Classical Probability Distributions

Usage in Simulation

Radu T. Trîmbițaș

UBB

1st Semester 2010-2011

Bernoulli

Experiment: n independent trials, each is a succes or a failure, the probability of succes remains constant from trial to trial – Bernoulli process

$$X_k : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, \qquad k = 1, \ldots, n, q = 1 - p$$

$$E(X_k) = 0 \cdot q + 1 \cdot p = p$$
$$V(X_k) = E(X_k^2) - E(X_k)^2 = 0^2 \cdot q + 1^2 \cdot p - p^2 = pq$$

Binomial

(Bernoulli, 1713)
A r.v. X has a binomial distribution if its mass function is

$$p(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0, 1, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

X denotes the number of successes in n Bernoulli trials
X can be written as a sum of n independent Bernoulli r.v.

$$X = X_1 + X_2 + \cdots + X_n$$

So,

$$E(X) = np$$
$$V(X) = npq$$

Applications

Examples

1. Plot pdf and cdf of a binomial distribution with $p = 0.2$ and $n = 10$.
2. A Quality Assurance inspector tests 200 circuit boards a day. If 2% of the boards have defects, what is the probability that the inspector will find no defective boards on any given day? What is the most likely number of defective boards the inspector will find?
3. Suppose that a lot of 300 electrical fuses contains 5% defectives. If a sample of five fuses is tested, find the probability of observing at least one defective.
4. Experience has shown that 30%of all persons afflicted by a certain illness recover. A drug company has developed a new medication. Ten people with the illness were selected at random and injected with the medication; nine recovered shortly thereafter. Suppose that the medication was absolutely worthless. What is the probability that at least nine of ten injected with the medication will recover?

Geometric distribution

- The number of trials to achieve the first success

$$p(x) = \begin{cases} q^{x-1}p, & x = 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

- Mean and variance

$$E(X) = \sum_{k=1}^{\infty} kpq^{k-1} = \frac{1}{p}$$

$$V(X) = \frac{q}{p^2}$$

- The geometric distribution is useful for modelling the runs of consecutive successes (or failures) in repeated independent trials of a system. The geometric distribution models the number of successes before one failure in an independent succession of tests where each test results in success or failure.

Examples

1. Plot a pdf and a cdf of a geometric distribution with $p = 0.5$.
2. Suppose you toss a fair coin repeatedly. If the coin lands face up (heads), that is a success. What is the probability of observing exactly three tails before getting a heads? What is the probability of observing three or fewer tails before getting a heads?
3. 40% of the assembled ink-jet printers are rejected at the inspection station. Find the probability that the first acceptable ink-jet printer is the thirs one inspected.

Negative binomial distribution

- X has a negative binomial distribution with parameters $r \in \mathbb{N}$ and $p \in (0,1)$ if its mass function is

$$f(x|r,p) = \binom{r+x-1}{x} p^r q^x, \qquad x \in \mathbb{N}$$

- It can be thought of as modelling the total number of failures that occured before the nth success

$$E(X) = \frac{rq}{p}$$
$$V(X) = \frac{rq}{p^2}$$

- Sometimes X is considered to be the number of trials to amass a total o r successes

$$P(X = n) = \binom{n-1}{r-1} p^r q^{n-r}, \qquad n = r, r+1, \ldots$$

Applications

Examples

1. Plot the pdf and the cdf of a negative binomial distribution for $r = 3$ and $p = 0.5$
2. A geological study indicates that an exploratory oil well drilled in a particular region should strike oil with the probability 0.2. Find the probability that the oil strike comes on the fifth well drilled and on the first five wells drilled.
3. 40% of the assembled ink-jet printers are rejected at the inspection station. Find the probability that the first printer inspected is the second acceptable printer.

Poisson Distribution

- A random variable X is said to have a Poisson distribution with parameter > 0 iff its mass function is

$$f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}, \qquad x \in \mathbb{N}.$$

- The Poisson distribution is appropriate for applications that involve counting the number of times a random event occurs in a given amount of time, distance, area, etc.
- Mean, variance

$$E(X) = V(X) = \lambda.$$

- As Poisson showed, the Poisson distribution is the limiting case of a binomial distribution where n approaches infinity and p goes to zero while $np = \lambda$.
- First used to model deaths from the kicks of horses in the Prussian Army.

