k^2\dfrac{\alpha}{\alpha - 2} - \left(k\dfrac{\alpha}{\alpha - 1}\right)^2 & \alpha > 2 \\ \infty & \alpha \leq 2 \end{cases}$$

$$\alpha = \frac{\sigma + \sqrt{\sigma^2 + \mu^2}}{\sigma} \qquad k = \mu\,\frac{\alpha - 1}{\alpha} = \mu\,\frac{\sqrt{\sigma^2 + \mu^2}}{\sigma + \sqrt{\sigma^2 + \mu^2}}$$

$$y = 1 - k^{\alpha} x^{-\alpha}$$

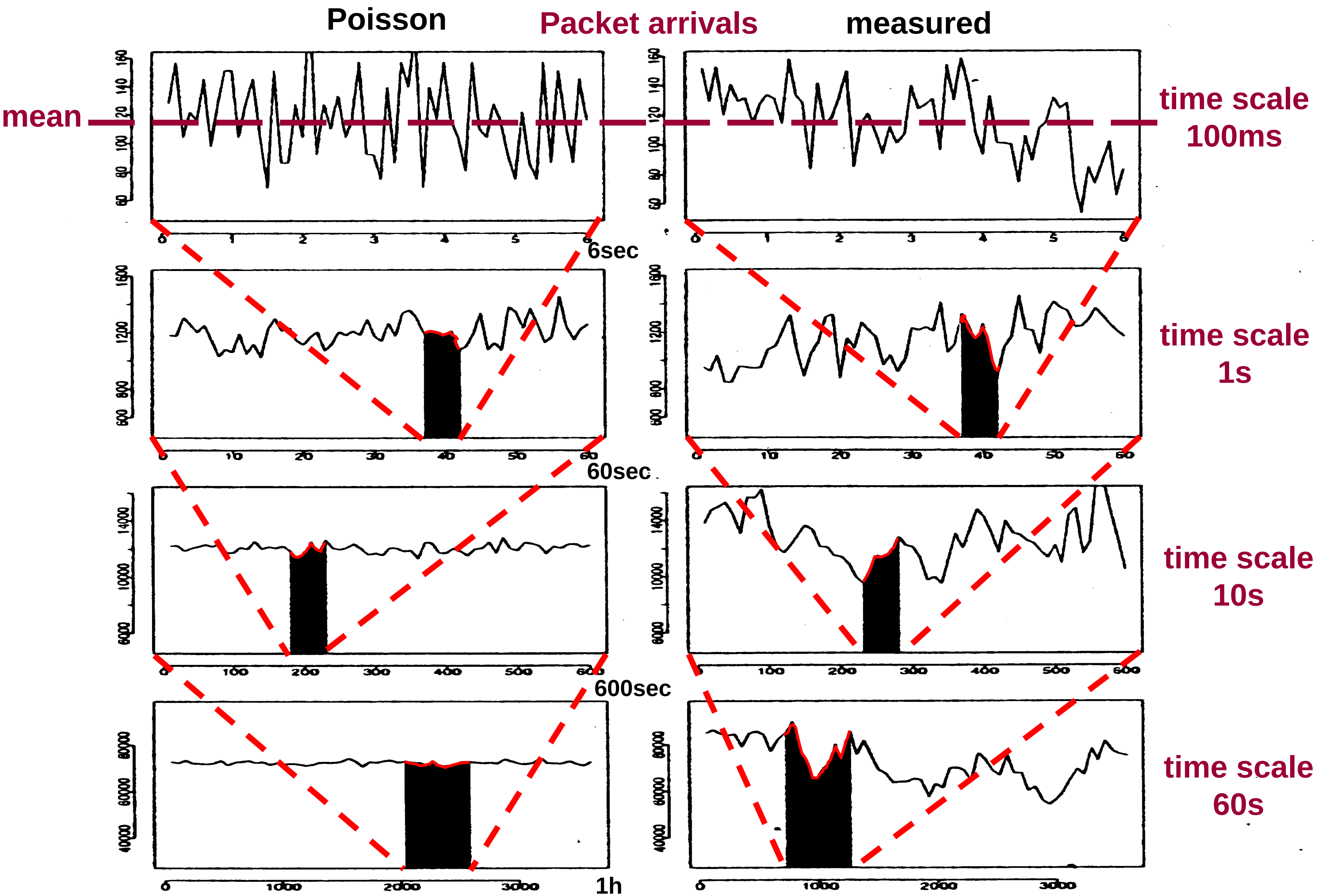
$$\begin{aligned} 1 - y &= k^{\alpha} x^{-\alpha} \\ y\,k^{-\alpha} &= x^{-\alpha} \\ \sqrt[\alpha]{y\,k^{-\alpha}} &= \sqrt[\alpha]{x^{-\alpha}} \\ x &= \frac{1}{\sqrt[\alpha]{y}}\;k \end{aligned}$$

fractals (self similarity)

