International Series on Actuarial Science

Nonlife Actuarial Models
Theory, Methods and Evaluation

Yiu-Kuen Tse

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521764650

This page intentionally left blank

Nonlife Actuarial Models

Actuaries must pass exams, but more than that: they must put knowledge into practice. This coherent book gives complete syllabus coverage for Exam C of the Society of Actuaries (SOA) while emphasizing the concepts and practical application of nonlife actuarial models. Ideal for those approaching their professional exams, it is also a class-tested textbook for undergraduate university courses in actuarial science.

All the topics that students need to prepare for Exam C are here, including modeling of losses, risk, and ruin theories, credibility theory and applications, and empirical implementation of loss models. The book also covers more recent topics, such as risk measures and bootstrapping. Readers are assumed to have studied statistical inference and probability at the introductory undergraduate level.

Numerous examples and exercises are provided, with many exercises adapted from past Exam C questions. Computational notes on the use of Excel are included. Teaching slides are available for download.

International Series on Actuarial Science

Christopher Daykin, Independent Consultant and Actuary
Angus Macdonald, Heriot-Watt University

The *International Series on Actuarial Science*, published by Cambridge University Press in conjunction with the Institute of Actuaries and the Faculty of Actuaries, contains textbooks for students taking courses in or related to actuarial science, as well as more advanced works designed for continuing professional development or for describing and synthesizing research. The series is a vehicle for publishing books that reflect changes and developments in the curriculum, that encourage the introduction of courses on actuarial science in universities, and that show how actuarial science can be used in all areas where there is long-term financial risk.

NONLIFE ACTUARIAL MODELS

Theory, Methods and Evaluation

YIU-KUEN TSE
Singapore Management University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521764650

First published in print format 2009

ISBN-13 978-0-511-65198-4 eBook (NetLibrary)

ISBN-13 978-0-521-76465-0 Hardback

To Vicky

Contents

Preface

This book is on the theory, methods, and empirical implementation of nonlife actuarial models. It is intended for use as a textbook for senior undergraduates. Users are assumed to have done one or two one-semester courses on probability theory and statistical inference, including estimation and hypothesis testing. The coverage of this book includes all the topics found in Exam C of the Society of Actuaries (Exam 4 of the Casualty Actuarial Society) as per the 2007 Basic Education Catalog. In addition, it covers some topics (such as risk measures and ruin theory) beyond what is required by these exams, and may be used by actuarial students in general.

This book is divided into four parts: loss models, risk and ruin, credibility, and model construction and evaluation. An appendix on the review of statistics is provided for the benefit of students who require a quick summary. Students may read the appendix prior to the main text if they desire, or they may use the appendix as a reference when required. In order to be self contained, the appendix covers some of the topics developed in the main text.

Some features of this book should be mentioned. First, the concepts and theories introduced are illustrated by many practical examples. Some of these examples explain the theory through numerical applications, while others develop new results. Second, several chapters of the book include a section on numerical computation using Excel. Students are encouraged to use Excel to solve some of the numerical exercises. Third, each chapter includes some exercises for practice. Many of these exercises are adapted from past exam questions of the Society of Actuaries.

I would like to thank Tao Yang for painstakingly going through the manuscript and for providing many useful comments and suggestions. Diana Gillooly has professionally guided me through the publication process with admirable patience and efficiency. Clare Dennison has performed a superb job of

coordinating the copy editing. I am also grateful to the Society of Actuaries for allowing me to use its past exam questions.

Resources are available at: www.mysmu.edu/faculty/yktse/NAM/NAMbase.htm. Slides in pdf format can be downloaded from this site, which will facilitate classroom teaching by instructors adopting this book. An errata file will be provided, and the solution manual for instructors is obtainable from the author on request.

Yiu-Kuen Tse, Ph.D. FSA
Singapore Management University
yktse@smu.edu.sg

Notation and convention

1 Abbreviations are used in this book without periods. For example, "probability density function" is referred to as pdf (not p.d.f.) and "moment generating function" is referred to as mgf (not m.g.f.).
2 We do not make distinctions between a random variable and the distribution that describes the random variable. Thus, from time to time we make statements such as: "X denotes the binomial distribution".
3 We use calligraphic fonts to denote commonly used distributions. Discrete distributions are denoted with two alphabets and continuous distributions are denoted with one alphabet. For example, $\mathcal{PN}$ stands for Poisson, $\mathcal{BN}$ stands for binomial, $\mathcal{N}$ stands for normal, and $\mathcal{L}$ stands for lognormal.
4 The following conventions are generally used:
   (a) Slanted upper case for random variables, e.g. X.
   (b) Slanted lower case for fixed numbers, e.g. x.
   (c) Slanted bold-faced upper case for vectors of random variables, e.g. $\boldsymbol{X}$.
   (d) Slanted bold-faced lower case for vectors of fixed numbers (observations), e.g. $\boldsymbol{x}$.
   (e) Upright bold-faced upper case for matrices of fixed numbers (observations), e.g. $\mathbf{X}$.
5 Natural logarithm is denoted by log, not ln.

Computation notes

1 In some chapters we include a section of Excel computation notes to discuss the use of Excel functions to facilitate computation. These functions require the Excel add-ins Analysis ToolPak and Solver Add-in.
2 Other computer languages for more advanced statistical analysis include R, C^{++}, Splus, Gauss, and Matlab. All graphs in this book were produced using Matlab, and many of the computations were performed using Gauss.

Part I

Loss models

In this part of the book we discuss actuarial models for claim losses. The two components of claim losses, namely claim frequency and claim severity, are modeled separately, and are then combined to derive the aggregate-loss distribution. In Chapter 1, we discuss the modeling of claim frequency, introducing some techniques for modeling nonnegative integer-valued random variables. Techniques for modeling continuous random variables relevant for claim severity are discussed in Chapter 2, in which we also consider the effects of coverage modifications on claim frequency and claim severity. Chapter 3 discusses the collective risk model and individual risk model for analyzing aggregate losses. The techniques of convolution and recursive methods are used to compute the aggregate-loss distributions.

1

Claim-frequency distribution

This book is about modeling the claim losses of insurance policies. Our main interest is nonlife insurance policies covering a fixed period of time, such as vehicle insurance, workers compensation insurance, and health insurance. An important measure of claim losses is the claim frequency, which is the number of claims in a block of insurance policies over a period of time. Though claim frequency does not directly show the monetary losses of insurance claims, it is an important variable in modeling the losses.

In this chapter we first briefly review some tools in modeling statistical distributions, in particular the moment generating function and probability generating function. Some commonly used discrete random variables in modeling claim-frequency distributions, namely the binomial, geometric, negative binomial, and Poisson distributions, are then discussed. We introduce a family of distributions for nonnegative, integer-valued random variables, called the $(a, b, 0)$ class, which includes all the four distributions aforementioned. This class of discrete distributions has found important applications in the actuarial literature. Further methods of creating new nonnegative, integer-valued random variables are introduced. In particular, we discuss the zero-modified distribution, the $(a, b, 1)$ class of distributions, the compound distributions, and the mixture distributions.

Learning objectives

1 Discrete distributions for modeling claim frequency
2 Binomial, geometric, negative binomial, and Poisson distributions
3 The $(a, b, 0)$ and $(a, b, 1)$ class of distributions
4 Compound distribution
5 Convolution
6 Mixture distribution

1.1 Claim frequency, claim severity, and aggregate claim

We consider a block of nonlife insurance policies with coverage over a fixed period of time. The **aggregate claim** for losses of the block of policies is the sum of the monetary losses of all the claims. The number of claims in the block of policies is called the **claim frequency**, and the monetary amount of each claim is called the **claim severity** or **claim size**. A general approach in loss modeling is to consider claim frequency and claim severity separately. The two variables are then combined to model the aggregate claim. Naturally claim frequency is modeled as a nonnegative discrete random variable, while claim severity is continuously distributed.

In this chapter we focus on the claim-frequency distribution. We discuss some nonnegative discrete random variables that are commonly used for modeling claim frequency. Some methods for constructing nonnegative discrete random variables that are suitable for modeling claim frequency are also introduced. As our focus is on short-term nonlife insurance policies, the time value of money plays a minor role. We begin with a brief review of some tools for modeling statistical distributions. Further discussions on the topic can be found in the Appendix, as well as the references therein.

1.2 Review of statistics

Let X be a random variable with **distribution function (df)** $F_X(x)$, which is defined by

$$F_X(x) = \Pr(X \le x). \tag{1.1}$$

If $F_X(x)$ is a continuous function, X is said to be a **continuous random variable**. Furthermore, if $F_X(x)$ is differentiable, the **probability density function (pdf)** of X, denoted by $f_X(x)$, is defined as

$$f_X(x) = \frac{dF_X(x)}{dx}. \tag{1.2}$$

If X can only take discrete values, it is called a **discrete random variable**. We denote $\Omega_X = \{x_1, x_2, \cdots\}$ as the set of values X can take, called the **support** of X. The **probability function (pf)** of a discrete random variable X, also denoted by $f_X(x)$, is defined as

$$f_X(x) = \begin{cases} \Pr(X = x), & \text{if } x \in \Omega_X, \\ 0, & \text{otherwise.} \end{cases} \tag{1.3}$$

We assume the support of a continuous random variable to be the real line, unless otherwise stated. The rth moment of X about zero (also called the rth

raw moment), denoted by $\mathrm{E}(X^r)$, is defined as

$$\mathrm{E}(X^r) = \int_{-\infty}^{\infty} x^r f_X(x)\, dx, \qquad \text{if } X \text{ is continuous,} \tag{1.4}$$

and

$$\mathrm{E}(X^r) = \sum_{x \in \Omega_X} x^r f_X(x), \qquad \text{if } X \text{ is discrete.} \tag{1.5}$$

For convenience, we also write $\mathrm{E}(X^r)$ as μ'_r. The **moment generating function (mgf)** of X, denoted by $M_X(t)$, is a function of t defined by

$$M_X(t) = \mathrm{E}(e^{tX}), \tag{1.6}$$

if the expectation exists. If the mgf of X exists for t in an open interval around $t = 0$, the moments of X exist and can be obtained by successively differentiating the mgf with respect to t and evaluating the result at $t = 0$. We observe that

$$M_X^r(t) = \frac{d^r M_X(t)}{dt^r} = \frac{d^r}{dt^r}\mathrm{E}(e^{tX}) = \mathrm{E}\left[\frac{d^r}{dt^r}(e^{tX})\right] = \mathrm{E}(X^r e^{tX}), \tag{1.7}$$

so that

$$M_X^r(0) = \mathrm{E}(X^r) = \mu'_r. \tag{1.8}$$

If $X_1, X_2, \cdots, X_n$ are **independently and identically distributed (iid)** random variables with mgf $M(t)$, and $X = X_1 + \cdots + X_n$, then the mgf of X is

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}(e^{tX_1 + \cdots + tX_n}) = \mathrm{E}\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} \mathrm{E}(e^{tX_i}) = [M(t)]^n. \tag{1.9}$$

The mgf has the important property that it uniquely defines a distribution. Specifically, if two random variables have the same mgf, their distributions are identical.[1]

If X is a random variable that can only take nonnegative integer values, the **probability generating function (pgf)** of X, denoted by $P_X(t)$, is defined as

$$P_X(t) = \mathrm{E}(t^X), \tag{1.10}$$

[1] See Appendix A.8 for more details.

if the expectation exists. The mgf and pgf are related through the following equations

$$M_X(t) = P_X(e^t), \tag{1.11}$$

and

$$P_X(t) = M_X(\log t). \tag{1.12}$$

Given the pgf of X, we can derive its pf. To see how this is done, note that

$$P_X(t) = \sum_{x=0}^{\infty} t^x f_X(x). \tag{1.13}$$

The rth order derivative of $P_X(t)$ is

$$P_X^r(t) = \frac{d^r}{dt^r}\left(\sum_{x=0}^{\infty} t^x f_X(x)\right) = \sum_{x=r}^{\infty} x(x-1)\cdots(x-r+1)t^{x-r} f_X(x). \tag{1.14}$$

If we evaluate $P_X^r(t)$ at $t = 0$, all terms in the above summation vanish except for $x = r$, which is $r! f_X(r)$. Hence, we have

$$P_X^r(0) = r! f_X(r), \tag{1.15}$$

so that given the pgf, we can obtain the pf as

$$f_X(r) = \frac{P_X^r(0)}{r!}. \tag{1.16}$$

In sum, given the mgf of X, the moments of X can be computed through equation (1.8). Likewise, given the pgf of a nonnegative integer-valued random variable, its pf can be computed through equation (1.16). Thus, the mgf and pgf are useful functions for summarizing a statistical distribution.

1.3 Some discrete distributions for claim frequency

We now review some key results of four discrete distributions, namely binomial, geometric, negative binomial, and Poisson. As these random variables can only take nonnegative integer values, they may be used for modeling the distributions of claim frequency. The choice of a particular distribution in practice is an empirical question to be discussed later.

1.3.1 Binomial distribution

A random variable X has a binomial distribution with parameters n and θ, denoted by $\mathcal{BN}(n,\theta)$, where n is a positive integer and θ satisfies $0 < \theta < 1$, if the pf of X is

$$f_X(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad \text{for } x = 0, 1, \cdots, n, \tag{1.17}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}. \tag{1.18}$$

The mean and variance of X are

$$\mathrm{E}(X) = n\theta \qquad \text{and} \qquad \mathrm{Var}(X) = n\theta(1-\theta), \tag{1.19}$$

so that the variance of X is always smaller than its mean.

The mgf of X is

$$M_X(t) = (\theta e^t + 1 - \theta)^n, \tag{1.20}$$

and its pgf is

$$P_X(t) = (\theta t + 1 - \theta)^n. \tag{1.21}$$

The expression in equation (1.17) is the probability of obtaining x successes in n independent trials each with probability of success θ. The distribution is symmetric if $\theta = 0.5$. It is positively skewed (skewed to the right) if $\theta < 0.5$, and is negatively skewed (skewed to the left) if $\theta > 0.5$. When n is large, X is approximately normally distributed. The convergence to normality is faster the closer θ is to 0.5.

There is a recursive relationship for $f_X(x)$, which can facilitate the computation of the pf. From equation (1.17), we have $f_X(0) = (1-\theta)^n$. Now for $x = 1, \cdots, n$, we have

$$\frac{f_X(x)}{f_X(x-1)} = \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{\binom{n}{x-1}\theta^{x-1}(1-\theta)^{n-x+1}} = \frac{(n-x+1)\theta}{x(1-\theta)}, \tag{1.22}$$

so that

$$f_X(x) = \left[\frac{(n-x+1)\theta}{x(1-\theta)}\right] f_X(x-1). \tag{1.23}$$

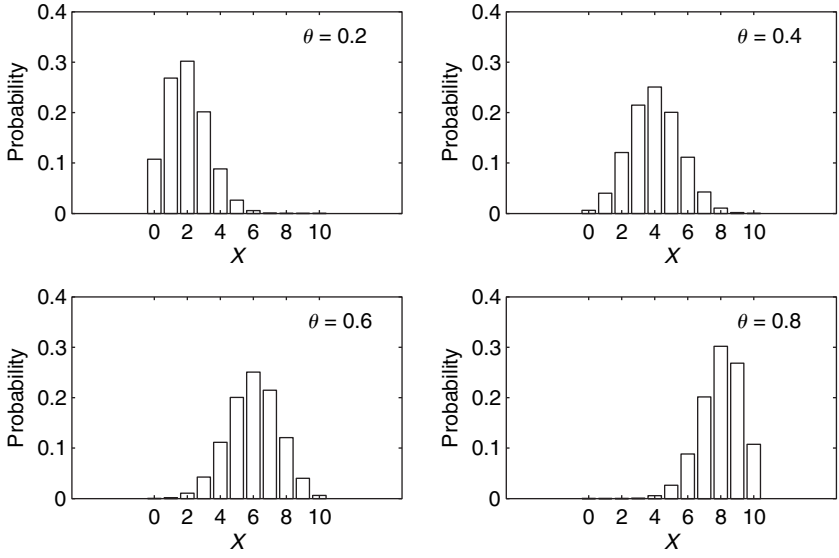

Figure 1.1 Probability function of $\mathcal{BN}(10, \theta)$

Example 1.1 Plot the pf of the binomial distribution for $n = 10$, and $\theta = 0.2$, 0.4, 0.6, and 0.8.

Solution Figure 1.1 plots the pf of $\mathcal{BN}(n, \theta)$ for $\theta = 0.2, 0.4, 0.6$, and 0.8, with $n = 10$.

It can be clearly seen that the binomial distribution is skewed to the right for $\theta = 0.2$ and skewed to the left for $\theta = 0.8$. □

1.3.2 Geometric distribution

A nonnegative discrete random variable X has a geometric distribution with parameter θ for $0 < \theta < 1$, denoted by $\mathcal{GM}(\theta)$, if its pf is given by

$$f_X(x) = \theta(1-\theta)^x, \quad \text{for } x = 0, 1, \cdots. \tag{1.24}$$

The mean and variance of X are

$$\mathrm{E}(X) = \frac{1-\theta}{\theta} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{1-\theta}{\theta^2}, \tag{1.25}$$

so that, in contrast to the binomial distribution, the variance of a geometric distribution is always larger than its mean.

The expression in equation (1.24) is the probability of having x failures prior to the first success in a sequence of independent Bernoulli trials with probability of success θ.

The mgf of X is

$$M_X(t) = \frac{\theta}{1-(1-\theta)e^t}, \tag{1.26}$$

and its pgf is

$$P_X(t) = \frac{\theta}{1-(1-\theta)t}. \tag{1.27}$$

The pf of X is decreasing in x. It satisfies the following recursive relationship

$$f_X(x) = (1-\theta) f_X(x-1), \tag{1.28}$$

for $x = 1, 2, \cdots$, with starting value $f_X(0) = \theta$.

1.3.3 Negative binomial distribution

A nonnegative discrete random variable X has a negative binomial distribution with parameters r and θ, denoted by $\mathcal{NB}(r, \theta)$, if the pf of X is

$$f_X(x) = \binom{x+r-1}{r-1}\theta^r(1-\theta)^x, \qquad \text{for } x = 0, 1, \cdots, \tag{1.29}$$

where r is a positive integer and θ satisfies $0 < \theta < 1$. The geometric distribution is a special case of the negative binomial distribution with $r = 1$. We may interpret the expression in equation (1.29) as the probability of getting x failures prior to the rth success in a sequence of independent Bernoulli trials with probability of success θ. Thus, $\mathcal{NB}(r, \theta)$ is just the sum of r independently distributed $\mathcal{GM}(\theta)$ variates. Hence, using equation (1.25), we can conclude that if X is distributed as $\mathcal{NB}(r, \theta)$, its mean and variance are

$$\mathrm{E}(X) = \frac{r(1-\theta)}{\theta} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{r(1-\theta)}{\theta^2}, \tag{1.30}$$

so that its variance is always larger than its mean.

Furthermore, using the results in equations (1.9), (1.26), and (1.27), we obtain the mgf of $\mathcal{NB}(r, \theta)$ as

$$M_X(t) = \left[\frac{\theta}{1-(1-\theta)e^t}\right]^r, \tag{1.31}$$

and its pgf as

$$P_X(t) = \left[\frac{\theta}{1-(1-\theta)t}\right]^r. \tag{1.32}$$

Note that the binomial coefficient in equation (1.29) can be written as

$$\begin{aligned}\binom{x+r-1}{r-1} &= \frac{(x+r-1)!}{(r-1)!x!} \\ &= \frac{(x+r-1)(x+r-2)\cdots(r+1)r}{x!}.\end{aligned} \tag{1.33}$$

The expression in the last line of the above equation is well defined for any number $r > 0$ (not necessarily an integer) and any nonnegative integer x.[2] Thus, if we define

$$\binom{x+r-1}{r-1} = \frac{(x+r-1)(x+r-2)\cdots(r+1)r}{x!}, \tag{1.34}$$

we can use equation (1.29) as a pf even when r is not an integer. Indeed, it can be verified that

$$\sum_{x=0}^{\infty}\binom{x+r-1}{r-1}\theta^r(1-\theta)^x = 1, \tag{1.35}$$

for $r > 0$ and $0 < \theta < 1$, so that the extension of the parameter r of the negative binomial distribution to any positive number is meaningful. We shall adopt this extension in any future applications.

The recursive formula of the pf follows from the result

$$\frac{f_X(x)}{f_X(x-1)} = \frac{\binom{x+r-1}{r-1}\theta^r(1-\theta)^x}{\binom{x+r-2}{r-1}\theta^r(1-\theta)^{x-1}} = \frac{(x+r-1)(1-\theta)}{x}, \tag{1.36}$$

so that

$$f_X(x) = \left[\frac{(x+r-1)(1-\theta)}{x}\right]f_X(x-1), \tag{1.37}$$

with starting value

$$f_X(0) = \theta^r. \tag{1.38}$$

[2] As factorials are defined only for nonnegative integers, the expression in the first line of equation (1.33) is not defined if r is not an integer.

Table 1.1. *Results of Example 1.2*

x	$r = 0.5$	$r = 1.0$	$r = 1.5$	$r = 2.0$
0	0.6325	0.4000	0.2530	0.1600
1	0.1897	0.2400	0.2277	0.1920
2	0.0854	0.1440	0.1708	0.1728
3	0.0427	0.0864	0.1195	0.1382

Example 1.2 Using the recursion formula, calculate the pf of the negative binomial distribution with $r = 0.5$, 1, 1.5, and 2, and $\theta = 0.4$, for $x = 0$, 1, 2, and 3. What is the mode of the negative binomial distribution?

Solution From the recursion formula in equation (1.37), we have

$$f_X(x) = \left[\frac{0.6(x + r - 1)}{x}\right] f_X(x - 1), \qquad \text{for } x = 1, 2, \cdots,$$

with starting value $f_X(0) = (0.4)^r$. We summarize the results in Table 1.1.

Note that the mode for $r = 0.5$, 1, and 1.5 is 0, and that for $r = 2$ is 1. To compute the mode in general, we note that, from equation (1.37)

$$f_X(x) > f_X(x - 1) \qquad \text{if and only if} \qquad \frac{(x + r - 1)(1 - \theta)}{x} > 1,$$

and the latter inequality is equivalent to

$$x < \frac{(r - 1)(1 - \theta)}{\theta}.$$

Therefore, the mode of the negative binomial distribution is equal to the nonnegative integer part of $(r - 1)(1 - \theta)/\theta$. We can verify this result from Table 1.1. For example, when $r = 2$,

$$\frac{(r - 1)(1 - \theta)}{\theta} = \frac{0.6}{0.4} = 1.5,$$

and its integer part (the mode) is 1. □

1.3.4 Poisson distribution

A nonnegative discrete random variable X is said to have a Poisson distribution with parameter λ, denoted by $\mathcal{PN}(\lambda)$, if the pf of X is given by

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad \text{for } x = 0, 1, \cdots, \tag{1.39}$$

where $\lambda > 0$. The mean and variance of X are

$$\mathrm{E}(X) = \mathrm{Var}(X) = \lambda. \tag{1.40}$$

The mgf of X is

$$M_X(t) = \exp\left[\lambda(e^t - 1)\right], \tag{1.41}$$

and its pgf is

$$P_X(t) = \exp\left[\lambda(t - 1)\right]. \tag{1.42}$$

The Poisson distribution is one of the most widely used discrete distributions in empirical applications. It is commonly applied to model the number of arrivals of certain events within a period of time (such as the number of insurance claims in a year), the number of defective items in production, and as an approximation of the binomial distribution, among others. We now introduce two properties of the Poisson distribution, which make it convenient to use.

Theorem 1.1 *If $X_1, \cdots, X_n$ are independently distributed with $X_i \sim \mathcal{PN}(\lambda_i)$, for $i = 1, \cdots, n$, then $X = X_1 + \cdots + X_n$ is distributed as a Poisson with parameter $\lambda = \lambda_1 + \cdots + \lambda_n$.*

Proof To prove this result, we make use of the mgf. Note that the mgf of X is

$$\begin{aligned}
M_X(t) &= \mathrm{E}(e^{tX}) \\
&= \mathrm{E}(e^{tX_1 + \cdots + tX_n}) \\
&= \mathrm{E}\left(\prod_{i=1}^{n} e^{tX_i}\right) \\
&= \prod_{i=1}^{n} \mathrm{E}(e^{tX_i}) \\
&= \prod_{i=1}^{n} \exp\left[\lambda_i(e^t - 1)\right] \\
&= \exp\left[(e^t - 1)\sum_{i=1}^{n} \lambda_i\right] \\
&= \exp\left[(e^t - 1)\lambda\right],
\end{aligned} \tag{1.43}$$

which is the mgf of $\mathcal{PN}(\lambda)$. Thus, by the uniqueness of mgf, $X \sim \mathcal{PN}(\lambda)$. □

It turns out that the converse of the above result is also true, as summarized in the following theorem.

Theorem 1.2 *Suppose an event A can be partitioned into m mutually exclusive and exhaustive events A_i, for $i = 1, \cdots, m$. Let X be the number of occurrences of A, and X_i be the number of occurrences of A_i, so that $X = X_1 + \cdots + X_m$. Let the probability of occurrence of A_i given A has occurred be p_i, i.e. $Pr(A_i \mid A) = p_i$, with $\sum_{i=1}^{m} p_i = 1$. If $X \sim \mathcal{PN}(\lambda)$, then $X_i \sim \mathcal{PN}(\lambda_i)$, where $\lambda_i = \lambda p_i$. Furthermore, $X_1, \cdots, X_m$ are independently distributed.*

Proof To prove this result, we first derive the marginal distribution of X_i. Given $X = x$, $X_i \sim \mathcal{BN}(x, p_i)$ for $i = 1, \cdots, m$. Hence, the marginal pf of X_i is (note that $x_i \le x$)

$$\begin{aligned} f_{X_i}(x_i) &= \sum_{x=x_i}^{\infty} \Pr(X_i = x_i \mid X = x) \Pr(X = x) \\ &= \sum_{x=x_i}^{\infty} \binom{x}{x_i} p_i^{x_i} (1-p_i)^{x-x_i} \left[\frac{e^{-\lambda} \lambda^x}{x!} \right] \\ &= \left[\frac{e^{-\lambda} (\lambda p_i)^{x_i}}{x_i!} \right] \sum_{x=x_i}^{\infty} \frac{[\lambda(1-p_i)]^{x-x_i}}{(x-x_i)!} \\ &= \left[\frac{e^{-\lambda} (\lambda p_i)^{x_i}}{x_i!} \right] e^{\lambda(1-p_i)} \\ &= \frac{e^{-\lambda p_i} (\lambda p_i)^{x_i}}{x_i!}, \end{aligned} \tag{1.44}$$

which is the pf of $\mathcal{PN}(\lambda p_i)$.

We now consider the joint pf of $X_1, \cdots, X_m$. Note that given $X = x$, the joint distribution of $X_1, \cdots, X_m$ is multinomial with parameters $x, p_1, \cdots, p_m$.[3] Thus

$$\Pr(X_1 = x_1, \cdots, X_m = x_m \mid X = x) = \frac{x!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m}. \tag{1.45}$$

By the multiplication rule of probability, we have

$$\begin{aligned} f_{X_1 X_2 \cdots X_m}(x_1, \cdots, x_m) &= \Pr(X_1 = x_1, \cdots, X_m = x_m \mid X = x) \Pr(X = x) \\ &= \left(\frac{x!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m} \right) \frac{e^{-\lambda} \lambda^x}{x!} \end{aligned}$$

[3] The multinomial distribution is a generalization of the binomial distribution; see DeGroot and Schervish (2002, p. 309).

$$= \prod_{i=1}^{m} \frac{e^{-\lambda p_i}(\lambda p_i)^{x_i}}{x_i!}$$

$$= \prod_{i=1}^{m} f_{X_i}(x_i), \tag{1.46}$$

so that the joint pf of $X_1, \cdots, X_m$ is the product of their marginal pf. This completes the proof that $X_1, \cdots, X_m$ are independent. □

Readers may verify the following recursive relationship of the pf of $\mathcal{PN}(\lambda)$

$$f_X(x) = \left(\frac{\lambda}{x}\right) f_X(x-1), \tag{1.47}$$

with $f_X(0) = e^{-\lambda}$. Finally, we add that when λ is large, $\mathcal{PN}(\lambda)$ is approximately normally distributed.

Example 1.3 The average number of female employees taking sick leave is 1.3 per week, and the average number of male employees taking sick leave is 2.5 per week. What is the probability of finding fewer than two sick leaves in a week? You may assume that the numbers of sick leaves in a week for the female and male employees are independent Poisson distributions.

Solution From Theorem 1.1, the number of sick leaves in each week is Poisson with mean 1.3 + 2.5 = 3.8. Thus, the required probability is

$$f_X(0) + f_X(1) = e^{-3.8} + 3.8e^{-3.8} = 0.1074. \quad \square$$

Example 1.4 Bank A has two blocks of loans: housing loans and study loans. The total number of defaults is Poisson distributed with mean 23, and 28% of the defaults are study loans. Bank B has three blocks of loans: housing loans, study loans, and car loans. The total number of defaults is Poisson distributed with mean 45, where 21% of the defaults are study loans and 53% are housing loans. The defaults of the two banks are independent. If the loan portfolios of the two banks are merged, find the distribution of the defaults of study loans and car loans.

Solution From Theorem 1.2, defaults of study loans of Bank A is Poisson with mean (23)(0.28) = 6.44, and that of Bank B is Poisson with mean (45)(0.21) = 9.45. Thus, in the merged portfolio, defaults of study loans, by Theorem 1.1, is Poisson with mean 6.44 + 9.45 = 15.89.

As Bank A has no car loans, defaults of car loans come from Bank B only, which is Poisson distributed with mean (45)(1 − 0.21 − 0.53) = 11.70. □

1.4 The $(a, b, 0)$ class of distributions

The binomial, geometric, negative binomial, and Poisson distributions belong to a class of nonnegative discrete distributions called the $(a, b, 0)$ class in the actuarial literature. Below is the definition of this class of distributions.

Definition 1.1 A nonnegative discrete random variable X is in the $(a, b, 0)$ class if its pf $f_X(x)$ satisfies the following recursion

$$f_X(x) = \left(a + \frac{b}{x}\right) f_X(x-1), \quad \text{for } x = 1, 2, \cdots, \tag{1.48}$$

where a and b are constants, with given $f_X(0)$.

As an example, we consider the binomial distribution. Equation (1.23) can be written as follows

$$f_X(x) = \left[-\frac{\theta}{1-\theta} + \frac{\theta(n+1)}{(1-\theta)x}\right] f_X(x-1). \tag{1.49}$$

Thus, if we let

$$a = -\frac{\theta}{1-\theta} \quad \text{and} \quad b = \frac{\theta(n+1)}{(1-\theta)}, \tag{1.50}$$

the pf of the binomial distribution satisfies equation (1.48) and thus belongs to the $(a, b, 0)$ class. Readers may verify the results in Table 1.2, which show that the four discrete distributions discussed in the last section belong to the $(a, b, 0)$ class.[4]

The $(a, b, 0)$ class of distributions is defined by recursions starting at $x = 1$, given an initial value $f_X(0)$. By analogy, we can define a class of nonnegative discrete distributions, called the $(a, b, 1)$ class, with recursions starting at $x = 2$ given an initial value $f_X(1)$.

Definition 1.2 A nonnegative discrete random variable X belongs to the $(a, b, 1)$ class if its pf $f_X(x)$ satisfies the following recursion

$$f_X(x) = \left(a + \frac{b}{x}\right) f_X(x-1), \quad \text{for } x = 2, 3, \cdots, \tag{1.51}$$

where a and b are constants, with given initial value $f_X(1)$.

Note that the $(a, b, 0)$ and $(a, b, 1)$ classes have the same recursion formulas, and they differ only at the starting point of the recursion. Also, the probability

[4] This class of distributions has only four members and no others. See Dickson (2005, Section 4.5.1), for a proof of this result.

Table 1.2. *The* $(a, b, 0)$ *class of distributions*

Distribution	a	b	$f_X(0)$
Binomial: $\mathcal{BN}(n, \theta)$	$-\frac{\theta}{1-\theta}$	$\frac{\theta(n+1)}{1-\theta}$	$(1-\theta)^n$
Geometric: $\mathcal{GM}(\theta)$	$1-\theta$	0	θ
Negative binomial: $\mathcal{NB}(r, \theta)$	$1-\theta$	$(r-1)(1-\theta)$	θ^r
Poisson: $\mathcal{PN}(\lambda)$	0	λ	$e^{-\lambda}$

$f_X(0)$ of a random variable belonging to the $(a, b, 1)$ class needs not be zero. Thus, it is possible to create a distribution with a specific probability at point zero and yet a shape similar to one of the $(a, b, 0)$ distributions. This flexibility is of considerable importance because insurance claims of low-risk events are infrequent. It may be desirable to obtain a good fit of the distribution at zero claim based on empirical experience and yet preserve the shape to coincide with some simple parametric distributions. This can be achieved by specifying the zero probability while adopting the $(a, b, 1)$ recursion to mimic a selected $(a, b, 0)$ distribution.

Let $f_X(x)$ be the pf of a $(a, b, 0)$ distribution called the base distribution. We denote $f_X^M(x)$ as the pf that is a modification of $f_X(x)$, where $f_X^M(x)$ belongs to the $(a, b, 1)$ class. The probability at point zero, $f_X^M(0)$, is specified and $f_X^M(x)$ is related to $f_X(x)$ as follows

$$f_X^M(x) = c f_X(x), \qquad \text{for } x = 1, 2, \cdots, \tag{1.52}$$

where c is an appropriate constant. For $f_X^M(\cdot)$ to be a well-defined pf, we must have

$$\begin{aligned} 1 &= f_X^M(0) + \sum_{x=1}^{\infty} f_X^M(x) \\ &= f_X^M(0) + c \sum_{x=1}^{\infty} f_X(x) \\ &= f_X^M(0) + c[1 - f_X(0)]. \end{aligned} \tag{1.53}$$

Table 1.3. *Results of Example 1.5*

x	$\mathcal{BN}(4, 0.3)$	Zero-modified	Zero-truncated
0	0.2401	0.4000	0
1	0.4116	0.3250	0.5417
2	0.2646	0.2089	0.3482
3	0.0756	0.0597	0.0995
4	0.0081	0.0064	0.0107

Thus, we conclude that

$$c = \frac{1 - f_X^M(0)}{1 - f_X(0)}. \tag{1.54}$$

Substituting c into equation (1.52) we obtain $f_X^M(x)$, for $x = 1, 2, \cdots$. Together with the given $f_X^M(0)$, we have a $(a, b, 1)$ distribution with the desired zero-claim probability and the same recursion as the base $(a, b, 0)$ distribution. This is called the **zero-modified distribution** of the base $(a, b, 0)$ distribution. In particular, if $f_X^M(0) = 0$, the modified distribution cannot take value zero and is called the **zero-truncated distribution**. The zero-truncated distribution is a particular case of the zero-modified distribution.

Example 1.5 X is distributed as $\mathcal{BN}(4, 0.3)$. Compute the zero-modified pf with the probability of zero equal to 0.4. Also compute the zero-truncated pf.

Solution The results are summarized in Table 1.3. The second column of the table gives the pf of $\mathcal{BN}(4, 0.3)$, the third column gives the pf of the zero-modified distribution with $f_X^M(0) = 0.4$, and the fourth column gives the zero-truncated distribution. Note that, using equation (1.54), the constant c of the zero-modified distribution with $f_X^M(0) = 0.4$ is

$$c = \frac{1 - 0.4}{1 - 0.2401} = 0.7896.$$

Thus, $f_X^M(1) = (0.7896)(0.4116) = 0.3250$, and other values of $f_X^M(x)$ are obtained similarly. For the zero-truncated distribution, the value of c is

$$c = \frac{1}{1 - 0.2401} = 1.3160,$$

and the pf is computed by multiplying the second column by 1.3160, for $x = 1$, 2, 3, and 4.

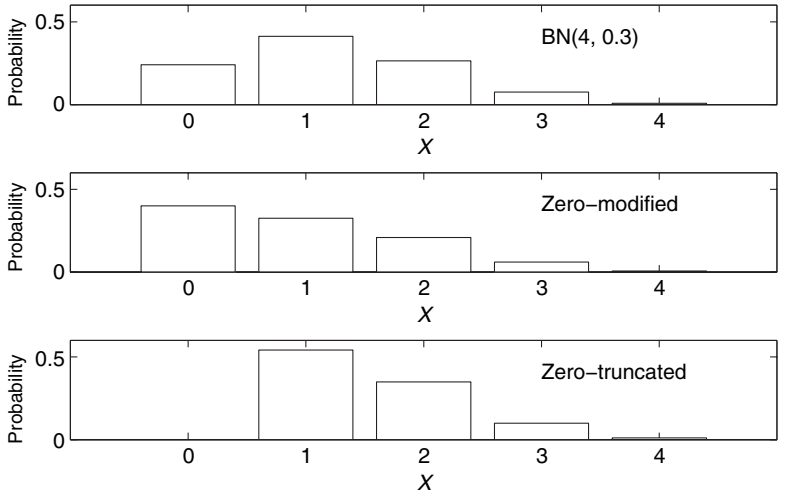

Figure 1.2 Probability functions of $\mathcal{BN}(4, 0.3)$ and its modifications

Figure 1.2 plots the pf of $\mathcal{BN}(4, 0.3)$, the zero-modified distribution and zero-truncated distribution.

From the plots, we can see that the zero-modified and zero-truncated distributions maintain similar shapes as that of the binomial distribution for $x \geq 1$. □

Example 1.6 X is distributed as $\mathcal{NB}(1.8, 0.3)$. What is the recursion formula for the pf of X? Derive the recursion formula for the zero-modified distribution of X with $f_X^M(0) = 0.6$. Compute the mean and variance of the zero-modified distribution.

Solution From Table 1.2, we have

$$a = 1 - \theta = 1 - 0.3 = 0.7,$$

and

$$b = (r - 1)(1 - \theta) = (1.8 - 1)(1 - 0.3) = 0.56.$$

Thus, from equation (1.48), the recursion is

$$f_X(x) = \left(0.7 + \frac{0.56}{x}\right) f_X(x - 1), \qquad \text{for } x = 1, 2, \cdots,$$

with the initial value of the recursion equal to

$$f_X(0) = (0.3)^{1.8} = 0.1145.$$

The recursion of the pf of the zero-modified distribution has the same formula as that of X, except that the starting point is $x = 2$ and the initial value is different. To compute the initial value, we first calculate c, which, from equation (1.54), is

$$c = \frac{1 - 0.6}{1 - 0.1145} = 0.4517.$$

Thus, the initial value is

$$f_X^M(1) = cf_X(1) = (0.4517)\left[1.8(0.3)^{1.8}(0.7)\right] = 0.0652,$$

and the recursion of the zero-modified distribution is

$$f_X^M(x) = \left(0.7 + \frac{0.56}{x}\right) f_X^M(x-1), \qquad \text{for } x = 2, 3, \cdots,$$

with $f_X^M(1) = 0.0652$.

To compute the mean and variance of the zero-modified distribution, we first note that its rth moment is

$$\begin{aligned}\sum_{x=0}^{\infty} x^r f_X^M(x) &= \sum_{x=1}^{\infty} x^r f_X^M(x) \\ &= c\sum_{x=1}^{\infty} x^r f_X(x) \\ &= c\,\mathrm{E}(X^r).\end{aligned}$$

From equation (1.30), the mean and variance of X are

$$\frac{r(1-\theta)}{\theta} = \frac{(1.8)(0.7)}{0.3} = 4.2,$$

and

$$\frac{r(1-\theta)}{\theta^2} = \frac{(1.8)(0.7)}{(0.3)^2} = 14,$$

respectively. Thus, $\mathrm{E}(X^2) = 14 + (4.2)^2 = 31.64$. Hence, the mean of the zero-modified distribution is $(0.4517)(4.2) = 1.8971$ and its raw second moment is $(0.4517)(31.64) = 14.2918$. Finally, the variance of the zero-modified distribution is $14.2918 - (1.8971)^2 = 10.6928$. □

We have seen that the binomial, geometric, negative binomial, and Poisson distributions can be unified under the $(a, b, 0)$ class of distributions. We shall

conclude this section with another unifying theme of these four distributions, as presented in the theorem below.

Theorem 1.3 *Let X denote the binomial, geometric, negative binomial, or Poisson distributions. The pgf $P_X(t)$ of X can be written as follows*

$$P_X(t \mid \beta) = Q_X[\beta(t-1)], \tag{1.55}$$

where β is a parameter of X and $Q_X(\cdot)$ is a function of $\beta(t-1)$ only. In other words, β and t appear in the pgf of X only through $\beta(t-1)$, although the function may depend on other parameters.

Proof First, we show this result for the binomial distribution. From equation (1.21), the pgf of $\mathcal{BN}(n, \theta)$ is

$$P_X(t) = [1 + \theta(t-1)]^n. \tag{1.56}$$

Thus, equation (1.55) is satisfied with $\theta = \beta$. Next, we prove the result for the negative binomial distribution, as the geometric distribution is a special case of it. From equation (1.32), the pgf of $\mathcal{NB}(r, \theta)$ is

$$P_X(t) = \left[\frac{\theta}{1-(1-\theta)t}\right]^r. \tag{1.57}$$

If we define

$$\beta = \frac{1-\theta}{\theta}, \tag{1.58}$$

then equation (1.57) becomes

$$P_X(t) = \left[\frac{1}{1-\beta(t-1)}\right]^r, \tag{1.59}$$

which satisfies equation (1.55). Finally, the pgf of $\mathcal{PN}(\lambda)$ is $P_X(t) = \exp[\lambda(t-1)]$, which satisfies equation (1.55) with $\lambda = \beta$. □

This theorem will be used in the next chapter.

1.5 Some methods for creating new distributions

We now introduce two models of statistical distributions through which we can create new distributions. These are the compound distributions and mixture distributions. We may use these methods to create discrete as well as continuous distributions. Our main purpose in this section, however, is to use them to create discrete distributions which may be used to model claim frequency.

1.5.1 Compound distribution

Let $X_1, \cdots, X_N$ be iid nonnegative integer-valued random variables, each distributed like X. We denote the sum of these random variables by S, so that

$$S = X_1 + \cdots + X_N. \tag{1.60}$$

If N is itself a nonnegative integer-valued random variable distributed independently of $X_1, \cdots, X_N$, then S is said to have a **compound distribution**. The distribution of N is called the **primary distribution**, and the distribution of X is called the **secondary distribution**. We shall use the *primary–secondary* convention to name a compound distribution. Thus, if N is Poisson and X is geometric, S has a Poisson–geometric distribution. A **compound Poisson** distribution is a compound distribution where the primary distribution is Poisson, for *any* secondary distribution. Terminology such as compound geometric distribution is similarly defined. While N is always integer valued (being the number of summation terms of X_i), X in general may be continuous. In this chapter we shall only consider the case of nonnegative integer-valued X, in which case S is also nonnegative integer valued.

Let us consider the simple case where N has a degenerate distribution taking value n with probability 1. S is thus the sum of n terms of X_i, where n is fixed. Suppose $n = 2$, so that $S = X_1 + X_2$. Then

$$\begin{aligned} f_S(s) &= \Pr(X_1 + X_2 = s) \\ &= \sum_{x=0}^{s} \Pr(X_1 = x \text{ and } X_2 = s - x). \end{aligned} \tag{1.61}$$

As the pf of X_1 and X_2 are $f_X(\cdot)$, and X_1 and X_2 are independent, we have

$$f_S(s) = \sum_{x=0}^{s} f_X(s) f_X(s - x). \tag{1.62}$$

The above equation expresses the pf of S, $f_S(\cdot)$, as the **convolution** of $f_X(\cdot)$, denoted by $(f_X * f_X)(\cdot)$, i.e.

$$f_{X_1+X_2}(s) = (f_X * f_X)(s) = \sum_{x=0}^{s} f_X(x) f_X(s - x). \tag{1.63}$$

Convolutions can be evaluated recursively. When $n = 3$, the 3-fold convolution is

$$f_{X_1+X_2+X_3}(s) = (f_{X_1+X_2} * f_{X_3})(s) = (f_{X_1} * f_{X_2} * f_{X_3})(s) = (f_X * f_X * f_X)(s). \tag{1.64}$$

The last equality holds as X_i are identically distributed as X. For $n \geq 2$, the pf of S is the convolution $(f_X * f_X * \cdots * f_X)(\cdot)$ with n terms of f_X, and is denoted by $f_X^{*n}(\cdot)$.