Applications I

1. Plot the pdf and the cdf of a Poisson distribution for $\lambda = 5$.
2. Suppose that a random system of police patrol is devised so that a patrol officer may visit a given beat location $Y = 0, 1, 2, \ldots$ times per half-hour period, with each location being visited an average of once per time period. Assume that Y possesses, approximately, a Poisson probability distribution. Calculate the probability that the patrol officer will miss a given location during a half-hour period. What is the probability that it will be visited once? Twice? At least once?
3. A computer hard disk manufacturer has observed that flaws occur randomly in the manufacturing process at the average rate of two flaws in a 4 Gb hard disk and has found this rate to be acceptable. What is the probability that a disk will be manufactured with no defects?

Applications II

4. Consider a Quality Assurance department that performs random tests of individual hard disks. Their policy is to shut down the manufacturing process if an inspector finds more than four bad sectors on a disk. What is the probability of shutting down the process if the mean number of bad sectors (λ) is two?
5. A repair person is "beeped" each time there is a call service. The number of beeps is Poisson with $\lambda = 2$ per hour. Find the probability of three beeps in the next hour, and the probability of more than three beeps i an 1-hour period.
6. The lead time demand in an inventory system is the accumulation of demand for an item at which an order is placed until the order is received

$$L = \sum_{i=1}^{T} D_i,$$

Applications III

where L is the lead-time demand, D_i is the demand during the ith time period, and T is the number of time periods during the lead time. Both D_i and T may be random variables. An inventory manager desires that the probability of a stockout not exceed a certain fraction (e.g. 5%) during the lead time. The roerder point is the level of inventory at which a new order is placed. Assume that the lead-time demand is poisson distributed with a mean of $\lambda = 10$ units and that a 95% protection of a stockout is desired. That is, find the smallest x such that the probability that the lead-time demand does not exceed x is $\geq 95\%$.

Hypergeometric distribution

- Parameters (they have physical interpretations): N is the size of the population, m is the number of items with the desired characteristic in the population, $m \leq N$. n is the number of samples drawn. Sampling "without replacement".
- The pmf

$$f(x|N,m,n) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, \qquad x = 0,1,\ldots,n, x \geq K, n-x \geq M$$

- mean, variance

$$E(X) = \frac{nm}{N}.$$

$$V(X) = \frac{nm(N-n)(N-m)}{N^2(N-1)}.$$

Applications

1. Plot the pdf and the cdf of an experiment taking 20 samples from a group of 1000 where there are 50 items of the desired type.
2. Suppose you have a lot of 100 floppy disks and you know that 20 of them are defective. What is the probability of drawing 0 through 5 defective floppy disks if you select 10 at random? What is the probability of drawing zero to two defective floppies if you select 10 at random?
3. Suppose you are the Quality Assurance manager for a hard disk manufacturer. The production line turns out disks in batches of 1,000. You want to sample 50 disks from each batch to see if they have defects. You want to accept 99% there are no more than 10 defective disks in the batch. What is the maximum number of defective disks should you allow in your sample of 50?

Uniform distribution

- X has a $U(a,b)$ distribution it its pdf is

$$f(x|a,b) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$$

- cdf

$$F(x|a,b) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases}$$

- mean, variance

$$E(X) = \frac{a+b}{2}$$

$$V(X) = \frac{(b-a)^2}{12}$$

Applications

1. Plot the graphs of pdf and cdf of a $U[0,1]$ distribution.
2. What is the probability that an observation from a uniform distribution with $a = -1$ and $b = 1$ will be less than 0.75?
3. What is the 99th percentile of the uniform distribution between -1 and 1?
4. A bus arrives every 20 minutes at a specified stop beginning at 6:40 $\mathrm{A.M.}$ and continuing until 8:40 $\mathrm{A.M.}$ A certain passenger does not know the schedule, but arrives randomly (uniformly distributed) between 7:00 A.M. and 7:30 A.M. every morning. What is the probability that the passenger waits more than 5 minutes for a bus.

