Using Test Statistics to Estimate Non-normal Data

Adrian Manhey

amanhey@oxy.edu

Occidental College

Abstract

Statistics is the science of collecting, analyzing, and interpreting data. Much of the technologies that are popular today either are guided by or built from statistical principles, seen in the bustling field of data science and machine learning. However, those fields aren't so different from statistics, Nate Silver statistician behind the media site FiveThirtyEight – and the guy who famously and correctly predicted the electoral outcome of 49 of 50 states in the 2008 US Presidential election, and a perfect 50 for 50 in 2012 – thinks of data science and statistics being one in the same, saying "I think data-scientist is a sexed up term for a statistician" [5]. Using statistics this paper gives definitions of different statistical concepts and walks through an example to show how we might apply statistics in modeling the distribution of failure times of light-bulbs and answer the question of whether these new light-bulbs work and if they are better than current ones.

1 Problem Context

Generally reliability testing is a process that checks whether the product can perform a failure-free operation for a specified time period in a particular environment. Consider a light-bulb manufacturer that is attempting to create a more efficient and environmentally-friendly kind of light-bulb. What is the best way of checking the reliability of these light-bulbs? Should we record the time that a series of light-bulbs takes to fail? Should we take the mean time of a set? Although this question may seem daunting, let's go through some methods in determining how to perform this test [4].

2 Background

To understand how to tackle this problem let's go through some foundational topics in mathematical statistics such as random variables, the exponential distributions, and the Maximum Likelihood estimator.

2.1 Random Variables

To begin it is important to understand random event. Let's consider the tossing of a normal, two-sided coin. The phenomenon of flipping the coin has two outcomes, either "heads" or "tails", each with a given probability of having. Since this is a standard coin, each outcome has a $\frac{1}{2}$ opportunity of happening. Since we do not know which will happen, we describe this as a *random events*. The potential opportunity of each outcome is describes as the *probability* of the event. A random variable is a variable whose possible values are numerical outcomes of a random phenomenon with an associated set of outcome events and probabilities [6].

A function whose domain is a sample space S and whose values form a finite or countably infinite set of real numbers is called a discrete random variable [3]. Discrete random variables are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family, the number of patients in a doctor's surgery, or the number of defective light bulbs in a box of ten.

A continuous random variable takes inputs from a sample space S to the real numbers. Continuous random variables are usually measurements. Examples include height, weight, the amount of sodium in a mixture, or the time required to run a mile.

Each random variable has a distribution of the probabilities associated with the event. The function used to model this distribution is the Probability Density Function (PDF), i.e. $P(X = x)$, the probability that the output will be x.

2.2 Exponential Distributions

Exponential distributions are used to model the distribution the amount of time it takes to wait for certain events to occur. Intuitively this is because the distribution has a higher density of smaller values than larger ones. For example, if we're waiting at a bus stop for our bus, it is more likely that our bus comes in the next 10 minutes than 60 minutes.

To denote a random variable as being exponentially distributed with parameter λ, we write $X \sim \text{Exp}[\lambda]$. Exponential distributions have a PDF of $f_Y(y;\lambda) = \lambda e^{-y\lambda}$ where y is a value of our random variable and λ is the decay parameter. Looking at Figure 1 we can see the decreasing exponential distribution for various values of λ.

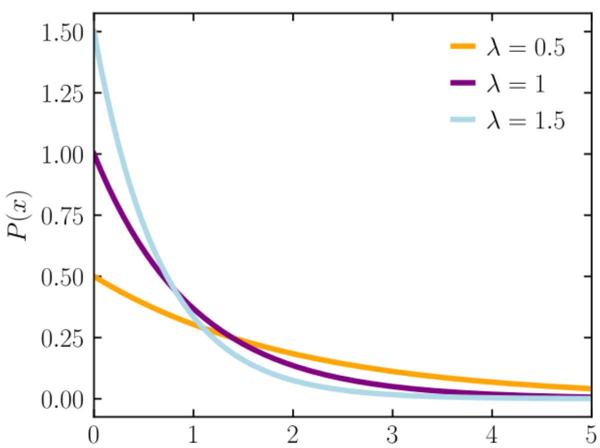

Figure 1. PDF of the Exponential Distribution [2]

The CDF of an exponential distribution, seen in Figure 2, is strictly increasing as it approaches 1. The exponential distribution is notable in that it possesses a *memory-less* property, only shared with the geometric distribution. In terms of waiting times, it means that the conditional probability of an event happening 10 seconds after the current point in time is the same as having waited 30 seconds prior.