Convolution is in general tedious to evaluate, especially when n is large. The following example illustrates the complexity of the problem.

Example 1.7 Let the pf of X be $f_X(0) = 0.1, f_X(1) = 0, f_X(2) = 0.4$, and $f_X(3) = 0.5$. Find the 2-fold and 3-fold convolutions of X.

Solution We first compute the 2-fold convolution. For $s = 0$ and 1, the probabilities are

$$(f_X * f_X)(0) = f_X(0) f_X(0) = (0.1)(0.1) = 0.01,$$

and

$$(f_X * f_X)(1) = f_X(0) f_X(1) + f_X(1) f_X(0) = (0.1)(0) + (0)(0.1) = 0.$$

Other values are similarly computed as follows

$$\begin{aligned}
(f_X * f_X)(2) &= (0.1)(0.4) + (0.4)(0.1) = 0.08, \\
(f_X * f_X)(3) &= (0.1)(0.5) + (0.5)(0.1) = 0.10, \\
(f_X * f_X)(4) &= (0.4)(0.4) = 0.16, \\
(f_X * f_X)(5) &= (0.4)(0.5) + (0.5)(0.4) = 0.40,
\end{aligned}$$

and

$$(f_X * f_X)(6) = (0.5)(0.5) = 0.25.$$

For the 3-fold convolution, we show some sample workings as follows

$$\begin{aligned}
f_X^{*3}(0) &= [f_X(0)]\left[f_X^{*2}(0)\right] = (0.1)(0.01) = 0.001, \\
f_X^{*3}(1) &= [f_X(0)]\left[f_X^{*2}(1)\right] + [f_X(1)]\left[f_X^{*2}(0)\right] = 0,
\end{aligned}$$

and

$$\begin{aligned}
f_X^{*3}(2) &= [f_X(0)]\left[f_X^{*2}(2)\right] + [f_X(1)]\left[f_X^{*2}(1)\right] + [f_X(2)]\left[f_X^{*2}(0)\right] \\
&= 0.012.
\end{aligned}$$

Other calculations are similar. Readers may verify the results summarized in Table 1.4.

Table 1.4. *Results of Example 1.7*

x	$f_X(x)$	$f_X^{*2}(x)$	$f_X^{*3}(x)$
0	0.1	0.01	0.001
1	0	0	0
2	0.4	0.08	0.012
3	0.5	0.10	0.015
4		0.16	0.048
5		0.40	0.120
6		0.25	0.139
7			0.240
8			0.300
9			0.125

□

Having illustrated the computation of convolutions, we now get back to the compound distribution in which the primary distribution N has a pf $f_N(\cdot)$. Using the total law of probability, we obtain the pf of the compound distribution S as

$$\begin{aligned} f_S(s) &= \sum_{n=0}^{\infty} \Pr(X_1 + \cdots + X_N = s \,|\, N = n) f_N(n) \\ &= \sum_{n=0}^{\infty} \Pr(X_1 + \cdots + X_n = s) f_N(n), \end{aligned} \tag{1.65}$$

in which the term $\Pr(X_1 + \cdots + X_n = s)$ can be calculated as the n-fold convolution of $f_X(\cdot)$. However, as the evaluation of convolution is usually quite complex when n is large, the use of equation (1.65) may be quite tedious.

We now discuss some useful properties of S, which will facilitate the computation of its pf.

Theorem 1.4 *Let S be a compound distribution. If the primary distribution N has mgf $M_N(t)$ and the secondary distribution X has mgf $M_X(t)$, then the mgf of S is*

$$M_S(t) = M_N[\log M_X(t)]. \tag{1.66}$$

If N has pgf $P_N(t)$ and X is nonnegative integer valued with pgf $P_X(t)$, then the pgf of S is

$$P_S(t) = P_N[P_X(t)]. \tag{1.67}$$

Proof The proof makes use of results in conditional expectation, which can be found in Appendix A.11.[5] We note that

$$\begin{aligned} M_S(t) &= \mathrm{E}\left(e^{tS}\right) \\ &= \mathrm{E}\left(e^{tX_1+\cdots+tX_N}\right) \\ &= \mathrm{E}\left[\mathrm{E}\left(e^{tX_1+\cdots+tX_N} \mid N\right)\right] \\ &= \mathrm{E}\left\{\left[\mathrm{E}\left(e^{tX}\right)\right]^N\right\} \\ &= \mathrm{E}\left\{[M_X(t)]^N\right\} \\ &= \mathrm{E}\left\{\left[e^{\log M_X(t)}\right]^N\right\} \\ &= M_N[\log M_X(t)]. \end{aligned} \tag{1.68}$$

For the pgf, we have

$$\begin{aligned} P_S(t) &= \mathrm{E}\left(t^S\right) \\ &= \mathrm{E}(t^{X_1+\cdots+X_N}) \\ &= \mathrm{E}\left[\mathrm{E}(t^{X_1+\cdots+X_N} \mid N)\right] \\ &= \mathrm{E}\left\{\left[\mathrm{E}(t^X)\right]^N\right\} \\ &= \mathrm{E}\left\{[P_X(t)]^N\right\} \\ &= P_N[P_X(t)]. \end{aligned} \tag{1.69}$$

□

Equation (1.69) provides a method to compute the pf of S. We note that

$$f_S(0) = P_S(0) = P_N[P_X(0)], \tag{1.70}$$

and, from equation (1.16), we have

$$f_S(1) = P_S'(0). \tag{1.71}$$

[5] Appendix A.11 discusses other results in conditional expectations, which will be used in later parts of this section.

The derivative $P'_S(t)$ may be computed by differentiating $P_S(t)$ directly, or by the chain rule using the derivatives of $P_N(t)$ and $P_X(t)$, i.e.

$$P'_S(t) = \left\{P'_N[P_X(t)]\right\} P'_X(t). \tag{1.72}$$

Other values of the pf of S may be calculated similarly, although the complexity of the differentiation becomes more involved.

Example 1.8 Let $N \sim \mathcal{PN}(\lambda)$ and $X \sim \mathcal{GM}(\theta)$. Calculate $f_S(0)$ and $f_S(1)$.

Solution The pgf of N is

$$P_N(t) = \exp[\lambda(t-1)],$$

and the pgf of X is

$$P_X(t) = \frac{\theta}{1-(1-\theta)t}.$$

From equation (1.67), the pgf of S is

$$P_S(t) = P_N[P_X(t)] = \exp\left[\lambda\left(\frac{\theta}{1-(1-\theta)t} - 1\right)\right],$$

from which we obtain

$$f_S(0) = P_S(0) = \exp\left[\lambda\left(\theta - 1\right)\right].$$

Note that as $\theta - 1 < 0, f_S(0) < 1$. To calculate $f_S(1)$, we differentiate $P_S(t)$ directly to obtain

$$P'_S(t) = \exp\left[\lambda\left(\frac{\theta}{1-(1-\theta)t} - 1\right)\right] \frac{\lambda\theta(1-\theta)}{[1-(1-\theta)t]^2},$$

so that

$$f_S(1) = P'_S(0) = \exp\left[\lambda\left(\theta - 1\right)\right] \lambda\theta(1-\theta).$$

□

Suppose the primary distribution of a compound Poisson distribution S has parameter λ, and the secondary distribution has a pgf $P(t)$, then using equation (1.69) the pgf of S is

$$P_S(t) = P_N[P(t)] = \exp\{\lambda[P(t) - 1]\}. \tag{1.73}$$

By the uniqueness of the pgf, equation (1.73) also *defines* a compound Poisson distribution; that is, if a distribution S has a pgf given by equation (1.73), where λ is a constant and $P(t)$ is a well-defined pgf, then S is a compound Poisson distribution. In particular, the secondary distribution of S has pgf $P(t)$.

We now introduce a recursive method for computing the pf of S, which applies to the case when the primary distribution N belongs to the $(a, b, 0)$ class. This method is called the Panjer (1981) recursion.

Theorem 1.5 *If N belongs to the $(a, b, 0)$ class of distributions and X is a nonnegative integer-valued random variable, then the pf of S is given by the following recursion*

$$f_S(s) = \frac{1}{1 - af_X(0)} \sum_{x=1}^{s} \left(a + \frac{bx}{s}\right) f_X(x) f_S(s - x), \qquad \text{for } s = 1, 2, \cdots, \tag{1.74}$$

with initial value $f_S(0)$ given by equation (1.70).

Proof See Dickson (2005, Section 4.5.2). □

The recursion formula in equation (1.74) applies to the $(a, b, 0)$ class only. Similar formulas, however, can be derived for other classes of primary distributions, such as the $(a, b, 1)$ class. Readers will find more details in Dickson (2005).

Example 1.9 Let $N \sim \mathcal{PN}(2)$ and $X \sim \mathcal{GM}(0.2)$. Calculate $f_S(0), f_S(1), f_S(2)$ and $f_S(3)$ using the Panjer recursion.

Solution From Example 1.8, we have $f_S(0) = \exp[\lambda(\theta - 1)] = \exp[(2)(0.2 - 1)] = 0.2019$. Evaluating the pf of the geometric distribution, we have $f_X(0) = 0.2$, $f_X(1) = (0.2)(1 - 0.2) = 0.16$, $f_X(2) = (0.2)(1 - 0.2)^2 = 0.128$, and $f_X(3) = (0.2)(1 - 0.2)^3 = 0.1024$. From Table 1.2, the parameters a and b of the Poisson distribution are: $a = 0$ and $b = \lambda$. Hence, from equation (1.74) we have

$$\begin{aligned} f_S(1) &= \lambda f_X(1) f_S(0) \\ &= (2)(0.16)(0.2019) = 0.0646. \end{aligned}$$

This agrees with the answer of $f_S(1) = [f_S(0)]\lambda\theta(1 - \theta) = 0.0646$ from Example 1.8. Similarly, we have

$$\begin{aligned} f_S(2) &= \frac{\lambda}{2} f_X(1) f_S(1) + \lambda f_X(2) f_S(0) \\ &= (0.16)(0.0646) + (2)(0.128)(0.2019) = 0.0620, \end{aligned}$$

and

$$f_S(3) = \frac{\lambda}{3} f_X(1) f_S(2) + \frac{2\lambda}{3} f_X(2) f_S(1) + \lambda f_X(3) f_S(0) = 0.0590.$$

□

Apart from the pf, it may be of interest to calculate the moments of the compound distribution. It turns out that the mean and variance of a compound distribution can be obtained from the means and variances of the primary and secondary distributions. Thus, the first two moments of the compound distribution can be obtained without computing its pf. The theorem below provides the results.

Theorem 1.6 *Consider the compound distribution defined in equation (1.60). We denote* $\mathrm{E}(N) = \mu_N$ *and* $\mathrm{Var}(N) = \sigma_N^2$, *and likewise* $\mathrm{E}(X) = \mu_X$ *and* $\mathrm{Var}(X) = \sigma_X^2$. *The mean and variance of S are then given by*

$$E(S) = \mu_N \mu_X, \tag{1.75}$$

and

$$Var(S) = \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2. \tag{1.76}$$

Proof We use the results in Appendix A.11 on conditional expectations to obtain[6]

$$\mathrm{E}(S) = \mathrm{E}[\mathrm{E}(S \mid N)] = \mathrm{E}[\mathrm{E}(X_1 + \cdots + X_N \mid N)] = \mathrm{E}(N\mu_X) = \mu_N \mu_X. \tag{1.77}$$

From (A.115), we have

$$\begin{aligned} \mathrm{Var}(S) &= \mathrm{E}[\mathrm{Var}(S \mid N)] + \mathrm{Var}[\mathrm{E}(S \mid N)] \\ &= \mathrm{E}[N\sigma_X^2] + \mathrm{Var}(N\mu_X) \\ &= \mu_N \sigma_X^2 + \sigma_N^2 \mu_X^2, \end{aligned} \tag{1.78}$$

which completes the proof. □

Note that if S is a compound Poisson distribution with $N \sim \mathcal{PN}(\lambda)$, so that $\mu_N = \sigma_N^2 = \lambda$, then

$$\mathrm{Var}(S) = \lambda(\sigma_X^2 + \mu_X^2) = \lambda\, \mathrm{E}(X^2). \tag{1.79}$$

[6] See Appendix A.11 for the interpretation of the expectation operators in equation (1.77).

Theorem 1.6 holds for any compound distribution, whether X is discrete or continuous. This result will be found useful in later chapters when X is the claim severity rather than the claim frequency.

Example 1.10 Let $N \sim \mathcal{PN}(2)$ and $X \sim \mathcal{GM}(0.2)$. Calculate $\mathrm{E}(S)$ and $\mathrm{Var}(S)$. Repeat the calculation for $N \sim \mathcal{GM}(0.2)$ and $X \sim \mathcal{PN}(2)$.

Solution As $X \sim \mathcal{GM}(0.2)$, we have

$$\mu_X = \frac{1-\theta}{\theta} = \frac{0.8}{0.2} = 4,$$

and

$$\sigma_X^2 = \frac{1-\theta}{\theta^2} = \frac{0.8}{(0.2)^2} = 20.$$

If $N \sim \mathcal{PN}(2)$, from equation (1.75) we have $\mathrm{E}(S) = (4)(2) = 8$. Since N is Poisson, we use equation (1.79) to obtain

$$\mathrm{Var}(S) = 2(20 + 4^2) = 72.$$

For $N \sim \mathcal{GM}(0.2)$ and $X \sim \mathcal{PN}(2)$, $\mu_N = 4$, $\sigma_N^2 = 20$, and $\mu_X = \sigma_X^2 = 2$. Thus, $\mathrm{E}(S) = (4)(2) = 8$, and from equation (1.76), we have

$$\mathrm{Var}(S) = (4)(2) + (20)(4) = 88. \qquad \square$$

In Theorem 1.1 we see that the sum of independently distributed Poisson distributions is also Poisson. It turns out that the sum of independently distributed *compound* Poisson distributions has also a *compound* Poisson distribution. Theorem 1.7 states this result.

Theorem 1.7 *Suppose $S_1, \cdots, S_n$ have independently distributed compound Poisson distributions, where the Poisson parameter of S_i is λ_i and the pgf of the secondary distribution of S_i is $P_i(\cdot)$. Then $S = S_1 + \cdots + S_n$ has a compound Poisson distribution with Poisson parameter $\lambda = \lambda_1 + \cdots + \lambda_n$. The pgf of the secondary distribution of S is $P(t) = \sum_{i=1}^{n} w_i P_i(t)$, where $w_i = \lambda_i/\lambda$.*

Proof The pgf of S is

$$\begin{aligned} P_S(t) &= \mathrm{E}\left(t^{S_1+\cdots+S_n}\right) \\ &= \prod_{i=1}^{n} P_{S_i}(t) \\ &= \prod_{i=1}^{n} \exp\{\lambda_i[P_i(t)-1]\} \\ &= \exp\left\{\sum_{i=1}^{n}\lambda_i P_i(t) - \sum_{i=1}^{n}\lambda_i\right\} \\ &= \exp\left\{\sum_{i=1}^{n}\lambda_i P_i(t) - \lambda\right\} \\ &= \exp\left\{\lambda\left[\sum_{i=1}^{n}\frac{\lambda_i}{\lambda}P_i(t) - 1\right]\right\} \\ &= \exp\{\lambda[P(t)-1]\}. \end{aligned} \tag{1.80}$$

Thus, by the uniqueness property of the pgf, S has a compound Poisson distribution with Poisson parameter λ, and the pgf of its secondary distribution is $P(t)$. Note that $P(t)$ is a well-defined pgf as it is a convex combination (with positive weights that sum to 1) of n well-defined pgf. □

Note that if $f_i(\cdot)$ is the pf of the secondary distribution of S_i, then $f(\cdot)$ defined by

$$f(x) = \sum_{i=1}^{n} w_i f_i(x), \qquad \text{for } x = 0, 1, \cdots, \tag{1.81}$$

is the pf of the secondary distribution of S. This can be seen by comparing the coefficients of t^x on both sides of the equation $P(t) = \sum_{i=1}^{n} w_i P_i(t)$. It is easy to see that if all the compound Poisson distributions S_i have the same secondary distribution, then this is also the secondary distribution of S.

Example 1.11 The claim frequency of a block of insurance policies follows a compound Poisson distribution with Poisson parameter 2 and a secondary distribution $\mathcal{BN}(4, 0.4)$. The claim frequency of another block of policies follows a compound Poisson distribution with Poisson parameter 4 and a secondary distribution $\mathcal{GM}(0.1)$. If these two blocks of policies are independent, what is the probability that their aggregate claim frequency is less than 3?

Solution Note that the aggregate claim frequency S of the two blocks of policies is the sum of two compound Poisson distributions. Thus, S is itself a compound Poisson distribution with $\lambda = 2 + 4 = 6$. We shall apply the Panjer recursion to compute the distribution of S. To do this, we first consider the pf of the secondary distributions of the blocks of policies. The relevant probabilities of the secondary distribution of the first block of policies are

$$f_1(0) = (0.6)^4 = 0.1296,$$
$$f_1(1) = 4(0.4)(0.6)^3 = 0.3456,$$
$$f_1(2) = 6(0.4)^2(0.6)^2 = 0.3456,$$

and the relevant probabilities of the secondary distribution of the second block of policies are

$$f_2(0) = 0.1, \quad f_2(1) = (0.1)(0.9) = 0.09, \quad f_2(2) = (0.1)(0.9)^2 = 0.081.$$

The relevant probabilities of the secondary distribution of S are

$$f(0) = \frac{1}{3}f_1(0) + \frac{2}{3}f_2(0) = 0.1099,$$
$$f(1) = \frac{1}{3}f_1(1) + \frac{2}{3}f_2(1) = 0.1752,$$

and

$$f(2) = \frac{1}{3}f_1(2) + \frac{2}{3}f_2(2) = 0.1692.$$

The pgf of the secondary distribution of S is

$$P(t) = \frac{1}{3}(0.4t + 0.6)^4 + \frac{2}{3}\left(\frac{0.1}{1 - 0.9t}\right),$$

from which we obtain

$$P(0) = \frac{1}{3}(0.6)^4 + \frac{2}{3}(0.1) = 0.1099.$$

This is the probability of zero of the secondary distribution of S, i.e. $f(0)$. Now we have

$$f_S(0) = \exp\{\lambda[P(0) - 1]\} = \exp[6(0.1099 - 1)] = 0.004793,$$

and using the Panjer recursion in equation (1.74), we obtain

$$f_S(1) = 6f(1)f_S(0) = 6(0.1752)(0.004793) = 0.005038,$$

and

$$f_S(2) = \frac{6}{2} f(1) f_S(1) + \frac{(6)(2)}{2} f(2) f_S(0) = 0.007514.$$

Thus, the probability of fewer than three aggregate claims is

$$0.004793 + 0.005038 + 0.007514 = 0.017345. \qquad \square$$

1.5.2 Mixture distribution

New distributions can also be created by mixing distributions. Let $X_1, \cdots, X_n$ be random variables with corresponding pf or pdf $f_{X_1}(\cdot), \cdots, f_{X_n}(\cdot)$ in the common support Ω. A new random variable X may be created with pf or pdf $f_X(\cdot)$ given by

$$f_X(x) = p_1 f_{X_1}(x) + \cdots + p_n f_{X_n}(x), \qquad x \in \Omega, \tag{1.82}$$

where $p_i \geq 0$ for $i = 1, \cdots, n$ and $\sum_{i=1}^{n} p_i = 1$. Thus, $\{p_i\}$ form a well-defined probability distribution and we may define a random variable Y by $\Pr(Y = i) = f_Y(i) = p_i$. Hence, X may be regarded as a random variable which is equal to X_i with probability p_i, and Y may be interpreted as a random variable which takes value i if and only if $X = X_i$.

We can check that $f_X(x)$ defined in equation (1.82) is a well-defined pf or pdf. As we are interested in claim-frequency distributions, we shall focus on cases in which $X_1, \cdots, X_n$ are nonnegative integer valued. Let the mean and variance of X_i be μ_i and σ_i^2, respectively. The following theorem gives the mean and variance of X.

Theorem 1.8 *The mean of X is*

$$\mathrm{E}(X) = \mu = \sum_{i=1}^{n} p_i \mu_i, \tag{1.83}$$

and its variance is

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p_i \left[(\mu_i - \mu)^2 + \sigma_i^2 \right]. \tag{1.84}$$

Proof We use the conditional expectation formulas in Appendix A.11. By equation (A.111), we have

$$\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X \mid Y)]. \tag{1.85}$$

We denote $\mathrm{E}(X \mid Y) = \mu(Y)$, where $\mu(Y) = \mu_i$ if $Y = i$. Thus

$$\begin{aligned}\mathrm{E}(X) &= \mathrm{E}[\mu(Y)] \\ &= \sum_{i=1}^{n} p_i \mu_i. \end{aligned} \tag{1.86}$$

Note that this result can also be derived using equation (1.82) as follows

$$\begin{aligned}\mathrm{E}(X) &= \sum_{x=0}^{\infty} x f_X(x) \\ &= \sum_{x=0}^{\infty} x \left[\sum_{i=1}^{n} p_i f_{X_i}(x) \right] \\ &= \sum_{i=1}^{n} p_i \left[\sum_{x=0}^{\infty} x f_{X_i}(x) \right] \\ &= \sum_{i=1}^{n} p_i \mu_i. \end{aligned} \tag{1.87}$$

We now denote $\mathrm{Var}(X \mid Y) = \sigma^2(Y)$, where $\sigma^2(Y) = \sigma_i^2$ if $Y = i$. To compute the variance of X, we use equation (A.115) to obtain

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}[\mathrm{E}(X \mid Y)] \\ &= \mathrm{E}[\sigma^2(Y)] + \mathrm{Var}[\mu(Y)] \\ &= \sum_{i=1}^{n} p_i \sigma_i^2 + \sum_{i=1}^{n} p_i (\mu_i - \mu)^2 \\ &= \sum_{i=1}^{n} p_i \left[(\mu_i - \mu)^2 + \sigma_i^2 \right]. \end{aligned} \tag{1.88}$$

□

A random variable X with pf or pdf given by equation (1.82) has a **mixture distribution**. The random variable Y which determines the probability of occurrence of X_i is called the **mixing distribution**. As Y is discrete, X is called a **discrete mixture**, notwithstanding the fact that it has a continuous distribution if X_i are continuous. Note that X is different from the random variable $X^* = p_1 X_1 + \cdots + p_n X_n$. The exact distribution of X^* is in general difficult to compute. Assuming $\{X_i\}$ are independent,[7] to calculate the distribution of X^* we need to

[7] Note that this assumption is not relevant for the definition of X with pf or pdf defined in equation (1.82), for which only the *marginal* distributions of X_i are relevant.

use the convolution method. Also, while the mean of X^* is the same as that of X its variance is $\sum_{i=1}^{n} p_i^2 \sigma_i^2$, which is smaller than that of X.

Example 1.12 The claim frequency of a bad driver is distributed as $\mathcal{PN}(4)$, and the claim frequency of a good driver is distributed as $\mathcal{PN}(1)$. A town consists of 20% bad drivers and 80% good drivers. What are the mean and variance of the claim frequency of a randomly selected driver from the town?

Solution The mean of the claim frequency is

$$(0.2)(4) + (0.8)(1) = 1.6,$$

and its variance is

$$(0.2)\left[(4 - 1.6)^2 + 4\right] + (0.8)\left[(1 - 1.6)^2 + 1\right] = 3.04. \qquad \square$$

We have so far considered discrete mixing distributions. Continuous distributions, however, can also be used as the mixing distribution. For example, consider a Poisson distribution $\mathcal{PN}(\lambda)$. Let $h(\cdot)$ be a function such that $h(\lambda) > 0$ for $\lambda > 0$, and

$$\int_0^\infty h(\lambda)\, d\lambda = 1. \tag{1.89}$$

In other words, $h(\lambda)$ is a properly defined pdf with the Poisson parameter λ treated as the realization of a *random variable*. A new mixture distribution can be created by defining a random variable X with pf given by

$$f_X(x) = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}\, h(\lambda)\, d\lambda. \tag{1.90}$$

It can be checked that $f_X(\cdot)$ is a well-defined pf. Specifically, $f_X(x) > 0$ for $x = 0, 1, \cdots$, and

$$\begin{aligned}\sum_{x=0}^{\infty} f_X(x) &= \sum_{x=0}^{\infty} \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}\, h(\lambda)\, d\lambda = \int_0^\infty \left(\sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}\right) h(\lambda)\, d\lambda \\ &= \int_0^\infty h(\lambda)\, d\lambda = 1. \end{aligned} \tag{1.91}$$

We now replace the Poisson pf in equation (1.90) by an arbitrary pf $f(x \mid \theta)$, where θ is the parameter of the distribution. With $h(\theta)$ satisfying the conditions of a pdf (now treating θ as a random variable), a general formula for a mixture

distribution for which the mixing distribution is *continuous* can be obtained. The pf of the mixture distribution is

$$f_X(x) = \int_0^\infty f(x \mid \theta) h(\theta)\, d\theta. \tag{1.92}$$

Note that as $f(x \mid \theta)$ is a pf, $f_X(x)$ is also a pf and X is discrete. On the other hand, equation (1.92) can also be applied to a pdf $f(x \mid \theta)$, in which case $f_X(x)$ is a pdf and X is continuous. With this extension, equation (1.92) defines a **continuous mixture**, notwithstanding the fact that X is discrete if $f(x \mid \theta)$ is a pf. We will see some examples of continuous mixing in later chapters.

1.6 Excel computation notes

Excel provides some functions to compute the pf and df of the discrete distributions discussed in this chapter. Table 1.5 summarizes the use of these functions. Some of these functions can be used to compute the pf as well as the df, and their use is determined by the input variable `ind`, which is set to `FALSE` to compute the pf and `TRUE` to compute the df. While the pf of these distributions are generally straightforward to compute, the Excel functions facilitate the calculation of the df for which multiple values of the pf are required.

As the geometric distribution is a special case of the negative binomial distribution with $r = 1$, `NEGBINOMDIST` can be used to compute the pf of the geometric distribution by setting `x2` equal to 1. An example is shown in the table.

1.7 Summary and conclusions

We have discussed some standard nonnegative integer-valued random variables that may be applied to model the claim-frequency distributions. Properties of the binomial, geometric, negative binomial, and Poisson distributions are discussed in detail. These distributions belong to the class of $(a, b, 0)$ distributions, which play an important role in the actuarial science literature. Further methods of creating new distributions with the flexibility of being able to fit various shapes of empirical data are introduced, including the compound-distribution and mixture-distribution methods. The computation of the compound distribution requires the convolution technique in general, which may be very tedious. When the primary distribution of the compound distribution belongs to the $(a, b, 0)$ class, the Panjer recursion may be used to facilitate the computation.

Claim events are dependent on the policy terms. Policy modifications such as deductibles may impact the claim events and thus their distributions. The

Table 1.5. *Some Excel functions for the computation of the pf $f_X(x)$ and df $F_X(x)$ of discrete random variable X*

		Example	
X	Excel function	input	output
$\mathcal{BN}(n,\theta)$	`BINOMDIST(x1,x2,x3,ind)`	`BINOMDIST(4,10,0.3,FALSE)`	0.2001
	`x1` $= x$	`BINOMDIST(4,10,0.3,TRUE)`	0.8497
	`x2` $= n$		
	`x3` $= \theta$		
$\mathcal{PN}(\lambda)$	`POISSON(x1,x2,ind)`	`POISSON(4,3.6,FALSE)`	0.1912
	`x1` $= x$	`POISSON(4,3.6,TRUE)`	0.7064
	`x2` $= \lambda$		
$\mathcal{NB}(r,\theta)$	`NEGBINOMDIST(x1,x2,x3)`	`NEGBINOMDIST(3,1,0.4)`	0.0864
	`x1` $= x$	`NEGBINOMDIST(3,3,0.4)`	0.1382
	`x2` $= r$		
	`x3` $= \theta$		

Note: Set `ind` to `FALSE` for pf and `TRUE` for df.

techniques discussed in this chapter may be applied to any policies without reference to whether there are policy modifications. In later chapters we shall examine the relationship between claim-frequency distributions with and without claim modifications.

Exercises

1.1 Let $X \sim \mathcal{BN}(n,\theta)$, prove that $\mathrm{E}(X) = n\theta$ and $\mathrm{E}[X(X-1)] = n(n-1)\theta^2$. Hence, show that $\mathrm{Var}(X) = n\theta(1-\theta)$. Derive the mgf of X, $M_X(t)$.

1.2 Let $X \sim \mathcal{GM}(\theta)$, prove that $\mathrm{E}(X) = (1-\theta)/\theta$ and $\mathrm{E}[X(X-1)] = 2(1-\theta)^2/\theta^2$. Hence, show that $\mathrm{Var}(X) = (1-\theta)/\theta^2$. Derive the mgf of X, $M_X(t)$.

1.3 Let $X \sim \mathcal{PN}(\lambda)$, prove that $\mathrm{E}(X) = \lambda$ and $\mathrm{E}[X(X-1)] = \lambda^2$. Hence, show that $\mathrm{Var}(X) = \lambda$. Derive the mgf of X, $M_X(t)$.

1.4 Let X^* be the zero-truncated distribution of X, which has pgf $P_X(t)$ and pf $f_X(x)$. Derive an expression for the pgf of X^* in terms of $P_X(\cdot)$ and $f_X(\cdot)$.

1.5 Let $S = X_1 + X_2 + X_3$, where X_i are iid $\mathcal{BN}(2, 0.4)$ for $i = 1, 2$, and 3. Compute the pf of S.

1.6 What are the supports of the following compound distributions?
- (a) Primary distribution: $N \sim \mathcal{NB}(r,\theta)$, secondary distribution: $X \sim \mathcal{PN}(\lambda)$.
- (b) Primary distribution: $N \sim \mathcal{NB}(r,\theta)$, secondary distribution: $X \sim \mathcal{BN}(n,\theta)$.
- (c) Primary distribution: $N \sim \mathcal{BN}(n,\theta)$, secondary distribution: $X \sim \mathcal{GM}(\theta)$.
- (d) Primary distribution: $N \sim \mathcal{BN}(n,\theta)$, secondary distribution: $X \sim \mathcal{BN}(m,\theta)$.

1.7 S_1 and S_2 are independent compound Poisson distributions. The Poisson parameter of S_1 is 3 and its secondary distribution is $\mathcal{GM}(0.6)$. The Poisson parameter of S_2 is 4 and its secondary distribution is Bernoulli with probability of success 0.2. If $S = S_1 + S_2$, what is $\Pr(S < 4)$?

1.8 Suppose $X \sim \mathcal{GM}(0.1)$, calculate the probability that the zero-modified distribution of X with $f_X^M(0) = 0.3$ is less than 4.

1.9 Let $X_1, \cdots, X_n$ be nonnegative integer-valued random variables with identical pf $f_X(\cdot)$. A discrete mixture distribution W is created with pf $f_W(x) = p_1 f_{X_1}(x) + \cdots + p_n f_{X_n}(x)$, where $p_i \geq 0$ for $i = 1, \cdots, n$ and $\sum_{i=1}^n p_i = 1$. Another random variable Y is defined by $Y = p_1 X_1 + \cdots + p_n X_n$.

(a) Compare the mean of W and Y.
(b) If $X_1, \cdots, X_n$ are independent, compare the variance of W and Y.

1.10 S has a compound Poisson distribution, for which the Poisson parameter of the primary distribution is λ_1. Suppose the secondary distribution of S is also Poisson, with parameter λ_2.
(a) Find the mgf of S.
(b) Find the mean and the variance of S.
(c) Find the probabilities of $S = 0$ and $S = 1$.

1.11 S has a compound Poisson distribution with Poisson parameter λ. The secondary distribution X of S follows a logarithmic distribution with parameter β, with pf given by

$$f_X(x) = \frac{\beta^x}{x(1+\beta)^x \log(1+\beta)}, \qquad \beta > 0,\ x = 1, 2, \cdots.$$

(a) Derive the pgf of X.
(b) Derive the pgf of S and show that S has a negative binomial distribution. What are the parameters of the negative binomial distribution?

1.12 Construct the zero-modified distribution X from the Poisson distribution with $\lambda = 2.5$ so that the probability of zero is 0.55. What is the probability of $X \geq 4$? Find the mean and the variance of X.

1.13 Let $X_1 \sim \mathcal{PN}(2)$, $X_2 \sim \mathcal{PN}(3)$, and $S = X_1 + 2X_2$, where X_1 and X_2 are independent. Calculate $\Pr(S = s)$, for $s = 0$, 1, and 2.

1.14 S_1 has a compound distribution with primary distribution $\mathcal{PN}(1)$ and secondary distribution $\mathcal{GM}(\theta_1)$. Likewise, S_2 has a compound distribution with primary distribution $\mathcal{PN}(2)$ and secondary distribution $\mathcal{GM}(\theta_2)$, and S_1 and S_2 are independent. Let $S = S_1 + S_2$. Calculate $\Pr(S = s)$, for $s = 0$, 1, and 2, if
(a) $\theta_1 = \theta_2 = 0.2$,
(b) $\theta_1 = 0.2$ and $\theta_2 = 0.4$.

1.15 Let $N_i \sim \mathcal{PN}(\lambda_i)$, $i = 1, 2, \cdots, n$, be independently distributed. Define a random variable S by

$$S = x_1 N_1 + x_2 N_2 + \cdots + x_n N_n,$$

where x_i are n different positive numbers.
(a) Derive the pgf of S.
(b) Calculate $\Pr(S = 0)$.

1.16 S_1 has a compound distribution with primary distribution $\mathcal{PN}(2)$ and secondary distribution $\mathcal{NB}(4, 0.5)$. S_2 has a compound distribution

with primary distribution $\mathcal{NB}(4, 0.5)$ and secondary distribution $\mathcal{PN}(2)$. Calculate the mean and the variance of S_1 and S_2.

1.17 X is a mixture of $\mathcal{PN}(\lambda)$ distributions, where $\lambda - 1$ has a $\mathcal{BN}(2, 0.2)$ distribution. Calculate the pgf, the mean and the variance of X.

1.18 The number of trips a courier needs to make into the business district each day is distributed as $\mathcal{BN}(2, 0.7)$. The number of stops he has to make in front of traffic lights in each trip follows a $\mathcal{PN}(4)$ distribution. What is the probability that the courier will not make any stop in front of traffic lights on a working day?

1.19 The number of sick leaves for male workers is on average three times that of female workers. The number of sick leaves of all workers in a day is Poisson distributed with mean 4.5.

(a) Find the probability of the event of having no sick leave in a day.

(b) Find the probability of the event of having one male sick leave and one female sick leave in a day.

1.20 Let $X_1 \sim \mathcal{PN}(1)$ and $X_2 \sim \mathcal{PN}(2)$, where X_1 and X_2 are independent. Calculate $\Pr(X_1 = x)$ and $\Pr(X_2 = x)$ for $x = 0, 1$, and 2. Hence find $\Pr(X_1 + X_2 \leq 2)$. Can you suggest an alternative method to compute $\Pr(X_1 + X_2 \leq 2)$ without calculating the probabilities of X_1 and X_2?

1.21 Suppose the nonnegative integer-valued random variable X has pf $f_X(x)$ for $x = 0, 1, \cdots$, and pgf $P_X(t)$. A zero-modified distribution X^* of X has probability at zero of $f_X^M(0)$, where $f_X^M(0) > f_X(0)$. Prove that the pgf of X^*, denoted by $P_{X^*}(t)$, is given by

$$P_{X^*}(t) = 1 - c + cP_X(t),$$

where

$$c = \frac{1 - f_X^M(0)}{1 - f_X(0)}.$$

By recognizing that the pgf of a Bernoulli distribution with probability of success θ is $P(t) = 1 - \theta + \theta t$, show that the above results can be interpreted as saying that any zero-modified distribution is a compound distribution. What are the primary and secondary distributions of this compound distribution? Also, how would you interpret X^* as a mixture distribution?

1.22 Show that any geometric-geometric compound distribution can be interpreted as a Bernoulli–geometric compound distribution, and vice-versa.

1.23 Show that any binomial–geometric compound distribution can be interpreted as a negative binomial–geometric compound distribution, and vice-versa.

1.24 A diversified portfolio of bonds consists of investment-grade and noninvestment-grade bonds. The number of defaults of investment-grade bonds in each month is Poisson distributed with parameter 0.2, and the number of defaults of noninvestment-grade bonds in each month is independently Poisson distributed with parameter 1. When there is a default, whether it is an investment-grade or noninvestment-grade bond, the loss is \$1 million or \$2 million with equal probability. Derive the recursive formula for calculating the distribution of the losses of the portfolio in a month.

1.25 Let S be a compound distribution, where the primary distribution is $\mathcal{NB}(r,\theta)$ and the secondary distribution is $\mathcal{PN}(\lambda)$. Suppose S^* is a mixture of distributions, with pf

$$f_{S^*}(x) = \sum_{i=0}^{n} f_{Y_i}(x)\, p_i,$$

where $p_i \geq 0$, $\sum_{i=0}^{n} p_i = 1$ (n may be infinite) and $Y_i \sim \mathcal{PN}(\lambda i)$ (note that $Y_0 = 0$ with probability 1). Let X be a random variable such that $\Pr(X = i) = p_i$, with mgf $M_X(\cdot)$.

(a) What is the pgf of S, $P_S(t)$?

(b) Show that the pgf of S^*, $P_{S^*}(t)$, is given by

$$P_{S^*}(t) = \sum_{i=0}^{n} e^{(t-1)\lambda i} p_i = \mathrm{E}[e^{(t-1)\lambda X}] = M_X[\lambda(t-1)].$$

(c) If p_i are such that $X \sim \mathcal{NB}(r,\theta)$, show that $P_{S^*}(t) = P_S(t)$. How would you interpret this result?

1.26 X is distributed as $\mathcal{GM}(0.8)$. What is the recursion formula for the pf of X? Derive the recursion formula for the zero-modified distribution of X with $f_X^M(0) = 0.4$. Calculate the mean and the variance of the zero-modified distribution.

1.27 S has a binomial–Poisson compound distribution. What is the pgf of the zero-truncated distribution of S?

1.28 S has a compound distribution with primary distribution N and secondary distribution X. If $N \sim \mathcal{GM}(0.5)$ and $X \sim \mathcal{PN}(3)$, calculate $f_S(s)$ for $s = 0, 1$, and 2 using Panjer recursion.

1.29 Business failures are due to three mutually exclusive risks: market risk, credit risk, and operation risk, which account for 20%, 30%, and 50%, respectively, of all business failures. Suppose the number of business failures each year is Poisson distributed with mean 4.6.

(a) What is the chance that there are two business failures due to operation risk in a year?

(b) What is the chance that the business failures due to market risk and credit risk are both fewer than two in a year?

(c) Given that there are four business failures in a year, what is the probability that two of these are due to market risk?

1.30 What are the mgf and pgf of the following distributions?

(a) Geometric–binomial compound distribution, $\mathcal{GM}(\theta_N)-\mathcal{BN}(n,\theta_X)$.

(b) Binomial–Poisson compound distribution, $\mathcal{BN}(n,\theta)-\mathcal{PN}(\lambda)$.

(c) Negative binomial–Poisson compound distribution, $\mathcal{NB}(r,\theta)-\mathcal{PN}(\lambda)$.

2

Claim-severity distribution

Claim severity refers to the monetary loss of an insurance claim. Unlike claim frequency, which is a nonnegative integer-valued random variable, claim severity is usually modeled as a nonnegative continuous random variable. Depending on the definition of loss, however, it may also be modeled as a mixed distribution, i.e. a random variable consisting of probability masses at some points and continuous otherwise.

We begin this chapter with a brief review of some statistical tools for analyzing continuous distributions and mixed distributions. The use of the survival function and techniques of computing the distribution of a transformed random variable are reviewed. Some standard continuous distributions for modeling claim severity are summarized. These include the exponential, gamma, Weibull, and Pareto distributions. We discuss methods for creating new claim-severity distributions such as the mixture-distribution method. As losses that are in the extreme right-hand tail of the distribution represent big losses, we examine the right-hand tail properties of the claim-severity distributions. In particular, measures of tail weight such as limiting ratio and conditional tail expectation are discussed. When insurance loss payments are subject to coverage modifications such as deductibles, policy limits, and coinsurance, we examine their effects on the distribution of the claim severity.

Learning objectives

1 Continuous distributions for modeling claim severity
2 Mixed distributions
3 Exponential, gamma, Weibull, and Pareto distributions
4 Mixture distributions
5 Tail weight, limiting ratio, and conditional tail expectation
6 Coverage modification and claim-severity distribution

2.1 Review of statistics

In this section we review some results in statistical distributions relevant for analyzing claim severity. These include the survival function, the hazard function and methods for deriving the distribution of a transformed random variable.

2.1.1 Survival function and hazard function

Let X be a continuous random variable with df $F_X(x)$ and pdf $f_X(x)$. The **survival function (sf)** of X, denoted by $S_X(x)$, is the complement of the df,[1] i.e.

$$S_X(x) = 1 - F_X(x) = \Pr(X > x). \tag{2.1}$$

The pdf can be obtained from the sf through the equation

$$f_X(x) = \frac{dF_X(x)}{dx} = -\frac{dS_X(x)}{dx}. \tag{2.2}$$

While the df $F_X(x)$ is monotonic nondecreasing, the sf $S_X(x)$ is monotonic nonincreasing. Also, we have $F_X(-\infty) = S_X(\infty) = 0$ and $F_X(\infty) = S_X(-\infty) = 1$. If X is nonnegative, then $F_X(0) = 0$ and $S_X(0) = 1$.

The **hazard function (hf)** of a nonnegative random variable X, denoted by $h_X(x)$, is defined as[2]

$$h_X(x) = \frac{f_X(x)}{S_X(x)}. \tag{2.3}$$

If we multiply both sides of the above equation by dx, we obtain

$$\begin{aligned} h_X(x)\,dx &= \frac{f_X(x)\,dx}{S_X(x)} \\ &= \frac{\Pr(x \le X < x + dx)}{\Pr(X > x)} \\ &= \frac{\Pr(x \le X < x + dx \text{ and } X > x)}{\Pr(X > x)} \\ &= \Pr(x < X < x + dx \,|\, X > x), \end{aligned} \tag{2.4}$$

where we have made use of the results in Appendix A.2 to obtain the second line of the equation. Thus, $h_X(x)\,dx$ can be interpreted as the conditional probability

[1] The survival function is also called the **decumulative distribution function**. In this book we use the term survival function.

[2] The hazard function is also called the *hazard rate* or the *failure rate*. In the survival analysis literature, it is called the *force of mortality*.

of X taking value in the infinitesimal interval $(x, x+dx)$ given $X > x$. In the life contingency and survival analysis literature, the hf of the age-at-death random variable is a very important tool for analyzing mortality and life expectancy. We shall see that it is an important determinant of the tail behavior of claim-severity distributions.

Given the pdf or sf, we can compute the hf using equation (2.3). The reverse can be obtained by noting that equation (2.3) can be written as

$$h_X(x) = -\frac{1}{S_X(x)}\left(\frac{dS_X(x)}{dx}\right) = -\frac{d\log S_X(x)}{dx}, \tag{2.5}$$

so that

$$h_X(x)\,dx = -d\log S_X(x). \tag{2.6}$$

Integrating both sides of the equation, we obtain

$$\int_0^x h_X(s)\,ds = -\int_0^x d\log S_X(s) = -\log S_X(s)]_0^x = -\log S_X(x), \tag{2.7}$$

as $\log S_X(0) = \log(1) = 0$. Thus, we have

$$S_X(x) = \exp\left(-\int_0^x h_X(s)\,ds\right), \tag{2.8}$$

from which the sf can be calculated given the hf. The integral in the above equation is called the **cumulative hazard function**.

