Nick T. Thomopoulos

Statistical Distributions

Applications and Parameter Estimates

Springer

Statistical Distributions

Nick T. Thomopoulos

Statistical Distributions

Applications and Parameter Estimates

Springer

Nick T. Thomopoulos
Stuart School of Business
Illinois Institute of Technology
Burr Ridge, IL, USA

ISBN 978-3-319-65111-8 ISBN 978-3-319-65112-5 (eBook)
DOI 10.1007/978-3-319-65112-5

Library of Congress Control Number: 2017949365

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

For my wife, my children, and my grandchildren.

Preface

A statistical distribution is a mathematical function that defines the probable occurrence of a random variable over its admissible space. Understanding statistical distributions is a fundamental requisite to researchers in almost all disciplines. The informed researcher will select the statistical distribution that best fits the data in the study at hand. This book gives a description of the group of statistical distributions that have ample application to studies in statistics and probability. Some of the distributions are well known to the general researcher and are in use in a wide variety of ways. Other useful distributions are less understood and are not in common use. This book describes when and how to apply each of the distributions in research studies, with a goal to identify the distribution that best applies to the study. The distributions are for continuous, discrete, and bivariate random variables. In most studies, the parameter values are not known a priori, and sample data is needed to estimate the parameter values. In other scenarios, no sample data is available, and the researcher seeks some insight that allows the estimate of the parameter values to be gained. This book is easy to read and includes many examples to guide the reader; it will be a highly useful reference to anyone who does statistical and probability analysis. This includes management scientists, market researchers, engineers, mathematicians, physicists, chemists, economists, social science researchers, and students in many disciplines.

Burr Ridge, IL, USA — Nick T. Thomopoulos

Acknowledgments

Thanks especially to my wife, Elaine Thomopoulos, who encouraged me to write this book, and who gave consultation whenever needed. Daniel Sussman assisted in proofing the text. Thanks also to the many people who have helped and inspired me over the years, including some former Illinois Institute of Technology (Illinois Tech) Ph.D. students. I can name only a few here: Emanuel Betinis (National University of Health Science), Fred Bock (IIT Research Institute), Dick Chiapetta (Chiapetta and Welch), Al Endres (Tampa University), John Garofalakis (Patras University), James Hall (Caywood Schiller Associates), Montira Jantaravareerat (Illinois Tech), Arvid Johnson (St. Francis University), Carol Lindee (Panduit), Anatol Longinow (Illinois Tech), Fotis Mouzakis (Frynon Research), George Resnikoff (California State University), and Paul Spirakis (Patras University).

Contents

About the Author

Nick T. Thomopoulos has degrees in Business (B.S.) and in Mathematics (M.A.) from the University of Illinois, and in Industrial Engineering (Ph.D.) from Illinois Institute of Technology. He was supervisor of operations research at International Harvester; senior scientist at IIT Research Institute; Professor in Industrial Engineering, and in the Stuart School of Business at Illinois Tech. He is the author of eleven books including: *Fundamentals of Queuing Systems*, Springer; *Essentials of Monte Carlo Simulation,* Springer; *Applied Forecasting Methods*, Prentice Hall; and *Fundamentals of Production*, *Inventory and the Supply Chain*, Atlantic. He has published many papers, and has consulted in a wide variety of industries in the United States, Europe, and Asia. Nick has received honors over the years, such as the *Rist Prize* from the Military Operations Research Society for new developments in queuing theory; the *Distinguished Professor Award* in Bangkok, Thailand, from the Illinois Tech Asian Alumni Association; and the *Professional Achievement Award* from the Illinois Tech Alumni Association.

Chapter 1
Statistical Concepts

1.1 Introduction

A statistical distribution is a mathematical function that defines how outcomes of an experimental trial occur randomly in a probable way. The outcomes are called random variables, and their admissible region lies in a specified sample space that is associated with each individual distribution. The statistical distributions are mostly of two type: continuous and discrete. The continuous probability distributions apply when the random variable can fall anywhere between two limits, such as the amount of rain-water that accumulates in a five-gallon container after a rainfall. The discrete probability distribution pertains when the outcomes of the experiment are specific values, like the number of dots that appear on a roll of two dice. The distributions may also be classified as univariate or multivariate. The univariate is when the distribution has only one random variable; multivariate is when two or more random variables are associated with the distribution. The statistical distributions in this book pertain to the commonly used univariate continuous and discrete probability distributions, and to the most frequently applied bivariate continuous statistical distributions, where bivariate distributions have two jointly related random variables.

1.1.1 Probability Distributions, Random Variables, Notation and Parameters

The distributions with the designation and parameters on each are listed below:

Continuous Distributions:

Continuous Uniform	$x \sim CU(a,b)$
Exponential	$x \sim Exp(\theta)$

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_1

Erlang	$x \sim Erl(k, \theta)$
Gamma	$x \sim Gam(k, \theta)$
Beta	$x \sim Beta(k_1,k_2,a,b)$
Weibull	$x \sim We(k_1,k_2,\gamma)$
Normal	$x \sim N(\mu,\sigma^2)$
Lognormal	$x \sim LN(\mu_y,\sigma_y^2)$
Left-Truncated Normal	$t \sim LTN(k)$
Right-Truncated Normal	$t \sim RTN(k)$
Triangular	$x \sim TR(a, b, \tilde{x})$

Discrete Distributions:

Discrete Uniform	$x \sim DU(a,b)$
Binomial	$x \sim Bin(n,p)$
Geometric	$x \sim Ge(p)$
Pascal	$x \sim Pa(k,p)$
Hyper Geometric	$x \sim HG(n,N,D)$
Poisson	$x \sim Po(\theta)$

Bivariate Distributions:

Bivariate Normal	$x_1,x_2 \sim BVN(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$
Bivariate Lognormal	$x_1,x_2 \sim BVLN(\mu_{y1},\mu_{y2},\sigma_{y1},\sigma_{y2},\rho_y)$

The continuous distributions, with a brief on each, are the following:

Continuous uniform:	density is horizontal throughout.
Exponential:	density peaks at zero and tails down thereafter.
Erlang:	many shapes ranging from exponential to normal.
Gamma:	many shapes ranging from exponential to normal.
Beta:	many shapes that skew left, right, bathtub and symmetrical.
Weibull:	many shapes from exponential to normal.
Normal:	symmetrical bell shaped.
Lognormal:	peaks near zero and skews far to right.
Left-truncated normal:	normal is truncated on left and skews to right.
Right-truncated normal:	normal is truncated on right and skews to left.
Triangular:	density ramps to a peak, and then down to zero.

The discrete distributions, with a brief on each, are the following:

Discrete uniform:	probability is horizontal throughout.
Binomial:	n trials with constant probability of success per trial.
Geometric:	number of trials till a success.
Pascal:	number of trials till k successes.
Poisson:	number of events when event rate is constant.
Hypergeometric:	n samples without replacement from a lot of size N.

The bivariate distributions, with a short brief, are the following:

Bivariate normal: the marginal distributions are normally shaped.
Bivariate lognormal: the marginal distributions are lognormal.

1.2 Fundamentals

This chapter describes the commonly used properties pertaining to the continuous and the discrete statistical distributions. These are the probability functions, the mean, variance, standard deviation, mode and median. When sample data is available, it is used to aid the analyst to estimate the parameter values of the statistical distribution under study. Sample estimates of the measures are the min, max, average, variance, standard deviation, mode and median.

1.3 Continuous Distribution

Admissible Range The continuous distribution has a random variable, x, with an admissible range as follows:

$a \leq x \leq b$

where a could be minus infinity and b could be plus infinity.

Probability Density The probability density function of x is:

$f(x) \geq 0 \qquad a \leq x \leq b$

where

$$\int_a^b f(x)dx = 1.0$$

Cumulative Distribution The cumulative distribution function of x is:

$$F(x) = \int_a^x f(w)dw$$

This gives the cumulative probability of x less or equal to x_o, say, as below:

$$F(x_o) = P(x \leq x_o) = \int_a^{xo} f(x)dx$$

Complementary Probability The complementary probability at x greater than x_o is obtained as follows:

$$H(x_o) = 1 - F(x_o) = P(x > x_o)$$

Expected Value The expected value of x, E(x), also called the *mean* of x, μ, is derived as below:

$$\mu = E(x) = \int_a^b xf(x)dx$$

Variance and Standard Deviation The variance of x, σ^2, is obtained in the following way:

$$\sigma^2 = E\left[(x - \mu)^2\right] = \int_a^b (x - \mu)^2 f(x)dx$$

The ***standard deviation***, σ, is merely,

$$\sigma = \sqrt{\sigma^2}$$

Median The median on x, denoted by $\mu_{0.5}$, is the value of x with cumulative probability of 0.50 as shown below:

$$F(\mu_{0.5}) = 0.5$$

Mode The mode, $\widetilde{\mu}$, is the most-likely value of x and is located where the probability density is highest in the admissible range, as the following:

$$f\left(\widetilde{\mu}\right) = \max\{f(x) \;\; a \leq x \leq b\}$$

The *α-percent-point* of x, denoted as xα, is obtained by the inverse function of F (x), where,

$$F(x\alpha) = \alpha$$

Note, for example, the median of x is $x_{0.50} = \mu_{0.5}$.

Coefficient-of-Variation The coefficient-of-variation, cov, of x is the ratio of the standard deviation over the mean, as below:

$$\text{cov} = \sigma/\mu$$

1.4 Discrete Distributions

Admissible Range The discrete distribution has a random variable, x, with admissible range,

$$a \leq x \leq b$$

For simplicity, the range in this book is limited to vary in increments of one, i.e.,

$$\text{range} = (a, a+1, \ldots.., b-1, b)$$

Probability Function The probability function of x, P(x), is below:

$$P(x) \geq 0 \qquad a \leq x \leq b$$

where

$$\sum_{a}^{b} P(x) = 1.0$$

Cumulative Probability The cumulative probability function of x, F(x), is the following:

$$F(x) = \sum_{a}^{x} P(w)$$

This gives the cumulative probability of $x = x_o$ or less, as below:

$$F(x_o) = P(x \leq x_o) = \sum_{a}^{xo} P(x)$$

Complementary Probability The complementary probability of x_o is the probability of x larger than x_o as follows:

$$H(x_o) = 1 - F(x_o) = P(x > x_o)$$

Expected Value and Mean The expected value of x, E(x), also called the mean of x, μ, is derived as below:

$$\mu = E(x) = \sum_{a}^{b} xP(x)$$

Variance and Standard Deviation The variance of x, σ^2, is obtained in the following way:

$$\sigma^2 = E\left[(x - \mu)^2\right] = \sum_{a}^{b} (x - \mu)^2 P(x)$$

The standard deviation, σ, is computed below:

$$\sigma = \sqrt{\sigma^2}$$

Median The median on x, denoted by $\mu_{0.5}$, is the value of x with cumulative probability of 0.50 as shown below:

$$F(\mu_{0.5}) = 0.5$$

Mode The mode**,** $\widetilde{\mu}$, is the most-likely value of x and is located where the probability is highest in the admissible range, as the following:

$$P(\widetilde{\mu}) = \max\{P(x) \quad a \leq x \leq b\}$$

Lexis Ratio The Lexis Ratio, τ, of x is the ratio of the variance over the mean, as below:

$$\tau = \frac{\sigma^2}{\mu}$$

1.5 Sample Data Basic Statistics

When n sample data, $(x_1, \ldots, x_n)$, is collected, various statistical measures can be computed as described below:

n = sample size
$x(1)$ = minimum of $(x_1, \ldots, x_n)$
$x(n)$ = maximum of $(x_1, \ldots, x_n)$
$\bar{x} = \sum_{i=1}^{n} x_i/n$ = sample average
$s^2 = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2/(n-1)$ = sample variance
$s = \sqrt{s^2}$ = sample standard deviation
$cov = s/\bar{x}$ = sample coefficient of variation
$\tilde{x}$ = sample mode
$x_{0.5}$ = sample median
$\tau = s^2/\bar{x}$ = sample lexis ratio

The *sample median* is the mid value of the sorted data set $(x_1, \ldots, x_n)$. The sorted entries are listed as: $x(1), x(2), \ldots, x(n)$. If the number of samples is odd, the sample median is:

$$x_{0.5} = x\left(n^{'}\right) \text{ where } n^{'} = (n+1)/2.$$

If n is even, the median is:

$$x_{0.5} = [x(n/2) + x(1+n/2)]/2.$$

The sample mode, $\tilde{x}$, is the most frequent data value from the sample data. Sometimes two or more modes may appear, and sometime no mode is found. To find the mode, the analyst should sort the data and choose the value that appears the most. If no value appears more than others, the data could be grouped and the average of the group with most entries could be chosen as the mode.

1.6 Parameter Estimating Methods

When an analyst wants to apply a probability distribution in a research study, often the parameter value(s) are not known and need to be estimated. The estimates are generally obtained from sample data $(x_1, \ldots, x_n)$ that has been collected. Two popular methods to estimate the parameter from the data entries are the maximum-likelihood-estimator and the method-of-moments. A brief description on each follows.

1.6.1 Maximum-Likelihood-Estimator (MLE)

This method formulates a likelihood function using sample data ($x_1, \ldots, x_n$) and the parameter(s) of the distribution under study, and seeks the value of the parameter (s) that maximize this function. For example, when a statistical distribution has one parameter, θ, the search is for the value of θ that maximizes the likelihood the n samples would have produced those numbers, and this value is called the maximum-likelihood estimate.

1.6.2 Method-of-Moments (MoM)

This method finds the theoretical moments of a distribution from the probability function, and the counterpart sample moments from the sample data ($x_1, \ldots, x_n$) are obtained from the same probability function. A combination of the theoretical moments produces the population parameters (μ, σ, θ, etc.). Substituting the corresponding sample moments into the theoretical moments yields the sample estimates of the parameters estimates ($\bar{x}$, s, $\widehat{\theta}$, etc.), and this estimate is called the method-of-moments estimate.

1.7 Transforming Variables

It is sometimes necessary to convert the original sample data ($x_1, \ldots, x_n$) in a manner that allows an easier identification of the probability distribution that best fits the sample data. Two methods are especially helpful, the conversion to a data set of zero and larger, and the conversion to a set of zero and one. Both of these methods are described below.

1.7.1 Transform Data to Zero or Larger

The analyst with sample data ($x_1, \ldots, x_n$) may find it useful to convert the data to a new set where all the entries are zero or larger. This is by applying:

$$y = x - x(1)$$

to each entry, and where x(1) is the min value of the x data set. The new set of n data entries become: ($y_1, \ldots, y_n$). The mean and standard deviation of the new y set of data become the following:

$$\bar{y} = \bar{x} - x(1)$$

$$s_y = s_x$$

Sometimes the coefficient-of-variation, cov_y, from the y set of data is needed to help identify the probability distribution that fits the sample data. The coefficient-of-variation for the new y set of data becomes:

$$cov_y = s_y/\bar{y}$$

1.7.2 Transform Data to Zero and One

Sometimes it is advantageous to convert the sample data to a new set that lies within the range of zero to one. This is accomplished by applying the following:

$$w = [x - x(1)]/[x(n) - x(1)]$$

to each entry in the sample data set, and where x(1) is the min value, and x(n) the max value. The new set of data becomes: $(w_1, \ldots, w_n)$. This method yields the mean and standard deviation of w as below:

$$\bar{w} = [\bar{x} - x(1)/[x(n) - x(1)]$$

$$s_w = s_x/[x(n) - x(1)]$$

The coefficient-of-variation from the w data set becomes:

$$cov_w = s_w/\bar{w}$$

With w lying within (0, 1), the cov that emerges is sometimes useful to identify the probability distribution of the data set.

Example 1.1 Consider the sample data with 11 entries listed as: $(x_1, \ldots, x_{11}) = (23, 14, 26, 31, 27, 22, 15, 17, 31, 29, 34)$. The basic statistical measures of the data are listed below:

$n = 11$
$x(1) = 14$
$x(11) = 34$
$\bar{x} = 24.45$
$s_x^2 = 46.87$
$s_x = 6.85$
$cov_x = 0.28$
$\tilde{x} = 31$
$x_{0.5} = 26$

Example 1.2 Assume the analyst of the data in Example 1.1 wants to convert the data to yield a new set where $y \geq 0$. This is accomplished by taking the min = x (1) = 14, and applying y = (x – 14) to each entry. The set of 11 data entries with variable x and the counterpart with variable y are listed below:

$$(x_1, \ldots, x_{11}) = (23, 14, 26, 31, 27, 22, 15, 17, 31, 29, 34)$$
$$(y_1, \ldots, y_{11}) = (9, 0, 12, 17, 13, 8, 1, 3, 17, 15, 20)$$

The basic statistical measures for the revised data set of y are listed below. Note, the only measures of the y set that remain the same as the x set are the number, variance and the standard deviation.

$n = 11$
$y(1) = 0$
$y(11) = 20$
$\bar{y} = 10.45$
$s_y^2 = 46.87$
$s_y = 6.85$
$cov_y = 0.65$
$\tilde{y} = 17$
$y_{0.5} = 12$

Example 1.3 Suppose the analyst of the data from Example 1.1 wants to transform the data to lie between 0 and 1. This is achieved by taking the min = x(1) = 14 and the max = x(11) = 34 values and applying w = (x − 14)/(34 − 14) to each of the entries. Below lists the original values of x and the transformed set of w.

$$(x_1, \ldots, x_{11}) = (23, 14, 26, 31, 27, 22, 15, 17, 31, 29, 34)$$
$$(w_1, \ldots, w_{11}) = (0.45, 0.00, 0.60, 0.85, 0.65, 0.40, 0.05, 0.15, 0.85, 0.75, 1.00)$$

The basic statistical measures are the following:

$n = 11$
$w(1) = 0.00$
$w(11) = 1.00$
$\bar{w} = 0.52$
$s_w^2 = 0.34$
$s_w = 0.65$
$cov_w = 1.25$
$\tilde{w} = 0.85$
$w_{0.5} = 0.60$

1.7.3 Continuous Distributions and Cov

When an analyst has sample data and is seeking the statistical distribution to apply to the data, the coefficient-of-variation (cov) is sometimes helpful in determining the continuous distributions that may best fit sample data. Below lists some candidate distributions for selected values of the cov.

When cov $\geq$ 1.00: Exponential, Erlang ($k_1 = 1$), gamma ($k_1 \leq 1.00$); Weibull ($k_1 \leq 1.00$), lognormal, left-truncated normal, right-truncated normal and beta.

When cov $\leq$ 0.33: Normal, Erlang ($k_1 = 9$), gamma ($k_1 \geq 9.00$), Weibull ($k_1 \geq 4.00$) and beta.

When cov between (0.33 and 1.00): Continuous uniform, Erlang, gamma, beta, left-truncated normal, right-truncated normal and Weibull.

1.7.4 Discrete Distributions and Lexis Ratio

When a researcher has discrete sample data and is seeking the discrete distribution that best fits the data, the lexis ratio is sometimes helpful in identifying the candidate distributions to use. Below is a list of the lexis ratio values and the candidate distributions:

When $\tau < 1$: Binomial.
When $\tau = 1$: Poisson.

When $\tau \geq 1$: Geometric with number fails as the variable; and Pascal with number of fails as the variable.

1.8 Summary

The univariate probability distributions have one variable and are of two type, continuous and discrete. The bivariate normal and bivariate lognormal distributions each have two jointly related variables. The continuous distributions have a probability function that defines how the outcomes are possible in a probability manner. The discrete have a probability mass function that specifies the probability of each particular outcome in the sample space. The common statistical measures are the mean, variance, standard deviation, median, mode, coefficient of variation, and lexis ratio. When the parameter values are not known and sample data is available, estimates of the statistical measures are computed.

Chapter 2
Continuous Uniform

The continuous uniform distribution applies when the random variable can fall anywhere equally likely between two limits. The distribution is often called when an analyst does not have definitive information on the range and shape of the random variable. For example, management may estimate the time to finish a project is equally likely between 50 and 60 h. A baseball is hit for a homerun and the officials estimate the ball traveled somewhere between 410 and 430 feet. The amount of official snowfall at a location on a wintry day is predicted between 1 and 5 inches. The chapter lists the probability density, cumulative distribution, mean, variance and standard deviation of the random variable. Also described is the α-percent-point of x that identifies the value of x where the cumulative probability is α. When the parameter limit values are not known, and sample data is available, estimates of the parameter values are obtained. Two estimates are described, one by way of the maximum-likelihood method, and the other by the method-of-moments. When sample data is not available, experts are called to obtain the estimates. Often both of the limits are unknown and estimates on both are needed. Sometimes only the low limit is known, and on other occasions, only the upper limit is unknown. The way to estimate the parameter values is described for all three scenarios.

2.1 Fundamentals

The random variable, x, of the continuous uniform distribution, has an admissible range of (a to b), where any value in this range is equally likely to occur. The parameters of the distribution define the range interval and are:

$a = \min$
$b = \max$

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_2

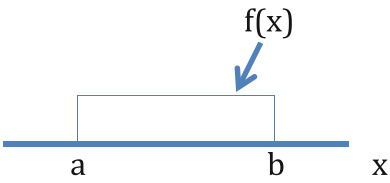

Fig. 2.1 The continuous uniform density

The probability density is below, and a depiction is in Fig. 2.1.

$$f(x) = 1/(b - a) \qquad a \leq x \leq b$$

The mean, variance and standard deviation are listed below:

$$\mu = (a + b)/2$$
$$\sigma^2 = (b - a)^2/12$$
$$\sigma = (b - a)/\sqrt{12}$$

The coefficient-of-variation is obtained as below:

$$cov = \sigma/\mu$$

In the special case when $a = 0$ and $b = 1$, cov becomes

$$cov = 2/\sqrt{12} = 0.578$$

The cumulative distribution function, F(x), becomes:

$$F(x) = (x - a)/(b - a) \qquad a \leq x \leq b$$

The α-percent-point on x, denoted as xα, is related to the cumulative probability, α, as below:

$$P(x \leq x\alpha) = \alpha$$

To find the value of xα for the continuous uniform distribution, apply the following:

$$x\alpha = a + \alpha(b - a)$$

Example 2.1 In a production system, the amount of liquid poured into a container varies and could be anywhere from 7.0 to 7.2 ounces. Hence, for a randomly sampled container, the amount of liquid in the container, noted as x, has an admissible range of: $7.0 \leq x \leq 7.2$. The probability density becomes:

$$f(x) = 1/(7.2 - 7.0) = 5.0 \qquad 7.0 \leq x \leq 7.2$$

The average amount of liquid in a container is $\mu = (7.0 + 7.2)/2 = 7.10$. The variance becomes $\sigma^2 = (7.2–7.0)^2/12 = 0.0033$; and the standard deviation is $\sigma = \sqrt{0.0033} = 0.0577$.

The cumulative probability distribution for any value of x in the admissible range becomes:

$$F(x) = (x - 7.0)/(0.20) \qquad 7.0 \leq x \leq 7.2$$

Note, for example, the probability of x less or equal to 7.15 is obtained as below:

$$P(x \leq 7.15) = F(7.15) = (7.15 - 7.00)/(0.20) = 0.75$$

As an example, the 0.25%-point value of x is computed as follows:

$$x_{.25} = 7.0 + 0.25(7.20 - 7.00) = 7.05$$

2.2 Sample Data

When an analyst wants to apply the continuous uniform distribution in a study and the parameters values, (a, b), are not known, sample data is used to estimate the values of the parameters. The sample data are n randomly drawn observations denoted as: $(x_1, \ldots, x_n)$. To apply the estimates, the following statistics are drawn from the sample data:

$\bar{x}$ = average
s = standard deviation
x(1) = min
x(n) = max

Example 2.2 An experiment yields the following n = 10 observations: (9.1, 3.1, 17.1. 15.8, 12.6. 5.9, 5.1, 14.2, 19.8, 7.3). The analyst assumes the data comes from a continuous uniform distribution and would like to find the mid 50% interval of values. An initial step requires estimates of the parameters, (a, b); and to achieve, the first task is to measure the stats from the data. These are:

$\bar{x} = 11.00$
s = 5.69
x(1) = 3.1
x(n) = 19.8

2.3 Parameter Estimates from Sample Data

Two ways to estimate the parameters are available. One is by the maximum-likelihood, and the other is from the method-of-moments.

The parameter estimates using the maximum-likelihood estimate method become:

$$\hat{a} = \mathrm{x}(1)$$
$$\hat{b} = \mathrm{x}(\mathrm{n})$$

The parameter estimates from the method-of-moments are below:

$$\hat{a} = \bar{x} - \sqrt{12}\mathrm{s}/2$$
$$\hat{b} = \bar{x} + \sqrt{12}\mathrm{s}/2$$

Example 2.3 Continuing with Example 2.2, the estimates for the maximum-likelihood method and by the method-of-moment method are below:

Using the maximum-likelihood method, the parameter estimates become:

$$\hat{a} = \mathrm{x}(1) = 3.1$$
$$\hat{b} = \mathrm{x}(\mathrm{n}) = 19.8$$

The mid-50% interval is computed as below:

$$(\mathrm{x}_{.25} \leq \mathrm{x} \leq \mathrm{x}_{.75})$$

where $\mathrm{x}_{.25}$ and $\mathrm{x}_{.75}$ are the 0.25% and 0.75%-point values of x, respectively. Using the maximum-likelihood estimate method, these become:

$$\mathrm{x}_{.25} = \hat{a} + 0.25\left(\hat{b} - \hat{a}\right) = 3.1 + 0.25(19.8 - 3.1) = 7.27$$
$$\mathrm{x}_{.75} = \hat{a} + 0.75\left(\hat{b} - \hat{a}\right) = 3.1 + 0.75(19.8 - 3.1) = 15.62$$

Hence, the mid-50% interval becomes:

$$(7.27 \leq \mathrm{x} \leq 15.62)$$

and the associated probability estimate on the range is:

$$\mathrm{P}(7.27 \leq \mathrm{x} \leq 15.62) = 0.50$$

Using the data from Example 2.2, the parameter estimates by use of the method-of-moments method are computed below:

$$\hat{a} = \bar{x} - \sqrt{12}s/2 = 11.00 - \sqrt{12} \times 5.69/2 = 1.14$$
$$\hat{b} = \bar{x} + \sqrt{12}s/2 = 11.00 + \sqrt{12} \times 5.69/2 = 20.86$$

Applying the method-of-moment estimates, the mid-50% range is obtained as below:

$$x_{.25} = \hat{a} + 0.25\left(\hat{b} - \hat{a}\right) = 1.14 + 0.25(20.86 - 1.14) = 6.07$$
$$x_{.75} = \hat{a} + 0.75\left(\hat{b} - \hat{a}\right) = 1.14 + 0.75(20.86 - 1.14) = 15.69$$

2.4 Parameter Estimates When No Data

Consider the situation when an analyst wishes to apply the continuous uniform distribution but has no estimates on the parameters (a, b) and has no sample data to draw the estimates. In this situation, the analyst seeks advise from experts who give some estimates on the range of the variable, x.

2.5 When (a, b) Not Known

When both parameter values are not known, the analyst seeks an expert who gives two estimates on the percent-points of x, denoted as: ($x_1 = \alpha_1$-percent-point, and $x_2 = \alpha_2$-percent-point). Note:

$$P(x < x_1) = \alpha_1$$
$$P(x < x_2) = \alpha_2$$

Should $\alpha_1 = 0.0$, the estimate on the min is $\hat{a} = x_1$; and if $\alpha_2 = 1.0$, the estimate on the max is $\hat{b} = x_2$.

Below shows how to estimate the min and max parameter values when $\alpha_1 \geq 0.0$, $\alpha_2 \leq 1.0$ and $\alpha_1 < \alpha_2$. Using x_1, x_2, α_1 and α_2, the estimates of the parameters are obtained as shown below. First observe how x_1 and x_2 are related to the parameters (a, b):

$$x_1 = a + \alpha_1(b - a)$$
$$x_2 = a + \alpha_2(b - a)$$

Also note, the equivalent relations below:

$$x_1 = a(1 - \alpha_1) + \alpha_1 b$$
$$x_2 = a(1 - \alpha_2) + \alpha_2 b$$

Now using some algebra, the estimates on the parameters (a, b) become:

$$\hat{a} = [x_2\alpha_1 - x_1\alpha_2)]/[\alpha_1 - \alpha_2]$$
$$\hat{b} = [x_2 - \hat{a}(1 - \alpha_2)]/\alpha_2$$

Example 2.4 Suppose an analyst wants to use the continuous uniform distribution on some data where experts give estimate values of $x_1 = 10$ and $x_2 = 100$.

If $\alpha_1 = 0.0$ and $\alpha_2 = 1.0$, the above equations yield:

$$\hat{a} = [100 \times 0.0 - 10 \times 1.0]/[0.0 - 1.0] = 10$$
$$\hat{b} = [100 - 10(1 - 1.0)]/1.0 \quad = 100$$

If $\alpha_1 = 0.0$ and $\alpha_1 = 0.8$, the estimates become:

$$\hat{a} = [100 \times 0.0 - 10 \times 0.8]/[0.0 - 0.8] \quad = 10$$
$$\hat{b} = [100 - 10(1 - 0.8)]/0.8 \quad = 122.5$$

If $\alpha_1 = 0.2$ and $\alpha_2 = 1.0$, the estimates become:

$$\hat{a} = [100 \times 0.2 - 10 \times 1.0]/[0.2 - 1.0] \quad = -12.5$$
$$\hat{b} = [100 - (-12.5)(1 - 1.0)]/1.0 \quad = 100$$

If $\alpha_1 = 0.2$ and $\alpha_2 = 0.8$, the estimates become:

$$\hat{a} = [100 \times 0.2 - 10 \times 0.8/[0.2 - 0.8] \quad = -20$$
$$\hat{b} = [100 - (-20)(1 - 0.8)]/0.8 \quad = 130$$

Example 2.5 In a manufacturing system, management needs as estimate on the time to design the configuration on a new product. Inquiring with the engineers, the estimate of the mid-50% interval will be 100–120 h. This implies, the 0.25% point and the 0.75% point are the following:

$$x_{.25} = 100$$
$$x_{.75} = 120$$

Also,

$$\alpha_1 = 0.25$$
$$\alpha_2 = 0.75$$

So now, the estimates of the parameter values are computed as below:

$$\hat{a} = [120 \times 0.25 - 100 \times 0.75]/[0.25 - 0.75] = 90$$
$$\hat{b} = [120 - 90\ (1 - 0.75)]/0.75 = 130$$

Finally, the estimate of the time to complete the design is 90–130 h.

2.6 Summary

The continuous uniform distribution is called when the random variable can equally fall anywhere equally likely between two limits. The distribution is often used when the analyst has little information on the distribution. Estimates on the limits, (a, b), are needed to apply. When sample data is available, the estimates of the parameter values are readily computed, either by the maximum-likelihood method, or by the method-of-moments. When sample data is not available, expert knowledge is taken from which the estimates are obtained.

Chapter 3
Exponential

The exponential distribution peaks when the random variable is zero and gradually decreases as the variable value increases. The distribution has one parameter and has easy computations on the probability density and cumulative distribution. The distribution has a memory-less property where the probability of the next event occurring in a fixed interval is the same no matter the start time of the interval. This distribution is fundamental in queuing theory since it is used as the variable for the time between arrivals to a system, and also the time of service. The distribution also applies in studying reliability where it is assigned as the time to fail for an item. When the parameter value is not known, sample data is used to obtain an estimate, and when no sample data is available, an approximate measure on the distribution allows the analyst to estimate the parameter value.

3.1 Fundamentals

The exponential random variable, x, has an admissible range of zero and larger, where the occurrence peaks at zero and steadily declines as x increases. The probability density, f(x), with one parameter, θ, is listed below, and a depiction is given in Fig. 3.1.

$$f(x) = \theta e^{-\theta x} \qquad x \geq 0$$

The mean, variance and standard deviation of x are listed below:

$$\mu = 1/\theta$$
$$\sigma^2 = 1/\theta^2$$
$$\sigma = 1/\theta$$

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_3

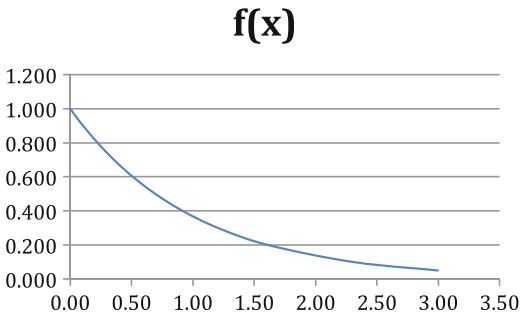

Fig. 3.1 The probability density of the exponential distribution when $\theta = 1$

Because the mean and standard deviation are the same, the coefficient-of-variation is as follows:

$$\text{cov} = \sigma/\mu = 1.0$$

The cumulative probability of x is listed below:

$$F(x) = 1 - e^{(-\theta x)} \qquad x \geq 0$$

Note, at x = x‘:

$$F(x') = P(x \leq x')$$

and thereby:

$$F(0) = P(x' = 0) = 0$$

and

$$F(\infty) = P(x' \leq \infty) = 1$$

The α-percent –point on the distribution, denoted as x_α, is the value of x where the probability of x less or equal to x_α is α, as the following:

$$P(x \leq x_\alpha) = \alpha$$

The value of xα is obtained as below:

$$x_\alpha = -(1/\theta)\ln(1 - \alpha)$$

where ln is the natural logarithm.

