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.
- Means of normal variables are normally distributed.
- Central Limit Theorem: Means of non-normal variables are approximately normally distributed.
- “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 $\rightarrow$ “amounts” (John Tukey)

Example: HydroxyMethylFurfurol (HMF) in honey (Renner 1970)
Measurements:
- size, weight, concentration, intensity, duration, price, activity
- All $> 0 \rightarrow$ “amounts”

Distribution is skewed: left steep, right flat, skewness $> 0$
- unless coefficient of variation $cv(X) = sd(X)/E(X)$ 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] $\rightarrow$ log scale!
The 95% Range Check

For every normal distribution, negative values have a probability $> 0$. → normal distribution inadequate for positive variables.

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

Then, the distribution is noticeably skewed.
2. Laws of Nature

(a) Physics

\[ E = m \cdot c^2 \]

Stopping distance \( s = \frac{v^2}{2 \cdot a} \); Velocity \( v = \frac{F \cdot t}{m} \)

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 \( \Delta t \rightarrow +10^0C \quad \Rightarrow \quad v \rightarrow \cdot 2 \)

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) \(1 \rightarrow 2 \rightarrow 4 \rightarrow 8 \rightarrow 16\)

Growth, size \(s_t = s_0 \cdot k^t\)

Hagen-Poiseuille Law; Volume:
\[ V_t = \frac{(\Delta P \cdot r^4 \cdot \pi)}{(8 \cdot \eta \cdot L)}; \quad \Delta P : \text{pressure difference} \]

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
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: \( \mu, \sigma \): 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
  \( \sigma^* \) (or \( \sigma \)) determines the shape of the distribution.

Contrast to

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

Less useful!
### Ranges

<table>
<thead>
<tr>
<th>Probability</th>
<th>normal</th>
<th>log-normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/3 (68%)</td>
<td>( \mu \pm \sigma )</td>
<td>( \mu^* \times/ \sigma^* )</td>
</tr>
<tr>
<td>95%</td>
<td>( \mu \pm 2\sigma )</td>
<td>( \mu^* \times/ \sigma^{*2} )</td>
</tr>
</tbody>
</table>

\( \times/ \) : “times-divide”

---

![Diagram showing the distribution of normal and log-normal probabilities.](image-url)
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.
- Means of normal variables are normally distributed.
- Central Limit Theorem: Means of non-normal variables are approximately normally distributed.
- “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.
- Geometric means of log-normal variables are log-normally distributed.
- Multiplicative Central Limit Theorem: Geometric means of (non-log-normal) variables are approx. log-normally distributed.
- 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
Back to Properties

- **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 $\sigma^*$ 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:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + E \]

Regressors \( X_j \) may be functions of original input variables

\( \rightarrow \) model also describes nonlinear relations, interactions, ...

Categorical (nominal) input variables = “factors”

\( \rightarrow \) “dummy” binary regressors

\( \rightarrow \) Model includes Analysis of Variance (ANOVA)!

Linear in the coefficients \( \beta_j \)

\( \rightarrow \) “simple”, exact theory, exact inference

estimation by Least Squares \( \rightarrow \) simple calculation
Characteristics of the model:

Formula:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + E \]

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 \( Y \)

\[ \longrightarrow \text{transform } Y \longrightarrow \log(Y) ! \]

\[ \log(\tilde{Y}) = Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + E \]
### 4. Regression

<table>
<thead>
<tr>
<th>Ordinary, additive model</th>
<th>Multiplicative model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Formula</strong></td>
<td></td>
</tr>
<tr>
<td>[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + E ]</td>
<td>[ \log(\tilde{Y}) = Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + E ]</td>
</tr>
<tr>
<td>[ \tilde{Y} = \tilde{\beta}_0 \cdot \tilde{X}_1^{\beta_1} \cdot \tilde{X}_2^{\beta_2} \cdot ... \cdot \tilde{E} ]</td>
<td></td>
</tr>
<tr>
<td>additive effects, additive error</td>
<td>multiplicative effects, mult. errors</td>
</tr>
<tr>
<td><strong>Error term</strong></td>
<td></td>
</tr>
<tr>
<td>[ E \sim \mathcal{N}(0, \sigma^2) ]</td>
<td>[ \tilde{E} \sim \mathcal{LN}(1, \sigma^*) ]</td>
</tr>
<tr>
<td>– constant variance</td>
<td>– constant relative error</td>
</tr>
<tr>
<td>– symmetric error distribution</td>
<td>– skewed error distribution</td>
</tr>
</tbody>
</table>
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 \( \sigma^* < 1.2 \iff cv < 0.18 \)

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

\( \rightarrow \) We discuss case of larger \( \sigma^* \).
More meaningful parameters

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

\( \rightarrow \) 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
- **Loss of power!**  → wasted effort!

coverage prob. of confidence intervals are **o.k.**
5. Advantages of using the log-normal distribution

- Loss of power! \(\rightarrow\) wasted effort!

Difference between 2 groups (samples)
5. Advantages of using the log-normal distribution

More informative graphics
More informative graphics

More significance
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 → normal distribution

Multiplication, Geometric means, ... → log-normal distribution
Applications

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

$\rightarrow$ uses normal distribution $\quad \rightarrow$ yields log-normal distribution

Scientists (and applied statisticians)

*add logarithms!*

use the normal distribution for $\log(\text{data})$ and theory

use log-normal distribution for data

Thank you for your attention!