The normal distribution is the log-normal distribution

Werner Stahel, Seminar für Statistik, ETH Zürich

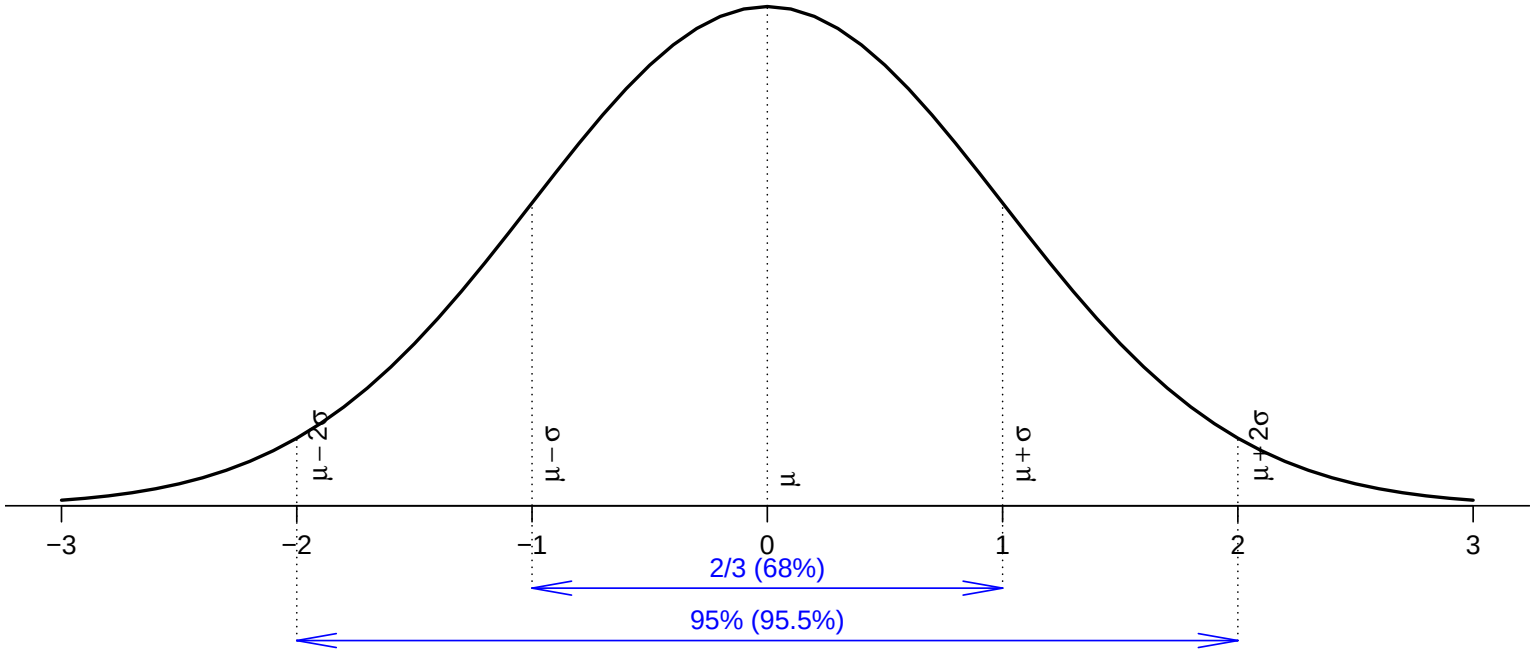
and Eckhard Limpert

2 December 2014

The normal Normal distribution

We like it!

- Nice shape.
- Named after Gauss. Decorated the 10 DM bill.
- We know it. Passed the exam.



Why it is right.

It is given by mathematical theory.

- Adding normal random variables gives a normal sum.
- Linear combinations $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots$ remain normal.
- $\longrightarrow$ Means of normal variables are normally distributed.
- Central Limit Theorem: Means of non-normal variables are approximately normally distributed.
- $\longrightarrow$ "Hypothesis of Elementary Errors": If random variation is the sum of many small random effects, a normal distribution must be the result.
- Regression models assume normally distributed errors.

Is it right?

Mathematical statisticians believe(d) that it is prevalent in Nature.

Well, it is not. Purpose of this talk: What are the consequences?

1. Empirical Distributions
2. Laws of Nature
3. Logarithmic Transformation, the Log-Normal Distribution
4. Regression
5. Advantages of using the log-normal distribution
6. Conclusions

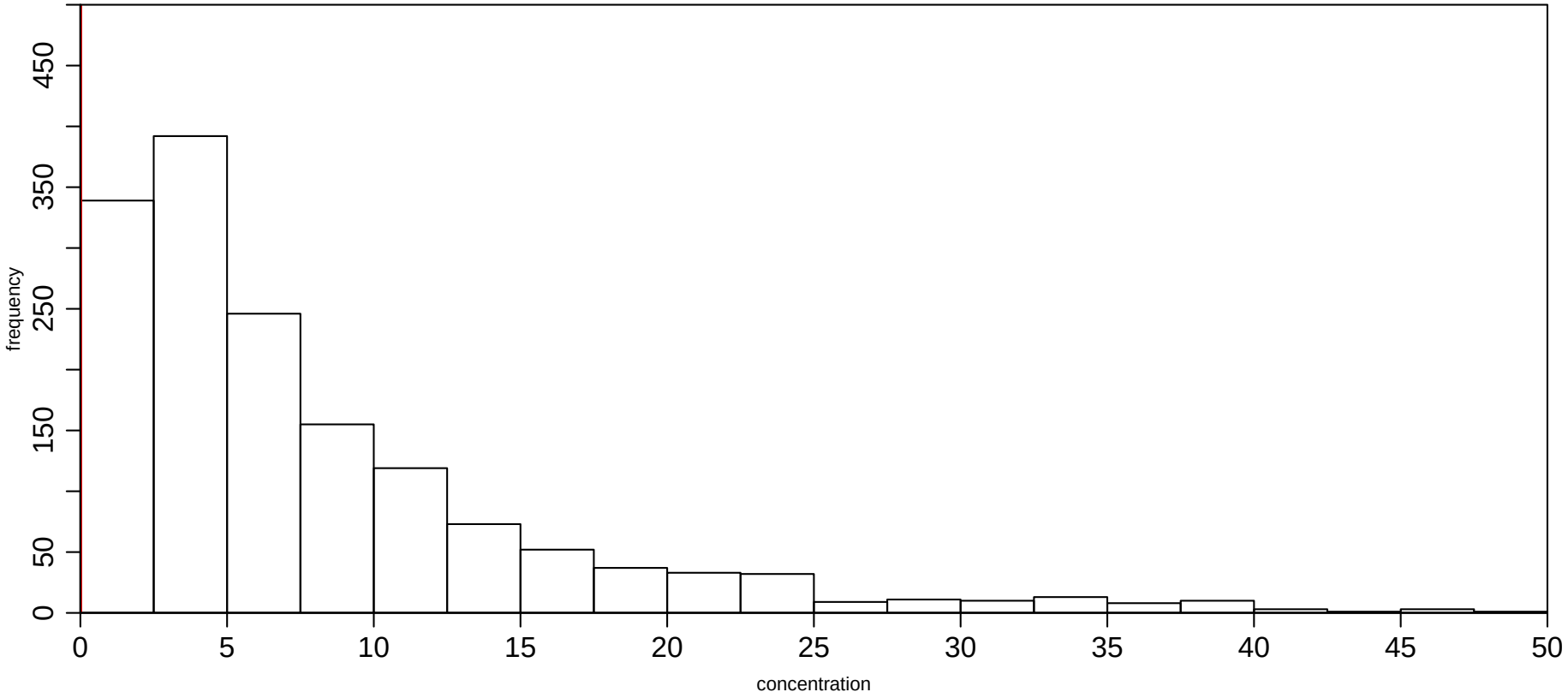
1. Empirical Distributions

Measurements:

size, weight, concentration, intensity, duration, price, activity

All $> 0 \longrightarrow$ "amounts" (John Tukey)

Example: HydroxyMethylFurfurol (HMF) in honey (Renner 1970)



Measurements:

size, weight, concentration, intensity, duration, price, activity

All > 0 $\longrightarrow$ "amounts"

Distribution is skewed: left steep, right flat, skewness > 0

unless coefficient of variation $\text{cv}(X) = \text{sd}(X)/E(X)$ is small.