Normal distribution

- X has a $N(\mu, \sigma^2)$ distribution, $\mu \in \mathbb{R}$, $\sigma > 0$ if its pdf is

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

- cdf

$$F(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

- for $\mu = 0$, $\sigma^2 = 1$, standard normal distribution
- Laplace's function – cdf of standard normal

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

- mean, variance

$$E(X) = \mu$$
$$V(X) = \sigma^2$$

Gauss

Pierre Simon Laplace

Central Limit Theorem

Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variable having a finite mean μ and a finite variance σ^2. Then

$$\lim_{n\to\infty} P\left(\frac{X_1 + \cdots X_n - n\mu}{\sigma\sqrt{n}} < x\right) = \Phi(x)$$

Applications

1. Plot the pdf and the cdf of the standard normal distribution
2. Plot on the same graph the pdf of $N(0, 1^2)$, $N(0, 2^2)$, $N(0, 3^2)$. Do the same for the cdf.
3. Check the 3σ rule for a given normal distribution.
4. Find an interval that contains 95% of the values from a standard normal distribution. Note this interval is not the only such interval, but it is the shortest. Find another, longer.
5. Lead-time demand X for an item is approximated by a $N(25, 9)$ distribution. It is desired to compute the value for led time that will be exceeded only 5% of the time, that is find x_0 such that $P(X > x_0) = 0.05$.
6. The time to pass through a queue to begin self-service at a cafeteria has been found to be $N(15, 9)$. Compute the probability that an arriving customer waits between 14 and 17 minutes.

Lognormal distribution I

- If Y is $N(\mu, \sigma^2)$ then $X = e^Y$ has a lognormal distribution with parameters μ and σ (X is $\log N(\mu, \sigma^2)$)
- pdf

$$f(x|\mu,\sigma) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right], \qquad x > 0.$$

- cdf

$$F(x|\mu,\sigma) = \frac{1}{2}\mathrm{erfc}\left[-\frac{(\ln x - \mu)}{\sigma\sqrt{2}}\right] = \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$$

unde Φ este funcția lui Laplace, iar

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,dt$$

Lognormal distribution II

- Mean, variance

$$E(X) = e^{\mu+\sigma^2/2}$$
$$V(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$$

- If the mean and the variance are known to be μ_L and σ_L^2, respectively, then

$$\mu = \ln\left(\frac{\mu_L}{\sqrt{\mu_L^2+\sigma_L^2}}\right)$$
$$\sigma^2 = \ln\left(\frac{\mu_L^2+\sigma_L^2}{\mu_L^2}\right).$$

Lognormal distribution III

- median, mode

$$Mode(X) = e^{\mu - \sigma^2}$$
$$Median(X) = e^{\mu}$$

- A variable might be modeled as log-normal if it can be thought of as the multiplicative product of many independent random variables each of which is positive. For example, in finance, a long-term discount factor can be derived from the product of short-term discount factors. In wireless communication, the attenuation caused by shadowing or slow fading from random objects is often assumed to be log-normally distributed. Economists often model the distribution of income using a lognormal distribution.

