# ENCYCLOPEDIA OF STATISTICAL SCIENCES

## UPDATE VOLUME 2

A WILEY-INTERSCIENCE PUBLICATION

John Wiley & Sons, Inc.

NEW YORK • CHICHESTER • WEINHEIM • BRISBANE • SINGAPORE • TORONTO

This book is printed on acid-free paper. ♾



Published simultaneously in Canada.



***Library of Congress Cataloging in Publication Data:***

Encyclopedia of statistical sciences. Update/[edited by] Samuel
Kotz, Campbell B. Read, David L. Banks.
p. cm.
"A Wiley-Interscience publication."
Includes index.
1. Mathematical statistics—Encyclopedias. 2. Statistics
—Encyclopedias. I. Kotz, Samuel. II. Read, Campbell B.
III. Banks, David L.
QA276.14.E5 1982 Suppl.
519.5′03—dc20

96-25879
CIP

This is Volume 2 of the Encyclopedia of Statistical Sciences Update Series. ISBN: 0-471-11939-3.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

## Bibliography

Hastie, T. and Loader, C. (1993). Local regression: automatic kernel carpentry. *Statist. Sci.*, **8**, 120–143. (A readable introduction to local polynomial fitting from an applied point of view.)

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*, Lecture Notes in Statistics *46*. Springer-Verlag, Berlin. (Describes both convolution kernel estimators and local polynomials.)


(BANDWIDTH SELECTION
GRADUATION
LOCAL REGRESSION
NONPARAMETRIC REGRESSION
REGRESSION, POLYNOMIAL)


BURKHARDT SEIFERT
THEO GASSER


# LOG-GAMMA DISTRIBUTION

This distribution can be derived by using a transformation* of the form $X = \log(\theta Y)$, where $Y$ follows a gamma distribution* with scale and shape parameters $\theta$ and $k$, respectively. In that case, $X$ follows a log-gamma distribution (Bartlett and Kendall [3]) with probability density function given by

$$g_1(x; k) = \frac{1}{\Gamma(k)} \log(kx - e^x),$$
$$-\infty < x < \infty, \qquad k > 0,$$

where $\Gamma(\cdot)$ is the gamma function. The form of $g_1$ reduces to the standard extreme-value distribution* on setting $k = 1$. Some important properties of the log-gamma distribution are given below.

1. *Shape Properties:* $g_1$ is negatively skewed, with skewness decreasing as $k$ increases. Figure 1 shows the shape of the distribution for several values of $k$.
2. *Moment Generating Function:*
$$M_x(t) = \Gamma(t + k)/\Gamma(k).$$
3. *Cumulative Distribution Function:*
$$F^*(x) = I_{e^x}(k), \qquad -\infty < x < \infty, \quad k > 0,$$
where $I_t(\cdot)$ is the incomplete gamma function.
4. *Cumulants:*
$$k_r^* = d^r \log \Gamma(k)/dk^r.$$