Other variables may have other ranges and negative skewness.

They may have a normal distribution.

They are usually derived variables, not original measurements.

Any examples?

Our examples: Position in space and time, angles, directions. That's it!

For some, 0 is a probable value: rain, expenditure for certain goods, ...

pH, sound and other energies [dB] $\longrightarrow$ log scale!

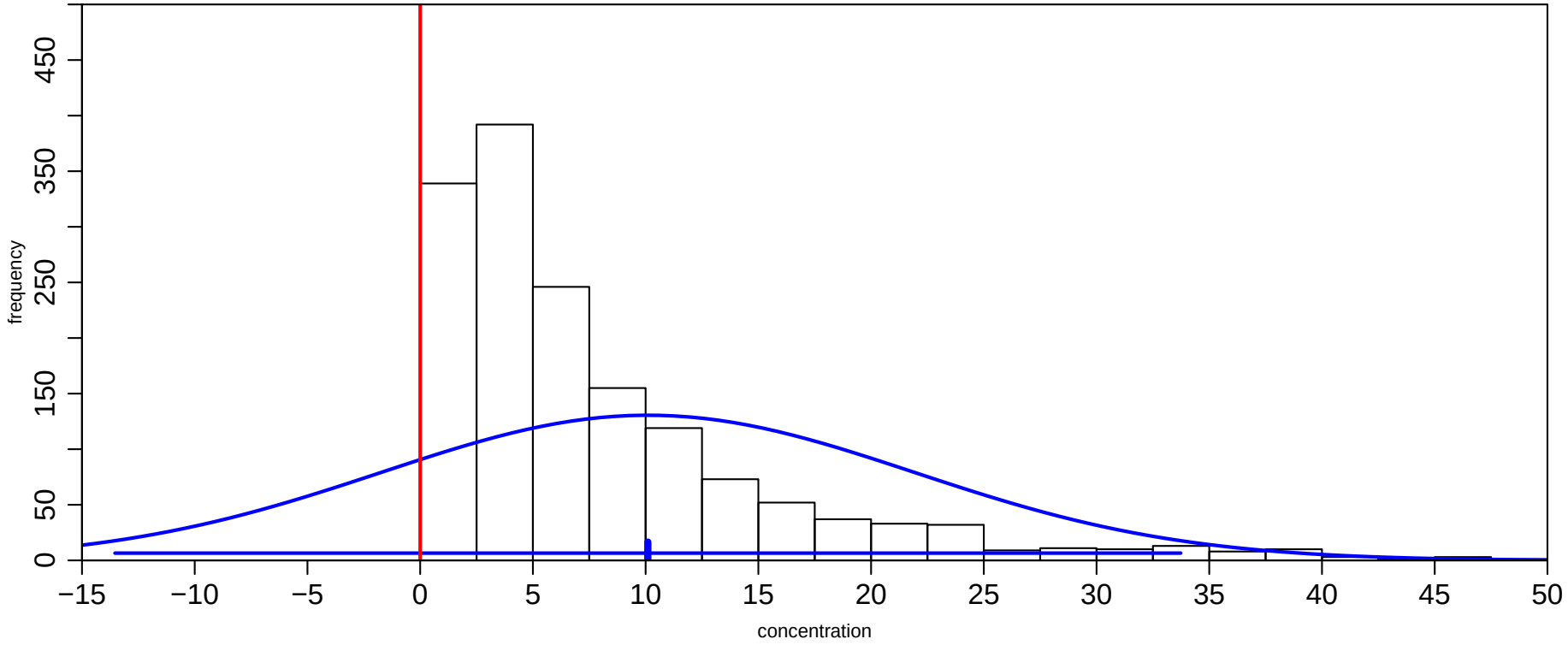
The 95% Range Check

For every normal distribution, negative values have a probability > 0.

$\longrightarrow$ normal distribution inadequate for positive variables.

Becomes relevant when 95% range $\bar{x} \pm 2\widehat{\sigma}$ reaches below 0.

Then, the distribution is noticeably skewed.



2. Laws of Nature

(a) Physics $E = m \cdot c^2$

Stopping distance $s = \cdot v^2/(2 \cdot a)$; Velocity $v = F \cdot t/m$

Gravitation $F = G \cdot m_1 \cdot m_2/r^2$

Gas laws $p \cdot V = n \cdot R \cdot T$; $R = p_0 \cdot V_0/T_0$

Radioactive decay $N_t = N_0 \cdot e^{-kt}$

(b) Chemistry

Reaction velocity $v = k \cdot [A]^{n_A} \cdot [B]^{n_B}$

change with changing temperature $\Delta t \rightarrow +10^0 C \Longrightarrow v \rightarrow \cdot 2$

based on Arrhenius' law $k = A \cdot e^{-E_A/R \cdot T}$

$E_A =$ activation energy; $R =$ gas constant

Law of mass action: $A + B \leftrightarrow C + D : K_c = [A] \cdot [B]/[C] \cdot [D]$

(c) Biology

Multiplication (of unicellular organisms) $1 - 2 - 4 - 8 - 16$

Growth, size $s_t = s_0 \cdot k^t$

Hagen-Poiseuille Law; Volume:

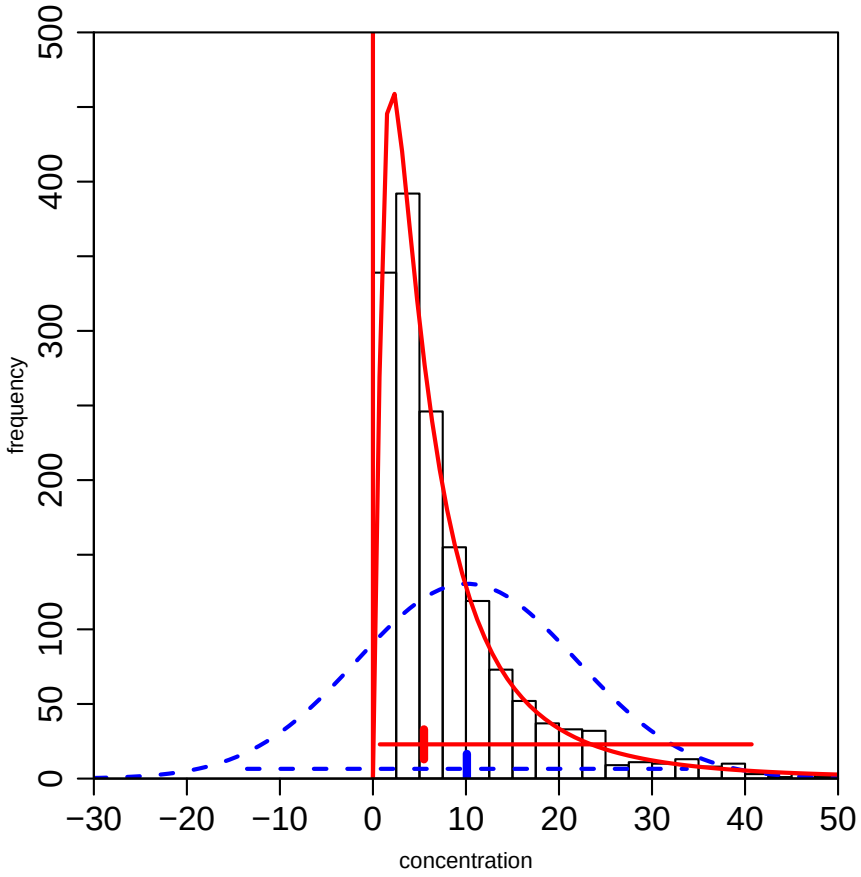
$V_t = (\Delta P \cdot r^4 \cdot \pi)/(8 \cdot \eta \cdot L)$; ΔP : pressure difference

