Exponential Distributions & its real-life Applications

Dr. Tanujit Chakraborty

Sorbonne University

Code: https://github.com/tanujit123/MATH-260

Probability Distribution

- Probability distributions are a fundamental concept in statistics.
- A probability distribution is a mathematical function that gives the probabilities of occurrence of different possible outcomes of an experiment.
- Probability distributions are broadly categorized into two classes: Discrete Probability Distribution and Continuous Probability Distribution.
- Exponential Distribution is an example of Continuous Probability Distribution and is primarily used in reliability applications and analysis of lifetime and survival data.
- The exponential distribution is used to model data with a constant failure rate (indicated by the hazard rate which is simply equal to a constant).
- The exponential distribution is the probability distribution of the time between events in Poisson process and continuous analogue of the geometric distribution.

Exponential Distribution

- In probability theory and statistics, the Cumulative Distribution Function (CDF) of a random variable X, evaluated at x, is the probability that X will take a value less than or equal to x.

- A continuous random variable X whose CDF $F(x)$ is given by

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

  is said to follow an Exponential Distribution.

- The Probability Density Function (PDF) is the derivative of the cumulative distribution function.

- The PDF $f(x)$ for Exponential distribution is obtained by differentiating the CDF $F(x)$ and is given by:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Maximum Likelihood Estimation

Let $x_1, x_2, \ldots, x_n$ be independent random observations from the same Poisson distribution with parameter λ. Then, the likelihood function is

$$L(\lambda \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i, \lambda)$$
$$= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

The log-likelihood function is

$$\log L(\lambda \mid x_1, x_2, \ldots, x_n) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

After equating the partial derivative of log-likelihood function w.r.t λ to 0, we obtain the MLE of λ as

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

Real-life Applications of Exponential Distribution

Exponential Distribution is used for analyzing lifetime data. Exponential Distribution is used to predict the amount of waiting time until the next event,i.e., success, failure, arrival, etc.
For example, we can use Exponential distribution to predict the following events:

- The amount of time until the customer finishes browsing and actually purchases something from an online store (success).
- The amount of time until the hardware of any computer fails (failure).
- The amount of time you need to wait until the train arrives (arrival).

The Memoryless Property

The memoryless property tells us about the conditional behavior of exponential random variables. Let X be exponentially distributed with parameter λ. Suppose we know $X > t$. What is probability that X is also greater than some value $s + t$? That is, we want to know

$$P(X > s + t \mid X > t).$$

Using the definition of conditional probability, we have

$$P(X > s + t \mid X > t) = \frac{P(X > s + t \text{ and } X > t)}{P(X > t)}.$$

If $X > s + t$, then $X > t$ is redundant, so we can simplify the numerator as

$$P(X > s + t \mid X > t) = \frac{P(X > s + t)}{P(X > t)}.$$

Using the CDF of the exponential distribution,

$$P(X > s + t \mid X > t) = \frac{1 - P(X \leq s + t)}{1 - P(X \leq t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}}.$$

The Memoryless Property

The $e^{-\lambda t}$ terms cancel, giving the surprising result

$$P(X > s + t \,|\, X > t) = e^{-\lambda s}.$$

- It turns out that the conditional probability does not depend on t!.
- The probability of an exponential random variable exceeding the value $s + t$ given t is same as the variable originally exceeding that value, s regardless of t.
- The exponential distribution is *memoryless* because the past has no bearing on its future behavior. Every instant is like the beginning of a new random period, which has the same distribution regardless of how much time has already elapsed.
- The exponential is the only memoryless continuous random variable.

Implications of the Memoryless Property

- Suppose we're observing a stream of events with exponentially distributed interarrival times. Because of memoryless property, the expected time until the next event is always $\frac{1}{\lambda}$, no matter how long we've been waiting for a new arrival to occur.

- This behavior is a bit counterintuitive. We might expect that arrivals get *more* likely the longer we wait. For example, if a bus is supposed to come every ten minutes, and we have been waiting for nine minutes without seeing a bus, we expect that the next bus should be along very soon. If the interarrival time of bus is exponentially distributed, the memoryless property tells us that our waiting time is of no use in predicting when the next bus will arrive [1]

- Suppose we have a queue with exponentially distributed service times. If a new customer arrives to the queue to find someone in service, the *residual service time* is the time until the currently running customer finishes service and departs the queue. Because of the memoryless property, the distribution of the residual service times does not depend on how long the customer has been in service. The probability that the current customer runs for an additional minute and then departs is same as the probability that a new customer just entering service runs for one minute. Likewise, the average remaining service time is simply the expected time for a new customer just entering service.

---

[1]Of course, the time between real bus arrivals is never exponentially distributed. People want to know exactly when their buses will come, so bus schedules are nearly deterministic.