For an exponential distribution the mean and variance are given by the decay parameter and the decay parameter squared, respectively. Now that we understand some way to represent the probability, let's try to understand how to use that to connect this distribution to observations of data.

2.3 Maximum Likelihood Estimator

It is widely known that the probability of two independent events is the probability of one multiplied by the probability of the other. Similarly if we have two PDFs of events, the product of them, called the *joint densities*, represents the probability of both events happening together. The likelihood function $L(\theta)$, is the joint densities of each observed data point. The Maximum Likelihood Estimator (MLE) is the value of the parameters, e.g. θ, of the probabilistic model under which the observed data is most probable [3].

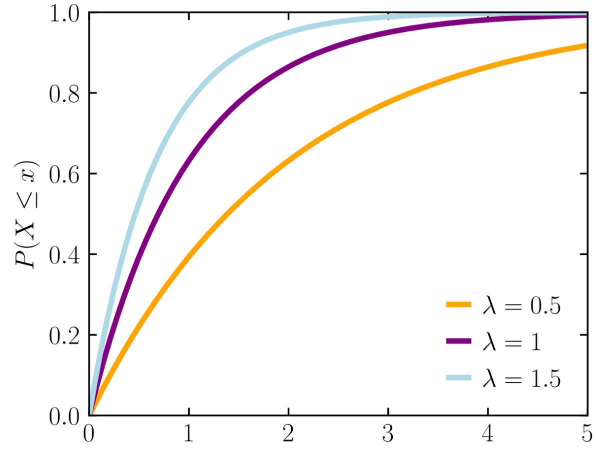

Figure 2. CDF of the Exponential Distribution [2]

In order to find the likelihood function, we take the joint densities of our observed valued. If we have a series of n random variables, $y_1, y_2, ..., y_n$, from a continuous pdf $f_Y(y;\theta)$, where θ is an unknown parameter, the likelihood function

$$L(\theta) = \Pi_{i=0}^{n} f_Y(y;\theta)$$

To make this a bit easier, let's take the log of $L(\theta)$ before finding the derivative. Since $y = \ln(x)$ is strictly increasing $x > 0$, since $\frac{1}{x}$ is positive for $x > 0$ and $\ln(x)$ is only defined for $x > 0$. So, it won't change the relationship between θ and our max. In many instances our PDFs will involve exponentials so taking the natural log helps in those cases.

2.4 Hypothesis Testing

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter, meaning seeing how well a distribution fits an educated guess. In hypothesis testing we have a null hypothesis, our educated guess, and an alternative hypothesis, which specifies how we're going to check if that educated guess is incorrect. In hypothesis testing we use significance levels, α, to denote the probability of us rejecting our null hypothesis despite it being correct, $P(\text{fail to reject} H_0 | H_0 \text{is true})$.

Our decision on whether the null hypothesis can be rejected is based on some condition and comparison:"Reject $H_0 : \lambda = \lambda_0$ if [condition]", where the condition is some comparison $\{<, >, \neq\}$ against c, a critical value.

Our critical value represents the point at which we know that the observation no longer fits the supposed distribution. For example, let's say that we believe the average number of cars on the road is 5, so we can say the population mean is 5, $\mu_0 = 5$. However, we want to test to see if that is true so we're going to challenge it by checking to see if μ being higher than 5 is more likely. Our hypothesis test might look something like this:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Let's suppose our critical value was $c = 15$ at a significance level of $\alpha = 0.05$. Now if we were to go out and see more than 15 cars we would be able to reject the null hypothesis and say that the true mean of the number of cars on the road is greater than 5.

