On generalisations of the log-Normal distribution by means of new product definition the Kapteyn process

Sílvio M. Duarte Queirós
Centro Brasileiro de Pesquisas Físicas
and
National Institute of Science and Technology for Complex Systems

Research carried out at: *Unilever R&D PortSunlight*

What's the Kapteyn process?

'[The log-Normal distribution] is one of the most important classes occurring in Nature.'

JC Kapteyn in "Skew Frequency Curves in Biology and Statistics" (1903)

K Pearson (1895)

$$\frac{p'(x)}{p(x)} + \frac{a + x - \lambda}{b_2(x-\lambda)^2 + b_1(x-\lambda) + b_0} = 0$$

The distributions can be interpreted as limiting forms of the Hyperbolic distribution.
Complete disinterest in possible natural causes.

The origin of the log-normal :

Pearson vs Galton and McAlister

Pearson vs Kapteyn

Theory of proporcionate effect

Let x *be a positive random variable that is the outcome of discrete random process.*

H Cramér (1923) :

> **If our random variable is the size of some specified organ that we are observing, the actual size of this organ in a particular individual may often be regarded as the joint effect of a large number of mutually independent causes, acting in an ordered sequence during the time of growth of the individual.**

$$x_t - x_{t-1} = \varepsilon_t \, \phi\left(x_{t-1}\right), \qquad \textit{Law of the proportionate effect}$$

$\phi\left(.\right)$ proportion function

$$\left\langle \varepsilon_t \, \varepsilon_{t'} \right\rangle \sim \delta_{t\,t'}$$

$$\left\langle \varepsilon_t \, x_{t'} \right\rangle = 0$$

$$\phi(x) = x$$

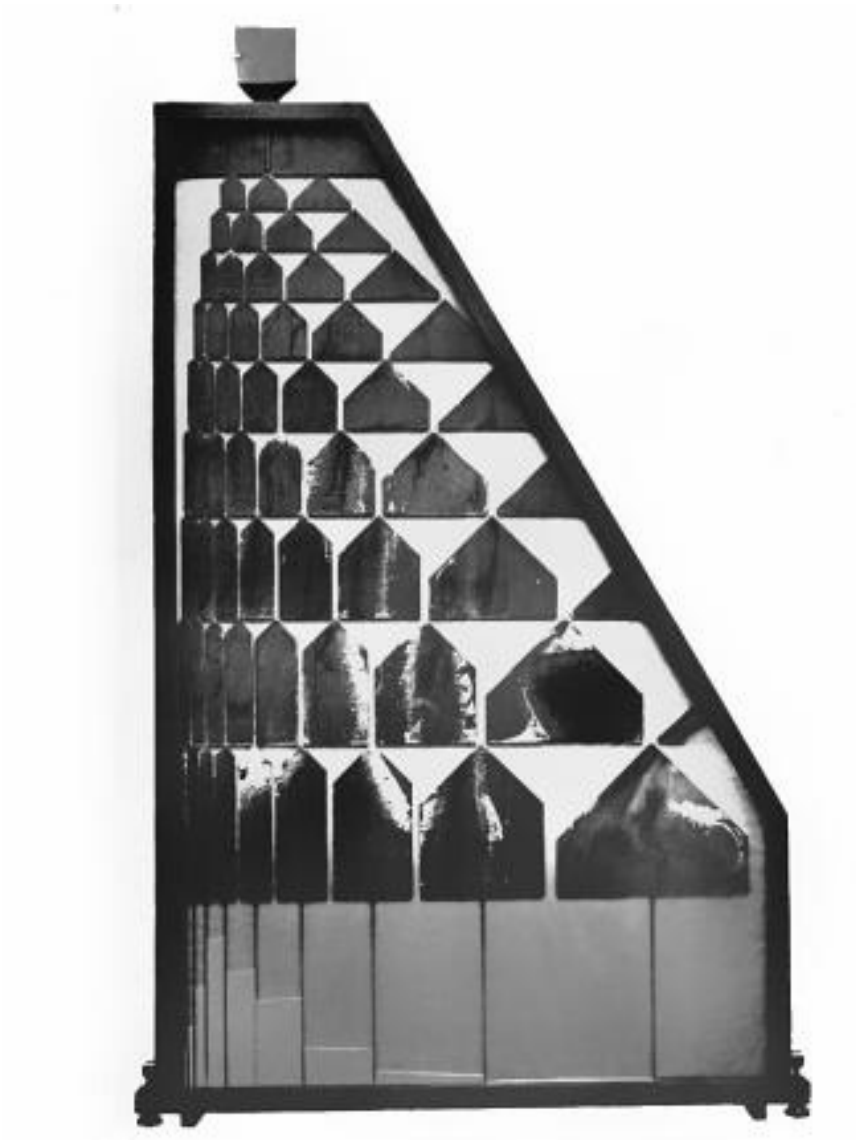
$$\frac{x_t - x_{t-1}}{x_{t-1}} = \varepsilon_t \qquad \Longrightarrow \qquad \sum_{t=1}^{N} \frac{x_t - x_{t-1}}{x_{t-1}} = \sum_{t=1}^{N} \varepsilon_t$$

