INFO 7374 Machine Learning in Finance Yizhen Zhao

Lecture 1 Basic Statistics

Random Variables

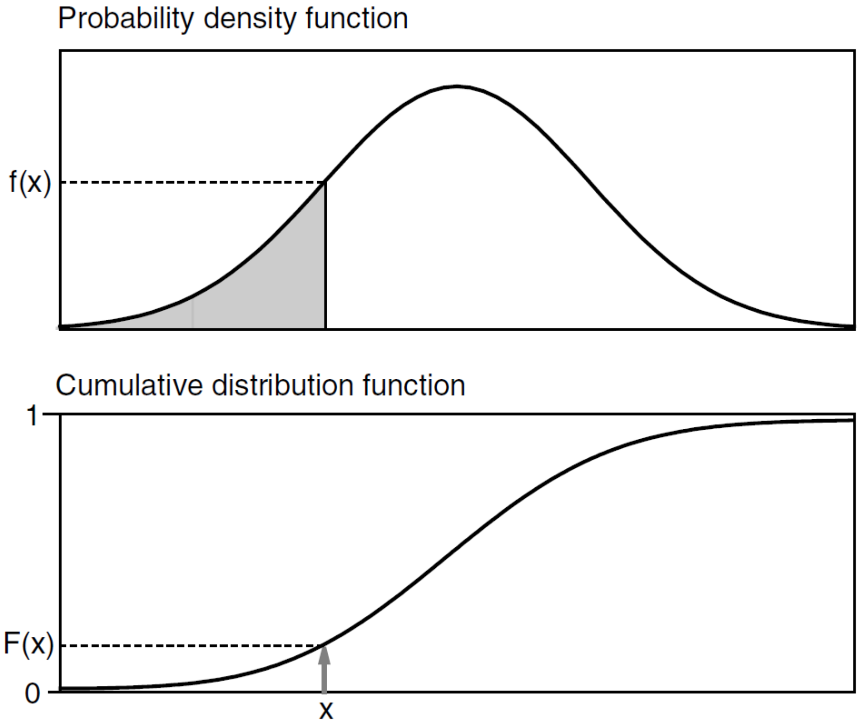
- A **random experiment** is defined as a process or action whose outcome cannot be predicted with certainty and would likely change when the experiment is repeated.
- The **sample space** is the set of all outcomes from an experiment.
- The outcomes from random experiments are often represented by an uppercase variable such as X. This is called a **random variable**, and its value is subject to the uncertainty intrinsic to the experiment.
- Random variables can be discrete or continuous. A ***discrete random variable*** can take on values from a finite or countably infinite set of numbers. Examples of discrete random variables are the number of defective parts or the number of typographical errors on a page. A ***continuous random variable*** is one that can take on values from an interval of real numbers.

  Discrete Random Variables: Letting 1 represent the *bull market* and letting 0 represent the *bear market*, then the probability of the event that *we are in the bull market* would be written as

$$\boldsymbol{P(X=1)}$$

  Continuous Random Variables: Let X denote the price of crude oil future: \$/barrel. The probability that the transaction price is in the range \$30 to \$50 is expressed as

$$\boldsymbol{P(\$30 < X \leq \$50)}.$$



Moments

Notation

- *n is the number of data points.*
- *x_i is each individual data point.*
- *$\bar{x}$ is the mean of the discrete data points. μ* is standard deviation of a continuous random variable.
- *s is the standard deviation of the discrete data points. σ* is standard deviation of a continuous random variable.

First Moment: The ***mean*** of a discrete data sample is defined as

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

The ***expected value*** of a random variable is defined using the probability density (mass) function. It provides a measure of central tendency of the distribution.

For a discrete random variable X with possible values $x_1, x_2, x_3, \dots, x_n$ and corresponding probabilities $p_1, p_2, p_3, \dots, p_n$, ***the first moment*** (the expected value) is defined as:

$$\mu = E[X] = \sum_{i=1}^{n} x_i \cdot p_i$$

We see from the definition that the expected value is a sum of all possible values of the random variable where each one is weighted by the probability that X will take on that value.

For a random variable X with *mean* μ, ***the first moment about the mean*** is defined as*:*

$$E[X - \mu]$$

The ***first moment about the mean*** of a random variable is a measure related to the average deviation of the variable from its mean.

For a continuous random variable X with a probability density function (PDF) $f(x)$, the first moment is defined as:

$$\mu = E[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

Here $f(x)$ is the probability density function of X. The integral is taken over the entire range of X.

Second Moment: The second moment of a random variable is a measure related to the variance and the mean of the random variable. It is the expected value of the square of the random variable. For a discrete random variable X with possible values $x_1, x_2, x_3, \dots, x_n$ and corresponding probabilities $p_1, p_2, p_3, \dots, p_n$, the second moment is defined as:

$$E[X^2] = \sum_{i=1}^{n} x_i^2 \cdot p_i$$

This measure captures the average of the squares of the values of X and is used to understand the dispersion and magnitude of X itself.

The ***second moment about the mean*** of a random variable X is defined as:

$$E[(X - \mu)^2]$$

This quantity represents the ***variance*** of the random variable $Var(X)$, which measures how much the values of X deviate from the mean. The variance is a measure of the spread or dispersion of the distribution around the mean.

The ***variance*** of a discrete random variable is defined as follow:

$$Var(X) = \sigma^2 = \frac{1}{n-1} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\mathbf{2}}$$

Variance is a measure of dispersion in the distribution. If a random variable has a large variance, then an observed value of the random variable is more likely to be far from the mean μ. The standard deviation σ, or alternatively denoted as s, is the square root of the variance.

Correspondingly, the ***standard deviation*** is defined as

$$s = \sqrt{\frac{1}{n-1} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\mathbf{2}}}$$

The second moment is related to the variance $Var(X)$ and the mean μ of the random variable. Specifically, the variance is given by:

$$Var(X) = E[X^2] - \mu^2$$

For a continuous random variable X with a probability density function (PDF) $f(x)$, the second moment is defined as:

$$E[X^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx$$

Here, $E[X^2]$ is the second moment of the random variable X. $f(x)$ is the probability density function of X. The integral is taken over the entire range of X.

Third Moment: The third moment of a random variable measures the asymmetry or skewness of the distribution. It is the expected value of the cube of the random variable. For a discrete random variable X with possible values $x_1, x_2, x_3, \dots, x_n$ and corresponding probabilities $p_1, p_2, p_3, \dots, p_n$, the third moment is defined as:

$$E[X^3] = \sum_{i=1}^{n} x_i^3 \cdot p_i$$

The ***skewness*** of the distribution can be derived from **the third moment about the mean**

$$E[(X-\mu)^3]$$

The skewness γ is defined as:

$$\gamma = \frac{E[(X-\mu)^3]}{\sigma^3}$$

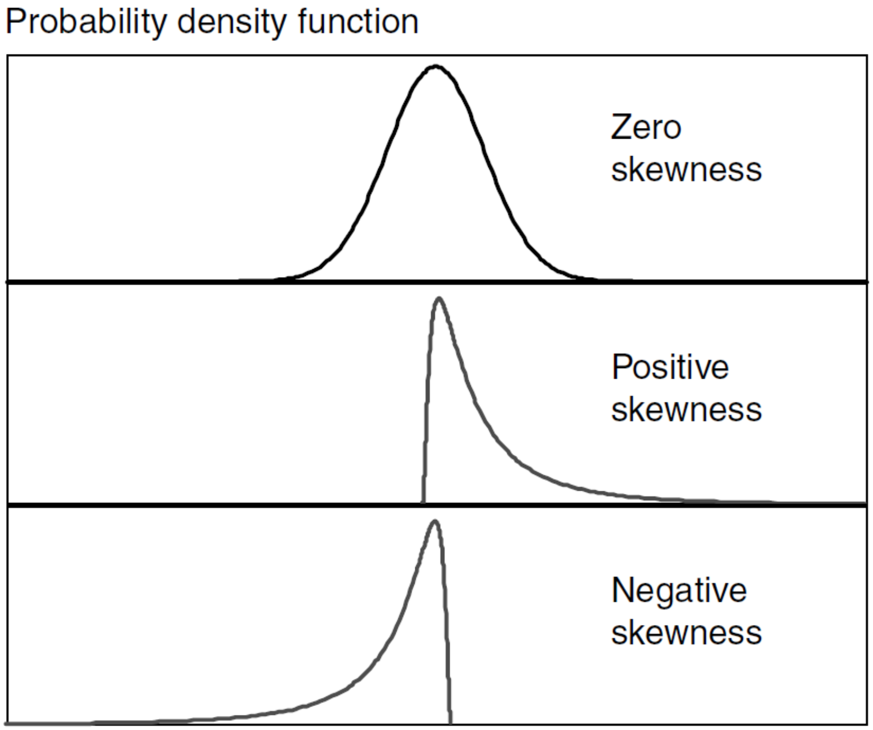
In the simplest case,

$$E[(X-\mu)^3] = \frac{1}{n}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \boldsymbol{\mu})^{\boldsymbol{3}}$$