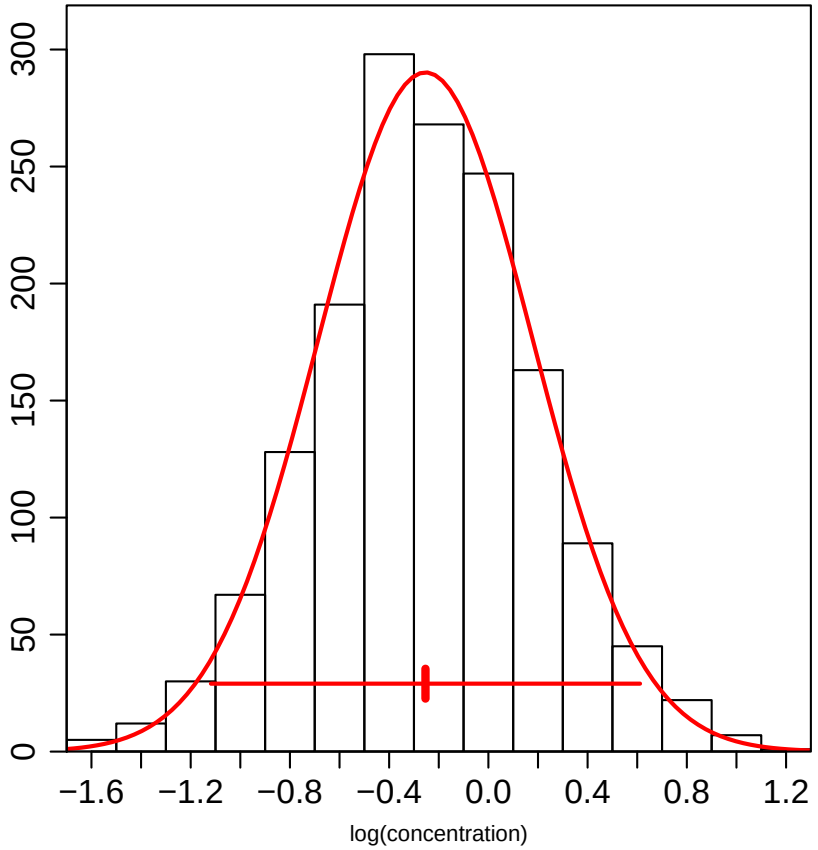
Permeability

Other laws in biology?

3. Logarithmic Transformation, Log-Normal Distribution

Transform data by log transformation





The log transform $Z = \log(X)$

- turns multiplication into addition,
- turns variables $X > 0$ into Z with unrestricted values,
- reduces (positive) skewness (may turn it negatively skewed)
- Often turns skewed distributions into normal ones.

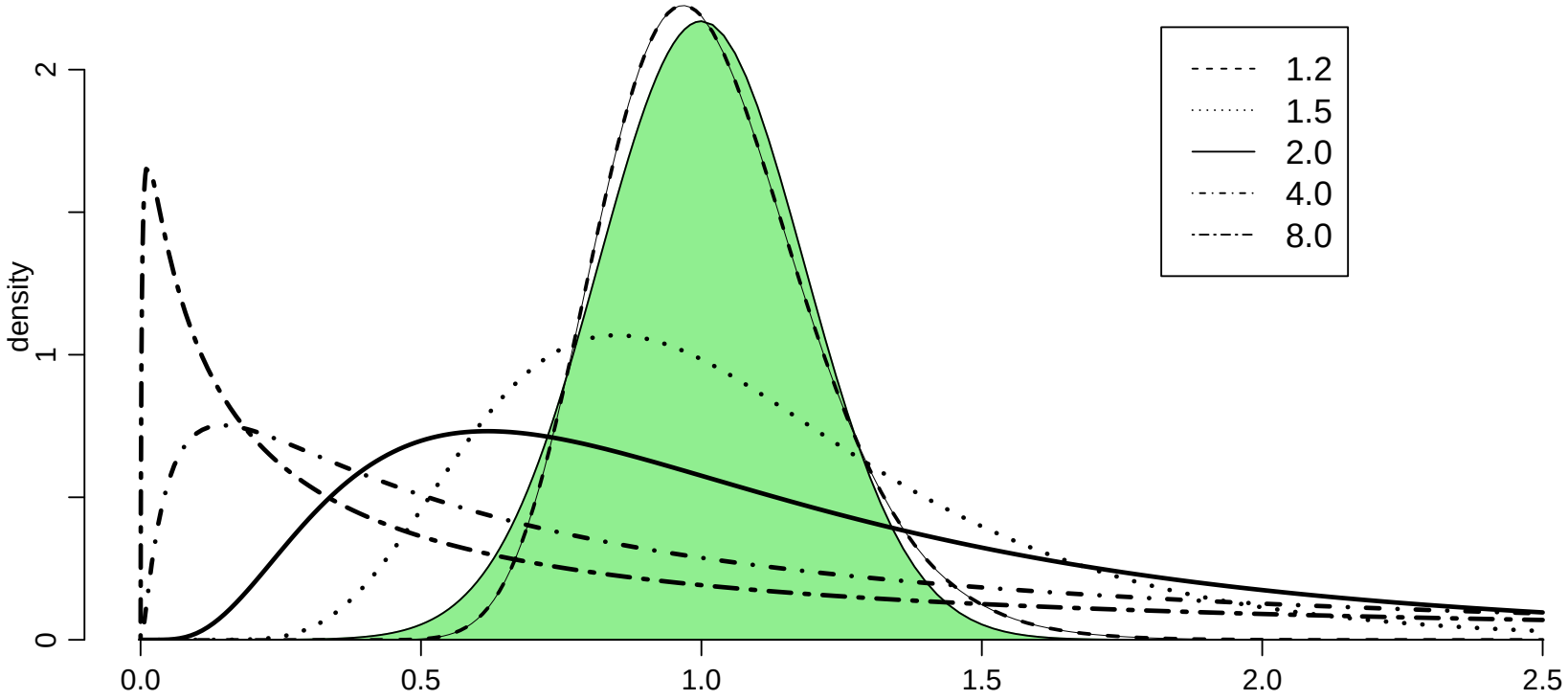
Note: Base of logarithm is not important.

- natural log for theory,
- $\log_{10}$ for practice.

The Log-Normal Distribution

If $Z = \log(X)$ is normally distributed (Gaussian), then the distribution of X is called log-normal.

Densities



green: normal distribution

Density: $\frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x}\exp\left(-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2\right)$

Parameters: μ, σ: Expectation and st.dev. of $\log(X)$

More useful:

- $e^{\mu} = \mu^*$: median, geometric "mean", scale parameter
- $e^{\sigma} = \sigma^*$: multiplicative standard deviation, shape parameter

  σ^* (or σ) determines the shape of the distribution.