$$\sum_{t=1}^{N} \frac{x_t - x_{t-1}}{x_{t-1}} \approx \int_{x_0}^{x_N} \frac{1}{x} dx = \ln x_N - \underset{(=0)}{\ln x_0} = \sum_{t=1}^{N} \varepsilon_t$$

$$x_N = \prod_{t=1}^{N} \varepsilon'_t$$

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

Kapteyn's analogue machine



Extending Kapteyn via q-algebra

L Nivanen (2003), EP Borges (2004)

$$\prod_{t=1}^{N} \varepsilon'_t \rightarrow \bigotimes_{q|t=1}^{N} \varepsilon'_t = \varepsilon'_1 \otimes_q \varepsilon'_2 \otimes_q \cdots \otimes_q \varepsilon'_N$$

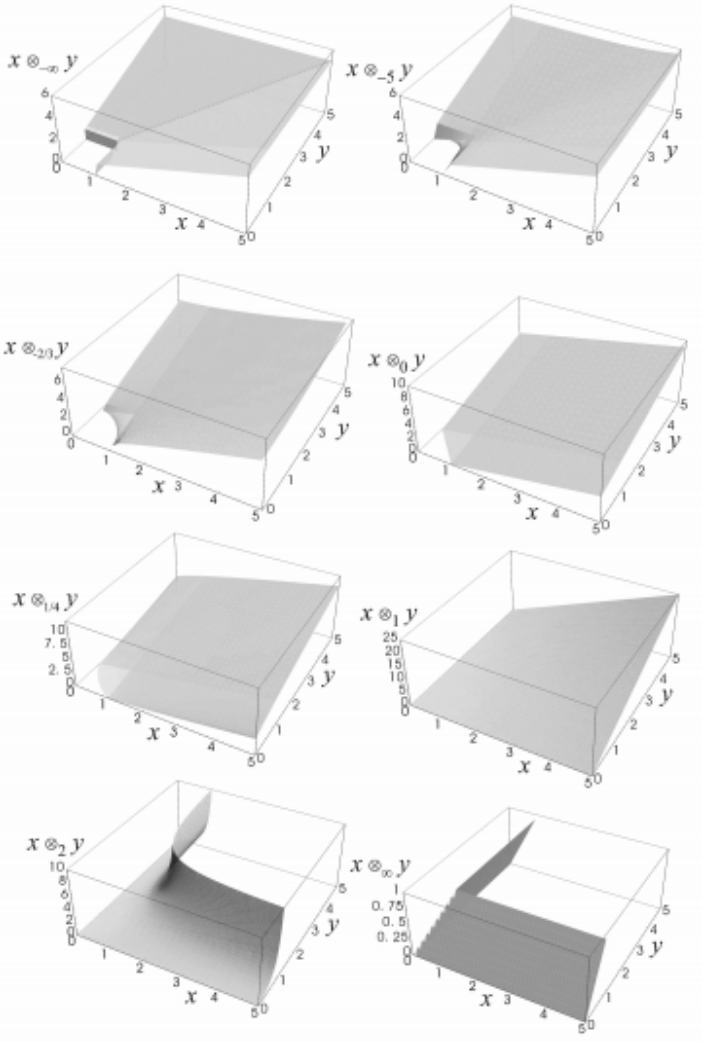
q-product

$$x \otimes_q y \equiv \exp_q \left[\ln_q x + \ln_q y\right]$$

$$\exp_q \equiv \left[1 + (1-q)\, x\right]^{\frac{1}{1-q}}, \qquad \ln_q x \equiv \frac{x^{1-q} - 1}{1-q}$$