Here σ is the standard deviation and defined above:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \boldsymbol{\mu})^{\boldsymbol{2}}}$$

Distributions that are skewed to the left will have a negative coefficient of skewness, and distributions that are skewed to the right will have a positive value. The coefficient of skewness is zero for symmetric distributions. However, a coefficient of skewness equal to zero does not mean that the distribution must be symmetric. The uniform distribution and the normal distribution are examples of symmetric distributions. The gamma and the exponential are examples of skewed or asymmetric distributions.



For a continuous random variable X with a probability density function (PDF) $f(x)$, the ***third moment about the mean*** is defined as:

$$E[(X-\mu)^3] = \int_{-\infty}^{+\infty} (x-\mu)^3 f(x) dx$$

Here, μ is the mean of the random variable X. $f(x)$ is the probability density function of X. The integral is taken over the entire range of X.

Fourth Moment: The ***fourth moment about the mean*** of a discrete random variable measures the "thickness of tails" or kurtosis of the distribution, which provides information about the extremity or outliers of the data. For a discrete random variable X with possible values $x_1, x_2, x_3, \dots, x_n$ and corresponding probabilities $p_1, p_2, p_3, \dots, p_n$, the ***fourth moment about the mean*** is defined as:

$$E[(X-\mu)^4] = \sum_{i=1}^{n} (x_i-\mu)^4 \cdot p_i$$

Kurtosis is a measure that quantifies the shape of the distribution's tails. It is derived from the fourth moment and is standardized to compare with the normal distribution.

$$\delta = \frac{E[(X-\mu)^4]}{\sigma^4}$$

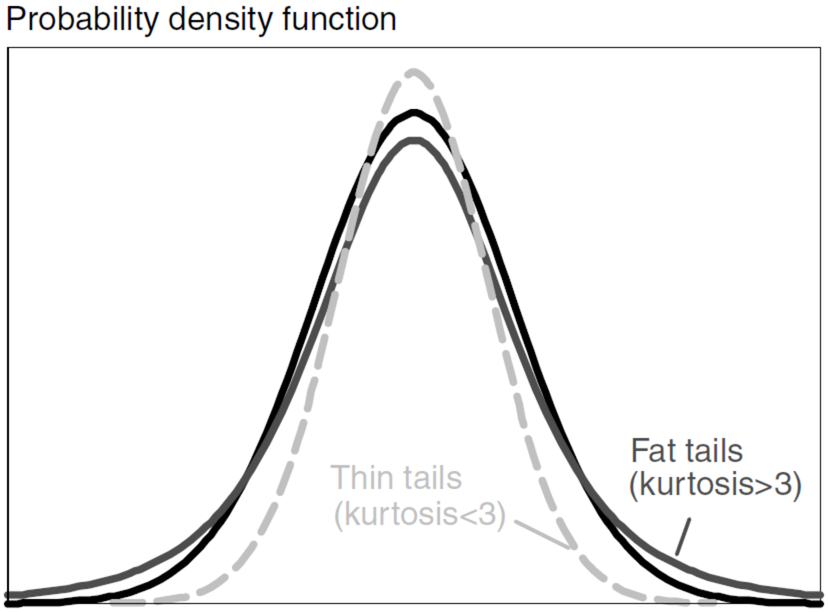
The excess kurtosis is defined as:

$$\delta' = \frac{E[(X-\mu)^4]}{\sigma^4} - 3$$

Here σ^2 is the variance of the distribution. We see that this is the ratio of the fourth central moment divided by the square of the variance. Because of the fourth power, observations far away from the mean will have a larger weight and hence create large kurtosis.

For a normal distribution, the excess kurtosis is 0, and the kurtosis is 3. This serves as a baseline for comparing other distributions.

- Heavy-Tailed Distributions (***leptokurtic***): Distributions with heavy tails (e.g., Cauchy distribution) have positive excess kurtosis, indicating more extreme values than a normal distribution.
- Light-Tailed Distributions: Distributions with light tails (e.g., uniform distribution) have negative excess kurtosis, indicating fewer extreme values than a normal distribution.



Quantile: The distribution can also be described by its ***Quantile***, which is the cutoff point x with an associated probability c:

$$F(x) = \int_{-\infty}^{x} f(u)du = c$$

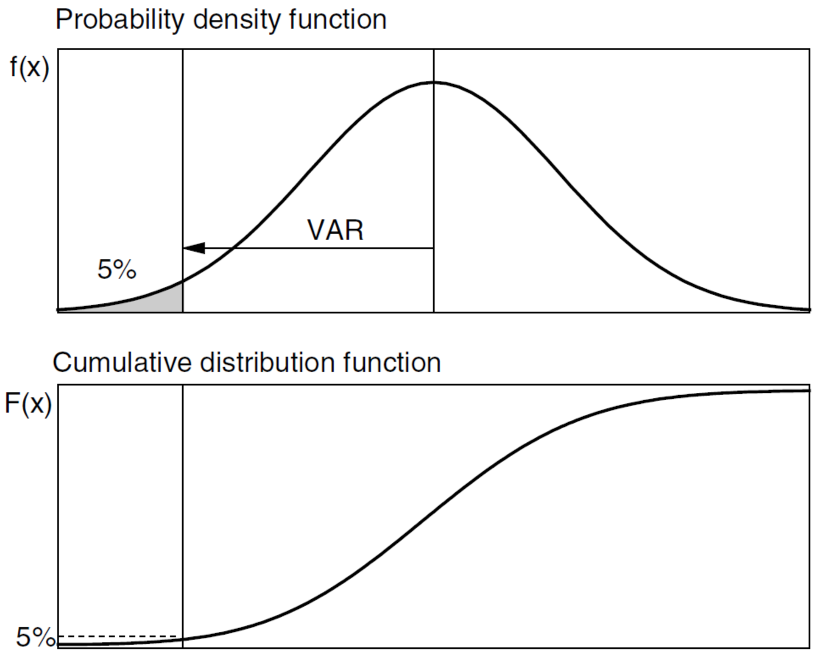
So, there is a probability of c that the random variable will fall below x. Because the total probability adds up to one, there is a probability of $p = 1 - c$ that the random variable will fall above x. Define this quantile as $Q(X, c)$. The 50% quantile is known as the ***median***.

Value At Risk (VAR) can be interpreted as the cutoff point such that a loss will not happen with probability greater than $p = 95\%$ percent, say. If $f(u)$ is the distribution of profit and losses on the portfolio, VAR is defined from

$$F(x) = \int_{-\infty}^{x} f(u)du = 1 - p$$

where is the right-tail probability, and c usual left-tail probability. VAR can then be defined as the deviation between the expected value and the quantile,

$$VAR(c) = E(X) - Q\ (X, c)$$



Common Distributions

Uniform

Perhaps one of the most important distributions is the uniform distribution for continuous random variables. One reason is that the uniform $(0,1)$ distribution is used as the basis for simulating most random variables.

