MATH 580: FINANCIAL STOCHASTIC PROCESSES

Paul Fearnhead (B23 Fylde College)
Department of Mathematics and Statistics
Lancaster University
Email: p.fearnhead@lancaster.ac.uk

October 2013

Aims: Due to their inherent randomness, it is natural to model financial and economic systems using probability models and stochastic processes. Analysis of appropriate stochastic models has become extremely important in recent years, such as for accurately pricing options. This module gives a thorough (but not too rigorous) introduction to stochastic processes in general and their use in modeling in business, finance and economic applications. Students will gain understanding about how both simulation and mathematical techniques can be used to learn about stochastic processes.

Measurable Objectives: At the end of this course students should be able to

- recognise the contexts when certain random variables occur;
- simulate a range of random variables;
- derive a range of properties for random variables using analytic and simulation methods;
- transform random variables;
- identify a range of stochastic models and derive a range of their behaviours including Poisson/Counting Processes, Markov processes.
- simulate a range of stochastic processes;
- use R to tackle a range of problems with stochastic processes which analytical methods cannot give tractable answers to.

Organisation and Assessment:

- The assessment for this course is 70% examination, together with 10% course work; and 20% project.
- The credit for this course is 15 credits.

Recommended books: The lecture notes are designed to be self-contained. However, the following books are recommended for additional reading.

Grimmett, G. and D. Welsh (1988) Probability An Introduction.
Ross, S. (2002) A First Course in Probability.
Grimmett, G. and D. Stirzaker (1992) Probability and Random Processes, OUP.
Morgan, B.J. T. (1984). Elements of Simulation, Chapman and Hall.

Glossary

P probability.

Ω sample space.

$\mathsf{P}(A)$ probability of event A.

$\mathsf{P}(A \mid B)$ probability of event A given event B.

A random variable is a function $X : \Omega \to \mathbb{R}$. It is an indicator of the outcome of a probability experiment that always takes numerical values.

$p_X(x)$ is the probability mass function (pmf) of a discrete random variable X evaluated at x, i.e. $\mathsf{P}(X = x)$.

$F_X(x)$ is the cumulative distribution function of a random variable X evaluated at x, i.e. $\mathsf{P}(X \leq x)$.

$\mathsf{E}(X)$ is the expectation of random variable X.

$\mathsf{E}(g(X))$ is the expectation of the function $g(X)$ of the random variable X.

$\mathsf{var}(X)$ is the variance of the random variable X.

$\mathsf{std}(X)$ is the standard deviation of the random variable X.

$\bar{F}_X(x)$ is the survivor function of the random variable X evaluated at x, i.e. $\mathsf{P}(X > x)$.

$f_X(x)$ is the probability density function (pdf) of random variable X evaluated at x.

μ_X is the expectation of variable X.

σ_X is the standard deviation of variable X.

$\mu_r = \mathsf{E}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^r\right]$ is the rth standardized central moment.

μ_3 is the coefficient of skewness.

$\mu_4 - 3$ is the kurtosis.

x_p is the 100p% quantile of random variable, i.e. $F_X(x_p) = p$.

$x_{0.5}$ is the median.

$x_{0.75} - x_{0.25}$ is the inter quartile range.

$X \sim \text{Uniform}(a, b)$, shows the random variable X follows the Uniform distribution on the interval $[a, b]$.

$X \sim \text{Exp}(\beta)$, shows the random variable X follows the Exponential distribution with rate β and mean β^{-1}.

$X \sim \text{Gamma}(\alpha, \beta)$ shows the random variable X follows the Gamma distribution with rate β and shape α and mean α/β.

$X \sim \text{Normal}(\mu, \sigma^2)$, usually written as $X \sim \text{N}(\mu, \sigma^2)$, shows the random variable X follows a Normal distribution with mean μ and standard deviation σ.

Φ is the cumulative distribution function for the standard Normal distribution $N(0,1)$.

The distribution of a random variable is either the name e.g. $\text{Exp}(\beta)$, the probability density function f or the cumulative distribution function F.

Probability integral transform is a result that allows one to transform from a Uniform random variable to a random variable with any specified cumulative distribution function.

$F_{XY}(x,y)$ is the joint cumulative distribution function of two random variables X and Y evaluated at (x,y), i.e. $\mathsf{P}(X \leq x, Y \leq y)$.

$p_{XY}(x,y)$ is the joint probability mass function (pmf) of two discrete random variables X and Y, i.e. $\mathsf{P}(X = x, Y = y)$.

$f_{XY}(x,y)$ is the joint probability density function of two continuous random variables X and Y evaluated at (x,y).

$X \mid Y = y$ is the conditional distribution of X given $Y = y$.

$p_{X|Y}(x \mid y)$ is the conditional probability mass function of X given $Y = y$.

$f_{X|Y}(x \mid y)$ is the conditional probability density function of X given $Y = y$.

$\mathsf{E}(g(X,Y))$ is the expectation of the function $g(X,Y)$ of the random variables X and Y.

$\mathsf{E}(X \mid Y = y)$ is the conditional expectation of X given $Y = y$.

$\mathsf{var}(X \mid Y = y)$ is the conditional variance of X given $Y = y$.

$\mathsf{cov}(X,Y)$ is the covariance between X and Y.

$\rho = \mathsf{corr}(X,Y)$ is the correlation between X and Y.

$\boldsymbol{X} = (X_1, \ldots, X_n)'$ is a random vector. It is a column vector.

$\mathsf{E}(\boldsymbol{X}) = (\mathsf{E}(X_1), \ldots, \mathsf{E}(X_n))'$ is the mean vector of $\boldsymbol{X}$.

$\mathsf{var}(\boldsymbol{X})$ is the variance matrix of $\boldsymbol{X}$, also called the variance-covariance matrix.

$\boldsymbol{X} \sim \text{MVN}_d(\boldsymbol{\mu}, \Sigma)$, shows the random vector $\boldsymbol{X}$ follows the multivariate Normal distribution of d dimensions with mean vector $\boldsymbol{\mu}$ and variance matrix Σ.

R is the statistical software package used to evaluate probabilities from standard distributions.

Chapter 1

The Axiomatic Approach

The theory of probability is used to provide mathematical models of situations affected by chance. In this chapter we set up the framework on which probability is defined.

1.1 The Sample Space

An experiment is a process of measurement or observation. We are interested in experiments that involve randomness, so the result cannot be predicted exactly. A trial is a single performance of an experiment.

The sample space Ω is the set (possibly infinite or uncountable) of all outcomes of the experiment. A particular outcome $\omega \in \Omega$ is called a sample point. A subset of Ω is called an event. An event $A \subseteq \Omega$ occurs if, when the experiment is performed, the outcome $\omega \in \Omega$ satisfies $\omega \in A$.

Suppose A and B are events in a sample space Ω. The union $A \cup B$ of A and B consists of all points in A or B or both. The intersection $A \cap B$ consists of all points that are in both A and B. If A and B have no points in common we call A and B mutually exclusive or disjoint. We write $A \cap B = \emptyset$, where $\emptyset$ is the empty set or set containing no elements. The complement A^c of A is the set of all points in Ω that are not in A.

If $A_1, A_2, \ldots \subseteq \Omega$, we write $\bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup \cdots$ and $\bigcap_{i=1}^{\infty} A_i = A_1 \cap A_2 \cap \ldots$.

Example: Suppose the experiment is to record the price of a share over a period of length T. Then

$$\Omega = \{f : [0,T] \to [0,\infty)\}.$$

An example of an event is that the share price does not exceed a price level L,

$$A = \{f : [0,T] \to [0,L]\}.$$

Suppose that the particular outcome of this experiment turns out to be the sample point $f(t) = c$ for all $t \in [0,T]$ (in this case the share price has not changed during the time period). Then if $c \leq L$, the event A has occurred and if $c > L$, the event A^c has occurred.

1.2 The Axioms of Probability

Let Ω be a sample space. Then probability P is a real valued function defined on subsets of Ω satisfying the following axioms of probability.

Axiom 1 (positivity)	$\mathsf{P}(A) \geq 0$ for all $A \subseteq \Omega$.
Axiom 2 (finitivity)	$\mathsf{P}(\Omega) = 1$.
Axiom 3 (additivity)	$\mathsf{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathsf{P}(A_i)$ for any finite or infinite sequence $A_1, A_2, \ldots \subseteq \Omega$ of disjoint events (i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$).

The number $\mathsf{P}(A)$ is called the probability of the event A. Probability can be thought of as being a measure of the chance that an event may occur, in the same way that length is a measure of the magnitude of an object.

1.3 The Laws of Probability

An immediate consequence of the axioms of probability are the following results, where A and B are any events.

Law of Complementary events:

$$\mathsf{P}(A^c) = 1 - \mathsf{P}(A).$$

A consequence of this is that $\mathsf{P}(\emptyset) = 0$.

Partition Law:

$$\mathsf{P}(A) = \mathsf{P}(A \cap B) + \mathsf{P}(A \cap B^c).$$

Addition Law:

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B).$$

Order preserving: If $A \subseteq B$, then $\mathsf{P}(A) \leq \mathsf{P}(B)$.

1.4 Conditional Probability

Often the probability of an event depends not just on the experiment itself but on other information you are given about the experiment. Conditional probability forms a framework in which this additional information can be incorporated.

If A and B are two events then, provided $\mathsf{P}(B) > 0$, the conditional probability of A given B is written as $\mathsf{P}(A|B)$ and calculated from

$$\mathsf{P}(A|B) = \mathsf{P}(A \cap B)/\mathsf{P}(B).$$

For evaluating $\mathsf{P}(A \cap B)$ it is often easiest to use

$$\mathsf{P}(A \cap B) = \mathsf{P}(A|B)\,\mathsf{P}(B) = \mathsf{P}(B|A)\,\mathsf{P}(A).$$

Bayes theorem inverts the ordering of conditioning for events A and B, with $\mathsf{P}(A), \mathsf{P}(B) > 0$:

$$\mathsf{P}(B|A) = \mathsf{P}(A|B)\,\mathsf{P}(B)/\,\mathsf{P}(A).$$

The Law of Total Probability follows from the Partition Law giving

$$\mathsf{P}(A) = \mathsf{P}(A|B)\,\mathsf{P}(B) + \mathsf{P}(A|B^c)\,\mathsf{P}(B^c).$$

1.5 Independent Events

Two events A and B are independent events if and only if the multiplication law holds:

$$\mathsf{P}(A \cap B) = \mathsf{P}(A)\,\mathsf{P}(B).$$

When events A and B are independent and $\mathsf{P}(A), \mathsf{P}(B) > 0$, then $\mathsf{P}(A|B) = \mathsf{P}(A)$ and $\mathsf{P}(B|A) = \mathsf{P}(B)$. This means that the occurrence of one event does not influence the chance of occurrence of the other event.

Chapter 2

Discrete Univariate Random Variables

We are not always interested in an experiment itself, but rather in some consequence of its random outcome. Random variables give us a way to think about these consequences in those situations when they take real values. The theory is richer than that of ordinary events because of the additional structure imposed by the number system.

2.1 Random Variables

A random variable X is a function $X : \Omega \to \mathbb{R}$. For each outcome in the sample space, $\omega \in \Omega$, it associates a unique real number $X(\omega)$.

NB: A random variable is neither random nor a variable. It is a function.

Every time the experiment is conducted, exactly one value of the random variable is observed; this is called a realisation of the random variable.

The range of values taken by the random variable is the induced sample space $\mathcal{S} = \{X(\omega) : \omega \in \Omega\}$. If $\mathcal{S}$ is finite or countable then the random variable is discrete. For simplicity we will take the sample space to be the integers in the remainder of this chapter.

2.2 Probabilities

Suppose that $A \subseteq \mathcal{S}$ is an event in the induced sample space. Then we write

$$\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}.$$

The right hand side is an event in Ω and hence has a probability assigned to it. This induces a probability on the induced sample space

$$\mathsf{P}(X \in A) = \mathsf{P}(\{\omega \in \Omega : X(\omega) \in A\}).$$

The probability of outcome r (r an integer) for a discrete random variable X is given by the probability mass function (pmf) defined by

$$p_X(r) = p(r) = \mathsf{P}(X = r) \quad \text{for } r \text{ an integer.}$$

If $p(r)$ is a probability mass function then

- $0 \leq p(r)$ for all r,
- $\sum_{r=-\infty}^{\infty} p(r) = 1$.

For any event A

$$\mathsf{P}(A) = \mathsf{P}(X \in A) = \sum_{r \in A} p(r).$$

The cumulative distribution function (cdf) of the random variable X is the function

$$F_X(m) = \mathsf{P}(X \leq m) = \sum_{r=-\infty}^{m} p(r).$$

Example: Given an event $A \subseteq \Omega$ define a function $I_A : \Omega \to \{0, 1\}$ by $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ otherwise. Since I_A is a real-valued function on Ω, it is a random variable. It is called the indicator function of A.

The pmf of I_A is given by $p(0) = \mathsf{P}(A^c) = 1 - \mathsf{P}(A)$, $p(1) = \mathsf{P}(A)$.

Indicator functions satisfy the following properties:

$$\begin{aligned} I_A &= 1 - I_{A^c}, \\ I_{A \cap B} &= I_A I_B \\ I_{A \cup B} &= I_A + I_B - I_A I_B. \end{aligned}$$

2.3 Summary Measures

There are a number of ways of summarising the distribution of a discrete random variable. Here are the most common.

Expectation: a measure of the location/mean of the random variable (in the units of the random variable). The expected value of a discrete random variable X is

$$\mathsf{E}(X) = \sum_{r=-\infty}^{\infty} r p(r).$$

The expected value of a real-valued function g of a discrete random variable X is

$$\mathsf{E}[g(X)] = \sum_{r=-\infty}^{\infty} g(r)\, p(r).$$

Variance: a measure of the variability between outcomes of the random variable (in squared units). The variance of the random variable X, $\mathsf{var}(X)$, is defined as

$$\mathsf{var}(X) = \mathsf{E}[(X - \mathsf{E}(X))^2],$$

which is most easily evaluated using:

$$\begin{aligned} \mathsf{var}(X) &= \mathsf{E}(X^2) - [\mathsf{E}(X)]^2 \\ &= \mathsf{E}[X(X-1)] + \mathsf{E}(X) - [\mathsf{E}(X)]^2. \end{aligned}$$

Standard deviation: a measure of the variability between outcomes of the random variable (in units of the random variable). The standard deviation of the random variable X, is defined as $\mathsf{std}(X) = \sqrt{\mathsf{var}(X)}$.

2.4 Properties of Summary Measures

Expectation obeys two rules of linearity which follow directly from the definition. For arbitrary functions g and h, and a constant c:

$$\begin{aligned} \mathsf{E}[g(X) + h(X)] &= \mathsf{E}[g(X)] + \mathsf{E}[h(X)], \\ \mathsf{E}[cg(X)] &= c\,\mathsf{E}[g(X)]. \end{aligned}$$

The expectation, variance and standard deviation respectively of the linear function $aX + b$ of the random variable X for constants a and b are:

$$\begin{aligned} \mathsf{E}(aX + b) &= a\,\mathsf{E}(X) + b, \\ \mathsf{var}(aX + b) &= a^2\,\mathsf{var}(X), \\ \mathsf{std}(aX + b) &= |a|\,\mathsf{std}(X). \end{aligned}$$

2.5 Probability Generating Functions

The distribution of a random variable is determined by its pmf, $p(r)$ for $r = 0, 1, \ldots$. A concise way to store this information is by means of a generating function.

The probability generating function (pgf) of a discrete random variable X is

$$G(x) = \mathsf{E}[x^X] = \sum_{r=0}^{\infty} x^r p(r).$$

Note that $G(x)$ is a polynomial or a power series and the coefficient of x^r is $p(r) = \mathsf{P}(X = r)$.

A pgf satisfies the following two properties:

- Since $p(r) \geq 0$ for all r, the coefficients of the powers of x in $G(x)$ are ≥ 0;
- $G(1) = \sum_{r=0}^{\infty} p(r) = 1$. (Note that in certain cases, $G(1)$ may not be defined. In this case the property still holds if $G(1)$ is replaced by $\lim_{x \uparrow 1} G(x)$ and L'hôpital's rule is used.)

Any (infinitely differentiable) function that satisfies the above two conditions is a pgf of some random variable. The pmf can be obtained either by just reading off the values of the coefficients of the powers of x, or by repeated differentiation to get

$$p(r) \quad = \quad \frac{1}{r!} G^{(r)}(0).$$

The pgf gives a convenient way to calculate the mean and variance of a random variable. Differentiating term by term gives

$$G'(x) = \sum_{r=0}^{\infty} r x^{r-1} p(r)$$

and so

$$G'(1) = \sum_{r=0}^{\infty} r p(r) = \mathsf{E}[X].$$

Similarly

$$G''(x) = \sum_{r=0}^{\infty} r(r-1) x^{r-2} p(r)$$

and so

$$G''(1) = \sum_{r=0}^{\infty} r(r-1) p(r) = \mathsf{E}[X(X-1)].$$

Using results about the variance

$$\begin{aligned} \mathsf{var}[X] &= \mathsf{E}[X(X-1)] + \mathsf{E}[X] - (\mathsf{E}[X])^2 \\ &= G''(1) + G'(1) - G'(1)^2. \end{aligned}$$

(If $G(1)$ is not defined then, as before, the above identities still hold if $G'(1)$ and $G''(1)$ are replaced by $\lim_{x \uparrow 1} G'(x)$ and $\lim_{x \uparrow 1} G''(x)$ respectively.)

Exercise 2.1 Find the pmf, the mean and the variance of the random variable with pgf given by $G(x) = \exp(x-1)$.

Sol:

Differentiating gives

$$G'(x) = \exp(x-1) = G(x),$$

and so

$$G^{(r)}(x) = \exp(x-1).$$

Therefore

$$p(r) = \frac{1}{r!}G^{(r)}(0) = \frac{e^{-1}}{r!} \quad \text{for} \quad r = 0, 1, \ldots.$$

(Note that since $p(r) \geq 0$ for all r and $G(1) = 1$, this is indeed a valid pgf of some random variable. In fact it is a Poisson random variable, which is defined in the next section.)

To find the mean and variance

$$\mathsf{E}[X] = G'(1) = 1$$

and

$$\mathsf{var}[X] = G''(1) + G'(1) - G'(1)^2 = 1 + 1 - 1^2 = 1.$$

□

2.6 Probability Models

Although probability mass functions are studied in general, it is helpful to focus on a subset of functions which describe the distribution of outcomes of broad classes of experiment. In this section we present some of the most widely used. Where the value of the pmf is not defined it can be assumed to be 0.

Uniform Random Variables

The uniform random variable is a model for outcomes of experiments which involve selecting at random from a sample space $\{0, 1, \ldots, m\}$. Here

$$p(r) = \frac{1}{m+1} \quad \text{for} \quad r = 0, 1, \ldots, m,$$

and

$$\begin{aligned} \mathsf{E}(X) &= \frac{m}{2}, \\ \mathsf{var}(X) &= \frac{m(m+2)}{12}, \\ G(x) &= \frac{1}{m+1}\,\frac{1-x^{m+1}}{1-x}. \end{aligned}$$

Bernoulli Random Variables

The Bernoulli random variable is a model for outcomes of experiments where the sample space is $\{0, 1\}$ and the outcomes are not necessarily equi-probable. Here

$$p(0) = 1 - \theta, \quad p(1) = \theta$$

and

$$\begin{aligned} \mathsf{E}(X) &= \theta, \\ \mathsf{var}(X) &= \theta(1-\theta), \\ G(x) &= 1 - \theta + \theta x. \end{aligned}$$

Binomial Random Variables

The Binomial random variable is a model for outcomes of experiments which count the number of 1 values (successes) in a sequence of n independent Bernoulli trials (each with probability θ of a 1). The sample space is $\{0, 1, \ldots, n\}$ and

$$p(r) = \binom{n}{r} \theta^r (1-\theta)^{n-r} \qquad \text{for} \quad r = 0, 1, 2, \ldots, n,$$

and

$$\begin{aligned} \mathsf{E}(X) &= n\theta, \\ \mathsf{var}(X) &= n\theta(1-\theta), \\ G(x) &= (1 - \theta + \theta x)^n. \end{aligned}$$

We write $X \sim \text{Binomial}(n, \theta)$.

The software package R can simulate m independent $\text{Binomial}(n, \theta)$ random variables using the command `rbinom(m, n, theta)` and can generate the value of the pmf at r using the command `dbinom(r, n, theta)` . Note that by setting $n = 1$ it is possible to simulate Bernoulli random variables.

Geometric Random Variables

The Geometric random variable is a model for the outcomes of experiments which count the number of 0 values before the first 1 in a sequence of independent Bernoulli trials (each with probability θ of a 1). The sample space is $\{0, 1, 2, \ldots\}$, and

$$p(r) = (1-\theta)^r \theta \qquad \text{for} \quad r = 0, 1, \ldots$$

and

$$\begin{aligned} \mathsf{E}(X) &= \frac{1-\theta}{\theta}, \\ \mathsf{var}(X) &= \frac{1-\theta}{\theta^2}, \\ G(x) &= \frac{\theta}{1 - x(1-\theta)}. \end{aligned}$$

We write $X \sim \text{Geometric}(\theta)$.

The software package R can simulate m independent $\text{Geometric}(\theta)$ random variables using the command `rgeom(m, theta)` and can generate the value of the pmf at r using the command `dgeom(r, theta)` .

Poisson Random Variables

The Poisson random variable is a model for the outcomes of experiments which count the number of points in an interval $[0, t]$ of a random process where points occur at random at a given rate $\phi > 0$ per unit interval. The sample space is $\{0, 1, \ldots\}$, and

$$p(r) = \frac{\lambda^r \exp(-\lambda)}{r!} \qquad \text{for} \qquad r = 0, 1, 2, \ldots$$

where $\lambda = \phi t$ and

$$\begin{aligned} \mathsf{E}(X) &= \lambda \\ \mathsf{var}(X) &= \lambda \\ G(x) &= \exp(\lambda(x-1)). \end{aligned}$$

We write $X \sim \text{Poisson}(\lambda)$.

The software package R can simulate m independent $\text{Poisson}(\lambda)$ random variables using the command `rpois(m, lambda)` and can generate the value of the pmf at r using the command `dpois(r, lambda)` .

The Poisson distribution also has the interpretation as the limit distribution of a $\text{Binomial}(\theta)$ distribution with $n \to \infty$, $\theta \to 0$ and $n\theta \to \lambda$.

Chapter 3

Continuous Univariate Random Variables

Discrete random variables describe the outcomes of experiments which are in a countable set of numbers. Focusing only on discrete random variables is too restrictive for many situations, in particular in cases when the outcome of the experiment is a measurement on a continuous scale.

3.1 Cumulative Distribution Functions

To describe continuous random variables we need slightly different mathematical tools to those we used for discrete random variables. For example, for a discrete random variable X we used the probability mass function $p(x) = \mathsf{P}(X = x)$. However, if X is a continuous random variable then $\mathsf{P}(X = x) = 0$ for all x. We focus on probabilities of events instead of probabilities of single outcomes. In particular we focus on events of the form

$$\{X \leq x\}$$

for fixed x and consider these as x takes on different values. The cumulative distribution function (cdf) of a random variable, X, is defined, for all real values of x, by

$$F_X(x) = F(x) = \mathsf{P}(X \leq x),$$

i.e. the probability that a random variable X takes a value less than or equal to x. We call X a continuous random variable if $F(x)$ is a differentiable (and hence continuous) function.

Properties of $F_X(x)$:

- $F_X(x)$ is a non-decreasing function of x,
- $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.

The distribution function is particularly useful for continuous random variables as we often want to know the probability of events that can be related by the laws of probability into probability statements about the event $\{X \leq x\}$ for some x.

The Survivor function: The survivor function of a random variable X is defined as

$$\bar{F}_X(x) \quad = \quad \mathsf{P}(X > x).$$

Using the law of complementary events

$$\mathsf{P}(X > x) = 1 - \mathsf{P}(X \leq x) = 1 - F_X(x).$$

Probabilities of Intervals: Often the probability of the random variable X falling in the interval $(a, b]$ is of interest for real numbers a, b with $a < b$.

This corresponds to the event $\{a < X \leq b\}$. By using Axiom 3,

$$\mathsf{P}(X \leq b) = \mathsf{P}(X \leq a) + \mathsf{P}(a < X \leq b),$$

so the probability of the interval event is

$$\mathsf{P}(a < X \leq b) = \mathsf{P}(X \leq b) - \mathsf{P}(X \leq a) = F_X(b) - F_X(a).$$

3.2 Probability Density Functions

The probability density function (pdf) $f_X(x)$ of a continuous random variable, X, is defined by

$$f_X(x) = \frac{d}{dx} F_X(x)$$

so that it satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\, ds.$$

Note that the pdf is zero in regions where there are no outcomes. Also the pdf can exceeds 1, so cannot be interpreted as a probability despite some of the mathematical properties of $f_X(x)$ being similar to those of the probability mass function (pmf).

Properties of $f_X(x)$:

- Positivity: $f_X(x) \geq 0$ for all x,
- Unit-integrability: $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

The probability that an observation on a continuous random variable X lies in the interval $(a, b]$ may be calculated as

$$\begin{aligned} \mathsf{P}(a < X \leq b) &= F_X(b) - F_X(a) \\ &= \int_{-\infty}^{b} f_X(x)\, dx - \int_{-\infty}^{a} f_X(x)\, dx \\ &= \int_{a}^{b} f_X(x)\, dx. \end{aligned}$$

In fact, for any event A the probability of that event is given by

$$\mathsf{P}(X \in A) = \int_{x \in A} f_X(x)\, dx.$$

To illustrate the equivalence of the definition of $f_X(x)$ with this property we start with the property and argue that the definition follows. Using the above property on intervals, we have for any x and a small δ

$$\begin{aligned} \mathsf{P}(x < X \leq x + \delta) &= F_X(x + \delta) - F_X(x) \\ &= \int_x^{x+\delta} f_X(s)\, ds \\ &\approx f_X(x)\delta, \end{aligned}$$

so that

$$\{F_X(x + \delta) - F_X(x)\}/\delta \approx f_X(x)$$

i.e. $dF_X(x)/dx = f_X(x)$ in the limit as $\delta \to 0$.

3.3 Summary Measures

All the information about the distribution of a continuous random variable X is contained in the cdf $F_X(x)$ and the pdf $f_X(x)$. However it is often helpful to summarise the main characteristics of the distribution in terms of a few values. Here we consider the standard summary measures.

Expectation and Variance

The expected value (or mean) of a continuous random variable X can be thought of as the average of the different values that X may take, according to their chance of occurrence.

The expected value of a continuous random variable X is

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx,$$

provided that not both

$$\int_{-\infty}^{0} x f_X(x)\, dx \quad \text{and} \quad \int_{0}^{\infty} x f_X(x)\, dx$$

are infinite.

Similarly for a real valued function $g(X)$ of a continuous random variable X the expected value is

$$\mathsf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx,$$

provided that

$$\int_{-\infty}^{\infty} |g(x)| f_X(x)\, dx < \infty.$$

For continuous variables, linearity properties of expectation are the same as those stated in Chapter 2 for discrete random variables.

