University *of*
Leicester

Lecture Notes

MA2404/MA7404

Markov Processes

Module MA2404/MA7404
Markov Processes
Edition 3, March 2022

Preface

Welcome to module MA2404/MA7404 *Markov Processes*.

The module has no formal prerequisites apart from that you are familiar with standard mathematics seen in first year undergraduate courses.

An important thread through all actuarial and financial disciplines is the use of appropriate models for various types of risks. The aim of this module is to provide an introduction to risk modelling, with emphasis on Markov models.

We begin the module with short review of probability theory, basic statistics, and stochastic processes, then study risk models, theory of Markov processes and their application to actuarial and financial modelling.

Dr Tetiana Grechuk

Contents

The following book has been used as the basis for the lecture notes

Faculty and Institute of Actuaries, CS2 Core Reading

Chapter 1
Probability theory and stochastic processes

The aim of this chapter is to give the review of probability theory, basic statistics, and give a very short introduction to stochastic processes. This background material is necessary for understanding the models that will be developed in later chapters. While the rigorous development of probability theory and stochastic processes can be very technical, we have attempted to avoid unnecessary technicalities in this text. The focus is that you understand the material on the intuitive level, and also have a level of technical knowledge able to solve quantitative problems when necessary.

1.1 Probability space

Intuitively, a random variable is a function with random values. For example, it may be a lifetime of an individual, or the number of car accidents in the next year. To model such randomness mathematically it is convenient to assume that a random variable is a function "defined somewhere" on the space of all possible "states of the world", ω. The randomness comes from the fact that "we do not know exactly what ω we are in".

As the simplest example, consider tossing a fair coin (i.e. that the probability of a head is equal to the probability of a tail, and so these probabilities are equal to 1/2). Then we can take $\Omega = \{H, T\}$, and $P(H) = P(T) = 1/2$, where P denotes the probability. In experiment with throwing a dice, we can take $\Omega = \{1, 2, 3, 4, 5, 6\}$, and the probability of each outcome is 1/6. In general, any experiment with finite number of possible outcomes can be studied using probability space

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\},$$

with probability $P(\omega_i) \geq 0$, $i = 1, \ldots, n$, and $\sum_{i=1}^{n} P(\omega_i) = 1$. Any subset $A \subset \Omega$ is called *event*, and the probability of A is $P(A) = \sum_{\omega \in A} P(\omega)$. For example, in the experiment with dice, $A = \{2, 4, 6\}$ corresponds to the event "dice shows an even number", and $P(A) = P(2) + P(4) + P(6) = 3/6 = 1/2$.

However, some random quantities, such as temperature outside tomorrow at 2pm (measured in some units) or a lifetime of an individual (measured, for example, in years), may take any real number from some interval. To model such random quantities, we need to consider probability space Ω with arbitrary, possibly infinitely many of elements. In this case, some subsets of Ω has very complicated structure, and the probability for such subsets cannot even be defined. Hence, we divide subsets of Ω into two groups: subsets for

which we will define the probability (such subsets we will call "events"), and subsets for which the probability is undefined. The set of all events will be denoted $\mathcal{F}$. We assume that $\mathcal{F}$ satisfies the following properties

1. $\Omega \in \mathcal{F}$
2. If $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ and $\bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$
3. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

Here, $\bigcup$ denotes the union of the sets, $\bigcap$ denotes the intersection, and $A^c = \{\omega \in \Omega \,|\, \omega \notin A\}$ denotes the complement of set A.

We now define a probability as a function $P : \mathcal{F} \to [0, 1]$ which satisfies the following properties

1. $P(\Omega) = 1$
2. $P\big(\bigcup_{n=1}^{\infty} A_n\big) = \sum_{n=1}^{\infty} P(A_n)$, provided $A_1, A_2, \cdots \in \mathcal{F}$ is a collection of pairwise disjoint sets, i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Then we call such P a *probability measure* and the triple $(\Omega, \mathcal{F}, P)$ a *probability space*.

An important example of a probability space is the so-called standard probability space $\big([0, 1], \mathcal{B}, \lambda\big)$, where $\mathcal{B}$ is the smallest set satisfying the properties 1-3 above and containing all intervals of the type $(a, b), (a, b], [a, b), [a, b] \subset [0, 1]$ and λ denotes the *Lebesgue measure*, that is, a natural extension of the functional which assigns to each interval its length:

$$\lambda(a, b] = b - a;$$

to the whole set $\mathcal{B}$.

1.2 Random variables and their expectations

Given a probability space $(\Omega, \mathcal{F}, P)$ we call a function

$$\xi : \Omega \to \mathbb{R}$$

a *random variable* if it is *measurable*, i.e. satisfies

$$\big\{\omega : \xi(\omega) \leq r\big\} \in \mathcal{F} \qquad \text{for all } \; r \in \mathbb{R}.$$

We may interpret a random variable as a number produced by an experiment, such as temperature or the oil price tomorrow. The mathematical formalism allows us to be accurate. In particular, the property of measurability is important to make sense of the probability of the event that a random variable does not exceed a threshold: $P\big(\xi \le r) := P(\{\omega : \xi(\omega) \le r\}\big)$.

For example, consider standard probability space $\big([0,1], \mathcal{B}, \lambda\big)$, and a function ξ such that $\xi(\omega) = a$, $\omega \le p$, and $\xi(\omega) = b$, $\omega > p$, where $p \in (0,1)$ and $a \ne b$. Then ξ is a random variable assuming just two different values a and b with probabilities $p \in (0,1)$ and $q := 1 - p$ respectively. It is called a *Bernoulli random variable*. If $a = 1$ and $b = 0$, ξ is called a standard Bernoulli variable.

Another example is linear function

$$\xi(\omega) = a + (b - a)\omega, \quad a < b, \tag{1}$$

on the standard probability space. This is the random variable whose possible values are all real numbers in the interval $[a, b]$.

The *expectation* of a random variable is, informally, the average of its possible values weighted according to their probabilities. Sometimes in the literature, the expectation is also called the *mathematical expectation* or *mean*. For example, in experiment with throwing a dice, the possible outcomes of the experiments are $\Omega = \{1, 2, 3, 4, 5, 6\}$ with equal probabilities, and the average outcome is $(1+2+3+4+5+6)/6 = 3.5$. More generally, if the possible values of ξ are $x_1, x_2, \ldots, x_n$, and they happen with probabilities $p_1, p_2, \ldots, p_n$, respectively, then

$$E[\xi] = \sum_{i=1}^{n} x_i p_i. \tag{2}$$

For example, if ξ is a Bernoulli random variable with outcomes a, b with probabilities $p, q = 1 - p$, then

$$E[\xi] = ap + bq = ap + b(1 - p).$$

Similarly, if ξ can take infinitely many possible values

$$x_1, x_2, \ldots, x_n, \ldots$$

with probabilities

$$p_1, p_2, \ldots, p_n, \ldots,$$

respectively, where each $p_n \geq 0$ and $\sum_{n=1}^{\infty} p_n = 1$, then

$$E[\xi] = \sum_{i=1}^{\infty} x_i p_i. \tag{3}$$

To calculate the average of possible values for general function (for example, one that may take any values from some interval, such as (1)), the summation in (2) and (3) is replaced by *integration*. By definition, the expectation of a random variable is its *integral on* Ω *with respect to* P, so that

$$E[\xi] \equiv \int_{\Omega} \xi \, dP \equiv \int_{\Omega} \xi(\omega) \, dP(\omega).$$

In particular, on the standard probability space

$$E[\xi] = \int_0^1 \xi(\omega) d\omega.$$

For example, for ξ defined in (1),

$$E[\xi] = \int_0^1 (a + (b-a)\omega) d\omega = a + (b-a)/2 = (a+b)/2.$$

It is easy to check that for random variable taking finitely many values this integral reduces to (2). For example, for Bernoulli random variable,

$$E[\xi] = \int_0^1 \xi(\omega) d\omega = \int_0^p a d\omega + \int_p^1 b d\omega = ap + b(1-p).$$

For some functions, for example $\xi(\omega) = 1/\omega$, $0 < \omega \leq 1$, the integral $\int_0^1 \xi(\omega) d\omega$ is not finite. If $E[|\xi|]$ exists and is finite, then ξ is called *integrable*. The class of all integrable random variables is denoted by $L^1(\Omega, \mathcal{F}, P)$ or just L^1.

The key property of the expectation is linearity

$$E[a\xi + b\eta] = aE[\xi] + bE[\eta], \quad \text{for all } \xi, \eta \in L^1 \text{ and constants } a, b \in \mathbb{R}.$$

1.3 Variance, covariance, and correlation

The *variance* of a random variable ξ is defined by

$$\mathrm{Var}(\xi) := E[(\xi - E[\xi])^2] \equiv E[\xi^2] - E[\xi]^2.$$

If ξ measures some real life phenomenon, such as remaining lifetime of an individual, $E[\xi]$ indicates "how big" ξ to expect on average, and may serve as a forecast how long is an individual expected to survive. Variance measures how big is the (square of) the difference $\xi - E[\xi]$, and therefore indicates how "close" is the prediction $E[\xi]$ to the reality. Therefore mean and variance are two fundamental quantities associated with a random variable.

Not every $\xi \in L^1$ has a finite variance. The class of random variables with finite variance is denoted by $L^2(\Omega, \mathcal{F}, P)$ or just L^2. A random variable with finite variance is called *square-integrable*. Variance is always non-negative and is equal to zero only for constants.

The square root of the variance $\sigma(\xi) = \sqrt{\mathrm{Var}(\xi)}$ is called the *standard deviation* of ξ.

Example 1.1. For ξ defined in (1),

$$E[\xi^2] = \int_0^1 (a + (b-a)\omega)^2 d\omega = (a^2 + ab + b^2)/3,$$

and

$$\sigma(\xi) = \sqrt{\mathrm{Var}(\xi)} = \sqrt{(a^2 + ab + b^2)/3 - (a+b)^2/4} = (b-a)/\sqrt{12}.$$

Note that

$$\mathrm{Var}(a\xi) = a^2 \mathrm{Var}(\xi), \quad \sigma(a\xi) = a\sigma(\xi), \quad \text{for all } \xi \in L^2 \text{ and constant } a \in \mathbb{R}.$$

For two random variables $\xi, \eta \in L^2$ their *covariance* is defined by

$$\mathrm{Cov}(\xi, \eta) := E\big[(\xi - E[\xi])(\eta - E[\eta])\big] \equiv E[\xi\eta] - E[\xi]E[\eta].$$

For a sum of two random variables we have

$$\begin{aligned} \mathrm{Var}(\xi + \eta) &= E\big[\big((\xi - E[\xi]) + (\eta - E[\eta])\big)^2\big] \\ &= E\big[(\xi - E[\xi])^2\big] + 2E\big[(\xi - E[\xi])(\eta - E[\eta])\big] + E\big[(\eta - E[\eta])^2\big] \\ &= \mathrm{Var}(\xi) + 2\mathrm{Cov}(\xi, \eta) + \mathrm{Var}(\eta). \end{aligned}$$

Sometimes it is convenient to normalise the covariance. For non-constant random variables ξ and η, define the *correlation* of ξ and η by

$$\mathrm{Corr}(\xi, \eta) := \frac{\mathrm{Cov}(\xi, \eta)}{\sqrt{\mathrm{Var}(\xi)}\sqrt{\mathrm{Var}(\eta)}}.$$

Cauchy-Schwarz inequality says

$$\big|E[\xi\eta]\big|^2 \le E[\xi^2]E[\eta^2] \qquad \text{for all } \xi,\eta \in L^2,$$

and one can easily deduce that

$$-1 \le \mathrm{Corr}(\xi,\eta) \le 1 \qquad \text{for all } \xi,\eta \in L^2.$$

The correlation equals 1 if and only if $\xi = a\eta$ with some constant $a > 0$, and is -1 if and only if $\xi = a\eta$ for $a < 0$. If ξ and η are independent, then the covariance $\mathrm{Cov}(\xi,\eta)$ and correlation $\mathrm{Corr}(\xi,\eta)$ are 0. If $\mathrm{Corr}(\xi,\eta) > 0$, the random variables ξ and η are called positively correlated, and the intuition is that higher values of ξ is an indication of higher values of η. If $\mathrm{Corr}(\xi,\eta) < 0$, ξ and η are called negatively correlated. For example, if ξ is temperature outside and η is the number of old people who died on the street, then ξ and η may be positively correlated during the summer (the higher temperature the hotter summer, and old people may be affected by very hot weather) but negatively correlated during the winter (the higher temperature the less cold the winter).

1.4 Probability distribution

Given a random variable ξ we define its *cumulative distribution function* (or CDF in short) by

$$F_\xi(x) := P(\xi \le x).$$

Clearly the CDF is non-decreasing in x and $\lim_{x\to-\infty} F_\xi(x) = 0$, $\lim_{x\to+\infty} F_\xi(x) = 1$.

A random variable ξ is called *discretely distributed* if it can take values $x_1, x_2, \ldots, x_n, \ldots$ with probabilities $p_1, p_2, \ldots, p_n, \ldots$, respectively. The CDF of such random variable is a piecewise constant function.

Example 1.2. An n-sided dice has numbers $1, 2, \ldots, n$ on its sides, each can be shown on its upper surface with the same probability $1/n$ when the dice is thrown or rolled. If ξ is the corresponding random variable, it is discretely distributed, and its CDF $F_\xi(x)$ is a piecewise constant function given by

$$F_\xi(x) = \begin{cases} 0 & \text{if } x < 1 \\ i/n & \text{if } i \le x < i+1,\ i = 1,2,\ldots,n-1 \\ 1 & \text{if } n \le x \end{cases}$$

By (2),

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\cdot\frac{n(n+1)}{2} = \frac{n+1}{2},$$

and

$$E[X^2] = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{1}{n}\cdot\frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6},$$

hence

$$Var[X] = E[X^2] - (E[X])^2 = \frac{n^2-1}{12}. \tag{4}$$

A random variable ξ is called *continuously distributed* if its CDF can be represented as

$$F_\xi(x) = \int_{-\infty}^{x} \rho_\xi(z)\,dz,$$

for some non-negative function ρ_ξ, which is called the *probability density function* (or PDF in short) of ξ.

The expectation of a random function can be calculated using its CDF or its PDF (if the latter exists) by

$$E[\xi] = \int_{-\infty}^{\infty} x\,dF_\xi(x) = \int_{-\infty}^{\infty} x\rho_\xi(x)\,dx.$$

Or more generally, if for some random variable ξ and function $f : \mathbb{R} \to \mathbb{R}$ the expectation $E[f(\xi)]$ exists, then

$$E\big[f(\xi)\big] = \int_{-\infty}^{\infty} f(x)\,dF_\xi(x) = \int_{-\infty}^{\infty} f(x)\rho_\xi(x)\,dx.$$

In particular, for the variance we have

$$\mathrm{Var}(\xi) = \int_{-\infty}^{\infty} \Big(x - E[\xi]\Big)^2\,dF_\xi(x) = \int_{-\infty}^{\infty} \Big(x - E[\xi]\Big)^2\rho_\xi(x)\,dx.$$

Example 1.3. For ξ defined in (1),

$$F_\xi(x) = P(\xi \le x) = 0 \quad \text{if} \quad x < a,$$

$$F_\xi(x) = P(\xi \le x) = \lambda(\omega : a + (b-a)\omega \le x) = \frac{x-a}{b-a} \quad \text{if} \quad a \le x \le b,$$

where λ denotes the length of the interval, and

$$F_\xi(x) = P(\xi \le x) = 1 \quad \text{if} \quad b < x.$$

It is easy to check that

$$F_\xi(x) = \int_{-\infty}^{x} \rho_\xi(z)\, dz,$$

for function $\rho_\xi(z) = 1/(b-a)$, $z \in [a,b]$ (and $\rho_\xi(z) = 0$, $z \notin [a,b]$). Hence, ξ is continuously distributed with PDF ρ_ξ. In fact, random variable ξ with this density is called *uniformly* distributed on $[a,b]$. In this case we usually write $\xi \sim U(a,b)$.

We have

$$E[\xi] = \int_{-\infty}^{\infty} x\rho_\xi(x)\, dx = \int_a^b \frac{x}{b-a}\, dx = \frac{a+b}{2},$$

and

$$\mathrm{Var}(\xi) = \int_{-\infty}^{\infty} \Big(x - E[\xi]\Big)^2 \rho_\xi(x)\, dx = \int_a^b \frac{(x-E[\xi])^2}{b-a}\, dx = \frac{(b-a)^2}{12},$$

which coincides with formulas obtained in Example 1.1.

Note that random variable $\eta(\omega) = b + (a-b)\omega$ has the same CDF (and PDF) as random variable ξ in Example 1.3, despite on the fact that $\xi \neq \eta$. If random variables ξ and η have identical CDFs $F_\xi = F_\eta$, we say that they are *identically distributed* (i.d.). It can be shown that the random variables ξ and η are identically distributed if and only if $E[f(\xi)] = E[f(\eta)]$ for any $f : \mathbb{R} \to \mathbb{R}$.

The inverse function to $F_\xi(x)$, defined as

$$F_\xi^{-1}(\alpha) = \inf\{x | F_\xi(x) > \alpha\}, \quad 0 < \alpha < 1,$$

is called the α-quantile of ξ. For example, random variables $\eta(\omega) = b + (a-b)\omega$ has CDF described in Example 1.3, and its inverse can be found by solving equation $\frac{x-a}{b-a} = \alpha$, resulting in $x = a + (b-a)\alpha$. Hence, $F_\eta^{-1}(\alpha) = a + (b-a)\alpha$.

For several random variables $\xi_1, \dots \xi_d$ their joint CDF is defined as

$$F_{\xi_1,\dots,\xi_d}(x_1,\dots,x_d) = P[\xi_1 \le x_1, \xi_2 \le x_2, \dots, \xi_d \le x_d].$$

If it can be represented as

$$F_{\xi_1,\dots,\xi_d}(x_1,\dots,x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} \rho_{\xi_1,\dots,\xi_d}(z_1,\dots,z_d)\, dz_1 \dots dz_d,$$

for some non-negative function $\rho_{\xi_1,\dots,\xi_d}(z_1,\dots,z_d)$, the latter is called the joint PDF of $\xi_1,\dots,\xi_d$.

If for some random variable $\xi_1,\dots\xi_d$ and function $f:\mathbb{R}^d\to\mathbb{R}$ the expectation $E[f(\xi_1,\dots\xi_d)]$ exists, then

$$E\big[f(\xi_1,\dots\xi_d)\big]=\int_{\mathbb{R}^d} f(x_1,\dots,x_d)\rho_{\xi_1,\dots,\xi_d}(x_1,\dots,x_d)\,dx_1\dots dx_d.$$

1.5 Examples of discrete probability distributions

One example of discrete probability distribution is studied in Example 1.2. This section provides more examples, with emphasis to ones useful in insurance modelling. The most common use of discrete probability distribution in insurance is to count the number of claims. The following discrete distribution often used for this.

- ξ is a Bernoulli variable if it takes values a and b with probabilities p and $1-p$, respectively, where a and b are arbitrary real numbers and $p\in[0,1]$. By (2),

  $$E[\xi]=ap+b(1-p),\quad E[X^2]=a^2p+b^2(1-p)$$

  and

  $$Var[\xi]=E[X^2]-(E[X])^2=(a-b)^2p(1-p).$$

  If $a=1$ and $b=0$, ξ is called standard Bernoulli variable. In this case,

  $$E[\xi]=p,\quad Var[\xi]=p(1-p).$$

  Standard Bernoulli variable is used to model the situation when only 1 claim can happen within a policy, and the probability of this claim is p.

- ξ follows Binomial distribution with parameters n (non-negative integer) and p (real number on $[0,1]$) if it takes values $0,1,2,\dots,n$ with probabilities

  $$p_k=\frac{n!}{k!(n-k)!}\cdot p^k(1-p)^{n-k},\quad k=0,1,\dots,n.$$

  In this case, we write $\xi\sim Bin(n,p)$. For example, if $n=1$, then

  $$p_0=\frac{1!}{0!(1-0)!}\cdot p^0(1-p)^{1-0}=1-p,\quad p_1=\frac{1!}{1!(1-1)!}\cdot p^1(1-p)^{1-1}=p,$$

hence in this case ξ is just a standard Bernoulli variable. In general, if $\xi_1, \ldots, \xi_n$ is a sequence of i.i.d. standard Bernoulli variables with parameter p, then

$$\xi = \xi_1 + \xi_2 + \cdots + \xi_n$$

has binomial distribution $Bin(n, p)$. In particular,

$$E[\xi] = \sum_{k=1}^{n} E[\xi_k] = \sum_{k=1}^{n} p = np,$$

and

$$Var[\xi] = \sum_{k=1}^{n} Var[\xi_k] = \sum_{k=1}^{n} p(1-p) = np(1-p).$$

Binomial distribution arises when there are n independent policies such that each can produce a claim with the same probability p. Then the total number of claims from all policies is $Bin(n, p)$.

- ξ follows geometric distribution with parameter p, $0 \leq p \leq 1$, if it takes values $0, 1, 2, \ldots$ with probabilities

$$p_n = (1-p)^n p, \quad n = 0, 1, 2, \ldots$$

If $\xi_1, \xi_2, \ldots$ is an infinite sequence of i.i.d. standard Bernoulli variables with parameter p, then the number of 0-s in this sequence before the first 1 follows the geometric distribution. This is a valid distribution because

$$\sum_{n=1}^{\infty} p_n = p \sum_{n=1}^{\infty} (1-p)^n = p \frac{1}{1-(1-p)} = 1.$$

Differentiating both sides of equation $\sum_{n=1}^{\infty} (1-p)^n = \frac{1}{p}$, we get

$$-\sum_{n=1}^{\infty} n(1-p)^{n-1} = -\frac{1}{p^2}.$$

This can be used to calculate the expectation of geometric distribution

$$E[\xi] = \sum_{n=1}^{\infty} n \cdot p_n = p(1-p) \sum_{n=1}^{\infty} n(1-p)^{n-1} = p(1-p)\frac{1}{p^2} = \frac{1-p}{p}.$$

By similar (but more involved) calculation, we can get

$$Var[\xi] = \frac{1-p}{p^2}.$$

- ξ follows negative Binomial distribution with parameters k (positive integer) and p (real number on $[0,1]$) if it takes values $0,1,2,\ldots$ with probabilities

$$p_n = \frac{(k+n-1)!}{n!(k-1)!} \cdot p^k (1-p)^n, \quad n = 0,1,2,\ldots$$

In this case, we write $\xi \sim NB(k,p)$. For example, if $k=1$, then

$$p_n = \frac{(1+n-1)!}{n!(1-1)!} \cdot p^1 (1-p)^n = p(1-p)^n, \quad n = 0,1,2,\ldots$$

hence $NB(1,p)$ is just a geometric distribution. In general, if $\xi_1,\ldots,\xi_n$ is a sequence of i.i.d. geometric variables with parameter p, then

$$\xi = \xi_1 + \xi_2 + \cdots + \xi_k$$

has negative binomial distribution $NB(k,p)$. In particular,

$$E[\xi] = \sum_{i=1}^{k} E[\xi_i] = \sum_{i=1}^{k} \frac{1-p}{p} = \frac{k(1-p)}{p},$$

and

$$Var[\xi] = \sum_{i=1}^{k} Var[\xi_i] = \sum_{i=1}^{k} \frac{1-p}{p^2} = \frac{k(1-p)}{p^2}.$$

Negative Binomial distribution can serve as a model for the total number of claims if this number is not bounded from above.

- ξ follows Poisson distribution with parameter $\lambda > 0$ if it takes values $0,1,2,\ldots$ with probabilities

$$p_n = \frac{\lambda^n e^{-\lambda}}{n!}$$

Using series expansion for exponential function $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, one can easily compute expectation of Poisson distribution

$$E[\xi] = \sum_{n=1}^{\infty} n \cdot p_n = \sum_{n=1}^{\infty} n \frac{\lambda^n e^{-\lambda}}{n!} = \lambda e^{-\lambda} \sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

Similar calculation shows that $E[\xi^2] = \lambda^2 + \lambda$, hence

$$Var[\xi] = E[\xi^2] - (E[\xi])^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda.$$

Poisson distribution is a natural model for the total number of claims if the claims arrives at "uniform rate λ claims per unit of time". We will study this model in details later in this course.

1.6 Examples of continuous probability distributions

One example (uniform probability distribution) is studied in Example 1.3. This section provides more examples, with emphasis to ones useful in insurance modelling.

Example 1.4. Question: We say that random variable ξ has the exponential distribution with parameter $\lambda \in \mathbb{R} > 0$, and write $\xi \sim \text{Exp}(\lambda)$, if its probability density function is

$$\rho_\xi(x) = \lambda e^{-\lambda x} \qquad \text{for } x \geq 0, \tag{5}$$

and $\rho_\xi(x) = 0$ for $x < 0$. Calculate $F_\xi(x)$, $F_\xi^{-1}(\alpha)$, $E[\xi]$ and $\text{Var}(\xi)$.

Answer: For any $x \geq 0$,

$$F_\xi(x) = \int_{-\infty}^{x} \rho_\xi(z)\, dz = \int_0^x \lambda e^{-\lambda z}\, dz = 1 - e^{-\lambda x}.$$

Equation $1 - e^{-\lambda x} = \alpha$ has a solution $x = -\ln(1-\alpha)/\lambda$, hence $F_\xi^{-1}(\alpha) = -\ln(1-\alpha)/\lambda$.
The expectation is calculated as

$$E[\xi] = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}.$$

The variance is calculated as

$$\text{Var}(\xi) = \int_0^\infty \left(x - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2}.$$

Some elementary calculus was required in each case.

Other examples of important probability distributions include:

- ξ has a normal distribution with parameters μ and σ^2 if it has PDF

$$\rho_\xi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{6}$$

  In that case we write $\xi \sim N(\mu, \sigma^2)$. The mean and variance of ξ are $E[\xi] = \mu$ and $\text{Var}(\xi) = \sigma^2$.

- ξ has a lognormal distribution if $\log(\xi)$ has a normal distribution, that is, $\log(\xi) \sim N(\mu, \sigma^2)$. In this case, we will write $\xi \sim LogN(\mu, \sigma^2)$. Equivalently, ξ has a lognormal distribution if its PDF is

$$\rho_\xi(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}}, \quad x > 0. \tag{7}$$

The mean and variance of lognormal distribution are

$$E[\xi] = e^{\mu+\sigma^2/2}, \quad \mathrm{Var}(\xi) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}.$$

- ξ has a gamma distribution with parameters $\alpha > 0$ and $\lambda > 0$ if it has PDF

$$\rho_\xi(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0, \tag{8}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$, $\alpha > 0$ is the gamma function. In this case we write $\xi \sim Ga(\alpha, \lambda)$. The mean and variance of ξ are

$$E[\xi] = \frac{\alpha}{\lambda}, \quad \mathrm{Var}(\xi) = \frac{\alpha}{\lambda^2}.$$

- ξ has the Pareto distribution with parameters $\alpha > 0$ and $\lambda > 0$ if it has PDF

$$\rho_\xi(x) = \frac{\alpha\lambda^\alpha}{(\lambda + x)^{\alpha+1}}, \quad x > 0. \tag{9}$$

In this case we write $\xi \sim Pa(\alpha, \lambda)$. Pareto distribution has CDF

$$F_\xi(x) = \int_{-\infty}^{x} \rho_\xi(z)\, dz = \int_0^x \frac{\alpha\lambda^\alpha}{(\lambda + z)^{\alpha+1}}\, dz = 1 - \left(\frac{\lambda}{\lambda + x}\right)^\alpha.$$

By solving equation $1 - \left(\frac{\lambda}{\lambda+x}\right)^\alpha = \beta$, we find its quantile function

$$F_\xi^{-1}(\beta) = \lambda\left[(1-\beta)^{-1/\alpha} - 1\right], \quad 0 < \beta < 1.$$

The mean of ξ is

$$E[\xi] = \frac{\lambda}{\alpha - 1}, \quad \alpha > 1$$

and is infinite if $\alpha \le 1$. The variance of ξ is

$$\mathrm{Var}(\xi)] = \frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2,$$

and is infinite if $\alpha \le 2$.

- ξ has the Burr distribution with parameters $\alpha > 0$, $\lambda > 0$, and $\gamma > 0$ if it has PDF

$$\rho_\xi(x) = \frac{\gamma\alpha\lambda^\alpha x^{\gamma-1}}{(\lambda + x^\gamma)^{\alpha+1}}, \quad x > 0. \tag{10}$$

In this case we write $\xi \sim Burr(\alpha, \lambda, \gamma)$. Burr distribution has CDF

$$F_\xi(x) = 1 - \left(\frac{\lambda}{\lambda + x^\gamma}\right)^\alpha, \quad x > 0.$$

By solving equation $1 - \left(\frac{\lambda}{\lambda+x^\gamma}\right)^\alpha = \beta$, we find quantile function of the Burr distribution

$$F_\xi^{-1}(\beta) = \left[\lambda(1-\beta)^{-1/\alpha} - \lambda\right]^{1/\gamma}, \quad 0 < \beta < 1.$$

Pareto distribution is the special case of Burr distribution with $\gamma = 1$.

- ξ has the generalized Pareto distribution with parameters $\alpha > 0$, $\delta > 0$, and $k > 0$ if it has PDF

$$\rho_\xi(x) = \frac{\Gamma(\alpha+k)\delta^\alpha}{\Gamma(\alpha)\Gamma(k)} \frac{x^{k-1}}{(\delta+x)^{\alpha+k}}, \quad x > 0. \tag{11}$$

The mean of ξ is

$$E[\xi] = \frac{d\Gamma(\alpha-1)\Gamma(k+1)}{\Gamma(\alpha)\Gamma(k)}, \quad \alpha > 1,$$

and is infinite if $\alpha \leq 1$. The variance exists if $\alpha > 2$, and is equal to $E[\xi^2] - (E[\xi])^2$, where

$$E[\xi^2] = \frac{d^2\Gamma(\alpha-2)\Gamma(k+2)}{\Gamma(\alpha)\Gamma(k)}, \quad \alpha > 2.$$

- ξ has the Weibull distribution with parameters $c > 0$ and $\gamma > 0$ if it has PDF

$$\rho_\xi(x) = c\gamma x^{\gamma-1} e^{-cx^\gamma}, \quad x > 0. \tag{12}$$

In this case we write $\xi \sim W(c, \gamma)$. Weibull distribution has CDF

$$F_\xi(x) = 1 - e^{-cx^\gamma}, \quad x > 0.$$

If $0 < \gamma < 1$, the upper tail of Weibull distribution $P[\xi > x] = e^{-cx^\gamma}$ decays faster than that for Pareto distribution (for which $P[\xi > x] = \left(\frac{\lambda}{\lambda+x}\right)^\alpha$) but slower than that for exponential distribution (for which $P[\xi > x] = e^{-\lambda x}$). This makes Weibull distribution a very flexible distribution, which is extensively used as a model for losses in insurance. By solving equation $1 - e^{-cx^\gamma} = \alpha$, we find quantile function

$$F_\xi^{-1}(\alpha) = \left(-\frac{\log(1-\alpha)}{c}\right)^{1/\gamma}, \quad 0 < \alpha < 1.$$

The mean of ξ is

$$E[\xi] = c^{-1/\gamma}\Gamma\left(\frac{1+\gamma}{\gamma}\right).$$

The variance is $E[\xi^2] - (E[\xi])^2$, where

$$E[\xi^2] = c^{-2/\gamma}\Gamma\left(\frac{2+\gamma}{\gamma}\right).$$

1.7 Independence

The notion of independence is one of the most important in Probability Theory. Intuitively we would like to call two events or random variables independent if there is no mutual dependency. For example, if we toss a coin (or roll a die) twice, the outcomes are, intuitively, independent from each other.

Two events A and B are called independent if

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

This property is consistent with the intuition of independence.

For random variables we have the following definition.

Definition 1.1. Random variables ξ and η are called *independent* if the events

$$\begin{aligned} A &= \{\omega \in \Omega : a < \xi(\omega) < b\} \\ B &= \{\omega \in \Omega : c < \eta(\omega) < d\} \end{aligned}$$

are independent for all real numbers a, b, c, d.

However, it is not sufficient to define independence only for pairs of events or random variables. It is possible that all the pairs of events (A, B), (A, C) and (B, C) are independent, but the triple (A, B, C) is not *mutually independent*.

Example 1.5. Consider a non-traditional dice with four faces. We number the first three faces with numbers 1, 2 and 3 respectively, and on the fourth face we put all the three numbers 1, 2 and 3. Now let us through the dice, and let

A be the event that 1 is on the face down;
B be the event that 2 is on the face down;
C be the event that 3 is on the face down.

Simple logic says that A depends on (B, C), since if B and C happen simultaneously then A happens too with probability one. But one can easily check that all the pairs (A, B), (A, C) and (B, C) are independent.

Definition 1.2. A collection of events are called *mutually independent*, if for every finite subset $A_1, \ldots, A_n$ of the collection we have

$$P(A_1 \cap \cdots \cap A_n) = \Pi_{s=1}^{n} P(A_i).$$

Similarly, we may define mutual independence for random variables.

Definition 1.3. Random variables $\xi_1, \ldots, \xi_n$ are called *mutually independent* if the events

$$A_1 = \{\omega \in \Omega : a_1 < \xi_1(\omega) < b_1\}$$
$$\cdots\cdots$$
$$A_n = \{\omega \in \Omega : a_n < \xi_n(\omega) < b_n\}$$

are mutually independent for all real numbers $\{a_k, b_k\}_{k=1}^n$. An infinite set of random variables $\{\xi_\alpha\}$ is called mutually independent if any finite subset $\{\xi_{\alpha_1}, \ldots, \xi_{\alpha_n}\}$ is mutually independent.

This definition can be reformulated in terms of the joint probability distribution. The random variables $\xi_1, \ldots \xi_d$ are mutually independent if and only if their joint distribution (i.e. the distribution of the random vector $(\xi_1, \ldots, \xi_d)$) is the product of the distribution functions of ξ_k's:

$$F_{\xi_1,\ldots,\xi_d}(x_1, \ldots, x_d) = \prod_{k=1}^d F_{\xi_k}(x_k).$$

If the variables $\xi_1, \ldots, \xi_d$ have probability densities, then they are mutually independent if and only if the random vector X also has a density which can be factorised as $\rho_X(x_1, \ldots, x_d) = \prod_{k=1}^d \rho_{\xi_k}(x_k)$.

For independent random variables ξ and η we have

$$E\big[f(\xi)g(\eta)\big] = E[f(\xi)]E[g(\eta)]$$

for any functions f and $g : \mathbb{R} \to \mathbb{R}$ (such that $f(\xi)$ and $g(\eta) \in L^1$). Indeed,

$$\begin{aligned} E\big[f(\xi)g(\eta)\big] &= \int_{\mathbb{R}^2} f(x)g(y)\rho_{\xi,\eta}(x,y)dxdy = \int_{\mathbb{R}^2} f(x)g(y)\rho_\xi(x)\rho_\eta(y)dxdy \\ &= \int_{-\infty}^{+\infty} f(x)\rho_\xi(x)dx \int_{-\infty}^{+\infty} \rho_\eta(y)g(y)dy = E[f(\xi)]E[g(\eta)] \end{aligned}$$

More generally, for mutually independent $\xi_1, \ldots, \xi_n$ we have

$$E\Big[\prod_{k=1}^n f_k(\xi_k)\Big] = \prod_{k=1}^n E[f_k(\xi_k)]$$

for any functions $f_k : \mathbb{R} \to \mathbb{R}$ (such that $f_k(\xi_k) \in L^1$).

This is a very convenient property. For instance, it allows us to prove that

$$\mathrm{Cov}(\xi, \eta) := E\big[(\xi - E[\xi])(\eta - E[\eta])\big] = E\big[\xi - E[\xi]\big]E\big[\eta - E[\eta]\big] = 0$$

for independent ξ and η. So the covariance could serve as a proxy measure for a degree of independence: the closer the covariance (or correlation) is to zero, the less dependent are the random variables. However, it is not a very good measure, since there exists dependent random variables with zero covariance.

For any random variables $\xi, \eta \in L^2$ such that $\mathrm{Cov}(\xi, \eta) = 0$ (so, in particular, for any independent ξ and η) we have

$$\mathrm{Var}(\xi + \eta) = \mathrm{Var}(\xi) + 2\mathrm{Cov}(\xi, \eta) + \mathrm{Var}(\eta) = \mathrm{Var}(\xi) + \mathrm{Var}(\eta).$$

More generally,

$$\mathrm{Var}\Big(\sum_{k=1}^{n} \xi_k\Big) = \sum_{k=1}^{n} \mathrm{Var}(\xi_k)$$

for all $\xi_1, \ldots, \xi_n \in L^2$ provided $\mathrm{Cov}(\xi_j, \xi_k) = 0$ for all $j \neq k$ (in particular if the random variables are mutually independent). Random variables with zero covariance are called *uncorrelated*.

Note that random variables that are independent and identically distributed are often denoted i.i.d..

1.8 Conditional Probability and Expectation

Sometimes, while estimating the probability of some event A, we can use the information that some other event B happened.

Example 1.6. An insurance company, which provides regular payments while a client is ill and unable to work, needs to estimate the probability of event $A = \{$a new customer will be ill during next year$\}$. They have a statistics, according to which, out of 100 current customers, 20 was ill at least once during the last year (10 of them smokes) and 80 was not ill during the last year (20 of them smokes). Based on this, they can estimate $P(A) = \frac{20}{100} = 0.2$ (a ratio of the number of customers that was ill to the total number of customers). However, if they know that the new customer smokes, they can estimate the probability in question as $\frac{10}{10+20} = \frac{1}{3} > 0.2$ (a ratio of the number of smokers that was ill to the total number of smokers).

In general, the probability of event A, when event B is known to occur, is called *conditional probability of A given B*, denoted $P[A|B]$, and can be evaluated as

$$P[A|B] = \frac{P[A \cap B]}{P[B]}.$$

In the example above, with $B = \{\text{a customer is a smoker}\}$, $P(A \cap B) = \frac{10}{100}$, $P(B) = \frac{30}{100} = 0.3$, and $P[A|B] = \frac{P(A \cap B)}{P(B)} = \frac{0.1}{0.3} = \frac{1}{3}$.

Events A and B are independent, if and only if

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[A]P[B]}{P[B]} = P[A].$$

Intuitively, this means that information about event B does not have any influence on the probability of the event A.

Knowing $P[A]$, $P[B]$, and $P[A|B]$, one can estimate $P[B|A]$ as follows

$$P[B|A] = \frac{P[A \cap B]}{P[A]} = \frac{P[A|B]P[B]}{P[A]},$$

which is called the *Bayes' theorem*.

If $\{B_n,\, n = 1, 2, \dots\}$ is a finite or countably infinite partition of Ω (that is, $B_i \cap B_j = \emptyset$, $i \neq j$ and $\bigcup_n B_n = \Omega$), then, for any event A,

$$P[A] = \sum_n P[A \cap B_n] = \sum_n P[A|B_n]P[B_n]. \tag{13}$$

The last relation is called *the law of total probability*.

Example 1.7. A car insurance company, which classifies drivers as "new" and "experienced", wants to estimate the probability of the event $A = \{$a randomly selected driver will cause a car accident during next year$\}$. From their past data, they estimate that the probability of A is 0.3 for "new" drivers, and 0.05 for "experienced" ones. If 40% of their current customers are new, then

$$P[A] = P[A|N]P[N] + P[A|E]P[E] = 0.3 \cdot 0.4 + 0.05 \cdot 0.6 = 0.15$$

(here N and E are the events that the driver is "new" and "experienced", correspondingly).

A continuous version of the law of total probability (13) is

$$p_X(x) = \int_y p_{X,Y}(x, y)\, \mathrm{d}y = \int_y p_{X|Y}(x \mid y)\, p_Y(y)\, \mathrm{d}y, \tag{14}$$

where $p_X(x)$ and $p_Y(y)$ are densities of X and Y, $p_{X,Y}(x, y)$ is the joint density of X and Y, and $p_{X|Y}(x \mid y) := \frac{p_{X,Y}(x,y)}{p_Y(y)}$ is called a conditional density of X given Y.

Let A be any event with positive probability. Let I_A be the indicator function of A, that is, $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ otherwise. The conditional expectation of random variable X given A is denoted as $E(X|A)$ and defined as

$$E(X|A) := \frac{E(I_A X)}{P(A)}.$$

Intuitively, $E(X|A)$ is the average value of X given that event A happened.