Example 3.1 Customers arrive to a store with the average time between arrivals in 5 min via an exponential distribution with random variable x. Hence,

$\mu = 5$ = average minutes between arrivals,
$\theta = 1/5 = 0.2$ = average number of arrivals per minute

The probability density of x is below:

$$f(x) = 0.2e^{-0.2x} \qquad x \geq 0$$

and the cumulative probability distribution is:

$$F(x) = 1 - e^{-0.2x} \qquad x \geq 0$$

Note for example, at $\alpha = 0.5$,

$$x_{0.5} = (-1/0.2)\ln(1 - 0.5) = 3.47 \text{ min}$$

and thereby,

$$F(3.47) = P(x \leq 3.47) = 0.50$$

Also, at $\alpha = 0.9$,

$$x_{0.9} = (-1/0.2)\ln(1 - 0.9) = 11.51 \text{ min}$$

and,

$$F(11.51) = P(x \leq 11.51) = 0.90$$

3.2 Table Values

Table 3.1 lists comparative values of the $x\alpha = \alpha$-percent –point when α varies from 0.00 to 0.95, and θ is from 0.1 to 10. Also listed is the corresponding value of the mean of x, μ, for each value of θ. Note when $\theta = 1$, the low to high values in this range is 0.00 to 3.00. For $\theta < 1$, the range increases, and when $\theta > 1$, the range decreases. Fig. 3.1 depicts the distribution when $\theta = 1$.

Table 3.1 The body lists the α-percent-points, xα, for exponential parameter $\theta =$ (0.1 to 10), α = (0.00 to 0.95) and μ = (10 to 0.1).

	θ	0.1	0.5	1	5	10
α	μ	10	2	1	0.2	0.1
0.00		0.00	0.00	0.00	0.00	0.00
0.05		0.51	0.10	0.05	0.01	0.01
0.10		1.05	0.21	0.11	0.02	0.01
0.20		2.23	0.45	0.22	0.04	0.02
0.30		3.57	0.71	0.36	0.07	0.04
0.40		5.11	1.02	0.51	0.10	0.05
0.50		6.93	1.39	0.69	0.14	0.07
0.60		9.16	1.83	0.92	0.18	0.09
0.70		12.04	2.41	1.20	0.24	0.12
0.80		16.09	3.22	1.61	0.32	0.16
0.90		23.03	4.61	2.30	0.46	0.23
0.95		29.96	5.99	3.00	0.60	0.30

3.3 Memory-Less Property

The exponential distribution has a memory-less property. This is because the probability of the next event occurring after time interval x, is the same no matter when the start time of the interval happens. The start time could be right after the last event, or any time after. To show this relation, note where the probability of the time being larger than x is the following:

$$\begin{aligned} \mathrm{H(x)} &= 1 - \mathrm{F(x)} \\ &= 1 - \left[1 - e^{-\theta x}\right] \\ &= e^{-\theta x} \end{aligned}$$

Suppose a time interval of length $\mathrm{x_o}$ passes after the last event, and the probability of the next event occurring after the next interval length x is sought. The conditional probability that the arrival time is larger than x for the next event, given the starting point since the last event, ($\mathrm{x_o}$ + x), is the following:

$$\begin{aligned} \mathrm{H(x|x_o)} &= \mathrm{H(x_o + x)/H(x_o)} \\ &= e^{-\theta(x_o+x)}/e^{-\theta x_o} \\ &= e^{-\theta x} \end{aligned}$$

Since,

$$\mathrm{H(x)} = \mathrm{H(x|x_o)} = e^{-\theta x}$$

the probability is the same no matter where the starting point is, and thereby the distribution has a memory-less property.

Example 3.2 In Example 3.1, the average time between arrivals is $\mu = 5$ min, and the exponential parameter is $\theta = 0.2$. Also recall, the probability that x is 11.51 min or less is 0.90. Hence, the probability of the next event being larger than 11.51 min is:

$$\begin{aligned} \mathrm{P}(\mathrm{x} > 11.51) &= 1 - \mathrm{F}(11.51) \\ &= 1 - 0.90 \\ &= \mathrm{H}(11.51) \\ &= 0.10 \end{aligned}$$

Suppose after $\mathrm{x_o} = 2.0$ min elapses, the analyst starts counting the time till the next event, and of interest is to find the probability that the next event will occur after the next 11.51 min. Note, the time after the last event is (2.00 + 11.51) = 13.51 min. The probability sought is computed as below:

$$\begin{aligned} \mathrm{H}(13.51|2.00) &= \mathrm{H}(13.51)/\mathrm{H}(2.00) \\ &= e^{-(0.2\times 13.51)}/e^{-(0.2\times 2.00)} \\ &= 0.067/0.670 \\ &= 0.10 \end{aligned}$$

Hence,

$$\mathrm{H}(13.51|2.00) = \mathrm{H}(11.51) = 0.10$$

3.4 Poisson Relation

The Poisson distribution is described in Chap. 17. The random variable of the Poisson, n, is the number of events that will occur during a specified time interval when the rate of occurrence, θ, is constant. When the parameter of the exponential is also the same θ, the random variable, x, is the time between Poisson events occurring. This relation between the Poisson and the exponential plays an important role in the development for queuing theory.

The probability of n from the Poisson is below:

$$\mathrm{P}(\mathrm{n}) = \theta^n e^{-\theta}/n! \qquad n = 0, 1, \ldots$$

and the probability density of the exponential is:

$$\mathrm{f}(\mathrm{x}) = \theta e^{-\theta x} \qquad x \geq 0$$

Also, the expected value of each of the two variables is below:

$$E(n) = \theta$$
$$E(x) = 1/\theta$$

Example 3.3 Continuing with Example 3.1 where x is exponential and represents the minutes between arrivals with $\mu = 5.0$ and parameter $\theta = 0.20$. Since the exponential is related to the Poisson distribution, the random variable becomes n representing the number of arrivals in a minute. The probability of n becomes:

$$P(n) = e^{-0.2}0.2^n/n! \qquad n = 0, 1, 2, \ldots$$

and the expected value of n is below:

$$E(x) = 0.2$$

Note,

$$P(0) = e^{-0.2}0.2^0/0! = 0.819$$
$$P(1) = e^{-0.2}0.2^1/1! = 0.164$$
$$P(2) = e^{-0.2}0.2^2/2! = 0.016$$
$$P(3) = e^{-0.2}0.2^3/3! = 0.001$$

3.5 Sample Data

When an analyst wants to apply the exponential distribution in a study and the parameter value, θ, is not known, sample data can be used to estimate the value of the parameter. The sample data are n randomly drawn observations denoted as: (x_1, . . ., x_n). To apply the estimate, the following statistic is drawn from the sample data:

$\bar{x}$ = average

3.6 Parameter Estimate from Sample Data

Using the sample average, $\bar{x}$, the estimate of the parameter value is obtained as shown below:

$$\widehat{\theta} = 1/\bar{x}$$

Example 3.4 In a study to determine the air traffic at a local airport, an analyst collects the following n = 10 aircraft inter-arrival times: 0.5, 1.7, 2.2, 1.9, 4.3, 8.3, 0.4, 1.8, 3.8, 2.7. The average, standard deviation and coefficient-of-variation are listed below:

$\bar{x} = 2.80$
$\mathrm{s} = 2.27$
$\mathrm{cov} = 0.81$

Note, the sample cov $= \mathrm{s}/\bar{x} = 0.81$, is reasonably close to an exponential theoretical cov $= \sigma/\mu = 1.00$.

The estimate of the exponential parameter becomes:

$$\widehat{\theta} = 1/\bar{x} = 1/2.80 = 0.357$$

and the estimated exponential density of x is the following:

$$\mathrm{f(x)} = 0.357\mathrm{e}^{-0.357\mathrm{x}} \qquad \mathrm{x} \geq 0$$

3.7 Parameter Estimate When No Data

When an analyst wants to apply the exponential distribution in a study, but has no estimate on the parameter θ and also has no sample data, the analyst seeks guidance from an expert who gives an approximation on an α-percent-point of x, denoted as x_α. Note:

$$\begin{aligned} \mathrm{P}(\mathrm{x} \leq x_\alpha) &= \alpha \\ &= 1 - e^{-\theta x_\alpha} \end{aligned}$$

Hence,

$$(1 - \alpha) = e^{-\theta x_\alpha}$$

and with some algebra, the estimate of the parameter value becomes:

$$\widehat{\theta} = -(1/x_\alpha)\ln(1 - \alpha)$$

Note also, the estimate on the mean of x turns out to be:

$$\widehat{\mu} = 1/\widehat{\theta}$$

Example 3.5 In a study to determine how long it takes to perform a task where the variation resembles an exponential distribution, the analyst has no data to estimate the parameter value. With consultation from an expert on the system, the expert estimates 50% of the tasks will take 10 min or less. Using this approximation, the following data is noted:

$$x_{0.50} = 10 \text{ min}$$
$$\alpha = 0.50$$

Hence, the estimate of the exponential parameter becomes:

$$\widehat{\theta} = -(1/10)\ \ln(1 - 0.50) = 0.069$$

and thereby the probability density and the cumulative distribution are the following:

$$f(x) = 0.069e^{-0.069x} \qquad x \geq 0$$
$$F(x) = 1 - e^{-0.069x} \qquad x \geq 0$$

Example 3.6 Faults occur in an electrical system from time to time and engineers are sent to repair. The engineer estimates the repair time will take 100 min or less on 90% of the faults. Assuming the time to repair the fault follows an exponential distribution, the estimate of the parameter value is obtained as shown below.

First note:

$$\alpha = 0.90$$
$$x0.90 = 100 \text{ min}$$

Hence,

$$\widehat{\theta} = -(1/100)\ln(1 - 0.90) = 0.023$$
$$\widehat{\mu} = 1/0.023 = 43.43 \text{ min}$$

and thereby,

$$f(x) = 0.023e^{-0.023x} \qquad x > 0$$
$$F(100) = 1 - e^{-0.023(100)} = 0.90$$

3.8 Summary

The exponential distribution has one parameter, θ, and the probability density peaks at zero and gradually decreases afterwards. The mean and standard deviation have the same value, and thus the coefficient-of-variation is one. When sample data is available, an estimate of the parameter value is easily computed, and when no sample data, an approximation of the parameter is obtained using a measure provided from an expert type person.

Chapter 4
Erlang

4.1 Introduction

The origin of the Erlang distribution is attributed to Agner Erlang, a Danish engineer for the Copenhagen Telephone Company, who in 1908 was seeking how many circuits are needed to accommodate the voice traffic on their telephone system. The distribution has two parameters, k, θ, where k represents the number of exponential variables that are summed to form the Erlang variable. The exponential variables have the same parameter, θ, as the Erlang. The Erlang has shapes that range from exponential to normal. This distribution is heavily used in the study of queuing systems, representing the time between arrivals, and also the time to service a unit. The chapter shows how to compute the cumulative probability, F(x), where x is the random variable. When the parameter values are not known, and sample data is available, measures from the data is used to estimate the parameter values. When the parameter values are not known, and no sample data is available, approximates of some measures from the distribution, usually by an expert, are obtained and these allow estimates of the parameter values. The Erlang is a special case of the gamma distribution, as will be seen in Chap. 5, where the parameter k is a positive integer for the Erlang, and is any positive number for the gamma.

4.2 Fundamentals

The parameters of the Erlang distribution are the following:

$\theta > 0 =$ scale parameter
$\text{k}\ (1, 2, \ldots) =$ shape parameter

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_4

The parameter k is a positive integer and the random variable x is the sum of k exponential random variables, y, with the same scale parameter, θ. Note the following:

$$x = y_1 + \ldots + y_k$$

where x is Erlang and y are exponential. When $k = 1$, the Erlang is the same as the exponential. As k increases, the Erlang approaches a normal due to the central-limit theorem.

The probability density and cumulative distribution are below:

$$f(x) = x^{k-1}\theta^k/(k-1)!e^{-\theta x} \qquad x > 0$$
$$F(x) = 1 - e^{-\theta x}\left[1 + (\theta x) + (\theta x)^2/2! + \ldots + (\theta x)^{k-1}/(k\text{-}1)!\right] \qquad x > 0$$

The mean, variance and standard deviation are the following:

$$\mu = k/\theta$$
$$\sigma^2 = k/\theta^2$$
$$\sigma = \sqrt{k}/0$$

The coefficient-of-variation for the Erlang is the following:

$$cov = \sigma/\mu = 1/\sqrt{k}$$

Note the values of the cov as k goes from 1 to 9:

k	1	2	3	4	5	6	7	8	9
Cov	1.00	0.71	0.58	0.50	0.45	0.41	0.38	0.35	0.33

The most frequent value of x, $\widetilde{\mu}$, the mode, is the following;

$$\widetilde{\mu} = (k - 1)/\theta$$

When $k = 1$, $\widetilde{\mu} = 0$; and as k increases, $\widetilde{\mu}$ also increases.

4.3 Tables

Table 4.1 lists the values of the cumulative distribution, F(x), when the parameters are: $\theta = 1$, and $k = [1, (1), 9]$, while the random variable range is: [0, (0.5), 18]. The table can be used with interpolation to find the cumulative distribution for any Erlang when the parameter value is other than $\theta = 1$. Below shows how this applies. Let:

Table 4.1 Erlang cumulative distribution, F(x), when $\theta = 1$; x = 0–18; and k = 1–9

x/k	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0
0.5	0.39	0.09	0.01	0	0	0	0	0	0
1	0.63	0.26	0.08	0.02	0	0	0	0	0
1.5	0.78	0.44	0.19	0.07	0.02	0	0	0	0
2	0.86	0.59	0.32	0.14	0.05	0.02	0	0	0
2.5	0.92	0.71	0.46	0.24	0.11	0.04	0.01	0	0
3	0.95	0.80	0.58	0.35	0.18	0.08	0.03	0.01	0
3.5	0.97	0.86	0.68	0.46	0.27	0.14	0.07	0.03	0.01
4	0.98	0.91	0.76	0.57	0.37	0.21	0.11	0.05	0.02
4.5	0.99	0.94	0.83	0.66	0.47	0.30	0.17	0.09	0.04
5	0.99	0.96	0.88	0.73	0.56	0.38	0.24	0.13	0.07
5.5	1	0.97	0.91	0.80	0.64	0.47	0.31	0.19	0.11
6	1	0.98	0.94	0.85	0.71	0.55	0.39	0.26	0.15
6.5	1	0.99	0.96	0.89	0.78	0.63	0.47	0.33	0.21
7	1	0.99	0.97	0.92	0.83	0.70	0.55	0.40	0.27
7.5	1	1	0.98	0.94	0.87	0.76	0.62	0.48	0.34
8	1	1	0.99	0.96	0.90	0.81	0.69	0.55	0.41
8.5	1	1	0.99	0.97	0.93	0.85	0.74	0.61	0.48
9	1	1	0.99	0.98	0.95	0.88	0.79	0.68	0.54
9.5	1	1	1	0.99	0.96	0.91	0.84	0.73	0.61
10	1	1	1	0.99	0.97	0.93	0.87	0.78	0.67
10.5	1	1	1	0.99	0.98	0.95	0.90	0.82	0.72
11	1	1	1	1	0.98	0.96	0.92	0.86	0.77
11.5	1	1	1	1	0.99	0.97	0.94	0.89	0.81
12	1	1	1	1	0.99	0.98	0.95	0.91	0.84
12.5	1	1	1	1	0.99	0.99	0.97	0.93	0.88
13	1	1	1	1	1	0.99	0.97	0.95	0.90
13.5	1	1	1	1	1	0.99	0.98	0.96	0.92
14	1	1	1	1	1	0.99	0.99	0.97	0.94
14.5	1	1	1	1	1	1	0.99	0.98	0.95
15	1	1	1	1	1	1	0.99	0.98	0.96
15.5	1	1	1	1	1	1	0.99	0.99	0.97
16	1	1	1	1	1	1	1	0.99	0.98
16.5	1	1	1	1	1	1	1	0.99	0.98
17	1	1	1	1	1	1	1	0.99	0.99
17.5	1	1	1	1	1	1	1	1	0.99
18	1	1	1	1	1	1	1	1	0.99

$$F(x\alpha|\theta, k) = P(x \leq x\alpha|\theta, k)$$

Table 4.1 lists values of $F(x\alpha) = F(x\alpha\,|\theta = 1,k)$ when $\theta = 1$ and k = 1 to 9. With a bit of math, the following relation occurs:

Recall when $\theta = 1$, xα is the α-percent –point for a given k. Now let.

xα‘ = the α-percent-point for any θ and let k remain the same. The relation between the two α-percent–points is below:

$$x\alpha` = x\alpha/\theta$$

Also note, the Erlang variable for any combination of θ, x‘, is related to the table value, x, in the following way:

$$x` = x/\theta$$

Example 4.1 Observe from Table 4.1 where F(5.0) = 0.73 when $\theta = 1$ and k = 4. If $\theta = 0.5$, the 0.73%-point becomes: x‘ = 5.0/0.5 = 10.0, and hence, F (10.0) = 0.73 when $\theta = 0.5$ and k = 4.

Another way to compute F(x‘) when $x` = 10, \theta = 0.5$ and k = 4 is by noting θx‘ = 0.5 × 10 = 5.0, and thereby,

$$F(10) = 1 - e^{-5}\left[1 + 5 + 5^2/2 + 5^3/6\right] = 0.73$$

Figure 4.1 depicts the probability density, f(x), when parameter values are $\theta = 1$ and k = 1, 3, 6 and 9. At k = 1, the distribution is exponential, and at k = 9, the shape is like a normal distribution.

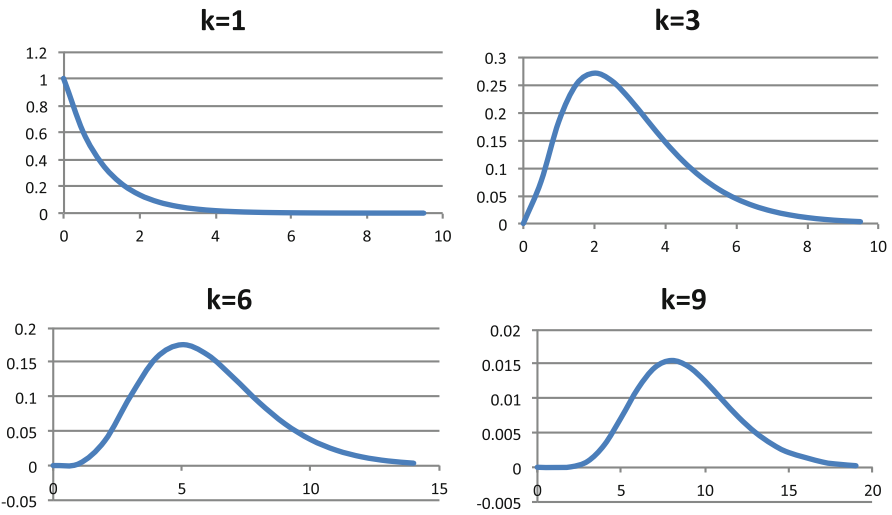

Fig. 4.1 Probability density, f(x), of the Erlang when $\theta = 1$ and k = 1, 3, 6 and 9

4.4 Sample Data

When sample data of size n is available, $(x_1, \ldots, x_n)$, the sample mean, variance and standard deviation are measured. These are the following:

$\bar{x}$ = sample mean
s^2 = sample variance
s = sample standard deviation

4.5 Parameter Estimates When Sample Data

Using the sample measures, the parameter values can readily by estimated by the method-of-moment method. To accomplish, recall the relation between the parameters and the population mean and variance as given earlier:

$\mu = k/\theta$
$\sigma^2 = k/\theta^2$

Substituting the corresponding sample estimates into the above relations yields the following:

$\bar{x} = k/\theta`$
$s^2 = k`/\theta`^2$

Using some algebra, the parameter estimates are the following:

$\theta` = \bar{x}/s^2$
$k` = \bar{x}\theta`$

But, since k is restricted to a positive integer, the closest integer to k‘ is selected and the parameter estimates are adjusted accordingly:

$\hat{k}$ = floor(k` + 0.5) and min of 1
$\widehat{\theta} = \hat{k}/\bar{x}$

Example 4.2 An engineer studying the time-to-fail for an auto component has sample data that yield the following average and variance.

$\bar{x} = 12.0\ (100 - \text{hours})$
$s^2 = 20.0$

Assuming the random variable, x, is Erlang distributed, the estimates of the parameters are obtained as below:

$\theta` = 12/20 = 0.60$
$\mathrm{k}` = 12 \times 0.60 = 7.2$

Adjusting k‘ to its closest integer, yields the Erlang parameter estimates,

$\hat{k} = \mathrm{floor}(7.2 + 0.5) = 7$
$\widehat{\theta} = 7/12 = 0.583$

Example 4.3 Assume the engineer from Example 4.2 also wants to estimate the 0.90%-point of the time-to-fail for the component. This is accomplished as follows: Table 4.1 is searched at $\mathrm{k} = 7$ and $\theta = 1.0$ to find the closest value of x with $\mathrm{F(x)} = 0.90$. This is observed at $\mathrm{x}‘_{0.90} = 10.5$. The corresponding value when $\theta = 0.583$ and $\mathrm{k} = 7$ becomes:

$\mathrm{x}_{0.90} = 10.5/0.583 = 18.0$ (100 hours)

Hence,

$\mathrm{P}(\mathrm{x} \leq 18.0) = 0.90$

4.6 Parameter Estimates When No Data

When no sample data is available, and an expert offers estimates on the shape of the Erlang distribution, the estimates on the parameter values are obtained in a search manner as described below. The expert data needed is the following:

$\mathrm{x}\alpha = \alpha$-percent-point
$\alpha = \mathrm{F}(x\alpha) =$ cumulative probability
$\tilde{\mathrm{x}} =$ most-likely value (mode)

For each value of k, from 1 to 9, find the corresponding value of θ, by the relation:

$\tilde{x} = (\mathrm{k} - 1)/\theta$

which yields:

$\theta = (\mathrm{k} - 1)/\tilde{x}$

So for each combination of (k, θ), the probability function, F(x), given earlier is applied to compute F(xα). The result yields nine combinations of the Erlang parameters that correspond with the data from the expert, and the combination that gives F(xα) close to α, is the estimate of the parameters to apply. In the event no combination is accepted, the expert data does not resemble an Erlang distribution.

Example 4.4 Table 4.1 lists four cases where expert data is used to estimate the parameter values for an Erlang distribution. The case information is listed in the table as: (1), (2), (3) and (4). With each case are the three expert data provided (xα, α, $\tilde{x}$). Using the expert data, the next step is to find and list the value of the cumulative probability, F(xα), for each value of k (1 to 9). The probability, F(x), is computed as given earlier for any x. Also listed is the corresponding value of θ, obtained from:

$$\theta = (\mathrm{k} - 1)/\tilde{x}$$

In Case 1, the parameter combination (k, θ) = (2, 1) yields:

$$\mathrm{F}(3) = 1 - \mathrm{e}^{-1\times 3}[1 + 1 \times 3] = 0.801$$

which is the closest of the nine candidate combinations to $\alpha = 0.80$, and thereby, the combination is chosen as the estimate of the Erlang parameters.

In Case 2, the combination (6, 0.55), give:

$$\mathrm{F}(17) = 1 - \mathrm{e}^{-0.11\times 17}\left[1 + 0.11 \times 17 + \ldots + 0.11 \times 17^5/5!\right] = 0.909$$

for the closest to $\alpha = 0.90$, and thereby become the parameter estimates for the Erlang.

At Case 3, the combination (5, 2) gives $(\mathrm{k} - 1)\theta = (5 - 1)2 = 8$ and hence:

$$\mathrm{F}(4) = 1 - \mathrm{e}^{-8}\left[1 + 8 + 8^2/2 + 8^3/6 + 8^4/24\right] = 0.900$$

for the Erlang.

Case 4 has no combination that is near $\alpha = 0.90$, indicating the expert's data does not represent a good fit for an Erlang distribution (Table 4.2).

4.7 Summary

The Erlang with parameters, k, θ, has many shapes from exponential to normal, and is highly used in queuing studies. It is the same as a gamma distribution when the parameter k is a positive integer. The Erlang variable is formed by the sum of k exponential variables with the same parameter, θ. At k = 1, the distribution is the same as the exponential, and as k increases, due to the central-limit-theorem, the

Table 4.2 Four cases: (1–4) of Erlang expert data: (xα, α and $\tilde{x}$) that yield F(xα) for k = 1 to 9; and the corresponding value of $\theta = (k-1)/\tilde{x}$

(1)									
xα	α	$\tilde{x}$							
3	0.8	1							
k	1	2	3	4	5	6	7	8	9
θ	0	1	2	3	4	5	6	7	8
F(xα)	0	0.801	0.938	0.979	0.992	0.997	0.999	1.000	1.000
(2)									
xα	α	$\tilde{x}$							
17	0.9	9							
k	1	2	3	4	5	6	7	8	9
θ	0	0.11	0.22	0.33	0.44	0.55	0.66	0.77	0.88
F(xα)	0	0.563	0.727	0.816	0.872	0.909	0.934	0.952	0.965
(3)									
xα	α	$\tilde{x}$							
4	0.9	2							
k	1	2	3	4	5	6	7	8	9
θ	0	0.5	1	1.5	2	2.5	3	3.5	4
F(xα)	0	0.594	0.762	0.849	0.900	0.933	0.954	0.968	0.978
(4)									
xα	α	$\tilde{x}$							
4	0.9	3							
k	1	2	3	4	5	6	7	8	9
θ	0	0.33	0.67	1.00	1.33	1.67	2.00	2.33	2.67
F(xα)	0.000	0.385	0.498	0.567	0.616	0.655	0.687	0.714	0.737

distributions approaches a normal shape. When sample data is obtained, the data is used to estimate the parameter values of the Erlang. When no sample data, approximate measures on the distribution are used to estimate the parameter values.

Chapter 5
Gamma

5.1 Introduction

Karl Pearson, a famous British Professor, introduced the gamma distribution in 1895. The distribution, originally called the Pearson type III distribution, was renamed in the 1930s to the gamma distribution. The gamma distribution has many shapes ranging from an exponential-like to a normal-like. The distribution has two parameters, k and θ, where both are larger than zero. When k is a positive integer, the distribution is the same as the Erlang. Also, when k is less or equal to one, the mode is zero and the distribution is exponential-like; and when k is larger than one, the mode is greater than zero. As k increases, the shape is like a normal distribution. There is no closed form solution to compute the cumulative probability, but quantitative methods have been developed and are available. Another method is developed in this chapter and applies when k ranges from 1 to 9. When sample data is available, estimates of the parameter values are obtained. When no sample data is available, estimates of the parameter values are obtained using approximations on some distribution measures.

5.2 Fundamentals

The parameters of the gamma distribution are the following:

$\theta > 0 =$ scale parameter
$\mathrm{k} > 0 =$ shape parameter

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_5

The probability density, f(x), and discussion of the cumulative distribution, F(x), are below:

$$f(x) = x^{k-1}\theta^k e^{-\theta x}/\Gamma(k) \qquad x > 0$$

5.3 Gamma Function

$\Gamma(k)$ is a gamma function, not a distribution, and is computed as below:

$$\Gamma(0.5) = \sqrt{\pi}$$
$$\Gamma(k) = (k\text{-}1)! \qquad \text{for } k = \text{a positive integer}$$
$$\Gamma(k) = \int_0^\infty w^{k-1} e^{-w} dw \qquad \text{for } k > 0$$

For other values of $k > 1$, the Stirling's formula [Abramowitz and Stegun (1964), p 257] listed below can be applied:

$$\Gamma(k) = k^{k-.5} e^{-k} \sqrt{2\pi}\left[1 + 1/12k + 1/288k^2 - 139/51840k^3 - 571/2488320k^4 + \ldots\right]$$

5.4 Cumulative Probability

There is no closed-form solution to the cumulative probability distribution of the gamma distribution. However, when k is a positive integer, the distribution is the same as the Erlang, as described in Chap. 4. In this special case, the cumulative probability is computed as below:

$$F(x) = 1 - e^{-\theta x}\left[1 + (\theta x) + (\theta x)^2/2! + \ldots + (\theta x)^{k-1}/(k\text{-}1)!\right] \quad x > 0$$

Table 4.1 lists values of F(x) for a selective range of x, and with k the integers from 1 to 9. Figure 5.1 depicts the shape of the distributions where $k = 1$ is the curve on the left-hand-side, and k is progressively increasing till $k = 9$ is the curve on the right-hand-side of the figure.

The mean, variance and standard deviation are the following:

$$\mu = k/\theta$$
$$\sigma^2 = k/\theta^2$$
$$\sigma = \sqrt{k}/\theta$$

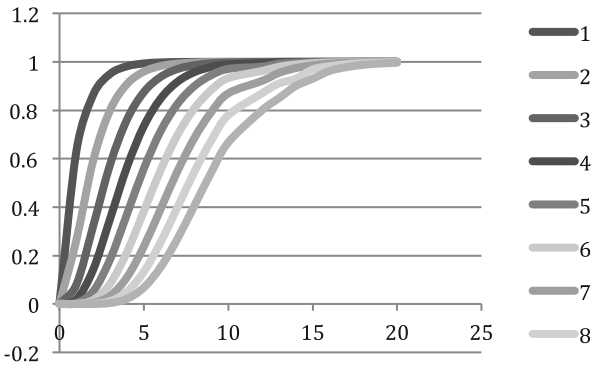

Fig. 5.1 The cumulative probability distribution for the gamma when k is the integers 1 to 9 from left to right

The coefficient-of-variation is the following:

$$\text{cov} = \sigma/\mu = 1/\sqrt{\text{k}}$$

The most frequent value of x, the mode, is the following;

$$\begin{aligned} \widetilde{\mu} &= (\text{k}-1)\theta \qquad \text{when k} \geq 1 \\ \widetilde{\mu} &= 0 \qquad\qquad\quad \text{when k} < 1 \end{aligned}$$

When $\text{k} \leq 1$, the mode is zero and the shape of the distribution is exponential-like, and when $\text{k} > 1$, the mode is greater than zero; and as k increases the distribution is normal-like.

Example 5.1 Suppose an analyst has data where the variable x is gamma distributed with $\text{k} = 3.0$ and $\theta = 0.8$. The probability density becomes:

$$\text{f}(\text{x}) = 0.256\text{x}^2\text{e}^{-0.8\text{x}} \qquad \text{x} > 0$$

Selected values of x (0 to 12) are listed in Table 5.1. Because k is an integer, the cumulative probability can be computed using F(x) as given in the Erlang chapter. When $\text{k} = 3$ and $\text{x} = 5$, say,

$$\text{F}(5) = 1 - \text{e}^{-4}\left[1 + 4 + 4^2/2\right] = 0.76$$

The mean, variance and standard deviation of x are listed below:

$$\mu = 3/0.8 = 3.75$$

$$\sigma^2 = 3/0.64 = 4.69$$

$$\sigma = 2.16$$

Table 5.1 The probability density, f(x), for selected values of the gamma distributed x with k = 3.0 and $\theta = 0.8$

x	f(x)
0	0
1	0.115
2	0.207
3	0.209
4	0.167
5	0.117
6	0.076
7	0.046
8	0.027
9	0.015
10	0.009
11	0.005
12	0.002

The most likely value, the mode, is:

$$\widetilde{\mu} = (3 - 1)/0.8 = 2.5$$

5.5 Estimating the Cumulative Probability

Assume the gamma parameter values are (k, θ) and the goal is to estimate the cumulative probability for a value x‘ given, k and θ, denoted here as F(x‘|k, θ).

Since x‘ = x/θ, x = x‘θ is the corresponding variable value associated with the Table 4.1 entries. Suppose:

$$k_1 < k < k_2$$
$$x_1 < x < x_2$$

where k_1, k_2, x_1, x_2, are the closest entries of k and x in Table 4.1.