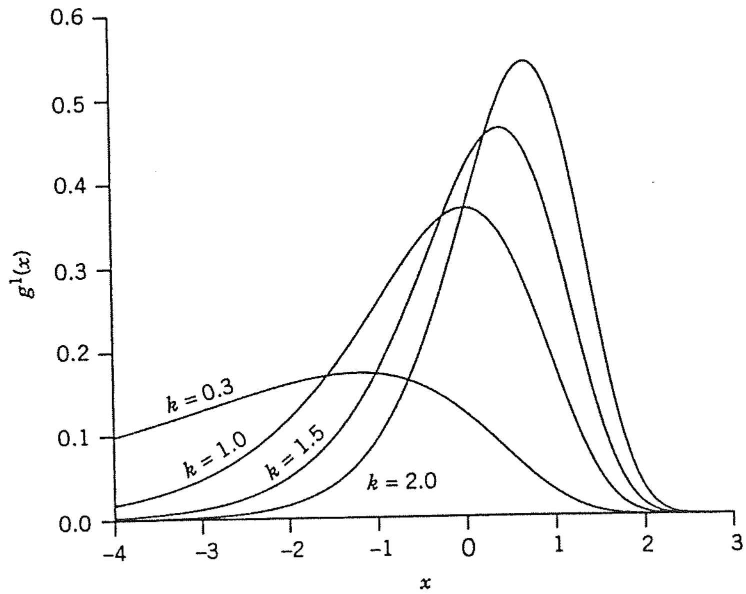


**Figure 1** Log-gamma probability density function with $k = 0.3$, 1.0, 1.5, and 2.0.

5. *Mean and Variance:*

$$\mu = E(X) = \psi(k),$$

$$\sigma^2 = \text{Var}(X) = \psi'(k),$$

where $\psi(k) = d \log \Gamma(k)/dk$ (*see* DIGAMMA FUNCTION) and $\psi'(k) = d^2 \log \Gamma(k)/dk^2$ (*see* TRIGAMMA FUNCTION).

6. *Infinite Divisibility*:* It follows from Shanbhag et al. [13] that the log-gamma distribution is self-decomposable and therefore infinitely divisible.

The log-gamma distribution with density function $g_1$ and various reparametrizations (given below) have been shown to be very useful as lifetime models. By using the asymptotic formulae $\psi(k) \sim \log k$ and $\psi'(k) \sim 1/k$, Prentice [10] has suggested a reparametrization of the density function $g_1$ as

$$g_2(x;k) = \frac{k^{k-1/2}}{\Gamma(k)} \exp\left[\sqrt{k}\,x - k \exp\left(\frac{x}{\sqrt{k}}\right)\right], \quad -\infty < x < \infty, \quad k > 0,$$

and showed that as $k \to \infty$, $g_2(x;k)$ converges to the standard normal density function. He has also considered a reparametrization which replaces $k$ with $q = k^{-1/2}$, and extended the family of models with density function $g_2$ to include distributions with $q < 0$. By setting $X = (Z - \mu)/\sigma$, a further reparametrization of the density function $g_2$ is given by

$$g_3(z;\mu,\sigma,k) = \frac{k^{k-1/2}}{\sigma\Gamma(k)} \exp\left[\frac{\sqrt{k}\,(z-\mu)}{\sigma} - k \exp\left(\frac{z-\mu}{\sigma\sqrt{k}}\right)\right], \quad -\infty < z < \infty,$$

where $-\infty < \mu < \infty$, $\sigma > 0$, and $k > 0$ are the location, scale, and shape parameters, respectively. Uesaka [15] has demonstrated graphically the relationship between skewness and kurtosis for the density function $g_3$, and has obtained an approximation to the log-gamma distribution by the generalized logistic. Henna [7] has studied the identifiability* of some countable mixtures of the density function $g_3$ (*see* MIXTURE DISTRIBUTIONS), and has obtained a sufficient condition for the identifiability of these mixtures provided that the supports of mixing distributions are well-ordered sets for a total ordering of the parameter space. He has also shown that all finite mixtures of distributions with density $g_3$ are identifiable. Moreover, he has studied the identifiability of countable and finite mixtures of the *reversed log-gamma distribution* with moment generating function $M_x(t) = e^{\mu t}\Gamma(-\sigma t + k)/\Gamma(k)$.