For arbitrary functions g and h, constants a, b and c, and a continuous random variable X

$$\begin{aligned} \mathsf{E}[g(X) + h(X)] &= \mathsf{E}[g(X)] + \mathsf{E}[h(X)], \\ \mathsf{E}[cg(X)] &= c\,\mathsf{E}[g(X)], \\ \mathsf{E}[aX + b] &= a\,\mathsf{E}[X] + b. \end{aligned}$$

The variance, measuring of the spread or dispersion of a random variable about the expectation, for a continuous distribution is

$$\mathsf{var}(X) = \mathsf{E}[(X - \mathsf{E}(X))^2] = \int_{-\infty}^{\infty} (x - \mathsf{E}(X))^2 f_X(x)\, dx.$$

For continuous random variables the easiest way to evaluate the variance is

$$\mathsf{var}(X) = \mathsf{E}(X^2) - [\mathsf{E}(X)]^2.$$

The standard deviation $\mathsf{std}(X)$ of a continuous random variable X is $\sqrt{\mathsf{var}(X)}$.

The variance and standard deviation of the linear function $aX + b$ are

$$\begin{aligned} \mathsf{var}(aX + b) &= a^2\,\mathsf{var}(X), \\ \mathsf{std}(aX + b) &= |a|\,\mathsf{std}(X). \end{aligned}$$

A measure of the typical size of a random variable to its variability is given by the coefficient of variation which is defined by $\mathsf{E}(X)/\,\mathsf{std}(X)$.

Warning: Sometimes the expectation and variance are infinite. This occurs when the probabilities of obtaining large values is too big. Examples are given later.

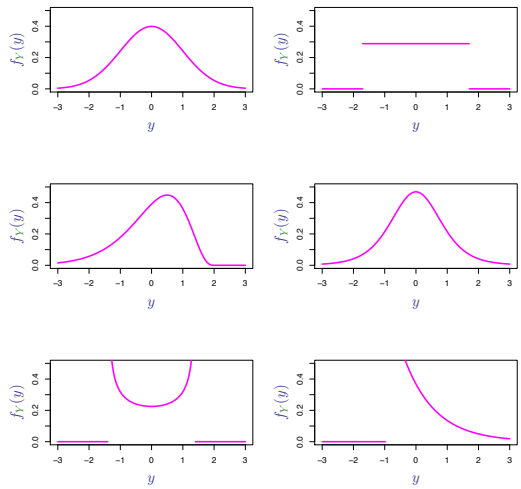
Higher Moments

If X has expectation μ_X and standard deviation σ_X then the random variable Y defined as

$$Y = \frac{X - \mu_X}{\sigma_X}$$

has $\mathsf{E}(Y) = 0$ and $\mathsf{var}(Y) = 1$ for any μ_X and $\sigma_X < \infty$.

The six pdfs below are for random variables Y for a range of density functions f_Y which all have $\mathsf{E}(Y) = 0$ and $\mathsf{var}(Y) = 1$. Despite having the same expectation and variance there are quite substantial differences between the six distributions. It is helpful to think how you would summarise such differences in the shape of the distributions.



Six pdfs: different rvs Y [$\mathsf{E}(Y) = 0$, $\mathsf{var}(Y) = 1$].

The most obvious differences in the shapes are
- skewness (lack of symmetry), and
- pointedness (light tailedness).

To be able to evaluate these shape characteristics we need to introduce higher moments.

The rth moment is evaluated as

$$\mathsf{E}(X^r) = \int_{-\infty}^{\infty} x^r f_X(x)\, dx.$$

This depends on the value of both μ_X and σ_X.

The rth standardized central moment is the expected value of the rth power of the standardized random variable $(X - \mu_X)/\sigma_X$. It is defined as

$$\mu_r = \mathsf{E}\left[\left(\frac{X - \mu_X}{\sigma_X}\right)^r\right] = \int_{-\infty}^{\infty} \left(\frac{x - \mu_X}{\sigma_X}\right)^r f_X(x)\, dx,$$

and does not depend on μ_X and σ_X.

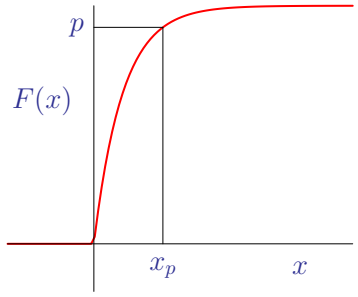
Skewness The third standardised central moment, μ_3, is a measure of the extent of the asymmetry and is called the coefficient of skewness. Positive (negative) values of the coefficient of skewness correspond to the distribution having a longer (shorter) upper tail than lower tail. If the distribution is symmetric the coefficient of skewness is zero.

Kurtosis The extent of the pointedness of the distribution is measured by *kurtosis*, $\mu_4 - 3$. The reason for subtracting 3 is to give the Normal distribution kurtosis 0. The Normal distribution is given in the top left panel of the figure. Positive (negative) kurtosis correspond to the distribution being more (less) pointed than the Normal distribution.

Quantiles

Often interest is in the values of a continuous random variable which are not exceeded with a given probability, such values are termed quantiles with x_p the $100p\%$ quantile defined by

$$F_X(x_p) = p.$$



Here p is known as the percentile corresponding to the quantile x_p.

Certain quantiles are of special interest:

Median: The median is the middle of the distribution in the sense that half the values of the variable (in probability) are less than the median, and half are more. The median is the 50% quantile, $x_{0.5}$, so that $F(x_{0.5}) = 0.5$. As a measure of location, the median has the advantage of existing for all distributions, unlike the expectation.

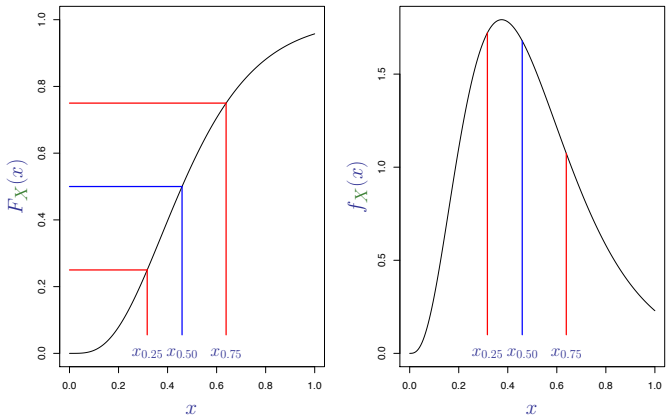
Quartile: The quartiles split the distribution into four equally likely regions, $x_{0.25}$ the lower quartile, $x_{0.5}$ the median and $x_{0.75}$ the upper quartile.

$$\begin{aligned} \mathsf{P}(X \leq x_{0.25}) &= \mathsf{P}(x_{0.25} < X \leq x_{0.5}) \\ &= \mathsf{P}(x_{0.5} < X \leq x_{0.75}) \\ &= \mathsf{P}(X > x_{0.75}) \\ &= 0.25. \end{aligned}$$

This is illustrated in the next figure.

Inter-quartile range: The difference in values of quartiles provides a measure of the variability of a random variable (measured in the units of the variable) that does not require the evaluation of the standard deviation (which can be infinite). The inter-quartile range is

$$x_{0.75} - x_{0.25}.$$



The cdf and pdf for a continuous random variable X and the three quartiles $x_{0.25}$, $x_{0.50}$ and $x_{0.75}$. Note that the quartiles split the area under pdf into four equally sized regions.

Quantiles of Monotone Functions of Random Variables: If $g(x)$ is a strictly increasing function of x, then the qth quantile of $g(X)$ is $g(x_q)$ where x_q is the qth quantile of X.

This follows as, because $g(x)$ is strictly increasing

$$\mathsf{P}(g(X) < y) = \mathsf{P}(X < g^{-1}(y)).$$

So if we want this y to be the qth quantile of $g(X)$, then we want the value such that this probability is q. However that gives $g^{-1}(y) = x_q$ by definition of tne x_q. Hence $y = g(x_q)$.

Similarly is If $g(x)$ is a strictly decreasing function of x, then the qth quantile of $g(X)$ is $g(x_{1-q})$ where x_{1-q} is the $1-q$th quantile of X.

This follows as, because $g(x)$ is strictly decreasing

$$\mathsf{P}(g(X) < y) = \mathsf{P}(X > g^{-1}(y)) = 1 - \mathsf{P}(X < g^{-1}(y)).$$

[Note change of inequality].

So if we want this y to be the qth quantile of $g(X)$, then we want the value such that this probability is q. However that gives $g^{-1}(y) = x_{1-q}$ by definition of the x_{1-q}. Hence $y = g(x_{1-q})$.

Exercise 3.1 A model for the distribution of the log-return of a stock, X, is Laplacian

$$F_X x = \begin{cases} 1 - 0.5\exp\{-200x\} & \text{if} x > 0 \\ 0.5\exp\{200x\} & \text{if} x \leq 0 \end{cases}$$

If you invest $£1,000$ in the stock, what is the value at risk at the 0.01 level?

Sol:

The value at risk is an amount, such that with probability 0.01 you would lose that amount (or more). This will correspond to the 0.01th quantile of the distribution of the log-return. So first find $x_{0.01}$.

$$F_X(x_{0.01}) = 0.01 \Rightarrow 0.5 \exp\{200 x_{0.01}\} = 0.01.$$

Solving gives $\exp\{200 x_{0.01}\} = 0.02$, and hence $x_{0.01} = \log(0.02)/200$. This is the log-return, so the actual return is $\exp\{x_{0.01}\} = 0.981$. Thus a $£1,000$ investment would be worth $£981$, which corresponds to a loss of $£19$. □

3.4 Moment Generating Functions

The analogue of the probability generating function for continuous random variables is the moment generating function.

The moment generating function (mgf) of a continuous random variable X is defined as

$$M_X(t) = \mathsf{E}[e^{tX}] = \int_{x=-\infty}^{\infty} e^{tx} f_X(x) dx$$

for all real values of t for which the integral is finite.

The mgf determines uniquely the distribution of the random variable X, provided it is finite for all values of t in some interval containing the 0.

By considering the series expansion of the exponential function, and using linearity of the expected value,

$$M_X(t) = \sum_{k=0}^{\infty} t^k \, \mathsf{E}[X^k]/k!.$$

So $M_X(t)$ is a power series and the coefficient of t^k is the kth moment of X divided by $k!$. Moments can be obtained by differentiating the mgf with respect to t and then evaluating it at zero i.e. $\mathsf{E}[X^k] = M_X^{(k)}(0)$.

Chapter 4

Standard Continuous Univariate Distributions

In this chapter we give the details of standard univariate continuous distributions that you are likely to see in subsequent study of Probability and Statistics.

The distributions arise in two ways, either as the probability distributions that are used in statistics modelling contexts or as the distributions that arise in statistical techniques. Because of the strong links to statistical modelling, where interest is in the effect on the distribution of the choice of parameter value, here we use the notation

$$f_X(x;\boldsymbol{\theta}) = f_X(x)$$

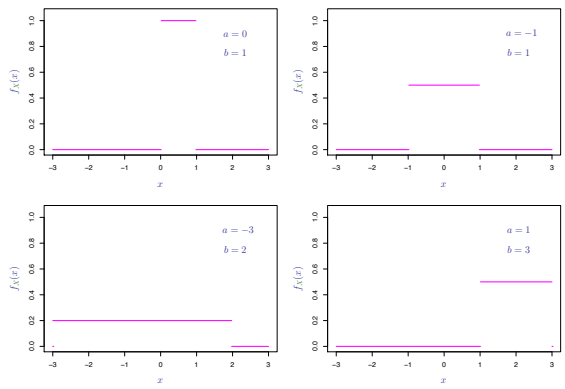
to emphasise the dependence of the probability on the parameters $\boldsymbol{\theta}$ of the probability model.

4.1 Uniform Distribution

A continuous random variable for which all outcomes in a given range have equal likelihood of occurring is said to be uniformly distributed. Specifically, a random variable X has a Uniform distribution over the interval (a, b) if the pdf is given by

$$f_X(x;\boldsymbol{\theta}) = \begin{cases} \frac{1}{b-a} & a < x < b, \\ 0 & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\theta} = (a, b)$, which we write $X \sim \text{Uniform}(a, b)$ or sometimes $X \sim \text{U}(a, b)$. This pdf for four different sets of parameter values is illustrated in the figure.



The pdfs for four uniformly distributed rvs X with different parameter values.

The reason that this pdf is appropriate for such random outcomes is that for all x and $x+\delta$ such that $a < x < x+\delta < b$

$$\mathsf{P}(x < X \leq x+\delta) = \delta/(b-a),$$

so the probability of X falling in any interval of length δ in the range (a,b) is the same for all x.

The cdf of the $\text{Uniform}(a,b)$ distribution is

$$F_X(x) = \begin{cases} 0 & \text{if} \quad x \leq a, \\ \frac{x-a}{b-a} & \text{if} \quad a < x < b, \\ 1 & \text{if} \quad x \geq b. \end{cases}$$

The rth moment of the $\text{Uniform}(a,b)$ distribution is

$$\begin{aligned} \mathsf{E}(X^r) &= \int_{-\infty}^{\infty} x^r f_X(x)\, dx \\ &= \int_{-\infty}^{a} x^r \cdot 0 \; dx + \int_a^b \frac{x^r}{b-a}\, dx + \int_b^{\infty} x^r \cdot 0 \; dx \\ &= \frac{b^{r+1} - a^{r+1}}{(r+1)(b-a)}. \end{aligned}$$

Hence the expectation (taking $r = 1$ in the result above) is

$$\mathsf{E}(X) = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2},$$

the variance is

$$\text{var}(X) = \frac{b^3 - a^3}{3(b-a)} - \left[\frac{b+a}{2}\right]^2 = \frac{(b-a)^2}{12},$$

and the mgf is

$$M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}.$$

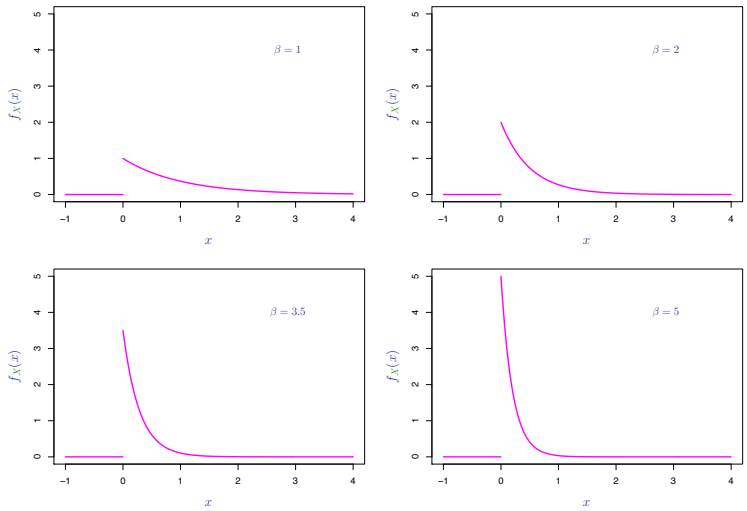
The software package R can simulate m independent $\text{Uniform}(a,b)$ random variables using the command `runif(m, a, b)` and can generate the value of the pdf and cdf at x using the commands `dunif(x, a, b)` and `punif(x, a, b)`, and the pth quartile using the command `qunif(p, a, b)`.

4.2 Exponential Distribution

A random variable X has an exponential distribution if its pdf is given by

$$f_X(x;\boldsymbol{\theta}) = \begin{cases} \beta \exp(-\beta x), & x \geq 0, \\ 0 & \text{otherwise}, \end{cases}$$

where $\boldsymbol{\theta} = \beta$, with $\beta > 0$, which we write $X \sim \text{Exp}(\beta)$. The pdf for four different values of β is shown in the figure.



The pdfs for four exponentially distributed random variables X with $\beta = 1$, 2, 3.5 and 5.

The exponential distribution arises in practice as the distribution of a waiting time when events occur at random with a rate of β per unit time in a Poisson process (see Chapter 9). The exponential distribution arises either as the distribution of the time between events or as the distribution of the time to an event from a given start time.

The cdf of the $\mathrm{Exp}(\beta)$ distribution is

$$F_X(x) = \begin{cases} 0 & \text{if} \quad x \leq 0, \\ 1 - \exp(-\beta x) & \text{if} \quad x > 0. \end{cases}$$

The survivor function for $x > 0$ is

$$\bar{F}(x) = \exp(-\beta x).$$

The mgf is

$$\begin{aligned} \mathsf{E}(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx \\ &= \int_{-\infty}^{0} x^r \cdot 0 \; dx + \int_{0}^{\infty} \beta \exp(-(\beta - t)x)\, dx \\ &= \frac{\beta}{\beta - t} \exp(-(\beta - t)x)|_{x=0}^{\infty} \\ &= \frac{\beta}{\beta - t} \quad \text{provided } t < \beta \\ &= \sum_{r=0}^{\infty} t^r \beta^{-r}. \end{aligned}$$

Hence

$$\mathsf{E}(X^r) = \frac{r!}{\beta^r},$$

for positive integers r. It follows that the expectation and variance are

$$\mathrm{E}(X) = \frac{1}{\beta} \qquad \text{and} \qquad \mathrm{var}(X) = \frac{2}{\beta^2} - \frac{1}{\beta^2} = \frac{1}{\beta^2}.$$

The expectation and standard deviation are the same, so the coefficient of variation is 1 for all β. Note that the expectation decreases with β; the higher the rate of event occurrence the shorter the expected waiting time to the next event.

The software package R can simulate m independent $\mathrm{Exp}(\beta)$ random variables using the command `rexp(m, beta)` and can generate the value of the pdf and cdf at x using the commands `dexp(x, beta)` and `pexp(x, beta)`, and the pth quartile using the command `qexp(p, beta)`.

Lack of memory A key property of the exponential distribution is its lack of memory property, which arises due to the way the exponential distribution is obtained (see Chapter 9). Exponential random variables are the only continuous random variables with this property. A random variable satisfies the memoryless property if

$$\mathrm{P}(X > s + t \mid X > t) = \mathrm{P}(X > s) \qquad \text{for} \qquad s > 0, t > 0,$$

i.e. the conditional probability that a variable exceeds $s + t$, given that it exceeds t, is independent of t, and so has no memory of how large it is already.

If we interpret X as a waiting time to an event, this means that the probability that you have to wait a further time s is independent of how long you have waited already.

4.3 Gamma Distribution

A random variable X has a gamma distribution if its pdf is given by

$$f_X(x; \boldsymbol{\theta}) = \begin{cases} \dfrac{\beta \exp(-\beta x)(\beta x)^{\alpha-1}}{\Gamma(\alpha)} & x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

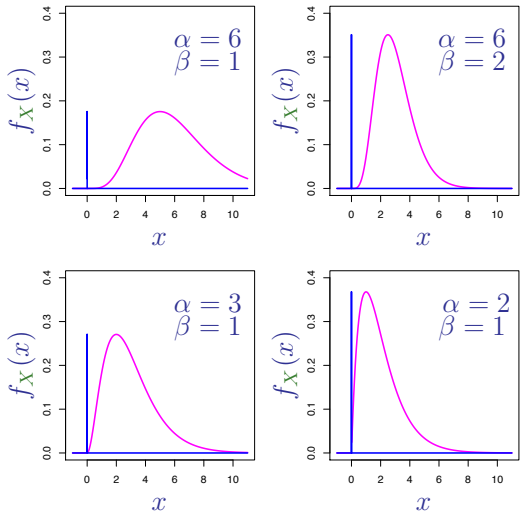
with $\boldsymbol{\theta} = (\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$, and $\Gamma(\alpha)$, called the gamma function, is defined as

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

We write $X \sim \mathrm{Gamma}(\alpha, \beta)$.

Properties of the gamma function are discussed in Appendix A, but for present purposes we only need think of it as a number, which is a function of the parameter α, which ensures the pdf satisfies the unit-integrability condition.

The parameter α is called the shape parameter and β is called the rate. The figure shows the pdf for four different sets of parameters.



Four gamma pdfs with different parameter values.

Convolution property: When α is a positive integer the $\text{Gamma}(\alpha, \beta)$ distribution is the distribution of the waiting time until a total of α events have occurred where the time between each event follows an $\text{Exp}(\beta)$ distribution. [This is proved later.]

More generally, the gamma distribution provides a flexible class of pdfs which may describe the distribution of a variable even when there is no strong probability based justification.

The software package R can simulate m independent $\text{Gamma}(\alpha, \beta)$ random variables using the command `rgamma(m, alpha, beta)` and can generate the value of the pdf and cdf at x using the commands `dgamma(x, alpha, beta)` and `pgamma(x, alpha, beta)` , and the pth quartile using the command `qgamma(p, alpha, beta)` .

Note that when $\alpha = 1$ the gamma distribution reduces to the exponential distribution. However, unlike the exponential distribution we cannot evaluate the cdf in closed form for a general (non-integer) value of α.

Standard calculations show that

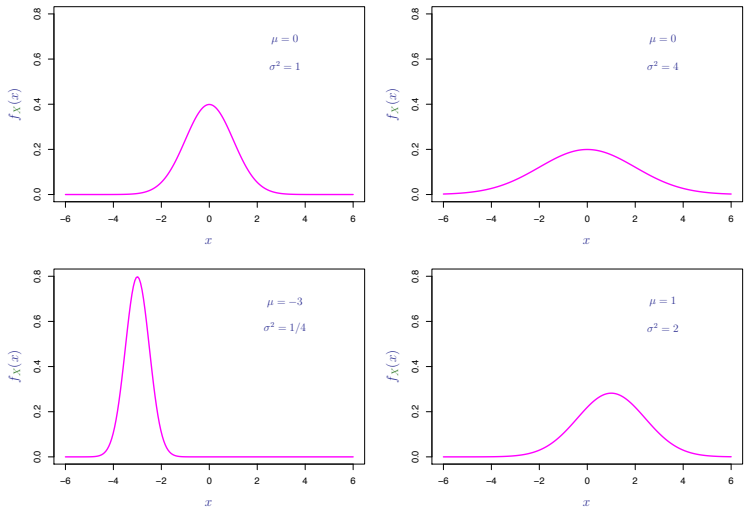
$$\begin{aligned} \mathsf{E}(X) &= \frac{\alpha}{\beta}, \\ \text{var}(X) &= \frac{\alpha}{\beta^2}, \\ M_X(t) &= (1 - t\beta)^{-\alpha}. \end{aligned}$$

4.4 Normal Distribution

A random variable X has a Normal distribution if its pdf is given by

$$f_X(x; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad \text{for} \quad -\infty < x < \infty,$$

with parameters $\boldsymbol{\theta} = (\mu, \sigma^2)$. We write $X \sim N(\mu, \sigma^2)$. Notice that all the curves are symmetric around μ with a characteristic bell-shape. The width of the bell is controlled by the value of σ^2.



Four Normal pdfs with different parameter values.

The Normal distribution has played a central role in the history of probability and statistics. It was introduced by the French mathematician Abraham de Moivre in 1733, who used it to approximate probabilities of winning in various game of chance involving coin tossing. It was later used by the German mathematician Carl Friedrich Gauss to predict the location of astronomical bodies and became known as the Gaussian distribution.

In statistics the Normal distribution is by far the most important distribution. Traditionally, it has been viewed as the natural distribution of (measurement) errors, yields from field experiments etc. The theoretical justification for this is the central limit theorem (see Chapter 8), which says that the sum of a large number of independent random variables each of which is small compared to the sum is approximately Normally distributed. The CLT is the reason why the Normal distribution often occurs as the approximate distribution of estimators in statistics.

First we need to show this is a valid pdf

$$\begin{aligned}
\int_{-\infty}^{\infty} f_X(x)\,dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-s^2/2)\,ds \qquad \text{substitution} \\
&= \frac{1}{\sqrt{2\pi}} 2 \int_{0}^{\infty} \exp(-s^2/2)\,ds \qquad \text{symmetry} \\
&= \frac{1}{\sqrt{2\pi}} 2 \int_{0}^{\infty} (2t)^{-\frac{1}{2}} \exp(-t)\,dt \qquad \text{substitution} \\
&= \frac{1}{\sqrt{\pi}} \Gamma(1/2) = 1 \qquad \text{see Appendix A} \quad .
\end{aligned}$$

Moments The expectation, variance and mgf of $X \sim N(\mu, \sigma^2)$ can be shown to be

$$\begin{aligned}
\mathsf{E}[X] &= \mu, \\
\mathsf{var}(X) &= \sigma^2, \\
M_X(t) &= e^{t\mu + \frac{1}{2}\sigma^2 t^2}.
\end{aligned}$$

Probabilities and quantiles The normal cdf cannot be expressed in closed form so numerical evaluation is required, if we want to obtain probabilities of the form $\mathsf{P}(X \leq a)$, or quantiles.

The software package R can simulate m independent $N(\mu, \sigma^2)$ random variables using the command `rnorm(m, mu, sigma)` and can generate the value of the pdf and cdf at x using the commands `dnorm(x, mu, sigma)` and `pnorm(x, mu, sigma)` , and the pth quartile using the command `qnorm(p, mu, sigma)` .

Note the R functions for the normal use the standard deviation σ, not the variance σ^2.

Standard Normal Distribution

Historically the evaluation of probabilities for Normal random variables could not be performed routinely for any different μ and σ^2 as access to computers with ability to perform integrals or to store functions was not possible.

However, it was noted that the calculation could be reduced and tables of integral values used for the evaluation. Critical to this step is the following theorem which is also of wider interest.

Theorem 4.1 *If $X \sim N(\mu, \sigma^2)$, then the random variable*

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

and conversely, if $Z \sim N(0, 1)$, then the random variable

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2).$$

We will formally prove this in Chapter 5 but for now it is sufficient to note that from previous results the transformation from X to Z ensures Z has $\mathsf{E}(Z) = 0$ and $\mathsf{var}(Z) = 1$. Furthermore shifting and scaling a random variable does not change the shape of the distribution so we would expect Z also to be Normally distributed.

A random variable Z is said to have a standard Normal distribution if it has a Normal distribution with expectation 0 and variance 1, i.e. $Z \sim N(0, 1)$. Thus Z has a standard Normal distribution if its pdf is given by

$$f_Z(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \qquad \text{for} \qquad -\infty < z < \infty.$$

The cdf of the standard Normal variable Z, and denoted by Φ, is given by

$$F_Z(z) = \mathsf{P}(Z \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx.$$

Values of $\Phi(z)$ can be obtained from the table of standard Normal probabilities, though we use R.

4.5 Beta Distribution

If $\alpha_1 > 0, \alpha_2 > 0$, the pdf of a Beta rv $X \sim \text{Beta}(\alpha_1, \alpha_2)$ is

$$f_X(x) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_1-1}(1-x)^{\alpha_2-1} \quad \text{for} \quad 0 < x < 1.$$

Calculation gives

$$\mathsf{E}(X) = \frac{\alpha_1}{\alpha_1+\alpha_2}.$$

Usage: The family of Beta distributions constitutes a flexible class of distributions on $[0,1]$ used for modelling. The Beta$(1,1)$ distribution is the uniform distribution on $[0,1]$.

The software package R can simulate m independent Beta(α_1, α_2) random variables using the command `rbeta(m, alpha1, alpha2)` and can generate the value of the pdf and cdf at x using the commands `dbeta(x, alpha1, alpha2)` and `pbeta(x, alpha1, alpha2)` , and the pth quartile using the command `qbeta(p, alpha1, alpha2)` .

4.6 Cauchy Distribution

$$\begin{aligned} f_X(x) &= \frac{1}{\pi(1+x^2)} \quad \text{for} \quad -\infty < x < \infty, \\ F_X(x) &= \frac{1}{\pi}\arctan(x) + \frac{1}{2}, \\ \mathsf{E}(X) &\quad \text{not defined, as} \\ & -\int_{-\infty}^{0} \frac{1}{\pi(1+x^2)} x dx = \int_0^{\infty} \frac{1}{\pi(1+x^2)} x dx = \infty. \end{aligned}$$

We write $X \sim$ Cauchy.