The sf is defined for both continuous and discrete random variables. If X is a discrete random variable, the last expression on the right-hand side of equation (2.2) is replaced by the difference of the sf, i.e. $f_X(x_i) = S_X(x_{i-1}) - S_X(x_i)$, where x_i are values in increasing order in the support of X. We will, however, use the hf only for continuous random variables.

Example 2.1 Let X be a uniformly distributed random variable in the interval [0, 100], denoted by $\mathcal{U}(0, 100)$.[3] Compute the pdf, df, sf, and hf of X.

Solution The pdf, df, and sf of X are, for $x \in [0, 100]$

$$f_X(x) = 0.01,$$

$$F_X(x) = 0.01x,$$

and

$$S_X(x) = 1 - 0.01x.$$

[3] See Appendix A.10.3 for a brief discussion of uniform distribution.

From equation (2.3), we obtain the hf as

$$h_X(x) = \frac{f_X(x)}{S_X(x)} = \frac{0.01}{1 - 0.01x},$$

which increases with x. In particular, $h_X(x) \to \infty$ as $x \to 100$. □

2.1.2 Mixed distribution

Some random variables may have a mixture of discrete and continuous parts. A random variable X is said to be of the **mixed type** if its df $F_X(x)$ is continuous and differentiable except for some values of x belonging to a countable set Ω_X.[4] Thus, if X has a mixed distribution, there exists a function $f_X(x)$ such that[5]

$$F_X(x) = \Pr(X \le x) = \int_{-\infty}^{x} f_X(x)\,dx + \sum_{x_i \in \Omega_X,\, x_i \le x} \Pr(X = x_i). \tag{2.9}$$

The functions $f_X(x)$ and $\Pr(X = x_i)$ together describe the density and mass function of the mixed random variable X. Whether a random variable X is of the continuous, discrete, or mixed type, we may use the convenient **Stieltjes integral** to state that, for any constants a and b[6]

$$\Pr(a \le X \le b) = \int_a^b dF_X(x), \tag{2.10}$$

which is equal to

$$\int_a^b f_X(x)\,dx, \qquad \text{if } X \text{ is continuous,} \tag{2.11}$$

$$\sum_{x_i \in \Omega_X,\, a \le x_i \le b} \Pr(X = x_i), \qquad \text{if } X \text{ is discrete with support } \Omega_X, \tag{2.12}$$

and

$$\int_a^b f_X(x)\,dx + \sum_{x_i \in \Omega_X,\, a \le x_i \le b} \Pr(X = x_i), \qquad \text{if } X \text{ is mixed.} \tag{2.13}$$

[4] We use the term *mixed* random variable to denote one consisting of discrete and continuous parts. This should be distinguished from the *mixture* distributions described in Sections 1.5.2 and 2.3.2.

[5] Note that $f_X(x)$ is the derivative of $F_X(x)$ at the points where $F_X(x)$ is continuous and differentiable, but it is not the pdf of X. In particular, $\int_{-\infty}^{\infty} f_X(x)\,dx \ne 1$.

[6] For the definition of Stieltjes integral, see Ross (2006, p. 404).

2.1.3 Expected value of function of random variable

Consider a function $g(\cdot)$. The expected value of $g(X)$, denoted by $\mathrm{E}[g(X)]$, is defined as

$$\begin{aligned}\mathrm{E}[g(X)] &= \int_{-\infty}^{\infty} g(x)\, dF_X(x) \\ &= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx + \sum_{x_i \in \Omega_X} g(x_i)\ \Pr(X = x_i),\end{aligned} \tag{2.14}$$

for a general mixed distribution X. If X is continuous, we have

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx, \tag{2.15}$$

and when it is discrete we have

$$\mathrm{E}[g(X)] = \sum_{x_i \in \Omega_X} g(x_i) f_X(x_i). \tag{2.16}$$

If X is continuous and nonnegative, and $g(\cdot)$ is a nonnegative, monotonic, and differentiable function, the following result holds[7]

$$\mathrm{E}[g(X)] = \int_0^{\infty} g(x)\, dF_X(x) = g(0) + \int_0^{\infty} g'(x)[1 - F_X(x)]\, dx, \tag{2.17}$$

where $g'(x)$ is the derivative of $g(x)$ with respect to x. Defining $g(x) = x$, so that $g(0) = 0$ and $g'(x) = 1$, the mean of X can be evaluated by

$$\mathrm{E}(X) = \int_0^{\infty} [1 - F_X(x)]\, dx = \int_0^{\infty} S_X(x)\, dx. \tag{2.18}$$

Example 2.2 Let $X \sim \mathcal{U}(0, 100)$. Calculate the mean of X using equation (2.15) and verify equation (2.18). Define a random variable Y as follows

$$Y = \begin{cases} 0, & \text{for } X \le 20, \\ X - 20, & \text{for } X > 20. \end{cases}$$

Determine the df of Y, and its density and mass function.

[7] See Appendix A.3 for a proof.

Solution From equation (2.15) and the results in Example 2.1, the mean of X is given by

$$\begin{aligned}\mathrm{E}(X) &= \int_0^{100} x f_X(x)\,dx = \int_0^{100} 0.01x\,dx \\ &= 0.01\left(\frac{x^2}{2}\Big]_0^{100}\right) = 0.01\left(\frac{100^2}{2}\right) = 50.\end{aligned}$$

Applying equation (2.18), we have

$$\mathrm{E}(X) = \int_0^{100} S_X(x)\,dx = \int_0^{100} (1 - 0.01x)\,dx = 100 - \int_0^{100} 0.01x\,dx = 50.$$

Thus, the result is verified.

To determine the distribution of Y, we note that

$$\Pr(Y = 0) = \Pr(X \le 20) = F_X(20) = 0.2.$$

For $0 < y \le 80$, we have

$$\begin{aligned}\Pr(Y \le y) &= \Pr(Y = 0) + \Pr(0 < Y \le y) \\ &= 0.2 + \Pr(20 < X \le y + 20) \\ &= 0.2 + 0.01y.\end{aligned}$$

Thus, the df of Y is

$$F_Y(y) = \begin{cases} 0, & \text{for } y < 0, \\ 0.2, & \text{for } y = 0, \\ 0.2 + 0.01y, & \text{for } 0 < y \le 80, \\ 1, & \text{for } y > 80. \end{cases}$$

Hence, Y has a probability mass of 0.2 at point 0, and has a density function of 0.01 in the interval (0, 80] and zero otherwise. See Figure 2.1 for the graphical illustration of the distribution functions of X and Y. □

2.1.4 Distribution of function of random variable

Let $g(\cdot)$ be a continuous and differentiable function, and X be a continuous random variable with pdf $f_X(x)$. We define $Y = g(X)$, which is also a random variable. Suppose $y = g(x)$ is a one-to-one transformation, i.e. for any value of y, there is a unique value x such that $y = g(x)$. We denote the value of x

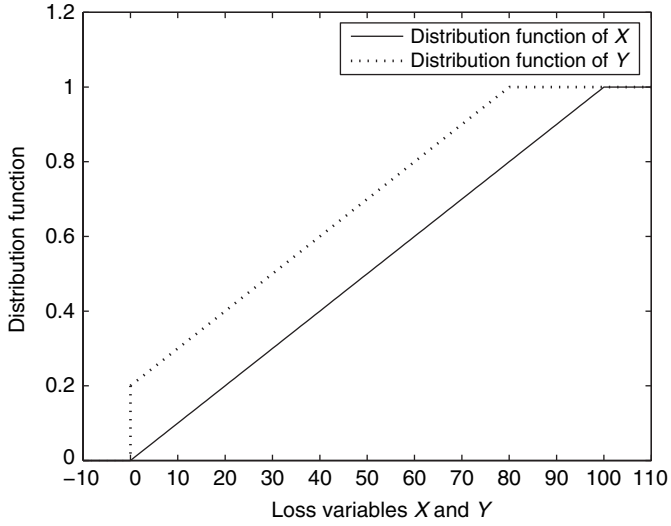

Figure 2.1 Distribution functions of X and Y in Example 2.2

corresponding to y by $g^{-1}(y)$, where $g^{-1}(\cdot)$ is called the inverse transformation. The theorem below gives the pdf of Y.

Theorem 2.1 *Let X be a continuous random variable taking values in $[a, b]$ with pdf $f_X(x)$, and let $g(\cdot)$ be a continuous and differentiable one-to-one transformation. Denote $a' = g(a)$ and $b' = g(b)$. The pdf of $Y = g(X)$ is*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \dfrac{dg^{-1}(y)}{dy} \right|, & \text{for } y \in [a', b'], \\ 0, & \text{otherwise.} \end{cases} \tag{2.19}$$

Proof See DeGroot and Schervish (2002, pp. 160–161). □

The restriction of one-to-one transformation may be relaxed by partitioning the domain of the transformation $g(\cdot)$ can be partitioned as the union of k mutually disjoint sets. Corresponding to each set there exists a function $g_i(\cdot)$ so that for a given value of $y \in [a', b']$, there is a unique x in the ith set with the property $y = g_i(x)$, for $i = 1, \cdots, k$. Then the pdf of Y is given by

$$f_Y(y) = \begin{cases} \displaystyle\sum_{i=1}^{k} f_X(g_i^{-1}(y)) \left| \frac{dg_i^{-1}(y)}{dy} \right|, & \text{for } y \in [a', b'], \\ 0, & \text{otherwise.} \end{cases} \tag{2.20}$$

The example below illustrates this result.

Example 2.3 Suppose $X \sim \mathcal{U}(-1, 1)$. Determine the pdf and the df of $Y = X^2$.

Solution The pdf of X is

$$f_X(x) = \frac{1}{2}, \qquad \text{for } -1 \le x \le 1 \text{ and 0 otherwise.}$$

Note that $Y = X^2 = g(X)$ is not a one-to-one transformation. We partition the domain of X into $A_1 = [-1, 0)$ and $A_2 = [0, 1]$, and define the functions $g_i(x) = x^2$, for $x \in A_i$, $i = 1, 2$. Then $g_i(x)$ are one-to-one transformations with respect to their domains, and we have

$$g_1^{-1}(y) = -\sqrt{y}, \qquad \text{for } y \in (0, 1],$$

and

$$g_2^{-1}(y) = \sqrt{y}, \qquad \text{for } y \in [0, 1].$$

Now

$$f_X(g_i^{-1}(y)) = \frac{1}{2}, \qquad \text{for } i = 1, 2,$$

and

$$\frac{dg_1^{-1}(y)}{dy} = -\frac{1}{2\sqrt{y}} \qquad \text{and} \qquad \frac{dg_2^{-1}(y)}{dy} = \frac{1}{2\sqrt{y}}, \qquad \text{for } y \in (0, 1].$$

Thus, using equation (2.20), we have

$$f_Y(y) = \frac{1}{2}\left|-\frac{1}{2\sqrt{y}}\right| + \frac{1}{2}\left|\frac{1}{2\sqrt{y}}\right| = \frac{1}{2\sqrt{y}}, \qquad \text{for } y \in (0, 1].$$

The df of Y is

$$F_Y(y) = \int_0^y f_Y(s)\, ds = \frac{1}{2}\int_0^y \frac{1}{\sqrt{s}}\, ds = \sqrt{s}\Big]_0^y = \sqrt{y}.$$

Note that the df can also be derived directly as follows

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) = \Pr(X^2 \le y) = \Pr(-\sqrt{y} \le X \le \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\, dx = \sqrt{y}. \end{aligned}$$

□

2.2 Some continuous distributions for claim severity

In this section we review some key results of four continuous random variables, namely exponential, gamma, Weibull, and Pareto. These random variables can only take nonnegative values, and may be used for distributions of claim severity. The choice of a particular distribution in practice is an empirical question to be discussed later.

2.2.1 Exponential distribution

A random variable X has an exponential distribution with parameter λ, denoted by $\mathcal{E}(\lambda)$, if its pdf is

$$f_X(x) = \lambda e^{-\lambda x}, \qquad \text{for } x \geq 0, \tag{2.21}$$

where $\lambda > 0$. The df and sf of X are

$$F_X(x) = 1 - e^{-\lambda x}, \tag{2.22}$$

and

$$S_X(x) = e^{-\lambda x}. \tag{2.23}$$

Thus, the hf of X is

$$h_X(x) = \frac{f_X(x)}{S_X(x)} = \lambda, \tag{2.24}$$

which is a constant, irrespective of the value of x. The mean and variance of X are

$$\mathrm{E}(X) = \frac{1}{\lambda} \quad \text{and} \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}. \tag{2.25}$$

The mgf of X is

$$M_X(t) = \frac{\lambda}{\lambda - t}. \tag{2.26}$$

The exponential distribution is often used to describe the inter-arrival time of an event, such as the breakdown of a machine. It is related to the Poisson distribution. If the inter-arrival time of an event is distributed as an exponential random variable with parameter λ, which is the reciprocal of the expected waiting time for the event (see equation (2.25)), then the number of occurrences of the event in a unit time interval is distributed as a Poisson with parameter λ.

2.2.2 Gamma distribution

X is said to have a gamma distribution with parameters α and β ($\alpha > 0$ and $\beta > 0$), denoted by $\mathcal{G}(\alpha, \beta)$, if its pdf is

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad \text{for } x \geq 0. \tag{2.27}$$

The function $\Gamma(\alpha)$ is called the gamma function, defined by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy, \tag{2.28}$$

which exists (i.e. the integral converges) for $\alpha > 0$. For $\alpha > 1$, $\Gamma(\alpha)$ satisfies the following recursion

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1). \tag{2.29}$$

In addition, if α is a positive integer, we have

$$\Gamma(\alpha) = (\alpha - 1)!. \tag{2.30}$$

Both the df and the hf of the gamma distribution are not in analytic form. However, it can be shown that the hf decreases with x if $\alpha < 1$, and increases with x if $\alpha > 1$.[8] The mean and variance of X are

$$\mathrm{E}(X) = \alpha\beta \quad \text{and} \quad \mathrm{Var}(X) = \alpha\beta^2, \tag{2.31}$$

and its mgf is

$$M_X(t) = \frac{1}{(1 - \beta t)^\alpha}, \quad \text{for } t < \frac{1}{\beta}. \tag{2.32}$$

From equation (2.30), $\Gamma(1) = 1$. Thus, from equation (2.27) we can see that the pdf of $\mathcal{G}(1, \beta)$ is the same as that of $\mathcal{E}(1/\beta)$, and the exponential distribution is a special case of the gamma distribution. Suppose $X_1, \cdots, X_n$ are independently and identically distributed as $X \sim \mathcal{E}(1/\beta)$, and we define $Y = X_1 + \cdots + X_n$, then the mgf of Y is

$$M_Y(t) = [M_X(t)]^n = \left[\frac{\frac{1}{\beta}}{\frac{1}{\beta} - t}\right]^n = \frac{1}{(1 - \beta t)^n}, \tag{2.33}$$

[8] See Klugman *et al.* (2004, p.51), for a proof of this result.

which is the mgf of $\mathcal{G}(n, \beta)$. Thus, the sum of iid exponential distributions follows a gamma distribution with a positive integer-valued α. A gamma distribution with α being a positive integer is referred to in the literature as an **Erlang distribution**.

2.2.3 Weibull distribution

A random variable X has a 2-parameter Weibull distribution if its pdf is

$$f_X(x) = \left(\frac{\alpha}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\alpha-1} \exp\left[-\left(\frac{x}{\lambda}\right)^{\alpha}\right], \quad \text{for } x \geq 0, \tag{2.34}$$

where α is the shape parameter and λ is the scale parameter. We denote the distribution by $\mathcal{W}(\alpha, \lambda)$, where both α and λ are positive. The mean and variance of X are

$$\mathrm{E}(X) = \mu = \lambda\,\Gamma\left(1 + \frac{1}{\alpha}\right) \quad \text{and} \quad \mathrm{Var}(X) = \lambda^2\,\Gamma\left(1 + \frac{2}{\alpha}\right) - \mu^2. \tag{2.35}$$

The df of X is

$$F_X(x) = 1 - \exp\left[-\left(\frac{x}{\lambda}\right)^{\alpha}\right], \quad \text{for } x \geq 0. \tag{2.36}$$

Due to its complexity, the mgf of the Weibull distribution is not presented here. A Weibull distribution with $\lambda = 1$ is called the standard Weibull, with pdf equal to $\alpha x^{\alpha-1} \exp(-x^{\alpha})$.

We shall see later that there is a close relationship between the exponential distribution and the Weibull distribution. Figure 2.2 shows the pdf of the standard Weibull distribution with different shape parameter α.

2.2.4 Pareto distribution

A random variable X has a Pareto distribution with parameters $\alpha > 0$ and $\gamma > 0$, denoted by $\mathcal{P}(\alpha, \gamma)$, if its pdf is

$$f_X(x) = \frac{\alpha\gamma^{\alpha}}{(x+\gamma)^{\alpha+1}}, \quad \text{for } x \geq 0. \tag{2.37}$$

The df of X is

$$F_X(x) = 1 - \left(\frac{\gamma}{x+\gamma}\right)^{\alpha}, \quad \text{for } x \geq 0. \tag{2.38}$$

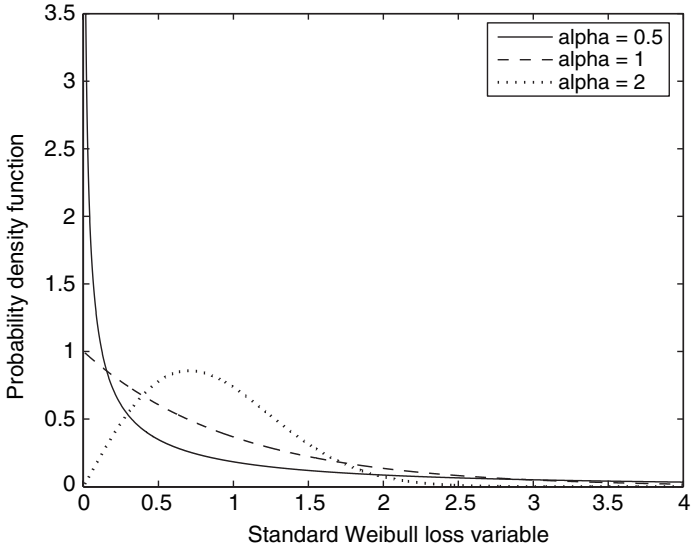

Figure 2.2 Probability density functions of standard Weibull distribution

The hf of X is

$$h_X(x) = \frac{f_X(x)}{S_X(x)} = \frac{\alpha}{x+\gamma}, \tag{2.39}$$

which decreases with x. The kth moment of X exists for $k < \alpha$. For $\alpha > 2$, the mean and variance of X are

$$\mathrm{E}(X) = \frac{\gamma}{\alpha - 1} \quad \text{and} \quad \mathrm{Var}(X) = \frac{\alpha\gamma^2}{(\alpha-1)^2(\alpha-2)}. \tag{2.40}$$

The Pareto distribution was first applied in economics to study income distribution. It does not have a mgf. We shall see in the next section that the Pareto distribution can be derived as a mixture of exponential distributions. In Section 2.4 we discuss some tail properties of the Pareto distribution.

2.3 Some methods for creating new distributions

In this section we discuss some methods for creating new distributions. In particular, we consider the methods of transformation, mixture distribution, and splicing.

2.3.1 Transformation of random variable

New distributions may be created by transforming a random variable with a known distribution. The easiest transformation is perhaps the multiplication or division by a constant. For example, let $X \sim \mathcal{W}(\alpha, \lambda)$. Consider the **scaling** of X by the scale parameter λ and define

$$Y = g(X) = \frac{X}{\lambda}. \tag{2.41}$$

Using Theorem 2.1, we have $x = g^{-1}(y) = \lambda y$, so that

$$\frac{dg^{-1}(y)}{dy} = \lambda. \tag{2.42}$$

Hence, from equations (2.19) and (2.34), we have

$$f_Y(y) = \frac{\alpha y^{\alpha-1}}{\lambda}[\exp\left(-y^{\alpha}\right)]\lambda = \alpha y^{\alpha-1}\exp\left(-y^{\alpha}\right), \tag{2.43}$$

which is the pdf of a standard Weibull distribution.

Another common transformation is the **power transformation**. To illustrate this application, assume $X \sim \mathcal{E}(\lambda)$ and define $Y = X^{\frac{1}{\alpha}}$ for an arbitrary constant $\alpha > 0$. Thus, $x = g^{-1}(y) = y^{\alpha}$, and we have

$$\frac{dg^{-1}(y)}{dy} = \alpha y^{\alpha-1}. \tag{2.44}$$

Applying Theorem 2.1 we obtain

$$f_Y(y) = \lambda\alpha y^{\alpha-1}\exp(-\lambda y^{\alpha}). \tag{2.45}$$

If we let

$$\lambda = \frac{1}{\beta^{\alpha}}, \tag{2.46}$$

equation (2.45) can be written as

$$f_Y(y) = \frac{\alpha}{\beta^{\alpha}}\, y^{\alpha-1}\exp\left[-\left(\frac{y}{\beta}\right)^{\alpha}\right], \tag{2.47}$$

from which we can conclude $Y \sim \mathcal{W}(\alpha, \beta) \equiv \mathcal{W}(\alpha, 1/\lambda^{\frac{1}{\alpha}})$. Thus, the Weibull distribution can be obtained as a power transformation of an exponential distribution.

We now consider the **exponential transformation**. Let X be normally distributed with mean μ and variance σ^2, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$. As X can take negative values, it is not suitable for analyzing claim severity. However, a new random variable may be created by taking the exponential of X. Thus, we define $Y = e^X$, so that $x = \log y$. The pdf of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{2.48}$$

As

$$\frac{d \log y}{dy} = \frac{1}{y}, \tag{2.49}$$

applying Theorem 2.1, we obtain the pdf of Y as

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma y} \exp\left[-\frac{(\log y - \mu)^2}{2\sigma^2}\right]. \tag{2.50}$$

A random variable Y with pdf given by equation (2.50) is said to have a **lognormal distribution** with parameters μ and σ^2, denoted by $\mathcal{L}(\mu, \sigma^2)$. In other words, if $\log Y \sim \mathcal{N}(\mu, \sigma^2)$, then $Y \sim \mathcal{L}(\mu, \sigma^2)$. The mean and variance of $Y \sim \mathcal{L}(\mu, \sigma^2)$ are given by

$$\mathrm{E}(Y) = \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{2.51}$$

and

$$\mathrm{Var}(Y) = \left[\exp\left(2\mu + \sigma^2\right)\right]\left[\exp(\sigma^2) - 1\right]. \tag{2.52}$$

A lognormal distribution is skewed to the right. Figure 2.3 presents the pdf of some lognormal distributions with $\mu = 0$ and $\sigma = 1, 2$, and 4.

2.3.2 Mixture distribution

In Section 1.5.2 we discuss the creation of a new distribution as a finite mixture of pdf or pf. For continuous distributions a finite mixture can be created from n pdf. Thus, if $X_1, \cdots, X_n$ are random variables with corresponding pdf $f_{X_1}(\cdot), \cdots, f_{X_n}(\cdot)$, a new random variable X may be created with pdf $f_X(\cdot)$ given by

$$f_X(x) = p_1 f_{X_1}(x) + \cdots + p_n f_{X_n}(x), \tag{2.53}$$

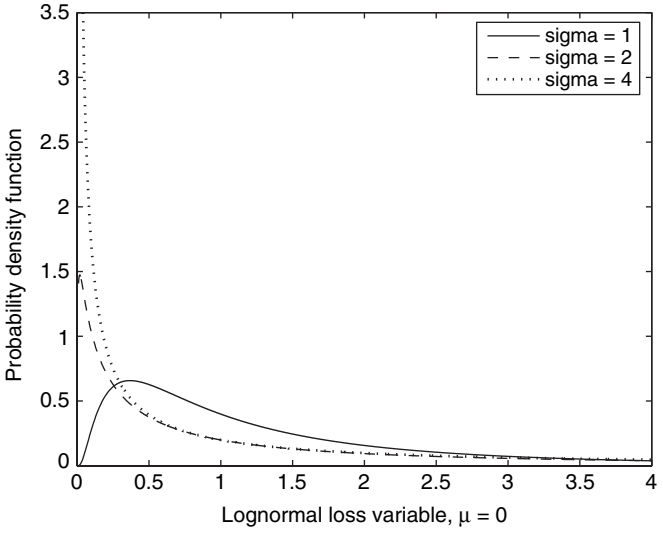

Figure 2.3 Probability density functions of $\mathcal{L}(0, \sigma^2)$

where $p_i \geq 0$ for $i = 1, \cdots, n$ and $\sum_{i=1}^{n} p_i = 1$. We now formally extend this to continuous mixing (a brief discussion of this is in Section 1.5.2).

Let X be a continuous random variable with pdf $f_X(x \mid \lambda)$, which depends on the parameter λ. We allow λ to be the realization of a random variable Λ with support Ω_Λ and pdf $f_\Lambda(\lambda \mid \theta)$, where θ is the parameter determining the distribution of Λ, sometimes called the **hyperparameter**. A new random variable Y may then be created by *mixing* the pdf $f_X(x \mid \lambda)$ to form the pdf

$$f_Y(y \mid \theta) = \int_{\lambda \in \Omega_\Lambda} f_X(y \mid \lambda) f_\Lambda(\lambda \mid \theta)\, d\lambda. \tag{2.54}$$

Thus, Y is a mixture distribution and its pdf depends on θ. Unlike equation (2.53), in which the mixing distribution is discrete, the distribution of Y given in equation (2.54) is a continuous mixture, as its mixing distribution is represented by the pdf $f_\Lambda(\lambda \mid \theta)$. The example below illustrates an application of continuous mixing.

Example 2.4 Assume $X \sim \mathcal{E}(\lambda)$, and let the parameter λ be distributed as $\mathcal{G}(\alpha, \beta)$. Determine the mixture distribution.

Solution We have

$$f_X(x \mid \lambda) = \lambda e^{-\lambda x},$$

and

$$f_\Lambda(\lambda \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\, \lambda^{\alpha-1} e^{-\frac{\lambda}{\beta}}.$$

Thus

$$\begin{aligned}
\int_0^\infty f_X(x \mid \lambda) f_\Lambda(\lambda \mid \alpha, \beta)\, d\lambda &= \int_0^\infty \lambda e^{-\lambda x} \left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\, \lambda^{\alpha-1} e^{-\frac{\lambda}{\beta}} \right] d\lambda \\
&= \int_0^\infty \frac{\lambda^\alpha \exp\left[-\lambda \left(x + \frac{1}{\beta} \right) \right]}{\Gamma(\alpha)\beta^\alpha}\, d\lambda \\
&= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\beta^\alpha} \left[\frac{\beta}{\beta x + 1} \right]^{\alpha+1} \\
&= \frac{\alpha}{\beta^\alpha} \left[\frac{\beta}{\beta x + 1} \right]^{\alpha+1}.
\end{aligned}$$

If we let $\gamma = 1/\beta$, the above expression can be written as

$$\frac{\alpha}{\beta^\alpha} \left[\frac{\beta}{\beta x + 1} \right]^{\alpha+1} = \frac{\alpha \gamma^\alpha}{(x+\gamma)^{\alpha+1}},$$

which is the pdf of $\mathcal{P}(\alpha, \gamma)$. Thus, the gamma–exponential mixture has a Pareto distribution. We also see that the distribution of the mixture distribution depends on α and β (or α and γ). □

In the above example we may consider the exponential distribution as a *conditional distribution* given the parameter Λ and denote this by $X \mid \Lambda \sim \mathcal{E}(\Lambda)$. The distribution of Λ is the *mixing distribution*, with the mixture distribution regarded as the *unconditional distribution* of X.

As the mixture distribution is Pareto, from equation (2.40) we may conclude that the unconditional mean and variance of X are (when $\alpha > 2$)

$$\mathrm{E}(X) = \frac{1}{(\alpha-1)\beta} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{\alpha}{(\alpha-1)^2(\alpha-2)\beta^2}. \tag{2.55}$$

However, as the mixing technique does not always give rise to a straightforward pdf its mean and variance may not be directly obtainable from a standard distribution. The example below illustrates the computation of the mean and variance of a continuous mixture using rules of **conditional expectation.** For the mean, we use the following result

$$\mathrm{E}(X) = \mathrm{E}\left[\mathrm{E}(X \mid \Lambda) \right]. \tag{2.56}$$

For the variance, we use the result in equation (A.115), which can be rewritten in the current context as[9]

$$\mathrm{Var}(X) = \mathrm{E}\left[\mathrm{Var}(X \mid \Lambda)\right] + \mathrm{Var}\left[\mathrm{E}(X \mid \Lambda)\right]. \tag{2.57}$$

Example 2.5 Assume $X \mid \Lambda \sim \mathcal{E}(\Lambda)$, and let the parameter Λ be distributed as $\mathcal{G}(\alpha, \beta)$. Calculate the unconditional mean and variance of X using rules of conditional expectation.

Solution As the conditional distribution of X is $\mathcal{E}(\Lambda)$, from equation (2.25) we have

$$\mathrm{E}(X \mid \Lambda) = \frac{1}{\Lambda}.$$

Thus, from equation (2.56), we have

$$\begin{aligned}
\mathrm{E}(X) &= \mathrm{E}\left(\frac{1}{\Lambda}\right) \\
&= \int_0^\infty \frac{1}{\lambda}\left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}\right]d\lambda \\
&= \frac{\Gamma(\alpha-1)\beta^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \\
&= \frac{1}{(\alpha-1)\beta}.
\end{aligned}$$

From equation (2.25), we have

$$\mathrm{Var}(X \mid \Lambda) = \frac{1}{\Lambda^2},$$

so that using equation (2.57) we have

$$\mathrm{Var}(X) = \mathrm{E}\left(\frac{1}{\Lambda^2}\right) + \mathrm{Var}\left(\frac{1}{\Lambda}\right) = 2\mathrm{E}\left(\frac{1}{\Lambda^2}\right) - \left[\mathrm{E}\left(\frac{1}{\Lambda}\right)\right]^2.$$

As

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{\Lambda^2}\right) &= \int_0^\infty \frac{1}{\lambda^2}\left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}\right]d\lambda \\
&= \frac{\Gamma(\alpha-2)\beta^{\alpha-2}}{\Gamma(\alpha)\beta^\alpha} \\
&= \frac{1}{(\alpha-1)(\alpha-2)\beta^2},
\end{aligned}$$

[9] See Appendix A.11 for a proof of the result, which has been used in the proof of Theorem 1.6.

we conclude

$$\text{Var}(X) = \frac{2}{(\alpha-1)(\alpha-2)\beta^2} - \left[\frac{1}{(\alpha-1)\beta}\right]^2 = \frac{\alpha}{(\alpha-1)^2(\alpha-2)\beta^2}.$$

These results agree with those in equation (2.55) obtained directly from the Pareto distribution. □

2.3.3 Splicing

Splicing is a technique to create a new distribution from standard distributions using different pdf in different parts of the support. Suppose there are k pdf, denoted by $f_1(x), \cdots, f_k(x)$, defined on the support $\Omega_X = [0, \infty)$, a new pdf $f_X(x)$ can be defined as follows

$$f_X(x) = \begin{cases} p_1 f_1^*(x), & x \in [0, c_1), \\ p_2 f_2^*(x), & x \in [c_1, c_2), \\ \vdots & \vdots \\ p_k f_k^*(x), & x \in [c_{k-1}, \infty), \end{cases} \tag{2.58}$$

where $p_i \geq 0$ for $i = 1, \cdots, k$ with $\sum_{i=1}^k p_i = 1$, $c_0 = 0 < c_1 < c_2 \cdots < c_{k-1} < \infty = c_k$, and $f_i^*(x)$ is a legitimate pdf based on $f_i(x)$ in the interval $[c_{i-1}, c_i)$ for $i = 1, \cdots, k$. For the last condition to hold, we define

$$f_i^*(x) = \frac{f_i(x)}{\int_{c_{i-1}}^{c_i} f_i(x)\, dx}, \qquad \text{for } x \in [c_{i-1}, c_i). \tag{2.59}$$

It is then easy to check that $f_X(x) \geq 0$ for $x \geq 0$, and

$$\int_0^\infty f_X(x)\, dx = 1. \tag{2.60}$$

Note that while $f_X(x)$ is a legitimate pdf, it is in general not continuous, as the splicing causes jumps at the points $c_1, \cdots, c_{k-1}$. The example below illustrates the method.

Example 2.6 Let $X_1 \sim \mathcal{E}(0.5)$, $X_2 \sim \mathcal{E}(2)$ and $X_3 \sim \mathcal{P}(2, 3)$, with corresponding pdf $f_i(x)$ for $i = 1, 2$, and 3. Construct a spliced distribution using $f_1(x)$ in the interval $[0, 1)$, $f_2(x)$ in the interval $[1, 3)$, and $f_3(x)$ in the interval $[3, \infty)$, so that each interval has a probability content of one third. Also, determine the spliced distribution so that its pdf is continuous, without imposing equal probabilities for the three segments.

Solution We first compute the probability of each pdf $f_i(x)$ in their respective interval. For $x \in [0, 1)$, we have

$$\int_0^1 f_1(x)\,dx = \int_0^1 0.5e^{-0.5x}\,dx = 1 - e^{-0.5} = 0.3935.$$

For $x \in [1, 3)$, we have

$$\int_1^3 f_2(x)\,dx = \int_1^3 2e^{-2x}\,dx = e^{-0.2} - e^{-0.6} = 0.2699,$$

and for $x \in [3, \infty)$, we have, from equation (2.38)

$$\int_3^\infty f_3(x)\,dx = \left(\frac{3}{3+3}\right)^2 = 0.25.$$

Now $p_1 = p_2 = p_3 = 1/3$. Thus, from equations (2.58) and (2.59), $f_X(x)$ is equal to

$$\frac{1}{3}\left(\frac{0.5e^{-0.5x}}{0.3935}\right) = 0.4235e^{-0.5x}, \qquad \text{for } x \in [0, 1),$$

$$\frac{1}{3}\left(\frac{2e^{-2x}}{0.2699}\right) = 2.4701e^{-2x}, \qquad \text{for } x \in [1, 3),$$

and

$$\frac{1}{3}\left[\frac{(2)(3)^2}{(0.25)(x+3)^3}\right] = \frac{24}{(x+3)^3}, \qquad \text{for } x \in [3, \infty).$$

If the spliced pdf is to be continuous, we require $p_1 f_1^*(1) = p_2 f_2^*(1)$, i.e.

$$p_1\left(\frac{f_1(1)}{0.3935}\right) = p_2\left(\frac{f_2(1)}{0.2699}\right),$$

and similarly

$$p_2\left(\frac{f_2(3)}{0.2699}\right) = (1 - p_1 - p_2)\left(\frac{f_3(3)}{0.25}\right).$$

Solving for the above simultaneous equations, we obtain $p_1 = 0.5522$, $p_2 = 0.4244$, and $p_3 = 0.0234$. Figure 2.4 plots the spliced pdf $f_X(x)$ with equal-probability restriction and continuity restriction. □

2.4 Tail properties of claim severity

Claims with large losses undermine the viability of insurance contracts. Thus, in modeling losses special efforts must be made in analyzing the behavior of

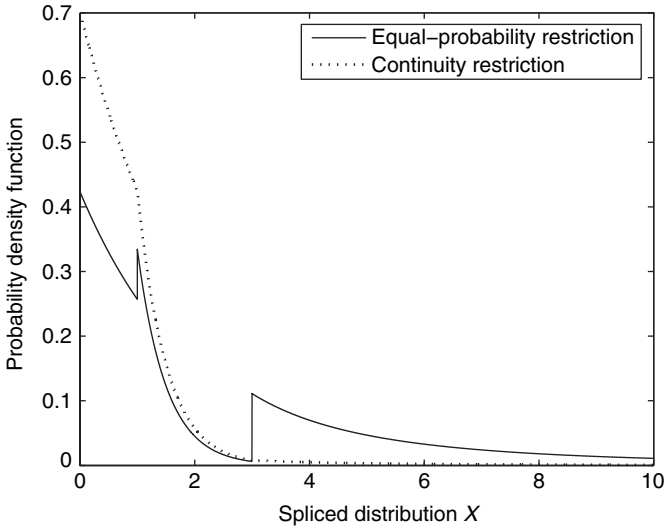

Figure 2.4 Spliced distributions of Example 2.6

extreme values. For claim severity, the extreme values occur in the upper (right-hand) tail of the distribution. A distribution with high probability of heavy loss is said to have either a **fat tail**, **heavy tail**, or **thick tail**, which may be interpreted in the relative or absolute sense. While it is difficult to define *thickness*, some measures may be used as indicators. First, the existence of moments is an indication of whether the distribution has a thick tail. For example, the gamma distribution has moments of all order, which indicates that the probability of extreme values dies down quite fast. In contrast, the Pareto distribution has moments only up to order α. Hence, if $\alpha < 2$, the Pareto distribution has *no variance*. This is an indication of a thick-tail distribution.

To compare the tail behavior of two distributions we may take the **limiting ratio** of their sf. The faster the sf approaches zero, the thinner is the tail. However, as the sf of a distribution always tends to zero, the ratio of two sf at the tail end has to be computed using the l'Hôpital rule. Thus, if $S_1(x)$ and $S_2(x)$ are the sf of the random variables X_1 and X_2, respectively, with corresponding pdf $f_1(x)$ and $f_2(x)$, we have

$$\lim_{x\to\infty} \frac{S_1(x)}{S_2(x)} = \lim_{x\to\infty} \frac{S_1'(x)}{S_2'(x)} = \lim_{x\to\infty} \frac{f_1(x)}{f_2(x)}. \tag{2.61}$$

The example below compares the limiting ratio of the Pareto and gamma distributions.

Example 2.7 Let $f_1(x)$ be the pdf of the $\mathcal{P}(\alpha,\gamma)$ distribution, and $f_2(x)$ be the pdf of the $\mathcal{G}(\theta,\beta)$ distribution. Determine the limiting ratio of these distributions, and suggest which distribution has a thicker tail.

Solution The limiting ratio of the Pareto versus the gamma distribution is

$$\lim_{x\to\infty}\frac{f_1(x)}{f_2(x)} = \lim_{x\to\infty}\frac{\dfrac{\alpha\gamma^\alpha}{(x+\gamma)^{\alpha+1}}}{\dfrac{1}{\Gamma(\theta)\beta^\theta}x^{\theta-1}e^{-\frac{x}{\beta}}}$$

$$= \alpha\gamma^\alpha\Gamma(\theta)\beta^\theta\lim_{x\to\infty}\frac{e^{\frac{x}{\beta}}}{(x+\gamma)^{\alpha+1}x^{\theta-1}}.$$

As the exponential function tends to infinity faster than the power function, the ratio of the right-hand side of the above equation tends to infinity as x tends to infinity. Thus, we conclude that the Pareto distribution has a thicker tail than the gamma distribution. This conclusion is congruent with the conclusion drawn based on the comparison of the moments. □

Another important determinant of tail thickness is the hf. Consider a random variable X with sf $S_X(x)$. For any $d > 0$, the ratio

$$\frac{S_X(x+d)}{S_X(x)} < 1. \tag{2.62}$$

This ratio measures the rate of decrease of the upper tail, and, using equation (2.8), it can be expressed in terms of the hf as follows

$$\frac{S_X(x+d)}{S_X(x)} = \frac{\exp\left(-\int_0^{x+d}h_X(s)\,ds\right)}{\exp\left(-\int_0^{x}h_X(s)\,ds\right)}$$

$$= \exp\left(-\int_x^{x+d}h_X(s)\,ds\right)$$

$$= \exp\left(-\int_0^{d}h_X(x+s)\,ds\right). \tag{2.63}$$

Thus, if the hf is a decreasing function in x, $\exp\left(-\int_0^d h_X(x+s)\,ds\right)$ is increasing in x, which implies the ratio $S_X(x+d)/S_X(x)$ increases in x. This is an indication of a thick-tail distribution. However, if the hf is an increasing function in x, the ratio $S_X(x+d)/S_X(x)$ decreases in x, suggesting low probability of extreme values. As mentioned in Section 2.2.2 the gamma distribution with $\alpha > 1$ has an increasing hf, the exponential distribution has a constant hf,

and both the gamma distribution with $\alpha < 1$ and the Pareto distribution have decreasing hf. Thus, in terms of increasing tail thickness based on the hf, we have the ordering of: (a) gamma with $\alpha > 1$, (b) exponential, and (c) gamma with $\alpha < 1$ and Pareto.

We may also quantify extreme losses using quantiles in the upper end of the loss distribution. The **quantile function (qf)** is the inverse of the df. Thus, if

$$F_X(x_\delta) = \delta, \tag{2.64}$$

then

$$x_\delta = F_X^{-1}(\delta). \tag{2.65}$$

$F_X^{-1}(\cdot)$ is called the quantile function and x_δ is the δ-quantile (or the 100δth percentile) of X. Equation (2.65) assumes that for any $0 < \delta < 1$ a unique value x_δ exists.[10]

Example 2.8 Let $X \sim \mathcal{E}(\lambda)$ and $Y \sim \mathcal{L}(\mu, \sigma^2)$. Derive the quantile functions of X and Y. If $\lambda = 1$, $\mu = -0.5$, and $\sigma^2 = 1$, compare the quantiles of X and Y for $\delta = 0.95$ and 0.99.

Solution From equation (2.22), we have

$$F_X(x_\delta) = 1 - e^{-\lambda x_\delta} = \delta,$$

so that $e^{-\lambda x_\delta} = 1 - \delta$, implying

$$x_\delta = -\frac{\log(1-\delta)}{\lambda}.$$

For Y we have

$$\begin{aligned}
\delta &= \Pr(Y \le y_\delta) \\
&= \Pr(\log Y \le \log y_\delta) \\
&= \Pr(\mathcal{N}(\mu, \sigma^2) \le \log y_\delta) \\
&= \Pr\left(Z \le \frac{\log y_\delta - \mu}{\sigma}\right),
\end{aligned}$$

where Z follows the standard normal distribution.

[10] If $F_X(\cdot)$ is not continuous, the inverse function may not exist. However, if $F_X(\cdot)$ is flat in some neighborhoods, there may be multiple solutions of the inverse. To get around these difficulties, we may define the quantile function of X by $F_X^{-1}(\delta) = \inf\{x : F_X(x) \ge \delta\}$.

Thus

$$\frac{\log y_\delta - \mu}{\sigma} = \Phi^{-1}(\delta),$$

where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal. Hence, $y_\delta = \exp\left[\mu + \sigma\Phi^{-1}(\delta)\right]$.