3 Methods

Now in order to model the distribution of time between light-bulb failures we are going to consider two methods, using the MLE and the mean, $\bar{X}$. Each are detailed in Test 1 and Test 2, respectively. We'll see that despite the MLE and $\bar{X}$ having the same distribution, $\bar{X}$ results in a considerably longer wait time than the MLE.

3.1 Test 1: Taking an [m:n]-element Sample

Typically a reliability inspection policy puts n items "on test," and the first m failure times $X_1, X_2, ..., X_m$, are recorded with $X_i = X_{i:n}$ indicating the ith order statistic, i.e. the ith observation in increasing order among the n observations. So $X \sim$ Exponential[λ]. The failure times can be modeled by an exponential distribution with parameter λ:

$$f_X(x) = \frac{1}{\lambda}e^{-x/\lambda}, \text{if} x > 0.$$

Since X_m is the last failure in our observation, we can represent it as the maximum failure time. The probability that any observation is less than X_m is given by the CDF at X_m. So,

$$\begin{aligned} F_X(x) &= P(x \leq x_m) \\ &= \int_0^{x_m} \frac{1}{\lambda}e^{\frac{-x}{\lambda}} \\ &= 1 - e^{\frac{-x_m}{\lambda}} \\ &= 1 - P(x > x_m) \end{aligned}$$

The resulting probability for an observation to be greater than x_m is $P(x > x_m) = e^{\frac{-x_m}{\lambda}}$. Using this as the probability of the unobserved elements in our sample, our likelihood function for λ is given by

$$L(\lambda) = \prod_{i=1}^{m} f_X(x) * \prod_{i=m+1}^{n} P(x > x_m)$$

. Substituting in our PDF for X and the probability equation for X_m and reducing the equation,

$$L(\lambda) = \frac{1}{\lambda^m}e^{\frac{-1}{\lambda}\sum_{i=1}^m x_i} * e^{(n-m)(\frac{-x_m}{\lambda})}$$

.

Since the natural log function is strictly increasing over the $x > 0$ domain, taking the log of our likelihood function will not change the behavior. Using the simplified form we can find the maximum value using the first derivative,

$$\frac{\partial}{\partial\lambda}(\ln(L(\lambda))) = \frac{-m}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{m} x_i + (n-m)(\frac{x_m}{\lambda^2})$$

$$\lambda_e = \frac{1}{m}(\sum_{i=1}^{m} x_i + (n-m)x_m)$$

Since $x_{i+1} - x_i \sim \text{Exp}[\frac{\lambda}{n-i}]$ and $(n-i)(x_{i+1} - x_i) \sim \text{Exp}[\lambda]$, λ_e can be rewritten as the sum of exponential variables.

$$\begin{aligned} m\lambda_e &= S_m \\ &= \sum_{i=1}^{m} x_i + (n-m)x_m \\ &= (n-m)x_m + (x_1 + ... + x_m) \\ &= (n-m+1)x_m + ... + x_1 \\ &= (n-m+1)(x_m - x_{m-1}) \\ &+ ... + (n-1)(x_2 - x_1) \\ S_m &= \sum_{i=1}^{m} \text{Exp}[\lambda] \end{aligned}$$

Thus, $\lambda_e = \sum_{i=1}^m \text{Exp}[\frac{\lambda}{m}]$ and $\lambda_e \sim \Gamma[m, \frac{\lambda}{m}]$, thus yielding $E\hat{\lambda} = m\lambda/m = \lambda$, i.e., the MLE $\hat{\lambda}$ is an unbiased estimator of the parameter λ. Looking at the expected value of Test 1 we have,

$$\begin{aligned} E(X_{[m:n]}) &= \sum_{i=1}^{m} E(X_{[i:n]} - X_{[(i-1):n]}) \\ &= \lambda \sum_{i=1}^{m} \frac{1}{n-i+1} \\ &= \lambda(H_n - H_{n-m}) \\ &\approx \lambda \ln(\frac{n}{n-m}) \\ &\approx \lambda\frac{m}{n} \end{aligned}$$