q-properties and graphical representation

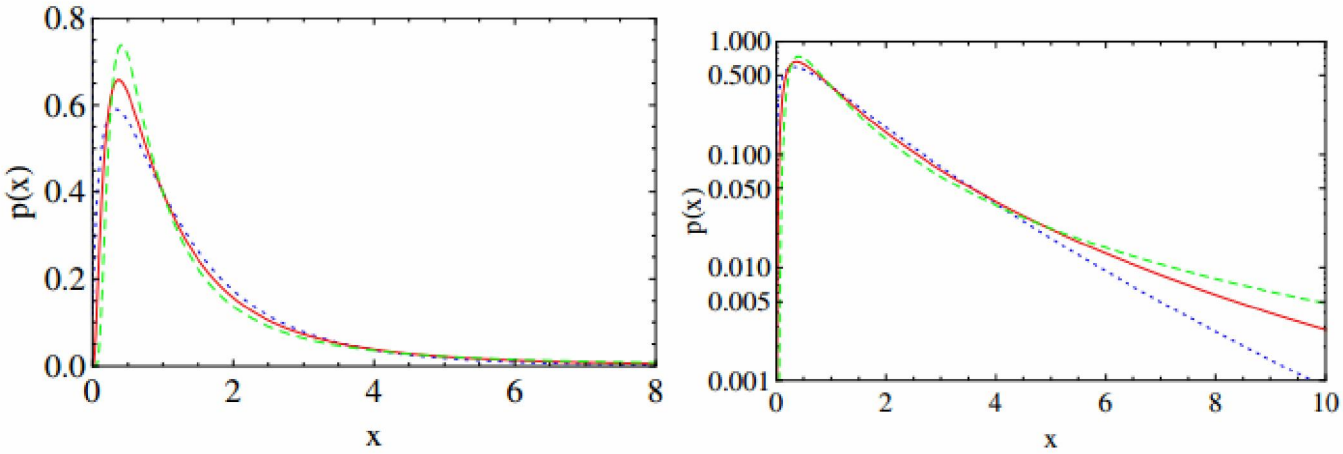
1. $x \otimes_1 y = x\ y$;
2. $x \otimes_q y = y \otimes_q x$;
3. $(x \otimes_q y) \otimes_q z = x \otimes_q (y \otimes_q z) = [x^{1-q} + y^{1-q} - 2]^{\frac{1}{1-q}}$;
4. $(x \otimes_q 1) = x$;
5. $\ln_q [x \otimes_q y] \equiv \ln_q x + \ln_q y$;
6. $\ln_q (x\ y) = \ln_q (x) + \ln_q (y) + (1-q) \ln_q (x) \ln_q (y)$;
7. $(x \otimes_q y)^{-1} = x^{-1} \otimes_{2-q} y^{-1}$;
8. $(x \otimes_q 0) = \begin{cases} 0 & \text{if } (q \geq 1 \text{ and } x \geq 0) \text{ or if } (q < 1 \text{ and } 0 \leq x \leq 1) \\ (x^{1-q} - 1)^{\frac{1}{1-q}} & \text{otherwise} \end{cases}$



From modifying the multiplicand in the proportionate law the distribution reads,

$$p_q(x) = \frac{1}{\mathcal{Z}_q\, x^q} \exp\left[-\frac{(\ln_q x - \mu)^2}{2\sigma^2}\right]$$

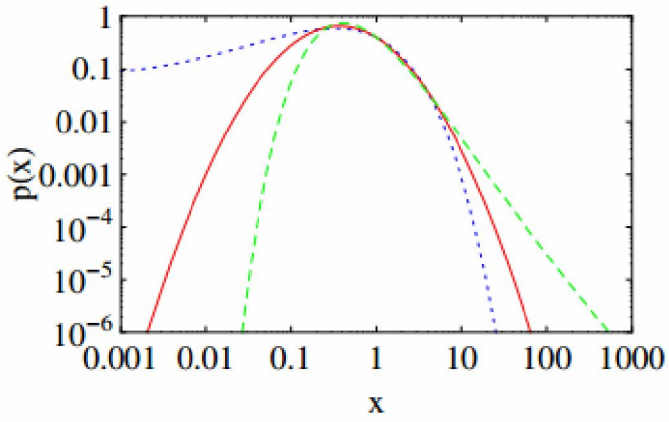
$$\mathcal{Z}_q = \begin{cases} \sqrt{\frac{\pi}{2}}\,\mathrm{erfc}\left[-\frac{1}{\sqrt{2}\sigma}\left(\frac{1}{1-q}+\mu\right)\right]\sigma & if \quad q<1 \\ \sqrt{\frac{\pi}{2}}\,\mathrm{erfc}\left[\frac{1}{\sqrt{2}\sigma}\left(\frac{1}{1-q}+\mu\right)\right]\sigma & if \quad q>1. \end{cases}$$