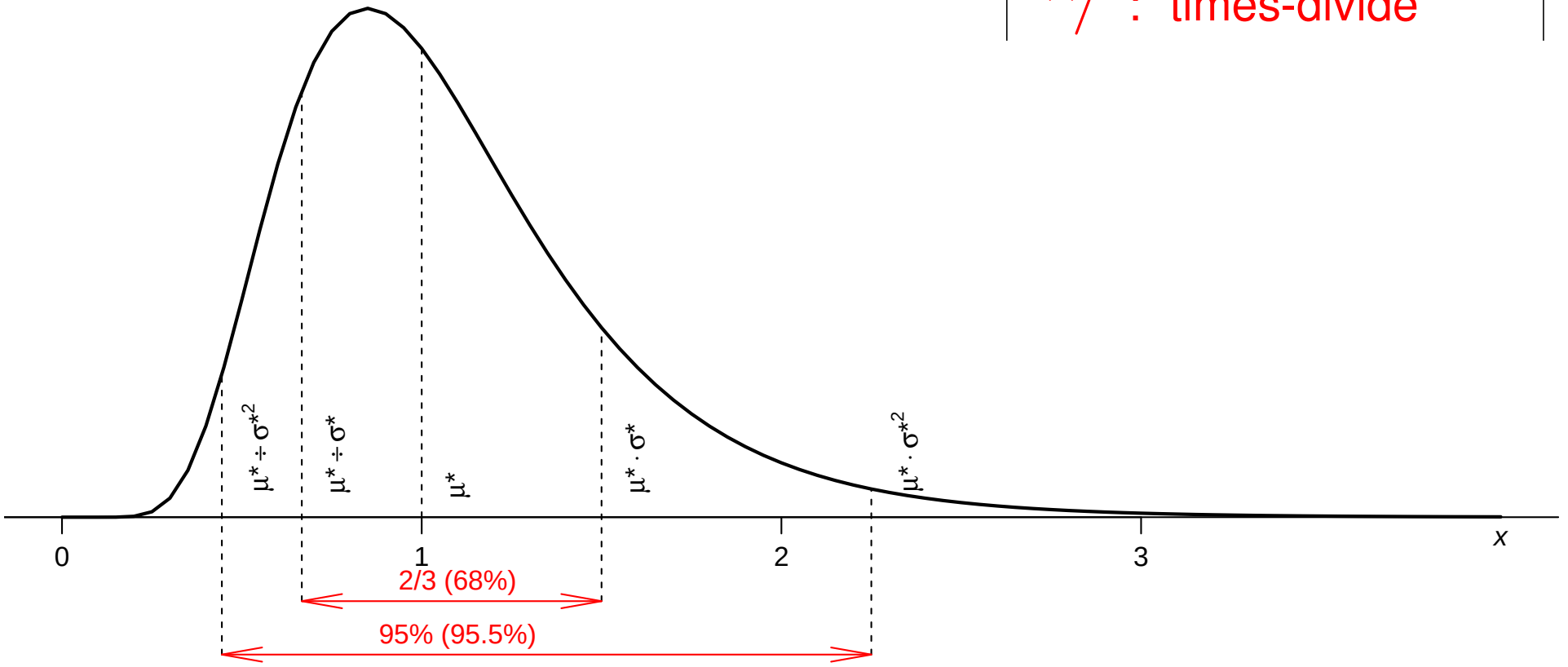
Contrast to

- expectation $E(X) = e^{\mu} \cdot e^{\sigma^2/2}$
- standard deviation $\mathrm{sd}(X)$ from $\mathrm{var}(X) = e^{\sigma^2}\left(e^{\sigma^2}-1\right)e^{2\mu}$

Less useful!

Ranges

Probability	normal	log-normal
2/3 (68%)	$\mu \pm \sigma$	$\mu^* \,{}^{\times}\!/\, \sigma^*$
95%	$\mu \pm 2\sigma$	$\mu^* \,{}^{\times}\!/\, \sigma^{*2}$
		${}^{\times}\!/$: “times-divide”



Properties

We had for the normal distribution:

- Adding normal random variables gives a normal sum.
- Linear combinations $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots$ remain normal.
- $\longrightarrow$ Means of normal variables are normally distributed.
- Central Limit Theorem: Means of non-normal variables are approximately normally distributed.
- $\longrightarrow$ "Hypothesis of Elementary Errors": If random variation is the sum of many small random effects, a normal distribution must be the result.
- Regression models assume normally distributed errors.

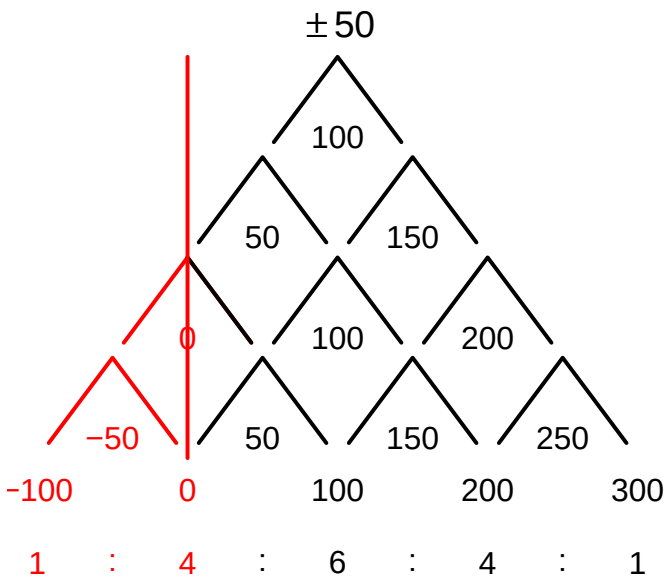
Properties: We have for the log-normal distribution:

- Multiplying log-normal random variables gives a log-normal product.
- $\longrightarrow$ Geometric means of log-normal var.s are log-normally distr.
- Multiplicative Central Limit Theorem: Geometric means of (non-log-normal) variables are approx. log-normally distributed.
- $\longrightarrow$ Multiplicative "Hypothesis of Elementary Errors": If random variation is the product of several random effects, a log-normal distribution must be the result.

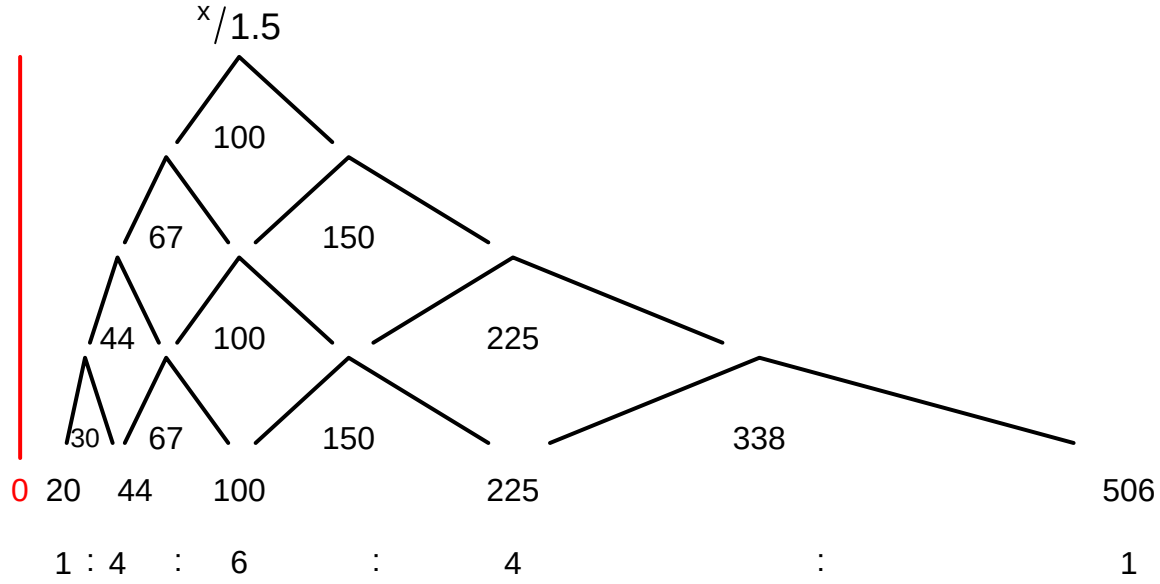
Better name: **Multiplicative** normal distribution!