For example, if Y is a discrete random variable which takes some value y with non-zero probability, then

$$E(X|Y = y) := \frac{E(I_{Y=y} X)}{P(Y = y)}.$$

We also define $E(X|Y)$ as a random variable, such that

$$E(X|Y)(\omega) = E(X|Y = Y(\omega)), \quad \forall \omega \in \Omega.$$

Fundamental formulas involving conditional expectation are law of total expectation

$$E(X) = E[E(X|Y)], \tag{15}$$

and law of total variance

$$Var(X) = E[Var(X|Y)] + Var(E(X|Y)), \tag{16}$$

where

$$Var(X|Y) = E(X^2|Y) - (E(X|Y))^2.$$

Example 1.8. Let X and Y be random variables taking values 0 and 1, such that

$$P[X = Y = 1] = P[X = Y = 0] = 0.4$$

and

$$P[X = 1, Y = 0] = P[X = 0, Y = 1] = 0.1.$$

Then $P[X = 0] = P[X = 1] = P[Y = 0] = P[Y = 1] = 0.5$, hence,

$$E(X) = E(Y) = 0 \cdot 0.5 + 1 \cdot 0.5 = 0.5$$

and

$$Var(X) = Var(Y) = (0^2 \cdot 0.5 + 1^2 \cdot 0.5) - (0.5)^2 = 0.25.$$

Now assume that we know that $Y = 0$. Then

$$P[X = 0|Y = 0] = \frac{P[X = Y = 0]}{P[Y = 0]} = \frac{0.4}{0.5} = 0.8,$$

$$P[X = 1|Y = 0] = \frac{P[X = 1, Y = 0]}{P[Y = 0]} = \frac{0.1}{0.5} = 0.2,$$

and

$$E[X|Y = 0] = 0 \cdot P[X = 0|Y = 0] + 1 \cdot P[X = 1|Y = 0] = 0.2.$$

Similarly,

$$E[X|Y = 1] = 0 \cdot P[X = 0|Y = 1] + 1 \cdot P[X = 1|Y = 1] = 0.8.$$

Hence,

$$E(E[X|Y]) = 0.2 \cdot 0.5 + 0.8 \cdot 0.5 = 0.5 = E[X],$$

in agreement with the law of total expectation (15).
Similarly we can calculate that

$$Var(X|Y = 0) = E(X^2|Y = 0) - (E(X|Y = 0))^2 = 0.2 - 0.2^2 = 0.16,$$

$$Var(X|Y = 1) = E(X^2|Y = 1) - (E(X|Y = 1))^2 = 0.8 - 0.8^2 = 0.16,$$

hence

$$E[Var(X|Y)] = 0.16.$$

Also,

$$Var(E[X|Y]) = 0.2^2 \cdot 0.5 + 0.8^2 \cdot 0.5 - (0.2 \cdot 0.5 + 0.8 \cdot 0.5)^2 = 0.09,$$

hence

$$E[Var(X|Y)] + Var(E(X|Y)) = 0.16 + 0.09 = 0.25 = Var(X),$$

in agreement with the law of total variance (16).

1.9 Stochastic processes

A random variable is a suitable model for a number produced by a single experiment, at a specified moment of time, for example "a temperature tomorrow at 14.00". If we are interesting how temperature will change in time, we actually need to consider the whole collection of random variables $\{X_t : t \in \mathcal{T}\}$, where X_t is the (random) temperature at a specified moment

t, and $\mathcal{T}$ is the set of all times we are interested in. We will call such family of random variables a *stochastic process* (or *random process*).

Formally, a stochastic process is a family of random variables $\{X_t : t \in \mathcal{T}\}$, where $\mathcal{T}$ is an arbitrary index set. For example, any random variable is a stochastic process with one-element set $\mathcal{T}$. But typically parameter t represents time, and the most common examples for $\mathcal{T}$ are $\mathcal{T} = \{0, 1, 2, \dots\}$ and $\mathcal{T} = \mathbb{R}$ (or $[0, \infty)$). In the first case the stochastic process is called *discrete time*, and is actually just a sequence of random variables; in the second case the random process is called *continuous time*.
There are many classifications of stochastic processes. One of the most basic is to classify them with respect to the time (index) set $\mathcal{T}$ and with respect to the *state space*. By definition the state space $\mathcal{S}$ is the set of possible values of a random process X_t.

Discrete state spaces with discrete time changes. Most typically $\mathcal{T}$ is $\{0, 1, 2, \dots\}$ in this case and the state space $\mathcal{S}$ is a discrete set. For example, a motor insurance company reviews the status of each customer yearly with respect to three possible levels of discount $\mathcal{S} = \{0, 10\%, 25\%\}$.

It is not necessary that $\mathcal{S}$ should be a set of numbers. For example, it may be the credit rating, or information from an applicant form. Typical examples of discrete state processes with discrete time change are Markov chains which are discussed later in this course.

Discrete state spaces with continuous time changes. In this case typically $\mathcal{T} = [0, \infty)$ and the state space $\mathcal{S}$ is a discrete set. For instance, an individual insurance policy holder can be classified as healthy, sick or dead. So $\mathcal{S} = \{\text{healthy, sick, dead}\}$. It is natural to take the time set as continuous as illness or death can happen at any time.

An important special case of this class are so-called *counting processes*. A process $(N_t)_{t\in[0,\infty)}$ is counting, if it is increasing and takes values in $\{0, 1, 2, \dots\}$. For instance, it can be the cumulative number of claims reported at random times.

Continuous state spaces with continuous time changes. Typically in this case $\mathcal{T}$ is $[0, \infty)$ or $\mathbb{R}$ and $\mathcal{S} = [0, \infty)$, $(0, \infty)$ or $\mathbb{R}$. For instance, it is natural to consider the exchange rate GBP/USD as a random process with the state space $(0, \infty)$ and continuous time.

Continuous state spaces with discrete time changes. The typical example of these is when an essentially continuously valued process such

as price or temperature is measured only at certain time intervals (days, months, quarters, years). For example, if we do not care about intra-day changing of the GBP/USD exchange rate, then we can consider this as a discrete-time process $\{X_0, X_1, X_2, \dots\}$ where X_i indicates the exchange rate in the morning of i-th day.

Mixed type processes. There are special types of continuous-time processes, with continuous or discrete state spaces, which have some specifically structured changes at predetermined times. For example, the market price of a coupon-paying bond changes at the deterministic times of the coupon payments, but it also changes randomly all the time before its maturity, due to the current situation in the market.

For every real-life process to be analysed, it is important to establish whether it is most usefully modelled using a discrete, a continuous, or a mixed time domain. Usually the choice of state space will be clear from the nature of the process being studied (as, for example, with the Healthy-Sick-Dead model), but whether a continuous or discrete time set is used will often depend on the specific aspects of the process which are of interest, and upon practical issues like the time points for which data are available. Continuous time and continuous state space stochastic processes, although conceptually more difficult than discrete ones, are often more convenient to deal with, in the same way as it is often easier to calculate an integral than to sum an infinite series.

Sample pathes. To determine a particular value for a general stochastic process $\{X_t : t \in \mathcal{T}\}$, we need to specify time $t \in \mathcal{T}$, and a particular realization $\omega \in \Omega$. From this perspective we can interpret a stochastic process as a function of two variables: t and ω. If we fix a particular state of nature $\omega \in \Omega$, we get a particular *realization* of stochastic process, which is a deterministic function from $\mathcal{T}$ to the state space $\mathcal{S}$. This function is called a *sample path* of the process.

For example, consider the exchange rate GBP/USD during a particular month. In advance, it is hard to predict the exact exchange rate at every moment of time during this month, so it is natural to model it as a random process. During the month, we can observe a realization of this process as a function of time, which is an example of a sample path. If we use the model of continuous time process, a sample path is a continuous function, defined for every moment t during the month. If we are interested only in exchange rates at (say) 9.00 every day, the suitable model is a discrete-time process, and the sample path is just a sequence of numbers. In this case we will sometimes refer to it as to *sample sequence* of the discrete-time process.

Describing a stochastic process. To describe the stochastic process $\{X_t : t \in \mathcal{T}\}$, we need to specify the joint distributions of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ for all $t_1, t_2, \ldots, t_n$ in $\mathcal{T}$ and all integers n. The collection of the joint distributions above is called *the family of finite dimensional probability distributions* (f.f.d. for short). To describe a stochastic process in practice, we will rarely give the exact formulas for its f.f.d., but will rather use some indirect intuitive descriptions. For example, take the familiar Bernoulli trials of consecutive tosses of a fair coin. A sequence of i.i.d. Bernoulli variables $(\xi_t)_{t=1}^{\infty}$ is a stochastic process, and its f.f.d. is fully determined by this description. Indeed, for any sequences of times $t_1, t_2, \ldots, t_n$ in $\mathcal{T} = \{0, 1, 2, \ldots\}$ and "results" $x_1, x_2, \ldots, x_n$ in $\mathcal{S}$, we are able to estimate the probability $P(\xi_{t_1} = x_1 \cup \xi_{t_2} = x_2 \cup \cdots \cup \xi_{t_n} = x_n)$, and it is equal to 2^{-n}.

Stationarity. In the example above, the probability to "meet" any sequences of results $x_1, x_2, \ldots, x_n$ does not depend on times $t_1, t_2, \ldots, t_n$. This means that the statistical properties of the process remain unchanged over the time, which is intuitively obvious for tosses of a fair coin. If, however, a stochastic process describes the tomorrow's temperature, it would be reasonable to expect a lower temperature during the morning than at noon.

Formally, a stochastic process $\{X_t : t \in \mathcal{T}\}$ is said to be *stationary*, or *strictly stationary*, if for all integers n and all $t, t_1, t_2, \ldots, t_n$ in $\mathcal{T}$ the joint distributions of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ and $X_{t+t_1}, X_{t+t_2}, \ldots, X_{t+t_n}$ coincides. Substituting $n = 1$, we can see that, in particular, all distribution functions $\{F_{X_t}(x) : t \in \mathcal{T}\}$ are the same for all t. Consequently, all parameters depending only on distribution (such as mean and variance), if they exist, also do not change over time.

Strict stationarity is a strong requirement which may be difficult to test fully in real life. Actually, a much weaker condition, known as weak stationarity, is often already very useful for applications. A stochastic process $\{X_t : t \in \mathcal{T}\}$ is said to be *weakly stationary*, if the mean of the process, $m(t) = E[X_t]$, is constant, and the covariance of the process, $C(s,t) = E[(X_s - m(s))(X_t - m(t))]$, depends only on the time difference $t - s$. Obviously, any strictly stationary stochastic process with finite mean and variance is also weakly stationary.

Increments. If stochastic process $\{X_t : t \in \mathcal{T}\}$ describes the tomorrow's temperature, we are interested in the value of X_t itself. Sometimes, however, the dynamic of how the value changes over the time is much more interesting. For example, if X_t is the price of a stock share, a "forecast" $X_t = 100$ provides almost no information by itself. If the current stock price is $X_0 = 60$, the

above forecast is very optimistic; if, however, $X_0 = 120$, it is pessimistic. What we are really interested in is the price dynamics, whether it increases or decreases and how much.

Formally, an *increment* of the stochastic process $\{X_t : t \in \mathcal{T}\}$ is the quantity $X_{t+u} - X_t$, $u > 0$. Many processes are the most naturally defined through their increments. For example, let X_t be total amount of money on a bank account of a person A at the first day of month t. Assume that monthly salary of A is a fixed amount C, and monthly expenses Y_t are random. Then the stochastic process X_t is naturally defined through its increments $X_{t+1} - X_t = C - Y_t$.

In the above example, the process X_t is not stationary (even weakly) unless $Y_t \equiv C$, $\forall t$. For example, if $EY_t < C$, $\forall t$, the total amount of money on the bank account increases (on average) with time. However, if Y_t are identically distributed, the *rate* of growth of X_t remains unchanged over the time. Such processes are said to have *stationary increments*. If, moreover, monthly expenses Y_t are (jointly) independent, the rate of growth of X_t does not depend of its history, and we say that X_t has *independent increments*.

Formally, a stochastic process $\{X_t : t \in \mathcal{T}\}$ has stationary (or time-homogeneous) increments, if for every $u > 0$ the increment $Z_t = X_{t+u} - X_t$ is a stationary process; a process $\{X_t : t \in \mathcal{T}\}$ is said to have independent increments if for any $a, b, c, d \in \mathcal{T}$ such that $a < b < c < d$, random variables $X_a - X_b$ and $X_c - X_d$ are independent.

1.10 Summary

A random variable is a measurable function $\xi : \Omega \to \mathbb{R}$ from probability space Ω to real line.

For a random variable ξ, its cumulative distribution function (cdf) is $F_\xi(x) := P(\xi \le x)$. If $F_\xi(x)$ can be represented as

$$F_\xi(x) = \int_{-\infty}^{x} \rho_\xi(z)\, dz,$$

for some non-negative function ρ_ξ, the latter is called the probability density function (pdf) of ξ.

The expectation of a random variable ξ is defined as $E[\xi] \equiv \int_\Omega \xi\, dP$. It can be calculated as

$$E[\xi] = \int_{-\infty}^{\infty} x\, dF_\xi(x) = \int_{-\infty}^{\infty} x\rho_\xi(x)\, dx.$$

The variance of a random variable ξ is defined by $\mathrm{Var}(\xi) := E[(\xi - E[\xi])^2] \equiv E[\xi^2] - E[\xi]^2$. For two random variables ξ, η, their covariance is defined by $\mathrm{Cov}(\xi, \eta) := E\big[(\xi - E[\xi])(\eta - E[\eta])\big] \equiv E[\xi\eta] - E[\xi]E[\eta]$.

Two events A and B are called independent if $P(A \cap B) = P(A)P(B)$. Random variables ξ and η are called *independent* if the events $A = \{\omega \in \Omega : a < \xi(\omega) < b\}$ and $B = \{\omega \in \Omega : c < \eta(\omega) < d\}$ are independent for all real numbers a, b, c, d. If ξ and η are independent, $\mathrm{Cov}(\xi, \eta) = 0$, but the converse is not always true.

The conditional probability of A given B, denoted $P[A|B]$, can be evaluated as $P[A|B] = \frac{P[A \cap B]}{P[B]}$.

A stochastic process is a family of random variables indexed in time, $\{X_t : t \in \mathcal{T}\}$. The time set $\mathcal{T}$ can be discrete or continuous, as can the state space $\mathcal{S}$ in which the variables take their values.

Stochastic process can be roughly classified into the following groups:

- Discrete $\mathcal{S}$ and discrete $\mathcal{T}$;
- Continuous $\mathcal{S}$ and discrete $\mathcal{T}$;
- Discrete $\mathcal{S}$ and continuous $\mathcal{T}$;
- Continuous $\mathcal{S}$ and continuous $\mathcal{T}$; or

- Mixed processes.

A stochastic process $\{X_t : t \in \mathcal{T}\}$ is said to be *stationary*, if for all integers n and all $t, t_1, t_2, \ldots, t_n$ in $\mathcal{T}$ the joint distributions of $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ and $X_{t+t_1}, X_{t+t_2}, \ldots, X_{t+t_n}$ coincides. It is called *weakly stationary*, if the mean of the process, $m(t) = E[X_t]$, is constant, and the covariance of the process, $C(s,t) = E[(X_s - m(s))(X_t - m(t))]$, depends only on the time difference $t - s$.

An *increment* of the stochastic process $\{X_t : t \in \mathcal{T}\}$ is the quantity $X_{t+u} - X_t$, $u > 0$. A stochastic process has stationary (or time-homogeneous) increments, if for every $u > 0$ the increment $Z_t = X_{t+u} - X_t$ is a stationary process; a process $\{X_t : t \in \mathcal{T}\}$ is said to have independent increments if for any $a, b, c, d \in \mathcal{T}$ such that $a < b < c < d$, random variables $X_a - X_b$ and $X_c - X_d$ are independent.

1.11 Questions

1. A sample space consists of five elements $\Omega = \{a_1, a_2, a_3, a_4, a_5\}$. For which of the following sets of probabilities does the corresponding triple $(\Omega, \mathcal{A}, P)$ become a probability space? Why?

   (a) $p(a_1) = 0.3; p(a_2) = 0.2; p(a_3) = 0.1; p(a_4) = 0.1; p(a_5) = 0.1;$

   (b) $p(a_1) = 0.4; p(a_2) = 0.3; p(a_3) = 0.1; p(a_4) = 0.1; p(a_5) = 0.1:$

   (c) $p(a_1) = 0.4; p(a_2) = 0.3; p(a_3) = 0.2; p(a_4) = -0.1; p(a_5) = 0.2.$

2. Let X be a random variable from the continuous uniform distribution, $X \sim U(0.5, 1.0)$. Starting with the probability density function, derive expressions for the cumulative distribution function, expectation and variance of X.

3. Assets A and B have the following distribution of returns in various states:

State	Asset A	Asset B	Probability
1	10%	-2%	0.2
2	8%	15%	0.2
3	25%	0%	0.3
4	-14%	6%	0.3

   Show that the correlation between the returns on asset A and asset B is equal to -0.3830.

4. Formalise Example 1.5 as $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, $P(\omega_1) = P(\omega_2) = P(\omega_3) = P(\omega_4) = 1/4$ and

   $$A := \{\omega_1, \omega_4\}, \quad B := \{\omega_2, \omega_4\}, \quad C := \{\omega_3, \omega_4\}.$$

   Prove that the pairs (A, B), (A, C) and (B, C) are independent, but the triple (A, B, C) is not mutually independent according to Definition 1.2.

5. You intend to model the maximum daily temperature in your office as a stochastic process. What time set and state space would you use?

Chapter 2
Claim size estimation in insurance and reinsurance

2.1 Basic principles of insurance risk modelling

In general, for a risk to be insurable, the following conditions must be satisfied:

- The policyholder must have an interest in the risk being insured, to distinguish between insurance and a bet; and
- a risk must be of a financial and reasonably quantifiable nature

In addition, the following conditions are desirable:

- Individual risk events should be independent of each other.
- The probability of the event should be relatively small. In other words, an event that is nearly certain to occur is not conducive to insurance.
- Large numbers of potentially similar risks should be pooled to reduce the variance and achieve more certainty.
- There should be an ultimate limit on the liability undertaken by the insurer.
- Moral hazards should be eliminated as far as possible because these are difficult to quantify, result in selection against the insurer and lead to unfairness in treatment between one policyholder and another.

However, the desire for income means that an insurer will usually be found to provide cover when these ideal criteria are not met.

Other characteristics that most general insurance products share are:

- Cover is normally for a fixed period, most commonly one year, after which it has to be renegotiated. There is normally no obligation on insurer or insured to continue the arrangement thereafter, although in most cases a need for continuing cover may be assumed to exist.
- Claims are not of fixed amounts, and the amount of loss as well as the fact needs to be proved before a claim can be settled.
- A claim occurring does not bring the policy to an end.

- Claims may occur at any time during the policy period.

Although there is normally a contractual obligation on the policyholder to report a claim to the insurer as quickly as possible, notification may take some time if the loss is not evident immediately. Settlement of the claim may take a long time if protracted legal proceedings are needed or if it is not straightforward to determine the extent of the loss. However, from the moment of the event giving rise to the claim the ultimate settlement amount is a liability of the insurer. Estimating the amounts of money that need to be reserved to settle these liabilities is one of the most important areas of actuarial involvement in general insurance. Classes of insurance in which claims tend to take a long time to settle are known as long-tail. Those which tend to take a short time to settle are known as short-tail, although the dividing line between the two categories is not always distinct.

Many forms of non-life insurance can be regarded as short-term contracts, for example motor insurance Some forms of life insurance also fall into this category, for example group life and one-year term assurance policies.

A short-term insurance contract can be defined as having the following attributes:

- The policy lasts for a fixed, and relatively short, period of time, typically one year.
- The insurance company receives from the policyholder(s) a premium.
- In return, the insurer pays claims that arise during the term of the policy.

At the end of the policy's term the policyholder may or may not renew the policy; if it is renewed, the premium payable by the policyholder may or may not be the same as in the previous period.

The insurer may choose to pass part of the premium to a reinsurer; in return, the reinsurer will reimburse the insurer for part of the cost of the claims during the policy's term according to some agreed formula.

An important feature of a short-term insurance contract is that the premium is set at a level to cover claims arising during the (short) term of the policy only. This contrasts with life assurance policies, where mortality rates increasing with age mean that the (level) annual premium in the early years would be more than sufficient to cover the expected claims in the early years. The excess amount would then be accumulated as a reserve to be used in the later years, when the premium on its own would be insufficient to meet the expected cost of claims.

The profit of any company, including insurance company, during a certain time period, e.g. a month, can be calculated as income of the company during this time period minus its expenses/losses.

The profit of insurance company is due to premiums paid by customers. At the beginning of a month, a company knows the number of customers and what premiums they are paying. Of course, the company cannot predict the number of new customers coming during next month as well as the number of customers who stop paying premiums. However, because the premium paid by every individual customer is usually small, and the number of customers changes not too much during each month, these are the minor issues. Hence, the company can estimate its profit for next month with good accuracy.

The expenses/losses for the insurance company can be divided into two parts: expenses to cover the claims and other expenses. Other expenses, such as taxes, staff salaries, etc., can also be predicted. The main problem for any insurance company is to estimate expenses/losses to cover the claims.

If there will be N claims during a month with sizes $X_1, X_2, \ldots, X_N$, then the total losses to cover all claims are

$$X_1 + X_2 + \cdots + X_N.$$

Here, the number N of claim is unpredictable, hence it is modelled as a random variable. The sizes $X_1, X_2, \ldots, X_N$ of claims are random variables as well. Our first key assumption is that the number of claims and the sizes of claims are independent random variables and can be studied separately. To justify this assumption, let us consider motor insurance as an example. A prolonged spell of bad weather may have a significant effect on claim numbers but little or no effect on the distribution of individual claim amounts. On the other hand, inflation may have a significant effect on the cost of repairing cars, and hence on the distribution of individual claim amounts, but little or no effect on claim numbers.

Our second key assumption is that sizes $X_1, X_2, \ldots, X_N$ of claims are independent and identically distributed (i.i.d). If the distribution of X_i is known, the model is complete and can be used to answer various questions of central importance for insurance company, for example, with what probability the claims above certain level arrive.

In practise, however, the claim distribution is rarely known. The insurance company usually has a sequence

$$x_1, x_2, \ldots, x_n$$

of past claims and use this sequence to estimate the claim distribution. In most cases, this process works as follows.

1. The company assumes that claim distribution belongs to certain family, but with unknown parameters. For example, it may assume that claims follow normal distribution with unknown mean and variance.

2. The company estimate unknown parameters to fit the data of past claims $x_1, x_2, \ldots, x_n$ as good as possible.

3. Steps 1-2 can then be repeated for different families of distributions. Then there are "goodness of fit" tests in statistics, for example ξ^2 test, which allow to select family which fits the data best.

We next focus on step 2: estimating unknown parameters of some given family of distributions.

2.2 Method of moments

For a random variable X and positive integer j, the j-th moment of X is $E[X^j]$. For example, the first moment ($j = 1$) is just expectation $m = E[X]$. The difference $X - m$ has expectation $E[X - m] = E[X] - m = 0$ and is called centralized random variable. The second moment $E[(X - m)^2]$ of centralized random variable $X - m$ is just a variance of X, usually denoted as σ^2, where σ is standard deviation. The ratio $\frac{X-m}{\sigma}$ has mean 0 and standard deviation 1 and is called standardized random variable. The third moment

$$E\left[\left(\frac{X-m}{\sigma}\right)^3\right]$$

of $\frac{X-m}{\sigma}$ is called the *skewness* of X, while the forth moment

$$E\left[\left(\frac{X-m}{\sigma}\right)^4\right]$$

is known as *kurtosis* of X.

Sometimes it is convenient to calculate moments using moment generating function. For a random variable X, its moment generating function is

$$M_X(t) := E\left[e^{tX}\right].$$

If we have $M_X(t)$, we can find n-th moment of X as n-th derivative of $M_X(t)$ at 0:

$$E\left[X^n\right] = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}$$

If X and Y are independent random variables and $S = X + Y$, then

$$M_S(t) = M_X(t) \cdot M_Y(t),$$

which is a very convenient property.

If X belongs to certain family of distributions with r parameters $a_1, a_2, \ldots, a_r$, its j-th moment can be explicitly calculated as a function of the parameters, that is

$$E[X^j] = f_j(a_1, a_2, \ldots, a_r), \quad j = 1, 2, \ldots$$

If past data $x_1, x_2, \ldots, x_n$ of i.i.d. realizations of X are available, its j-th moment can also be estimated from data as

$$E[X^j] \approx \frac{1}{n} \sum_{i=1}^{n} x_i^j, \quad j = 1, 2, \ldots$$

Now, the method of moments suggests to select parameters $a_1, a_2, \ldots, a_r$ in such a way that first r moments estimated from the data match the first r moments estimated from the formulas for the distribution, that is,

$$\frac{1}{n} \sum_{i=1}^{n} x_i^j = f_j(a_1, a_2, \ldots, a_r), \quad j = 1, 2, \ldots, r. \tag{17}$$

If we denote

$$m_j = \frac{1}{n} \sum_{i=1}^{n} x_i^j, \quad j = 1, 2, \ldots, r$$

to be the moments estimated from the data, then (17) simplifies to

$$m_j = f_j(a_1, a_2, \ldots, a_r), \quad j = 1, 2, \ldots, r. \tag{18}$$

This is a system of r equations with r unknowns which often has a unique solution.

We next consider concrete examples with specific families of distributions.

- Assume that X follow the exponential distribution with parameter λ, see (5). In this case, we have only 1 parameter, so it suffices to consider the 1-st moment only, that is, expectation. The expectation $E[X]$ of exponential distribution is $1/\lambda$, and (17) reduces to

  $$\frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{\lambda},$$

  which results in the estimate

  $$\lambda = \frac{n}{\sum_{i=1}^{n} x_i}.$$

- Assume that X follow a normal distribution with parameters μ and σ, see (6). Because we have 2 parameters, it suffices to consider 2 moments. The first 2 moments of normal distribution are $E[X] = m$ and $E[X^2] = \sigma^2 + m^2$. Hence, parameters μ and σ can be found from system of equations

$$\frac{1}{n}\sum_{i=1}^{n} x_i = m, \quad \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma^2 + m^2.$$

  The solution is

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}.$$

- If X follow a log-normal distribution with parameters μ and σ, that is, $\log(X) \sim N(\mu, \sigma^2)$, and past data $x_1, \ldots, x_n$ are available, then the logarithms $y_i = \log x_i$ of past data are i.i.d sample from normal distribution $N(\mu, \sigma^2)$, and its parameters can be estimated from these data exactly as above:

$$m = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2}.$$

- If X follow a gamma distribution (8) with parameters $\alpha > 0$ and $\lambda > 0$, we need $r = 2$ moments, estimated from the data as

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \tag{19}$$

  Then (18) reduces to

$$m_1 = \frac{\alpha}{\lambda}, \quad m_2 = \frac{\alpha}{\lambda^2} + \left(\frac{\alpha}{\lambda}\right)^2,$$

  and the solution is

$$\alpha = \frac{m_1^2}{m_2 - m_1^2}, \quad \lambda = \frac{m_1}{m_2 - m_1^2}.$$

- If X follow a Pareto distribution (9) with parameters $\alpha > 0$ and $\lambda > 0$, then two moments exists if $\alpha > 2$, and, in this case, system (18) reduces to

$$m_1 = \frac{\lambda}{\alpha - 1}, \quad m_2 = \frac{\alpha\lambda^2}{(\alpha - 1)^2(\alpha - 2)} + m_1^2,$$

where m_1 and m_2 are defined in (19). The solution is

$$\alpha = \frac{2m_2 - 2m_1^2}{m_2 - 2m_1^2}, \quad \lambda = \frac{m_1 m_2}{m_2 - 2m_1^2},$$

provided that $m_2 - 2m_1^2 > 0$.

For other families of distributions, like Burr distribution (10), the generalized Pareto distribution (11) or Weibull distribution (12), explicit expressions for moments may be too complicated to solve system (18) analytically. However, it may be solved numerically using appropriate computer software.

2.3 Method of maximum likelihood

Method of moments is not always appropriate because, for some families of distributions, moments may not exist for some parameters. An alternative, and even more intuitive method, is the method of maximal likelihood. For simplicity, we first introduce this method on a very simple example. Imagine a coin which can show 2 outcomes, head H and tail T, with probabilities p and $1-p$, respectively, but assume that parameter p is unknown. To estimate p, we have tossed this coin 10 times, and the outcomes are $H, H, T, H, T, H, H, H, T, H$. The method of moments is not applicable here, because the outcomes are not even numerical values. Instead, let us just calculate the probability of getting exactly this sequence. The probability of getting the first head is p, the second head is p, the next tail is $1-p$, and so on. All 10 tosses are independent, hence the probability of getting this sequence is the product of corresponding probabilities

$$p \cdot p \cdot (1-p) \cdot p \cdot (1-p) \cdot p \cdot p \cdot p \cdot (1-p) \cdot p = p^7(1-p)^3.$$

For example, if $p = 0$ or $p = 1$, then $p^7(1-p)^3 = 0$, hence we would *never* get this sequence. If $p = 1/2$, then $p^7(1-p)^3 = 1/1024$, so getting this sequence is unlikely but possible. However, can we do better? Does there exist p for which such sequence is not so unlikely? After all, why not select p for which this sequence is as likely as possible. In other word, we want to find p which maximizes the function $L(p) = p^7(1-p)^3$. To find such p, we can differentiate and solve equation $L'(p) = 0$.

In fact, the following trick can be used to simplify calculations. Because the logarithm is an increasing function, maximizing $L(p)$ is equivalent to maximizing the logarithm $\log(L(p))$. In our case, $\log(L(p)) = 7\log p + 3\log(1-p)$. The derivative is $\frac{7}{p} - \frac{3}{1-p}$, which is equal to 0 if $\frac{7}{p} = \frac{3}{1-p}$, or $7(1-p) = 3p$,

or $p = 0.7$. With this parameter, the result of the experiment which we actually observed is as likely as it possibly can. This is the idea of the method of maximum likelihood.

More generally, let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d random variable whose distribution belongs to some family of discrete distributions with parameter θ. Given historical data $x_1, x_2, \ldots, x_n$ we actually obtained, we can ask how "likely" it was to get this data, and the answer is

$$L(\theta) = \prod_{i=1}^{n} P(X_i = x_i \,|\, \theta),$$

where $P(X_i = x_i \,|\, \theta)$ is the conditional probability of event $X_i = x_i$ given θ. We then find θ which maximizes $L(\theta)$, or, equivalently, maximizes its logarithm

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log[P(X_i = x_i \,|\, \theta)].$$

The optimal $\hat{\theta}$ can be found from equation

$$\frac{d}{d\theta} l(\hat{\theta}) = 0. \tag{20}$$

In fact, θ can be vector of r parameters. Then $\frac{d}{d\theta}$ in (20) should be understood as r partial derivatives, and (20) reduces to system of r equations with r unknowns.

Given sample $x_1, x_2, \ldots, x_n$ from a continuous distribution with density $f(x \,|\, \theta)$ which depends on vector θ of parameters, the likelihood function $L(\theta)$ takes the form

$$L(\theta) = \prod_{i=1}^{n} f(x_i \,|\, \theta),$$

and its logarithm is

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log[f(x_i \,|\, \theta)].$$

The optimal $\hat{\theta}$ maximizing this function can be found from the same system of equations (20).

We next consider concrete examples with specific families of distributions.

- Assume that X follow the exponential distribution with parameter λ, that is, has density given by (5). Then

$$l(\lambda) = \sum_{i=1}^{n} \log[f(x_i \,|\, \lambda)] = \sum_{i=1}^{n} \log[\lambda e^{-\lambda x_i}] = \sum_{i=1}^{n} (\log[\lambda] - \lambda x_i) = n \log[\lambda] - \lambda \sum_{i=1}^{n} x_i,$$

and (20) reduces to

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0,$$

from which we find $\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$. Note that in this case the result is the same as with method of moments.

- Assume that X follow a normal distribution with parameters μ and σ^2, see (6). Then

$$l(\mu, \sigma^2) = \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

and (20) reduces to

$$\frac{d}{d\mu} l(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

and

$$\frac{d}{d\sigma^2} l(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

from which we find that

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

the same solution as with method of moments.

- If X follow a log-normal distribution with parameters μ and σ^2, then logarithms $y_i = \log x_i$ are i.i.d sample from normal distribution $N(\mu, \sigma^2)$, hence:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \log x_i, \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (\log x_i - \mu)^2,$$

- If X follow a gamma distribution (8) with parameters $\alpha > 0$ and $\lambda > 0$, then

$$l(\alpha, \lambda) = \sum_{i=1}^{n} \log \left[\frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \right]$$

or

$$l(\alpha, \lambda) = n\alpha \log[\lambda] - n \log[\Gamma(\alpha)] + (\alpha - 1) \sum_{i=1}^{n} \log[x_i] - \lambda \sum_{i=1}^{n} x_i.$$

Then

$$\frac{d}{d\lambda} l(\alpha, \lambda) = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i = 0,$$

and

$$\frac{d}{d\alpha} l(\alpha, \lambda) = n \log[\lambda] - n \frac{d}{d\alpha} \log[\Gamma(\alpha)] + \sum_{i=1}^{n} \log[x_i] = 0.$$

From the first equation, $\lambda = \frac{n\alpha}{\sum_{i=1}^{n} x_i}$. We can substitute this into the second equation and solve it for α numerically. We remark that in this case the solution from maximum likelihood method is *different* from the one from method of moments.

For other families of distributions the method is the same, and the resulting equations, if impossible to solve analytically, can be solved numerically using appropriate computer software.

2.4 Method of percentiles

Let $x_1, x_2, \ldots x_n$ be i.i.d sample from some distribution. Given $\alpha \in (0, 1)$, we would like to find an estimate for the number x such that $F_X(x) = P[X \leq x] = \alpha$, where X is a random variable with this distribution. This can be done using the following procedure. First, let j be the smallest integer greater than $n\alpha$. Then sort the sequence $x_1, x_2, \ldots x_n$ is non-decreasing order, and then the j-th number is the answer. We denote this answer as $q(\alpha, x_1, \ldots x_n)$.

Example. Let the data be $1, 5, 6, 4, 3, 5, 6$ and $\alpha = 1/4$. Then $n = 7$, $n\alpha = 7/4$, and the smallest integer greater than $7/4$ is $j = 2$. Then sort the data in non-decreasing order to get $1, 3, 4, 5, 5, 6, 6$. The 2-nd number in this sequence is 3, and this is the answer.

Now consider family of distributions witch depends on vector of parameters λ and has cumulative distribution function $F(x, \lambda)$. We assume that F is strictly increasing, as has an inverse function defined on $(0, 1)$:

$$F^{-1}(\alpha, \lambda) = \inf\{x | F(x, \lambda) > \alpha\}, \quad 0 < \alpha < 1,$$

If there are r parameters, let us select r different numbers $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_r < 1$ and require that $F^{-1}(\alpha_i, \lambda)$ "agree with data", that is

$$F^{-1}(\alpha_i, \lambda) = q_i, \quad i = 1, 2, \ldots, r,$$

where

$$q_i = q(\alpha_i, x_1, \ldots x_n), \quad i = 1, 2, \ldots, r, \tag{21}$$

Applying function F to both sides of these equations, we get

$$\alpha_i = F(q_i, \lambda), \quad i = 1, 2, \ldots, r. \tag{22}$$

Finding parameters from this system is called the method of percentiles.

Some examples of its application are presented below.

- Assume that X follow the exponential distribution (5) with parameter λ. Because we have just 1 parameter, it suffice to choose one $\alpha_1 \in (0, 1)$, for example, $\alpha_1 = 0.5$. Then, given data $x_1, \ldots, x_n$, we need to estimate

  $$q_1 = q(\alpha_1, x_1, \ldots x_n)$$

  as explained above. Finally, λ is found from the equation

  $$\alpha_1 = F(q_1, \lambda) = 1 - e^{-\lambda q_1}.$$

  We get

  $$\lambda = -\frac{\log(1 - \alpha_1)}{q_1}.$$

- Assume that X follow the Pareto distribution (9) with parameters $\alpha > 0$ and $\lambda > 0$. Because we have 2 parameters, we need to select $0 < \alpha_1 < \alpha_2 < 1$, for example, $\alpha_1 = 1/4$, $\alpha_2 = 3/4$. Then we estimate

  $$q_1 = q(\alpha_1, x_1, \ldots x_n), \quad q_2 = q(\alpha_2, x_1, \ldots x_n)$$

  from data. Then system (22) becomes

  $$\alpha_1 = 1 - \left(\frac{\lambda}{\lambda + q_1}\right)^{\alpha}, \quad \alpha_2 = 1 - \left(\frac{\lambda}{\lambda + q_2}\right)^{\alpha}$$

  and can be solved numerically to find α and λ.

- Assume that X follow the Weibull distribution (12) with parameters $c > 0$ and $\gamma > 0$. Because we have 2 parameters, we need to select $0 < \alpha_1 < \alpha_2 < 1$, for example, $\alpha_1 = 1/4$, $\alpha_2 = 3/4$. Then we estimate

  $$q_1 = q(\alpha_1, x_1, \ldots x_n), \quad q_2 = q(\alpha_2, x_1, \ldots x_n)$$

  from data. Then system (22) becomes

  $$\alpha_1 = 1 - \exp(-cq_1^{\gamma}), \quad \alpha_2 = 1 - \exp(-cq_2^{\gamma}).$$

It can we rewritten as

$$-cq_1^\gamma = \log(1-\alpha_1), \quad -cq_2^\gamma = \log(1-\alpha_2).$$

Dividing first equation by the second one, we get

$$\left(\frac{q_1}{q_2}\right)^\gamma = \frac{\log(1-\alpha_1)}{\log(1-\alpha_2)},$$

hence

$$\gamma = \log\left[\frac{\log(1-\alpha_1)}{\log(1-\alpha_2)}\right] / \log\left[\frac{q_1}{q_2}\right], \tag{23}$$

and then

$$c = -\frac{\log(1-\alpha_1)}{q_1^\gamma}. \tag{24}$$

2.5 Reinsurance

To protect itself from large claims, an insurance company, let us call it I (insurer), may in turn take out an insurance policy in another company, which we call it R (reinsurer). Such a policy is called a reinsurance policy. Insurance company I received premiums from client C, and pay part of this premium to R. Then, if client C makes a claim, part of it may be covered by R, in accordance to a contract between I and R. In this section we consider reinsurance contracts of two very simple types: proportional reinsurance and individual excess of loss reinsurance.

In proportional reinsurance the insurer I pays a fixed proportion α of the claim, $0 < \alpha < 1$, whatever the size of the claim, and the reinsurer R pays the remaining proportion $1-\alpha$ of the claim. In other words, if claim amount is X, then I pays αX and R pays $(1-\alpha)X$. The parameter α is known as the retained proportion or retention level.

Imagine that claim arrives independently and identically distributed from some unknown distribution, and insurance company I has historical record of sizes $y_1, y_2, \ldots, y_n$ of payments they made. Then they may easy recover the actual sizes of past claims: they are $y_1/\alpha, y_2/\alpha, \ldots, y_n/\alpha$. Similarly, if reinsurance company R has historical record of sizes $z_1, z_2, \ldots, z_n$ of payments they made, then the actual sizes of past claims are $z_1/(1-\alpha), z_2/(1-\alpha), \ldots, z_n/(1-\alpha)$. These data may be used to estimate the claim distribution using methods described in sections 2.2-2.4.

In excess of loss reinsurance, the insurer I will pay any claim in full up to a certain amount M, which is called the retention level; any amount above M will be paid by the reinsurer R. Note that the term "retention level" is

used in both proportional reinsurance and excess of loss reinsurance, but it has completely different meaning!

With excess of loss reinsurance contract, if the claim for amount X arrives then the insurer will pay Y where:

$$Y = \begin{cases} X & \text{if } X \leq M \\ M & \text{if } X > M \end{cases}$$

The reinsurer pays the amount

$$Z = X - Y = \begin{cases} 0 & \text{if } X \leq M \\ X - M & \text{if } X > M \end{cases} \tag{25}$$

Because $Z \geq 0$, it is clear that $E[Z] \geq 0$, or, equivalently, $E[Y] \leq E[X]$. In fact, from formula for Z it is clear that $E[Z] > 0$ if $P[X > M] > 0$. Hence, in this case, $E[Y] < E[X]$. With a bit more work, one may prove similar inequality for variance as well. In conclusion,

$$E[Y] \leq E[X] \quad \text{and} \quad Var[Y] \leq Var[X],$$

and both inequality are strict if $P[X > M] > 0$. This means that, for the insurer, both mean amount paid and the variance of the amount paid are reduced.