For X, given the parameter value $\lambda = 1$, $\mathrm{E}(X) = \mathrm{Var}(X) = 1$ and $x_{0.95} = -\log(0.05) = 2.9957$. For Y with $\mu = -0.5$ and $\sigma^2 = 1$, from equations (2.51) and (2.52) we have $\mathrm{E}(Y) = 1$ and $\mathrm{Var}(Y) = \exp(1) - 1 = 1.7183$. Hence, X and Y have the same mean, while Y has a larger variance. For the quantile of Y we have $\Phi^{-1}(0.95) = 1.6449$, so that

$$y_{0.95} = \exp\left[\mu + \sigma\Phi^{-1}(0.95)\right] = \exp(1.6449 - 0.5) = 3.1421.$$

Similarly, we obtain $x_{0.99} = 4.6052$ and $y_{0.99} = 6.2109$. Thus, Y has larger quantiles for $\delta = 0.95$ and 0.99, indicating it has a thicker upper tail. Figure 2.5 presents a comparison of the pdf of the two loss random variables in the upper tail. □

Given the tolerance probability $1 - \delta$, the quantile x_δ indicates the loss which will be exceeded with probability $1 - \delta$. However, it does not provide information about how bad the loss might be if loss exceeds this threshold. To address this issue, we may compute the expected loss conditional on the threshold being exceeded. We call this the **conditional tail expectation (CTE)**

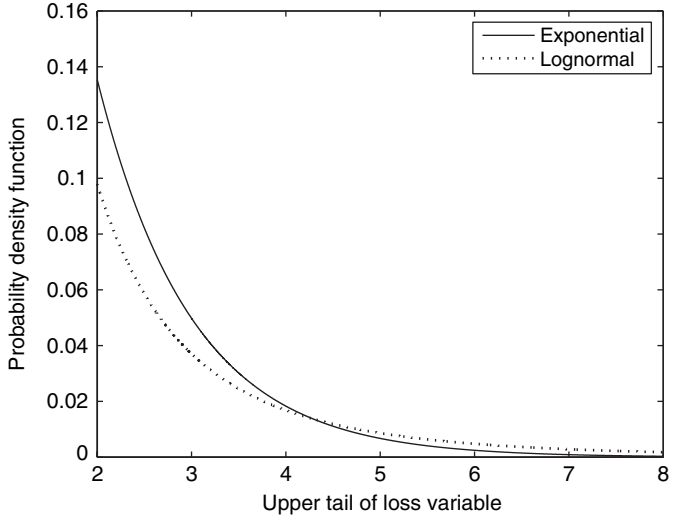

Figure 2.5 Upper tails of pdf of $\mathcal{E}(1)$ and $\mathcal{L}(-0.5, 1)$

with tolerance probability $1-\delta$, denoted by CTE_δ, which is defined as

$$\text{CTE}_\delta = \text{E}(X \mid X > x_\delta). \tag{2.66}$$

To compute CTE_δ we first define the conditional pdf of X given $X > x_\delta$, denoted by $f_{X \mid X > x_\delta}(x)$. Using conditional law of probability, this quantity is given by

$$f_{X \mid X > x_\delta}(x) = \frac{f_X(x)}{\Pr(X > x_\delta)} = \frac{f_X(x)}{S_X(x_\delta)}, \qquad \text{for } x \in (x_\delta, \infty). \tag{2.67}$$

Thus

$$\begin{aligned} \text{CTE}_\delta &= \int_{x_\delta}^{\infty} x f_{X \mid X > x_\delta}(x)\, dx \\ &= \int_{x_\delta}^{\infty} x \left[\frac{f_X(x)}{S_X(x_\delta)} \right] dx \\ &= \frac{\int_{x_\delta}^{\infty} x f_X(x)\, dx}{1-\delta}. \end{aligned} \tag{2.68}$$

Example 2.9 For the loss distributions X and Y given in Example 2.8, calculate $\text{CTE}_{0.95}$.

Solution We first consider X. As $f_X(x) = \lambda e^{-\lambda x}$, the numerator of the last line of equation (2.68) is

$$\begin{aligned} \int_{x_\delta}^{\infty} \lambda x e^{-\lambda x}\, dx &= -\int_{x_\delta}^{\infty} x\, de^{-\lambda x} \\ &= -\left(x e^{-\lambda x} \Big]_{x_\delta}^{\infty} - \int_{x_\delta}^{\infty} e^{-\lambda x}\, dx \right) \\ &= x_\delta e^{-\lambda x_\delta} + \frac{e^{-\lambda x_\delta}}{\lambda}, \end{aligned}$$

which, for $\delta = 0.95$ and $\lambda = 1$, is equal to

$$3.9957 e^{-2.9957} = 0.1997876.$$

Thus, $\text{CTE}_{0.95}$ of X is

$$\frac{0.1997876}{0.05} = 3.9957.$$

The pdf of the lognormal distribution is given in equation (2.50). Thus, the numerator of (2.68) is

$$\int_{y_\delta}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2} \right] dx.$$

To compute this integral, we define the transformation

$$z = \frac{\log x - \mu}{\sigma} - \sigma.$$

As

$$\exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right] = \exp\left[-\frac{(z+\sigma)^2}{2}\right]$$
$$= \exp\left(-\frac{z^2}{2}\right)\exp\left(-\sigma z - \frac{\sigma^2}{2}\right),$$

and

$$dx = \sigma x\, dz = \sigma \exp(\mu + \sigma^2 + \sigma z)\, dz,$$

we have

$$\int_{y_\delta}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right] dx = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$
$$\times \int_{z^*}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$$
$$= \exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 - \Phi(z^*)\right],$$

where $\Phi(\cdot)$ is the df of the standard normal and

$$z^* = \frac{\log y_\delta - \mu}{\sigma} - \sigma.$$

Now we substitute $\mu = -0.5$ and $\sigma^2 = 1$ to obtain

$$z^* = \log y_{0.95} - 0.5 = \log(3.1421) - 0.5 = 0.6449,$$

so that $\text{CTE}_{0.95}$ of Y is

$$\text{CTE}_{0.95} = \frac{e^0\left[1 - \Phi(0.6449)\right]}{0.05} = 5.1900,$$

which is larger than that of X. Thus, Y gives rise to more extreme losses compared to X, whether we measure the extreme events by the upper quantile or CTE. □

2.5 Effects of coverage modifications

To reduce risks and/or control problems of **moral hazard**,[11] insurance companies often modify the policy coverage. Examples of such modifications are **deductibles**, **policy limits**, and **coinsurance**. These modifications change the amount paid by the insurance companies in case of a loss event. For example, with deductibles the insurer does not incur any payment in a loss event if the loss does not exceed the deductible. Thus, we need to distinguish between a **loss event** and a **payment event**. A loss event occurs whenever there is a loss, while a payment event occurs only when the insurer is liable to pay for (some or all of) the loss.

In this section we study the distribution of the *loss to the insurer* when there is coverage modification. To begin with, we define the following notations:

1. X = amount paid in a loss event when there is no coverage modification, also called the **ground-up loss**
2. X_L = amount paid in a loss event when there is coverage modification, also called the **cost per loss**
3. X_P = amount paid in a payment event when there is coverage modification, also called the **cost per payment**

Thus, X and X_P are positive and X_L is nonnegative. Now we consider some coverage modifications and their effects on the loss-amount variable X_L and the payment-amount variable X_P.

2.5.1 Deductible

An insurance policy with a per-loss deductible of d will not pay the insured if the loss X is less than or equal to d, and will pay the insured $X - d$ if the loss X exceeds d. Thus, the amount paid in a loss event, X_L, is given by[12]

$$X_L = \begin{cases} 0, & \text{for } X \le d, \\ X - d, & \text{for } X > d. \end{cases} \tag{2.69}$$

If we adopt the notation

$$x_+ = \begin{cases} 0, & \text{for } x \le 0, \\ x, & \text{for } x > 0, \end{cases} \tag{2.70}$$

[11] Moral hazard refers to the situation in which the insured behaves differently from the way he would behave if he were fully exposed to the risk, such that the risk of burden is transferred to the insurer. For example, in vehicle insurance, an insured may drive less carefully, knowing that the insurer will pay for the damages in accidents.

[12] The deductible defined in equation (2.69) is called an **ordinary deductible**. A deductible policy may also pay 0 when $X \le d$ and X when $X > d$, in which case it is called a **franchise deductible**.

then X_L may also be defined as

$$X_L = (X - d)_+. \tag{2.71}$$

Note that $\Pr(X_L = 0) = F_X(d)$. Thus, X_L is a mixed-type random variable. It has a probability mass at point 0 of $F_X(d)$ and a density function of

$$f_{X_L}(x) = f_X(x + d), \qquad \text{for } x > 0. \tag{2.72}$$

Furthermore

$$F_{X_L}(x) = F_X(x + d) \qquad \text{and} \qquad S_{X_L}(x) = S_X(x + d), \qquad \text{for } x > 0. \tag{2.73}$$

The random variable X_P, called the **excess-loss variable**, is defined only when there is a payment, i.e. when $X > d$. It is a conditional random variable, defined as $X_P = X - d \mid X > d$. X_P follows a continuous distribution if X is continuous, and its pdf is given by[13]

$$f_{X_P}(x) = \frac{f_X(x + d)}{S_X(d)}, \qquad \text{for } x > 0. \tag{2.74}$$

Its sf is computed as

$$S_{X_P} = \frac{S_X(x + d)}{S_X(d)}, \qquad \text{for } x > 0. \tag{2.75}$$

Figure 2.1 plots the df of X, X_L, and X_P.

Note that in empirical applications, only data on payments are available. When the loss X is less than or equal to d, a claim is not made and the loss information is not captured. Thus, for a policy with deductible, only X_P is observed, and some information about X and X_L is lost.[14]

[13] In the life-contingency literature with X being the **age-at-death** variable, X_P is the future lifetime variable *conditional* on an entity reaching age d, and its expectation given in equation (2.77) is called the **expected future lifetime** or **mean residual lifetime**.

[14] Using the terminology in the statistics literature, X_L has a **censored distribution**, while X_P has a **truncated distribution**. Empirically, loss data are truncated.

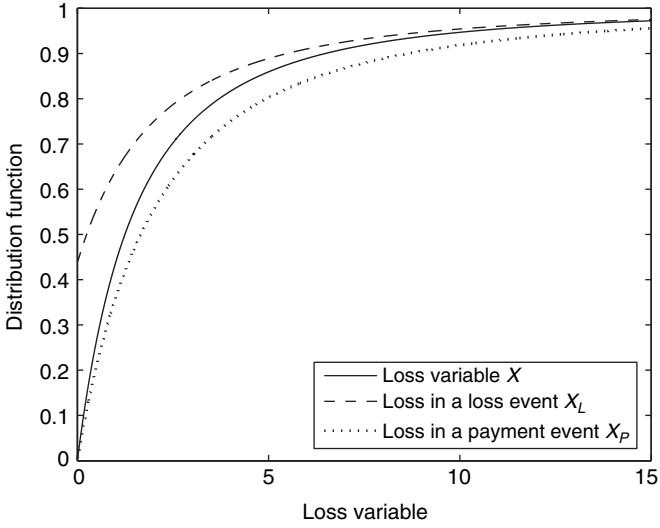

Figure 2.6 Distribution functions of X, X_L, and X_P

The mean of X_L can be computed as follows

$$\begin{aligned}
\mathrm{E}(X_L) &= \int_0^\infty x f_{X_L}(x)\,dx \\
&= \int_d^\infty (x-d) f_X(x)\,dx \\
&= -\int_d^\infty (x-d)\,dS_X(x) \\
&= -\left[(x-d)S_X(x)]_d^\infty - \int_d^\infty S_X(x)\,dx\right] \\
&= \int_d^\infty S_X(x)\,dx. \qquad (2.76)
\end{aligned}$$

On the other hand, the mean of X_P, called the **mean excess loss**, is given by the following formula

$$\begin{aligned}
\mathrm{E}(X_P) &= \int_0^\infty x f_{X_P}(x)\,dx \\
&= \int_0^\infty x\left[\frac{f_X(x+d)}{S_X(d)}\right] dx
\end{aligned}$$

$$= \frac{\int_0^\infty x f_X(x+d)\,dx}{S_X(d)}$$
$$= \frac{\int_d^\infty (x-d) f_X(x)\,dx}{S_X(d)}$$
$$= \frac{\mathrm{E}(X_L)}{S_X(d)}. \tag{2.77}$$

Note that we can also express X_P as $X_P = X_L \,|\, X_L > 0$. Now using conditional expectation, we have

$$\begin{aligned}\mathrm{E}(X_L) &= \mathrm{E}(X_L \,|\, X_L > 0)\Pr(X_L > 0) + \mathrm{E}(X_L \,|\, X_L = 0)\Pr(X_L = 0)\\ &= \mathrm{E}(X_L \,|\, X_L > 0)\Pr(X_L > 0)\\ &= \mathrm{E}(X_P)\Pr(X_L > 0),\end{aligned} \tag{2.78}$$

which implies

$$\mathrm{E}(X_P) = \frac{\mathrm{E}(X_L)}{\Pr(X_L > 0)} = \frac{\mathrm{E}(X_L)}{S_{X_L}(0)} = \frac{\mathrm{E}(X_L)}{S_X(d)}, \tag{2.79}$$

as proved in equation (2.77).

There is also a relationship between $\mathrm{E}(X_P)$ and CTE_δ. From equation (2.68) and the fourth line of equation (2.77), we have

$$\begin{aligned}\mathrm{E}(X_P) &= \frac{\int_d^\infty x f_X(x)\,dx - d\int_d^\infty f_X(x)\,dx}{S_X(d)} = \frac{\int_d^\infty x f_X(x)\,dx - d[S_X(d)]}{S_X(d)}\\ &= \mathrm{CTE}_\delta - d,\end{aligned} \tag{2.80}$$

where $\delta = 1 - S_X(d)$. This equation can also be derived by observing

$$\mathrm{E}(X_P) = \mathrm{E}(X - d \,|\, X > d) = \mathrm{E}(X \,|\, X > d) - \mathrm{E}(d \,|\, X > d) = \mathrm{CTE}_\delta - d. \tag{2.81}$$

Thus, CTE_δ and $\mathrm{E}(X_P)$ are mathematically equivalent. An important difference is that δ in CTE_δ is typically large (say, 0.95 or 0.99) as it measures the upper-tail behavior of the loss distribution. In contrast, the deductible d is typically in the lower tail of the loss distribution.

Example 2.10 For the loss distributions X and Y given in Examples 2.8 and 2.9, assume there is a deductible of $d = 0.25$. Calculate $\mathrm{E}(X_L)$, $\mathrm{E}(X_P)$, $\mathrm{E}(Y_L)$, and $\mathrm{E}(Y_P)$.

Solution For X, we compute $\mathrm{E}(X_L)$ from equation (2.76) as follows

$$\mathrm{E}(X_L) = \int_{0.25}^\infty e^{-x}\,dx = e^{-0.25} = 0.7788.$$

Now $S_X(0.25) = e^{-0.25} = 0.7788$. Thus, from equation (2.77), $\mathrm{E}(X_P) = 1$. For Y, we use the results in Example 2.9. First, we have

$$\mathrm{E}(Y_L) = \int_d^\infty (y-d) f_Y(y)\,dy = \int_d^\infty y f_Y(y)\,dy - d\,[S_Y(d)]\,.$$

Replacing y_δ in Example 2.9 by d, the first term of the above expression becomes

$$\begin{aligned}\int_d^\infty y f_Y(y)\,dy &= \int_d^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log y - \mu)^2}{2\sigma^2}\right] dy \\ &= \exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 - \Phi(z^*)\right],\end{aligned}$$

where

$$z^* = \frac{\log d - \mu}{\sigma} - \sigma = \log(0.25) - 0.5 = -1.8863.$$

As $\Phi(-1.8863) = 0.0296$, we have

$$\int_d^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log y - \mu)^2}{2\sigma^2}\right] dy = 1 - 0.0296 = 0.9704.$$

Now

$$S_Y(d) = \Pr\left(Z > \frac{\log d - \mu}{\sigma}\right) = \Pr(Z > -0.8863) = 0.8123.$$

Hence

$$\mathrm{E}(Y_L) = 0.9704 - (0.25)(0.8123) = 0.7673,$$

and

$$\mathrm{E}(Y_P) = \frac{0.7673}{0.8123} = 0.9446. \qquad \square$$

It turns out that the mean loss in a loss event for the lognormal distribution has important applications in the finance literature.[15] We summarize this result in the following theorem.

Theorem 2.2 *Let $Y \sim \mathcal{L}(\mu, \sigma^2)$, then for a positive constant d*

$$E\left[(Y-d)_+\right] = \exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 - \Phi(z^*)\right] - d[1 - \Phi(z^* + \sigma)], \quad (2.82)$$

where

$$z^* = \frac{\log d - \mu}{\sigma} - \sigma. \quad (2.83)$$

[15] This result is closely related to the celebrated Black–Scholes option pricing formula (see Exercise 2.28).

Proof The derivation in Example 2.10 shows that

$$\mathrm{E}[(Y-d)_+] = \int_d^\infty (y-d) f_Y(y)\,dy = \int_d^\infty y f_Y(y)\,dy - d\,[S_Y(d)], \quad (2.84)$$

and

$$\int_d^\infty y f_Y(y)\,dy = \exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 - \Phi(z^*)\right]. \quad (2.85)$$

Now

$$S_Y(d) = \Pr\left(Z > \frac{\log d - \mu}{\sigma}\right) = 1 - \Phi(z^* + \sigma). \quad (2.86)$$

Combining equations (2.84) through (2.86), we obtain equation (2.82). □

Comparing the results in Examples 2.8 through 2.10, we can see that although the lognormal loss distribution has thicker tail than the exponential loss distribution (note that they are selected to have the same mean), its mean loss in a loss event and mean amount paid in a payment event are smaller than those of the exponential loss distribution. This echoes the point that the deductible is set at the lower tail of the loss distribution, which influences the overall mean of the modified loss.

If we subtract the loss amount with deductible from the loss amount without deductible, the result is the reduction in loss due to the deductible, which is equal to $X - (X-d)_+$. Thus, the expected reduction in loss due to the deductible is

$$\mathrm{E}(X) - \mathrm{E}[(X-d)_+] = \mathrm{E}(X) - \mathrm{E}(X_L). \quad (2.87)$$

Now we define the **loss elimination ratio** with deductible d, denoted by $\mathrm{LER}(d)$, as the ratio of the expected reduction in loss due to the deductible and the expected loss without the deductible, which is given by

$$\mathrm{LER}(d) = \frac{\mathrm{E}(X) - \mathrm{E}(X_L)}{\mathrm{E}(X)}. \quad (2.88)$$

Example 2.11 Calculate $\mathrm{LER}(0.25)$ for the loss distributions X and Y given in Examples 2.8 through 2.10.

Solution For X, we have

$$\mathrm{LER}(0.25) = \frac{1 - 0.7788}{1} = 0.2212.$$

Similarly, $\mathrm{LER}(0.25)$ for Y is $1 - 0.7673 = 0.2327$. Thus, the deductible of amount 0.25 has caused a bigger percentage reduction in the loss for the lognormal loss distribution than for the exponential loss distribution. □

Finally, we note that for any loss distribution X, higher raw moments of X_L and X_P can be computed as follows

$$\mathrm{E}(X_L^k) = \int_d^\infty (x-d)^k f_X(x)\,dx, \tag{2.89}$$

and

$$\mathrm{E}(X_P^k) = \frac{\int_d^\infty (x-d)^k f_X(x)\,dx}{S_X(d)} = \frac{\mathrm{E}(X_L^k)}{S_X(d)}. \tag{2.90}$$

2.5.2 Policy limit

For an insurance policy with a **policy limit**, the insurer compensates the insured up to a pre-set amount, say, u. If a policy has a policy limit but no deductible, then the amount paid in a loss event is the same as the amount paid in a payment event. We denote the amount paid for a policy with a policy limit by X_U. If we define the binary operation $\wedge$ as the minimum of two quantities, so that

$$a \wedge b = \min\{a, b\}, \tag{2.91}$$

then

$$X_U = X \wedge u, \tag{2.92}$$

i.e.

$$X_U = \begin{cases} X, & \text{for } X < u, \\ u, & \text{for } X \geq u. \end{cases} \tag{2.93}$$

X_U defined above is called the **limited-loss variable**. It is interesting to note that the loss amount with a policy limit and the loss amount with a deductible are closely related. Specifically, for any arbitrary positive constant q, the following identity holds

$$X = (X \wedge q) + (X - q)_+. \tag{2.94}$$

This relationship is verified in Table 2.1, from which we can see that the two sides of equation (2.94) are equal, whether $X < q$ or $X \geq q$.

Thus, $(X \wedge u) = X - (X-u)_+$, which can be interpreted as the saving to the insurer if the policy has a *deductible* of amount u. With the identity in equation

Table 2.1. *Proof of equation (2.94)*

	$X < q$	$X \geq q$
$X \wedge q$	X	q
$(X-q)_+$	0	$X-q$
$(X \wedge q)+(X-q)_+$	X	X

(2.94), LER can be written as

$$\begin{aligned} \mathrm{LER}(d) &= \frac{\mathrm{E}(X)-\mathrm{E}\left[(X-d)_+\right]}{\mathrm{E}(X)} = \frac{\mathrm{E}(X)-[\mathrm{E}(X)-\mathrm{E}(X\wedge d)]}{\mathrm{E}(X)} \\ &= \frac{\mathrm{E}(X\wedge d)}{\mathrm{E}(X)}. \end{aligned} \tag{2.95}$$

2.5.3 Coinsurance

An insurance policy may specify that the insurer and insured share the loss in a loss event, which is called **coinsurance**. We consider a simple coinsurance policy in which the insurer pays the insured a fixed portion c of the loss in a loss event, where $0 < c < 1$. Under pure coinsurance the insurer pays damages whenever there is a loss. We denote X_C as the payment made by the insurer with coinsurance. Thus

$$X_C = c\,X, \tag{2.96}$$

where X is the loss without policy modification. Using Theorem 2.1, the pdf of X_C is

$$f_{X_C}(x) = \frac{1}{c} f_X\left(\frac{x}{c}\right), \tag{2.97}$$

and it is also true that

$$\mathrm{E}(X_C) = c\,\mathrm{E}(X). \tag{2.98}$$

Now we consider a policy with a deductible of amount d and a **maximum covered loss** of amount u ($u > d$). Thus, the policy limit is of amount $u-d$. In addition, assume the policy has a coinsurance factor c ($0 < c < 1$). We denote the loss random variable in a loss event of this insurance policy by X_T, which is given by

$$X_T = c\left[(X\wedge u)-(X\wedge d)\right] = c\left[(X-d)_+-(X-u)_+\right]. \tag{2.99}$$

It can be checked that X_T defined above satisfies

$$X_T = \begin{cases} 0, & \text{for } X < d, \\ c(X-d), & \text{for } d \le X < u, \\ c(u-d), & \text{for } X \ge u. \end{cases} \tag{2.100}$$

From equation (2.99), we have

$$\mathrm{E}(X_T) = c\left\{\mathrm{E}\left[(X-d)_+\right] - \mathrm{E}\left[(X-u)_+\right]\right\}, \tag{2.101}$$

which can be computed using equation (2.76).

Example 2.12 For the exponential loss distribution X and lognormal loss distribution Y in Examples 2.8 through 2.11, assume there is a deductible of $d = 0.25$, maximum covered loss of $u = 4$, and coinsurance factor of $c = 0.8$. Calculate the mean loss in a loss event of these two distributions.

Solution We use equation (2.101) to calculate $\mathrm{E}(X_T)$ and $\mathrm{E}(Y_T)$. $\mathrm{E}[(X-d)_+]$ and $\mathrm{E}[(Y-d)_+]$ are computed in Example 2.10 as 0.7788 and 0.7673, respectively. We now compute $\mathrm{E}[(X-u)_+]$ and $\mathrm{E}[(Y-u)_+]$ using the method in Example 2.10, with u replacing d. For X, we have

$$\mathrm{E}\left[(X-u)_+\right] = \int_u^\infty e^{-x}\,dx = e^{-4} = 0.0183.$$

For Y, we have $z^* = \log(4) - 0.5 = 0.8863$ so that $\Phi(z^*) = 0.8123$, and

$$S_Y(u) = \Pr\left(Z > \frac{\log(u)-\mu}{\sigma}\right) = \Pr(Z > 1.8863) = 0.0296.$$

Thus

$$\mathrm{E}\left[(Y-u)_+\right] = (1-0.8123) - (4)(0.0296) = 0.0693.$$

Therefore, from equation (2.101), we have

$$\mathrm{E}(X_T) = (0.8)\,(0.7788 - 0.0183) = 0.6084,$$

and

$$\mathrm{E}(Y_T) = (0.8)(0.7673 - 0.0693) = 0.5584.$$

We concluded from Example 2.10 that the mean loss with a deductible of 0.25 is lower for the lognormal distribution than the exponential. Now the maximum covered loss brings about a bigger reduction in loss for the lognormal distribution as it has a thicker tail than the exponential. Thus, the resulting mean loss for the lognormal distribution is further reduced compared to that of the exponential. □

To compute the variance of X_T, we may use the result $\mathrm{Var}(X_T) = \mathrm{E}(X_T^2) - [\mathrm{E}(X_T)]^2$. When $c = 1$, $\mathrm{E}(X_T^2)$ can be evaluated from the following result

$$\mathrm{E}(X_T^2) = \mathrm{E}\left[(X \wedge u)^2\right] - \mathrm{E}\left[(X \wedge d)^2\right] - 2d\left\{\mathrm{E}\left[(X \wedge u)\right] - \mathrm{E}\left[(X \wedge d)\right]\right\}. \tag{2.102}$$

To prove the above equation, note that

$$\begin{aligned}\mathrm{E}(X_T^2) &= \mathrm{E}\left\{[(X \wedge u) - (X \wedge d)]^2\right\} \\ &= \mathrm{E}\left[(X \wedge u)^2\right] + \mathrm{E}\left[(X \wedge d)^2\right] - 2\mathrm{E}\left[(X \wedge u)(X \wedge d)\right]. \end{aligned} \tag{2.103}$$

Now we have

$$\begin{aligned}\mathrm{E}\left[(X \wedge u)(X \wedge d)\right] &= \int_0^d x^2 f_X(x)\,dx + d\int_d^u x f_X(x)\,dx + du[1 - F_X(u)] \\ &= \int_0^d x^2 f_X(x)\,dx + d\left[\int_0^u x f_X(x)\,dx - \int_0^d x f_X(x)\,dx\right] \\ &\quad + du[1 - F_X(u)] \\ &= \int_0^d x^2 f_X(x)\,dx + d^2[1 - F_X(d)] \\ &\quad + d\left[\int_0^u x f_X(x)\,dx + u[1 - F_X(u)]\right] \\ &\quad - d\left[\int_0^d x f_X(x)\,dx + d[1 - F_X(d)]\right] \\ &= \mathrm{E}\left[(X \wedge d)^2\right] + d\left\{\mathrm{E}\left[(X \wedge u)\right] - \mathrm{E}\left[(X \wedge d)\right]\right\}. \end{aligned} \tag{2.104}$$

Substituting equation (2.104) into (2.103), we obtain the result in equation (2.102). Finally, if the policy has a deductible d but no policy limit (i.e. $u = \infty$), we have

$$\begin{aligned}\mathrm{E}(X_L^2) &= \mathrm{E}\left[(X - d)_+^2\right] \\ &= \mathrm{E}\left\{[X - (X \wedge d)]^2\right\} \\ &= \mathrm{E}(X^2) - \mathrm{E}\left[(X \wedge d)^2\right] - 2d\left\{\mathrm{E}(X) - \mathrm{E}\left[(X \wedge d)\right]\right\}. \end{aligned} \tag{2.105}$$

2.5.4 Effects of inflation

While loss distributions are specified based on current experience and data, inflation may cause increases in the costs. However, policy specifications (such as deductible) remain unchanged for the policy period. To model the effects of inflation, we consider a one-period insurance policy and assume the rate of price increase in the period to be r. We use a tilde to denote inflation-adjusted losses. Thus, the inflation-adjusted loss distribution is denoted by $\tilde{X}$, which is equal to $(1+r)X$. For an insurance policy with deductible d, the loss in a loss event and the loss in a payment event with inflation adjustment are denoted by $\tilde{X}_L$ and $\tilde{X}_P$, respectively. As the deductible is not inflation adjusted, we have

$$\tilde{X}_L = (\tilde{X} - d)_+ = \tilde{X} - (\tilde{X} \wedge d), \tag{2.106}$$

and

$$\tilde{X}_P = \tilde{X} - d \,|\, \tilde{X} - d > 0 = \tilde{X}_L \,|\, \tilde{X}_L > 0. \tag{2.107}$$

For any positive constants k, a, and b, we have

$$k(a-b)_+ = (ka - kb)_+ \qquad \text{and} \qquad k(a \wedge b) = (ka) \wedge (kb). \tag{2.108}$$

Thus, the mean inflation-adjusted loss is given by

$$\begin{aligned} \mathrm{E}(\tilde{X}_L) &= \mathrm{E}\left[(\tilde{X} - d)_+\right] \\ &= \mathrm{E}\left[(1+r)\left(X - \frac{d}{1+r}\right)_+\right] \\ &= (1+r)\,\mathrm{E}\left[\left(X - \frac{d}{1+r}\right)_+\right]. \end{aligned} \tag{2.109}$$

Similarly, we can also show that

$$\mathrm{E}(\tilde{X}_L) = (1+r)\left\{\mathrm{E}(X) - \mathrm{E}\left[X \wedge \left(\frac{d}{1+r}\right)\right]\right\}. \tag{2.110}$$

From equation (2.107), we have

$$\mathrm{E}(\tilde{X}_P) = \mathrm{E}(\tilde{X}_L \,|\, \tilde{X}_L > 0) = \frac{\mathrm{E}(\tilde{X}_L)}{\Pr(\tilde{X}_L > 0)}. \tag{2.111}$$

As

$$\Pr(\tilde{X}_L > 0) = \Pr(\tilde{X} > d) = \Pr\left(X > \frac{d}{1+r}\right) = S_X\left(\frac{d}{1+r}\right), \tag{2.112}$$

we conclude

$$\mathrm{E}(\tilde{X}_P) = \frac{\mathrm{E}(\tilde{X}_L)}{S_X\left(\frac{d}{1+r}\right)}. \tag{2.113}$$

For a policy with a policy limit u, we denote the loss in a loss event by $\tilde{X}_U$. Thus

$$\tilde{X}_U = \tilde{X} \wedge u = (1+r)\left[X \wedge \left(\frac{u}{1+r}\right)\right], \tag{2.114}$$

and

$$\mathrm{E}(\tilde{X}_U) = (1+r)\,\mathrm{E}\left[X \wedge \left(\frac{u}{1+r}\right)\right]. \tag{2.115}$$

If we consider a policy with deductible d, maximum covered loss u and coinsurance factor c, and denote the loss in a loss event with inflation by $\tilde{X}_T$, then, similar to equation (2.99), we have

$$\tilde{X}_T = c\left[(\tilde{X} \wedge u) - (\tilde{X} \wedge d)\right] = c\left[(\tilde{X} - d)_+ - (\tilde{X} - u)_+\right]. \tag{2.116}$$

The mean loss in a loss event is then given by

$$\begin{aligned}\mathrm{E}(\tilde{X}_T) &= c(1+r)\left\{\mathrm{E}\left[X \wedge \left(\frac{u}{1+r}\right)\right] - \mathrm{E}\left[X \wedge \left(\frac{d}{1+r}\right)\right]\right\} \\ &= c(1+r)\left[\mathrm{E}\left(X - \frac{d}{1+r}\right)_+ - \mathrm{E}\left(X - \frac{u}{1+r}\right)_+\right].\end{aligned} \tag{2.117}$$

2.5.5 Effects of deductible on claim frequency

If the loss incurred in a loss event is less than the deductible, no claim will be made. Thus, the deductible affects the distribution of the claim frequency. Suppose N denotes the claim frequency when the insurance policy has no deductible, and N_D denotes the claim frequency when there is a deductible of amount d. Let $P_N(t)$ and $P_{N_D}(t)$ be the pgf of N and N_D, respectively. We define I_i as the random variable taking value 1 when the ith loss is larger than d (i.e. the loss gives rise to a claim event) and 0 otherwise, and assume I_i to be iid as I. Furthermore, let $\Pr(I = 1) = v$, so that the pgf of I is

$$P_I(t) = 1 - v + vt. \tag{2.118}$$

We note that

$$N_D = I_1 + \cdots + I_N, \tag{2.119}$$

so that N_D is a compound distribution, with N being the primary distribution and I the secondary distribution. Hence, from equation (1.69), we have

$$P_{N_D}(t) = P_N\left[P_I(t)\right] = P_N\left[1 + v(t-1)\right]. \tag{2.120}$$

The above equation can be further simplified if N belongs to the $(a, b, 0)$ class. From Theorem 1.3, if N belongs to the $(a, b, 0)$ class of distributions, the pgf of N can be written as

$$P_N(t \mid \beta) = Q_N\left[\beta(t-1)\right], \tag{2.121}$$

where β is a parameter of the distribution of N and $Q_N(\cdot)$ is a function of $\beta(t-1)$ only. Thus

$$\begin{aligned} P_{N_D}(t) &= P_N\left[P_I(t)\right] \\ &= Q_N\left\{\beta\left[P_I(t) - 1\right]\right\} \\ &= Q_N\left[\beta v(t-1)\right] \\ &= P_N(t \mid \beta v), \end{aligned} \tag{2.122}$$

so that N_D has the same distribution as N, with the parameter β replaced by βv (while the values of other parameters, if any, remain unchanged).

Example 2.13 Consider the exponential loss distribution $\mathcal{E}(1)$ and lognormal loss distribution $\mathcal{L}(-0.5, 1)$ discussed in Examples 2.8 through 2.12. Assume there is a deductible of $d = 0.25$, and the claim frequencies for both loss distributions without deductibles are (a) $\mathcal{PN}(\lambda)$ and (b) $\mathcal{NB}(r, \theta)$. Find the distributions of the claim frequencies with the deductibles.

Solution We first consider the probability v of the deductible being exceeded. From Example 2.10, $v = 0.7788$ for the $\mathcal{E}(1)$ loss, and $v = 0.8123$ for the $\mathcal{L}(-0.5, 1)$ loss. For (a), when $N \sim \mathcal{PN}(\lambda)$, $N_D \sim \mathcal{PN}(0.7788\lambda)$ if the loss is $\mathcal{E}(1)$, and $N_D \sim \mathcal{PN}(0.8123\lambda)$ if the loss is $\mathcal{L}(-0.5, 1)$. This results from the fact that λ equals β defined in equation (2.121) (see the proof of Theorem 1.3). For (b), when $N \sim \mathcal{NB}(r, \theta)$, from equation (1.59), we have

$$P_N(t) = \left[\frac{1}{1 - \beta(t-1)}\right]^r,$$

where

$$\beta = \frac{1-\theta}{\theta}$$

so that

$$\theta = \frac{1}{1+\beta}.$$

Thus, when the loss is distributed as $\mathcal{E}(1)$, $N_D \sim \mathcal{NB}(r, \theta^*)$ with θ^* given by

$$\theta^* = \frac{1}{1+0.7788\beta} = \frac{\theta}{\theta + 0.7788(1-\theta)}.$$

Likewise, when the loss is distributed as $\mathcal{L}(-0.5, 1)$, $N_D \sim \mathcal{NB}(r, \theta^*)$ with θ^* given by

$$\theta^* = \frac{\theta}{\theta + 0.8123(1-\theta)}. \qquad \square$$

2.6 Excel computation notes

Excel provides some functions for the computation of the pdf and df of some continuous distributions. These functions are summarized in Table 2.2. If `ind` is an argument of the function, setting it to `FALSE` computes the pdf and setting it to `TRUE` computes the df. For $\mathcal{E}(\lambda)$ and $\mathcal{W}(\alpha, \lambda)$, their pdf and df are available in closed form. The df of $\mathcal{G}(\alpha, \beta)$ and $\mathcal{N}(\mu, \sigma^2)$, however, are not available in closed form, and the Excel functions are useful for the computation. For the functions `NORMSDIST` and `LOGNORMDIST`, only the df are computed.

Table 2.3 summarizes the inverse df (i.e. the qf) of the gamma and normal distributions. As these functions have no closed-form solutions, the Excel functions are particularly useful. In contrast, for the Weibull and Pareto distributions, the qf are available in closed form. Another useful function in Excel is `GAMMALN(x1)`, which computes $\log[\Gamma(x)]$, where $x =$ `x1`, from which $\Gamma(x)$ can be computed.

Excel has a Solver which can be used to compute the root (or the maximum or minimum) of a nonlinear equation (or function). For example, suppose we wish to compute the root of the cubic equation

$$x^3 - 2x^2 - x + 2 = 0. \tag{2.123}$$

With the initial guess set to a given value, the Solver computes the roots as desired. Figure 2.7 illustrates the computation. An initial guess value of 0.9 is entered in cell A1, and the expression in equation (2.123) is entered in cell A2, resulting in a value of 0.209. The Solver is called (from `Tools`, followed by `Solver`). In the Solver Parameters window the Target Cell is set to A2 with a target Value of 0. The solution is computed by changing the value in cell A1. The tolerance limits and convergence criteria of the Solver may be varied by

Table 2.2. *Some Excel functions for the computation of the pdf $f_X(x)$ and df $F_X(x)$ of continuous random variable X*

		Example	
X	Excel function	Input	Output
$\mathcal{E}(\lambda)$	`EXPONDIST(x1,x2,ind)` `x1` $= x$ `x2` $= \lambda$	`EXPONDIST(4,0.5,FALSE)` `EXPONDIST(4,0.5,TRUE)`	0.0677 0.8647
$\mathcal{G}(\alpha,\beta)$	`GAMMADIST(x1,x2,x3,ind)` `x1` $= x$ `x2` $= \alpha$ `x3` $= \beta$	`GAMMADIST(4,1.2,2.5,FALSE)` `GAMMADIST(4,1.2,2.5,TRUE)`	0.0966 0.7363
$\mathcal{W}(\alpha,\lambda)$	`WEIBULL(x1,x2,x3,ind)` `x1` $= x$ `x2` $= \alpha$ `x3` $= \lambda$	`WEIBULL(10,2,10,FALSE)` `WEIBULL(10,2,10,TRUE)`	0.0736 0.6321
$\mathcal{N}(0,1)$	`NORMSDIST(x1)` `x1` $= x$ output is $\Pr(\mathcal{N}(0,1) \le x)$	`NORMSDIST(1.96)`	0.9750
$\mathcal{N}(\mu,\sigma^2)$	`NORMDIST(x1,x2,x3,ind)` `x1` $= x$ `x2` $= \mu$ `x3` $= \sigma$	`NORMDIST(3.92,1.96,1,FALSE)` `NORMDIST(3.92,1.96,1,TRUE)`	0.0584 0.9750
$\mathcal{L}(\mu,\sigma^2)$	`LOGNORMDIST(x1,x2,x3)` `x1` $= x$ `x2` $= \mu$ `x3` $= \sigma$ output is $\Pr(\mathcal{L}(\mu,\sigma^2) \le x)$	`LOGNORMDIST(3.1424,-0.5,1)`	0.9500

Note: Set `ind` to `FALSE` for pdf and `TRUE` for df.

Table 2.3. *Some Excel functions for the computation of the inverse of the df $F_X^{-1}(\delta)$ of continuous random variable X*

		Example	
X	Excel function	Input	Output
$\mathcal{G}(\alpha,\beta)$	`GAMMAINV(x1,x2,x3)` `x1` $= \delta$ `x2` $= \alpha$ `x3` $= \beta$	`GAMMAINV(0.8,2,2)`	5.9886
$\mathcal{N}(0,1)$	`NORMSINV(x1)` `x1` $= \delta$	`NORMSINV(0.9)`	1.2816
$\mathcal{N}(\mu,\sigma^2)$	`NORMINV(x1,x2,x3)` `x1` $= \delta$ `x2` $= \mu$ `x3` $= \sigma$	`NORMINV(0.99,1.2,2.5)`	7.0159

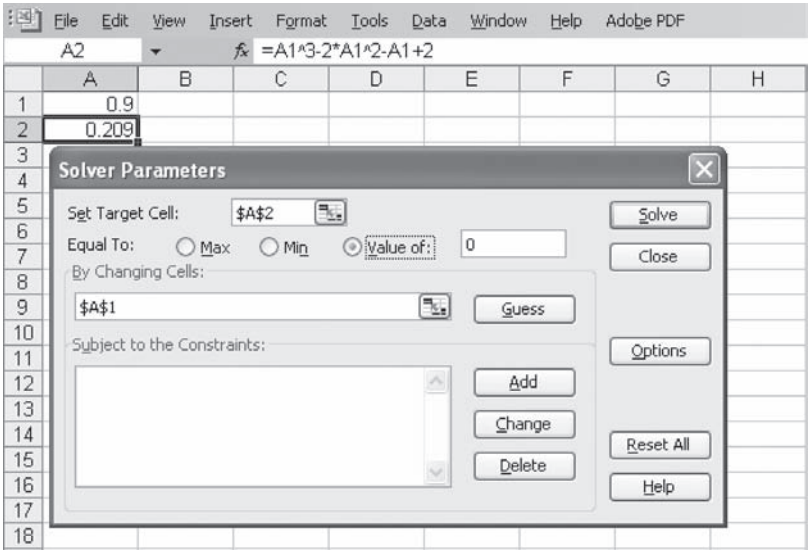

Figure 2.7 Use of the Excel Solver

the Options key. For the initial value of 0.9, the solution found is 1. Readers may try a different initial value in A1, such as 1.8, for solving other roots of equation (2.123).

2.7 Summary and conclusions

We have discussed the use of some standard continuous distributions for modeling claim-severity distributions. Some methods for creating new distributions are introduced, and these include the methods of transformation, mixing, and splicing. The right-hand tail properties of these distributions are important for modeling extreme losses, and measures such as quantiles, limiting ratios, and conditional tail expectations are often used. We derive formulas for calculating the expected amount paid in a loss event and in a payment event. Policies may be modified to give rise to different payment functions, and we consider policy modifications such as deductible, policy limit, and coinsurance. Methods of calculating the mean loss subject to policy modifications are presented. For the $(a, b, 0)$ class of claim-frequency distributions, we derive the effects of deductibles on the distribution of the claim frequency.

We have assumed that the ground-up loss distribution is independent of the policy modification. For example, we assume the same claim-severity distribution whether there is a deductible or not. This assumption may not

be valid as the deductible may affect the behavior of the insured and thus the claim-severity distribution. Such limitations of the model, however, are not addressed in this chapter.

Exercises

2.1 Prove that raw moments of all positive orders exist for the gamma distribution $\mathcal{G}(\alpha, \beta)$, and that only raw moments of order less than α exist for the Pareto distribution $\mathcal{P}(\alpha, \gamma)$.

2.2 The inverse exponential distribution X has the following pdf

$$f_X(x) = \frac{\theta e^{-\frac{\theta}{x}}}{x^2}, \qquad \text{for } \theta > 0.$$

(a) Find the df, sf, and hf of X.
(b) Derive the median and the mode of X.

2.3 The inverse Weibull distribution X has the following df

$$F_X(x) = e^{-\left(\frac{\theta}{x}\right)^{\tau}}, \qquad \text{for } \theta, \tau > 0.$$

(a) Find the sf, pdf, and hf of X.
(b) Derive the median and the mode of X.

2.4 Suppose X has the following hf

$$h_X(x) = \frac{1}{100 - x}, \qquad \text{for } 0 \leq x < 100.$$

(a) Find the sf, df, and pdf X.
(b) Calculate the mean and the variance of X.
(c) Calculate the median and the mode of X.
(d) Calculate the mean excess loss for $d = 10$.

2.5 Suppose X has the following pdf

$$f_X(x) = \frac{3x(20 - x)}{4{,}000}, \qquad \text{for } 0 < x < 20 \text{ and 0 otherwise.}$$

(a) Find the sf, df, and hf of X.
(b) Calculate the mean and the variance of X.
(c) Calculate the median and the mode of X.
(d) Calculate the mean excess loss for $d = 8$.

2.6 A Pareto distribution has mean 4 and variance 32. What is its median?

2.7 Let $X \sim \mathcal{E}(2)$ and $Y \sim \mathcal{P}(3, 4)$. Suppose Z is a mixture of X and Y with equal weights. Find the mean and the median of Z. [*Hint:* Use the Excel Solver to compute the median.]

2.8 If the pdf of X is $f_X(x) = 2xe^{-x^2}$, for $0 < x < \infty$ and 0 otherwise, what is the pdf of $Y = X^2$?

2.9 If $X \sim \mathcal{U}(-\pi/2, \pi/2)$ (see Appendix A.10.3), what is the pdf of $Y = \tan X$? [*Hint:* $d \tan^{-1} y/dy = 1/(1+y^2)$.]

2.10 There is a probability of 0.2 that the loss X is zero. Loss occurs with density proportional to $1 - x/20$ for $0 < x \leq 20$.
(a) What is the df of X?
(b) Find the mean and the variance of X.
(c) What is $x_{0.8}$?