Prentice [10] and Farewell and Prentice [4] have fitted the above distributions to data sets from industrial and medical failure-time studies in order to model distributional shape and to discriminate between special cases. They have also considered maximum likelihood estimation* for regression models based on the density function $g_3$. The usual approach for obtaining the maximum likelihood estimators (MLEs) $\hat{\mu}$, $\hat{\sigma}$, and $\hat{k}$ would be to maximize $\log L(\mu, \sigma, k)$, where $L$ is the likelihood based on a sample $z_1, z_2, \ldots, z_n$. This would be achieved by simultaneously solving the equations $\partial \log L/\partial\mu = 0$, $\partial \log L/\partial\sigma = 0$, and $\partial \log L/\partial k = 0$. However, this presents problems in this case, since it involves the calculation of the derivatives of $\log L$ with respect to $k$. This obstacle is overcome by performing interactions in two stages, treating $k$ as fixed in the first case. For a single value of $k$, one finds the values $\tilde{\mu}(k)$ and $\tilde{\sigma}(k)$ that maximize $\log L$ by solving $\partial \log L/\partial\mu = 0$ and $\partial \log L/\partial\sigma = 0$. By repeating the procedure for different values of $k$, the maximized likelihood function $L_{\max}(k) = L(\tilde{\mu}(k), \tilde{\sigma}(k), k)$ can be determined sufficiently accurately to obtain the MLE $\hat{k}$, which is the value that maximizes $L_{\max}(k)$. Thus, one obtains the MLEs $\hat{\mu} = \tilde{\mu}(\hat{k})$, $\hat{\sigma} = \tilde{\sigma}(\hat{k})$, and $\hat{k}$, and the maximized relative likelihood function $R_{\max}(k) = L_{\max}(k)/L(\hat{\mu}, \hat{\sigma}, \hat{k})$ can be determined. A graph of $R_{\max}(k)$ portrays plausible $k$-values and is useful with likelihood ratio tests.

Balakrishnan and Chan [2] have studied MLEs for $\mu$, $\sigma$, and $k$ under doubly Type II censored samples (*see* PROGRESSIVE CENSORING SCHEMES). They have presented the second derivatives of $\log L$ with respect to all three parameters, so that the Newton–Raphson method can be used to obtain the estimates. They have also derived the expected Fisher information* matrix through which the asymptotic variances and covariances of the MLEs are tabulated for various proportions of censoring.

When the samples are uncensored and $k$ is known, a different picture emerges. Lawless [8, 9] has studied exact inference procedures for parameters, and also for quantiles, when $k$ is known. Such results are important because firstly, good inference procedures are difficult to obtain with $k$ assumed unknown, and secondly, in real situations, a model with a particular value of $k$ often is actually chosen for analysis. In that case, convenient estimators are the MLEs $\tilde{\mu}(k)$ and $\tilde{\sigma}(k)$, obtained by solving the following equations (arising from $\partial \log L/\partial \mu = 0$ and $\partial \log L/\partial \sigma = 0$):

$$\exp(\tilde{\mu}) = \left(\frac{\sum_{i=1}^{n} \exp(z_i/\tilde{\sigma}\sqrt{k})}{n}\right)^{\tilde{\sigma}\sqrt{k}}$$

$$\bar{z} + \frac{\tilde{\sigma}}{\sqrt{k}} = \frac{\sum_{i=1}^{n} z_i \exp(z_i/\tilde{\sigma}\sqrt{k})}{\sum_{i=1}^{n} \exp(z_i/\tilde{\sigma}\sqrt{k})}.$$

These are solved using the given value of $k$. Another pair of convenient estimators are the sample mean $\bar{\mu} = \hat{\mu}$ and the scaled standard deviation $\bar{\sigma} = [\sum_{i=1}^{n}(z_i - \bar{z})^2/n]^{1/2}$, which are also the MLEs of $\mu$ and $\sigma$ for the standard normal distribution ($k = \infty$). To give confidence intervals for the parameters $\mu$, $\sigma$ and the quantiles, Lawless [8] has shown that the $p$th quantile $x_{k,p}$ of the random variable $X$ with density function $g_2$ can be expressed in terms of the $p$th quantile $\chi^2_{(2k),p}$ of a chi-square distribution* with $2k$ degrees of freedom by the formula $x_{k,p} = \sqrt{k}\log[(2k)^{-1}\chi^2_{2k,p}]$. Therefore, the $p$th quantile $z_p$ of the random variable $Z$ with density function $g_3$ is expressed as $z_p = \mu + \sigma x_{k,p}$. By considering the pivotal quantities $W_1 = (\tilde{\mu} - \mu)/\tilde{\sigma}$, $W_2 = \tilde{\sigma}/\sigma$, and $W_p = (\tilde{\mu} - z_p)/\tilde{\sigma}$, noting that $W_p = W_1 - x_{k,p}W_2^{-1}$ and further that the quantities $a_i = (z_i - \tilde{\mu})/\tilde{\sigma}$, $i = 1, 2, \ldots, n$, are ancillary statistics*, confidence intervals and tests for $\mu$, $\sigma$, or $z_p$ can be based on the conditional distributions of $W_1$, $W_2$, and $W_p$ given $\mathbf{a}' = (a_1, \ldots, a_n)$.