If X is a non-negative continuous random variable with density function $f(x)$, then

$$E[X] = \int_0^\infty x f(x) dx.$$

This is the mean amount the insurance company I would pay without reinsurance. With reinsurance, the mean amount for I to pay is

$$E[Y] = \int_0^M x f(x) dx + MP(X > M),$$

while the mean amount for re-insurer R to pay is

$$E[Z] = \int_M^\infty (x - M) f(x) dx.$$

Without the reinsurence, if the claim amount is inflated by a factor of k, then so is the mean amount for insurer to pay:

$$E[kX] = kE[X].$$

This is not the case for excess of loss reinsurance if the retention level M is fixed. In this case, the amount for insurer I po pay after inflation becomes

$$Y' = \begin{cases} kX & \text{if } kX \leq M \\ M & \text{if } kX > M \end{cases}$$

and the mean amount is

$$E[Y'] = \int_0^{M/k} kxf(x)dx + MP(X > M/k).$$

One can easily check that in general $E[Y'] \neq kE[Y]$.

Many insurance companies, especially those working in motor insurance and many kinds of property and accident insurance, offer policies which require from clients to cover their loss themselves up to some limit L, called the excess. If the amount X of loss is less than L, the claim cannot happen, and if X is greater than L, then the client will claim only for $X - L$. This policy is called "policy with excess". If Y is the amount actually paid by the insurer, then

$$Y = \begin{cases} 0 & \text{if } X \leq L \\ X - L & \text{if } X > L \end{cases}$$

Note that this equation completely coincides with (25), just M is replaced by L. Hence, the position of the insurer in a policy with excess is exactly the same as that of an reinsurer with an excess of loss reinsurance contract. Similarly, the position of the policyholder/client buying policy with excess is exactly the same as that of an insurer with an excess of loss reinsurance contract.

2.6 Claim size estimation with excess of loss reinsurance

With excess of loss reinsurance, the problem of estimation of unknown claim distribution may be challenging due to incomplete data. Imagine that claims arrives independently and identically distributed from some unknown distribution, and insurance company I has historical record of sizes of payments they made. A typical record has the form

$$x_1, x_2, M, x_4, x_5, x_6, M, x_8, \ldots$$

In this example, the actual sizes of 3-rd and 7-th claims are not recorded, the company only known that they are greater than or equal to M. In general,

if some data are missing or incomplete, we say that we have a "censored sample". So, with censored sample as above, can we still use the methods described in sections 2.2-2.4 to estimate the parameters of the unknown claim distribution?

- The method of moments, see Section 2.2, is not available, because the moments cannot be reliably estimated from the censored sample;

- The method of percentiles, see Section 2.4, can be used without modification, provided that all q_i in (21) are less than M. This is the case if the retention level M is high, so that only few highest claims are unknown. For example, let $M = 1000$, $n = 9$, the data are

  $$500, 300, 1000, 100, 800, 500, 1000, 700, 300.$$

  and the percentiles levels in Section 2.4 are $\alpha_1 = 1/4$, $\alpha_2 = 3/4$. Then the lowest integers grater than $n\alpha_1 = 9/4$ and $n\alpha_2 = 27/4$ are 3 and 7, respectively. If we sort the data in non-decreasing order

  $$100, 300, 300, 500, 500, 700, 800, 1000, 1000$$

  the third and seventhes terms are $q_1 = 300$ and $q_2 = 800$, respectively. These values of q_1 and q_2 are all that we need to proceed with method of percentiles, and unknown claims over 1000 has no influence on this calculation. This is the big advantage of method of percentiles.

- The method of maximum likelihood, described in Section 2.3, can be used but require modification. Let J be the set of claims less than M for which information is available. The contribution of these data to the maximal likelihood function is exactly like in Section 2.3:

  $$\prod_{i \in J} f(x_i \,|\, \theta),$$

  where $f(x \,|\, \theta)$ is the density function if claim distribution. Next, assume that m other claims are referred to the re-insurer, and insurer only knows that they are greater than m. These censored claims contribute to the likelihood function a factor of

  $$[P(X > M)]^m.$$

  If $F(x \,|\, \theta)$ is the cumulative distribution function if claim distribution, then $P(X > M) = 1 - P(X \leq M) = 1 - F(M \,|\, \theta)$. Hence, the complete likelihood function is

  $$L(\theta) = \prod_{i \in J} f(x_i \,|\, \theta) \cdot (1 - F(M \,|\, \theta))^m,$$

and its logarithm is

$$l(\theta) = \log(L(\theta)) = \sum_{i \in J} \log[f(x_i \,|\, \theta)] + m \log(1 - F(M \,|\, \theta)).$$

The optimal $\hat{\theta}$ maximizing this function can be found from the same system of equations (20).

Let us consider the same problem of estimating unknown claim distribution from the point of view of reinsurer. The historical records

$$w_1, w_2, \ldots, w_k$$

of reinsurer expenses consist on differences $w_i = x_i - M$ for claims whose size x_i is greater than M. For claims with $x_i \leq M$, the reinsurer even does not know that such claims occur. This implies that in fact $w_1, w_2, \ldots, w_k$ is an i.i.d. sample for random variable

$$W = X - M | X > M.$$

Let us express the cdf $G(w)$ and pdf $g(w)$ of W using the cdf $F(x)$ and pdf $f(x)$ of the original claim size distribution X. For any $w \geq 0$,

$$G(w) = P[W \leq w] = P[X - M \leq w | X > M] = \frac{P[X \leq w + M \text{ and } X > M]}{P[X > M]} =$$

$$= \frac{P[M < X \leq w + M]}{1 - P[X \leq M]} = \frac{F(w + M) - F(M)}{1 - F[M]}$$

Differentiating with respect to w, we get

$$g(w) = G'(w) = \frac{f(w + M)}{1 - F(M)}.$$

If X belongs to some distribution family with unknown parameters, we can use these formulas to express cdf and pdf of W as a function of these parameters, and then use methods from sections 2.2-2.4 to estimate the parameters base on data $w_1, w_2, \ldots, w_k$.

Example 2.1. Assume that claim size X has Pareto distribution (9) with parameters $\alpha > 2$ and $\lambda > 0$. Then the pdf of W is

$$g(w) = \frac{f(w + M)}{1 - F(M)} = \frac{\alpha \lambda^\alpha}{(\lambda + w + M)^{\alpha+1}} : \left(1 - \left(1 - \left(\frac{\lambda}{\lambda + M}\right)^\alpha\right)\right) =$$

$$= \frac{\alpha\lambda^\alpha}{(\lambda + w + M)^{\alpha+1}} \cdot \left(\frac{\lambda + M}{\lambda}\right)^\alpha = \frac{\alpha(\lambda + M)^\alpha}{(\lambda + w + M)^{\alpha+1}}.$$

Let us use, for example, method of moments to estimate parameters. It states that

$$\frac{1}{k}\sum_{i=1}^{k} w_i = E[W] = \frac{\lambda + M}{\alpha - 1}$$

and

$$\frac{1}{k}\sum_{i=1}^{k} w_i^2 - \left(\frac{1}{k}\sum_{i=1}^{k} w_i\right)^2 = Var[W] = \frac{\alpha(\lambda + M)^2}{(\alpha - 1)^2(\alpha - 2)}.$$

This gives a system of two equations to find two unknown parameters λ and α.

2.7 Summary

In this chapter we focus on modelling the claim size distribution. We assume that claim distribution belongs to certain family, but with unknown parameters. The company estimate unknown parameters $a_1, a_2, \ldots, a_r$ to fit the data of past claims $x_1, x_2, \ldots, x_n$ as good as possible.

The method of moments suggests to select parameters in such a way that first r moments estimated from the data match the first r moments $f_j(a_1, a_2, \ldots, a_r)$, $j = 1, 2, \ldots, r$ estimated from the formulas for the distribution, that is,

$$m_j = f_j(a_1, a_2, \ldots, a_r), \quad j = 1, 2, \ldots, r$$

where $m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j,\ j = 1, 2, \ldots, r.$

Method of maximum likelihood suggests to select vector of parameters $theta = (a_1, a_2, \ldots, a_r)$ to maximize (the logarithm of) the likelihood function

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log[f(x_i \,|\, \theta)].$$

The optimal $\hat{\theta} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_r)$ can be found from the system of equations

$$\frac{d}{da_i} l(\hat{\theta}) = 0, \quad i = 1, 2, \ldots, r.$$

Method of percentiles suggests to find vector of parameters $\lambda = (a_1, a_2, \ldots, a_r)$ from the system of equations

$$\alpha_i = F(q_i, \lambda), \quad i = 1, 2, \ldots, r,$$

where F is a cdf which depends on parameters, $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_r < 1$ are some pre-specified numbers, and q_i is the estimate of the percentile at level α_i based on data $x_1, x_2, \ldots, x_n$. To find it, we first find the smallest integer j greater than $n\alpha_i$, then sort the sequence $x_1, x_2, \ldots x_n$ is non-decreasing order, and then the j-th smallest number is q_i.

To protect itself from large claims, an insurance company may in turn take out an insurance policy in another company, called reinsurer. We consider reinsurance contracts of two very simple types: proportional reinsurance and individual excess of loss reinsurance. In proportional reinsurance, if claim amount is X, then insurer pays αX and reinsurer pays $(1 - \alpha)X$, where α is

the parameter known as the retention level. With excess of loss reinsurance contract, if the claim for amount X arrives then the insurer will pay

$$Y = \begin{cases} X & \text{if } X \leq M \\ M & \text{if } X > M \end{cases}$$

and the reinsurer pays $Z = X - Y$.

2.8 Questions

1. The number of claims a company received during the last 12 months are

$$10, 8, 15, 10, 7, 3, 20, 14, 5, 12, 8, 8.$$

Assuming that these numbers are i.i.d. realizations of

(a) Poisson distribution with parameter λ

(b) negative binomial distribution with parameters p and k,

use method of moments to estimate unknown parameters.

2. Assume that the same data as in question 1 are i.i.d. realizations of geometric distribution with parameter p. Use method of maximum likelihood to estimate p.

3. The history of $n = 18$ most recent claim sizes (rounded to integer number of pounds) are

$$937, 342, 150, 1080, 401, 3500, 7970, 1400, 530,$$

$$1106, 847, 899, 3076, 2837, 315, 2560, 390, 2950.$$

Assuming that these are i.i.d data from Weibull distribution, use the method of percentiles with $\alpha_1 = 1/4$ and $\alpha_2 = 3/4$ to estimate parameters of the distribution.

4. Assume that the history of claim sizes are the same as in the previous question, but the company order a reinsurance policy with excess of loss reinsurance above the level $M = 2000$.

(a) Write down the history of expenses of the reinsurer;

(b) Assuming that the original claim size distribution is Pareto distribution with parameters $\alpha > 2$ and $\lambda > 0$, estimate the unknown parameters using method of moments with data available to the reinsurer.

(c) Comment whether do you think the Pareto distribution is a good model to fit these data.

Chapter 3
Estimation of aggregate claim distribution

3.1 The collective risk model

As discussed in the previous chapter, if there will be N claims during a month (or week, or year, or other fixed period of time) with sizes $X_1, X_2, \ldots, X_N$, then the total losses to cover all claims are

$$S = X_1 + X_2 + \cdots + X_N,$$

and $S = 0$ if $N = 0$. If all X_i are independent, identically distributed, and also independent of N, then we say that S has *compound distribution.*

In the previous chapter we focused on estimating the cumulative distribution function $F(x)$ of individual claims X_i based on historical data.

In this chapter we focus on estimation the cumulative distribution function $G(x)$ of total claim size S. By definition, $G(x)$ is equal to the probability of event $\{S \le x\}$. This event can happen if either

- $\{S \le x$ and $N = 0\}$, that is, no claims occurred, or
- $\{S \le x$ and $N = 1\}$, that is, one claims of amount $\le x$ occurred, or
- $\{S \le x$ and $N = 2\}$, that is, 2 claims of total amount $\le x$ occurred, or
- ...
- $\{S \le x$ and $N = n\}$, that is, n claims of total amount $\le x$ occurred, or
- ...

Hence, by the law of total probability (13)

$$G(x) = P(S \le x) = \sum_{n=0}^{\infty} P(S \le x \text{ and } N = n) = \sum_{n=0}^{\infty} P(N = n) \cdot P(S \le x \,|\, N = n). \tag{26}$$

The term

$$P(S \le x \,|\, N = n) = P(X_1 + X_2 + \cdots + X_n \le x)$$

represents the distribution function of sum of n i.i.d. random variables with distribution function $F(x)$ each. This function is known as n-fold convolution of F and is denoted as $F^{n*}(x)$.

If independent random variables X and Y have densities f and g, the density h of their sum $X+Y$ is given by

$$h(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt,$$

and the cdf of $X+Y$ is $H(x) = \int_{-\infty}^{x} h(t)dt$. The n-fold convolution of any continuous distribution F can be computed by applying these formulas n times. The n-fold convolution of discrete distribution F can be computed by definition, as demonstrated in Example below.

Example 3.1. Let F be a distribution of discrete random variable, taking values 0 and 1 with equal chances. Calculate $F^{3*}(x)$.

Answer: Let $S = X_1 + X_2 + X_3$ where each X_i is 0 or 1 with equal chances. Then S can takes values 0, 1, 2, and 3 with probabilities

$$P(S=0) = P(X_1 = X_2 = X_3 = 0) = 1/8,$$

$$P(S=1) = P(X_1 = 1, X_2 = X_3 = 0) + P(X_2 = 1, X_1 = X_3 = 0) +$$
$$+P(X_3 = 1, X_1 = X_2 = 0) = 3/8,$$

and similarly

$$P(S=2) = 3/8, \quad P(S=3) = 1/8.$$

Hence,

$$F^{3*}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/8 & \text{if } 0 \le x < 1 \\ 1/2 & \text{if } 1 \le x < 2 \\ 7/8 & \text{if } 2 \le x < 3 \\ 1 & \text{if } 3 \le x \end{cases}$$

We remark that $F^{1*}(x) = F(x)$. For convenience, we also introduce notation

$$F^{0*}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \le x. \end{cases}$$

With notation $F^{n*}(x)$ equation (26) becomes

$$G(x) = P(S \le x) = \sum_{n=0}^{\infty} P(N=n) \cdot F^{n*}(x). \tag{27}$$

For example, if X_i are discrete random variables taking values in non-negative integers only, then, for every non-negative integer x,

$$P(S = x) = G(x) - G(x-1) = \sum_{n=0}^{\infty} P(N = n) \cdot (F^{n*}(x) - F^{n*}(x-1)).$$

Example 3.2. Let $S = X_1 + X_2 + \cdots + X_N$ where X_i are i.i.d. random variables taking values 0 or 1 with equal chances, and N can be 0, 1, 2, or 3 with equal chances. Find the probability that $S = 2$. **Answer:**

$$P(S = 2) = \sum_{n=0}^{3} P(N = n) \cdot (F^{n*}(2) - F^{n*}(2-1)) = \frac{1}{4} \sum_{n=0}^{3} P(N = n) \cdot (F^{n*}(2) - F^{n*}(1)).$$

In Example 3.1, we calculated that

$$F^{3*}(2) - F^{3*}(1) = \frac{7}{8} - \frac{1}{2} = \frac{3}{8}.$$

Similar calculation shows that $F^{2*}(2) - F^{2*}(1) = 1 - \frac{3}{4} = \frac{1}{4}$, $F^{1*}(2) - F^{1*}(1) = F^{0*}(2) - F^{0*}(1) = 1 - 1 = 0$. Hence,

$$P(S = 2) = \frac{1}{4} \cdot \left(0 + 0 + \frac{1}{4} + \frac{3}{8}\right) = \frac{5}{32}.$$

We now find mean and variance of $S = X_1 + X_2 + \cdots + X_N$, where $X_1, \ldots, X_N$ are i.i.d. copies of some random variable X.

We denote μ_X, μ_N, μ_S the means of random variables X, N, S and by σ_X^2, σ_N^2, σ_S^2 the corresponding variances. We have

$$E(S|N = n) = E(X_1 + X_2 + \cdots + X_n) = nE(X),$$

or, in other words, $E(S|N) = NE(X)$. Hence, by the law of total expectation (15)

$$\mu_S = E(S) = E[E(S|N)] = E[NE(X)] = E[N]E[X] = \mu_N \cdot \mu_X. \qquad (28)$$

In words, the expected size of total claim is equal to expected number of claims times the average size of one claim.

Similarly,

$$Var(S|N = n) = Var(X_1 + X_2 + \cdots + X_n) = nVar(X),$$

or $Var(S|N) = NVar(X)$. Then by the law of total variance (16),

$$\sigma_S^2 = Var(S) = E[Var(S|N)] + Var[E(S|N)] = E[NVar(X)] + Var[NE(X)],$$

hence

$$\sigma_S^2 = E[N]Var(X) + Var[N](E(X))^2 = \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2. \tag{29}$$

We can also calculate moment generating function $M_S(t)$ of S. By the law of total expectation (15)

$$M_S(t) = E(e^{tS}) = E[E(e^{tS}|N)].$$

But

$$E(e^{tS}|N = n) = E[e^{t(X_1 + X + 2 + \cdots + X_n)}] = \prod_{i=1}^{n} E[e^{tX_i}] = (M_X(t))^n.$$

Hence,

$$M_S(t) = E[(M_X(t))^N] = E[e^{N \cdot \log M_X(t)}] = M_N(\log M_X(t)). \tag{30}$$

Example 3.3. Consider the special case when all claims are for the same fixed amount B. That is, $P(X_i = B) = 1$ for all i. Then

$$S = X_1 + X_2 + \cdots + X_N = B + B + \cdots + B = NB.$$

Hence, $E[S] = E[NB] = BE[N]$ and $Var[S] = Var[NB] = B^2 Var[N]$. Because $\mu_X = B$ and $\sigma_X^2 = 0$, the same expressions follow from formulas (28) and (29).

In the next sections we consider compound distributions S for various models for the distribution of the number of claims N.

3.2 The compound Poisson distribution

In this section we assume that the number of claims N follow a Poisson distribution with parameter λ, that is,

$$P[N = n] = \frac{\lambda^n e^{-\lambda}}{n!}$$

This is a natural model if we assume that claims arrives "uniformly at rate λ" as we explain in later chapters.

Using series expansion for exponential function $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, one can compute moment generating function of Poisson distribution

$$M_N(t) = E[e^{tX}] = \sum_{n=0}^{\infty} e^{tn} P[N=n] = \sum_{n=0}^{\infty} e^{tn} \frac{\lambda^n e^{-\lambda}}{n!} = \lambda e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} = e^{-\lambda} e^{\lambda e^t}.$$

Differentiating it, we can find moments

$$\frac{dM_N(t)}{dt} = \exp(\lambda(e^t-1)) \cdot \lambda e^t, \quad \frac{d^2 M_N(t)}{dt^2} = \exp(\lambda(e^t-1))((\lambda e^t)^2 + \lambda e^t)$$

$$\mu_N = E[N] = \left.\frac{dM_N(t)}{dt}\right|_{t=0} = \lambda,$$

and

$$E[N^2] = \left.\frac{d^2 M_N(t)}{dt^2}\right|_{t=0} = \lambda^2 + \lambda,$$

hence

$$\sigma_N^2 = E[N^2] - (E[N])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Now we can use formulas derived in the previous section to calculate mean and variance of $S = X_1 + X_2 + \cdots + X_N$. By (28),

$$\mu_S = E[S] = \mu_N \cdot \mu_X = \lambda \mu_X. \tag{31}$$

By (29),

$$\sigma_S^2 = \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2 = \lambda(\sigma_X^2 + \mu_X^2) = \lambda E[X^2]. \tag{32}$$

Also, by (30),

$$M_S(t) = M_N(\log M_X(t)) = \exp(\lambda(e^{\log M_X(t)} - 1)) = \exp(\lambda(M_X(t) - 1)). \tag{33}$$

Formulas for μ_S and σ_S^2 above can also be derived by differentiating $M_S(t)$ once and twice. By differentiating it three times we can also derive the skewness:

$$E\left[(S-\mu_S)^3\right] = \lambda E[X^3],$$

hence

$$\text{skew}[S] := E\left[\left(\frac{S-\mu_S}{\sigma_S}\right)^3\right] = \frac{\lambda E[X^3]}{(\lambda E[X^2])^{3/2}}. \tag{34}$$

Because claims X_i are positive random variables, $E[X^3] > 0$, hence S is positively skewed even if the distribution of X_i are negatively skewed. Also note that $\lim_{\lambda \to \infty} \text{skew}[S] = 0$, hence the distribution of S is almost symmetric if λ is large.

Example 3.4. Assume that the number of claims during a year has the Poisson distribution with parameter λ and the size of each claim is a random variable uniformly distributed on $[a, b]$. All claims sizes are independent. What is the mean and variance of the cumulative size of the claims from all policies?

Answer: If X is the size of a claim, then

$$E[X] = \frac{1}{b-a}\int_a^b x dx = \frac{a+b}{2};$$

$$E[X^2] = \frac{1}{b-a}\int_a^b x^2 dx = \frac{a^2+ab+b^2}{3},$$

which gives

$$E[S] = \lambda E[X] = \lambda(a+b)/2,$$

and

$$Var[S] = \lambda E[X^2] = \lambda(a^2+ab+b^2)/3.$$

Now assume that we have n types of claims. The number N_i of claims of i-th type has Poisson distribution with parameter λ_i, and the sizes of such claims are i.i.d. with cdf $F_i(x)$. The total size of such claims is

$$S_i = X_1 + X_2 + \cdots + X_{N_i}, \quad i = 1, 2, \ldots, n,$$

and the total size off all claims in then

$$S = S_1 + S_2 + \cdots + S_n.$$

We claim that S has a compound Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2 + \ldots \lambda_n$ and the cdf of individual claim amounts

$$F(x) = \frac{1}{\lambda}\sum_{i=1}^{n} \lambda_i F_i(x).$$

This result is very important because it reduces sum of n independent compound Poisson distribution to just one compound Poisson distribution with different parameters.

Let us prove this result. Let N follow a Poisson distribution with parameter λ, let $Y_1, Y_2, \ldots, Y_N$ be i.i.d. random variables with cdf F, and let

$$S' = Y_1 + Y_2 + \cdots + Y_n.$$

We want to prove that S and S' have the same distributions. Because there is a one-to-one relationship between distributions and moment-generation functions, it is sufficient to prove that S and S' has the same moment-generation functions.

We first calculate $M_S(t)$ of S. By independence of S_i,

$$M_S(t) = E[e^{tS}] = E[e^{t(S_1+S_2+\cdots+S_n)}] = \prod_{i=1}^{n} E[e^{tS_i}].$$

Now, by (33)

$$E[e^{tS_i}] = \exp(\lambda_i(M_i(t) - 1)), \quad i = 1, 2, \ldots, n,$$

where $M_i(t)$ is the moment generation function which corresponds to cdf $F_i(x)$. Then

$$M_S(t) = \prod_{i=1}^{n} \exp(\lambda_i(M_i(t) - 1)) = \exp\left[\sum_{i=1}^{n} \lambda_i(M_i(t) - 1)\right]$$

Now let us calculate the moment generation function $M'(t)$ of S'. By by (33),

$$M'(t) = \exp(\lambda(M_Y(t) - 1)),$$

where $M_Y(t)$ is the moment generation function which corresponds to cdf $F(x)$. By definition, it is equal to

$$M_Y(t) = \int_{-\infty}^{\infty} e^{tx} dF(x) = \frac{1}{\lambda} \sum_{i=1}^{n} \lambda_i \int_{-\infty}^{\infty} e^{tx} dF_i(x) = \frac{1}{\lambda} \sum_{i=1}^{n} \lambda_i M_i(t).$$

Hence,

$$M'(t) = \exp(\lambda M_Y(t) - \lambda) = \exp\left(\sum_{i=1}^{n} \lambda_i M_i(t) - \sum_{i=1}^{n} \lambda_i\right) = M_S(t).$$

This implies that S and S' have the same distribution and finishes the proof.

3.3 The compound binomial distribution

In this section we assume that the number of claims N follow a binomial distribution with parameter n and p, that is,

$$P[N = k] = \frac{n!}{k!(n-k)!} \cdot p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

As mentioned in Chapter , this is the case if

$$N = N_1 + N_2 + \cdots + N_n$$

where N_i are i.i.d standard Bernoulli variables (that is, take value 1 with probability p and 0 otherwise). This is a natural model if we assume that the company covers n independent policies such that each one may issue a claim with the same probability p.

Because $E[N_i] = p$, $Var[N_i] = p(1-p)$ and the moment generating function $M_i(t)$ of N_i is

$$E[e^{tN_i}] = e^{t\cdot 0} \cdot (1-p) + e^{t\cdot 1} \cdot p = pe^t + 1 - p,$$

we have $E[N] = np$, $Var[N] = np(1-p)$, and $M_N(t) = (pe^t + 1 - p)^n$.

By (28),

$$\mu_S = E[S] = \mu_N \cdot \mu_X = np\mu_X. \tag{35}$$

By (29),

$$\sigma_S^2 = \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2 = np(\sigma_X^2 + (1-p)\mu_X^2) = np(E[X^2] - p(E[X])^2). \tag{36}$$

Also, by (30),

$$M_S(t) = M_N(\log M_X(t)) = (pe^{\log M_X(t)} + 1 - p)^n = (pM_X(t) + 1 - p)^n. \tag{37}$$

Differentiating $M_S(t)$ three times we can derive the skewness:

$$E\left[(S - \mu_S)^3\right] = npE[X^3] - 3np^2E[X^2]E[X] + 2np^3(E[X])^3,$$

hence

$$\text{skew}[S] = \frac{npE[X^3] - 3np^2E[X^2]E[X] + 2np^3(E[X])^3}{(npE[X^2] - np^2(E[X])^2)^{3/2}}.$$

We can see that S can be positively or negatively skewed, depending on the parameters.

Example 3.5. Assume that all claims are for the same amount B. Then $E[X^k] = B^k$, $k = 1, 2, 3$, and

$$\text{skew}[S] = \frac{npB^3 - 3np^2B^3 + 2np^3B^3}{(npB^2 - np^2B^2)^{3/2}} = \frac{0.5 - p}{\sqrt{np(1-p)}}.$$

In particular, $\text{skew}[S] > 0$ if $p < 0.5$ but $\text{skew}[S] < 0$ if $p > 0.5$.

Example 3.6. Assume that the number of claims during a year has the binomial distribution with parameters n and p the size of each claim is a

random variable X uniformly distributed on $[a, b]$. All claims sizes are independent. What is the mean and variance of the cumulative size of the claims from all policies?

Answer:

$$E[S] = npE[X] = np(a+b)/2,$$

and

$$Var[S] = npE[X^2] - np^2(E[X])^2 = np(a^2+ab+b^2)/3 - np^2(a+b)^2/4.$$

3.4 The compound negative binomial distribution

In this section we assume that the number of claims N follow a negative binomial distribution with parameters k and p, that is,

$$P[N=n] = \frac{(k+n-1)!}{n!(k-1)!} \cdot p^k(1-p)^n, \quad n = 0, 1, 2, \ldots$$

As mentioned in Chapter , this is the case if

$$N = N_1 + N_2 + \cdots + N_k$$

where N_i are i.i.d variables with geometric distribution.

Because $E[N_i] = (1-p)/p$, $Var[N_i] = (1-p)/p^2$ and the moment generating function $M_i(t)$ of N_i is

$$E[e^{tN_i}] = \sum_{n=1}^{\infty} e^{tn} \cdot (1-p)^n p = \frac{p}{1-(1-p)e^t}$$

we have

$$E[N] = k(1-p)/p, \quad Var[N] = k(1-p)/p^2, \quad M_N(t) = p^k(1-(1-p)e^t)^{-k}.$$

You may note that $Var[N] > E[N]$, while for Poisson distribution $Var[N] = E[N]$. This is the advantage of negative binomial distribution: it can better fit the data if sample variance is greater than sample mean, which is often the case in practise.

By (28),

$$\mu_S = E[S] = \mu_N \cdot \mu_X = \frac{k(1-p)}{p}\mu_X \tag{38}$$

By (29),

$$\sigma_S^2 = \mu_N\sigma_X^2 + \sigma_N^2\mu_X^2 = \frac{k(1-p)}{p}E[X^2] + \frac{k(1-p)^2}{p^2}(E[X])^2. \tag{39}$$

Also, by (30),

$$M_S(t) = M_N(\log M_X(t)) = \frac{p^k}{(1-(1-p)M_X(t))^k}. \tag{40}$$

Differentiating $M_S(t)$ three times we can derive the third centralized moment:

$$E\left[(S-\mu_S)^3\right] = \frac{3k(1-p)^2E[X]E[X^2]}{p^2} + \frac{2k(1-p)^3(E[X])^3}{p^3} + \frac{k(1-p)E[X^3]}{p}.$$

Because all terms are positive, the compound negative binomial distribution is always positively skewed.

3.5 Aggregate claim distribution under reinsurance

Under proportional reinsurance, the insurer and reinsurer each pays a defined proportion of every claim, and therefore their aggregate claim is proportional to the aggregate claim with no reinsurance, whose distribution has been derived in the previous sections. For a retention level α $(0 \leq \alpha \leq 1)$, the i-th individual claim amount for the insurer is αX_i and for the reinsurer is $(1-\alpha)X_i$. The aggregate claims amounts are αS and $(1-\alpha)S$, respectively.

Under the excess of loss reinsurance with retention level M, the amount that an insurer pays on the i-th claim is $Y_i = \min(X_i, M)$, while the amount that the reinsurer pays is $Z_i = X_i - Y_i = \max(0, X_i - M)$. Thus, the insurer's aggregate claims net of reinsurance can be represented as:

$$S_I = Y_1 + Y_2 + \cdots + Y_N,$$

and the reinsure's aggregate claims is:

$$S_R = Z_1 + Z_2 + \cdots + Z_N. \tag{41}$$

We remark that if $X_i < M$, which is usually the case, then $Z_i = 0$, and, in reality, the reinsurer will not even see this claim. However, this formula for S_R with lots of zero terms, while somewhat artificial, is convenient for calculations. In particular, we can use formulas (28), (29), and (30) to estimate mean, variance, and moment generating function for S_I and S_R.

Example 3.7. The number N of claims has Poisson distribution with parameter $\lambda = 10$. Individual claim amounts are uniformly distributed on $(0, 2000)$. The insurer of this risk has effected excess of loss reinsurance with retention level 1600. Calculate the mean, variance and coefficient of skewness

of both the insurer's and reinsurer's aggregate claims under this reinsurance arrangement.

Answer: In this case, $X_i \sim U(0, 2000)$, $M = 1600$. As usual, denote $Y_i = \min(X_i, M)$ and $Z_i = X_i - Y_i = \max(0, X_i - M)$. Then

$$E[Y_i] = \int_0^M x f(x) dx + M \cdot P(X_i > M),$$

where $f(x) = 0.0005$ is the $U(0, 2000)$ density. This gives

$$E[Y_i] = \frac{0.0005(M^2 - 0^2)}{2} + 0.2M = 960.$$

Hence, by (31),

$$E[S_I] = \lambda E[Y_i] = 10 \cdot 960 = 9600.$$

Further

$$E[Y_i^2] = \int_0^M x^2 f(x) dx + M^2 \cdot P(X_i > M) = 1,194,666.7,$$

and by (32)

$$Var[S_I] = \lambda E[Y_i^2] = 11,946,667.$$

Next,

$$E[Y_i^3] = \int_0^M x^3 f(x) dx + M^3 \cdot P(X_i > M) = 1,638,400,000,$$

hence by (34)

$$\text{skew}[S_I] = \frac{\lambda E[Y_i^3]}{(\lambda E[Y_i^2])^{3/2}} = \frac{16,384,000,000}{(11,946,667)^{3/2}} \approx 0.397.$$

Let us now do similar calculation for the reinsurer. Because $X_i \sim U(0, 2000)$, we have $E[X_i] = 1000$, hence

$$E[Z_i] = E[X_i - Y_i] = E[X_i] - E[Y_i] = 1000 - 960 = 40,$$

and by (31)

$$E[S_R] = \lambda E[Z_i] = 10 \cdot 40 = 400.$$

Further,

$$E[Z_i^2] = \int_M^{2000} (x - M)^2 f(x) dx = \frac{0.0005(2000 - M)^3}{3} \approx 10666.7,$$

and by (32)

$$Var[S_R] = \lambda E[Z_i^2] = 106,667.$$

Next,

$$E[Z_i^3] = \int_M^{2000} (x-M)^3 f(x)dx = \frac{0.0005(2000-M)^4}{4} = 3,200,000,$$

and by (34)

$$\text{skew}[S_I] = \frac{\lambda E[Y_i^3]}{(\lambda E[Y_i^2])^{3/2}} = \frac{32,000,000}{(106,667)^{3/2}} \approx 0.92.$$

The reinsure's aggregate claim can alternatively be represented as:

$$S_R = W_1 + W_2 + \cdots + W_{NR}, \tag{42}$$

where NR is the number of actual (non-zero) claims to the re-insurer, and W_i are the sizes of these claims. For example, if in the example above there are 8 claims of sizes

$$403, 1490, 1948, 443, 1866, 1704, 1221, 823,$$

the (41) reduces to

$$S_R = 0 + 0 + 348 + 0 + 266 + 104 + 0 + 0,$$

while (42) reduces to more natural expression

$$S_R = 348 + 266 + 104.$$

Random variables W_i in (42) have density function

$$g(w) = \frac{f_X(w+M)}{1 - F_X(W)}, \quad w > 0,$$

where f_X and F_X are density and cdf of the original claim size distribution. To find the distribution of NR note that

$$NR = I_1 + I_2 + \cdots + I_N,$$

where N is the total number of claims, and I_j is an indicator random variable which takes the value 1 if the reinsurer makes a (non-zero) payment on the j-th claim and takes the value 0 otherwise. Thus, NR gives the number of

payments made by the reinsurer. Denore π the probability that $X_j > M$. Since I_j takes the value 1 only if $X_j > M$, we have

$$P(I_j = 1) = P(X_j > M) = \pi, \quad \text{and} \quad P(I_j = 0) = 1 - \pi.$$

Hence, $E[I_j] = \pi$, $E[I_j^2] = \pi$, and moment generating function

$$M_I(t) = E[e^{tI_j}] = e^t\pi + 1 - \pi.$$

By formulas (28), (29), and (30), this implies that

$$E[NR] = \pi E[N],$$

$$Var[NR] = E[N](\pi - \pi^2) + Var[N]\pi,$$

and

$$M_{NR}(t) = M_N(\log M_I(t)) = M_N(\log(e^t\pi + 1 - \pi))$$

Example 3.8. In Example 3.7 above, we can use formula (42) to analyse the reinsurance aggregate claim S_R. In this case, NR follow Poisson distribution with parameter $10 \cdot 0.2 = 2$, and individual claims W_i have density function

$$g(w) = \frac{f_X(w + M)}{1 - F_X(W)} = \frac{0.0005}{0.2} = 0.0025, \quad 0 < w < 400,$$

that is, $W_i = U(0, 400)$. Then $E[W_i] = 200$, $E[W_i^2] = 53,333.33$ and $E[W_i^3] = 16,000,000$, giving the same result for mean, variance, and skewness of S_R as above.

Thus, there are two ways to specify and evaluate the distribution of S_R.

3.6 The individual risk model

Under this model a portfolio consisting of a fixed number of risks is considered. It will be assumed that

- these risks are independent;
- claim amounts from these risks are not identically distributed random variables; and
- the number of risks does not change over the period of insurance cover.

As before, aggregate claims from this portfolio are denoted by S. So:

$$S = Y_1 + Y_2 + \cdots + Y_n,$$

where Y_j denotes the claim amount under the j-th risk and n denotes the number of risks. It is possible that some risks will not give rise to claims. Thus, some of the observed values of Y_j may be 0.

For each risk, the following assumptions are made:

(a) The number of claims from the j-th risk, N_j, is either 0 or 1.

(b) The probability of a claim from the j-th risk is q_j.

If a claim occurs under the j-th risk, the claim amount is denoted by the random variable X_j. Let $F_j(x)$, μ_j and σ_j^2 denote the distribution function, mean and variance of X_j, respectively.

Assumption (a) is very restrictive. It means that a maximum of one claim from each risk is allowed for in the model. This includes risks such as one-year term assurance, but excludes many types of general insurance policy. For example, there is no restriction on the number of claims that could be made in a policy year under household contents insurance.

There are three important differences between this model and the collective risk model:

- The number of risks in the portfolio has been specified. In the collective risk model, this number N was not specified and was modelled as a random variable.

- The number of claims from each individual risk has been restricted. There was no such restriction in the collective risk model.

- It is assumed that individual risks are independent but not necessarily identically distributed. In the collective risk model, individual claim amounts are independent and identically distributed.

Assumptions (a) and (b) say that N_j are standard Bernoulli variables, or, equivalently, $N_j \sim Bin(1, q_j)$. Thus, the distribution of Y_j is compound binomial, with individual claim amount random variable X_j. From formulas (35) and (36) it follows that

$$E[Y_j] = q_j \mu_j$$

and

$$E[Y_j] = q_j \sigma_j^2 + q_j(1 - q_j)\mu_j^2.$$

The aggregate claim amount S is the sum of n independent compound binomial random variables. It is easy to find the mean and variance of S.

$$E[S] = E\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} E[Y_j] = \sum_{j=1}^{n} q_j \mu_j, \tag{43}$$

and

$$Var[S] = Var\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} Var[Y_j] = \sum_{j=1}^{n} (q_j \sigma_j^2 + q_j(1-q_j)\mu_j^2). \tag{44}$$

The distribution of S can be computed only under certain conditions, for example, when the compound binomial variables Y_j are identically distributed. In the special case for each policy the values of q_j, μ_j and σ_j^2 are identical, say q, μ and σ_2. Since $F_j(x)$ is independent of j, we can denote it simply $F(x)$. Hence, S is compound binomial, with binomial parameters n and q, and individual claims have distribution function $F(x)$. In this special case, it reduces to the collective risk model, and it can be seen from (43) and (44) that

$$E[S] = nq\mu, \quad var[S] = nq\sigma^2 + nq(1-q)\mu^2.$$

3.7 Aggregate claim estimation under uncertainty in parameters

In the previous sections we have assumed that the distributions of the number of claims and of the amounts of individual claims are known with certainty. In general, parameters of these distributions are not known and should be estimated from appropriate sets of data. In this section we will see how the models introduced earlier can be extended to allow for parameter uncertainty/variability. We will do this by looking at a series of examples. All the examples will be based on claim numbers having a Poisson distribution.

In our first Example we assume that insurance company has n independent policies,

$$S = S_1 + S_2 + \cdots + S_n,$$

and the aggregate claim S_i from i-th policy has a compound Poisson distribution with parameter λ_i, and the CDF of the individual claim amounts distribution is $F(x)$. We assume that $F(x)$ is fixed and identical for all policies, but λ_i are not known with certainty. In contrast, we assume that $\lambda_1, \lambda_2, \ldots, \lambda_n$ is an i.i.d. sample from some known distribution. If we then choose a policy i at random, we can make probability statements about λ_i, such that "there is a 50% chance that λ_i lies between 3 and 5", etc.

Example 3.9. Question: Suppose that the Poisson parameters λ_i of policies are not known but are equally likely to be 0.1 or 0.3. Let n be the number of policies and m_1 and m_2 be the first and second moments of the claim size X.

(i) Find the mean and variance of the aggregate claim S_i from a random policy i;

(ii) Find the mean and variance of the aggregate claim S.

Explanation: The situation described in this Example may arise in, for example, motor insurance. It may be that there are n drivers insured, and some of them are "good" drivers and some are "bad" drivers. The individual claim amount distribution is the same for all drivers but "good" drivers make fewer claims (0.1 p.a. on average) than "bad" drivers (0.3 p.a. on average). It is assumed that it is known, possibly from national data, that a policyholder is equally likely to be a "good" driver or a "bad" driver.