3.2 Test 2: Taking an m-element Sample

Using this expected value we can compare it to our second testing method to see which one will take longer. For

Test 2, we take the sample mean as our estimator. By taking the sample mean $\bar{X}$ of an *arbitrary* random sample of size m from the same distribution, $X \sim$ Exponential[λ], we can derive that $\bar{X} \sim \Gamma[m, \frac{\lambda}{m}]$, i.e., the distribution of $\bar{X}$ and $\hat{\lambda}$ are identical.

$$
\begin{aligned}
E(X_{[m:m]}) &= \sum_{i=1}^{m} E(X_{[i:n]} - X_{[(i-1):n]}) \\
&= \lambda H_m \\
&\approx \lambda(\ln m + \gamma)
\end{aligned}
$$

if n, m, and n/m are large enough, and $\gamma \approx 0.5772$ being the Euler-Mascheroni constant as in $H_n = \ln n + \gamma + O(\frac{1}{n})$ as $n \to \infty$.

3.3 Experiment

The experiment conducted was to investigate the distribution of the time between light-bulb failures using Python. The implementation for each process is very similar except with Test 1 we have an additional loop to represent the n light-bulbs that are being simultaneously tested. The probability, p, that a light-bulb failed at any point was 0.01 and the n was fixed at 20. Different values of m were tested with respect to percentages of n, i.e. $m = \{0.2 * n, 0.4 * n, 0.6 * n, , 0.8 * n, 1 * n\}$. Our implementation uses m as the number of failures we want, n as the size of the Test 1 sample, E as an integer representing the current number of episodes, L as an integer representing the episode with the last failure, V as a list representing every time between failures, k as an integer representing the current number of failures, T as the returned tuple containing the total number of episodes taken for a trial, the statistics of the trial, and the time take between each failure.

Figure 3. Code for Test 1

```
def test1(m, n, p):
    V = [0]
    L, E, k = 0
    while(k < m):
        E += 1
        for i in range(n):
            if(k >= m): return T
            if (random.random() <= p):
                V.append(E - L)
                L = E
                k += 1
    return T
```

In Figure 3, the implementation for Test 1 can be seen. There are variables created to hold information about current and past times and episodes to track the intervals of time between failures. The test is run until m failures are reached, $m = k$. To represent a set of n light-bulbs being put on test, a for loop is used to test each light-bulb in the same episode, as if they were happening in the same time-stamp. Each time a light-bulb fails, represented by the generation of a random number less than the probability of a light bulb failing, the the time taken is added to V, L is updated to the current episode, and the failure counter is incremented.

Figure 4. Code for Test 2

```
def test2(m, p):
    V = [0]
    L, E, k = 0
    while(k < m):
        E += 1
        if random.random() <= p:
            V.append(E - L)
            L = E
            k += 1
    return T
```

The implementation for Test 2, seen in Figure 4, is very similar to Test 1 except for the lack of the addition for-loop. This small change affects the amount of times that a light-bulb is being tested. This change is also reflected in the amount of episodes needed to complete each test. Considering Figure 5, we see that the range of the average amount of time between failures is between 0 and 8.

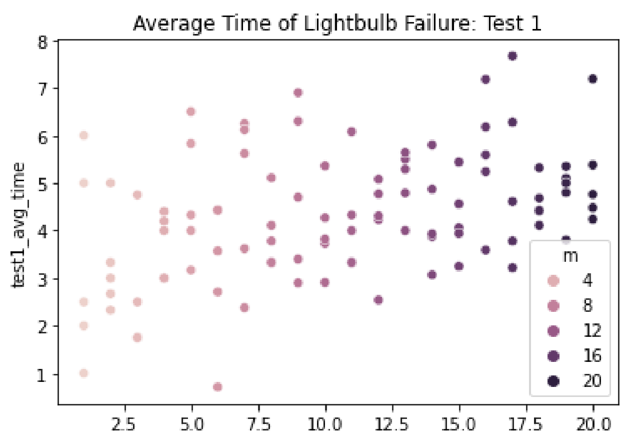

Figure 5. Scatter plot of the Average Time Between Failures for Test 1

On the other hand, for Test 2, Figure 6, the range is much wider and many values are above 50 episodes.