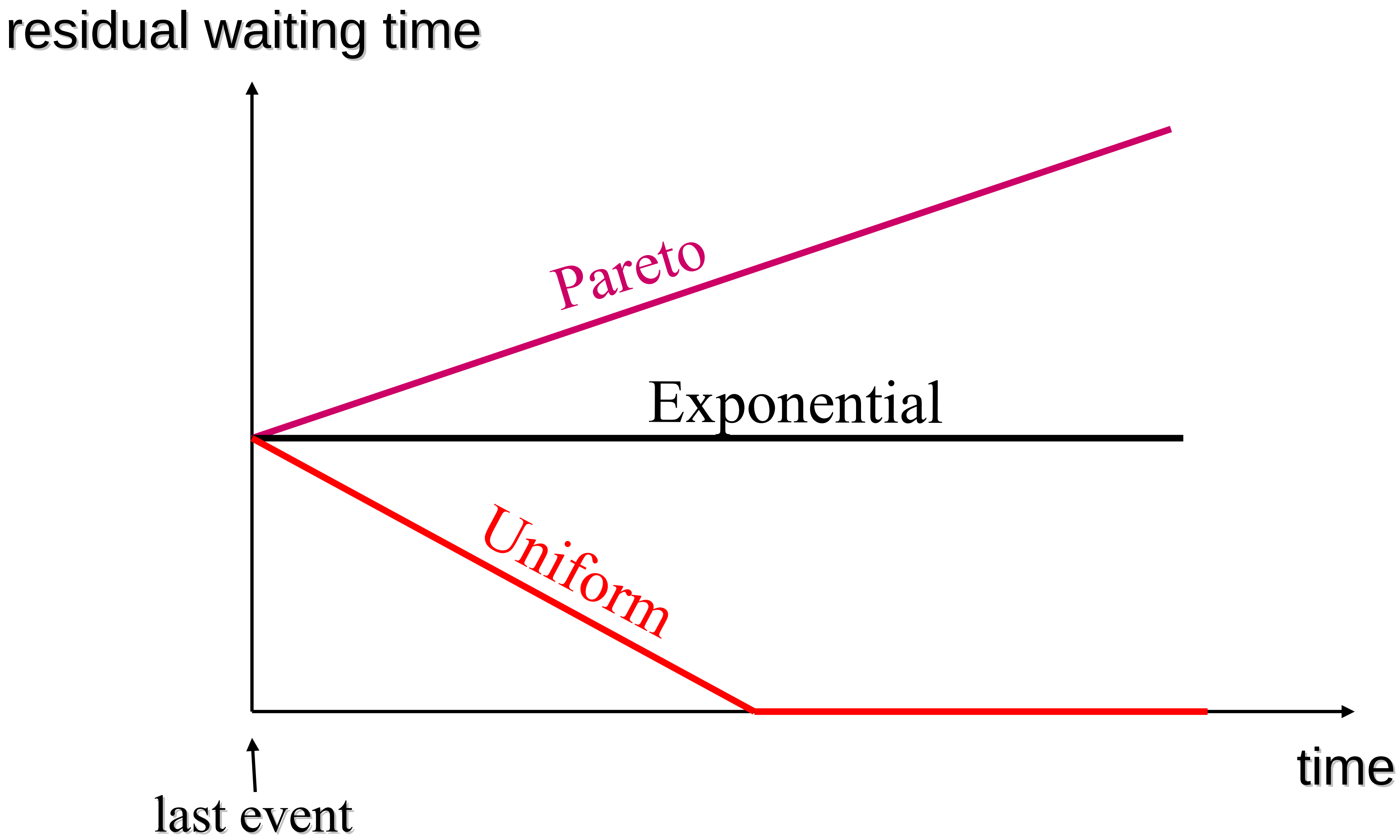


self-similarity of web traffic

from Willinger-Paxson, *Where Mathematics meets the Internet*, Notices of the American Mathematical Society, 45(8), pp.961-970, 1998

