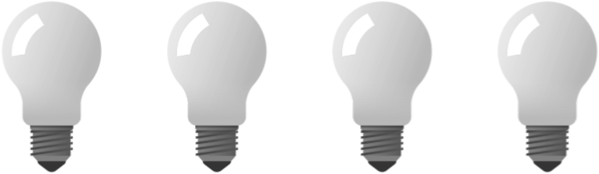
Using Test Statistics to Estimate Non-normal Data

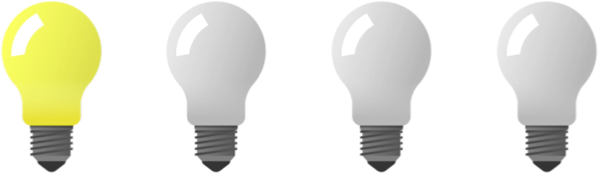
Adrian Manhey

April 2022

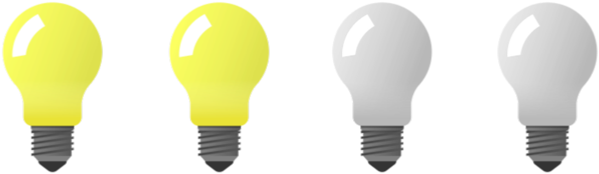
Introduction



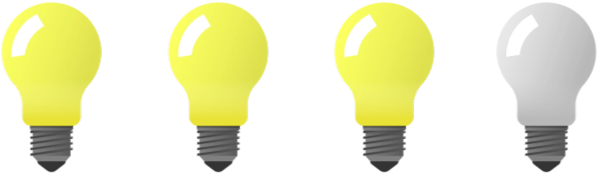
Introduction



Introduction



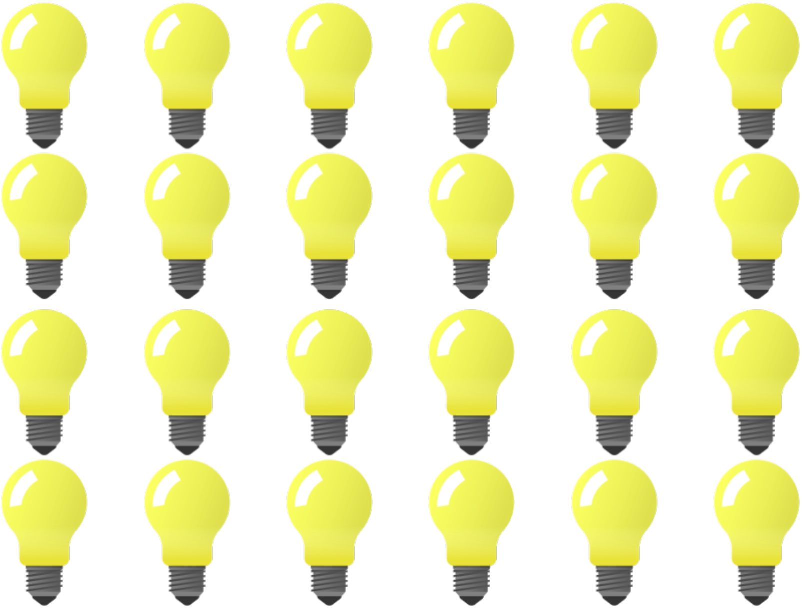
Introduction



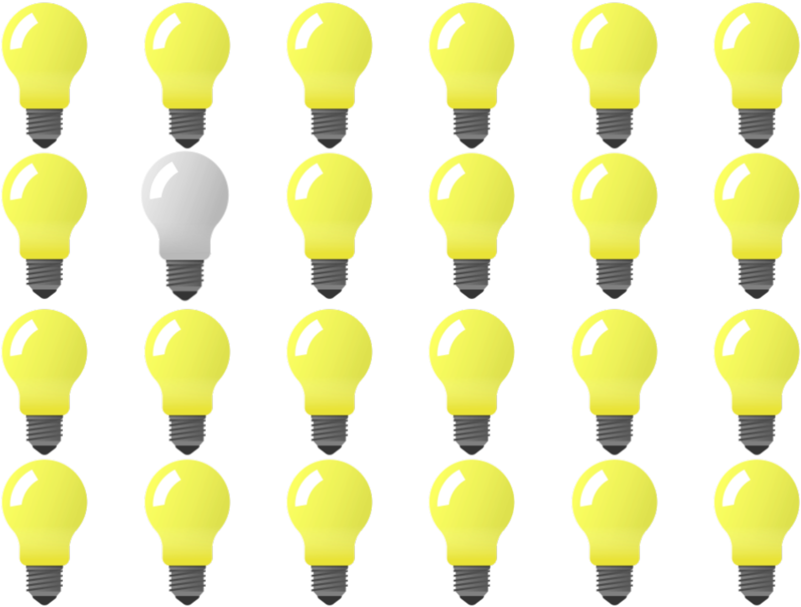
Problem Context

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?