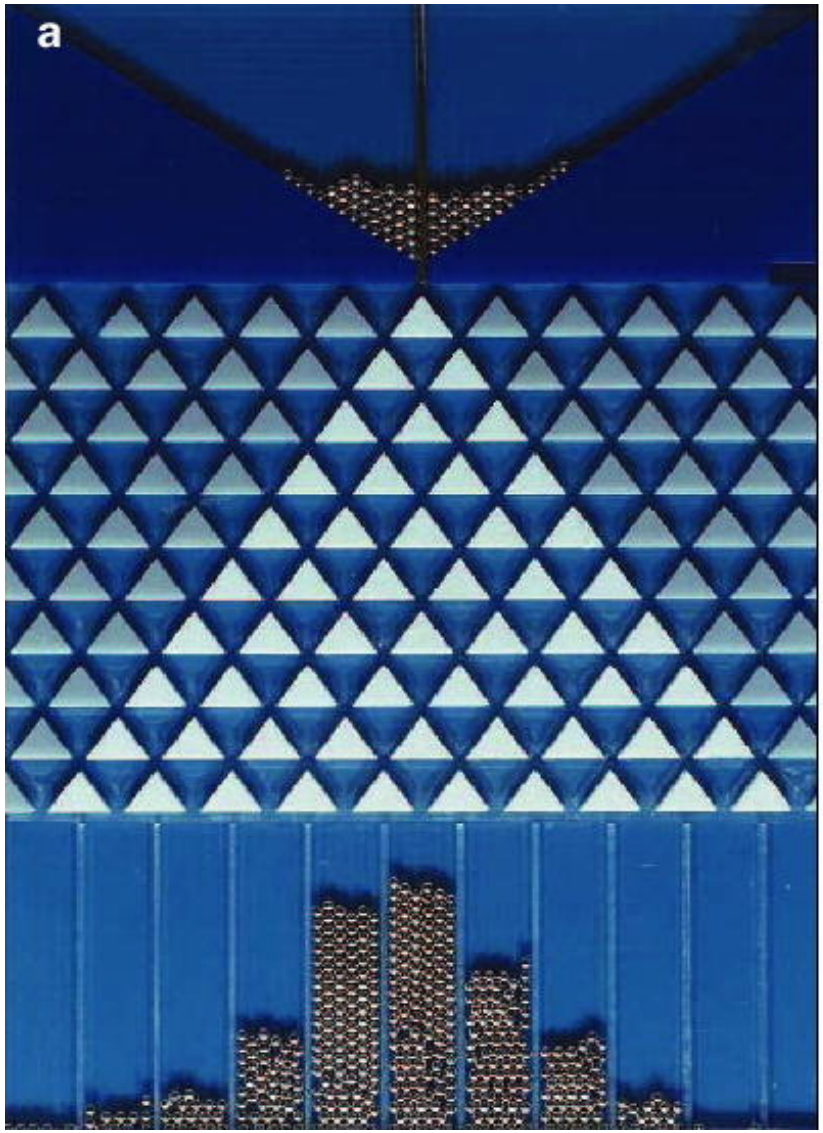
Qunicunx

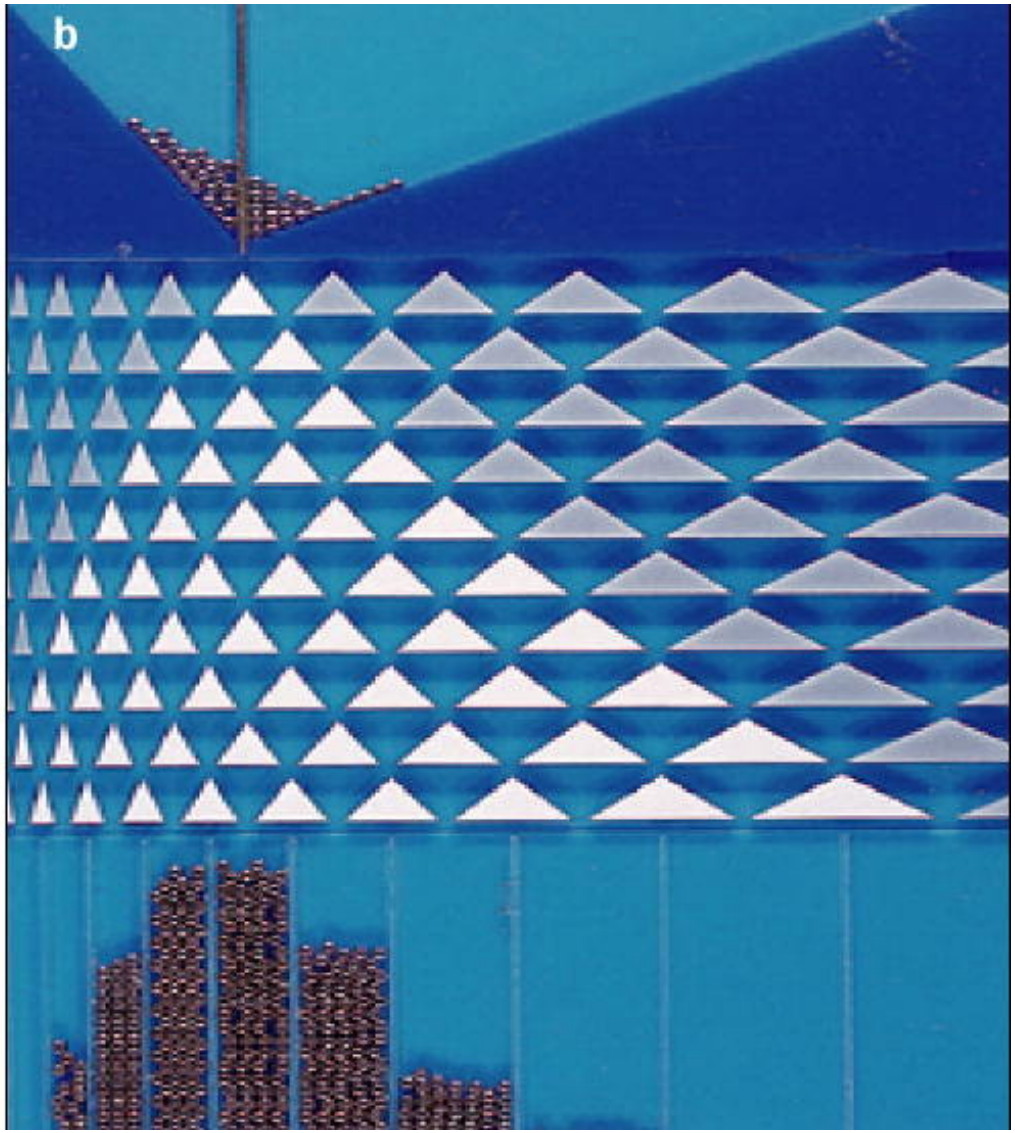
Galton: Additive



Limpert (improving on Kaptayn): Multiplicative




a


b

Back to Properties

- $\longrightarrow$ Multiplicative "Hypothesis of Elementary Errors":
  If random variation is the product of several random effects,
  a log-normal distribution must be the result.

  Note: For "many small" effects, the geometric mean will have
  a small σ^* $\longrightarrow$ approx. normal AND log-normal!

  Such normal distributions are "intrinsically log-normal".
  Keeping this in mind may lead to new insight!

- Regression models assume normally distributed errors! ???

4. Regression

Multiple linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + E$$

Regressors X_j may be functions of original input variables

$\longrightarrow$ model also describes nonlinear relations, interactions, ...

Categorical (nominal) input variables = "factors"

$\longrightarrow$ "dummy" binary regressors

$\longrightarrow$ Model includes Analysis of Variance (ANOVA)!

Linear in the coefficients β_j

$\longrightarrow$ "simple", exact theory, exact inference

estimation by Least Squares $\longrightarrow$ simple calculation

Characteristics of the model:

Formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + E$$

additive effects, additive error

Error term $E \sim \mathcal{N}(0, \sigma^2) \longrightarrow$

- constant variance
- symmetric error distribution

Target variable has skewed (error) distribution,

standard deviation of error increases with Y

$\longrightarrow$ transform $Y \longrightarrow \log(Y)$!

$$\log(\widetilde{Y}) = Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + E$$

Ordinary, additive model	**Multiplicative model**
Formula	
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + E$	$\log(\widetilde{Y}) = Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + E$ $\widetilde{Y} = \widetilde{\beta}_0 \cdot \widetilde{X}_1^{\beta_1} \cdot \widetilde{X}_2^{\beta_2} \cdot ... \cdot \widetilde{E}$
additive effects, additive error	multiplicative effects, mult. errors
Error term	
$E \sim \mathcal{N}(0, \sigma^2) \longrightarrow$	$\widetilde{E} \sim \ell\mathcal{N}(1, \sigma^*) \longrightarrow$
– constant variance	– constant relative error
– symmetric error distribution	– skewed error distribution