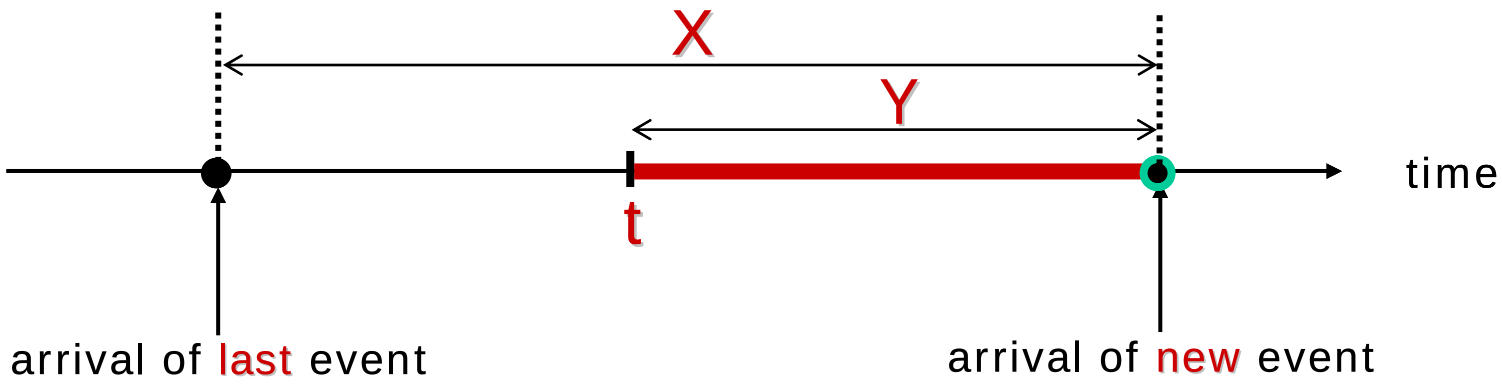


residual waiting time



residual waiting time Y

- X: time between two consecutive events
- given that *t* time units are elapsed since the last event occurred, we want to compute the distribution of Y, the residual waiting time
- Y = X - t



$$\phi(x) = \lim_{y \to 0^+} \frac{P[X \le t + y \mid X > t]}{y}$$

residual waiting time

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\phi(x) = \lim_{y \to 0^+} \frac{P[t < X \leq t + y]}{y} \frac{1}{P[X > t]} = \frac{f(x)}{1 - F(x)}$$

- **$G_Y (y|t)$:** conditional probability of the event $Y \leq y$ given that the event $X>t$ has occurred

$$G_Y(y|t) = P[Y \leq y \,|\, X > t] = P[X - t \leq y \,|\, X > t] =$$

$$P[X \leq t + y \,|\, X > t] = \frac{P[(X \leq t + y) \cap (X > t)]}{P[X > t]} = \frac{P[t < X \leq t + y]}{P[X > t]} =$$

$$\frac{P[t < X \leq t + y]}{1 - F(t)} = \frac{\int_t^{t+y} f(x)\, dx}{1 - F(t)}$$

residual waiting time

uniform distrib. $$T(t) = \frac{X_{\max} - t}{2}$$

exp. distrib. $$T(t) = \frac{1}{\lambda}$$

Pareto distrib. $$T(t) = \frac{t}{\alpha - 1}$$

residual wait. time: uniform and exp. distributions

$$\int_Y^{X_{\max}} (t-Y)\frac{f(t)}{1-F(Y)}\,dt = \frac{1}{1-F(Y)}\int_Y^{X_{\max}} (t-Y)\; f(t)\,dt =$$

$$\frac{1}{1-Y/X_{\max}}\int_0^{X_{\max}-Y} z\; f(Y+z)\;dz = \frac{X_{\max}}{X_{\max}-Y}\;\frac{1}{X_{\max}}\;\frac{(X_{\max}-z)^2}{2} = \frac{X_{\max}-z}{2}$$

$$\int_0^{\infty}(t-Y)\frac{f(t)}{1-F(Y)}\,dt = \frac{1}{1-F(Y)}\int_0^{\infty} z\; f(Y+z)\;dz =$$

$$\frac{1}{e^{-\lambda Y}}\int_0^{\infty} z\lambda e^{-\lambda Y}\, e^{-\lambda z}\;dz = \frac{e^{-\lambda Y}}{e^{-\lambda Y}}\int_0^{\infty} z\lambda\, e^{-\lambda z}\;dz = \frac{1}{\lambda}$$

residual waiting time: Pareto distrib.