From these graphs we can see that Test 1 takes much less iterations to observe m failures than Test 2. However, the question still remains of whether these light-bulbs are better than regular light-bulbs.

The full implementation can be found at *Destructive Sampling Repo* [1].

Table 1. Results from 20 Trials of Varying Values of m

	Test 1			Test 2		
m	Total Episodes	Average Time	Max Time	Total Episodes	Average Time	Max Time
1	12	6.0	12	49	24.5	49
5	26	4.33	9	485	80.83	301
10	32	2.91	10	857	77.91	219
15	52	3.25	16	1850	115.62	335
20	94	4.48	15	1929	91.86	194

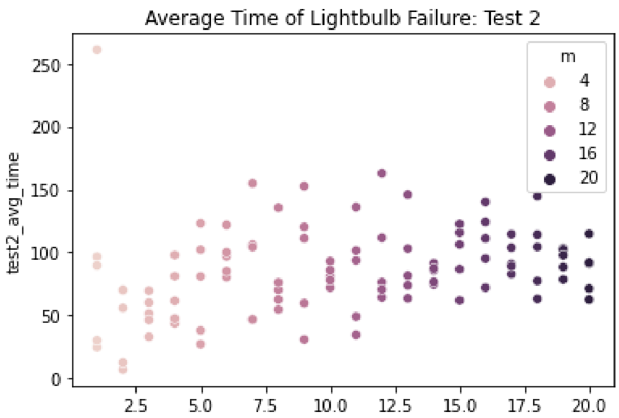

Figure 6. Scatter plot of the Average Time Between Failures for Test 2

3.4 Testing Against the Standard

Let's say that from our tests we found that the maximum time for the new light-bulbs to fail, from a sample of 20, is 12 months and we want to create a hypothesis test against standard bulbs. The standard light-bulb lasts about 4 months. Since we know that the average time between failures are exponentially distributed we can model this with $X \sim \text{Exp}[\lambda = 4months]$, since the parameter of an exponential distribution is the mean. Since we want to see if our light-bulbs perform better, we're going to check to see if the mean of our distribution is greater than the standard. We have,

$$H_0 : \lambda = 4$$

$$H_1 : \lambda > 4$$

Since we're using the maximum failure time, we need to use the respective PDF, by theorem

$$f_{Y_m}(y; \lambda_0 = 4) = n[F_Y(y)]^{n-1} f_Y(y)]$$

To find the critical value associated with an $\alpha = 0.05$ significance level, we need to find the probability of rejecting H_0 given it is true.

$$\begin{aligned}
\alpha &= \int_c^\infty f_{Y_m} dy \\
&= F_Y(y)|_c^\infty \\
&= -e^{\frac{-\infty}{4}} + e^{\frac{-c}{4}} \\
&= e^{\frac{-c}{4}} \\
\ln(\alpha) &= \frac{-c}{4} \\
c &= -4\ln(0.05) \\
&= 11.98
\end{aligned}$$

Since the critical value is less than our observed maximum, we fail to reject the null hypothesis, despite it being so close. Thus we can say that we have statistically significant evidence that the new environmentally friendly lightbulbs last longer than the standard.

References

[1] *Destructive Sampling Repo*. URL: https://github.com/aamanhey/DestructiveSampling.

[2] *Exponential Distribution*. URL: https://en.wikipedia.org/wiki/Exponential_distribution.

[3] Larsen, Richard and Marx, Morris. *An introduction to mathematical statistics and its applications*. 5th ed. Pearson Education, 2012. ISBN: 978-0-321-69394-5.

[4] Lengyel, Thomas. Personal communication. 2022.

[5] Megahan, Justin. *This is the difference between statistics and data science*. Oct. 2020. URL: https://mixpanel.com/blog/this-is-the-difference-between-statistics-and-data-science/.

[6] *Random Variables*. URL: https://en.wikipedia.org/wiki/Random_variable.