2.11 Suppose X has a probability mass of 0.4 at $x = 0$ and has a density proportional to x^3 for $0 < x \leq 1$, and 0 elsewhere.
(a) What is the df of X?
(b) Find the mean and the variance of X.
(c) What is $x_{0.8}$?

2.12 Use the mgf to calculate the skewness (see equation (A.28)) of $\mathcal{G}(\alpha, \beta)$.

2.13 Construct a two-component spliced distribution, where the density of its first component, for $0 \leq x < 0.8$, is proportional to an exponential density with parameter $\lambda = 2$ and the density of its second component, for $0.8 \leq x < \infty$, is proportional to a gamma density with $\alpha = 2$ and $\beta = 0.5$. Apply the continuity restriction to the spliced distribution.

2.14 Suppose $X \mid \Lambda \sim \mathcal{E}(\Lambda)$ and Λ is uniformly distributed in the interval [1,5].
(a) Calculate the unconditional mean and variance of X using equations (2.56) and (2.57).
(b) Determine the unconditional pdf of X, and hence calculate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$. [*Hint:* You may use the result $\int_0^\infty \frac{e^{-ax} - e^{-bx}}{x} dx = \log\left(\frac{b}{a}\right)$.]

2.15 Suppose $X \sim \mathcal{G}(\alpha, \beta)$. If $\beta = 2$ and $\alpha - 1$ is distributed as $\mathcal{PN}(2.5)$, calculate the unconditional mean and variance of X.

2.16 Let $X \sim \mathcal{P}(4, 6)$, calculate $x_{0.9}$ and $\mathrm{CTE}_{0.9}$.

2.17 Let $X \sim \mathcal{E}(\lambda)$, and d be the amount of the deductible.
(a) What is the df of $X_L = (X - d)_+$?
(b) What is the df of $X_P = X - d \mid X > d$?
(c) Find the pdf of X_P and $\mathrm{E}(X_P)$.

2.18 Let $X \sim \mathcal{P}(2, 5)$, and $d = 1$ be the amount of the deductible.
(a) Calculate $\mathrm{E}(X_P)$.
(b) Calculate $\mathrm{E}(X \wedge d)$.
(c) Calculate $\mathrm{LER}(d)$.

2.19 X_i are iid $\mathcal{G}(\alpha, \beta)$, for $i = 1, 2, \cdots, n$. Using the mgf, show that $S = X_1 + X_2 + \cdots + X_n$ is distributed as $\mathcal{G}(n\alpha, \beta)$.

2.20 X is a mixture of exponential distributions $\mathcal{E}(\lambda_1)$ and $\mathcal{E}(\lambda_2)$ with the following pdf

$$f_X(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}, \quad 0 < x,\ 0 < p < 1.$$

(a) Derive the df of X.
(b) Calculate $\mathrm{E}\big[(X-d)_+\big]$.

2.21 The pdf of the claim severity X is $f_X(x) = 0.02x$, for $0 \le x \le 10$. An insurance policy has a deductible of $d = 4$, calculate the expected loss payment in a payment event.

2.22 Policy loss X is distributed as $\mathcal{U}(0, 100)$, with a deductible of $d = 20$ and a maximum covered loss of $u = 80$. Calculate $\mathrm{E}(X_P)$ and $\mathrm{Var}(X_P)$.

2.23 Policy loss X is distributed as $\mathcal{P}(5, 100)$, with a deductible of $d = 10$ and a maximum covered loss of $u = 50$. Calculate the expected loss in a loss event.

2.24 Policy loss X is distributed as $\mathcal{E}(0.01)$. If there is a deductible of $d = 8$, calculate the mean and the variance of X_L.

2.25 Suppose claim severity X is distributed as $\mathcal{E}(0.01)$. The policy has a deductible of $d = 20$, maximum covered loss of $u = 200$ and coinsurance factor of $c = 0.8$. Calculate the expected loss in a loss event. Subject to inflation adjustment of 5%, with no adjustments in policy factors, what is the expected payment per loss?

2.26 Policy loss X_T has a deductible of d and maximum covered loss of u. If the coinsurance factor is c, the rate of inflation is r and the loss in a loss event with inflation is $\tilde{X}_T$, prove that

$$\mathrm{E}(\tilde{X}_T^2) = c^2(1+r)^2 \left\{ \mathrm{E}[(X \wedge \tilde{u})^2] - \mathrm{E}[(X \wedge \tilde{d})^2] - 2\tilde{d}\left(\mathrm{E}[(X \wedge \tilde{u})] - \mathrm{E}[(X \wedge \tilde{d})]\right) \right\},$$

where $\tilde{u} = u/(1+r)$ and $\tilde{d} = d/(1+r)$.

2.27 Assume $X \mid \Lambda \sim \mathcal{PN}(\Lambda)$, and let the parameter Λ be distributed as $\mathcal{G}(\alpha, \beta)$. Show that $X \sim \mathcal{NB}(\alpha, 1/(1+\beta))$.

2.28 The Black–Scholes model of pricing a European call option on a nondividend-paying stock states that the price C of the European call option with exercise price x and time to maturity t is equal to the discounted expected payoff at maturity, i.e. we have

$$C = e^{-rt}\,\mathrm{E}\left[(\tilde{S} - x)_+\right],$$

where $\tilde{S}$ is the stock price at maturity and r is the riskfree rate of interest. Here $\tilde{S}$ is assumed to be lognormally distributed.

Specifically, we assume

$$\log \tilde{S} \sim \mathcal{N}\left(\log S + \left(r - \frac{\sigma_S^2}{2}\right)t, \sigma_S^2 t\right),$$

where S is the current stock price and σ_S is the volatility parameter of the stock. Using Theorem 2.2 and the above assumption, prove that

$$C = S\Phi(d_1) - xe^{-rt}\Phi(d_2),$$

where

$$d_1 = \frac{\log\left(\frac{S}{x}\right) + \left(r + \frac{\sigma_S^2}{2}\right)t}{\sigma_S\sqrt{t}},$$

and

$$d_2 = \frac{\log\left(\frac{S}{x}\right) + \left(r - \frac{\sigma_S^2}{2}\right)t}{\sigma_S\sqrt{t}}.$$

This is the celebrated Black–Scholes formula.

Questions adapted from SOA exams

2.29 Let $X \sim \mathcal{P}(\alpha, \gamma)$. Derive the pdf of $Y = \log(1 + X/\gamma)$.

2.30 The claim frequency of a policy with no deductible is distributed as $\mathcal{NB}(3, 0.2)$. The claim severity is distributed as $\mathcal{W}(0.3, 100)$. Determine the expected number of claim payments when the policy has a deductible of 20.

2.31 Suppose $X \sim \mathcal{E}(0.001)$. Calculate the coefficient of variation, i.e. the ratio of the standard deviation to the mean, of $(X - 2{,}000)_+$.

3

Aggregate-loss models

Having discussed models for claim frequency and claim severity separately, we now turn our attention to modeling the aggregate loss of a block of insurance policies. Much of the time we shall use the terms aggregate loss and aggregate claim interchangeably, although we recognize the difference between them as discussed in the last chapter. There are two major approaches in modeling aggregate loss: the individual risk model and the collective risk model. We shall begin with the individual risk model, in which we assume there are n independent loss prospects in the block. As a policy may or may not have a loss, the distribution of the *loss* variable in this model is of the mixed type. It consists of a probability mass at point zero and a continuous component of positive losses. Generally, exact distribution of the aggregate loss can only be obtained through the convolution method. The De Pril recursion, however, is a powerful technique to compute the exact distribution recursively when the block of policies follow a certain set-up.

On the other hand, the collective risk model treats the aggregate loss as having a compound distribution, with the primary distribution being the claim frequency and the secondary distribution being the claim severity. The Panjer recursion can be used to compute the distribution of the aggregate loss if the claim-frequency distribution belongs to the $(a, b, 0)$ class and the claim-severity distribution is *discretized* or approximated by a discrete distribution. In particular, the compound Poisson distribution has some useful properties in applications, and it can also be used as an approximation for the individual risk model. Finally, we consider the effects of a stop-loss reinsurance on the distribution of the aggregate loss.

Learning objectives

1 Individual risk model
2 Collective risk model

3 De Pril recursion
4 Compound processes for collective risks
5 Approximation methods for individual and collective risks
6 Stop-loss reinsurance

3.1 Individual risk and collective risk models

The aggregate loss of a block of insurance policies is the sum of all losses incurred in the block. It may be modeled by two approaches: the **individual risk model** and the **collective risk model**. In the individual risk model, we denote the number of policies in the block by n. We assume the loss of each policy, denoted by X_i, for $i = 1, \cdots, n$, to be *independently and identically distributed* as X. The aggregate loss of the block of policies, denoted by S, is then given by

$$S = X_1 + \cdots + X_n. \tag{3.1}$$

Thus, S is the sum of n iid random variables each distributed as X, where n is a fixed number. Note that typically most of the policies have zero loss, so that X_i is zero for these policies. In other words, X follows a mixed distribution with a probability mass at point zero. Although X_i in equation (3.1) are stated as being iid, the assumption of identical distribution is not necessary. The individual risk model will be discussed in Section 3.3.

The aggregate loss may also be computed using the collective risk model, in which the aggregate loss is assumed to follow a compound distribution. Let N be the number of losses in the block of policies, and X_i be the amount of the ith loss, for $i = 1, \cdots, N$. Then the aggregate loss S is given by

$$S = X_1 + \cdots + X_N. \tag{3.2}$$

The compound process as stated above was introduced in equation (1.60), in which $X_1, \cdots, X_N$ are assumed to be iid nonnegative integer-valued random variables representing claim frequencies. The purpose was then to create a new nonnegative integer-valued compound distribution to be used to model the claim-frequency distribution. In equation (3.2), however, $X_1, \cdots, X_N$ are assumed to be iid as the claim-severity random variable X, which is the secondary distribution of the compound distribution; while N is the claim-frequency random variable representing the primary distribution. Furthermore, N and X are assumed to be independent. The distributions of the claim frequency and claim severity have been discussed extensively in the last two chapters. Note that X_i are defined differently in equations (3.1) versus (3.2). In equation (3.1) X_i is the loss (which may be zero) of the ith policy, whereas in equation (3.2)

X_i is the loss amount in the ith loss event and is positive with probability 1. The collective risk model will be discussed in Section 3.4.[1]

There are some advantages in modeling the claim frequency and claim severity separately, and then combining them to determine the aggregate-loss distribution. For example, expansion of insurance business may have impacts on the claim frequency but not on the claim severity. On the other hand, cost control (or general cost increase) and innovation in technology may affect the claim severity with no effects on the claim frequency. Furthermore, the effects of coverage modifications may impact the claim-frequency distribution and claim-severity distribution differently. Modeling the two components separately would enable us to identify the effects of these modifications on the aggregate loss.

3.2 Individual risk model

The basic equation of the individual risk model stated in equation (3.1) specifies the aggregate loss S as the sum of n iid random variables each distributed as X. Thus, the mean and variance of S are given by

$$\mathrm{E}(S) = n\,\mathrm{E}(X) \quad \text{and} \quad \mathrm{Var}(S) = n\,\mathrm{Var}(X). \tag{3.3}$$

To compute the mean and variance of S, we need the mean and variance of X. Let the probability of a loss be θ and the probability of no loss be $1-\theta$. Furthermore, we assume that when there is a loss, the loss amount is Y, which is a positive continuous random variable with mean μ_Y and variance σ_Y^2. Thus, $X = Y$ with probability θ, and $X = 0$ with probability $1-\theta$. We can now write X as

$$X = IY, \tag{3.4}$$

where I is a Bernoulli random variable distributed independently of Y, so that

$$I = \begin{cases} 0, & \text{with probability } 1-\theta, \\ 1, & \text{with probability } \theta. \end{cases} \tag{3.5}$$

Thus, the mean of X is

$$\mathrm{E}(X) = \mathrm{E}(I)\mathrm{E}(Y) = \theta\mu_Y, \tag{3.6}$$

[1] In Chapter 1 we use X to denote the claim frequency. From now onwards, however, we shall use N to denote the claim frequency.

and its variance, using equation (A.118) in Appendix A.11, is

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(IY)\\ &= [\mathrm{E}(Y)]^2\,\mathrm{Var}(I) + \mathrm{E}(I^2)\mathrm{Var}(Y)\\ &= \mu_Y^2\,\theta(1-\theta) + \theta\,\sigma_Y^2. \end{aligned} \tag{3.7}$$

Equations (3.6) and (3.7) can be plugged into equation (3.3) to obtain the mean and variance of S.

Example 3.1 Assume there is a chance of 0.2 that there is a claim. When a claim occurs the loss is exponentially distributed with parameter $\lambda = 0.5$. Find the mean and variance of the claim distribution. Suppose there are 500 independent policies with this loss distribution, compute the mean and variance of their aggregate loss.

Solution The mean and variance of the loss in a loss event are

$$\mu_Y = \frac{1}{\lambda} = \frac{1}{0.5} = 2,$$

and

$$\sigma_Y^2 = \frac{1}{\lambda^2} = \frac{1}{(0.5)^2} = 4.$$

Thus, the mean and variance of the loss incurred by a random policy are

$$\mathrm{E}(X) = (0.2)(2) = 0.4,$$

and

$$\mathrm{Var}(X) = (2)^2(0.2)(1-0.2) + (0.2)(4) = 1.44.$$

The mean and variance of the aggregate loss are

$$\mathrm{E}(S) = (500)(0.4) = 200,$$

and

$$\mathrm{Var}(S) = (500)(1.44) = 720.$$ □

3.2.1 Exact distribution using convolution

The general technique to compute the exact distribution of the sum of independent random variables is by convolution. We discussed the convolution for sums of discrete nonnegative random variables in Section 1.5.1. We now consider the case of convolution of the continuous and mixed-type random variables.

Let $X_1, \cdots, X_n$ be n independently distributed nonnegative continuous random variables with pdf $f_1(\cdot), \cdots, f_n(\cdot)$, respectively. Here we have relaxed the assumption of identically distributed losses, as this is not required for the computation of the convolution. We first consider the distribution of $X_1 + X_2$, the pdf of which is given by the 2-fold convolution

$$f^{*2}(x) = f_{X_1+X_2}(x) = \int_0^x f_1(x-y) f_2(y)\, dy = \int_0^x f_2(x-y) f_1(y)\, dy. \quad (3.8)$$

The pdf of $X_1 + \cdots + X_n$ can be calculated recursively. Suppose the pdf of $X_1 + \cdots + X_{n-1}$ is given by the $(n-1)$-fold convolution $f^{*(n-1)}(x)$, then the pdf of $X_1 + \cdots + X_n$ is the n-fold convolution given by

$$\begin{aligned} f^{*n}(x) &= f_{X_1+\cdots+X_n}(x) = \int_0^x f^{*(n-1)}(x-y) f_n(y)\, dy \\ &= \int_0^x f_n(x-y) f^{*(n-1)}(y)\, dy. \end{aligned} \quad (3.9)$$

It is clear that the above formulas do not assume identically distributed components X_i.

Now we consider the case where X_i are mixed-type random variables, which is typical of an individual risk model. We assume that the pf–pdf of X_i is given by

$$f_{X_i}(x) = \begin{cases} 1-\theta_i, & \text{for } x = 0, \\ \theta_i f_{Y_i}(x), & \text{for } x > 0, \end{cases} \quad (3.10)$$

in which $f_{Y_i}(\cdot)$ are well-defined pdf of some positive continuous random variables. The df of $X_1 + X_2$ is given by the 2-fold convolution in the Stieltjes-integral form, i.e.[2]

$$F^{*2}(x) = F_{X_1+X_2}(x) = \int_0^x F_{X_1}(x-y)\, dF_{X_2}(y) = \int_0^x F_{X_2}(x-y)\, dF_{X_1}(y). \quad (3.11)$$

The df of $X_1 + \cdots + X_n$ can be calculated recursively. Suppose the df of $X_1 + \cdots + X_{n-1}$ is given by the $(n-1)$-fold convolution $F^{*(n-1)}(x)$, then the df of $X_1 + \cdots + X_n$ is the n-fold convolution

$$\begin{aligned} F^{*n}(x) &= F_{X_1+\cdots+X_n}(x) = \int_0^x F^{*(n-1)}(x-y)\, dF_{X_n}(y) \\ &= \int_0^x F_{X_n}(x-y)\, dF^{*(n-1)}(y). \end{aligned} \quad (3.12)$$

[2] See (A.2) and (A.3) in the Appendix for the use of Stieltjes integral.

For the pf–pdf given in equation (3.10), we have[3]

$$F^{*n}(x) = \int_0^x F^{*(n-1)}(x-y) f_{X_n}(y)\, dy + (1-\theta_n) F^{*(n-1)}(x). \tag{3.13}$$

In particular, if $X_1, \cdots, X_n$ are iid, with $\theta_i = \theta$ and $f_{Y_i}(x) = f_Y(x)$, for $i = 1, \cdots, n$, then

$$F^{*n}(x) = \theta \int_0^x F^{*(n-1)}(x-y) f_Y(y)\, dy + (1-\theta) F^{*(n-1)}(x). \tag{3.14}$$

While the expressions of the pf–pdf of the sums of X_i can be written in convolution form, the computations of the integrals are usually quite complex. To implement the convolution method in practice, we may first *discretize* the continuous distribution, and then apply convolution to the discrete distribution on the computer. Assume the discretized approximate distribution of X_i has the pf $f_i(x)$ for $x = 0, \cdots, m$ and $i = 1, \cdots, n$.[4] Then the 2-fold convolution $X_1 + X_2$ is

$$f^{*2}(x) = \sum_{y=0}^{x} f_1(x-y) f_2(y) = \sum_{y=0}^{x} f_2(x-y) f_1(y), \quad \text{for } x = 0, \cdots, 2m, \tag{3.15}$$

and the n-fold convolution is given by

$$\begin{aligned} f^{*n}(x) &= \sum_{y=0}^{x} f^{*(n-1)}(x-y) f_n(y) \\ &= \sum_{y=0}^{x} f_n(x-y) f^{*(n-1)}(y), \qquad \text{for } x = 0, \cdots, nm. \end{aligned} \tag{3.16}$$

In the equations above, some of the probabilities in the summation are zero. For example, in equation (3.15) $f_1(\cdot)$ and $f_2(\cdot)$ are zero for $y > m$ or $x - y > m$.

Example 3.2 For the block of insurance policies defined in Example 3.1, approximate the loss distribution by a suitable discrete distribution. Compute the df $F_S(s)$ of the aggregate loss of the portfolio for s from 110 through 300 in steps of 10, based on the discretized distribution.

[3] Compare this equation with equation (A.15) in Appendix A.3.

[4] We assume that the maximum loss of a policy is m, which is a realistic assumption in practice, as insurance policies may set maximum loss limits. We shall not distinguish between the continuous claim-severity distribution and its discretized version by different notations. The discrete distribution may also be an empirical distribution obtained from claim data.

Table 3.1. *Discretized probabilities*

x	$f_X(x)$	x	$f_X(x)$
0	0.8442	6	0.0050
1	0.0613	7	0.0031
2	0.0372	8	0.0019
3	0.0225	9	0.0011
4	0.0137	10	0.0017
5	0.0083		

Solution We approximate the exponential loss distribution by a discrete distribution taking values $0, 1, \cdots, 10$. As the df of $\mathcal{E}(\lambda)$ is $F_X(x) = 1 - \exp(-\lambda x)$, we approximate the pf by

$$f_X(x) = (0.2)\left\{\exp\left[-\lambda(x-0.5)\right] - \exp\left[-\lambda(x+0.5)\right]\right\}, \quad \text{for } x = 1, \cdots, 9,$$

with

$$f_X(0) = 0.8 + (0.2)[1 - \exp(-0.5\lambda)],$$

and

$$f_X(10) = (0.2)\,\exp(-9.5\lambda).$$

The discretized approximate pf of the loss is given in Table 2.1.

Note that the mean and variance of the loss random variable, as computed in Example 3.1, are 0.4 and 1.44, respectively. In comparison, the mean and variance of the discretized approximate loss distribution given in Table 3.1 are 0.3933 and 1.3972, respectively.

Using the convolution method, the df of the aggregate loss S for selected values of s is given in Table 3.2. □

The computation of the convolution is very intensive if n is large. In the next section we present an efficient exact solution due to De Pril (1985, 1986) using a recursive method when the block of insurance policies follow a specific set-up.

3.2.2 Exact distribution using the De Pril recursion

The De Pril recursion provides a method to compute the aggregate-loss distribution of the individual risk model. This method applies to blocks of insurance policies which can be stratified by the sum assured and the claim probability. Specifically, we assume that the portfolio of insurance policies consists of risks with J different claim probabilities θ_j, for $j = 1, \cdots, J$, and each policy has an insured amount of $i = 1, \cdots, I$ benefit units, which is a

Table 3.2. *The df of S by convolution*

s	$F_S(s)$	s	$F_S(s)$
110	0.0001	210	0.7074
120	0.0008	220	0.8181
130	0.0035	230	0.8968
140	0.0121	240	0.9465
150	0.0345	250	0.9746
160	0.0810	260	0.9890
170	0.1613	270	0.9956
180	0.2772	280	0.9984
190	0.4194	290	0.9994
200	0.5697	300	0.9998

suitably standardized monetary amount. We assume there are n_{ij} independent policies with insured amount of i benefit units and claim probability of θ_j, for $i = 1, \cdots, I$ and $j = 1, \cdots, J$. If we denote the aggregate loss of the block by S and the pf of S by $f_S(s)$, $s = 0, \cdots, n$, where $n = \sum_{i=1}^{I} \sum_{j=1}^{J} i\, n_{ij}$, then $f_S(s)$ can be computed using the following theorem.[5]

Theorem 3.1 *The pf of S, $f_S(s)$, satisfies the following equation, known as the De Pril (1985, 1986) recursion formula*

$$f_S(s) = \frac{1}{s} \sum_{i=1}^{min\{s,I\}} \sum_{k=1}^{[\frac{s}{i}]} f_S(s - ik) h(i,k), \qquad for\ s = 1, \cdots, n, \tag{3.17}$$

where $[x]$ denotes the integer part x

$$h(i,k) = \begin{cases} i(-1)^{k-1} \sum_{j=1}^{J} n_{ij} \left(\frac{\theta_j}{1-\theta_j} \right)^k, & for\ i = 1, \cdots, I, \\ 0, & otherwise, \end{cases} \tag{3.18}$$

and the recursion has the following starting value

$$f_S(0) = \prod_{i=1}^{I} \prod_{j=1}^{J} (1 - \theta_j)^{n_{ij}}. \tag{3.19}$$

Proof See Dickson (2005, Section 5.3) for the proof. □

The De Pril recursion formula is computationally intensive for large values of s and I. However, as θ_j is usually small, $h(i,k)$ is usually close to zero for large k.

[5] Note that n is the maximum loss of the block of policies.

An approximation can thus be obtained by choosing a truncation parameter K in the summation over k in equation (3.17).[6] We denote the resulting pf by $f_S^K(s)$, which can be computed as

$$f_S^K(s) = \frac{1}{s} \sum_{i=1}^{\min\{x,I\}} \sum_{k=1}^{\min\{K,\lfloor \frac{s}{i} \rfloor\}} f_S^K(s - ik)h(i,k), \tag{3.20}$$

with $f_S^K(0) = f_S(0)$.

3.2.3 Approximations of the individual risk model

As the aggregate loss S in equation (3.1) is the sum of n iid random variables, its distribution is approximately normal when n is large, by virtue of the Central Limit Theorem. The (exact) mean and variance of S are given in equations (3.3), (3.6), and (3.7), which can be used to compute the approximate distribution of S. Thus

$$\begin{aligned} \Pr(S \le s) &= \Pr\left(\frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}} \le \frac{s - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}} \right) \\ &\simeq \Pr\left(Z \le \frac{s - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}} \right) \\ &= \Phi\left(\frac{s - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}} \right). \end{aligned} \tag{3.21}$$

The normal approximation holds even when the individual risks are not identically distributed. As in the case of the De Pril set-up, the claim probability varies. The mean and variance of S, however, can be computed as follows under the assumption of independent risks

$$\mathrm{E}(S) = \sum_{i=1}^{I} \sum_{j=1}^{J} i\, n_{ij}\, \theta_j, \tag{3.22}$$

and

$$\mathrm{Var}(S) = \sum_{i=1}^{I} \sum_{j=1}^{J} i^2\, n_{ij}\, \theta_j (1 - \theta_j). \tag{3.23}$$

Equation (3.21) can then be used to compute the approximate df of the aggregate loss.

[6] Dickson (2005) suggested that $K = 4$ is usually sufficient for a good approximation.

Example 3.3 A portfolio of insurance policies has benefits of 1, 2, or 3 thousand dollars. The claim probabilities may be 0.0015, 0.0024, 0.0031, or 0.0088. The numbers of policies with claim amount i thousands and claim probability θ_j are summarized in Table 3.3. Calculate the distribution of the aggregate claim using (a) the De Pril recursion with the exact formula, (b) the De Pril recursion with truncation at $K = 2$, and (c) the normal approximation.

Table 3.3. *Numbers of policies for Example 3.3*

	Claim probability group			
Benefit units i ($,000)	$j = 1$	$j = 2$	$j = 3$	$j = 4$
1	30	40	50	60
2	40	50	70	80
3	70	60	80	90
Claim probability θ_j	0.0015	0.0024	0.0031	0.0088

Solution Using equations (3.22) and (3.23), the mean and variance of the aggregate loss S are computed as 6.8930 and 16.7204, respectively. These values are plugged into equation (3.21) to obtain the normal approximation (with the usual continuity correction). Table 3.4 gives the results of the exact De Pril method, the truncated De Pril method, and the normal approximation. The numbers in the table are the df at selected values of aggregate loss s.

Table 3.4. *Aggregate-loss df at selected values*

	Method		
Aggregate loss s	De Pril (exact)	De Pril (truncated, $K = 2$)	Normal approximation
0	0.03978	0.03978	0.05897
5	0.40431	0.40431	0.36668
10	0.81672	0.81670	0.81114
15	0.96899	0.96895	0.98235
20	0.99674	0.99668	0.99956
25	0.99977	0.99971	1.00000

It can be seen that the truncated De Pril approximation works very well, even for a small truncation value of $K = 2$. On the other hand, the normal approximation is clearly inferior to the truncated De Pril approximation. □

Although the computation of the exact De Pril recursion is very efficient, its application depends on the assumption that the portfolio of insurance policies can be stratified according to the particular set-up.

Example 3.4 For the portfolio of policies in Example 3.1, calculate the df of the aggregate loss at $s = 180$ and 230 using normal approximation. Compare the answers against the answers computed by convolution in Example 3.2.

Solution From Example 3.1, the mean and standard deviation of the aggregate loss S are, respectively, 200 and $\sqrt{720} = 26.8328$. Using normal approximation, we have

$$\Pr(S \leq 180) \simeq \Pr\left(Z \leq \frac{180.5 - 200}{26.8328}\right) = \Phi(-0.7267) = 0.2337,$$

and

$$\Pr(S \leq 230) \simeq \Pr\left(Z \leq \frac{230.5 - 200}{26.8328}\right) = \Phi(1.1367) = 0.8722.$$

The corresponding results obtained from the convolution method in Example 3.2 are 0.2772 and 0.8968. The differences in the answers are due to (a) discretization of the exponential loss in the convolution method and (b) the use of normal approximation assuming sufficiently large block of policies. □

3.3 Collective risk model

As stated in equation (3.2), the collective risk model specifies the aggregate loss S as the sum of N losses, which are independently and identically distributed as the claim-severity variable X. Thus, S follows a compound distribution, with N being the primary distribution and X the secondary distribution. We have discussed compound distributions in Section 1.5, in which both the primary and secondary distributions are nonnegative discrete random variables. Some properties of the compound distributions derived in Section 1.5 are applicable to the collective risk model in which the secondary distribution is continuous. We shall review some of these properties, and apply the results in Chapter 1 to study the aggregate-loss distribution. Both recursive and approximate methods will be considered for the computation of the aggregate-loss distribution.

3.3.1 Properties of compound distributions

Firstly, as proved in Theorem 1.4, the mgf $M_S(t)$ of the aggregate loss S is given by

$$M_S(t) = M_N\left[\log M_X(t)\right], \tag{3.24}$$

where $M_N(t)$ and $M_X(t)$ are, respectively, the mgf of N and X. Furthermore, if the claim-severity takes nonnegative discrete values, S is also nonnegative and discrete, and its pgf is

$$P_S(t) = P_N[P_X(t)], \tag{3.25}$$

where $P_N(t)$ and $P_X(t)$ are, respectively, the pgf of N and X. Secondly, Theorem 1.6 also applies to the aggregate-loss distribution. Thus, the mean and variance of S are

$$\mathrm{E}(S) = \mathrm{E}(N)\mathrm{E}(X), \tag{3.26}$$

and

$$\mathrm{Var}(S) = \mathrm{E}(N)\mathrm{Var}(X) + \mathrm{Var}(N)[\mathrm{E}(X)]^2. \tag{3.27}$$

These results hold whether X is continuous or discrete. Thirdly, if S_i has a compound Poisson distribution with claim-severity distribution X_i, which may be continuous or discrete, for $i = 1, \cdots, n$, then $S = S_1 + \cdots + S_n$ has also a compound Poisson distribution. As shown in Theorem 1.7, the Poisson parameter λ of S is the sum of the Poisson parameters $\lambda_1, \cdots, \lambda_n$ of $S_1, \cdots, S_n$, respectively. In addition, the secondary distribution X of S is the mixture distribution of $X_1, \cdots, X_n$, where the weight of X_i is λ_i/λ.

When X is continuous, we extend the result in equation (1.65) to obtain the pdf of S as

$$\begin{aligned} f_S(s) &= \sum_{n=1}^{\infty} f_{X_1+\cdots+X_n\,|\,n}(s) f_N(n) \\ &= \sum_{n=1}^{\infty} f^{*n}(s) f_N(n), \end{aligned} \tag{3.28}$$

where $f^{*n}(\cdot)$ is the n-fold convolution given in equation (3.9). Thus, the exact pdf of S is a weighted sum of convolutions, and the computation is highly complex.

There are some special cases for which the compound distribution can be analytically derived. Theorem 3.2 provides an example.

Theorem 3.2 *For the compound distribution specified in equation (3.2), assume $X_1, \cdots, X_N$ are iid $\mathcal{E}(\lambda)$ and $N \sim \mathcal{GM}(\theta)$. Then the compound distribution S is a mixed distribution with a probability mass of θ at 0 and a continuous component of $\mathcal{E}(\lambda\theta)$ weighted by $1-\theta$.*

Proof From Section 2.2, the mgf of $X \sim \mathcal{E}(\lambda)$ is

$$M_X(t) = \frac{\lambda}{\lambda - t},$$

and from Section 1.3, the mgf of $N \sim \mathcal{GM}(\theta)$ is

$$M_N(t) = \frac{\theta}{1 - (1-\theta)e^t}.$$

Thus, using equation (3.24), we conclude that the mgf of S is

$$\begin{aligned} M_S(t) &= M_N\left[\log M_X(t)\right] \\ &= \frac{\theta}{1 - (1-\theta)\left(\dfrac{\lambda}{\lambda - t}\right)} \\ &= \frac{\theta(\lambda - t)}{\lambda\theta - t} \\ &= \frac{\theta(\lambda\theta - t) + (1-\theta)\lambda\theta}{\lambda\theta - t} \\ &= \theta + (1-\theta)\left(\frac{\lambda\theta}{\lambda\theta - t}\right). \end{aligned}$$

Thus, $M_S(t)$ is the weighted average of 1 and $\lambda\theta/(\lambda\theta - t)$, which are the mgf of a degenerate distribution at 0 and the mgf of $\mathcal{E}(\lambda\theta)$, respectively, with the weights being θ for the degenerate distribution and $1-\theta$ for $\mathcal{E}(\lambda\theta)$. Hence, the aggregate loss is a mixed distribution, with a probability mass of θ at 0 and a continuous component of $\mathcal{E}(\lambda\theta)$ weighted by $1-\theta$. □

Analytic solutions of compound distributions are not common. In what follows we discuss some approximations making use of recursions, as well as normal approximations assuming sufficiently large samples.

3.3.2 Panjer recursion

Theorem 1.5 provides the efficient Panjer recursion method to compute the exact distribution of a compound process which satisfies the conditions that (a) the primary distribution belongs to the $(a, b, 0)$ class and (b) the secondary distribution is discrete and nonnegative integer valued. Thus, if a continuous claim-severity distribution can be suitably discretized, and the primary distribution belongs to the $(a, b, 0)$ class, we can use the Panjer approximation to compute the distribution of the aggregate loss.

Example 3.5 Consider the block of insurance policies in Examples 3.1 and 3.2. Approximate the distribution of the aggregate loss using the collective risk model. You may assume a suitable discretization of the exponential loss, and that the primary distribution is Poisson.

Solution Unlike the case of Example 3.2, the loss variable X in the collective risk model is the loss in a loss event (not the loss of a random policy). While the discretized distribution should take positive values, we use the following formulas to compute the pf of the discretized distribution[7]

$$f_X(x) = \exp\left[-\lambda(x-0.5)\right] - \exp\left[-\lambda(x+0.5)\right], \qquad \text{for } x = 1, \cdots, 9,$$

with

$$f_X(0) = 1 - \exp\left(-0.5\lambda\right),$$

and

$$f_X(10) = \exp\left(-9.5\lambda\right).$$

The pf is summarized in Table 3.5.

Table 3.5. *Discretized probabilities*

x	$f_X(x)$	x	$f_X(x)$
0	0.2212	6	0.0252
1	0.3064	7	0.0153
2	0.1859	8	0.0093
3	0.1127	9	0.0056
4	0.0684	10	0.0087
5	0.0415		

As the probability of a claim is 0.2, and there are 500 policies in the portfolio, the expected number of claims is $(0.2)(500) = 100$. Thus, we assume the primary distribution to be Poisson with mean 100. From Table 1.2, the parameters of the $(a, b, 0)$ class are: $a = 0$, $b = 100$, and $f_X(0) = 0.2212$. Thus, from equation (1.70), we have

$$f_S(0) = \exp\left[(100)(0.2212 - 1)\right] = \exp(-77.88).$$

The Panjer recursion in equation (1.74) becomes

$$f_S(s) = \sum_{x=1}^{s} \frac{100x}{s} f_X(x) f_S(s-x).$$

[7] It may seem contradictory to have a nonzero probability of zero loss when X is defined as the loss in a loss event. We should treat this as an approximation of the continuous loss distribution rather than an assumption of the model.

As $f_X(x) = 0$ for $x > 10$, the above equation can be written as

$$f_S(s) = \sum_{x=1}^{\min\{s,10\}} \frac{100x}{s} f_X(x) f_S(s-x).$$

The df $F_S(s)$ of the aggregate loss of the portfolio for s from 110 through 300 in steps of 10 is presented in Table 3.6. □

Table 3.6. *The df of S by the Panjer recursion*

s	$F_S(s)$	s	$F_S(s)$
110	0.0003	210	0.6997
120	0.0013	220	0.8070
130	0.0052	230	0.8856
140	0.0164	240	0.9375
150	0.0426	250	0.9684
160	0.0932	260	0.9852
170	0.1753	270	0.9936
180	0.2893	280	0.9975
190	0.4257	290	0.9991
200	0.5684	300	0.9997

Note that the results in Tables 3.2 and 3.6 are quite similar. The results in Table 3.2 are exact (computed from the convolution) for the individual risk model given the discretized loss distribution. On the other hand, the results in Table 3.6 are also exact (computed from the Panjer recursion) for the collective risk model, given the assumptions of Poisson primary distribution and the discretized loss distribution. These examples illustrate that the individual risk model and the collective risk model may give rise to similar results when their assumptions concerning the claim frequency and claim severity are compatible.

3.3.3 Approximations of the collective risk model

Under the individual risk model, if the number of policies n is large, by virtue of the Central Limit Theorem the aggregate loss is approximately normally distributed. In the case of the collective risk model, the situation is more complex, as the number of summation terms N in the aggregate loss S is random. However, if the mean number of claims is large, we may expect the normal approximation to work. Thus, using the mean and variance formulas of

S in equations (3.26) and (3.27), we may approximate the df of S by

$$\Pr(S \le s) = \Pr\left(\frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}} \le \frac{s - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}}\right) \simeq \Phi\left(\frac{s - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}}\right). \tag{3.29}$$

Example 3.6 Assume the aggregate loss S in a collective risk model has a primary distribution of $\mathcal{PN}(100)$ and a secondary distribution of $\mathcal{E}(0.5)$. Approximate the distribution of S using the normal distribution. Compute the df of the aggregate loss for $s = 180$ and 230.

Solution From equation (3.26) we have

$$\mathrm{E}(S) = \mathrm{E}(N)\mathrm{E}(X) = (100)\left(\frac{1}{0.5}\right) = 200,$$

and

$$\begin{aligned}\mathrm{Var}(S) &= \mathrm{E}(N)\mathrm{Var}(X) + \mathrm{Var}(N)\,[\mathrm{E}(X)]^2 \\ &= (100)\left[\frac{1}{(0.5)^2}\right] + (100)\left(\frac{1}{0.5}\right)^2 \\ &= 800.\end{aligned}$$

Thus, we approximate the distribution of S by $\mathcal{N}(200, 800)$. Note that this model has the same mean as that of the portfolio of policies in the individual risk model in Example 3.4, while its variance is larger than that of the individual risk model, which is 720. Also, this model has the same expected number of claims and the same claim-severity distribution in a loss event as the individual risk model. Using the normal approximation the required probabilities are

$$\Pr(S \le 180) \simeq \Pr\left(Z \le \frac{180.5 - 200}{\sqrt{800}}\right) = \Phi(-0.6894) = 0.2453,$$

and

$$\Pr(S \le 230) \simeq \Pr\left(Z \le \frac{230.5 - 200}{\sqrt{800}}\right) = \Phi(1.0783) = 0.8596. \qquad \square$$

The normal approximation makes use of only the first two moments of S. Higher-order moments can be derived for S and may help improve the approximation. Dickson (2005, Section 4.3) shows that the compound Poisson process is skewed to the right (i.e. positively skewed) regardless of the shape of the secondary distribution. Improvements in the approximation may be achieved using more versatile approximations, such as the translated gamma distribution, which can capture the skewness of the distribution. An example can be found in Dickson (2005, Section 4.8.2).

3.3.4 Compound Poisson distribution and individual risk model

In Example 3.5 we compute the distribution of the aggregate loss using the collective risk model by way of the Panjer recursion, whereas the parameter of the primary distribution is selected based on the individual risk model and the secondary distribution is a discretized version of the random loss in a loss event. We now provide a more formal justification of this approximation, as well as some extensions.

We consider the individual risk model

$$S = X_1 + \cdots + X_n, \tag{3.30}$$

with n policies. The policy losses are independently distributed with pf–pdf $f_{X_i}(\cdot)$, which are not necessarily identical. Note that X_i can be regarded as having a compound Bernoulli distribution. The primary distribution N_i of X_i takes value 0 with probability $1 - \theta_i$ and value 1 with probability θ_i, and the secondary distribution has a pdf $f_{Y_i}(\cdot)$, so that

$$f_{X_i}(x) = \begin{cases} 1 - \theta_i, & \text{for } x = 0, \\ \theta_i f_{Y_i}(x), & \text{for } x > 0, \end{cases} \tag{3.31}$$

Thus, S is the sum of n compound Bernoulli distributions. This distribution is in general intractable and convolution has to be applied to compute the exact distribution. From Theorem 1.7, however, we know that the sum of n compound Poisson distributions has also a compound Poisson distribution. As the Panjer recursion provides an efficient method to compute the compound Poisson distribution, an approximate method to compute the distribution of S is available by approximating the compound Bernouli distributions X_i by compound Poisson distributions $\tilde{X}_i$. Thus, we define $\tilde{X}_i$ as a compound Poisson distribution, where the Poisson parameter is $\lambda_i = \theta_i$ and the secondary distribution has a pdf $f_{Y_i}(\cdot)$. Thus, the means of the primary distributions of X_i and $\tilde{X}_i$ are the same, and they have the same secondary distributions. We now define

$$\tilde{S} = \tilde{X}_1 + \cdots + \tilde{X}_n, \tag{3.32}$$

which, by virtue of Theorem 1.7, has a compound Poisson distribution with Poisson parameter

$$\lambda = \lambda_1 + \cdots + \lambda_n = \theta_1 + \cdots + \theta_n. \tag{3.33}$$

The pdf of the secondary distribution of $\tilde{S}$ is

$$f_{\tilde{X}}(x) = \frac{1}{\lambda} \sum_{i=1}^{n} \lambda_i f_{Y_i}(x). \tag{3.34}$$

With a suitable discretization of $f_{\tilde{X}}(\cdot)$, the distribution of $\tilde{S}$ can be computed using Panjer's recursion, which provides an approximation of the distribution of S.

Finally, if X_i are identically distributed with $\theta_i = \theta$ and $f_{Y_i}(\cdot) = f_Y(\cdot)$, for $i = 1, \cdots, n$, then we have

$$\lambda = n\theta \quad \text{and} \quad f_{\tilde{X}}(\cdot) = f_Y(\cdot). \tag{3.35}$$

This approximation was illustrated in Example 3.5.

3.4 Coverage modifications and stop-loss reinsurance

In Section 2.5 we discuss the effects of coverage modifications on the distributions of the claim frequency and claim severity. Depending on the model used for the aggregate-loss distribution, we may derive the effects of coverage modifications on aggregate loss through their effects on the claim frequency and severity.

We first consider the effects of a deductible of amount d. For the individual risk model, the number of policies n remains unchanged, while the policy loss X_i becomes the loss amount of the claim after the deductible, which we shall denote by $\tilde{X}_i$. Thus, the pf–pdf of $\tilde{X}_i$ is[8]

$$f_{\tilde{X}_i}(x) = \begin{cases} 1 - \theta_i + \theta_i F_{Y_i}(d), & \text{for } x = 0, \\ \theta_i f_{Y_i}(x + d), & \text{for } x > 0. \end{cases} \tag{3.36}$$

For the collective risk model the primary distribution of the compound distribution, i.e. the distribution of the claim frequency, is now modified, as discussed in Section 2.5.5. Also, the secondary distribution is that of the claim after the deductible, i.e. $\tilde{X}$ with pdf given by

$$f_{\tilde{X}}(x) = \frac{f_X(x + d)}{1 - F_X(d)}, \quad \text{for } x > 0. \tag{3.37}$$

Second, we consider the effects of a policy limit u. For the individual risk model, the number of policies again remains unchanged, while the claim-severity distribution is now capped at u. If we denote the modified claim-severity

[8] The definition of Y_i is in equation (3.31).

distribution by $\tilde{X}_i$, then the pf–pdf of $\tilde{X}_i$ is given by

$$f_{\tilde{X}_i}(x) = \begin{cases} 1-\theta_i, & \text{for } x = 0, \\ \theta_i f_{Y_i}(x), & \text{for } 0 < x < u, \\ \theta_i \left[1 - F_{Y_i}(u)\right], & \text{for } x = u, \\ 0, & \text{otherwise.} \end{cases} \tag{3.38}$$

For the collective risk model, the primary distribution is not affected, while the secondary distribution $\tilde{X}$ has a pf–pdf given by

$$f_{\tilde{X}}(x) = \begin{cases} f_X(x), & \text{for } 0 < x < u, \\ 1 - F_X(u), & \text{for } x = u, \\ 0, & \text{otherwise.} \end{cases} \tag{3.39}$$

Insurance companies may purchase reinsurance coverage for a portfolio of policies they own. The coverage may protect the insurer from aggregate loss S exceeding an amount d, called **stop-loss reinsurance**. From the reinsurer's point of view this is a policy with a deductible of amount d. Thus, the loss to the reinsurer is $(S-d)_+$. As shown in equation (2.76), we have

$$\mathrm{E}\left[(S-d)_+\right] = \int_d^\infty \left[1 - F_S(s)\right] ds, \tag{3.40}$$

which can be computed as

$$\mathrm{E}\left[(S-d)_+\right] = \int_d^\infty (s-d) f_S(s)\, ds \tag{3.41}$$

when S is continuous, or

$$\mathrm{E}\left[(S-d)_+\right] = \sum_{s>d} (s-d) f_S(s) \tag{3.42}$$

when S is discrete.