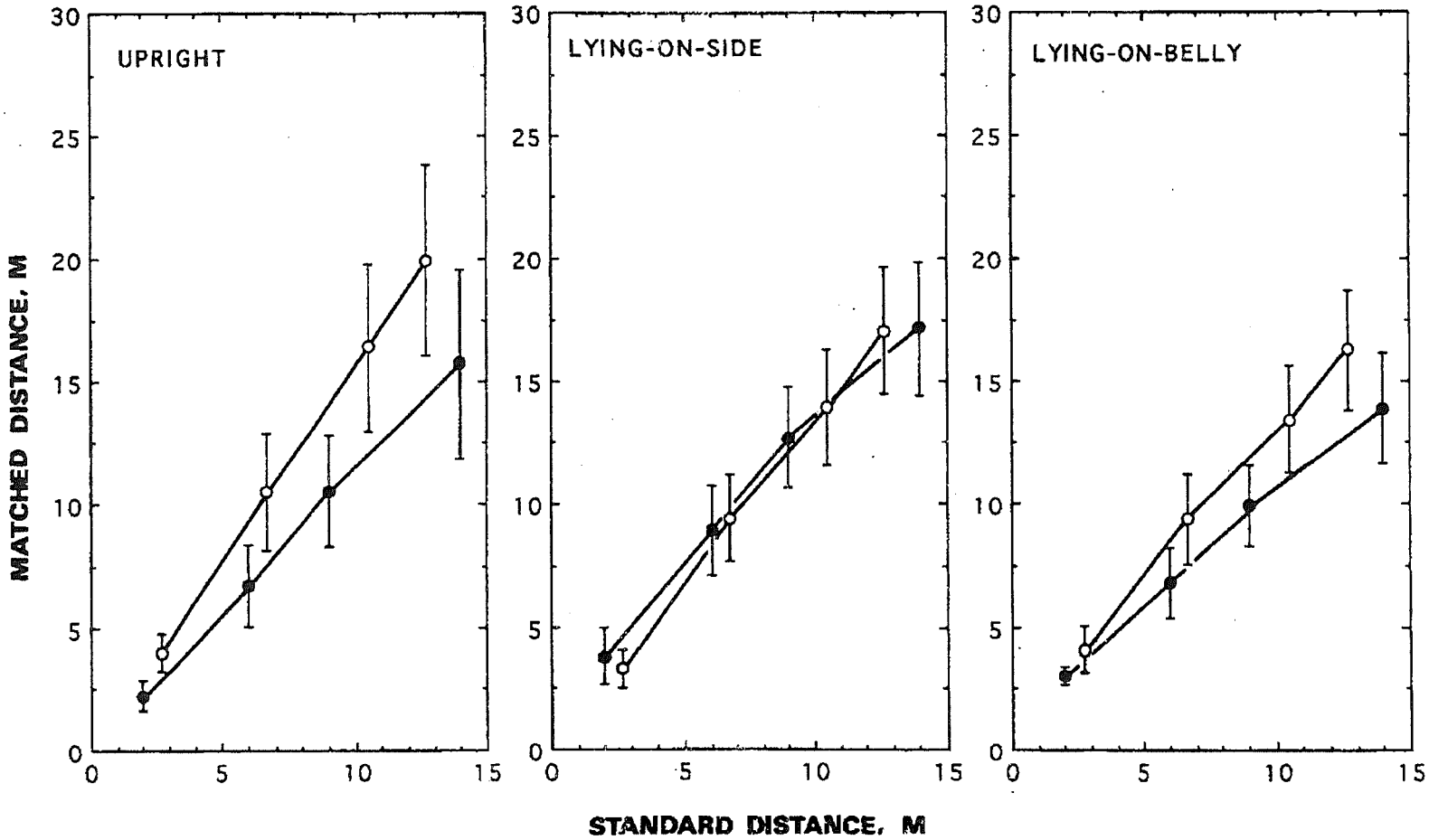
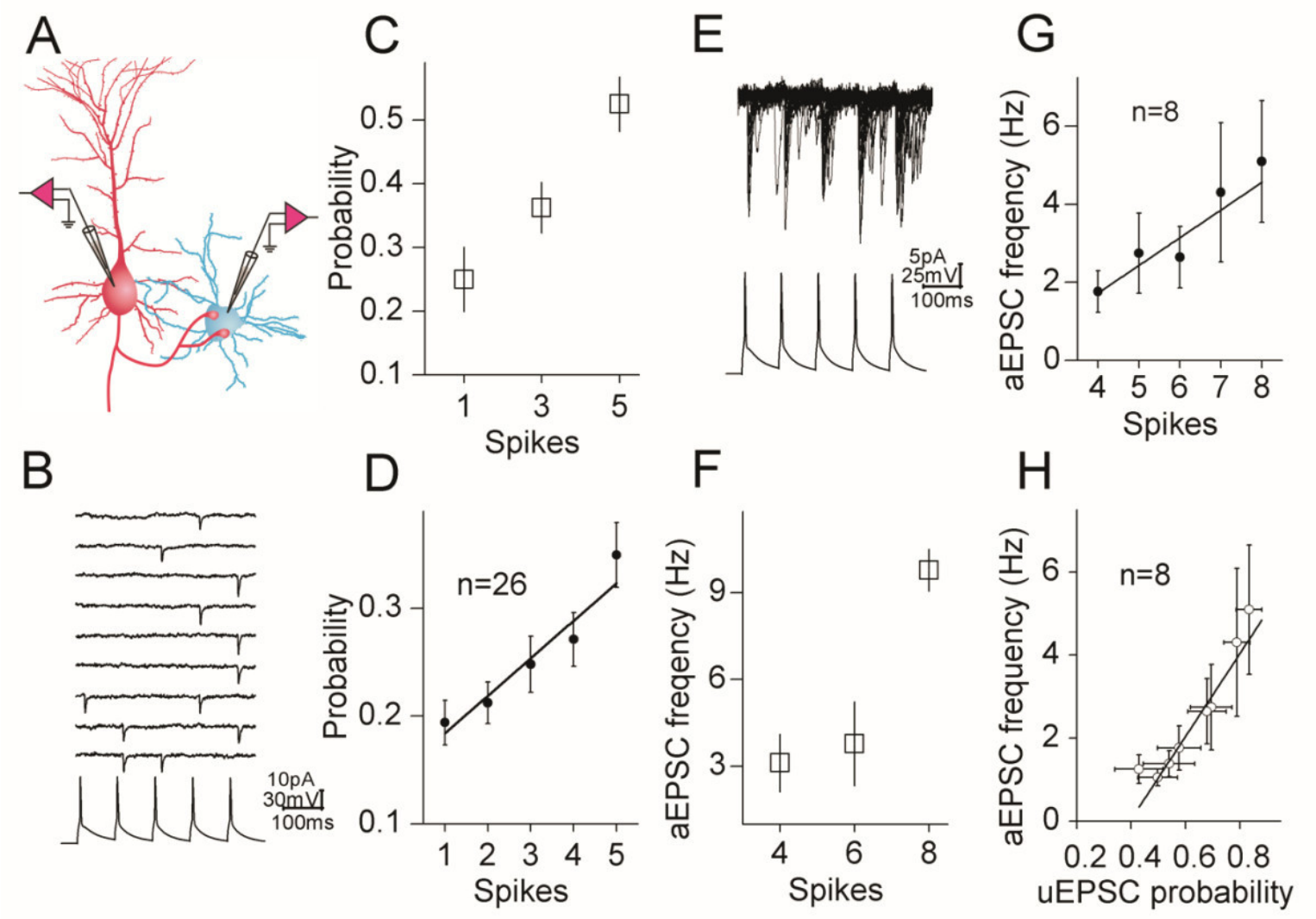
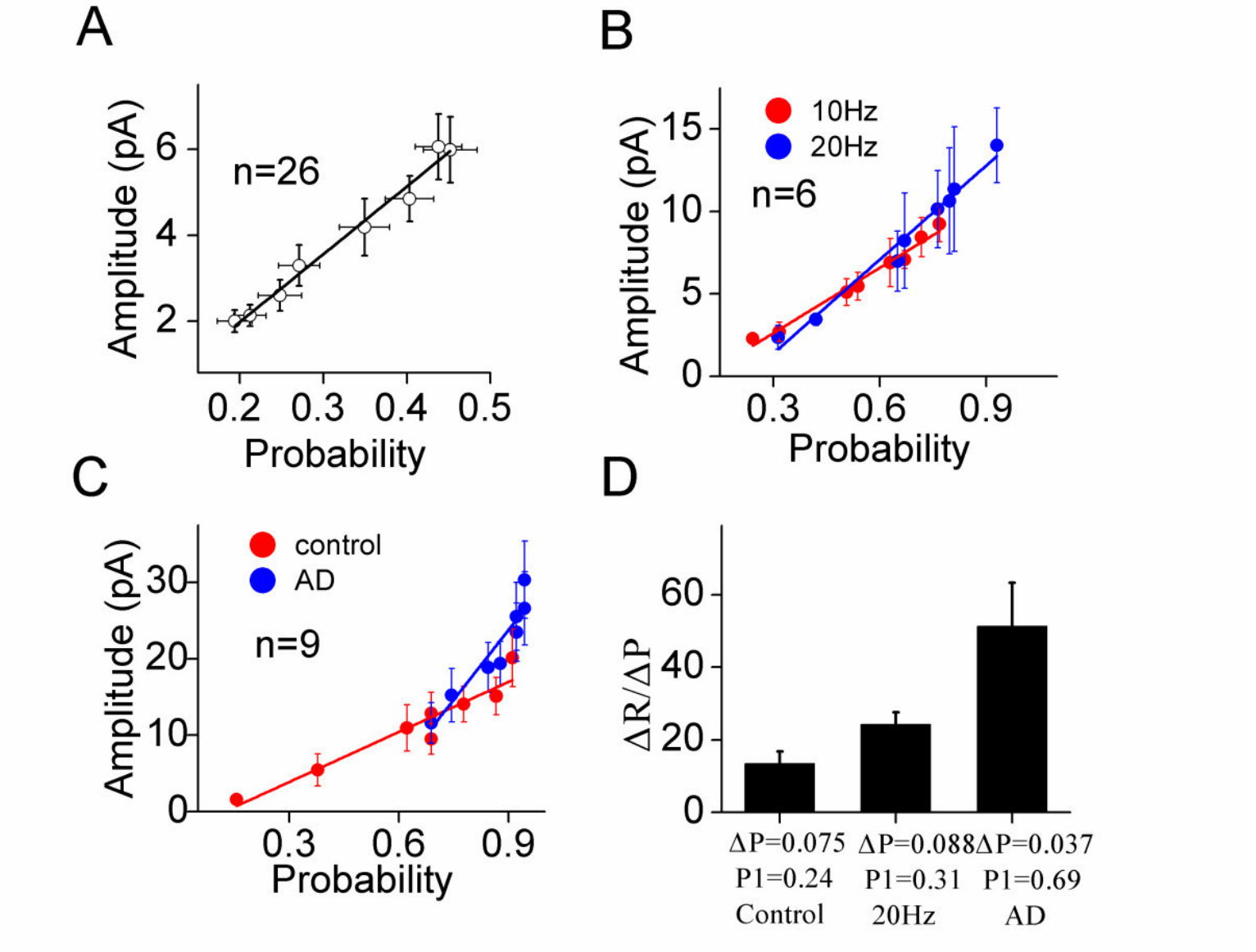