Failure Rates

In the field of reliability theory, it's common to use a random variable to represent the lifespan of a component. One of the main problem in this area is predicting the likelihood that a component fails in the very near future given its current age. This likelihood is summarized by the *failure rate* (also called the hazard rate) of the component.

For most components, the failure rate changes with time. There are three possible relationships.

- increasing failure rate
- decreasing failure rate
- constant failure rate

Failure Rates

If the failure rate is *increasing*, then failures become more likely as the component ages. Most manufactured products behave in this way - they're built to last for a certain amount of time, then fail. Failures become more likely as the product wears out and the end of its lifespan approaches.

A *decreasing* failure rate implies that the probability of failure decreases with the passage of time. In other words, the longer a component has worked, the more likely it is to continue working. UNIX processes have been shown to have decreasing failure rates - the longer a job runs, the more likely it is to continue running. Human lifespans also have decreasing failure rate, as surviving further into adulthood makes it more likely that you will live to old age, at least until you reach the upper limit of your natural lifespan.

The final category corresponds to components with exponentially distributed lifespans. Because of the memoryless property, the length of time a component has functioned in the past has no bearing on its future behavior, so the probability that the component fails in the near future is always the same and doesn't depend on its current age.

Need for generalizations of Exponential Distribution

- The exponential distribution models the behavior of units that fail at a constant rate, regardless of the accumulated age.
- Although this property greatly simplifies the analysis, but it makes the distribution inappropriate for most "good" reliability analyses because it does not apply to most real world applications.
- Due to this reason, several generalizations of Exponential Distribution have been suggested in the previous literature.
- Various modifications are made in the Exponential Distribution function as to optimize its performance by adding new parameters.
- Gompertz (1825) used the following distribution function to represent mortality growth

$$G(t) = \left(1 - \rho e^{-\lambda t}\right)^{\alpha}; \quad t > \frac{1}{\lambda}\ln\rho.$$

Exponentiated Exponential (EE) Distribution

- The EE Distribution is formed by adding a new parameter α to the already existing CDF of Exponential Distribution.

- The Cumulative Distribution Function of Exponentiated Exponential Distribution is

$$F_{EE}(x; \alpha, \lambda) = (1 - e^{-\lambda x})^{\alpha}; \ x \geq 0.$$

- On differentiation of this CDF, we get the Probability Density Function (PDF) as :

$$f_{EE}(x; \alpha, \lambda) = \alpha\lambda(1 - e^{-\lambda x})^{\alpha-1}e^{-\lambda x}; \ x \geq 0.$$

- Here $\alpha \ (> 0)$ is the shape and λ is the scale parameter.
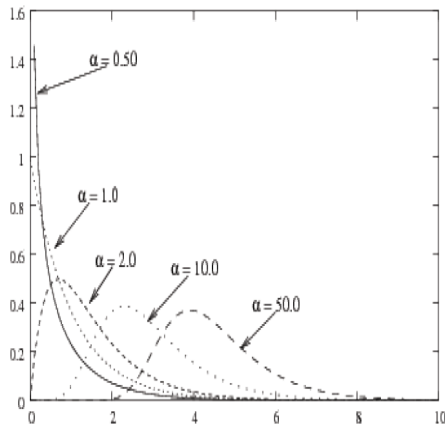
Fig: PDF's of EE distribution for different α.

Basic Properties of EE & Applications

- It can used for analyzing skewed lifetime data.
- It has nice physical interpretations.
- EE is very useful when the data are from a regular maintenance environment.
- Hazard function of EE can be increasing, decreasing or constant.
- It is a member of the proportional reversed hazard model.
- The Maximum Likelihood Estimates (values of α and λ) can be directly obtained from the log likelihood function.
- It can also be used for analyzing censored data.

Application

We are now moving on to the experimental part. Two data sets are analyzed and the results obtained from exponential and exponentiated exponential (EE) are compared.

We have analysed Coalmine data set and Guinea Pigs data set in this study. The parameters of the fitted distributions are estimated using maximum likelihood estimation (MLE) procedure. We have also performed Kolmogorov-Smirnov tests for testing the goodness of fit of the distributions.