$q = 4 / 5$

$q = 1$

$q = 5 / 4$



Some probabilistic properties

$q > 1$:

$$\lim_{x\to\infty} \exp\left[-\frac{(\ln_q x - \mu)^2}{2\sigma^2}\right] = \exp\left[-\frac{\gamma^2}{2}\right]$$

The vanishing for large x *is guaranteed by the factor* x^{-q} *in the distribution*

$$\mathcal{P}(x) \equiv \int_0^x p(z)\, dz,$$

$$\mathcal{P}_{q>1}(x) = \frac{1 + \mathrm{erf}\left[\frac{\ln_q(x)-\mu}{\sqrt{2}\sigma}\right]}{1 + \mathrm{erf}\left[\frac{\frac{1}{q-1}-\mu}{\sqrt{2}\sigma}\right]},$$

$$\mathcal{P}_{q<1}(x) = \frac{\mathrm{erf}\left[\frac{\ln_q(x)-\mu}{\sqrt{2}\sigma}\right] - \mathrm{erf}\left[-\frac{1}{\sqrt{2}\sigma}\left(\frac{1}{1-q}+\mu\right)\right]}{1 + \mathrm{erfc}\left[-\frac{1}{\sqrt{2}\sigma}\left(\frac{1}{1-q}+\mu\right)\right]},$$

Some statistical properties

$q < 1$

$$\langle x^n \rangle = \frac{\Gamma[\nu] \exp\left[-\frac{\gamma^2}{8\beta}\right] D_{-\nu}\left[\frac{\gamma}{\sqrt{2\beta}}\right]}{\sqrt{\beta^\nu \pi}\,\sigma\,(1-q)\,\mathrm{erfc}\left[-\frac{1}{\sqrt{2}\sigma}\left(\frac{1}{1-q}+\mu\right)\right]},$$

$$\beta = \frac{1}{2\sigma^2(1-q)^2}; \qquad \gamma = -\frac{1+\mu(1-q)}{\sigma^2(1-q)^2}; \qquad \nu = 1 + \frac{n}{1-q}$$

$q > 1$

$$\frac{1}{\sqrt{2}\sigma}\left(\frac{1}{1-q}+\mu\right)$$

Usefulness

Dynamical mechanism leading to the truncated Gaussian

$$\mathcal{G}_b(y) = \sqrt{\frac{2}{\pi\,\sigma}}\,\mathrm{erfc}\left[\mathrm{sgn}(b)\,\frac{1}{\sqrt{2}\sigma}\,(\mu - b)\right]^{-1}\exp\left[-\frac{(y-\mu)^2}{2\,\sigma^2}\right]$$

Picking the distribution I've introduced in the talk and assuming ,

$$y = \ln_q x$$

It's possible to establish a relation between both distributions and define,

$$b = \frac{1}{q-1}$$

But what about something ACTUALLY useful ?

Flux networks in metabolic graphs

IOP Publishing — Physical Biology

Phys. Biol. **6** (2009) 046006 (9pp) — doi:10.1088/1478-3975/6/4/046006

Flux networks in metabolic graphs

P B Warren, S M Duarte Queiros and J L Jones

Unilever R&D Port Sunlight, Bebington, Wirral, CH63 3JW, UK

E-mail: patrick.warren@unilever.com

Received 9 February 2009
Accepted for publication 25 August 2009
Published 22 September 2009
Online at stacks.iop.org/PhysBio/6/046006

Molecular Systems Biology 9; Article number 661; doi:10.1038/msb.2013.18
Citation: *Molecular Systems Biology* 9:661
www.molecularsystemsbiology.com

OPEN ACCESS

molecular systems biology

REVIEW

Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*

Douglas McCloskey[1], Bernhard Ø Palsson[1,2] and Adam M Feist[1,2,*]

[1] Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA and
[2] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark
* Corresponding author. Dr A M Feist, Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0412, USA. Tel.: +1 858 822 3181; Fax: +1 858 822 3120; E-mail: afeist@ucsd.edu

translation into computational models that can be used to calculate metabolic phenotypes (Palsson, 2009; Pfau *et al*, 2011; Lewis *et al*, 2012). In addition, other omics data types that have been generated can be interpreted in the context of a reconstruction and computational model to analyze cellular functions under specific conditions. Taken together, this information becomes a *de facto* knowledge base. Genome-scale models (GEMs) are a structured format of such a knowledge base that can be used to perform computational and quantitative queries to answer various questions about the