The software package R can simulate m independent Cauchy random variables using the command `rcauchy(m)` and can generate the value of the pdf and cdf at x using the commands `dcauchy(x)` and `pcauchy(x)` , and the pth quartile using the command `qcauchy(p)` .

Convolution property: If $X_1, \ldots, X_n$ are independent Cauchy, then $(X_1 + \ldots + X_n)/n \sim$ Cauchy.

Transformation property: If $X \sim \text{Uniform}(-\pi/2, \pi/2)$, then $\tan(X)$ is Cauchy-distributed. If X_1 and X_2 are independent $N(0,1)$-distributed, then X_1/X_2 is Cauchy-distributed.

Reciprocal: if $X \sim$ Cauchy then $\frac{1}{X} \sim$ Cauchy.

Other: The Cauchy distribution is the t_1 distribution.

4.7 Weibull Distribution

Parameters: $\boldsymbol{\theta} = (\beta, \eta)$ with a shape parameter $\beta > 0$ and a scale parameter $\eta > 0$.

$$f_X(x; \boldsymbol{\theta}) = (\beta/\eta)(x/\eta)^{\beta-1} \exp\{-(x/\eta)^{\beta}\} \quad \text{for} \quad 0 < x < \infty,$$

$$\begin{aligned} F_X(x) &= 1 - \exp(-(x/\eta)^\beta), \\ \mathsf{E}(X) &= \Gamma(1+\beta^{-1})\eta, \\ \mathsf{var}(X) &= \{\Gamma(1+2\beta^{-1}) - \Gamma(1+\beta^{-1})^2\}\eta^2. \end{aligned}$$

We write $X \sim \text{Weibull}(\beta, \eta)$.

The software package R can simulate m independent Weibull(β, η) random variables using the command `rweibull(m, beta, eta)` and can generate the value of the pdf and cdf at x using the commands `dweibull(x, beta, eta)` and `pweibull(x, beta, eta)` , and the pth quartile using the command `qweibull(p, beta, eta)` .

Other: The Weibull$(1, \eta)$ distribution is the Exponential(η^{-1}) distribution.
Usage: for modelling lifetimes, or times until failure. The failure rate h (or hazard rate) is given by

$$h(x) = \beta/\eta\,(x/\eta)^{\beta-1}\,.$$

Notice that when $\beta = 1$ (the Exponential case) the hazard rate is constant.

4.8 Other Distributions

In this section we present some properties of other, less common, distributions. We give these distributions in much less detail, but provide a range of properties about these distributions.

The cdf is only listed when it is available in closed form.

Chi-squared Distribution

Parameters: $\boldsymbol{\theta} = \lambda$, with $\lambda > 0$ called the degrees of freedom.

$$\begin{aligned} f_X(x;\boldsymbol{\theta}) &= \frac{1}{2^{\frac{\lambda}{2}}\Gamma(\frac{\lambda}{2})} x^{\frac{\lambda}{2}-1}\exp(-x/2) \qquad \text{for} \quad 0 < x < \infty, \\ \mathsf{E}(X) &= \lambda, \\ \mathsf{var}(X) &= 2\lambda. \end{aligned}$$

We write $X \sim \chi^2_\lambda$.

Derivation (property): If $X_1, \ldots, X_n$ are independent $N(0,1)$-distributed, then $X_1^2 + \ldots + X_n^2$ is χ^2_n-distributed.

Usage: Used in statistics as the distribution of the sum of square deviations (SSD) of a Normal sample from its mean.

Other: The χ^2_λ distribution is the Gamma$(\lambda/2, 1/2)$ distribution.

The F Distribution

Parameters : $\boldsymbol{\theta} = (\lambda_1, \lambda_2)$ with $\lambda_1 > 0$ and $\lambda_2 > 0$ called the degrees of freedom.

$$\begin{aligned} f_X(x;\boldsymbol{\theta}) &= \frac{\lambda_1^{\frac{\lambda_1}{2}} \lambda_2^{\frac{\lambda_2}{2}} x^{\frac{\lambda_1}{2}-1}}{B(\frac{\lambda_1}{2}, \frac{\lambda_2}{2})(\lambda_1 x + \lambda_2)^{\frac{\lambda_1+\lambda_2}{2}}} \qquad \text{for} \quad 0 < x < \infty, \\ \mathsf{E}(X) &= \frac{\lambda_2}{\lambda_2 - 2} \quad \text{when } \lambda_2 > 2, \\ \mathsf{var}(X) &= \frac{2\lambda_2^2(\lambda_1 + \lambda_2 - 2)}{\lambda_1(\lambda_2 - 2)^2(\lambda_2 - 4)} \quad \text{when } \lambda_2 > 4. \end{aligned}$$

We write $X \sim F_{\lambda_1,\lambda_2}$.

Transformation property: If X_1 and X_2 are independent with X_1 $\chi^2_{\lambda_1}$-distributed and X_2 $\chi^2_{\lambda_2}$-distributed, then

$$\frac{X_1/\lambda_1}{X_2/\lambda_2}$$

is F_{λ_1,λ_2}-distributed.

Usage: Used in statistics as the distribution of the test statistic in analysis of variance (ANOVA).

Other: Named after the eminent English statistician R.A. Fisher.

Log Normal Distribution

Parameters: $\boldsymbol{\theta} = (\xi, \sigma^2)$ with $\xi \in \mathbb{R}$ and $\sigma^2 > 0$.

$$\begin{aligned} f_X(x;\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi}} \frac{1}{x\sigma} \exp\{-\frac{(\log x - \xi)^2}{2\sigma^2}\} \qquad \text{for} \quad 0 < x < \infty, \\ \mathsf{E}(X) &= \exp(\frac{\sigma^2}{2} + \xi), \\ \mathsf{var}(X) &= \{\exp(\sigma^2) - 1\} \exp(\sigma^2 + 2\xi). \end{aligned}$$

We write $X \sim \log N(\xi, \sigma^2)$.

Transformation property: If X is $N(\xi, \sigma^2)$-distributed, then $\exp(X)$ is log Normal distributed with parameters (ξ, σ^2).

Useage: Used to model prices on the stock market and the size of particles during crushing processes.

Other: A random variable X follows a log-Normal distribution if and only if $\log(X)$ follows a Normal distribution.

Student's t Distribution

Parameters: $\boldsymbol{\theta} = (\mu, \lambda)$ with $\lambda > 0$ called the degrees of freedom.

$$\begin{aligned} f_X(x;\boldsymbol{\theta}) &= \frac{1}{\sqrt{\lambda}B(\frac{\lambda}{2},\frac{1}{2})\{1+\frac{(x-\mu)^2}{\lambda}\}^{\frac{\lambda+1}{2}}} \quad \text{for} \quad -\infty < x < \infty, \\ \mathrm{E}(X) &= \mu \quad \text{when } \lambda > 1 \text{ (otherwise not defined)}, \\ \mathrm{var}(X) &= \frac{\lambda}{\lambda-2} \quad \text{when } \lambda > 2. \end{aligned}$$

We write $X \sim t_\lambda$.

Derivation (property): If X and Y are independent with Y $N(0,1)$-distributed and X χ^2_λ-distributed, then

$$\mu + \frac{Y}{\sqrt{X/\lambda}}$$

is t_λ-distributed.

Usage: Used in statistics as the distribution of the test statistic for test of a hypothesis about the mean in a Normal sample with unknown variance, a so-called t test.

Other: The t distribution looks like a Normal distribution but has heavier tails. For λ going to infinity the t distribution converges to a Normal distribution.

Extreme Value Distribution

Parameters : $\boldsymbol{\theta} = (\alpha, \beta)$ with a location parameter $\alpha \in \mathbb{R}$ and a scale parameter $\beta > 0$.

$$\begin{aligned} f_X(x;\boldsymbol{\theta}) &= \frac{1}{\beta}\exp\{-(x-\alpha)/\beta\}\exp[-\exp\{-(x-\alpha)/\beta\}] \quad \text{for} \quad -\infty < x < \infty, \\ F_X(x) &= \exp[-\exp\{-(x-\alpha)/\beta\}], \\ \mathrm{E}(X) &= \alpha + \beta\gamma, \quad \text{where } \gamma \approx 0.5772 \text{ is Euler's constant}, \\ \mathrm{var}(X) &= \frac{\beta^2\pi^2}{6}. \end{aligned}$$

We write $X \sim \mathrm{GEV}(\alpha, \beta, 0)$.

Transformation property: If X is Weibull(β, η)-distributed, then $-\log(X)$ has an extreme value distribution with location parameter $\log(\eta^{-1})$ and scale parameter $1/\beta$. In particular, if X is Exp(λ)-distributed, i.e. X is Weibull$(1, \lambda^{-1})$-distributed, then $-\log(X)$ has an extreme value distribution with location parameter $\log(\lambda)$ and scale parameter 1.

Usage: Used to model extreme events.

Chapter 5

Univariate Transformations

5.1 Introduction

Sometimes we are interested in a function of a random variable X, say $Y = g(X)$. For example, we have already discussed interest in the linear transformation

$$Y = \frac{X - \mu}{\sigma}.$$

It is easy to show that in general the expectation of a function is not equal to the function of the expectation, i.e.

$$\mathsf{E}(Y) = \mathsf{E}[g(X)] \neq g(\mathsf{E}[X]).$$

For example $\mathsf{E}(X^2) > [\mathsf{E}(X)]^2$ unless X is constant, that is unless $\mathsf{Var}(X) = 0$. Therefore what can we say about the new random variable $Y = g(X)$?

In the discrete case, finding the distribution of the transformed random variable $Y = g(X)$ is a simple matter of adding up the corresponding probabilities for X i.e. the pmf

$$p_Y(r) = \sum_{s:g(s)=r} p_X(s).$$

For continuous random variables, however, we have $\mathsf{P}(X = x) = 0$ for all x so this method does not work.

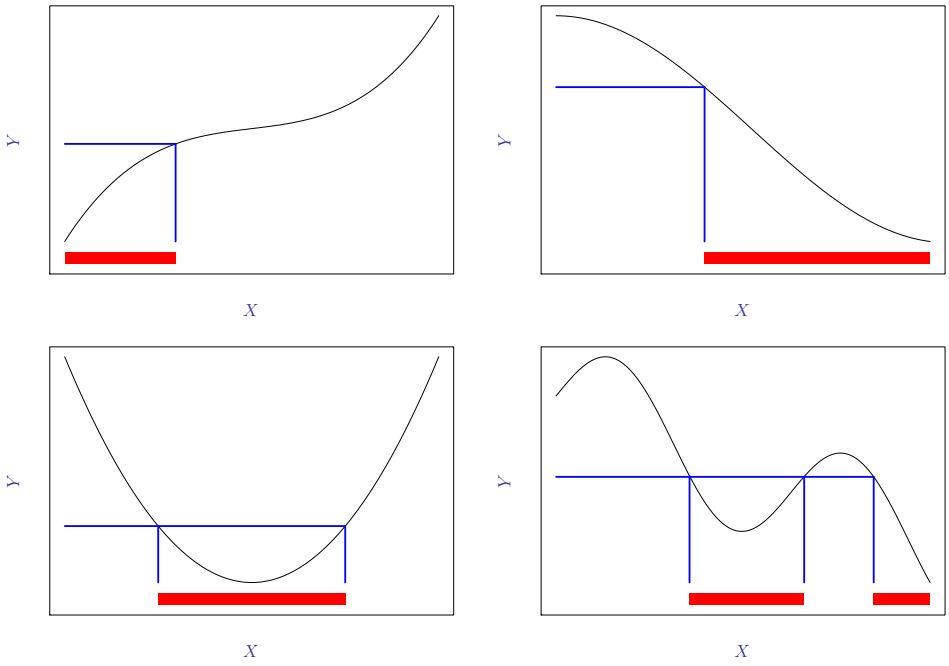
In the following we will see different methods of obtaining the cdf $F_Y(y)$ and pdf $f_Y(y)$ of the transformed random variable $Y = g(X)$ when X is a continuous random variable with cdf $F_X(x)$ and pdf $f_X(x)$.

5.2 The Distribution Function Method

The distribution function method for evaluating the distribution of a transformation is simply a technique whereby the cdf and pdf of Y is evaluated as follows:

- find the values of X which correspond to the event $Y \leq y$, let this correspond to the event $X \in A_y$ say,
- evaluate the probability $\mathsf{P}(Y \leq y) = \mathsf{P}(X \in A_y)$,
- differentiate $\mathsf{P}(Y \leq y)$ to obtain the pdf of Y.

The figure illustrates the sets A_y for various transformations $Y = g(X)$.

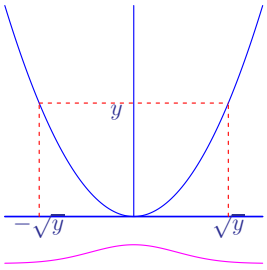


The set A_y for four different transformations $g(\cdot)$ (the curves).

When $g(\cdot)$ is monotonically increasing or decreasing A_y is an interval of the form $(-\infty, x]$ or $[x, \infty)$ and the method is particularly easy to apply in these cases. The method, however, holds whatever the properties of the transformation $g(\cdot)$.

Exercise 5.1 $X \sim N(0,1)$. Show that $Y = X^2$ has a Gamma distribution. In fact it is also a χ^2_1 distribution (see Section 4.8).

Sol:



$$F_Y(y) \quad = \quad \mathsf{P}(Y \leq y) = \mathsf{P}(X^2 \leq y) \quad \text{for} \quad y > 0$$

$$
\begin{aligned}
&= \mathrm{P}(-\sqrt{y} \le X \le \sqrt{y}) \\
&= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\
&= 2F_X(\sqrt{y}) - 1. \qquad \text{symmetry}
\end{aligned}
$$

To obtain the pdf we differentiate

$$
\begin{aligned}
f_Y(y) &= 2\frac{d}{dy}F_X(\sqrt{y}) \\
&= 2f_X(\sqrt{y})\frac{d(\sqrt{y})}{dy} \qquad \text{chain rule} \\
&= 2\frac{1}{\sqrt{2\pi}}\exp(-y/2)\frac{1}{2}y^{-1/2} \\
&= y^{-1/2}\frac{1}{\sqrt{2\pi}}\exp(-y/2) \qquad \text{for} \quad y > 0,
\end{aligned}
$$

so that $Y \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$. This also matches the χ^2_1 pdf. □

5.3 The Probability Integral Transform

The probability integral transformation is one of the most useful results in the theory of random variables. It provides a transformation for moving between $\text{Uniform}(0,1)$ distributed random variables and any continuous random variable (in either direction). By repeated use, the probability integral transformation can be used to transform any continuous random variable to any other continuous random variable. This property makes the result invaluable for simulation of random variables.

Theorem 5.1 Probability Integral Transformation. *Let Y be a continuous random variable with cdf $F_Y(y)$ and inverse cdf ${F_Y}^{-1}$ and U be a $\text{Uniform}(0,1)$ random variable. Then*

(i) $F_Y(Y)$ is a $\text{Uniform}(0,1)$ random variable and

(ii) ${F_Y}^{-1}(U)$ is a random variable with distribution function F_Y.

Proof:

$$
\begin{aligned}
\mathrm{P}\{F_Y(Y) \le y\} &= \mathrm{P}\{Y \le {F_Y}^{-1}(y)\} \\
&= F_Y\{{F_Y}^{-1}(y)\} \\
&= y
\end{aligned}
$$

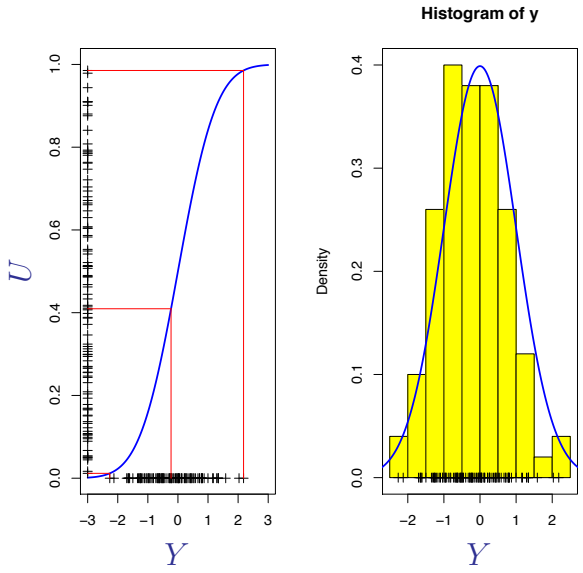
for all $0 < y < 1$, so the cdf of $F_Y(Y)$ is that of a $\text{Uniform}(0,1)$ random variable, hence $F_Y(Y) \sim \text{Uniform}(0,1)$.

Similarly,

$$
\mathrm{P}\{{F_Y}^{-1}(U) \le y\} = \mathrm{P}\{U \le F_Y(y)\} = F_Y(y)
$$

for all $-\infty < y < \infty$, so the cdf of ${F_Y}^{-1}(U)$ is F_Y. □

The transformation ${F_Y}^{-1}(U)$ with $F_Y = \Phi$ is illustrated on the figure.



Left: The distribution function $F_Y = \Phi$ for a standard Normal distribution. The vertical crosses are 100 replicates of U and the horizontal crosses are the corresponding transformed values, $Y = {F_Y}^{-1}(U)$. Right: A histogram of the 100 replicates of Y with the pdf for a standard Normal distribution superimposed.

A messy version of this theorem is available for a discrete random variable. Let Y be a discrete random variable taking values $y_1 < y_2 < \cdots$ with cdf $F_Y(y)$ and U be a $\text{Uniform}(0,1)$ random variable. Define a random variable X by setting

$$X = y_j \text{ if } F_Y(y_{j-1}) \leq U < F_Y(y_j).$$

Then X is a random variable with distribution function F_Y.

These results are very useful for simulating random variables.

5.4 The Density Method for One-to-one Transformations

Now we restrict attention to general one-to-one transformations $Y = g(X)$ such that $X = g^{-1}(Y)$ exists.

Theorem 5.2 The pdf of $1:1$ transformations. *If X has pdf $f_X(x)$ and $Y = g(X)$ defines a one-to-one transformation, then Y has pdf*

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

evaluated at $x = g^{-1}(y)$.

Proof: If g is increasing then

$$\begin{aligned} F_Y(y) &= \mathsf{P}[g(X) \leq y] \\ &= \mathsf{P}[X \leq g^{-1}(y)] \quad \text{as 1:1} \\ &= F_X(g^{-1}(y)), \end{aligned}$$

so by differentiating wrt y

$$\begin{aligned} f_Y(y) &= f_X[g^{-1}(y)]\frac{dg^{-1}(y)}{dy} \\ &= f_X[x]\frac{dx}{dy}. \end{aligned}$$

If g is decreasing then

$$\begin{aligned} F_Y(y) &= \mathsf{P}[g(X) \leq y] \\ &= \mathsf{P}[X \geq g^{-1}(y)] \\ &= 1 - \mathsf{P}[X \leq g^{-1}(y)] \\ &= 1 - F_X[g^{-1}(y)], \end{aligned}$$

so

$$\begin{aligned} f_Y(y) &= -f_X[g^{-1}(y)]\frac{dg^{-1}(y)}{dy} \\ &= f_X[x]\left(-\frac{dx}{dy}\right). \end{aligned}$$

Combining these results gives the stated result. □

Notation and hints:

- We call $|dx/dy|$ the Jacobian of the transformation.
- Sometimes it is most easy to evaluate $|dx/dy|$ by $|dy/dx|^{-1}$.
- It is often hard to remember which way up the Jacobian term should be. It is helpful to think in terms of probabilities of small sets, i.e. for suitable x and y

$$\begin{aligned} \mathsf{P}(y < Y \leq y + \delta y) &= \mathsf{P}(x < X \leq x + \delta x) \\ f_Y(y)\delta y &\approx f_X(x)\delta x \\ f_Y(y) &= f_X(x)\frac{\delta x}{\delta y}. \end{aligned}$$

In practice the procedure for finding the pdf of Y from this type of transformation is:

- Check it is a one-to-one transformation g over the range of X.
- Invert it - find x as a function of y. (This gives a way of checking it is a one-to-one transformation: can it be inverted?)
- Find dx/dy (as a function of y).
- Use the theorem, replacing x in $f_X(x)$ with $g^{-1}(y)$.
- Summarise, including the range of Y.

Exercise 5.2 Show that if $X \sim N(\mu, \sigma^2)$, then $Y = (X - \mu)/\sigma \sim N(0, 1)$.

Sol:

The range of X is $(-\infty, \infty)$ and so the range of Y is the same.

$$\begin{aligned} x &= \mu + \sigma y = g^{-1}(y) \\ \frac{dx}{dy} &= \sigma \\ f_Y(y) &= f_X(\mu + \sigma y)\sigma \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(\mu + \sigma y - \mu)^2\right] \sigma \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2) \qquad \text{for} \qquad -\infty < y < \infty. \end{aligned}$$

□

Links between Standard Distributions

Below we list some of the relationships that arise from transformations of the standard univariate continuous distributions.

X	$g(X)$	Y
$\text{Uniform}(a, b)$	$(X - a)/(b - a)$	$\text{Uniform}(0, 1)$
$\text{Uniform}(0, 1)$	$a + (b - a)X$	$\text{Uniform}(a, b)$
$\text{Exp}(\beta)$	βX	$\text{Exp}(1)$
$\text{Exp}(1)$	X/β	$\text{Exp}(\beta)$
$\text{Exp}(1)$	$\eta X^{1/\beta}$	$\text{Weibull}(\beta, \eta)$
$\text{Weibull}(\beta, \eta)$	$(X/\eta)^{\beta}$	$\text{Exp}(1)$
$\text{Uniform}(0, 1)$	$-\log(X)$	$\text{Exp}(1)$
$\text{Uniform}(0, 1)$	$-\log(1 - X)$	$\text{Exp}(1)$
$\text{Uniform}(0, 1)$	${F_Y}^{-1}(X)$	F_Y
F_X	$F_X(X)$	$\text{Uniform}(0, 1)$
$\Gamma(\alpha, \beta)$	βX	$\Gamma(\alpha, 1)$
$\Gamma(\alpha, 1)$	X/β	$\Gamma(\alpha, \beta)$
$\text{N}(\mu, \sigma^2)$	$(X - \mu)/\sigma$	$N(0, 1)$
$\text{N}(0, 1)$	$\mu + \sigma X$	$\text{N}(\mu, \sigma^2)$
$\text{N}(0, 1)$	X^2	χ^2_1
$\text{N}(\mu, \sigma^2)$	$\exp(X)$	$\log \text{N}(\mu, \sigma^2)$

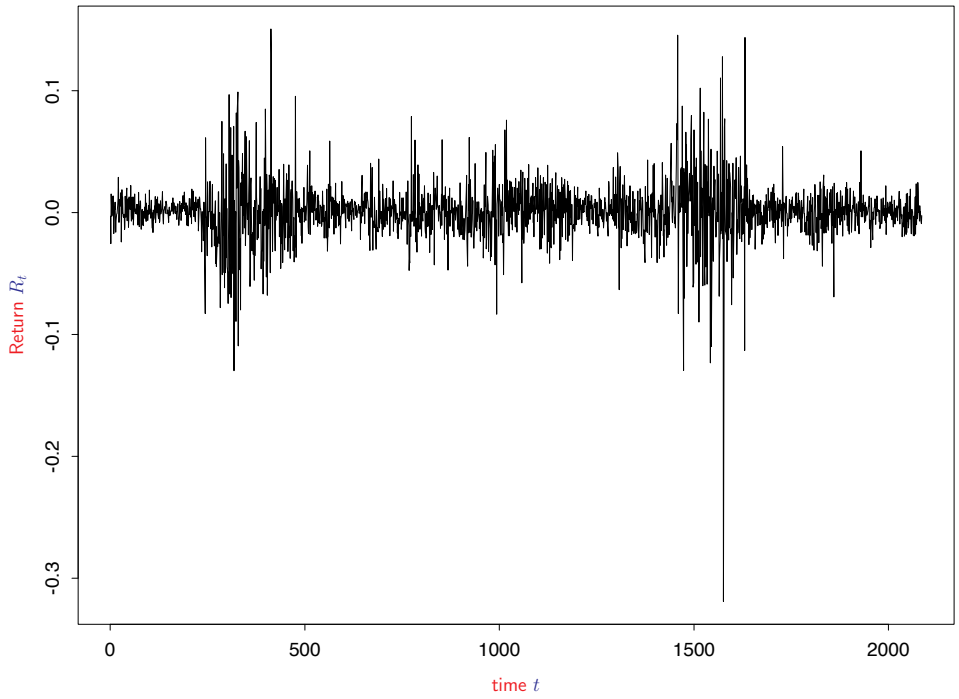
Chapter 6

Multivariate Distributions

6.1 Introduction and Motivation

Often we are interested in the joint behaviour of a number of random variables and the relationships between them. This chapter describes methods of dealing with these multivariate distributions. We consider a motivating data set which illustrates why, in many applications, it is not enough to consider the random variables individually.

Stock Market Movements



Plot of the series of daily returns, R_t, through time t.

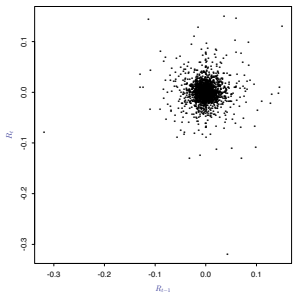
Much of the financial activity linked to the Stock Market is concerned with predicting the movements of stock prices. The better you are able to make predictions the more money that can be earned with less financial risk. The vital term to be able to predict is the variable R_t, the return of the stock on day t, this is

$$R_t = (P_t - P_{t-1})/P_{t-1}$$

where P_t is the price of the stock on day t. The figure above shows this variable on consecutive days.

Clearly there is some relationship between returns on consecutive days as the variability of values is similar for neighbouring values.

The next figure shows consecutive values, i.e. R_t plotted against R_{t-1}. Given R_{t-1} say we want to predict R_t, or at least know what this distribution of R_t will be.



Plot of returns on consecutive days.

The practical issue here is to identify association between these return values and what its effect is if we know the return on the previous day. The probability concepts in this chapter provide a basis for addressing this type of problem.

6.2 Discrete Random Variables

If $X_1, \ldots, X_n$ are discrete random variables, their joint probability mass function is

$$p_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \mathsf{P}(X_1 = x_1, \ldots, X_n = x_n).$$

As in the earlier chapters we will concentrate on discrete random variables taking integer values.

Properties of the bivariate pmf: $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ satisfies

- For all $x_1, \ldots, x_n$: $0 \leq p_{X_1 \ldots X_n}(x_1, \ldots, x_n)$.
- $\sum_{x_1,\ldots,x_n} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = 1$

- $\mathsf{P}[(X_1, \ldots, X_n) \in A] = \sum_{(x_1,\ldots,x_n)\in A} p_{X_1 \ldots X_n}(x_1, \ldots, x_n)$.

For notational simplicity we give the definitions below for two random variables, however they can all be extended to the general case.

If X and Y are discrete random variables their marginal pmfs are

$$p_X(x) = \sum_{y=-\infty}^{\infty} p_{XY}(x, y), \quad p_Y(y) = \sum_{x=-\infty}^{\infty} p_{XY}(x, y).$$

If X and Y are discrete random variables, the conditional pmfs are

$$p_{X|Y}(x \mid y) = \frac{p_{XY}(x, y)}{p_Y(y)}, \quad p_{Y|X}(y \mid x) = \frac{p_{XY}(x, y)}{p_X(x)}.$$

Independence is the simplest form for joint behaviour of two (or more) random variables. Informally, two random variables X and Y are independent if knowing the value of one of them gives no information about the value of the other.