Coalmine Dataset

The uncensored dataset corresponding to intervals in days between 109 successive coal-mining disasters in Great Britain, for the period 1875-1951. The sorted data are given as follows:
1 4 4 7 11 13 15 15 17 18 19 19 20 20 22 23 28 29 31 32 36 37 47 48 49 50 54 54 55 59 59 61 61 66 72 72 75 78 78 81 93 96 99 108 113 114 120 120 120 123 124 129 131 137 145 151 156 171 176 182 188 189 195 203 208 215 217 217 217 224 228 233 255 271 275 275 275 286 291 312 312 312 315 326 326 329 330 336 338 345 348 354 361 364 369 378 390 457 467 498 517 566 644 745 871 1312 1357 1613 1630

Histogram of Coalmine Data

Combined plot

Legend: Exponential (red), Exponentiated exponential (blue)
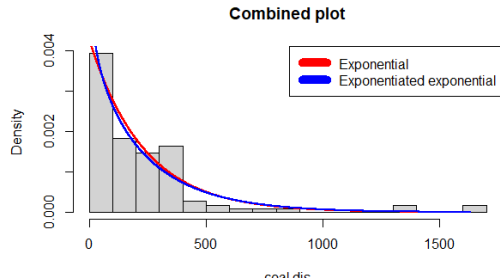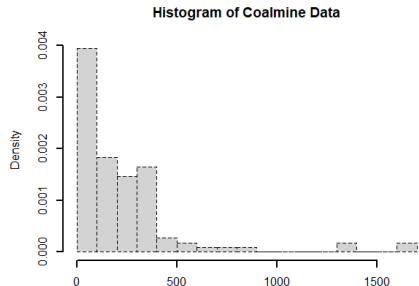
Guinea Pigs Dataset

The data set consists of survival times of guinea pigs injected with different amount of tubercle bacilli. Guinea pigs are known to have high susceptibility of human tuberculosis, which is one of the reasons for choosing this species. The data represents the survival times of Guinea pigs in days. The data are given below:
12 15 22 24 24 32 32 33 34 38 38 43 44 48 52 53 54 54 55 56 57 58 58 59 60 60 60 60 61 62 63 65 65 67 68 70 70 72 73 75 76 76 81 83 84 85 87 91 95 96 98 99 109 110 121 127 129 131 143 146 146 175 175 211 233 258 258 263 297 341 341 376
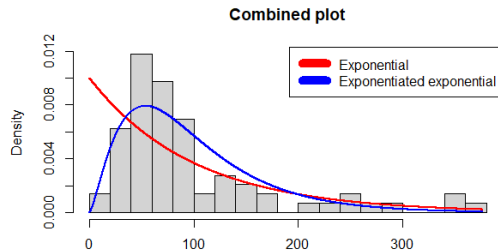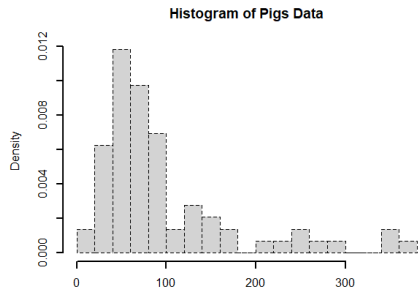
Results

In order to check, which distribution performed better, the log-likelihood (LL) function is computed. Smaller the value of the log-likelihood function, better is the distribution. Also, larger the **p-value**, better is the distribution for that data set.

The distributions used for comparisons are:

- Exponential Distribution
- Exponentiated Exponential Distribution

Results

Table: Coalmine

Distribution	$\hat{\alpha}$	$\hat{\lambda}$	LL	p-value
Exponential		0.0043	-703.3133	0.5158
Exponentiated Exponential	0.8588	0.0039	-702.5525	0.4422

Exponential distribution performed better than exponentiated exponential distribution for this data set.

Table: Pigs

Distribution	$\hat{\alpha}$	$\hat{\lambda}$	LL	p-value
Exponential		0.01002	-403.4421	0.00317
Exponentiated Exponential	2.48431	0.01702	-393.11059	0.15977

Exponentiated exponential distribution performed better than exponential distribution for this data set.

Discussions

- Here, a comparative study has been performed where, we have checked whether the observed data can be modelled better with Exponential distribution or with Exponentiated Exponential distribution.

- Based on the experimentation we can conclude that none of them is an universal distribution. However, we can gather some idea about the potentially better fitting distribution from the histogram plot.

- Can you think about a better model that these two discussed models for the Guinea Pigs Data set?

References

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-202.

Ristić, M. M., & Kundu, D. (2015). Marshall-Olkin generalized exponential distribution. Metron, 73(3), 317-333.

Mahdavi, A., & Kundu, D. (2017). A new method for generating distributions with an application to exponential distribution. Communications in Statistics-Theory and Methods, 46(13), 6543-6557.

Nadarajah, S., & Kotz, S. (2006). The beta exponential distribution. Reliability engineering & system safety, 91(6), 689-697.

Gupta, R. D., & Kundu, D. (2009). A new class of weighted exponential distributions. Statistics, 43(6), 621-634.

Gupta, R. D., & Kundu, D. (1999). Generalized exponential distributions. Australian & New Zealand Journal of Statistics, 41(2), 173-188.

Gupta, R. D., & Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and Weibull distributions. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 43(1), 117-130.

Marshall, A. W., & Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. Biometrika, 84(3), 641-652.