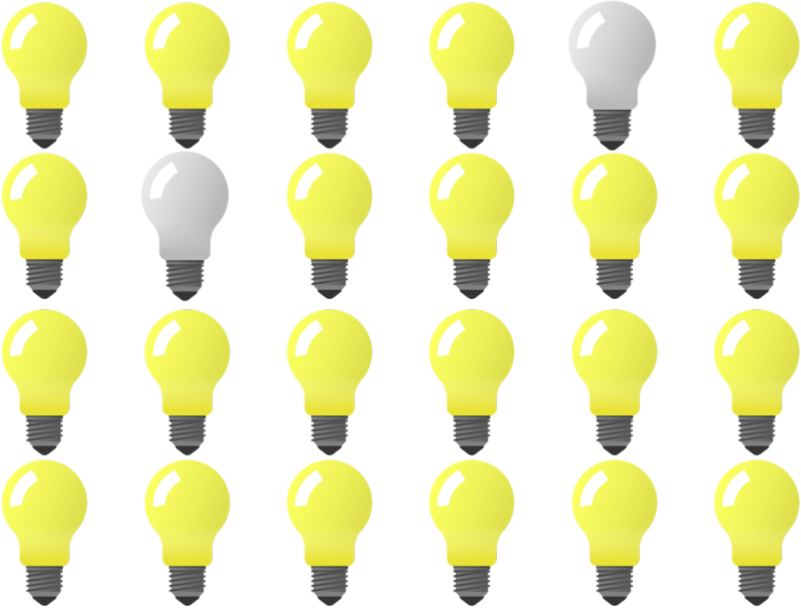
Test 1



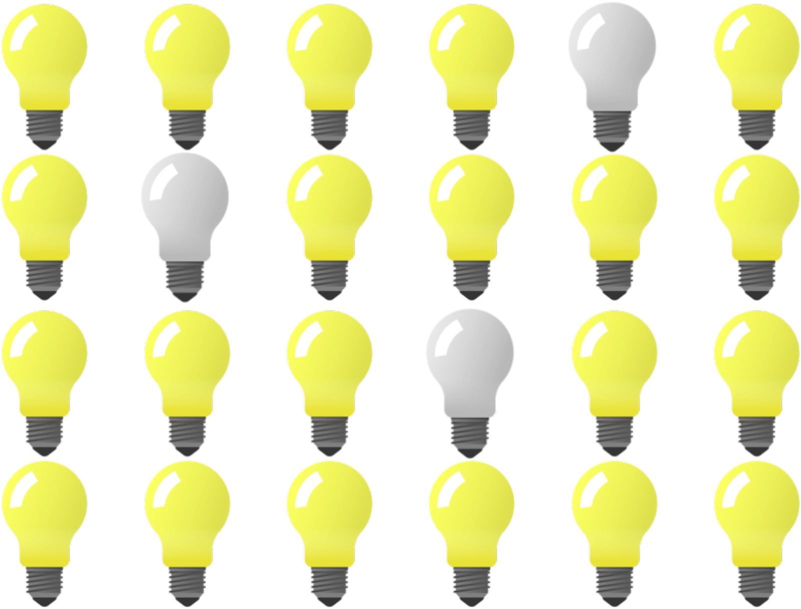
Test 1



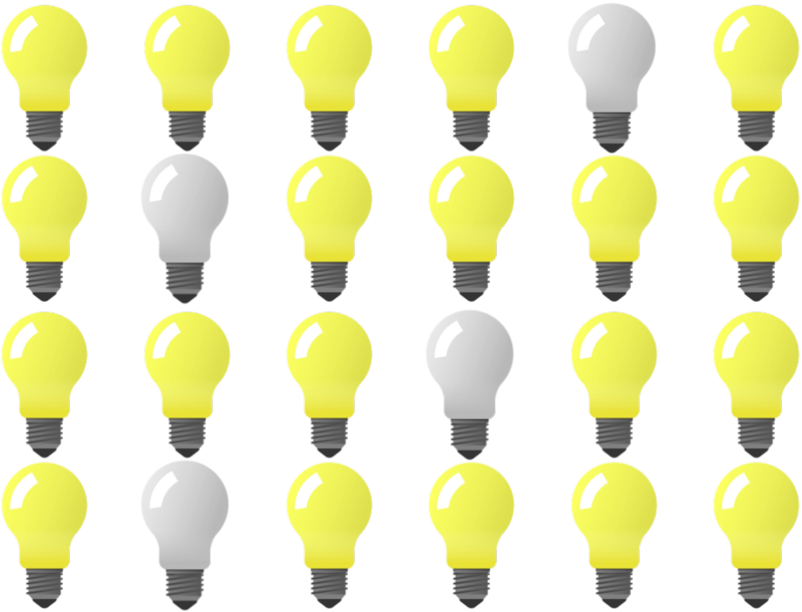
Test 1



Test 1



Test 1



Test 2

Test 2

Test 2

Test 2

Test 2

Test 2

Statistics

"I think data-scientist is a sexed up term for a statistician" - Nate Silver [4]