A random variable that is uniformly distributed over the interval (a, b) follows the probability density function given by

$$f(x; a, b) = \frac{1}{b - a}$$
$$a < x < b$$

The parameters for the uniform are the interval endpoints, a and b. The mean and variance of a uniform random variable are given by

$$E[X] = \frac{a+b}{2}$$

and

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

The cumulative distribution function for a uniform random variable is

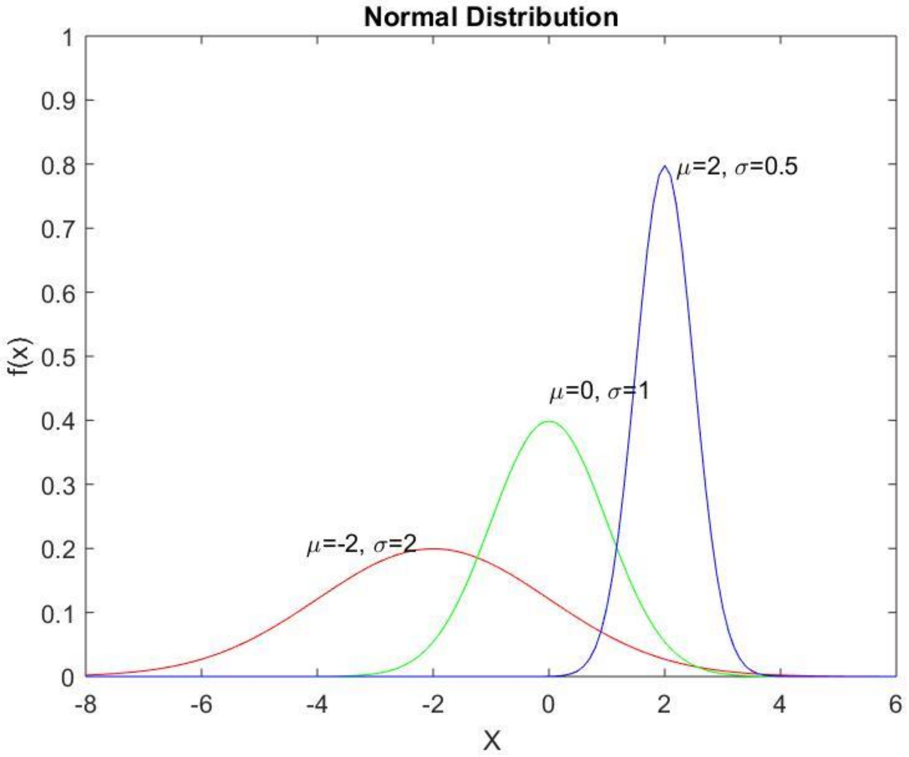
$$F(x) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}$$

Normal

A well-known distribution in statistics and engineering is the normal distribution. Also called the Gaussian distribution, it has a continuous probability density function given by

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

The normal distribution is completely determined by its parameters (μ and σ^2), which are also the expected value and variance for a normal random variable. The notation $X \sim N(\mu,\sigma^2)$ is used to indicate that a random variable X is normally distributed with mean μ and variance σ^2. Several normal distributions with different parameters are shown below.



Some special properties of the normal distribution are given here.

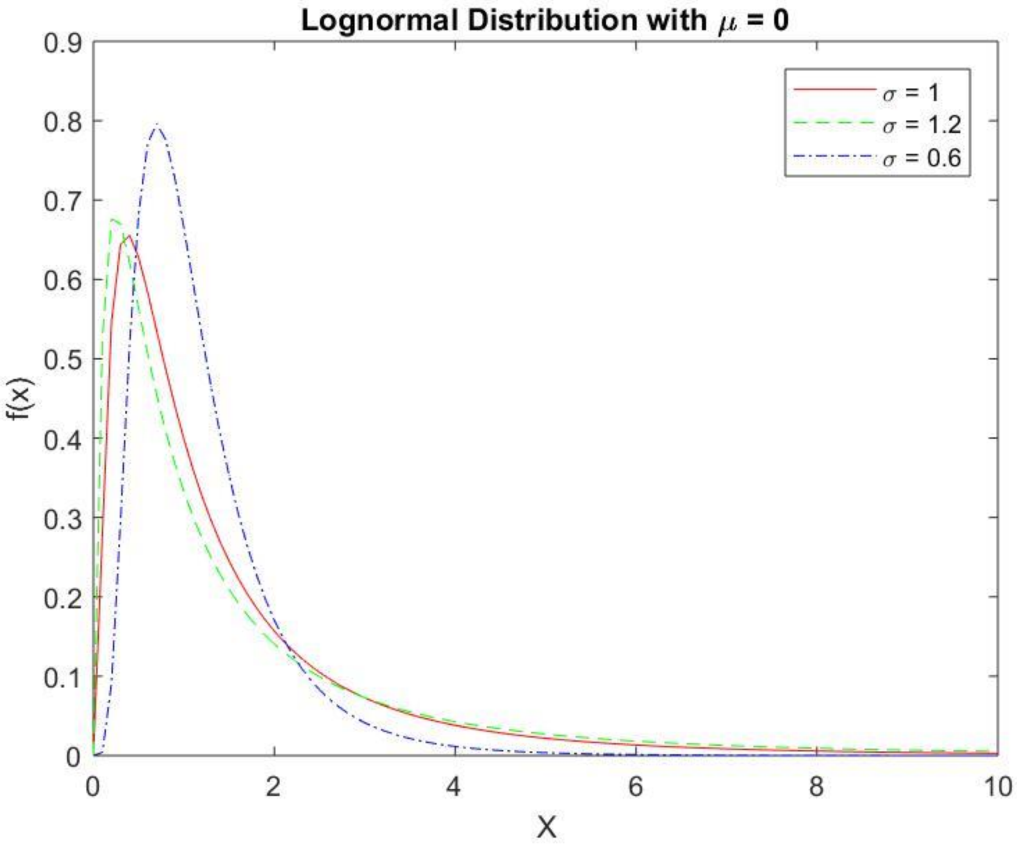
- The value of the probability density function approaches zero as x approaches positive and negative infinity.
- The probability density function is centered at the mean , and the maximum value of the function occurs at $x = \mu$.
- The probability density function for the normal distribution is symmetric about the mean μ.

The special case of a standard normal random variable is one whose mean is zero ($\mu = 0$), and whose standard deviation is one ($\sigma = 1$). If X is normally distributed, then

$$Z = \frac{X - \mu}{\sigma} \sim \text{ N } (0,1)$$

is a standard normal random variable.

Log-Normal Distribution



The normal distribution is a good approximation for many financial variables, such as the rate of return on a stock, $r = (P_1 - P_0)/P_0$, where P_0 and P_1 are the stock prices at time 0 and 1.

Strictly speaking, this is inconsistent with reality since a normal variable has infinite tails on both sides. Due to the limited liability of corporations, stock prices cannot turn negative. This rules out returns lower than minus unity and distributions with infinite left tails, such as the normal distribution. In many situations, however, this is an excellent approximation. For instance, with short horizons or small price moves, the probability of having a negative price is so small as to be negligible.

If this is not the case, we need to resort to other distributions that prevent prices from going negative. One such distribution is the lognormal.

A random variable X is said to have a **lognormal distribution** if its logarithm $Y = \ln(X)$ is normally distributed. This is often used for continuously compounded returns, defining $Y = \ln(P_1/P_0)$. Because the argument X in the logarithm function must be positive, the price can never go below zero. Large and negative large values of Y correspond to P_1 converging to, but staying above, zero.

The lognormal density function has the following expression

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

Note that this is more complex than simply plugging $\ln(x)$ in Equation above, because x also appears in the denominator. Its mean is

$$E[X] = \exp[\mu + \frac{1}{2}\sigma^2]$$

and variance

$$V[X] = \exp[2\mu + 2\sigma^2] - \exp[2\mu + \sigma^2].$$

The parameters were chosen to correspond to those of the normal variable,

$$E[Y] = E[\ln(X)] = \mu$$

and

$$V[Y] = V[\ln(X)] = \sigma^2.$$

Conversely, if we set $E[X] = \exp[r]$, the mean of the associated normal variable is

$$E[Y] = E[\ln(X)] = (r - \sigma^2/2).$$

This adjustment is also used in the Black-Scholes option valuation model, where the formula involves a trend in $(r - \sigma^2/2)$ for the log-price ratio.

We also note that the distribution of the bond price, resembles a lognormal distribution. Using continuous compounding instead of annual compounding, the price function is

$$V = 100\exp(-rT)$$

which implies $\ln\left(\frac{V}{100}\right) = -rT$. Thus if r is normally distributed, V has a log normal distribution.

Students$-t$ Distribution

Another important distribution is the Student's t-distribution. This arises in hypothesis testing, because it describes the distribution of the ratio of the estimated coefficient to its standard error.

