

Statistical Assumptions of an Exponential Distribution

Table of Contents

- Introduction
- Putting the Problem in Perspective
- Statistical Assumptions and their Implications
- Practical Methods to Verify Exponential 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 an Exponential distribution and presents several numerical and graphical examples showing how these methods are used. Most statistical methods (of parametric statistics) assume an underlying distribution in deriving results (methods that do not assume an underlying distribution are called non-parametric, or distribution free, and will be the topic of a separate paper).

Whenever we assume that the data follow a specific distribution we also assume risk. If the assumption is invalid, then the confidence levels of the confidence intervals (or the hypotheses tests) will be incorrect. The consequences of assuming the wrong distribution may prove very costly. The way to deal with this problem is to check distribution assumptions carefully, using the practical methods discussed in this paper.

There are two approaches to checking distribution assumptions. One is to use the Goodness of Fit (GoF) tests. These are numerically convoluted, theoretical tests such as the Chi Square, Anderson-Darling, Kolmogorov-Smirnov, etc. They are all based on complex statistical theory and usually require lengthy calculations. In turn, these calculations ultimately require the use of specialized software, not always readily available.

Alternatively, there are many practical procedures, easy to understand and implement. They are based on intuitive and graphical properties of the distribution that we wish to assess and can thus be used to check and validate these distribution assumptions. The implementation and interpretation of such procedures, for the important case of the Exponential distribution, so prevalent in quality and reliability theory and practice, are discussed in this paper.

Also addressed in this START sheet are some problems associated with checking the Exponential distribution assumption. First, a numerical example is given to illustrate the seriousness of this problem. Then, additional numerical and graphical examples are developed that illustrate how to implement such distribution checks and related problems.

Putting the Problem in Perspective

Assume that we are given the task of estimating the mean life of a device. We may provide a simple point estimator, the sample mean, which will not provide much useful information. Or we may provide a more useful estimator: the confidence interval (CI). This latter estimator consists of two values, the CI upper and lower limits, such that the unknown mean life, μ , will be in this range, with a 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 true mean life, μ).

The accuracy of CI estimators depends on the quality and quantity of the available data. However, we also need a statistical model that is consistent with and appropriate for the data. For example, to establish a CI for the Mean Life of a device we need, in addition to sufficiently good test data, to know or assume a statistical distribution (e.g., Normal, Exponential, Weibull) that actually fits these data and problem.

Every parametric statistical model is based upon certain assumptions that must be met, for it to hold true. In our discussion, and for the sake of illustration, consider only two possibilities: that the distribution of the lives (times to failure) is Normally distributed (Figure 1) and that it is

Exponentially distributed (Figure 2). The figures were obtained using 2000 data points, generated from each of these two distributions, with the same mean = 100 (and for the Normal, with a Standard Deviation of 20).

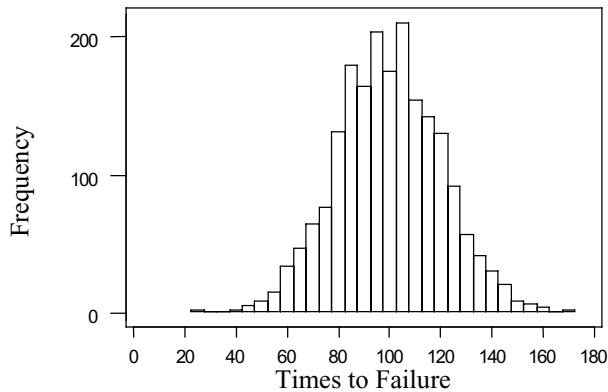


Figure 1. Normal Distribution of Times to Failure

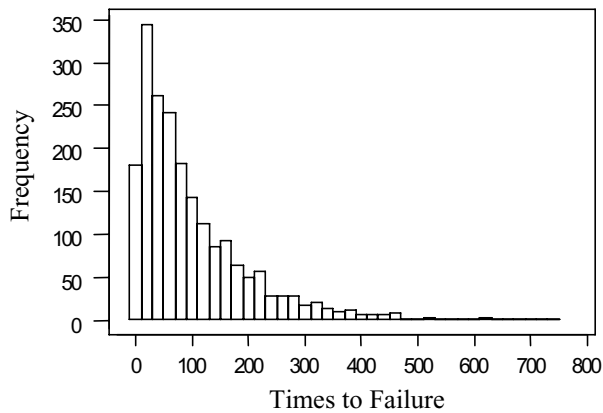


Figure 2. Exponential Distribution of Times to Failure

There are practical consequences of data fitting one or the other of these two different distributions. “Normal lives” are symmetric about 100 and concentrated in the range of 40 to 160 (three standard deviations, on each side of the mean, which comprises 99% of the population). “Exponential lives”, on the other hand, are right-skewed, with a relatively large proportion of device lives much smaller than 40 units and a small proportion of device lives larger than 200 units.