Metabolic network: *Set of chemical reactions and metabolites that relate one another by means of a system of equations:*

$$\frac{dc_i}{dt} = \sum_{\alpha} S_{i\alpha} v_{\alpha}$$

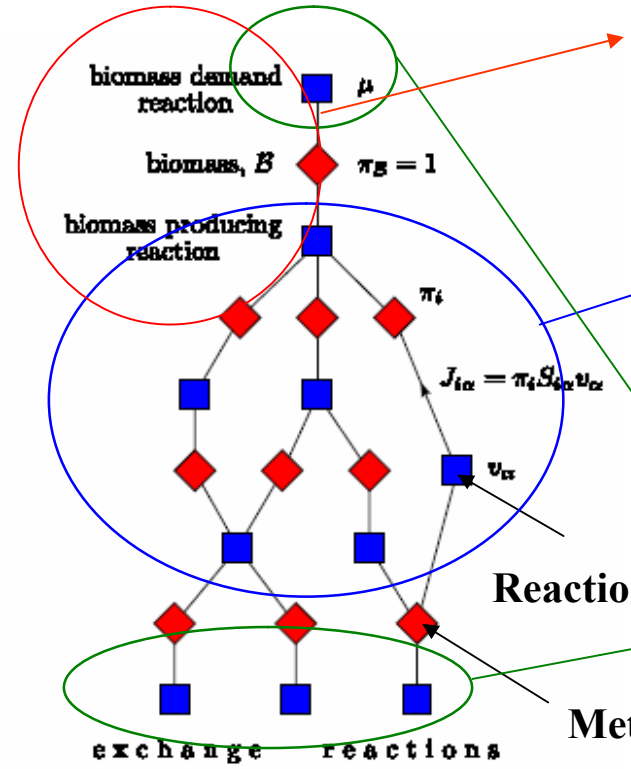
c_i *concentration of metabolite i;*

v_α *flux in reaction* α*;*

$S_{i\alpha}$ *stecheometric matrix.*

This system quickly attains a stationary state.

A metabolic network can be represented by a bi-partite graph as well,



Consumes the end-points of metabolism *e.g.* amino-acids, nucleotides, lipids and co-factors.

Internal reactions representing bio-chemical changes and transport processes throught the membrane.

Non-balanced reactions introduced within the context of constraint-based modelling.

Linear programming approach

In this problem the main variable is the set of the chemical reactions fluxes v_α *which maximise the growth rate* μ.

As any linear programming problem one can have a dual approach. Here, I'll focus on the primal

$$Z = \mu + \sum_{i\alpha} \pi_i S_{i\alpha} v_\alpha + \sum_{\alpha}^{'} y_\alpha \left(v_\alpha - u_\alpha^2\right) + \sum_{\alpha}^{''} y_\alpha \left(v_\alpha + v_i^{\max} - u_\alpha^2\right)$$

$$v_\alpha \geqslant 0 \Rightarrow v_\alpha = u_\alpha^2,$$

$$v_\alpha \geqslant -v_i^{\max} \Rightarrow v_\alpha + v_i^{\max} = u_\alpha^2.$$

Objective function

Lagrange multiplier associated with the stationary state condition.

In blending the reaction fluxes with the properties of the metabolites we can interpret the edges as fluxes, $\boldsymbol{J_{i\alpha}}$ *,that are conserved at every node (either reaction or metabolite).*

MOREOVER,

from the dual approach we proved that $\boldsymbol{\pi_i}$ ***is the chemical potential of the metabolite i.***

Organism	Model	Substrate	Molar yield	M_W	Mass yield
E. coli	iAF1260	D-glucose	96.3 gDW/mol	180 g/mol	0.535 gDW/g
— ,, —	— ,, —	D-malate	42.6	132	0.323
— ,, —	— ,, —	succinate	49.0	116	0.423
— ,, —	— ,, —	acetate	25.0	59	0.423
— ,, —	— ,, —	D-glucose (anaerobic)	31.1	180	0.173
— ,, —	iJR904	D-glucose	95.7	180	0.532
S. cerevisiae	iND750	D-glucose	97.3	180	0.541
M. barkeri	iAF692	H_2	4.45	2	2.23