Definition: Formally, we say that two random variables X and Y are independent if the events $\{X \in A\}$ and $\{Y \in B\}$ are independent for all sets A and B, i.e.

$$\mathsf{P}(X \in A, Y \in B) = \mathsf{P}(X \in A)\,\mathsf{P}(Y \in B) \quad \text{for all sets } A, B.$$

Theorem 6.1 Independence in terms of the pmfs. *Two discrete random variables X and Y are independent if and only if*

$$p_{XY}(x, y) = p_X(x) p_Y(y) \qquad \text{for all} \qquad x, y.$$

Proof: If X and Y are independent, discrete random variables we get by letting $A = \{x\}$ and $B = \{y\}$ that

$$\begin{aligned} p_{XY}(x, y) &= \mathsf{P}(X \in A, Y \in B) \\ &= \mathsf{P}(X \in A)\,\mathsf{P}(Y \in B) \\ &= p_X(x) p_Y(y). \end{aligned}$$

Conversely, if the joint pmf factorises we get for arbitrary sets A and B

$$\begin{aligned} \mathsf{P}(X \in A, Y \in B) &= \sum_{x\in A} \sum_{y\in B} p_{XY}(x, y) \\ &= \sum_{x\in A} \sum_{y\in B} p_X(x) p_Y(y) \quad \text{def indep} \\ &= \sum_{x\in A} p_X(x) \sum_{y\in B} p_Y(y) \quad \text{common factors} \\ &= \mathsf{P}(X \in A)\,\mathsf{P}(Y \in B). \end{aligned}$$

□

When the discrete variables (X, Y) are independent then for all x, y:

$$p_{X|Y}(x \mid y) = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x), \quad p_{Y|X}(y \mid x) = p_Y(y).$$

The converse is also true: if the conditional distribution of X given $Y = y$ is independent of y or, equivalently, the conditional distribution of Y given $X = x$ is independent of x, then X and Y are independent.

6.3 Cumulative Distribution Functions

The joint cumulative distribution function of X and Y is defined as

$$F(x, y) = F_{XY}(x, y) = \mathsf{P}(X \leq x, Y \leq y),$$

and gives the probability that a random variable X takes a value less than x and that the random variable Y takes a value less than y.

Properties of the joint cdf:

- $F_{XY}(x, y)$ is defined for all random variables, i.e. discrete, continuous or a mixture of these.
- Since it is a probability: $0 \leq F_{XY}(x, y) \leq 1$ for all x and y, and

$$\lim_{x \to -\infty} F_{XY}(x, y) = 0, \quad \lim_{y \to -\infty} F_{XY}(x, y) = 0, \quad \lim_{x,y \to \infty} F_{XY}(x, y) = 1.$$

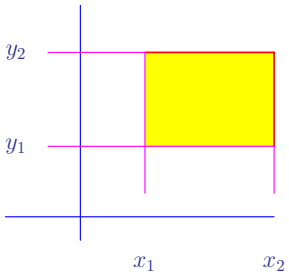
- $F_{XY}(x, y)$ is non-decreasing in both x and y, i.e. for all $h \geq 0$

$$F_{XY}(x + h, y) \geq F_{XY}(x, y) \qquad \text{and} \qquad F_{XY}(x, y + h) \geq F_{XY}(x, y).$$

- The marginal cdf $F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)$, and similarly for Y.

The probability of (X, Y) falling in a rectangle with opposite corners (x_1, y_1) and (x_2, y_2), with $x_1 < x_2$ and $y_1 < y_2$, can be found from F_{XY} using

$$\mathsf{P}(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1).$$



Exercise 6.1 Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$, with $\mathsf{P}(\{\omega_1\}) = \mathsf{P}(\{\omega_2\}) = \mathsf{P}(\{\omega_3\}) = 1/3$. Define $X, Y, Z : \Omega \to \mathbb{R}$ by

$$\begin{array}{lll} X(\omega_1) = 1, & X(\omega_2) = 2, & X(\omega_3) = 3 \\ Y(\omega_1) = 2, & Y(\omega_2) = 3, & Y(\omega_3) = 1 \\ Z(\omega_1) = 2, & Z(\omega_2) = 2, & Z(\omega_3) = 1. \end{array}$$

Show that X and Y have the same pmfs. Find the conditional pmfs $p_{Y|Z}$ and $p_{Z|Y}$.

Sol:

The random variables X and Y have the same marginal pmfs since

$$\begin{aligned} p_X(1) &= \mathsf{P}(\{\omega_1\}) = \frac{1}{3} = \mathsf{P}(\{\omega_3\}) = p_Y(1) \\ p_X(2) &= \mathsf{P}(\{\omega_2\}) = \frac{1}{3} = \mathsf{P}(\{\omega_1\}) = p_Y(2) \\ p_X(3) &= \mathsf{P}(\{\omega_3\}) = \frac{1}{3} = \mathsf{P}(\{\omega_2\}) = p_Y(3). \end{aligned}$$

The joint pmf p_{YZ} is given by

$$\begin{aligned} p_{YZ}(2,2) &= \mathsf{P}(\{\omega_1\}) = \frac{1}{3} \\ p_{YZ}(3,2) &= \mathsf{P}(\{\omega_2\}) = \frac{1}{3} \\ p_{YZ}(1,1) &= \mathsf{P}(\{\omega_3\}) = \frac{1}{3}, \end{aligned}$$

and the marginal pmf p_Z is

$$\begin{aligned} p_Z(1) &= \mathsf{P}(\{\omega_3\}) = \frac{1}{3} \\ p_Z(2) &= \mathsf{P}(\{\omega_1, \omega_2\}) = \frac{2}{3}. \end{aligned}$$

Hence

$$p_{Y|Z}(1|1) = \frac{p_{YZ}(1,1)}{p_Z(1)} = 1$$

$$
\begin{aligned}
p_{Y|Z}(2|2) &= \frac{1}{2} \\
p_{Y|Z}(3|2) &= \frac{1}{2},
\end{aligned}
$$

and

$$
\begin{aligned}
p_{Z|Y}(1|1) &= 1 \\
p_{Z|Y}(2|2) &= 1 \\
p_{Z|Y}(2|3) &= 1.
\end{aligned}
$$

□

6.4 Continuous Random Variables

If $X_1, \ldots, X_n$ are continuous random variables their joint probability density function (pdf) is defined from

$$
F_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \int_{t_n=-\infty}^{x_n} \cdots \int_{t_1=-\infty}^{x_1} f_{X_1 \ldots X_n}(t_1, \ldots, t_n) dx_1 \ldots dx_n
$$

or equivalently

$$
f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \frac{\partial^n F_{X_1 \ldots X_n}(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n}.
$$

For simplicity, we usually only state $F_{X_1 \ldots X_n}(x_1, \ldots, x_n)$ for values of $(x_1, \ldots, x_n)$ such that $f_{X_1 \ldots X_n}(x_1, \ldots, x_n) > 0$. So if $F_{X_1 \ldots X_n}$ is not defined for a particular n-tuple $(x_1, \ldots, x_n)$, then $f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = 0$ at that point.

Properties of the multivariate pdf:

- Positivity: $f_{X_1 \ldots X_n}(x_1, \ldots, x_n) \geq 0$ for all $(x_1, \ldots, x_n)$,
- Summability: $\int_{\mathbb{R}^n} f_{X_1 \ldots X_n}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = 1$.
- The probability of event A, i.e. $\mathsf{P}[(X_1, \ldots, X_n) \in A]$, is obtained by integrating the pdf over the event A:

$$
\mathsf{P}[(X_1, \ldots, X_n) \in A] = \int_{(x_1, \ldots, x_n) \in A} f_{X_1 \ldots X_n}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n.
$$

Exercise 6.2 The random variables (X, Y) have joint pdf

$$
f_{XY}(x, y) = \begin{cases} (x+y)/8 & \text{for} \quad 0 < x < 2,\ 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}
$$

(a) Find $\mathsf{P}(X > Y)$.

(b) Explain why you could have obtained this answer without doing any integration.

Sol:

$$\begin{aligned}
\mathsf{P}(X > Y) &= \int_{x=0}^{2} \int_{y=0}^{x} \frac{x+y}{8}\, dydx \\
&= \int_{x=0}^{2} \frac{1}{8} \left[xy + y^2/2\right]_{y=0}^{x} dx \\
&= \int_{x=0}^{2} \frac{3x^2}{16}\, dx \\
&= \left[\frac{x^3}{16}\right]_{x=0}^{2} \\
&= 1/2.
\end{aligned}$$

(b) By the symmetry of $f_{XY}(x, y)$ about $x = y$ line there is equal chance of $X > Y$ and $Y < X$. □

6.5 Marginal Distributions

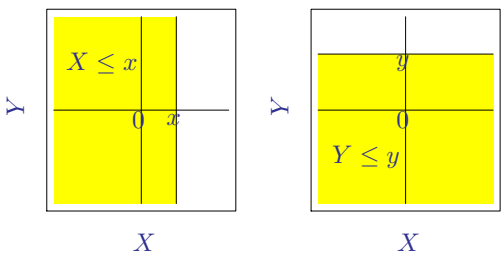
For notational simplicity, we again give the definitions below for two random variables, however they can all be extended to the general case.

Given the joint distribution of (X, Y) we may want to find the (marginal) distribution of X or Y alone. The marginal distribution tells us about the behaviour of one random variable alone, i.e. irrespective of the other. We have been studying such distributions in the earlier chapters on univariate variables.

If we have the joint cdf F_{XY} defined for $(x, y) \in (-\infty, \infty) \times (-\infty, \infty)$, the marginal cdfs are obtained as follows:

$$\begin{aligned}
F_X(x) &= \mathsf{P}(X \le x) = \mathsf{P}(X \le x, Y < \infty) = \lim_{y \to \infty} F_{XY}(x, y), \\
F_Y(y) &= \mathsf{P}(Y \le y) = \mathsf{P}(X < \infty, Y \le y) = \lim_{x \to \infty} F_{XY}(x, y),
\end{aligned}$$

because the marginal event $\{X \le x\}$ is the same as the joint event $\{X \le x, Y < \infty\}$ and the event $\{Y \le y\}$ is the same as the event $\{X < \infty, Y \le y\}$, as illustrated in the Figure.



Left: The event $\{X \le x\}$. Right: The event $\{Y \le y\}$.

If we have the joint pdf the marginal pdfs are obtained by integrating over the other variable.

Theorem 6.2 *If X and Y are* continuous random variables *their* marginal pdfs *are*

$$f_X(x) = \int_{t=-\infty}^{\infty} f_{XY}(x,t)\,dt, \quad f_Y(y) = \int_{s=-\infty}^{\infty} f_{XY}(s,y)\,ds.$$

Proof: For continuous random variables X and Y we have

$$\begin{aligned} F_X(x) &= F_{XY}(x,\infty) \\ &= \int_{s=-\infty}^{x}\int_{t=-\infty}^{\infty} f_{XY}(s,t)\,dtds \\ &= \int_{s=-\infty}^{x}\left\{\int_{t=-\infty}^{\infty} f_{XY}(s,t)\,dt\right\}ds, \end{aligned}$$

and by differentiating both sides wrt. x we get

$$f_X(x) = \int_{t=-\infty}^{\infty} f_{XY}(x,t)\,dt.$$

Similarly for Y. □

6.6 Independence

It turns out that for the independence property to hold it is *enough* that the events $\{X \le x\}$ and $\{Y \le y\}$ are independent for all x and y. Thus, two random variables X and Y are independent if and only if their joint distribution function factorises as

$$F_{XY}(x,y) = F_X(x)F_Y(y) \qquad \text{for all} \qquad x,y$$

where $F_X(x)$ and $F_Y(y)$ are the marginal cdfs of X and Y respectively.

Similarly, when X and Y are both continuous they are independent if and only if their joint pdf can be factorised as a product of the marginal pdfs.

Theorem 6.3 *Two* continuous *random variables X and Y are* independent *if and only if*

$$f_{XY}(x,y) = f_X(x)f_Y(y) \qquad \text{for all} \qquad x,y.$$

Proof: If X and Y are independent, continuous random variables we know that

$$F_{XY}(x,y) = F_X(x)F_Y(y)$$

and by differentiating both sides with respect to x and y we get

$$f_{XY}(x,y) = f_X(x)f_Y(y).$$

Conversely, if the joint pdf factorises we get for arbitrary sets A and B

$$\begin{aligned}
\mathsf{P}(X \in A, Y \in B) &= \int_A \int_B f_{XY}(x,y)dxdy \\
&= \int_A \int_B f_X(x) f_Y(y) dydx \\
&= \int_A f_X(x)dx \int_B f_Y(y)dy \\
&= \mathsf{P}(X \in A)\,\mathsf{P}(Y \in B).
\end{aligned}$$

□

Factorisation: Note that if we have the joint pdf it is enough to check that it can be factorised as a function of x, $g(x)$ say, times a function of y, $h(y)$ say:

$$f_{XY}(x,y) = g(x)h(y), \qquad \text{for all} \qquad x, y,$$

and that the range of X does not depend on Y. In other words, we do not have to show that the functions g and h are themselves densities. Note that if the range of X does not depend on Y, then the range of Y does not depend on X, so that the range condition is in fact symmetric in the way it treats x and y.

Variational independence: If the ranges of X does not depend on Y, we say that X and Y are variationally independent.

Two point method: Note that f_{XY} can be factorised as a function of x times a function of y if and only if

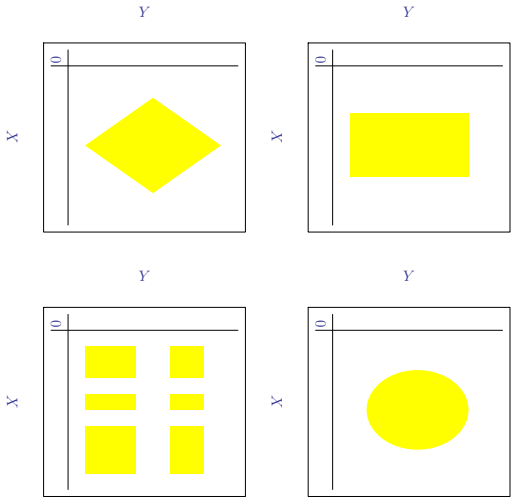
$$f_{XY}(x_1,y_1)f_{XY}(x_2,y_2) = f_{XY}(x_1,y_2)f_{XY}(x_2,y_1) \qquad \text{for all} \quad x_1, x_2, y_1, y_2.$$

This is particularly useful to prove that a given joint pdf f_{XY} does not correspond to independent random variables. Simply find (x_1, y_1), and (x_2, y_2), such that the two sides above are different.

Exercise 6.3 There are four joint density functions of the form

$$f_{XY}(x,y) = \frac{1}{|A|} 1_A(x,y),$$

illustrated in the figure. The function $1_A(x,y)$ is 1 when $(x,y) \in A$ and zero otherwise. In which cases are X and Y independent?



Four different shaded regions A for the pdf.

Sol:

Bottom left and top right: variationally independent. Top left and bottom right: not variationally independent. □

6.7 Conditional Distributions

Suppose we know the joint distribution of (X, Y) but then we find out the value of one of the random variables. What can we say about the other random variable?

We consider the conditional distributions $X \,|\, Y = y$, i.e. the distribution of X given that $Y = y$, and $Y \,|X = x$, i.e. the distribution of Y given that $X = x$. If X and Y are continuous random variables the conditional pdfs are

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}, \quad f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

Note that since we can only condition on possible values, we don't have to worry about zeros in the denominators: the marginal pdf has to be positive for the value to occur.

Also note that the conditional pdfs are themselves valid pdfs: they are non-negative and they sum/integrate to 1. For instance,

$$\begin{aligned} \int_{x=-\infty}^{\infty} f_{X|Y}(x \mid y)dx &= \int_{x=-\infty}^{\infty} \frac{f_{XY}(x,y)}{f_Y(y)}dx \\ &= \frac{1}{f_Y(y)} \int_{x=-\infty}^{\infty} f_{XY}(x,y)dx \\ &= \frac{1}{f_Y(y)} f_Y(y) = 1. \end{aligned}$$

When the continuous variables (X, Y) are independent then for all x, y:

$$f_{X|Y}(x \mid y) = \frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x), \quad f_{Y|X}(y \mid x) = f_Y(y).$$

Decomposition using conditionals

Suppose X, Y are random variables with joint pdf $f_{X,Y}(x, y)$. Often it can be difficult to simulate directly from $f_{XY}(x, y)$. Realisations of jointly distributed random variables may be obtained instead by simulating X from the marginal $f_X(x)$ (using simulation techniques for univariate random variables) and then, for the realisation $X = x$, simulating Y from the conditional $f_{Y|X}(y|x)$ (again using simulation techniques for univariate random variables).

This idea can be extended to random variables $X_1, \ldots, X_n$ with joint pdf $f_{X_1 \ldots X_n}(x_1, \ldots, x_n)$ by using their decomposition as a series of conditionals. Simulate X_1 from the marginal $f_{X_1}(x_1)$. Using the realisation $X_1 = x_1$, simulate X_2 from the conditional $f_{X_2|X_1}(x_2|x_1)$. Using the realisations $X_1 = x_1$ and $X_2 = x_2$, simulate X_3 from the conditional $f_{X_3|X_1,X_2}(x_3|x_1, x_2)$ etc.

Conversely, given a series of conditionals

$$f_{X_1}(x_1), \quad f_{X_2|X_1}(x_2|x_1), \quad \ldots, \quad f_{X_n|X_1,\ldots,X_{n-1}}(x_n|x_1, \ldots, x_{n-1}),$$

it is possible to recover the joint pdf

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) \cdots f_{X_n|X_1,\ldots,X_{n-1}}(x_n|x_1, \ldots, x_{n-1}).$$

Exercise 6.4 The random variable $X \sim \mathrm{N}(0, 1)$ and $Y|X = x \sim \mathrm{N}(\alpha x, 1)$.

(a) Write down the conditional pdf of Y given $X = x$.

(b) Write down the joint pdf of X and Y.

(c) For what values of α are X and Y independent?

Sol:

(a) The conditional pdf for $Y|X = x$ is

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(y - \alpha x)^2]$$

for $-\infty < y < \infty$.

(b) The joint pdf for X, Y is

$$\begin{aligned} f_{XY}(x, y) &= f_{Y|X}(y|x) f_X(x) \\ &= \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(y - \alpha x)^2] \times \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}x^2] \\ &= \frac{1}{2\pi} \exp[-\frac{1}{2}(y^2 - 2\alpha xy + (1 + \alpha^2)x^2)] \qquad \text{for} \quad -\infty < x < \infty, \ -\infty < y < \infty. \end{aligned}$$

(c) If $\alpha = 0$ then X and Y independent as the joint distribution factorises and the ranges are variationally independent.

If $\alpha \neq 0$, then

$$\begin{aligned} f_{XY}(0,1)f_{XY}(1,0) &= \frac{1}{(2\pi)^2}\exp[-\frac{1}{2}(2+\alpha^2)] \\ &\neq \frac{1}{(2\pi)^2}\exp[-\frac{1}{2}(2-2\alpha+\alpha^2)] \\ &= f_{XY}(0,0)f_{XY}(1,1). \end{aligned}$$

So, by the two-point method, X and Y are not independent. □

6.8 Multivariate Expectations

We know how to obtain expectations for univariate random variables. The definition extends easily to multivariate random variables. The expectation of any function $g(X_1, \ldots, X_n)$ is defined by:

Discrete random variables

$$\mathsf{E}[g(X_1,\ldots,X_n)] = \sum_{x_1=-\infty}^{\infty} \cdots \sum_{x_n=-\infty}^{\infty} g(x_1,\ldots,x_n)p_{X_1\ldots X_n}(x_1,\ldots,x_n),$$

Continuous random variables

$$\mathsf{E}[g(X_1,\ldots,X_n)] = \int_{x_1=-\infty}^{\infty} \cdots \int_{x_n=-\infty}^{\infty} g(x_1,\ldots,x_n)f_{X_1\ldots X_n}(x_1,\ldots,x_n)\,dx_n\ldots dx_1.$$

In the rest of this section results are given for the bivariate continuous random variable case only, however these extend immediately to multivariate and discrete random variables.

Notation: when the range of the integral is understood, we often write $\int_x$ instead of $\int_{x=a}^{b}$.

Moments of either variable alone can be obtained from the joint distribution or from the relevant marginal.

$$\begin{aligned} \mathsf{E}(X) &= \int_x\int_y x f_{XY}(x,y)\,dydx \\ &= \int_x x\left\{\int_y f_{XY}(x,y)\,dy\right\}dx \\ &= \int_x x f_X(x)\,dx, \end{aligned}$$

and, more generally, for a function g

$$\mathsf{E}[g(X)] = \int_x\int_y g(x)f_{XY}(x,y)\,dydx = \int_x g(x)f_X(x)\,dx.$$

Similarly for Y and a function h

$$\begin{aligned}
\mathsf{E}(Y) &= \int_y \int_x y f_{XY}(x,y)\,dxdy = \int_y y f_Y(y)\,dy, \\
\mathsf{E}[h(Y)] &= \int_y \int_x h(y) f_{XY}(x,y)\,dxdy = \int_y h(y) f_Y(y)\,dy.
\end{aligned}$$

Using linearity of integrals we also have for any functions g and h

$$\begin{aligned}
\mathsf{E}[g(X)+h(Y)] &= \int_x \int_y [g(x)+h(y)] f_{XY}(x,y)\,dydx \\
&= \int_x \int_y g(x) f_{XY}(x,y)\,dydx \\
&\quad + \int_x \int_y h(y) f_{XY}(x,y)\,dydx \\
&= \mathsf{E}[g(X)] + \mathsf{E}[h(Y)].
\end{aligned}$$

In particular

$$\mathsf{E}(X+Y) = \mathsf{E}(X) + \mathsf{E}(Y),$$

regardless of the joint distribution of (X,Y).

If X and Y are independent we also have for any functions g and h

$$\begin{aligned}
\mathsf{E}[g(X)h(Y)] &= \int_x \int_y g(x)h(y) f_{XY}(x,y)\,dydx \\
&= \int_x \int_y g(x)h(y) f_X(x) f_Y(y)\,dydx \quad \text{factor} \\
&= \left\{ \int_x g(x) f_X(x)\,dx \right\} \left\{ \int_y h(y) f_Y(y)\,dy \right\} \quad \text{v.indep} \\
&= \mathsf{E}[g(X)]\,\mathsf{E}[h(Y)].
\end{aligned}$$

Note that it is not true in general that the expectation of the product $\mathsf{E}(XY) = \mathsf{E}(X)\,\mathsf{E}(Y)$, unless X and Y are independent.

6.9 Conditional Expectations

Expectations for conditional random variables are defined in the obvious way:

$$\begin{aligned}
\mathsf{E}(X \mid Y=y) &= \int_x x f_{X|Y}(x \mid y)\,dx, \\
\mathsf{E}(Y \mid X=x) &= \int_y y f_{Y|X}(y \mid x)\,dy.
\end{aligned}$$

Sometimes conditioning provides an easy way to obtain the expectations of the marginal variables.

$$\begin{aligned}
\mathsf{E}[\,\mathsf{E}(X|Y)] &= \int_y \mathsf{E}(X \mid Y=y) f_Y(y)\,dy \\
&= \int_y \left(\int_x x f_{X|Y}(x \mid y)\,dx \right) f_Y(y)\,dy \\
&= \int_y \int_x x f_{XY}(x,y)\,dxdy \\
&= \mathsf{E}(X).
\end{aligned}$$

The conditional variances are given by

$$\begin{aligned}
\mathsf{var}(X \mid Y=y) &= \int_x [x - \mathsf{E}(X \mid Y=y)]^2 f_{X|Y}(x \mid y)\,dx \\
&= \mathsf{E}(X^2 \mid Y=y) - [\,\mathsf{E}(X \mid Y=y)]^2, \\
\mathsf{var}(Y \mid X=x) &= \int_y [y - \mathsf{E}(Y \mid X=x)]^2 f_{Y|X}(y \mid x)\,dy \\
&= \mathsf{E}(Y^2 \mid X=x) - [\,\mathsf{E}(Y \mid X=x)]^2.
\end{aligned}$$

If X and Y are independent the conditional distributions are the same as the marginal distributions, such that in particular

$$\begin{aligned}
\mathsf{E}(X \mid Y=y) &= \mathsf{E}(X), \quad \mathsf{var}(X \mid Y=y) = \mathsf{var}(X), \\
\mathsf{E}(Y \mid X=x) &= \mathsf{E}(Y), \quad \mathsf{var}(Y \mid X=x) = \mathsf{var}(Y).
\end{aligned}$$

6.10 Other Properties of Conditional Distributions

In the previous section we saw that the marginal expectations can be obtained from the conditional expectations. We can also obtain the marginal variances from the conditional means and variances by the following formula:

$$\begin{aligned}
\mathsf{E}[\,\mathsf{var}(X \mid Y)] + \mathsf{var}[\,\mathsf{E}(X \mid Y)] &= \mathsf{E}[\,\mathsf{E}(X^2 \mid Y) - (\,\mathsf{E}(X \mid Y))^2] \\
&\quad + \mathsf{E}[(\,\mathsf{E}(X \mid Y))^2] - (\,\mathsf{E}[\,\mathsf{E}(X \mid Y)])^2 \\
&= \mathsf{E}(X^2) - [\,\mathsf{E}(X)]^2 \\
&= \mathsf{var}(X).
\end{aligned}$$

These formulae are particularly useful when a random variable X is given as a mixture of distributions.

Exercise 6.5 The number of calls received by a call centre on a given day has mean μ_N and variance σ_N^2. The time taken to deal with each call has mean μ_T and variance σ_T^2 and is distributed

independently of the other calls and of the total number of calls. Find the mean and variance of the total time spent dealing with calls on that day.