The distribution is characterized by a parameter k known as the degrees of freedom. Its density is

$$f(x) = \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)} \frac{1}{\sqrt{k\pi}} \frac{1}{\left(1+\frac{x^2}{k}\right)^{(k+1)/2}}$$

where Γ is the gamma function, defined as

$$\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx.$$

As k increases, this function converges to the normal p.d.f.

The mean and standard deviation of the Student's t-distribution depend on the degrees of freedom k.

- Mean: The mean of the Student's t-distribution is 0, provided that the degrees of freedom are greater than 1 ($k > 1$). For the degrees of freedom less than 1 ($k \leq 1$), the mean is undefined because the distribution does not have a finite mean.
- Standard Deviation: The standard deviation σ of the Student's t-distribution is given by:

  $$\sigma = \sqrt{\frac{k}{k-2}}$$

  This is valid for degrees of freedom greater than 2 ($k > 2$). For degrees of freedom less than or equal to 2 ($k \leq 2$), the standard deviation is undefined because the distribution does not have a finite variance.

Correspondingly, the variance is (provided $k > 2$):

$$V[X] = \frac{k}{k-2}$$

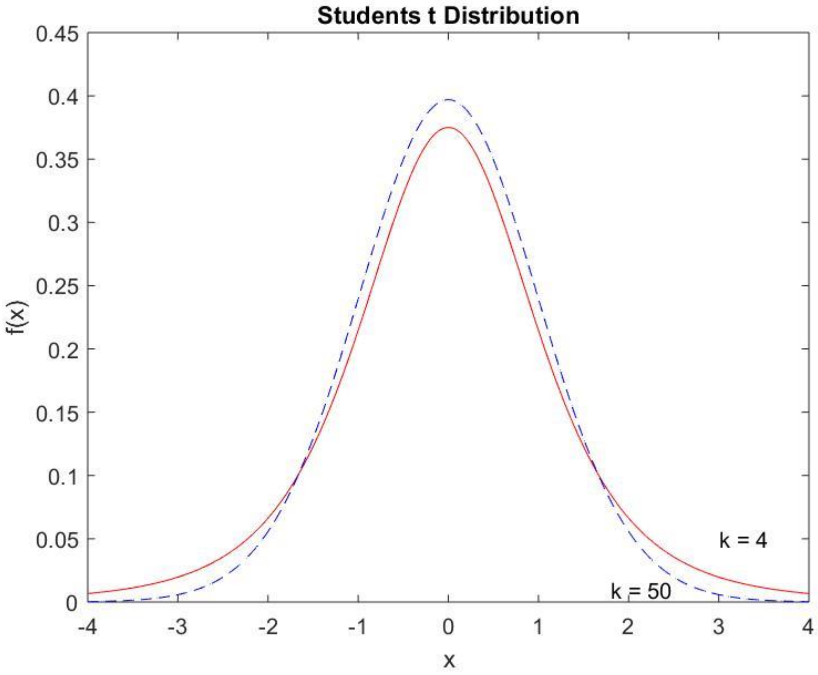
The kurtosis is (provided $k > 4$):

$$\delta = 3 + \frac{6}{k-4}$$

. It has fatter tail than the normal distribution, which often provides a better representation of typical financial variables. Typical estimated values of k are around 4 to 6 for stock returns. We can also use the Student's t to compute Value-At-Risk as function of volatility.

$$VAR = \alpha_k \sigma$$

where the multiplier α_k now depends on the degrees of freedom k.



Exponential

The exponential distribution can be used to model the amount of time until a specific event occurs or to model the time between independent events. Some examples where an exponential distribution could be used as the model are

- the time until the computer locks up,
- the time between arrivals of telephone calls, or
- the time until a part fails.

The exponential probability density function with parameter λ is:

$$\mathrm{f(x;\lambda) = \lambda e^{-\lambda x}}$$
$$\mathrm{x \geq 0}\ ;$$
$$\lambda > 0.$$

The mean and variance of an exponential random variable are given by the following:

$$E[X] = \frac{1}{\lambda},$$

and

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

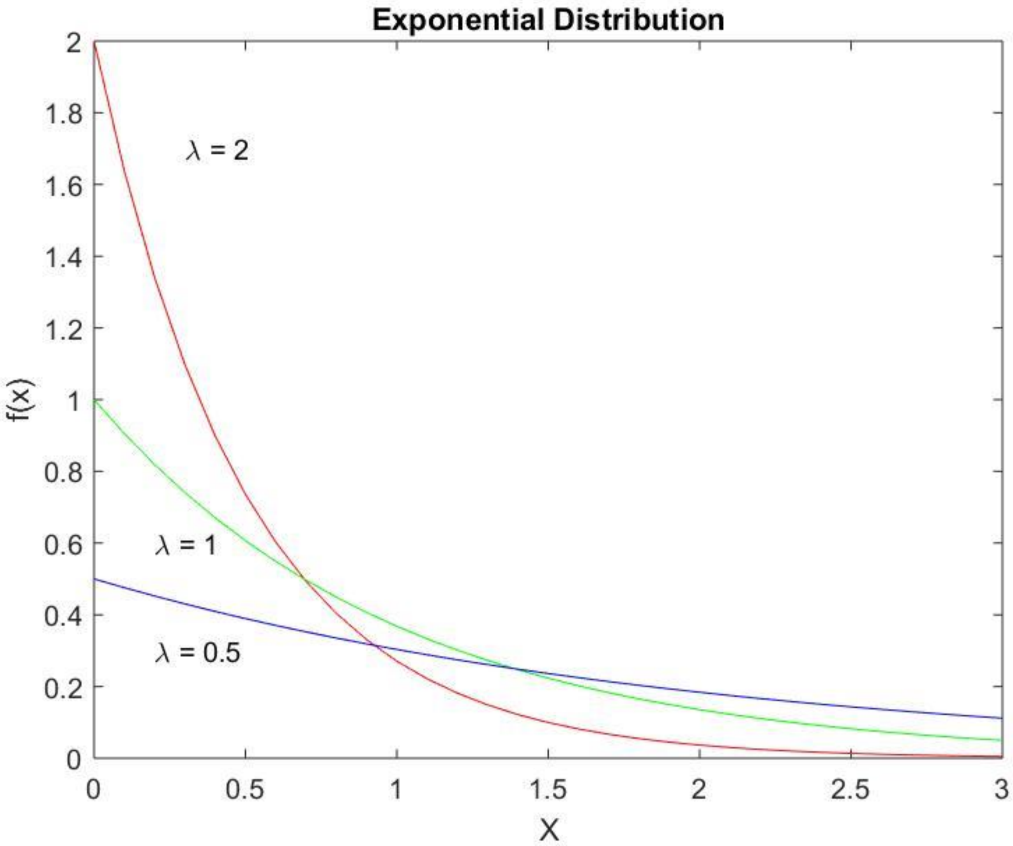
The cumulative distribution function of an exponential random variable is given by

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

The exponential distribution is the only continuous distribution that has the memory less property. This property describes the fact that the remaining lifetime of an object (whose lifetime follows an exponential distribution) does not depend on the amount of time it has already lived. This property is represented by the following equality, where $s \geq 0$ and $t \geq 0$:

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

In words, this means that the probability that the object will operate for time, given it has already operated for time s, is simply the probability that it operates for time t. When the exponential is used to represent inter-arrival times, then the parameter is a rate with units of arrivals per time period. When the exponential is used to model the time until a failure occurs, then is the failure rate. Several examples of the exponential distribution are shown in Example below.



Gamma

The gamma probability density function with parameters $\lambda > 0$ and $t > 0$ is

$$f(x; \lambda, t) = \frac{\lambda e^{-\lambda x}(\lambda x)^{t-1}}{\Gamma(t)}$$

$$x \geq 0$$

where t is a shape parameter, and λ is the scale parameter. The gamma function $\Gamma(t)$ is defined as

$$\Gamma(t) = \int_0^\infty e^{-y} y^{t-1} dy.$$

For integer values of t, Equation above becomes:

$$\Gamma(t) = (t-1)!.$$

Note that for $t = 1$, the gamma density is the same as the exponential. When t is a positive integer, the gamma distribution can be used to model the amount of time one has to wait until $t$events have occurred, if the inter-arrival times are exponentially distributed.