Plot of lognormal pdfs

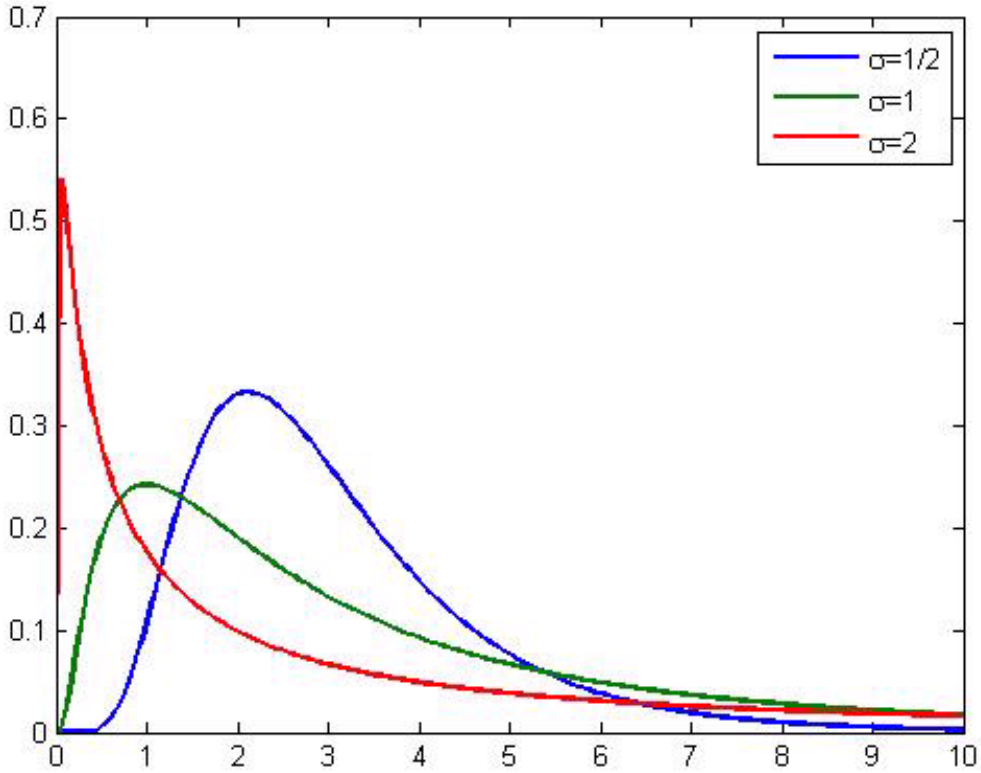


Figure: Plot of three lognormal pdfs for $\mu = 1$ and $\sigma = 1/2, 1, 2$

Applications

1. Suppose the income of a family of four in the United States follows a lognormal distribution with $\mu = ln(20,000\$)$ and $\sigma^2 = 1$. Plot the income density. What is the probability that the income be larger than 60000\$.
2. The rate of return on a volatile investment is modeled as having a lognormal distribution with mean 20% and standard deviation 5%. Compute the parameters for the lognormal distribution.

Beta distribution I

- A random variable X is said to have a beta distribution with parameters $\alpha, \beta > 0$ if and only if its density function is

$$f(x|\alpha,\beta) = \begin{cases} \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & x \in (0,1) \\ 0, & \text{otherwise} \end{cases}$$

where

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx$$

is the Euler's beta function.