Answer: (i) Let as choose policy i at random. From problem formulation,

$$P[\lambda_i = 0.1] = 0.5 \quad \text{and} \quad P[\lambda_i = 0.3] = 0.5.$$

Hence,

$$E[\lambda_i] = (0.1 + 0.3)0.5 = 0.2, \quad \text{and} \quad Var[\lambda_i] = (0.1^2 + 0.3^2)0.5 - 0.2^2 = 0.01.$$

The conditional distribution of S_i if λ_i is known is compound Poisson, hence formulas (31) and (32) imply

$$E[S_i|\lambda_i] = \lambda_i m_1, \quad \text{and} \quad Var[S_i|\lambda_i] = \lambda_i m_2.$$

Hence, by the law of total expectation (15)

$$E[S_i] = E[E[S_i|\lambda_i]] = E[\lambda_i m_1] = m_1 E[\lambda_i] = 0.2m_1.$$

Similarly, by the law of total variance (16)

$$Var[S_i] = E[Var(S_i|\lambda_i)] + Var(E(S_i|\lambda_i)) = E[\lambda_i m_2] + Var(\lambda_i m_1) =$$

$$m_2 E[\lambda_i] + m_1^2 Var[\lambda_i] = 0.2m_2 + 0.01m_1^2.$$

(ii) Because S_i are independent and identically distributed,

$$E[S] = E[S_1 + S_2 + \cdots + S_n] = nE[S_i] = 0.2nm_1,$$

and

$$Var[S] = Var[S_1 + S_2 + \cdots + S_n] = nVar[S_i] = 0.2nm_2 + 0.01nm_1^2.$$

In the previous example, we assumed that all S_i are independent. We now consider a modification of this example when the number of claims from any policy i follow a Poisson distribution with *the same* but unknown parameter λ. In this case higher S_1 is indication for a higher λ and therefore higher S_2, hence it is not reasonable to assume that S_i are independent. In this situation, the correct assumption is that conditional random variables $S_i|\lambda$ are independent (and identically distributed).

Example 3.10. Question: Suppose that the Poisson parameter λ is unknown but is the same for all policies and is equally likely to be 0.1 or 0.3. Let $S_i|\lambda$ be i.i.d, and let everything else be as in Example 3.9.

(i) Find the mean and variance of the aggregate claim S_i from a random policy i;

(ii) Find the mean and variance of the aggregate claim S.

Explanation: The situation described in this Example may arise in, for example, buildings insurance in a certain area. The number of claims could depend on, among other factors, the weather during the year; an unusually high number of storms resulting in a high expected number of claims (i.e. a high value of λ) and vice versa for all the policies together.

Answer: (i) For a random policy i,

$$P[\lambda = 0.1] = 0.5 \quad \text{and} \quad P[\lambda = 0.3] = 0.5,$$

and exactly the same calculation as in Example 3.9 gives the same result:

$$E[S_i|\lambda] = \lambda m_1, \quad \text{and} \quad Var[S_i|\lambda] = \lambda m_2,$$

and

$$E[S_i] = 0.2m_1 \quad \text{and} \quad Var[S_i] = 0.2m_2 + 0.01m_1^2.$$

(ii) Even when S_i are not independent, expectation of a sum is still the sum of expectations, and

$$E[S] = E[S_1 + S_2 + \cdots + S_n] = nE[S_i] = 0.2nm_1.$$

However, for dependent S_i,

$$Var[S] = Var[S_1 + S_2 + \cdots + S_n] \neq nVar[S_i].$$

Instead, we should use independence of $S_i|\lambda$ to get

$$E[S|\lambda] = nE[S_i|\lambda] = n\lambda m_1, \quad \text{and} \quad Var[S|\lambda] = nVar[S_i|\lambda] = n\lambda m_2,$$

and then the law of total variance (16) implies that

$$Var[S] = E[Var(S|\lambda)] + Var[E(S|\lambda)] =$$

$$= E[n\lambda m_2] + Var[n\lambda m_1] = 0.2nm_2 + 0.01n^2m_1^2.$$

We remark that this variance is grater than that in Example 3.9.

Example 3.11. Question: Suppose that Poisson parameters λ_i are drawn from gamma distribution with known parameters α and δ. Find a distribution for the number of claims N_i from a random policy i.

Answer: By problem formulation, conditional distribution $N_i|\lambda_i$ is Poisson distribution with parameter λ_i, while λ_i follows $Ga(\alpha, \delta)$, that is, has density $f(\lambda) = \frac{\delta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\delta\lambda}$, see (8). Then by (14), for any $m = 0, 1, 2, \ldots,$

$$P[N_i = m] = \int_0^\infty e^{-\lambda}\frac{\lambda^m}{m!}\frac{\delta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\delta\lambda}d\lambda = \frac{\delta^\alpha}{\Gamma(\alpha)m!}\int_0^\infty e^{-\lambda(\delta+1)}\lambda^{m+\alpha-1}d\lambda$$

The integral is proportional to density of gamma distribution with parameters $m+\alpha$ and $\delta+1$, hence the integration gives the corresponding cdf which can be found from tables. We get

$$P[N_i = m] = \frac{\delta^\alpha}{\Gamma(\alpha)m!}\frac{\Gamma(m+\alpha)}{(\delta+1)^{x+\alpha}}.$$

One can check that if $\alpha = k$ is positive integer, then this formula reduced for the formula for negative binomial distribution with parameters α and $\frac{\delta}{\delta+1}$.

3.8 Summary

Let N be a (random) number of claims during some fixed period of time. If the sizes of claims are $X_1, X_2, \ldots, X_N$, then the total cost to cover all claims is

$$S = X_1 + X_2 + \cdots + X_N,$$

and $S = 0$ if $N = 0$. If all X_i are independent, identically distributed, and also independent of N, then we say that S has *compound distribution.*

The cdf of S is given by

$$G(x) = P(S \leq x) = \sum_{n=0}^{\infty} P(N = n) \cdot F^{n*}(x),$$

where $F(x)$ is the cdf of the individual claim and $F^{n*}(x)$ its n-fold convolution.

We denote μ_X, μ_N, μ_S the means of random variables X_i, N, S and by σ_X^2, σ_N^2, σ_S^2 the corresponding variances. Then

$$\mu_S = \mu_N \cdot \mu_X, \quad \text{and} \quad \sigma_S^2 = \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2.$$

For example,

- If N follow a Poisson distribution with parameter λ, then S is called compound Poisson and

$$\mu_S = \lambda \mu_X, \quad \text{and} \quad \sigma_S^2 = \lambda E[X^2].$$

- If N follow a binomial distribution with parameter n and p, then S is called compound binomial and

$$\mu_S = np\mu_X. \quad \text{and} \quad \sigma_S^2 = npE[X^2] - np^2\mu_X^2.$$

- If N follow a negative binomial distribution with parameters k and p, then S is called compound binomial and

$$\mu_S = \frac{k(1-p)}{p}\mu_X \quad \text{and} \quad \sigma_S^2 = \frac{k(1-p)}{p}E[X^2] + \frac{k(1-p)^2}{p^2}\mu_X^2.$$

Under proportional reinsurance with retention level α, the i-th individual claim amount for the insurer is $Y_i = \alpha X_i$. Under the excess of loss reinsurance

with retention level M, it is $Y_i = \min(X_i, M)$. In both cases, the aggregate claim for the insurer is

$$S_I = Y_1 + Y_2 + \cdots + Y_N,$$

and all the formulas above work with Y_i in place of X_i.

For the reinsurer, the individual claim amount is $Z_i = X_i - Y_i = (1-\alpha)X_i$ under proportional reinsurance and $Z_i = X_i - Y_i = \max(0, X_i - M)$ under the excess of loss reinsurance. In both cases, the aggregate claim for the reinsurer is

$$S_R = Z_1 + Z_2 + \cdots + Z_N,$$

and all the formulas above work with Z_i in place of X_i.

In the individual risk model, the aggregate claim is

$$S = Y_1 + Y_2 + \cdots + Y_n,$$

where $Y_j = N_j X_j$, $N_j \sim Bin(1, q_j)$ is the number of claims from j-th policy, and X_j is the individual claim amount with mean m_j and variance σ_j^2. Then

$$E[S] = \sum_{j=1}^{n} q_j \mu_j, \quad \text{and} \quad Var[S] = \sum_{j=1}^{n} (q_j \sigma_j^2 + q_j(1-q_j)\mu_j^2).$$

3.9 Questions

1. Assume that the number N of claims can be any integer from 1 to 100 with equal chances, and the claim sizes $X_1, \ldots, X_N$ are i.i.d. from Pareto distribution with parameters $\alpha = 3$ and $\lambda = 2$. Estimate mean and variance of the aggregate claim S.

2. The number of claims insurance company receives in April follows Poisson distribution with $\lambda = \frac{45}{7}$, the sizes of claims are i.i.d. and follow a uniform distribution on $[1,000; 2,000]$.

   (a) Estimate the probability that the company will receive at least 3 claims in April.

   (b) Estimate the mean and variance for the total size of all April's claims.

   (c) Estimate the probability that the total size of all April's claims will be strictly less than $3,000$.

3. The claim size X follows a log-normal distribution with parameters μ and σ, where σ is known but μ is not. Instead, we model μ as another random variable such that $\lambda = e^{\mu+\sigma^2/2}$ has mean p and variance s^2. Estimate the mean and variance of X.

4. The number N of claims to be received by insurance company next year follow a negative binomial distribution with parameters $k = 20$ and $p = 0.25$. The claim sizes $X_1, \ldots, X_N$ are i.i.d. and follow exponential distribution with parameter $\lambda = 0.005$. Assuming that the aggregate claim size S is approximately normally distributed, estimate the probability that S with not exceed $20,000$.

Chapter 4
Tails and dependence analysis of claims distributions

4.1 How likely very large claims to occur?

Low frequency events involving large losses can have a devastating impact on companies and investment funds. The financial crisis that started in 2007 was an example of this. It generated more extreme movements in share prices than had been seen for over 20 years previously.

So, it is important to ensure that we model the form of the distribution in the tails correctly. However, the low frequency of these events also means that there is relatively little data to model their effects accurately.

Many types of financial data tend to be much more narrowly peaked in the centre of the distribution and to have fatter tails than the normal distribution. This shape of distribution is known as leptokurtic. For example, when share prices are modelled, large price movements occur more frequently than predicted by the normal distribution. So, the normal distribution may be unsuitable for modelling the large movements in the tails. One reason for these fat tails is that the volatility of financial variables does not remain constant but varies stochastically over time. This property is known as heteroscedasticity.

Even if we select an appropriate form of fat-tailed distribution, if we attempt to fit the distribution using the whole of our dataset, this is unlikely to result in a good model for the tails, since the parameter estimates will be heavily influenced by the main bulk of the data in the central part of the distribution.

Fortunately, better modelling of the tails of the data can be done through the application of extreme value theory. The key idea of extreme value theory is that the asymptotic behaviour of the tails of most distributions can be accurately described by certain families of distributions. More specifically, the maximum values of a distribution (when appropriately standardised) and the values exceeding a specified threshold (called threshold exceedances) converge to two families of distributions as the sample size increases.

There are a number of measures we can use to quantify the tail weight of a particular distribution, that is, how likely very large values are to occur. Depending on the context, an exponential, normal or log-normal distribution may be a suitable baseline to use for comparison.

- **Existence of moments.**

Recall that the k-th moment of a continuous positive-valued distribution with density function $f(x)$ is

$$M_k = \int_0^\infty x^k f(x) dx$$

For some distributions, like normal distribution or Gamma distribution, the density $f(x)$ decreases so fast that this integral exists and finite for every k. In other words, the k-th moment exists for all values of k. This is an indication of a relatively light tail, meaning that large claims are unlikely to happen.

However, for some distributions, the integral $\int_0^\infty x^k f(x) dx$ does not converge (becomes infinite) for all k greater than or equal to some value k_0. In this case, we say that the k-th moment does not exist for $k \geq k_0$. For example, for the Pareto distribution with density function (9) the k-th moment only exists when $k < \alpha$. So a Pareto distribution with a low value of the parameter α has a much thicker tail. If claims follow this distribution, very large claims may occur with non-negligible probability.

- **Limiting density ratios.**

  We can compare the thickness of the tail of two distributions by calculating the relative values of their density functions at the far end of the upper tail. For example, if we compare the Pareto distributions (9) with parameters $\alpha = 2$ and $\alpha = 3$ both with the same value of λ, we find that:

$$\lim_{x\to\infty} \frac{f_{\alpha=2}(x)}{f_{\alpha=3}(x)} = \lim_{x\to\infty} \left(\frac{2\lambda^2}{(\lambda+x)^3} : \frac{3\lambda^3}{(\lambda+x)^4} \right) = \frac{2}{3\lambda} \lim_{x\to\infty} (\lambda + x) = \infty.$$

  This confirms that the distribution with $\alpha = 2$ has a much thicker tail. If we compare the gamma distribution with the Pareto distribution, we find that the presence of the exponential factor in the gamma density results in a limiting density ratio of zero, which confirms that the gamma distribution has a lighter tail.

- **Hazard rate.**

  The hazard rate of a distribution with density function $f(x)$ and cumulative distribution function $F(x)$ is defined as:

$$h(x) = \frac{f(x)}{1 - F(x)}$$

The hazard rate is the analogy of force of mortality which you study in parallel module. If the force of mortality increases as a person's age increases, relatively few people will live to old age (corresponding to a light tail). If, on the other hand, the force of mortality decreases as the person's age increases, there is the potential to live to a very old age (corresponding to a heavier tail).

For example, for exponential distribution with parameter $\lambda > 0$ we have

$$f(x) = \lambda e^{-\lambda x}, \quad \text{and} \quad F(x) = 1 - e^{-\lambda x}, \quad x > 0,$$

hence the hazard rate

$$h(x) = \frac{\lambda e^{-\lambda x}}{1 - (1 - e^{-\lambda x})} = \lambda$$

is a constant. Exponential distribution corresponds to the $\alpha = 1$ case of the gamma distribution (8). Numerical calculations shows that, for gamma distribution, the hazard rate is decreasing if $\alpha > 1$ (which indicates for a heavier tail than that for the exponential distribution) and increasing if $\alpha > 1$ (which indicates for a lighter tail than that for the exponential distribution).

For the Pareto distribution (9), we find that the hazard rate is always a decreasing function of x (see the end of chapter question for the proof), confirming that it has a heavy tail.

- **Mean residual life.**

  The mean residual life of a distribution with density function $f(x)$ and cumulative distribution function $F(x)$ is defined as:

$$e(x) = \frac{\int_x^\infty (y - x) f(y) dy}{\int_x^\infty f(y) dy} = \frac{\int_x^\infty (1 - F(y)) dy}{1 - F(x)}.$$

  Again, we can interpret this in terms of mortality as the expected future lifetime. If the expected future lifetime decreases with age, relatively few people will live to old age (corresponding to a light tail), but if it increases, there is the potential to live to a very old age (corresponding to a heavier tail).

  For the gamma distribution, we find that, if $\alpha = 1$ (exponential distribution), the mean residual life is constant, but if $\alpha < 1$, it is an increasing function of x (indicating a heavier tail than the exponential distribution) and if $\alpha > 1$, it is a decreasing function (indicating a lighter tail than the exponential distribution).

For the Pareto distribution, we find that the mean residual life is an increasing function of x (see the end of chapter question for the proof), confirming that it has a heavy tail.

4.2 The distribution of large claims

As an alternative to focusing on "heavy" or "light" tail, we can consider the distribution of all the values of random variable X that exceed some (large) specified threshold, e.g. all claims exceeding 1 million pounds. As we will see, for large samples and large thresholds, the distribution of these extreme values converges to the Generalised Pareto Distribution. This enables us to model the tail of a distribution by selecting a suitably high threshold and then fitting a generalised Pareto distribution to the observed values in excess of that threshold.

Let X be a random variable with cumulative distribution function F, $u \in \mathbb{R}$ be a threshold, and let

$$F_u(x) = P(X - u \leq x | X > u)$$

denotes the conditional probability that the threshold exceedance $X - u$ is at most x under the condition that the claim X exceeded the threshold u. Then

$$F_u(x) = \frac{P(X \leq x + u, X > u)}{P(X > u)} = \frac{F(x+u) - F(u)}{1 - F(u)}.$$

For example, if the individual losses are distributed exponentially with $F(x) = 1 - e^{-\lambda x}$, we have:

$$F_u(x) = \frac{(1 - e^{-\lambda(x+u)}) - (1 - e^{-\lambda u})}{1 - (1 - e^{-\lambda u})} = \frac{e^{-\lambda u} - e^{-\lambda(x+u)}}{e^{-\lambda u}} = 1 - e^{-\lambda x} = F(x)$$

So, in this case, the threshold exceedances follow the same exponential distribution as X, irrespective of the threshold we choose.

In general, there is a theorem stating that, whatever the underlying distribution F is, the distribution of the threshold exceedances will converge to a Generalised Pareto distribution as the threshold u increases, that is,

$$\lim_{u \to \infty} F_u(x) = G(x),$$

where $G(x)$ belongs to a two-parameter family of Generalised Pareto distributions

$$G(x) = \begin{cases} 1 - \left(1 + \frac{x}{\gamma\beta}\right)^{-\gamma} & \text{if } \gamma \neq 0, \\ 1 - \exp\left(-\frac{x}{\beta}\right) & \text{if } \gamma = 0, \end{cases}$$

Parameter β is called scale parameter, while γ is called a shape parameter. When $\gamma = 0$, $G(x)$ reduces to exponential distribution with parameter $\lambda = \frac{1}{\beta}$.

4.3 The distribution of maximal claim

In the previous section we studied tails of individual random variable. In this section we assume that we have a sequence of claim sizes $X_1, X_2, \ldots, X_n, \ldots$, which are independent and identically distributed (i.i.d.), and we are interested to estimate the *maximal* claim out of first n:

$$M_n = \max\{X_1, X_2, \ldots, X_n\}.$$

This problem is of important practical interest. For example, if a company receives on average 3 claims per day, it might except to receive about $3 \cdot 365 = 1095$ claims per year, and then M_{1095} gives an intuition of how large the maximal claim in the year they should expect.

For any real x,

$$P(M_n \le x) = P(X_1 \le x, X_2 \le x, \ldots X_n \le x) =$$

$$= P(X_1 \le x)P(X_2 \le x) \cdot P(X_n \le x) = P(X_1 \le x)^n = [F_X(x)]^n,$$

where we have used that $X_1, X_2, \ldots, X_n, \ldots$ are i.i.d. and denote $F_X(x)$ their common cdf. This formula can be used to compute the distribution of M_n directly for small values of n.

However, if n is large, which is often the case in applications, that $[F_X(x)]^n$ is essentially 0 unless x is such that $F_X(x)$ is very close to 1. However, for such values of x it is difficult to estimate $F_X(x)$.

Fortunately, for large n we can apply famous Fisher-Tippett-Gnedenko theorem, also known as "extreme value theorem", which studies the distribution of M_n in the limit as $n \to \infty$. Namely, it states that if $a_1, a_2, \ldots, a_n, \ldots$ and $b_1, b_2, \ldots, b_n, \ldots$ are two sequence of real numbers such that the limit

$$F(x) = \lim_{n\to\infty} P\left(\frac{M_n - a_n}{b_n} \le x\right) = \lim_{n\to\infty} [F_X(a_n + b_n x)]^n \tag{45}$$

exists and depends only on x, then $F(x)$ must be of the form

$$F(x) = \begin{cases} \exp\left(-\left(1 + \frac{\gamma(x-\alpha)}{\beta}\right)^{-1/\gamma}\right) & \text{if } \gamma \neq 0, \\ \exp\left(-\exp(-\frac{x-\alpha}{\beta})\right) & \text{if } \gamma = 0, \end{cases} \tag{46}$$

where α, $\beta > 0$, and γ are some real parameters, and x is such that $1 + \frac{\gamma(x-\alpha)}{\beta} > 0$. If $1 + \frac{\gamma(x-\alpha)}{\beta} \le 0$, (46) is undefined, and we set $F(x) = 0$ if

$\gamma > 0$ and $F(x) = 1$ if $\gamma < 0$. With this convention, $F(x)$ becomes a cdf of some distribution, and this distribution is known as **Generalized Extreme Value (GEV)** distribution. Parameter α is called a location parameter, $\beta > 0$ is scale parameter, and γ is known as shape parameter.

Example 4.1. Assume that $X_1, X_2, \ldots, X_n, \ldots$ are i.i.d. r.v.s witch follow exponential distribution with parameter λ. Then, for the limit in (45) to exist, we should choose $a_n = \frac{1}{\lambda} \ln n$ and $\beta_n = \frac{1}{\lambda}$. In this case,

$$F(x) = e^{-e^{-x}},$$

which corresponds to (46) with parameters $\alpha = \gamma = 0$, $\beta = 1$.

The details for this example are given in an end-of-chapter question.

The parameters α and β just rescale (shift and stretch) the GEV distribution (46), in a similar way as changing mean and standard deviation shifts and stretches the normal distribution. The parameter γ determines the overall shape of the distribution (analogous to the skewness) and its sign (positive, negative or zero) results in three different shaped distributions.

- When $\gamma < 0$, the GEV distribution (46) reduces to

$$F(x) = \begin{cases} \exp\left(-\left(1 + \frac{\gamma(x-\alpha)}{\beta}\right)^{-1/\gamma}\right) & \text{if } x < \alpha - \frac{\beta}{\gamma}, \\ 1 & \text{if } x \geq \alpha - \frac{\beta}{\gamma}, \end{cases}$$

  and it is known as the Weibull distribution. For example, if the underlying distribution (the distribution F_X in (45)) follows the uniform distribution, or beta distribution, or triangular distribution, then the extreme value distribution (45) has the form (46) with $\gamma < 0$, that is, it is a Weibull distribution.

- When $\gamma = 0$, the GEV distribution reduces to

$$F(x) = \exp\left(-\exp(-\frac{x-\alpha}{\beta})\right),$$

  and it is known as the Gumbel distribution. As demonstrated in Example 4.1, the Gumbel distribution arises as the extreme value distribution when the underlying distribution (the distribution of individual claims) is exponential. Similarly, one can show that if the underlying distribution (the distribution F_X of X_i in (45)) is Chi-square, or Gamma, or Log-normal, or Normal, or Weibull, then the extreme value distribution (45) has the form (46) with $\gamma = 0$, that is, it is a Gumbel distribution.

- When $\gamma > 0$, the GEV distribution (46) reduces to

$$F(x) = \begin{cases} 0 & \text{if } x \leq \alpha - \frac{\beta}{\gamma}, \\ \exp\left(-\left(1 + \frac{\gamma(x-\alpha)}{\beta}\right)^{-1/\gamma}\right) & \text{if } x > \alpha - \frac{\beta}{\gamma}, \end{cases}$$

  and it is known as the Fretchet distribution. For example, if the underlying distribution (the distribution F_X in (45)) is Burr distribution, or F-distribution, or Log-gamma, or Pareto, or t-distribution, then the extreme value distribution (45) has the form (46) with $\gamma > 0$, that is, it is a Fretchet distribution.

If we know the form of the underlying distribution, it is possible to work out the limiting distribution of the maximum value. We can then use the appropriate member of the GEV family to model the tail of the distribution. The underlying distribution will determine which of the three different types of GEV distribution will arise. Mathematicians have determined criteria that can be used to predict which family a distribution belongs to. As a rough guide:

- Underlying distributions that have finite upper limits (e.g. the uniform distribution) are of the Weibull type (which also has a finite upper limit).
- "Light tail" distributions that have finite moments of all orders (e.g. exponential, normal, log-normal) are typically of the Gumbel type.
- "Heavy tail" distributions whose higher moments can be infinite are of the Frechet type.

4.4 Dependence, correlation, and concordance

On the previous section we assumed that claim sizes $X_1, X_2, \ldots, X_n, \ldots$, which are independent. However, this is not always an adequate and realistic assumption.

Imagine that an insurance company sells policies against fires. Most of claims comes from small fires (say, affecting just one room in a house), and the probability of a large claim, arising from fire which completely destroy the house, is estimated as about 10^{-3} for every individual policy. The company sold 10 such policies for houses on the same street, and estimated that they have reserved capital to pay at most 9 large claims. If all 10 policies make a large claim, the company become bankrupt. However, because the

probability of one claim is 10^{-3}, the company estimated that the chance that all 10 claims happen is about $(10^{-3})^{10} = 10^{-30}$. This is so tiny that can be ignored, and the company felt completely safe.

After several months, a large fire happened in one of the houses. The fire quickly spread on the neighbour houses, destroying them all. All 10 houses were affected, the company received 10 large claims and becomes bankrupt.

The companies mistake is that formula $(10^{-3})^{10} = 10^{-30}$ works only if all fires would be independent. However, they are not. Because houses were on one street, fire on one of them could course fire on others. Even if houses would be on different streets, fire on one of them could be a reason of, for example, extremely hot weather, which could increase the chance of another fire. So, understanding the dependency between different events and random variables is of fundamental importance.

A popular way to calculate "how much" two random variables are dependent, or correlated, is Pearson correlation coefficient

$$\mathrm{Corr}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}},$$

where $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$. In particular, if X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$ and therefore $\mathrm{Corr}(X,Y) = 0$.

However, one disadvantage of $\mathrm{Corr}(X,Y)$ is that the converse does not hold: we may have $\mathrm{Corr}(X,Y) = 0$ even if X and Y depend from each other in a strong way. For example, assume that there are 3 equiprobable possibilities:

$$X = -3 \quad \text{and} \quad Y = 1$$

or

$$X = 1 \quad \text{and} \quad Y = -5$$

or

$$X = 2 \quad \text{and} \quad Y = 4$$

We denote these scenarios as w_1, w_2, w_3, respectively. Then $E[X] = \frac{1}{3}(-3 + 1 + 2) = 0$, $E[Y] = \frac{1}{3}(1 - 5 + 4) = 0$, and

$$E[XY] = \frac{1}{3}(-3 \cdot 1 + 1 \cdot (-5) + 2 \cdot 4) = 0,$$

hence $\mathrm{Corr}(X,Y) = 0$. However, random variables X and Y clearly depends from each other. For example, if we know that $X = 1$ then we can conclude that $Y = -5$.

In applications, it is crucial to understand whether the dependence is direct (the higher X the higher Y) or converse (the higher X, the lower Y).

Let us think about this in our example. Consider 2 scenarios, w_1 and w_2. We see that X is higher in scenario w_2 while Y is higher in w_1. So, in this case we have converse dependence (the higher X, the lower Y). We call such pair of scenario "discordant".

Now, consider other pair of scenarios, w_1 and w_3. In this case, X is higher in scenario w_3, and Y is higher in w_3 as well. So, we have "the higher X the higher Y" case. We call such pair of scenario "concordant".

Similarly, pair of scenarios w_2 and w_3 is "concordant". Because there are more concordant pairs than discordant ones, we conclude that the dependence between X and Y is (mostly) "direct".

In general case, assume that there are n scenarios on which random variables X and Y assume different values. Then the can consider all pairs of scenarios (there are $n(n-1)/2$ pairs), and count how many concordant and discordant pairs are. If there are C concordant pairs and D discordant ones, the ratio

$$\tau = \frac{C-D}{C+D} = \frac{C-D}{n(n-1)/2}$$

is called Kendall coefficient of concordance. If $\tau > 0$, this means that dependence "the higher X the lower Y" is observed, and the closer τ to its maximal value 1, the stronger this dependence is. Conversely, if $\tau < 0$, then the dependence "the higher X the lower Y" is observed, and the closer τ to its minimal value -1, the stronger is this dependence.

In the special example considered above, $n = 3$, $n(n-1)/2 = 3$, $C = 2$, $D = 1$, and

$$\tau = \frac{2-1}{3} = \frac{1}{3}.$$

There are other ways to measure concordance, for example, Spearman coefficient (which we will not define here). If X and Y are independent, then their concordance is 0. However, as with Pearson correlation coefficient, the problem is in the different direction: we may have Kendall (or Spearman) coefficient equal to 0 even if the random variables are dependent. This just means that "positive" and "negative" dependences happens equally often and on average compensate each other.

4.5 Joint distributions and copulas

To fully understand if random variables X and Y are dependent or not, and if so, how they are dependent, we need to look at their joint distribution function. Recall that for random variable (r.v.) X its cumulative distribution function (cdf) is $F_X(x) = P[X \leq x]$, while for pair of r.v.s X and Y their

joint cdf is

$$F_{X,Y}(x,y) = P[X \le x, Y \le y].$$

If X and Y are independent, then

$$F_{X,Y}(x,y) = P[X \le x, Y \le y] = P[X \le x] \cdot P[Y \le y] = F_X(x) \cdot F_Y(y). \quad (47)$$

In the context of joint distribution functions, the individual distribution of each of the variables in isolation is known as its *marginal distributions*. (47) states that, for independent r.v.s, their joint distribution is the product of marginal distributions.

In contrast, imagine X and Y are dependent in the strongest possible way: the higher X, the higher Y (for all pairs of scenarios!). This means that $Y = f(X)$ for some strictly increasing function f. Then

$$F_{X,Y}(x,y) = P[X \le x, Y \le y] = P[f(X) \le f(x), Y \le y] = P[Y \le \min(f(x), y)] =$$

$$= \min(P[Y \le f(x)], P[Y \le y]) = \min(P[X \le x], P[Y \le y]) = \min(F_X(x), F_Y(y)).$$

Hence,

$$F_{X,Y}(x,y) = \min(F_X(x), F_Y(y)). \quad (48)$$

Now imagine the converse dependence: the higher X, the lower Y, that is, $Y = f(X)$ for some strictly *decreasing* function f. Then

$$F_{X,Y}(x,y) = P[X \le x, Y \le y] = P[f(X) \ge f(x), Y \le y] = P[f(x) \le Y \le y]$$

$$= \max(P[Y \le y] - P[Y \le f(x)], 0) = \max(P[Y \le y] - P[X \ge x], 0).$$

Hence,

$$F_{X,Y}(x,y) = \max(F_Y(y) + F_X(x) - 1, 0). \quad (49)$$

We can see that in all cases $F_{X,Y}(x,y)$ is some function of $F_X(x)$ and $F_Y(y)$:

$$F_{X,Y}(x,y) = C(F_X(x), F_Y(y)), \quad (50)$$

where $C(u,v) = u \cdot v$ in (47), $C(u,v) = \min(u,v)$ in (48), and $C(u,v) = \max(u+v-1, 0)$ in (49).

This is not a coincidence. Famous Sklar's theorem states that, for *any* pair of r.v. X and Y, (50) holds for some function C. This function is called a *copula*, and represents the way X and Y depends on each other. Specifically, copula shows how joint distribution depends on marginal distributions.

In particular,

- $C(u,v) = u \cdot v$ is called *independence copula*, which represents no dependence whatsoever.

- $C(u,v) = \min(u,v)$ is called co-monotonic (or minimum) copula, and represents the dependence in the form "the higher X, the higher Y".
- $C(u,v) = \max(u+v-1,0)$ is called counter-monotonic, or maximum copula, which represents the dependence of the form "the higher X, the lower Y".

Three copulas listed above are called *fundamental copulas*. They are specific cases of a more general family of copulas called Frechet-Hoeffding copulas.

Another important example is:

- Gaussian copula

$$C(u,v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)),$$

where Φ is the distribution function of the standard normal distribution and Φ_ρ is the distribution function of a bivariate normal distribution with correlation ρ. Applying this Gaussian copula to normal marginal distributions will result in a bivariate normal distribution with correlation ρ. Gaussian copula can be written in the integral form

$$C(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \exp\left\{-\frac{1}{2(1-\rho^2)}(s^2+t^2-2\rho st)\right\} dsdt,$$

or equivalently,

$$C(u,v) = \int_0^u \Phi\left(\frac{\Phi^{-1}(v) - \rho\Phi^{-1}(t)}{\sqrt{1-\rho^2}}\right) dt$$

Many of the commonly used copulas are special cases of the important family of copulas which is called *the Archimedean family*. Let $\psi : (0,1] \to [0,\infty)$ be a continuous, strictly decreasing, convex function with $\psi(1) = 0$. Properties of ψ imply that it has an inverse function $\psi^{-1} : [0,L) \to (0,1]$, where $L = \lim_{t\to 0} \psi(t)$. If $L < \infty$, we also define, by convention, $\psi^{-1}(x) = 0$ for $x \geq L$, so that ψ^{-1} is defined everywhere on $[0,\infty)$.

Then the Archimedean family of copulas are all copulas of the form

$$C(u,v) = \psi^{-1}(\psi(u) + \psi(v)). \tag{51}$$

Several examples are presented below.

- For
$$\psi(t) = -\ln t, \qquad 0 < t \leq 1.$$
the $C(u, v)$ in (51) reduces to
$$C(u, v) = uv.$$
Hence, in this case $C(u, v)$ is the independence copula.

- For
$$\psi(t) = (-\ln t)^{\alpha}, \qquad 0 < t \leq 1,$$
where $1 \leq \alpha$ is a parameter, the $C(u, v)$ in (51) reduces to
$$C(u, v) = \exp\left\{-\left((-\ln u)^{\alpha} + (-\ln v)^{\alpha}\right)^{1/\alpha}\right\}.$$
This copula is known as *the Gumbel copula*.

- For
$$\psi(t) = -\ln\left(\frac{e^{-\alpha t} - 1}{e^{-\alpha} - 1}\right), \qquad 0 < t \leq 1,$$
where $\alpha \neq 0$ is a parameter, the $C(u, v)$ in (51) reduces to
$$C(u, v) = -\frac{1}{\alpha}\ln\left(1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{(e^{-\alpha} - 1)}\right).$$
This copula is known as *the Frank copula*.

- For
$$\psi(t) = \frac{1}{\alpha}(t^{-\alpha} - 1), \qquad 0 < t \leq 1,$$
where $\alpha \neq 0$ is a parameter, the $C(u, v)$ in (51) reduces to
$$C(u, v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}.$$
This copula is known as *the Clayton copula*.

See the end of chapter questions for the details of all calculations.

In all cases, the value of parameter α represents the strength of the dependency between the variables.

Sklar's theorem also works for any number of random variables. It states that the joint distribution of n random variables is always a function of individual cumulative distribution functions:
$$F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)),$$

for some function C which depends on n variables, and is called the (n-variable) copula. For example,

$$C(x_1, x_2, \ldots, x_n) = x_1 \cdot x_2 \cdot \cdots \cdot x_n$$

implies that all n random variables are jointly independent, while

$$C(x_1, x_2, \ldots, x_n) = \min(x_1, x_2, \ldots, x_n)$$

is the n-variable version of the co-monotonic copula, and corresponds to the case when all variables are directly dependent (the higher is one variable, the higher are all).

The Archimedean family (51) can also be extended to the n variable case in a straighforward way:

$$C(x_1, x_2, \ldots, x_n) = \psi^{-1}(\psi(x_1) + \psi(x_2) + \cdots + \psi(x_n)), \tag{52}$$

where $\psi : (0, 1] \to [0, \infty)$ is a function satisfying the same conditions as above (continuous, strictly decreasing, convex function with $\psi(1) = 0$). Substituting different examples of ψ in (52), one can produce as many examples of n variable copulas as she/he wants.

4.6 Dependence of distribution tails

It is often the case in insurance and investment applications that large losses tend to occur together. For example, a hurricane could result in large losses on several different property insurance policies sold by the same company, or a stock market crash could lead to large losses on a number of investments in the same investment portfolio. So, the relationships between the variables at the extremes of the distributions, i.e. in the upper and lower tails, are of particular importance. These can be measured using the coefficients of upper and lower tail dependence, which can be calculated directly from the copula function.

The coefficient of lower tail dependence is defined as:

$$\lambda_L = \lim_{u \to 0+} P(X \leq F_X^{-1}(u) | Y \leq F_Y^{-1}(u)) = \lim_{u \to 0+} \frac{C(u, u)}{u}.$$

To define the upper tail dependence, we need to look at the opposite end of the marginal distributions. For any r.v. X, let

$$\bar{F}_X(x) = P[X > x] = 1 - P[X \leq x] = 1 - F_X(x).$$

Then for any two r.v.s X and Y and any two numbers x and y, we have

$$P[X > x, Y > y] = \bar{C}(\bar{F}_X(x), \bar{F}_X(y)),$$

for some function $\bar{C}$ which is called a survival copula function.

By the principle of inclusion/exclusion, we have:

$$1-P[X > x, Y > y] = P[X \leq x \text{ or } Y \leq y] = P[X \leq x]+P[Y \leq y]-P[X \leq x, Y \leq y],$$

hence,

$$1 - \bar{C}(\bar{F}_X(x), \bar{F}_X(y)) = F_X(x) + F_Y(y) - C(F_X(x), F_Y(y)),$$

or with notation $u = F_X(x)$, $v = F_Y(y)$,

$$\bar{C}(1 - u, 1 - v) = 1 - u - v + C(u, v).$$

The last equation allows to easily compute survival copula $\bar{C}$ if we know the (usual) copula C, and vice versa.

We can then define the coefficient of upper tail dependence as:

$$\lambda_U = \lim_{u \to 1-} P(X > F_X^{-1}(u) | Y > F_Y^{-1}(u)) = \lim_{u \to 0+} \frac{\bar{C}(u, u)}{u}.$$

The tail dependence can take values between 0 (no dependence) and 1 (full dependence).

Different copulas result in different levels of tail dependence. For example, the Frank copula and the Gaussian copula have zero dependence in both tails, while the Gumbel copula with parameter α has zero lower tail dependence but upper tail dependence of $2 - 2^{1/\alpha}$. The Clayton copula, on the other hand, has zero upper tail dependence but lower tail dependence of $2^{-1/\alpha}$. See the end of chapter question for the details of all calculations.

Hence, the Gumbel copula, with an appropriate value for the parameter α, might be a suitable copula to use when modelling large general insurance claims resulting from a common underlying cause.

4.7 Summary

The "heavier" is the tail of claim distribution, the more likely very large claims to occur. We can "quantify" how heavy are the tails by analysing the following:

- Existence of moments $M_k = \int_0^\infty x^k f(x)dx$, where $f(x)$ is a density of a non-negative r.v.
- Limiting density ratio: $\lim_{x\to\infty} \frac{f(x)}{g(x)}$, where $f(x)$ and $g(x)$ are two densities.
- Hazard rate $h(x) = \frac{f(x)}{1-F(x)}$, where $f(x)$ is density and $F(x)$ is cdf.
- Mean residual life $e(x) = \frac{\int_x^\infty (y-x)f(y)dy}{\int_x^\infty f(y)dy}$.

Let X be a random variable with cumulative distribution function F, $u \in \mathbb{R}$ be a threshold, and let

$$F_u(x) = P(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}.$$

Then

$$\lim_{u\to\infty} F_u(x) = G(x),$$

where $G(x)$ belongs to a two-parameter family of Generalised Pareto distributions.

Let

$$M_n = \max\{X_1, X_2, \ldots, X_n\}.$$

where $X_1, X_2, \ldots, X_n$ are i.i.d. random variables. The extreme value theorem states that if $a_1, a_2, \ldots, a_n, \ldots$ and $b_1, b_2, \ldots, b_n, \ldots$ are two sequence of real numbers such that the limit

$$F(x) = \lim_{n\to\infty} P\left(\frac{M_n - a_n}{b_n} \leq x\right)$$

exists and depends only on x, then $F(x)$ must follow the Generalized Extreme Value (GEV) distribution. It has 3 parameters: location parameter α, scale parameter $\beta > 0$, and shape parameter γ.

If $\gamma < 0$, $\gamma = 0$, and $\gamma > 0$, the GEV distribution reduces to the Weibull distribution, the Gumbel distribution, and the Fretchet distribution, respectively.

For 2 random variables X and Y, the "concordance" measures to what extend we have direct dependence of the form "the higher X the higher

Y" (this corresponds to positive concordance), and to what extend we have the opposite dependence "the higher X the lower Y" (corresponding to the negative concordance). If X and Y are independent, the concordance is 0. There are several ways to measure concoradance, one of them is the Kendall coefficient

$$\tau = \frac{C - D}{C + D},$$

where C and D are the numbers on concordant and discordant pairs of scenarios, respectively.

The joint cdf F of n random variables can be written in the form

$$F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)),$$

where $F_i(x_i)$ are cdfs of individual random variables (known as marginal distributions), and function C is called a copula and represents the way random variables depend on each other.

The Archimedean family is the family of copulas in the form:

$$C(x_1, x_2, \ldots, x_n) = \psi^{-1}(\psi(x_1) + \psi(x_2) + \cdots + \psi(x_n)),$$

where $\psi : (0, 1] \to [0, \infty)$ is a continuous, strictly decreasing, convex function with $\psi(1) = 0$.

For example,

$$C(x_1, x_2, \ldots, x_n) = x_1 \cdot x_2 \cdot \cdots \cdot x_n$$

is the independence copula, which implies that all n random variables are jointly independent, while

$$C(x_1, x_2, \ldots, x_n) = \min(x_1, x_2, \ldots, x_n)$$

is the co-monotonic copula, which corresponds to the case when all variables are directly dependent (the higher is one variable, the higher are all). For $n = 2$,

$$C(x_1, x_2) = \max(x_1 + x_2 - 1, 0)$$

is called counter-monotonic copula, which represents the inverse dependence of the form "the higher one variable, the lower another one".