Balakrishnan and Chan [1] have studied MLEs under doubly Type II censored samples of the parameters $\mu$ and $\sigma$ of the density function $g_3$ when $k$ is known. They have obtained expressions of the likelihood equations for $\mu$ and $\sigma$, and have given simulated values of the bias, variances, and covariances of the MLEs for various sample sizes, choices of censoring, and $k$-values. They have also derived the expected Fisher information matrix through which the asymptotic variances and covariances of the MLEs are tabulated for various proportions of censoring. Moreover, they have discussed how one is able to construct confidence intervals or carry out tests of hypotheses concerning the parameters $\mu$ and $\sigma$, based on the pivotal quantities $P_1 = \sqrt{n}(\hat{\mu} - \mu)/\hat{\sigma}$ and $P_2 = \sqrt{n}\,\hat{\sigma}/\sigma$. Since the small-sample distributions of $P_1$ and $P_2$ are intractable, they have simulated the percentage points of $P_1$ and $P_2$ (based on 3001 Monte Carlo runs) for sample sizes $n = 20, 25, 40$, various proportions of censoring, and different $k$-values. They have also applied asymptotic normal approximations to the distributions of $P_1$ and $P_2$; the normal approximation to the distribution of $P_2$ is fairly good even for a sample of size 40, but the approximation to the distribution of $P_1$ requires a much larger sample size.

In the statistical literature another distribution is also referred to as a log-gamma distribution. If $Y$ follows a gamma distribution with scale and shape parameters $\theta$ and $k$, respectively, then $X = e^{-Y}$ follows a log-gamma distribution (sometimes it is called a *unit-gamma distribution*; see Ratnaparkhi [11]) with probability density function given by

$$g_4(x;\theta,k) = \frac{\theta^k}{\Gamma(k)} x^{\theta-1}(-\log x)^{k-1},$$

$$0 < x < 1,$$

where $\theta, k > 0$. The form of $g_4$ reduces to the uniform distribution when $\theta = k = 1$ and represents power-function distributions when $\theta > 0$ and $k = 1$. The fact that a suitable choice of $\theta$ and $k$ gives almost any form corresponding to the beta distribution* has led to the density function $g_4$ being considered as an alternative to the beta [6, 11, 15]. Distributional properties of $g_4$ have been given by Grassia [6]; it is useful where inoculation is used to estimate bacteria or virus density in dilution assay with host variability to infection, and could be considered as a prior density in conjunction with the binomial or a zero-truncated binomial distribution.

Taguchi et al. [14] have used the density function $g_4$ in conjunction with income distribution* models, and Schultz [12] has studied it in the context of splitting models as a mass–size distribution. Fosam and Sapatinas [5] have used $g_4$ as a survival distribution in the context of multiplicative damage models* and have obtained characterizations of the Pareto distribution* based on power-type regression functions (*see* CHARACTERIZATIONS OF DISTRIBUTIONS).

## References


[1] Balakrishnan, N. and Chan, P. S. (1995). Maximum likelihood estimation for the log-gamma distribution under type II censored samples and associated inference. In *Recent Advances in Life-Testing and Reliability, A Volume in Honor of Alonzo Clifford Cohen, Jr.*, N. Balakrishnan ed. CRC Press, Boca Raton, Fl., pp. 409–422.

[2] Balakrishnan, N. and Chan, P. S. (1995). Maximum likelihood estimation for the three-parameter log-gamma distribution under type II censoring. In *Recent Advances in Life-Testing and Reliability, A Volume in Honor of Alonzo Clifford Cohen, Jr.*, N. Balakrishnan, ed., CRC Press, Boca Raton, Fl., pp. 439–452.