The mean and variance of a gamma random variable are:

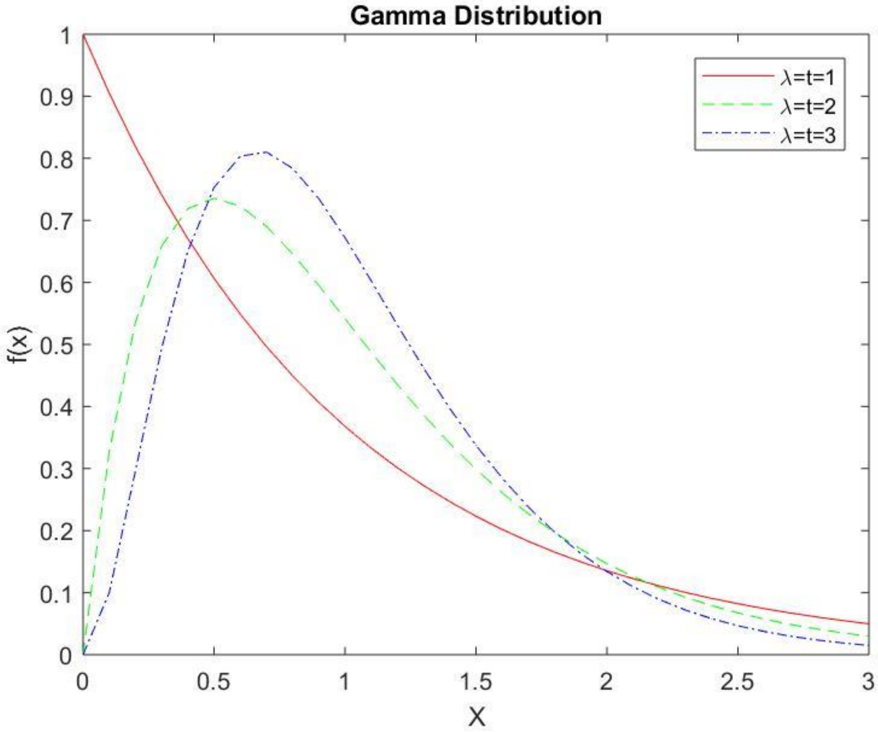
$$E[X] = \frac{t}{\lambda}$$

and

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

The cumulative distribution function for a gamma random variable is calculated using:

$$F(x; \lambda, t) = \begin{cases} 0 & x \leq 0 \\ \dfrac{1}{\Gamma(t)} \displaystyle\int_0^{\lambda x} y^{t-1} e^{-y}\, dy & x > 0 \end{cases}$$



Chi-Square

A gamma distribution where $\lambda = 0.5$and $t = \frac{\nu}{2}$, with ν a positive integer, is called a chi-square distribution (denoted as χ^2_ν) with ν degrees of freedom. The chi-square distribution is used to derive the distribution of the sample variance and is important for goodness-of-fit tests in statistical analysis.

The probability density function for a chi-square random variable with ν degrees of freedom is:

$$f(x;\nu) = \frac{1}{\Gamma(\nu/2)}\left(\frac{1}{2}\right)^{\nu/2} x^{\frac{\nu}{2}-1}e^{-\frac{1}{2}x}$$

$$x \geq 0.$$

The mean and variance of a chi-square random variable can be obtained from the gamma distribution. These are given by:

$$E[X] = \nu$$

and

$$V(X) = 2\nu$$

Weibull

The Weibull distribution has many applications in engineering. In particular, it is used in reliability analysis. It can be used to model the distribution of the amount of time it takes for objects to fail. For the special case where $\nu = 0$ and $\beta = 1$, the Weibull reduces to the exponential with $\lambda = 1/\alpha$.

The Weibull density for $\alpha > 0$ and $\beta > 0$ is given by

$$f(x;\nu,\alpha,\beta) = \left(\frac{\beta}{\alpha}\right)\left(\frac{x-\nu}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x-\nu}{\alpha}\right)^{\beta}}$$

$$x > \nu$$

and the cumulative distribution is

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

The location parameter is denoted by ν and the scale parameter is given by α. The shape of the Weibull distribution is governed by the parameter β. The mean and variance of a random variable from a Weibull distribution are given by

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

and

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

Beta

The beta distribution has support on unit interval. It can be used to model a random variable that takes on values over a bounded interval. It has two parameters α and β that determines the shape of the density. A random variable has a beta distribution with parameters $\alpha > 0$ and $\beta > 0$ if its probability density function is given by

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$0 < x < 1$$

where

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The mean and variance of a beta random variable are

$$E[X] = \frac{\alpha}{\alpha+\beta}$$

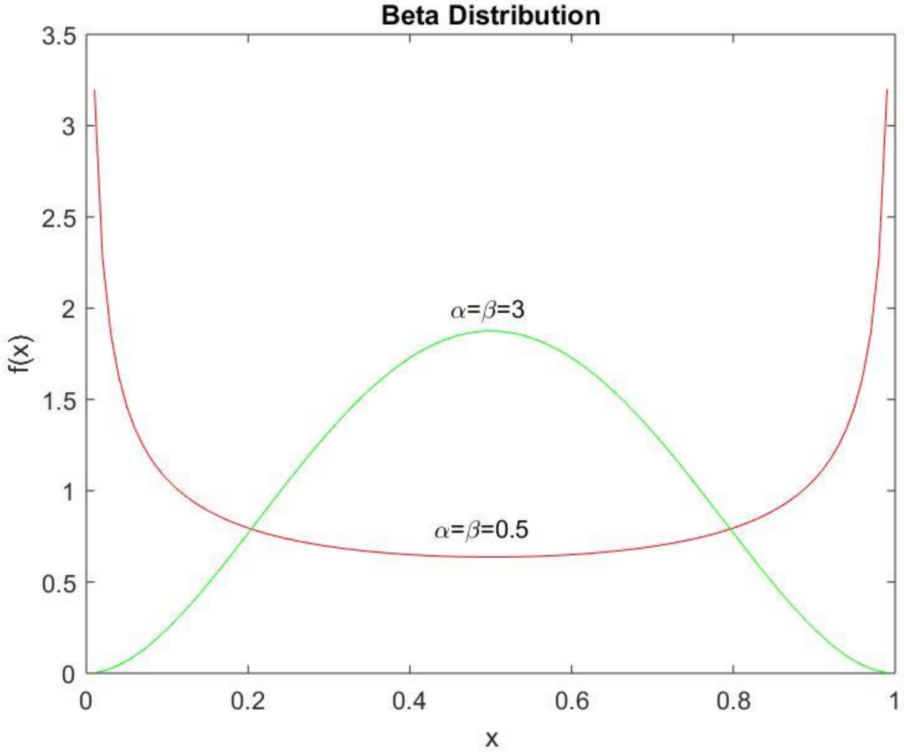
and

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

The cumulative distribution function for a beta random variable is given by integrating the beta probability density function as follows

$$F(x;\alpha,\beta) = \int_0^x \frac{1}{B(\alpha,\beta)} y^{\alpha}(1-y)^{\beta-1}dy.$$

The integral in Equation above is called the incomplete beta function.



Multivariate Normal

So far, we have discussed several univariate distributions for discrete and continuous random variables. In this section, we describe one of the important and most commonly used multivariate densities: the multivariate normal distribution. This distribution is used throughout the rest of the text.

Some examples of where we use it are in exploratory data analysis, in probability density estimation, and in statistical pattern recognition. The probability density function for a general multivariate normal density for d dimensions is given by

$$f(x,\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

where x is a d-component column vector, μ is the $d \times 1$ column vector of means, and Σ is the $d \times d$covariance matrix.

For example, when $d = 2$, $\mu = (\mu_1, \mu_2)^T$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

The superscript T represents the transpose of an array.