If S takes on integer values only, but d is not an integer, then interpolation is required to compute the expected loss. Given a stop-loss amount d, let $\underline{d}$ be the largest integer and $\overline{d}$ be the smallest integer such that $\underline{d} \le d < \overline{d}$ and

$\underline{d} + 1 = \overline{d}$. Then[9]

$$\begin{aligned}
\mathrm{E}[(S-d)_+] &= \int_d^\infty [1 - F_S(s)]\, ds \\
&= \int_d^{\overline{d}} [1 - F_S(s)]\, ds + \int_{\overline{d}}^\infty [1 - F_S(s)]\, ds \\
&= \left(\overline{d} - d\right)[1 - F_S(\underline{d})] + \mathrm{E}\left[(S-\overline{d})_+\right]. \qquad (3.43)
\end{aligned}$$

Note that if we consider $d = \underline{d}$ in equation (3.43), we have, after re-arranging terms

$$\mathrm{E}[(S-(d+1))_+] = \mathrm{E}[(S-d)_+] - [1 - F_S(d)], \qquad (3.44)$$

for any integer $d > 0$. This is a convenient recursive formula to use for computing the expected loss at integer-valued deductibles.

From equation (3.44) we have

$$1 - F_S(\underline{d}) = \mathrm{E}[(S-\underline{d})_+] - \mathrm{E}\left[(S-\overline{d})_+\right], \qquad (3.45)$$

so that equation (3.43) can be written as

$$\begin{aligned}
\mathrm{E}[(S-d)_+] &= \left(\overline{d} - d\right)\left\{\mathrm{E}[(S-\underline{d})_+] - \mathrm{E}\left[(S-\overline{d})_+\right]\right\} + \mathrm{E}\left[(S-\overline{d})_+\right] \\
&= \left(\overline{d} - d\right)\mathrm{E}[(S-\underline{d})_+] + \left(d - \underline{d}\right)\mathrm{E}\left[(S-\overline{d})_+\right]. \qquad (3.46)
\end{aligned}$$

Thus, $\mathrm{E}[(S-d)_+]$ is obtained by interpolating $\mathrm{E}[(S-\underline{d})_+]$ and $\mathrm{E}\left[(S-\overline{d})_+\right]$.

Example 3.7 Assume the aggregate loss S follows a compound Poisson distribution with Poisson parameter λ, and the claim-severity distribution follows a Pareto distribution with parameters α and γ, namely $\mathcal{P}(\alpha, \gamma)$, where $\alpha > 2$. Compute the mean and the variance of S. If policies are modified with a deductible of d, compute the mean and the variance of the aggregate loss $\tilde{S}$.

Solution Without the deductible, the claim-frequency distribution N has mean and variance λ, and the mean and variance of the claim severity X are, from equation (2.40)

$$\mathrm{E}(X) = \frac{\gamma}{\alpha - 1} \quad \text{and} \quad \mathrm{Var}(X) = \frac{\alpha\gamma^2}{(\alpha-1)^2(\alpha-2)}.$$

[9] Note that $F_S(s) = F_S(\underline{d})$ for $\underline{d} \le s < \overline{d}$.

Hence, from equations (3.26) and (3.27), we have

$$\mathrm{E}(S) = \mathrm{E}(N)\mathrm{E}(X) = \frac{\lambda\gamma}{\alpha - 1},$$

and

$$\begin{aligned}
\mathrm{Var}(S) &= \lambda\left\{\mathrm{Var}(X) + [\mathrm{E}(X)]^2\right\} \\
&= \lambda\left[\frac{\alpha\gamma^2}{(\alpha-1)^2(\alpha-2)} + \left(\frac{\gamma}{\alpha-1}\right)^2\right] \\
&= \frac{\lambda\gamma^2}{(\alpha-1)^2}\left[\frac{\alpha}{\alpha-2} + 1\right] \\
&= \frac{2\lambda\gamma^2}{(\alpha-1)(\alpha-2)}.
\end{aligned}$$

With a deductible of amount d, from equation (2.118) the claim-frequency distribution $\tilde{N}$ is Poisson with parameter λv, where $v = \Pr(X > d)$. Using the distribution function of $\mathcal{P}(\alpha, \gamma)$ given in equation (2.38), we have

$$v = \left(\frac{\gamma}{d+\gamma}\right)^{\alpha}.$$

From equations (2.37) and (3.37), the pdf of the modified claim severity $\tilde{X}$ is

$$\begin{aligned}
f_{\tilde{X}}(x) &= \frac{\dfrac{\alpha\gamma^{\alpha}}{(x+d+\gamma)^{\alpha+1}}}{\left(\dfrac{\gamma}{d+\gamma}\right)^{\alpha}} \\
&= \frac{\alpha(d+\gamma)^{a}}{[x+(d+\gamma)]^{\alpha+1}}, \quad \text{for } x > 0.
\end{aligned}$$

Thus, the modified claim severity $\tilde{X}$ is distributed as $\mathcal{P}(\alpha, d+\gamma)$. Now we can conclude

$$\begin{aligned}
\mathrm{E}(\tilde{S}) &= \mathrm{E}(\tilde{N})\mathrm{E}(\tilde{X}) \\
&= \lambda\left(\frac{\gamma}{d+\gamma}\right)^{\alpha}\left(\frac{d+\gamma}{\alpha-1}\right) \\
&= \frac{\lambda\gamma^{\alpha}}{(\alpha-1)(d+\gamma)^{\alpha-1}} \\
&= \mathrm{E}(S)\left(\frac{\gamma}{d+\gamma}\right)^{\alpha-1},
\end{aligned}$$

which is less than $\mathrm{E}(S)$. Furthermore

$$\begin{aligned}
\mathrm{Var}(\tilde{S}) &= \lambda v \left\{ \mathrm{Var}(\tilde{X}) + [\mathrm{E}(\tilde{X})]^2 \right\} \\
&= \lambda \left(\frac{\gamma}{d+\gamma} \right)^{\alpha} \left[\frac{\alpha(d+\gamma)^2}{(\alpha-1)^2(\alpha-2)} + \left(\frac{d+\gamma}{\alpha-1} \right)^2 \right] \\
&= \lambda \left(\frac{\gamma}{d+\gamma} \right)^{\alpha} \left(\frac{d+\gamma}{\alpha-1} \right)^2 \left(\frac{\alpha}{\alpha-2} + 1 \right) \\
&= \lambda \left(\frac{\gamma}{d+\gamma} \right)^{\alpha} \left[\frac{2(d+\gamma)^2}{(\alpha-1)(\alpha-2)} \right] \\
&= \mathrm{Var}(S) \left(\frac{\gamma}{d+\gamma} \right)^{\alpha-2},
\end{aligned}$$

which is less than $\mathrm{Var}(S)$. □

Example 3.8 An insurer pays 80% of the aggregate loss in excess of the deductible of amount 5 up to a maximum payment of 25. Aggregate claim amounts S are integers and the following expected losses are known: $\mathrm{E}[(S-5)_+] = 8.52$, $\mathrm{E}[(S-36)_+] = 3.25$ and $\mathrm{E}[(S-37)_+] = 2.98$. Calculate the expected amount paid by the insurer.

Solution If d is the stop-loss amount, then

$$0.8(d-5) = 25,$$

which implies $d = 36.25$.

Thus, the amount paid by the insurer is $0.8\,[(S-5)_+ - (S-36.25)_+]$, so that the expected loss is $0.8\,(\mathrm{E}[(S-5)_+] - \mathrm{E}[(S-36.25)_+])$. From equation (3.46), we have

$$\begin{aligned}
\mathrm{E}[(S-36.25)_+] &= 0.75\,\mathrm{E}[(S-36)_+] + 0.25\,\mathrm{E}[(S-37)_+] \\
&= (0.75)(3.25) + (0.25)(2.98) \\
&= 3.1825.
\end{aligned}$$

Thus, the insurer's expected loss is

$$0.8(8.52 - 3.1825) = 4.27.$$ □

Example 3.9 Aggregate loss S takes values in multiples of 10. If $\mathrm{E}[(S-20)_+] = 12.58$ and $F_S(20) = 0.48$, calculate $\mathrm{E}[(S-24)_+]$ and $\mathrm{E}[(S-30)_+]$.

Table 3.7. *Methods for computing the aggregate-loss distribution*

Model	Exact methods	Approximate methods
Individual risk	1 Convolution: with discretized claim-severity distribution 2 De Pril recursion: with specific set-up of policy stratification	1 Normal approximation 2 Compound Poisson distribution and Panjer recursion
Collective risk	1 Convolution: with discretized claim-severity distribution and assumed primary distribution 2 Panjer recursion: primary distribution follows $(a, b, 0)$ class, secondary distribution discretized 3 Some limited analytic results	1 Normal approximation

Solution Using equation (3.44) by modifying the increments of claim amounts as 10, we obtain

$$\mathrm{E}[(S-30)_+] = \mathrm{E}[(S-20)_+] - 10[1-F_S(20)] = 12.58 - (10)(0.48) = 7.78.$$

Thus, from equation (3.46), we have

$$\mathrm{E}[(S-24)_+] = (0.6)(12.58) + (0.4)(7.78) = 10.66. \qquad \square$$

3.5 Summary and conclusions

We have discussed the individual risk and collective risk models for analyzing the distribution of aggregate loss. Both models build upon assumptions of the claim-frequency and claim-severity distributions. While exact distributions are available for both models their computation may be very intensive. Table 3.7 summarizes the exact as well as approximate methods for the computation of the aggregate-loss distribution.

The Panjer recursion can be extended to other classes of primary distributions, and some results are provided in Dickson (2005). When the portfolio size of insurance policies is large, computation of the exact distribution often encounters underflow or overflow problems; Klugman *et al.* (1998) discuss some of the computational stability issues.

Exercises

3.1 The primary distribution of a compound distribution S is $\mathcal{NB}(2, 0.25)$, and its secondary distribution is $\mathcal{PN}(\lambda)$. If $f_S(0) = 0.067$, calculate λ.

3.2 The primary distribution of a compound distribution S is $\mathcal{BN}(4, \theta)$, and its secondary distribution is $\mathcal{GM}(\beta)$. If $f_S(0) = 0.4$ and $f_S(1) = 0.04$, find θ and β.

3.3 The primary distribution of a compound distribution S is $\mathcal{PN}(\lambda_1)$, and its secondary distribution is $\mathcal{PN}(\lambda_2)$. If the mean and the variance of S are 2 and 5, respectively, determine λ_1 and λ_2.

3.4 A portfolio has 100 independent insurance policies. Each policy has a probability 0.8 of making no claim, and a probability 0.2 of making a claim. When a claim is made, the loss amounts are 10, 50, and 80 with probabilities of 0.4, 0.4, and 0.2, respectively. Calculate the mean and the variance of the aggregate claim of the portfolio.

3.5 A portfolio has n independent insurance policies. Each policy has a probability $1 - \theta$ of making no claim, and a probability θ of making a claim. When a claim is made, the loss amount is distributed as $\mathcal{G}(\alpha, \beta)$. Determine the mgf of the aggregate claim of the portfolio.

3.6 There are two independent insurance policies. The claim frequency of each policy follows a $\mathcal{BN}(2, 0.1)$ distribution. When there is a claim, the claim amount follows the zero-truncated $\mathcal{BN}(5, 0.4)$ distribution. What is the probability that the aggregate claim of these two policies is (a) zero and (b) one?

3.7 There are two independent insurance policies. The claim frequency of each policy follows a $\mathcal{NB}(2, 0.2)$ distribution. When there is a claim, the claim amount follows the $\mathcal{BN}(4, 0.5)$ distribution. What is the probability that the aggregate claim of these two policies is (a) zero and (b) one?

3.8 A portfolio has five independent policies. Each policy has equal probability of making 0, 1, and 2 claims. Each claim is distributed as $\mathcal{E}(0.05)$. Determine the mgf of the aggregate loss of the portfolio.

3.9 An insurance policy has probabilities of 0.6, 0.2, 0.1, and 0.1 of making 0, 1, 2, and 3 claims, respectively. The claim severity is 1 or 2 with equal probability. Calculate the mean and the variance of the aggregate loss of the policy.

3.10 An insurance policy has claims of amount 0, 10, and 20 with probabilities of 0.7, 0.2, and 0.1, respectively. Calculate the probability that the aggregate claim of 4 independent policies is not more than 50.

3.11 In a collective risk model, S has a binomial compound distribution with primary distribution $\mathcal{BN}(2, 0.6)$. The individual claims are 1, 2, and 3, with probabilities of 0.5, 0.3, and 0.2, respectively. Determine the pgf of S.

3.12 An insurer has five independent policies. Each policy has a probability 0.2 of making a claim, which takes a value of 1 or 2 with probability of 0.3 and 0.7, respectively.
(a) Using Panjer's recursion, determine the probability that the aggregate claim is less than 5.
(b) Using convolution, determine the probability that the aggregate claim is less than 5.

3.13 X_1 and X_2 are iid mixed-type random variables, with $f_{X_1}(0) = f_{X_2}(0) = 0.5$ and a constant density in the interval (0, 2]. Determine the df of $X_1 + X_2$.

3.14 A portfolio has 500 independent policies. Each policy has a probability of 0.6 making no claim, and a probability of 0.4 making a claim. Claim amount is distributed as $\mathcal{P}(3, 20)$. Determine an approximate chance that the aggregate claim is larger than 2,500.

3.15 The number of claims N of an insurance portfolio has the following distribution: $f_N(0) = 0.1, f_N(1) = 0.4, f_N(2) = 0.3$, and $f_N(3) = 0.2$. The claim amounts are distributed as $\mathcal{E}(0.4)$ and are independent of each other as well as the claim frequency. Determine the variance of the aggregate claim.

3.16 The number of claims per year of an insurance portfolio is distributed as $\mathcal{PN}(10)$. The loss amounts are distributed as $\mathcal{U}(0, 10)$. Claim frequency and claim amount are independent, and there is a deductible of 4 per loss. Calculate the mean of the aggregate claim per year.

3.17 Claim severity X is distributed as $\mathcal{U}(0, 10)$. An insurance policy has a policy limit of $u = 8$. Calculate the mean and the variance of the aggregate loss of 20 independent policies, if each policy has a probability of 0.2 of a claim and 0.8 of no claim.

3.18 Aggregate loss S follows a compound distribution with primary distribution $\mathcal{NB}(3, 0.3)$. The claim severity is distributed as $\mathcal{P}(3, 20)$. Compute the mean and the variance of S. If the policies are modified with a deductible of $d = 5$, compute the mean and the variance of the aggregate loss $\tilde{S}$.

3.19 Let N_1 and N_2 be independent Poisson distributions with $\lambda = 1$ and 2, respectively. How would you express $S = -2N_1 + N_2$ as a compound Poisson distribution?

3.20 Aggregate loss S follows a compound distribution. The primary distribution is $\mathcal{PN}(20)$ and the secondary distribution is $\mathcal{U}(0, 5)$. If the policies are modified with a deductible of $d = 1$ and a maximum covered loss of $u = 4$, compute the mean and the variance of the modified aggregate loss $\tilde{S}$.

3.21 In a portfolio of 100 independent policies the probabilities of one claim and no claim are 0.1 and 0.9, respectively. Suppose claim severity is distributed as $\mathcal{E}(0.2)$ and there is a maximum covered loss of 8, approximate the probability that the aggregate loss is less than 50.

3.22 Portfolio A consists of 50 insurance policies, each with a probability 0.8 of no claim and a probability 0.2 of making one claim. Claim severity is distributed as $\mathcal{E}(0.1)$. Portfolio B consists of 70 insurance policies, each with a probability 0.7 of no claim and a probability 0.3 of making one claim. Claim severity is distributed as $\mathcal{P}(3, 30)$.

(a) Calculate the mean and the variance of the aggregate loss of the combined portfolio based on the individual risk model.

(b) How would you approximate the aggregate loss of the combined portfolio using a collective risk model? Determine the mean and the variance of the aggregate loss of the combined portfolio based on the collective risk model.

Questions adapted from SOA exams

3.23 The aggregate loss S is distributed as a compound binomial distribution, where the primary distribution is $\mathcal{BN}(9, 0.2)$. Claim severity X has pf: $f_X(1) = 0.4$, $f_X(2) = 0.4$, and $f_X(3) = 0.2$. Calculate $\Pr(S \leq 4)$.

3.24 Aggregate claim S can only take positive integer values. If $\mathrm{E}[(S-2)_+] = 1/6$, $\mathrm{E}[(S-3)_+] = 0$, and $f_S(1) = 1/2$, calculate the mean of S.

3.25 You are given that $\mathrm{E}[(S-30)_+] = 8$ and $\mathrm{E}[(S-20)_+] = 12$, and the only possible aggregate claim in (20, 30] is 22, with $f_S(22) = 0.1$. Calculate $F_S(20)$.

3.26 Aggregate losses follow a compound Poisson distribution with parameter $\lambda = 3$. Individual losses take values 1, 2, 3, and 4 with probabilities 0.4, 0.3, 0.2, and 0.1, respectively. Calculate the probability that the aggregate loss does not exceed 3.

3.27 Aggregate losses follow a compound distribution. The claim frequency has mean 100 and standard deviation 25. The claim severity has mean 20,000 and standard deviation 5,000. Determine the normal approximation of the probability that the aggregate loss exceeds 150% of the expected loss.

Part II

Risk and ruin

This part of the book is about two important and related topics in modeling insurance business: measuring risk and computing the likelihood of ruin. In Chapter 4 we introduce various measures of risk, which are constructed with the purpose of setting premium or capital. We discuss the axiomatic approach of identifying risk measures that are coherent. Specific measures such as Value-at-Risk, conditional tail expectation, and the distortion-function approach are discussed. Chapter 5 analyzes the probability of ruin of an insurance business in both discrete-time and continuous-time frameworks. Probabilities of ultimate ruin and ruin before a finite time are discussed. We show the interaction of the initial surplus, premium loading, and loss distribution on the probability of ruin.

4

Risk measures

As insurance companies hold portfolios of insurance policies that may result in claims, it is a good management practice to assess the exposure of the company to such risks. A risk measure, which summarizes the overall risk exposures of the company, helps the company evaluate if there is sufficient capital to overcome adverse events. Risk measures for blocks of policies can also be used to assess the adequacy of the premium charged. Since the Basel Accords I and II, financial institutions such as banks and insurance companies have elevated their efforts to assess their internal risks as well as communicating the assessments to the public.

In this chapter we discuss various measures of risks. We introduce the axioms proposed by Artzner *et al.* (1999), which define the concept of a *coherent* risk measure. Risk measures based on the premium principle, such as the expected-value principle, variance principle, and standard-deviation principle, are discussed. This is followed by the capital-based risk measures such as the Value-at-Risk and the conditional tail expectation. Many of the risk measures used in the actuarial literature can be viewed as the integral of a distortion function of the survival function of the loss variable, or the mean of the risk-adjusted loss. Further risk measures that come under this category are risk measures defined by the hazard transform and the Wang (2000) transform.

Learning objectives

1 Axioms of coherent risk measures
2 Risk measures based on premium principles
3 Risk measures based on capital requirements
4 Value-at-Risk and conditional tail expectation
5 Distortion function
6 Proportional hazard transform and Wang transform

4.1 Uses of risk measures

Risk management is of paramount importance for the management of a firm. The risks facing a firm can be generally classified under **market risk** (exposure to potential loss due to changes in market prices and market conditions), **credit risk** (risk of customers defaulting), and **operational risk** (any business risk that is not a market nor credit risk). The risk management process should be a holistic process covering the analysis of risk incidents, assessment of management control, reporting procedures, and prediction of risk trends. While a risk management process is multi-dimensional, measuring risk is the core component of the process.

In this chapter we focus on measuring the risks of an insurance company. A major risk an insurance company encounters is the loss arising from the insurance policies. In other words, our focus is on the operational risk of the firm. We shall discuss various measures that attempt to summarize the potential risks arising from the possible claims of the insurance policies. Thus, the measures will be based on the loss random variables. Such measures may be used by an insurance company in the following ways:

Determination of economic capital

Economic capital is the capital a firm is required to hold in order to avoid insolvency. It is a buffer against unexpected losses, and may differ from the available capital of the firm. The size of the economic capital is dependent on the level of credit standing the firm expects to achieve or the probability of insolvency the firm is prepared to tolerate. The first step in the calculation of economic capital often involves the quantification of the possible risks of the firm.

Determination of insurance premium

Insurance premium is the price demanded by the insurance company for transferring the risk of loss from the insured to the insurer. The premium charged should vary directly with the potential loss. Thus, appropriately measuring the risk is important for the determination of the insurance premium.

It should be noted that in practice the determination of premium often depends on other factors such as competition in the industry and strategic marketing concerns. In this book, however, we consider premium determination principles purely from the point of view of compensation for potential losses.

Internal risk management

Risk measures are important inputs to internal risk management and control. A firm may set targets on different segments of the business

based on certain risk measures. Internal evaluation will be much easier if clear and well-defined targets for risk measures are available.

External regulatory reporting

Concerned about the solvency of insurance companies, various regulatory bodies have attempted to institutionalize the regulatory framework of reporting, as well as step up the supervision of such reports. Risk measures form a main part of the reporting system.

Since the Basel Accord I in 1988 and the Basel Accord II in 2004, risk assessment and reporting have assumed important profiles in many financial institutions. A survey of the best practice in the industry in enterprise risk management can be found in Lam (2003). McNeil *et al.* (2005) provides an introductory description of the Basel Accords as well as some regulatory developments in the insurance industry.

In what follows we first describe some simple risk measures based on the premium principle. We then introduce the axiomatic approach of Artzner *et al.* (1999) for identifying desirable properties of a risk measure. Some risk measures based on capital requirements are then discussed, and a unifying theme on risk measures using distortion functions concludes this chapter.

Prior to introducing some premium-based risk measures, we first provide a formal definition of a risk measure based on a random loss X, which is the aggregate claim of a block of insurance policies.[1]

Definition 4.1 A risk measure of the random loss X, denoted by $\varrho(X)$, is a real-valued function $\varrho : X \to \mathbb{R}$, where $\mathbb{R}$ is the set of real numbers.

As a loss random variable, X is nonnegative. Thus, the risk measure $\varrho(X)$ may be imposed to be nonnegative for the purpose of measuring insurance risks. However, if the purpose is to measure the risks of a portfolio of assets, X may stand for the *change* in portfolio value, which may be positive or negative. In such cases, the risk measure $\varrho(X)$ may be positive or negative.

4.2 Some premium-based risk measures

We denote the mean and the variance of the random loss X by μ_X and σ_X^2, respectively. The **expected-value principle premium** risk measure is defined as

$$\varrho(X) = (1 + \theta)\mu_X, \tag{4.1}$$

[1] In this chapter we use X to denote the aggregate loss random variable rather than S, which has other notational use.

where $\theta \geq 0$ is the **premium loading factor**. Thus, the loading in excess of the mean loss μ_X is $\theta\mu_X$. In the special case of $\theta = 0$, $\varrho(X) = \mu_X$, and the risk measure is called the **pure premium**.

Note that in the expected-value premium risk measure, the risk depends only on the mean μ_X and the loading factor θ. Thus, two loss variables with the same mean and same loading will have the same risk, regardless of the higher-order moments such as the variance. To differentiate such loss distributions, we may consider the **variance principle premium** risk measure defined by

$$\varrho(X) = \mu_X + \alpha\sigma_X^2, \tag{4.2}$$

or the **standard-deviation principle premium** risk measure defined by

$$\varrho(X) = \mu_X + \alpha\sigma_X, \tag{4.3}$$

where $\alpha \geq 0$ in equations (4.2) and (4.3) is the loading factor. Under the variance premium and standard-deviation premium risk measures, the loss distribution with a larger dispersion will have a higher risk. This appears to be a reasonable property. However, these two risk measures have quite different properties, as we shall see later. Thus, the choice of a risk measure is not a trivial task. In the next section we introduce some properties of risk measures that are deemed to be desirable. The selection of risk measures may then be focused on the set of risk measures that satisfy these properties.

4.3 Axioms of coherent risk measures

Artzner *et al.* (1999) suggest four axioms of measures of risk. They argue that these axioms "should hold for any risk measure that is to be used to effectively regulate or manage risks." A risk measure that satisfies these four axioms is said to be **coherent**. We summarize these axioms as follows.[2]

Axiom 4.1 Translational invariance (T) *For any loss variable X and any nonnegative constant a, $\varrho(X + a) = \varrho(X) + a$.*

Axiom T states that if the loss X is increased by a fixed amount a, then the risk increases by the same amount.

Axiom 4.2 Subadditivity (S) *For any loss variables X and Y, $\varrho(X + Y) \leq \varrho(X) + \varrho(Y)$.*

[2] Artzner *et al.* (1999) consider X that can be positive or negative. As we have restricted our interest to nonnegative X for insurance losses, we have modified their axioms accordingly.

Axiom S implies that an insurance company cannot reduce its risk by splitting its business into smaller blocks. It also says that consolidating blocks of policies does not make the company more risky.

Axiom 4.3 Positive homogeneity (PH) *For any loss variable X and any nonnegative constant a, $\varrho(aX) = a\varrho(X)$.*

This axiom is reasonable as it ensures that changing the monetary units of the risks does not alter the risk measure.

Axiom 4.4 Monotonicity (M) *For any loss variables X and Y such that $X \leq Y$ under all states of nature, $\varrho(X) \leq \varrho(Y)$.*

Axiom M states that if the loss of one risk is no more than that of another risk under all states of nature, the risk measure of the former risk cannot be more than that of the latter.

Example 4.1 Show that, under Axiom PH, $\varrho(0) = 0$. Hence, prove that if Axioms M and PH hold, $\varrho(X) \geq 0$ for $X \geq 0$.

Solution First we note that $\varrho(0) = \varrho(a0)$ for all a. Now with Axiom PH, we have $\varrho(a0) = a\varrho(0)$ for all $a \geq 0$. Thus, we conclude $\varrho(0) = a\varrho(0)$ for all $a \geq 0$, which implies $\varrho(0) = 0$.

Now for $X \geq 0$, we have, from Axiom M, $\varrho(X) \geq \varrho(0)$, and we conclude $\varrho(X) \geq 0$. □

If we apply Axiom T to a coherent risk measure and assume $X = 0$, then for any nonnegative constant a we have $\varrho(a) = \varrho(X + a) = \varrho(X) + a = \varrho(0) + a = a$. This result says that if a risk takes a constant value, a coherent risk measure of the risk must be equal to this constant. Thus, for a coherent risk measure based on the premium principle, the loading for a constant risk must be equal to zero. Consequently, we say that a coherent risk measure has **no unjustified loading**.

If the loss X has a finite support with maximum value x_U, then a risk defined by $Y = x_U$ satisfies $X \leq Y$. From Axiom M, a coherent risk measure must satisfy $\varrho(X) \leq \varrho(Y) = \varrho(x_U) = x_U$. Thus, a coherent risk is bounded above by the maximum loss. A premium that satisfies this condition is said to have the property of **no ripoff**.[3]

Example 4.2 Show that the expected-value premium risk measure satisfies Axioms S, PH, and M, but not T.

[3] It can also be shown that a coherent risk measure satisfies the condition $\varrho(X) \geq \mu_X$. A proof of this result is given in Artzner (1999).

Solution For any risks X and Y, we have

$$\begin{aligned}\varrho(X+Y) &= (1+\theta)\mathrm{E}(X+Y)\\ &= (1+\theta)\mathrm{E}(X)+(1+\theta)\mathrm{E}(Y)\\ &= \varrho(X)+\varrho(Y).\end{aligned}$$

Thus, Axiom S holds. Now for $Y = aX$ with $a \geq 0$, we have

$$\varrho(Y) = (1+\theta)\mathrm{E}(Y) = (1+\theta)\mathrm{E}(aX) = a(1+\theta)\mathrm{E}(X) = a\varrho(X),$$

which proves Axiom PH. For two risks X and Y, $X \geq Y$ implies $\mu_X \geq \mu_Y$. Thus

$$\varrho(X) = (1+\theta)\mu_X \geq (1+\theta)\mu_Y = \varrho(Y),$$

and Axiom M holds. To examine Axiom T, we consider an arbitrary constant $a > 0$. Note that, if $\theta > 0$

$$\varrho(X+a) = (1+\theta)\mathrm{E}(X+a) > (1+\theta)\mathrm{E}(X)+a = \varrho(X)+a.$$

Thus, Axiom T is not satisfied if $\theta > 0$, which implies the expected-value premium is in general not a coherent risk measure. However, when $\theta = 0$, Axiom T holds. Thus, the pure premium risk measure is coherent. □

It can be shown that the variance premium risk measure satisfies Axiom T, but not Axioms S, M, and PH. On the other hand, the standard-deviation premium risk measure satisfies Axioms S, T, and PH, but not Axiom M. Readers are invited to prove these results (see Exercises 4.2 and 4.3).

The axioms of coherent risk narrow down the set of risk measures to be considered for management and regulation. However, they do not specify a unique risk measure to be used in practice. Some risk measures (such as the pure premium risk measure) that are coherent may not be suitable for some reasons. Thus, the choice of which measure to use depends on additional considerations.

4.4 Some capital-based risk measures

We now introduce some risk measures constructed for the purpose of evaluating economic capital.

4.4.1 Value-at-Risk (VaR)

Value-at-Risk (VaR) is probably one of the most widely used measures of risk. Simply speaking, the VaR of a loss variable is the minimum value of the distribution such that the probability of the loss larger than this value is not

more than a given probability. In statistical terms, VaR is a quantile as defined in Section 2.4. We now define VaR formally as follows.

Definition 4.2 Let X be a random variable of loss with continuous df $F_X(\cdot)$, and δ be a probability level such that $0 < \delta < 1$, the Value-at-Risk at probability level δ, denoted by $\text{VaR}_\delta(X)$, is the δ-quantile of X. That is

$$\text{VaR}_\delta(X) = F_X^{-1}(\delta) = x_\delta. \tag{4.4}$$

The probability level δ is usually taken to be close to 1 (say, 0.95 or 0.99), so that the probability of loss X exceeding $\text{VaR}_\delta(X)$ is not more than $1 - \delta$, and is thus small. We shall write VaR at probability level δ as VaR_δ when the loss variable is understood.

If $F_X(\cdot)$ is a step function (as when X is not continuous), there may be some ambiguity in the definition of $F_X^{-1}(\delta)$. Thus, a more general definition of $\text{VaR}_\delta(X)$ is

$$\text{VaR}_\delta(X) = \inf\{x \in [0, \infty) : F_X(x) \geq \delta\}. \tag{4.5}$$

Example 4.3 Find VaR_δ of the following loss distributions X: (a) $\mathcal{E}(\lambda)$, (b) $\mathcal{L}(\mu, \sigma^2)$, and (c) $\mathcal{P}(\alpha, \gamma)$.

Solution For (a), from Example 2.8, we have

$$\text{VaR}_\delta = -\frac{\log(1-\delta)}{\lambda}.$$

For (b), from Example 2.8, the VaR is

$$\text{VaR}_\delta = \exp\left[\mu + \sigma\Phi^{-1}(\delta)\right].$$

For (c), from equation (2.38), the df of $\mathcal{P}(\alpha, \gamma)$ is

$$F_X(x) = 1 - \left(\frac{\gamma}{x+\gamma}\right)^\alpha,$$

so that its quantile function is

$$F_X^{-1}(\delta) = \gamma(1-\delta)^{-\frac{1}{\alpha}} - \gamma,$$

and

$$\text{VaR}_\delta = F_X^{-1}(\delta) = \gamma\left[(1-\delta)^{-\frac{1}{\alpha}} - 1\right]. \qquad \square$$

Example 4.4 Find VaR_δ, for $\delta = 0.95, 0.96, 0.98$, and 0.99, of the following discrete loss distribution

$$X = \begin{cases} 100, & \text{with prob } 0.02, \\ 90, & \text{with prob } 0.02, \\ 80, & \text{with prob } 0.04, \\ 50, & \text{with prob } 0.12, \\ 0, & \text{with prob } 0.80. \end{cases}$$

Solution As X is discrete, we use the definition of VaR in equation (4.5). The df of X is plotted in Figure 4.1. The dotted horizontal lines correspond to the probability levels 0.95, 0.96, 0.98, and 0.99. Note that the df of X is a step function. For VaR_δ we require the value of X corresponding to the probability level equal to or next-step higher than δ. Thus, VaR_δ for $\delta = 0.95, 0.96, 0.98$, and 0.99, are, respectively, 80, 80, 90, and 100. □

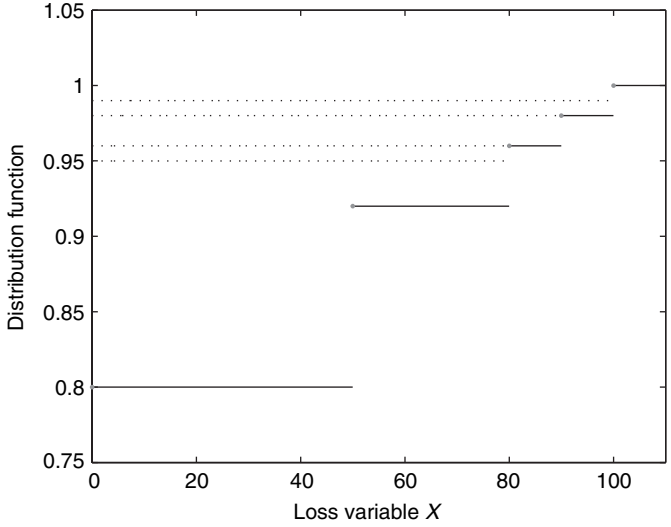

Figure 4.1 Distribution function of loss and VaR of Example 4.4

As the distribution function is monotonic, it is easy to see that VaR satisfies Axiom M for coherency. Thus, if $X \leq Y$ under all states of nature

$$\Pr(X \leq \text{VaR}_\delta(Y)) \geq \Pr(Y \leq \text{VaR}_\delta(Y)) \geq \delta, \tag{4.6}$$

which implies $\text{VaR}_\delta(X) \leq \text{VaR}_\delta(Y)$ and Axiom M holds. Now for any risk X and any positive constant a, let $Y = X + a$. We have

$$\begin{aligned}
\text{VaR}_\delta(X + a) &= \text{VaR}_\delta(Y) \\
&= \inf\{y : \Pr(Y \leq y) \geq \delta\} \\
&= \inf\{x + a : \Pr(X + a \leq x + a) \geq \delta\} \\
&= a + \inf\{x : \Pr(X + a \leq x + a) \geq \delta\} \\
&= a + \inf\{x : \Pr(X \leq x) \geq \delta\} \\
&= a + \text{VaR}_\delta(X), \qquad (4.7)
\end{aligned}$$

so that Axiom T holds. Furthermore, if we let $Y = aX$ for $a \geq 0$, we have

$$\begin{aligned}
\text{VaR}_\delta(aX) &= \text{VaR}_\delta(Y) \\
&= \inf\{y : \Pr(Y \leq y) \geq \delta\} \\
&= \inf\{ax : \Pr(aX \leq ax) \geq \delta\} \\
&= \inf\{ax : \Pr(X \leq x) \geq \delta\} \\
&= a\,[\inf\{x : \Pr(X \leq x) \geq \delta\}] \\
&= a\,\text{VaR}_\delta(X), \qquad (4.8)
\end{aligned}$$

and Axiom PH holds.

While VaR satisfies Axioms M, T, and PH, it does not satisfy Axiom S and is thus not coherent. A counter-example which illustrates that VaR is not subadditive can be found in Artzner *et al.* (1999).

4.4.2 Conditional tail expectation and related measures

A drawback of VaR is that it only makes use of the cut-off point corresponding to the probability level δ and does not use any information about the tail distribution beyond this point. The conditional tail expectation (CTE) corrects for this. As stated in equation (2.66), the CTE at probability level δ, denoted by $\text{CTE}_\delta(X)$ (or CTE_δ when the loss variable is understood) is defined as

$$\text{CTE}_\delta(X) = \text{E}(X \mid X > x_\delta). \qquad (4.9)$$

When X is continuous, the above can be written as

$$\text{CTE}_\delta(X) = \text{E}\left[X \mid X > \text{VaR}_\delta(X)\right], \qquad (4.10)$$

which will be used as the definition of CTE as a risk measure, and this definition also applies to discrete losses. Analogous to the excess-loss variable defined in Section 2.5.1, we consider the loss in excess of the VaR conditional on it being exceeded, i.e.

$$X - \text{VaR}_\delta(X) \,|\, X > \text{VaR}_\delta(X). \tag{4.11}$$

The mean of this conditional excess, called the **conditional VaR**, is denoted by $\text{CVaR}_\delta(X)$ (or CVaR_δ when the loss variable is understood) and defined as[4]

$$\text{CVaR}_\delta(X) = \text{E}\left[X - \text{VaR}_\delta(X) \,|\, X > \text{VaR}_\delta(X)\right]. \tag{4.12}$$

The above equation can be written as

$$\begin{aligned}\text{CVaR}_\delta(X) &= \text{E}\left[X \,|\, X > \text{VaR}_\delta(X)\right] - \text{E}\left[\text{VaR}_\delta(X) \,|\, X > \text{VaR}_\delta(X)\right] \\ &= \text{CTE}_\delta(X) - \text{VaR}_\delta(X),\end{aligned} \tag{4.13}$$

which is analogous to equation (2.81), with the deductible replaced by VaR.

If we use VaR_δ as the economic capital, the **shortfall** of the capital is

$$(X - \text{VaR}_\delta)_+. \tag{4.14}$$

When X is continuous, $\text{VaR}_\delta = x_\delta$ and the **mean shortfall** is

$$\begin{aligned}\text{E}\left[(X - x_\delta)_+\right] &= \text{E}\left[X - x_\delta \,|\, X > x_\delta\right] \Pr(X > x_\delta) \\ &= (1 - \delta)\,\text{CVaR}_\delta,\end{aligned} \tag{4.15}$$

and we have, from equations (4.13) and (4.15)

$$\begin{aligned}\text{CTE}_\delta &= x_\delta + \text{CVaR}_\delta \\ &= x_\delta + \frac{1}{1-\delta}\text{E}\left[(X - x_\delta)_+\right],\end{aligned} \tag{4.16}$$

which relates CTE_δ to the mean shortfall.[5] To evaluate CTE_δ, we consider

$$\text{CTE}_\delta = \text{E}(X \,|\, X > x_\delta) = \frac{1}{1-\delta}\int_{x_\delta}^{\infty} x f_X(x)\,dx. \tag{4.17}$$

[4] This definition follows Denuit *et al.* (2005). CVaR is also alternatively taken as synonymous with CTE.

[5] Note that mathematically CVaR is equivalent to $\text{E}(X_P)$ (see equation (2.80)), and the mean shortfall is equivalent to $\text{E}(X_L)$ (see equation (2.78)).

Using change of variable $\xi = F_X(x)$, the integral above can be written as

$$\int_{x_\delta}^{\infty} x f_X(x)\,dx = \int_{x_\delta}^{\infty} x\,dF_X(x) = \int_{\delta}^{1} x_\xi\,d\xi, \tag{4.18}$$

which implies

$$\text{CTE}_\delta = \frac{1}{1-\delta}\int_{\delta}^{1} x_\xi\,d\xi. \tag{4.19}$$

Thus, CTE_δ can be interpreted as the *average* of the quantiles exceeding x_δ.

On the other hand, when X is not necessarily continuous, equation (4.17) is replaced by the Stieltjes integral

$$\text{CTE}_\delta = \text{E}(X \mid X > \text{VaR}_\delta) = \frac{1}{1-\bar{\delta}}\int_{x\,\in\,(\text{VaR}_\delta,\,\infty)} x\,dF_X(x), \tag{4.20}$$

where

$$\bar{\delta} = \Pr(X \le \text{VaR}_\delta). \tag{4.21}$$

However, as $\bar{\delta} \ge \delta$, equation (4.20) implies that we may be using less than the worst $1-\delta$ portion of the loss distribution in computing CTE_δ. To circumvent this problem, we use the following formula (see Hardy, 2003, p. 164)

$$\text{CTE}_\delta = \frac{(\bar{\delta}-\delta)\text{VaR}_\delta + (1-\bar{\delta})\text{E}(X \mid X > \text{VaR}_\delta)}{1-\delta}, \tag{4.22}$$

which is a weighted average of VaR_δ and $\text{E}(X \mid X > \text{VaR}_\delta)$. We shall adopt this formula for CTE_δ in subsequent discussions. When $\bar{\delta} = \delta$ (as when X is continuous), (4.22) reduces to $\text{CTE}_\delta = \text{E}(X \mid X > \text{VaR}_\delta)$.

An expression analogous to the right-hand side of equation (4.19) is

$$\frac{1}{1-\delta}\int_{\delta}^{1} \text{VaR}_\xi\,d\xi, \tag{4.23}$$

which is sometimes called the **tail Value-at-Risk**, denoted by $\text{TVaR}_\delta(X)$ (or TVaR_δ when the loss variable is understood). Expression (4.23) is an alternative way of writing equation (4.22), whether or not X is continuous. Hence, TVaR_δ and CTE_δ (as defined by equation (4.22)) are equivalent.

Example 4.5 Find CTE_δ and CVaR_δ of the following loss distributions X: (a) $\mathcal{E}(\lambda)$, (b) $\mathcal{L}(\mu, \sigma^2)$, and (c) $\mathcal{P}(\alpha, \gamma)$ with $\alpha > 1$.

Solution For $\mathcal{E}(\lambda)$, CTE_δ was computed in Examples 2.8 and 2.9, i.e.

$$\text{CTE}_\delta = \frac{e^{-\lambda x_\delta}}{1-\delta}\left[x_\delta + \frac{1}{\lambda}\right] = x_\delta + \frac{1}{\lambda}.$$

Thus, CVaR_δ is given by

$$\text{CVaR}_\delta = \text{CTE}_\delta - x_\delta = \frac{1}{\lambda}.$$

For $\mathcal{L}(\mu, \sigma^2)$, we have, from Example 2.9

$$\text{CTE}_\delta = \frac{1}{1-\delta}\left\{\exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 - \Phi(z^*)\right]\right\},$$

where

$$z^* = \frac{\log x_\delta - \mu}{\sigma} - \sigma,$$

and x_δ is obtained from Example 2.8 as

$$x_\delta = \exp\left[\mu + \sigma\,\Phi^{-1}(\delta)\right].$$

Thus

$$\begin{aligned} z^* &= \frac{\mu + \sigma\,\Phi^{-1}(\delta) - \mu}{\sigma} - \sigma \\ &= \Phi^{-1}(\delta) - \sigma. \end{aligned}$$

CVaR_δ of $\mathcal{L}(\mu, \sigma^2)$ is

$$\text{CVaR}_\delta = \frac{1}{1-\delta}\left\{\exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 - \Phi(z^*)\right]\right\} - \exp\left[\mu + \sigma\,\Phi^{-1}(\delta)\right].$$

For $\mathcal{P}(\alpha, \gamma)$ we compute the integral in equation (4.17) as

$$\begin{aligned} \int_{x_\delta}^{\infty} x f_X(x)\,dx &= \alpha\gamma^\alpha \int_{x_\delta}^{\infty} \frac{x}{(x+\gamma)^{\alpha+1}}dx \\ &= -\gamma^\alpha\left\{\frac{x}{(x+\gamma)^\alpha}\right]_{x_\delta}^{\infty} - \int_{x_\delta}^{\infty}\frac{dx}{(x+\gamma)^\alpha}\right\} \\ &= -\gamma^\alpha\left\{-\frac{x_\delta}{(x_\delta+\gamma)^\alpha} + \left[\frac{1}{(\alpha-1)(x+\gamma)^{\alpha-1}}\right]_{x_\delta}^{\infty}\right\} \\ &= x_\delta\left(\frac{\gamma}{x_\delta+\gamma}\right)^\alpha + \frac{\gamma^\alpha}{(\alpha-1)(x_\delta+\gamma)^{\alpha-1}}. \end{aligned}$$

Substituting the result

$$\delta = 1 - \left(\frac{\gamma}{x_\delta + \gamma}\right)^\alpha$$

in Example 4.3 into the above equation, we obtain

$$\int_{x_\delta}^{\infty} x f_X(x)\, dx = (1-\delta)\left[x_\delta + \frac{x_\delta + \gamma}{\alpha - 1}\right].$$

Thus, from equation (4.17) we conclude

$$\begin{aligned} \text{CTE}_\delta &= x_\delta + \frac{x_\delta + \gamma}{\alpha - 1} \\ &= \frac{\gamma}{\alpha - 1} + \frac{\alpha x_\delta}{\alpha - 1}, \end{aligned}$$

and CVaR_δ is given by

$$\text{CVaR}_\delta = \text{CTE}_\delta - x_\delta = \frac{x_\delta + \gamma}{\alpha - 1}. \qquad \square$$

Example 4.6 Calculate CTE_δ for the loss distribution given in Example 4.4, for $\delta = 0.95, 0.96, 0.98$, and 0.99. Also, calculate TVaR corresponding to these values of δ.