To highlight the consequences of choosing the wrong distribution, consider a sample of $n = 10$ data points (Table 1). We will obtain a 95% CI for the mean of these data, using two different distribution assumptions: Exponential and Normal.

Table 1. Small Sample Data Set

| | | | | |
|--------|---------|---------|---------|---------|
| 5.950 | 119.077 | 366.074 | 155.848 | 30.534 |
| 20.615 | 15.135 | 3.590 | 103.713 | 120.859 |

The statistic “sample average”, $\bar{x} = 94.14$, will follow a different sampling distribution, according to whether the Normal or the Exponential distributions are assumed for the population. Hence, the data will be processed twice, each time using a different “formula”. This, in turn, will produce two different CI that will exhibit different confidence probabilities.

Normal Assumption. If the original device lives are assumed distributed Normal (with $\sigma = 20$), the 95% CI for the device mean life μ , based on the Normal distribution is:

$$(81.7, 106.5)$$

Exponential Assumption. If, however, the device lives are assumed Exponential, then the 95% CI for the mean life θ , based on the Exponential, is:

$$(55.11, 196.3)$$

We leave the details of obtaining these two specific statistics or “formulas” for another paper.

Since in reality the ten data points come from the Exponential, only the CI (55.11, 196.3) is correct and its coverage probability (95%) is the one prescribed. Had we erroneously assumed Normality, the CI obtained under this assumption, for this small sample, would have been incorrect. Moreover, its true coverage probability (confidence) would be unknown and every policy, derived under such unknown probability, is at risk.

This example illustrates and underlines how important it is to establish the validity (or at least the strong plausibility) of the underlying statistical distribution of the data.

Statistical Assumptions and their Implications

Every statistical model has its own “assumptions” that have to be verified and met, to provide valid results. In the Exponential case, the CI for the mean life of a device requires two “assumptions”: that the lives of the tested devices are (1) independent, and (2) Exponentially distributed. These two statistical assumptions must be met (and verified) for the corresponding CI to cover the true mean with the prescribed probability. But if the data do not follow the assumed distribution, the CI coverage probability (or its confidence) may be totally different than the one prescribed.

Fortunately, the assumptions for all distribution models (e.g., Normal, Exponential, Weibull, Lognormal, etc.) have practical and useful implications. Hence, having some background information about a device may help us assess its life distribution.

A case in question occurs with the assumption that the distribution of the lives of a device is Exponential. An implication of the Exponential is that the device failure rate is constant. In practice, the presence of a constant failure rate may be confirmed

from observing the times between failures of a process where failures occur at random times.

In general, if we observe any process composed of events that occur at random times (say lightning strikes, coal mine accidents, earthquakes, fires, etc.), the times between these events will be Exponentially distributed. The probability of occurrence of the next event is independent of the occurrence time of the past event. As a result, phrases such as “old is as good as new” have a valid meaning. [It is important to note that although failures may occur at random times, they do not occur for “no reason.” Every failure has an underlying cause.]

In what follows, we will use statistical properties derived from Exponential distribution implications to validate the Exponential assumption.

Practical Methods to Verify Exponential Assumptions

Several empirical and practical methods can be used to establish the validity of the Exponential distribution. We will illustrate the process of validating the Exponential assumptions using the life test data in Table 2. This larger sample (n = 45) was generated following the same process used to generate the previous smaller sample (n = 10) presented in Table 1.