The mean and covariance are calculated using the following formulas:

$$\mu = \mathrm{E}[\mathrm{x}],$$

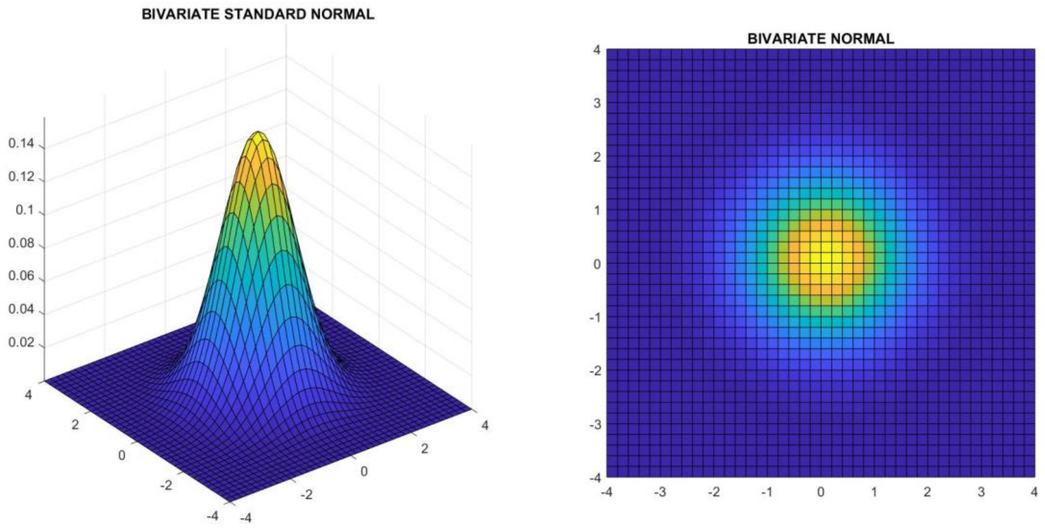
and

$$\Sigma = \mathrm{E}[(\mathrm{x} - \mu)(\mathrm{x} - \mu)^{\mathrm{T}}],$$

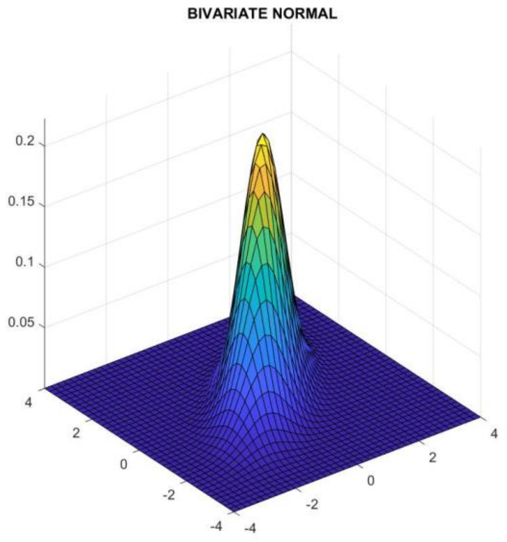
where the expected value of an array is given by the expected values of its components. The covariance matrix is symmetric ($\Sigma^T = \Sigma$) positive definite (all eigenvalues of Σ are greater than zero) for most applications of interest to statisticians and engineers.

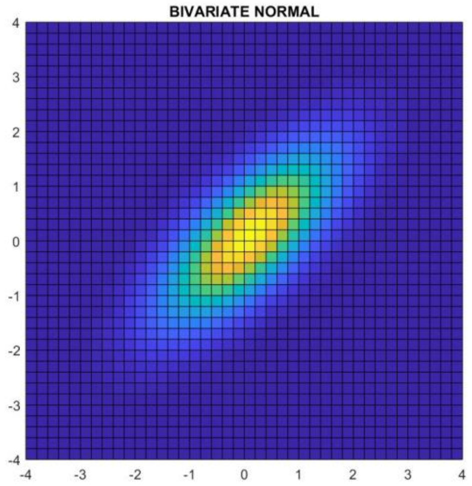
We illustrate some properties of the multivariate normal by looking at the bivariate ($d = 2$) case. The probability density function for a bivariate normal is represented by a bell-shaped surface. The center of the surface is determined by the mean μ and the shape of the surface is determined by the covariance Σ.

- If the covariance matrix is diagonal (allof the off-diagonal elements are zero), and the diagonal elements are equal, then the shape is circular.
- If the diagonal elements are not equal, then we get an ellipse with the major axis vertical or horizontal. If the covariance matrix is not diagonal, then the shape is elliptical with the axes at an angle. Some of these possibilities are illustrated in the next example.



This figure shows a standard bivariate normal probability density function that is centered at the origin. The covariance matrix is given by the identity matrix. Notice that the shape of the surface looks circular. The plot on the right is for a viewpoint looking down on the surface.





This shows a bivariate normal density where the covariance matrix has non-zero off-diagonal elements. Note that the surface has an elliptical shape. The plot on the right is for a viewpoint looking down on the surface.

Generating Random Variables

Uniform Random Numbers

Most methods for generating random variables start with random numbers that are uniformly distributed on the interval $(0, 1)$. We will denote these random variables by the letter u. With the advent of computers, we now have the ability to generate uniform random variables very easily.

It should be noted that random numbers that are uniformly distributed over an interval a to b may be generated by a simple transformation, as follows

$$X = (b - a) \bullet u + a$$

where

$$u \sim U\ (0, 1)\ and\ X \sim U\ (a, b).$$

Normal Random Variables

Given a standard normal random variable $Z \sim N\ (0, 1)$, we can obtain any normally distributed random variable X with mean $\mu$and variance$\sigma^2$by means of a transformation:

$$X = \mu\ +\ Z \cdot \sigma$$

Then $X \sim N\ (\mu, \sigma^2)$.

Jointly Normal Random Variables

Suppose we generate standard normal random variables X and Y with correlation ρ

- Let X be standard normal.
- Let U be standard normal (independent of X).
- Let $Y = \rho X + \sqrt{1-\rho^2}U$
- $E(Y) = 0,\ Var(Y) = \rho^2 + 1 - \rho^2 = 1,$
- $Cov(X, Y) = E(XY) = \rho Var(X) = \rho$

So X and Y are standard normal with correlation ρ.

Inverse Transform Method

The inverse transform method can be used to generate random variables from a continuous distribution. It uses the fact that the cumulative distribution function $F$is uniform $(0, 1)$.

$U = F\ (X)$.

If U is a uniform (0,1) random variables, then we can obtain the desired random variable X from the following relationship.

$$X = F^{-1}(U).$$

We see an example of how to use the inverse transform method when we discuss generating random variables from the exponential distribution (see the following Example). The general procedure for the inverse transformation method is outlined here.

PROCEDURE - INVERSE TRANSFORM METHOD (CONTINUOUS)

1. Derive the expression for the inverse distribution function $F^{-1}(U)$.
2. Generate a uniform random number U.
3. Obtain the desired X from $X = F^{-1}(U)$.

This same technique can be adapted to the discrete case. Say we would like to generate a discrete random variable X that has a probability mass function given by

$$P(X = x_i) = p_i; \qquad x_1 < x_2 < \cdots; \qquad \sum_i p_i = 1.$$

We get the random variables by generating a random number U and then deliver the random number X according to the following

$$X = x_i \qquad \text{if } F(x_{i-1}) < U \le F(x_i).$$

Example: Exponential Distribution

The inverse transform method can be used to generate random variables from the exponential distribution and serves as an example of this procedure. The distribution function for an exponential random variable with parameter $\lambda$is given by

$$F(x) = 1 - e^{-\lambda x} \qquad 0 < x < \infty.$$

Letting

$$u = F(x) = 1 - e^{-\lambda x},$$

we can solve for x, as follows

$$\begin{gathered} u = 1 - e^{-\lambda x} \\ e^{-\lambda x} = 1 - u \\ -\lambda x = \ log(1-u) \\ x = \ -\frac{1}{\lambda} log(1-u). \end{gathered}$$

By making note of the fact that $1 - u$ is also uniformly distributed over the
interval $(0,1)$, we can generate exponential random variables with parameter λ
using the transformation
$X = \ -\frac{1}{\lambda}$log (U).

Acceptance-Rejection Method

In some cases, we might have a simple method for generating a random variable from one density, say $g(y)$, instead of the density we are seeking. We can use this density to generate from the desired continuous density $f(x)$. We first generate a random number Y from $g(y)$ and accept the value with a probability proportional to the ratio $f(Y)/g(Y)$.

If we define *c* as a constant that satisfies

$$\frac{f(y)}{g(y)} \leq c; \qquad\qquad for\ all\ y,$$

then we can generate the desired variates using the procedure outlined below. The constant *c* is needed because we might have to adjust the height of $g(y)$ to ensure that it is above $f(y)$. We generate points from $cg(y)$, and those points that are inside the curve $f(y)$ are accepted as belonging to the desired density. Those that are outside are rejected. It is best to keep the number of rejected variates small for maximum efficiency.

