Andrew Gard - equitable.equations@gmail.com

The log-normal distribution

The **log-normal distribution** is used to model phenomenon where

- Only positive values are possible
- Observations may differ by several degrees of magnitude.

The **log-normal distribution** is used to model phenomenon where

- Only positive values are possible
- Observations may differ by several degrees of magnitude.

If X has a log-normal distribution, then $\log(X)$ is normally-distributed. While the base of the logarithm doesn't matter, the natural logarithm is standard.

Many real-world variables can be modeled with a log-normal distribution.

Many real-world variables can be modeled with a log-normal distribution. For instance,

- Home prices

Many real-world variables can be modeled with a log-normal distribution. For instance,

- Home prices
- Numbers of moves in games of chess

Many real-world variables can be modeled with a log-normal distribution. For instance,

- Home prices
- Numbers of moves in games of chess
- Lengths of comments to YouTube videos

Many real-world variables can be modeled with a log-normal distribution. For instance,

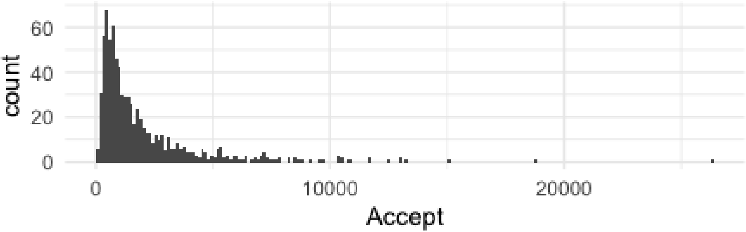
- Home prices
- Numbers of moves in games of chess
- Lengths of comments to YouTube videos
- Household incomes

Many real-world variables can be modeled with a log-normal distribution. For instance,

- Home prices
- Numbers of moves in games of chess
- Lengths of comments to YouTube videos
- Household incomes
- Population sizes

Example. In the U.S. News `College` data set, included with the `ISLR2` R package, many of the variables can be modeled with a log-normal distribution.

Example. In the U.S. News `College` data set, included with the `ISLR2` R package, many of the variables can be modeled with a log-normal distribution. For instance, the `Accept` variable, which represents the number of students accepted, looks like this:



While acceptance at most schools is in the hundreds or thousands, a few accepted less than 100 or more than 10,000.

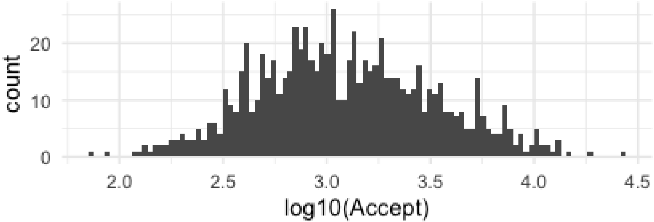
The following table shows a breakdown of the number of digits in each college's acceptance count.

Enroll: number of digits	number of schools
2	2
3	346
4	413
5	16

The following table shows a breakdown of the number of digits in each college's acceptance count.

Enroll: number of digits	number of schools
2	2
3	346
4	413
5	16

Here's a histogram showing counts of the base-10 logarithm of the acceptance counts.



Notice how bell-shaped it looks!

Like the normal distribution, the log-normal distribution is determined by two parameters, μ and σ^2.

Like the normal distribution, the log-normal distribution is determined by two parameters, μ and σ^2. Despite the notation, these do **not** represent the mean and variance of the distribution. Rather, they are the mean and variance of the corresponding normal distribution, the one obtained when the natural logarithm of the values are taken.

Like the normal distribution, the log-normal distribution is determined by two parameters, μ and σ^2. Despite the notation, these do **not** represent the mean and variance of the distribution. Rather, they are the mean and variance of the corresponding normal distribution, the one obtained when the natural logarithm of the values are taken.

We write $X \sim \text{Lognormal}(\mu, \sigma^2)$ for a random variable with this distribution.

Like the normal distribution, the log-normal distribution is determined by two parameters, μ and σ^2. Despite the notation, these do **not** represent the mean and variance of the distribution. Rather, they are the mean and variance of the corresponding normal distribution, the one obtained when the natural logarithm of the values are taken.

We write $X \sim \text{Lognormal}(\mu, \sigma^2)$ for a random variable with this distribution.

It turns out that the mean and variance of such an X are given by

$$\mu_X = e^{\mu+\sigma^2/2} \quad \text{and} \quad \sigma_X^2 = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$$

Probabilities in the log-normal distribution can be calculated from normal probabilities using the fact that $X \sim \textit{Lognormal}(\mu, \sigma^2)$ exactly when $\ln X \sim N(\mu, \sigma^2)$.

Probabilities in the log-normal distribution can be calculated from normal probabilities using the fact that $X \sim \mathit{Lognormal}(\mu, \sigma^2)$ exactly when $\ln X \sim N(\mu, \sigma^2)$.

$$P(X \leq x) = P(\ln X \leq \ln x)$$

Probabilities in the log-normal distribution can be calculated from normal probabilities using the fact that $X \sim \mathit{Lognormal}(\mu, \sigma^2)$ exactly when $\ln X \sim N(\mu, \sigma^2)$.

$$\begin{aligned} P(X \leq x) &= P(\ln X \leq \ln x) \\ &= P(Y \leq \ln x) \qquad \text{in } N(\mu, \sigma^2) \end{aligned}$$

Probabilities in the log-normal distribution can be calculated from normal probabilities using the fact that $X \sim \mathit{Lognormal}(\mu, \sigma^2)$ exactly when $\ln X \sim N(\mu, \sigma^2)$.

