

Empirical Assessment of Normal and Lognormal Distribution Assumptions

Table of Contents

- Introduction
- Illustrating the Problem
- Statistical Assumptions and Their Implications
- Practical Methods to Verify Normal Assumptions
- Summary
- Bibliography
- About the Author
- Other START Sheets Available

Introduction

This START sheet discusses some empirical and practical methods for checking and verifying the statistical assumptions of Normal and Lognormal distributions. It presents several numerical and graphical examples showing how this is done with data. The Normal and Lognormal distributions are addressed together because they are intimately related. A random variable (e.g., life of a device) follows the Lognormal distribution if the natural logarithm (base e) of this variable (e.g., Log of the device life) follows a Normal distribution. Hence, by dealing with one distribution, we are also dealing with the other.

It is important to correctly assess statistical distributions. Most parametric statistical methods assume an underlying distribution in the derivation of their results (methods that do not assume an underlying distribution are called non-parametric or distribution-free, and will be the topic of a separate START sheet). The consequences of specifying the wrong distribution may prove very costly. If such distribution does not hold, then the confidence levels of the confidence intervals (or of hypotheses tests) may be completely off. To avoid such problem, distribution assumptions must be carefully checked. Presenting practical methods for doing so for the Normal and Lognormal distributions is the objective of this START sheet.

Two approaches can be used to check distribution assumptions. One is by implementing a Goodness of Fit (GoF) test such as the Chi Square, Anderson Darling or Kolmogorov-Smirnov. These are numerically convoluted theoretical tests, based on statistical theory and usually requiring lengthy calculations. In turn, these calculations ultimately require the use of specialized software, not always readily available. On the other hand, there also exist practical procedures, easy to understand and

implement, based on intuitive and graphical properties of a distribution. These properties can be used to check and assess these distribution assumptions. This START sheet demonstrates the implementation and interpretation of such assessment procedures for the Normal and the Lognormal, two distributions widely used in quality and reliability.

Illustrating the Problem

Assume we must estimate the mean life of a device using a confidence interval (CI). This estimator consists of two values, the CI upper and lower limits, such that the unknown mean life μ will be in this range, with the prescribed coverage probability $(1-\alpha)$. For example, we say that the life of a device is between 90 and 110 hours with probability 0.95 (or that there is a 95% chance that the interval 90 to 110, covers the device's true mean life μ). The actual coverage of such a CI essentially depends on the assumptions of the statistical model used.

Every parametric statistical procedure is based on distribution assumptions that must be met for it to hold true. For the sake of illustration, let's consider that we want to establish a valid CI for the mean life of a device. To do this, we need, in addition to a set of good data, to assume an underlying statistical distribution (e.g., Normal, Lognormal, and Weibull) that actually fits these data and problem.

First, we will assume that the distribution of the lives (times to failure) of a device is Normal (Figure 1), and second, we will assume it is Cauchy (Figure 2). Figures 1 and 2 were obtained from 5000 data points from each of these two distributions, having the same mean = 20. The Normal has a Standard Deviation $\sigma = 7.6$ and the Cauchy has a scale parameter $\alpha = 5$.

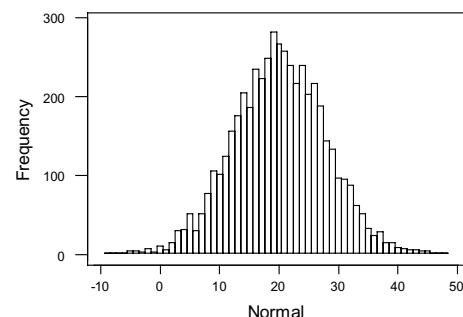


Figure 1. Normal ($\mu = 20$, $\sigma = 7.6$)

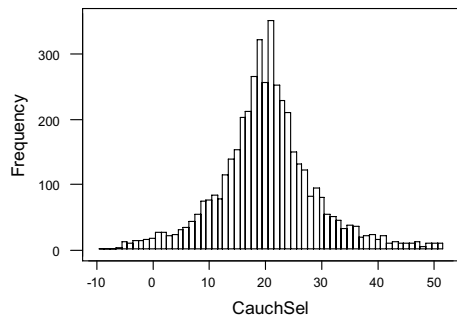


Figure 2. Cauchy ($\mu = 20, \alpha = 5$)