1. Choose a density $g(y)$ that is easy to sample from.
2. Find a constant c such that Equation $\frac{f(y)}{g(y)} \leq c$ is satisfied.
3. Generate a random number Y from the density $g(y)$.
4. Generate a uniform random number U.
5. If

$$U \leq \frac{f(Y)}{cg(Y)},$$

then accept , else go to step 3.

Metropolis-Hastings Algorithms

The Metropolis-Hastings method is a generalization of the Metropolis technique of Metropolis, et al. [1953], which had been used for many years in the physics community. The paper by Hastings [1970] further generalized the technique in the context of statistics. The Metropolis sampler, the independence sampler and the random-walk are all special cases of the Metropolis-Hastings method. Thus, we cover the general method first, followed by the special cases.

These methods share several properties, but one of the more useful properties is that they can be used in applications where is known up to the constant of proportionality. Another property that makes them useful in a lot of applications is that the analyst does not have to know the conditional distributions, which is the case with the Gibbs sampler. While it can be shown that the Gibbs sampler is a special case of the Metropolis-Hastings algorithm. We include it in the next section because of this difference.

Metropolis-Hastings Sampler

The Metropolis-Hastings sampler obtains the state of the chain at by sampling a ***candidate point Y*** from a ***proposal distribution*** . Note that this depends only on the previous state and can have any form, subject to regularity conditions. An example for is the multivariate normal with mean and fixed covariance matrix. One thing to keep in mind when selecting is that the proposal distribution should be easy to sample from.

The required regularity conditions for are irreducibility and aperiodicity [Chib and Greenberg, 1995]. ***Irreducibility*** means that there is a positive probability that the Markov chain can reach any non-empty set from all starting points. ***Aperiodicity*** ensures that the chain will not oscillate between different sets of states. These conditions are usually satisfied if the proposal distribution has a positive density on the same support as the target distribution. They can also be satisfied when the target distribution has a restricted support. For example, one could use a uniform distribution around the current point in the chain.

The candidate point is accepted as the next state of the chain with probability given by

$$\alpha(X_t, Y) = \min\{1, \frac{\pi(Y)q(X_t|Y)}{\pi(X_t)q(Y|X_t)}\}.$$

If the point ***Y*** is not accepted, then the chain does not move and $X_{t+1} = X_t$. The steps of the algorithm are outlined below. It is important to note that the distribution of interest $\pi(x)$ appears as a ratio, so the constant of proportionality cancels out. This is one of the appealing characteristics of the Metropolis-Hastings sampler, making it appropriate for a wide variety of applications.

1. Initialize the chain to X_o and set t = 0.
2. Generate a candidate point $\boldsymbol{Y}$ from $q(.|X)$.
3. Generate U from a uniform $(0, 1)$ distribution.
4. If $U \leq \alpha(X_t, Y)$ then set $X_{t+1} = Y$, else set $X_{t+1} = X_t$.
5. Set $t = t + 1$ and repeat steps 2 through 5.

The Metropolis-Hastings procedure is implemented in Example, where we use it to generate random variables from a standard Cauchy distribution. As we will see, this implementation is one of the special cases of the Metropolis-Hastings sampler described later.

Example

We show how the Metropolis-Hastings sampler can be used to generate random variables from a standard Cauchy distribution given by

$$f(x) = \frac{1}{\pi(1 + x^2)} \qquad\qquad -\infty < x < \infty.$$

From this, we see that

$$f(x) \propto \frac{1}{1 + x^2}$$

We will use the normal as our proposal distribution, with a mean given by the previous value in the chain and a standard deviation given by σ.

Non-parametric Kernel Density

A kernel distribution is a nonparametric representation of the probability density function (pdf) of a random variable. You can use a kernel distribution when a parametric distribution cannot properly describe the data, or when you want to avoid making assumptions about the distribution of the data. A kernel distribution is defined by a smoothing function and a bandwidth value, which control the smoothness of the resulting density curve.

Kernel Density Estimator

The kernel density estimator is the estimated pdf of a random variable. For any real values of *x*, the kernel density estimator's formula is given by

$$\hat{f}_h = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

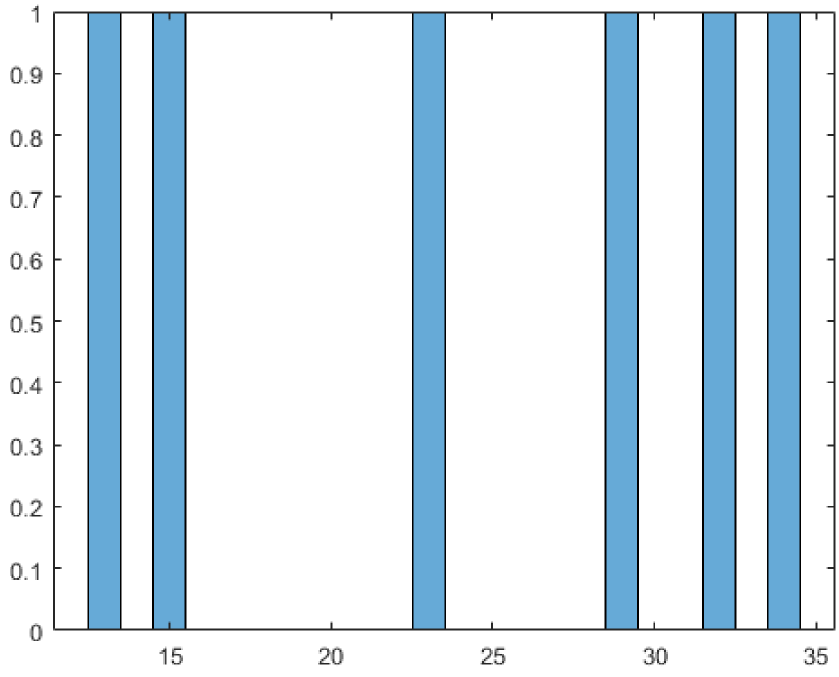
where $x_1, x_2, \dots, x_n$are random samples from an unknown distribution, n is the sample size, $K(\bullet)$ is the kernel smoothing function, and *h* is the bandwidth.

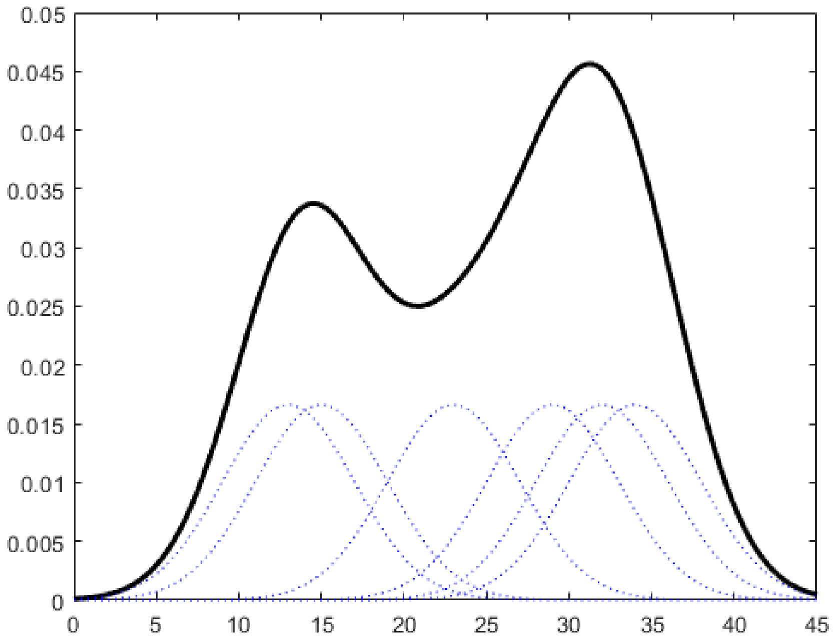
Kernel Smoothing Function

The kernel smoothing function defines the shape of the curve used to generate the pdf. Similar to a histogram, the kernel distribution builds a function to represent the probability distribution using the sample data. But unlike a histogram, which places the values into discrete bins, a kernel distribution sums the component smoothing functions for each data value to produce a smooth, continuous probability curve. You may plot a visual comparison of a histogram and a kernel distribution generated from the same sample data.