Solution As X is not continuous we use equation (4.22) to calculate CTE_δ. Note that $\text{VaR}_{0.95} = \text{VaR}_{0.96} = 80$. For $\delta = 0.95$, we have $\bar{\delta} = 0.96$. Now

$$\text{E}(X \mid X > \text{VaR}_{0.95} = 80) = \frac{90(0.02) + 100(0.02)}{0.04} = 95,$$

so that from equation (4.22) we obtain

$$\text{CTE}_{0.95} = \frac{(0.96 - 0.95)80 + (1 - 0.96)95}{1 - 0.95} = 92.$$

For $\delta = 0.96$, we have $\bar{\delta} = 0.96$, so that

$$\text{CTE}_{0.96} = \text{E}(X \mid X > \text{VaR}_{0.96} = 80) = 95.$$

For TVaR_δ, we use equation (4.23) to obtain

$$\begin{aligned} \text{TVaR}_{0.95} &= \frac{1}{1 - 0.95}\int_{0.95}^{1} \text{VaR}_\xi\, d\xi \\ &= \frac{1}{0.05}\left[(80)(0.01) + (90)(0.02) + (100)(0.02)\right] = 92, \end{aligned}$$

and

$$\text{TVaR}_{0.96} = \frac{1}{1-0.96}\int_{0.96}^{1} \text{VaR}_\xi \, d\xi = \frac{1}{0.04}\left[(90)(0.02) + (100)(0.02)\right] = 95.$$

For $\delta = 0.98$, we have $\bar{\delta} = 0.98$, so that

$$\text{CTE}_{0.98} = \text{E}(X \mid X > \text{VaR}_{0.98} = 90) = 100,$$

which is also the value of $\text{TVaR}_{0.98}$. Finally, for $\delta = 0.99$, we have $\bar{\delta} = 1$ and $\text{VaR}_{0.99} = 100$, so that $\text{CTE}_{0.99} = \text{VaR}_{0.99} = 100$. On the other hand, we have

$$\text{TVaR}_{0.99} = \frac{1}{1-0.99}\int_{0.99}^{1} \text{VaR}_\xi \, d\xi = \frac{(100)(0.01)}{0.01} = 100. \qquad \square$$

When X is continuous, CTE satisfies Axioms M, T, PH, and S, and is thus coherent. Let a be any positive constant. We have

$$\begin{aligned}
\text{CTE}_\delta(X + a) &= \text{E}\left[X + a \mid X + a > \text{VaR}_\delta(X + a)\right] \\
&= \text{E}\left[X + a \mid X > \text{VaR}_\delta(X)\right] \\
&= a + \text{E}\left[X \mid X > \text{VaR}_\delta(X)\right] \\
&= a + \text{CTE}_\delta(X),
\end{aligned} \tag{4.24}$$

so that CTE is translational invariant. Likewise

$$\begin{aligned}
\text{CTE}_\delta(aX) &= \text{E}\left[aX \mid aX > \text{VaR}_\delta(aX)\right] \\
&= \text{E}\left[aX \mid X > \text{VaR}_\delta(X)\right] \\
&= a\,\text{E}\left[X \mid X > \text{VaR}_\delta(X)\right] \\
&= a\,\text{CTE}_\delta(X),
\end{aligned} \tag{4.25}$$

so that CTE is positively homogeneous. If two continuous loss distributions X and Y satisfy the condition $X \leq Y$, we have $x_\delta \leq y_\delta$ for $\delta \in (0, 1)$. Thus, from equation (4.21)

$$\begin{aligned}
\text{CTE}_\delta(X) &= \frac{1}{1-\delta}\int_{\delta}^{1} x_\xi \, d\xi \\
&\leq \frac{1}{1-\delta}\int_{\delta}^{1} y_\xi \, d\xi \\
&= \text{CTE}_\delta(Y),
\end{aligned} \tag{4.26}$$

and CTE is monotonic. Finally, a proof of the subadditivity of CTE can be found in Denuit *et al.* (2005).

4.5 More premium-based risk measures

In this section we discuss further risk measures based on the premium principle. Various features of these measures will be explored.

4.5.1 Proportional hazard transform and risk-adjusted premium

The premium-based risk measures introduced in Section 4.2 define risk based on a loading of the expected loss. As shown in equation (2.18), the expected loss μ_X of a nonnegative continuous random loss X can be written as

$$\mu_X = \int_0^\infty S_X(x)\,dx. \tag{4.27}$$

Thus, instead of adding a *loading* to μ_X to obtain a premium we may *re-define* the distribution of the losses by shifting more probability weighting to the high losses. Suppose $\tilde{X}$ is distributed with sf $S_{\tilde{X}}(x) = [S_X(x)]^{\frac{1}{\rho}}$, where $\rho \geq 1$, then the mean of $\tilde{X}$ is[6]

$$\mathrm{E}(\tilde{X}) = \mu_{\tilde{X}} = \int_0^\infty S_{\tilde{X}}(x)\,dx = \int_0^\infty [S_X(x)]^{\frac{1}{\rho}}\,dx. \tag{4.28}$$

The parameter ρ is called the **risk-aversion index**. Note that

$$\frac{d\mathrm{E}(\tilde{X})}{d\rho} = -\frac{1}{\rho^2}\int_0^\infty [S_X(x)]^{\frac{1}{\rho}} \log[S_X(x)]\,dx > 0, \tag{4.29}$$

(as $\log[S_X(x)] < 0$), so that the premium increases with ρ, justifying the risk-aversion index interpretation of ρ.

The distribution of $\tilde{X}$ is called the **proportional hazard (PH) transform** of the distribution of X with parameter ρ.[7] If we denote $h_X(x)$ and $h_{\tilde{X}}(x)$ as the hf of X and $\tilde{X}$, respectively, then from equations (2.2) and (2.3) we have

$$\begin{aligned} h_{\tilde{X}}(x) &= -\frac{1}{S_{\tilde{X}}(x)}\left(\frac{dS_{\tilde{X}}(x)}{dx}\right) \\ &= -\frac{1}{\rho}\left(\frac{[S_X(x)]^{\frac{1}{\rho}-1}\,S'_X(x)}{[S_X(x)]^{\frac{1}{\rho}}}\right) \end{aligned}$$

[6] Given $S_X(x)$ is a well-defined sf, $S_{\tilde{X}}(x)$ is also a well-defined sf.

[7] In general, the PH transform only requires the parameter ρ to be positive. In the context of risk loading, however, we consider PH transforms with $\rho \geq 1$. It can be shown that $\mathrm{E}(\tilde{X}) \geq \mathrm{E}(X)$ if and only if $\rho \geq 1$.

$$= -\frac{1}{\rho}\left(\frac{S'_X(x)}{S_X(x)}\right)$$
$$= \frac{1}{\rho} h_X(x), \tag{4.30}$$

so that the hf of $\tilde{X}$ is *proportional to* the hf of X. As $\rho \geq 1$, the hf of $\tilde{X}$ is less than that of X, implying that $\tilde{X}$ has a thicker tail than that of X. Also, $S_{\tilde{X}}(x) = [S_X(x)]^{\frac{1}{\rho}}$ declines slower than $S_X(x)$ so that $\mu_{\tilde{X}} > \mu_X$, the difference of which represents the loading.

Example 4.7 If $X \sim \mathcal{E}(\lambda)$, find the PH transform of X with parameter ρ and the risk-adjusted premium.

Solution The sf of X is $S_X(x) = e^{-\lambda x}$. Thus, the sf of the PH transform is $S_{\tilde{X}}(x) = \left(e^{-\lambda x}\right)^{\frac{1}{\rho}} = e^{-\frac{\lambda}{\rho}x}$, which implies $\tilde{X} \sim \mathcal{E}(\lambda/\rho)$. Hence, the risk-adjusted premium is $\mathrm{E}(\tilde{X}) = \rho/\lambda \geq 1/\lambda = \mathrm{E}(X)$.

Figure 4.2 plots the pdf of $X \sim \mathcal{E}(1)$ and its PH transforms for $\rho = 1.5$ and 2.0. It can be seen that the PH transforms have thicker tails than the original loss distribution.

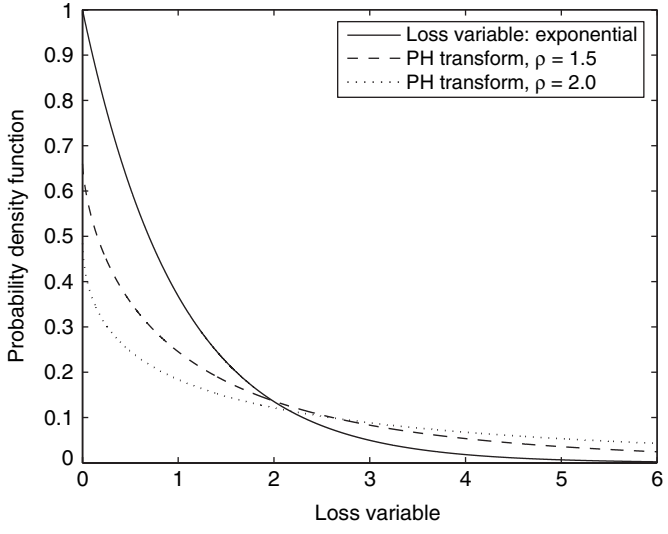

Figure 4.2 Probability density functions of $\mathcal{E}(1)$ and its PH transforms □

Example 4.8 If $X \sim \mathcal{P}(\alpha, \gamma)$ with $\alpha > 1$, find the PH transform of X with parameter $\rho \in [1, \alpha)$ and the risk-adjusted premium.

Solution The sf of X is

$$S_X(x) = \left(\frac{\gamma}{\gamma + x}\right)^{\alpha},$$

with mean

$$\mu_X = \frac{\gamma}{\alpha - 1}.$$

The sf of $\tilde{X}$ is

$$S_{\tilde{X}}(x) = [S_X(x)]^{\frac{1}{\rho}} = \left(\frac{\gamma}{\gamma + x}\right)^{\frac{\alpha}{\rho}},$$

so that $\tilde{X} \sim \mathcal{P}(\alpha/\rho, \gamma)$. Hence, the mean of $\tilde{X}$ (the risk-adjusted premium) is

$$\mu_{\tilde{X}} = \frac{\gamma}{\frac{\alpha}{\rho} - 1} = \frac{\rho\gamma}{\alpha - \rho} > \frac{\gamma}{\alpha - 1} = \mu_X.$$

Figure 4.3 plots the pdf of $X \sim \mathcal{P}(4, 2)$ and its PH transform for $\rho = 2$ and 3. It can be seen that the PH transforms have thicker tails than the original loss distribution.

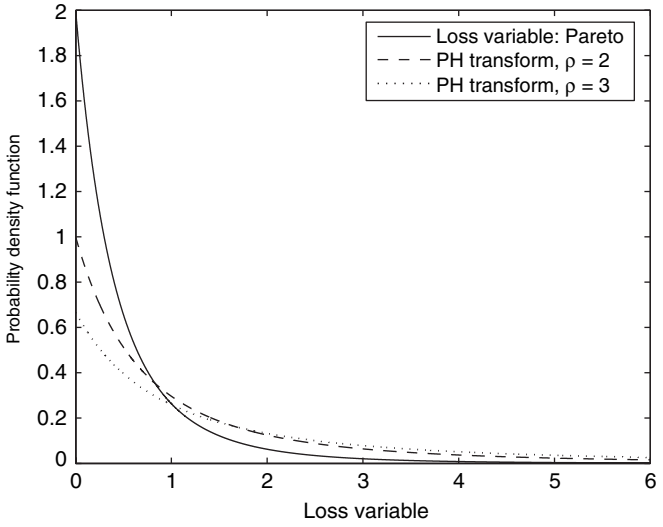

Figure 4.3 Probability density functions of $\mathcal{P}(4, 2)$ and its PH transforms □

It can be shown that $\mu_{\tilde{X}}$ as a risk measure satisfies the properties of positive homogeneity, monotonicity, and translational invariance. It also has the property of no ripoff. Readers are invited to prove these results (see Exercise 4.4). By

virtue of Theorem 4.1 presented later, it is also subadditive. Hence, the PH risk measure is coherent.

4.5.2 Esscher transform and risk-adjusted premium

The PH transform puts more weights on the right-hand tail of the loss distribution through the transformed sf. An alternative method to shift the weights to the right is to transform the pdf directly. Thus, if X has pdf $f_X(x)$, we may define a loss distribution $\tilde{X}$ with pdf $f_{\tilde{X}}(x)$ by

$$f_{\tilde{X}}(x) = w(x) f_X(x). \tag{4.31}$$

To put more weights on the right-hand tail of the loss distribution, we require $w'(x)$ to be positive and, in addition, $f_{\tilde{X}}(x)$ must be a well-defined pdf. Thus, we consider the following weighting function

$$w(x) = \frac{e^{\rho x}}{M_X(\rho)} = \frac{e^{\rho x}}{\int_0^\infty e^{\rho x} f_X(x)\,dx}, \qquad \rho > 0, \tag{4.32}$$

where

$$M_X(\rho) = \int_0^\infty e^{\rho x} f_X(x)\,dx = \mathrm{E}(e^{\rho X}) \tag{4.33}$$

is the mgf of X. It is easy to see that

$$w'(x) = \frac{\rho e^{\rho x}}{M_X(\rho)} > 0, \qquad \rho > 0, \tag{4.34}$$

and that

$$\int_0^\infty f_{\tilde{X}}(x)\,dx = \int_0^\infty w(x) f_X(x)\,dx = \int_0^\infty \left[\frac{e^{\rho x}}{\int_0^\infty e^{\rho x} f_X(x)\,dx}\right] f_X(x)\,dx = 1. \tag{4.35}$$

Thus

$$f_{\tilde{X}}(x) = \frac{e^{\rho x} f_X(x)}{\int_0^\infty e^{\rho x} f_X(x)\,dx} = \frac{e^{\rho x} f_X(x)}{M_X(\rho)}, \qquad \rho > 0, \tag{4.36}$$

is a well-defined pdf. The distribution of $\tilde{X}$ defined by the pdf in equation (4.36) is called the **Esscher transform** of X with parameter ρ. A risk measure based on the premium principle can be constructed as the expected value of the

Esscher transform of X, i.e. the risk-adjusted premium. Specifically, we define the Esscher premium as

$$\varrho(X) = \mathrm{E}(\tilde{X}) = \mu_{\tilde{X}} = \int_0^\infty x f_{\tilde{X}}(x)\,dx = \frac{\int_0^\infty x e^{\rho x} f_X(x)\,dx}{M_X(\rho)} = \frac{\mathrm{E}(Xe^{\rho X})}{\mathrm{E}(e^{\rho X})}. \tag{4.37}$$

It can be shown that $d\varrho(X)/d\rho \geq 0$ (see Denuit *et al.*, 2005, Section 2.5.5) so that ρ can be interpreted as the risk-aversion index. To identify the distribution of $\tilde{X}$, we may use its mgf, which is given by

$$M_{\tilde{X}}(t) = \mathrm{E}(e^{t\tilde{X}}) = \int_0^\infty e^{tx} f_{\tilde{X}}(x)\,dx = \frac{\int_0^\infty e^{tx} e^{\rho x} f_X(x)\,dx}{M_X(\rho)} = \frac{M_X(\rho + t)}{M_X(\rho)}. \tag{4.38}$$

Example 4.9 If $X \sim \mathcal{E}(\lambda)$, calculate the Esscher transform of X with parameter $\rho \in (0, \lambda)$ and the risk-adjusted premium.

Solution From equation (2.26), the mgf $M_X(\rho)$ of X is

$$M_X(\rho) = \frac{\lambda}{\lambda - \rho},$$

so that the mgf $M_{\tilde{X}}(t)$ of the Esscher transform $\tilde{X}$ with parameter ρ is

$$M_{\tilde{X}}(t) = \frac{M_X(\rho + t)}{M_X(\rho)} = \frac{\lambda - \rho}{\lambda - \rho - t}.$$

Thus, $\tilde{X} \sim \mathcal{E}(\lambda - \rho)$. The risk-adjusted premium is

$$\varrho(X) = \mu_{\tilde{X}} = \frac{1}{\lambda - \rho} > \frac{1}{\lambda} = \mu_X. \qquad \square$$

We conclude this section by stating that the Esscher premium is translational invariant and does not allow ripoff. However, this risk measure is not positively homogeneous and violates monotonicity. Thus, it is not coherent. Readers are invited to prove these results (see Exercise 4.5).

4.6 Distortion-function approach

The **distortion function** is a mathematical device to construct risk measures. We define below a distortion function and its associated risk measure, and then show that some of the risk measures we have discussed belong to this class of risk measures.

Definition 4.2 A distortion function is a nondecreasing function $g(\cdot)$ satisfying $g(1) = 1$ and $g(0) = 0$.

Suppose X is a loss random variable with sf $S_X(x)$. As the distortion function $g(\cdot)$ is nondecreasing and $S_X(\cdot)$ is nonincreasing, $g(S_X(x))$ is a nonincreasing function of x. This can be seen by noting that the derivative of $g(S_X(x))$ is (assuming $g(\cdot)$ is differentiable)

$$\frac{dg(S_X(x))}{dx} = g'(S_X(x))S'_X(x) \leq 0. \tag{4.39}$$

Together with the property that $g(S_X(0)) = g(1) = 1$ and $g(S_X(\infty)) = g(0) = 0$, $g(S_X(x))$ is a well-defined sf over the support $[0, \infty)$. We denote the random variable with this sf as $\tilde{X}$, which may be interpreted as a risk-adjusted loss random variable, and $g(S_X(x))$ is the risk-adjusted sf.

We further assume that $g(\cdot)$ is concave down (i.e. $g''(x) \leq 0$ if the derivative exists), then the pdf of $\tilde{X}$ is

$$f_{\tilde{X}}(x) = -\frac{dg(S_X(x))}{dx} = g'(S_X(x))f_X(x), \tag{4.40}$$

and we note that

$$\frac{dg'(S_X(x))}{dx} = g''(S_X(x))S'_X(x) \geq 0. \tag{4.41}$$

Thus, $g'(S_X(x))$ is nondecreasing. Comparing equation (4.40) with equation (4.31), we can interpret $g'(S_X(x))$ as the weighting function to *scale up* the pdf of the loss at the right-hand tail.

Definition 4.3 Let X be a nonnegative loss random variable. The distortion risk measure based on the distortion function $g(\cdot)$, denoted by $\varrho(X)$, is defined as

$$\varrho(X) = \int_0^\infty g(S_X(x))\,dx. \tag{4.42}$$

Thus, the distortion risk measure $\varrho(X)$ is the mean of the risk-adjusted loss $\tilde{X}$. The class of distortion risk measures includes the following measures we have discussed:[8]

Pure premium risk measure

This can be seen easily by defining

$$g(u) = u, \tag{4.43}$$

[8] More examples of distortion risk measures can be found in Wirch and Hardy (1999).

which satisfies the conditions $g(0) = 0$ and $g(1) = 1$, and $g(\cdot)$ is nondecreasing. Now

$$\varrho(X) = \int_0^\infty g(S_X(x))\,dx = \int_0^\infty S_X(x)\,dx = \mu_X, \tag{4.44}$$

which is the pure premium risk measure.

Proportional hazard risk-adjusted premium risk measure

This can be seen by defining

$$g(u) = u^{\frac{1}{\rho}}, \qquad \rho \geq 1. \tag{4.45}$$

VaR risk measure

For VaR_δ we define the distortion function as

$$g(S_X(x)) = \begin{cases} 0, & \text{for } 0 \leq S_X(x) < 1-\delta, \\ 1, & \text{for } 1-\delta \leq S_X(x) \leq 1, \end{cases} \tag{4.46}$$

which is equivalent to

$$g(S_X(x)) = \begin{cases} 0, & \text{for } x > \text{VaR}_\delta, \\ 1, & \text{for } 0 \leq x \leq \text{VaR}_\delta. \end{cases} \tag{4.47}$$

Hence

$$\varrho(X) = \int_0^\infty g(S_X(x))\,dx = \int_0^{\text{VaR}_\delta} dx = \text{VaR}_\delta. \tag{4.48}$$

CTE risk measure

For CTE_δ we define the distortion function as (subject to the condition X is continuous)

$$g(S_X(x)) = \begin{cases} \dfrac{S_X(x)}{1-\delta}, & \text{for } 0 \leq S_X(x) < 1-\delta, \\ \\ 1, & \text{for } 1-\delta \leq S_X(x) \leq 1, \end{cases} \tag{4.49}$$

which is equivalent to

$$g(S_X(x)) = \begin{cases} \dfrac{S_X(x)}{1-\delta}, & \text{for } x > x_\delta, \\ \\ 1, & \text{for } 0 \leq x \leq x_\delta. \end{cases} \tag{4.50}$$

Hence

$$\varrho(X) = \int_0^\infty g(S_X(x))\,dx = \int_0^{x_\delta} dx + \int_{x_\delta}^\infty \frac{S_X(x)}{1-\delta}\,dx$$
$$= x_\delta + \int_{x_\delta}^\infty \frac{S_X(x)}{1-\delta}\,dx. \tag{4.51}$$

Now using integration by parts, we have

$$\int_{x_\delta}^\infty S_X(x)\,dx = xS_X(x)]_{x_\delta}^\infty + \int_{x_\delta}^\infty xf_X(x)\,dx$$
$$= -x_\delta(1-\delta) + \int_{x_\delta}^\infty xf_X(x)\,dx. \tag{4.52}$$

Substituting equation (4.52) into equation (4.51), we obtain

$$\varrho(X) = \frac{1}{1-\delta}\int_{x_\delta}^\infty xf_X(x)\,dx, \tag{4.53}$$

which is equal to CTE_δ by equation (4.17).

Risk measures based on distortion functions form a very important class. This approach is very easy to use to create new risk measures. More importantly, this class of risk measures has very desirable properties, as summarized in the following theorem.

Theorem 4.1 *Let $g(\cdot)$ be a concave-down distortion function. The risk measure of the loss X defined in equation (4.42) is translational invariant, monotonic, positively homogeneous, and subadditive, and is thus coherent.*[9]

Proof See Denuit *et al.* (2005, Section 2.6.2.2). □

4.7 Wang transform

Wang (2000) proposed the following distortion function

$$g(u) = \Phi\left[\Phi^{-1}(u) + \rho\right], \tag{4.54}$$

where $\Phi(\cdot)$ is the df of the standard normal and ρ is the risk parameter taking positive values. Note that $\Phi(\cdot)$ is only used to define the transform and no normality assumption is made for the loss distribution. Equation (4.54) is known as the Wang transform.

[9] Note that VaR_δ is a step function and is thus not concave-down. In contrast, CTE_δ is concave-down. See Wirch and Hardy (1999).

We can easily verify that $g(0) = 0$ and $g(1) = 1$. In addition, denoting $\phi(\cdot)$ as the pdf of the standard normal and $x = \Phi^{-1}(u)$, we have

$$\frac{dg(u)}{du} = \frac{\phi(x+\rho)}{\phi(x)} = \exp\left(-\rho x - \frac{\rho^2}{2}\right) > 0, \tag{4.55}$$

and

$$\frac{d^2 g(u)}{du^2} = -\frac{\rho\phi(x+\rho)}{[\phi(x)]^2} < 0. \tag{4.56}$$

Thus, the Wang transform is increasing and concave down. Denoting $\tilde{X}$ as the Wang-transformed variable of the loss distribution X, the risk measure of X based on the Wang transform is defined as the risk-adjusted premium

$$\varrho(X) = \mathrm{E}(\tilde{X}) = \int_0^\infty \Phi\left[\Phi^{-1}(S_X(x)) + \rho\right] dx. \tag{4.57}$$

It can be seen that

$$\frac{d\varrho(X)}{d\rho} = \int_0^\infty \phi\left[\Phi^{-1}(S_X(x)) + \rho\right] dx > 0. \tag{4.58}$$

This implies the risk measure $\varrho(X)$ increases with ρ, which represents the risk aversion.

Example 4.10 If $X \sim \mathcal{N}(\mu, \sigma^2)$, find the distribution of the loss under the Wang transform, and the risk-adjusted premium.

Solution The sf of X is

$$S_X(x) = 1 - \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(-\frac{x-\mu}{\sigma}\right).$$

The sf of the Wang-transformed variable $\tilde{X}$ is

$$\begin{aligned} S_{\tilde{X}}(x) &= g(S_X(x)) \\ &= \Phi\left[\Phi^{-1}(S_X(x)) + \rho\right] \\ &= \Phi\left[\Phi^{-1}\left(\Phi\left(-\frac{x-\mu}{\sigma}\right)\right) + \rho\right] \end{aligned}$$

$$
\begin{aligned}
&= \Phi\left[\left(-\frac{x-\mu}{\sigma}\right)+\rho\right] \\
&= \Phi\left[-\frac{x-(\mu+\rho\sigma)}{\sigma}\right] \\
&= 1-\Phi\left[\frac{x-(\mu+\rho\sigma)}{\sigma}\right].
\end{aligned}
$$

Thus, $\tilde{X} \sim \mathcal{N}(\mu+\rho\sigma, \sigma^2)$ and the risk-adjusted premium is $\varrho(X) = \mathrm{E}(\tilde{X}) = \mu+\rho\sigma$. □

Example 4.11 If $X \sim \mathcal{L}(\mu, \sigma^2)$, find the distribution of the loss under the Wang transform, and the risk-adjusted premium.

Solution The sf of X is

$$
S_X(x) = 1-\Phi\left[\frac{\log x-\mu}{\sigma}\right] = \Phi\left[-\frac{\log x-\mu}{\sigma}\right].
$$

The sf of the Wang-transformed variable $\tilde{X}$ is

$$
\begin{aligned}
S_{\tilde{X}}(x) &= g(S_X(x)) \\
&= \Phi\left[\Phi^{-1}\left(\Phi\left[-\frac{\log x-\mu}{\sigma}\right]\right)+\rho\right] \\
&= \Phi\left[-\frac{\log x-\mu}{\sigma}+\rho\right] \\
&= 1-\Phi\left[\frac{\log x-(\mu+\rho\sigma)}{\sigma}\right].
\end{aligned}
$$

Thus, $\tilde{X} \sim \mathcal{L}(\mu+\rho\sigma, \sigma^2)$ and the risk-adjusted premium is

$$
\varrho(X) = \mathrm{E}(\tilde{X}) = \exp\left(\mu+\rho\sigma+\frac{\sigma^2}{2}\right). \qquad \square
$$

Examples 4.10 and 4.11 show that the Wang-transformed loss remains in the same family of the original loss distribution for the case of normal and lognormal losses. Another advantage of the Wang transform is that it can be applied to measure risks of assets as well, in which case ρ will take negative values. More details of this can be found in Wang (2000).

4.8 Summary and conclusions

We have discussed the uses of risk measures for insurance business. The risk measures may be constructed for the determination of economic capital or

for the setting of insurance premium. The four axioms of desirable properties proposed by Artzner *et al.* (1999) help to narrow down the choice of risk measures, although they do not specifically identify a unique choice. Risk measures satisfying these axioms are said to be coherent. A class of risk measures that is coherent is constructed based on concave-down distortion functions. This class of risk measures includes the conditional tail expectation, the PH transform risk-adjusted premium, and the Wang transform risk-adjusted premium. On the other hand, the commonly used risk measure Value-at-Risk is not coherent.

Many distortion functions depend on a risk-aversion parameter, which in turn determines the risk-adjusted premium. Theory is often unable to identify the risk-aversion parameter. Some recent works (Jones *et al.*, 2006; Jones and Zitikis, 2007) consider the estimation of the risk-aversion parameter, and test the equality of the risk measures.

Quantile-based risk measures have received much attention since the emergence of VaR. In the actuarial science and risk management literature, however, the terminologies for quantile-based risk measures such as VaR, CVaR, and CTE have not been standardized. Further details and extensions can be found in Denuit *et al.* (2005) and Dowd and Blake (2006).

Exercises

4.1 When the loss random variable X is not continuous, show that $\mathrm{CTE}_\delta(X)$ computed according to equation (4.20) is larger than or equal to $\mathrm{TVaR}_\delta(X)$ given in equation (4.23).

4.2 Prove that the risk measure based on the variance premium satisfies Axiom T, but not Axioms S, M, and PH.

4.3 Prove that the risk measure based on the standard-deviation premium satisfies Axioms S, T, and PH, but not Axiom M.

4.4 Prove that the risk measure based on the expected value of the PH transform satisfies Axioms PH, M, and T. It also has the property of no ripoff.

4.5 Prove that the risk measure based on the Esscher premium satisfies Axioms T, but not Axiom PH.

4.6 Let X be a nonnegative loss random variable distributed as $\mathcal{G}(\alpha, \beta)$. Find the Esscher premium risk measure with risk parameter ρ, where $\rho \in [0, 1/\beta)$.

4.7 Let $X \sim \mathcal{W}(\alpha, \lambda)$. Determine $\mathrm{VaR}_\delta(X)$ and the PH premium with parameter ρ.

4.8 Suppose $c \in (0, 1)$ and the df of the loss random variable X is $F_X(x) = cx$ for $x \in [0, 1)$ and 1 when $x \geq 1$.

(a) Plot the df of X.
(b) Compute $\mathrm{E}(X)$.
(c) Determine $\mathrm{VaR}_\delta(X)$ for $\delta \in (0, 1)$.
(d) Determine $\mathrm{CTE}_\delta(\mathrm{X})$ for $\delta \in (0, c)$.

4.9 The loss random variable X has the following pf:

x	$f_X(x)$
60	0.04
50	0.04
40	0.22
30	0.30
20	0.30
10	0.10

Calculate $\mathrm{VaR}_\delta(X)$ for $\delta = 0.90$, 0.95, and 0.99, and $\mathrm{CTE}_\delta(X)$ for $\delta = 0.90$ and 0.95.

4.10 Calculate the risk-adjusted premium risk measure of the PH transform with risk-aversion index ρ for the following loss distributions
(a) $\mathcal{U}(0, 2b)$,
(b) $\mathcal{E}(1/b)$,
(c) $\mathcal{P}(2, b)$,
where $b > 0$ and $1 < \rho < 2$. Note that all the above distributions have expected loss b. Now consider $\rho = 1.2, 1.5$, and 1.8, and compute the PH premium for the above loss distributions as a function of b. Comment on your results.

4.11 Let $X \sim \mathcal{U}(0, b)$.
(a) Determine the pure premium and the expected-value premium with loading θ.
(b) Determine the variance premium and standard-deviation premium with loading α.
(c) Determine $\mathrm{VaR}_\delta(X)$, and calculate $\mathrm{CTE}_\delta(X)$ using equation (4.19).
(d) Calculate $\mathrm{CTE}_\delta(X)$ using equation (4.17) and verify that the answer is the same as in (c).
(e) Calculate $\mathrm{CVaR}_\delta(X)$ using equation (4.13) and $\mathrm{TVaR}_\delta(X)$ using equation (4.23).

4.12 Let $X \sim \mathcal{E}(\lambda)$.
(a) Determine the pure premium and the expected-value premium with loading θ.

(b) Determine the variance premium and standard-deviation premium with loading α.

(c) Determine $\text{VaR}_\delta(X)$, and calculate $\text{CTE}_\delta(X)$ using equation (4.19).

(d) Calculate $\text{CTE}_\delta(X)$ using equation (4.17) and verify that the answer is the same as in (c).

(e) Calculate $\text{CVaR}_\delta(X)$ using equation (4.13) and $\text{TVaR}_\delta(X)$ using equation (4.23).

4.13 If loss follows a compound Poisson distribution with parameter λ and $\mathcal{G}(\alpha, \beta)$ as the secondary distribution, determine the expected-value premium with a loading of 20%. If the same premium is charged under the variance principle, what is the loading used?

4.14 If loss follows a compound Poisson distribution with parameter λ and $\mathcal{G}(\alpha, \beta)$ as the secondary distribution, determine the Esscher premium with parameter ρ.

4.15 Losses X and Y have the following distribution

$$\Pr(X = 0, Y = 0) = \Pr(X = 0, Y = 3) = \Pr(X = 6, Y = 6) = \frac{1}{3}.$$

Show that $\Pr(X \leq Y) = 1$, but that the Esscher premium of X with parameter 0.5 is larger than that of Y.

4.16 The losses in two portfolios, denoted by P_1 and P_2, have identical distributions of 100 with probability 4% and zero with probability 96%. Suppose P_1 and P_2 are independent.

(a) Determine $\text{VaR}_{0.95}(P_1)$ and $\text{VaR}_{0.95}(P_2)$.

(b) Determine the distribution of $P_1 + P_2$.

(c) Calculate $\text{VaR}_{0.95}(P_1 + P_2)$ and $\text{VaR}_{0.95}(P_1) + \text{VaR}_{0.95}(P_2)$. Comment on your results.

4.17 The loss random variable X has the following pf:

x	$f_X(x)$
500	0.08
400	0.12
300	0.35
200	0.45

Calculate $\text{VaR}_{0.9}(X)$, $\text{CTE}_{0.9}(X)$, $\text{CVaR}_{0.9}(X)$, and $\text{TVaR}_{0.9}(X)$.

4.18 Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Show that

$$\text{E}[(X - x_\delta)_+] = \sigma\phi\left(\Phi^{-1}(\delta)\right) - \sigma\Phi^{-1}(\delta)(1 - \delta),$$

for $0 < \delta < 1$. Note that the expression on the right-hand side depends on σ and δ only.

4.19 Let $X \sim \mathcal{L}(\mu, \sigma^2)$. Show that

$$\mathrm{E}[(X - x_\delta)_+] = \exp\left(\mu + \frac{\sigma^2}{2}\right) \Phi\left(\sigma - \Phi^{-1}(\delta)\right) - \exp\left(\mu + \sigma\Phi^{-1}(\delta)\right)(1 - \delta),$$

for $0 < \delta < 1$.

4.20 U and V are two loss distributions. The sf of U is $S_U(x) = 0.25$ for $0 \le x < 4$ and zero for $4 \le x$. The sf of V is $S_V(x) = (2/(2 + x))^3$, for $x \ge 0$.

(a) Show that $\mathrm{E}(U) = \mathrm{E}(V) = 1$, and $\mathrm{Var}(U) = \mathrm{Var}(V) = 3$.

(b) Determine the PH risk-adjusted premium of U and V with parameter $\rho < 3$.

5

Ruin theory

We consider models for analyzing the surplus of an insurance portfolio. Suppose an insurance business begins with a start-up capital, called the initial surplus. The insurance company receives premium payments and pays claim losses. The premium payments are assumed to be coming in at a constant rate. When there are claims, losses are paid out to policy holders. Unlike the constant premium payments, losses are random and uncertain, in both timing and amount. The net surplus through time is the excess of the initial capital and aggregate premiums received over the losses paid out. The insurance business is in ruin if the surplus falls to or below zero. The main purpose of this chapter is to consider the probability of ruin as a function of time, the initial surplus and the claim distribution. Ultimate ruin refers to the situation where ruin occurs at finite time, irrespective of the time of occurrence.

We first consider the situation in which premium payments and claim losses occur at discrete time. We derive recursive formulas for the probability of ultimate ruin given the initial surplus. These recursive formulas require the value of the probability of ultimate ruin when the start-up capital is zero. Formulas for the probability of ruin before fixed finite times are also derived. To obtain bounds for the probability of ultimate ruin, we introduce Lundberg's inequality. In the continuous-time set-up, we assume the claims follow a Poisson process. Lundberg's bound for the probability of ultimate ruin in continuous time is then presented.

Ruin probabilities are measures of risk. They express risk as a dynamic process and relate ruin to the counteracting factors of premium rates and claim-loss distributions. Our discussions, however, will be restricted only to discrete-time models and the Poisson process in the continuous time.

Learning objectives

1 Surplus function, premium rate, and loss process
2 Probability of ultimate ruin

3 Probability of ruin before a finite time
4 Adjustment coefficient and Lundberg's inequality
5 Poisson process and continuous-time ruin theory

5.1 Discrete-time surplus and events of ruin

We assume that an insurance company establishes its business with a start-up capital of u at time 0, called the **initial surplus**. It receives premiums of one unit per period at the end of each period. Loss claim of amount X_i is paid out at the end of period i for $i = 1, 2, \cdots$. We assume X_i are independently and identically distributed as the loss random variable X. Thus, we have a discrete-time model in which the **surplus** at time n with initial capital u, denoted by $U(n; u)$, is given by[1]

$$U(n; u) = u + n - \sum_{i=1}^{n} X_i, \quad \text{for } n = 1, 2, \cdots. \tag{5.1}$$

Note that the *numeraire* of the above equation is the amount of premium per period, or the premium rate. All other variables are measured as multiples of the premium rate. Thus, the initial surplus u may take values of $0, 1, \cdots$, times the premium rate. Likewise, X_i may take values of j times the premium rate with pf $f_X(j)$ for $j = 0, 1, \cdots$. We denote the mean of X by μ_X and its variance by σ_X^2. Furthermore, we assume X is of finite support, although in notation we allow j to run to infinity.

If we denote the premium loading by θ, then we have

$$1 = (1 + \theta)\mu_X, \tag{5.2}$$

which implies

$$\mu_X = \frac{1}{1 + \theta}. \tag{5.3}$$

We shall assume positive loading so that $\mu_X < 1$. Also, we have a model in which expenses and the time value of money are not considered. The business is said to be in **ruin** if the surplus function $U(n; u)$ falls to or below zero sometime after the business started, i.e. at a point $n \geq 1$. Specifically, we have the following definition of ruin.

[1] We may regard the premium and the loss as the total amount received and paid, respectively, over each period. Surplus is only computed at the end of each period, and equation (5.1) is appropriate as there is no interest in our model.

Definition 5.1 Ruin occurs at time n if $U(n;u) \leq 0$ for the first time at n, for $n \geq 1$.

Note that an insurance business may begin with zero start-up capital. According to the above definition, the insurance business is not in ruin at time 0 even if the initial surplus is zero. A main purpose of the model is to analyze the surplus and the probability of ruin. Given the initial surplus u, we define $T(u)$ as the **time of ruin** as follows.

Definition 5.2 The time-of-ruin random variable $T(u)$ is defined as

$$T(u) = \min\{n \geq 1 : U(n;u) \leq 0\}. \tag{5.4}$$

As long as there exists a finite n such that $U(n;u) \leq 0$, the event of ruin has occurred. However, for some realizations, such a finite value of n may not exist, in which case ruin does not occur. Thus, $T(u)$ may not have a finite value and is, in this sense, an *improper* random variable. A key interest in analyzing the surplus function is to find the probability that $T(u)$ has a finite value, i.e. the **probability of ultimate ruin**.

Definition 5.3 Given an initial surplus u, the probability of ultimate ruin, denoted by $\psi(u)$, is

$$\psi(u) = \Pr(T(u) < \infty). \tag{5.5}$$

Apart from the probability of ultimate ruin, it may also be of interest to find the probability of ruin at or before a finite time. We define the probability of ruin at or before a finite time as follows.

Definition 5.4 Given an initial surplus u, the probability of ruin by time t, denoted by $\psi(t;u)$, is

$$\psi(t;u) = \Pr(T(u) \leq t), \quad \text{for } t = 1, 2, \cdots. \tag{5.6}$$

Both $\psi(u)$ and $\psi(t;u)$ are important measures of the risks of ruin. In the following section, we present some recursive methods for the computation of these functions and some bounds for their values in discrete time.

5.2 Discrete-time ruin theory

In this section we first derive recursive formulas for the computation of the probability of ultimate ruin. We then consider the calculation of the probability of ruin by a finite time, and finally we derive the Lundberg inequality in discrete time.

5.2.1 Ultimate ruin in discrete time

We first consider the probability of ultimate ruin when the initial surplus is 0, i.e. $\psi(0)$. At time 1, if there is no claim, which occurs with probability $f_X(0)$, then the surplus accumulates to 1 and the probability of ultimate ruin is $\psi(1)$. On the other hand, if $X_1 \geq 1$, which occurs with probability $S_X(0) = 1 - F_X(0) = \Pr(X \geq 1)$, then the business ends in ruin. Thus, we have

$$\psi(0) = f_X(0)\psi(1) + S_X(0), \tag{5.7}$$

from which $\psi(1)$ can be computed given $\psi(0)$. Similarly, for $u = 1$, we have

$$\psi(1) = f_X(0)\psi(2) + f_X(1)\psi(1) + S_X(1). \tag{5.8}$$

The above equation can be generalized to larger values of u as follows

$$\psi(u) = f_X(0)\psi(u+1) + \sum_{j=1}^{u} f_X(j)\psi(u+1-j) + S_X(u), \quad \text{for } u \geq 1. \tag{5.9}$$

Re-arranging equation (5.9), we obtain the following recursive formula for the probability of ultimate ruin

$$\psi(u+1) = \frac{1}{f_X(0)}\left[\psi(u) - \sum_{j=1}^{u} f_X(j)\psi(u+1-j) - S_X(u)\right], \quad \text{for } u \geq 1. \tag{5.10}$$

To apply the above equation we need the starting value $\psi(0)$, which is given by the following theorem.

Theorem 5.1 *For the discrete-time surplus model,* $\psi(0) = \mu_X$.