Table 2. Large Sample Life Data Set

| | | | | | |
|---------|---------|---------|---------|---------|---------|
| 12.411 | 58.526 | 46.684 | 49.022 | 77.084 | 7.400 |
| 21.491 | 28.637 | 16.263 | 53.533 | 93.241 | 43.911 |
| 33.771 | 78.954 | 399.071 | 102.947 | 118.077 | 61.894 |
| 72.435 | 108.561 | 46.252 | 40.479 | 95.291 | 10.291 |
| 27.668 | 116.729 | 149.432 | 59.067 | 199.458 | 45.771 |
| 272.005 | 60.266 | 233.254 | 87.592 | 137.149 | 50.668 |
| 89.601 | 313.879 | 150.011 | 173.580 | 220.413 | 182.737 |
| 6.171 | 162.792 | 82.273 | | | |

In this 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 an Exponential.

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 devices in warehouses, etc., should be randomly selected so they become representative of these characteristics and of the environment in which devices will normally be operated.

Having knowledge about the product and its testing procedure, can help in assessing that the observations are independent and representative of the population from which they come, and establishes the first of the two distribution assumptions.

To assess the exponentiality of the data, we use several informal methods, based on the properties of the Exponential distribution.

They are practical for the engineer because they are largely intuitive and easy to implement.

To assess the data in Table 2, using this more practical approach, we first obtain their descriptive statistics (Table 3). Then, we analyze and plot the raw data in several ways, to check (empirically but efficiently) if the Exponential assumption holds.

Table 3. Descriptive Statistics of Data in Table 2

| Variable | n | Mean | Median | Std. Dev. |
|-----------|----|------|--------|-----------|
| Exp. Data | 45 | 99.9 | 77.1 | 85.6 |

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

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

It is worthwhile to notice that the values of the sample mean and standard deviation are the same, irrespective of the underlying distribution. What will change are the properties of such values, a fact that can be used to help identify the distribution in question.

There are a number of useful and easy to implement procedures, based on well-known statistical properties of the Exponential distribution, which help us to informally assess this assumption. These properties are summarized in Table 4.

Table 4. Some Properties of the Exponential Distribution

1. The theoretical mean and standard deviation are equal¹; hence, the sample values of mean and standard deviation should be close.
2. Histogram should show that the distribution is right-skewed (Median < Mean).
3. A plot of Cumulative-Failure vs. Cumulative-Time should be close to linear.
4. The regression slope of Cum-Failure vs. Cum-Time is close to the failure rate.
5. A plot of Cum-Rate vs. Cum-Failure should decrease/stabilize at the failure rate level.
6. Plots of the Exponential probability and its scores should also be close to linear.

¹Although the Exponential is a one-parameter distribution, it has a standard deviation. All distributions, except for the Cauchy, have a standard deviation.

First, from the descriptive statistics in Table 2, we verify that the Mean (99.9) and Standard Deviation (85.6) are close, and that the Median (77.1) is smaller than the Mean. This agrees with the Exponential distribution Property No. 1 and Property No. 2, from Table 4.

The theoretical Exponential standard deviation σ is equal to the mean. Hence, one standard deviation above and below the mean yields a range of population values 0 to 2σ , which comprises the majority (86% of the values) of the population (see Figure 3). For reference, in the Normal distribution, one standard deviation above and below the Mean comprise only 68% of the population. The corresponding sample points under these ranges should be commensurate to these percentages and provide an indication to which distribution they come from (especially in large samples).

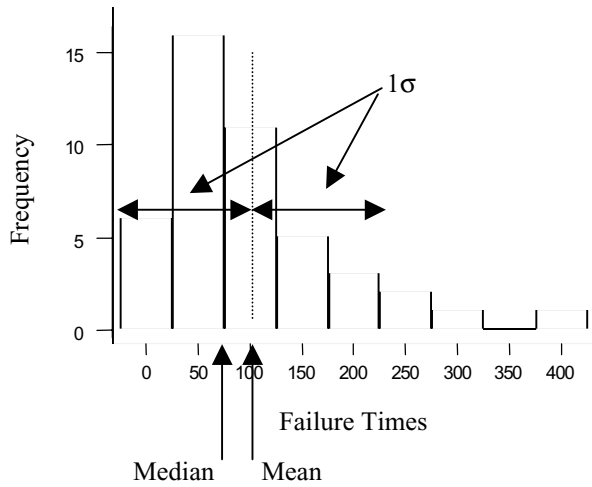


Figure 3. Histogram is Skewed to the Right, as in an Exponential Distribution (Property No. 2)

If we regress the Cumulative Failures on Cumulative Test Time (Table 5), the result is a straight line (Figure 4), whose slope (0.00931) is close to the true (0.01) failure rate (Property 4).