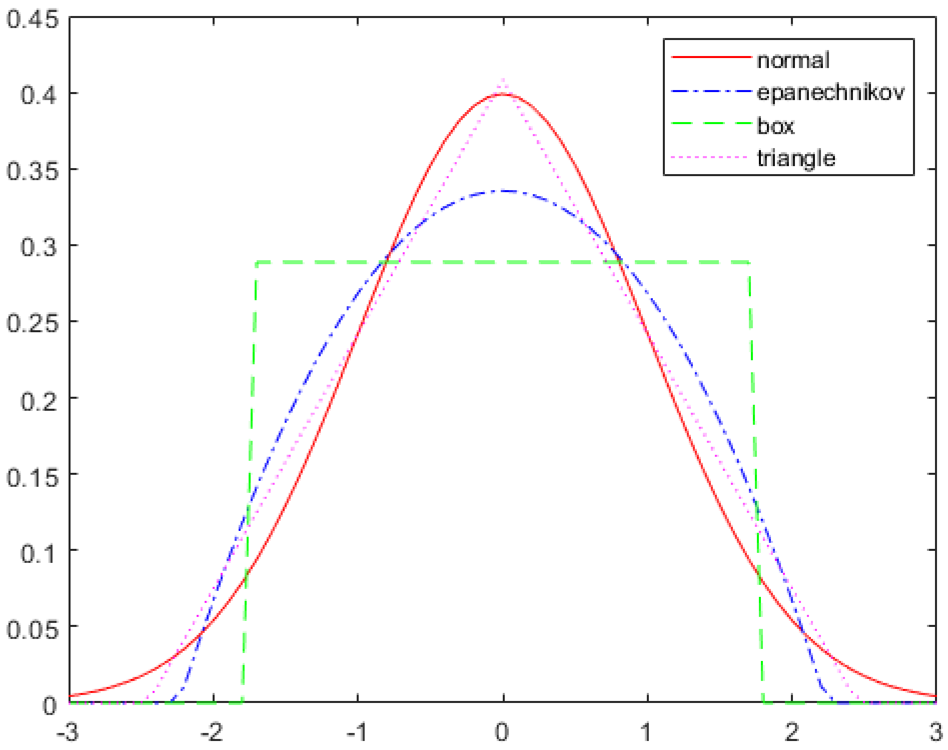
Because of this bin count approach, the histogram produces a discrete probability density function. This might be unsuitable for certain applications, such as generating random numbers from a fitted distribution.

Alternatively, the kernel distribution builds the pdf by creating an individual probability density curve for each data value, then summing the smooth curves. This approach creates one smooth, continuous probability density function for the data set.





Examples of Kernels for Density Estimation



- Triangle $K(t) = (1 - |t|) - 1 \leq t \leq 1$
- Epanechnikov

$$K(t) = \frac{3}{4}(1 - t^2) - 1 \leq t \leq 1$$

- Biweight

$$K(t) = \frac{15}{16}(1 - t^2)^2 - 1 \leq t \leq 1$$

- Triweight

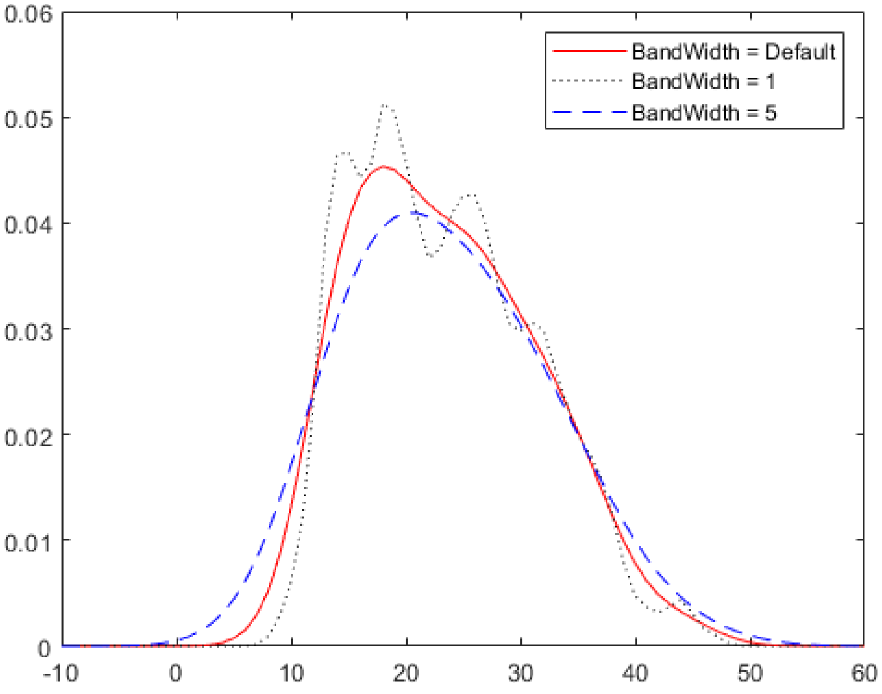
$$K(t) = \frac{35}{32}(1 - t^2)^3 - 1 \leq t \leq 1$$

- Normal

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\{\frac{-t^2}{2}\}$$

Bandwidth

The choice of bandwidth value controls the smoothness of the resulting probability density curve. This plot shows the density estimate for the same MPG data, using a normal kernel smoothing function with three different bandwidths.



Bootstrap

Central Limit Theorem:

If $X_1 \ldots X_N$ are random samples drawn from a population with overall mean μ and finite variance σ^2. Then the mean of the drawn sample $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ follows a normal distribution with mean and variance to be

$$N(\mu, \frac{\sigma^2}{N})$$

The standard deviation of sample mean $\bar{X}$ is $\sigma/\sqrt{N}$. The above expression is often written as

$$\sqrt{N}(\bar{X} - \mu) \sim N(0, \sigma^2)$$

The bootstrap is a simulation method for forming confidence intervals and obtaining standard errors using only information from the sample. Like a Monte-Carlo simulation but uses only the data.

Advantages

1. Applicable in a wide range of contexts
2. Easy.
3. The bootstrap in some cases may produce better approximations (knocks out an extra term in the Edgeworth expansion).

Hall justifies the bootstrap with a “Russian Dolls” analogy

- Doll zero: population that we do not get to see
- Doll one: sample we observe
- Doll two: bootstrap sample

<u>Bootstrap: Sample Code in Matlab</u>

```
X=[79 73 68 77 86 71 69] ';
[T, N] = size(X);
x_mu = mean(X);
x_se = std(X)/sqrt(T);
B = 1000;
x_boot_mu = zeros(B,1);
 for i = 1:B;
      x_boot=x(ceil(T*rand(T,1)+0.0001)); % draw x with replacement.
      x_boot_mu(i) = mean(x_boot);
      x_boot_se = std(xboot)/sqrt(T);
      t_stat_boot(i)=(x_boot_mu(i)-x_mu)/x_boot_se;
 end;

xbootmu = sort(x_boot_mu);
tstatboot=sort(t_stat_boot);
```

```
% confidence interval of x_boot_mu:
[x_boot_mu (25) x_boot_mu (975)]
% confidence interval of x:
[x_mu-(tstatboot(975)*x_se) x_mu-(tstatboot(25)*x_se)]
```

The last two lines give the OP and Percentile t CIs for mean

The bootstrap in a regression model

Suppose I have a linear regression model

$$y = \beta' x_i + \varepsilon_i$$

The most standard implementation of the bootstrap entails the following steps:

1. Estimate the parameter vector $\beta$and work out the residuals

   $$e_i = \ y - \hat{\beta}' x_i$$

2. Resample from the residuals with replacement and from the regressors with replacement.
3. Build up a new dataset of the dependent variables as

   $$\mathrm{y}_\mathrm{i}^{\mathrm{BOOT}} = \hat{\beta}' \mathrm{x}_\mathrm{i}^{\mathrm{BOOT}} + \mathrm{e}_\mathrm{i}^{\mathrm{BOOT}}$$

4. Work out the quantity of interest in this new dataset
5. Repeat (2)-(4) many times.

Other percentile, percentile or percentile-t confidence intervals can then be worked out.

Disadvantage

- Destroy conditional Heteroskedasticity
- Destroy autocorrelation, as it assumes all observations are independent of each other.