Sol:

Suppose N calls are received and the time taken to deal with the ith call is given by T_i. Then the total time spent dealing with calls is

$$T = T_1 + \cdots + T_N.$$

Now

$$\begin{aligned} \mathsf{E}(T|N=n) &= \mathsf{E}[T_1 + \cdots + T_n] = n\mu_T, \\ \mathsf{var}(T|N=n) &= \mathsf{var}(T_1 + \cdots + T_n) = n\sigma_T^2. \end{aligned}$$

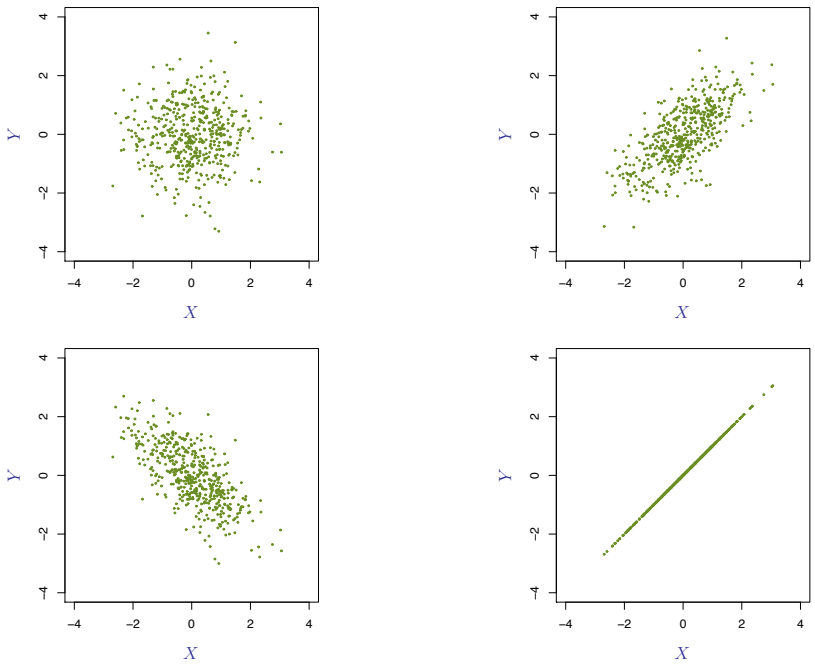
Hence

$$\begin{aligned} \mathsf{E}[T] &= \mathsf{E}[\,\mathsf{E}(T|N)] = \mathsf{E}[N\mu_T] = \mu_N\mu_T, \\ \mathsf{var}(T) &= \mathsf{E}[\,\mathsf{var}(T|N)] + \mathsf{var}[\,\mathsf{E}(T|N)] = \mathsf{E}[N\sigma_T^2] + \mathsf{var}[N\mu_T] \\ &= \mu_N\sigma_T^2 + \sigma_N^2\mu_T^2. \end{aligned}$$

□

6.11 Covariance and Correlation

The figure shows samples from four different joint distributions. In all cases the variables have the same $\mathrm{N}(0,1)$ marginal distribution for both X and Y, however the joint distributions have very different forms as they have different dependence structures. In this section we will try to characterise the dependence through a summary measure.



Four different joint distributions of X and Y. The marginal distributions are the same in all cases.

Throughout this section we use the notation

$$\mathsf{E}(X) = \mu_X, \quad \mathsf{E}(Y) = \mu_Y,$$

$$\mathsf{std}(X) = \sigma_X, \quad \mathsf{std}(Y) = \sigma_Y.$$

The most common way of describing the relationship between two random variables is through the covariance or correlation.

The covariance between X and Y is

$$\begin{aligned}
\mathsf{cov}(X,Y) &= \mathsf{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \mathsf{E}[XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y] \\
&= \mathsf{E}[XY] - \mu_X\,\mathsf{E}[Y] - \mathsf{E}[X]\mu_Y + \mu_X\mu_Y \\
&= \mathsf{E}(XY) - \mathsf{E}(X)\,\mathsf{E}(Y) \\
&= \mathsf{E}(XY) - \mu_X\mu_Y.
\end{aligned}$$

Note that $\mathsf{cov}(X,Y) = \mathsf{cov}(Y,X)$ and $\mathsf{cov}(X,X) = \mathsf{var}(X)$.

The covariance occurs in the variance of sums of random variables. Consider

$$\begin{aligned}
\mathsf{var}(X+Y) &= \mathsf{E}(X+Y)^2 - [\mathsf{E}(X+Y)]^2 \\
&= \mathsf{E}(X^2+2XY+Y^2) - [\mathsf{E}(X) + \mathsf{E}(Y)]^2 \\
&= \mathsf{E}(X^2) + \mathsf{E}(2XY) + \mathsf{E}(Y^2) - [\mathsf{E}(X^2) + 2\,\mathsf{E}(X)\,\mathsf{E}(Y) + \mathsf{E}(Y)^2] \\
&= \mathsf{E}(X^2) - \mathsf{E}(X)^2 + \mathsf{E}(Y^2) - \mathsf{E}(Y)^2 + 2[\mathsf{E}(XY) - \mathsf{E}(X)\,\mathsf{E}(Y)] \\
&= \mathsf{var}(X) + \mathsf{var}(Y) + 2\,\mathsf{cov}(X,Y).
\end{aligned}$$

Covariance has units = (units of X)$\times$(units of Y) and changes if we change the scale of either X or Y,

$$\mathsf{cov}(aX+b, cY+d) = ac\ \mathsf{cov}(X,Y).$$

The correlation, $\mathsf{corr}(X,Y)$, between X and Y is

$$\mathsf{corr}(X,Y) = \frac{\mathsf{cov}(X,Y)}{\sqrt{\mathsf{var}(X)\,\mathsf{var}(Y)}} = \frac{\mathsf{cov}(X,Y)}{\sigma_X\sigma_Y}.$$

It can be shown that $-1 \le \mathsf{corr}(X,Y) \le 1$.

Proof: Replacing X by $X/\sqrt{\mathsf{var}(X)}$ and Y by $Y/\sqrt{\mathsf{var}(Y)}$ leaves the correlation unchanged. So we may suppose $\mathsf{var}(X) = 1 = \mathsf{var}(Y)$. Then

$$\mathsf{var}(X+Y) = \mathsf{var}(X) + \mathsf{var}(Y) + 2\,\mathsf{cov}(X,Y) = 2 + 2\,\mathsf{corr}(X,Y).$$

As the variance must be positive, $-1 \le \mathsf{corr}(X,Y)$. By considering $\mathsf{var}(X-Y)$, it can be shown that $\mathsf{corr}(X,Y) \le 1$. □

The correlation $\mathrm{corr}(X, Y)$ is often denoted by ρ_{XY}, in the same way that $\mathrm{E}(X)$ is denoted by μ_X.

Correlation has the benefit of being invariant to location and scale changes, which aids interpretation,

$$\mathrm{corr}(aX + b, cY + d) = \mathrm{sgn}\,(ac)\,\mathrm{corr}(X, Y).$$

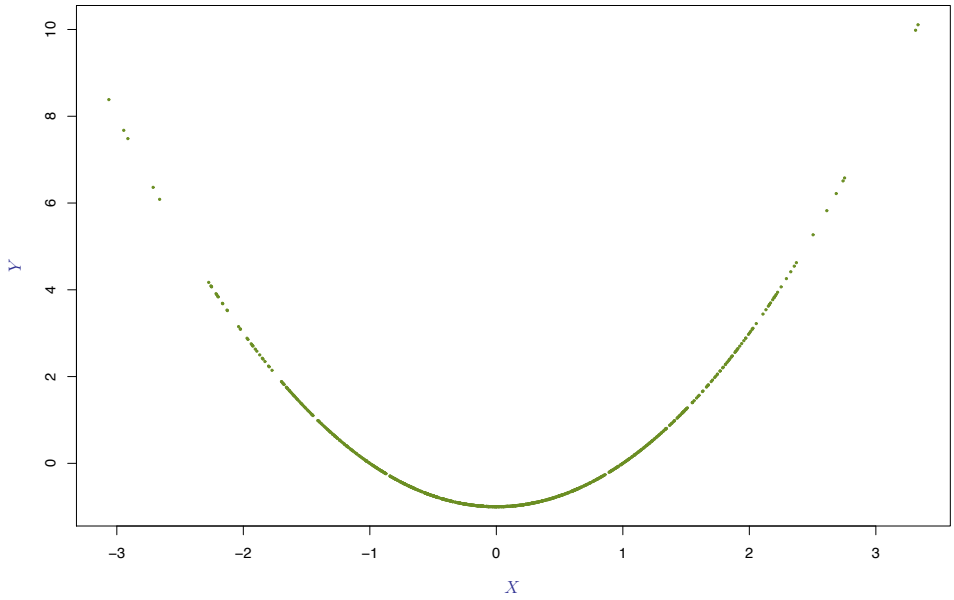
The interpretation of the covariance/correlation between X and Y is that if one variable tends to increase when the other does then both the covariance and the correlation will be positive, and the stronger the association between X and Y the larger the value of the covariance and correlation, with $\rho_{XY} = 1$ corresponding to perfect positive linear association. If one variable tends to decrease when the other increases then both the covariance and the correlation will be negative, with $\rho_{XY} = -1$ corresponding to perfect negative linear association. The figure shows four joint distributions with different correlations.

When X and Y are independent we have $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$ so the covariance

$$\mathrm{cov}(X, Y) \quad = \quad \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = 0.$$

and the correlation is $\rho_{XY} = 0$.

The converse, however, is not true: $\rho_{XY} = 0$ does not imply that X and Y are independent as the following example shows.



A 1000 realisations of (X, Y), where $X \sim N(0, 1)$ and $Y = X^2 - 1$. While X and Y are uncorrelated ($\rho_{XY} = 0$), they are not independent.

Exercise 6.6 Let $X \sim \mathrm{N}(0, 1)$ and $Y = X^2 - 1$. The joint distribution of (X, Y) is illustrated in joint distribution. Clearly the variables X and Y are strongly related, as given X we know Y exactly. Show that $\rho_{XY} = \mathrm{cov}(X, Y) = 0$.

Sol:

$$\begin{aligned} \mathrm{cov}(X,Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) \\ &= \mathrm{E}(X^3 - X) = \mathrm{E}(X^3) - \mathrm{E}(X) = 0 - 0 = 0, \end{aligned}$$

since $\mathrm{E}(X^r) = 0$ for r an odd integer. □

In general, we must be careful not to interpret too much into the value of these summary measures as both covariance and correlation measure linear association only and not association.

6.12 Expectation and Variance Matrices

Matrix notation provides an excellent method of handling summary information of multivariate distributions.

If $X_1, \ldots, X_n$ are random variables, we can define a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$. The expectation, $\mathrm{E}(\boldsymbol{X})$, is the $n \times 1$ vector with elements $\mathrm{E}(X_i)$ and the variance, $\mathrm{var}(\boldsymbol{X})$, is the $n \times n$ matrix with elements $\mathrm{cov}(X_i, X_j)$ i.e. the variance matrix is defined by

$$\mathrm{var}(\boldsymbol{X}) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \ldots & \mathrm{cov}(X_1,X_n) \\ \mathrm{cov}(X_2,X_1) & \mathrm{var}(X_2) & \ldots & \mathrm{cov}(X_2,X_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_2,X_n) & \ldots & \mathrm{var}(X_n) \end{pmatrix}$$

Notes: Variance matrices are sometimes called variance-covariance matrices.

- The mean vector is simply the vector of means; the variance matrix has variances down the diagonal, covariances as off-diagonals.
- The variance matrix is always symmetric, because $\mathrm{cov}(X_i, X_j) = \mathrm{cov}(X_j, X_i)$, and positive (semi) definite.

The linearity properties of expectation and covariance can be expressed very succinctly in matrix notation.

Expectation

Suppose Y is a linear combination of a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$,

$$Y = \boldsymbol{a}'\boldsymbol{X} = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

where $\boldsymbol{a}' = (a_1, \ldots, a_n)$ is a vector of known constants.

Since expectation is linear the expectation of Y is given by

$$\mathrm{E}(Y) = a_1\,\mathrm{E}(X_1) + a_2\,\mathrm{E}(X_2) + \ldots + a_n\,\mathrm{E}(X_n) = \boldsymbol{a}'\,\mathrm{E}(\boldsymbol{X}).$$

Note that this holds whatever the dependence structure between the variables is.

Now suppose we are interested in several linear combinations, say m. We can collect these into a vector and write

$$\boldsymbol{Y} = A\boldsymbol{X}$$

where A is an $m \times n$ matrix of constants.

By similar arguments as above it can be shown that the mean vector of $\boldsymbol{Y}$ is

$$\mathsf{E}(\boldsymbol{Y}) = A\,\mathsf{E}(\boldsymbol{X}).$$

An important special case is the formula for the expectation of the mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\mathsf{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}(X_i).$$

The expectation of the mean is the mean of the expectations. If the sample is independent identically distributed (iid) then $\mathsf{E}(X_i) = \mu$ for all i, so $\mathsf{E}(\bar{X}) = \mu$ as well. In vector notation $\bar{X} = \frac{1}{n}1'\boldsymbol{X}$ and $\mathsf{E}(\bar{X}) = \frac{1}{n}1'\,\mathsf{E}(\boldsymbol{X})$ where 1 is a vector of ones.

Covariance

Suppose there are two linear transforms

$$\begin{aligned} Y_1 &= \boldsymbol{a}'\boldsymbol{X} = a_1X_1 + a_2X_2 + \ldots + a_nX_n \\ Y_2 &= \boldsymbol{b}'\boldsymbol{X} = b_1X_1 + b_2X_2 + \ldots + b_nX_n \end{aligned}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors of known constants. The main result is

Theorem 6.4 The covariance sandwich theorem.

$$\mathsf{cov}(Y_1, Y_2) = \mathsf{cov}(\boldsymbol{a}'\boldsymbol{X}, \boldsymbol{b}'\boldsymbol{X}) = \boldsymbol{a}'\,\mathsf{var}(\boldsymbol{X})\boldsymbol{b}.$$

Proof: Case $n = 2$. We first observe that the covariance is bilinear, that is linear in both its arguments:

$$\begin{aligned} \mathsf{cov}(a_1X_1 + a_2X_2, Y_2) &= \mathsf{E}[(a_1X_1 + a_2X_2)Y_2] - \mathsf{E}(a_1X_1 + a_2X_2)\,\mathsf{E}(Y_2) \\ &= \mathsf{E}(a_1X_1Y_2) + \mathsf{E}(a_2X_2Y_2) - \mathsf{E}(a_1X_1)\,\mathsf{E}(Y_2) - \mathsf{E}(a_2X_2)\,\mathsf{E}(Y_2) \\ &= a_1\,[\mathsf{E}(X_1Y_2) - \mathsf{E}(X_1)\,\mathsf{E}(Y_2)] + a_2\,[\mathsf{E}(X_2Y_2) - \mathsf{E}(X_2)\,\mathsf{E}(Y_2)] \\ &= a_1\,\mathsf{cov}(X_1, Y_2) + a_2\,\mathsf{cov}(X_2, Y_2). \end{aligned}$$

Similarly,

$$\mathsf{cov}(Y_1, b_1X_1 + b_2X_2) = b_1\,\mathsf{cov}(Y_1, X_1) + b_2\,\mathsf{cov}(Y_1, X_2).$$

These linearity relations generalise immediately from $n = 2$ to arbitrary n. Consequently

$$\begin{aligned}
\mathrm{cov}(Y_1, Y_2) &= \mathrm{cov}(\boldsymbol{a}'\boldsymbol{X}, Y_2) \\
&= \boldsymbol{a}'\,\mathrm{cov}(\boldsymbol{X}, Y_2) \\
&= \boldsymbol{a}'\,\mathrm{cov}(\boldsymbol{X}, \boldsymbol{b}'\boldsymbol{X}) \\
&= \boldsymbol{a}'\,\mathrm{cov}(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{b} \\
&= \boldsymbol{a}'\,\mathrm{var}(\boldsymbol{X})\boldsymbol{b}.
\end{aligned}$$

□

Another way of writing the covariance between two linear combinations is

$$\begin{aligned}
\mathrm{cov}(\boldsymbol{a}'\boldsymbol{X}, \boldsymbol{b}'\boldsymbol{X}) &= \mathrm{cov}(a_1X_1 + \ldots + a_nX_n, b_1X_1 + \ldots + b_nX_n) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_ib_j\,\mathrm{cov}(X_i, X_j).
\end{aligned}$$

Remembering that $\mathrm{var}(Y) = \mathrm{cov}(Y, Y)$ we can now obtain the formula for the variance of the linear combination $Y = \boldsymbol{a}'\boldsymbol{X}$

$$\begin{aligned}
\mathrm{var}(\boldsymbol{a}'\boldsymbol{X}) &= \mathrm{cov}(\boldsymbol{a}'\boldsymbol{X}, \boldsymbol{a}'\boldsymbol{X}) \quad \text{def of var and cov} \\
&= \boldsymbol{a}'\,\mathrm{cov}(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{a} \quad \text{bilinearity of cov} \\
&= \boldsymbol{a}'\,\mathrm{var}(\boldsymbol{X})\boldsymbol{a}. \quad \text{def again}
\end{aligned}$$

The long hand version is

$$\begin{aligned}
\mathrm{var}(\boldsymbol{a}'\boldsymbol{X}) &= \mathrm{cov}(\boldsymbol{a}'\boldsymbol{X}, \boldsymbol{a}'\boldsymbol{X}) \\
&= \mathrm{cov}(a_1X_1 + \ldots + a_nX_n, a_1X_1 + \ldots + a_nX_n) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\,\mathrm{cov}(X_i, X_j).
\end{aligned}$$

For $n = 1$ we get back the familiar expression

$$\mathrm{var}(a_1X_1) = a_1^2\,\mathrm{var}(X_1).$$

For $n = 2$ we get

$$\begin{aligned}
&\mathrm{var}(a_1X_1 + a_2X_2) \\
&= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\
&= a_1^2\,\mathrm{var}(X_1) + a_2^2\,\mathrm{var}(X_2) + 2a_1a_2\,\mathrm{cov}(X_1, X_2).
\end{aligned}$$

Now suppose we have several linear combinations

$$\boldsymbol{Y} = A\boldsymbol{X}$$

where A is an $m \times n$ matrix of constants.

By similar arguments as above it can be shown that the variance matrix of $\boldsymbol{Y}$ becomes

$$\text{var}(\boldsymbol{Y}) = A \, \text{var}(\boldsymbol{X}) A'.$$

Independence: When $X_1, \ldots, X_n$ are independent and $Y = \boldsymbol{a}'\boldsymbol{X}$ then

$$\begin{aligned} \text{var}(Y) &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \, \text{cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} a_i a_i \, \text{cov}(X_i, X_i) + 0 \\ &= \sum_{i=1}^{n} a_i^2 \, \text{var}(X_i), \end{aligned}$$

because $\text{cov}(X_i, X_j) = 0$ for $i \neq j$.

In particular, when $X_1, \ldots, X_n$ are independent,

$$\text{var}(X_1 + \ldots + X_n) = \text{var}(X_1) + \ldots + \text{var}(X_n).$$

The variance of the sum is the sum of the variances, when $X_1, \ldots, X_n$ are independent.

When $X_1, \ldots, X_n$ are independent we also get a simple formula for the variance of the mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

$$\text{var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{var}(X_i),$$

and if further $X_1, \ldots, X_n$ have the same variance this simplifies to

$$\text{var}(\bar{X}) = \frac{1}{n} \text{var}(X_1).$$

In particular, this formula holds when $X_1, \ldots, X_n$ are iid (independent, identically distributed).

Exercise 6.7 Find $\text{var}(X + Y)$, $\text{var}(X - Y)$,and $\text{cov}(X + Y, X - Y)$, when the variances are σ_X^2 and σ_Y^2 and their correlation is ρ_{XY}.

Sol:

Let $\boldsymbol{W} = (X, Y)'$. Then

$$\text{var}(\boldsymbol{W}) = \begin{bmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}.$$

Let $\boldsymbol{Z} = (X + Y, X - Y)$. Then

$$\boldsymbol{Z} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \boldsymbol{W}$$

and so

$$\begin{aligned}
\mathsf{var}(\boldsymbol{Z}) &= \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \mathsf{var}(\boldsymbol{W}) \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_X^2 + \sigma_Y^2 + 2\sigma_X\sigma_Y\rho_{XY} & \sigma_X^2 - \sigma_Y^2 \\ \sigma_X^2 - \sigma_Y^2 & \sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y\rho_{XY} \end{bmatrix}.
\end{aligned}$$

□

Chapter 7

Multivariate Transformations

Suppose we are interested in the distribution of a function of several random variables. For example, at the end of the last chapter, we looked at how the joint distribution of $F_X^{-1}(X)$ and $F_Y^{-1}(Y)$ gives rise to copulas. We considered this problem in one dimension in Chapter 5 and gave various methods for obtaining the cdf and pdf. The distribution function method extends immediately to higher dimensions, but in practice is hard to use, so in this chapter we focus on the density method. To keep the presentation simple, for the general detail we focus on bivariate transformations. (The extension to higher dimensions is straight-forward conceptually but messy mathematically.)

7.1 One-to-one Bivariate Transformations

Suppose that there are two random variables X and Y which have joint pdf f_{XY}. We are interested in the joint distribution of two new random variables,

$$S = g_1(X, Y) \qquad \text{and} \qquad T = g_2(X, Y)$$

which are functions of (X, Y).

We assume that the transformation from $(X, Y) \to (S, T)$ is a one-to-one bivariate transformation, so that there exist functions h_1 and h_2 such that $X = h_1(S, T)$ and $Y = h_2(S, T)$. Then the joint pdf of (S, T) is

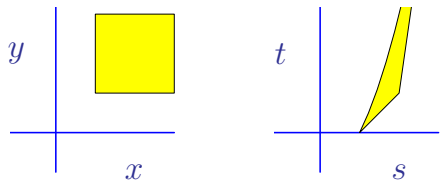
$$f_{ST}(s, t) = f_{XY}(x, y) \mid \det J \mid |_{x=h_1(s,t), y=h_2(s,t)}$$

where $\mid \det J \mid$ is the absolute value of the determinant of J; and J is the Jacobian matrix of the transformation:

$$J = \begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{dt} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{dt} \end{bmatrix};$$

and where both $f_{XY}(x, y)$ and J are evaluated as functions of s and t, as indicated by $x = h_1(s, t), y = h_2(s, t)$. In the diagram the joint density $f_{XY}(x, y) > 0$ on the region indicated and leads to a joint density

$f_{ST}(s,t) > 0$ on its region induced by the transformation.



Summary: the transformation procedure is:

- Check one-to-one bivariate transformation. (Given x and y can we find s and t uniquely, and given s and t can we find x and y uniquely?)
- Invert the transformation: find s and t as functions of x and y. (Again this might be an easy way of checking whether it is a one-to-one transformation).
- Find the Jacobian.
- Use the formula, replacing x and y in $f_{XY}(x,y)$ and J by the appropriate functions of s and t, $x = h_1(s,t)$ and $y = h_2(s,t)$.
- Summarise, taking care with the ranges of S and T.

As in the univariate case it is sometimes easier to calculate the inverse $|\det J|^{-1}$ using

$$|\det J|^{-1} = \left|\det \begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{ds} & \frac{\partial y}{\partial t} \end{bmatrix}\right|^{-1} = \left|\det \begin{bmatrix} \frac{\partial s}{\partial x} & \frac{\partial s}{\partial y} \\ \frac{\partial t}{\partial x} & \frac{\partial t}{\partial y} \end{bmatrix}\right|.$$

Exercise 7.1 Suppose X and Y are independent, X with an $\mathrm{Exp}(1)$ distribution and Y with a $\mathrm{Uniform}(0, 2\pi)$ distribution. Find the joint and marginal pdfs of

$$(S,T) = (\sqrt{2X}\cos(Y), \sqrt{2X}\sin(Y)),$$

i.e. if $(\sqrt{2X}, Y)$ are the polar coordinates of a point in the plane then (S,T) are the corresponding Cartesian coordinates.

Sol:

First note that the sample space of (S,T) is $-\infty < s < \infty$ and $-\infty < t < \infty$.

The joint pdf of (X,Y) is

$$\begin{aligned} f_{XY}(x,y) &= f_X(x)\, f_Y(y) \\ &= \exp(-x)\,\frac{1}{2\pi} \qquad \text{for} \quad 0 < x < \infty, 0 < y < 2\pi. \end{aligned}$$

In this case it is easier to find the inverse $|\det J|^{-1}$

$$
\begin{aligned}
|\det J|^{-1} &= \left|\det \begin{bmatrix} \frac{\partial s}{\partial x} & \frac{\partial s}{\partial y} \\ \frac{\partial t}{\partial x} & \frac{\partial t}{\partial y} \end{bmatrix}\right| \\
&= \left|\det \begin{bmatrix} \frac{1}{\sqrt{2x}}\cos(y) & -\sqrt{2x}\sin(y) \\ \frac{1}{\sqrt{2x}}\sin(y) & \sqrt{2x}\cos(y) \end{bmatrix}\right| \\
&= 1.
\end{aligned}
$$

Since $X = (S^2 + T^2)/2$ we get

$$
\begin{aligned}
f_{ST}(s,t) &= \frac{1}{2\pi}\exp\left(-\frac{s^2+t^2}{2}\right) \\
&= \frac{1}{\sqrt{2\pi}}\exp(-s^2/2)\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)
\end{aligned}
$$

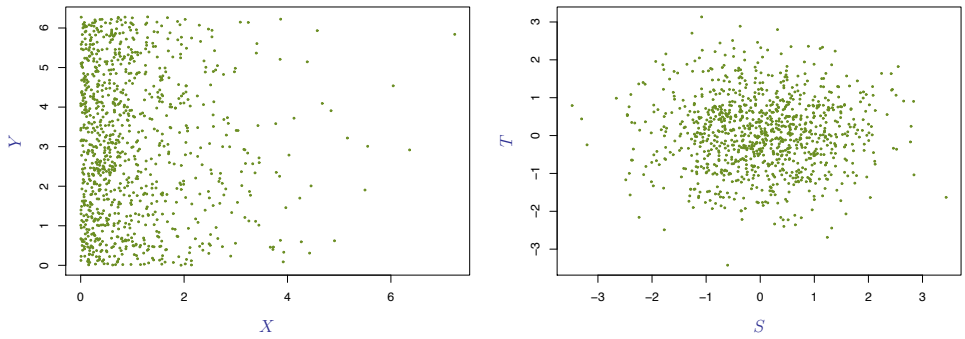
with the range given by $-\infty < s < \infty$ and $-\infty < t < \infty$. So S and T are independent and identically distributed $N(0,1)$ random variables. □

The transformation in this exercise is called the Box-Muller transformation, which is useful for simulating Normal random variables.

Remember that we can generate an $\mathrm{Exp}(1)$ random variable from a $\mathrm{Uniform}(0,1)$ by the transformation $X = -\log(1-U)$, thus we can generate two independent $\mathrm{N}(0,1)$ random variables X_1 and X_2 from two independent $\mathrm{Uniform}(0,1)$ random variables U_1 and U_2 by

$$
(X_1, X_2) = (\sqrt{-2\log(1-U_1)}\cos(2\pi U_2), \sqrt{-2\log(1-U_1)}\sin(2\pi U_2)).
$$

The Box-Muller transformation for generating standard Normal random variables is illustrated here.



7.2 Use of Dummy Variables

Often we are interested in not two new variables, S and T, but in just one, S say. To obtain the pdf of S alone we have to create a dummy variable T, obtain the joint pdf of S and T, then integrate to get the marginal distribution of S.

Suppose interest lies in $S = g_1(X, Y)$.

- Define a new variable $T = g_2(X, Y)$ which makes a one-to-one bivariate transformation between (X, Y) and (S, T). The choice of T is essentially arbitrary and can be made for convenience. Sometimes some trial and error is required.
- Find the joint pdf $f_{ST}(s, t)$ of S and T using the methods in the previous section.
- Find the marginal pdf of S:

$$f_S(s) = \int_t f_{ST}(s, t)dt$$

  taking care with the range of integration.

Convolution A transformation of general interest is $S = X + Y$. We use the dummy variable method to obtain the pdf of S. We make the transformation

$$S = X + Y \qquad \text{and} \qquad T = X,$$

with T as the dummy variable. It follows that the inverse transformation is

$$X = T \qquad \text{and} \qquad Y = S - T,$$

so

$$|\det J| = \left|\det \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}\right| = |-1| = 1.$$

Thus

$$f_{ST}(s,t) = f_{XY}(t, s-t),$$

so the marginal pdf of $S = X + Y$ is

$$f_S(s) = \int_{t=-\infty}^{\infty} f_{XY}(t, s-t)\, dt.$$

This formula is known as the convolution formula. It is finding the probability of $S = X + Y$ by summing the probabilities, over all possible t, for the pairs $(t, s-t)$ in (X, Y).