Table 5. Cum-Fail vs. Cum-Time Regression Analysis

| Predictor | Coeff. | Std. Dev. | T | P |
|-----------|-----------|-----------|-------|-------|
| Constant | 5.8048 | 0.5702 | 10.18 | 0.000 |
| Cum-Time | 0.0093135 | 0.0002478 | 37.59 | 0.000 |

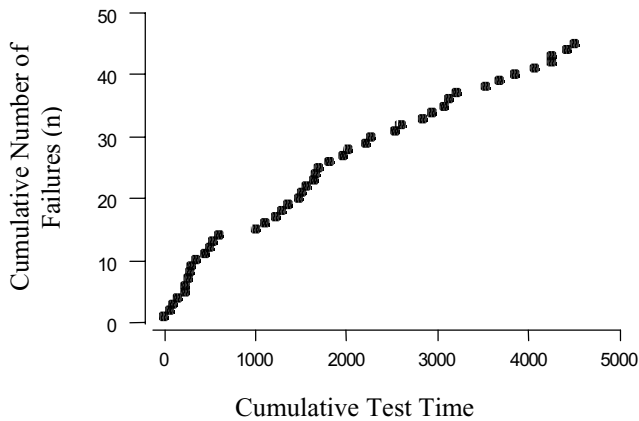


Figure 4. Plot of Cum-Failure vs. Cum-Time is Close to Linear, as Expected in an Exponential Distribution (Property No. 3)

The regression equation is:

$$\text{Cum-Fail} = 5.80 + 0.00931 \text{ Cum-Time}$$

$$S = 2.283 \quad R\text{-Sq} = 97.0\%$$

Notice how the regression in Table 5 is significant, as shown by the large T (= 37.59) test value for the regression coefficient and by the large Index of Fit ($R^2 = 0.97$). Both results suggest that a linear regression with slope equal to failure rate is plausible.

Cumulative Failure Rate (at failure i) is obtained by dividing Cumulative Test Time (at failure i) by Cumulative Failures (up to i), for $i = 2, \dots, n$. If the Cumulative Failure Rate is plotted vs. the Cumulative Failures, it then soon stabilizes to a constant value (the true Failure Rate = 0.01) as expected in the case of the Exponential distribution (Property 5), see Figure 5.

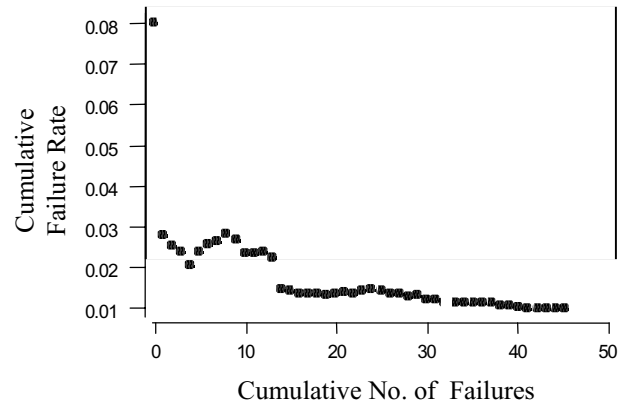


Figure 5. Plot of Cum-Rate vs. Cum-Fail Stabilizes (flat) and Converges to the Exponential Failure Rate (Close to Value 0.01) (Property No. 5)

The Probability Plot is one where the Exponential 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 Exponential probability of the corresponding failure data, X_i using the sample mean (see Figure 6). For example, the first sorted (smallest) data point is 6.17 and the sample average = 99.9:

$$P_{99.9}(6.17) = 1 - \exp(-6.17/99.9) = 1 - 0.94 = 0.06$$

which is then plotted against the corresponding $i/(n + 1)$ value: $1/46 = 0.0217$ and so on, until all other sorted sample elements $i = 1, \dots, 45$, have been considered.