$$\int_Y^\infty (t-Y)\frac{f(t)}{1-F(Y)}\,dt = \frac{1}{1-F(Y)}\int_0^\infty z\, f(Y+z)\,dz =$$

$$\frac{1}{(k/Y)^\alpha}\int_0^\infty z[\alpha k^\alpha (Y+z)^{-\alpha-1}]\,dz =$$

$$\frac{1}{(k/Y)^\alpha}\left[\int_0^\infty (Y+z)[\alpha k^\alpha (Y+z)^{-\alpha-1}]\,dz - Y\int_0^\infty \alpha k^\alpha (Y+z)^{-\alpha-1}\,dz\right] =$$

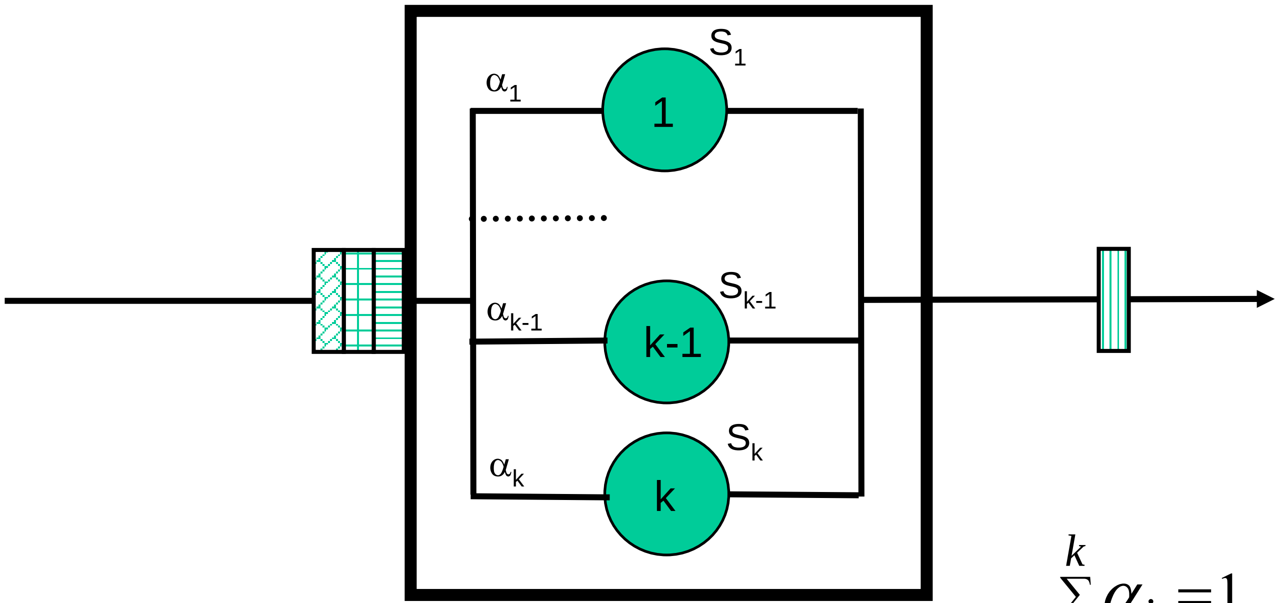
$$\frac{Y^\alpha}{k^\alpha}\alpha k^\alpha\left(\frac{Y^{-\alpha+1}}{\alpha-1} - Y\frac{Y^{-\alpha}}{\alpha}\right) = \alpha Y^\alpha\left(\frac{\alpha Y^{-\alpha+1} - Y^{-\alpha+1}(\alpha-1)}{\alpha(\alpha-1)}\right) = Y^\alpha\frac{Y^{-\alpha+1}}{(\alpha-1)} =$$

$$\frac{Y}{\alpha-1}$$

hyperexponential distribution

hyperexponential distribution

- a process consists of alternate phases that have exponential distributions
- during a single visit the process experiences one and only one of the many alternate phases



$$\sum_{i=1}^{k} \alpha_i = 1 \qquad 1 \le \alpha_i < 0$$

hyperexponential distribution

$$f(t) = \sum_{i=1}^{k} \alpha_i \ \lambda_i \ e^{-\lambda_i t} \qquad \sum_{i=1}^{k} \alpha_i = 1 \quad k = 1,2,\ldots \ \alpha_i, \lambda_i, t > 0$$

$$F(t) = P[T \le t] = \sum_{i=1}^{k} \alpha_i \ (1 - e^{-\lambda_i t})$$

hyperexponential distribution

$$mean \quad E[X] = \sum_{i=1}^{k} \alpha_i \, E[X_i] = \sum_{i=1}^{k} \alpha_i \, \frac{1}{\lambda_i}$$