For convenience, let F(x|k) represent the cumulative probability of x when the parameters are k and $\theta = 1$. The Table 4.1 values of the cumulative probability, F(x|k), that are listed are the following:

$$F(x_1|k_1)$$
$$F(x_2|k_1)$$
$$F(x_1|k_2)$$
$$F(x_2|k_2)$$

Using interpolation, the following probability estimates are now obtained:

$$F(x_1|k) \approx F(x_1|k_1) + (k - k_1)/(k_2 - k_1)[F(x_1|k_2) - F(x_1|k_1)]$$
$$F(x_2|k) \approx F(x_2|k_1) + (k - k_1)/(k_2 - k_1)[F(x_2|k_2) - F(x_2|k_1)]$$

and

$$F(x|k) \approx F(x_1|k) + (x - x_1)/(x_2 - x_1)\,[F(x_2|k) - F(x_1|k)]$$

Finally for $x` = x/\theta$, the cumulative probability becomes:

$$F\left(x`|k, \theta\right) \approx F(x|k)$$

Example 5.2 Assume an analyst has a scenario with gamma data and parameter values of $k = 5.7$ and $\theta = 0.8$. The analyst is seeking the estimate of the cumulative probability of $x`$ less or equal to 12. Note $x = x`\theta = 12 \times 0.8 = 9.6$. To accomplish, the data from Table 4.1 is searched to find the closest entries to $k = 5.7$ and $x = 9.6$ as below:

$$k_1, k_2 = 5, 6$$
$$x_1, x_2 = 9.5, 10.0$$

Next the associated probability values from the table are the following:

$$F(9.5, 5) = 0.96$$
$$F(10.0, 5) = 0.97$$
$$F(9.5, 6) = 0.91$$
$$F(10.0, 6) = 0.93$$

Now applying interpolation,

$$F(9.5|\,5.7) \approx 0.96 + (5.7 - 5)/(6 - 5)[0.91 - 0.96] = 0.925$$
$$F(10.0|\,5.7) \approx 0.97 + (5.7 - 5)/(6 - 5)[0.93 - 0.97] = 0.942$$

and

$$F(9.6|\,5.7) \approx 0.925 + (9.6 - 9.5)/(10.0 - 9.5)[0.942 - 0.925] = 0.9284$$

Hence,

$$F\left(x` = 12|\,9.6, 5.7\right) = P\left(x` < 12\right)\,0.9284$$

5.6 Sample Data

Assume the sample data that is available is the following: $(x_1, \ldots, x_n)$, and the statistics measured from the data are the average, variance and standard deviation as follows:

$\bar{x}$ = sample average
s^2 = sample variance
s = sample standard deviation

5.7 Parameter Estimates When Sample Data

Using the sample measures, the parameter values can readily by estimated by the method-of-moment method. To accomplish, recall the relation between the parameters and the population mean and variance as given earlier:

$\mu = k/\theta$
$\sigma^2 = k/\theta^2$

Substituting the corresponding sample estimates into the above relations yields the following:

$\bar{x} = \hat{k}/\widehat{\theta}$
$s^2 = \hat{k}/\widehat{\theta}^2$

Using some algebra, the parameter estimates are below:

$\widehat{\theta} = \bar{x}/s^2$
$\hat{k} = \bar{x}\widehat{\theta}$

Example 5.3 Assume an analyst has the following sample data and wishes to apply the gamma distribution:

$(5.1, 7.2, 11.8, 3.1, 7.4, 15.4, 2.1\ 6.4, 7.3, 4.5)$

The sample average, variance and standard deviation are the following:

$\bar{x} = 7.03$
$s^2 = 15.925$
$s = 2.236$

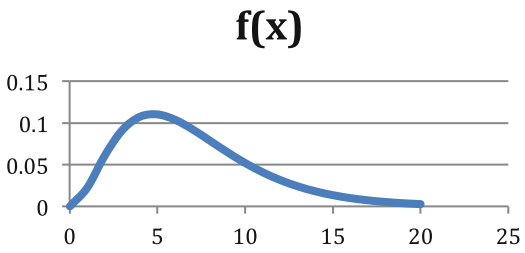

Fig. 5.2 Gamma plot when k = 3.10 and $\theta = 0.44$

Example 5.4 Using the data from Example 5.3, the estimates of the gamma parameters are computed as follows:

$\widehat{\theta} = \bar{x}/\mathrm{s}^2 = 7.03/15.925 = 0.44$
$\hat{\mathrm{k}} = \bar{x}\widehat{\theta} = 7.03 \times 0.44 = 3.10$

The plot of the gamma distribution is depicted in Fig. 5.2.

5.8 Parameter Estimate When No Data

When no sample data is available, and an approximation on the shape of the gamma distribution is provided, the estimates on the parameter values are obtained in a search manner as described below. The approximation data needed is the following:

$x\alpha$ = α-percent-pointa
α = F($x\alpha$) = cumulative probability
$\tilde{\mathrm{x}}$ = most-likely value (mode)

The method described here assumes the value of k falls somewhere between (1 and 9). Hence, for each value of k, from 1 to 9, find the corresponding value of θ, by the relation:

$\theta = (\mathrm{k} - 1)/\tilde{x}$

With each combination of (k, θ), the probability function, F(x), given via the Erlang is applied to compute F(xα). The result yields nine cumulative probabilities for the combinations (k, θ). The two combinations that yield the closest probabilities to α are identified where:

$\mathrm{F}(x\alpha|\mathrm{k}_1, \theta) \leq \alpha \leq \mathrm{F}(x\alpha|\mathrm{k}_2, \theta)$

and $\mathrm{k}_2 = \mathrm{k}_1 + 1$.

So now, the estimate of k is obtained by interpolation as below:

$$\hat{k} = \mathrm{k}_1 + [\alpha - \mathrm{F}(x\alpha|\mathrm{k}_1, \theta)/[\mathrm{F}(x\alpha|\mathrm{k}_2, \theta) - \mathrm{F}(x\alpha|\mathrm{k}_1, \theta)](\mathrm{k}_2 - \mathrm{k}_1)$$

The corresponding value of θ, denoted as $\widehat{\theta}$, is computed as below: $\widehat{\theta} = (\hat{k} - 1)/\tilde{x}$.

Finally, the parameter estimates are:

$$(\hat{k}, \widehat{\theta}).$$

In the event no combination is accepted, the approximation data does not resemble a gamma distribution.

Example 5.5 A simulation is being designed with a need for a gamma variable when no sample data is available to estimate the parameter values. An expert is called and approximates the following measures on the distribution: 90% of the observations will fall below 100, and the most likely value is 40. Hence, the measures are the following:

$\mathrm{x}_{0.9} = 100$
$\alpha = 0.90$
$\tilde{x} = 40$

To estimate the parameter values, the analyst sets k = 1–9, and for each, the associated parameter θ is measured, and also the cumulative probability of less or equal to 100, F(100). Below shows how F(100) is computed when x = 100, k = 3 and $\theta = 0.05$, say. Note θ x = 0.05 × 100 = 5.0. Hence,

$$\mathrm{F}(100) = 1 - \mathrm{e}^{-5}\left[1 + 5 + 5^2/2\right] = 0.875$$

The results are listed below:

k	1	2	3	4	5	6	7	8	9
θ	0	.025	.050	.075	.100	.125	.150	.175	.200
F(100)	0	.717	.875	.941	.971	.985	.992	.996	.998

Because $\alpha = 0.90$ falls between k = 3 and 4, $\mathrm{k}_1 = 3$ and $\mathrm{k}_2 = 4$. The estimate of k is obtained by interpolation as below:

$$\hat{k} \tilde{} 3 + [0.900 - 0.875]/[0.941 - 0.875] = 3.38$$

The corresponding value of θ becomes:

$$\widehat{\theta} = (\hat{k} - 1)/\tilde{x} = 2.38/40 = 0.059$$

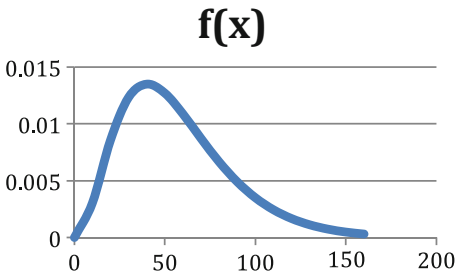

Fig. 5.3 Depiction of gamma density when k = 3.38 and $\theta = 0.059$

Hence, the gamma probability density is:

$$f(x) = x^{2.38}(0.059)^{3.38}e^{(-0.059x)}/\Gamma(3.38) \qquad x > 0$$

and the resulting plot of the gamma is in Fig. 5.3.

5.9 Summary

The gamma has many shapes from exponential-like to normal. The distribution does not have a closed-form solution for the cumulative probability, but quantitative methods are available for use. The chapter presents another quantitative method on computing the cumulative probability. When available, sample data is gathered to estimate the parameter values; and when no sample data is available, an expert provides approximation measures that allow estimates of the parameter values.

Chapter 6
Beta

6.1 Introduction

The beta distribution, introduced in 1895 by Karl Pearson a famous British mathematician, was originally called the Pearson type 1 distribution. The name was changed in the 1940s to the beta distribution. Thomas Bayes also applied the distribution in 1763 as a posterior distribution to the parameter of the Bernoulli distribution. The beta distribution has many shapes that range from exponential, reverse exponential, right triangular, left triangular, skew right, skew left, normal and bathtub. The only fault is that it is a bit difficult to apply to real applications. The beta has two main parameters, k_1 and k_2 that are both larger than zero; and two location parameters a and b that define the limits of the admissible range. When the parameters are both larger than one, the beta variable is skewed to the left or to the right. These are the most used shapes of the distribution, and for this reason, the chapter concerns mainly these shapes. The random variable is denoted here as w where ($a \leq w \leq b$). A related variable, called the standard beta, x, has a range of 0–1. The mathematical properties of the probability density are defined with the standard beta. This includes the beta function and the gamma function, which are not easy to calculate. An algorithm is listed and needs to be calculated via a computed. There is no closed form solution to the cumulative probability, and thereby, a quantitative method is shown in the chapter examples. A straightforward process is used to convert from w to x and from x to w. When sample data is available, the sample average and mode are needed to estimate the parameter values of k_1 and k_2. A regression fit is developed to estimate the average of x from the mode of x. When sample data is not available, best estimates of the limits (a, b) and the most-likely value of w are obtained to estimate the parameter values.

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_6

6.2 Fundamentals

The random variable of the beta distribution is listed as w and has a range of (a to b), while a counterpart standard beta random variable x has a range from (0 to 1). The parameters for the beta are the following:

(k_1, k_2) = parameters
(a, b) = range for the beta variable w
$(0, 1)$ = range for the standard beta variable x

The relation between w and x are below:

$$x = (w - a)/(b - a)$$
$$w = a + x(b - a)$$

6.3 Standard Beta

The probability density for the standard beta is below:

$$f(x) = x^{k_1 - 1}(1 - x)^{k_2 - 1}/B(k_1, k_2)$$

where the denominator is the beta function, not a distribution, and is computed as the following:

$$B(k_1, k_2) = \Gamma(k_1)\Gamma(k_2)/\Gamma(k_1 + k_2) \qquad 0 < x < 1$$

The components in the beta function are the gamma function as described in Chap. 5.

The mean and variance of the beta are below:

$$\mu = k_1/(k_1 + k_2)$$
$$\sigma^2 = k_1 k_2/\left[(k_1 + k_2)^2(k_1 + k_2 + 1)\right]$$

When $k_1 > 1$ and $k_2 > 1$, the most likely value, the mode, is listed below:

$$\widetilde{\mu} = (k_1 - 1)/(k_1 + k_2 - 2)$$

There is no closed-form solution for a cumulative probability function, F(x), but a quantitative method that yields an estimate is shown in the examples that follow.

6.4 Beta Has Many Shapes

The beta distribution takes on many shapes depending on the value of the two parameters (k_1,k_2) where $k_1 > 0$ and $k_2 > 0$.

Parameters	Shape
$k_1 < 1$ and $k_2 > 1$	exponential like (right skewed)
$k_1 > 1$ and $k_2 < 1$	exponential like (left skewed)
$k_1 = 1$ and $k_2 < 1$	ramp down (right skewed)
$k_1 < 1$ and $k_2 = 1$	ramp up (left skewed)
$k_1 < 1$ and $k_2 < 1$	bathtub shape
$k_1 > 1$ and $k_2 > 1$	gamma like (skewed left & right)
$k_1 > 1$ and $k_2 > 1$ & $k_1 = k_2$	symmetrical, normal like
$k_1 = k_2 = 1$	uniform

The discussion of this chapter pertains only when $k_1 > 1$ and $k_2 > 1$.

Example 6.1 Assume the standard beta distribution with parameters $k_1 = 1.4$ and $k_2 = 2.0$ where the random variable ranges from 0 to 1. The mean and mode of this distribution is computed below:

$$\mu = 1.4/3.4 = 0.412$$
$$\widetilde{\mu} = 0.4/1.4 = 0.286$$

The probability density is the following;

$$f(x) = x^{0.4}(1-x)^{1.0}/B(1.4, 2.0)$$

where the beta function is obtained as below:

$$\begin{aligned} B(1.4, 2.0) &= \Gamma(1.4)\Gamma(2.0)/\Gamma(3.4) \\ &= 0.887 \times 1.000/2.981 \\ &= 0.298 \end{aligned}$$

The shape of the probability density is depicted in Fig. 6.1.

Example 6.2 Applying the same parameters as in Example 6.1, assume the limits are $a = 10$ and $b = 50$, and thereby the variable becomes:

$$w = 10 + x(50 - 10)$$

The mean and mode of w are obtained below:

$$\mu = 10 + 0.412(40) = 26.46$$
$$\widetilde{\mu} = 10 + 0.286(40) = 21.44$$

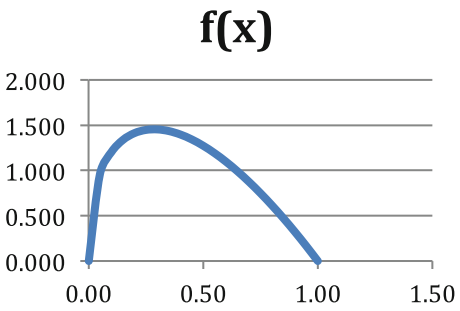

Fig. 6.1 Standard Beta when $k_1 = 1.4$ and $k_2 = 2.0$

The following table lists the probability density, f(w), and an estimate of the cumulative probability, F(w), for selective values of w. F(w) is obtained by the following:

$$F(w_o) \approx \sum_{w=10}^{wo} f(w) / \sum_{w=10}^{50} f(w)$$

where the summation spans all of the probability densities listed from 10 to 50. Note f(w) is the same as f(x) when w = (10 + 40x).

w	f(w)	Σf(w)	F(w)
10	0.000	0	0.000
12	0.963	0.963	0.049
14	1.204	2.167	0.110
16	1.337	3.505	0.178
18	1.412	4.917	0.249
20	1.448	6.364	0.322
22	1.453	7.818	0.396
24	1.435	9.253	0.469
26	1.398	10.650	0.540
28	1.343	11.993	0.608
30	1.273	13.267	0.672
32	1.191	14.457	0.732
34	1.096	15.553	0.788
36	0.990	16.543	0.838
38	0.874	17.417	0.882
40	0.749	18.166	0.920
42	0.615	18.781	0.951
44	0.472	19.253	0.975
46	0.322	19.575	0.992
48	0.165	19.740	1.000
50	0.000	19.740	1.000

Using the results in the table, note for example, the 0.95%-point of w is approximately at $w = 42.0$, and hence,

$P(w \leq 42.0) \approx 0.95.$

More accurate estimate of F(w) can be obtained by applying more values of w, f (w) and F(w).

6.5 Sample Data

When sample data from a beta with limits (a, b) is available as: $(w_1, \ldots, w_n)$, the measures gathered are the following:

$\bar{w} =$ sample average
$\tilde{w} =$ sample mode

The corresponding measures on the standard beta x with limits (0, 1) are obtained by:

$$\bar{x} = (\bar{w} - a)/(b - a)$$
$$\tilde{x} = (-a)/(b - a)$$

6.6 Parameter Estimates When Sample Data

When sample data from a beta with limits (a, b) is available, the sample measures that are needed are the average, $\bar{w}$, and the mode, $\tilde{w}$. These measures are converted to the average and mode of the standard beta ($\bar{x}$ and $\tilde{x}$), and they are then inserted into the population mean and mode equations listed earlier to form the below relations:

$$\bar{x} = \hat{k}_1/(\hat{k}_1 + \hat{k}_2)$$
$$\tilde{x} = (\hat{k}_1 - 1)/(\hat{k}_1 + \hat{k}_2 - 2)$$

With a bit of algebra, the above relations are used to estimate the parameters (k_1, k_2) of the beta as listed below:

$$\hat{k}_1 = \bar{x}[2\tilde{x} - 1/[\tilde{x} - \bar{x}]$$
$$\hat{k}_2 = (1 - \bar{x})\hat{k}_1/\bar{x}$$

Example 6.3 A researcher has beta sample data with limits (a, b) = (0, 100) that yields the following measures:

$\bar{w} = 30$
$\tilde{w} = 20$

Converting the measures to the standard beta counterparts yields:

$$\bar{x} = (30 - 0)/(100 - 0) = 0.3$$
$$\tilde{x} = (20 - 0)/(100 - 0) = 0.2$$

The beta parameters are now computed as below:

$$k_1 = 0.3(2 \times 0.2 - 1)/(0.2 - 0.3) = 1.8$$
$$k_2 = (1 - 0.3)1.8/0.3 = 4.2$$

Finally, the probability density is the following:

$$f(x) = x^{0.8}(1 - x)^{3.2}/B(1.8, 4.2)$$

where

$$\begin{aligned} B(1.8, 4.2) &= \Gamma(1.8)\Gamma(4.2)/\Gamma(6.0) \\ &= 0.931 \times 7.757/120.000 \\ &= 0.060 \end{aligned}$$

The table below lists selective values of the probability density, f(x), and cumulative probability, F(x), for selective values of x. Note F(x) is computed as described in Example 6.2. A plot of the probability density is in Fig. 6.2.

x	f(x)	Σf(x)	F(x)
0.00	0.000	0	0.000
0.05	1.283	1.283	0.065
0.10	1.879	3.162	0.160
0.15	2.165	5.327	0.269
0.20	2.244	7.571	0.382
0.25	2.182	9.754	0.492
0.30	2.025	11.779	0.594
0.35	1.807	13.586	0.685
0.40	1.556	15.142	0.764
0.45	1.295	16.437	0.829
0.50	1.038	17.475	0.882
0.55	0.800	18.275	0.922
0.60	0.588	18.863	0.952
0.65	0.409	19.272	0.972
0.70	0.265	19.537	0.986
0.75	0.156	19.693	0.994
0.80	0.081	19.774	0.998
0.85	0.034	19.807	0.999
0.90	0.010	19.817	1.000
0.95	0.001	19.818	1.000
1.00	0.000	19.818	1.000

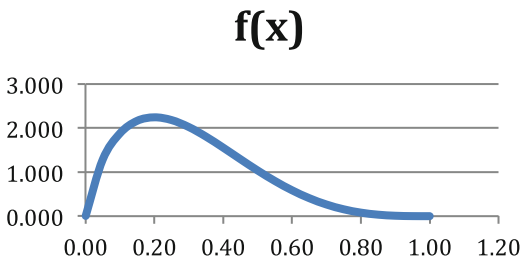

Fig. 6.2 Beta plot when $k_1 = 1.8$ and $k_2 = 4.2$

6.7 Regression Estimate of the Mean from the Mode

The table below lists a variety of beta parameters (k_1,k_2) from the typical range of values; and also lists the computed values of the mode, $\widetilde{\mu}$, and mean, μ. Using this data, a linear regression was run to predict the value of the mean from the mode. Figure 6.3 is a plot of the data showing the mode on the x-axis and the mean on the y-axis. The result gives a high correlation and the fit is the following;

$$\mu = 0.175 + 0.645\widetilde{\mu}$$

K_1	K_2	$\widetilde{\mu}$	μ
1.5	1.5	0.50	0.50
2	1.5	0.67	0.57
3	1.5	0.80	0.67
4	1.5	0.86	0.73
5	1.5	0.89	0.77
1.5	2	0.33	0.43
2	2	0.50	0.50
3	2	0.67	0.60
4	2	0.75	0.67
5	2	0.80	0.71
1.5	3	0.20	0.33
2	3	0.33	0.40
3	3	0.50	0.50
4	3	0.60	0.57
5	3	0.67	0.63
1.5	4	0.14	0.27
2	4	0.25	0.33
3	4	0.40	0.43
4	4	0.50	0.50
5	4	0.57	0.56
1.5	5	0.11	0.23
2	5	0.20	0.29
3	5	0.33	0.38
4	5	0.43	0.44
5	5	0.50	0.50

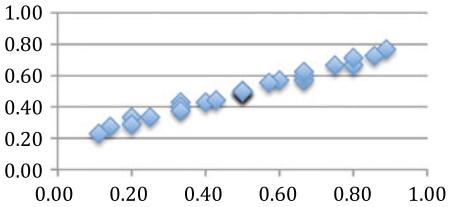

Fig. 6.3 Plot of mode on x-axis vs. mean on y-axis

6.8 Parameter Estimates When No Data

When sample data is not available, the analyst seeks an expert to provide some information on the shape of the distribution. One way is to estimate the most-likely value, the mode, denoted as $\tilde{w}$, and also the min and max (a, b) limits. Using these estimates, the mode of the beta distribution is obtained from the relation below:

$$\tilde{x} = (\tilde{w} - \mathrm{a})/(\mathrm{b} - \mathrm{a})$$

The analyst next applies the regression result to estimate the value of the average of x as follows:

$$\bar{x} = 0.175 + 0.645\tilde{x}$$

This step allows the analyst to have estimates of the key standard beta measures: $\tilde{x}$,$\bar{x}$. The analyst is now able to compute the estimates of the beta parameters as follows:

$$\hat{k}_1 = \bar{x}[2\tilde{x} - 1/[\,\tilde{x} - \bar{x}]$$
$$\hat{k}_2 = (1 - \bar{x})\hat{k}_1/\bar{x}$$

Example 6.5 A researcher is developing a study and needs to use a beta variable, w, but has no sample data to estimate the parameters. In particular, the researcher is seeking the value of w where the cumulative probability is 0.95. The researcher is provided with the following estimates on the beta distribution.

$\mathrm{a} = 50$ is the min
$\mathrm{b} = 100$ is the max
$\tilde{w} = 80$ is the most likely value

The counterpart standard beta mode becomes:

$$\tilde{x} = (80 - 50)/(100 - 50)$$
$$= 0.60$$

Using the regression fit, the estimate of the standard beta average is computed as follows:

$$\begin{aligned}\bar{x} &= 0.175 + 0.645 \times 0.60 \\ &= 0.562\end{aligned}$$

With these measures, the estimate of the beta parameters are obtained as below:

$$\begin{aligned}k_1 &= 0.562[2 \times 0.60 - 1/[0.60 - 0.562] \\ &= 2.958 \\ k_2 &= [1 - 0.562]\ 2.958/0.562 \\ &= 2.305\end{aligned}$$

Finally, the probability density of x is the following:

$$f(x) = x^{1.958}(1 - x)^{1.305}/B(2.958, 2.305)$$

where

$$\begin{aligned}B(2.958, 2.305) &= \Gamma(2.958)\Gamma(2.305)/\Gamma(5.263) \\ &= 1.924 \times 1.1702/35.933 \\ &= 0.063\end{aligned}$$

Table 6.1 lists the probability density, f(x), and the cumulative probability, F(x), for selected values of x. Fig. 6.4 has a depiction of the probability density of x. The 0.9%-point of w is obtained using interpolation as shown below:

$$\begin{aligned}x_{0.95} &= 0.85 + 0.05(0.950 - 0.934)/(0.973 - 0.934) \\ &= 0.871\end{aligned}$$

The counterpart 0.95% point for w is:

$$\begin{aligned}w_{0.95} &= 50 + 0.871(100 - 50) \\ &= 93.55\end{aligned}$$

Hence, $P(w \leq 93.55) \approx 0.95$.

6.9 Summary

The beta distribution has many shapes. This chapter concerns only the shapes that are skewed left and right. The beta variable has a range from a to b, and its counterpart variable, the standard beta, has a range from 0 to 1. The main computations are with the standard beta, and these are easily converted to the beta

Table 6.1 List of the standard beta, x, with f(x) and F(x)

x	f(x)	∑f(x)	F(x)
0.00	0.00	0.00	0.00
0.05	0.04	0.04	0.00
0.10	0.15	0.20	0.01
0.15	0.31	0.51	0.03
0.20	0.51	1.02	0.05
0.25	0.73	1.75	0.09
0.30	0.95	2.69	0.13
0.35	1.16	3.86	0.19
0.40	1.36	5.22	0.26
0.45	1.53	6.75	0.34
0.50	1.66	8.41	0.42
0.55	1.75	10.16	0.51
0.60	1.77	11.93	0.60
0.65	1.74	13.68	0.68
0.70	1.65	15.33	0.77
0.75	1.49	16.82	0.84
0.80	1.26	18.08	0.90
0.85	0.98	19.05	0.95
0.90	0.64	19.70	0.99
0.95	0.29	19.99	1.00
1.00	0.00	19.99	1.00

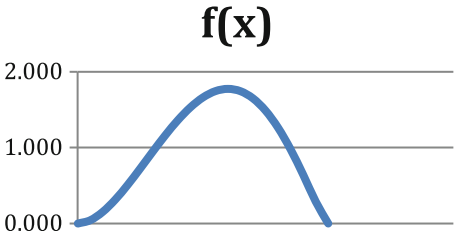

Fig. 6.4 Beta probability density when $k_1 = 2.958$, $k_2 = 2.305$, a = 50 and b = 100

variable. The parameter values are obtained from the mean and mode of the distribution. This allows defining the probability density, which also includes a beta function. The beta function is composed of the gamma function whose value is not easily computed except by a computer algorithm. The cumulative probability of the beta does not have a closed form solution, and quantitative methods are needed to compute. When sample data is available, the average and mode are used to estimate the parameter values. When sample data is not available, the limits, (a, b) and the most-likely value are provided by a knowledgeable person. With this data, the parameters are estimated.

Chapter 7
Weibull

7.1 Introduction

The Weibull distribution was formally introduced by Waloddi Weibull, a Swedish mathematician in 1939. The distribution was earlier used by a Frenchman, Maurice Frechet in 1927, and applied by R. Rosin and E. Rammler in 1933. The Weibull distribution has shapes that range from exponential-like to normal-like, and the random variable, w, takes on values of γ or larger. A related distribution, the standard Weibull with variable, x, has values of zero or larger. Both distributions have the same parameters (k_1, k_2) and these form the shape of the distribution. When $k_1 \leq 1$, the mode of the standard Weibull is zero and the shape is exponential-like; when $k_1 > 1$, the mode is larger than zero, and when k_1 is 3 or larger, the shape is normal–like. The mathematical equations for the probability density and the cumulative probability are shown and are easy to compute. However, the calculation of the mean and variance of the distribution are not so easy to compute and require use of the gamma function. Methods to estimate the parameters, γ, k_1, k_2, are described when sample data is available. When no data is available, and an expert type person provides approximations of some measure of the distribution, methods are shown how to estimate the parameter values.

7.2 Fundamentals

The random variable of the Weibull distribution is listed as w and has a range of (γ and above), while a counterpart standard Weibull random variable x has a range from (0 and above). The parameters for the Weibull are the following:

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_7

$(k_1 > 0, k_2 > 0) =$ parameters
$\gamma =$ min location parameter for w
$0 =$ min location parameter for x

The relations between w and x are below:

$$x = w - \gamma$$
$$w = x + \gamma$$

7.3 Standard Weibull

The probability density of the standard Weibull is below:

$$f(x) = k_1 {k_2}^{-k1} x^{k1-1} \exp\left[-(x/k_2)^{k1}\right] \qquad x > 0$$

and the cumulative distribution becomes:

$$F(x) = 1 - \exp\left[-(x/k_2)^{k1}\right] \qquad x > 0$$

The mean and variance of x are listed below:

$$\mu = k_2/k_1 \Gamma(1/k_1)$$
$$\sigma^2 = {k_2}^2/k_1 \left[2\Gamma(2/k_1) - 1/k_1 \Gamma(1/k_1)^2\right]$$

The mode is the following:

$$\begin{aligned} \widetilde{\mu} \quad &= k_2[(k_1 - 1)/k_1]^{1/k1} && \text{when } k_1 \geq 1 \\ &= 0 && \text{when } k_1 < 1 \end{aligned}$$

The α-percent point of x is denoted as xα, and is obtained as below:

$$F(x\alpha) = \alpha = 1 - \exp\left[-(x\alpha/k_2)^{k1}\right]$$

and applying some algebra,

$$x\alpha = -k_2 \ln (1 - \alpha)^{1/k1}$$

where ln is the natural logarithm.

Example 7.1 An analyst is using a Weibull distribution with the following parameters:

$$\gamma = 100$$
$$k_1 = 1.97$$
$$k_2 = 28.66$$

The related standard Weibull becomes:

$$x = w - 100,$$

and the probability density of x is:

$$f(x) = 1.97 \times 28.66^{1.97} x^{0.97} \exp - (x/28.66)^{1.97} \qquad x > 0$$

The cumulative probability of less or equal to $x = 50$, say, becomes:

$$F(50) = 1 - \exp - (50/28.66)^{1.97}$$
$$= 0.95$$

Hence, the probability of w less than 150 is:

$$P(w \leq 150) = 0.95.$$

The 0.95%–point of x is computed as below:

$$x_{0.95} = -28.66 \times \ln (1 - 0.95)^{1/1.97} = 50$$

and the counterpart for the random variable w is:

$$w_{0.95} = 100 + 50 = 150$$

Figure 7.1 depicts the probability density of x.

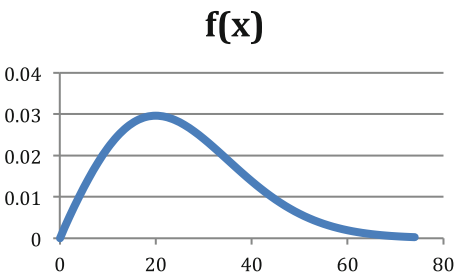

Fig. 7.1 Weibull probability density when $k_1 = 1.97$ and $k_2 = 28.66$

7.4 Sample Data

When sample data ($w_1, \ldots, w_n$), is available, the pertinent measures of the Weibull are the following:

γ = location parameter of w
$\tilde{w}$ = mode of w
$w_{0.5}$ = median of w (0.5-percent point)

7.5 Parameter Estimate of γ When Sample Data

[Zanakis (1979) p 101–116] provides a way to estimate the (min) location parameter of w as listed below:

$$\gamma = [w(1)w(n) - \quad w(k)^2]/[w(1) + w(n) - 2w(k)]$$

where

$w(1) = \min(w_1, \ldots, w_n)$
$w(n) = \max(w_1, \ldots, w_n)$
k is smallest integer where $w(k) > w(1)$

Example 7.2 A researcher has n = 10 sample observations and wants to use the Weibull distribution in a research study. The observations are the following:

$$49, 37, 63, 84, 71, 55, 48, 62, 58, 43$$

A first concern is to estimate the min location parameter, γ. Applying the Zanakis formula,

$w(1) = 37$
$w(10) = 84$
$w(k) = 43$

and thereby the min location parameter is estimated as:

$$\gamma = \left(37 \times 84 - 43^2\right)/(37 + 84 - 2 \times 43) = 35.97$$

7.6 Parameter Estimate of (k_1, k_2) When Sample Data

It is possible to estimate the parameters (k_1, k_2) when $k_1 \geq 1$, and hence, where the mode is larger than zero. To accomplish, the Weibull sample measures ($\tilde{w}$, $w_{0.5}$) are converted to the counterpart standard Weibull measures as below:

$$\tilde{x} = \tilde{w} - \gamma = \text{mode of x}$$

$$x_{0.5} = w_{0.5} - \gamma = \text{median of x (0.50\%-point of x)}$$

When $k_1 < 1$, the mode of x is at $x = 0$. The analysis here is when $k_1 \geq 1$ and the mode of x is greater than zero. For this situation, the mode is measured as below.

$$\tilde{x} = k_2[(k_1 - 1)/k_1]^{1/k1}$$

Using some algebra, k_2 becomes

$$k_2 = \tilde{x}/[(k_1 - 1)/k_1]^{1/k_1}$$

The α-percent-point of x is x_α and is obtained by the following,

$$F(x_\alpha) = 1 - \exp\left[-(x_\alpha/k_2)^{k1}\right] = \alpha$$

Hence,

$$\ln(1 - \alpha) = -[x_\alpha/k_2]^{k1}$$

Applying more algebra and solving for k_2,

$$k_2 = x_\alpha/\ln[1/(1 - \alpha)]^{1/k1}$$

Note, the only unknown in the equation below is k_1,

$$\tilde{x}/[(k_1 - 1)/k_1]^{1/k_1} = x_\alpha/\ln[1/(1 - \alpha)]^{1/k_1}$$

whereby,

$$\tilde{x}/x_\alpha = \{(k_1 - 1)/[k_1 \times \ln[1/(1 - \alpha)]]\}^{1/k_1}$$

Substituting $\alpha = 0.50$ and $x_{0.5} = x\alpha$, the following results:

$$\tilde{x}/x_{0.5} = \{(k_1 - 1)/[k_1 \times \ln(2)]\}^{1/k_1}$$

Solving for k_1 An iterative search is made to find the value of $k_1$where the right side of the above equation is equal to the left side. The result is $\hat{k}_1$.

Solving for k_2 Having found $\hat{k}_1$, the other parameter, k_2, is now obtained from

$$\hat{k}_2 = \tilde{x}/\left[\left(\hat{k}_1 - 1\right)/\hat{k}_1\right]^{1/\hat{k}_1}$$

Example 7.3 Using the same data from Example 7.2, the researcher next needs to estimate the Weibull parameters, (k_1, k_2), and this requires estimates of the mode, $\tilde{x}$, and median, $x_{0.5}$ from the standard Weibull.