Beta distribution II

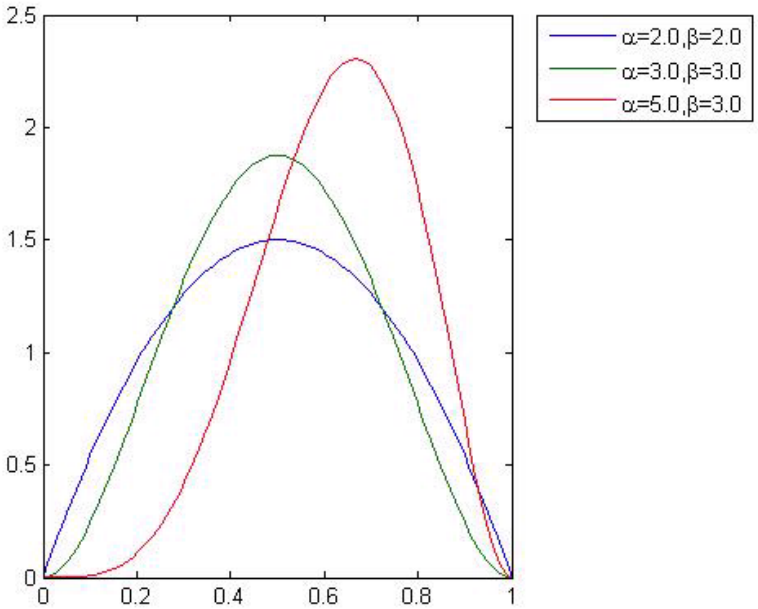
- mean, variance

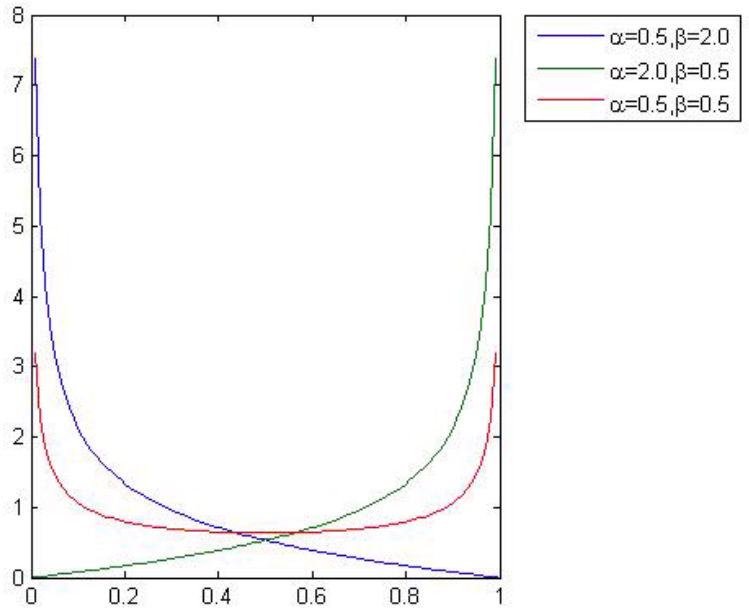
$$E(X) = \frac{\alpha}{\alpha + \beta}$$

$$V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

- The graphs of beta density exhibit a large variety of shapes for various values of the two parameters α and β (see Figure below). Beta distribution can be defined on an arbitrary interval (c, d): if $x \in (c, d)$ then $x* = (x - c)/(d - c)$ defines a new variable such that $x^* \in (0, 1)$.

Beta distribution III





Applications I

1. A gasoline wholesale distributor has bulk storage tanks that hold fixed supplies and are filled every Monday. Of interest to the wholesaler is the proportion of this supply that is sold during the week. Over many weeks of observation, the distributor that this proportion could be modelled by a beta distribution with $\alpha = 4$ and $\beta = 2$. Find the probability that the wholesaler will sent at least 90% of her stock in a given week.
2. **Rule of succession**. A classic application of the beta distribution is the rule of succession, introduced in the 18th century by Pierre-Simon Laplace in the course of treating the sunrise problem. It states that, given s successes in n conditionally independent Bernoulli trials with probability p, that p should be estimated as $\frac{s+1}{n+2}$. This estimate may be regarded as the expected value of the posterior distribution over p, namely $Beta(s+1, n-s+1)$, which is given by Bayes' rule if one

assumes a uniform prior over p (i.e., $Beta(1,1)$) and then observes that p generated s successes in n trials.

3. **Task duration modeling**. The beta distribution can be used to model events which are constrained to take place within an interval defined by a minimum and maximum value. For this reason, the beta distribution — along with the triangular distribution — is used extensively in PERT, critical path method (CPM) and other project management / control systems to describe the time to completion of a task. In project management, shorthand computations are widely used to estimate the mean and standard deviation of the beta distribution:

$$E(X) = \frac{a + 4b + c}{6}$$

$$\sigma(X) = \sqrt{V(X)} = \frac{c - a}{6}$$

Applications III

where a is the minimum, c is the maximum, and b is the most likely value. Using this set of approximations is known as three-point estimation and are exact only for particular values of α and β, specifically when:

$$\alpha = 3 - \sqrt{2}$$
$$\beta = 3 + \sqrt{2}$$

or vice versa. These are notably poor approximations for most other beta distributions exhibiting average errors of 40% in the mean and 54% in the variance.

Triangular distribution I

- X has a triangular distribution with parameters a, b, c, $a \leq b \leq c$ if its pdf is

$$f(x|a,b,c) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)}, & x \in [a,b] \\ \dfrac{2(c-x)}{(c-b)(c-a)}, & x \in (b,c] \\ 0, & \text{elsewhere} \end{cases}$$

- mean, mode

$$E(X) = \frac{a+b+c}{3}$$
$$Mode = b = 3E(X) - (a+c)$$

Triangular distribution II

- cdf

$$F(x|a,b,c) = \begin{cases} 0 & x \leq a \\ \dfrac{(x-a)^2}{(b-a)(c-a)} & x \in (a,b] \\ 1 - \dfrac{(c-x)^2}{(c-b)(c-a)} & x \in (b,c] \\ 1 & x > c \end{cases}$$

Applications I

1. A central processor unit requirements, for programs that will execute, have a triangular distribution with $a = 0.05$ ms, $b = 1.1$ ms, and $c = 6.5$ ms. Find the probability that the CPU requirements for a random program is 2.5 ms or less.
2. Implement MATLAB functions for the pdf, cdf, icdf of a triangular distribution.
3. An electronic sensor evaluates the quality of memory chips, rejecting those thai fail. Upon demand the sensor will give the minimum and maximum number of rejects over the past 24 hours. The mean is also given. Without further information, the quality control department has assumed thet the number of rejected chips can be approximated by a triangular distribution. The current dump data indicates that the minimum number of rejected chips during any hour was zero, the maximum was 10, and the mean was 4. Find a, b, and c and the median.