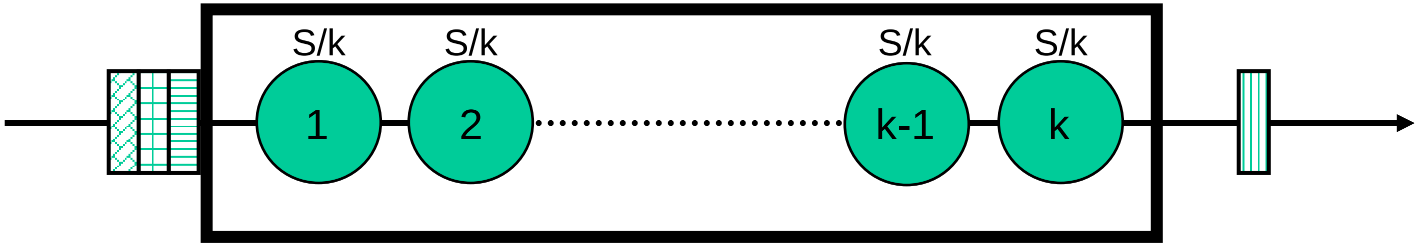
$$Var[X] = 2 \sum_{i=1}^{k} \frac{\alpha_i}{\lambda_i^2} - \left[\sum_{i=1}^{k} \frac{\alpha_i}{\lambda_i^2} \right]$$

$$coeff.\ of \ \ variation > 1$$

hypoexponential distribution (Erlang)

Erlang (hypoexponential) distribution

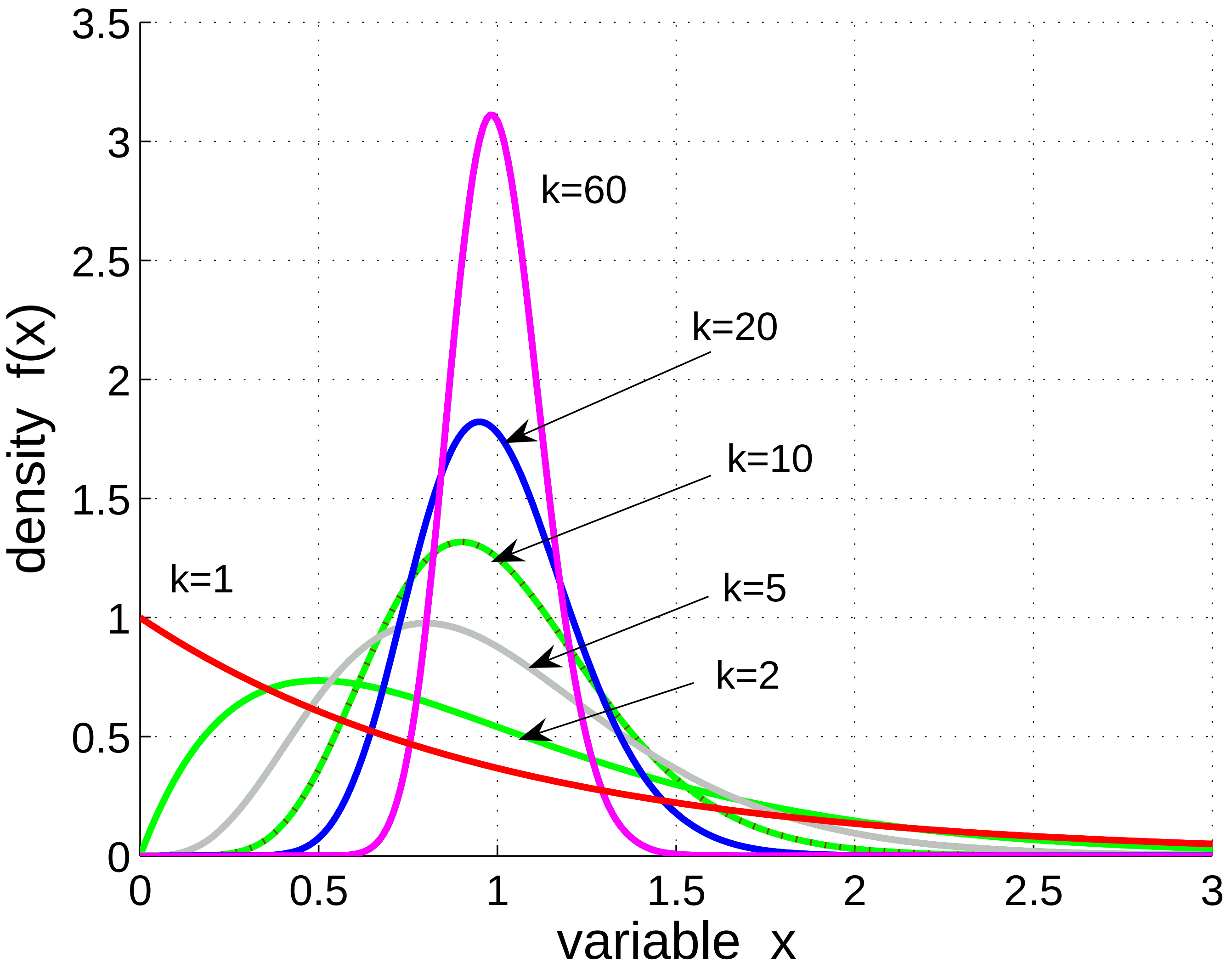
- a process with *K* sequential phases with identical exponential distributions (*k stage* Erlang)