One way to estimate the mode is to arrange the data into the six subgroups of size 10 as below, and count the number of observations in each subgroup. The subgroup with the highest number is the location of the mode. The average of the observation in the selected subgroup is a way to estimate the value of the mode. In the example, the computations for the Weibull and the standard Weibull estimates are below:

$$\tilde{w} = (48 + 49 + 43)/3 = 46.67$$
$$\tilde{x} = (46.67 - 35.97) = 10.70$$

Subgroup	30–29	40–49	50–59	60–69	70–79	80–89
Number	1	3	2	2	1	1

The next step is to estimate the median of the data. This is the 0.50%-point denoted as $x_{0.50}$. To accomplish, the $n = 10$ data points are sorted from low to high as below:

$$37, 43, 48, 49, 55, 58, 62, 63, 71, 84$$

The mid value is halfway between the 5th and 6th sorted observation. The estimates for the Weibull median and the counterpart standard Weibull median are below:

$$w_{0.5} = (55 + 58)/2 = 56.50$$
$$x_{0.5} = (56.50 - 35.97\ ) = 20.53$$

With the two standard Weibull measures ($\tilde{x}$, $x_{0.5}$) from the data, it is now possible to estimate the two parameter values of the Weibull, k_1, k_2, using the iterative method described earlier. The ratio of the mode over the median becomes:

$$\tilde{x}/x_{0.5} = 10.70/20.53 = 0.52$$

Recall the equation listed earlier:

$$\tilde{x}/x_\alpha = \{(k_1 - 1)/[k_1 \times \ln[1/(1 - \alpha)]]\}^{1/k_1}$$

Since the left-hand-side (LHS) of the equation is equal to 0.52, an iterative search on the right-hand-side (RHS) is performed with $\alpha = 0.50$, and values of k_1 from 1.1 and higher. When the RHS is reasonably close to 0.52, the value of k_1 is identified. In the table below, this occurs when $\hat{k}_1 = 1.55$.

k_1	RHS
1.1	0.128
1.2	0.233
1.3	0.325
1.4	0.408
1.5	0.487
1.6	0.562
1.7	0.636
1.8	0.707
1.9	0.779
2.0	0.849

Having an estimate on k_1, it is now possible to estimate k_2 by the equation:

$$\begin{aligned}\hat{k}_2 &= \tilde{x}/\left[\left(\hat{k}_1 - 1\right)/\hat{k}_1\right]^{1/\hat{k}_1} \\ &= 10.70/[0.55/1.55]^{1/1.55} \\ &= 20.87\end{aligned}$$

Hence, the Weibull parameters estimated from the sample data become:

$$\left(\hat{\gamma}, \hat{k}_1, \hat{k}_2\right) = (35.97, 1.55, 20.87)$$

Figure 7.2 depicts the Weibull plot with these parameters.

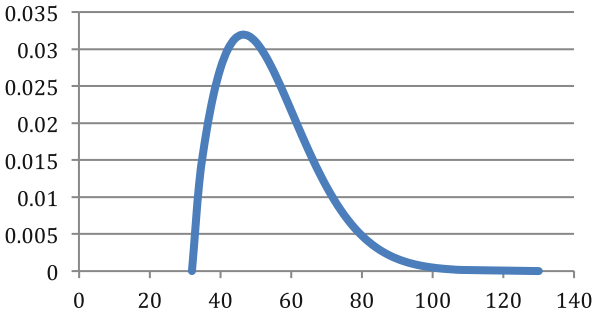

Fig. 7.2 Weibull plot when $\gamma = 35.97$ $k_1 = 1.55$ and $k_2 = 20.87$

7.7 Parameter Estimate When No Data

When no sample data is available and a researcher wants to apply the Weibull distribution, measures of the shape of the distribution are needed and best estimates are provided of the following:

$\gamma =$ location parameter of w
$\tilde{w} =$ most $-$ likely value of w (mode)
$(w\alpha, \alpha) = \alpha$-percent-point of w

Using these measures, the conversion to the standard Weibull follows:

$\tilde{x} =$ estimate of mode of x
$(x\alpha, \alpha) = \alpha$-percent-point of x

Using $(\tilde{x}, x\alpha, \alpha)$, the iterative algorithm listed above is again used to estimate the Weibull parameters (k_1, k_2).

Example 7.4 Assume an analyst is seeking to use the Weibull distribution and has no sample data to estimate the parameters. With the aid of an expert, the following estimates are given:

$\gamma = \min = 10$
$\tilde{w} =$ most $-$ likely value $= 20$
$w_{0.9} = 0.90\%$-point $= 30$
$\alpha = 0.90$

Using the above estimates, the mode for the standard Weibull becomes:

$$\tilde{x} = (20 - 10) = 10$$

In the equation below, the left-hand-side is equal to $10/20 = 0.50$, and the right-hand-side, with $\alpha = 0.90$, is searched for the value of k_1 (from 1.1 to 2.1) when the RHS is close to 0.50. With interpolation, $k_1 \approx 2.06$.

$$\tilde{x}/x_\alpha = \{(k_1 - 1)/[k_1 \times \ln[1/(1 - \alpha)]]\}^{1/k_1}$$

k_1	RHS
1.1	0.042
1.2	0.085
1.3	0.129
1.4	0.173
1.5	0.218
1.6	0.265
1.7	0.313
1.8	0.363
1.9	0.413
2.0	0.466
2.1	0.519

Using $\hat{k}_1 = 2.06$, the equation below yields the estimate of $\hat{k}_2$.

$$\begin{aligned}\hat{k}_2 &= \tilde{x}/\left[\left(\hat{k}_1 - 1\right)/\hat{k}_1\right]^{1/\hat{k}_1} \\ &= 10/[1.06/2.06]^{1/2.06} \\ &= 13.80\end{aligned}$$

Finally, the Weibull parameters become:

$$\left(\hat{\gamma}, \hat{k}_1, \hat{k}_2\right) = (10, 2.06, 13.80)$$

and a plot of the Weibull distribution is in Fig. 7.3.

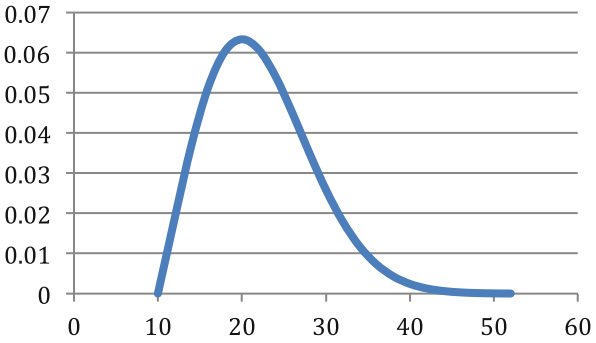

Fig. 7.3 Weibull plot when $\hat{\gamma} = 10$, $\hat{k}_1$, $= 2.06$ and $\hat{k}_2 = 13.80$

7.8 Summary

The Weibull random variable, w, ranges from γ and above, and a related distribution is the standard Weibull, whose variable, x, is zero or larger. Both also have the same parameters (k_1, k_2) that form the shape of the distribution. The mathematical manipulations are mostly with the standard Weibull. When k_1 is less or equal to one, the shape of the standard Weibull is exponential-like, and when k_1 is larger than one the mode is greater than zero. As k_1 increases to three or larger, the shape looks normal–like. A method is described on how sample data is used to estimate the location parameter γ for the Weibull. Also, when estimates of the mode and an α-percent-point are available, an iterative algorithm is used to estimate the parameters (k_1, k_2). The chapter gives examples on how to use the distribution when sample data is available and also when no sample data is available.

Chapter 8
Normal

8.1 Introduction

In 1809, Carl Friedrich Gauss introduced the method of least squares, the maximum likelihood estimator method and the normal distribution, which is often referred as the Gaussian distribution. The normal distribution is the most commonly used distribution in all disciplines. Tne normal has a random variable x with two parameters, μ is the mean, and σ is the standard deviation. A related distribution is the standard normal with random variable z whose mean is zero and standard deviation is one. An easy way to convert from x to z and also from z to x is shown. Tables on the standard normal distribution are available in almost all statistics books. There is no closed-form solution to the cumulative probability, denoted as F (z), and thereby various quantitative methods have been developed over the years. This chapter provides the Hasting's approximation formula to find F(z) from z: and also another Hasting's approximation formula to find z from F(z). When sample data is available, the sample average and sample standard deviation are used to estimate the mean and standard deviation of the normal. When sample data is not available, a method is shown on how to estimate the mean and standard deviation of the normal from some approximate measures of the distribution.

8.2 Fundamentals

The random variable of the normal distribution is denoted as x and has a range of $-\infty$ to $+\infty$. The probability density is the following:

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_8

$$f(x) = \sqrt{\left(\frac{1}{2\pi\sigma^2}\right)} e^{-0.5[(x-\mu)/\sigma]^2}$$

where

μ = mean
σ^2 = variance
σ = standard deviation

The notation for the normal distribution is below:

$$x \sim N(\mu, \sigma^2)$$

The cumulative probability of the random variable being less or equal to x is denoted as F(x), and when $x = x_o$, say, the probability is the following:

$$\begin{aligned} F(x_o) &= P(x \leq x_o) \\ &= \int_{-\infty}^{xo} f(x)dx \end{aligned}$$

There is no closed-form solution to the above, and quantitative methods have been developed to generate tables on F(x).

8.3 Standard Normal

The random variable of the standard normal distribution is denoted as z and has mean zero and standard deviation one. Figure 8.1 depicts the shape of the standard normal. The notation for z is listed below:

$$z \sim N(0, 1)$$

The way to convert from x (normal distribution) to z (standard normal distribution), and vice versa is shown here:

$$z = (x - \mu)/\sigma$$
$$x = \mu + z\sigma$$

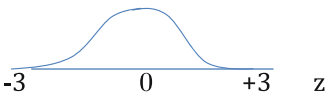

Fig. 8.1 The standard normal distribution

The probability density, cumulative probability and complementary probability are obtained as below, where k represents a particular value of z.

$$\begin{aligned} f(z) &= \left(1/\sqrt{2\pi}\right)e^{-z^2/2} && = \text{probability density of z} \\ F(k) &= P(z \le k) = \int_{-\infty}^{k} f(z)dz && = \text{cumulative probability of z = k} \\ H(k) &= P(z > k) = 1 - F(k) && = \text{complementary probability of z = k} \end{aligned}$$

Two integral identities on the standard normal are below:

$$\int_{-\infty}^{\infty} zf(z)dz = 0$$

$$\int_{-\infty}^{\infty} z^2 f(z)dz = 1$$

Note since z is a continuous variable, $F(k) = P(z \le k) = P(z < k)$.

8.4 Hastings Approximations

Since there is no closed-form solution for the cumulative distribution F(z); various quantitative ways to estimate the cumulative probability of the standard normal have been developed over the years. Two methods that concern the relation between F(z) and z are provided by C.Hastings [Abramowitz and Stegun (1964). p 931–936] and are described in this chapter.

Approximation of F(z) from z For a given z, to find F(z), the following Hastings routine is run.

1. $d_1 = 0.0498673470$
   $d_2 = 0.0211410061$
   $d_3 = 0.0032776263$
   $d_4 = 0.0000380036$
   $d_5 = 0.0000488906$
   $d_6 = 0.0000053830$
2. If $z \ge 0$: $w = z$
   If $z < 0$: $w = -z$
3. $F = 1–0.5[1 + d_1w + d_2w^2 + d_3w^3 + d_4w^4 + d_5w^5 + d_6w^6]^{-16}$
4. if $z \ge 0$: $F(z) = F$
   If $z < 0$: $F(z) = 1 - F$

Return F(z).

Approximation of z from F(z) Another useful approximation also comes from Hastings, and gives a routine that yields a random z from a value of F(z). The routine is listed below.

1. $c_0 = 2.515517$
   $c_1 = 0.802853$
   $c_2 = 0.010328$
   $d_1 = 1.432788$
   $d_2 = 0.189269$
   $d_3 = 0.001308$
2. $H(z) = 1 - F(z)$
   If $H(z) \leq 0.5$: $H = H(z)$
   If $H(z) > 0.5$: $H = 1 - H(z)$
3. $t = \sqrt{\ln\left(1/H^2\right)}$ where ln = natural logarithm.
4. $w = t - [c_0 + c_1 t + c_2 t^2]/[1 + d_1 t + d_2 t^2 + d_3 t^3]$
5. If $H(z) \leq 0.5$: $z = w$
   If $H(z) > 0.5$: $z = -w$

Return z.

8.5 Tables of the Standard Normal

Table 8.1 lists values of k, F(k), H(k), and f(k), from the standard normal distribution that fall in the range: [−3.0, (0.1), +3.0]. Figure 8.2 depicts F(k), f(k) and k for standard normal variable z.

Example 8.1 A researcher has data that is normally distribution with mean 50 and standard deviation 4 and is seeking the probability of the variable less or equal to 60. Hence,

$$x \sim N\left(50, 4^2\right)$$
$$z = (60 - 50)/4 = 2.5$$

Searching Table 8.1 when $z = 2.5$ yields:

$$F(2.5) = 0.994$$

Hence,

$$P(x \leq 60) = F(2.5) = 0.994$$

Example 8.2 Assume the same normal variable as Example 8.1 and now the researcher is seeking the probability of the variable falling between 42 and 58. To accomplish, the following five steps are taken:

Table 8.1 The standard normal statistics sorted by k; with cumulative distribution, F(k); complement probability, H(k); and probability density, f(k)

k	F(k)	H(k)	f(k)	k	F(k)	H(k)	f(k)
−3.0	0.001	0.999	0.004	0.0	0.500	0.500	0.399
−2.9	0.002	0.998	0.006	0.1	0.540	0.460	0.397
−2.8	0.003	0.997	0.008	0.2	0.579	0.421	0.391
−2.7	0.003	0.997	0.010	0.3	0.618	0.382	0.381
−2.6	0.005	0.995	0.014	0.4	0.655	0.345	0.368
−2.5	0.006	0.994	0.018	0.5	0.691	0.309	0.352
−2.4	0.008	0.992	0.022	0.6	0.726	0.274	0.333
−2.3	0.011	0.989	0.028	0.7	0.758	0.242	0.312
−2.2	0.014	0.986	0.035	0.8	0.788	0.212	0.290
−2.1	0.018	0.982	0.044	0.9	0.816	0.184	0.266
−2.0	0.023	0.977	0.054	1.0	0.841	0.159	0.242
−1.9	0.029	0.971	0.066	1.1	0.864	0.136	0.218
−1.8	0.036	0.964	0.079	1.2	0.885	0.115	0.194
−1.7	0.045	0.955	0.094	1.3	0.903	0.097	0.171
−1.6	0.055	0.945	0.111	1.4	0.919	0.081	0.150
−1.5	0.067	0.933	0.130	1.5	0.933	0.067	0.130
−1.4	0.081	0.919	0.150	1.6	0.945	0.055	0.111
−1.3	0.097	0.903	0.171	1.7	0.955	0.045	0.094
−1.2	0.115	0.885	0.194	1.8	0.964	0.036	0.079
−1.1	0.136	0.864	0.218	1.9	0.971	0.029	0.066
−1.0	0.159	0.841	0.242	2.0	0.977	0.023	0.054
−0.9	0.184	0.816	0.266	2.1	0.982	0.018	0.044
−0.8	0.212	0.788	0.290	2.2	0.986	0.014	0.035
−0.7	0.242	0.758	0.312	2.3	0.989	0.011	0.028
−0.6	0.274	0.726	0.333	2.4	0.992	0.008	0.022
−0.5	0.309	0.691	0.352	2.5	0.994	0.006	0.018
−0.4	0.345	0.655	0.368	2.6	0.995	0.005	0.014
−0.3	0.382	0.618	0.381	2.7	0.997	0.003	0.010
−0.2	0.421	0.579	0.391	2.8	0.997	0.003	0.008
−0.1	0.460	0.540	0.397	2.9	0.998	0.002	0.006
				3.0	0.999	0.001	0.004

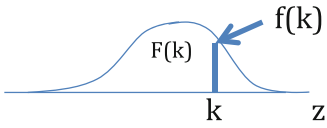

Fig. 8.2 F(k), f(k), k and z of the standard normal distribution

1. $z_1 = (58–50)/4 = 2.0$
2. $z_2 = (42–50)/4 = -2.0$
3. $F(2.0) = 0.977$
4. $F(-2.0) = 0.023$
5. $P(42 \leq x \leq 58) = F(2.0) – F(-2.0) = 0.954$

8.6 Sample Data

When sample data ($x_1, \ldots, x_n$) is available, the sample average, $\bar{x}$, variance, s^2, and standard deviation are computed as below:

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$$

$$s = \sqrt{s^2}$$

Example 8.3 An analyst has the following $n = 10$ randomly selected observations: (123, 109, 94, 87, 105, 111, 89, 106, 99, 136), and computes the sample average, variance and standard deviation as below:

$$\bar{x} = 1059/10 = 105.9$$

$$s^2 = 2066.9/9 = 117.66$$

$$s = \sqrt{117.66} = 10.85$$

8.7 Parameter Estimates When Sample Data

When sample data is available, the sample average, $\bar{x}$, and sample standard deviation, s, are the estimates of the population mean, μ, and standard deviation, σ, respectively.

Example 8.4 Using the data from Example 8.3, the estimates of the mean and standard deviation become:

$$\hat{\mu} = \bar{x} = 105.9$$

$$\hat{\sigma} = s = 10.85$$

Assuming the data is normally distributed,

$$x \sim N(105.9, 10.85^2)$$

8.8 Parameter Estimates When No Data

In the event an analyst has no sample data to estimate the parameter values, an expert is called to give some approximate measures on the distribution. Two methods are described.

The first method requires the approximate measures listed below:

$x\alpha_1 = \alpha_1$-percent-point on x
$x\alpha_2 = \alpha_2$-percent-point on x

With the above data, the following two equations are listed:

$$x\alpha_1 = \mu + z\alpha_1\sigma$$
$$x\alpha_2 = \mu + z\alpha_2\sigma$$

where

$\alpha_1 = \alpha_1$-percent-point on $z \sim N(0, 1)$
$z\alpha_2 = \alpha_2$-percent-point on $z \sim N(0, 1)$

With some algebra, the estimates of μ and σ are below:

$$\widehat{\sigma} = (x\alpha_2 - x\alpha_1)/(z\alpha_2 - z\alpha_1)$$
$$\widehat{\mu} = x\alpha_1 - z\alpha_1\widehat{\sigma}$$

Example 8.5 A researcher is designing a simulation model and needs a normal distribution, but has no sample data. The following approximations are given on the random variable:

$$(x\alpha_1, \alpha_1) = (50, 0.05)$$
$$(x\alpha_2, \alpha_2) = (100, 0.95)$$

Table 8.1 is searched to find the values of z that correspond to $\alpha_1 = F(z_{0.05}) = 0.05$ and $\alpha_2 = F(z_{0.95}) = 0.95$. These are below:

$$z_{0.05} = -1.645$$
$$z_{0.95} = 1.645$$

Finally, the estimates of the normal standard deviation and mean are below:

$$\widehat{\sigma} = [100 - 50]/[1.645 - (-1.645)] = 15.20$$

$$\widehat{\mu} = 50 - (-1.645) \times 15.20 = 75.00$$

The second method merely needs approximation on the following measures of the distribution:

$$\begin{aligned} &\mathrm{L} = \text{low limit} \\ &\mathrm{H} = \text{high limit} \end{aligned}$$

Since the interval from L to H on a normal distribution is six standard deviations, and the mean is half-way between L and H, the estimates of the parameters are the following:

$$\begin{aligned} \widehat{\mu} &= (\mathrm{L} + \mathrm{H})/2 \\ \widehat{\sigma} &= (\mathrm{H} - \mathrm{L})/6 \end{aligned}$$

Example 8.6 A researcher wants to apply the normal distribution in a study but has no sample data. However best estimates on the min and max of the variable are available and are the following:

$$\begin{aligned} &\mathrm{L} = 10 \\ &\mathrm{H} = 90 \end{aligned}$$

Hence, the parameter estimates become the following:

$$\begin{aligned} \widehat{\mu} &= (10 + 90)/2 = 50 \\ \widehat{\sigma} &= (90 - 10)/6 = 13.3 \end{aligned}$$

8.9 Summary

The normal distribution is bell-shaped with random variable x and parameters, μ and σ. A related distribution is the standard normal with random variable z whose mean is zero and standard deviation is one. Most statistics books have tables on the standard normal distribution. An easy conversion from x to z and also from z to x is shown in the chapter. Because there is no closed-form solution to the cumulative probability F(z), an approximate formula is given in the chapter. There also is an approximate formula to find z from a given value of F(z). When sample data is available, the sample average and standard deviation are used to estimate the mean and standard deviation of the normal distribution. When no sample data is available, approximation measures of the distribution are provided that allow estimates of the mean and standard deviation of the normal.

Chapter 9
Lognormal

9.1 Introduction

Two British mathematicians, Francis Galton and Donald McAlister, introduced the lognormal distribution in 1879. The lognormal distribution is sometimes referred as the Galton distribution. The lognormal variable begins at zero, its density peaks soon after and thereafter tails down to higher x values. The variable x is lognormal distributed when another variable, y, formed by the logarithm of x, becomes normally distributed. The probability density of x is listed in the chapter, while the associated cumulative distribution function, F(x), is not since there is no closed–form solution. A method to compute the cumulative probability of any x is provided. When sample data is available, the measures from the sample are used to estimate the parameters of the lognormal. In the event no sample data is available and estimates of the lognormal variable are needed, two methods are described on how to compute the estimates. The lognormal distribution is not as popularly known as the normal, but applies easily in research studies of all kinds. It has applications in many disciplines, such as weather, engineering and economics.

9.2 Fundamentals

The variable x is lognormal when the natural log, ln, of x is normally distributed. The relation between x and y are below:

$$y = \ln(x)$$
$$x = e^{y}$$

The parameters of y are the mean, μ_y, and standard deviation, $\sigma_{y.}$ and of x they are μ_x and σ_x. The symbols for the lognormal distribution of x and the

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_9

corresponding normal distribution of y are below. Note, the parameters that describe the distribution of x are the mean and standard deviation of y.

$$x \sim LN\left(\mu_y, {\sigma_y}^2\right)$$
$$y \sim N\left(\mu_y, {\sigma_y}^2\right)$$

The probability density of x, f(x), is listed below, but not the cumulative distribution, F(x), since there is not a closed-form solution for the latter.

$$f(x) = \sqrt{\left(\frac{1}{2\pi x^2 {\sigma_y}^2}\right)} e^{-0.5\left[\left(y-\mu_y\right)/\sigma_y\right]^2}$$

The relation between the parameters of x and y are listed below:

$$\mu_x = \exp\left[\mu_y + {\sigma_y}^2/2\right]$$
$${\sigma_x}^2 = \exp\left[2\mu_y + {\sigma_y}^2\right]\left[\exp\left({\sigma_y}^2\right) - 1\right]$$
$$\mu_y = \ln\left[{\mu_x}^2/\sqrt{\mu_x^2 + \sigma_x^2}\right]$$
$${\sigma_y}^2 = \ln\left[1 + {\sigma_x}^2/{\mu_x}^2\right]$$

9.3 Lognormal Mode

The mode of the lognormal variable, x, denoted as $\widetilde{\mu}_x$, is obtained as below:

$$\widetilde{\mu}_x = \exp\left(\mu_y - {\sigma_y}^2\right)$$

9.4 Lognormal Median

The median of lognormal x is obtained as follows:

$$\mu_{0.5} = \exp\left(\mu_y\right)$$

Example 9.1 In a lab experiment, one of the variables, x, is detected as lognormal with LN(2.5, 1.1^2). The corresponding mean and variance of the x is computed as below:

$$\mu_x = \exp\left(2.5 + 1.1^2/2\right) = 22.31$$

$$\sigma_x{}^2 = \exp\left(2 \times 2.5 + 1.1^2\right)\left[\exp\left(1.1^2\right) - 1\right] = 1171.3$$

The mode and the median are obtained in the following way:

$$\widetilde{\mu}_x = \exp\left(2.5 - 1.1^2\right) = 3.63$$

$$\mu_{0.5} = \exp(2.5) = 12.18$$

Assume the researcher is seeking the probability that x is less or equal to 30. To accomplish, the computations are below:

$$\begin{aligned} P(x \le 30) &= P[y \le \ln(30)] \\ &= P(y \le 3.40) \\ &= P\left(z \left[\le (3.4 - 2.5)/1.1\right]\right. \\ &= P\left(z \le 0.82\right) \\ &= F_z(0.82) = 0.794 \\ &\cong 0.794 \end{aligned}$$

For clarity, $F_z(z)$ is the cumulative probability of variable z and is listed in Table 8.1.

Further assume the researcher wants to find the 0.90%-point of x. This is obtained by finding the corresponding 0.90%-points of $z \sim N(0, 1)$; $x \sim N(2.5, 1.1^2)$; and then $x \sim LN(2.5, 1.1^2)$, as shown below:

$$z_{0.90} = 1.282$$
$$y_{0.90} = 2.5 + 1.282 \times 1.1 = 3.91$$
$$x_{0.90} = \exp(3.91) = 49.9$$

Note, $F(49.9) = P(x \le 49.9) = 0.90$.

A plot depicting the distribution is in Fig. 9.1.

9.5 Sample Data

When sample data is available as: $(x_1, \ldots, x_n)$, and an analyst is seeking to apply the lognormal distribution, a first step is to convert the data to their natural logarithm counterparts: $(y_1, \ldots, y_n)$, where $y_i = \ln(x_i)$ for i = 1 to n, and ln is the natural log. If the converted data appears as normally distributed, the distribution of the original x

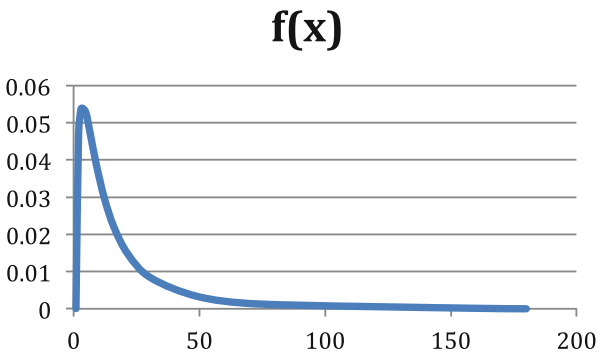

Fig. 9.1 Lognormal plot when x ~ LN(2.5, 1.1^2)

Table 9.1 Sample data (x_1, . . ., x_{13}) and the corresponding $y_i = \ln(x_i)$ i = 1–13

x	y
12	2.485
215	5.371
23	3.135
2	0.693
35	3.555
27	3.296
11	2.398
89	4.489
13	2.565
45	3.807
82	4.407
28	3.332
8	2.079

data is lognormally distributed. The average and standard deviation of the converted data is computed and listed as below:

$\bar{y}$ = sample average
s_y = sample standard deviation

Example 9.2 An analyst has 13 observations of sample data, x_i (i = 1–13), as listed in Table 9.1, and assumes the data is shaped as a lognormal distribution. The table also lists the associate values of $y_i = \ln(x_i)$ for each of the 13 observations.

The sample average of the 13 entries of y, along with the sample variance and sample standard deviation are computed below:

$$\bar{y} = \sum_{i=1}^{13} y_i/13 = 3.201$$
$$s_y{}^2 = \sum_{i=1}^{13} \left[\left(y_i - \bar{y} \right)^2 / 12 \right] = 1.454$$
$$s_y = 1.206$$

Note also, the min, y(1), and max, y(13), of the y entries are the following:

$$y(1) = 0.693$$
$$y(13) = 5.371$$

9.6 Parameter Estimates When Sample Data

When sample data is available and the average, $\bar{y}$, and standard deviation, s_y, are computed, the estimate of the lognormal parameters become:

$$\widehat{\mu}_y = \bar{y}$$
$$\widehat{\sigma}_y = s_y$$

Of further analysis here is to determine if the distribution of the n sample data entries of y are similar to a normal distribution. Recall, if the y entries are normal, the x entries are lognormal. One way to check for normality is by the spread ratio test described in Chap. 10. The spread ratio is computed as below:

$$\theta = [\bar{y} - y(1)]/[y(n) - \bar{y}]$$

where

$$y(1) = \min[y_1, \ldots, y_n]$$
$$y(n) = \max[y_1, \ldots, y_n]$$

If θ is close to 1.00, the distribution of the y entries is deemed similar to a normal distribution. An approximate range for acceptance of the normal is when ($0.70 < \theta < 1.30$).

Example 9.3 Using the same sample data of Example 9.3, the researcher wants to test whether the x sample data is sufficiently similar to a lognormal distribution. When so, the estimates of the parameters are proper to apply in the analysis of the lognormal application. To begin, the researcher computes the spread ratio, θ, as shown below:

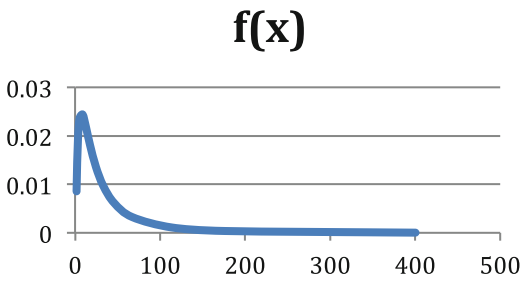

Fig. 9.2 Lognormal plot when x ~ LN(3.201, 1.206^2)

$$\theta = (3.201 - 0.693)/(5.371 - 3.201) = 1.15$$

Since the spread ratio is near 1.00, the y entries are distributed similarly to a normal distribution. Thereby, the y entries are assumed normally shaped, and the associated x entries are assumed lognormal shaped. Figure 9.2 depicts the shape of the lognormal.

Finally, the estimates of the lognormal parameters are the following:

$$\widehat{\mu}_y = \bar{y} = 3.201$$
$$\widehat{\sigma}_y = s_y = 1.206$$

9.7 Parameter Estimates When No Data

When no sample data is available and the researcher is seeking to apply the lognormal distribution in a study, methods are available to estimate the parameters when various approximations are provided on the shape of the lognormal distribution. Two such methods are described here.

1. The first method applies when the following 2%-point approximations are given on the shape of the lognormal:

$(x\alpha_i, \alpha_i) = \alpha_i$-percent-point on x
$(x\alpha_2, \alpha_2) = \alpha_2$-percent point on x

From the standard normal, the corresponding percent-points on z are the following:

$z\alpha_i = \alpha_i$-percent-point on z
$z\alpha_2 = \alpha_2$-percent-point on z

Recall, the two normal relations below that form two equations and two unknowns:

$$y\alpha_i = \mu_y + z\alpha_i\sigma_y$$
$$y\alpha_2 = \mu_y + z\alpha_2\sigma_y$$

Finally, the estimates of the two lognormal parameters are obtained as below:

$$\widehat{\sigma}_y = (y\alpha_2 - y\alpha_i)/(z\alpha_2 - z\alpha_i)$$
$$\widehat{\mu}_y = \left(y\alpha_i - z\alpha_i\widehat{\sigma}_y\right)$$

2. A second method to estimate the lognormal parameters when there is no sample data, requires the following type of approximations to the log-normal distribution:

$$\tilde{x} = \text{most-likely value of x}$$
$$x_{0.5} = \text{mid-value of x}$$

Recall, the lognormal relations on the mode and median below:

$$\tilde{x} = e^{\mu_y - \sigma_y^2}$$
$$x_{0.5} = e^{\mu_y}$$

Applying some algebra, the two lognormal parameters are estimated as follows:

$$\widehat{\mu}_y = \ln(x_{0.5})$$
$$\widehat{\sigma}_y = \left[\widehat{\mu}_y - \ln\left(\tilde{x}\right)\right]^{0.5}$$

Example 9.4 An engineer wants to apply the lognormal in a study but has no sample data to estimate the parameter values. With some analysis, the engineer has approximates on the following 2%-points of x:

$$x_{0.2} = 10$$
$$x_{0.8} = 70$$

The corresponding percent-points on the counterpart normal are obtained as below:

$$y_{0.2} = \ln(10) = 2.302$$
$$y_{0.8} = \ln(70) = 4.248$$

Also, the related percent-points on the standard normal are the following:

$$z_{0.2} = -0.841$$
$$z_{0.8} = 0.841$$

The estimates of the lognormal parameters are the following:

$$\widehat{\sigma}_y = (4.248 - 2.302)/[0.841 - (-0.841)] = 1.157$$
$$\widehat{\mu}_y = 2.302 - (-0.841) \times 1.157 = 3.275$$

Finally, the variable x has the lognormal distribution as below:

$$\mathrm{x} \sim \mathrm{LN}\left(3.275, 1.157^2\right)$$

Example 9.5 In a simulation study, the analyst needs to apply a lognormal distribution and has no sample data, but has the following two approximations on the distribution:

$$\tilde{x} = 100$$
$$\mathrm{x}_{0.5} = 200$$

Using the method described above, the estimate of the two lognormal parameters is the following:

$$\widehat{\mu}_y = \ln(200) = 5.298$$
$$\widehat{\sigma}_y = [5.298 - \ln(100)]^{0.5} = 0.832$$

The lognormal variable to use in the simulation study is below:

$$\mathrm{x} \tilde{} \mathrm{LN}\left(5.298, 0.832^2\right)$$

9.8 Summary

The variable x is lognormally distributed when another variable, y, the log of x, is normally distributed. With easy transformations, x and y are readily related. Equations showing how to convert the mean and variance of x to the mean and variance of y, are listed. The parameters that describe the lognormal variable, x, are the mean and variance of y. The probability density of x is listed, but the associated cumulative distribution function is not shown since there is no closed-form solution to the latter. In the event the cumulative probability of a particular value of x is needed, the method to compute is easily applied. When sample data is available, measures from the data is used to estimate the lognormal parameters needed. When no sample data is available, two methods are described that yield estimates of the lognormal parameters.

Chapter 10
Left Truncated Normal

10.1 Introduction

In an earlier book [Thomopoulos (1980) p 318–324], the author shows how to use the left-truncated normal distribution to applications in inventory control.