Applications II

4. Find conditions on a, b, c such that mean, mode and median be equal.
5. Find the variance and the median of a $Triang(a, b, c)$ r.v.

Exponential distribution I

- X is said to be exponential distributed with parameter $\lambda > 0$ if its pdf is

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

- cdf

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

- mean, variance, median

$$E(X) = \frac{1}{\lambda}$$

$$V(X) = \frac{1}{\lambda^2}$$

$$Me(X) = \frac{\ln 2}{\lambda}$$

Exponential distribution II

- variant (in statistical packages)

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$E(X) = \lambda, \qquad V(X) = \lambda^2$$

- Exponential distribution models interarrival times when arrivals are completely random and to model service time that are highly variable. In this instances λ is a rate: arrivals per hour or services per minute
- model the lifetime of a component that fails catastrophically (instantaneously), luch as a light bulb; λ is the failure rate.

Graphs

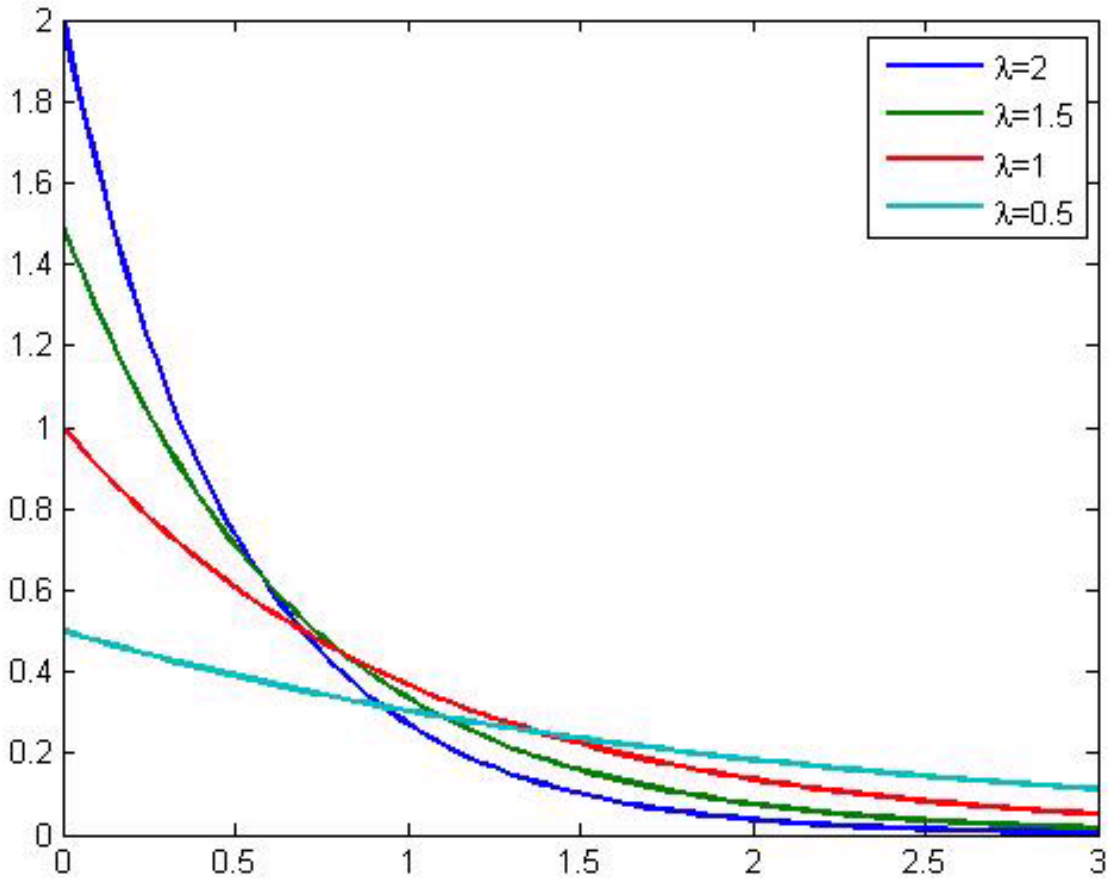


Figure: Exponential pdfs for several values of λ

Properties I

- Memoryless property:

$$P(X > s+t|X > s) = P(X > t)$$

- Proof

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

- The exponential distributions and the geometric distributions are the only memoryless probability distributions.
- An exponential r.v. remains exponential when multiplied by a constant

$$P(cX \leq x) = P\left(X \leq \frac{x}{c}\right) = 1 - e^{-\frac{\lambda}{c}x}$$