$$f(t) = \lambda \; e^{-\lambda t} \; \frac{(\lambda t)^{k-1}}{(k-1)!} \qquad k = 1,2,3,\ldots \quad \lambda, t > 0$$

$$F(t) = P[T \le t] = 1 - e^{-\lambda t} \sum_{r=0}^{k-1} \frac{(\lambda t)^r}{r!}$$

example of Erlang distribution (mean=1)



Erlang (hypoexponential) distribution

$$mean \quad E[X] = \sum_{i=1}^{k} E[X_i] = \sum_{i=1}^{k} \frac{1}{\lambda_i} = k\frac{S}{k} = S$$

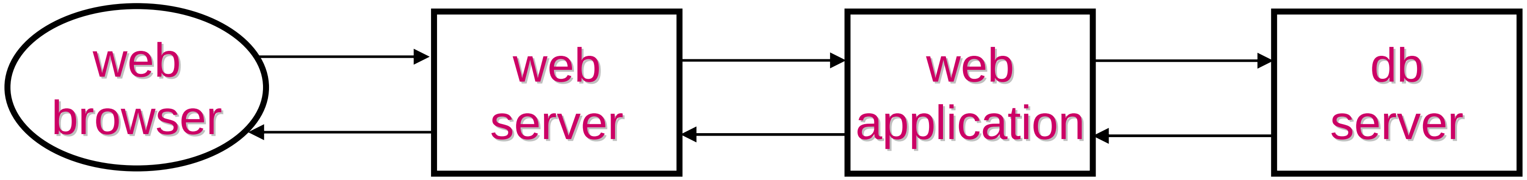
$$Var[X] = \sum_{i=1}^{k} Var[X_i] = \sum_{i=1}^{k} \frac{1}{\lambda_i^2} = k\left(\frac{S}{k}\right)^2 = \frac{S^2}{k}$$

$$coeff.of\ variation = \frac{\sqrt{S^2/k}}{S} = \frac{1}{\sqrt{k}} \quad < 1$$

as k increases the variance decreases

problem: execution time of a web application

- consider the global execution time T of a transaction of a web application on an intranet



- each one of the three software components (mutually independent) have an average execution time (response time) of 10 ms, exponentially distributed
- a complete execution of a command require the sequential execution of 5 sw components (web server, web appl., db server, web appl., web server)
- compute the probability that the complete execution time requires more than 60 ms, more than 90 ms

problem: execution time of a web application

- the distribution of the global execution time T is an Erlang-5

$$mean \quad E[T] = 50\ ms = \frac{1}{\lambda} \qquad \lambda = \frac{1}{50} = 0.02$$