2.7. Statistical analysis

We undertook statistical analysis of the shadow price distributions for selected conditions and organisms although the results are somewhat inconclusive. We attempted to fit the observed distributions, using maximum likelihood estimators, to a log-Normal, a χ distribution, an inverted χ distribution and a general distribution $\sim \exp\left[-\left(\pi_i^2 + \omega\beta\right)/\omega\pi_i\right]$ which has tunable exponential asymptotic behaviour for large and small π_i (ω and β are parameters). However, none of these distributions could be said to fit the observed distributions, as judged by the Kolmogorov–Smirnov test [30]. Aside from

[PB Warren, SMDQ, JL Jones (2009)]

Why not check the π distribution with the extended log-Normal?

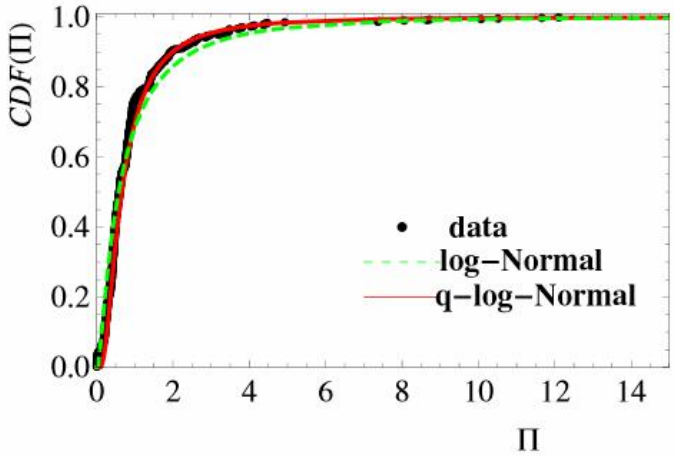


FIG. 4: Cumulative density function of the shadow prices vs shadow price of the metabolic network of the E. coli (iJR 904) growing on a D-glucose substrate. The symbols are obtained from the data and the lines the best fits with the q-log-Normal distribution and log-Normal. The values of the parameters and error are mentioned in the text.

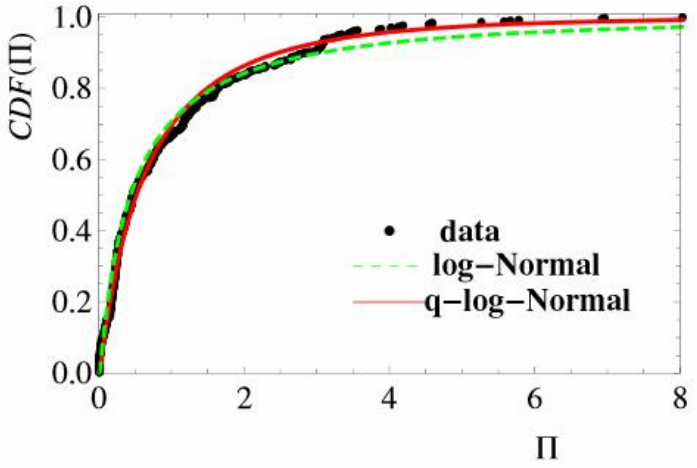


FIG. 5: Cumulative density function of the shadow prices vs shadow price of the metabolic network of the M. barkeri (iAF 692 model) growing in Hydrogen medium [Feist, 2006]. The symbols are obtained from the data and the lines the best fits with the q-log-Normal distribution and log-Normal. The values of the parameters and error are mentioned in the text.

[SMDQ (2012)]

"bah... Of course the new fit is better, you're adding a parameter!!!"

Not quite like that....

Better, but how much better? The Akeike Information Criterion

$$AIC = 2\,k + n \ln\left[\frac{RSS}{n}\right],$$

The AIC for the extended log-Normal based on the q-product fits are systematically smaller than the AIC for the log-Normal.

Interestingly,

Anaerobic environment → $q < 1$

Aerobic environment → $q > 1$

Take-home messages

- *Assuming the q-product in the Kapteyn multiplicative process it's possible to extend the log-Normal distribution;*

- *This new distribution either favours small or large value asymptotics depending on the value of the multiplicative parameter q;*

- *This new distribution describes the optimising distribution of chemical potentials of metabolites in bacterial growth with statistical significance and does it efficiently;*

- *From the q value (greater/less than 1) one is apparently capable of checking whether the growth is aerobic or anaerobic.*

Thanks.