$$\begin{aligned} P(X \le x) &= P(\ln X \le \ln x) \\ &= P(Y \le \ln x) \qquad \text{in } N(\mu, \sigma^2) \\ &= F(\ln x) \end{aligned}$$

where F is the cumulative distribution function (cdf) of $N(\mu, \sigma^2)$.

$$F(x) = P(Y \le x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

Probabilities in the log-normal distribution can be calculated from normal probabilities using the fact that $X \sim \mathit{Lognormal}(\mu, \sigma^2)$ exactly when $\ln X \sim N(\mu, \sigma^2)$.

$$
\begin{aligned}
P(X \leq x) &= P(\ln X \leq \ln x) \\
&= P(Y \leq \ln x) \qquad \text{in } N(\mu, \sigma^2) \\
&= F(\ln x)
\end{aligned}
$$

where F is the cumulative distribution function (cdf) of $N(\mu, \sigma^2)$.

$$F(x) = P(Y \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

We get the cdf of the log-normal distribution by replacing x with $\ln x$.

Differentiating the cumulative distribution function gives the densify function (pdf) of the log-normal distribution.

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

Differentiating the cumulative distribution function gives the densify function (pdf) of the log-normal distribution.

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

As with any continuous distribution, we can calculate probabilities by integrating this pdf over the appropriate range.

$$P(a \leq X \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b \frac{1}{x} \cdot \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) dx$$

Differentiating the cumulative distribution function gives the densify function (pdf) of the log-normal distribution.

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

As with any continuous distribution, we can calculate probabilities by integrating this pdf over the appropriate range.

$$P(a \leq X \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b \frac{1}{x} \cdot \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) dx$$

Typically we use technology such as R to do this.

$$P(a \leq X \leq b) = \texttt{plnorm}(\texttt{b}, \mu, \sigma) - \texttt{plnorm}(\texttt{a}, \mu, \sigma)$$

Estimating μ and σ^2 from data

The maximum likelihood estimators for μ and σ^2 in the log-normal distribution are

$$\hat{\mu} = \frac{1}{n}\sum \ln x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum(\ln x_i - \hat{\mu})^2$$

Estimating μ and σ^2 from data

The maximum likelihood estimators for μ and σ^2 in the log-normal distribution are

$$\hat{\mu} = \frac{1}{n}\sum \ln x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum(\ln x_i - \hat{\mu})^2$$

The latter is biased, however. On average, it will slightly underestimate the parameter σ^2.

Estimating μ and σ^2 from data

The maximum likelihood estimators for μ and σ^2 in the log-normal distribution are

$$\hat{\mu} = \frac{1}{n}\sum \ln x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum (\ln x_i - \hat{\mu})^2$$

The latter is biased, however. On average, it will slightly underestimate the parameter σ^2. Similarly to the normal distribution, an unbiased estimator of population variance is given by

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum (\ln x_i - \hat{\mu})^2$$

Example. Freshman undergraduate enrollments at 8 randomly-selected schools in the `College` data set are given below.

$$605,\ 97,\ 736,\ 143,\ 622,\ 2408,\ 345,\ 324$$

Example. Freshman undergraduate enrollments at 8 randomly-selected schools in the `College` data set are given below.

$$605, \ 97, \ 736, \ 143, \ 622, \ 2408, \ 345, \ 324$$

- $\hat{\mu} = \frac{1}{n}\sum \ln x_i = \frac{1}{8}(\ln 605 + \cdots + \ln 324) \approx 6.05$

Example. Freshman undergraduate enrollments at 8 randomly-selected schools in the `College` data set are given below.

$$605,\ 97,\ 736,\ 143,\ 622,\ 2408,\ 345,\ 324$$

- $\hat{\mu} = \frac{1}{n}\sum \ln x_i = \frac{1}{8}(\ln 605 + \cdots + \ln 324) \approx 6.05$
- $\hat{\sigma}^2 = \frac{1}{n-1}\sum(\ln x_i - \hat{\mu})^2 = \frac{1}{7}\Big((\ln 605 - 6.05)^2 + \cdots + (\ln 324 - 6.05)^2\Big) \approx 1.01$

Example. Freshman undergraduate enrollments at 8 randomly-selected schools in the `College` data set are given below.

$$605,\ 97,\ 736,\ 143,\ 622,\ 2408,\ 345,\ 324$$

- $\hat{\mu} = \frac{1}{n}\sum \ln x_i = \frac{1}{8}(\ln 605 + \cdots + \ln 324) \approx 6.05$
- $\hat{\sigma}^2 = \frac{1}{n-1}\sum(\ln x_i - \hat{\mu})^2 = \frac{1}{7}\Big((\ln 605 - 6.05)^2 + \cdots + (\ln 324 - 6.05)^2\Big) \approx 1.01$

These correspond to a mean of 701.5 and standard deviation of 922.2. Bear in mind, however, that these numbers are unlikely to be representative of the center and spreads of that distribution.

Example. Freshman undergraduate enrollments at 8 randomly-selected schools in the `College` data set are given below.

$$605,\ 97,\ 736,\ 143,\ 622,\ 2408,\ 345,\ 324$$

- $\hat{\mu} = \frac{1}{n}\sum \ln x_i = \frac{1}{8}(\ln 605 + \cdots + \ln 324) \approx 6.05$
- $\hat{\sigma}^2 = \frac{1}{n-1}\sum(\ln x_i - \hat{\mu})^2 = \frac{1}{7}\Big((\ln 605 - 6.05)^2 + \cdots + (\ln 324 - 6.05)^2\Big) \approx 1.01$

These correspond to a mean of 701.5 and standard deviation of 922.2. Bear in mind, however, that these numbers are unlikely to be representative of the center and spreads of that distribution.

The actual parameters in the `College` data set are $\mu = 6.18$ and $\sigma = 0.91$.