Background

To understand how to tackle this problem let's go through some foundational topics

- random variables
- the exponential distributions
- the Maximum Likelihood estimator

Random Variables

”heads” or ”tails”

H - H - H - H - H

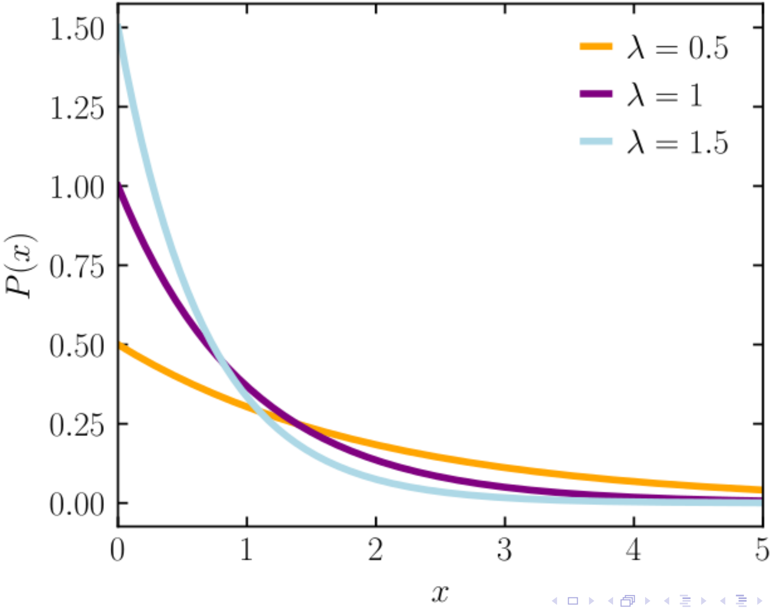
H - T - H - T - H

T - T - T - T - T

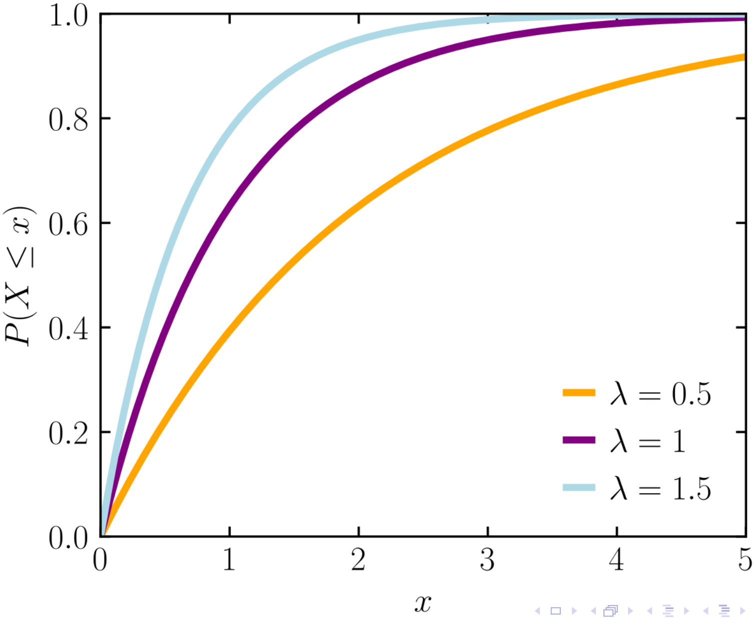
Exponential Distributions

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.

Exponential Distributions Probability Density Function



Exponential Distribution Cumulative Density Function



Exponential Distribution

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.

Maximum Likelihood Estimators

$$P(A \cap B) = P(A) * P(B)$$

Maximum Likelihood Estimators

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 [2].

Maximum Likelihood Estimators

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

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

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.

Hypothesis Testing

”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.

Hypothesis Testing

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:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Hypothesis Testing

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.

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.

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

Suppose a reliability inspection policy puts n items "on test," and the first m failure times are represented by $X_1, X_2, ..., X_m$.