The coefficients of lower and upper tail dependence of 2 random variables are

$$\lambda_L = \lim_{u \to 0+} \frac{C(u, u)}{u}, \quad \text{and} \quad \lambda_U = \lim_{u \to 0+} \frac{\bar{C}(u, u)}{u},$$

where $\bar{C}$ is the survival copula defined by $\bar{C}(1-u, 1-v) = 1-u-v+C(u, v)$.

4.8 Questions

1. (a) Calculate the Hazard rate of the Pareto distribution. Check if it is an increasing or decreasing function.

   (b) Calculate the Mean residual life of the Pareto distribution. Check if it is an increasing or decreasing function.

   (c) What conclusion about tails of Pareto distribution can we make based on items (a) and (b).

2. Prove the formula for $F(x)$ is Example 4.1.

3. (a) Check that function $\psi(t) = -\ln t$, $0 < t \leq 1$, is continuous, strictly decreasing, convex, and $\psi(1) = 0$;

   (b) Find its inverse function ψ^{-1};

   (c) Find $C(u, v) = \psi^{-1}(\psi(u) + \psi(v))$, see (51).

4. Repeat the previous question for

   (a) $\psi(t) = (-\ln t)^{\alpha}$, $0 < t \leq 1$, where $\alpha \geq 1$ is a parameter;

   (b) $\psi(t) = -\ln\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$, $0 < t \leq 1$, where $\alpha \neq 0$ is a parameter;

   (c) $\psi(t) = \frac{1}{\alpha}(t^{-\alpha} - 1)$, $0 < t \leq 1$, where $\alpha \neq 0$ is a parameter.

5. Calculate the coefficients of lower and upper tail dependence of two random variables with

   (a) the independence copula,

   (b) the Gumbel copula with $\alpha \geq 1$,

   (c) the Frank copula with $\alpha \neq 0$,

   (d) the Clayton copula with $\alpha > 0$.

Chapter 5
Markov Chains

The Markov property is used extensively in the Actuarial Mathematics to develop two-state and multi-state Markov models of mortality and other decrements. The rest of this course is devoted to a thorough description of the Markov property in a general context and its applications to actuarial modelling.

We will distinguish between two types of stochastic process that possess the Markov property: Markov chains and Markov jump processes. Both have a discrete state space, but Markov chains have a discrete time set and Markov jump processes have a continuous time set.

We begin with Markov chains and discuss the mathematical formulation of such process, leading to one important actuarial application: the no-claims-discount process used in motor insurance. We then move onto Markov jump processes.

The practical considerations of applying these models in the Actuarial Mathematics will be discussed in detail in later sections. In this chapter we focus on the mathematical development of Markov models without reference to their calibration to real data.

5.1 The Markov property

A major simplification to the general stochastic processes discussed in Section 1.9 occurs if the development of a process can be predicted from its current state, i.e. without any reference to its past history. This is the *Markov property*.

In this chapter we are concerned with the stochastic process $\{X_t\}$ defined on a state space $\mathcal{S}$ and time set $t \geq 0$.

The Markov property can be stated mathematically as

$$P\left[X_t \in \mathcal{A} | \, X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n, X_s = x\right] = P[X_t \in \mathcal{A} | X_s = x], \tag{53}$$

for all $s_1 < s_2 < \cdots < s_n < s < t$ in the time set, all states $x_1, x_2, \ldots, x_n, x$ in the state space $\mathcal{S}$ and all subsets of $\mathcal{A}$ of $\mathcal{S}$.

Working with subsets $\mathcal{A} \subseteq \mathcal{S}$ is necessary so that the above definition of the Markov property covers both the discrete and continuous state spaces. Recall from Chapter 1 that in the continuous case the probability that X_t is a particular value is zero, and so it is necessary to work with probabilities of X_t lying in some subset of the state space in any general definition.

Although we are entirely concerned with discrete state spaces in this chapter, it is important to realise that the Markov property can be possessed by general stochastic processes.

For discrete state spaces the Markov property is written as

$$P\left[X_t = a | \, X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n, X_s = x\right] = P[X_t = a | X_s = x], \tag{54}$$

for all $s_1 < s_2 < \cdots < s_n < s < t$ and all states $a, x_1, x_2, \ldots, x_n, x$ in $\mathcal{S}$.

An important result is that *any process with independent increments has the Markov property.*

Example 5.1. Question: Prove that any process with independent increments has the Markov property.

Answer: We begin with equation (53) and use the fact that $X_t = X_t - X_s + x$ to introduce an increment

$$\begin{aligned}
&P[\, X_t \in \mathcal{A} |\, X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n, X_s = x], \\
&= P\left[\, X_t - X_s + x \in \mathcal{A} |\, X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n, X_s = x\right], \\
&= P\left[\, X_t - X_s + x \in \mathcal{A} |\, X_s = x\right] = P\left[\, X_t \in \mathcal{A} |\, X_s = x\right],
\end{aligned}$$

the second equality arises from the definition of independent increments and the fact that x is known.

A Markov process with a discrete state space and a discrete time set is called a Markov chain, these are consider in this chapter. A Markov process with discrete state space and continuous time set is called a Markov jump process, these are considered in the next chapter.

5.2 Definition of Markov Chains

A *Markov chain* is a discrete-time stochastic process with a countable state space $\mathcal{S}$, obeying the Markov property. It is therefore a sequence of random variables $\{X_t\}$ with the property given by equation (54) which we rewrite for notational convenience as

$$P\left[\, X_n = j |\, X_0 = i_0, X_1 = i_1, \ldots, X_{m-1} = i_{m-1}, X_m = i\right] = P[X_n = j | X_m = i],$$

for all *integer* times $n > m$ and states $\{i_0, i_1, \ldots, i_{m-1}, i, j\} \in \mathcal{S}$.

We define the *transition probabilities* as

$$p_{ij}(n, n+1) = P\big[X_{n+1} = j |\, X_n = i\big]. \tag{55}$$

Therefore, $p_{ij}(n, n+1)$ is the probability of being in state j at time $n+1$, having been in state i at time n.

For each fixed $n = 0, 1, \ldots$, we can form a matrix of transition probabilities from time n to the next time step $n+1$:

$$\mathbf{P}(n, n+1) = \big[p_{ij}(n, n+1)\big]_{i,j \in \mathcal{S}}. \tag{56}$$

Note that $\mathbf{P}(n, n+1)$ is a finite matrix in the case of a finite number of states, and an infinite matrix in the case of an infinite number of states.

Example 5.2. Question: Consider a no claims discount (NCD) model for car-insurance premiums. The insurance company offers discounts of 0%, 30% and 60% of the full premium, determined by the following rules:

1. All new policyholders start at the 0% level.
2. If no claim is made during the current year the policyholder moves up one discount level, or remains at the 60% level.
3. If one or more claims are made the policyholder moves down one level, or remains at the 0% level.

The insurance company believes that the chance of claiming each year is independent of the current discount level and has a probability of 1/4. Why can this process be modeled as a Markov chain? Give the state space and transition matrix.

Answer: The model can be considered as a Markov chain since the future discount depends only on the current level, not the entire history. The state space is $\mathcal{S} = \{0\%, 30\%, 60\%\}$, which is convenient to denote as $\mathcal{S} = \{0, 1, 2\}$ (where state 0 is the 0% state, 1 is the 30% state and 2 is the 60% state). The transition probability matrix between two states in a unit time is given by

$$\mathbf{P} = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}. \tag{57}$$

A matrix $\mathbf{A}$ is called a *stochastic matrix* if

1. All its entries are non-negative, and
2. The sum of entries in any row is one.

It is clear that the transition matrix in Example 5.2 is a stochastic matrix by this definition. More generally, *every* transition matrix $\mathbf{P}(n, n+1)$ of a Markov chain is a stochastic matrix. Indeed, all the transition probabilities $p_{ij}(n, n+1)$ are by definition non-negative, and $\sum_{j \in \mathcal{S}} p_{ij} = 1$ for all i since the system must move to some state from any state i.

A clear way of representing Markov chains is by a *transition graph*. The states are represented by circles linked by arrows indicating each possible transition. Next to each arrow is the corresponding transition probability.

Example 5.3. Question: Draw the transition graph for the NCD system defined in Example 5.2.

Answer: See Figure 1

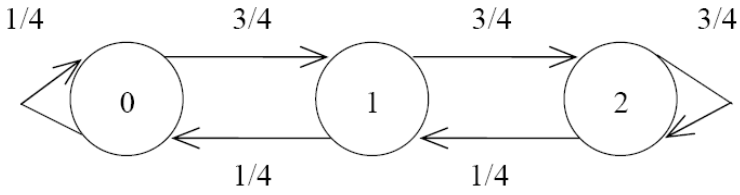


Figure 1: Transition graph for the NCD system of Example 5.2. *Reproduced with permission of the Faculty and Institute of Actuaries.*

5.3 The Chapman-Kolmogorov equations

Equation (55) defines the probabilities of transition over a single time step. Similarly, the n-step transition probabilities $p_{ij}(m, m+n)$ denote the probability that a process in state i at time m will be in state j at time $m+n$. That is:

$$p_{ij}(m, m+n) = P\left[X_{m+n} = j \middle| X_m = i\right].$$

The transition probabilities of a Markov process satisfy the system of equations called the *Chapman–Kolmogorov equations*

$$p_{ij}(m, n) = \sum_{k \in \mathcal{S}} p_{ik}(m, l) p_{kj}(l, n), \tag{58}$$

for all states $i, j \in \mathcal{S}$ and all integer times $m < l < n$. This can be expressed in terms of n-step stochastic matrices as

$$\mathbf{P}(m, n) = \mathbf{P}(m, l)\mathbf{P}(l, n),$$

where $\mathbf{P}(m, l)\mathbf{P}(l, n)$ is the product of matrices in the usual sense.

Example 5.4. Question: Prove equation (58).

Answer: We use the Markov property and the law of total probability.

$$\begin{aligned} p_{ij}(n,m) &= P\left(X_m = j \middle| X_n = i\right), \\ &= \sum_{k\in\mathcal{S}} P\left(X_m = j, X_l = k \middle| X_n = i\right), \\ &= \sum_{k\in\mathcal{S}} P\left(X_m = j \middle| X_l = k, X_n = i\right) P\left(X_l = k \middle| X_n = i\right), \\ &= \sum_{k\in\mathcal{S}} P\left(X_m = j \middle| X_l = k\right) P\left(X_l = k \middle| X_n = i\right). \end{aligned}$$

Which is the required result.

The Chapman–Kolmogorov equations provide a method for computing the n-step transition probabilities from the one-step transition probabilities. The distribution of a Markov chain is therefore fully determined once the following are specified:

- The one-step transition probabilities $p_{ij}(n, n+1)$.
- The initial probability distribution $\alpha_{j_0} = P\left(X_0 = j_0\right)$.

The probability of any path can then be determined from

$$P\left[X_0 = j_0, X_1 = j_1, \ldots, X_N = j_N\right] = \alpha_{j_0} \prod_{n=0}^{N-1} p_{j_n, j_{n+1}}(n, n+1). \tag{59}$$

This should be intuitively clear but a formal proof is left as a question at the end of the chapter.

5.4 Time dependency of Markov chains

For a *time-inhomogeneous* Markov chain, the transition probabilities $p_{ij}(t, t+1)$ change with time t. The transition probabilities will therefore have a sequence of stochastic matrices denoted by $\mathbf{P}(t)$:

$$\mathbf{P}(t) = \left[p_{ij}(t,t+1)\right]_{i,j\in\mathcal{S}} = \begin{bmatrix} p_{00}(t,t+1) & p_{01}(t,t+1) & \ldots \\ p_{10}(t,t+1) & p_{11}(t,t+1) & \ldots \\ \vdots & \vdots & \end{bmatrix}$$

The value of t can represent many factors such as time of year, age of policyholder or the length of time the policy has been in force. For example,

young drivers and very old drivers may have more accidents than middle-aged drivers and therefore t might represent the age or age group of the driver purchasing a motor insurance policy.

Although time-inhomogeneous models are important in practical modelling, a further analysis is beyond the scope of this course.
A Markov chain is called *time homogeneous* if transition probabilities do not depend on time. This is a significant simplification to any Markov-chain model. In particular, for a time-homogeneous Markov chain, equation (59) becomes

$$P\left[X_n = j_n, n = 0, 1, 2 \ldots, N\right] = P\left[X_0 = j_0\right] \prod_{n=0}^{N-1} p_{j_n j_{n+1}}. \tag{60}$$

It is therefore clear that the matrix of the n-step transition probabilities is the n-th power of the matrix of 1-step transition probabilities $\{p_{ij}\}$:

$$P\left[X_{n+m} = j \middle| X_m = i\right] := p_{ij}^{(n)} = \sum_{k_1, k_2, \ldots, k_{n-1}} \underbrace{p_{ik_1} p_{k_1 k_2} \cdots p_{k_{n-2} k_{n-1}} p_{k_{n-1} j}}_{n \text{ terms}}.$$

If we let $\mathbf{P}^{(n)}$ denote the n-step transition matrix, then

$$\mathbf{P}^{(n)} = \mathbf{P}^n,$$

where $\mathbf{P}$ is the one-step matrix of transition probabilities.

Example 5.5. Question: Calculate the 2-step transition matrix for the NCD system from Example 5.2 and confirm that it is a stochastic matrix.

Answer: The 1-step transition matrix is given by equation (57) and so we can compute that

$$\mathbf{P}^{(2)} = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix} \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix} = \frac{1}{16} \begin{pmatrix} 4 & 3 & 9 \\ 1 & 6 & 9 \\ 1 & 3 & 12 \end{pmatrix}.$$

We note that the two conditions for $\mathbf{P}^{(2)}$ to be a stochastic matrix are satisfied.

Example 5.6. Question: Using the 2-step transition matrix from Example 5.5 state the probabilities that

(a) A policyholder initially in the 0%-state is in the 60%-state after 2 years.

(b) A policyholder initially in the 60%-state is in the 30%-state after 2 years.

(c) A policyholder initially in the 0%-state is in the 0%-state after 2 years.

Answer:

(a) Element $(\mathbf{P}^{(2)})_{1,3}$ gives the required probability, 9/16.

Note that this is consistent with the path $0\% \to 30\% \to 60\%$, i.e. no claims for two years, therefore the probability is $3/4 \times 3/4 = 9/16$.

(b) $(\mathbf{P}^{(2)})_{3,2} = 3/16$.

Note that this is consistent with the path $60\% \to 60\% \to 30\%$, therefore the probability is $3/4 \times 1/4 = 3/16$.

(c) $(\mathbf{P}^{(2)})_{1,1} = 4/16$.

Note that this is consistent with either path $0\% \to 0\% \to 0\%$ or $0\% \to 30\% \to 0\%$, therefore the probability is $1/4 \times 1/4 + 3/4 \times 1/4 = 4/16$.

5.5 Further applications

The simple NCD system of Example 5.2 gives a practical example of a time-homogeneous Markov chain. We now consider three further examples.

5.5.1 The simple (unrestricted) random walk

A *simple random walk* is a stochastic process $\{X_t\}$ with state space $\mathcal{S} = \mathbb{Z}$ i.e. the integers. The process is defined by

$$X_n = Y_1 + Y_2 + \cdots + Y_n,$$

where $Y_1, Y_2, \ldots$ are a sequence of i.i.d. Bernoulli variables such that

$$P(Y_i = 1) = p \quad \text{and} \quad P(Y_i = -1) = 1 - p.$$

The simple random walk has the Markov property, that is:

$$\begin{aligned}
&P\left(X_{m+n} = j \middle| X_1 = i_1, X_2 = i_2, \ldots, X_m = i\right), \\
&\quad = P\left(X_m + Y_{m+1} + Y_{m+2} + \cdots + Y_{m+n} = j \middle| X_1 = i_1, X_2 = i_2, \ldots, X_m = i\right), \\
&\quad = P\left(Y_{m+1} + Y_{m+2} + \cdots + Y_{m+n} = j - i\right), \\
&\quad = P\left(X_{m+n} = j \middle| X_m = i\right).
\end{aligned}$$

Hence a simple random walk is a time-homogeneous Markov chain with transition probabilities:

$$p_{ij} = \begin{cases} p & \text{if} \quad j = i+1 \\ 1-p & \text{if} \quad j = i-1 \\ 0 & \text{otherwise} \end{cases}$$

Since we are considering an unrestricted simple random walk, the transition graph (Figure 2) and 1-step transition matrix are infinite.

In particular the 1-step transition matrix is given by

$$\mathbf{P} = \begin{bmatrix} \ddots & \ddots & & & & & & \\ \ddots & 0 & p & & & & & \\ & 1-p & 0 & p & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & 1-p & 0 & p & & \\ & & & & 1-p & 0 & \ddots & \\ & & & & & \ddots & \ddots & \end{bmatrix}.$$

It is clear that this is a stochastic matrix.

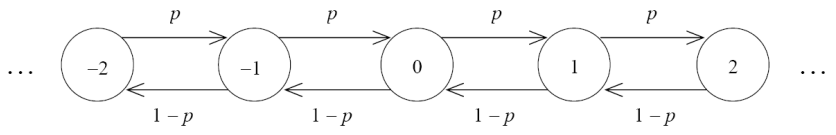


Figure 2: Transition graph for the unrestricted random walk. *Reproduced with permission of the Faculty and Institute of Actuaries.*

To determine the n-step transition probabilities, consider moving from state i to state j in n steps. Let the number of positive steps be r, (that is, r is the total number of steps where $X_{i+1} - X_i = 1$), and the number of negative steps be l, (that is, l is the total number of steps where $X_{i+1} - X_i = -1$).

Since there are n steps in total, it follows that $r + l = n$ and that $r - l = j - i$, the excess of positive steps over negative steps. Solving these simultaneous

equations for r and l gives

$$r = \frac{1}{2}(n + j - i) \quad \text{and} \quad l = \frac{1}{2}(n - j + i).$$

From this we can see that the n-step transition probabilities are

$$p_{ij}^{(n)} = \binom{n}{\frac{1}{2}(n + j - i)} p^{\frac{1}{2}(n+j-i)}(1-p)^{\frac{1}{2}(n-j+i)},$$

where $\binom{n}{r}$ is the number of possible paths with r positive steps, each of which occurs with probability $p^r(1-p)^{n-r}$. The expression arises since the distribution of the number of positive steps in n steps is Binomial with parameters n and p. Since r and l must be non-negative integers, it follows that both $n + j - i$ and $n - j + i$ must be non-negative even numbers.

In addition to being time-homogeneous, a simple random walk is spatially-homogeneous, that is

$$p_{ij}^{(n)} = P\left(X_n = j \middle| X_0 = i\right) = P\left(X_n = j + r \middle| X_0 = i + r\right).$$

A simple random walk with $p = q = \frac{1}{2}$ is called a *symmetric simple random walk*.

5.5.2 The restricted random walk

We introduce the *restricted random walk* with an example:

A man needs to raise $£N$ to fund a specific project and asks his very rich friend to accompany him to a casino where he hopes to win this money. The man plays the following game: A fair coin is tossed. If it lands heads-up the man wins $£1$; if it lands tails-up the man loses $£1$. If he loses all his money he will borrow $£1$ from his friend and continue to play until he has the required $£N$. Once he has accumulated $£N$ he will stop playing the game.

The restricted random walk is therefore a simple random walk with boundary conditions. In this example the boundary conditions are specified at 0 and N. At N the barrier is an *absorbing barrier*, while at 0 it is called a *reflecting barrier*.

More formally, an *absorbing barrier* is a value b such that:

$$P\left(X_{n+s} = b \middle| X_n = b\right) = 1 \quad \text{for all } s > 0.$$

In other words, once state b is reached, the random walk stops and remains in this state thereafter.

A *reflecting barrier* is a value c such that:

$$P\left(X_{n+1} = c + 1 \middle| X_n = c\right) = 1.$$

In other words, once state c is reached, the random walk is "pushed away".

A *mixed barrier* is a value d such that:

$$P\left(X_{n+1} = d \middle| X_n = d\right) = \alpha \quad \text{and} \quad P\left(X_{n+1} = d + 1 \middle| X_n = d\right) = 1 - \alpha,$$

for all $s > 0$ and $\alpha \in [0, 1]$. In other words, once state d is reached, the random walk remains in this state with probability α or moves to the neighboring state $d + 1$ with probability $1 - \alpha$, i.e. it is an absorbing barrier with probability α and a reflecting barrier with probability $1 - \alpha$.

If, in the example, above the man does not take his rich friend, he will continue to gamble until either his money reaches the target N or he runs out of money. In each case reaching the boundary means that the wealth will remain there forever; the barriers therefore become absorbing barriers.

The transition graph for the general case of a restricted random walk with two mixed barriers is given in Figure 3. The special cases of reflecting and absorbing boundary conditions are obtained by taking α or β equal to 0 or 1.

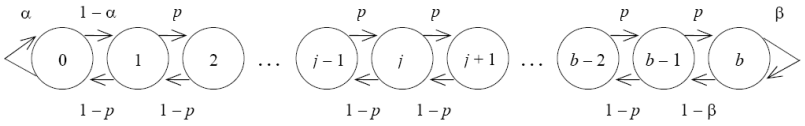


Figure 3: Transition graph for the restricted random walk with mixed boundary conditions. *Reproduced with permission of the Faculty and Institute of Actuaries.*

The 1-step transition matrix is given by

$$\mathbf{P} = \begin{bmatrix} \alpha & 1-\alpha & & & & & & \\ 1-p & 0 & p & & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & 1-p & 0 & p & & \\ & & & & 1-p & 0 & p & \\ & & & & & 1-\beta & \beta \end{bmatrix}.$$

Note that the matrix is finite, which is in contrast to the transition matrix for the unrestricted random walk.

The simple NCD model given in Example 5.2 is a practical example of a restricted random walk.

5.5.3 The modified NCD model

The simple NCD model given in Example 5.2 can be modified with a number of improvements. One such improvement is to have the following states:

State 0: no discount
State 1: 25% discount
State 2: 40% discount
State 3: 60% discount

The transition rules are as before except that when there is a claim during the current year, the discount status moves down either two levels if there was a claim in the previous year, or one level if the previous year was claim-free.

It is clear that the discount status of a policyholder at time n, X_n, does not form a Markov chain since the future discount status does not only depend on the current status but also on the previous year's status.

For example

$$P\left[X_{n+1} = 1 \middle| X_n = 2, X_{n-1} = 1\right] > 0, \tag{61}$$

whereas

$$P\left[X_{n+1} = 1 \middle| X_n = 2, X_{n-1} = 3\right] = 0. \tag{62}$$

A Markov chain can be constructed from this non-Markov chain by splitting state 2 into two states defined as:

- 2^+: 40% discount and no claim in the previous year, that is, the state corresponding to $\{X_n = 2, X_{n-1} = 1\}$.

- 2^-: 40% discount and claim in the previous year, that is, the state corresponding to $\{X_n = 2, X_{n-1} = 3\}$.

Assuming that the probability of making no claims in any year is still 3/4, the Markov chain on the modified state space $\mathcal{S}' = \{0, 1, 2^+, 2^-, 3\}$ has transition graph given by Figure 4, and 1-step transition matrix given by

$$\mathbf{P} = \begin{bmatrix} 1/4 & 3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 1/4 & 0 & 0 & 3/4 \\ 1/4 & 0 & 0 & 0 & 3/4 \\ 0 & 0 & 0 & 1/4 & 3/4 \end{bmatrix}.$$

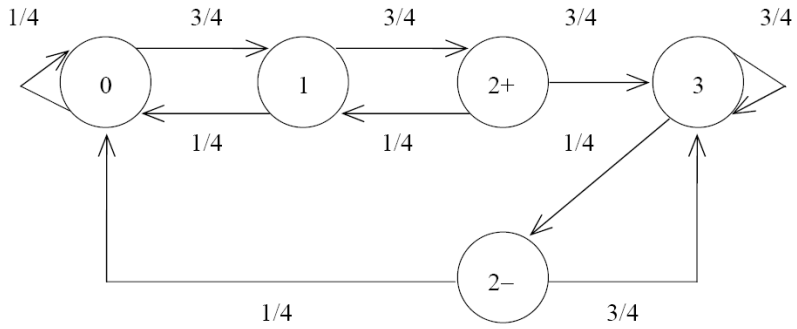


Figure 4: Transition graph for the modified NCD process. *Reproduced with permission of the Faculty and Institute of Actuaries.*

Note that a policyholder can only be in state 2^+ by moving up from state 1; and in state 2^- by moving down from state 3. Hence equations (61) and (62) become

$$P\left[X_{n+1} = 1 | \, X_n = 2^+\right] = 1/4,$$

and

$$P\left[X_{n+1} = 1 | \, X_n = 2^-\right] = 0.$$

respectively. The transition probabilities are now determined by the current discount status only and the process is Markov.

5.5.4 A model of accident proneness

An insurance company may want to use the whole history of the claims from a given driver to estimate his/her accident proneness. Let Y_i be a number of claims during period i. In the simplest model, we may assume that it can be no more than 1 claim per period, so Y_i is either 0 or 1. By time t a driver has a history of the form $Y_1 = y_1, Y_2 = y_2, \dots, Y_t = y_t$, where $y_i \in \{0, 1\}, i = 1, \dots, t$. Based of this history, the probability of future claim can be estimated, say, as

$$P[Y_{t+1} = 1 | Y_1 = y_1, Y_2 = y_2, \dots, Y_t = y_t] = \frac{f(y_1 + y_2 + \cdots + y_t)}{g(t)},$$

where f, g are two given increasing functions satisfying $0 \le f(m) \le g(m), \forall m$.

The stochastic process $\{Y_t, t = 0, 1, 2 \dots\}$ does not have the Markov property (54). However, the cumulative number of claims from the driver, given by

$$X_t = \sum_{i=1}^{t} Y_t$$

satisfies (54). Indeed,

$$\begin{aligned} &P[X_{t+1} = 1 + x_t | X_1 = x_1, X_2 = x_2, \dots, X_t = x_t] \\ &= P[Y_{t+1} = 1 | Y_1 = x_1, Y_2 = x_2 - x_1, \dots, Y_t = x_t - x_{t-1}] = \frac{f(x_t)}{g(t)}, \end{aligned}$$

which does not depend on the past history $x_1, x_2, \dots, x_{t-1}$. Thus, $\{X_t, t = 0, 1, 2 \dots\}$ is a Markov chain.

5.5.5 A model for credit rating dynamics

Credit rating dynamics of a bond is often represented by a Markov chain. There are finitely many possible ratings, such as AAA, AA, A, BBB, etc., with AAA being the highest rating, corresponding to the lowest probability of default, AA is the next one, and so on. There may be subdivisions inside any class, like AA (high), AA, and AA (low), etc. In any case, we can associate a state of Markov chain to every rating, and there is also a special state (say, D), corresponding to the default. The default state is absorbing, that is, transition probability from D to any state $i \neq D$ is 0.

Example 5.7. Assume that there are just 2 ratings for a bond B: I - investment grade, and J - junk grade. This can be modelled as a Markov

chain with states I, J, and D - default. Assume that the transition matrix is given by

$$\mathbf{P} = \begin{pmatrix} 0.90 & 0.05 & 0.05 \\ 0.10 & 0.80 & 0.10 \\ 0 & 0 & 1 \end{pmatrix}$$

Assume that bond B returns yearly profit 9% of the investment in state I, and 10% of the investment in state J. However, in case of default you will be able to get back only 40% of your investment. Assume also that risk-free rate (in a bank) is 2%.

In this case, investing capital C in a bond in state I, we will get back $1.09C$ in case we stay at I or move to J, and $0.4(1.09C)$ if the process moves to state D. Hence, our expected profit is $0.90(1.09C)+0.05(1.09C)+0.05(0.4\cdot 1.09C) - C = 0.0573C$, that is, 5.73%. This is higher than risk-free rate 2%. The difference is called *premium for risk*.

A risk neutral probability measure is an "adjusted" probability measure such that the expected profit (with respect to this measure) from a bond is the same as the profit from the bank. The "adjusted" transition probabilities from state I to states I, J, and D are given by $1-0.1\pi_I$, $0.05\pi_I$, $0.05\pi_I$, where π_I is the *adjustment coefficient*. By definition of risk neutral probability measure, $(1-0.10\pi_I)(1.09C) + 0.05\pi_I(1.09C) + 0.05\pi_I(0.4\cdot 1.09C) - C = 0.02C$, from which we can find π_I. Similarly, "adjusted" transition probabilities from state J to states I, J, and D are given by $0.10\pi_J$, $1-0.20\pi_J$, $0.10\pi_J$, where π_J can be found from a similar argument (this is left as an exercise).

5.5.6 General principles of modelling using Markov chains

In this section we summarise the examples above and identify the key step in modelling real-life situations using Markov chains. For simplicity, we discuss only time-homogeneous models here.

- **Step 1. Setting up the state space**: The most natural choice is usually to identify the state space of the Markov chain with a set of observations. For example, this is the case for the NCD system from Example 5.2, where we setted up the state space $\mathcal{S} = \{0, 1, 2\}$ in correspondence with possible discounts $\{0\%, 30\%, 60\%\}$. However, as we saw in the example with the modified NCD system (see section 5.5.3), such natural state space may be not suitable to form a Markov chain, because the Markov property may fail. In the modified NCD example, a small modification of the state space allowed us to construct a Markov chain.

- **Step 2. Estimation transition probabilities**: Once the state space is determined, the Markov model must be fitted to the data by estimating the transition probabilities. In the NCD model (Example 5.2) we have just claimed that "the company believes that the chance of claiming each year ... has a probability of 1/4". In practice, however, the transition probabilities should be estimated from the data. Naturally, the probability p_{ij} of transition from state i to state j should be estimated as number of transitions from i to j, divided by the total number of transitions from state i. More formally, let $x_1, x_2, \ldots, x_N$ be a set of available observations, n_i be the number of times t $(1 \leq t \leq N-1)$ such that $x_t = i$, and n_{ij} be the number of times t $(1 \leq t \leq N-1)$ such that $x_t = i$ and $x_{t+1} = j$. Then the best estimate for p_{ij} is $\hat{p_{ij}} = \frac{n_{ij}}{n_i}$, and the 95% confidence interval can be approximated as

$$\left(\hat{p_{ij}} - 1.96\sqrt{\frac{\hat{p_{ij}}(1-\hat{p_{ij}})}{n_i}},\ \hat{p_{ij}} + 1.96\sqrt{\frac{\hat{p_{ij}}(1-\hat{p_{ij}})}{n_i}} \right)$$

  This follows from the fact that the conditional distribution of N_{ij} given N_i is binomial with parameters N_i and p_{ij}.

- **Step 3. Checking the Markov property**: Once the state space and transition probabilities are found, the model is fully determined. But, to ensure that the fit of the model to the data is adequate, we need to check that the Markov property seems to hold. In practice, it is often considered sufficient to look at triplets of successive observations. For a set of observations $x_1, x_2, \ldots, x_N$, let n_{ijk} be the number of times t $(1 \leq t \leq N-2)$ such that $x_t = i$, $x_{t+1} = j$, and $x_{t+2} = k$. If the Markov property holds, n_{ijk} is an observation from a Binomial distribution with parameters n_{ij} and p_{jk}. An effective test to check this is a χ^2 test: the statistic

$$X^2 = \sum_i \sum_j \sum_k \frac{(n_{ijk} - n_{ij}\hat{p_{jk}})^2}{n_{ij}\hat{p_{jk}}}$$

  should approach the χ^2 distribution with $r = |\mathcal{S}|^3$ degrees of freedom. For example, if $|\mathcal{S}| = 4$, the statistic X^2 does not exceed the critical level 83.675 with probability 95%. Thus, exceeding this level is a strong indication that the Markov property does not hold. The Chi-square distribution table up to the level $r = 1000$ can be found, say, at http://www.medcalc.org/manual/chi-square-table.php

- **Step 4. Using the model**: Once the model parameters are determined, and Markov property checked, we can use the established

model to estimate different quantities of interest. In particular, we have used the Markov model for the Example 5.2 to address questions like "What is the probability that a policyholder initially in the 0%-state is in the 0%-state after 2 years?" (see Example 5.6). If the Markov model is too complicated to answer questions of this type analytically, we can use Monte-Carlo simulation (see chapter 2). Simulating a time-homogeneous Markov chain is relatively straightforward. In addition to commercial simulation packages, even standard spreadsheet software can easily cope with the practical aspects of estimating transition probabilities and performing a simulation.

5.6 Stationary distributions

In many cases the distribution of X_n converges to a limit π in the sense that

$$P\left(X_n = j \middle| X_0 = i\right) \to \pi_j, \tag{63}$$

and the limit is the same regardless of the starting point.

The distribution $\{\pi_j\}_{j\in\mathcal{S}}$ is said to be a *stationary distribution* of a Markov chain with transition matrix $\mathbf{P}$ if

1. $\pi_j = \sum_{i\in\mathcal{S}} \pi_i p_{ij}$ for all j, which can be expressed as $\pi = \pi\mathbf{P}$ where π is a row vector and $\pi\mathbf{P}$ is the usual vector-matrix product; and
2. $\pi_j \geq 0$ for all j and $\sum_{j\in\mathcal{S}} \pi_j = 1$.

The interpretation of π is that, if the initial probability distribution of π, i.e. $\pi_i = P\left(X_0 = i\right)$, then at time 1 the probability distribution of X_1 is again given by π. Mathematically

$$\begin{aligned} P\left(X_1 = j\right) &= \sum_{i\in\mathcal{S}} P\left(X_1 = j \middle| X_0 = i\right) P\left(X_0 = i\right), \\ &= \sum_{i\in\mathcal{S}} \pi_i p_{ij} = \pi_j. \end{aligned}$$

By induction we have that

$$\begin{aligned} P\left(X_n = j\right) &= \sum_{i\in\mathcal{S}} P\left(X_n = j \middle| X_{n-1} = i\right) P\left(X_{n-1} = i\right), \\ &= \sum_{i\in\mathcal{S}} \pi_i p_{ij} = \pi_j. \end{aligned}$$

Hence if the *initial distribution* for a Markov chain is a stationary distribution, then X_n has the same probability distribution for all n.

A general Markov chain does not necessarily have a stationary probability distribution, and if it does it need not be unique. For instance, the unrestricted random walk discussed in §5.5 has no stationary distribution, and the uniqueness of the stationary distribution in the restricted random walk depends on the parameters α and β.

However it is known that a Markov chain with finite state space has at least one stationary probability distribution. This is stated without proof.

Whether the stationary distribution is unique is more subtle and requires that we consider only *irreducible* chains. This is defined by the property that any state j can be reached from any other state i in a finite number of steps. In other words, a chain is irreducible if for any pair of states i and j there exists an integer n such that $p_{ij}^{(n)} > 0$. It is often sufficient to view the transition graph to determine whether a Markov chain is irreducible or not.

Example 5.8. **Question:** Are the simple NCD, modified NCD, unrestricted and restricted random walk processes irreducible?

Answer: It is clear from Figures 1, 2 & 4 that both NCD processes and the unrestricted random walks are irreducible as all states have a non-zero probability of being reached from any other state in a finite number of steps. For the restricted random walk, Figure 3 shows that it is irreducible unless either boundary is absorbing, i.e. it is irreducible for $\alpha \neq 1$ or $\beta =\neq 1$.

An irreducible Markov chain with a finite state space has a unique stationary probability distribution. This is stated without proof.

Example 5.9. **Question:** Do the simple NCD, modified NCD, unrestricted and restricted random walk processes have a unique stationary distribution?

Answer: The simple NCD process is irreducible and has a finite state space. It therefore has a unique stationary distribution.

The modified NCD process is irreducible and has a finite state space. It therefore has a unique stationary distribution.

The unrestricted random walk is irreducible but does not have a finite state space. It therefore does not have a unique stationary distribution.

The restricted random walk has a finite state space and is irreducible for $\alpha \neq 1$ and $\beta \neq 1$. It therefore has a unique stationary distribution for $\alpha \neq 1$ and $\beta \neq 1$.

Example 5.10. **Question:** Compute the stationary distribution for the modified NCD model defined in §5.5.

Answer: The conditions for a stationary distribution defined above lead to the following expressions

$$\begin{aligned}
\pi_0 &= \frac{1}{4}\pi_0 + \frac{1}{4}\pi_1 + \frac{1}{4}\pi_{2^-}, \\
\pi_1 &= \frac{3}{4}\pi_0 + \frac{1}{4}\pi_{2^+}, \\
\pi_{2^+} &= \frac{3}{4}\pi_1, \\
\pi_{2^-} &= \frac{1}{4}\pi_3, \\
\pi_3 &= \frac{3}{4}\pi_{2^+} + \frac{3}{4}\pi_{2^-} + \frac{3}{4}\pi_3.
\end{aligned}$$

This system of equations is not linearly independent since adding all the equations results in an identity. This is a general feature of $\pi = \pi\mathbf{P}$ due to the property $\sum_{j\in\mathcal{S}} p_{ij} = 1$.

We therefore discard one of the equations (discarding the last one will simplify the system) and work in terms of a working variable, say π_1.

$$\begin{aligned}
3\pi_0 - \pi_{2^-} &= \pi_1, \quad & 3\pi_0 + \pi_{2^+} &= 4\pi_1, \\
\pi_{2^+} &= \frac{3}{4}\pi_1, \quad & 4\pi_{2^-} - \pi_3 &= 0.
\end{aligned}$$

This system is solved with

$$\begin{aligned}
\pi_{2^+} &= \frac{3}{4}\pi_1, & \pi_0 &= \frac{13}{12}\pi_1, \\
\pi_{2^-} &= \frac{9}{4}\pi_1, & \pi_3 &= 9\pi_1.
\end{aligned}$$

Using the requirement that $\sum_{j\in\mathcal{S}} \pi_j = 1$, we arrive at the stationary distribution

$$\pi = \left(\frac{13}{169}, \frac{12}{169}, \frac{9}{169}, \frac{27}{169}, \frac{108}{169}\right).$$

5.7 The long-term behaviour of Markov chains

It is natural to expect the distribution of a Markov chain to tend to the stationary distribution π for large times if π exists. However, certain phenomena can complicate this. For example, a state i is said to be *periodic* with period $d > 1$ if a return to i is possible only in a number of steps that is a multiple of d. More specifically, $p_{ii}^{(n)} = 0$ unless $n = md$ for some integer m.

Any periodic behaviour is usually evident from the transition graph. For example, both NCD models considered above are aperiodic; the unrestricted random walk has period 2 and restricted random walk is aperiodic unless α and β are either 0 or 1.

We state the following result about convergence of a Markov chain without proof:

Let $p_{ij}^{(n)}$ be the n-step transition probability of an irreducible aperiodic Markov chain on a finite state space. Then, $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ for each i and j.

Example 5.11. Question: An insurance company has 10,000 policyholders on the modified NCD system defined in §5.5. Estimate the number of policyholders on each discount rate.

Answer: The model is irreducible and aperiodic, therefore, assuming that the policies have been held for a sufficient length of time, the distribution of policyholders amongst states is given by the stationary distribution computed in Example 5.10. We would therefore expect the following distribution:

State 0:	no discount	$10,000 \times 13/169 \approx 769$
State 1:	25% discount	$10,000 \times 12/169 \approx 710$
State 2:	40% discount	$10,000 \times (9/169 + 27/169) \approx 2,130$
State 3:	60% discount	$10,000 \times 108/169 \approx 6,391$

References

The following texts were used in the preparation of this chapter and you are referred there for further reading if required.

- Faculty and Institute of Actuaries, CT4 Core Reading;
- D. R. Cox & H. D. Miller, The Theory of Stochastic Processes;
- S. Ross, Stochastic Processes.

5.8 Summary

For discrete state spaces the Markov property is written as

$$P\left[X_t = a \middle| X_{s_1} = x_1, X_{s_2} = x_2, \ldots, X_{s_n} = x_n, X_s = x\right] = P[X_t = a | X_s = x],$$

for all $s_1 < s_2 < \cdots < s_n < s < t$ and all states $a, x_1, x_2, \ldots, x_n, x$ in $\mathcal{S}$.

Any process with independent increments has the Markov property.

Markov chains are discrete-time and discrete-state-space stochastic processes satisfying the Markov property. You should be familiar with the simple NCD, modified NCD, unrestricted random walk and restricted random walk processes.

In general, the n-step transition probabilities $p_{ij}(m, m+n)$ denote the probability that a process in state i at time m will be in state j at time $m+n$.

The transition probabilities of a Markov process satisfy the Chapman–Kolmogorov equations:

$$p_{ij}(m, n) = \sum_{k \in \mathcal{S}} p_{ik}(m, l) p_{kj}(l, n),$$

for all states $i, j \in \mathcal{S}$ and all integer times $m < l < n$. This can be expressed in terms of n-step stochastic matrices as

$$\mathbf{P}(m, n) = \mathbf{P}(m, l)\mathbf{P}(l, n).$$

An irreducible time-homogeneous Markov chain with a finite state space has a unique stationary probability distribution, π, such that

$$\pi = \pi \mathbf{P}^{(n)}.$$

Aperiodic processes will converge to the stationary distribution as $n \to \infty$.