Figure 2. Mean longitudinal distance (in meters) as a function of standard distance under natural binocular viewing. The open circles represent the vertical standard and the filled circles represent the horizontal standard. The left panel is for the upright subjects; the center panel for the lying-on-side subjects; and the right panel for the lying-on-belly subjects. The bars passing through the data points represent the standard deviations.



Yu et al (2012): Upregulation of transmitter release probability improves a conversion of synaptic analogue signals into neuronal digital spikes

Figure 1. The probability of releasing glutamates increases during sequential presynaptic spikes...



Yu et al (2012): Upregulation of transmitter release probability improves a conversion of synaptic analogue signals into neuronal digital spikes

Figure 4. Presynaptic Ca 2+ enhances an efficiency of probability-driven facilitation.

5. Advantages of using the log-normal distribution

... or of applying the log transformation to data.

The normal and log-normal distributions are difficult to distinguish

for $\sigma^* < 1.2 \leftrightarrow$ cv < 0.18

where the coef. of variation cv $\approx \sigma^* - 1$

$\longrightarrow$ We discuss case of larger σ^*.

More meaningful parameters

- The expected value of a skewed distribution is less typical than the median.
- (cv or) σ^* characterizes size of relative error
- Characteristic σ^* found in diseases:
  latent periods for different infections: $\sigma^* \approx 1.4$;
  survival times after diagnosis of cancer, for different types: $\sigma^* \approx 3$
  $\longrightarrow$ Deeper insight?

Fulfilling assumptions, power

What happens to inference based on the normal distribution

if the data is log-normal?

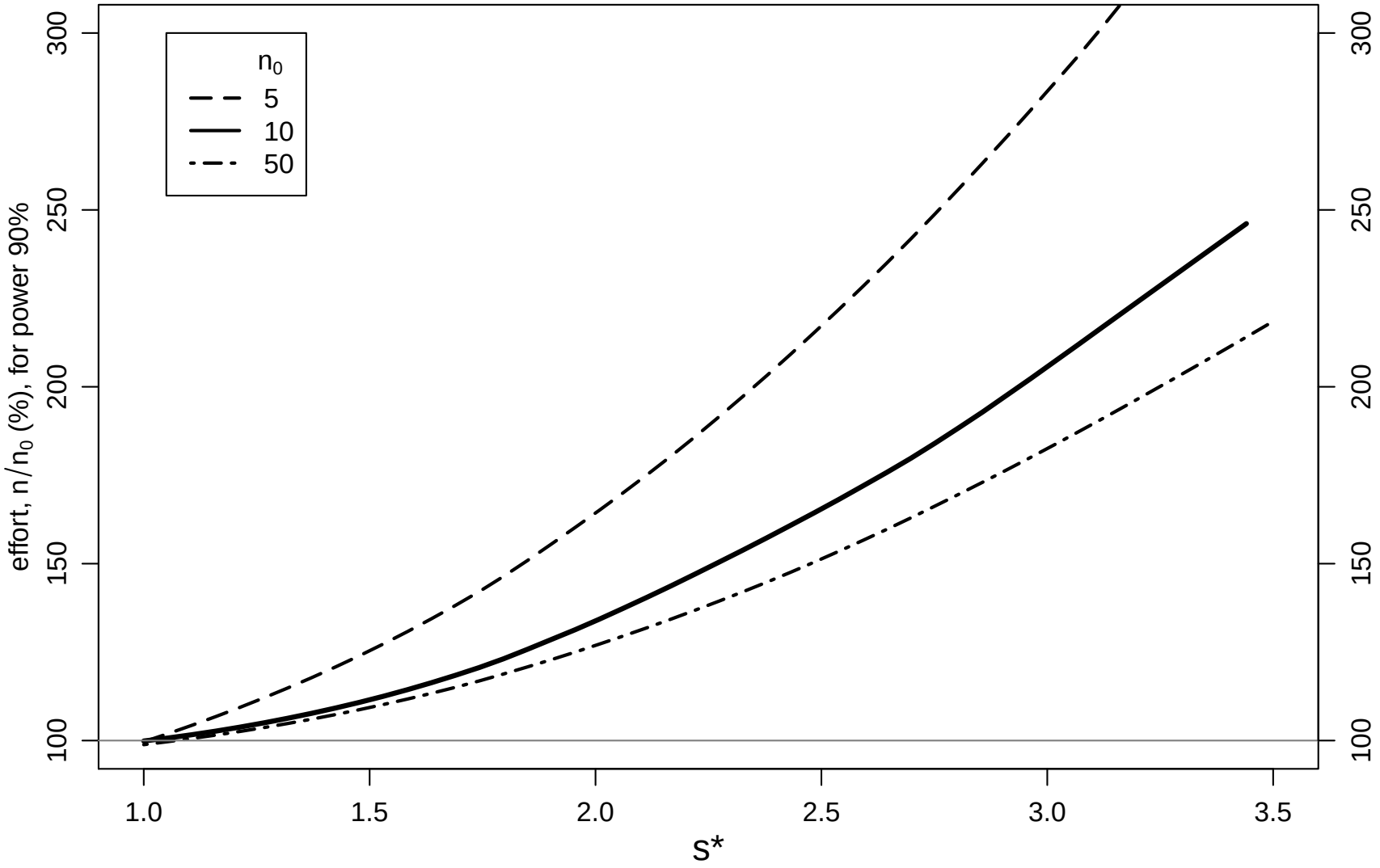
- Level = prob. of falsely rejecting the null hypothesis

  coverage prob. of confidence intervals are o.k.