7.3 Multivariate Normal Distribution

Let $X_1, \ldots, X_n$ be independent $N(0,1)$ random variables. The joint density is then

$$\begin{aligned} f_{X_1 \ldots X_n}(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_i^2} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}}, \end{aligned}$$

where $\boldsymbol{x} = (x_1 \ldots x_n)'$. Using vector notation, $\boldsymbol{X} = (X_1, \ldots, X_n)'$, let $\boldsymbol{Y} = \boldsymbol{\mu} + A\boldsymbol{X}$, where A is an invertible matrix.

Then since the transformation is invertible (with inverse transform $\boldsymbol{X} = A^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$), the density method can be used to evaluate the pdf of $\boldsymbol{Y}$

$$\begin{aligned} f_{Y_1 \ldots Y_n}(y_1, \ldots, y_n) &= \frac{1}{|A|(2\pi)^{n/2}} e^{-\frac{1}{2}(A^{-1}(\boldsymbol{y}-\boldsymbol{\mu}))' A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})} \\ &= \frac{1}{|A|(2\pi)^{n/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'(A^{-1})' A^{-1}(\boldsymbol{y}-\boldsymbol{\mu})} \\ &= \frac{1}{|\Sigma|^{1/2}(2\pi)^{n/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}, \end{aligned}$$

where $\Sigma = AA'$. We say that Y has a multivariate Normal distribution of n dimensions, or

$$\boldsymbol{Y} \sim \mathrm{MVN}_n(\boldsymbol{\mu}, \Sigma).$$

Note that

$$\mathrm{E}[\boldsymbol{Y}] = \mathrm{E}[\boldsymbol{\mu} + A\boldsymbol{X}] = \boldsymbol{\mu} + A\,\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu}$$

and

$$\mathrm{var}(\boldsymbol{Y}) = \mathrm{E}[(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})'] = \mathrm{E}[(A\boldsymbol{X})(A\boldsymbol{X})'] = A\,\mathrm{E}[\boldsymbol{X}\boldsymbol{X}']A' = AI_nA' = \Sigma.$$

So Σ is the covariance matrix of $\boldsymbol{Y}$. Since linear combinations of Normal random variables are themselves Normal, the univariate marginal distributions are given by

$$X_i \sim N(\mu_i, \sigma_i^2),$$

where $\sigma_i^2 = \Sigma_{ii}$.

Suppose now that $\boldsymbol{X} \sim \text{MVN}_n(\boldsymbol{\mu}, \Sigma)$. If X_i, X_j are uncorrelated for all $i \neq j$, then the covariance matrix is diagonal, say

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix}.$$

The pdf becomes

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right\} \\ &= \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right\}\right\} \\ &= \prod_{i=1}^{n} f_{X_i}(x_i), \end{aligned}$$

where f_{X_i} is the pdf of a $N(\mu_i, \sigma_i^2)$ random variable. Hence the components of the random vector $\boldsymbol{X}$ are independent. (Note that this is not necessarily true if the distributions are not Normal).

The Bivariate Normal Distribution

In the case $n = 2$, two continuous random variables X and Y are said to have a bivariate Normal distribution if their joint pdf is given for all x and y by

$$f_{XY}(x, y; \boldsymbol{\theta}) = \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2(1-\rho_{XY}^2)}} \exp\left\{-\frac{1}{2(1-\rho_{XY}^2)}Q(x,y)\right\},$$

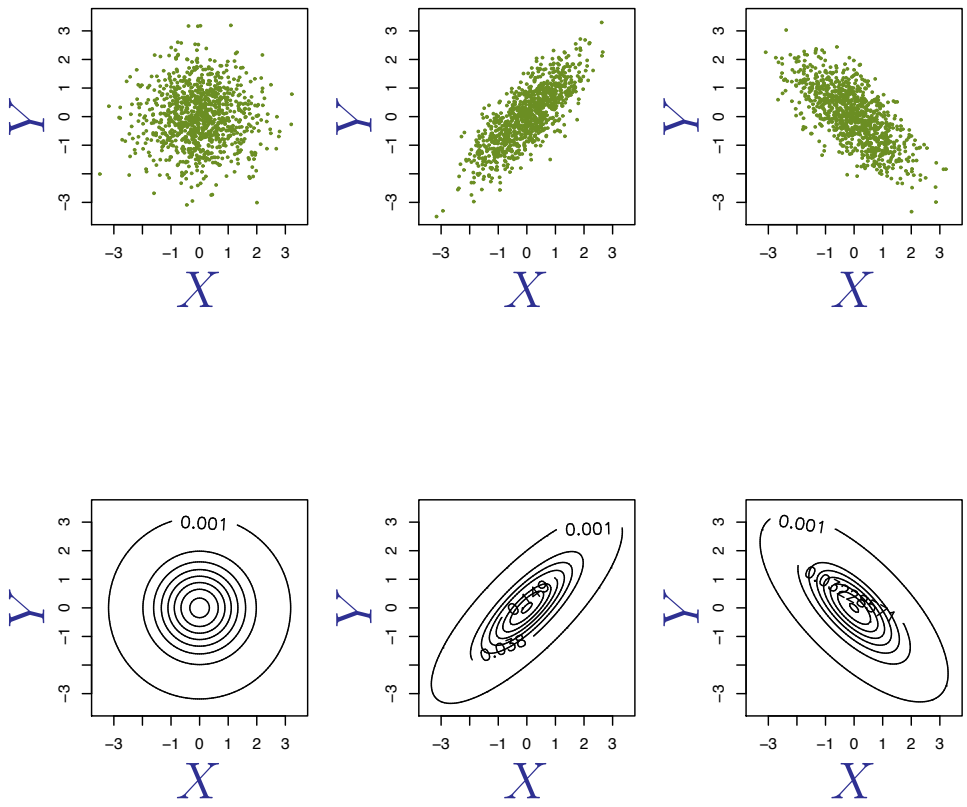
where $Q(x, y)$ is given by

$$\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{XY}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2$$

with parameters $\boldsymbol{\theta} = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY})$, where $\sigma_X^2 > 0$, $\sigma_Y^2 > 0$ and $-1 < \rho_{XY} < 1$.

The marginal distributions of X and Y are both Normal with

$$X \sim N(\mu_X, \sigma_X^2), \quad Y \sim N(\mu_Y, \sigma_Y^2) \qquad \text{and} \qquad \text{corr}(X, Y) = \rho_{XY}.$$



Top row: 1000 realisations of three different bivariate Normal distributions with $\rho_{XY} = 0$, 0.8 and -0.7 respectively. The marginal distributions are standard Normal in each case. Bottom row: Contour plots of the corresponding pdfs.

Simulation of the multivariate Normal

The bivariate Normal distribution can be represented as a transformation of two independent standard Normal random variables $S \sim \mathrm{N}(0,1)$ and $T \sim \mathrm{N}(0,1)$, as follows.

Consider the linear transformation

$$\begin{aligned} X &= \mu_X + \sigma_X S \\ Y &= \mu_Y + \rho_{XY}\sigma_Y S + \sqrt{(1-\rho_{XY}^2)}\sigma_Y T. \end{aligned}$$

In matrix notation this is

$$\begin{aligned} \begin{bmatrix} X \\ Y \end{bmatrix} &= \boldsymbol{\mu} + A \begin{bmatrix} S \\ T \end{bmatrix} \text{ where} \\ A &= \begin{bmatrix} \sigma_X & 0 \\ \rho_{XY}\sigma_Y & \sqrt{(1-\rho_{XY}^2)}\sigma_Y \end{bmatrix}. \end{aligned}$$

As linear combinations of independent Normal random variables give rise to the multivariate Normal distribution, it is sufficient to check that (X, Y) has the required mean and variance

matrix. It is easy to show that the mean is

$$\mathsf{E}\begin{bmatrix} X \\ Y \end{bmatrix} = \boldsymbol{\mu} + A\,\mathsf{E}\begin{bmatrix} S \\ T \end{bmatrix} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$

as $\mathsf{E}(S) = 0$ and $\mathsf{E}(T) = 0$.

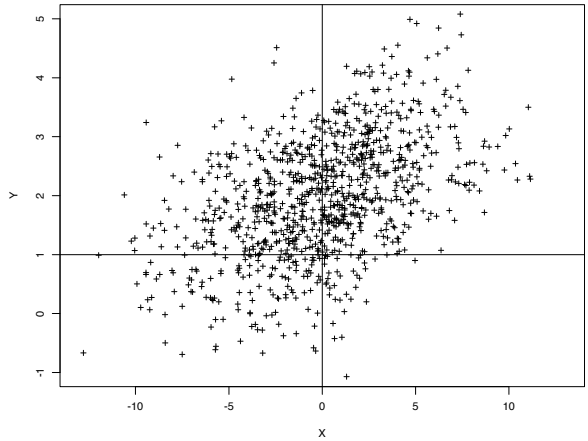
The variance is

$$\begin{aligned} \mathsf{var}\begin{bmatrix} X \\ Y \end{bmatrix} &= A\,\mathsf{var}\begin{bmatrix} S \\ T \end{bmatrix} A' \\ &= \begin{bmatrix} \sigma_X & 0 \\ \rho_{XY}\sigma_Y & \sqrt{(1-\rho_{XY}^2)}\sigma_Y \end{bmatrix} \begin{bmatrix} \sigma_X & \rho_{XY}\sigma_Y \\ 0 & \sqrt{(1-\rho_{XY}^2)}\sigma_Y \end{bmatrix} \\ &= \begin{bmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}. \end{aligned}$$

Independent standard Normal random variables can be generated from uniform random variables using the Box-Muller transformation, or directly using the `rnorm` command in R. Using the above method, it is possible to simulate any bivariate Normal distribution as a linear combination of these independent random variables. This method can be extended to generate MVN random variables in higher dimensions.

Alternatively, R has functions for generating multivariate Normal random variables and for evaluating the joint cdf and pdf, although it is necessary to download the multivariate Normal package `mvtnorm` into your R library from `http://cran.r-project.org` in order to use them. To upload this package, use the command `library(mvtnorm)` .

R can simulate m independent $MVN_n(\boldsymbol{\mu}, \Sigma)$ random variables using the command given by `rmvnorm(m, mu, sigma)` (where $\boldsymbol{\mu}$ is entered as a n-dimensional vector and Σ as a $n \times n$ dimensional matrix). R can generate the value of the pdf and cdf at $\boldsymbol{x}$ using the commands `dmvnorm(x, mu, sigma)` and `pmvnorm(x, mu, sigma)` (where x is entered as a n-dimensional vector).



1000 realisations from a bivariate Normal distributions with $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY}) = (0, 2, 16, 1, 0.5)$. The lines illustrate the events $(X \leq 0)$ and $(Y \leq 1)$.

Example: The following list of R commands generates and plots (in 1000 realisations from a bivariate Normal distribution with parameters

$$(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY}) = (0, 2, 16, 1, 0.5).$$

It also calculates the probability $\mathsf{P}(X \leq 0, Y \leq 1)$ and for comparison the product of the two marginal probabilities $\mathsf{P}(X \leq 0)\,\mathsf{P}(Y \leq 1)$. These probabilities are also approximated from the random sample using Monte-Carlo techniques (discussed in detail in the next chapter).

```
library(mvtnorm)              # loads up various multivariate functions
varx=16; vary=1               # defines the variances
covxy=sqrt(16)*sqrt(1)*0.5    # defines the covariance
sig=matrix(c(varx,covxy,covxy,vary),2)
                              # sets up the variance matrix
xy=rmvnorm(1000,mean=c(0,2),sigma=sig)
                              # R needs the full variance matrix, here
plot(xy,xlab='X',ylab='Y')    # Plot the realisations
abline(v=0)                   # Add vertical line
abline(h=1)                   # Add horizontal line
pmvnorm(upper=c(0,1),mean=c(0,2),sigma=sig)
                              # for P(X<0,Y<1)
pnorm(0,0,4)*pnorm(1,2,1)     # for P(X<0)P(Y<1)
mean((xy[,1]<0)*(xy[,2]<1))   # MC approximation to P(X<0,Y<1)
mean(xy[,1]<0)*mean(xy[,2]<1) # MC approximation to P(X<0)P(Y<1)
```

Note that the joint probability $\mathsf{P}(X \leq 0, Y \leq 1)$ is larger than the product of the marginal probabilities because of the positive correlation between X and Y which increases the chance of small values of X and Y occurring together.

Conditional multivariate Normal distributions

The linear transformation described above can be used to find the conditional distribution of $Y|X$, where $(X, Y) \sim MVN_2(\boldsymbol{\mu}, \Sigma)$. Since $X = \mu_X + \sigma_X S$, conditioning on $X = x$ corresponds to the condition $S = (x - \mu_X)/\sigma_X$ and inserting this into the expression for Y gives

$$Y = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X) + \sqrt{(1 - \rho_{XY}^2)\sigma_Y^2}\,T.$$

This shows that

$$Y \mid X = x \sim N\left(\mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1 - \rho_{XY}^2)\right).$$

Notice that the conditional expectation is linear in x, while the conditional variance is constant. Also note how the conditional variance depends on ρ_{XY}. For ρ_{XY} close to ± 1 the term $1 - \rho_{XY}^2$ is close to 0 and the conditional variance thus small. This is saying that when X and Y are very correlated knowing X gives us a lot of information about Y.

Representing Y as a linear transformation of X plus an independent Normal random variable ($[(1-\rho_{XY}^2)\sigma_Y^2]^{1/2}T$ above) is called regressing Y on X and is the idea behind the regression models used in Statistics.

The conditional distribution of X given $Y = y$ can be found analogously to be

$$X \mid Y = y \sim N\left(\mu_X + \rho_{XY}\frac{\sigma_X}{\sigma_Y}(y-\mu_Y), \sigma_X^2(1-\rho_{XY}^2)\right).$$

7.4 Copulas

Often, when dealing with multivariate distributions, it is useful to be able to describe the dependence structure between the random variables, in a way that does not depend on the marginal distributions. One way to do this is to transform the random variables in such a way that each transformed marginal variable has a standard uniform distribution. The dependence structure can then be expressed as a multivariate distribution on the obtained uniforms. In this section we only describe the bivariate case, but the ideas can be extended to the general case.

Properties of copulas: A copula is a joint cdf with the property that every marginal distribution is uniform on the interval $[0,1]$. More precisely, it is a function $C : [0,1]^2 \to [0,1]$ satisfying

- $C(x,y)$ is non-decreasing in both x and y;
- $C(0,y) = 0$ and $C(x,0) = 0$;
- $C(1,y) = y$ and $C(x,1) = x$.

Calculating copulas: The dependence between two continuous random variables can be expressed as a copula using the following method. Consider two random variables X and Y, with continuous cumulative distribution functions F_X and F_Y. Let $S = F_X(X)$ and $T = F_Y(Y)$. By the probability integral transform theorem, S and T both have standard uniform distributions but are, in general, dependent if X and Y were already dependent.

The copula expressing the dependence between X and Y is then given by

$$\begin{aligned} C(s,t) &= \mathsf{P}(S \le s, T \le t) \\ &= \mathsf{P}(F_X(X) \le s, F_Y(Y) \le t) \\ &= \mathsf{P}(X \le F_X^{-1}(s), Y \le F_Y^{-1}(t)) \\ &= F_{XY}(F_X^{-1}(s), F_Y^{-1}(t)). \end{aligned}$$

The joint cdf F_{XY} can be recovered from the copula by

$$F_{XY}(x,y) = C(F_X(x), F_Y(y)).$$

This provides a way of generating a joint distribution, given the marginal distributions and the dependence structure.

In particular, the copula can be used to calculate various properties of the joint distribution. For example

$$\mathsf{E}[g(X,Y)] = \int_0^1 \int_0^1 g(F_X^{-1}(s), F_Y^{-1}(t)) \frac{\partial}{\partial s}\frac{\partial}{\partial t} C(s,t) ds dt.$$

Simulation using copulas

Suppose X, Y are random variables with joint cdf F_{XY}. A realisation of X, Y can be obtained by simulating from the copula as follows.

- Simulate $S \sim \text{Uniform}(0,1)$.
- For realisation $S = s$, simulate (using, for example, PIT) T from the conditional cdf

$$F_{T|S}(t|s) = \frac{\partial}{\partial s} C(s,t),$$

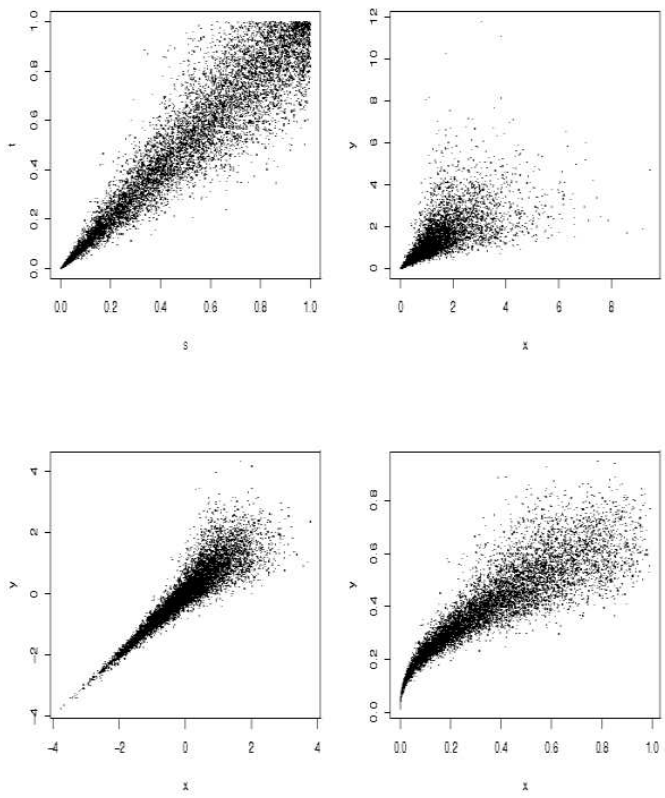
  where $C(s,t) = F_{XY}(F_X^{-1}(s), F_Y^{-1}(t))$ is the copula.
- Set $X = F_X^{-1}(S)$ and $Y = F_Y^{-1}(T)$.

Example: The Clayton copula is given by

$$C(s,t) = (s^{-\alpha} + t^{-\alpha} - 1)^{-1/\alpha},$$

where $\alpha \geq 0$. It can be used to generate joint distributions that exhibit greater dependence in the negative tail than in the positive.

In the figure below, the Clayton copula is shown with $\alpha = 8$ (top left). Joint random variables X, Y are shown with this copula but with marginal distributions $X, Y \sim \text{Exp}(1)$ (top right), $X, Y \sim N(0,1)$ (bottom left) and $X \sim \text{Beta}(1,2)$, $Y \sim \text{Beta}(3,5)$ (bottom right).



Four different joint distributions of X and Y. The copula is the same in all cases.

7.5 Links Between Standard Distributions

Transformations of multivariate random variables generalise the ones studied above. However in this section we list some of the links between standard distributions obtained by multivariate transformations of independent random variables. Many of these results are important in statistics.

Random variables that are related by transformation. All the variables in the left hand column are assumed independent.

Distribution	Transformation	Distribution of Y
$Z_i \sim N(0,1)$, $i = 1, \ldots, n$	$Y = Z_1^2 + \ldots + Z_n^2$	$\text{Gamma}(n/2, 1/2) = \chi^2_n$
$X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$	$Y = X_1 + \ldots + X_n$	$\text{N}(\mu_1 + \ldots + \mu_n, \sigma_1^2 + \ldots + \sigma_n^2)$
$X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$	$Y = \boldsymbol{a}'\boldsymbol{X}$	$\text{N}(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}' \text{ diag } (\Sigma)\boldsymbol{a})$
$X_i \sim \text{Exp}(\beta)$, $i = 1, \ldots, n$	$Y = X_1 + \ldots + X_n$	$\text{Gamma}(n, \beta)$
$X_i \sim \text{Gamma}(\alpha_i, \beta)$, $i = 1, \ldots, n$	$Y = X_1 + \ldots + X_n$	$\text{Gamma}(\alpha_1 + \ldots + \alpha_n, \beta)$
$X_i \sim \text{Gamma}(\alpha_i, \beta)$, $i = 1, 2$	$Y = \frac{X_1}{X_1 + X_2}$	$\text{Beta}(\alpha_1, \alpha_2)$

Chapter 8

Limit Theorems

In this Chapter we study two important limit results, the Law of Large Numbers and the Central Limit Theorem, both of which tell us about the behaviour of the mean of n independent identically distributed (iid) random variables as n gets larger and larger. While these results are mathematically interesting in their own right they give theoretical justification for the whole practise of statistics.

The results provide a theoretical justifications

to assert large samples are good, and provide criterion to say how large is large;

to motivate the Normal distribution as a probability model;

for using Monte Carlo methods to construct approximations to unknown mathematical integrals.

8.1 The Law of Large Numbers

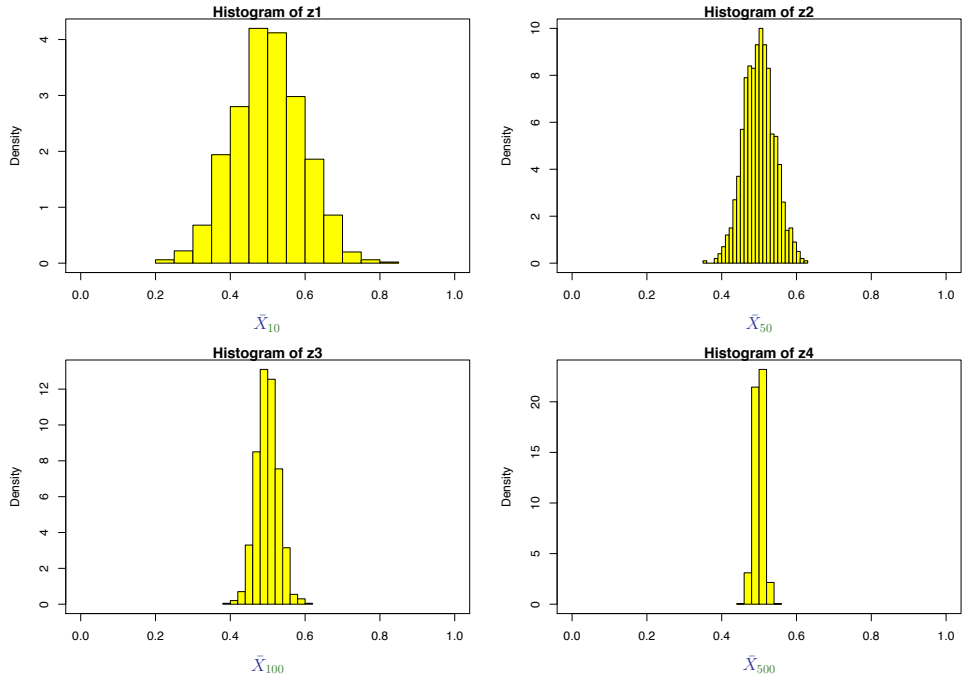
Assume that $X_1, X_2, \ldots$ are iid random variables with common mean μ and common variance σ^2, and let $\bar{X}_n$ denote the mean of the first n of these,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The expected value and variance of $\bar{X}_n$ are

$$\mathsf{E}(\bar{X}_n) = \mu, \quad \mathsf{var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

The figure illustrates the distribution of $\bar{X}_n$ for varying n, when each X_i has a $\mathrm{Uniform}(0,1)$ distribution. The plots are histograms of 1000 realisations of $\bar{X}_n$ for $n = 10$, 50, 100 and 500, i.e. to make the plot in the lower left-hand corner we have taken the mean of 100 $\mathrm{Uniform}(0,1)$-distributed random variables 1000 times.



Histograms of 1000 realisations of $\bar{X}_n$ for $n = 10$, 50, 100 and 500 when $X_i \sim \text{Uniform}(0, 1)$.

It is clear to see from the figure how the distribution of $\bar{X}_n$ concentrates more and more around $\mu = 0.5$ as n gets larger reflecting the fact that the variance decreases to 0 as n increases. This is the subject of the Law of Large Numbers to be shown below.

We first need a bound on the probability that a random variable deviates more than ϵ from its mean in terms of its variance. We will need the indicator function $I_{(|Y-\mu|>\epsilon)}$, i.e. the function

$$I_{(|Y-\mu|>\epsilon)} = \begin{cases} 1 & \text{if } |Y-\mu| > \epsilon, \\ 0 & \text{if } |Y-\mu| \leq \epsilon. \end{cases}$$

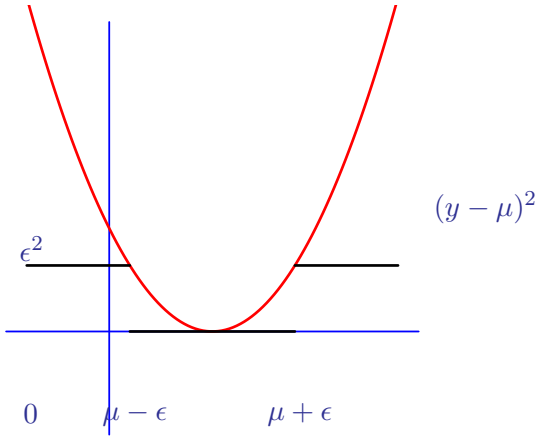
Theorem 8.1 Chebyshev's Inequality. *If Y is a random variable with mean μ and finite variance then for any $\epsilon > 0$*

$$\mathsf{P}(|Y-\mu| > \epsilon) \leq \frac{1}{\epsilon^2}\,\mathsf{var}(Y).$$

Proof: First note that

$$\epsilon^2 I_{(|Y-\mu|>\epsilon)} \leq (Y-\mu)^2,$$

since when the left-hand side is ϵ^2 then $|Y-\mu| > \epsilon$ and therefore $(Y-\mu)^2 \geq \epsilon^2$. See the picture.



Taking expectations of the LHS and the RHS gives

$$\epsilon^2 \,\mathsf{P}(|Y-\mu|>\epsilon) \;\leq\; \mathsf{E}[(Y-\mu)^2] = \mathsf{var}(Y),$$

as $\mathsf{E}[I_{(Y\in A)}] = \mathsf{P}(Y\in A)$. □

With this at hand we can easily show:

Theorem 8.2 The Weak Law of Large Numbers. *Suppose $X_1, X_2, \ldots$ is a sequence of iid random variables with mean μ and finite variance σ^2. Then, for any $\epsilon > 0$,*

$$\mathsf{P}(|\bar{X}_n - \mu| > \epsilon) \to 0, \qquad \text{as} \quad n\to\infty.$$

Proof: Using Chebyshev's inequality we have for any $\epsilon > 0$

$$\begin{aligned} \mathsf{P}(|\bar{X}_n - \mu| > \epsilon) &\leq \frac{1}{\epsilon^2}\,\mathsf{var}(\bar{X}_n) \\ &= \frac{1}{\epsilon^2}\frac{\sigma^2}{n} \\ &\to 0, \qquad \text{as} \quad n\to\infty. \end{aligned}$$

□

We say that $\bar{X}_n$ converges in probability to μ.

The theorem is a mathematical restatement of the figure above. The distribution of $\bar{X}_n$ concentrates more and more around μ, in the sense that no matter how small an interval $[\mu-\epsilon, \mu+\epsilon]$ we take around μ the probability of $\bar{X}_n$ falling in this interval tends to 1.