The left-truncated normal (LTN) has many shapes that range from normal to exponential-like. The distribution has one parameter k that is a particular value of the standard normal z, and the distribution includes all values of z larger than k. The variable is denoted as t where $t = (z - k)$, and since z is equal or larger than k, t is zero or larger. The key measures of the distribution are the mean, standard deviation, coefficient-of-variation, and a new measure called the spread ratio. The spread ratio is always larger than zero. Another measure is $t\alpha = \alpha$-percent –point of t that gives the value of t with cumulative probability α. Of immediate interest is $t_{0.01}$ and $t_{0.99}$. All of these measures are listed in a table for each k ranging from -3.0 to $+3.0$. When an analyst has sample data $(x_1, \ldots, x_n)$ and draws certain stats from the data including an estimate of the spread-ratio, the analyst can identify if any of the LTN distributions are a better fit to the data than the normal distribution. If any is selected, the low-limit, denoted as γ, where $x \geq \gamma$ is caluculated. The chapter also shows how to compute $x\alpha = \alpha$-percent-point of x for any α.

10.2 Fundamentals

The left-truncated normal distribution takes many shapes from normal to exponential-like, and has one parameter k that forms the shape of the distribution. A quick review follows on the standard normal distribution with variable z. Since almost all of the z values fall between -3.0 and $+3.0$, the analysis in this chapter includes only this range.

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_10

10.3 Standard Normal

The standard normal distribution is perhaps the most utilized probability distribution by researchers in all type of disciplines. This distribution is described fully in Chap. 8. The random variable, z, has limits from minus to plus infinity, with mean zero and variance one, and the designation is listed below:

$$z \sim N(0, 1)$$

The expected value of z and variance are the following:

$$E(z) = 0$$
$$V(z) = 1$$

Below is a listing of the probability density of z, the cumulative probability function and the complementary probability. There is no closed form solution to the cumulative probability and over the years various quantitative methods to compute F(x) have been developed. One such method is described in Chap. 8 and is used again in some of the computations of this chapter.

$$f(z) = \left(1/\sqrt{2\pi}\right)e^{-z^2/2} \quad = \text{probability density of z}$$
$$F(k) = P(z \leq k) = \int_{-\infty}^{k} f(z)dz \quad = \text{cumulative probability of z = k}$$
$$H(k) = P(z > k) = 1 - F(k) \quad = \text{complementary probability of z = k}$$

Note also the integral identity of the standard normal that is listed below:

$$\int_{k}^{\infty} zf(z)dz = f(k)$$

10.4 Left-Truncated Normal

The left-truncated normal random variable, t, is formed by the standard normal variable, z, and a parameter, denoted as k, that is a particular value of z. In this situation, z is greater or equal to k and thereby, t is greater or equal to zero, as shown below:

$$t = z - k \qquad z \geq k$$
$$t \geq 0$$

The designation of the left-truncated-normal distribution is the following:

$$t \tilde{} LTN(k)$$

The probability density of t, g(t), and the cumulative distribution of t, G(t), are obtained as below:

$$g(t) = f(z)/H(k)$$
$$G(t) = [F(z) - F(k)]/H(k)$$

The way to compute the mean and variance of t with parameter k is described here. To begin, the identities for the partial expectations, $E(z > k)$, and $E[(z > k)^2]$, are needed, and are obtained as below:

$$E(z > k) = E(z > k) = \int_k^\infty (z - k) f(z) dz = f(k) - kH(k)$$
$$E\left[(z > k)^2\right] = \int_k^\infty (z - k)^2 f(z) dz = -kf(k) + H(k)\left(1 + k^2\right)$$

The following shows how the expected value of t and t^2 with parameter k are obtained.

$$E(t)_k = E(z > k)/H(k) \quad = \text{expected value of t given k}$$
$$E(t^2)_k = E\left[(z > k)^2\right]/H(k) \quad = \text{expected value of } t^2 \text{ given k}$$
$$V(t)_k = E(t^2)_k - E(t)_k{}^2 \quad = \text{variance of t given k}$$

Finally, the mean and standard deviation of t with parameter k are the following:

$$\mu_t(k) = E(t)_k \quad = \text{mean of t given k}$$
$$\sigma_t(k) = \sqrt{V(t)_k} \quad = \text{standard deviation of t given k}$$

The coefficient-of-variation of variable t with parameter k is obtained as follows:

$$\text{cov}_t(k) = \sigma_t(k)/\mu_t(k)$$

10.5 Cumulative Probability of t

Below shows how to find the value of $t\alpha = \alpha$-percent-point of t. Recall G(t) is the cumulative probability of $t\alpha$ where:

$$G(t\alpha) = P(t \le t\alpha) = \alpha$$

Recall the following relation between, α and k of LTN, and F(z), F(k) and H(k), of the standard normal:

$$G(t) = \alpha = [F(z) - F(k)]/H(k)$$

With some algebra the results yield:

$$\begin{aligned}(1-\alpha) &= 1-[F(z)-F(k)]/H(k)\\ &= [H(k)-F(z)+F(k)]/H(k)\\ &= [1-F(z)]/H(k)\\ &= H(z)/H(k)\end{aligned}$$

Hence, the complementary probability of z from the standard normal is computed as below:

$$H(z) = (1-\alpha)H(k)$$

With a value of H(z) available, the next step is to find the associated value of z, denoted as z`, that corresponds to H(z); and then obtain $t\alpha = z` - k$. This is shown in the five steps below:

1. From k, find H(k) using Table 8.1
2. $H(z`) = (1-\alpha)H(k)$
3. $F(z`) = [1 - H(z`)]$
4. Find z` using Table 8.1 and interpolation
5. $t\alpha = z` - k$

When $\alpha = 0.01$ and 0.99, the following percent-points are obtained:

$t_{0.01}$ = 0.01-percent-point
$t_{.0.99}$ = 0.99-percent-point

Recall, the mean and standard deviation of t with parameter k are the following:

$\mu_{t(k)}$ = mean of variable t with location parameter k
$\sigma_{t(k)}$ = standard deviation of variable t with location parameter k

The spread ratio, denoted as θ, is unique for each k and is computed as shown below:

$$\theta = \left[\mu_{t(k)} - t_{0.01}\right] / \left[t_{0.99} - \mu_{t(k)}\right] = \text{spread-ratio}$$

Table 10.1 lists various measures ($\mu_t(k)$, $\sigma_t(k)$, $cov_t(k)$, $t_{0.01}$, $t_{0.99}$, θ) of t from the LTN for each k = [−3.0, (0.1), 3.0].

Table 10.1 Left truncated normal by location parameter k, percent-points $t_{0.01}$, $t_{0.99}$, mean, $\mu_t(k)$, standard deviation, $\sigma_t(k)$, coefficient-of-variation, cov(k), and spread-ratio, θ

k	$t_{0.01}$	$t_{0.99}$	$\mu_{t(k)}$	$\sigma_{t(k)}$	cov(k)	θ
−3.0	0.720	5.327	3.004	0.993	0.33	0.98
−2.9	0.637	5.227	2.906	0.991	0.34	0.98
−2.8	0.559	5.127	2.808	0.989	0.35	0.97
−2.7	0.486	5.028	2.710	0.986	0.36	0.96
−2.6	0.419	4.928	2.614	0.982	0.38	0.95
−2.5	0.359	4.829	2.518	0.978	0.39	0.93
−2.4	0.305	4.729	2.423	0.972	0.40	0.92
−2.3	0.258	4.630	2.329	0.966	0.41	0.90
−2.2	0.218	4.532	2.236	0.959	0.43	0.88
−2.1	0.183	4.433	2.145	0.951	0.44	0.86
−2.0	0.155	4.335	2.055	0.942	0.46	0.83
−1.9	0.130	4.237	1.968	0.931	0.47	0.81
−1.8	0.110	4.140	1.882	0.920	0.49	0.78
−1.7	0.093	4.043	1.798	0.907	0.50	0.76
−1.6	0.080	3.947	1.717	0.894	0.52	0.73
−1.5	0.068	3.852	1.639	0.879	0.54	0.71
−1.4	0.059	3.758	1.563	0.863	0.55	0.69
−1.3	0.051	3.664	1.490	0.847	0.57	0.66
−1.2	0.044	3.572	1.419	0.830	0.58	0.64
−1.1	0.039	3.481	1.352	0.812	0.60	0.62
−1.0	0.034	3.390	1.288	0.794	0.62	0.60
−0.9	0.030	3.302	1.226	0.775	0.63	0.58
−0.8	0.027	3.214	1.168	0.756	0.65	0.56
−0.7	0.024	3.128	1.112	0.736	0.66	0.54
−0.6	0.022	3.044	1.059	0.717	0.68	0.52
−0.5	0.020	2.962	1.009	0.697	0.69	0.51
−0.4	0.018	2.881	0.962	0.678	0.70	0.49
−0.3	0.017	2.802	0.917	0.659	0.72	0.48
−0.2	0.015	2.724	0.875	0.640	0.73	0.47
−0.1	0.014	2.649	0.835	0.621	0.74	0.45
0.0	0.012	2.576	0.798	0.603	0.76	0.44
0.1	0.011	2.504	0.763	0.585	0.77	0.43
0.2	0.010	2.435	0.729	0.568	0.78	0.42
0.3	0.010	2.367	0.698	0.551	0.79	0.41
0.4	0.009	2.301	0.669	0.534	0.80	0.40
0.5	0.008	2.238	0.641	0.518	0.81	0.40
0.6	0.008	2.176	0.615	0.503	0.82	0.39
0.7	0.007	2.117	0.590	0.488	0.83	0.38
0.8	0.007	2.059	0.567	0.473	0.83	0.38
0.9	0.007	2.003	0.546	0.460	0.84	0.37
1.0	0.007	1.949	0.525	0.446	0.85	0.36
1.1	0.006	1.897	0.506	0.433	0.86	0.36

(continued)

Table 10.1 (continued)

k	$t_{0.01}$	$t_{0.99}$	$\mu_{t(k)}$	$\sigma_{t(k)}$	cov(k)	θ
1.2	0.006	1.846	0.488	0.421	0.86	0.35
1.3	0.006	1.797	0.470	0.409	0.87	0.35
1.4	0.006	1.750	0.454	0.398	0.88	0.35
1.5	0.005	1.704	0.439	0.387	0.88	0.34
1.6	0.005	1.660	0.424	0.376	0.89	0.34
1.7	0.005	1.617	0.410	0.366	0.89	0.34
1.8	0.005	1.575	0.397	0.356	0.90	0.33
1.9	0.005	1.534	0.385	0.347	0.90	0.33
2.0	0.004	1.495	0.373	0.338	0.91	0.33
2.1	0.004	1.456	0.362	0.330	0.91	0.33
2.2	0.004	1.417	0.351	0.321	0.91	0.33
2.3	0.004	1.379	0.341	0.313	0.92	0.33
2.4	0.004	1.340	0.332	0.306	0.92	0.33
2.5	0.003	1.301	0.323	0.298	0.92	0.33
2.6	0.003	1.260	0.314	0.291	0.93	0.33
2.7	0.003	1.218	0.306	0.284	0.93	0.33
2.8	0.002	1.173	0.298	0.277	0.93	0.34
2.9	0.002	1.124	0.291	0.270	0.93	0.35
3.0	0.001	1.070	0.283	0.264	0.93	0.36

Example 10.1 The following five steps describe the computations to find $t_{0.01}$ at k = 1.0.

1. From Table 8.1, H(1.0) = 0.159
2. H(z`) = (1–0.01) $_{\times}$ 0.159 = 0.157
3. F(z`) = 1–0.157 = 0.843
4. z` $_{\approx}$ 1.008 via Table 8.1 and interpolation
5. $t_{0.01} \approx$ 1.008–1.000 = 0.008

Note, Table 10.1 shows $t_{0.01}$ = 0.007 at k = 1.0. The difference in the table and in the computations here is due to rounding.

Example 10.2 The following five steps describe the computations for $t_{0.99}$ at k = 1.0.

1. From Table 8.1, H(1.0) = 0.159
2. H(z`) = (1–0.99)× 0.159 = 0.0016
3. F(z`) = 1–0.0016 = 0.9984
4. z` ≈ 2.94 via Table 8.1 and interpolation
5. $t_{0.01} \approx$ 2.94–1.00 = 1.94

Note, Table 10.1 shows $t_{0.99}$ = 1.949 at k = 1.0. The difference in the table and in the computations here is due to rounding.

Example 10.3 Show the computations to obtain $\mu_t(k)$, standard deviation, $\sigma_t(k)$, coefficient-of-variation, cov(k), spread-ratio, θ of t at k = 1.0.

The mean, standard deviation and coefficient-of-variation are computed as below:

$$\begin{aligned}\mu_t(1.0) &= [f(1.0) - 1.0 \times H(1.0)]/H(1.0)\\ &= [0.242 - 1.0 \times 0.159]/0.159\\ &= 0.522\end{aligned}$$

$$\begin{aligned}E(t^2)_{1.0} &= \left[-1.0f(1.0) + H(1.0)\left(1 + 1.0^2\right)\right]/H(1.0)\\ &= [-1 \times 0.242 + 0.159 \times (1 + 1)]/0.159\\ &= 0.478\end{aligned}$$

$$\begin{aligned}V(t)_{1.0} &= 0.478 - 0.522^2\\ &= 0.453\end{aligned}$$

$$\begin{aligned}\sigma_t(k) &= \sqrt{0.205}\\ &= 0.453\end{aligned}$$

$$cov(1.0) = 0.453/0.522 = 0.868$$

Note the difference in rounding verses the table values listed below at k = 1.0:

$$\mu_t(k) = 0.525$$
$$\sigma_t(k) = 0.446$$
$$cov(k) = 0.85$$

The computation for the spread ratio is shown below using the table values instead of the computed values.

$$\begin{aligned}\theta &= (0.525 - 0.007)/(1.949 - 0.525)\\ &= 0.363\end{aligned}$$

10.6 Sample Data

When sample data (x_1, ..., x_n) is available, the analyst can obtain the following statistical measures on the variable x:

$\bar{x}$ = average
s = standard deviation
$x(1) = \min(x_1, \ldots, x_n)$
$x(n) = \max(x_1, \ldots, x_n)$

10.7 Parameter Estimates When Sample Data

After gathering the statistical measures on the sample data, the estimate of the spread ratio, denoted as $\widehat{\theta}$, is computed in the following way:

$$\widehat{\theta} = [\bar{x} - \mathrm{x}(1)/[\mathrm{x}(\mathrm{n}) - \bar{x}]$$

With the spread ratio now estimated, the analyst searches Table 10.1 to obtain the following measures on the LTN:

$$\begin{aligned} \mathrm{k} &= \text{LTN parameter} \\ \mu_{\mathrm{t}}(\mathrm{k})\big) &= \text{mean of t at k} \\ \sigma_{\mathrm{t}}(\mathrm{k}) &= \text{standard deviation of t at k} \\ \mathrm{t}_{0.01} &= 0.01 - \text{percent} - \text{point of t at k} \end{aligned}$$

As a rule-of-thumb, if $\widehat{\theta}$ is in the range of 0.70 to 1.00, the sample data is not much different than the normal and the normal distribution applies. If $\widehat{\theta} < 0.70$, the LTN distribution with parameter k is deemed the best fit to the sample data.

When LTN is selected, and with k identified, the shape of the distribution becomes clearer, allowing the analyst to seek an estimate on the low-limit of x; and also on estimating various values of the α-percent-point on x. Both of these are described below.

First, the estimate of the low-limit on x is obtained using the following equation:

$$\widehat{\gamma} = \bar{x} + \mathrm{s}[(\mathrm{t}_{0.01} - \mu_{\mathrm{t}}(\mathrm{k}))/\sigma_{\mathrm{t}}(\mathrm{k})]$$

Second, the estimate of the α-percent-point on x, xα, is obtained. To acquire this estimated, the value of its counterpart, tα = α-percent-point of t at k, is needed. This value is obtained by applying the five steps described earlier in this chapter. With tα, now known, the estimate on xα is computed using the following equation:

$$x\alpha = \bar{x} + \mathrm{s}[(t\alpha - \mu_{\mathrm{t}}(\mathrm{k}))/\sigma_{\mathrm{t}}(\mathrm{k})]$$

Example 10.4 An analyst has n random samples from an experiment and obtains the statistical measures listed below.

$$\begin{aligned} \bar{x} &= 100 \\ \mathrm{s} &= 25 \\ \mathrm{x}(1) &= 70 \\ \mathrm{x}(\mathrm{n}) &= 150 \end{aligned}$$

The analyst suspects the data is not normally distributed and seeks to apply the LTN distribution. The estimate of the spread ratio is computed below:

$$\widehat{\theta} = (100 - 70)/(150 - 100) = 0.60$$

The estimated spread ratio is applied in Table 10.1 to find the closest table value of θ. At $\theta = 0.60$, the table yields, $\mathrm{k} = -1.00$ and the following measures for the LTN:

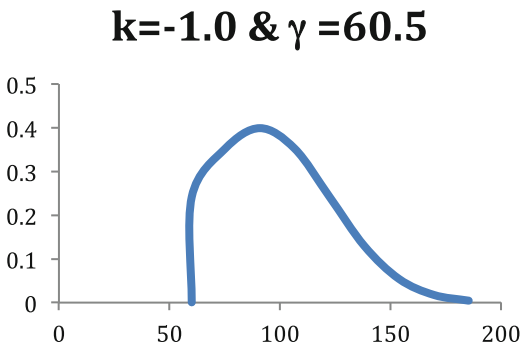

Fig. 10.1 Plot of LTN with $k = -1.00$ and $\gamma = 60.5$

$t_{0.01} = 0.034$
$\mu_t(k) = 1.288$
$\sigma_t(k) = 0.794$

With these measures, the low-limit on x is now estimated as below:

$$\widehat{\gamma} = 100 + 25(0.034 - 1.288)/0.794$$
$$= 60.5$$

Figure 10.1 depicts the plot of the LTN where $k = -1.00$ and $\gamma = 60.5$.

Example 10.5 Using the same data as in Example 10.4, assume the analyst is seeking the value of x with a 0.95 cumulative probability. This is the same as the $x_{0.95}$-percent-point of x and this is shown in the 6 steps below:

1. Find $H(-1.00) = 0.841$ from Table 8.1
2. $H(z`) = (1–0.95)0.841 = 0.042$
3. $F(z`) = (1–0.042) = 0.958$
4. $z` = 1.73$ from Table 8.1
5. $t_{0.95} = 1.73 –(-1.00) = 2.73$
6. $x_{0.95} = 100 + 25[(2.73–1.288)/0.794] = 145.4$

Hence,

$P(x \leq 145.4) \approx 0.95$

10.8 LTN in Inventory Control

Controlling the inventory, when to buy and how much, is a vital function in all distribution centers, retailers, dealers and stores. A typical distribution center has 100,000 part numbers, and a dealer has 20,000. A forecast F is generated each

month on the demands of each part at every location. The standard deviation of the forecast error, s, is computed and the coefficient-of-variation, (cov), is measured as cov = s/F. This cov is a measure of the monthly forecast error on each of the parts and is updated each month. The goal of the inventory management is to provide the minimum amount of stock that is needed to achieve a desired service level (SL) where SL = demand filled/total demand, and the typical service level desired is 0.95. The forecast and the forecast errors are needed to determine the exact amount of stock to have. In most inventory systems, the normal distribution is assumed as the monthly demands for each part. This is not correct, the most typical distribution on the monthly demands is the LTN with the low limit, $\gamma = 0$. It is important to apply the left-truncated normal distribution on the inventory computations for those parts whose monthly demands are non-normal. When the amount of stock to provide for each part by location is always based on the normal, the results are not correct since most of the demands are not normal, but are LTN. Below describes the author's experience in the auto industry and in the retail industry.

10.9 Distribution Center in Auto Industry

This is a large distribution center with over 100,000 parts. The forecasts are generated each month and the buy decisions are monitored each day. The cov of the monthly demands by part are computed each month and are listed in the table below with two columns, cov, and % parts. The % parts is the portion of all parts with a cov as listed from 0.00 to over 1.00. Any part with a cov of 0.50 or less has monthly demands that are similar to a normal distribution, and any part with cov of 0.50 or larger has monthly demands that are LTN. The chart shows that 62 percent of the parts are not normal and are LTN. To provide the best computations on when to buy and how much, the computations using the LTN should be used on the parts with cov of 0.50 or larger.

cov	% parts
0.00–0.30	26
0.30–0.50	12
0.50–0.80	12
0.80–1.00	10
1.00 –	40
Sum	100

10.10 Dealer, Retailer or Store

A typical dealer has 20,000 parts and is limited with the amount of inventory to hold due to space and budget constraints. Their goal is to provide the minimum investment of stock to achieve a high service level. When they run out of stock, their supplier is the distribution center. The monthly demands for each dealer (store, retailer) are much lower than the demands of the same part at the distribution center. The monthly demands do not follow the normal distribution, but have distributions that are either Poisson of LTN. Almost all of the parts in this scenario do not have normal monthly demands. The inventory control computations should be computed using the LTN distribution.

10.11 Summary

The standard left-truncated normal has one parameter k that forms the shape of the distribution ranging from normal to exponential-like. The standard variable is denoted as t and has values of zero and larger. A table is generated listing the key statistical measures of t for selective values of k between -3.0 and $+3.0$. The statistical measures of t are the α-percent-points at $\alpha = 0.01$ and 0.99, the mean, standard deviation, coefficient of variation, and the spread ratio. When sample data from a variable x is available, statistical measures from the data allow the analyst to estimate the parameter k that best fits the sample data; and also shows how to estimate the low-limit γ where $x \geq \gamma$, and any $x\alpha$ where $P(x \leq x\alpha) = \alpha$.

Chapter 11
Right Truncated Normal

11.1 Introduction

In 2001, Arvid Johnson and Nick Thomopoulos generated tables on the right-truncated normal distribution. The right-truncated normal (RTN) takes on a variety of shapes from normal to exponential-like. The distribution has one parameter k where the range includes all values of the standard normal that is less than a value of $z = k$. In this way, the distribution has the shape of the standard normal on the left and is truncated on the right. The variable is denoted as t and t is zero or negative throughout. With k specified, the following statistics are computed: the mean, standard deviation, coefficient-of-variation, and 0.01% and 0.99% points of t. The spread ratio of the RTN is also computed for each parameter value of k. A table is generated that lists all these statistics for values of k ranging from −3.0 to +3.0. When sample data is available, the analyst computes the following statistics: sample average, standard deviation, min and max. From these, the estimate of the spread ratio is computed, and this estimate is compared to the table values to locate the parameter k that has the closest value of θ. With k estimated for the sample data, the analyst identifies the RTN distribution that best fits the data. From here, the high-limit δ can be estimated, and also any α-percent-point on x that may be needed. The spread-ratio test sometimes indicates the sample data is best fit by the normal distribution, and on other situations by the RTN.

11.2 Fundamentals

The right-truncated normal distribution allows the analyst to apply a variety of shapes from normal to exponential-like. The distribution has one parameter k that determines its shape since only the portion of the standard normal distribution less

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_11

than z = k applies. Below gives a quick review on the standard normal, and this follows with a full description.

11.3 Standard Normal

The random variable z of the standard normal has limits from minus to plus infinity, and has mean zero and variance one. The designation is listed below:

$$z \sim N(0, 1)$$

Below is a listing of the probability density of z, the cumulative probability function, and the complementary probability. There is no closed form solution to the cumulative probability and as a result, various quantitative methods to compute F(z) have been developed. One such method is described in Chap. 8 and is used again in some of the computations of this chapter.

$$\begin{aligned} f(z) &= \left(1/\sqrt{2\pi}\right)e^{-z^2/2} && = \text{probability density of z} \\ F(k) &= P(z \le k) = \int_{-\infty}^{k} f(z)dz && = \text{cumulative probability of z = k} \\ H(k) &= P(z > k) = 1 - F(k) && = \text{complementary probability of z = k} \end{aligned}$$

Some related integrals on the standard normal are below:

$$\int_{-\infty}^{k} zf(z)dz = -f(k)$$

$$\int_{-\infty}^{k} z^2 f(z)dz = -kf(k) + F(k)$$

11.4 Right-Truncated Normal

The right-truncated normal random variable, t, is formed by the standard normal variable z and a parameter k that is a particular value of z. In this situation, the range on z is less or equal to k and thereby t is negative throughout as shown below:

$$\begin{aligned} t &= z - k \qquad z \le k \\ t &\le 0 \end{aligned}$$

The designation of the right-truncated-normal distribution is the following:

$$t \sim RTN(k)$$

The probability density of t, denoted as g(t); and the cumulative distribution of t, denoted as G(t), are obtained as below:

$$g(t) = f(z)/F(k)$$
$$G(t) = F(z)/F(k)$$

where f(z) is the probability density at z, F(z) is the cumulative probability of z, and F(k) is the cumulative probability when $z = k$.

The way to compute the mean and variance of t with parameter k is described here. First, $E(z < k)$, $E[(z < k)^2]$, $E(t)_k$ and $E(t^2)_k$ are derived as below:

$$E(z < k) = \int_{-\infty}^{k} (z-k) f(z) dz \qquad = -f(k) - kF(k)$$
$$E\left[(z < k)^2\right] = \int_{-\infty}^{k} (z-k)^2 f(z) dz \qquad = kf(k) + F(k)\left(1 + k^2\right)$$
$$E(t)_k = E(z < k)/F(k) \qquad = \text{expected value of t at k}$$
$$E(t^2)_k = E\left[(z < k)^2\right]/F(k) \qquad = \text{expected value of } t^2 \text{at k}$$

Next, the mean, variance and standard deviation of the variable t with parameter k are obtained in the following way:

$$\mu_t(k) = E(t)_k \qquad = \text{mean of t at k}$$
$$V(t)_k = E(t^2)_k - E(t)_k{}^2 = \text{variance of t at k}$$
$$\sigma_t(k) = \sqrt{V(t)_k} \qquad = \text{standard deviation of t at k}$$

Although the true range on z is from minus to plus infinity, almost all the probability falls between −3.0 and +3.0. For simplicity sake in computing the statistical measure on the RTN, the latter range of ±3 on z is used. With this adjusted range, the high and low limits on t become: [0,-(3 + k)]. Note, when $k = 3$, the range on t is (0 to −6), when $k = 2$, the range is (0 to −5), and so forth.

11.5 Cumulative Probability of k

The α-percent-point on t represents the value of t where the probability of t less or equal to tα is α, i.e.,

$$P(t \leq t\alpha) = \alpha$$

This is the same as the cumulative probability of tα as below:

$$G(t\alpha) = \alpha$$

The four steps that follow describe how to find tα, the α−percent-point value of t, for each combination of k and G(t).

1. For the given k, Table 8.1 lists F(k).
2. Since, $G(t) = \alpha = F(z)/F(k)$:

$$F(z) = \alpha F(k)$$

3. With F(z), Table 8.1 gives the corresponding value of z.
4. Finally,

$$t\alpha = (z - k)$$

When $\alpha = 0.01$ and 0.99, the following percent-points are obtained:

$t_{0.01}$ = 0.01-percent-point
$t_{.0.99}$ = 0.99-percent-point

11.6 Mean and Standard Deviation of t

Recall, the mean and standard deviation of t with parameter k are the following:

$\mu_{t(k)}$ = mean of variable t with location parameter k
$\sigma_{t(k)}$ = standard deviation of variable t with location parameter k

11.7 Spread Ratio of RTN

The spread ratio, denoted as θ, is unique for each k and is computed as shown below:

$$\theta = \left[\mu_{t(k)} - t_{0.01}\right] / \left[t_{0.99} - \mu_{t(k)}\right] = \text{spread} - \text{ratio}$$

11.8 Table Values

Table 11.1 is a list of key measures on the RTN where k ranges as: [−3.0, (0.1), +3.0]. For each entry of k, the measures listed are the following: the α-percent-points $t_{0.01}$, $t_{0.99}$, the mean, standard deviation, coefficient-of-variation and the spread ratio.

Table 11.1 Right-truncated normal by location parameter k, percent-points $t_{0.01}$, $t_{0.99}$, mean, $\mu_t(k)$, standard deviation, $\sigma_t(k)$, coefficient-of-variation, cov(k), spread-ratio, θ

k	$t_{0.01}$	$t_{0.99}$	$\mu_{t(k)}$	$\sigma_{t(k)}$	cov(k)	θ
−3.0	−1.494	−0.006	−0.283	0.264	−0.93	4.36
−2.9	−1.397	−0.005	−0.291	0.270	−0.93	3.88
−2.8	−1.365	−0.005	−0.298	0.277	−0.93	3.64
−2.7	−1.359	−0.005	−0.306	0.284	−0.93	3.50
−2.6	−1.366	−0.005	−0.314	0.291	−0.93	3.40
−2.5	−1.381	−0.005	−0.323	0.298	−0.92	3.32
−2.4	−1.401	−0.005	−0.332	0.306	−0.92	3.27
−2.3	−1.426	−0.005	−0.341	0.313	−0.92	3.22
−2.2	−1.454	−0.005	−0.351	0.321	−0.91	3.18
−2.1	−1.485	−0.005	−0.362	0.330	−0.91	3.14
−2.0	−1.518	−0.005	−0.373	0.338	−0.91	3.11
−1.9	−1.553	−0.005	−0.385	0.347	−0.90	3.07
−1.8	−1.590	−0.005	−0.397	0.356	−0.90	3.04
−1.7	−1.629	−0.005	−0.410	0.366	−0.89	3.01
−1.6	−1.670	−0.005	−0.424	0.376	−0.89	2.98
−1.5	−1.713	−0.006	−0.439	0.387	−0.88	2.94
−1.4	−1.757	−0.006	−0.454	0.398	−0.88	2.91
−1.3	−1.803	−0.006	−0.470	0.409	−0.87	2.87
−1.2	−1.851	−0.006	−0.488	0.421	−0.86	2.83
−1.1	−1.901	−0.006	−0.506	0.433	−0.86	2.79
−1.0	−1.953	−0.007	−0.525	0.446	−0.85	2.75
−0.9	−2.006	−0.007	−0.546	0.460	−0.84	2.71
−0.8	−2.062	−0.007	−0.567	0.473	−0.83	2.67
−0.7	−2.119	−0.008	−0.590	0.488	−0.83	2.62
−0.6	−2.179	−0.008	−0.615	0.503	−0.82	2.58
−0.5	−2.240	−0.008	−0.641	0.518	−0.81	2.53
−0.4	−2.303	−0.009	−0.669	0.534	−0.80	2.48
−0.3	−2.369	−0.010	−0.698	0.551	−0.79	2.43
−0.2	−2.436	−0.010	−0.729	0.568	−0.78	2.37
−0.1	−2.506	−0.011	−0.763	0.585	−0.77	2.32
0.0	−2.577	−0.013	−0.798	0.603	−0.76	2.27
0.1	−2.650	−0.014	−0.835	0.621	−0.74	2.21
0.2	−2.726	−0.015	−0.875	0.640	−0.73	2.15
0.3	−2.803	−0.017	−0.917	0.659	−0.72	2.09
0.4	−2.882	−0.018	−0.962	0.678	−0.70	2.03
0.5	−2.963	−0.020	−1.009	0.697	−0.69	1.97
0.6	−3.045	−0.022	−1.059	0.717	−0.68	1.91
0.7	−3.129	−0.024	−1.112	0.736	−0.66	1.86
0.8	−3.215	−0.027	−1.168	0.756	−0.65	1.80
0.9	−3.303	−0.030	−1.226	0.775	−0.63	1.74
1.0	−3.391	−0.034	−1.288	0.794	−0.62	1.68
1.1	−3.481	−0.039	−1.352	0.812	−0.60	1.62

(continued)

Table 11.1 (continued)

k	$t_{0.01}$	$t_{0.99}$	$\mu_{t(k)}$	$\sigma_{t(k)}$	cov(k)	θ
1.2	−3.573	−0.044	−1.419	0.830	−0.58	1.57
1.3	−3.665	−0.051	−1.490	0.847	−0.57	1.51
1.4	−3.759	−0.059	−1.563	0.863	−0.55	1.46
1.5	−3.853	−0.068	−1.639	0.879	−0.54	1.41
1.6	−3.948	−0.080	−1.717	0.894	−0.52	1.36
1.7	−4.044	−0.094	−1.798	0.907	−0.50	1.32
1.8	−4.141	−0.110	−1.882	0.920	−0.49	1.28
1.9	−4.238	−0.131	−1.968	0.931	−0.47	1.24
2.0	−4.336	−0.155	−2.055	0.942	−0.46	1.20
2.1	−4.434	−0.184	−2.145	0.951	−0.44	1.17
2.2	−4.532	−0.218	−2.236	0.959	−0.43	1.14
2.3	−4.631	−0.259	−2.329	0.966	−0.41	1.11
2.4	−4.730	−0.305	−2.423	0.972	−0.40	1.09
2.5	−4.829	−0.359	−2.518	0.978	−0.39	1.07
2.6	−4.929	−0.419	−2.614	0.982	−0.38	1.06
2.7	−5.028	−0.486	−2.710	0.986	−0.36	1.04
2.8	−5.128	−0.559	−2.808	0.989	−0.35	1.03
2.9	−5.228	−0.638	−2.906	0.991	−0.34	1.02
3.0	−5.328	−0.721	−3.004	0.993	−0.33	1.02

As a rule-of-thumb, when the sample data estimate of $\theta$is 1.30 or less, the RTN is not significantly different from the normal, and as such, the normal distribution applies; else, the RTN is deemed the better fit.

Example 11.1 At k = 1.0, Table 11.1 lists $t_{0.01} = -3.391$ and $t_{0.99} = -0.034$. Below shows how these results are derived:

At k = 1.00, Table 8.1 shows F(1.0) = 0.841.
Using $\alpha = 0.01$, $F(z) = 0.01 \times 0.841 = 0.0084$
Using Table 8.1, z = −2.4
Hence, $t_{0.01} = -2.40 - (1.00) = -3.40$

At k = 1.00, Table 8.1 shows F(1.0) = 0.841.
Using $\alpha = 0.99$, $F(z) = 0.99 \times 0.841 = 0.0.833$
Using Table 8.1, z = 1.07
Hence, $t_{099} = 1.07 - (1.00) = -0.07$

Any difference from the computed value and the table value is due to rounding.