5.9 Questions

1. Consider a Markov chain with state space $\mathcal{S} = \{0, 1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} p & q & 0 \\ 1/4 & 0 & 3/4 \\ p - 1/2 & 7/10 & 1/5 \end{pmatrix}.$$

   (a) Calculate values for p and q.

   (b) Draw the transition graph for the process.

   (c) Calculate the transition probabilities $p_{ij}^{(3)}$.

   (d) Find any stationary distributions for the process.

2. Prove equation (59) relating the probability of a particular path occurring in a Markov chain.

3. A No-Claims Discount system operated by a motor insurer has the following four levels:

   Level 1: 0% discount;

   Level 2: 25% discount;

   Level 3: 40% discount;

   Level 4: 60% discount.

   The rules for moving between these levels are as follows:

   - Following a year with no claims, move to the next higher level, or remain at level 4.
   - Following a year with one claim, move to the next lower level, or remain at level 1.
   - Following a year with two or more claims, move down two levels, or move to level 1 (from level 2) or remain at level 1.

   For a given policyholder in a given year the probability of no claims is 0.85 and the probability of making one claim is 0.12. X_t denotes the level of the policyholder in year t.

   (i) Explain why X_t is a Markov chain. Write down the transition matrix of this chain.

   (ii) Calculate the probability that a policyholder who is currently at level 2 will be at level 2 after:

i. one year.

ii. two years.

iii. three years.

(iii) Explain whether the chain is irreducible and/or aperiodic.

(iv) Does this Markov chain converge to a stationary distribution?

(v) Calculate the long-run probability that a policyholder is in discount level 2.

Chapter 6
Markov Jump Processes

A *Markov jump process* is a stochastic process with discrete state space and *continuous time set*, which has Markov property.

The mathematical development of Markov jump processes is similar to Markov chains considered in the previous chapter. For example, the Chapman–Kolmogorov equations have the same format. However, Markov jump processes are in continuous time and so the notion of a one-step transition probability does not exist and we are forced to consider time intervals of arbitrarily small length. Taking the limit of these intervals to zero leads to the reformulation of the Chapman–Kolmogorov equations in terms of differential equations.

We begin my discussing the Poisson process which is the simplest example of a Markov jump process. In doing so we will encounter some general features of Markov jump processes.

6.1 Poisson process

The Poisson process $\{N_t\}_{t\in[0,\infty)}$, is an example of a *counting process*. That is, it has state space $\mathcal{S} = \{0, 1, 2, \dots, n, \dots\}$ corresponding to the number of occurrences of some event. The events occur singly and can occur at any time. Counting process are useful in modelling customers in a queue, insurance claims or car accidents, for example.

Informally, a counting process (with counts, for example, customers in a queue) is a Poisson process, if customers arrives independently, and "uniformly" in time, i.e. with constant rate of λ customers per time unit. Thus, in time interval of length h we would expect on average about λh customers.

To make the above intuition more formal, let us assume that time interval $(t, t+h)$ is very short, such that the probability of two or more events during this interval can be neglected. In this case the expected number of events (which should be about λh by the intuition above) is $0 \cdot (1-p) + 1 \cdot p = p$, where p is the probability of an event to occur.

Formally, the probability of an event in any short time interval $(t, t+h)$ is $\lambda h + o(h)$, where a function f is said to be $o(h)$ if

$$\lim_{h\to 0} \frac{f(h)}{h} = 0.$$

The Poisson process can then be defined as follows:

The counting process $\{N_t\}_{t\in[0,\infty)}$ is said to be a Poisson process with rate $\lambda > 0$, if

1. $N_0 = 0$;
2. The process has stationary and independent increments.
3. $P\left(N_{t+h} - N_t = 1\right) = \lambda h + o(h)$;
4. $P\left(N_{t+h} - N_t > 1\right) = o(h)$.

Example 6.1. Question: Prove that a Poisson process is a Markov jump process.

Answer: A Poisson process has independent increments (property 2), therefore it has the Markov property. The state state $\mathcal{S} = \{0, 1, 2, \dots, n, \dots\}$ of the process is discrete, and time set $t \in [0, \infty)$ is continuous, thus it is a Markov jump process by definition.

It is possible to show that the Poisson process defined above coincides with the other standard definition of the Poisson process, that is, a process having independent stationary Poisson-distributed increments. Or more formally, for any $t > 0$, N_t follows a Poisson distribution with parameter λt, that is

$$P(N_t = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \qquad \text{for any } n = 0, 1, 2, \dots \tag{64}$$

More generally, for any $t, s > 0$, $N_{t+s} - N_s$ has the same probability distribution as N_t.

Example 6.2. Question: Prove that the two definitions of a Poisson process are consistent.

Answer: Define the probability that there have been n events by time t as $p_n(t) = P(N_t = n)$. Then,

$$\begin{aligned} p_0(t+h) &= P(N_{t+h} = 0), \\ &= P(N_t = 0, N_{t+h} - N_t = 0), \\ &= P(N_t = 0)P(N_{t+h} - N_t = 0), \\ &= p_0(t)(1 - \lambda h + o(h)). \end{aligned}$$

Rearranging this equation and dividing by h yields

$$\frac{p_0(t+h) - p_0(t)}{h} = -\lambda p_0(t) - \frac{o(h)p_0(t)}{h}.$$

Taking the limit as $h \to 0$, leads to the differential equation

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t),$$

with the initial condition, $p_0(0) = 1$. It is clear that this has solution

$$p_0(t) = e^{-\lambda t}. \tag{65}$$

Similarly, for $n \geq 1$;

$$\begin{aligned}
p_n(t+h) &= P(N_{t+h} = n), \\
&= P(N_t = n, N_{t+h} - N_t = 0) + P(N_t = n-1, N_{t+h} - N_t = 1) + o(h), \\
&= P(N_t = n)P(N_{t+h} - N_t = 0) + P(N_t = n-1)P(N_{t+h} - N_t = 1) + o(h), \\
&= p_n(t)p_0(h) + p_{n-1}(t)p_1(h) + o(h), \\
&= (1 - \lambda h)p_n(t) + \lambda h p_{n-1}(t) + o(h).
\end{aligned}$$

Rearranging this for $p_n(t+h)$, and again taking the limit as $h \to 0$, we obtain the differential equation

$$\frac{dp_n(t)}{dt} = -\lambda p_n(t) + \lambda p_{n-1}(t), \tag{66}$$

for $n = 1, 2, 3, \ldots$.

It can be shown by mathematical induction, or using generating functions, that the solution to the differential equation (66), with initial condition $p_n(0) = 0$ yields equation (64). As required.

A Poisson process has positive integer values and can jump at any time $t \in [0, \infty)$. However, since time is continuous, the probability of a jump is zero at specific time point t. The process can be pictured as an "upwards staircase" shown in Figure 5.

6.1.1 Interarrival times

Since the Poisson process changes only by unit upward jumps, its sample paths are fully characterised by the times at which the jumps take place.

Consider a Poisson process and let τ_1 be the time at which the first event occurs and let τ_n for $n > 1$ denote the time between the $(n-1)$th and the nth event. It is clear that τ_n for $n \geq 1$ is a continuous random variable which takes values in the range $[0, \infty)$.

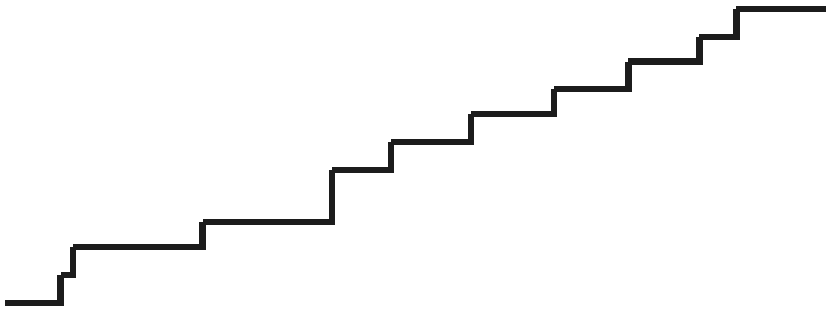


Figure 5: Sample Poisson process. Horizontal distance is time.

The sequence $\{\tau_n\}_{n\geq 1}$ is called the sequence of *interarrival times* (or *holding times*). These are the horizontal distances between each step in Figure 5.

The random variables $\tau_1, \tau_2, \ldots$ are i.i.d., each having the exponential distribution with parameter λ. They therefore each have the density function

$$f_\tau(t) = \lambda e^{-\lambda t} \quad \text{for } t > 0. \tag{67}$$

To demonstrate this for general τ_n, first consider τ_1 and note that the event $\tau_1 > t$ occurs if and only if there are zero events of the Poisson process in the fixed interval $(0, t]$, that is

$$P(\tau_1 > t) = P(N_t = 0) = e^{-\lambda t}.$$

The distribution function of τ_1 is therefore

$$P(\tau_1 \leq t) = 1 - e^{-\lambda t},$$

and so τ_1 is exponentially distributed with parameter λ.

Now consider the distribution of τ_2 conditional on τ_1:

$$\begin{aligned} P(\tau_2 > t | \tau_1 = s) &= P(0 \text{ events in } (s, s+t] | \tau_1 = s), \\ &= P(N_{t+s} - N_s = 0 | \tau_1 = s), \\ &= P(N_{t+s} - N_s = 0), \quad \text{(by independent increments)} \\ &= p_0(t) = e^{-\lambda t}. \end{aligned}$$

Therefore τ_2 is independent of τ_1 and has the same exponential distribution as τ_1.

The same argument can be repeated for $\tau_3, \tau_4, \ldots$ leading to the conclusion that the interarrival times are i.i.d. random variables that are exponentially distributed with parameter λ.

Further, it can be shown using similar arguments that if $\hat{N}_t$ and $\tilde{N}_t$ are two independent Poisson processes with parameters λ_1 and λ_2 respectively, then their sum $N_t = \hat{N}_t + \tilde{N}_t$ is a Poisson process with parameter $\lambda_1 + \lambda_2$. This result follows immediately from our intuitive interpretation of a Poisson process: assume that male customers are arriving uniformly with rate λ_1, and female customers are arriving independently and uniformly with rate λ_2. Then $\hat{N}_t$ describes the cumulative number of male customers, $\tilde{N}_t$ - female customers, thus $N_t = \hat{N}_t + \tilde{N}_t$ is the total number of customers, which clearly also arriving uniformly with rate $\lambda_1 + \lambda_2$.

This can be extended to the sum of any number of Poisson processes and is a very useful result.

Example 6.3. An insurance company assumes that the number of claims on an individual motor insurance policy in a year is a Poisson random variable with parameter q. Claims in successive time intervals are assumed to be independent. The company holds $10,000$ such motor insurance policies which are assumed to be independent.

For $10,000$ independent policies, the total number of claims in any year will therefore be Poisson with mean $10,000q$.

The total number of claims on a policy in a two-year period is a Poisson random variable with mean $2q$.

6.1.2 Compound Poisson process

The Poisson process $\{N_t\}_{t\in[0,\infty)}$ is a natural model for counting *number* of claims reaching an insurance company during time period $[0, t]$. In practice, however, the cumulative *size* of the claims is more important. If Y_i is the size of claim i, the cumulative size is given by $X_t = \sum_{i=0}^{N_t} Y_i$. The simplest model is to assume that all claims Y_i are independent and identically distributed. In this case the stochastic process $\{X_t\}_{t\in[0,\infty)}$ is called a *compound Poisson process*.

Formally, a compound Poisson process with rate $\lambda > 0$ and jump size distribution F is a continuous-time stochastic process given by

$$X_t = \sum_{i=0}^{N_t} Y_i \tag{68}$$

where $\{N_t\}_{t\in[0,\infty)}$ is a Poisson process with rate λ, and $\{Y_i,\ i \geq 1\}$ are independent and identically distributed random variables, with distribution function F, which are also independent of N_t.

The expected value and variance of the compound Poisson process are given by

$$E[X_t] = \lambda t E[Y], \qquad Var[X_t] = \lambda t E[Y^2], \tag{69}$$

where Y is a random variable with distribution function F.

Example 6.4. In Example 6.3 assume that size of each claim is a random variable uniformly distributed on $[a, b]$. All claims sizes are independent. What is the mean and variance of the cumulative size of the claims from all policies during 3 years?

Answer: The cumulative size of the claims is the compound Poisson process $X_t = \sum_{i=0}^{N_t} Y_i$, where N_t is the number of claims from all policies, which is the Poisson process with parameter $\lambda = 10,000q$, and Y_i is size of claim i. Then $EY_i = \frac{1}{b-a}\int_a^b x dx = \frac{a+b}{2}$; $EY_i^2 = \frac{1}{b-a}\int_a^b x^2 dx = \frac{a^2+ab+b^2}{3}$, which gives

$$E[X_3] = 3\lambda E[Y_i] = 30,000q\frac{a+b}{2} = 15,000q(a+b),$$

and

$$Var[X_3] = 3\lambda E[Y^2] = 30,000q\frac{a^2+ab+b^2}{3} = 10,000q(a^2+ab+b^2).$$

Assume that a company has initial capital u, premium rate c, and the cumulative claims size X_t is given by (68). Then the basic problem in risk theory is to estimate the *probability of ruin* at time $t > 0$, defined as

$$\Phi_t(u) = P[u + ct - X_t < 0]. \tag{70}$$

6.2 The time-inhomogeneous Markov jump process

Similar to the Markov chain, we introduce transition probabilities for a general Markov jump process

$$p_{ij}(s,t) = P\left[X_t = j | X_s = i\right], \text{ where } p_{ij}(s,t) \geq 0 \text{ and } s < t. \tag{71}$$

The transition probabilities must also satisfy the *Chapman-Kolmogorov equations*

$$p_{ij}(t_1,t_3) = \sum_{k\in\mathcal{S}} p_{ik}(t_1,t_2)\ p_{kj}(t_2,t_3), \quad \text{for} \quad t_1 < t_2 < t_3. \tag{72}$$

In matrix form, these are expressed as

$$\mathbf{P}(t_1, t_3) = \mathbf{P}(t_1, t_2)\mathbf{P}(t_2, t_3).$$

The proof of these is analogous to that for equation (58) in discrete time, and is left as a question at the end of the chapter.

We require that the transition probabilities satisfy the continuity condition

$$\lim_{t \to s^+} p_{ij}(s,t) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{73}$$

This condition means that as the time difference between two observations approach zero, the process will very likely not change its state with probability approaching one in the limit.

It is easy to see that this condition is consistent with Chapman-Kolmogorov equation. Indeed, taking the limits $t_2 \to t_3^-$ or $t_2 \to t_1^+$ in equation (72) we obtain the identity.

However, this condition does not follow from the Chapman-Kolmogorov equations. For example, $p_{ij}(s,t) = \frac{1}{2}$ for $i, j = 1, 2$ satisfy equation (72), since

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

6.3 Transition rates

Let us assume that transition probabilities $p_{ij}(s,t)$ for $t > s$ have derivatives with respect to t and s. Also, assume for simplicity that the state space $\mathcal{S}$ is finite. Then by the standard definition of a derivative we have

$$\begin{aligned} \frac{\partial p_{ij}(s,t)}{\partial t} &= \lim_{h \to 0} \frac{p_{ij}(s, t+h) - p_{ij}(s,t)}{h}, \\ &= \lim_{h \to 0} \frac{\sum_k p_{ik}(s,t) p_{kj}(t, t+h) - p_{ij}(s,t)}{h,} \\ &= \lim_{h \to 0} \left(\sum_{k \neq j} p_{ik}(s,t) \frac{p_{kj}(t, t+h)}{h} \right. \left. + p_{ij}(s,t) \frac{p_{jj}(t, t+h) - 1}{h} \right). \\ &:= \lim_{h \to 0} \alpha_{ij} \end{aligned} \tag{74}$$

It follows from equation (74) that $\{\alpha_{ij}\}$ approach certain limits as $h \to 0$. In particular, we define

$$\begin{aligned} \lim_{h\to 0} \frac{p_{jj}(t, t+h) - 1}{h} &:= q_{jj}(t), \\ \lim_{h\to 0} \frac{p_{kj}(t, t+h)}{h} &:= q_{kj}(t), \quad \text{for} \quad k \neq j. \end{aligned} \tag{75}$$

The quantities $q_{jj}(t), q_{kj}(t)$ are called *transition rates*. They correspond to the rate of transition from state k to state j in a small time interval h, given that state k is occupied at time t.

Transition probabilities $p_{kj}(t, t+h)$ can be expressed through the transition rates as

$$p_{kj}(t, t+h) = \begin{cases} hq_{kj}(t) + o(h), & k \neq j \\ 1 + hq_{jj}(t) + o(h), & k = j \end{cases} \tag{76}$$

It follows from equation (74) that

$$\frac{\partial p_{ij}(s,t)}{\partial t} = \sum_{k\in\mathcal{S}} p_{ik}(s,t) q_{kj}(t). \tag{77}$$

These differential equations are called *Kolmogorov's forward equations*. In matrix form, they can be written as

$$\frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{P}(s,t)\mathbf{Q}(t),$$

where $\mathbf{Q}(t)$ is called the *generator matrix* with entries $q_{ij}(t)$.

Repeating the procedure but differentiating with respect to s, we have

$$\begin{aligned} \frac{\partial p_{ij}(s,t)}{\partial s} &= \lim_{h\to 0} \frac{p_{ij}(s+h,t) - p_{ij}(s,t)}{h}, \\ &= \lim_{h\to 0} \frac{p_{ij}(s+h,t) - \sum_k p_{ik}(s,s+h) p_{kj}(s+h,t)}{h}, \\ &= -\lim_{h\to 0} \left(\sum_{k\neq i} \frac{p_{ik}(s,s+h)}{h} p_{kj}(s+h,t) \; + \frac{p_{ii}(s,s+h) - 1}{h} p_{ij}(s+h,t) \right). \end{aligned}$$

Therefore

$$\frac{\partial p_{ij}(s,t)}{\partial s} = -\sum_{k\in\mathcal{S}} q_{ik}(s) p_{kj}(s,t), \tag{78}$$

and we see that the derivative with respect to s can also be expressed in terms of the transition rates. The differential equations (78) are called *Kolmogorov's backward equations*. In matrix form these are written as

$$\frac{\partial \mathbf{P}(s,t)}{\partial s} = -\mathbf{Q}(s)\mathbf{P}(s,t).$$

Therefore if transition probabilities $p_{ij}(s,t)$ for $t > s$ have derivatives with respect to t and s, transition rates are well-defined and given by equation (75).

Alternatively, if we can assume the existence of transition rates, then it follows that transition probabilities $p_{ij}(s,t)$ for $t > s$ have derivatives with respect to t and s, given by equations (77) and (78). These equations are compatible, and we may ask whether we can find transition probabilities, given transition rates, by solving equations (77) and (78).

It can be shown that each row of the generator matrix $\mathbf{Q}(s)$ has zero sum. That is,

$$q_{ii}(s) = -\sum_{j \neq i} q_{ij}(s).$$

The *residual holding time* for a general Markov jump process is denoted R_s. This is the random amount of time between time s and the next jump:

$$\{R_s > w, X_s = i\} = \{X_u = i, s \leq u \leq s + w\}.$$

It can be proved that

$$P(R_s > w | \, X_s = i) = e^{\int_s^{s+w} q_{ii}(t)dt}.$$

Similarly, the *current holding time* is denoted C_t. This is the time between the last jump and time t:

$$\{C_t \geq w, X_t = i\} = \{X_u = i, t - w \leq u \leq t\}.$$

We will not study these questions further for general Markov processes, but will investigate such and related questions for *time-homogeneous* Markov processes below.

6.4 Time-homogeneous Markov jump processes

Just as we defined time-homogeneous Markov chains (equation (60)), we can define time-homogeneous Markov jump processes.

Consider the transition probabilities for a Markov process given by equation (71), a Markov process in continuous time is called *time-homogeneous* if the transition probabilities $p_{ij}(s,t) = p_{ij}(0, t-s)$ for all $i, j \in \mathcal{S}$ and $s, t > 0$.

In other words, a Markov process in continuous time is called time-homogeneous if the probability $P(X_t = j | X_s = i)$ depends only on the time interval $t - s$. In this case we can write

$$\begin{aligned} p_{ij}(s,t) &= P(X_t = j | X_s = i) = p_{ij}(t-s), \\ p_{ij}(t,t+s) &= P(X_{t+s} = j | X_t = i) = p_{ij}(s), \\ p_{ij}(0,t) &= P(X_t = j | X_0 = i) = p_{ij}(t). \end{aligned}$$

Here, for example, $p_{ij}(s)$ form a stochastic matrix for every s, that is

$$p_{ij}(s) \geq 0 \quad \text{and} \quad \sum_{j \in \mathcal{S}} p_{ij}(s) = 1,$$

and is assumed to satisfy continuity conditions at $s = 0$

$$\lim_{s \to 0^+} p_{ij}(s) = p_{ij}(0) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

Also $p_{ij}(s)$ satisfy the Chapman-Kolmogorov equations, which, for a time-homogeneous Markov process take the form

$$p_{ij}(t+s) = \sum_{k \in \mathcal{S}} p_{ik}(t)\ p_{kj}(s). \tag{79}$$

In matrix form, the Chapman–Kolmogorov equations become

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s). \tag{80}$$

Note that $\mathbf{P}(0) = \mathbf{I}$ is the identity matrix.

If a time-homogeneous Markov process is currently in state i, it follows from equation (73) that the probability of remaining in i is non-zero for all t, that is, $p_{ii}(t) > 0$. Indeed, from equation (79) it follows that $p_{ii}(t) \geq p_{ii}(t/n)^n$ for any integer n. For example,

$$p_{ii}(t) = \sum_{k \in \mathcal{S}} p_{ik}(t/2)\ p_{ki}(t/2) \geq p_{ii}(t/2) p_{ii}(t/2).$$

The argument for different values of n is similar. So, if for some t we would have $p_{ii}(t) = 0$, this would imply $p_{ii}(t/n) = 0$ for all n, contradiction with (73).

The following properties of transition functions and transition rates for a time-homogeneous process are stated without proof:

1. Transition rates $q_{ij} = \dfrac{dp_{ij}(t)}{dt}\Big|_{t=0} = \lim\limits_{h\to 0} \dfrac{p_{ij}(h) - \delta_{ij}}{h}$ exist for all i, j. Equivalently, as $h \to 0, h > 0$

$$p_{ij}(h) = \begin{cases} hq_{ij} + o(h), & i \neq j \\ 1 + hq_{ii} + o(h), & i = j \end{cases} \tag{81}$$

   Comparing this to equation (76) we see that the only difference between the time-homogeneous and time-inhomogeneous cases is that the transition rates q_{ij} are not allowed to change over time.

2. Transition rates are non-negative and finite for $i \neq j$, and are non-positive when $i = j$, that is

$$q_{ij} \geq 0 \quad \text{for } i \neq j \qquad \text{but} \qquad q_{ii} \leq 0 \quad \text{for } i = j.$$

   Differentiating $\sum\limits_{j\in\mathcal{S}} p_{ij}(t) = 1$ with respect to t at $t = 0$ yields that

$$q_{ii} = -\sum_{j\neq i} q_{ij}.$$

3. If the set of states $\mathcal{S}$ is finite, all transition rates are finite.

Kolmogorov's forward equations for a time-homogeneous process takes the form

$$\frac{dp_{ij}(t)}{dt} = \sum_{k\in\mathcal{S}} p_{ik}(t)q_{kj},$$

and in matrix form

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{Q}$$

where $\mathbf{Q}$ is the generator matrix with entries q_{kj}.

Similarly, Kolmogorov's backward equations are:

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{Q}\mathbf{P}(t).$$

Note that since $q_{ii} = -\sum_{j \neq i} q_{ij}$, each row of the matrix $\mathbf{Q}$ has zero sum.

Example 6.5. Consider the Poisson process again. The rate at which events occur is a constant λ, leading to

$$q_{ij} = \begin{cases} \lambda, & j = i+1 \\ 0, & j \neq i, i+1 \\ -\lambda, & j = i \end{cases} \tag{82}$$

and $p_{ij}(t) = P\left(N_{t+s} = j \middle| N_s = i\right)$.

The Kolmogorov forward equations are

$$\begin{aligned} \frac{dp_{i0}(t)}{dt} &= -\lambda p_{i0}(t), \\ \frac{dp_{ij}(t)}{dt} &= -\lambda p_{ij}(t) + \lambda p_{ij-1}(t), \end{aligned}$$

with $p_{ij}(0) = \delta_{ij}$. These equations are essentially the same as equations (65) and (66).

The backward equations are

$$\frac{dp_{ij}(t)}{dt} = -\lambda p_{ij}(t) + \lambda p_{i+1,j}(t).$$

6.5 Applications

In this section we briefly discuss a number of applications of Markov jump processes to actuarial modelling. In each case the models can be made time-homogeneous by insisting that the transition rates are independent of time. A more detailed discussion of the survival model is postponed to the next chapters.

6.5.1 Survival model

Consider a two-state model where the two states are alive and dead, i.e. transition is in one direction only, from the state alive (A) to the state dead (D) with transition rate $\mu(t)$. This is the *survival model* and has discrete state space $\mathcal{S}$={A,D}. The transition graph is given in Figure 6.

In actuarial notation, the transition rate $\mu(t)$ is identified with the *force of mortality* at age t.

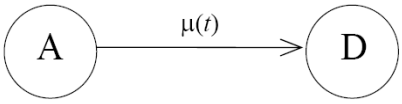


Figure 6: Transition graph for the survival model. *Reproduced with permission of the Faculty and Institute of Actuaries.*

It is clear that the generator matrix $\mathbf{Q}(t)$ is given by

$$\mathbf{Q}(t) = \begin{bmatrix} -\mu(t) & \mu(t) \\ 0 & 0 \end{bmatrix}.$$

The Kolmogorov forward equations therefore become

$$\frac{\partial p_{AA}(s,t)}{\partial t} = -\mu(t) p_{AA}(s,t),$$

and it is clear that the solution corresponding to the initial condition $p_{AA}(s,s) = 1$ is

$$p_{AA}(s,t) = e^{-\int_s^t \mu(x)\,dx}.$$

Note that $p_{AA}(s,t)$ is the probability that an individual alive at time (age) s will still be alive at time (age) t.

Equivalently, consider the probability that an individual now aged s will survive until at least age $s+w$, denoted ${}_wp_s$ in the standard mortality notation

$${}_wp_s = p_{AA}(s, s+w) = e^{-\int_s^{s+w} \mu(x)\,dx} = e^{-\int_0^w \mu(s+u)\,du}.$$

6.5.2 Sickness-death model

The survival model can be extended to include the state of health of an individual. In this so-called *sickness-death* model, the state of an individual is described as being *healthy* (H), *sick* (S), or *dead* (D). The discrete state space is therefore $\mathcal{S}$={H,S,D}.

An individual in state H can jump to either state S or state D. Similarly, an individual in state S can jump to either state H or state D. Time-inhomogeneity arises through the following age-dependent transitions rates

$$\begin{aligned} H &\to S : \sigma(t) \\ H &\to D : \mu(t) \\ S &\to H : \rho(t) \\ S &\to D : \nu(t) \end{aligned}$$

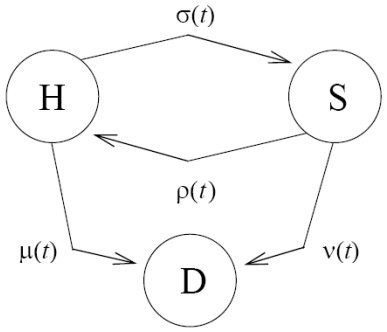


Figure 7: Transition graph for the sickness-death model. *Reproduced with permission of the Faculty and Institute of Actuaries.*

The transition graph is given in Figure 7.

The generator matrix is:

$$\mathbf{Q}(t) = \begin{bmatrix} -(\sigma(t)+\mu(t)) & \sigma(t) & \mu(t) \\ \rho(t) & -(\rho(t)+\nu(t)) & \nu(t) \\ 0 & 0 & 0 \end{bmatrix}.$$

Under this formulation it is possible to calculate probabilities such as

- the probability that an individual who is healthy at time s will still be healthy at time t; or
- the probability that an individual who is sick at time s will still be sick at time t.

These are in terms of the residual holding times as

$$P(R_s > t-s|\, X_s = H) = e^{-\int_s^t (\sigma(u)+\mu(u))du},$$

and

$$P(R_s > t-s|\, X_s = S) = e^{-\int_s^t (\rho(u)+\nu(u))du},$$

respectively.

We note that transition probabilities can be related to each other. For example, the probability of a transition from state H at time s to S at time t would be

$$p_{HS}(s,t) = \int_0^{t-s} \left(e^{-\int_s^{s+w}(\sigma(u)+\mu(u))du}\right) \sigma(s+w) p_{SS}(s+w,t)dw.$$

This is interpreted as "the individual remains in the *healthy* state from time s to time $s+w$ and then jumps to the state *sick* at time $s+w$ where he remains". The derivation of this equation is beyond the scope of the course, however similar expressions can be written down intuitively.

This sickness-death model can be extended to include the length of time an individual has been in state S. This leads to the so-called *long term care model* where the rate of transition out of state S will depend on the current holding time in state S.

6.5.3 Marriage model

A further example of a time-inhomogeneous model is the *marriage model* under which an individual can be either never married (B), married (M), divorced (D), widowed (W) or dead (Δ). A Markov jump process can be formulated on the state space $\mathcal{S}$ ={B, M, D, W, Δ}.

The transition graph is given in Figure 8, where we can see that the death rate has been taken to be independent of the marital status for simplicity.

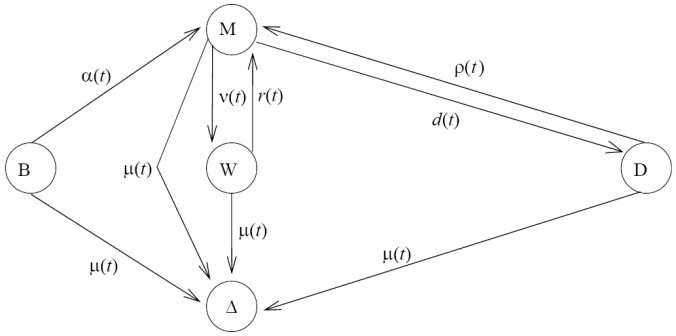


Figure 8: Transition graph for the marriage model. *Reproduced with permission of the Faculty and Institute of Actuaries.*

Example 6.6. Question: State an expression for the probability of being married at time t and of having been so for at least w given that you have never been married at time s ($w < t-s$).

Answer: If C_t is the current holding time, we have

$$P[X_t = M, C_t > w | X_s = B] = \int_s^{t-w} \big(p_{BB}(s,t-v)\alpha(t-v) + p_{BW}(s,t-v)r(t-v) + p_{BD}(s,t-v)\rho(t-v)\big) e^{-\int_{t-v}^{t}(\mu(u)+\nu(u)+d(u))du} dv.$$

This mathematical statement can be read as "the individual is in state B at time s where he either remains until time $(t-v)$, or jumps to states W or D by time $(t-v)$. At time $(t-v)$ he then jumps to state M and remains there until time t".

References

The following texts were used in the preparation of this chapter and you are referred there for further reading if required.

- Faculty and Institute of Actuaries, CT4 Core Reading;
- D. R. Cox & H. D. Miller, The Theory of Stochastic Processes;
- S. Ross, Stochastic Processes.

6.6 Summary

Markov jump processes are continuous-time and discrete-state-space stochastic processes satisfying the Markov property. You should be familiar with the Poisson, survival, sickness-death and marriage models.

The Poisson process is a simple Markov jump process. It is time-homogeneous with stationary increments that are Poisson distributed with mean $\lambda > 0$. Waiting times between jumps are exponentially distributed with mean $1/\lambda$.

As with Markov chains, transition probabilities exist for a general Markov jump process

$$p_{ij}(s,t) = P\left[X_t = j | X_s = i\right], \text{ where } p_{ij}(s,t) \geq 0 \text{ and } s < t,$$

which must also satisfy the Chapman-Kolmogorov equations.

The quantities $q_{jj}(t), q_{kj}(t)$ are the transition rates, such that

$$\lim_{h\to 0} \frac{p_{jj}(t,t+h) - 1}{h} := q_{jj}(t),$$
$$\lim_{h\to 0} \frac{p_{kj}(t,t+h)}{h} := q_{kj}(t), \quad \text{for} \quad k \neq j.$$

Kolmogorov's forward and backwards equations are respectively

$$\frac{\partial p_{ij}(s,t)}{\partial t} = \sum_{k\in\mathcal{S}} p_{ik}(s,t) q_{kj}(t) \quad \text{and} \quad \frac{\partial p_{ij}(s,t)}{\partial s} = -\sum_{k} q_{ik}(s) p_{kj}(s,t).$$

These can be written in matrix form as

$$\frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{P}(s,t)\mathbf{Q}(t) \quad \text{and} \quad \frac{\partial \mathbf{P}(s,t)}{\partial t} = -\mathbf{Q}(s)\mathbf{P}(s,t),$$

where $\mathbf{Q}(t)$ is the generator matrix with entries $q_{ij}(t)$.

In time-homogeneous models the time-dependence of the transition probabilities and transition rates (therefore generator matrices) is removed.

The residual holding time R_s is the random amount of time between time s and the next jump:

$$\{R_s > w, X_s = i\} = \{X_u = i, s \leq u \leq s + w\}.$$

It can be proved that

$$P(R_s > w | \, X_s = i) = e^{\int_s^{s+w} q_{ii}(t)dt},$$

.

6.7 Questions

1. Claims are known to follow a Poisson process with a uniform rate of 3 per day.

   (a) Calculate the probability that there will be fewer than 1 claim on a given day.

   (b) Estimate the probability that another claim will be reported during the next hour. State all assumptions made.

   (c) If there have not been any claims for over a week, calculate the expected time before a new claim occurs.

2. Prove equation (72) which gives the Chapman–Kolmogorov equations for a Markov jump process.

3. Consider the sickness-death model given in Figure 7, write down an integral expression for $p_{HD}(s,t)$.

4. Let $\{X_t,\ t \geq 0\}$ be a time-homogeneous Markov process with state space $\mathcal{S} = \{0,1\}$ and transition rates $q_{01} = \alpha$, $q_{10} = \beta$.

   (a) Write down the generator matrix for this process.

   (b) Solve the Kolmogorov's forward equations for this Markov jump process to find all transition probabilities.

   (c) Check that the Chapman–Kolmogorov equations hold.

   (d) What is the probability that the process will be in state 0 in the long term? Does it depend on the initial state?

Chapter 7
Machine Learning

7.1 A motivating example

Machine learning can be defined as the study of systems and algorithms that improve their performance with experience. Machine learning methods have become increasingly popular in recent decades, because of increase both in computing power and in the amount of data available. You use machine learning systems every day, even without noticing. For example, you may use Google Translate website to translate the text, and the current best methods in automated translation use machine learning. As another example, you mailbox most probably has some kind of spam filter, and most of nowadays spam filters are build based on machine learning technology.

Below is the example of report from one popular spam filter, SpamAssassin.

- 0.6 HTML IMAGE RATIO 02 BODY: HTML has a low ratio of text to image area
- 0.0 HTML MESSAGE BODY: HTML included in message
- 2.0 URIBL BLACK Contains an URL listed in the URIBL blacklist
- -0.9 AWL AWL: From: address is in the auto white-list

This is a typical report from analysing an e-mail. The system noticed that it has "low ratio of text to image area", which is typical for spam e-mails. It assign to this factor a positive but not very high weight 0.6. The system also noticed that some HTML included in message, but assign weight 0 to this fact. Most importantly, the e-mail contains a URL listed in some blacklist, and this gives positive weight 2.0. On the other hand, the message came from a trusted sender (included in the auto white-list), which is an indication that it is not a spam, and therefore contributes a negative weight -0.9. The total weight is $0.6 + 0 + 2.0 - 0.9 = 1.7$. SpamAssassin classifies as "spam" e-mails with total weight at least 5, so this particular e-mail is classified as a good one.

But how SpamAssassin formed the list of criteria/factors to look at and how assigned specific weights for these factors. As you can imagine, the number of factors is large, and it would be very difficult for the developers of spam filter to select weights "by common sense". Instead, they programme the filter to "learn" weights from data. We can collect a large amount of

examples of spam e-mails and mark them as "spam". We can also collect a large amount of normal e-mails and mark them as "not spam". We can then use these e-mails as a "training set", and find weights such that as many as possible of these e-mails would be classified correctly.

Example 7.1. As an oversimplified example, imagime that we have just two factors: "low ratio of text to image area" and "containing URL from the blacklist". Let us introduce variables x_1 and x_2, and, for every e-mail, write $x_i = 1$ if the corresponding factors is true and $x_i = 0$ otherwise. Imagine that we have the training set with 3 e-mails:

- (a) e-mail with low ratio of text to image area, but not containing URL from the blacklist. In other words, $x_1 = 1$ but $x_2 = 0$. The e-mail marked as "not spam".
- (b) e-mail with low ratio of text to image area, and with URL from the blacklist. In other words, $x_1 = 1$ and $x_2 = 1$. The e-mail marked as "spam".
- (c) e-mail with bad URL only: $x_1 = 0$ but $x_2 = 1$. The e-mail marked as "not spam".

Assume that we classify e-mail as spam if and only if $w_1x_1 + w_2x_2 \geq 5$. Then, to classify e-mail (a) correctly, we should have

$$w_1x_1 + w_2x_2 = w_1 \cdot 1 + w_2 \cdot 0 = w_1 < 5.$$

To classify e-mail (b) correctly, we should have

$$w_1x_1 + w_2x_2 = w_1 \cdot 1 + w_2 \cdot 1 = w_1 + w_2 \geq 5.$$

And finally, to classify e-mail (c) correctly, we should have

$$w_1x_1 + w_2x_2 = w_1 \cdot 0 + w_2 \cdot 1 = w_2 < 5.$$

So, in this example, w_1 and w_2 can be any numbers less than 5 with sum at least 5. For example, $w_1 = w_2 = 3$ works.

We can represent the analysis in Example 7.1 geometrically, by introducing a coordinate plane with coordinates (x_1, x_2). Then the three e-mails in the training set are points with coordinates $(1, 0)$, $(1, 1)$, and $(0, 1)$, which we denote A, B, and C, respectively. We can draw the points (A and C) corresponding to non-spam e-mails as blue, and the point (B) representing the spam e-mail in red. The "spam region" $w_1x_1 + w_2x_2 \geq 5$ is a half-plane

whose boundary is the line $w_1x_1 + w_2x_2 = 5$. So, geometrically our task was do draw a line on the plane which separates the blue and red points. One possible such line is the one with equation $3x_1 + 3x_2 = 5$, corresponding to the solution $w_1 = w_2 = 3$ we have chosen in Example 7.1.

If we would use three factors to make a decision (for example, the above two plus whether the sender address is in the auto white-list), then the third factor can be represented as a coordinate x_3, the training set could be depicted as a set of points in the 3-dimensional space, and the problem would be to separate red and blue points by a plane with equation $w_1x_1 + w_2x_2 + w_3x_3 = 5$. In general, we may have n factors and m e-mails in the training set, which, geometrically, corresponds to set of m red and blue points in the n-dimensional space. We then need to separate the points by $n-1$-dimensional set of points satisfying equation $\sum_{i=1}^{n} w_i x_i = 5$ (or any other constant instead of 5). Set of points satisfying this equation is called a hyperplane.

In addition to finding the "best" weights for the given list of factors, the system may learn which new factors it is good to include. For example, it can compute the frequency of various words in e-mails and find that the word "lottery" appears in a lot of spam e-mails and in almost none good e-mails. Based on this statistic, the system may introduce a new, $n+1$-th factor indicating whether an e-mail contains the word "lottery" or not. The red and blue points then get a new coordinate x_{n+1} (with $x_{n+1} = 1$ for e-mails with "lottery" word and $x_{n+1} = 0$ for all other e-mails), and then we separate these points in $(n+1)$-dimensional space and find weights $w_1, w_2 \ldots, w_n, w_{n+1}$, including the weight w_{n+1} of the new factor.

We can see that this problem (find the weights based on training data) is just a problem in algebra (find w_i from some system of inequities), or, equivalently, in geometry (separate red and blue points by a hyperplane). However, we call the branch of science studying this and similar problems "machine learning", because it allows the system to improve with experience. In this specific example with spam filter, it may initially work badly for my specific mailbox, because I, for example, may:

- receive a lot of good e-mails with "low ratio of text to image area" because of nature of my work, but

- receive a lot of spam messages with no images because my e-mail may be in the database of spammers who send such spam messages.