The Exponential scores X_i are the percentiles corresponding to the values $i/(n + 1)$, for $i = 1, \dots, n$; calculated under the Exponential distribution (assuming the sample mean). For the same example, the first $i/(n + 1)$ is $1/46 = 0.0217$ and the sample average = 99.9. Then:

$$P_{99.9}(X_i) = 1 - \exp(-X_i/99.9) \Rightarrow X_i = -99.9 \ln(1 - P_{99.9}(X_i))$$

where:

$$P_{99.9}(X_i) \approx \frac{i}{n+1}$$

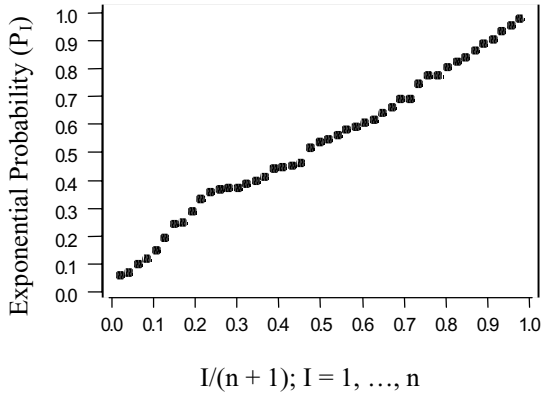


Figure 6. Plot of Exponential Probability (P_i) vs. $i/(n+1)$; $i = 1, \dots, n$ is Close to Linear, as Expected When the Data Come from an Exponential Distribution (Property 6)

Substituting in the above formula $i/(n + 1)$ for $i = 1$, we get the first exponential score:

$$X_i = -99.9 \ln \left(1 - \frac{1}{46} \right) = -99.9 \ln(0.9783) = -99.9 \times (-0.022) = 2.2$$

The scores are then plotted vs. their corresponding sorted real data values (in the case above, 2.2 is plotted against 6.17, the smallest data point). When the data come from an Exponential Distribution, this plot is close to a straight line (Property 6), see Figure 7.

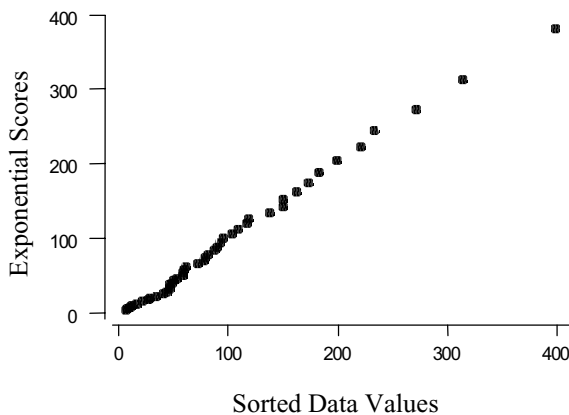


Figure 7. Plot of the Exponential Scores vs. the Sorted Real Data Values

All of the preceding empirical results contribute to support the plausibility of the assumption of the Exponentiality of the given life data. If, at this point, a stronger case for the validity of the Exponential distribution is required, then a number of theoretical GoF tests can be carried out with the data.

A final comment about distribution assumptions and engineering work is due. In practice, engineers do not solve ideal problems -but real and complicated ones, whose settings are not perfect. Consequently, some statistical assumptions may not be met. This does not, however, necessarily preclude the use of statistical procedures.

In such cases, some assumptions may have to be relaxed and some of the inferences (results) may have to be interpreted with care and used with special caution. The best criteria to establish such relaxation and interpretation of the rules (e.g., which assumptions can be relaxed and by how much) often come from a thorough knowledge of the underlying engineering and statistical theories, from extensive professional experience and from a deep understanding of the specific processes under consideration.

Summary

This START sheet discussed the important problem of (empirically) assessing the Exponential distribution assumptions. Several numerical and graphical examples were provided, together with some related theoretical and practical issues, and some background information and references to further readings.

Other, very important, reliability analysis topics were mentioned in this paper. 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. A Practical Guide to Statistical Analysis of Material Property Data, Romeu, J.L. and C. Grethlein. AMPTIAC, 2000.
3. Mechanical Applications in Reliability Engineering, Sadlon, R.J., RAC, 1993.
4. Reliability and Life Testing Handbook (Vols. 1 & 2), Kececioglu, D., Editor, Prentice Hall, NJ, 1993.

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.