There are some practical connotations of belonging to one or the other, of these two distributions. The Normal values are symmetric about 20 and concentrated in the range of -2.8 to 42.8 (for, three standard deviations on each side of the mean, comprises 99% of the population). Cauchy values, on the other hand, are similar to the Normal in the centered 50% of the distribution. Notice how the respective interquartile ranges (between Q1, the 25th percentile and Q3, the 75th percentile) are very close. For example, IQR lower quartiles Q1 = 14.8 and 14.9 and upper quartiles, Q3 = 25.1 and 25.3, differ in 0.02 or less. This proximity becomes even more evident when analyzing the differences between the corresponding sorted Normal and Cauchy data points (data set Diff).

The Cauchy distribution however, departs from the Normal in the ranges below Q1 and above Q3, achieving large maximum (21607.0) and minimum (-27796.8) values. Such extremes are highly improbable in a Normal ($\mu = 20, \sigma = 7.6$). Hence, to obtain comparable graphs (Figures 1 and 2), 300 of the higher and lower original Cauchy values (CauchSel) were discarded. Descriptive statistics of the data sets are presented in Table 1.

Table 1. Descriptive Statistics for the Data Sets

Statistics	Distribution			
	Cauchy	Normal	CauchSel	Diff
N	5000	5000	4466	5000
Mean	17.2	20.087	20.323	-2.9
Median	20.0	20.046	20.137	-0.0
Std. Dev.	518.8	7.6	8.6	517.6
Minimum	-27796.8	-8.52	-5.59	-27788.3
Maximum	21607.0	47.58	50.97	21559.4
Q1	14.8	14.9	15.8	-0.5
Q3	25.1	25.3	24.4	0.2

The Cauchy distribution does not have a variance (meaning it is infinite). This allows the inclusion of unusually large and small values in the sample, which can seriously bias the estimates. The present example illustrates how, by incorrectly specifying a Cauchy distribution as Normal, we can commit serious errors in parameter confidence interval (CI) estimation.

To further illustrate this situation, we select a small sample of $n = 10$ devices, from a Cauchy (20, 5) distribution (Table 2) and obtain a 95% CI for the mean, under the (wrong) assumption that the sample comes from a Normal (20, 7.6)

Table 2. Small Sample Data Set [from Cauchy (20, 5)]

36.31	21.65	22.41	-0.29	26.13
54.59	22.59	3.590	11.53	23.66

The descriptive statistics yield a sample average $\bar{x} = -370$, a sample median = 23 and a sample variance $s^2 = 1247^2$. From these results, we can conclude the following:

1. If the original device lives are assumed distributed Normal (with known $\sigma = 7.6$), the 95% CI for the mean life m , based on the Normal distribution is (-374.75, -365.32).
2. If the data are assumed Normal with variance unknown and estimated from this small ($n = 10$) sample, the resulting statistical distribution follows a Student-t with $n - 1 = 9$ degrees of freedom. The corresponding CI obtained for the population mean is then (-1262, 522).
3. Finally, if the device lives are assumed Cauchy, a 95% CI for the mean life cannot be obtained. For this is a small ($n = 10$) sample and the Cauchy has an infinite variance. Our example includes a large outlier (-3918.92) whose net effect is to bias the CI.

Since in reality the 10 data points come from the Cauchy distribution, the Normal CI is incorrect and its coverage probability (95%) is inaccurate. Had we erroneously assumed Normality for this small sample, the CI true coverage probability (confidence) would be unknown and policies, derived under such unknown probability, are at risk.

Statistical Assumptions and Their Implications

Every statistical model has its own “assumptions” that have to be verified and met, to provide valid results. For example, deriving the small sample Student-t CI for the mean life of a device requires two “assumptions:” that the device lives are independent and Normally distributed. These two statistical assumptions must be met (and verified) in order for such CI to cover the true mean with the prescribed probability. If, as in data in Table 2, the data does not follow the assumed distribution, the CI is invalid and its coverage (confidence) may be different from the one prescribed. Fortunately, distribution model assumptions are associated with very practical and useful implications - the Normal and Lognormal distributions are no exceptions.