Because of this, I initially may see some spam messages in the main mailbox, and, conversely, some good e-mails in the spam folder. However, I can then mark the spam e-mails from main mailbox as "spam" and the good e-mails

in the spam folder as "not spam", which provides new training data for the spam filter, which are specific for my particular situation. The filter will then find new weights based on new data and can quickly "learn" that, in my particular case, "low ratio of text to image area" is not an indication of spam, so the corresponding weight should be low, or zero, or even negative. Instead, it can put the large positive weights for some other factors which my spammers often use. In this way, the system may quickly learn and adapt itself to the need of each particular user.

Moreover, if spammers are smart enough, they may try to develop e-mails which avoid typical features of spam message, such as word "lottery" or low ratio of text to image area. This may help them to pass the spam filters, but only temporary. After user marks that e-mails as "spam", the system will use this to update the weights, "understand" how these new spam messages looks like, and filter them out next time.

7.2 The problems machine learning can solve

In the previous section we studied one specific example of the problem which is solved by machine learning technique: spam filtering. The filter divides all e-mails into two classes: spam and non-spam, and therefore this problem is an example of *classification problem*. More generally, one may want to automatically classify all e-mails into more than two groups: spam, work e-mails, private e-mails, etc., which is an example of **multi-class classification**. As another example of the same problem, we may want to develop an image recognition technology, which classify images detected by a web-cam as "human", "animal", "bird", etc. Such image recognition is critical for, for example, self-driving cars, because it helps to estimate the level of danger and select action (if we see a human ahead on the road - stop urgently, but if this is a bird, the car may continue to move, possibly with decreased speed).

In some applications of multi-class classification (like spam filtering) it may be clear in advances what classes to consider (spam, non-spam, etc.). However, in some other applications this is unclear. For example, in the image recognition problem, it is difficult to list in advance all possible "objects" the webcam can detect. In this case, we may ask the system to do this automatically. It may represent all observed objects as points in some n-dimensional space, and then automatically detect that these points form m groups, or classes. After this, every new object detected is classified into one of these classes. "Similar" objects go to one class, and "dissimilar" ones - into different classes. This problem is called **clustering**.

Sometimes, we would prefer not to classify objects into finite number of "classes" but rather give them a "score", a real number. For example, we

may want for the spam filter to automatically assign to every e-mail its "importance", so that we can sort e-mails by it and answer the most important e-mails first. Clear spam e-mails can have zero or negative score, the "borderline" e-mails for which the filter is unsure if this a spam or not can have small positive score, then work e-mails may be prioritised over the personal ones, etc. Similarly, for self-driving car, instead of classifying detected objects into classes, it may be more convenient to just assign to every object a score, indicating how dangerous/important it is, so that the higher score the more urgently we need to stop. This task is called **regression**. Mathematically, (linear) regression problem is typically reduces to the following task: given m points in the coordinate plane $\mathbb{R}^{n+1}$ with coordinates $x_1, x_2, \ldots, x_n, y$, approximate the points by linear function $y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$ as good as possible. Typically, the quality of approximation is measured as the sum of squares of differences between the y coordinates of data points and the function.

Example 7.2. Approximate points $(0,0)$, $(1,1)$, $(2,3)$, $(3,3)$ by a line $y = ax + b$ to minimize the sum of squares error.

Solution. For $x = 0$, $y = a \cdot 0 + b = b$, and the data point is $(0,0)$ so the (squared) error is $(b-0)^2$. Similarly, for $x = 1$, $y = a+b$, and the data point is $(1,1)$, the squared error is $(a+b-1)^2$. Continuing this way, we write down the error as

$$e(a,b) = (b-0)^2 + (a+b-1)^2 + (2a+b-3)^2 + (3a+b-3)^2.$$

In optimality

$$\frac{\partial e(a,b)}{\partial a} = 2(a+b-1) + 2(2a+b-3)2 + 2(3a+b-3)3 = 0$$

and

$$\frac{\partial e(a,b)}{\partial b} = 2b + 2(a+b-1) + 2(2a+b-3) + 2(3a+b-3) = 0$$

This simplifies to

$$4(7a+3b-8) = 2(6a+4b-7) = 0$$

and the solution is $a = 11/10$, $b = 1/10$.

Another example is the analysis of **association rules**, patterns which are popular in marketing applications. For example, when you view any popular book on Amazon website, you can see a list at the bottom of the page entitled

"Customers who viewed this item also viewed", followed by the list of similar books. There can be also another list entitled "customers who bought this also bough". Also, when you read information about any film, you often get a list of suggestions for similar films you may be interested, etc. To create such lists, the program should learn what books/products/films are "similar" or "associated" with each other, and then use these associations to provide you with best recommendations.

When recommending the best film for you, the system often looks for **hidden** or **latent variables**, that is, some structure which helps to understand your choice. For example, you may assign scores to the films you saw, and these scores looks unstructured even for you. However, automatic analysis of your scores may reveal the "structure" that you typically assign a high score to films of particular genre that have at least one of three particular actors.

Machine learning can also be used in a variety of other applications, for example, in playing games, from board games like poker, Chess, or Go, to real-life application with "resemble" the game, like developing optimal strategies for trading on financial markets.

In all applications, it is important to understand how to evaluate the performance of machine learning system, e.g. of algorithm for calculating weights for spam filter. If we have N e-mails marked as spam or not, we can use these N e-mails as test set to find weights, but it might be that these weights are good only for these specific N e-mails and will not work well for any other e-mails. The situation when weights are so adapted to the specific test set that the system fails to do anything else is called **overfitting**, and is one of the most serious problems in the area. To test that it did not occur, one may divide our N e-mails into two groups, and use one group (say, 90% of e-mails) for training, and the remaining 10% of e-mails for testing. On the test stage, we may count the number of e-mails (both spam and non-spam) which are classified correctly and divide it by the total number of e-mails - this ratio is called the **accuracy** of the classifier.

The set of 10% of e-mails we select for testing is called **the test set**, and it is usually selected randomly. However, as in any random experiment, we may get a very different result each time we repeat the procedure. One time the test set may contain some "typical" e-mails for which the filter performs well, while another time it may contain many "atypical" e-mails (e.g. good e-mails which happen to have many spam-like features) and the filter may make a lot of errors. Because of this, it is a good idea to repeat the test several times and average the result. For example, we may randomly divide out initial set of N e-mails into 10 group, use one of them as test set and the remaining 9 as training set, and then repeat this procedure 10 times, with

each group being the test set once. The final accuracy is the average of 10 accuracies obtained. This procedure is an example of **cross validation**.

Sometimes the set of data used for training is further divided into two groups: a **training data set** and **validation data set**. The training set is used to estimate the parameters of the model (such as weights w_i in Example 7.1), while the validation set is used to decide some more global questions such as number of parameters to consider, number of categories to classify (should it just spam and non-spam or maybe 3 or more categories), rate at which the model should learn from the data, etc. Such "global parameters" are called **hyper-parameters**.

In some applications, the "accuracy" as defined above (the total number of correct classifications divided by the sum of the correct and incorrect ones), is not the best way to evaluate the system performance, because there are two very different types of errors. The first mistake is when the e-mail is spam but the system does not recognise it and puts it to the main folder. This is called a False negative (FN). The second type of mistake is when a good e-mail is putted into the spam folder, and this is called False Positive (FP). If you never check the spam folder, then FN is not a big problem, while FP may be a big issue. We can also define true positive (TP) when the e-mail is spam and it is in the spam folder, and true negative (TN) if the e-mail is non-spam and it goes to the main folder. Then we can consider the following measures of system performance:

- Precision $Pre = \frac{TP}{TP+FP}$;
- Recall $Rec = \frac{TP}{TP+FN}$;
- F1 score $= \frac{2 \cdot Pre \cdot Rec}{Pre+Rec}$;
- False Positive Rate $= \frac{FP}{TN+FP}$.

In many cases, there is a trade-off between recall and false positive rate. If we improves one of this, the other one may become worse.

7.3 Models, methods, and techniques

In Example 7.1, the e-mails in training set were market as "spam" and "non-spam", and this information was used to calculate weights. We assume that this classification of the training set was done by human. This approach is called **supervised learning**.

However, in reality we can have millions of e-mails in training set and it may take months of hard work for a human to classify it. What id our

training set is just set of e-mails, and it is unknown which e-mail is spam. Can we program computer to learn something from this data? In fact, we can! The computer can still note that some e-mails are somewhat similar (for example, use similar words like "lottery" or have low ratio of text to image area), and "guess" that maybe this group of similar e-mails are the spam ones. This is an example of **unsupervised learning**.

In example Example 7.1, we can find weights algebraically, from a system of inequalities, or geometrically, as a line which separate red and blue points on the plane. More generally, we are looking for a hyperplane which separate points in the n-dimensional space, where n is the number of factors to be weighted. This n-dimensional space is called the **instance space**, and the whole method is known as **geometric model**.

Hyperplane that perfectly separate the data may not exists, and, even if it exists, its coefficients are not always easy to compute. Here is the method which is very easy to compute. We can calculate the mean of all red points (call if point A), the mean of all blue points (call if point B), and then select a hyperplane which is perpendicular to the line segment AB and crosses it in the midpoint. This hyperplane is known as **basic linear classifier**, and may separate red and blue points quite well, although not perfectly.

In Example 7.1, there are many possible solutions, for example, $w_1 = w_2 = 3$, or $w_1 = w_2 = 4$, etc. Geometrically, one can draw many possible lines which separates read and blue points. Some of such lines may be very close to blue points, some - to the red ones. Intuitively, one may prefer a line which separates points with maximal **margin**, that is, as far from the data points as possible. This idea forms the basis of the method called **support vector machines**. In Example 7.1, such "best" line corresponds to the solution $w_1 = w_2 = \frac{10}{3}$.

In many applications, the key task is the similarity testing: is this e-mail similar to the spam e-mails in the database? What are the films most similar to the one a customer likes and put high rating? If every e-mail (or film, or any other object) is described by n parameters and is represented as a point in n-dimensional space, one easy way to define "similarly" is to say that objects are similar if and only if the corresponding points $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ are close in the usual Euclidean distance

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

A very simple classification algorithm is, for every new point A to be classified, find an already classified point B at the smallest distance from A, and

then assign A the same class as B. This method is known as the **nearest-neighbour classifier**. In particular, imagine that in Example 7.1 new e-mail arrives with no bad url and no high image to text area ratio. It corresponds to new point $(0, 0)$ on the plane. The distance from it to the "spam" point $(1, 1)$ is

$$\sqrt{(0-1)^2 + (0-1)^2} = \sqrt{2},$$

while the distances to both "non-spam" points are

$$\sqrt{(0-0)^2 + (0-1)^2} = \sqrt{(0-1)^2 + (0-0)^2} = 1 < \sqrt{2}.$$

Hence, the nearest-neighbour classifier would classify this new point as non-spam.

Another distance-based method can be used for clustering task. Imagine that we need to cluster the data into K clusters, and we have some initial guess how to do this. For each cluster $1 = 1, 2, \ldots, K$, we can calculate the mean M_i of all points in cluster i. After this, for each point X, we can compute the distance from X to $M_1, M_2, \ldots, M_K$, select the minimal out of these distances, and re-assign X to the corresponding cluster. We can then repeat this procedure until no point will be re-assigned. This method is called K**-means**, and it is very popular and powerful method for clustering.

Example 7.3. Consider 3 points A, B, C on the plane with coordinates $(0, 0)$, $(0, 1)$, $(4, 0)$, initially classified such that A and C are red while B is blue. Then the mean of red cluster is $M_1 = (2, 0)$, while the mean of blue cluster is $M_2 = B = (0, 1)$. The distances from A to M_i are

$$d(A, M_1) = 2,\ d(A, M_2) = 1,$$

hence M_2 is the nearest one, and A moves to the blue cluster. Similarly,

$$d(B, M_1) = \sqrt{5},\ d(B, M_2) = 0,\ d(C, M_1) = 2,\ d(C, M_2) = \sqrt{17},$$

hence B and C stays in blue and red clusters, respectively.

We now repeat the procedure: the new cluster means are $M_1 = C = (4, 0)$ and $M_2 = (0, 0.5)$, respectively. It is easy to check that the closest M_i is M_2 for A and B and M_1 for C, hence all points stay in the same class and the algorithm terminates.

Of course, for nearest-neighbour classifier, K-means, and other related methods, it is not necessary to use the Euclidean distance. For example, we may instead use the **Manhattan distance** between points $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$, given by

$$\sum_{i=1}^{n} |x_i - y_i|.$$

7.4 Probabilistic analysis

In classification task, models using **probabilistic** analysis may be useful. For example, assume that we try to decide whether an e-mail a spam or not based on the information whether it contain words "lottery" and "tablet".

Example 7.4. Imagine that we have the following data:

- 60 e-mails without these words. 20 of them are spam, and 40 - not;
- 15 e-mails with word "lottery" only. 10 of them are spam, and 5 - not;
- 24 e-mails with word "tablet" only. 20 of them are spam, and 4 - not;
- 1 e-mail with both these words, and it is not a spam.

Based on this data we can calculate various conditional probabilities. For example, if an e-mail contains no "lottery" and no "tablet", we may estimate that it has about $\frac{20}{60} = \frac{1}{3}$ chance to be a spam e-mail and about $\frac{40}{60} = \frac{2}{3}$ chance to be a non-spam. In notation, let S be a random variable indicating that e-mail is spam if $S = 1$ and non-spam if $S = 0$, L be a random variable such that $L = 1$ if this random e-mail contains the word Lottery (and $L = 0$ if not), and T is a similar random variable about word "tablet". Then

$$P(S = 1|T = L = 0) = \frac{1}{3}, \quad P(S = 0|T = L = 0) = \frac{2}{3}.$$

Because $P(S = 0|T = L = 0) > P(S = 1|T = L = 0)$, we classify e-mail with $T = L = 0$ as non-spam. This method is called **maximum a posteriori (MAP) decision rule**.

In a similar way, we can apply MAP decision rule to all possible combinations of value of L and T, and write a computer program which gives the answers in all possible cases:

- If $T = 0$ (no "tablet" word in e-mail), then:
  - If $L = 0$ (no "lottery" word in e-mail), then: NON-SPAM
  - If $L = 1$ (there is a "lottery" word in e-mail), then: SPAM
- If $T = 1$ (there is a "tablet" word in e-mail), then:
  - If $L = 0$ (no "lottery" word in e-mail), then: SPAM
  - If $L = 1$ (there is a "lottery" word in e-mail), then: NON-SPAM

This computer program is an example of **decision tree**. In general, decision three may contain any number of this nested "if... then" structure. Instead of final decision (SPAM or NON-SPAM) the program may return, for example, a real number which represents the probability that an e-mail is spam.

In Example 7.4, suppose that part of e-mail is encoded and the filter cannot read it. In the open part, the filter see the word "lottery" but no "tablet", and it is not clear if "tablet" is present in the encoded part or not. In this case, the MAP decision rule classifies e-mail as spam if $P(S = 1|L = 1) > 0.5$. This probability can be estimated from the law of total probability (13), or directly from data as

$$P(S = 1|L = 1) = \frac{10 + 0}{15 + 1} = \frac{10}{16} = 0.625,$$

because in total there are $15 + 1 = 16$ e-mails with word "lottery", 10 of them are spam.

In fact, in the situation like in Example 7.4, statisticians often makes decisions based on different conditional probabilities, such as $P(T = L = 0|S = 1)$ and $P(T = L = 0|S = 0)$, which is an example of **likelihood function**. The logic is that one asks herself: how likely I would find e-mail looking like this (in our case, with no words "lottery" and "tablet") in the spam folder? And how likely I would find it in the non-spam folder? In our example, there are 50 spam e-mails, 20 of them are with $T = L = 0$, and also there are 50 non-spam e-mails, 40 of them are with $T = L = 0$, hence

$$P(T = L = 0|S = 1) = \frac{20}{50} < \frac{40}{50} = P(T = L = 0|S = 0).$$

Thus, observing an e-mail like this in a spam folder is about twice less likely than finding such an e-mail in the non-spam folder. Hence, the e-mail should be classified as non-spam.

The two methods above (MAP and likelihood methods) are related by *Bayes' theorem*

$$P[B|A] = \frac{P[A \cap B]}{P[A]} = \frac{P[A|B]P[B]}{P[A]},$$

which in our case says that

$$P(S = 1|T = L = 0) = \frac{P(T = L = 0|S = 1)P(S = 1)}{P(T = L = 0)}.$$

The same formulas work for all other possible values of T and L, for example

$$P(S = 1|T = L = 1) = \frac{P(T = L = 1|S = 1)P(S = 1)}{P(T = L = 1)}.$$

In fact, our data-based estimate $P(S = 1|T = L = 1) = P(T = L = 1|S = 1) = 0$ is unjustified, because there is just one e-mail with $T = L = 1$, and it is strange to calculate any probabilities based on the sample with ONE experiment. This situation is very typical: there are may be very few e-mails with the word pattern exactly as prescribed, or, in general case, very few objects with values of parameters exactly equal to some values. An alternative ways to estimate probabilities like $P(T = L = 1|S = 1)$ is to assume that words occur independently (or at least independently conditional on the event $S = 1$), and then

$$P(T = L = 1|S = 1) = P(T = 1|S = 1) \cdot P(L = 1|S = 1),$$

and then

$$P(S = 1|T = L = 1) = \frac{P(T = 1|S = 1) \cdot P(L = 1|S = 1) \cdot P(S = 1)}{P(T = L = 1)}.$$

The classification based on probabilities calculated in this way is called **naive Bayes** classification, with the word "naive" reflecting the assumption of independence.

Another probabilistic learning method which have been used recently with dramatic success is **reinforcement learning**. In reinforcement learning the learner is not given a target output in the same way as with supervised learning. The learner uses the input data to choose some output, and is then told how well it is doing, or how close the chosen output is to the desired output. The learner can then use this information as well as the input data to choose another hypothesis. This method has been recently apply to playing games, such as Chess and Go, in which a machine first plays against itself at random, and then automatically learn from experience, increasing the chance of selecting moves similar to ones that lead to the success in previous games. A generic program Alpha Zero designed to play in many games with the same algorithm, having just rules of the games as an input, quickly learn from self-play and beat all human and all specifically designed programs which human developed for many decades!

7.5 Stages of analysis in Machine Learning

Machine Learning tasks can often be broken down into a series of steps.

- **Collecting data**

  The data must be assembled in a form suitable for analysis using computers. Several different tools are useful for achieving this: a

spreadsheet may be used, or a database such as Microsoft Access. Data may come from a variety of sources, including sample surveys, population censuses, company administration systems, databases constructed for specific purposes (such as the Human Mortality Database, www.mortality.org). During the last 20-30 years the size of datasets available for analysis by actuaries and other researchers has increased enormously. Datasets, such as those on purchasing behaviour collected by supermarkets, relate to millions of transactions.

- **Exploring and preparing the data**

  The data need to be prepared in such a way that a computer is able to access it and apply a range of algorithms. If the data are already in a spreadsheet, this may be a simple matter of importing the data into whatever computer package is being used to develop the algorithms. If the data are stored in complex file formats, it will be useful to convert the data to rectangular format, with one line per case and one column per variable. It is also important here to recognise the nature of the variables being analysed: are they nominal, ordinal or continuous? Next we should do cleaning of the data, which includes replacing missing values, and checking the data for obvious errors.

- **Feature scaling**

  Some Machine Learning techniques will only work effectively if the variables are of similar scale. If, for example, one variable (say x_1) is measured in, say, kilometres, and other variables are in centimetres, the value of x_1 will be 1000 times larger than it would be with the same data if it were measured in centimetres as well. This may lead to inadequate results in a number of machine learning methods such as linear regression.

- **Splitting the data into the training, validation and testing data sets**

  typical split might be to use 60% of the data for training, 20% for validation and 20% for testing. However it depends on the problem and not on the data. A guide might be to select enough data for the validation data set and the testing data set so that the validation and testing processes can function, and to allocate the rest of the data to the training data set. In practice, this often leads to around a 60%/20%/20% split.

- **Training a model on the data**

This involves choosing a suitable Machine Learning algorithm using a subset of the data. The algorithm will typically represent the data as a model and the model will have parameters which need to be estimated from the data. This stage is analogous to the process of fitting a model to data when using linear regression and generalised linear models.

- **Validation and testing**

  The model should then be validated using the 20% of the data set aside for this purpose. This should indicate, for example, whether we are at risk of over-fitting our data. The results of the validation exercise may mean that further training is required. Once the model has been trained on a set of data, its performance should be evaluated. How this is done may depend on the purpose of the analysis. If the aim is prediction, then one obvious approach is to test the model on a set of data different from the one used for development. If the aim is to identify hidden patterns in the data, other measures of performance may be needed.

- **Improving model performance**

  We can measure the performance of the model by testing it on the 20% of the data we have reserved for this purpose. The hope is that the performance on the "test" data set is similar to that achieved on the training data set. This amounts to stating that the difference between the in-sample error and the out-of-sample error will be generally small.

  If the performance of the model is not sufficient for the task to hand, it may be possible to improve its performance. Sometimes the combination of several different algorithms applied to the same data set will produce a performance which is substantially better than any individual model. In other cases, the use of more data might provide a boost to performance. However, except when considering very simple combinations of models, care should be taken not to overfit the evaluation set.

- **The reproducibility of research**

  It is important that data analysis be reproducible. This means that someone else can take the same data, analyse it in the same way, and obtain the same results. In order that an analysis be reproducible the following criteria are necessary:

  – The data used should be fully described and available to other researchers.

- Any modifications to the data (e.g. recoding or transformation of variables, or computation of new variables) should be clearly described, ideally with the computer code used. In Machine Learning this is often called "features engineering", whereby combinations of features are used to create something more meaningful.
- The selection of the algorithm and the development of the model should be described, again with computer code being made available. This should include the parameters of the model and how and why they were chosen.

There is an inherent problem with reproducing stochastic models, in that those of necessity have a random element. Of course, details of the random number generator seeds chosen, and the precise command and package used to generate any randomness, could be presented. However, since stochastic models are typically run many times to produce a distribution of results, it normally suffices that the distribution of the results is reproducible.

7.6 Summary

Machine learning can be defined as the study of systems and algorithms that improve their performance with experience.

Machine learning typically used to solve classification problems, clustering, regression problems, analysis of association rules, discovering hidden or latent variables, etc.

One of the main problems in machine learning is overfitting, when the system works perfectly on the data it was trained in but performs badly on any other data. To test that it did not occur, one may divide our data into two groups, and use one group for training, and another one for testing. We can then exchange the role of training and testing data, this procedure is called cross validation. In fact, data for training can be further divided into two groups: a training data set and validation data set, the first one is to find model parameters, and the second one is to find hyper-parameters. The training-validation-test data proportion may be, for example, $60\% - 20\% - 20\%$.

In the yes-no classification problem, the correct outcomes are true positive (TP), and true negative (TN), while the incorrect ones are False Positive (FP) and False negative (FN). This can be used to calculate various measures of performance, such as Precision, Recall, F1 score, and False Positive Rate.

You should be able to define and understand the following terms and methods:

- Supervised learning
- Unsupervised learning
- Linear classifier
- Support vector machines
- Nearest-neighbour classifier
- K-means algorithm
- Maximal a posteriori (MAP) decision rule
- Decision tree
- Likelihood function
- Naive Bayes classifier
- Reinforcement learning

Machine Learning tasks can often be broken down into a series of steps:

- Collecting data
- Exploring and preparing the data
- Feature scaling
- Splitting the data into the training, validation and testing data sets
- Training a model on the data
- Validation and testing
- Improving model performance
- The reproducibility of research

7.7 Questions

1. Approximate points $(0, 0, 0)$, $(1, 0, 2)$, $(0, 1, 3)$, $(1, 1, 4)$ in the coordinate space (x_1, x_2, y) by a plane $y = ax_1 + bx_2 + c$ to minimize the sum of squares error.

2. There are two points marked on the plane - red point A with coordinate $(0, 0)$ and blue point B with coordinates $(10, 10)$. Then 4 points C, D, E, F arrives in order, and each is coloured in the same way as its nearest neighbour. The coordinates of F is $(10, 8)$. Give examples of coordinates of points C, D, E such that point F will be coloured red.

3. Consider 4 points A, B, C, D on the plane with coordinates $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(2, 1)$, initially classified such that A and D are red while B and C are blue. What would be the outcome of the K-means algorithm (with $K = 2$) applied to these initial data?

4. A filter should classify e-mails into 3 categories: personal, work, and spam. The statistics shows that approximately 30% of e-mails are personal, 50% are work ones, and 20% are spam. It also shows that word "friend" is included into 20% of personal e-mails, 5% of work e-mails, and 30% of spam e-mails. In addition, the word "profit" is included into 5% of personal e-mails, 30% of work e-mails, and 25% of spam e-mails. A new e-mails arrives which contains both words "friend" and "profit". Use naive Bayes classification to decide if this e-mail is more likely to be personal, work, or spam?

Solutions of end-of-chapter questions

8.1 Chapter 1 solutions

1. A sample space consists of five elements $\Omega = \{a_1, a_2, a_3, a_4, a_5\}$. For which of the following sets of probabilities does the corresponding triple $(\Omega, \mathcal{A}, P)$ become a probability space? Why?

(a) $p(a_1) = 0.3$; $p(a_2) = 0.2$; $p(a_3) = 0.1$; $p(a_4) = 0.1$; $p(a_5) = 0.1$;

(b) $p(a_1) = 0.4$; $p(a_2) = 0.3$; $p(a_3) = 0.1$; $p(a_4) = 0.1$; $p(a_5) = 0.1$;

(c) $p(a_1) = 0.4$; $p(a_2) = 0.3$; $p(a_3) = 0.2$; $p(a_4) = -0.1$; $p(a_5) = 0.2$;

Answer: Since Ω is finite, we may assume that $\mathcal{A}$ is the set of all subsets of Ω. So we only have to look at the point probabilities $p(a_i) = P(\{a_i\})$ for $i = 1, \ldots, 5$. From the definition of a discrete probability distribution, we know that the sum of all point probabilities must be equal to 1, i.e. here we must have

$$p(a_1) + p(a_2) + \cdots + p(a_5) = 1.$$

In part (a), the values of this sum is equal to 0.8, which means that P is not a probability distribution and therefore $(\Omega, \mathcal{A}, P)$ is not a probability space.

We further know that probabilities can never be negative. In part (c), we have $p(a_4) = -0.1$, which means that $(\Omega, \mathcal{A}, P)$ is not a probability space.

In part (b), $(\Omega, \mathcal{A}, P)$ is indeed a probability space, since here all requirements are met.

2. Let X be a random variable from the continuous uniform distribution, $X \sim U(0.5, 1.0)$. Starting with the probability density function, derive expressions for the cumulative distribution function, expectation and variance of X.

Answer: The probability density function $\rho_X(x)$ of X is a positive constant on the interval $(0.5, 1.0)$ and zero otherwise. Further it has integral 1, so the right choice is given by

$$\rho_X(x) = \begin{cases} 2, & \text{if } x \in (1/2, 1), \\ 0, & \text{otherwise.} \end{cases}$$

Here $I_A(x)$ denotes the indicator function of a set A which is equal to 1 if $x \in A$ and equal to 0 otherwise. The cumulative distribution function is

given by

$$F_X(x) \;=\; \int_{-\infty}^{x} \rho_X(z)\,dz = \begin{cases} 0, & \text{if } x \le 1/2 \\ \int_{1/2}^{x} 2\,dz = 2x-1, & \text{if } x \in (1/2,1) \\ 1, & \text{if } x \ge 1. \end{cases}$$

Hence the expectation of X is

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x\rho_X(x)\,dx = \int_{1/2}^{1} 2x\,dx = x^2\Big|_{x=1/2}^{x=1} = 1 - \frac{1}{4} = \frac{3}{4} = 0.75.$$

The variance of X can be calculated like this:

$$\begin{aligned} \mathrm{Var}(X) \;&=\; \int_{-\infty}^{\infty} \Big(x-\frac{3}{4}\Big)^2 \rho_X(x)\,dx = \int_{1/2}^{1} 2\Big(x-\frac{3}{4}\Big)^2 dx = \frac{2}{3}\Big(x-\frac{3}{4}\Big)^3\Big|_{x=1/2}^{x=1} \\ &=\; \frac{2}{3}\Big[\frac{1}{4^3}+\frac{1}{4^3}\Big] = \frac{1}{3\times 16} = \frac{1}{48} = 0.020833... \end{aligned}$$

3. Assets A and B have the following distribution of returns in various states:

State	Asset A	Asset B	Probability
1	10%	−2%	0.2
2	8%	15%	0.2
3	25%	0%	0.3
4	−14%	6%	0.3

Show that the correlation between the returns on asset A and asset B is equal to -0.3830.

Answer: Let R_A and R_B be the returns on assets A and B, respectively. Then the correlation between R_A and R_B is given by

$$\mathrm{Corr}(R_A, R_B) = \frac{\mathrm{Cov}(R_A, R_B)}{\sqrt{\mathrm{Var}(R_A)\mathrm{Var}(R_B)}},$$

where $\mathrm{Cov}(R_A, R_A) = \mathrm{E}(R_A R_B) - \mathrm{E}(R_A)\mathrm{E}(R_B)$ is the covariance between

R_A and R_B. We have

$$\begin{aligned}
\mathrm{E}(R_A) &= (10 \times 0.2 + 8 \times 0.2 + 25 \times 0.3 + (-14) \times 0.3)\% = 6.9\%, \\
\mathrm{Var}(R_A) &= \mathrm{E}(R_A^2) - (\mathrm{E}(R_A))^2, \\
&= \Big(10^2 \times 0.2 + 8^2 \times 0.2 + 25^2 \times 0.3 + (-14)^2 \times 0.3 - 6.9^2\Big)\%\% \\
&= (15.2148)^2\%\%, \\
\sqrt{\mathrm{Var}(R_A)} &= 15.2148\%, \\
\mathrm{E}(R_B) &= (-2 \times 0.2 + 15 \times 0.2 + 0 \times 0.3 + 6 \times 0.3)\% = 4.4\%, \\
\mathrm{Var}(R_B) &= \mathrm{E}(R_B^2) - (\mathrm{E}(R_B))^2 \\
&= \Big(2^2 \times 0.2 + 15^2 \times 0.2 + 0^2 \times 0.3 + 6^2 \times 0.3 - 4.4^2\Big)\%\% \\
&= (6.1025)^2\%\%, \\
\sqrt{\mathrm{Var}(R_B)} &= 6.1025\%, \\
\mathrm{E}(R_A R_B) &= \Big(10 \times (-2) \times 0.2 + 8 \times 15 \times 0.2 + 25 \times 0 \times 0.3 \\
&\qquad + (-14) \times 6 \times 0.3\Big)\%\% \\
&= -5.2\%\%.
\end{aligned}$$

Note that % and %% stand for $1/100$ and $1/100^2$, respectively. Using the values above, we obtain

$$\mathrm{Corr}(R_A, R_B) = \frac{-5.2/100^2 - 6.9 \times 4.4/100^2}{0.152148 \times 0.061025} = -0.3830,$$

as required.

4. Formalise Example 8.5 as $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, $P(\{\omega_1\}) = P(\{\omega_2\}) = P(\{\omega_3\}) = P(\{\omega_4\}) = 1/4$ and

$$A := \{\omega_1, \omega_4\}, \quad B := \{\omega_2, \omega_4\}, \quad C := \{\omega_3, \omega_4\}.$$

Prove that the pairs (A, B), (A, C) and (B, C) are independent, but the triple (A, B, C) is not mutually independent according to Definition 8.2.

Answer: We have

$$\begin{aligned}
P(A) = P(B) = P(C) &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \\
P(A \cap B) = P(\{\omega_4\}) &= \frac{1}{4} = P(A)P(B), \\
P(A \cap C) = P(\{\omega_4\}) &= \frac{1}{4} = P(A)P(C), \\
P(B \cap C) = P(\{\omega_4\}) &= \frac{1}{4} = P(B)P(C),
\end{aligned}$$

which shows that the pairs (A, B), (A, C) and (B, C) are independent. However

$$P(A \cap B \cap C) = P(\{\omega_4\}) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C).$$

So the triple (A, B, C) is not mutually independent.

5. You intend to model the maximum daily temperature in your office as a stochastic process. What time set and state space would you use?

Answer: It is reasonable to use a suitable discrete time set such as $\mathcal{T} = \{0, 1, 2, \dots\}$ and a continuous state space such as $\mathcal{S} = \mathbb{R}$.

8.2 Chapter 2 solutions

1. The number of claims a company received during the last 12 months are

$$10, 8, 15, 10, 7, 3, 20, 14, 5, 12, 8, 8.$$

Assuming that these numbers are i.i.d. realizations of
(a) Poisson distribution with parameter λ
(b) negative binomial distribution with parameters p and k,
use method of moments to estimate unknown parameters.

Answer: Because Poisson distribution has only 1 parameter, it suffices to condider only first moment (expectation). The expectation of Poisson distribution is λ. From data, the estimate for the expectation is

$$\frac{1}{12}(10+8+15+10+7+3+20+14+5+12+8+8) = 10.$$

Hence, $\lambda \approx 10$.

For negative binomial distribution, the mean and variance are $\frac{k(1-p)}{p}$ and $\frac{k(1-p)}{p^2}$, respectively, and the corresponding estimates from the data are 10 and

$$\frac{1}{12}(10^2+8^2+15^2+10^2+7^2+3^2+20^2+14^2+5^2+12^2+8^2+8^2) - 10^2 = 20.$$

Hence, the system of equations is

$$\frac{k(1-p)}{p} = 10, \quad \frac{k(1-p)}{p^2} = 20.$$

The solution to this system is $p = \frac{1}{2}$, $k = 10$.

2. Assume that the same data as in question 1 are i.i.d. realizations of geometric distribution with parameter p. Use method of maximum likelihood to estimate p.

Answer: Denote the data as $k_1, k_2, \ldots, k_{12}$. The logarithm of the likelihood function is

$$l(p) = \sum_{i=1}^{12} \log[(1-p)^{k_i} p] = 12\log(p) + \log(1-p)\sum_{i=1}^{12} k_i.$$

Then

$$l'(p) = \frac{12}{p} - \frac{1}{1-p}\sum_{i=1}^{12} k_i = 0$$

if

$$p = \frac{12}{12 + \sum_{i=1}^{12} k_i} = \frac{12}{132} = \frac{1}{11}.$$

3. The history of $n = 18$ most recent claim sizes (rounded to integer number of pounds) are

$$937, 342, 150, 1080, 401, 3500, 7970, 1400, 530,$$

$$1106, 847, 899, 3076, 2837, 315, 2560, 390, 2950.$$

Assuming that these are i.i.d data from Weibull distribution, use the method of percentiles with $\alpha_1 = 1/4$ and $\alpha_2 = 3/4$ to estimate parameters of the distribution.

Answer: The sorted data are

$$150, 315, 342, 390, 401, 530, 847, 899, 937,$$

$$1080, 1400, 1106, 2560, 2837, 2950, 3076, 3500, 7970.$$

The smallest integers grater than $\alpha_1 n = 18/4 = 4.5$ and $\alpha_2 n = 13.5$ are 5 and 14, respectively. The 5-th and 14-th data points in the sorted sequence are $q_1 = 401$ and $q_2 = 2837$, respectively. Hence, by (23) and (24), the parameters of the Weibull distribution are

$$\gamma = \log\left[\frac{\log(1-\alpha_1)}{\log(1-\alpha_2)}\right] / \log\left[\frac{q_1}{q_2}\right] = \log\left[\frac{\log(1-0.25)}{\log(1-0.75)}\right] / \log\left[\frac{401}{2837}\right] \approx 0.8037,$$

and

$$c = -\frac{\log(1-\alpha_1)}{q_1^{\gamma}} \approx -\frac{\log(1-0.25)}{401^{0.8037}} \approx 0.002327.$$

4. Assume that the history of claim sizes are the same as in the previous question, but the company order a reinsurance policy with excess of loss reinsurance above the level $M = 2000$.

(a) Write down the history of expenses of the reinsurer;

(b) Assuming that the original claim size distribution is Pareto distribution with parameters $\alpha > 2$ and $\lambda > 0$, estimate the unknown parameters using method of moments with data available to the reinsurer.

(c) Comment whether do you think the Pareto distribution is a good model to fit these data.

Answer: (a) The history of reinsurer expenses is

$$1500, 5970, 1076, 837, 560, 950.$$

(b) Using formulas from Example 2.1 with $k = 6$ and w_i as above, we get

$$\frac{1}{k}\sum_{i=1}^{k} w_i = \frac{1}{6}(1500 + 5970 + 1076 + 837 + 560 + 950) = 1815.5$$

and

$$\frac{1}{k}\sum_{i=1}^{k} w_i^2 - \left(\frac{1}{k}\sum_{i=1}^{k} w_i\right)^2 = \frac{1}{6}(1500^2+5970^2+1076^2+837^2+560^2+950^2) - 1815.5^2 = 3531517.25,$$

so the system of equations to find the parameters become

$$1815.5 = \frac{\lambda + 2000}{\alpha - 1}$$

and

$$3531517.25 = \frac{\alpha(\lambda + 2000)^2}{(\alpha - 1)^2(\alpha - 2)}.$$

Squaring both parts of the first equation and dividing it on the second one, we get

$$\frac{(1815.5)^2}{3531517.25} = \left(\frac{\lambda + 2000}{\alpha - 1}\right)^2 : \frac{\alpha(\lambda + 2000)^2}{(\alpha - 1)^2(\alpha - 2)} = \frac{\alpha - 2}{\alpha}.$$

Hence, $\frac{\alpha-2}{\alpha} \approx 0.9333$ and $\alpha \approx \frac{2}{1-0.9333} \approx 30$. Then

$$\lambda = 1815.5(\alpha - 1) - 2000 \approx 50649.5.$$

(c) The resulting parameters does not look realistic. One would expect much smaller values such that $\alpha \approx 3$ or $\alpha \approx 4$. We may conclude that Pareto distribution is not the best model to fit these data.

8.3 Chapter 3 solutions

1. Assume that the number N of claims can be any integer from 1 to 100 with equal chances, and the claim sizes $X_1, \ldots, X_N$ are i.i.d. from Pareto distribution with parameters $\alpha = 3$ and $\lambda = 2$. Estimate mean and variance of the aggregate claim S.

Answer: The mean and variance of Pareto distribution $Pa(3, 2)$ are

$$\mu_X = \frac{\lambda}{\alpha - 1} = 1, \quad \text{and} \quad \sigma_X^2 = \frac{\alpha\lambda^2}{(\alpha-1)^2(\alpha-2)} = 3.$$

The mean and variance of N are

$$\mu_N = \frac{1}{100}\sum_{i=1}^{100} i = 50.5, \quad \text{and} \quad \sigma_N^2 = \frac{1}{100}\sum_{i=1}^{100}(i-\mu_N)^2 = \frac{100^2-1}{12} = 833.25,$$

see (4) for derivation of variance. Hence, the mean and variance of S are

$$\mu_S = \mu_N \cdot \mu_X = 50.5 \cdot 1 = 50.5$$

and

$$\sigma_S^2 = \mu_N\sigma_X^2 + \sigma_N^2\mu_X^2 = 50.5 \cdot 3 + 833.25 \cdot 1^2 = 974.75.$$

2. The number of claims insurance company receives in April follows Poisson distribution with $\lambda = \frac{45}{7}$, the sizes of claims are i.i.d. and follow a uniform distribution on $[1,000; 2,000]$.

(a) Estimate the probability that the company will receive at least 3 claims in April.

(b) Estimate the mean and variance for the total size of all April's claims.

(c) Estimate the probability that the total size of all April's claims will be strictly less than $3,000$.