Example 11.2 The Table 11.1 also shows $\mu_t(k) = -1.288$, $\sigma_t(k) = 0.794$, cov = −0.62 and $\theta = 1.68$ when k = 1.00. Below shows how these results are derived.

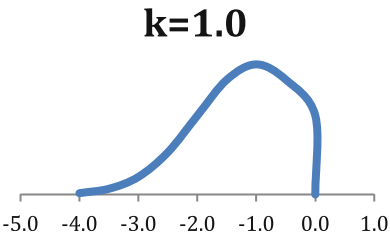

Fig. 11.1 RTN plot when k = 1.00

$$E(x < k) = -0.242 - 1 \times 0.841 = -1.083$$
$$E\left[(x < k)^2\right] = 1 \times 0.242 + 0.0842 \times 2 = 1.924$$
$$E(t)_k = -1.083/0.841 = -1.288$$
$$\mu_t(k) = -1.288$$
$$E(t^2)_k = 2.288/0.841 = 2.288$$
$$V(t)_k = \left[2.288 - (-1.288)^2\right] = 0.629$$
$$\sigma_t(k) = \sqrt{0.629} = 0.793$$
$$\text{cov} = 0.793/-1.288 = -0.676$$
$$\theta = [-1.288 - (-3.391)]/[-0.034 - (-1.288)] = 1.67$$

Any difference in the computed and table results is due to rounding. Figure 11.1 depicts the probability density on t.

11.9 Sample Data

When sample data ($x_1, \ldots, x_n$) is available, the analyst can obtain the following statistical measures on the variable x:

$\bar{x}$ = average
s = standard deviation
$x(1) = \min(x_1, \ldots, x_n)$
$x(n) = \max(x_1, \ldots, x_n)$

After gathering the statistical measures, the estimate of the spread ratio for the RTN, denoted as $\widehat{\theta}$, is computed in the following way:

$$\widehat{\theta} = [\bar{x} - x(1)/[x(n) - \bar{x}]$$

Recall when the RTN is in effect, $\widehat{\theta}$ is greater than 1.0. With the spread ratio now estimated, the analyst searches Table 11.1 to find the parameter k with the closest θ, and then obtain the following measures on the RTN:

k = RTN parameter
$\mu_t(k)$ = mean of t at k
$\sigma_t(k)$ = standard deviation of t at k
$t_{0.99}$ = 0.99 − percent − point of t at k

11.10 Parameter Estimates When Sample Data

When the RTN distribution is identified, the high-limit on x, denoted as δ, represents the max value of x, i.e., $x \leq \delta$. Often this high-limit is known a-priori, but sometimes it is not and needs to be estimated. A way to estimate δ is described here.

11.11 Estimate δ When RTN

When the high limit is not known and an estimate is needed, a way to approximate the value of δ is described. Gather the data listed below where some come from Table 11.1 and some from the sample data:

$t_{.0.99}$ = 0.99-percent-point
$\mu_{t(k)}$ = mean
$\sigma_{t(k)}$ = standard deviation
$\bar{x}$ = sample average
s = sample standard deviation
$x(n) = \max(x_1, \ldots, x_n)$

Note the approximate relation between the sample data and the table measures as shown below:

$$(\delta - \bar{x})/s = \left[t_{0.99} - \mu_{t(k)}\right]/\sigma_{t(k)}$$

Solving for δ yields:

$$\delta' = \bar{x} + s\left[t_{0.99} - \mu_{t(k)}\right]/\sigma_{t(k)}$$

$$\delta = \max\left[\delta', x(n)\right]$$

11.12 Estimate the α-Percent-Point of x

Using $t\alpha$ = α-percent-point of variable t, and the same data gathered for the high-limit on x, the following relation is now formed to find the α-percent-point on x:

$$t\alpha = \mu_t(k) + \sigma_t(k)\left[\left(x\alpha - \bar{x}\right)/s\right]$$

Solving for $x\alpha$,

$$x\alpha = \bar{x} + s[(t\alpha - \mu_t(k))/\sigma_t(k)]$$

where,
$x\alpha$ = α-percent-point of variable x.

Example 11.3 A researcher is conducting a study and obtains the following statistical measures from sample data:

$\bar{x} = 75$
$s = 15$
$x(1) = 40$
$x(n) = 90$

Example 11.4 Using the statistical measure from Example 11.3, the researcher is seeking to apply the RTN to the data and needs to estimate the parameters of the distribution. The spread ratio is computed as shown below:

$$\widehat{\theta} = (75 - 40)/(90 - 75) = 2.33$$

Searching Table 11.1 to find the closest spread ratio yields $\theta = 2.33$. The associated statistics are the following:

$k = -0.10$
$t_{\cdot 0.01} = -2.506$
$t_{\cdot 0.99} = -0.011$
$\mu_{t(k)} = -0.763$
$\sigma_{t(k)} = 0.585$

The estimate of the high-limit δ is below:

$$\widehat{\delta} = 75 + 15[-0.011 - (-0.763)]/0.585 = 94.3$$

In the quest to find the $x_{0.90} = 0.90\%$ point on x, it is first necessary to derive $t_{0.90} = 0.90\%$ point on t. Below shows how these two estimates are obtained.

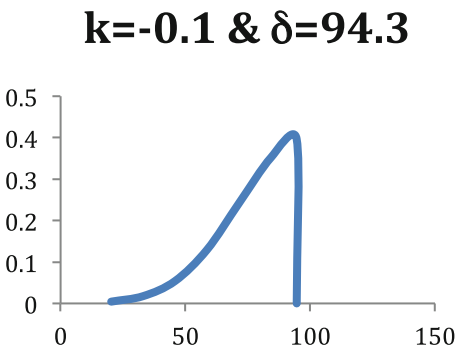

Fig. 11.2 RTN plot when k = −0.1 and δ = 94.3

Given $\alpha = 0.90$ and $k = -0.10$, below shows how to find $t_{0.90}$:
At $k = -0.10$, Table 8.1 shows $F(-0.10) = 0.460$

$$F(z) = 0.90 \times 0.460 = 0.414$$

At $F(z) = 0.414$, Table 8.1 yields $z = -0.22$.
Finally, $t_{0.90} = (-0.22 - (-0.10)) = -0.12$.
Given $t_{0.90} = -0.12$, $x_{0.90}$ is computed as below:

$$x_{0.90} = 70 + 15[(-0.12 - (-0.763))/0.585] = 91.5$$

Figure 11.2 depicts the probability density when $k = -0.10$ and $\delta = 94.3$.

11.13 Summary

The RTN is one of the few distributions that skews to the left. The high-limit is denoted as $\delta = 0$, and the variable, t includes negative values less than δ. The parameter of the distribution is k, where k ranges from −3.0 to +3.0. Table values are developed for selected values of k. With each k, the following statistical measures are listed: the 0.01% and 0.99% points on t, the mean, standard deviation and coefficient-of-variation of t, and the spread ratio. When sample data is available, the analyst computes the average, standard deviation, min and max values on x. With these measures, the spread ratio of the data is estimated. The estimated spread ratio is used to find the RTN distribution that best fits the data, and thereby allows estimating the parameter k. With k known, the analyst can estimate the high-limit δ, and also any value of $x\alpha$ = α-percent-point on x.

Chapter 12
Triangular

12.1 Introduction

When an analyst is in need of a continuous distribution in a project and has little information on the shape, the often practice is the employ the triangular distribution. The analyst seeks estimates on the min, max and most-likely values of the variable and from these forms the distribution. The probability density goes up linearly from the min to the mode, and linearly down to the max value. Another related distribution is the standard triangular that ranges from zero to one. This latter distribution in easier to manage in a mathematical sense, and is used to define the probability distribution, cumulative probability, mean, variance and standard deviation. The conversion of all statistics and probabilities from the standard triangular to the triangular distribution is readily obtained.

12.2 Fundamentals

The triangular distribution has two versions: the standard triangular with random variable y, and the triangular with random variable x. The triangular is the distribution the analyst employs in the research application, while the standard triangular is easier to apply mathematically. Below describes both distributions and shows how one is converted from the other.

12.3 Standard Triangular

The random variable of the standard triangular, y, has a range from zero to one, and the most-likely value is denoted as $\tilde{y}$, as noted below:

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_12

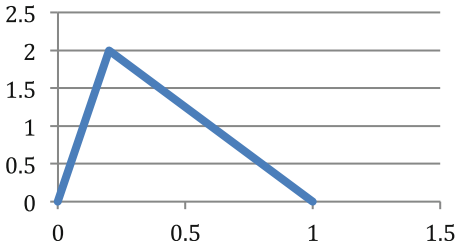

Fig. 12.1 Standard triangular T(0, 1, $\tilde{y}$)

$$y \qquad 0 \le y \le 1$$

$0 =$ min of y
$1 =$ max of y
$\tilde{y} =$ mode of y

The designation of y is the following:

$$y \sim T(0, 1, \tilde{y})$$

Figure 12.1 depicts the standard triangular density.

Below gives the probability of y, which linearly ramps up from 0 to $\tilde{y}$, and then down from $\tilde{y}$ to 1.

$$\begin{aligned} f(y) &= 2y/\tilde{y} && y = (0 \text{ to } \tilde{y}) \\ &= 2(1-y)/(1-\tilde{y}) && y = (\tilde{y} \text{ to } 1) \end{aligned}$$

The cumulative probability on y is the following:

$$\begin{aligned} F(y) &= y^2/\tilde{y} && y = (0 \text{ to } \tilde{y}) \\ &= 1-(1-y)^2/(1-\tilde{y}) && y = (\tilde{y} \text{ to } 1) \end{aligned}$$

Below lists the mean, variance and standard deviation on y.

$$\mu_y = (1+\tilde{y})/3$$

$$\sigma_y{}^2 = (1+\tilde{y}^2-\tilde{y})/18$$

$$\sigma_y = \sqrt{\sigma_{\tilde{y}}^2}$$

12.4 Triangular

The counterpart variable to y is denoted as x and comes from the triangular distribution. The range on x is from a to b and the most-likely value is denoted as $\tilde{x}$ as listed below:

$$\begin{aligned} &\mathrm{a} = \text{min of x} \\ &\mathrm{b} = \text{max of x} \\ &\tilde{\mathrm{x}} = \text{mode of x} \end{aligned}$$

Hence, the parameters are: a, b, $\tilde{x}$, and the designation on x is the following:

$$\mathrm{x} \sim \mathrm{T}\left(\mathrm{a}, \mathrm{b}, \tilde{x}\right)$$

Note the conversion from variable y to x, and also from x to y:

$$\begin{aligned} \mathrm{x} &= \mathrm{a} + \mathrm{y}(\mathrm{b} - \mathrm{a}) \\ \mathrm{y} &= (\mathrm{x} - \mathrm{a})/(\mathrm{b} - \mathrm{a}) \end{aligned}$$

Below shows how the mean, variance and standard deviation of x are converted from their counterparts on y, and also how they are obtained from their parameters, a, b and $\tilde{x}$.

$$\begin{aligned} \mu_{\mathrm{x}} &= \mathrm{a} + \mu_{\mathrm{y}}(\,\mathrm{b} - \mathrm{a}) \\ &= \left(\mathrm{a} + \mathrm{b} + \tilde{x}\right)/3 \\ \sigma_x^2 &= \sigma_y^2(\mathrm{b} - \mathrm{a})^2 \\ &= \left(\mathrm{a}^2 + \mathrm{b}^2 + \tilde{x}^2 - \mathrm{ab} - \mathrm{a}\tilde{x} - \mathrm{b}\tilde{x}\right)/18 \\ \sigma_{\mathrm{y}} &= \sqrt{\sigma_x^2} \end{aligned}$$

Example 12.1 Suppose x is triangular with parameters: x ∼ (10, 90, 70). The mean, variance and standard deviation of x are computed below.

$$\begin{aligned} \mu_{\mathrm{x}} &= (10 + 90 + 70)/3 = 56.67 \\ \sigma_{\mathrm{x}}{}^2 &= \left(\,10^2 + 90^2 + 70^2 - 10 \times 90 - 10 \times 70 - 90 \times 70\right)/18 = 289 \\ \sigma_{\mathrm{x}} &= \sqrt{289} = 17.0 \end{aligned}$$

Below shows how to convert to the standard triangular.

$$\begin{aligned} \tilde{y} &= (70 - 10)/(90 - 10) = 0.75 \\ \mathrm{y} &\sim \mathrm{T}(0, 1, 0.75) \end{aligned}$$

The probability density and cumulative probability of y are the following:

$$\begin{aligned} \mathrm{f}(\mathrm{y}) &= 2\mathrm{y}/0.75 && \mathrm{y} = (0 \text{ to } 0.75) \\ &= 2(1 - \mathrm{y})/0.25 && \mathrm{y} = (0.75 \text{ to} 1.00) \end{aligned}$$

$$\begin{aligned} F(y) &= y^2/0.75 && y = (0 \text{ to } 0.75) \\ &= 1-(1-y)^2/0.25 && y = (0.75 \text{ to } 1.00) \end{aligned}$$

The mean, variance and standard deviation of y are obtained below:

$$\mu_y = (1+0.75)/3 = 0.583$$

$$\sigma_y{}^2 = \left(1+0.75^2-0.75\right)/18 = 0.045$$

$$\sigma_y = \sqrt{0.045} = 0.212$$

The mean, variance, standard deviation and mode of x are obtained as below:

$$\mu_x = 10 + 0583(90-10) = 56.67$$

$$\sigma_x{}^2 = 0.045(90-10)^2 = 289$$

$$\sigma_x = \sqrt{289} = 17$$

$$\tilde{x} = 10 + 0.75(90-10) = 70$$

12.5 Table Values on y

Table 12.1 lists values of F(y) for selective values of y and $\tilde{y}$.

Example 12.2 When $\tilde{y} = 0.3$, Table 12.1 lists F(0.2) = 0.13, F(0.8) = 0.94, $\mu_y = 0.43$ and $\sigma_y = 0.21$. Below shows how these values are calculated:

Table 12.1 Cumulative Probability, F(y), of the standard triangular distribution for selective values of y, and selective values of the mode, $\tilde{y}$

$y/\tilde{y}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.19	0.10	0.05	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01
0.2	0.36	0.29	0.20	0.13	0.10	0.08	0.07	0.06	0.05	0.04	0.04
0.3	0.51	0.46	0.39	0.30	0.22	0.18	0.15	0.13	0.11	0.10	0.09
0.4	0.64	0.60	0.55	0.49	0.40	0.32	0.27	0.23	0.20	0.18	0.16
0.5	0.75	0.72	0.69	0.64	0.58	0.50	0.42	0.36	0.31	0.28	0.25
0.6	0.84	0.82	0.80	0.77	0.73	0.68	0.60	0.51	0.45	0.40	0.36
0.7	0.91	0.90	0.89	0.87	0.85	0.82	0.77	0.70	0.61	0.54	0.49
0.8	0.96	0.96	0.95	0.94	0.93	0.92	0.90	0.87	0.80	0.71	0.64
0.9	0.99	0.99	0.99	0.99	0.98	0.98	0.97	0.97	0.95	0.90	0.81
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
μ	0.33	0.37	0.40	0.43	0.47	0.50	0.53	0.57	0.60	0.63	0.67
σ	0.24	0.22	0.22	0.21	0.21	0.20	0.21	0.21	0.22	0.22	0.24

Also listed is the mean and standard deviation of y for each mode

$$F(0.2) = 0.2^2/0.3 = 0.133$$

$$F(0.8) = 1 - (1 - 0.8)^2/0.7 = 0.942$$

$$\mu_y = (1 + 0.3)/3 = 0.433$$

$$\sigma_y{}^2 = \left(1 + 0.3^2 - 0.3\right)/18 = 0.044$$

$$\sigma_y = 0.044 = 0.209$$

12.6 Deriving xα = α-Percent-Point on x

To find xα = α-percent point on x, the following three steps are taken.

1. The mode on variable y, $\tilde{y}$, is obtained from the parameters on x as shown below:

$$\tilde{y} = \left(\tilde{x} - a\right)/(b - a)$$

2. The α-percent-point on the variable y, yα, is computed in the following way depending on the relation between α and $\tilde{y}$.

$$\text{If } \alpha \leq \tilde{y} \qquad \alpha = y\alpha^2/\tilde{y}$$
$$y\alpha = \sqrt{\alpha\tilde{y}}$$

$$\text{If } \alpha > \tilde{y} \qquad \alpha = 1 - (1 - y\alpha)^2/\left(1 - \tilde{y}\right)\frac{n!}{r!(n-r)!}$$
$$y\alpha = 1 - \sqrt{(1 - \alpha)\left(1 - \tilde{y}\right)}$$

3. The α-percent-point on the variable x is obtained as shown below:

$$x\alpha = a + y\alpha(b - a)$$

Example 12.2 Consider a triangular distribution where x ~ (50, 100, 70), where the plot of the probability density in Fig. 12.2.

Note, the mode of the standard triangular is the following:

$$\tilde{y} = (70 - 50)/(100 - 50) = 0.40$$

and thereby,

$$y \sim T(0, 1, 0.40)$$

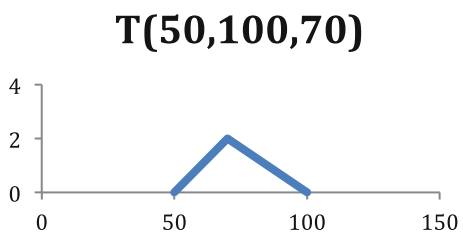

Fig. 12.2 Triangular plot of (50,100,70).

Assume the analyst is seeking $x_{0.10}$ and $x_{0.90}$. These are obtained below:

Since $0.10 \leq 0.40$: $y_{0,10} = \sqrt{0.10 \times 0.40} = 0.20$
$x_{0.10} = 50 + 0.20(100 - 50) = 60.00$

Also, since, $0.90 > 0.40$: $y_{0.90} = 1 - \sqrt{(1 - 0.90)(1 - 0.40)} = 0.755$
$x_{0.90} = 50 + 0.755(100 - 50) = 87.75$

12.7 Parameter Estimates When No Data

This distribution is primarily in use when the analyst has no data. Estimates on the min, max and most –likely values are given on a variable x and these yield the following:

$a = \min$
$b = \max$
$\tilde{x} = \text{most} - \text{likely}$

and thereby,

$x \sim T(a, b, \tilde{x})$

12.8 Summary

The standard triangular distribution has variable y with a range from zero to one. The distribution ramps up linearly from zero to the mode, and down from the mode to one. The probability density, cumulative probability, mean and standard deviation are listed for the standard triangular. The triangular distribution with variable x has a range of a to b with the mode in-between. Easy conversion is noted from variable y to x, and also from x to y. The distribution is used when the analyst needs to apply a distribution in a project and only has estimates of the min, max and most-likely values of x. When needed, the α-percent-point on x is readily derived from first obtaining the α-percent-point on y.

Chapter 13
Discrete Uniform

13.1 Introduction

The discrete uniform distribution occurs when the random variable x can take on any integer value from a to b with equal probability. A church raffle of 1000 numbers (1–1000) is such a system where the winning number, x, is equally likely to be any of the numbers in the admissible range. Often the parameters (a,b) are known apriori to an analyst who is seeking to apply the distribution. Other times the parameters are not known and sample data is provided to estimate the parameters. Still on other occasions when the parameters are not known and no sample data is available, the analyst seeks advise from an expert who provides some approximations on the range, and this information is used to estimate the parameter values.

13.2 Fundamentals

In applying the discrete uniform distribution, the variable x can equally be any integer value from a to b, where a is the min and b is the max. In this way (a, b) are the parameters of x where the range is:

$$x = a, a + 1, \ldots, b$$

Because all admissible values of x are equally likely to occur, the probability of any value x is the following:

$$P(x) = 1/(b - a + 1) \qquad x = a, a + 1, \ldots, b$$

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_13

The cumulative probability of x or smaller becomes:

$$F(x) = P(w \leq x) = (x - a + 1)/(b - a + 1) \qquad x = a, a + 1, \ldots, b$$

The mean and variance of x are obtained as below:

$$\mu = (a + b)/2$$
$$\sigma^2 = \left[(b - a + 1)^2 - 1\right]/12$$

Example 13.1 At apple harvest time, the apples are picked and placed in a basket until full. The number of apples in a basket varies depending on the size of the apples, and the range is from 50 to 60. Letting x represent the number of apples in a basket as the random variable, the range is from 50 to 60, and the probability of any value of x is the following:

$$P(x) = 1/(60 - 50 + 1) = 0.091 \qquad \text{for } x = 50 \text{ to } 60$$

The cumulative probability of 53 or less apples in a basket is computed as follows:

$$P(x \leq 53) = (53 - 50 + 1)/(60 - 50 + 1) = 0.364$$

The mean, variance and standard deviation on x are obtained below:

$$\mu = (50 + 60)/2 = 55$$
$$\sigma^2 = \left[(60 - 50 - 1)^2 - 1\right]/12 = 10$$
$$\sigma = 3.16$$

13.3 Lexis Ratio

When $a = 0$, the lexis ratio becomes:

$$\tau = \sigma^2/\mu$$
$$= \left[(b + 1)^2 - 1\right]/6b$$

Note when $b \geq 4$, $\tau \geq 1$.

13.4 Sample Data

When n sample observations, $(x_1, \ldots, x_n)$, are available, the following statistics are obtained:

$\bar{x}$ = sample average
s^2 = sample variance
$x(1) = \min(x_1, \ldots, x_n)$
$x(n) = \max(x_1, \ldots, x_n)$

13.5 Parameter Estimates When Sample Data

The above stats from sample data are used to estimate the parameters as shown below. The MLE method yields:

$\hat{a} = x(1)$
$\hat{b} = x(n)$

Another way to estimate the parameters is by the method-of-moments as shown below:

$\hat{a}$ = floor integer of $\left(\bar{x} + 0.5 - 0.5\sqrt{12s^2 + 1}\right)$
$\hat{b}$ = ceiling integer of $\left(\bar{x} - 0.5 + 0.5\sqrt{12s^2 + 1}\right)$

Example 13.2 Assume discrete integer data and an experiment yields the ten sample entries (13, 11, 4, 8, 16, 14, 8, 9, 7, 5) and estimates of the parameters (a, b) are needed. The following statistics are computed from the data:

$\bar{x} = 9.5$
$s^2 = 15.39$
$x(1) = 4$
$x(n) = 16$

Using the MLE method, the parameter estimates are:

$\hat{a} = 4$
$\hat{b} = 16$

The method-of-moment estimate on the parameters is as follows:

$$\hat{a} = \text{floor}\left(9.5 + 0.5 - 0.5\sqrt{12 \times 15.39 + 1}\right) = \text{floor}\,(3.19) = 3$$
$$\hat{b} = \text{ceiling}\left(9.5 - 0.5 + 0.5\sqrt{12 \times 15.39 + 1}\right) = \text{ceiling}(15.81) = 16$$

Example 13.3 The military intelligence command is concerned on how many units of a particular devastating weapon the enemy has produced. In their skirmishes, n units of this weapon have been retrieved and have serial numbers engraved. The captured numbers are listed as: (x_1, …, x_n), and with the numbers, the usual statistics are obtained as: $\bar{x}$, s, x(1) and x(n). With this data, and the methods described above the parameter estimates ($\hat{a}$,$\hat{b}$), are computed. Assuming the serial numbers are listed in integer order, the number of enemy weapons of this type is estimated as:

$$\mathrm{N} = \left(\hat{b} - \hat{a} + 1\right)$$

13.6 Parameter Estimates When No Data

When no sample data is available to estimate the parameters for the discrete uniform distribution, the analyst may seek an expert's advise on the distribution. One-way is when the expert gives the following type of estimates on the spread of the distribution:

$$x_1 = \alpha_1 - \text{percent point on x}$$
$$x_2 = \alpha_2 - \text{percent point on x}$$

where:

$$\alpha_1 = (x_1 - a + 1)/(b - a + 1)$$
$$\alpha_2 = (x_2 - a + 1)/(b - a + 1)$$

With this data and relations, and with some algebra, the estimates on the parameters are obtained in the following way:

$$\grave{a} = [\alpha_2(x_1 + 1) - \alpha_1(x_2 + 1)/[(\alpha_2 - \alpha_1)]\ \grave{b} = \left[x_1 + \grave{a}(\alpha_1 - 1) - (\alpha_1 - 1)\right]/\alpha_1$$
$$\hat{a} = \text{floor}\left(\grave{a}\right)$$
$$\hat{b} = \text{ceiling}\left(\grave{b}\right)$$

Example 13.4 An analyst has the need to use the discrete uniform distribution but has no sample data to estimate the parameter values. He asks an expert type person who gives the following approximations on the distribution:

"The mid 80 percent range falls from 5 to 20."which translates as follows:

$$P(5 \leq x \leq 20) = 0.80$$

and thereby:

$$x_1 = 5 \text{ and } \alpha_1 = 0.10$$
$$x_2 = 20 \text{ and } \alpha_2 = 0.90$$

The parameter values are estimated below:

$$\grave{a} = [0.9(5+1) - 0.1(20+1)]/[0.9 - 0.1] = 4.125$$
$$\grave{b} = [5 + 4.125(0.1 - 1) - (0.1 - 1)]/0.1 = 21.875$$
$$\hat{a} = \text{floor}(4.125) = 4$$
$$\hat{b} = \text{ceiling}(21.875) = 22$$

13.7 Summary

The discrete uniform distribution has two parameters (a, b) where the random variable includes all integers from a to b with equal probability. The chapter shows how to compute the probability of x, the cumulative probability of x, the mean and standard deviation. When the parameter values are not known, and sample data is available, the sample data is used to estimate the parameter values. When no sample data is available, and an expert type person gives some approximations on the spread of the variable, this information can be used to estimate the parameter values.

Chapter 14
Binomial

14.1 Introduction

Some historians give the first use of the binomial distribution to Jakob Bernoulli who was a prominent Swiss mathematician in the 1600s. The binomial distribution applies when a number of trials of an experiment is run and only two outcomes are noted on each trial, success and failure, and the probability of the outcomes remain the same over all of the trials. This happens, for example, when a roll of two dice is run five times and the success per run is when the number of dots is two, say. The probability of a success per run remains constant, the number of trials in five, and the probability of a success per trial is $p = 1/36$. The random variable, denoted as x, for this scenario is the number of successes in the five trials, and this could be: $x = 0, 1, 2, 3, 4, 5$. The chapter lists the probability distribution of the random variable x. The mean, variance, standard deviation and mode of x is also given. When p is not known, it can be estimated using sample data, and even when no sample data is provided, an estimate of p can be obtained.

14.2 Fundamentals

The parameters of the binomial distribution are the following:

$n = \text{number of trials}$
$p = \text{probability of a success per trial}$

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_14

When n trials with probability p of a success per trial, the probability of x successes becomes:

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, \ldots, n$$

The cumulative probability of x or smaller is denoted as F(x), and when. $x = x_o$,

$$F(x_o) = P(x \leq x_o).$$

The mean, variance, and standard deviation of x are computed as below:

$$\mu = np$$
$$\sigma^2 = np(1-p)$$
$$\sigma = \sqrt{np(1-p)}$$

The mode of x, $\widetilde{\mu}$, is derived in the following way:

If $p(n+1)$ is not an integer: $\widetilde{\mu} = \text{floor}[p(n+1)]$
If $p(n+1)$ is an integer: $\widetilde{\mu} = p(n+1) - 1$ and $p(n+1)$

14.3 Lexis Ratio

The lexis ratio for the binomial is below:

$$\tau = \sigma^2/\mu$$
$$= np(1-p)/np$$
$$= (1-p)$$

Note where τ is always less than one.

Example 14.1 An experiment with $n = 10$ trials follows the binomial distribution with $p = 0.3$. The mode on x (number of successes) is obtained as below:

Since $p(n+1) = 0.3(11) = 3.3$ is not an integer:

$$\widetilde{\mu} = \text{floor}[p(n+1)] = \text{floor}[0.3(11)] = \text{floor}[3.3] = 3$$

14.4 Normal Approximation

The random variable x is approximated by the normal distribution when n is large, and,

$p \leq 0.5$ with $np > 5$,

or when,

$p > 0.5$ with $n(1-p) > 5$,

The mean, variance and standard deviation of x are below:

$$\begin{aligned} \mu &= np && = \text{mean of x} \\ \sigma^2 &= np(1-p) && = \text{variance of x} \\ \sigma &= \sqrt{np(1-p)} && = \text{standard deviation of x} \end{aligned}$$

and thereby, the normal approximation becomes:

$$x \sim N\left(\mu, \sigma^2\right)$$

The probability of $x = x_o$ or below is approximated by converting the integer of x to the continuous standard normal z variable as below:

$$z_o = (x_o + 0.5 - np)/\sqrt{np(1-p)}$$

and thereby:

$$\begin{aligned} P(x \leq x_o) &\approx P(x \leq x_o + .5) \\ &= F(z_o) \end{aligned}$$

where F(z) is the cumulative probability from Table 8.1.

The probability that $x = x_o$ is obtained as below by converting the integer value of x to two continuous standard normal variables as below:

$$\begin{aligned} z_L &= (x_o + .5 - np)/\sqrt{np(1-p)} \\ z_H &= (x_o - .5 - np)/\sqrt{np(1-p)} \end{aligned}$$

Hence,

$$\begin{aligned} P(x = x_o) &\approx P(x_o - .5 \leq x \leq x_o + .5) \\ &= F(z_H) - F(z_L) \end{aligned}$$

14.5 Poisson Approximation

In the event the normal approximation does not apply, the Poisson approximation may be used when n is large and p is small. The Poisson parameter, representing the mean of x, becomes:

$$\theta = \text{np}$$

Using the Poisson distribution, the probability of x becomes:

$$\text{P(x)} = e^{-\theta}\theta^{x}/x! \qquad x = 0, 1, \ldots..$$

Example 14.2 A lot of n = 10 units are to be produced on a facility that yields 10% defective units. With x the number of defectives, the mean and standard deviation of x are below:

$$\begin{aligned}\mu &= \text{np}\\ &= 10 \times 0.1\\ &= 1.0\\ \sigma &= \sqrt{np(1-p)}\\ &= \sqrt{10(.1)(.9)}\\ &= 0.95\end{aligned}$$

The probability distribution on x is:

$$\text{P(x)} = \binom{10}{x} 0.1^{\text{x}} 0.9^{10-\text{x}} \qquad \text{x} = 0, 1, \ldots., 10$$

The probability of two or more defectives in the lot is computed below:

$$\begin{aligned}\text{P(x} \geq 2) &= 1 - \text{P(x} \leq 1)\\ &= 1 - [\text{P(0)} + \text{P(1)}]\end{aligned}$$

where

$$\text{P(0)} = \binom{10}{0} 0.1^{0} 0.9^{10} = 0.349$$

$$\text{P(1)} = \binom{10}{1} 0.1^{1} 0.9^{9} = 0.387$$

and thereby,

$$\begin{aligned} P(x \geq 2) &= 1 - [0.349 + 0.387] \\ &= 0.264 \end{aligned}$$

Example 14.3 A laboratory chemist conducts an experiment $n = 8$ times and the number of successes is $x = 6$. The estimate on the probability of a success becomes:

$$\hat{p} = x/n = 6/8 = 0.75$$

The standard deviation on the estimate of p, $s_{p.}$, is the following:

$$\begin{aligned} s_p &= \left(\hat{p}\left(1 - \hat{p}\right)/n\right)^{0.5} \\ &= \sqrt{0.75(0.25)/8} \\ &= 0.153 \end{aligned}$$

Example 14.4 A production batch of $n = 12$ units is run each day for five consecutive days with x_i defectives per batch and the results of the five batches are the following: (0, 1, 0, 2, 1). The estimate on the probability of a defective unit is computed below:

$$\begin{aligned} \hat{p} &= \sum x_i/(nm) \\ &= 4/(5 \times 12) = 0.067 \end{aligned}$$

and the standard deviation is:

$$\begin{aligned} s_p &= \sqrt{.067(.933)/60} \\ &= 0.032 \end{aligned}$$

Example 14.5 A supplier delivers lots of 200 units each week to a plant and the history shows the portion of faulty units is 0.05. The plant needs 185 non-faulty units each week and wants to determine the probability of this event. The probability of 185 or more non-faulty units is the same as 15 or less faulty units in the lot. Since

$$p = 0.05 < 0.50$$

and

$$np = 200 \times 0.05 = 10 > 5$$

the normal approximation can be used to approximate the probability. The mean and standard deviation of x = number of faulty units is below:

$$\mu = np = 200 \times 0.05 = 10$$
$$\sigma = \sqrt{200 \times 0.05(0.95)} = 3.08$$

The probability of x less or equal to 15 is estimated below:

$$\begin{aligned} P(x \leq 15) &= P(z \leq [(15.5 - 10.0)/3.08]) \\ &= P(z \leq 1.79) \end{aligned}$$

Applying Table 8.1 of Chap. 8, with some interpolation, the cumulative probability becomes:

$$\begin{aligned} P(x \leq 15) &\approx F(1.79) \\ &= 0.963 \end{aligned}$$

Hence,

$$P(185 \text{ or more non} - \text{faulty units}) \approx P(x \leq 15) = 0.963$$

Example 14.6 In phone book production, the estimated of a page error is 0.004. What is the probability of two or more page errors in a 1000 page book? Note, the normal approximation does not apply since p = 0.004 is less than 0.5 and,

$$np = 1000 \times 0.004 = 4 < 5$$

Instead the Poisson approximation can be used since n is large and p is small. In this situation, the Poisson parameter becomes:

$$\theta = np = 1000 \times 0.004 = 4.0$$

The probability sought is computed below:

$$\begin{aligned} P(x \geq 2) &= 1 - P(x \leq 1) \\ &= 1 - [P(0) + P(1)] \\ &= 1 - \left[e^{-4}4^0/0! + e^{-4}4^1/1!\right] \\ &= 1 - [0.0183 + 0.0733] \\ &= 0.908 \end{aligned}$$

14.6 Sample Data

When an experiment is conducted and repeated various times, and the number of successes is recorded, the sample data is the following:

n = number of trials
x = number of successes

14.7 Parameter Estimates with Sample Data

Using the sample data of n trials and x successes, the MLE of p is,

$\hat{p} = \mathrm{x}/\mathrm{n}$

The associated standard deviation of p becomes:

$\mathrm{s_p} = \sqrt{\hat{p}(1-\hat{p})/n}$

When the n-trial experiment is run m times, and the number of successes are ($\mathrm{x_1}, \ldots, \mathrm{x_m}$), the estimate of p becomes,

$\hat{p} = \sum x_i/(\mathrm{nm})$

In the event the i-th experiment has $\mathrm{n_i}$ trials and $\mathrm{x_i}$ successes, the estimate of p becomes:

$\hat{p} = \sum x_i / \sum n_i$

14.8 Parameter Estimates When No Data

Suppose a need to apply the binomial distribution occurs while the number of trials, n, is known, but not the probability of a success per trial, p. Also assume there is no sample data to estimate p. In this scenario, the analyst may seek opinion on the most-likely value of x in n trials. This estimate is denoted as $\tilde{x}$. Recalling the equation to measure the mode given earlier, the following relation is formed:

$$\tilde{x} = \mathrm{p}(\mathrm{n} + 1)$$

and therefore the estimate on p becomes:

$$\hat{p} = \tilde{x}/(\mathrm{n} + 1)$$

Example 14.7 Assume a scenario where n = 10 trials will be tested and the probability of a success, p, is not known. Also, no sample data is available, but an approximation on the most-likely number of successes in ten trials is 3. Hence, $\tilde{x}$ = 3 and the estimate of the probability per trial becomes:

$$\hat{p} = 3/(10 + 1) = 0.273$$

14.9 Summary

The binomial distribution applies when n trials of an experiment occur and the probability of a success in the outcome per trial remains the same on all trials. The random variable, x, is the number of successes in the n trials. The probability of x successes in n trials is described, along with the cumulative probability of x or less successes. The mean, variance, standard deviation and mode on the number of successes is listed. An estimate on the probability of a success per trial can be obtained when sample data, and also when no sample data.