Properties II

- if $X_1, \ldots, X_n$ are e.i.r.v. with parameters $\lambda_1, \ldots, \lambda_n$ then $M = \min(X_1, \ldots, X_n)$ is exponential with $\sum_i \lambda_i$

$$P\left(X_j = \min_i X_i | M > t\right) = P\left(X_j - t = \min_i (X_i - t) | M > t\right)$$
$$= P\left(X_j - t = \min_i (X_i - t) | X_i > t, i = \overline{1, n}\right)$$
$$= P\left(X_j = \min_i X_i\right)$$

(we used the memoryless property)

$$P(M > t) = P(X_i > t, i = 1, \ldots, n) = e^{-\sum_{i=1}^{n} \lambda_i t}$$

Properties III

- Probability that X_j is the smallest (limits $0, \infty$)

$$
\begin{aligned}
P(X_j = M) &= \int P(X_j = M | X_j = t)\, \lambda_j e^{-\lambda_j t} dt \\
&= \int P(X_i > t, i \neq j | X_j = t)\, \lambda_j e^{-\lambda_j t} dt \\
&= \int P(X_i > t, i \neq j)\, \lambda_j e^{-\lambda_j t} dt \\
&= \int \left(\prod_{i \neq j} e^{-\lambda_i t} \right) \lambda_j e^{-\lambda_j t} dt \\
&= \lambda_j \int e^{-\sum_i \lambda_i t} dt = \frac{\lambda_j}{\sum_i \lambda_i}
\end{aligned}
$$

Applications

1. The median of the exponential distribution is $\frac{\ln 2}{\lambda}$. Prove this fact this fact.
2. What is the probability that an exponential random variable will be less than or equal to $1/E(X)$?
3. Let the lifetime of light bulbs be exponentially distributed with $\beta = 700$ hours. What is the median lifetime of a bulb?

Weibull distribution I

- X has a Weibull distribution with parameters $\nu \in \mathbb{R}$, $\alpha > 0$, $\beta > 0$ if its pdf is

$$f(x|\nu,\alpha,\beta) = \begin{cases} \frac{\beta}{\alpha}\left(\frac{x-\nu}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{x-\nu}{\alpha}\right)^{\beta}\right], & x \geq \nu \\ 0, & \text{otherwise} \end{cases}$$

- cdf

$$F(x) = \begin{cases} 0, & x < \nu \\ 1 - \exp\left[-\left(\frac{x-\nu}{\alpha}\right)^{\beta}\right], & x \geq \nu \end{cases}$$

- ν - location parameter, α - scale parameter, β - shape parameter
- $\nu = 0$ two parameter Weibull
- $\nu = 0$, $\beta = 1$, exponential with parameter $\lambda = \frac{1}{\alpha}$

Weibull distribution II

- mean

$$E(X) = \nu + \alpha\Gamma\left(\frac{1}{\beta} + 1\right)$$

$$V(X) = \alpha^2\left[\Gamma\left(\frac{2}{\beta} + 1\right) - \left(\Gamma\left(\frac{1}{\beta} + 1\right)\right)^2\right]$$

- an appropriate analytical tool for modeling the breaking strength of materials. Current usage also includes reliability and lifetime modeling. The Weibull distribution is more flexible than the exponential for these purposes.