In practice, the assumption that the distribution of the lives of a device is Normal means that there are many, independent factors that are contributing to the final result. An analogy is the intelligence quotient (IQ). A human is the product of his or her socioeconomic level, upbringing, schooling, nutrition, inherited genes, health, etc. All these factors contribute to human intelligence. For this reason, IQ is usually Normally distributed, as are also (and for the same reason) height, weight, etc.

In addition, the Normal distribution has several specific characteristics. It is continuous, symmetric (mean = median = mode) and standardizable (by subtracting the mean and dividing by the

standard deviation, we always obtain a unique Normal with mean zero and variance unit). Finally, the ranges defined by one, two, and three standard deviations above and below the mean cover 68%, 95%, and 99% of the population, respectively.

In what follows, we will use these statistical Normal distribution properties and its implications, to check and empirically validate the Normal assumptions of our data.

Practical Methods to Verify Normal Assumptions

In this section we discuss several empirical and practical methods for assessing the validity of two important and widely used distributions: the Normal and Lognormal. We illustrate this validation process via the life test data in Table 3. This sample (n = 45) was taken from the Normal (20, 7.6) process that generated Figure 1, presented in Section 2.

Table 3. Large Sample Life Data Set (sorted)

6.1448	6.6921	6.7158	7.7342	9.6818	12.3317
12.5535	13.0973	13.6704	14.0077	14.7975	15.3237
15.5832	15.7808	15.7851	16.2981	16.3317	16.8147
16.8860	17.5166	17.5449	17.9186	18.5573	18.8098
19.2541	19.5172	19.7322	21.9602	23.2046	23.2625
23.7064	23.9296	24.8702	25.2669	26.1908	26.9989
27.4122	27.7297	28.0116	28.2206	28.5598	29.5209
30.0080	31.2306	32.5446			

In our data set, two distribution assumptions need to be verified or assessed: (1) that the data are independent and (2) that they are identically distributed as a Normal.

The assumption of independence implies that randomization (sampling) of the population of devices (and other influencing factors) must be performed before placing them on test. For example, device operators, times of operations, weather conditions, location of the devices in warehouses, etc. should be randomly selected so they become representative of these same characteristics and of the contexts in which devices will normally operate.

To assess the Normality of the data, we use informal methods, based on the properties of the Normal distribution. They seem appropriate for the practical engineer, since they are largely intuitive and easy to implement.

To assess data, we must first obtain their descriptive statistics (Table 4). Then, we analyze and plot the raw data in several ways, to check (empirically but efficiently) if the Normality assumption holds.

There are a number of useful and easy to implement procedures, based on well-known statistical properties of the Normal distribution, which help us to informally assess this assumption. These properties are summarized in Table 5.

Table 4. Descriptive Statistics of Data in Table 3

Statistics	Normal Sample
N	45
Mean	19.50
Median	18.56
Std. Dev.	7.05
Minimum	6.14
Maximum	32.54
Q1	15.06
Q3	25.73

Where Mean is the average of the data and the Standard Deviation is the square root of:

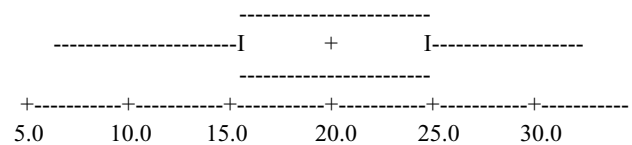
$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Table 5. Some Properties of the Normal Distribution

1. Mean, median, and mode coincide; hence, sample values should also be close.
2. Graphs should suggest that the distribution is symmetric about the mean.
3. About 70% of the data should be within one standard deviation of the mean.
4. About 95% of the data should be within two standard deviations of the mean.
5. About 1% of the data, should be beyond three standard deviations of the mean.
6. Plots of the Normal probability and Normal scores should be close to linear.
7. Regressions of these probability and score plots should yield Unit slope.

First, from the descriptive statistics in Table 4, we observe that the sample Mean (19.5) and Median (18.56) are close, and how the Standard Deviation is 7.05. This supports the Normality of the distribution by Property No. 1, in Table 5.

The distribution looks symmetric about mean = 19.5, as suggested by the following Box Plot (plot of minimum, Q1, median, Q3, and maximum). Observe how the centered 50% of the data (between Q1 = 15.06 and Q3 = 25.73) is dispersed about the mean.