Chapter 15
Geometric

15.1 Introduction

The geometric distribution applies when an experiment is run repeatedly until a successful outcome occurs, and the probability of a success is the same for all trials. The random variable could be defined as the number of fails till the first success, and has a range of integers zero and above. The random variable could also be labeled as the number of trials till the first success and the range is integers one and above. Both scenarios are described in the chapter. This situation occurs, for example, when a process produces units that need to meet acceptable engineering standards, and the process is repeated until an acceptable unit is produced. When the probability of a successful outcome is not known, sample data can be used to estimate the probability. Sometimes, no sample data is available, and a person of experience offers an approximation on the distribution and this data is used to estimate the probability of a successful outcome. The chapter also describes how the geometric distribution is the only discrete distribution that has a memory less property.

15.2 Fundamentals

The geometric distribution occurs when a process is repeated until a successful outcome occurs. The random variable is defined as the number of failures until a success happens. On other occasions, the random variable is listed as the number of trials till a success occurs. This chapter describes both scenarios.

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_15

15.3 Number of Failures

The parameter p represents the probability of a success per trial. The random variable is labeled as x and represents the number of fails till the first success, as listed below:

$x = 0, 1, 2, \ldots$

The function that defines the probability of x is below:

$P(x) = p(1-p)^x \qquad x = 0, 1, 2, \ldots$

The mean and variance of x are the following:

$\mu_x = (1-p)/p$
$\sigma_x{}^2 = (1-p)/p^2$

The cumulative distribution of x or less is obtained as follows:

$F(x) = 1 - (1-p)^{x+1} \qquad x = 0, 1, 2, \ldots$

15.4 Sample Data

Sample data is often used to estimate the parameter p. When the random variable is x = number of fails till a success, the experiment is run m times and the outcomes are labeled as: $(x_1, \ldots, x_m)$.

15.5 Parameter Estimate with Sample Data

The above sample data is used to estimate the probability of a success per trial in the following way. First the average value of x is computed from the m runs as shown below:

$\bar{x} = \sum x_i/m$

and, second, the estimate of p is obtained as follows:

$$\hat{p} = 1/(\bar{x} + 1)$$

Example 15.1 An oil exploration team has a success probability of p = 0.2 per drill. Using x as the number of fails to achieve a success, the probability of x is the following:

$$P(x) = 0.2(0.8^x) \qquad x = 0, 1, \ldots.$$

In this situation:

$$P(0) = 0.200$$
$$P(1) = 0.160$$
$$P(2) = 0.128$$

so on. The probability that two or less fails are needed to achieve a success is:

$$F(2) = P(x \leq 2) = 0.200 + 0.160 + 0.128 = 0.488$$

The mean, variance and standard deviation of x are computed below:

$$\mu_x = (1 - 0.2)/0.2 = 4.00$$
$$\sigma_x{}^2 = (1 - 0.2)/0.2^2 = 20.00$$
$$\sigma_x = \sqrt{5.00} = 4.47$$

Example 15.2 Suppose m = 5 samples from geometric data are obtained and yield the following values of x, the number of fails till a success: [3, 7, 5, 4, 5]. The sample average is: $\bar{x}$= 4.80, and thereby the estimate on p becomes:

$$\hat{p} = 1/(4.80 + 1) = 0.172$$

15.6 Number of Trials

The parameter of the geometric, p, represents the probability of a success per trial, and remains constant for all trials. When the number of trials, n, needed to achieve the first success, is the random variable, the range of n is the following:

$n = 1, 2, \ldots.$

The probability of n is listed below:

$$P(n) = p(1-p)^{n-1} \quad n = 1, 2, \ldots$$

and the cumulative probability of n or less is the following:

$$F(n) = 1 - (1-p)^{n} \quad n = 1, 2, \ldots$$

The mean and variance of n are below.

$$\mu_x = 1/p$$
$$\sigma_x{}^2 = (1-p)/p^2$$

Since, $n = x + 1$,

$$\mu_n = \mu_x + 1$$
$$\sigma_n{}^2 = \sigma_x{}^2$$

The cumulative probability of n or less is:

$$F(n) = 1 - (1-p)^{n} \qquad n = 1, 2, \ldots.$$

Example 15.3 Consider a process where the probability of a success is $p = 0.7$, and the analyst now wants to find the probability of n trials till a success. The probability becomes:

$$P(n) = 0.7(1-0.7)^{n-1} \quad n = 1, 2, \ldots$$

where,

$$P(1) = 0.700$$
$$P(2) = 0.210$$
$$P(3) = 0.063$$

so on.

The cumulative probability that n is 3 or less is the following:

$$\begin{aligned} F(3) &= 1 - (1-0.7)^3 \\ &= 1 - 0.027 \\ &= 0.973 \end{aligned}$$

The mean, variance and standard deviation of n are below.

$\mu_n = 1/0.7 = 1.428$
$\sigma_n{}^2 = (1 - 0.7)/0.7^2 = 0.612$
$\sigma_n = \sqrt{0.612} = 0.782$

15.7 Sample Data

Sample data is often used to estimate the parameter p. When the random variable is n (number of trials till a success), the experiment is run m times and the outcomes are labeled as: $(n_1, \ldots, n_m)$.

15.8 Parameter Estimate with Sample Data

The sample data is used to estimate the probability of a success per trial in the following way. First the average value of n is computed from the m runs as shown below:

$$\bar{n} = \sum n_i/\mathrm{m}$$

and, second, the estimate of p is obtained as follows:

$$\hat{p} = 1/\bar{n}$$

Example 15.4 Assume eight samples on n (number of trials till a success) are taken to estimate the probability of a success per trial with results: (1, 3, 2, 1, 1, 2, 1, 2). The average on the number of trials is: $\bar{n} = 13/8 = 1.615$, and thereby, the estimate of p is:

$$\hat{p} = 1/1.625 = 0.615$$

15.9 Parameter Estimate When No Sample Data

On some occasions the analyst wants to apply the geometric distribution but has no sample data to estimate the probability of a success. In this situation, a person of experience offers an approximation of the following type: 'The probability is α that the number of trials till a success is nα". This is the same as saying:

$$P(n \leq n\alpha) = \alpha$$

Note,

$$\alpha = F(n\alpha) = 1 - (1 - p)^{n}\alpha$$

and with some algebra, the estimate of p is:

$$\hat{p} \ = 1 - (\ 1 - \alpha)^{1/n}\alpha$$

Hence, with α and nα approximated, an estimate of p is obtained.

Example 15.5 A production facility wants to apply the geometric distribution in a simulation model and needs an estimate on the probability on a successful run of a new product. There is no sample data to base the estimate and the management seeks a design engineer's best opinion. The engineer states there should be 90% confidence that a successful run will take place in n = 3 trials or less. This approximation gives:

$$n\alpha = 3$$
$$\alpha = 0.90$$

Using the relation listed earlier, the estimate of p becomes:

$$\hat{p} = 1 - (1 - 0.9)^{1/3} = 0.536$$

15.10 Lexis Ratio

The lexis ratio for x (number of fails till a success), is:

$$\tau = 1/p$$

which is always larger than one.

But the lexis ratio for n (number of trials till a success), is:

$$\tau = (1 - p)/p$$

which is inconclusive since it could be above and below one.

15.11 Memory Less Property

The geometric distribution has a memory less property of the following type. The cumulative probability of $n = n_o$ or less remains the same regardless on where the counting begins. Below shows this relation. First observe the complementary probability on n larger than n_o.

$$\begin{aligned} P(n > n_o) &= 1 - P(n \leq n_o) \\ &= 1 - 1 + (1-p)^{no} \\ &= (1-p)^{no} \end{aligned}$$

Now note the complementary probability of n larger than $n_1 + n_o$, given n_1 trials have already taken place when the counting of trials begins:

$$\begin{aligned} P(n > n_1 + n_o | n_1) &= (1-p)^{n1+no} / (1-p)^{n1} \\ &= (1-p)^{no} \end{aligned}$$

Since the complementary probabilities are the same, the distribution has the memory less property.

15.12 Summary

The random variable of the geometric distribution could be the number of failures till a success, or the number of trials till the first success. The probability of each is given, along with the measures of the statistics on the mean and standard deviation. The probability of a success is estimated when sample data is available, and it is also estimated when no sample data but an approximation on the variable is offered. The geometric is the only discrete distribution with the memory less property.

Chapter 16
Pascal

16.1 Introduction

Blaise Pascal, a prominent French mathematician of the 1600s, was the first to formulate the Pascal distribution. The distribution is also often referred as the negative binomial distribution. When an experiment is run whose outcome could be a success or a failure with probabilities of p and (1 − p), respectively, and the analyst is seeking k successes of the experiment, the random variable is the minimum number of fails that occur to achieve the goal of k successes. This distribution is called the Pascal distribution. Some analysts working with the Pascal are interested when the random variable is the minimum number of trials to achieve the k successes. An example is when a production facility needs to produce k successful units for a customer order and the probability of a successful unit is less than one. The number of fails till the k successful units becomes the random variable. The chapter describes how to measure the probabilities for each situation. When the probability of a success per trial is not known, sample data may be used to estimate the probability. On other occasions, no sample data is available and an approximation on the distribution is used to estimate the probability.

16.2 Fundamentals

The random variable of the Pascal is sometimes referred as the number of fails till k successes are achieved; and sometimes as the number of trials till k successes. Below shows how to compute the probability measures for each situation.

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_16

16.3 Number of Failures

When the random variable is the number of failures till k successes, the variable, denoted as x, is as follows:

$x = 0, 1, 2, \ldots$

The probability, and the cumulative probability of x are obtained as below:

$$P(x) = \binom{k+x-1}{x} p^k (1-p)^x \qquad x = 0, 1, \ldots$$

$$F(x) = \sum_{j=0}^{x} \binom{k+j-1}{j} p^k (1-p)^j \qquad x = 0, 1, \ldots$$

The mean and variance of x are the following:

$\mu_x = k(1-p)/p$
$\sigma_x{}^2 = k(1-p)/p^2$

The mode of x, denoted as $\widetilde{\mu}$ is derived in the following way:
Let $w = [k(1-p) - 1]/p$

if w is an integer: $\widetilde{\mu} = w$ and $w + 1$
else: $\widetilde{\mu} = \text{floor}(w + 1)$

16.4 Parameter Estimate When Sample Data

When the parameter p is not known, sample data could be used to estimate the value. The data is obtained when m observations of x, (x_1, …, x_m), are gathered representing the number of fails from each of the m samples. The average value of x from the sample is computed as below:

$$\bar{x} = \sum x_i / m$$

and the MLE of p is the following:

$$\hat{p} = k/(\bar{x} + k)$$

16.5 Parameter Estimate When No Data

Sometimes, the analyst has a need to apply the Pascal distribution but has no sample data to estimate p (probability of a successs). In this situation, the analyst may seek the advice of an expert on the model and ask for an approximation on the most-likely value of x. The approximation of this type is the same as the mode of x, denoted as $\tilde{x}$. Using this estimate of the mode, and the value of k already known, the estimate of p is obtained as shown below. Recall how the mode is obtained earlier using the interim value of w listed below:

$$\mathrm{w} = [\mathrm{k}(1 - \mathrm{p}) - 1]/\mathrm{p}$$

Substituting the mode for w, $\tilde{x} = \mathrm{w}$, and applying some algebra gives the following estimate of p:

$$\hat{p} = (\mathrm{k} - 1)/\left(\mathrm{k} + \tilde{x}\right)$$

Example 16.1 A chemist seeks a mixture of chemicals and temperature that produces a positive result with a probability of p = 0.5 per trial. The chemist requires k = 3 such mixtures to complete an experiment. Letting x equal the number of mixtures that fail till the three positive ones are achieved is computed as below:

$$\mathrm{P}(\mathrm{x}) = \binom{3 + x - 1}{x} 0.5^3 (1 - 0.5)^{\mathrm{x}} \qquad \mathrm{x} = 0, 1, \ldots$$

The probability of x = 0 or 1 are obtained below:

$$\mathrm{P}(0) = \binom{3 + 0 - 1}{0} 0.5^3 (1 - 0.5)^0 = 0.125$$

$$\mathrm{P}(1) = \binom{3 + 1 - 1}{1} 0.5^3 (1 - 0.5)^1 = 0.187$$

and the probability on x less or equal to one is:

$$\mathrm{F}(1) = 0.125 + 0.187 = 0.312$$

The mean, variance and standard deviation are the following:

$$\mu_{\mathrm{x}} = 3(1 - 0.5)/0.5 = 3.00$$

$$\sigma_{\mathrm{x}}{}^2 = 3(1 - 0.5)/0.5^2 = 6.00$$

$$\sigma_{\mathrm{x}} = \sqrt{6.00} = 2.45$$

Example 16.2 An assembly requires two units and requests a shop to build as needed. The units are very sensitive and the machining does not always produce an acceptable unit, and the probability of an acceptable unit is not known. The shop wants an estimate on producing an acceptable unit to properly price the job request. Over the past months, the shop retrieves m = 8 the similar requests and gathers the number of failures that occurred with them. These are listed below:

$\mathrm{x_i} = (2, 5, 3, 4, 6, 7, 2, 5)$

The sample average on the number of failures is below:

$\bar{x} = 34/8 = 4.25$

Using the average, the probability of producing an acceptable unit is obtained below:

$\hat{p} = 2/(4.25 + 2) = 0.32$

Example 16.3 A researcher is building a simulation model and requires the use of the Pascal distribution where the required number of acceptable units is k = 4 and the probability, p, of obtaining an acceptable unit per trial is not known. The researcher seeks the opinion of several co-workers on the most-likely value of the number of non-acceptable units for this process. The typical approximation is $\tilde{x} = 3$. With this approximation, the estimate on p is obtained as below:

$\hat{p} = (4 - 1)/(\ 4 + 3) = 0.428$

16.6 Number of Trials

Consider the Pascal distribution when the random variable is the number of trials till k successes. In this situation, the variable is denoted as n where:

$\mathrm{n} = \mathrm{k}, \mathrm{k} + 1, \ldots.$

The probability and the cumulative probability of n are computed as below:

$$\mathrm{P(n)} = \binom{n-1}{k-1} \mathrm{p^k (1-p)^{n-k}} \qquad \mathrm{n} = \mathrm{k}, \mathrm{k}+1, \ldots$$

$$\mathrm{F(n)} = \sum_{y=k}^{n} P(y) \qquad \mathrm{n} = \mathrm{k}, \mathrm{k}+1, \ldots$$

The mean and variance of n are listed below.

$$\mu_n = k/p$$
$$\sigma_n{}^2 = k(1-p)/p^2$$

Since n = (x + k), the mean and standard deviation of n are related to their counterparts when x is the number of fails, as listed below:

$$\mu_n = \mu_x + k$$
$$\sigma_n{}^2 = \sigma_x{}^2$$

16.7 Lexis Ratio

The lexis ratio of x (number of fails till k successes) is:

$$\tau = 1/p$$

and thereby is always larger than one. The same measure for n (number of trials till k successes) is:

$$\tau = (1-p)/p$$

is sometimes below one and sometimes above, and as such is inconclusive.

16.8 Parameter Estimate When Sample Data

When the parameter p is not known, it could be estimated using sample data. Assuming m samples of variable n are observed randomly, with the results: $(n_1, \ldots, n_m)$. The data yields the sample average of n, and also the estimate of p as shown below:

$$\bar{n} = \sum n_i/m$$
$$\hat{p} = k/\bar{n}$$

Example 16.4 Forty percent of voters are in favor of a referendum and a reporter is seeking five of them for a feature article. Below shows how many voters he has to

interview to gain the five that he needs. In this situation, n represents the number of voters he interviews (number of trials), and p = 0.40 is the probability a voter is in favor of the referendum. The probability of n is listed below:

$$\mathrm{P(n)} = \binom{n-1}{5-1} 0.4^5 (0.6)^{\mathrm{n}-5} \qquad \mathrm{n} = 5, 6, \ldots..$$

Note:

$$\mathrm{P}(5) = \binom{5-1}{5-1} 0.4^5 (0.6)^0 \quad = 0.0102$$

$$\mathrm{P}(6) = \binom{6-1}{5-1} 0.4^5 (0.6)^1 \quad = 0.0307$$

so on.

The probability of n = 6 or less becomes:

$$\mathrm{P(n} \leq 6) = 0.0102 + 0.307 = 0.0409$$

The mean and variance of n are below:

$$\mu_{\mathrm{n}} = 5/0.4 = 12.50$$
$$\sigma_{\mathrm{n}}{}^2 = 5(1 - .4)/.4^2 = 5.56$$
$$\sigma_{\mathrm{n}} = \sqrt{5.56} = 2.36$$

Note, the counterpart mean and variance of x = number of failures are the following:

$$\mu_{\mathrm{x}} = (12.5 - 5) = 7.5$$
$$\sigma_{\mathrm{x}}{}^2 = 5.56$$

Example 16.5 Assume a home-run derby where m = 3 players are competing for the least number of trials till they hit five balls over the fence (k = 5) and assuming all are equally qualified. The number of trials for the three participants is the following: [15, 12, 14]. A reporter inquires on the probability of them hitting a ball over the fence. Since the average number of trials is $\bar{n} = 13.67$, the probability estimate becomes $\hat{p} = 5/13.67 = 0.365$.

16.9 Summary

The Pascal is a distribution where the random variable concerns the number of failures to achieve k successes of an event where the probability of a success is constant for all trials. The random variable could also be noted as the number of trials till the k successes. The probability of the number of fails (or trials) is readily computed, along with the mean and standard deviation. When the probability of a success is not known, sample data is used to estimate the probability. When no sample data is available, an estimate of the most-likely number of trials allows an estimate on the probability of a success per trial.

Chapter 17
Poisson

17.1 Introduction

The Poisson distribution is named after Simmeon Poisson, a leading French mathematician during the early 1800s. The distribution applies when the random variable is discrete and represents the number of events that occur in a unit-of-scale, such as unit-or-time or unit-of-area. The rate of events occurring is constant and the number of events are the integers of zero and larger. This distribution is used heavily in the study of queuing systems, and also in statistical process control. In queuing, the random variable pertains to the number of arriving units to a system in a unit-of-time; and also pertains to the number of departing units in a unit-of-time for a continuously busy service facility. In statistical process control, the random variable pertains to the number of defectives occurring in a unit of product. The distribution is mostly described in one unit-of-scale, but easily extends to multiple units-of-scale. This distribution is often used in place of the normal when the expected number of events is relatively low.

17.2 Fundamentals

The Poisson distribution has parameter, θ, that is a measure of the number of events that occur in a unit-of-scale. The scale could be time, area, volume, so on. The random variable is x which includes integers zero or larger. The probability function of x is listed below:

$$P(x) = \theta^x e^{-\theta}/x! \qquad x = 0, 1, 2, \ldots..$$

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_17

The mean and variance of x are the following:

$\mu = \theta$
$\sigma^2 = \theta$

The mode of x, denoted by $\widetilde{\mu}$, is obtained in the following way:

if θ is integer:	$\widetilde{\mu} = \theta - 1$ and θ
else	$\widetilde{\mu} = \text{floor}(\theta)$

17.3 Lexis Ratio

The lexis ratio of the Poisson is below:

$\tau = \theta/\theta = 1.00$

17.4 Parameter Estimate When Sample Data

When n samples of x are collected as: (x_1, …, x_n), and the sample average is $\bar{x}$, the MLE of θ is:

$\widehat{\theta} = \bar{x}$

17.5 Parameter Estimate When No Data

When a need to estimate the parameter value θ and no sample data is available, the relation to the mode may be used In the following way. The researcher seeks the advice from a person familiar with the system under study who gives an approximation on the most-likely value of x, denoted as $\tilde{x}$. Since the mode and parameter are related in the following way:

$\widetilde{\mu} = \theta - 1$ and θ

substituting $\tilde{x}$ for $\widetilde{\mu}$ yields an estimate of the mode as below:

$\widehat{\theta} \approx \tilde{x} + 0.5$

17.6 Exponential Connection

The Poisson discrete random variable x is related to the exponential continuous random variable t in the following way. When the unit-of-scale is unit-of-time, the expected value of x is:

$E(x) = \mu = \theta =$ expected number of events in a unit-of-time

The time between events, denoted as t, is exponentially distributed and the probability function is:

$$f(t) = \theta e^{-\theta t} \qquad t > 0$$

The expected value of t becomes:

$E(t) = 1/\theta =$ expected time between events

Example 17.1 Customers arrive to a restaurant on Saturday mornings at an average rate of 5 per 30 min. For this time duration, the mean and variance of the number of customers arriving are:

$$\mu = 5.0$$
$$\sigma^2 = 5.0$$

The probability of x arrivals in 30 min is computed as follows:

$$P(x) = 5^x e^{-5}/x! \quad x = 0, 1, 2, \ldots.$$

Note,

$$P(0) = 5^0 e^{-5}/0! = 0.007$$
$$P(1) = 5^1 e^{-5}/1! = 0.034$$
$$P(2) = 5^2 e^{-5}/2! = 0.084$$

so on.

The probability x is two or less becomes:

$$P(x \leq 2) = 0.007 + 0.034 + 0.084 = 0.125$$

Example 17.2 Suppose a Poisson system where the parameter value is not known; and the analyst obtains the following n = 8 samples of x: [3, 6, 2, 5, 3, 5, 2, 7]. From this sample data, the analyst computes the average of $\bar{x} = 4.125$, and thereby the parameter estimate becomes $\widehat{\theta} = 4.125$.

Example 17.3 A researcher wants to use the Poisson distribution in a simulation model but has no estimate on the Poisson parameter and has no sample data to assist. A person familiar with the system approximates the most-likely value of x as $\tilde{x} = 3.0$. Using this approximation, the estimate of the parameter becomes:

$$\widehat{\theta} \approx 3.0 + 0.5 = 3.5$$

Example 17.4 Customers arrive to a bank on Saturday mornings via a Poisson distribution with a rate of 6 per hour. The Poisson parameter could be expressed as:

$\theta = 6$ per hour

or as

$\theta = 1$ per 10 min

The time between customer arrivals, denoted as t, is exponential with an expected time between arrivals as follows:

$$\mathrm{E(t)} = 1/\theta$$

If the unit-of-time is an hour,

$$\mathrm{E(t)} = 1/6 = 0.167\text{-hour}$$

If the unit-of-time is 10-min,

$$\mathrm{E(t)} = 1/1 = 10\text{-min}$$

Using hours as the unit-of-time, the parameter is $\theta = 6$ per-hour, and the probability density function of t is:

$$\mathrm{f(t)} = 6\mathrm{e}^{-6\mathrm{t}} \qquad \mathrm{t} > 0$$

17.7 Poisson with Multi Units

The Poisson random variable is often applied with a multiple number of n-units-of-scale, where n is larger than zero. When $\mathrm{n} = 1$, the unit-of-scale is the same as the Poisson random variable described above. When θ is defined with one unit-of-scale, the parameter that applies with n-units-of-scale is denoted and obtained as below:

$$\theta_{\mathrm{n}} = \theta \times \mathrm{n}$$

The random variable remains as x where the admissible region is:

$x = 0, 1, \ldots$

The probability function of x in the n-units-of-scale is listed below:

$$Pn(x) = \theta_n e^{-\theta n}/x! \qquad x = 0, 1, 2, \ldots..$$

The mean and variance of x in the n-units-of-scale are the following:

$\mu = \theta_n$
$\sigma^2 = \theta_n$

Example 17.5 Consider a Poisson random variable where the parameter value for one-unit-of-scale is $\theta = 2.0$. Below shows how some of the statistical measures vary when n = 0.5, 1.0 and 1.5, say.

$\mu = 1.0$ at $n = 0.5$
$\mu = 2.0$ at $n = 1.0$
$\mu = 3.0$ at $n = 1.5$

$P_{0.5}(x) = 1.0e\text{-}^{1.0x}/x!$ $\quad x = 0, 1, 2, \ldots.$
$P_{1.0}(x) = 2.0e\text{-}^{2.0x}/x!$ $\quad x = 0, 1, 2, \ldots.$
$P_{1.5}(x) = 3.0e\text{-}^{3.0x}/x!$ $\quad x = 0, 1, 2, \ldots.$

$P_{0.5}(0) = 0.368$ $\quad P_{0.5}(1) = 0.368$
$P_{1.0}(0) = 0.135$ $\quad P_{1.0}(1) = 0.271$
$P_{1.5}(0) = 0.050$ $\quad P_{1.5}(1) = 0.149$

Example 17.6 In a glass manufacturing plant, the average rate of a flaw in a one-square-foot pane is 0.001. The probability of a flaw on a 16-square-feet pane is obtained in the following way. To begin, note the parameter settings for a pane of 1-square-foot and for16-square–feet are below:

$\theta = 0.001$ for a 1-square-foot pane
$\theta_{16} = 0.016$ for a 16-square-feet pane

The probability of zero flaws on a pane of 16-square-feet is obtained as below:

$$P_{16}(0) = 0.016^0 e^{-0.016}/0! = 0.984$$

Hence, the probability of one or more flaws on a pane of 16-square-feet becomes:

$$P_{16}(x \geq 1) = 1 - 0.984 = 0.016$$

17.8 Summary

The Poisson distribution applies when the number of events in a unit-of-scale is relatively low, The random variable is discrete and includes all integers of zero and larger. The Poisson discrete number of events in a unit-of-time is related to the exponential continuous time between events. The distribution is often described with one unit-of-scale, but easily extends to multiple units-of-scale. The Poisson parameter can be estimated using sample data. When sample data is not available, and estimate on the parameter value is also possible.

Chapter 18
Hyper Geometric

18.1 Introduction

The hyper geometric distribution applies when a population of size N has D marked items, and a sample of n items taken without replacement yields x marked items. This differs from the binomial distribution where the population size is infinite and the samples are taken with replacement. The hyper geometric applies often in quality applications when a lot of N items with D defectives (quantity unknown) and a sample of n without replacement is taken. The sample result of x defectives allow the management to gauge the quality of the lot.

18.2 Fundamentals

The probability function of x is the following:

$$\mathrm{P(x)} = \binom{N-D}{n-x}\binom{D}{x}/\binom{N}{n} \qquad \mathrm{x} = 0, \ldots., \min(\mathrm{n}, \mathrm{D})$$

The mean and variance of x are listed below:

$$\sigma^2 = \mathrm{n[D/N][1 - D/N][N - n]/[N - 1]}$$

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_18

18.3 Parameter Estimate When Sample Data

An analyst wants to apply the hyper geometric distribution in an application where the lot size and sample size is known, but not the number of defectives in the lot. Over the past m periods, with N and n fixed in size, the number of defectives from the samples are listed as: ($x_1, \ldots, x_m$). The average of the number of defectives per period becomes $\bar{x}$. Utilizing the relation between, D, N, x and n, the estimate of the probability of a defective is the following:

$$\frac{\hat{D}}{N} = \bar{x}/\mathrm{n}$$

Hence, for a lot size of N, the estimate on the average number of defectives becomes:

$$\bar{D} = \mathrm{N}\bar{x}/\mathrm{n}$$

18.4 Binomial Estimate

In the event, the sample size of n (samples without replacement) is small compared to N (population size), the binomial distribution can be used to estimate the probability of the hyper geometric by letting p = D/N with n the sample size.

Example 18.1 A shipment from a supplier arrives to a plant with a lot of N = 8 units, and a sample on n = 2 units are taken without replacement to seek out any defectives. If the number of defectives in the lot is D = 1, the probability of finding x defectives is below:

$$\mathrm{P(x)} = \binom{7}{2-x}\binom{1}{x}/\binom{8}{2} \qquad \mathrm{x} = 0, 1$$

The probabilities become:

$$\mathrm{P}(0) = \binom{7}{2-0}\binom{1}{0}/\binom{8}{2} = 0.75$$

$$\mathrm{P}(1) = \binom{7}{2-1}\binom{1}{1}/\binom{8}{2} = 0.25$$

The mean and variance of x are computed below:

$$\mu = 2 \times 1/8 = 0.25$$
$$\sigma^2 = 2 \times 1/8[1 - 1/8][8 - 2]/[8 - 1] = 0.1875$$

Example 18.2 Over 6 weeks, a lot arrives to a plant with N = 8 units and n = 2 are sampled without replacement. The history on the 6 weeks yields the following number of defectives: (0, 1, 0, 0,1, 0), yielding an average of $\bar{x} = 0.333$ defectives per lot. The estimate of the ratio of defectives to the lot size is below:

$$\frac{\hat{D}}{N} = \mathrm{x/n} = 0.333/2 = 0.167$$

Hence, $\bar{D} = 0.167 \times 8 = 1.33$ is the estimate of the average number of defectives in a lot of size N = 8.

Example 18.3 A park ranger wants to estimate the number of large fish in a small lake. The ranger catches D = 20 such fish and tags them. After tagging, the tagged fish are released back into the lake. Soon after, the ranger catches another n = 10 such fish and x = 2 have the tag. An estimate on the number of large fish, N, in the lake is obtained in the following way:

Since,

$$\mathrm{E(x)} = \mathrm{nD/N}$$
$$\hat{N} = \mathrm{nD/x}$$
$$= 10 \times 20/2 = 100$$

Example 18.4 A lot of size N = 100 is received from a supplier at a plant and the number of samples taken without replacement is n = 5. Assuming D = 8 defectives are in the lot, the analyst want to find the probability of x defectives in the sample. The hyper geometric probability function on x is below:

$$\mathrm{P(x)} = \binom{92}{5-x}\binom{8}{x}/\binom{100}{5}$$

The computations to the above are a bit difficult because of the large value of N = 100. Since the ratio of n over N is rather small, (5/100 = 0.05), the binomial distribution can be used to approximate the probabilities sought. The binomial probability function on x is below:

$$\mathrm{P(x)} = \binom{5}{x}0.08^{\mathrm{x}}0.92^{5-\mathrm{x}}$$

Below lists the comparison between the hyper geometric and binomial probabilities:

x	Hyper geometric	Binomial
0	0.653	0.659
1	0.297	0.287
2	0.047	0.050
3	0.003	0.004
4	0.000	0.000
5	0.000	0.000

18.5 Summary

The hyper geometric distribution applies when a finite lot size may have some items with a characteristic and a small sample is taken without replacement from the lot to seek the number of items with the characteristic. The distribution is often used in industry for quality assurance applications.