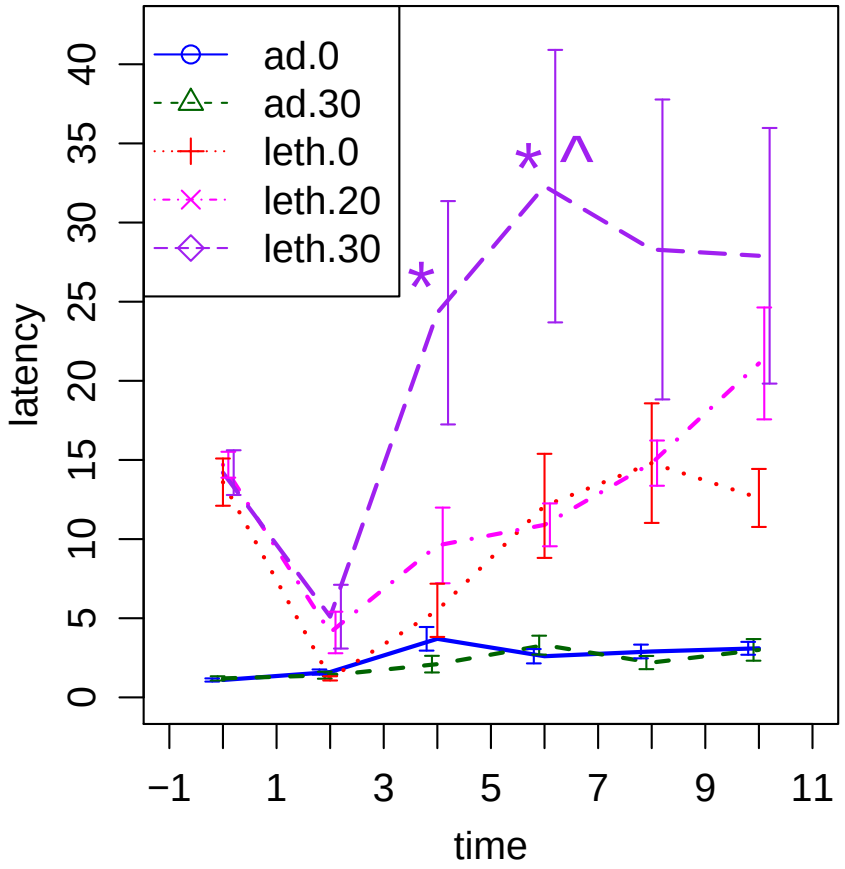
- Loss of power! $\longrightarrow$ wasted effort!

- Loss of power! $\longrightarrow$ wasted effort!

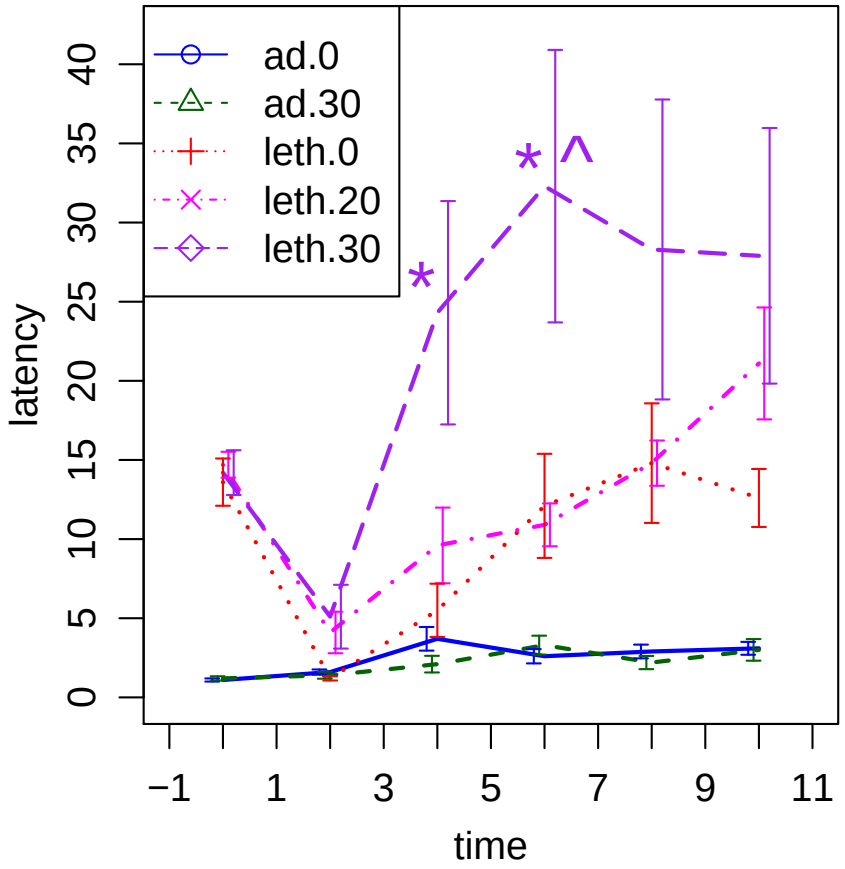
  Difference between 2 groups (samples)

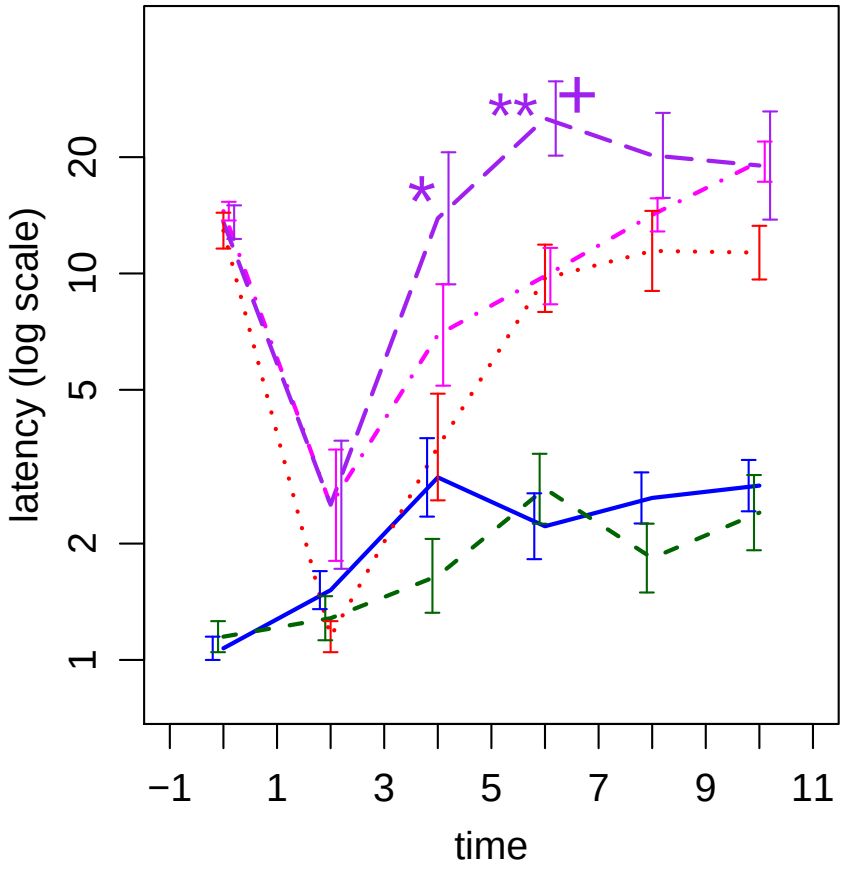


More informative graphics



More informative graphics





More signi-

ficance

6. Conclusions

Genesis

- The normal distribution is good for estimators, test statistics, data with small coef.of variation, and log-transformed data.

  The log-normal distribution is good for original data.

- Summation, Means, Central limit theorem, Hyp. of elem. errors

  $\longrightarrow$ normal distribution

  Multiplication, Geometric means, ...

  $\longrightarrow$ log-normal distribution

Applications

- Adequate ranges: $\mu^* \times / \sigma^{*2}$ covers $\approx 95\%$ of the data
- Gain of power of hypothesis tests $\longrightarrow$ save efforts for experiments (e.g., saves animals!)
- Regression models assume normally distributed errors.
  $\longrightarrow$ Regression model for $\log(Y)$ instead of Y.
  Back transformation: $Y = \widetilde{\beta}_0 \cdot X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot \ldots \cdot \widetilde{E}$
- Parameter σ^* may characterize a class of phenomena
  (e.g., diseases) $\longrightarrow$ new insight ?!

Mathematical Statistics *adds*.

$\longrightarrow$ uses normal distribution

Nature *multiplies*

$\longrightarrow$ yields log-normal distribution

Scientists (and applied statisticians)

add logarithms!

use the normal distribution for log(data) and theory

use log-normal distribution for data

Thank you for your attention!