The histogram (Figure 3) also suggests some symmetry about Mode = 18 (center of the interval with the highest frequency in Figure 3). All of which, by Property No. 2 in Table 5, suggests the validity of the Normal distribution.

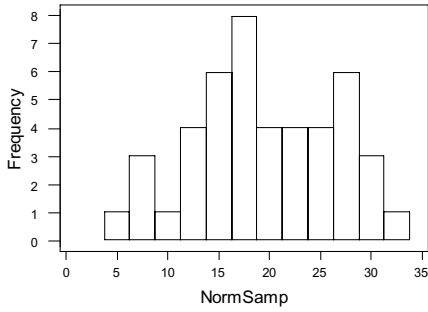


Figure 3. Histogram of the Normal Data Set (Mode is 18)

The interval defined by one standard deviation about the mean: $(\mu - \sigma, \mu + \sigma) = (19.5 - 7.05, 19.5 + 7.05) = (12.4, 26.1)$ includes 28 values (in ranks 7 to 34, of sorted Table 3) representing 62% of the total data set (close to the expected 68.25%). The interval $(\mu - 2\sigma, \mu + 2\sigma) = (5.4, 33.6)$ includes values in ranks 1 to 45 (i.e., all data) representing 100% of the data set (close to the expected 95%). There are zero values beyond $\mu \pm 3\sigma$, supporting the statement that about 1 point (about 1% of the values) would be outside the interval $(\mu - 3\sigma, \mu + 3\sigma)$. All these results support Properties 3 to 5 of Table 5.

In the Probability Plot, the Normal Probability (P_i) is plotted vs. $I/(n + 1)$ where I is the data sequence order, i.e., $I = 1, \dots, 45$. Each P_i is obtained by calculating the Normal probability of the corresponding failure data, X_i using the sample mean (19.5) and the standard deviation (7.05). For example, the first ($I = 1$) sorted (smallest) data point is 6.15:

$$P_{19.5,7.05}(6.15) = \text{Normal}\left(\frac{6.15 - 19.5}{7.05}\right) = \text{Normal}(-1.89) = 0.029$$

The data point is then plotted against the corresponding $I/(n + 1)$ value, $1/46 = 0.0217$ and so on, until done with all sorted sample elements $I = 1, \dots, 45$.

When the population is Normal, the Probability Plot (Figure 4) follows an upward linear trend, with unit slope. Hence, the linear regression of the Normal Probability vs. Data Rank must also reflect this one-to-one relation, via achieving a unit slope:

Predictor	Coef	Std. Dev.	T	P
Constant	-0.02282	0.01192	-1.91	0.062
NormRank	1.00783	0.02076	48.54	0.000
S = 0.03933		R-Sq = 98.2%		R-Sq(adj) = 98.2%

The regression Index of Fit ($R^2 = 98.2\%$) is very high (close to 100%). Also, the P-value (0.0) for the NormRank regression coefficient T-Test (48.54) is very small, thus suggesting a linear trend. The regression coefficient (slope) itself (1.00783) is close to Unit, suggesting the Normal as the data statistical distribution.

This regression slope Unit value serves as the “litmus test” of this graphical approach to assess Normality.

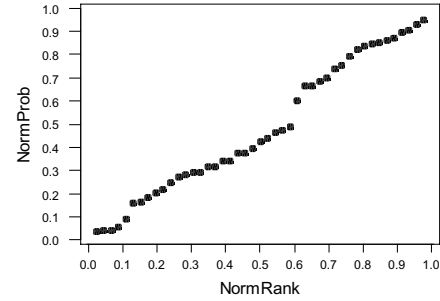


Figure 4. Plot of Normal Probability (P_i) vs. $I/(n + 1)$; $I = 1, \dots, n$; Close to Linear, as Expected from a Normal

The Normal scores X_i are the percentiles corresponding to the values $I/(n + 1)$, for $I = 1, \dots, n$; calculated under the Normal distribution (using mean = 19.5, std-dev = 7.05). For our example, the first $I/(n + 1)$ is $1/46 = 0.0217$ and the smallest data point = 6.15:

$$P_{19.5,7.05}(X_i) = \text{Normal}\left(\frac{X_i - 19.5}{7.05}\right) \approx \frac{i}{n+1} = 0.0217$$

$$\Rightarrow \text{Percentile}(0.0217) = -2.02 = \frac{X_i - 19.5}{7.05}$$

Solving in the above equation for scores X_i we get the first ($I = 1$) Normal score:

$$X_1 = -2.02 \times 7.05 + 19.5 = -14.24 + 19.5 = 5.26$$

These Normal scores are then plotted vs. their corresponding sorted data values (Figure 5). In the above example, score 5.26 is plotted against 6.15 (the smallest data point) and so on, for $I = 1, \dots, n$. When the data come from a Normal Distribution, the Normal Scores plot is close to a straight line (Property 6).