Chapter 19
Bivariate Normal

19.1 Introduction

Over the years a great many scholars have contributed to the literature concerning the bivariate normal distribution. In 1998, Montira Jantaravareerat and N. Thomopoulos describe a way to estimate the cumulative probability of the distribution. In this chapter, a new method is shown on computing the joint cumulative probability. The bivariate normal has two variables, x_1, x_2, that are jointly related, and has five parameters, μ_1, μ_2, σ_1, σ_2, ρ. The marginal distributions are normally distributed, and when the value of one of the variables is known, the distribution on the other is also normally distributed. The variables are converted to a new set, z_1, z_2, that are jointly related by the bivariate standard normal distribution. The latter two variables are easier to apply mathematically in the computations. An approximation method is developed here to compute the joint probability of the two variables. Table values are listed and examples are presented to demonstrate the application.

19.2 Fundamentals

The chapter begins by describing the bivariate normal distribution, the variables, x_1, x_2, and its five parameters. The parameters are taken from the marginal distributions of x_1 and x_2, which are normally distributed. When one of the variable values is given, the other variable has a distribution that is also normally distributed with its mean and standard deviation defined. Because the computations using x_1 and x_2 are difficult, the variables are converted to their counterparts, z_1 and z_2 that are from the bivariate standard normal distribution and are computationally easier to apply. The marginal and conditional distributions of z_1 and z_2 are also described. Since there is no closed-form solution to calculate the joint probability of the pair (k_1, k_2) that are

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_19

particular values of (z_1, z_2), an approximation method is developed in this chapter. Using the approximation, some table values are listed, and examples are presented to describe how to use the tables.

19.3 Bivariate Normal

When two variables, x_i, x_2, are bivariate normally distributed, their joint relation has five parameters, μ_1, μ_2, σ_1, σ_2, ρ, and their designation is the following where the mean and standard deviation parameters pertain to the marginal distribution of the variables.

$$(x_1, x_2) \sim \mathrm{BVN}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$$

The joint probability function is listed below:

$$f(x_1, x_2) = 1/\left[2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}\exp\left[-w/2\left(1-\rho^2\right)\right]\right.$$

$$w = \left[(x_1-\mu_1)/{\sigma_1}^2 + \left[(x_2-\mu_2)/\sigma_2\right]^2 - 2\rho[(x_1-\mu_1)\,(x_2-\mu_2)/\sigma_1\sigma_2\right]$$

19.4 Marginal Distributions

The marginal distribution on each of the variables is normally distributed and are designated as follows:

$$x_1 \sim N\left(\mu_1, {\sigma_1}^2\right)$$

$$x_2 \sim N\left(\mu_2, {\sigma_2}^2\right)$$

Note, μ_1 and σ_1 are the marginal parameters for x_1, while μ_2 and σ_2 are the same for x_2. The correlation between the two variables is ρ and is computed as follows:

$$\rho = \sigma_{12}/(\sigma_1\sigma_2)$$

where σ_{12} is the covariance between x_1 and x_2 and is obtained as below:

$$\sigma_{12} = E(x_1x_2) - E(x_1)E(x_2)$$

19.5 Conditional Distribution

An important relation occurs when the value of one of the variables is known, such as x_{1o} is a particular value of x_1. With this conditional situation, the distribution of x_2, denoted as $x_2|x_{1o}$, becomes normally distributed with designation:

$$x_2 \mid x_{1o} \sim N\left(\mu_{x2|x1o}, {\sigma_{x2|x1o}}^2\right)$$

The conditional mean and variance of $x_2|x_{1o}$ is below:

$$\mu_{x2|x1o} = \mu_2 + \rho(\sigma_2/\sigma_1)(x_{1o} - \mu_1)$$

$${\sigma_{x2|x1o}}^2 = {\sigma_2}^2\left(1 - \rho^2\right)$$

19.6 Bivariate Standard Normal

The easier way to study the bivariate normal is by the standard bivariate normal distribution with variables z_1 and z_2. The marginal distributions of the two standard variables each have a mean of zero and a variance of one; while the correlation ρ is a measure of the joint relation between the two variables. The designation between the two variables is listed below:

$$(z_1, z_2) \sim \text{BVN}(0, 0, 1, 1, \rho)$$

Note, the mean and standard deviation of each are as follows:

$$\mu_1 = 0, \sigma_1 = 1, \mu_2 = 0, \sigma_2 = 1$$

19.7 Marginal Distribution

The marginal distributions of the two variables are standard normal and are listed below:

$$z_1 \sim N(0, 1)$$
$$z_2 \sim N(0, 1)$$

19.8 Conditional Distributions

When one of the variable values is known, such as letting k_1 be a particular value of z_1, the conditional notation for variable z_2 becomes $z_2|k_1$, and the conditional distribution is also normally distributed, with designation below:

$$z_2 \mid k_1 \sim N\left(\mu_{z2|k1}, {\sigma_{z2|k1}}^2\right)$$

where the mean and variance are the following:

$$\mu_{z2|k_1} = \rho k_1$$

$${\sigma_{z2|k1}}^2 = \left(1 - \rho^2\right)$$

Note when $z_2 = k_2$, the conditional variable z_1 becomes $z_1|k_2$, and the conditional distribution is normal and is designated as below:

$$z_1 \mid k_2 \sim N\left(\mu_{z1|k2}, {\sigma_{z1|k2}}^2\right)$$

with mean and variance,

$$\mu_{z1|k2} = \rho k_2$$
$${\sigma_{z1|k2}}^2 = \left(1 - \rho^2\right)$$

19.9 Approximation to the Cumulative Joint Probability

The cumulative joint probability of z_1 less or equal to k_1, and z_2 less or equal to k_2 is expressed in the following way:

$$F(k_1, k_2) = P(z_1 \leq k_1 \cap z_2 \leq k_2)$$

However, since there is no closed-form solution to the above probability, an approximation is developed here and is described below. Mathematically, the joint probability is obtained as the following:

$$F(k_1, k_2) = \int_{-\infty}^{k_1} \int_{-\infty}^{k_2} f(z_1, z_2) dz_2 dz_1$$

An equivalent relation is below:

$$\begin{aligned} \mathrm{F}(\mathrm{k}_1, \mathrm{k}_2) &= \int_{-\infty}^{k_1} \left[\int_{-\infty}^{k_2} f(z_2|z_1) dz_2 \right] f(z_1) dz_1 \\ &= \int_{-\infty}^{k_1} F(k_2|z_1)\mathrm{f}(\mathrm{z}_1)\mathrm{dz}_1 \end{aligned}$$

The approximation uses the following two relations:

1. Hasting's standard normal approximation for obtaining the cumulative probability, F(z), from z as described in Chap. 8.
2. A discrete normal distribution where:

$$\mathrm{P}(\mathrm{k}) = \mathrm{f}(\mathrm{k}) / \sum_{z=-3.0}^{3.0} f(z) \text{ for } \mathrm{z} = [-3.0, (0.1), 3.0],$$

where f(z) is the probability density of z ~ N(0,1), and k is a particular value of z.

The above two relations are used in the computations below:

$$\mathrm{F}(\mathrm{k}_1, \mathrm{k}_2) \approx \sum_{z_1=-3.0}^{k_1-0.1} F(k_2|z_1)P(z_1) + 0.5\mathrm{F}(\mathrm{k}_2|\mathrm{k}_1)\mathrm{P}(\mathrm{k}_1)$$

For example, if $(\mathrm{k}_1, \mathrm{k}_2) = (1.0,1.0)$,

$$\begin{aligned} \mathrm{F}(1.0, 1.0) = \mathrm{F}(1.0| -3.0)\mathrm{P}(-3.0) + \mathrm{F}(1.0| -2.9)\mathrm{P}(-2.9) + \ldots \\ \ldots . + \mathrm{F}(1.0|0.9)\mathrm{P}(0.9) + 1/2\mathrm{F}(1.0|1.0)\mathrm{P}(1.0) \end{aligned}$$

Note, all of the cumulative functions are standard normally distributed where,

$$\mathrm{F}(\mathrm{k}_2|\, \mathrm{z}_1) = \mathrm{F}\left[\left(\mathrm{k}_2 - \mu_{z2|z1} \right) / \sigma_{z2|z1} \right]$$

and,

$$\mu_{\zeta 2|z1} = \rho \mathrm{z}_1$$

$$\sigma_{z2|z1} = \sqrt{(1 - \rho^2)}$$

Table 19.1 $F(k_1,k_2)$ for $\rho = -0.8$ and $\rho = 0.8$ and for $k_1 = [-3.0,(0.5), 3.0]$ and $k_2 = [-3.0, (0.5), 3.0]$.

$\rho = -0.8$													
k2/k1	−3	−2.5	−2	−1.5	−1	−0.5	0	0.5	1	1.5	2	2.5	3
3	0	0.01	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98	0.99	1.00
2.5	0	0.01	0.02	0.06	0.15	0.30	0.49	0.69	0.84	0.93	0.97	0.99	0.99
2	0	0	0.01	0.05	0.14	0.29	0.48	0.67	0.82	0.91	0.96	0.97	0.98
1.5	0	0	0.01	0.03	0.11	0.25	0.43	0.63	0.78	0.87	0.91	0.93	0.93
1	0	0	0	0.01	0.06	0.17	0.35	0.53	0.68	0.78	0.82	0.84	0.84
0.5	0	0	0	0	0.02	0.09	0.22	0.39	0.53	0.63	0.67	0.69	0.69
0	0	0	0	0	0.01	0.03	0.10	0.22	0.35	0.43	0.48	0.49	0.50
−0.5	0	0	0	0	0.00	0.01	0.03	0.09	0.17	0.25	0.29	0.30	0.31
−1	0	0	0	0	0.00	0.00	0.01	0.02	0.06	0.11	0.14	0.15	0.16
−1.5	0	0	0	0	0	0	0	0	0.01	0.03	0.05	0.06	0.07
−2	0	0	0	0	0	0	0	0	0	0.01	0.01	0.02	0.02
−2.5	0	0	0	0	0	0	0	0	0	0	0	0	0
−3	0	0	0	0	0	0	0	0	0	0	0	0	0
$\rho = 0.8$													
k2/k1	−3	−2.5	−2	−1.5	−1	−0.5	0	0.5	1	1.5	2	2.5	3
3	0	0.01	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98	0.99	1.00
2.5	0	0.01	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98	0.99	0.99
2	0	0.01	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.97	0.98	0.98
1.5	0	0.01	0.02	0.07	0.16	0.31	0.50	0.69	0.83	0.90	0.93	0.93	0.93
1	0	0.01	0.02	0.07	0.16	0.31	0.49	0.67	0.78	0.83	0.84	0.84	0.84
0.5	0	0.01	0.02	0.07	0.16	0.30	0.47	0.60	0.67	0.69	0.69	0.69	0.69
0	0	0.01	0.02	0.07	0.15	0.28	0.40	0.47	0.49	0.50	0.50	0.50	0.50
−0.5	0	0.01	0.02	0.06	0.13	0.22	0.28	0.30	0.31	0.31	0.31	0.31	0.31
−1	0	0.01	0.02	0.05	0.10	0.13	0.15	0.16	0.16	0.16	0.16	0.16	0.16
−1.5	0	0	0.02	0.03	0.05	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.07
−2	0	0	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
−2.5	0	0	0	0	0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
−3	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 19.2 $F(k_1,k_2)$ for $\rho = -1.0$ to 1.0 and for $k_1 = [-3.0, (1.0), 3.0]$ and $k_2 = [-3.0, (1.0), 3.0]$

k1/k2		−1.0	−0.9	−0.8	−0.7	−0.6	−0.5	−0.4	−0.3	−0.2	0.1	**0.0**
3	**3**	1.000	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.998	0.998	0.998
3	**2**	0.978	0.977	0.977	0.977	0.977	0.977	0.977	0.977	0.977	0.977	0.977
3	**1**	0.842	0.841	0.841	0.841	0.841	0.841	0.841	0.841	0.841	0.841	0.841
3	**0**	0.500	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499
3	**−1**	0.158	0.157	0.157	0.157	0.157	0.157	0.157	0.157	0.158	0.158	0.158
3	**−2**	0.022	0.021	0.021	0.021	0.021	0.021	0.021	0.022	0.022	0.022	0.022
3	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	**3**	0.978	0.978	0.978	0.978	0.978	0.977	0.977	0.977	0.977	0.977	0.977
2	**2**	0.956	0.957	0.956	0.956	0.956	0.956	0.956	0.956	0.956	0.956	0.956
2	**1**	0.820	0.820	0.820	0.820	0.820	0.820	0.820	0.820	0.821	0.822	0.823
2	**0**	0.478	0.478	0.478	0.478	0.479	0.480	0.481	0.483	0.484	0.487	0.489
2	**−1**	0.136	0.137	0.138	0.140	0.143	0.145	0.147	0.150	0.151	0.153	0.154
2	**−2**	0.001	0.009	0.013	0.015	0.017	0.018	0.019	0.020	0.020	0.021	0.021
2	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	**3**	0.842	0.842	0.842	0.842	0.842	0.842	0.841	0.841	0.841	0.841	0.841
1	**2**	0.820	0.820	0.820	0.820	0.820	0.820	0.820	0.821	0.821	0.822	0.823
1	**1**	0.684	0.684	0.684	0.685	0.686	0.688	0.690	0.694	0.698	0.703	0.708
1	**0**	0.342	0.343	0.348	0.355	0.364	0.373	0.383	0.392	0.402	0.411	0.421
1	**−1**	0.006	0.044	0.061	0.075	0.086	0.096	0.105	0.113	0.120	0.127	0.133
1	**−2**	0.000	0.000	0.002	0.004	0.007	0.009	0.012	0.014	0.015	0.017	0.018
1	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0	**3**	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
0	**2**	0.478	0.478	0.478	0.479	0.479	0.480	0.482	0.483	0.485	0.487	0.489
0	**1**	0.342	0.343	0.348	0.355	0.364	0.373	0.383	0.392	0.402	0.411	0.421
0	**0**	0.010	0.072	0.103	0.127	0.148	0.167	0.185	0.202	0.218	0.234	0.250

(continued)

Table 19.2 (continued)

0	**−1**	0.000	0.001	0.006	0.013	0.022	0.031	0.041	0.050	0.060	0.070	0.079
0	**−2**	0.000	0.000	0.000	0.000	0.001	0.002	0.003	0.005	0.007	0.009	0.011
0	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
−1	**3**	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.159	0.159
−1	**2**	0.136	0.136	0.138	0.140	0.143	0.146	0.148	0.150	0.152	0.154	0.155
−1	**1**	0.006	0.043	0.061	0.075	0.086	0.096	0.105	0.113	0.121	0.127	0.134
−1	**0**	0.000	0.001	0.006	0.013	0.022	0.031	0.041	0.050	0.060	0.070	0.079
−1	**−1**	0.000	0.000	0.000	0.001	0.002	0.004	0.007	0.010	0.015	0.020	0.025
−1	**−2**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.002	0.003
−1	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
−2	**3**	0.022	0.021	0.022	0.022	0.022	0.022	0.022	0.023	0.023	0.023	0.023
−2	**2**	0.001	0.009	0.013	0.015	0.017	0.019	0.020	0.021	0.021	0.022	0.022
−2	**1**	0.000	0.000	0.002	0.004	0.007	0.010	0.012	0.014	0.016	0.018	0.019
−2	**0**	0.000	0.000	0.000	0.000	0.001	0.002	0.003	0.005	0.007	0.009	0.011
−2	**−1**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.002	0.004
−2	**−2**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
−2	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
−3	**3**	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
−3	**2**	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001
−3	**1**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001
−3	**0**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001
−3	**−1**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
−3	**−2**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
−3	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

k1/k2		0 .0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
3	**3**	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999	0.999	0.999	1.000
3	**2**	0.977	0.977	0.977	0.977	0.977	0.977	0.978	0.978	0.978	0.978	0.978
3	**1**	0.841	0.841	0.841	0.841	0.841	0.842	0.842	0.842	0.842	0.842	0.842
3	**0**	0.499	0.499	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
3	**−1**	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158
3	**−2**	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022
3	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	**3**	0.977	0.977	0.977	0.977	0.977	0.977	0.978	0.978	0.978	0.978	0.978
2	**2**	0.956	0.956	0.957	0.957	0.958	0.960	0.961	0.963	0.965	0.969	0.977
2	**1**	0.823	0.824	0.826	0.828	0.830	0.832	0.835	0.838	0.840	0.842	0.842
2	**0**	0.489	0.491	0.493	0.495	0.497	0.498	0.499	0.500	0.500	0.500	0.500
2	**−1**	0.154	0.156	0.157	0.157	0.158	0.158	0.158	0.158	0.158	0.158	0.158
2	**−2**	0.021	0.021	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022
2	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	**3**	0.841	0.841	0.841	0.841	0.842	0.842	0.842	0.842	0.842	0.842	0.842
1	**2**	0.823	0.824	0.826	0.828	0.830	0.833	0.835	0.838	0.840	0.842	0.842
1	**1**	0.708	0.714	0.721	0.729	0.737	0.746	0.756	0.767	0.781	0.799	0.836
1	**0**	0.421	0.430	0.440	0.450	0.459	0.469	0.478	0.487	0.494	0.499	0.500
1	**−1**	0.133	0.138	0.143	0.148	0.151	0.154	0.156	0.158	0.158	0.158	0.158
1	**−2**	0.018	0.019	0.020	0.021	0.021	0.022	0.022	0.022	0.022	0.022	0.022
1	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0	**3**	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
0	**2**	0.489	0.491	0.493	0.495	0.497	0.498	0.499	0.500	0.500	0.500	0.500
0	**1**	0.421	0.430	0.440	0.450	0.459	0.469	0.478	0.487	0.494	0.499	0.500
0	**0**	0.250	0.266	0.282	0.298	0.315	0.333	0.352	0.373	0.397	0.428	0.490
0	**−1**	0.079	0.089	0.098	0.108	0.117	0.127	0.136	0.145	0.152	0.157	0.158

(continued)

Table 19.2 (continued)

0	**−2**	0.011	0.013	0.015	0.017	0.018	0.020	0.021	0.021	0.022	0.022	0.022
0	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
−1	**3**	0.159	0.159	0.159	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158
−1	**2**	0.155	0.156	0.157	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158
−1	**1**	0.134	0.139	0.144	0.148	0.152	0.154	0.156	0.157	0.158	0.158	0.158
−1	**0**	0.079	0.089	0.098	0.108	0.117	0.127	0.136	0.145	0.152	0.157	0.158
−1	**−1**	0.025	0.031	0.038	0.045	0.053	0.062	0.072	0.083	0.097	0.115	0.152
−1	**−2**	0.003	0.005	0.006	0.008	0.010	0.012	0.015	0.017	0.020	0.021	0.022
−1	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
−2	**3**	0.023	0.023	0.023	0.023	0.023	0.022	0.022	0.022	0.022	0.022	0.022
−2	**2**	0.022	0.022	0.023	0.023	0.023	0.022	0.022	0.022	0.022	0.022	0.022
−2	**1**	0.019	0.020	0.021	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022
−2	**0**	0.011	0.013	0.016	0.017	0.019	0.020	0.021	0.022	0.022	0.022	0.022
−2	**−1**	0.004	0.005	0.007	0.009	0.011	0.013	0.015	0.018	0.020	0.021	0.022
−2	**−2**	0.000	0.001	0.001	0.002	0.003	0.004	0.005	0.007	0.009	0.012	0.020
−2	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
−3	**3**	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
−3	**2**	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
−3	**1**	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
−3	**0**	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
−3	**−1**	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
−3	**−2**	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.000
−3	**−3**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

19.10 Statistical Tables

Table 19.1 lists the cumulative probability, $F(k_1,k_2)$, for selective values of k_1 and k_2, and for $\rho = -0.8$ and $\rho = 0.8$. Table 19.2 lists the cumulative probability, $F(k_1, k_2)$ for selective values of k_1 and k_2 and for ρ ranging from -1.0 to 1.0.

Example 19.1 An analyst has bivariate data of $(x_1,x_2) \sim BVN(5, 10, 1, 2, 0.8)$ and wants to find the probability of x_1 less or equal to 6, and x_2 less or equal to 13. For clarity and also simplicity, the cumulative probability for variables x_1, x_2, is listed as: $F_x(6,13)$. The counterpart cumulative probability for variables z_1, z_2 becomes: $F_z(1.0,1.5)$ where,

$$z_1 = (6.0 - 5.0)/1.0 = 1.0$$
$$z_2 = (13.0 - 10.0)/2.0 = 1.5$$

Using Table 19.1, the probability sought becomes:

$$F_x(6, 13) = F_z(1.0, 1.5) = 0.83$$

Example 19.2 Applying the same bivariate normal from Example 19.1, assume now that the analyst is seeking the probability of x_1 between 5 and 6, and x_2 between 10 and 13. Using Table 19.1 and the same notation as in the prior example, the probability is obtained as shown below:

$$\begin{aligned} &P(5 \le x_1 \le 6 \cap 10 \le x_2 \le 13) \\ &= F_x(6, 13) - F_x(6, 10) - F_x(5, 13) + F_x(5, 10) \\ &= F_z(1,0, 1,5) - F_z(1.0, 0.0) - F_z(0.0, 1.5) + F_z(0.0, 0.0) \\ &= 0.83 - 0.49 - 0.50 + 0.40 \\ &= 0.24 \end{aligned}$$

19.11 Summary

The bivariate normal distribution has two variables, x_1, x_2, that are jointly related, and is defined with five parameters. The marginal and conditional distributions with their mean and variances are described. A related set of variables, z_1, z_2, are converted from x_1, x_2, and are easier to apply in the computations. They are related by the bivariate standard normal distribution, and their corresponding marginal and conditional distributions are also described. An approximation method is developed to compute the joint cumulative probability of the variables, and some table listings are presented. Examples are presented to illustrate the computations.

Chapter 20
Bivariate Lognormal

20.1 Introduction

The author in [Thomopoulos and Longinow (1984). p 3045–3049] showed how to compute the cumulative probability for the bivariate lognormal distribution in a structural engineering reliability problem. The bivariate lognormal distribution with variables x_1, x_2 appears at first to be difficult to maneuver, but by taking the natural log of each of the two variables, the bivariate normal distribution emerges and this distribution is easier to handle. The five parameters of the bivariate normal distribution become the parameters to the bivariate lognormal distribution. The chapter shows how to convert the parameters from the bivariate lognormal to the bivariate normal and vice versa. Shown also is how to compute the correlation for the bivariate lognormal pair, and how to compute the joint probability of any pair (x_1, x_2). An example is given to aid the reader in the methodology.

20.2 Fundamentals

When variables x_1, x_2 are jointly related by the bivariate lognormal distribution, they have five parameters: $(\mu_{y1}, \mu_{y2}, \sigma_{y1}, \sigma_{y2}, \rho_y)$. The designation of the variables is listed below.

$$(x_1, x_2) \sim \text{BVLN}\left(\mu_{y1}, \mu_{y2}, \sigma_{y1}, \sigma_{y2}, \sigma_y\right)$$

The marginal distributions on the variables are lognormal, with designations below:

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5_20

$$x_1 \sim \mathrm{LN}\left(\left(\mu_{y1}, \sigma_{y1}\right)\right.$$
$$x_2 \sim \mathrm{LN}\left(\left(\mu_{y2}, \sigma_{y2}\right)\right.$$

The mean and variance parameters to the above are from the corresponding normal distributions of $y_1 = \ln(x_1)$ and $y_2 = \ln(x_2)$ where ln is the natural log and their designations are the following:

$$y_1 \sim \mathrm{N}\left(\mu_{y1}, {\sigma_{y1}}^2\right)$$
$$y_2 \sim \mathrm{N}\left(\mu_{y2}, {\sigma_{y2}}^2\right)$$

The correlation between y_1 and y_2 is ρ_y, and the two variables are bivariate normal and have the following designation:

$$(y_1, y_2) \sim \mathrm{BVN}\left(\mu_{y1}, \mu_{y2}, \sigma_{y1}, \sigma_{y2}, \rho_y\right)$$

The mean and variance of each x can be converted from the same on y as shown below:

$$\mu_x = \exp\left[\mu_y + {\sigma_y}^2/2\right] \qquad = \text{mean of x}$$
$$\sigma_x{}^2 = \exp\left[2\mu_y + {\sigma_y}^2\right]\left[\exp\left({\sigma_y}^2\right) - 1\right] \qquad = \text{variance of x}$$

In the same way, the mean and variance on each variable y can be obtained from the same on x in the following way:

$$\mu_y = \ln\left[{\mu_x}^2/\sqrt{\mu_x^2 + \sigma_x^2}\right] \qquad = \text{mean of y}$$
$$\sigma_y{}^2 = \ln\left[1 + {\sigma_x}^2/{\mu_x}^2\right] \qquad = \text{variance of y}$$

The covariance and correlation between normal variables y_1 and y_2 are listed below.

$$\sigma_{y1y2} = \mathrm{E}(y_1 y_2) - \mathrm{E}(y_1)\mathrm{E}(y_2)$$
$$\rho_{y1y2} = \sigma_{y1y2}/\sigma_{y1}\sigma_{y2}$$

The correlation between the lognormal variables x_1 and x_2 is obtained as follows::

$$\rho_{x1x2} = \left[\exp\left(\sigma_{y1y2}\right) - 1/\left\{\left[\exp\left({\sigma_{y1}}^2 - 1\right)\right]\left[\exp\left({\sigma_{y2}}^2 - 1\right)\right]\right\}^{0.5}\right.$$

20.3 Cumulative Probability

For clarity in listing the cumulative probability function, the following designations are in use here:

$F_x(x_1, x_2)$ is the cumulative probability for lognormal variables x.

$F_y(y_1, y_2)$ is the cumulative probability for normal variables y.

$F_z(z_1, z_2)$ is the cumulative probability for standard normal variables z.

The cumulative probability function of two values x_{1o} of x_1, and x_{2o} of x_2 is denoted as below:

$$F_x(x_{1o}, x_{2o}) = P(x_1 \leq x_{1o} \cap x_2 \leq x_{2o})$$

To obtain the above probability, the four steps are followed:

1. Convert the lognormal values of x to the corresponding normal values of y:

$$y_{1o} = \ln(x_{1o})$$
$$y_{2o} = \ln(x_{2o})$$

2. Change the normal values of y to the associate standard normal values of k:

$$k_1 = \left[y_{1o} - \mu_{y1} \right] / \sigma_{y1}$$
$$k_2 = \left[y_{2o} - \mu_{y2} \right] / \sigma_{y2}$$

3. Using the bivariate standard normal distribution, with ρ_y, compute $F_z(k_1,k_2)$ as shown in Chap. 19.
4. Finally,

$$F_x(x_{1o}, x_{2o}) = F_z(k_1, k_2)$$

Example 20.1 An analyst has bivariate lognormal data with $(x_1,x_2) \sim$ BVLN(2, 4, 1, 2, 0.8). For this scenario, Table 20.1 is generated listing the cumulative probabilities with selective values of k_1, k_2 and x_1, x_2 when $\rho_y = 0.8$. The table probabilities are taken from Table 19.1 of Chap. 19. The values for x_1 and x_2 are obtained in the following way:

$$x_1 = \exp(2.0 + k_1 \times 1.0)$$
$$x_2 = \exp(4.0 + k_2 \times 2.0)$$

Table 20.1 $F_x(x_1, x_2)$ for selected values of $(x_1,x_2) \sim$ BVLN(2.0, 4.0, 1.0, 2.0, 0.8), and $F_z(k_1,k_2)$ for values of $(k_1,k_2) \sim$ BVN(0, 0, 1, 1, 0.8)

$\rho = 0.8$														
k2/k1	−3	−2.5	−2	−1.5	−1	−0.5	0	0.5	1	1.5	2	2.5	3	×2
3	0	0.01	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98	0.99	1.00	22,026
2.5	0	0.01	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98	0.99	0.99	8103
2	0	0.01	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.96	0.98	0.98	2980
1.5	0	0.01	0.02	0.07	0.16	0.31	0.50	0.69	0.83	0.90	0.93	0.93	0.93	1096
1	0	0.01	0.02	0.07	0.16	0.31	0.49	0.67	0.78	0.83	0.84	0.84	0.84	403
0.5	0	0.01	0.02	0.07	0.16	0.30	0.47	0.60	0.67	0.69	0.69	0.69	0.69	148
0	0	0.01	0.02	0.07	0.15	0.28	0.40	0.47	0.49	0.50	0.50	0.50	0.50	55
−0.5	0	0.01	0.02	0.06	0.13	0.22	0.28	0.30	0.31	0.31	0.31	0.31	0.31	20
−1	0	0.01	0.02	0.05	0.10	0.13	0.15	0.16	0.16	0.16	0.16	0.16	0.16	7
−1.5	0	0	0.02	0.03	0.05	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.07	3
−2	0	0	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	1
−2.5	0	0	0	0	0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.4
−3	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
×1	0.4	0.6	1	2	3	4	7	12	20	33	55	90	148	

Assume the analyst now wants to find the cumulative joint probability when $x_1 = 30$ and $x_2 = 500$, i.e., $F_x(30, 500)$. To obtain, the following four steps are taken:

1. Convert the lognormal x values to normal values as below:

$$y_{1o} = \ln(30) = 3.40$$
$$y_{2o} = \ln(500) = 6.21$$

2. Covert the normal values to standard normal values as follows:

$$k_1 = (3.40 - 2.00)/1.00 = 1.40$$
$$k_2 = (6.21 - 4.00)/2.00 = 1.10$$

3. Using Table 20.1, and applying some interpolation, find the cumulative probability sought as below:

$$F_z(1.40, 1.1) \approx 0.82$$

4. Finally,

$$F_x(30, 500) = F_y(3.40, 6.21) = F_z(1.40, 1.1) \approx 0.82$$

20.4 Summary

The bivariate lognormal has two variables x_1, x_2 and the marginal distributions are lognormal. When the variables are converted by taking the natural log of each, the bivariate normal distribution emerges. The five parameters of the latter distribution become the parameters to the bivariate lognormal distribution. The way to derive the joint probability of a pair of bivariate lognormal variables is by computing the joint probability of the counterpart pair from the bivariate normal variables.

References

Abramowitz, M. & Stegun, I. (1964), *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C.

Aitchison, J. & Brown, J. (1969). *The Lognormal Distribution*. Cambridge: Cambridge University Press.

Beyer, W.H. (1968). *Handbook of Tables for Probability and Statistics*. Chemical Rubber Co.

Broadbent, S. (1956). *Lognormal Approximation to Products and Quotients*. 43, 404-417: Biometrika.

Brown, R.G. (1959). *Smoothing, Forecasting and Prediction of Discrete Time Series*. Engelwood Cliffs, N.J.: Prentice Hall.

Crow, E. & Shinizu, K. (1988). *Lognormal Distribution: Theory and Applications*, New York: Marcel Dekker.

Galton, F. (1909). *Memories of My Life*. New York: E.P. Dutton & Co.

Hasting, N.A.J. & Peacock, J.B. (1974). *Statistical Distributions*. New York: Wiley & Sons

Hines, W.W., Montgomery, L.D.C., Goldsman, D.M., & Burror, C.M. (2003). *Probability and Statistics for Engineers*. New York: Wiley & Sons.

Jantaravareerat, M. (1998). *Approximation of the Distribution Function for the Standard Bivariate Normal*. Doctoral Dissertation, Stuart School of Business: Illinois Institute of Technology.

Jantaravareerat, M. & Thomopoulos N. (1998). *Some New Tables on the Standard Normal Distribution*. 30, pp179-184: Computing Science and Statistics.

Johnson, A.C. (2001). *On The Truncated Normal Distribution*. Doctoral Dissertation. Stuart School of Business: Illinois Institute of Technology.

Johnson A.C., & Thomopoulos, N.T. (2002). *Characteristics and Tables of the Left-truncated Normal Distribution*. Proceedings of the Midwest Decision Sciences Institute, 133-139.

Law, A. & Kelton, W. (2000). *Simulation, Modeling and Analysis*. Boston: McGraw Hill.

Lindee, C., (2001). *The Multivariate Standard Normal Distribution*. Doctoral Dissertation. Stuart School of Business: Illinois Institute of Technology

Lindee, C. & Thomopoulos, N. (2001). *Values for the Cumulative Distribution Function of the Standard Multivariate Normal Distribution*. 48, pp 600-609: Proceedings of the Midwest Decision Sciences Institute.

Robson, D.S., & Whitlock, J.W. (1964). *Estimation of a Truncation Point*, 51, pp 33-39: Biometrical.

Schneider, J. (1986). *Truncated and Censored Samples from Normal Populations*. New York: Marcel Dekker.

Thomopoulos, N.T. (1980). *Applied Forecasting Methods*. Englewood Cliffs, N.J.: Prentice Hall.

Thomopoulos, N.T. (2013). *Essentials of Monte Carlo Simulation*. New York: Springer.

N.T. Thomopoulos, *Statistical Distributions*, DOI 10.1007/978-3-319-65112-5

Thomopoulos, N.T. (2016). Demand Forecasting for Inventory Control. New York: Springer.
Thomopoulos, N.T., & Longinow, A.C. (1984). *Bivariate Lognormal Probability Distribution*. 110: Journal of Structural Engineering.
Thomopoulos, N.T. & Johnson, A.C. (2003). *Tables and Characteristics of the Standardized Lognormal Distribution*. 103, pp 1-6: Proceeding of the Decision Sciences Institute
Thomopoulos, N.T., & Johnson, A.C. (2004*). Some Measures on the Standard Bivariate Lognormal Distribution*; pp 1721-1726: Proceedings of the Decision Sciences Institute.
Zanakis, S.H. (1979). A simulation study of some simple estimators for the three parameter Weibull distribution. J. Stat Comput. Simul.