Answer: (a) Let N be the number of April claims. Then $P(N = n) = e^{-\lambda}\frac{(\lambda)^n}{n!}$. Hence

$$P(N=0) = e^{-\frac{45}{7}}\frac{(\frac{45}{7})^0}{0!} \approx 0.0016$$

$$P(N=1) = e^{-\frac{45}{7}}\frac{(\frac{45}{7})^1}{1!} \approx 0.0104$$

$$P(N=2) = e^{-\frac{45}{7}}\frac{(\frac{45}{7})^2}{2!} \approx 0.0334$$

$$P(N \geq 3) = 1 - P(N = 0) - P(N = 1) - P(N = 2) \approx$$
$$\approx 1 - 0.0016 - 0.0104 - 0.0334 = 0.9546.$$

(b) Let Y be a claim size, uniformly distributed on $[a, b]$, where $a = 1,000$, $b = 2,000$. Then the density of Y is $f(y) = \frac{1}{b-a}$, $a \leq y \leq b$, and

$$E[Y] = \int_a^b y \frac{1}{b-a} dy = \frac{1}{b-a}(y^2/2)\Big|_a^b = \frac{a+b}{2} = 1,500,$$

$$E[Y^2] = \int_a^b y^2 \frac{1}{b-a} dy = \frac{1}{b-a}(y^3/3)\Big|_a^b = \frac{a^2+ab+b^2}{3} = \frac{7}{3} \cdot 10^6.$$

Then for the aggregate claim size S

$$E[S] = \lambda E[Y] = \frac{45}{7} \cdot 1,500 \approx 9,643,$$

$$Var[S] = \lambda E[Y^2] = \frac{45}{7} \cdot \frac{7}{3} \cdot 10^6 = 15 \cdot 10^6.$$

(c) If there are at least 3 claims, then the total size will be at least $3,000$. If there are 2 claims, there is a 50% chance to have a total size $3,000$. If there is 1 claim or no claim, the total size will surely be less than $3,000$. Hence, the answer is

$$P(N = 0) + P(N = 1) + P(N = 2)/2 \approx 0.0016 + 0.0104 + 0.0334/2 = 0.0287.$$

3. The claim size X follows a log-normal distribution with parameters μ and σ, where σ is known but μ is not. Instead, we model μ as another random variable such that $\lambda = e^{\mu+\sigma^2/2}$ has mean p and variance s^2. Estimate the mean and variance of X.

Answer: By the properties of log-normal distribution,

$$E[X|\lambda] = e^{\mu+\sigma^2/2} = \lambda,$$

and

$$Var[X|\lambda] = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2} = (e^{\sigma^2} - 1)\lambda^2.$$

By the law of total expectation,

$$E[X] = E[E[X|\lambda]] = E[\lambda] = p.$$

By the law of total variance,

$$Var(X) = E[Var(X|\lambda)] + Var(E(X|\lambda)) = E[(e^{\sigma^2} - 1)\lambda^2] + Var(\lambda) =$$

$$= (e^{\sigma^2} - 1)(p^2 + s^2) + s^2 = e^{\sigma^2}(p^2 + s^2) - p^2.$$

4. The number N of claims to be received by insurance company next year follow a negative binomial distribution with parameters $k = 20$ and $p = 0.25$. The claim sizes $X_1, \ldots, X_N$ are i.i.d. and follow exponential distribution with parameter $\lambda = 0.005$. Assuming that the aggregate claim size S is approximately normally distributed, estimate the probability that S with not exceed $20,000$.

Answer: If X follows exponential distribution with $\lambda = 0.005$, then $\mu_X = E[X] = \frac{1}{\lambda} = 200$ and $Var[X] = \frac{1}{\lambda^2} = 40,000$. Hence, $E[X^2] = Var[X] + (EX)^2 = 80,000$.

By (38),

$$\mu_S = E[S] = \frac{k(1-p)}{p}\mu_X = \frac{20(1-0.25)}{0.25}200 = 12,000.$$

By (39),

$$\sigma_S^2 = \frac{k(1-p)}{p}E[X^2] + \frac{k(1-p)^2}{p^2}(E[X])^2 =$$

$$= \frac{20(1-0.25)}{0.25}80,000 + \frac{20(1-0.25)^2}{0.25^2}(200)^2 = 12,000,000.$$

Hence,

$$P(S \leq 20,000) = P\left(\frac{S - 12,000}{\sqrt{12,000,000}} \leq \frac{20,000 - 12,000}{\sqrt{12,000,000}}\right) \approx P(Z \leq 2.3),$$

where Z follows standard normal distribution. From tables, $P(Z \leq 2.3) \approx 0.99$.

8.4 Chapter 4 solutions

1. (a) Calculate the Hazard rate of the Pareto distribution. Check if it is an increasing or decreasing function.

(b) Calculate the Mean residual life of the Pareto distribution. Check if it is an increasing or decreasing function.

(c) What conclusion about tails of Pareto distribution can we make based on items (a) and (b).

Answer: The Pareto distribution with parameters $\alpha > 0$ and $\lambda > 0$ has PDF

$$f(x) = \frac{\alpha\lambda^\alpha}{(\lambda + x)^{\alpha+1}}, \quad x > 0.$$

and CDF

$$F(x) = 1 - \left(\frac{\lambda}{\lambda + x}\right)^\alpha.$$

(a) The Hazard rate is

$$h(x) = \frac{f(x)}{1 - F(x)} = \frac{\alpha\lambda^\alpha}{(\lambda + x)^{\alpha+1}} : \frac{\lambda^\alpha}{(\lambda + x)^\alpha} = \frac{\alpha}{\lambda + x}$$

The derivative

$$h'(x) = -\frac{\alpha}{(\lambda + x)^2} < 0$$

is negative, hence $h(x)$ is a decreasing function.

(b) The Mean residual life is

$$e(x) = \frac{\int_x^\infty (1 - F(y))dy}{1 - F(x)} = \left(\frac{\lambda + x}{\lambda}\right)^\alpha \int_x^\infty \left(\frac{\lambda}{\lambda + y}\right)^\alpha dy =$$

$$= \left(\frac{\lambda + x}{\lambda}\right)^\alpha \frac{\lambda^\alpha(\lambda + x)^{1-\alpha}}{\alpha - 1} = \frac{\lambda + x}{\alpha - 1},$$

which is an increasing function of x provided that $\alpha > 1$.

(c) Because the Hazard rate is $h(x)$ is a decreasing function, and the Mean residual life is increasing function, this is an indication that The Pareto distribution has heavy tail. If the claims follow this distribution, we can expect some very large claim with not very small probability.

2. Prove the formula for $F(x)$ is Example 4.1.

Answer: Substituting $F_X(x) = 1 - \exp(-\lambda x)$, $a_n = \frac{1}{\lambda}\ln n$ and $\beta_n = \frac{1}{\lambda}$ into (45), we get

$$F(x) = \lim_{n\to\infty}\left[1 - \exp\left[-\lambda\left(\frac{1}{\lambda}\ln n + \frac{1}{\lambda}x\right)\right]\right]^n = \lim_{n\to\infty}(1 - \exp(-x - \ln n))^n =$$

$$\lim_{n\to\infty}\left(1-\frac{e^{-x}}{n}\right)^n = e^{-e^{-x}}.$$

3. (a) Check that function $\psi(t) = -\ln t$, $0 < t \le 1$, is continuous, strictly decreasing, convex, and $\psi(1) = 0$;

(b) Find its inverse function ψ^{-1};

(c) Find $C(u,v) = \psi^{-1}(\psi(u) + \psi(v))$, see (51).

Answer: (a) The continuity of $\psi(t)$ is obvious. The derivative $\psi'(t) = -\frac{1}{t}$ is negative, hence $\psi(t)$ is strictly decreasing. The second derivative $\psi''(t) = \frac{1}{t^2}$ is positive, hence $\psi(t)$ is a convex function. Finally, $\psi(1) = -\ln 1 = 0$.

(b) To find inverse function we solve equation $\psi(t) = -\ln t = x$ to get answer $t = e^{-x}$, hence $\psi^{-1}(x) = \exp(-x)$.

(c) By (51),

$$C(u,v) = \psi^{-1}(\psi(u) + \psi(v)) = exp(-(-\ln u - \ln v)) = \exp(\ln(uv)) = uv.$$

Hence, in this case $C(u,v)$ is the independence copula.

4. Repeat the previous question for

(a) $\psi(t) = (-\ln t)^\alpha$, $0 < t \le 1$, where $\alpha \ge 1$ is a parameter;

(b) $\psi(t) = -\ln\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$, $0 < t \le 1$, where $\alpha \ne 0$ is a parameter;

(c) $\psi(t) = \frac{1}{\alpha}(t^{-\alpha} - 1)$, $0 < t \le 1$, where $\alpha \ne 0$ is a parameter.

Answer: (a)

$$\psi(t) = (-\ln t)^\alpha, \qquad 0 < t \le 1,$$

where $1 \le \alpha$ is a parameter. Then

$$\psi'(t) = \alpha(-\ln t)^{\alpha-1}(-\frac{1}{t}) < 0$$

and

$$\psi''(t) = \alpha(\alpha-1)(-\ln t)^{\alpha-2}\left(-\frac{1}{t}\right)^2 + \alpha(-\ln t)^{\alpha-1}\left(\frac{1}{t^2}\right) > 0,$$

hence $\psi(t)$ is strictly decreasing and convex. We also have $\psi(1) = (-\ln 1)^\alpha = 0$. To find inverse function we solve equation $\psi(t) = (-\ln t)^\alpha = x$ to get answer $t = \exp(-x^{1/\alpha})$. Then by (51),

$$C(u,v) = \psi^{-1}(\psi(u) + \psi(v)) = \exp\left\{-\left((-\ln u)^\alpha + (-\ln v)^\alpha\right)^{1/\alpha}\right\},$$

which is exactly the Gumbel copula.

(b) Let

$$\psi(t) = -\ln\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right), \qquad 0 < t \le 1,$$

where $\alpha \neq 0$ is a parameter. Then

$$\psi'(t) = -\frac{e^{-\alpha}-1}{e^{-\alpha t}-1}\cdot\frac{-\alpha e^{-\alpha t}}{e^{-\alpha}-1} = \frac{\alpha e^{-\alpha t}}{e^{-\alpha t}-1}.$$

If $\alpha > 0$ then $e^{-\alpha t} - 1 < 0$, while if $\alpha < 0$ then $e^{-\alpha t} - 1 > 0$. In any case, $\psi'(t) < 0$.

Next,

$$\psi''(t) = \frac{-\alpha^2 e^{-\alpha t}(e^{-\alpha t}-1) - \alpha e^{-\alpha t}(-\alpha e^{-\alpha t})}{(e^{-\alpha t}-1)^2} = \frac{\alpha^2 e^{-2\alpha t}}{(e^{-\alpha t}-1)^2} > 0,$$

hence $\psi(t)$ is strictly decreasing and convex. We also have $\psi(1) = -\ln\left(\frac{e^{-\alpha}-1}{e^{-\alpha}-1}\right) = 0$, so all the conditions are satisfied. To find inverse function we solve equation $\psi(t) = -\ln\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right) = x$. Equivalently, $\frac{e^{-\alpha t}-1}{e^{-\alpha}-1} = e^{-x}$, or $e^{-\alpha t} - 1 = e^{-x}(e^{-\alpha}-1)$, hence $t = -\frac{1}{\alpha}\ln\left(e^{-x}(e^{-\alpha}-1)+1\right)$. For $x = \psi(u) + \psi(v)$,

$$\exp(-x) = \exp\left(\ln\left(\frac{e^{-\alpha u}-1}{e^{-\alpha}-1}\right) + \ln\left(\frac{e^{-\alpha v}-1}{e^{-\alpha}-1}\right)\right) = \frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{(e^{-\alpha}-1)^2},$$

and by (51),

$$C(u,v) = \psi^{-1}(x) = -\frac{1}{\alpha}\ln\left(1 + \frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{(e^{-\alpha}-1)}\right),$$

which is the Frank copula.

(c) Let

$$\psi(t) = \frac{1}{\alpha}(t^{-\alpha}-1), \qquad 0 < t \le 1,$$

where $\alpha \neq 0$ is a parameter. Then

$$\psi'(t) = -t^{-\alpha-1} < 0$$

and

$$\psi''(t) = -(-\alpha-1)t^{-\alpha-2},$$

hence $\psi(t)$ is strictly decreasing for all $\alpha \neq 0$ and convex for $\alpha > -1$. We also have $\psi(1) = \frac{1}{\alpha}(1^{-\alpha}-1) = 0$. To find inverse function we solve equation $\psi(t) = \frac{1}{\alpha}(t^{-\alpha}-1) = x$ to get answer $t = (\alpha x + 1)^{\alpha}$. Then by (51),

$$C(u,v) = \psi^{-1}(\psi(u)+\psi(v)) = (\alpha\psi(u) + \alpha\psi(v) + 1)^{\alpha} = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha},$$

which is the Clayton copula.

5. Calculate the coefficients of lower and upper tail dependence of two random variables with

(a) the independence copula,

(b) the Gumbel copula with $\alpha \geq 1$,

(c) the Frank copula with $\alpha \neq 0$,

(d) the Clayton copula with $\alpha > 0$.

Answer: (a) For the independence copula $C(u, v) = uv$, we have

$$C(u, u) = u^2,$$

$$\lambda_L = \lim_{u \to 0+} \frac{C(u, u)}{u} = \lim_{u \to 0+} \frac{u^2}{u} = 0,$$

$$\bar{C}(u, u) = -1 + u + u + C(1 - u, 1 - u) = -1 + 2u + (1 - u)^2 = u^2,$$

$$\lambda_U = \lim_{u \to 0+} \frac{\bar{C}(u, u)}{u} = \lim_{u \to 0+} \frac{u^2}{u} = 1$$

(b) For the Gumbel copula $C(u, v) = \exp\left\{-\left((-\ln u)^\alpha + (-\ln v)^\alpha\right)^{1/\alpha}\right\}$, we have

$$C(u, u) = \exp\left\{-\left(2(-\ln u)^\alpha\right)^{1/\alpha}\right\} = \exp(2^{1/\alpha} \ln u) = u^\beta,$$

where $\beta = 2^{1/\alpha}$. Because $\alpha \geq 1$, we have $1 < \beta \leq 2$.

$$\lambda_L = \lim_{u \to 0+} \frac{C(u, u)}{u} = \lim_{u \to 0+} u^{\beta - 1} = 0.$$

Next,

$$\bar{C}(u, u) = -1 + 2u + C(1 - u, 1 - u) = -1 + 2u + (1 - u)^\beta,$$

and

$$\lambda_U = \lim_{u \to 0+} \frac{\bar{C}(u, u)}{u} = 2 + \lim_{u \to 0+} \frac{(1 - u)^\beta - 1}{u} = 2 - \beta = 2 - 2^{1/\alpha}.$$

(c) For the Frank copula

$$C(u, v) = -\frac{1}{\alpha} \ln\left(1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{(e^{-\alpha} - 1)}\right),$$

we have

$$C(u,u) = -\frac{1}{\alpha}\ln\left(1+\frac{(e^{-\alpha u}-1)^2}{(e^{-\alpha}-1)}\right).$$

As $u \to 0^+$, $e^{-\alpha u} - 1 \approx -\alpha u$, and

$$C(u,u) \approx -\frac{1}{\alpha}\ln\left(1+\frac{\alpha^2 u^2}{(e^{-\alpha}-1)}\right) \approx -\frac{1}{\alpha}\cdot\frac{\alpha^2 u^2}{(e^{-\alpha}-1)} = -\frac{\alpha u^2}{(e^{-\alpha}-1)},$$

hence

$$\lambda_L = \lim_{u\to 0+}\frac{C(u,u)}{u} = -\frac{\alpha}{(e^{-\alpha}-1)}\lim_{u\to 0+} u = 0.$$

Next,

$$\bar{C}(u,u) = -1+2u+C(1-u,1-u) = -1+2u-\frac{1}{\alpha}\ln\left(1+\frac{(e^{-\alpha(1-u)}-1)^2}{(e^{-\alpha}-1)}\right),$$

and

$$\lambda_U = \lim_{u\to 0+}\frac{\bar{C}(u,u)}{u} = 0.$$

(d) For the Clayton copula $C(u,v) = (u^{-\alpha}+v^{-\alpha}-1)^{-1/\alpha}$, we have

$$C(u,u) = (2u^{-\alpha}-1)^{-1/\alpha}.$$

If $\alpha > 0$ and $u \to 0$, then $u^{-\alpha} \to \infty$, and therefore $2u^{-\alpha} - 1 \approx 2u^{-\alpha}$. Hence,

$$C(u,u) \approx (2u^{-\alpha})^{-1/\alpha} = 2^{-1/\alpha}\cdot u,$$

and

$$\lambda_L = \lim_{u\to 0+}\frac{C(u,u)}{u} = 2^{-1/\alpha}.$$

Also,

$$\bar{C}(u,u) = -1+2u+C(1-u,1-u) = -1+2u+(2(1-u)^{-\alpha}-1)^{-1/\alpha}.$$

If $u \to 0$, $(1+Au)^B \approx 1+ABu$ for any constants A, B, hence

$$(2(1-u)^{-\alpha}-1)^{-1/\alpha} \approx (2(1+\alpha u)-1)^{-1/\alpha} = (1+2\alpha u)^{-1/\alpha} \approx 1-2u,$$

where $\approx$ is up to the terms of order u^2. Hence,

$$\lambda_U = \lim_{u\to 0+}\frac{\bar{C}(u,u)}{u} = \lim_{u\to 0+}\frac{-1+2u+1-2u+O(u^2)}{u} = 0.$$

8.5 Chapter 5 solutions

1. Consider a Markov chain with state space $\mathcal{S} = \{0, 1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} p & q & 0 \\ 1/4 & 0 & 3/4 \\ p - 1/2 & 7/10 & 1/5 \end{pmatrix}.$$

(a) Calculate values for p and q.

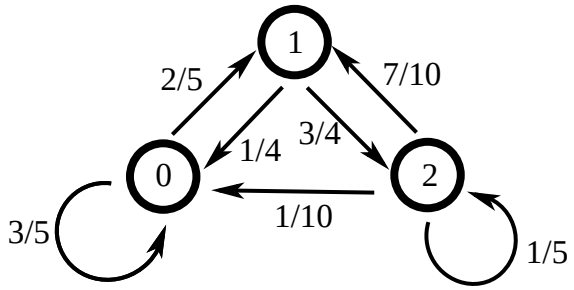
(b) Draw the transition graph for the process.

(c) Calculate the transition probabilities $p_{i,j}^{(3)}$.

(d) Find any stationary distributions for the process.

Answer: (a) The sum of all entries in the last row must be equal to 1, as a consequence of which $p = 1 - \frac{1}{5} - \frac{7}{10} + \frac{1}{2} = \frac{3}{5}$. In view of the first row, we see that $q = \frac{2}{5}$.

(b)



(c) With the help of (a) we get

$$\mathbf{P} = \begin{pmatrix} 3/5 & 2/5 & 0 \\ 1/4 & 0 & 3/4 \\ 1/10 & 7/10 & 1/5 \end{pmatrix}, \quad \mathbf{P}^2 = \begin{pmatrix} 23/50 & 6/25 & 3/10 \\ 9/40 & 5/8 & 3/20 \\ 51/200 & 9/50 & 113/200 \end{pmatrix},$$

$$\mathbf{P}^3 = \begin{pmatrix} 183/500 & 197/500 & 6/25 \\ 49/160 & 39/200 & 399/800 \\ 509/2000 & 199/400 & 31/125 \end{pmatrix} = \begin{pmatrix} 0.366 & 0.394 & 0.24 \\ 0.30625 & 0.195 & 0.49875 \\ 0.2545 & 0.4975 & 0.248 \end{pmatrix}.$$

The values $p_{i,j}^{(3)}$ are the entries of $\mathbf{P}^3$, i.e. $\mathbf{P}^3 = (p_{i,j}^{(3)})_{i,j \in \mathcal{S}}$. For example, we have $p_{1,2}^{(3)} = 0.49875$. It should be mentioned that higher powers of $\mathbf{P}$ can be evaluated using the property $\mathbf{P}^{k+\ell} = \mathbf{P}^k \mathbf{P}^\ell$, $(k, \ell \in \mathbb{N})$. E.g. the calculation of $\mathbf{P}^4 = (\mathbf{P}^2)^2$ does not require the calculation of $\mathbf{P}^3$.

(d) It can be shown that the only stationary distribution is given by

$$\pi = (\pi_1, \pi_2, \pi_3) = \Big(\frac{55}{179}, \frac{64}{179}, \frac{60}{179}\Big) \approx (0.30726, 0.35754, 0.33520).$$

Indeed this follows, if we solve the linear equations $\pi\mathbf{P} = \pi$ for $\pi_1, \pi_2, \pi_3 \in [0, 1]$ with $\pi_1 + \pi_2 + \pi_3 = 1$. More precisely, we have

$$\frac{3}{5}\pi_1 + \frac{1}{4}\pi_2 + \frac{1}{10}\pi_3 = \pi_1, \tag{83}$$
$$\frac{2}{5}\pi_1 + 0\pi_2 + \frac{7}{10}\pi_3 = \pi_2, \tag{84}$$
$$0\pi_1 + \frac{3}{4}\pi_2 + \frac{1}{5}\pi_3 = \pi_3, \tag{85}$$
$$\pi_1 + \pi_2 + \pi_3 = 1. \tag{86}$$

From (85) it follows that $\pi_3 = \frac{15}{16}\pi_2$. Using this in (84), we see that $\pi_2 = \frac{64}{55}\pi_1$ and, in turn, $\pi_3 = \frac{12}{11}\pi_1$. In view of (86), we then get $\pi_1(1 + \frac{64}{55} + \frac{12}{11}) = 1$, i.e. $\pi_1 = \frac{55}{179}$. From the above, we then obtain the remaining values π_2 and π_3 as indicated. We did not use (83). This equation must be valid, since $\mathbf{P}$ is a stochastic matrix. Therefore, (83) can be used to check our solution.

2. Prove equation (28) relating the probability of a particular path occuring in a Markov chain.

Answer: We have to show that, if $\{X_k\}_{k\in\mathbb{Z}_+}$ ($\mathbb{Z}_+ = \{0, 1, 2, \dots\}$) is a Markov chain, then

$$P(X_0 = j_0, X_1 = j_1, \dots, X_N = j_N) = P(X_0 = j_0) \prod_{n=0}^{N-1} p_{j_n, j_{n+1}}(n, n+1)$$

for $N \in \mathbb{N}$ and states $j_0, \dots, j_N \in \mathcal{S}$. Note that we do not assume a time-homogeneous chain, as a consequence of which the one-step transition probabilities $p_{i,j}(n, n+1)$ also depend on time n. The above equation can be proved by induction over N. Indeed, if $N = 1$, then the equation can be shown this way:

$$\begin{aligned} P(X_0 = j_0, X_1 = j_1) &= P(X_0 = j_0) P(X_1 = j_1 \,|\, X_0 = j_0) \\ &= P(X_0 = j_0) p_{j_0, j_1}(0, 1). \end{aligned}$$

Suppose the equation is true for a $N \in \mathbb{N}$, then, using the Markov property

of $\{X_k\}$, we get

$$\begin{aligned}
&P(X_0 = j_0, X_1 = j_1, \ldots, X_{N+1} = j_{N+1}) \\
&\quad = P(X_{N+1} = j_{N+1} \,|\, X_0 = j_0, X_1 = j_1, \ldots, X_N = j_N) \\
&\qquad \times P(X_0 = j_0, X_1 = j_1, \ldots, X_N = j_N) \\
&\quad = P(X_{N+1} = j_{N+1} \,|\, X_N = j_N) \times P(X_0 = j_0) \prod_{n=0}^{N-1} p_{j_n, j_{n+1}}(n, n+1) \\
&\quad = p_{j_N, j_{N+1}}(N, N+1) \times P(X_0 = j_0) \prod_{n=0}^{N-1} p_{j_n, j_{n+1}}(n, n+1) \\
&\quad = P(X_0 = j_0) \prod_{n=0}^{N} p_{j_n, j_{n+1}}(n, n+1),
\end{aligned}$$

which completes our induction proof.

3. A No-Claims Discount system operated by a motor insurer has the following four levels:

Level 1: 0% discount;

Level 2: 25% discount;

Level 3: 40% discount;

Level 4: 60% discount.

The rules for moving between these levels are as follows:

- Following a year with no claims, move to the next higher level, or remain at level 4.
- Following a year with one claim, move to the next lower level, or remain at level 1.
- Following a year with two or more claims, move down two levels, or move to level 1 (from level 2) or remain at level 1.

For a given policyholder in a given year the probability of no claims is 0.85 and the probability of making one claim is 0.12. X_t denotes the level of the policyholder in year t.

(i) Explain why X_t is a Markov chain. Write down the transition matrix of this chain.

(ii) Calculate the probability that a policyholder who is currently at level 2 will be at level 2 after:

(a) one year.

(b) two years.

(c) three years.

(iii) Explain whether the chain is irreducible and/or aperiodic.

(iv) Does this Markov chain converge to a stationary distribution?

(v) Calculate the long-run probability that a policyholder is in discount level 2.

Answer:

(i) It is clear that $X(t)$ is a Markov chain; knowing the present state, any additional information about the past is irrelevant for predicting the next transition.

Then the transition matrix is given by

$$P = \begin{pmatrix} .15 & .85 & 0 & 0 \\ .15 & 0 & .85 & 0 \\ .03 & .12 & 0 & .85 \\ 0 & .03 & .12 & .85 \end{pmatrix}.$$

(ii) (a) For the one year transition $p_{22} = 0$, since with probability 1, the chain will leave the state 2.

(b) The second order transition matrix is given by

$$P^{(2)} = P * P = \begin{pmatrix} 0.15 & 0.1275 & 0.7225 & 0 \\ 0.048 & 0.2295 & 0 & 0.7225 \\ 0.0225 & 0.051 & 0.204 & 0.7225 \\ 0.0081 & 0.0399 & 0.1275 & 0.8245 \end{pmatrix},$$

and thus $p_{22}^{(2)} = .2295$.

(c) The relevant entry from the third order transition matrix is .062475.

(iii) The chain is irreducible as any state is reachable by any other state.

It is also aperiodic. For states 1 and 4 the chain can simply remain there. This is not the case for states 2 and 3. However these are

also aperiodic, since starting from 2 the chain can return in 2 and 3 transitions from the previous part of the question. Similarly the chain started at 3 can return at 3 in two steps (look at P^2), and at three steps.

(iv) The chain is irreducible and has a finite state space and thus has a unique stationary distribution.

(v) To find the long run probability that the chain is at level 2 we need to calculate the unique stationary distribution π. This amounts to solving the matrix equation $\pi P = \pi$. This is a system of 4 equations in 4 unknowns given by

$$\pi_1 = .15\pi_1 + .15\pi_2 + .03\pi_3 \tag{87}$$
$$\pi_2 = .85\pi_1 + .12\pi_3 + .03\pi_4 \tag{88}$$
$$\pi_3 = .85\pi_2 + .12\pi_4 \tag{89}$$
$$\pi_4 = .85\pi_3 + .85\pi_4. \tag{90}$$

We discard the first equation and replace it by

$$\pi_1 + \pi_2 + \pi_4 + \pi_4 = 1.$$

Using substitutions or any other method we solve the system to obtain

$$(\pi_1, \pi_2, \pi_3, \pi_4) = (.01424, .05269, .13996, .79311).$$

Let $p_{ij}^{(n)}$ be the n-step transition probability of an irreducible aperiodic Markov chain on a finite state space. Then, $\lim p_{ij}^{(n)} = \pi_j$ for each i and j. Thus the long run probability that the chain is in state 2 is given by $\pi_2 = .05269$.

8.6 Chapter 6 solutions

1. Claims are known to follow a Poisson process with a uniform rate of 3 per day.

(a) Calculate the probability that there will be fewer than 1 claim on a given day.

(b) Estimate the probability that another claim will be reported during the next hour. State all assumptions made.

(c) If there have not been any claims for over a week, calculate the expected time before a new claim occurs.

Answer: Let $\{N_t\}_{t\in[0,\infty)}$ denote our Poisson process with rate $\lambda = 3$, where the time is measured in days.

(a) We have to evaluate $P(N_{t+1} - N_t < 1)$ for a fixed $t \geq 0$. But this is equal to

$$P(N_1 = 0) = \mathrm{e}^{-\lambda} = \mathrm{e}^{-3} = 0.04979.$$

(b) We look for the probability that, during the time interval $(t, t + \frac{1}{24}]$ for a fixed t, at least one claim will be reported, i.e.

$$\begin{aligned} P(N_{t+1/24} - N_t \geq 1) &= P(N_{1/24} \geq 1) = 1 - P(N_{1/24} = 0) = 1 - \mathrm{e}^{-\lambda/24} \\ &= 1 - \mathrm{e}^{-1/8} = 0.11750. \end{aligned}$$

(c) Conditional on $N_7 = 0$, we can assume that $\{N_{t+7}\}_{t\in[0,\infty)}$ behaves like a Poisson process $(\widetilde{N}_t)_{t\in[0,\infty)}$ with parameter $\lambda = 3$. But here the first jump (claim) occurs at a random time $\widetilde{\tau}_1$, which has an exponential distribution with parameter λ. It is well-known that the expectation is $\mathrm{E}(\widetilde{\tau}_1) = \frac{1}{\lambda} = \frac{1}{3}$.

2. Prove equation (38), which gives the Chapman-Kolmogorov equations for a Markov jump process.

Answer: Let $\{X_t\}_{t\in[0,\infty)}$ be a (not necessarily time-homogeneous) Markov process with discrete state space $\mathcal{S}$ and transition probabilities $p_{i,j}(s,t) = P(X_t = j \,|\, X_s = i)$ where $i, j \in \mathcal{S}$, $0 \leq s < t < \infty$ and we assume that $P(X_s = i) > 0$. We have to show that

$$p_{i,j}(t_1, t_3) = \sum_{k\in\mathcal{S}} p_{i,k}(t_1, t_2)\, p_{k,j}(t_2, t_3),$$

where $i, j \in \mathcal{S}$ and $0 \leq t_1 < t_2 < t_3 < \infty$. We have

$$\begin{aligned} p_{i,j}(t_1, t_3) &= P(X_{t_3} = j \,|\, X_{t_1} = i) \\ &= \sum_{k \in \mathcal{S}} P(X_{t_3} = j,\, X_{t_2} = k \,|\, X_{t_1} = i) \\ &= \sum_{k \in \mathcal{S}} P(X_{t_3} = j \,|\, X_{t_1} = i,\, X_{t_2} = k)\, P(X_{t_2} = k \,|\, X_{t_1} = i) \\ &= \sum_{k \in \mathcal{S}} P(X_{t_3} = j \,|\, X_{t_2} = k)\, P(X_{t_2} = k \,|\, X_{t_1} = i) \\ &= \sum_{k \in \mathcal{S}} p_{i,k}(t_1, t_2)\, p_{k,j}(t_2, t_3). \end{aligned}$$

3. Consider the sickness-death model given in Figure 9, write down an integral expression for $p_{HD}(s, t)$.

Answer: We have

$$\begin{aligned} p_{HD}(s,t) &= \int_0^{t-s} \Big(p_{SD}(s+w, t)\sigma(s+w) + p_{DD}(s+w, t)\mu(s+w) \Big) \\ &\quad \times \exp\Big(-\int_s^{s+w} (\sigma(u) + \mu(u))\, du \Big) dw \\ &= \int_0^{t-s} \Big(p_{SD}(s+w, t)\sigma(s+w) + \mu(s+w) \Big) \\ &\quad \times \exp\Big(-\int_s^{s+w} (\sigma(u) + \mu(u))\, du \Big) dw. \end{aligned}$$

The individual remains in the *healthy* state from time s to time $s+w$ and then jumps to the state *dead* (where he remains) or to the state *sick* (where he jumps to state *dead* by time t). Note that here $p_{DD}(s+w, t) = 1$. Further note that the formula for $p_{HS}(s, t)$ did not contain the term $p_{DS}(s+w, t)\mu(s+w)$, since the probability to jump from *dead* to *sick* is equal to zero.

4. Let $\{X_t,\ t \geq 0\}$ be a time-homogeneous Markov process with state space $\mathcal{S} = \{0, 1\}$ and transition rates $q_{01} = \alpha$, $q_{10} = \beta$.

(a) Write down the generator matrix for this process.

(b) Solve the Kolmogorov's forward equations for this Markov jump process to find all transition probabilies.

(c) Check that the Chapman–Kolmogorov equations hold.

(d) What is the probability that the process will be in state 0 in the long term? Does it depend on the initial state?

Answer: (a) The generator matrix $\mathbf{Q}$ is given by

$$\mathbf{Q} = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}.$$

(b) The Kolmogorov forward equations $\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{Q}$ therefore become

$$\frac{dp_{00}(t)}{dt} = -\alpha p_{00}(t) + \beta p_{01}(t),$$

$$\frac{dp_{01}(t)}{dt} = \alpha p_{00}(t) - \beta p_{01}(t),$$

$$\frac{dp_{10}(t)}{dt} = -\alpha p_{10}(t) + \beta p_{11}(t),$$

$$\frac{dp_{11}(t)}{dt} = \alpha p_{10}(t) - \beta p_{11}(t),$$

Substituting $p_{01}(t) = 1 - p_{00}(t)$ in the first equation, we get equation

$$\frac{dp_{00}(t)}{dt} = -\alpha p_{00}(t) + \beta(1 - p_{00}(t)) = -(\alpha + \beta)p_{00}(t) + \beta.$$

which has a general solution

$$p_{00}(t) = \frac{\beta}{\alpha + \beta} + Ce^{-(\alpha+\beta)t}.$$

Initial condition $p_{00}(0) = 1$ lealds to $C = \frac{\alpha}{\alpha+\beta}$, so finally we get

$$p_{00}(t) = \frac{\beta}{\alpha + \beta} + \frac{\alpha}{\alpha + \beta}e^{-(\alpha+\beta)t}.$$

Transition probabilities $p_{01}(t)$, $p_{10}(t)$, and $p_{11}(t)$ can be found similarly. They are:

$$p_{01}(t) = \frac{\alpha}{\alpha + \beta} - \frac{\alpha}{\alpha + \beta}e^{-(\alpha+\beta)t};$$

$$p_{10}(t) = \frac{\beta}{\alpha + \beta} - \frac{\beta}{\alpha + \beta}e^{-(\alpha+\beta)t};$$

$$p_{11}(t) = \frac{\alpha}{\alpha + \beta} + \frac{\beta}{\alpha + \beta}e^{-(\alpha+\beta)t}.$$

(c) For a time-homogeneous Markov process Chapman–Kolmogorov equations take the form

$$p_{ij}(t+s) = \sum_{k \in \mathcal{S}} p_{ik}(t)\ p_{kj}(s).$$

In our case, $\mathcal{S} = \{0, 1\}$, thus there are 4 equations. For example, for $i = j = 0$ we get

$$p_{00}(t+s) = p_{00}(t)\ p_{00}(s) + p_{01}(t)\ p_{10}(s).$$

So we should check that

$$\begin{aligned}
&\frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta} e^{-(\alpha+\beta)(t+s)} = \\
&\left(\frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta} e^{-(\alpha+\beta)t}\right)\left(\frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta} e^{-(\alpha+\beta)s}\right) + \\
&\left(\frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta} e^{-(\alpha+\beta)t}\right)\left(\frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta} e^{-(\alpha+\beta)s}\right),
\end{aligned}$$

which is a straighforward exersize. Three other Chapman–Kolmogorov equations can be checked similarly.

(d)

$$\lim_{t\to\infty} p_{00}(t) = \lim_{t\to\infty} p_{10}(t) = \frac{\beta}{\alpha+\beta},$$

so the the probability that the process will be in state 0 in the long term is $\frac{\beta}{\alpha+\beta}$, and it does not depend on the initial state.

8.7 Chapter 7 solutions

1. Approximate points $(0,0,0)$, $(1,0,2)$, $(0,1,3)$, $(1,1,4)$ in the coordinate space (x_1, x_2, y) by a plane $y = ax_1 + bx_2 + c$ to minimize the sum of squares error.

Answer: For $(x_1, x_2) = (0,0)$, $y = a \cdot 0 + b \cdot 0 + c = c$, and the data point is $(0,0,0)$ so the (squared) error is $(c-0)^2$. Similarly, for $(x_1, x_2) = (1,0)$, $y = a + c$, and the data point is $(1,0,2)$, the squared error is $(a+c-2)^2$. Continuing this way, we write down the error as

$$e(a,b,c) = (c-0)^2 + (a+c-2)^2 + (b+c-3)^2 + (a+b+c-4)^2.$$

In optimality

$$\frac{\partial e(a,b,c)}{\partial a} = 2(-6+2a+b+2c) = 0$$
$$\frac{\partial e(a,b,c)}{\partial b} = 2(-7+a+2b+2c) = 0$$
$$\frac{\partial e(a,b,c)}{\partial c} = 2(-9+2a+2b+4c) = 0$$

and the solution is $a = 3/2$, $b = 5/2$, $c = 1/4$.

2. There are two points marked on the plane - red point A with coordinate $(0,0)$ and blue point B with coordinates $(10,10)$. Then 4 points C, D, E, F arrives in order, and each is coloured in the same way as its nearest neighbour. The coordinates of F is $(10,8)$. Give examples of coordinates of points C, D, E such that point F will be coloured red.

Answer: For example, C may have coordinates $(5,4)$, D - $(8,6)$, and E - $(10,7)$. Then

$$|CA| = \sqrt{(5-0)^2 + (4-0)^2} < \sqrt{(5-10)^2 + (4-10)^2} = |CB|,$$

hence C will be coloured the same as A, that is, in red. Next,

$$|DC| = \sqrt{(8-5)^2 + (6-4)^2} < \sqrt{(8-10)^2 + (6-10)^2} = |DB|,$$

hence D will be coloured the same as C, that is, in red. Further,

$$|ED| = \sqrt{(10-8)^2 + (7-6)^2} < \sqrt{(10-10)^2 + (7-10)^2} = |EB|,$$

hence E will be coloured the same as D, that is, in red. Finally,

$$|FE| = 1 < 2 = |FB|,$$

hence F will be coloured the same as E: in red.

3. Consider 4 points A, B, C, D on the plane with coordinates $(0,0)$, $(1,0)$, $(0,1)$, and $(2,1)$, initially classified such that A and D are red while B and C are blue. What would be the outcome of the K-means algorithm (with $K=2$) applied to these initial data?

Answer: Them mean M_R of red points is $((0,0)+(2,1))/2 = (1,0.5)$. The mean M_B of blue points is $((1,0)+(0,1))/2 = (0.5,0.5)$. Now, for point A,

$$|AM_B| = \sqrt{(0.5)^2+(0.5)^2} < \sqrt{(0.5)^2+1^2} = |AM_R|,$$

hence point A becomes blue. By similar argument, we deduce that C stays blue, B becomes red, and D stays red. After this, the mean M_R of red points is $((1,0)+(2,1))/2 = (1.5,0.5)$, while the mean M_B of blue points is $((0,0)+(0,1))/2 = (0,0.5)$. Now, for point A,

$$|AM_B| = \sqrt{0^2+(0.5)^2} < \sqrt{(0.5)^2+(1.5)^2} = |AM_R|,$$

hence point A stays blue. By similar argument, we deduce that all points stay the same colour, and the algorithm terminates. Answer: A and C are blue, B and D are red.

4. A filter should classify e-mails into 3 categories: personal, work, and spam. The statistics shows that approximately 30% of e-mails are personal, 50% are work ones, and 20% are spam. It also shows that word "friend" is included into 20% of personal e-mails, 5% of work e-mails, and 30% of spam e-mails. In addition, the word "profit" is included into 5% of personal e-mails, 30% of work e-mails, and 25% of spam e-mails. A new e-mails arrives which contains both words "friend" and "profit". Use naive Bayes classification to decide if this e-mail is more likely to be personal, work, or spam?

Answer: Let X be a random variable equal to $X=1$, $X=2$, or $X=3$ if the random e-mail is personal, work, or spam, respectively. Let F we a random variable such that $F=1$ if e-mail contains word "friend" (and $F=0$) otherwise. Similarly, let R we a random variable such that $R=1$ if e-mail contains word "profit" (and $R=0$) otherwise. We need to compare three conditional probabilities

$$P(X=1|R=F=1), \quad P(X=2|R=F=1), \quad P(X=3|R=F=1),$$

and select the maximal one. For each $i=1,2,3$, the naive Bayes estimate for $P(X=i|R=F=1)$ is

$$P(X=i|R=F=1) = \frac{P(R=1|X=i)\cdot P(F=1|X=i)\cdot P(X=i)}{P(R=F=1)}.$$

As we can see, the denominators are the same foe all i, so it suffices to compare the numerators. We have

$$P(R = 1|X = 1) \cdot P(F = 1|X = 1) \cdot P(X = 1) = 0.05 \cdot 0.2 \cdot 0.3 = 0.003,$$

$$P(R = 1|X = 2) \cdot P(F = 1|X = 2) \cdot P(X = 2) = 0.3 \cdot 0.05 \cdot 0.5 = 0.0075,$$

$$P(R = 1|X = 3) \cdot P(F = 1|X = 3) \cdot P(X = 3) = 0.25 \cdot 0.3 \cdot 0.2 = 0.015.$$

Hence, the e-mail is most likely to be a spam.