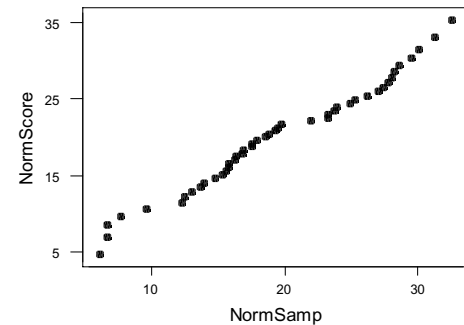


Figure 5. Plot of the Normal Scores vs. the Sorted Real Data Values, Close to Linear

We regress the Normal Scores vs. the corresponding data. The regression, if the data comes from the Normal distribution, should yield a unit slope:

$$\text{NormScore} = 0.487 + 1.00 \text{ NormSamp}$$

Predictor	Coef	Std. Dev.	T	P
Constant	0.4872	0.4554	1.07	0.291
NormSamp	1.00042	0.02199	45.50	0.000

S = 1.028 R-Sq = 98.0% R-Sq(adj) = 97.9%

An Index of Fit $R^2 = 97.9\%$ and a regression coefficient 1.0042, plus the Normal Probability and Normal Scores plots, suggest that the assumption of a Normal distribution is reasonable.

Assessing the Lognormal Distribution

The Lognormal distribution is widely used in reliability studies. Consequently, there is a strong interest in assessing whether a data set comes from such distribution. The results in the previous sections show how this is now easy to do. When a random variable (e.g., device life) is distributed Lognormal, the Logarithm (base e) of the random variable (e.g., Log life) is distributed Normal. This property carries on to data sets. When a data set comes from a Lognormal population, then the Logarithm of these data are distributed as a Normal.

In practice, to assess the Lognormality of a data set, we take the Logarithms of the original data and assess the Normality of the transformed data, as done in the sections above. For example, the data set in Table 6 comes from a Lognormal distribution with Location parameter 4 and Scale parameter 0.4.

Table 6. Original Data (Lognormal)

67.842	91.030	42.974	42.849	46.459	64.746
55.031	38.326	119.612	62.903	31.778	87.068
58.854	44.790	69.054	69.222	39.334	121.592
90.537	99.651	93.440	31.021	47.152	63.716
92.824	36.030	104.526	62.006	35.605	35.019
32.102	24.288	80.420	132.861	48.886	57.911
79.527	37.659	63.223	110.359	77.153	84.713
52.391	42.475	65.333			

We then transform these data by taking Logarithms of each element. For the first data point: $\text{Log}(67.842) = 4.21718$. The transformed data are shown in Table 7.

Table 7. Transformed Data (LogELN)

4.21718	4.51119	3.76060	3.75769	3.83858	4.17048
4.00790	3.64612	4.78425	4.14159	3.45877	4.46668
4.07506	3.80198	4.23489	4.23732	3.67210	4.80067
4.50576	4.60168	4.53732	3.43466	3.85338	4.15443
4.53071	3.58434	4.64944	4.12723	3.57250	3.55589
3.46892	3.18999	4.38726	4.88930	3.88949	4.05891
4.37610	3.62857	4.14667	4.70374	4.34579	4.43927
3.95874	3.74891	4.17949			

Descriptive statistics for both the Lognormal data set and its Logarithmic transformation are given in Table 8. For example, results for the sample Mean: $\text{Ln}(65.21) = 4.177 \approx 4.091$. We now assess the Normality of the transformed data set (LogELN) by repeating the work discussed in the previous sections. If these transformed data fulfill the Properties given in Table 5, then the original data (Table 6) are distributed Lognormal.