Graphs

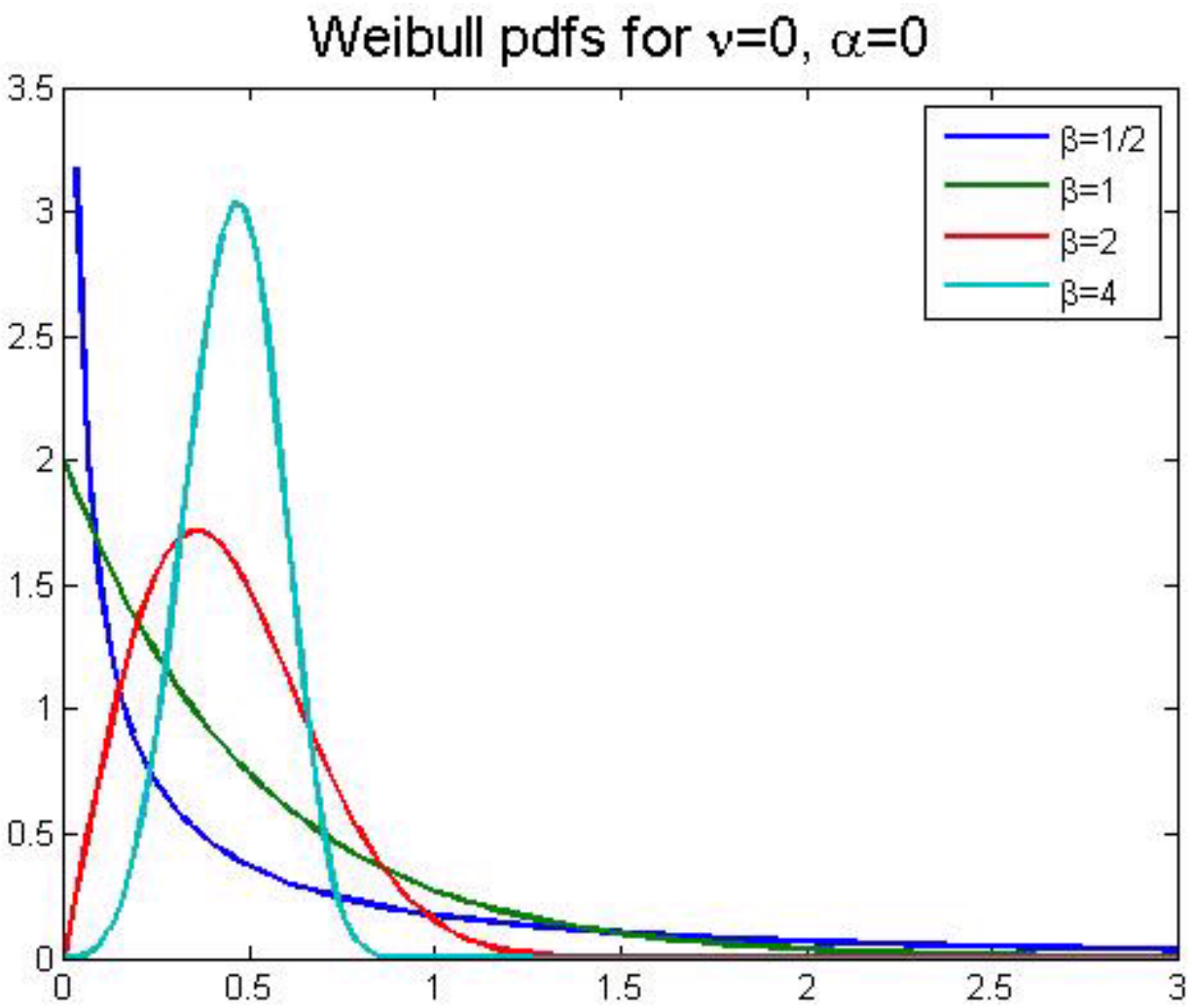


Figure: Weibul pdfs for $\nu = 0$ and $\alpha = \frac{1}{2}$ and various values for β

Applications

1. Reproduce Figure 3. Plot cdfs for the same distributions.
2. Find the formula for the median of a Weibull distribution.
3. The time to failure for a component screen is known to have a Weibull distribution with $\nu = 0$, $\beta = 1/3$, and $\alpha = 200$ hours. Find the mean, the variance and the probability that a unit fails before 2000 hours.
4. The time it takes for an aircraft to land and clear the runaway at a major international airport has a Weibull distribution with $\nu = 1.34$ minutes, $\beta = 0.5$ and $\alpha = 0.04$ minutes. Find the probability that an incoming airpalne will take more than 1.5 minutes to land and cler the runaway.