Frequencies converge to probabilities: The WLLN as stated concerns $\bar{X}_n$. It easily extends to frequencies. Consider the random variables $Y_i = I_{(X_i\in A)}$, where A is an event and I is its

indicator function. Application of the WLLN gives

$$\begin{aligned} \frac{\text{number of times } A \text{ occurs in } n \text{ trials}}{n} &= \frac{1}{n}\sum_{i=1}^{n} I_{(X_i \in A)} \\ &\to \mathsf{P}(X \in A), \qquad \text{as} \quad n \to \infty. \end{aligned}$$

The frequency of the event A occurring converges to the probability of A for all realisations $x_1, x_2, \ldots$ of the sequence $X_1, X_2, \ldots$. For instance, the proportion of times heads occur in n throws of a coin will converge to the probability of a head. This statement recovers our intuitive grasp of probability.

8.2 The Central Limit Theorem

The Central Limit Theorem is one of the most important results in probability theory and statistics and is the reason the Normal distribution plays such a prominent role. It asserts that the sum (or the mean) of many independent identically distributed random variables is approximately Normally distributed. The remarkable fact is true, whatever the common distribution of the random variables, as long as it has finite mean and variance.

Theorem 8.3 The Central Limit Theorem. *Suppose $X_1, X_2, \ldots$ is a sequence of iid random variables with mean μ and finite variance σ^2. Then for any number $-\infty < x < \infty$*

$$\mathsf{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) \to \Phi(x), \qquad \text{as} \quad n \to \infty.$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\Phi(x)$ is the cumulative distribution function for the standard Normal distribution $N(0,1)$ evaluated at x.

Whereas the WLLN only tells us that $\bar{X}_n$ converges to μ the CLT gives us the stronger information that the deviations of $\bar{X}_n$ from μ scaled by $\sqrt{n}$ follow a $N(0, \sigma^2)$ distribution in the limit. The practical use of this is that for reasonably large n we can assume that

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

approximately.

The proof of the Central Limit Theorem is not examinable, but we give a sketch below.

Proof: (Sketch) Suppose $X_1, X_2, \ldots$ is a sequence of iid random variables with common mean 0 and common variance 1 (to make it easy). Let S_n denote the standardised sum of the first n of these

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i, \quad \text{with } \mathsf{E}(S_n) = 0, \ \ \mathsf{var}(S_n) = 1.$$

Recall that the distribution of a random variable X is uniquely determined by it's mgf. Now

$$\begin{aligned} M_{S_n}(t) &= (M_{X_1}(t/\sqrt{n}))^n \\ &= \left(1 + \frac{t^2}{2n} + o(t^2/n)\right)^n \\ &\to \exp(t^2/2) \end{aligned}$$

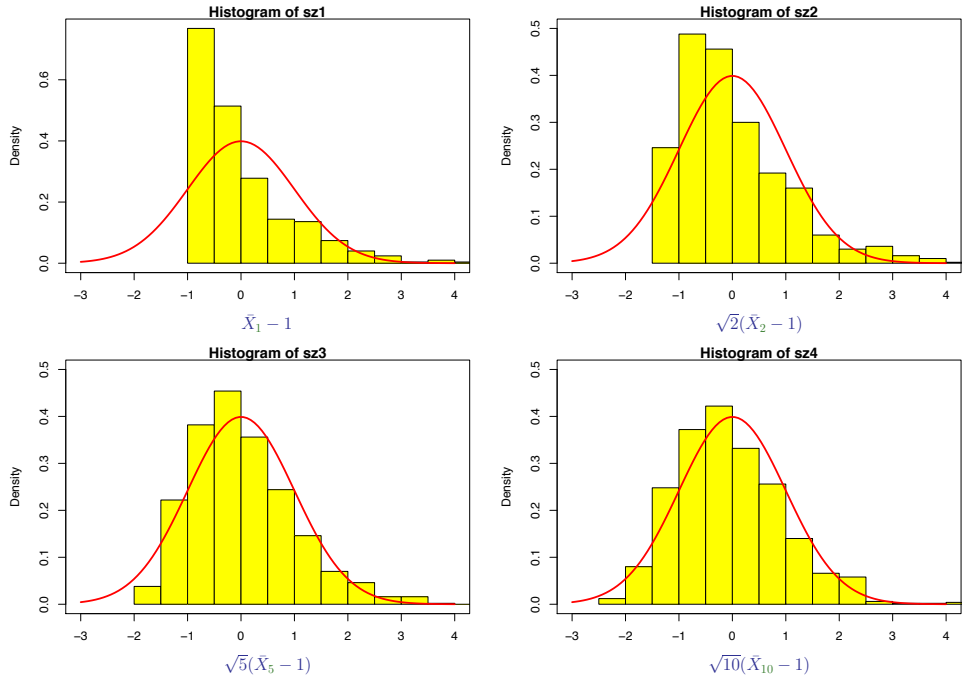
as $n \to \infty$.

But if $X \sim N(0,1)$, then

$$M_X(t) \quad = \quad \exp(t^2/2)$$

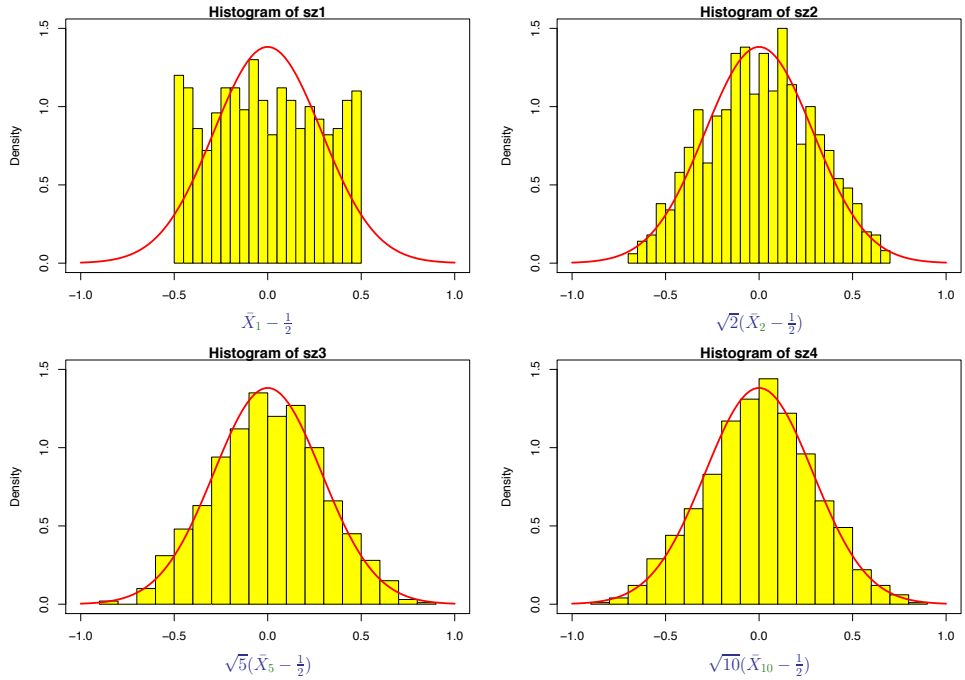
and hence if a limit distribution of the sum of iid random variables with finite variance exists, it has to be Normal. □

How large does n have to be for the Normal approximation to be valid? This depends on how close the original distribution of the X's is to Normal in the first place: the closer it is the quicker the approximation becomes accurate. For instance, if each X_i itself is Normal, then $\bar{X}_n$ is exactly Normal for every n. Almost always $n > 30$ will be enough to justify the approximation - sometimes much smaller n will do.



Histograms of 1000 realisations of $\sqrt{n}(\bar{X}_n - 1)$ for $n = 1$, 2, 5 and 10 when $X_i \sim \text{Exp}(1)$. The pdf of a $N(0,1)$ distribution is superimposed on each histogram.

The figures illustrate the distribution of $\sqrt{n}(\bar{X}_n - \mu)$ for $n = 1$, 2, 5 and 10 when the X's are exponential and uniform. The pdf for the approximating Normal distribution is superimposed on each of the histograms. Note the very fast convergence to a Normal in the uniform case and the somewhat slower convergence in the exponential case. In both cases however, the Normal approximation is very good for $n \geq 10$, say.



Histograms of 1000 realisations of $\sqrt{n}(\bar{X}_n - \frac{1}{2})$ for $n = 1$, 2, 5 and 10 when $X_i \sim \text{Uniform}(0, 1)$. The pdf of a $N(0, \frac{1}{12})$ distribution is superimposed on each histogram.

Example: Suppose an investor buys a commodity for $£1$ and wishes to sell it after some fixed time T. We are interested in the distribution of the price $£S_T$ that the commodity is worth at this time. The amount by which the price of a commodity rises or falls over a long time period is determined by random fluctuations in the price over small time periods. Consider splitting the time period T into n equal time periods. Let X_i denote the factor by which the price of the commodity changes from time $(i-1)T/n$ to time iT/n i.e. if an investor buys $£1$ of the commodity at time $(i-1)T/n$, then at time iT/n it is worth $£X_i$. If the time periods are sufficiently small, and the market is stable, we may assume that the X_i are iid random variables. Now $S_T = X_1 X_2 \dots X_n$ and so

$$\log S_T = \log X_1 + \cdots + \log X_n.$$

Since $\log X_1, \dots, \log X_n$ are iid, by the central limit theorem, the distribution of $\log S_T$ is approximately Normal and therefore S_T has a Log Normal distribution.

8.3 Monte Carlo Evaluation

Given a sequence of iid realisations $x_1, \dots, x_n$ of a random variable X we can approximate various properties of the distribution of X. This is because of the the following properties of long sequences.

• Limiting frequencies: the probability of an event is the long run proportion of times this event occurs in independent experiments, i.e. for any event A

$$\lim_{n\to\infty} \frac{\text{the number of times } x_i \in A}{n} = \mathrm{P}(X \in A).$$

• Limiting averages: the expectation is the long run average of independent replicates from an experiment, i.e.

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}x_i &= \mathsf{E}(X)\\ \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}g(x_i) &= \mathsf{E}[g(X)].\end{aligned}$$

The theoretical result behind this is the Law of Large Numbers. We can use these results to obtain the following approximations to the properties of the distribution of X.

pdf: As the pdf $f_X(x)$ is the probability that X belongs to a small interval around x divided by the length of the interval

$$f_X(x) \approx \mathsf{P}(x < X \leq x+\delta)/\delta$$

we can estimate the pdf by

$$f_X(x) \approx \frac{\text{the number of times } x_i \in [x, x+\delta)}{n\delta}.$$

This is the histogram of the simulated data $x_1, \ldots, x_n$.

cdf: As the cdf $F_X(x)$ is the probability that $\mathsf{P}(X \leq x)$ we can estimate it by

$$F_X(x) \approx \frac{\text{the number of times } x_i \leq x}{n}.$$

Probabilities of events: The probability of a general event A is approximately

$$\mathsf{P}(X \in A) \approx \frac{\text{the number of times } x_i \in A}{n}.$$

Expectations: We can estimate the mean of X and $g(X)$ by

$$\begin{aligned}\mathsf{E}(X) &\approx \frac{1}{n}\sum_{i=1}^{n}x_i,\\ \mathsf{E}[g(X)] &\approx \frac{1}{n}\sum_{i=1}^{n}g(x_i).\end{aligned}$$

Transformations: We can estimate properties of Y by obtaining a sample from the transformed variable $Y = g(X)$ as $y_1 = g(x_1), \ldots, y_n = g(x_n)$.

Confidence Intervals

Approximations of this type are called Monte Carlo approximations because of the randomness involved. If we ran the R functions again we would get slightly different results. The precision of the approximation also depends on how large n is. The larger n the higher the precision.

By the Central Limit Theorem

$$\mathsf{P}\left(-x \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) \to \Phi(x) - \Phi(-x) = 2\Phi(x) - 1,$$

as $n \to \infty$, so

$$\mathsf{P}\left(\mu \in \left[\bar{X}_n - x_p\sqrt{\frac{\sigma^2}{n}}, \bar{X}_n + x_p\sqrt{\frac{\sigma^2}{n}}\right]\right) \approx 2p - 1,$$

where x_p is the pth quantile of the $N(0,1)$ distribution.

A way of assessing the accuracy of the approximation, using tools we have already, is therefore to repeat the simulation a number of times to obtain m, say, different estimates $z_1, \ldots, z_m$ and compute a confidence interval of the required confidence by

$$\left[\mathsf{Mean}(z) - x_p\sqrt{\frac{\mathsf{Var}(z)}{m}},\ \mathsf{Mean}(z) + x_p\sqrt{\frac{\mathsf{Var}(z)}{m}}\right],$$

where $\mathsf{Mean}(z)$ and $\mathsf{Var}(z)$ are the mean and variance of the sample $z_1, \ldots, z_m$. Usually we take $x_p = 1.96$, or $p = 0.975$. In this case we can be 95% sure that the true value lies within the interval. Thus taking m as large as possible is sensible as it gives the best possible precision in the Monte Carlo estimate.

Example: For $X \sim N(0,1)$ estimate the expectation $\mathsf{E}[\cos(X)]$, and give a 95% confidence interval.

```
z = rep(0,1000)            # 1000 zeros
for (i in 1:1000){         # each estimate based on 10000
    x = rnorm(10000,0,1)   # replicates of a N(0,1) rv
    z[i] = mean(cos(x))
}
mean(z)                     # mean is best estimate
var(z)                      # variance in the sample
sd = sqrt(var(z)/1000)
mean(z)-qnorm(0.975,0,1)*sd
mean(z)+qnorm(0.975,0,1)*sd
```

Thus we get $\mathsf{E}[\cos(X)] \approx 0.6067$, with 95% confidence interval $= [0.6064, 0.6070]$. In fact it can be shown that

$$\begin{aligned}\mathsf{E}[\cos(X)] &= \int_{-\infty}^{\infty} \cos(x)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx \\ &= \exp(-1/2) = 0.60653.\end{aligned}$$

Chapter 9

Poisson Processes

Often, we are not just interested in the outcome of a single experiment, but in a sequence of outcomes that evolve in some random way. The theory of stochastic processes provides us with a way to describe such situations.

9.1 Stochastic Processes

A Stochastic Process X is a collection of random variables

$$\{X(t) \ : \ t \in T\}$$

where T is some index set.

Usually $T = \mathbb{N}$, in which case we say that the process is a discrete time stochastic process, or $T = [0, \infty)$ in which case we call it a continuous time stochastic process. We often use the notation X_t instead of $X(t)$.

The State Space of a stochastic process is the set of possible values of $X(t)$.

The distribution of a stochastic process is determined by the joint distributions of all finite subsets of

$$\{X(t) \ : \ t \in T\}.$$

Example: A Bernoulli process is a discrete time stochastic process where X_t, $t = 1, 2, \ldots$ are independent identically distributed Bernoulli random variables. It represents performing a sequence of independent experiments and recording $X_t = 1$ if the tth trial is a success and $X_t = 0$ if it is a failure.

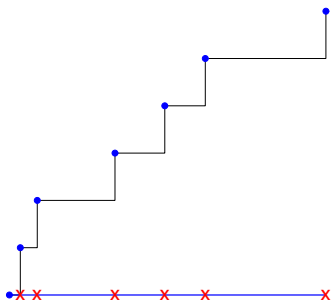
9.2 Counting Processes

A counting process is a stochastic process $\{N(t), t \geq 0\}$ that satisfies the following properties.

- $N(t) \in \mathbb{N}$.

- If $s \leq t$ then $N(s) \leq N(t)$

The process can be thought of as counting events as they occur in time. If $s < t$, then $N(t) - N(s)$ is the number of events that occurred during the interval $(s, t]$.



9.3 Poisson Processes

A common example of a counting process is the Poisson process. This describes processes in which events occur with a constant average rate irrespective of time or previous events.

The Poisson process provides the mathematical justification for statistical modelling using the Poisson distribution, the Exponential distribution, and the Gamma distribution. This also provides understanding of the inter-relationships between these distribution and some background on the memoryless property of the Exponential random variable.

Mathematical Model

Let $N(t)$ denote the number of events in $(0, t]$. This could also equally be the number of events in the interval $(t_1, t_1 + t]$ due to time homogeneity of the process.

Suppose events occur at random, independently, and with a fixed rate λ per unit time. We describe this, to first order, for small $\delta > 0$

$$\mathsf{P}\{N(t+\delta) = j | N(t) = i\} = \begin{cases} 0 & \text{if} \quad j < i, \\ 1 - \lambda\delta & \text{if} \quad j = i, \\ \lambda\delta & \text{if} \quad j = i+1, \\ 0 & \text{if} \quad j > i+1. \end{cases}$$

We let

$$p_i(t) = \mathsf{P}[N(t) = i].$$

So by the law of total probability

$$\mathsf{P}[N(t+\delta) = i+1]$$

$$
\begin{aligned}
&= \mathsf{P}[N(t+\delta) = i+1 | N(t) = i+1]\,\mathsf{P}[N(t) = i+1] \\
&+ \mathsf{P}[N(t+\delta) = i+1 | N(t) = i]\,\mathsf{P}[N(t) = i] + 0
\end{aligned}
$$

or

$$
p_{i+1}(t+\delta) \quad = \quad (1-\lambda\delta)p_{i+1}(t) + \lambda\delta p_i(t).
$$

Rearranging the above gives

$$
\frac{p_{i+1}(t+\delta) - p_{i+1}(t)}{\delta} = -\lambda p_{i+1}(t) + \lambda p_i(t).
$$

Taking the limit as $\delta \to 0$

$$
\frac{dp_{i+1}(t)}{dt} = -\lambda p_{i+1}(t) + \lambda p_i(t) \qquad \text{for all} \quad t, i.
$$

The solution of this equation is uniquely given by

$$
p_i(t) = \frac{(\lambda t)^i \exp(-\lambda t)}{i!}, \quad i = 0, 1, \ldots.
$$

This shows that the random variable $N(t) \sim \text{Poisson}(\lambda t)$.

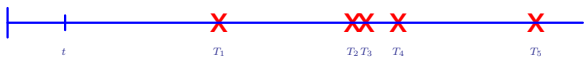
Distribution of the Number of Events

The mathematical model above says that the number of events of a Poisson process with rate λ in a fixed interval of time $(0, t]$ follows a Poisson distribution with expectation λt. That is $N(t) \sim \text{Poisson}(\lambda t)$.

The result holds whatever the value of t or whatever time interval. So as $N(t_2) - N(t_1)$ is the number of events that fall in the interval $(t_1, t_2]$,

$$
N(t_2) - N(t_1) \sim \text{Poisson}(\lambda[t_2 - t_1]).
$$

Waiting Time Distributions



Let the random variable T_k denote the time to the k-th event of the Poisson process from time 0. First we derive the distribution of T_1.

Note the equivalence of the two events

$$\{T_1 > t\} = \{N(t) = 0\},$$

for all $t > 0$, so that

$$\begin{aligned} \mathsf{P}(T_1 > t) &= \mathsf{P}[N(t) = 0] \\ &= \exp(-\lambda t) \qquad \text{for} \quad t > 0, \end{aligned}$$

using the property that $N(t)$ is a Poisson random variable with expectation λt for all $t > 0$. Thus $T_1 \sim \text{Exp}(\lambda)$.

As the Poisson process is homogeneous in time and independent of events in the past, the distribution of the time until the first event after time 0 is the same whether or not an event of the process occurred at time 0. Thus the Exponential distribution is also the distribution of waiting times between events of a Poisson process. This link to events occurring at random explains the memoryless property of the Exponential distribution. This also provides a way to simulate the Poisson process.

Now consider $k > 1$. Note the equivalence of the two events

$$\{T_k > t\} = \{N(t) \leq k - 1\},$$

for all $t > 0$. Hence

$$\begin{aligned} \mathsf{P}(T_k > t) &= \mathsf{P}[N(t) \leq k-1] \\ &= \sum_{r=0}^{k-1} \mathsf{P}[N(t) = r] \\ &= \sum_{r=0}^{k-1} \frac{(\lambda t)^r \exp(-\lambda t)}{r!} \qquad \text{for} \quad t > 0, \end{aligned}$$

using the property that $N(t)$ is a Poisson random variable with expectation λt for all $t > 0$. Differentiating with respect to t the right hand side has much cancellation giving

$$f_{T_k}(t) = \frac{\lambda(\lambda t)^{k-1} \exp(-\lambda t)}{(k-1)!} \qquad \text{for} \quad t > 0,$$

thus $T_k \sim \text{Gamma}(k, \lambda)$.

Note that we can break T_k down into

$$T_k = T_1 + (T_2 - T_1) + (T_3 - T_2) + \ldots + (T_k - T_{k-1}),$$

the time to the first event and the time between consecutive events. With this decomposition, the variables being summed are independent and identically distributed $\text{Exp}(\lambda)$ random variables. This characterises the Gamma random variable $\text{Gamma}(\alpha, \beta)$, when $\alpha \in \mathbb{N}$, as the sum of α iid Exponential variables.

Exercise 9.1 The occurrence of earthquakes in California can be modelled as a Poisson process with rate $\lambda = 8.35$ per year.

(a) Find the probability that the number of earthquakes in a year exceeds 12.

(b) If an earthquake has just happened what is the probability that no earthquake occurs in the next month?

Sol:

Let $N(t)$ be the number of earthquakes in t years, so that

$$N(t) \sim \text{Poisson}(\lambda t) \quad \text{and} \quad N(1) \sim \text{Poisson}(8.35 \times 1).$$

(a) Hence

$$\begin{aligned} \mathsf{P}(N(1) > 12) &= \sum_{r=13}^{\infty} \exp(-8.35) 8.35^r / r! \\ &= \texttt{1-ppois(12,8.35)} \\ &= 0.0821336. \end{aligned}$$

(b) The lack of memory property implies

$$\begin{aligned} &\mathsf{P}(\text{ No earthquake next month} | \text{earthquake now}) \\ &= \mathsf{P}(\text{ No earthquake next month}). \end{aligned}$$

No earthquakes next month corresponds to the event $\{N(\frac{1}{12}) = 0\}$. Hence

$$\begin{aligned} \mathsf{P}\left(N\left(\frac{1}{12}\right) = 0\right) &= \frac{(\frac{8.35}{12})^0 \exp(-\frac{8.35}{12})}{0!} \\ &= \exp\left(-\frac{8.35}{12}\right) \\ &= 0.4986. \end{aligned}$$

Alternatively use the distribution of waiting time, $T \sim \text{Exp}(8.35)$, so that to wait more that 1 month is $\mathsf{P}(T > \frac{1}{12}) = \exp(-8.35/12)$. □

Exercise 9.2 Data on major eruptions in the northern hemisphere between 1851 and 1985 suggest that it may be reasonable to model these events by a Poisson process, with mean number of eruptions per year being 2.39.

(a) Find the probability that in any fixed 2-year period there are exactly 2 eruptions.

(b) Find the probability of the waiting time between successive eruptions being at least t years.

(c) Deduce the median time (in years) between successive eruptions.

Sol:

Let $N(t)$ be the number of eruptions in time t years. Then $N(t) \sim \text{Poisson}(2.39t)$.

(a)

$$\mathsf{P}\{N(2) = 2\} = \frac{(2 \times 2.39)^2 \exp(-2 \times 2.39)}{2!} = 0.0959.$$

(b) From the Exponential distribution for waiting times of a Poisson process

$$\mathsf{P}(\text{waiting time} > t) = \exp(-2.39t).$$

(c) Waiting times are distributed as $\mathrm{Exp}(2.39)$ so the median m satisfies

$$\mathsf{P}(\text{waiting time} > m) = \exp(-2.39m) = \frac{1}{2}$$

so $m = \frac{\log 2}{2.39} = 0.29$.

□

Chapter 10

Markov Chains

Many stochastic process that arise naturally have the property that only the current state of the process influences where it goes next, and it retains no memory of the past. This property is known as the Markov property and these processes are known as Markov processes. In this chapter we shall consider Markov processes that have a finite or countable state space. These processes are called Markov chains.

10.1 Discrete Time Markov Chains

A discrete time Markov chain (MC) is a stochastic process $(X_t : t \in \{0, 1, \ldots\})$ for which

$$\mathsf{P}(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots, X_0 = x_0) = \mathsf{P}(X_t = x_t | X_{t-1} = x_{t-1})$$

whatever the values of x_{t-1}, x_{t-2}, $\ldots$, x_0.

Note that, although the index set $T = \{0, 1, \ldots\}$ is referred to as time, the subscripts t are just labels of successive experiments and the steps of the process are not necessarily equally spaced in real time.

The transition probability matrix P_t at time t of a MC is defined by its elements to be:

$$(P_t)_{i,j} = \mathsf{P}(X_{t+1} = j | X_t = i),$$

the probabilities of transition to state i from state j at time t. Observe that row i of P_t is the conditional pdf of X_{t+1} given $X_t = i$, so sums to 1.

A MC is homogeneous (in time) if $P_t = P$ does not depend on time t. We shall henceforth assume that all Markov chains we consider are homogeneous unless otherwise stated.

If $(X_t : t \in \{0, 1, \ldots\})$ is a Markov chain with initial distribution $\pi(0)$ i.e. $\mathsf{P}(X_0 = i) = \pi(0)_i$ for all i in the state space, and transition probability matrix P, we say that X_t is Markov$(\pi(0), P)$.

Example: The Gambler's Ruin problem: A fair coin is tossed repeatedly and at each toss a gambler wins £1 if a head appears and loses £1 otherwise. He starts with £x and continues playing until his capital reaches £m or he goes broke. Let X_t denote the amount of money that the gambler has after t tosses. Then X_t is a homogeneous discrete Markov chain with state space

$\{0, 1, \ldots, m\}$, $\pi(0)_x = 1$, and transition probability matrix given by

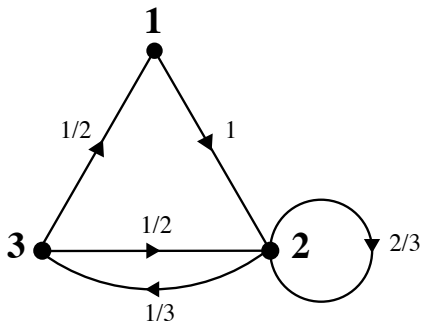
$$P_{ij} = \begin{cases} 1 & \text{if} \quad i = j = 0 \quad \text{or} \quad i = j = m, \\ \frac{1}{2} & \text{if} \quad 0 \leq j = i - 1 < m \quad \text{or} \quad 0 < j = i + 1 \leq m, \\ 0 & \text{otherwise.} \end{cases}$$

Analysing Markov chains

To aid the understanding of the dynamics of a Markov chain it is useful to draw a diagram which indicates all the possible states of the chain, with arrows denoting the possible transitions (i.e. an arrow joins state i to state j if $P_{ij} > 0$).

Example: The diagram below represents the Markov chain with transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$



A state i is persistent if

$$\mathrm{P}(X_t = i, \text{for some } t \geq 1 | X_0 = i) = 1.$$

Otherwise a state is transient. This definition states that a state, i, is persistent if the chain, started from state i, will always (eventually) return to that state, and transient if it does not necessarily return.

A set of states is irreducible (or intercommunicating) if for all ordered pairs of states, i and j, in that set

$$\mathrm{P}(X_t = i | X_0 = j) > 0 \text{ for some } t.$$

This is equivalent to saying that for all pairs of states i and j it is possible to get from j to i in some number of steps. Otherwise the set of states is reducible.

We say a Markov chain is irreducible if the complete state-space is irreducible. In diagrams of intercommunicating MCs, all states are connected.

In practice we will need to determine Markov chain's for which the set of all persistent states is irreducible.