Table 8. Descriptive Statistics for the Data and their Normal Transformation

Statistics	Lognormal	LogELN
N	45	45
Mean	65.21	4.0911
Median	62.90	4.1416
StDev	27.36	0.4244
Min	24.29	3.190
Max	132.86	4.889
Q1	42.66	3.753
Q3	85.89	4.453

These empirical results help assess the plausibility of the Normality or the Lognormality assumptions of a given life data set. If, at such point, a stronger case for the validity of these distributions is required, then a number of theoretical GoF tests can be carried out.

Summary

This START sheet discusses the important problem of (empirically) assessing both the Normal and Lognormal distribution assumptions of a data set. Several numerical and graphical examples were presented and some related theoretical and practical issues were discussed.

Some other, very important, reliability analysis topics were mentioned. Due to their complexity, these will be treated in more detail in separate, forthcoming START sheets.

Bibliography

1. Practical Statistical Tools for Reliability Engineers, Coppola, A., RAC, 1999.
2. Mechanical Applications in Reliability Engineering, Sadlon, R.J., RAC, 1993.
3. A Practical Guide to Statistical Analysis of Material Property Data, Romeu, J.L. and C. Grethlein, AMPTIAC, 2000.
4. Probability and Statistics for Engineers and Scientists (6th Edition), Walpole, R.E.; R.H. Myers, and S.L. Myers, Prentice Hall, NJ, 1998.
5. Statistical Assumptions of an Exponential Distribution, Romeu, J.L. RAC START, Volume 8, Number 2.
6. Statistical Confidence, Romeu, J.L., RAC START, Volume 9, Number 4.

About the Author

Dr. Jorge Luis Romeu has over thirty years of statistical and operations research experience in consulting, research, and teaching. He was a consultant for the petrochemical, construction, and agricultural industries. Dr. Romeu has also worked in statistical and simulation modeling and in data analysis of software and hardware reliability, software engineering and ecological problems.

Dr. Romeu has taught undergraduate and graduate statistics, operations research, and computer science in several American and foreign universities. He teaches short, intensive professional training courses. He is currently an Adjunct Professor of Statistics and Operations Research for Syracuse University and a Practicing Faculty of that school's Institute for Manufacturing Enterprises.

For his work in education and research and for his publications and presentations, Dr. Romeu has been elected Chartered Statistician Fellow of the Royal Statistical Society, Full Member of the Operations Research Society of America, and Fellow of the Institute of Statisticians.

Romeu has received several international grants and awards, including a Fulbright Senior Lectureship and a Speaker Specialist Grant from the Department of State, in Mexico. He has extensive experience in international assignments in Spain and Latin America and is fluent in Spanish, English and French.

Romeu is a senior technical advisor for reliability and advanced information technology research with Alion Science and

Technology. Since joining Alion, and its predecessor IIT Research Institute (IITRI) in 1998, Romeu has provided consulting for several statistical and operations research projects. He has written a State of the Art Report on Statistical Analysis of Materials Data, designed and taught a three-day intensive statistics course for practicing engineers, and written a series of articles on statistics and data analysis for the AMPTIAC Newsletter and RAC Journal.

Other START Sheets Available

Many Selected Topics in Assurance Related Technologies (START) sheets have been published on subjects of interest in reliability, maintainability, quality, and supportability. START sheets are available on-line in their entirety at <<http://rac.alionscience.com/rac/jsp/start/startsheet.jsp>>.

For further information on RAC START Sheets contact the:

Reliability Analysis Center
201 Mill Street
Rome, NY 13440-6916
Toll Free: (888) RAC-USER
Fax: (315) 337-9932

or visit our web site at:

<<http://rac.alionscience.com>>



About the Reliability Analysis Center

The Reliability Analysis Center is a world-wide focal point for efforts to improve the reliability, maintainability, supportability and quality of manufactured components and systems. To this end, RAC collects, analyzes, archives in computerized databases, and publishes data concerning the quality and reliability of equipments and systems, as well as the microcircuit, discrete semiconductor, electronics, and electromechanical and mechanical components that comprise them. RAC also evaluates and publishes information on engineering techniques and methods. Information is distributed through data compilations, application guides, data products and programs on computer media, public and private training courses, and consulting services. Alion, and its predecessor company IIT Research Institute, have operated the RAC continuously since its creation in 1968.