[3] Bartlett, M. S. and Kendall, D. G. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *J. R. Statistic. Soc. Suppl.*, **8**, 128–138.

[4] Farewell, V. T. and Prentice, R. L. (1977). A study of distributional shape in life testing. *Technometrics*, **19**, 69–75.

[5] Fosam, E. B. and Sapatinas, T. (1995). Characterisations of some income distributions based on multiplicative damage models. *Austra. J. Statist.*, **37**, 89–93.

[6] Grassia, A. (1977). On a family of distributions with argument between 0 and 1 obtained by transformations of the gamma and derived compound distributions. *Austra. J. Statist.*, **19**, 108–114.

[7] Henna, J. (1994). Examples of identifiable mixture. *J. Japan Statist. Soc.*, **24**, 193–200.

[8] Lawless, J. F. (1980). Inference in the generalized gamma and log-gamma distributions. *Technometrics*, **22**, 409–419.

[9] Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* Wiley, New York. (Various parametric models and their associated statistical methods, nonparametric and distribution-free methods, graphical procedures on the most important topics in lifetime data methodology.)

[10] Prentice, R. L. (1974). A log-gamma model and its maximum likelihood estimation. *Biometrika*, **61**, 539–544.

[11] Ratnaparkhi, M. V. (1981). Some bivariate distributions of $(X, Y)$ where the conditional distribution of $Y$, given $X$, is either beta or unit-gamma. In *Statistical Distributions in Scientific Work*, vol. 4, C. Taillie et al., eds. Reidel, Dordrecht, pp. 389–400.

[12] Schultz, D. M. (1975). Mass–size distributions. A review and a proposed new model. In *Statistical Distributions in Scientific Work*, vol. 2, G. P. Patil et al., eds. Reidel, Dordrecht, pp. 275–288.

[13] Shanbhag, D. N., Pestana, D., and Sreehari, M. (1977). Some further results in infinite divisibility. *Math. Proc. Cambridge Phil. Soc.*, **82**, 289–295.

[14] Taguchi, T., Sakurai, H., and Nakajima, S. (1993). A concentration analysis of income distribution model and consumption pattern—introduction of logarithmic gamma distribution and statistical analysis of Engel elasticity. *Statistica*, **LIII**, 33–57.

[15] Uesaka, H. (1989). Some properties of the generalized logistic and log-gamma distributions. *J. Japan Statist. Soc.*, **19**, 1–13.


(GAMMA DISTRIBUTION)

THEOFANIS SAPATINAS

# LYNDEN-BELL ESTIMATOR

Let $X_1, X_2, \ldots, X_N$ be i.i.d. positive random variables with a common cdf $F(\cdot)$, and $Y_1, Y_2, \ldots, Y_N$ be a sequence of such random variables with a common cdf $G(\cdot)$. Here $X_i$ is observable iff $X_i \geqslant Y_i$. The $n$ truncated observations are denoted by $X_i^0, Y_i^0$, $i = 1, \ldots, n$ for

$$n = \sum_{i=1}^{N} I_{\{X_i \geqslant Y_i\}},$$

where $I_{\{\}}$ is the indicator function.

Lynden-Bell [1] in his study of truncated data with applications to astronomy proposed to estimate $F(\cdot)$ and $G(\cdot)$ via

$$1 - \hat{F}_n(t) = \prod_{s \leqslant t} \left(1 - \frac{\Delta L_N(s)}{R_N(s)}\right),$$

$$\hat{G}_n(t) = \prod_{s > t} \left(1 - \frac{\Delta Q_N(s)}{R_N(s)}\right);$$

where $L_N(s) = \sum_{i=1}^{N} I_{\{Y_i \leqslant X_i \leqslant s\}}$, $R_N(s) = \sum_{i=1}^{N} I_{\{Y_i \leqslant s \leqslant X_i\}}$, and $Q_N(s) = \sum_{i=1}^{N} I_{\{Y_i \leqslant s,\, Y_i \leqslant X_i\}}$