$$Var[X] = \frac{E[T]^2}{k} = \frac{2500}{5} = 500\, ms^2$$

$$F(t) = 1 - e^{-t/10}\left[1 + \frac{t}{10} + \frac{1}{2}\left(\frac{t}{10}\right)^2 + \frac{1}{6}\left(\frac{t}{10}\right)^3 + \frac{1}{24}\left(\frac{t}{10}\right)^4\right]$$

$$P[T \le 60] = F(60) = 0.7149 \qquad P[T > 60] = 1 - F(60) = 0.2851$$

$$P[T \le 90] = F(90) = 0.945 \qquad P[T > 90] = 1 - F(90) = 0.055$$

fitting: Erlang-k distribution

- the payload of the messages of an application have the following distribution

lenght (m_i) crt	25	50	70	100	140
frequency (f_i)	0.4	0.3	0.1	0.15	0.05

$$mean \quad E[X] = \sum_{i=1}^{5} m_i \ f_i \ = 25\,x\,0.4 + 50\,x\,0.3 + ... = \ 54\,crt = \frac{1}{\lambda}$$

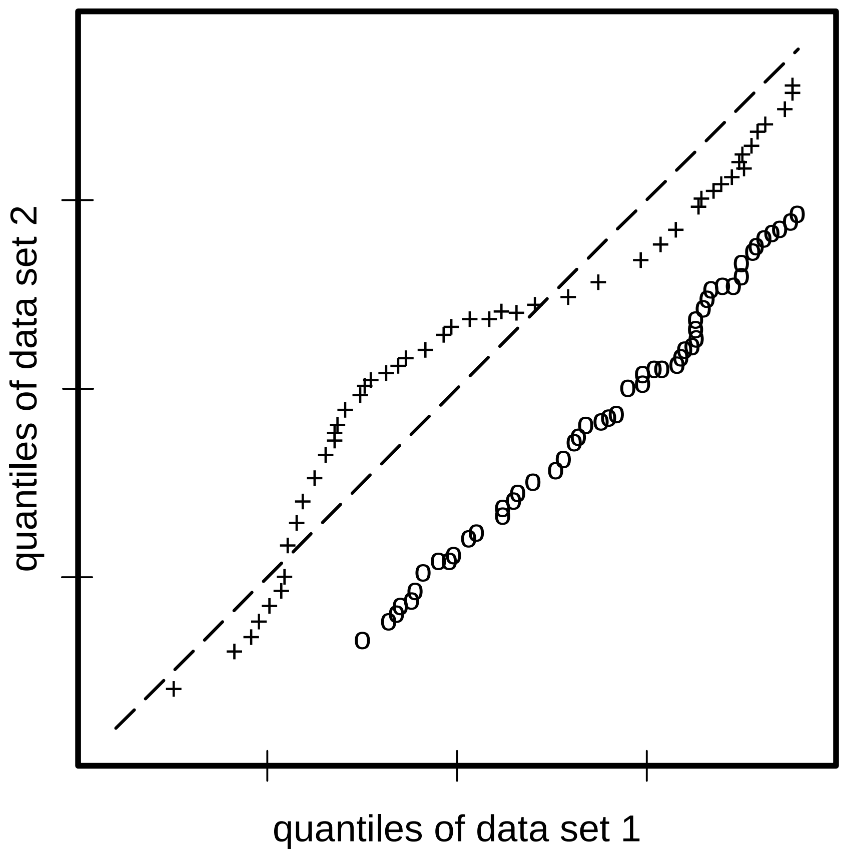
$$Var[X] = \sum_{i=1}^{5} (m_i - E[X])^2 \ = (25-54)^2 \ x\,0.4 + ... = 1054\,crt = \frac{E[X]^2}{k}$$

$$k = \frac{E[X]^2}{Var[X]} = 2.77 \quad \Rightarrow \quad Erlang\,2 \quad with \quad E[X] = \frac{1}{\lambda} = 54\,crt$$

$$Var[X] = \frac{1}{k\,\lambda^2} = \frac{54^2}{2} = 1458\,crt$$

fitting: Q-Q plot

graphical technique for determining if two data sets come from populations that have the same distribution



fitting: Q-Q plot

- provide insight into the nature of the difference between two samples
- the points of two data sets that come from two populations with the same distribution should fall approximatively along the reference line
- the sample sizes do not need to be equal
- probability plot: the values of one data set are replaced with the ones of a theoretical distribution
- e.g.: two data sets come from populations whose distributions differ by a shift in location, the points should lie along a straight line that is displaced up/down from the reference line