The period of a state i is defined as

$$d_i = \text{greatest common divider}\{t : \mathrm{P}(X_t = i | X_0 = i) > 0\}.$$

A state, i, has period d if the Markov chain, started at i, can only return to i after a multiple of d time-steps.

All states of an irreducible Markov chain have the same period.

A Markov chain is aperiodic if all recurrent states have period 1. (So an irreducible Markov chain is aperiodic if any recurrent state has period 1.)

A periodic chain exhibits some form of deterministic behaviour. For example a Markov chain with period 2 will alternate between two sets of states, one of which it can be in at odd time points, and the other at even time points.

If an irreducible Markov chain has $P_{ii} > 0$ for some i then it is aperiodic.

Example: Consider a MC with state space $\{1, 2, 3, 4\}$ and transition matrix given by

$$P = \begin{pmatrix} 0.5 & 0.2 & 0.2 & 0.1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

The chain is reducible since

$$\mathsf{P}(X_t = 1 | X_0 = 2) = 0 \text{ for all } t.$$

State 1 has period 1; states 2,3 and 4 have period 3.

However, the set of persistent states ($\{2, 3, 4\}$) is irreducible.

Multi-step transitions

We call $P^{(m)}$ the matrix of multi-step transition probabilities where

$$\left(P^{(m)}\right)_{i,j} = \mathsf{P}(X_{t+m} = j | X_t = i).$$

Theorem 10.1

$$P^{(m)} = P^m,$$

that is the mth power of the matrix P.

Proof: For the case $m = 2$; induction can be used for the general result.

$$\left(P^{(2)}\right)_{i,j} = \mathsf{P}(X_{t+2} = j | X_t = i) = \sum_k \mathsf{P}(X_{t+2} = j, X_{t+1} = k | X_t = i).$$

Now in general, $\mathsf{P}(A, B|C) = \mathsf{P}(A|B, C)\,\mathsf{P}(B|C)$, giving

$$\sum_k \mathsf{P}(X_{t+2} = j | X_{t+1} = k, X_t = i)\,\mathsf{P}(X_{t+1} = k | X_t = i).$$

By the Markov property this is

$$\sum_k \mathsf{P}(X_{t+2} = j | X_{t+1} = k)\,\mathsf{P}(X_{t+1} = k | X_t = i) = \sum_k P_{i,k} P_{k,j} = P^2_{i,j}.$$

□

It follows from the transitivity of matrix multiplication that

$$P^{(m+n)} \;=\; P^{(m)}\, P^{(n)}$$

or

$$\mathsf{P}(X_{t+m+n} = j | X_t = i) \;=\; \sum_k \mathsf{P}(X_{t+m+n} = j | X_{t+m} = k)\, \mathsf{P}(X_{t+m} = k | X_t = i).$$

This is called the Chapman-Kolmogorov equation.

Let the row vector $\pi(t)$ hold the pmf of X_t, i.e.

$$\pi(t)_i = \mathsf{P}(X_t = i)$$

so

$$\pi(t) = (\,\mathsf{P}(X_t = 1) \quad \mathsf{P}(X_t = 2) \quad \cdots).$$

Then

$$\pi(t+1) = \pi(t)P$$

and more generally

$$\pi(t+m) = \pi(t)P^{(m)} = \pi(t)P^m.$$

In particular, suppose that X_0 has initial distribution $\pi(0)$ and that the chain is homogeneous in time, then

$$\pi(2) = \pi(1)P,\;\; \pi(3) = \pi(2)P,\;\; \ldots\;\; \pi(t+1) = \pi(t)P \;=\; \pi(1)P^t.$$

This means that $\pi(t+1)$, $\pi(t+2)$, ...can be evaluated knowing only $\pi(t)$.

Example: Let X_t have four states representing either none, one year, two years or three years of no claims bonus on an automobile insurance. Let the transition probability matrix P be

$$\begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ \frac{1}{6} & \frac{1}{6} & 0 & \frac{2}{3} \\ 0 & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{pmatrix}$$

and take $\pi(0) = (\,1,\, 0,\, 0,\, 0\,)$. Then

$$\pi(1)' = \begin{pmatrix} 0.33 \\ 0.67 \\ 0 \\ 0 \end{pmatrix} \quad \pi(2)' = \begin{pmatrix} 0.33 \\ 0.22 \\ 0.44 \\ 0 \end{pmatrix} \quad \pi(3)' = \begin{pmatrix} 0.26 \\ 0.30 \\ 0.15 \\ 0.29 \end{pmatrix} \quad \pi(10)' = \begin{pmatrix} 0.16 \\ 0.21 \\ 0.21 \\ 0.42 \end{pmatrix} \quad \pi(20)' = \begin{pmatrix} 0.16 \\ 0.21 \\ 0.21 \\ 0.42 \end{pmatrix}.$$

Discrete Markov chains are easy to simulate, using techniques developed so far in this course. Suppose X_t is Markov$(\pi(0), P)$. Simulate X_0 from the discrete distribution with pmf given by $p(r) = \pi(0)_r$. Suppose the realisation is $X_0 = i$. Then simulate X_1 from the distribution with pmf $p(r) = P_{ir}$. Continue recursively i.e. suppose you have simulated $X_0, \ldots, X_t$ and the realisation of $X_t = j$, then X_{t+1} can be simulated from the distribution with pmf $p(r) = P_{jr}$. Simulation provides an alternative means of analysing MCs to matrix multiplication, which can be especially useful in situations where matrix multiplication may be very time-consuming.

Asymptotic and invariant distributions

In the example above, the distribution $\pi(t)$ appears to be converging to a stable distribution as $t \to \infty$. A homogeneous MC has an asymptotic (or steady state) distribution π if $\pi(t) \to \pi$ whatever the initial distribution $\pi(0)$.

Example: Let X_t be a MC with state space $\{1, 2\}$ and transition matrix

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

where $0 < a, b < 1$. Let $\pi(t) = (u_t \quad 1-u_t)$ for all t. Then $\pi(t+1) = \pi(t)P$ gives

$$u_{t+1} = (1-a-b)u_t + b.$$

The general solution for this recurrence relation is

$$u_t = \frac{b}{a+b} + (1-a-b)^t \left(u_0 - \frac{b}{a+b} \right).$$

Since $0 < a + b < 2$, the term $(1-a-b)^t \to 0$ as $t \to \infty$ so that $u_t \to b/(a+b)$ and

$$\pi(t) \to \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}$$

whatever the value of $\pi(0)$.

A homogeneous MC has an invariant distribution π if

$$\pi = \pi P,$$

i.e. if $\pi(t) = \pi$ then so also $\pi(t+1) = \pi, \ldots$.

Calculation of invariant distributions can be done by directly solving

$$\pi P = \pi \text{ or equivalently } \pi(P - I) = 0.$$

together with the condition that $\sum_i \pi_i = 1$.

These equations always have a non-zero solution as, when written out as a set of equations, any one of them is a combination of the others, because they all sum to give $1 = 1$, or equivalently $0 = 0$. Thus any one equation is redundant and can be removed. Furthermore, the equations are also homogeneous, i.e. given any solution, it is also a solution when multiplied by any constant. Therefore we can always find a solution with the condition that $\sum_i \pi_i = 1$.

Therefore, a chain always has one invariant distribution (at least on a finite state space), and may have more than one. If π_a and π_b are two different invariant distributions, then another is $p\pi_a + (1-p)\pi_b$.

A sufficient condition for π to be the invariant distribution is given by the so-called detailed-balance equations:

$$\pi_i P_{ij} = \pi_j P_{ji} \text{ for all } i \text{ and } j.$$

(These equations do not always hold, but when they do they are much easier to solve.)

Example: To find the invariant distribution for the example above, either directly solve

$$(\pi_1 \;\; \pi_2)\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} = (\pi_1 \;\; \pi_2),$$

or use the detailed balance equations to get $a\pi_1 = b\pi_2$ or $\pi_1 = b\pi_2/a$. Thus

$$\pi' = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} \propto \begin{pmatrix} \frac{b}{a} \\ 1 \end{pmatrix} \propto \begin{pmatrix} b \\ a \end{pmatrix} \Rightarrow \pi' = \begin{pmatrix} \frac{b}{a+b} \\ \frac{a}{a+b} \end{pmatrix}.$$

The last step was simply to divide by the sum of the components of a solution proportional to that required.

Calculating the asymptotic distribution

If P has an asymptotic distribution π, then π is also its unique invariant distribution.

To see this, let $t\to\infty$ in $\pi(t+1) = \pi(t)P$ which gives $\pi = P\pi$, so that π is also invariant. Secondly, if π is invariant, set $\pi_1 = \pi$, so that $\pi(t) = \pi$ for all t and $\pi(t)\to\pi$ from which π must be the asymptotic distribution.

This shows that, if it exists, the asymptotic distribution is the invariant distribution. The following theorem shows when a Markov chain has an asymptotic distribution. The proof is beyond the scope of this course.

Theorem 10.2 The Ergodic Theorem. *A Markov chain, with a finite number of states, has an asymptotic distribution if and only if it is aperiodic and the set of persistent states is irreducible.*

Exercise 10.1 The size of a simple queue can be described by a stochastic process X_t which takes a step up when a new customer arrives and a step down when one has been served and leaves. Let n be the maximum size of the queue. The process X_t can be modelled as a MC with state space $\{0, 1, \ldots, n\}$ and transition probabilities $P_{00} = P_{i\ i-1} = q$, for $i = 1, \ldots, n$, $P_{nn} = P_{i\ i+1} = p = 1-q$, for $i = 0, \ldots, n-1$ and $P_{ij} = 0$ otherwise, where $0 < q < 1$. Does the chain have an asymptotic distribution?

Sol:

If $i < j$, then $\mathrm{P}(X_{j-i} = i | X_0 = j) \geq P_{j\ j-1} \ldots P_{i+1\ i} > 0$, and if $i > j$, then the probability $\mathrm{P}(X_{i-j} = i | X_0 = j) \geq P_{j\ j+1} \ldots P_{i-1\ i} > 0$. Hence the MC is irreducible. Since $P_{00} > 0$, it is aperiodic, and so it has an asymptotic distribution.

To find the asymptotic distribution, solve $\pi P = \pi$ for the invariant distribution. Take these equations in turn. Firstly,

$$q\pi_0 + q\pi_1 = \pi_0 \;\Rightarrow\; q\pi_1 = p\pi_0 \Rightarrow\; \pi_1 = \frac{p}{q}\pi_0.$$

Substitute for $p\pi_0 = q\pi_1$ in the next equation,

$$p\pi_0 + q\pi_2 = \pi_1 \;\Rightarrow\; q\pi_1 + q\pi_2 = \pi_1 \;\Rightarrow\; q\pi_2 = p\pi_1 \;\Rightarrow\; \pi_2 = \frac{p}{q}\pi_1.$$

The general equation is

$$p\pi_{j-1} + q\pi_{j+1} = \pi_j$$

and by a similar (inductive) argument

$$\pi_{j+1} = \frac{p}{q}\pi_j.$$

Thus π_j is a geometric progression and

$$\pi = \pi_0 \left[1, \frac{p}{q}, \left(\frac{p}{q}\right)^2, \ldots, \left(\frac{p}{q}\right)^n\right].$$

Divide the vector by its sum

$$\pi_0 \frac{1 - \left(\frac{p}{q}\right)^{n+1}}{1 - \frac{p}{q}}$$

to get

$$\pi = \frac{1 - \frac{p}{q}}{1 - \left(\frac{p}{q}\right)^{n+1}} \left[1, \frac{p}{q}, \left(\frac{p}{q}\right)^2, \ldots, \left(\frac{p}{q}\right)^n\right].$$

This is known as a truncated geometric distribution and for large n is close to the geometric distribution. □

Exercise 10.2 Let X_t be the maximum reading obtained in the first t rolls of a fair 6-sided die. Show that X_t is a Markov chain and give the initial distribution $\pi(0)$ and the transition probability matrix P.

Sol:

Since the die is fair, the initial distribution is $\pi(0) = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$. Let Y_t be the value shown by the die on the tth roll. Then Y_{t+1} is independent of $X_1, \ldots X_t$ and $X_{t+1} = \max(X_t, Y_{t+1})$. Hence

$$\begin{aligned}
\mathsf{P}(X_{t+1} = i_{t+1} | X_1 = i_1, \ldots, X_t = i_t) &= \mathsf{P}\left(\max(i_t, Y_{t+1}) = i_{t+1}\right) \\
&= \begin{cases} \frac{i_t}{6} & \text{if} i_{t+1} = i_t, \\ \frac{1}{6} & \text{if} i_{t+1} > i_t, \\ 0 & \text{otherwise} \end{cases} \\
&= \mathsf{P}(X_{t+1} = i_{t+1} | X_t = i_t).
\end{aligned}$$

Hence X_t is a Markov chain with transition probability matrix

$$P = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 1/3 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 1/2 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 2/3 & 1/6 & 1/6 \\ 0 & 0 & 0 & 0 & 5/6 & 1/6 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

□

Chapter 11

Gaussian Processes

A stochastic process $\{X_t : t \in T\}$ is Gaussian if for every finite set of indices $t_1, \ldots, t_k$ in the index set T, $\boldsymbol{X}_{t_1,\ldots,t_k} = (X_{t_1}, \ldots, X_{t_k})$ has a multivariate Normal distribution.

As the MVN distribution is determined by the mean and covariance matrix, in order to specify a Gaussian distribution, it is sufficient to specify $\mathsf{E}[X_t]$ for all $t \in T$ and $\mathsf{cov}(X_s, X_t)$ for all $s, t \in T$. We shall look at four important Gaussian processes.

11.1 The Wiener Process

The Wiener process is possibly the most widely studied Gaussian process and has applications throughout the mathematical sciences. In physics it is used to study Brownian motion, the diffusion of minute particles suspended in fluid. It is also prominent in the mathematical theory of finance, in particular the Black-Scholes option pricing model.

The Wiener process has a parameter $\sigma > 0$, called the volatility. It is defined as a continuous Gaussian process with $\mathsf{E}[X_t] = 0$ for all $t \geq 0$ and $\mathsf{cov}(X_s, X_t) = \sigma^2 \min(s, t)$ for all $s, t \geq 0$.

The standard Wiener process has $\sigma = 1$.

If $0 \leq t_1 \leq t_2 \leq t_3 \leq t_4$, then

$$\begin{aligned} \mathsf{cov}(X_{t_4} - X_{t_3}, X_{t_2} - X_{t_1}) &= \mathsf{cov}(X_{t_4}, X_{t_2}) - \mathsf{cov}(X_{t_4}, X_{t_1}) - \mathsf{cov}(X_{t_3}, X_{t_2}) + \mathsf{cov}(X_{t_3}, X_{t_1}) \\ &= \sigma^2(t_4 - t_4 - t_3 + t_3) = 0. \end{aligned}$$

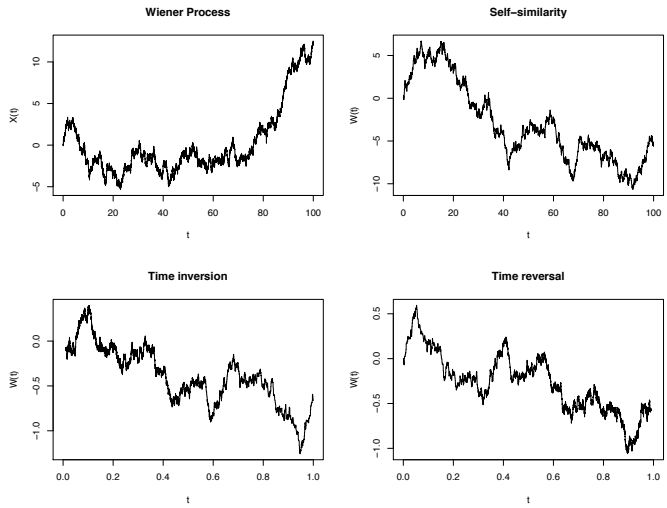
Because of the properties of the MVN distribution, this means that $X_{t_4} - X_{t_3}$ and $X_{t_2} - X_{t_1}$ are independent. So the Wiener process has independent increments. This provides an easy way to simulate the Wiener process.

The following properties of the Wiener process can all be verified by calculating the means and covariances.

- Stationary increments: for any $s > 0$, the process $W_t = X_{t+s} - X_s$ is another Wiener process.
- Self-similarity: for every $c > 0$, the process $W_t = (1/\sqrt{c})X_{ct}$ is another Wiener process.

- Time reversal: the process $W_t = X_1 - X_{1-t}$ for $0 \le t \le 1$ is distributed like X_t for $0 \le t \le 1$.
- Time inversion: the process $W_t = tX_{1/t}$ is another Wiener process.

The diagrams below illustrates these transformation properties.



Realisations of the Wiener processes generated from the top left process $X(t)$. The top right process is generated by $W_1(t) = 2X(t/4)$; the bottom left by $W_2(t) = tX(1/t)$; and the bottom right by $W_3(t) = W_2(1) - W_2(1-t)$.

Exercise 11.1 Suppose that the value of £1 of stock at time 0, is worth $P_t = \exp(\sigma W_t)$ at time t, where W_t is a standard Wiener process and $\sigma > 0$ is the volatility. Suppose that the interest rate is $\rho > 0$, so £1 in the bank at time 0, will have increased to $e^{\rho t}$ at time t. Show that there exists a value of ρ, such that $\mathsf{E}[P_t] = e^{\rho t}$. For this value of ρ, calculate $\mathsf{E}[e^{-\rho t} P_t | P_s]$, where $s < t$ and interpret your result. [Hint: Write $W_t = W_s + (W_t - W_s)$.]

Sol:

The random variable $W_t \sim N(0, t)$. Therefore

$$\begin{aligned}
\mathsf{E}[P_t] &= \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} e^{\sigma x} e^{-\frac{x^2}{2t}} dx \\
&= e^{\sigma^2 t/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-\sigma t)^2}{2t}} dx \\
&= e^{\sigma^2 t/2},
\end{aligned}$$

and so if $\rho = \sigma^2/2$, then $\mathsf{E}[P_t] = e^{\rho t}$. Now since $W_t - W_s \sim N(0, \sigma^2(t-s))$ is independent of W_s, and hence of P_s

$$\begin{aligned}
\mathsf{E}[e^{-\rho t} P_t | P_s] &= e^{-\rho t} \mathsf{E}[\exp(\sigma(W_t - W_s)) P_s | P_s] \\
&= e^{-\rho t} P_s \mathsf{E}[\exp(\sigma(W_t - W_s)) | P_s] \\
&= e^{-\rho t} P_s e^{\rho(t-s)} \\
&= e^{-\rho s} P_s.
\end{aligned}$$

The interpretation is that for interest rate $\rho = \sigma^2/2$, the discounted stock price is a fair price. □

11.2 The Ornstein-Uhlenbeck Process

The Ornstein-Uhlenbeck process has wide applications in financial mathematics and is used to model interest rates, currency exchange rates, and commodity prices stochastically. It is also used in physics to model, for example, the dynamics of springs.

The Ornstein-Uhlenbeck process has parameters $\sigma > 0$ and $\theta > 0$. It is defined as a continuous Gaussian process with $\mathsf{E}[X_t] = 0$ for all $t \geq 0$ and $\mathsf{cov}(X_s, X_t) = \frac{\sigma^2}{2\theta} e^{-\theta(s+t)}(e^{2\theta \min(s,t)} - 1)$ for all $s, t \geq 0$. Unlike the Wiener process, the Ornstein-Uhlenbeck process converges to a stationary distribution as $t \to \infty$.

Often, you may see this process expressed as the solution to:

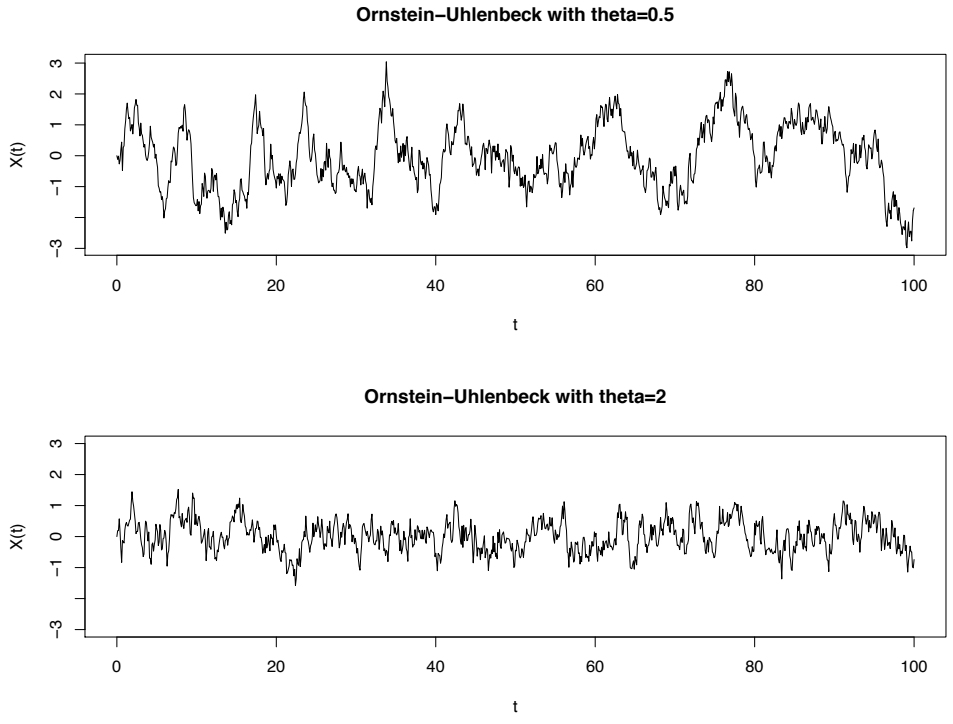
$$dX_s = -\theta X_s ds + \sigma dW_s.$$

with $X_0 = 0$.

The Ornstein-Uhlenbeck process can be expressed as a time-change of the standard Wiener process $W(t) = W_t$ by

$$X_t = \frac{\sigma}{\sqrt{2\theta}} W(e^{2\theta t}) e^{-\theta t}.$$

This provides a useful way to simulate the Ornstein-Uhlenbeck process.



Realisations of the Ornstein-Uhlenbeck process, with $\theta = 0.5$ and $\theta = 2$.

Exercise 11.2 Calculate the stationarity distribution of the Ornstein-Uhlenbeck Process

Sol:

Consider the above definition of X_t. It is a linear transformation of W_t, and hence has a marginal distribution that is Gaussian:

$$X_t \sim \mathsf{N}\left(0, \frac{\sigma^2}{2\theta} e^{-\theta(2t)}(e^{2\theta t} - 1)\right).$$

Now taking the limit as $t \to \infty$ gives the stationary distribution:

$$\mathsf{N}\left(0, \frac{\sigma^2}{2\theta}\right).$$

□

The above Ornstein-Uhlenbeck process has $X_0 = 0$. We can construct an Ornstein-Uhlenbeck process, R_s which has a non-zero starting value, r_0, by the following transformation

$$R_s = e^{-\theta s} r_0 + X_s.$$

We can also construct an Ornstein-Uhlenbeck process which is at stationarity, by assigning R_0 to be drawn from the stationary distribution of the process. This would lead to a Gaussian process with $\mathsf{E}[X_t] = 0$ for all $t \geq 0$ and $\mathsf{cov}(X_s, X_t) = \frac{\sigma^2}{2\theta} e^{-\theta|s-t|}$ for all $s, t \geq 0$.

Exercise 11.3 Let V_s be an Ornstein-Uhlenbeck process but with $\mathsf{E}[V_t] = \mu$, $V_0 = v_0$. Show how you can express V_s in terms of X_s.

Sol:

We have that $V_s = R_s + \mu$. Hence

$$V_s = e^{-\theta s} r_0 + X_s + \mu.$$

Now $V_0 = R_0 + \mu$ so $r_0 = v_0 - \mu$. So

$$V_s = e^{-\theta s}(v_0 - \mu) + X_s + \mu = e^{-\theta s} v_0 + (1 - e^{-\theta s})\mu + X_s.$$

□

Appendix A

Useful Integrals

A.1 Gamma Function

Let $h(x) = x^{\alpha-1}\exp(-x)$ for $0 < x < \infty$. The Gamma function determines how the integral of this function over the range $(0, \infty)$ varies with α.

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\exp(-x)\,dx, \text{ for } \alpha > 0.$$

The Gamma function in R is `gamma(a)`:

```
gamma(0.5)        # the gamma function at 0.5
   [1] 1.772454
x = seq(1,10)     # Let x = (1,2,3,4,5,6,7,8,9,10)
x
   [1]  1  2  3  4   5    6    7    8      9     10
gamma(x)          #  gamma function at these values
   [1]  1  1  2  6  24  120  720 5040  40320 362880
```

Properties:

- Recurrence relation $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ for $\alpha > 1$. This is easy to prove using integration by parts.
- Special cases: $\Gamma(1) = 1$, $\Gamma(\alpha) = (\alpha-1)!$ for α a positive integer, and $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(\alpha)\to\infty$ as $\alpha\to 0$ or $\alpha\to\infty$.

Related integrals: For $\alpha > 0$ and $\beta > 0$

$$\int_0^\infty x^{\alpha-1}\exp(-\beta x)\,dx = \beta^{-\alpha}\Gamma(\alpha).$$

To see this substitute $y = \beta x$ gives $dy = \beta\, dx$ so

$$\begin{aligned} & \int_0^\infty (y/\beta)^{\alpha-1} \exp(-y) dy/\beta \\ = \quad & \beta^{-\alpha} \int_0^\infty y^{\alpha-1} \exp(-y) dy = \beta^{-\alpha}\Gamma(\alpha). \end{aligned}$$

A.2 Normal Distribution Integrals

Let $h(x) = x^r \exp(-x^2/2)$ for $-\infty < x < \infty$. The function $J(r)$ determines how the integral of $h(x)$ over the range $(-\infty, \infty)$ varies over non-negative integer values of r. As $h(x)$ is an odd (even) function and all the integrals converge for all $r > 0$ we have that $J(r) = 0$ for all odd r, and for even r

$$\begin{aligned} J(r) &= \int_{-\infty}^{\infty} x^r \exp(-x^2/2)\, dx \\ &= 2\int_0^{\infty} x^r \exp(-x^2/2)\, dx. \end{aligned}$$

Substituting $s = x^2/2$ gives

$$\begin{aligned} J(r) &= 2\int_0^{\infty} (2s)^{(r-1)/2} \exp(-s)\, ds \\ &= 2^{(r+1)/2}\Gamma((r+1)/2). \end{aligned}$$

It follows that $J(0) = \sqrt{2\pi}$ and $J(2) = \sqrt{2\pi}$, and $J(4) = 3\sqrt{2\pi}$.

A.3 Beta Function

Let $h(x) = x^{\alpha_1-1}(1-x)^{\alpha_2-1}$ for $0 < x < 1$. The Beta function $B(\alpha_1, \alpha_2)$ determines how the integral of this function over the range $(0, 1)$ varies with $\alpha_1 > 0$ and $\alpha_2 > 0$

$$B(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1-1}(1-x)^{\alpha_2-1} dx.$$

Properties:

- $B(1, 1) = 1$.
- $B(\alpha_1, \alpha_2) = B(\alpha_2, \alpha_1)$
- $B(\alpha_1, \alpha_2) = \Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1 + \alpha_2)$

Thus the Beta function can easily be evaluated using the Gamma function. In R:

```
gamma(4)*gamma(0.5)/gamma(4+0.5)   # Calc Beta(4,0.5)
          [1] 0.9142857
```