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

The failure times can be modeled by an exponential distribution with parameter λ:

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

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

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

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

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

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

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

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

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

.

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

To maximize λ,

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

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

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

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 [3].

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

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

Looking at the expected value of Test 1 we have,

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

Test 2: Taking an m-element Sample

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

Test 2: Taking an m-element Sample

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

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 $n \to \infty$.

Experiment

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. $m = \{0.2 * n, 0.4 * n, 0.6 * n, , 0.8 * n, 1 * n\}$.

Experiment

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.

Experiment

Figure: 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
```

Experiment

Figure: 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
```

Experiment

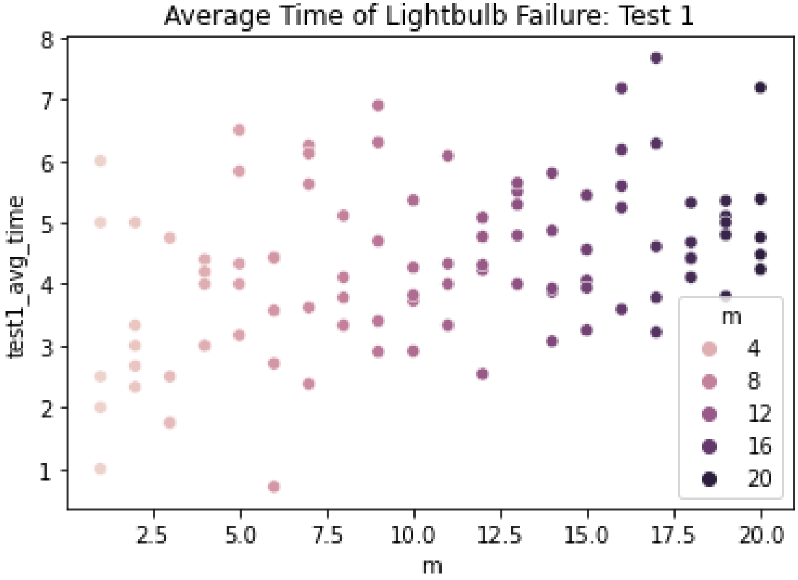


Figure: Scatter plot of the Average Time Between Failures for Test 1

Experiment

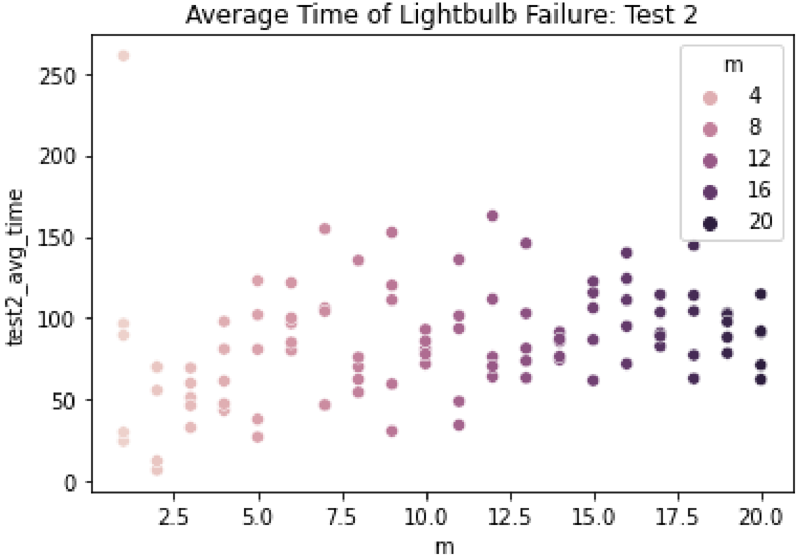


Figure: Scatter plot of the Average Time Between Failures for Test 2

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.

Testing Against the Standard

We can model this with $X \sim \text{Exp}[\lambda = 4 \textit{months}]$, 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.

Testing Against the Standard

Hypothesis Test

$$H_0 : \lambda = 4$$
$$H_1 : \lambda > 4$$

Testing Against the Standard

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

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

Testing Against the Standard

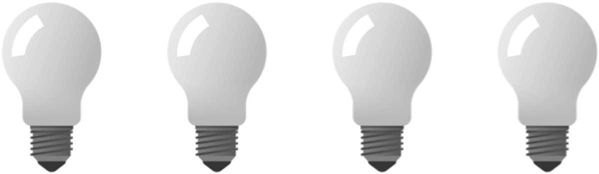
To find the critical value associated with an $\alpha = 0.05$ significance level, we need to find the probability of rejecting H_0 given it is true.

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

Testing Against the Standard

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 light-bulbs last longer than the standard.

Thank you!



References

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

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

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

[4] 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/.