Proof We first re-arrange equation (5.7) to obtain

$$f_X(0)\left[\psi(1) - \psi(0)\right] = \left[1 - f_X(0)\right]\psi(0) - S_X(0), \tag{5.11}$$

and equation (5.9) to obtain

$$f_X(0)\left[\psi(u+1) - \psi(u)\right] = \left[1 - f_X(0)\right]\psi(u) - \sum_{j=1}^{u} f_X(j)\psi(u+1-j) - S_X(u), \quad \text{for } u \geq 1. \tag{5.12}$$

Now adding equations (5.11) and (5.12) for $u = 0, \cdots, z$, we obtain

$$f_X(0)\sum_{u=0}^{z}[\psi(u+1)-\psi(u)] = [1-f_X(0)]\sum_{u=0}^{z}\psi(u) - \sum_{u=1}^{z}\sum_{j=1}^{u}f_X(j)\psi(u+1-j) - \sum_{u=0}^{z}S_X(u), \tag{5.13}$$

the left-hand side of which can be written as

$$f_X(0)\left[\psi(z+1)-\psi(0)\right]. \tag{5.14}$$

We now simplify the second term on the right-hand side of equation (5.13) as

$$\begin{aligned}\sum_{u=1}^{z}\sum_{j=1}^{u}f_X(j)\psi(u+1-j) &= \sum_{u=1}^{z}\sum_{r=1}^{u}f_X(u+1-r)\psi(r)\\ &= \sum_{r=1}^{z}\psi(r)\sum_{u=r}^{z}f_X(u+1-r)\\ &= \sum_{r=1}^{z}\psi(r)\sum_{u=1}^{z+1-r}f_X(u),\end{aligned} \tag{5.15}$$

where in the first line above we have applied the change of index $r = u+1-j$, and in the second line we have reversed the order of the summation indexes u and r. Now substituting equation (5.15) into (5.13), we have

$$\begin{aligned}f_X(0)\left[\psi(z+1)-\psi(0)\right] &= [1-f_X(0)]\sum_{u=0}^{z}\psi(u)\\ &\quad -\sum_{r=1}^{z}\psi(r)\sum_{u=1}^{z+1-r}f_X(u) - \sum_{u=0}^{z}S_X(u)\\ &= [1-f_X(0)]\,\psi(0) + \sum_{r=1}^{z}\psi(r)\\ &\quad \times\left[1-\sum_{u=0}^{z+1-r}f_X(u)\right] - \sum_{u=0}^{z}S_X(u)\end{aligned}$$

$$= [1 - f_X(0)]\,\psi(0) + \sum_{r=1}^{z} \psi(r) S_X(z+1-r)$$

$$- \sum_{u=0}^{z} S_X(u). \tag{5.16}$$

Now $\psi(z+1) \to 0$ as $z \to \infty$, and as X is of finite support, we have

$$\sum_{r=1}^{z} \psi(r) S_X(z+1-r) \to 0 \tag{5.17}$$

when $z \to \infty$. Thus, when $z \to \infty$, we conclude from equation (5.16) that

$$-f_X(0)\psi(0) = [1 - f_X(0)]\,\psi(0) - \sum_{u=0}^{\infty} S_X(u), \tag{5.18}$$

so that

$$\psi(0) = \sum_{u=0}^{\infty} S_X(u) = \mu_X, \tag{5.19}$$

where the last equality of the above equation is due to the discrete analogue of equation (2.18). □

Example 5.1 The claim variable X has the following distribution: $f_X(0) = 0.5$, $f_X(1) = f_X(2) = 0.2$, and $f_X(3) = 0.1$. Calculate the probability of ultimate ruin $\psi(u)$ for $u \geq 0$.

Solution The survival function of X is $S_X(0) = 0.2+0.2+0.1 = 0.5$, $S_X(1) = 0.2 + 0.1 = 0.3$, $S_X(2) = 0.1$, and $S_X(u) = 0$ for $u \geq 3$. The mean of X is

$$\mu_X = (0)(0.5) + (1)(0.2) + (2)(0.2) + (3)(0.1) = 0.9,$$

which can also be calculated as

$$\mu_X = \sum_{u=0}^{\infty} S_X(u) = 0.5 + 0.3 + 0.1 = 0.9.$$

Thus, from Theorem 5.1 $\psi(0) = 0.9$, and from equation (5.7), $\psi(1)$ is given by

$$\psi(1) = \frac{\psi(0) - S_X(0)}{f_X(0)} = \frac{0.9 - 0.5}{0.5} = 0.8.$$

From equation (5.8), we have

$$\psi(2) = \frac{\psi(1) - f_X(1)\psi(1) - S_X(1)}{f_X(0)} = \frac{0.8 - (0.2)(0.8) - 0.3}{0.5} = 0.68,$$

and applying equation (5.10) for $u = 2$, we have

$$\psi(3) = \frac{\psi(2) - f_X(1)\psi(2) - f_X(2)\psi(1) - S_X(2)}{f_X(0)} = 0.568.$$

As $S_X(u) = 0$ for $u \geq 3$, using equation (5.10) we have, for $u \geq 3$

$$\psi(u+1) = \frac{\psi(u) - f_X(1)\psi(u) - f_X(2)\psi(u-1) - f_X(3)\psi(u-2)}{f_X(0)}.$$

Using the above recursive equation we compute $\psi(u)$ for u up to 30. The results are plotted in Figure 5.1, from which it is clear that $\psi(u)$ is a monotonic decreasing function of u. To obtain a probability of ultimate ruin of less than 1%, the initial surplus must be at least 26.

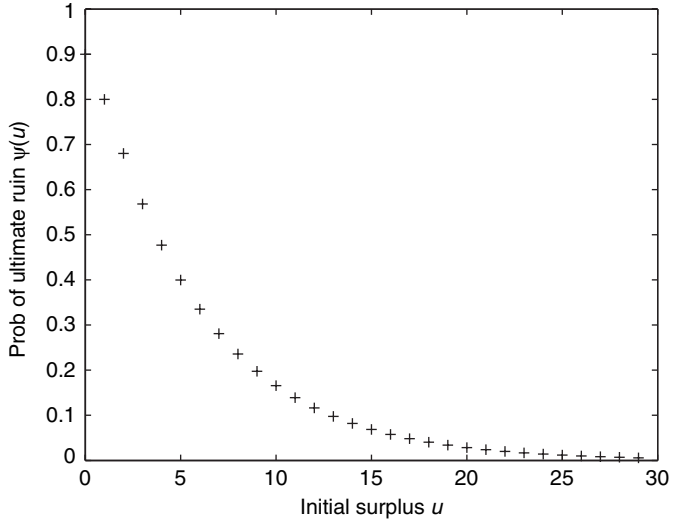

Figure 5.1 Probability of ultimate ruin in Example 5.1

□

Example 5.2 The claim variable X can take values 0 or 2 with the following probabilities: $f_X(0) = p$ and $f_X(2) = q = 1 - p$, where $p > 0.5$. Calculate the probability of ultimate ruin $\psi(u)$ for $u \geq 0$.

Solution The survival function of X is $S_X(0) = S_X(1) = q$ and $S_X(u) = 0$ for $u \geq 2$. Thus, $\psi(0) = \mu_X = S_X(0) + S_X(1) = 2q < 1$. For $u = 1$, we have, from equation (5.7)

$$\psi(1) = \frac{\psi(0) - S_X(0)}{f_X(0)} = \frac{2q - q}{p} = \frac{q}{p}.$$

When $u = 2$, we apply equation (5.10) to obtain

$$\begin{aligned}\psi(2) &= \frac{\psi(1) - S_X(1)}{f_X(0)} \\ &= \frac{1}{p}\left(\frac{q}{p} - q\right) \\ &= \left(\frac{q}{p}\right)^2.\end{aligned}$$

To derive a general formula for $\psi(u)$, we observe that

$$\psi(u) = \left(\frac{q}{p}\right)^u$$

holds for $u = 1$ and 2. Assuming the formula holds for $u - 1$ and u with $u \geq 2$, we can show that it also holds for $u + 1$. To do this we apply equation (5.10) to obtain

$$\begin{aligned}\psi(u+1) &= \frac{1}{f_X(0)}\left[\psi(u) - f_X(2)\psi(u-1)\right] \\ &= \frac{1}{p}\left[\left(\frac{q}{p}\right)^u - q\left(\frac{q}{p}\right)^{u-1}\right] \\ &= \frac{1}{p}\left(\frac{q}{p}\right)^{u-1}\left(\frac{q}{p} - q\right) \\ &= \left(\frac{q}{p}\right)^{u+1}.\end{aligned}$$

Thus, the general formula is established by induction. □

5.2.2 Finite-time ruin in discrete time

We now consider the probability of ruin at or before a finite time point t given an initial surplus u. First we consider $t = 1$ given initial surplus u. As defined

in equation (5.6), $\psi(t;u) = \Pr(T(u) \le t)$. If $u = 0$, the ruin event occurs at time $t = 1$ when $X_1 \ge 1$. Thus

$$\psi(1;0) = 1 - f_X(0) = S_X(0). \tag{5.20}$$

Likewise, for $u > 0$, we have

$$\psi(1;u) = \Pr(X_1 > u) = S_X(u). \tag{5.21}$$

Thus, $\psi(1;u)$ are easily obtained from equations (5.20) and (5.21) for $u \ge 0$. We now consider $\psi(t;u)$ for $t \ge 2$ and $u \ge 0$. Note that the event of ruin occurring at or before time $t \ge 2$ may be due to (a) ruin at time 1, or (b) loss of j at time 1 for $j = 0, 1, \cdots, u$, followed by ruin occurring within the next $t-1$ periods. When there is a loss of j at time 1, the surplus becomes $u + 1 - j$ at time 1, so that the probability of ruin within the next $t-1$ periods is $\psi(t-1;u+1-j)$. Thus, we conclude that

$$\psi(t;u) = \psi(1;u) + \sum_{j=0}^{u} f_X(j)\psi(t-1;u+1-j). \tag{5.22}$$

Hence, $\psi(t;u)$ can be computed as follows:

1 Construct a table with time t running down the rows for $t = 1, 2, \cdots$, and u running across the columns for $u = 0, 1, \cdots$.
2 Initialize the first row of the table for $t = 1$ with $\psi(1;u) = S_X(u)$. Note that if M is the maximum loss in each period, then $\psi(1;u) = 0$ for $u \ge M$.
3 Increase the value of t by 1 and calculate $\psi(t;u)$ for $u = 0, 1, \cdots$, using equation (5.22). Note that the computation requires the corresponding entry in the first row of the table, i.e. $\psi(1;u)$, as well as some entries in the $(t-1)$th row. In particular, the $u+1$ entries $\psi(t-1;1), \cdots, \psi(t-1;u+1)$ in the $(t-1)$th row are required.[2]
4 Re-do Step 3 until the desired time point.

The example below illustrates the computation of the probabilities.

Example 5.3 As in Example 5.1, the claim variable X has the following distribution: $f_X(0) = 0.5, f_X(1) = f_X(2) = 0.2$, and $f_X(3) = 0.1$. Calculate the probability of ruin at or before a finite time t given initial surplus u, $\psi(t;u)$, for $u \ge 0$.

Solution The results are summarized in Table 5.1 for $t = 1, 2$, and 3, and $u = 0, 1, \cdots, 6$.

[2] Note that if $f_X(j) = 0$ for $j > M$, we only require the entries $\psi(t-1;\max\{1, u+1-M\}), \cdots, \psi(t-1;u+1)$ in the $(t-1)$th row.

Table 5.1. *Results of Example 5.3*

	Initial surplus u						
Time t	0	1	2	3	4	5	6
1	0.500	0.300	0.100	0.000	0.000	0.000	0.000
2	0.650	0.410	0.180	0.050	0.010	0.000	0.000
3	0.705	0.472	0.243	0.092	0.030	0.007	0.001

The first row of the table is $S_X(u)$. Note that $\psi(1;u) = 0$ for $u \geq 3$, as the maximum loss in each period is 3. For the second row, the details of the computation are as follows. First, $\psi(2;0)$ is computed as

$$\psi(2;0) = \psi(1;0) + f_X(0)\psi(1;1) = 0.5 + (0.5)(0.3) = 0.65.$$

Similarly

$$\begin{aligned}\psi(2;1) &= \psi(1;1) + f_X(0)\psi(1;2) + f_X(1)\psi(1;1) \\ &= 0.3 + (0.5)(0.1) + (0.2)(0.3) = 0.41,\end{aligned}$$

and

$$\psi(2;2) = \psi(1;2) + f_X(0)\psi(1;3) + f_X(1)\psi(1;2) + f_X(2)\psi(1;1) = 0.18.$$

We use $\psi(3;3)$ to illustrate the computation of the third row as follows

$$\begin{aligned}\psi(3;3) &= \psi(1;3) + f_X(0)\psi(2;4) + f_X(1)\psi(2;3) \\ &\quad + f_X(2)\psi(2;2) + f_X(3)\psi(2;1) \\ &= 0 + (0.5)(0.01) + (0.2)(0.05) + (0.2)(0.18) + (0.1)(0.41) \\ &= 0.092.\end{aligned}$$

Figure 5.2 plots the probabilities of ruin given three values of initial surplus, $u = 0, 5$, and 10. □

5.2.3 Lundberg's inequality in discrete time

The recursive formulas presented in the last two sections compute the exact probability of ruin given the initial surplus. We now introduce the Lundberg inequality, which provides an upper bound for the probability of ultimate ruin as long as the mgf of the loss distribution exists. Prior to stating the Lundberg

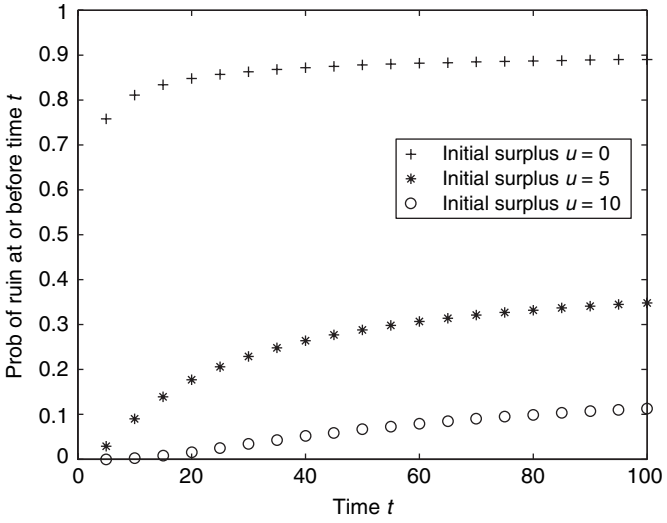

Figure 5.2 Probability of ruin by a finite time in Example 5.3

inequality, however, we first define an important quantity called the **adjustment coefficient** as follows.

Definition 5.5 Suppose X is the loss random variable. The adjustment coefficient, denoted by r^*, is the positive value of r that satisfies the following equation

$$\mathrm{E}\left[\exp\left\{r(X-1)\right\}\right] = 1. \tag{5.23}$$

Note that $\mathrm{E}\left[\exp\left\{r(X-1)\right\}\right]$ is the mgf of $X-1$ (i.e. the deficit per period) evaluated at r, or $M_{X-1}(r)$. To show that a positive root r^* exists, we first define the function $\phi(r)$ as follows

$$\phi(r) = \mathrm{E}\left[\exp\left\{r(X-1)\right\}\right]. \tag{5.24}$$

Now we note the following properties of $\phi(r)$:

1 $\phi(0) = 1$ and $\phi(r)$ is decreasing at $r = 0$. The latter result arises from the fact that

$$\phi'(r) = \mathrm{E}\left[(X-1)\exp\left\{r(X-1)\right\}\right], \tag{5.25}$$

so that

$$\phi'(0) = \mathrm{E}\left[X-1\right] = \mu_X - 1 < 0. \tag{5.26}$$

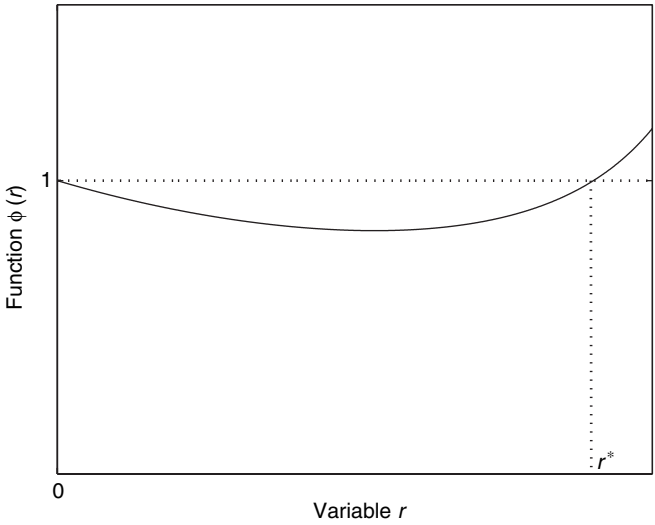

Figure 5.3 A plot of $\phi(r)$

2 $\phi(r)$ is concave upward for $r > 0$, as

$$\phi''(r) = \mathrm{E}\left[(X-1)^2 \exp\{r(X-1)\}\right] > 0, \quad \text{for } r > 0. \tag{5.27}$$

Furthermore, suppose there exits $x^* > 1$, such that $\Pr(X \geq x^*) > 0$. Then $\phi(r) \geq e^{r(x^*-1)} \Pr(X \geq x^*)$, which tends to ∞ as r tends to ∞. These results show that the typical shape of $\phi(r)$ is as in Figure 5.3, for which there is a unique value $r^* > 0$ satisfying equation $\phi(r^*) = 1$ for the existence of the adjustment coefficient.

Example 5.4 Assume the loss random variable X follows the distribution given in Examples 5.1 and 5.3. Calculate the adjustment coefficient r^*.

Solution Equation (5.23) is set up as follows

$$0.5e^{-r} + 0.2 + 0.2e^{r} + 0.1e^{2r} = 1,$$

which is equivalent to

$$0.1w^3 + 0.2w^2 - 0.8w + 0.5 = 0,$$

for $w = e^r$. We solve the above equation numerically to obtain $w = 1.1901$, so that $r^* = \log(1.1901) = 0.1740$. □

We now state Lundberg's inequality in the following theorem.

Theorem 5.2 *For the discrete-time surplus function, the probability of ultimate ruin satisfies the following inequality*

$$\psi(u) \leq \exp(-r^* u), \tag{5.28}$$

where r^* is the adjustment coefficient.

Proof We shall show that all finite-time ruin probabilities satisfy inequality (5.28), i.e.

$$\psi(t;u) \leq \exp(-r^* u), \quad \text{for } t \geq 1, \tag{5.29}$$

so that the inequality holds for the probability of ultimate ruin. We first note that $\exp[-r^*(u+1-j)] \geq 1$ for $j \geq u+1$. Thus, for $t = 1$, the probability of ruin is, from equation (5.21)

$$\begin{aligned}
\psi(1;u) &= \sum_{j=u+1}^{\infty} f_X(j) \\
&\leq \sum_{j=u+1}^{\infty} e^{-r^*(u+1-j)} f_X(j) \\
&\leq \sum_{j=0}^{\infty} e^{-r^*(u+1-j)} f_X(j) \\
&= e^{-r^* u} \sum_{j=0}^{\infty} e^{r^*(j-1)} f_X(j) \\
&= e^{-r^* u} \, \mathrm{E}\left[r^*(X-1)\right] \\
&= e^{-r^* u}.
\end{aligned} \tag{5.30}$$

To prove the result by induction, we assume inequality (5.29) holds for a $t \geq 1$. Then, for $t+1$ the probability of ruin is, from equation (5.22)

$$\psi(t+1;u) = \psi(1;u) + \sum_{j=0}^{u} \psi(t;u+1-j) f_X(j)$$

$$
\begin{aligned}
&\leq \sum_{j=u+1}^{\infty} f_X(j) + \sum_{j=0}^{u} e^{-r^*(u+1-j)} f_X(j) \\
&\leq \sum_{j=u+1}^{\infty} e^{-r^*(u+1-j)} f_X(j) + \sum_{j=0}^{u} e^{-r^*(u+1-j)} f_X(j) \\
&= \sum_{j=0}^{\infty} e^{-r^*(u+1-j)} f_X(j) \\
&= e^{-r^*u} \sum_{j=0}^{\infty} e^{r^*(j-1)} f_X(j) \\
&= e^{-r^*u} \, \mathrm{E}\left[r^*(X-1)\right] \\
&= e^{-r^*u}. \qquad (5.31)
\end{aligned}
$$

Hence, inequality (5.28) holds for all finite time. □

The upper bound of the probability of ultimate ruin e^{-r^*u} decreases with the initial surplus u, which is intuitive. The example below compares the upper bound with the exact values for the problem in Example 5.1.

Example 5.5 Assume the loss random variable X follows the distribution given in Examples 5.1 and 5.4. Calculate the Lundberg upper bound for the probability of ultimate ruin for $u = 0, 1, 2$, and 3.

Solution From Example 5.4, the adjustment coefficient is $r^* = 0.1740$. The Lundberg upper bound for $u = 0$ is 1, and for $u = 1, 2$, and 3, we have $e^{-0.174} = 0.8403$, $e^{-(2)(0.174)} = 0.7061$ and $e^{-(3)(0.174)} = 0.5933$, respectively. These figures may be compared against the exact values computed in Example 5.1, namely, 0.8, 0.68, and 0.568, respectively. □

Note that equation (5.23) can be written as

$$e^{-r} M_X(r) = 1, \qquad (5.32)$$

or

$$\log M_X(r) = r. \qquad (5.33)$$

We can establish that (see Exercise 5.1)

$$\left.\frac{d \log M_X(r)}{dr}\right|_{r=0} = \mu_X, \qquad (5.34)$$

and

$$\left.\frac{d^2 \log M_X(r)}{dr^2}\right|_{r=0} = \sigma_X^2, \tag{5.35}$$

so that the Taylor series approximation of $\log M_X(r)$ is

$$\log M_X(r) \simeq r\mu_X + \frac{r^2\sigma_X^2}{2}. \tag{5.36}$$

Thus, equation (5.33) can be written as

$$r \simeq r\mu_X + \frac{r^2\sigma_X^2}{2}, \tag{5.37}$$

the solutions of which are $r = 0$ and

$$r^* \simeq \frac{2(1-\mu_X)}{\sigma_X^2}. \tag{5.38}$$

For the loss distribution X in Example 5.4, its variance is 1.09. Based on the approximation in equation (5.38), the approximate adjustment coefficient is 0.1835, which is larger than the exact value of 0.1740.

5.3 Continuous-time surplus function

In a continuous-time model the insurance company receives premiums continuously, while claim losses may occur at any time. We assume that the initial surplus of the insurance company is u and the amount of premium received per unit time is c. We denote the number of claims (described as the number of occurrences of events) in the interval $(0, t]$ by $N(t)$, with claim amounts $X_1, \cdots, X_{N(t)}$, which are assumed to be independently and identically distributed as X. Also, we denote the aggregate losses up to (and including) time t by $S(t)$, which is given by

$$S(t) = \sum_{i=1}^{N(t)} X_i, \tag{5.39}$$

with the convention that if $N(t) = 0$, $S(t) = 0$. Thus, the surplus at time t, denoted by $U(t; u)$, is defined as

$$U(t; u) = u + ct - S(t). \tag{5.40}$$

Figure 5.4 illustrates an example of a realization of the surplus function $U(t; u)$. The surplus increases at a constant rate c until there is a claim, upon which the

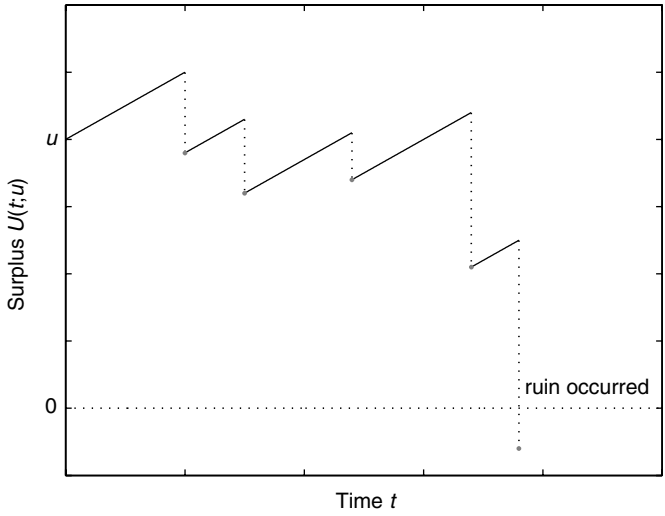

Figure 5.4 A realization of the continuous-time surplus process

surplus drops by the amount of the claim. The surplus then accumulates again at the rate c, and drops are repeated when claims occur. When a claim induces the surplus to fall to or below zero, ruin occurs.

Equation (5.40) defines a general surplus function. To analyze the behavior of $U(t;u)$ we make some assumptions about the claim process $S(t)$. In particular, we assume that the number of occurrences of (claim) events up to (and including) time t, $N(t)$, follows a **Poisson process**, which is defined as follows.

Definition 5.6 $N(t)$ is a Poisson process with parameter λ, which is the rate of occurrences of events per unit time, if (a) in any interval $(t_1, t_2]$, the number of occurrences of events, i.e. $N(t_2) - N(t_1)$, has a Poisson distribution with mean $\lambda(t_2 - t_1)$, and (b) over any non-overlapping intervals, the numbers of occurrences of events are independently distributed.

For a *fixed* t, $N(t)$ is distributed as a Poisson variable with parameter λt, i.e. $N(t) \sim \mathcal{PN}(\lambda t)$, and $S(t)$ follows a compound Poisson distribution (see equation (1.60)). Thus, for a fixed t, the distribution of $S(t)$ (and hence $U(t)$) can be analyzed as in Section 1.5.1. As a function of time t, $S(t)$ is a **compound Poisson process** and the corresponding surplus process $U(t;u)$ is a **compound Poisson surplus process**.

We assume that the claim random variable X has a mgf $M_X(r)$ for $r \in [0, \gamma)$, and consider the aggregate claim in a unit time interval $S(1)$. From equation

(1.75), the mean of $S(1)$ is

$$\mathrm{E}[S(1)] = \mathrm{E}[N(1)]\mathrm{E}(X) = \lambda\mu_X. \tag{5.41}$$

Thus, the premium per unit time is

$$c = (1+\theta)\mathrm{E}[S(1)] = (1+\theta)\lambda\mu_X, \tag{5.42}$$

where $\theta > 0$ is the loading factor.

5.4 Continuous-time ruin theory

As in the discrete-time case, the definitions of ultimate and finite-time ruin, as well as their associated probabilities, can be defined similarly for continuous-time models. We re-state their definitions as follows:

1 Ruin occurs at time t if $U(t;u) \leq 0$ for the first time at t, for $t > 0$.
2 Given the initial surplus u, $T(u)$ is defined as the time of ruin, i.e. $T(u) = \min\{t > 0 : U(t;u) \leq 0\}$.
3 Given an initial surplus u, the probability of ultimate ruin, denoted by $\psi(u)$, is $\psi(u) = \Pr(T(u) < \infty)$.
4 Given an initial surplus u, the probability of ruin by time t, denoted by $\psi(t;u)$, is $\psi(t;u) = \Pr(T(u) \leq t)$ for $t > 0$.

5.4.1 Lundberg's inequality in continuous time

We first define the adjustment coefficient in continuous time prior to presenting Lundberg's inequality. Analogous to the discrete-time case, in which the adjustment coefficient is the positive solution of the equation $M_{X-1}(r) = 1$, we consider the equation

$$M_{S(1)-c}(r) = \mathrm{E}\left\{e^{r[S(1)-c]}\right\} = e^{-rc}\,\mathrm{E}\left[e^{rS(1)}\right] = 1. \tag{5.43}$$

As $S(1)$ has a compound Poisson distribution, we have (see equation (1.66))

$$\begin{aligned}\mathrm{E}\left[e^{rS(1)}\right] &= M_{N(1)}\left[\log M_X(r)\right] \\ &= \exp\left[\lambda\left(e^{\log M_X(r)} - 1\right)\right] \\ &= \exp\left\{\lambda\left[M_X(r) - 1\right]\right\}. \end{aligned} \tag{5.44}$$

Combining equations (5.43) and (5.44), we have

$$\exp(rc) = \exp\left\{\lambda\left[M_X(r) - 1\right]\right\}, \tag{5.45}$$

so that

$$rc = \lambda \left[M_X(r) - 1 \right]. \tag{5.46}$$

The value $r^* > 0$ satisfying the above equation is called the adjustment coefficient. Substituting c in equation (5.42) into the above, we obtain the equivalent equation

$$1 + (1 + \theta)\, r\mu_X = M_X(r). \tag{5.47}$$

The existence of a positive solution to the above equation can be verified as follows. First, we denote the expression on the left-hand side of equation (5.47) by $\varphi(r)$, which is linear in r. Now $M_X'(r) = \mathrm{E}(Xe^{rX}) > 0$ and $M_X''(r) = \mathrm{E}(X^2 e^{rX}) > 0$. Thus, $M_X(r)$ is increasing and concave upward for $r > 0$. The gradient of the tangent to $M_X(r)$ at $r = 0$ is $M_X'(0) = \mathrm{E}(X) = \mu_X < (1+\theta)\mu_X$, which is the gradient of $\varphi(r)$. The graphs of $\varphi(r)$ and $M_X(r)$ can be illustrated in Figure 5.5, and they intersect at a point $r^* > 0$.[3] We note that the adjustment coefficient r^* depends on the loading factor θ and the distribution of X, but is not dependent on the parameter λ of the Poisson process.

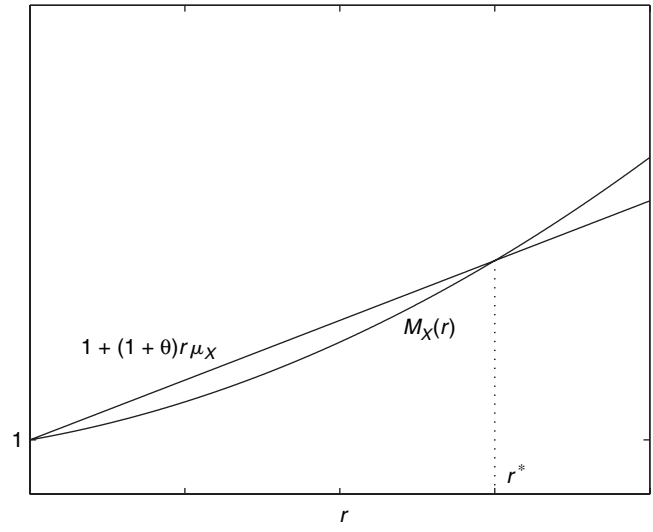

Figure 5.5 Illustration of the determination of r^*

[3] As the mgf of X is presumed to exist only for $r \in [0, \gamma)$, if $\varphi(r)$ intersects $M_X(r)$ at a point larger than γ, then solution to equation (5.47) does not exist. Hence, further technical conditions may be required to ensure the existence of r^*. In subsequent discussions, however, we assume that the adjustment coefficient is uniquely determined.

We now state Lundberg's inequality in continuous time in the following theorem.

Theorem 5.3 *If the surplus function follows a compound Poisson process defined in equation (5.40), the probability of ultimate ruin given initial surplus u, $\psi(u)$, satisfies the inequality*

$$\psi(u) \leq \exp(-r^* u), \tag{5.48}$$

where r^* is the adjustment coefficient satisfying equation (5.47).

Proof The proof of this theorem is similar to that for Theorem 5.2 using the method of induction. First, it can be shown that the inequality holds for ruin occurring at the first claim. Then, assuming the inequality holds for ruin occurring at the nth claim, it can be shown that it also holds for ruin occurring at the $(n+1)$th claim. The details can be found in Dickson (2005, Section 7.6). □

Solving equation (5.47) requires numerical methods. It is, however, possible to obtain an explicit approximate solution for r^*. If we take the logarithm on both sides of equation (5.47), the left-hand side can be approximated by[4]

$$(1+\theta) r \mu_X - \frac{(1+\theta)^2 r^2 \mu_X^2}{2}, \tag{5.49}$$

and from equation (5.36), we approximate the right-hand side by

$$r\mu_X + \frac{r^2 \sigma_X^2}{2}. \tag{5.50}$$

Equating expressions (5.49) and (5.50), we obtain

$$(1+\theta) r \mu_X - \frac{(1+\theta)^2 r^2 \mu_X^2}{2} = r\mu_X + \frac{r^2 \sigma_X^2}{2}, \tag{5.51}$$

so that the solutions are $r = 0$ and

$$r^* \simeq \frac{2\theta \mu_X}{\sigma_X^2 + (1+\theta)^2 \mu_X^2}. \tag{5.52}$$

Example 5.6 Let $U(t;u)$ be a compound Poisson surplus function with $X \sim \mathcal{G}(3, 0.5)$. Compute the adjustment coefficient and its approximate value using equation (5.52), for $\theta = 0.1$ and 0.2. Calculate the upper bounds for the probability of ultimate ruin for $u = 5$ and $u = 10$.

[4] This is due to the Taylor series expansion of $\log[1 + (1+\theta) r \mu_X]$, with $0 < (1+\theta) r \mu_X < 1$.

Solution The mgf of X is, from equation (2.32)

$$M_X(r) = \frac{1}{(1-\beta r)^{\alpha}} = \frac{1}{(1-0.5r)^3},$$

and its mean and variance are, respectively, $\mu_X = \alpha\beta = 1.5$ and $\sigma_X^2 = \alpha\beta^2 = 0.75$. From equation (5.47), the adjustment coefficient is the solution of r in the equation

$$\frac{1}{(1-0.5r)^3} = 1 + (1+\theta)(1.5)r,$$

from which we solve numerically to obtain $r^* = 0.0924$ when $\theta = 0.1$. The upper bounds for the probability of ultimate ruin are

$$\exp(-r^*u) = \begin{cases} 0.6300, & \text{for } u = 5, \\ 0.3969, & \text{for } u = 10. \end{cases}$$

When the loading is increased to 0.2, $r^* = 0.1718$, so that the upper bounds for the probability of ruin are

$$\exp(-r^*u) = \begin{cases} 0.4236, & \text{for } u = 5, \\ 0.1794, & \text{for } u = 10. \end{cases}$$

To compute the approximate values of r^*, we use equation (5.52) to obtain, for $\theta = 0.1$

$$r^* \simeq \frac{(2)(0.1)(1.5)}{0.75 + (1.1)^2(1.5)^2} = 0.0864,$$

and, for $\theta = 0.2$

$$r^* \simeq \frac{(2)(0.2)(1.5)}{0.75 + (1.2)^2(1.5)^2} = 0.1504.$$

Based on these approximate values, the upper bounds for the probability of ultimate ruin are, for $\theta = 0.1$

$$\exp(-r^*u) = \begin{cases} 0.6492, & \text{for } u = 5, \\ 0.4215, & \text{for } u = 10. \end{cases}$$

and, for $\theta = 0.2$

$$\exp(-r^*u) = \begin{cases} 0.4714, & \text{for } u = 5, \\ 0.2222, & \text{for } u = 10. \end{cases}$$

Thus, we can see that the adjustment coefficient increases with the premium loading θ. Also, the upper bound for the probability of ultimate ruin decreases with θ and u. We also observe that the approximation of r^* works reasonably well. □

5.4.2 Distribution of deficit

An insurance company is in deficit if the surplus function falls below the initial surplus. The theorem below presents an important result for the distribution of the deficit the *first time* the surplus function is below its initial level.

Theorem 5.4 *If the surplus function follows a compound Poisson surplus process as defined in equation (5.40), the probability that the surplus will ever fall below its initial level u, and will take value in the infinitesimal interval* $(u - x - dx, u - x)$ *the first time this happens, is*

$$\frac{\lambda S_X(x)}{c}dx = \frac{S_X(x)}{(1+\theta)\mu_X}dx, \qquad x > 0. \tag{5.53}$$

Proof See the proof of Theorem 13.5.1 in Bowers *et al.* (1997). □

Using equation (5.53) we can derive the probability that there is ever a deficit, which is obtained by integrating out x. Thus, the probability of a deficit is

$$\int_0^\infty \frac{S_X(x)}{(1+\theta)\mu_X}dx = \frac{1}{(1+\theta)\mu_X}\int_0^\infty S_X(x)dx = \frac{1}{1+\theta}. \tag{5.54}$$

Note that if the initial surplus is $u = 0$, a deficit is equivalent to ruin. Thus, the above probability is also the probability of ultimate ruin when the initial surplus is zero, i.e.

$$\psi(0) = \frac{1}{1+\theta}. \tag{5.55}$$

This result is analogous to Theorem 5.1 for the discrete-time model. Note that the Poisson parameter and the distribution of X does not determine $\psi(0)$.

We now define the conditional random variable L as the amount of deficit the *first time* deficit occurs, *given* that a deficit ever occurs. Thus, we consider

the following probability

$$\begin{aligned}\Pr\left(L \in (x, x+dx)\right) &= \frac{\Pr\left(\text{deficit occurs first time with amount } (x, x+dx)\right)}{\Pr(\text{deficit occurs})} \\ &= \frac{\dfrac{S_X(x)}{(1+\theta)\mu_X}dx}{\dfrac{1}{1+\theta}} \\ &= \frac{S_X(x)}{\mu_X}dx. \end{aligned} \tag{5.56}$$

Hence, the pdf of L, denoted by $f_L(\cdot)$, is

$$f_L(x) = \frac{S_X(x)}{\mu_X}. \tag{5.57}$$

The mgf of L, denoted by $M_L(\cdot)$, can be derived as follows

$$\begin{aligned} M_L(r) &= \int_0^\infty e^{rx}\left[\frac{S_X(x)}{\mu_X}\right]dx \\ &= \frac{1}{\mu_X}\left\{\frac{e^{rx}}{r}S_X(x)\right]_0^\infty + \frac{1}{r}\int_0^\infty e^{rx}f_X(x)dx\right\} \\ &= \frac{1}{r\mu_X}\left[M_X(r) - 1\right]. \end{aligned} \tag{5.58}$$

Example 5.7 Let $X \sim \mathcal{E}(\lambda)$. Compute the pdf of the conditional random variable of deficit L.

Solution The sf $S_X(x)$ of $\mathcal{E}(\lambda)$ is $e^{-\lambda x}$, and its mean is $\mu_X = 1/\lambda$. Hence, from equation (5.57), the pdf of L is

$$f_L(x) = \lambda e^{-\lambda x},$$

so that L is also distributed as $\mathcal{E}(\lambda)$ and $\mu_L = 1/\lambda$. □

Example 5.8 Show that the mean of the conditional random variable of deficit L is

$$\mathrm{E}(L) = \frac{\mathrm{E}(X^2)}{2\mu_X}.$$

If $X \sim \mathcal{G}(\alpha, \beta)$, calculate $\mathrm{E}(L)$.

Solution Using the mgf of L, we can compute the mean of L as $M_L'(0)$. Now differentiating $M_L(r)$ with respect to r, we obtain

$$M_L'(r) = \frac{rM_X'(r) - [M_X(r) - 1]}{r^2\mu_X}.$$

As the numerator and denominator of the expression on the right-hand side of the above equation evaluated at $r = 0$ are both equal to zero, we apply the l'Hôpital's rule (two times) to obtain

$$\mathrm{E}(L) = M_L'(0) = \frac{\lim_{r\to 0}\left[rM_X'''(r) + M_X''(r)\right]}{2\mu_X} = \frac{\mathrm{E}(X^2)}{2\mu_X}.$$

When $X \sim \mathcal{G}(\alpha, \beta)$, $\mu_X = \alpha\beta$ and $\sigma_X^2 = \alpha\beta^2$, so that

$$\mathrm{E}(L) = \frac{\alpha\beta^2 + \alpha^2\beta^2}{2\alpha\beta} = \frac{\beta + \alpha\beta}{2}.$$

Note that $\mathcal{E}(\lambda)$ is $\mathcal{G}(1, 1/\lambda)$. If we use the above result for $X \sim \mathcal{E}(\lambda)$, the mean of L is $\mu_L = \beta = 1/\lambda$, which is the same result as in Example 5.7. □

5.5 Summary and conclusions

The surplus of a block of insurance policies traces the excess of the initial surplus and premiums received over claim losses paid out. The business is said to be in ruin if the surplus falls to or below zero. We discuss the probabilities of ultimate ruin as well as ruin before a finite time. In the discrete-time case, we present recursive formulas to compute these probabilities, which depend on the initial surplus, the premium loading factor, and the distribution of the claim losses. An upper bound of the probability of ultimate ruin can be calculated using Lundberg's inequality.

In the continuous-time set-up, we consider a model in which claims following a compound Poisson process. Under this assumption the surplus function follows a compound Poisson surplus process. The compound Poisson process evaluated at a fixed time has a compound Poisson distribution. An upper bound of the probability of ultimate ruin is obtainable via Lundberg's inequality. We also discuss the distribution of the deficit when this first happens.

Exercises

5.1 Suppose the mgf of the claim-severity random variable X is $M_X(r)$ for $r \in [0, \gamma)$. Let the mean and variance of X be μ_X and σ_X^2, respectively.

Show that

$$\left.\frac{d \log M_X(r)}{dr}\right|_{r=0} = \mu_X,$$

and

$$\left.\frac{d^2 \log M_X(r)}{dr^2}\right|_{r=0} = \sigma_X^2.$$

5.2 You are given the following information about a block of insurance policies:

(a) There is an initial surplus of 2 at time zero. Premium per period is 1 and is paid at the beginning of each period. Surplus earns 5% interest in each period.

(b) The claims at the end of period 1, 2, and 3 are, 0.2, 1.5, and 0.6, respectively.

Find the surplus at the beginning of period 4 (prior to premium payment).

5.3 Claim severity in each period has the following distribution: $f_X(0) = 0.5, f_X(1) = 0.3, f_X(2) = 0$, and $f_X(3) = 0.2$. Find the probability of ultimate ruin if the initial surplus is 4.

5.4 Claim severity per period is distributed as $\mathcal{BN}(4, 0.2)$. Calculate the probability of ruin at or before time 3 if the initial surplus is 3.

5.5 In a discrete-time surplus model, the claim severity in each period is distributed as $\mathcal{GM}(0.6)$. Determine the adjustment coefficient and the maximum probability of ultimate ruin if the initial surplus is 3.

5.6 In a discrete-time surplus model, the claim severity in each period is distributed as $\mathcal{PN}(0.7)$. Determine the adjustment coefficient and the maximum probability of ultimate ruin if the initial surplus is 2.

5.7 In a discrete-time surplus model, the claim severity in each period is distributed as $\mathcal{BN}(2, 0.4)$. Determine the adjustment coefficient and the maximum probability of ultimate ruin if the initial surplus is 2.

5.8 In a continuous-time surplus model, the claim severity is distributed as $\mathcal{G}(4, 0.2)$. Determine the Lundberg upper bound for the probability of ultimate ruin if the initial surplus is 2 and the premium loading factor is 0.1.

5.9 In a continuous-time surplus model, the claim severity is distributed as $\mathcal{E}(2)$. Determine the Lundberg upper bound for the probability of ultimate ruin if the initial surplus is 4 and the premium loading factor is 0.2.

5.10 In a continuous-time surplus model, the claim severity is distributed as $\mathcal{GM}(0.6)$. Determine the Lundberg upper bound for the probability of ultimate ruin if the initial surplus is 3 and the premium loading factor is 0.5. Compare your results with those in Exercise 5.5.

5.11 In a continuous-time surplus model, the claim severity is distributed as $\mathcal{PN}(0.7)$. Determine the Lundberg upper bound for the probability of ultimate ruin if the initial surplus is 2 and the premium loading factor is 3/7. Compare your results with those in Exercise 5.6.

5.12 In a continuous-time surplus model, the claim severity is distributed as $\mathcal{BN}(2, 0.4)$. Determine the Lundberg upper bound for the probability of ultimate ruin if the initial surplus is 2 and the premium loading factor is 0.25. Compare your results with those in Exercise 5.7.

5.13 Claim data show that the claim in each period has mean 0.8 and variance 0.4. Assuming the claim process follows a Poisson process and the premium has a loading of 10%, determine the approximate bound of the probability of ultimate ruin if the initial surplus is 4. If it is desired to reduce this bound by 10%, what will be the revised premium loading without increasing the initial surplus? If the premium remains unchanged, what is the required initial surplus to achieve the desired probability bound of ultimate ruin?

5.14 Determine the mean of the first-time deficit L given that deficit occurs, if the loss random variable X is distributed as $\mathcal{U}(10, 20)$.

5.15 A discrete-time surplus function has initial surplus of 6 (exclusive of the first-year premium). Annual premiums of 3 are paid at the beginning of each year. Losses each year are 0 with probability 0.6 and 10 with probability 0.4, and are paid at the end of the year. Surplus earns interest of 6% annually. Determine the probability of ruin by the end of the second year.

5.16 A discrete-time surplus function has initial surplus of 8 (exclusive of the first-year premium). Annual losses X have the following distribution

x	$\Pr(X = x)$
0	0.50
10	0.30
20	0.10
30	0.10

Premiums equalling the annual expected loss are paid at the beginning of each year. If the surplus increases in a year, dividend equalling half of the increase is paid out at the end of the year. Determine the probability of ruin by the end of the second year.

Question adapted from SOA exams

5.17. You are given the following information about a block of insurance policies

(a) The initial surplus is 1. The annual premium collected at the beginning of each year is 2.

(b) The distribution of loss X each year is: $f_X(0) = 0.6, f_X(2) = 0.3$, and $f_X(4) = 0.1$.

(c) Capital at the beginning of the year earns 10% income for the year. Losses are paid and income is collected at the end of each year.

Calculate the finite-time probability of ruin $\psi(3; 1)$.

Part III

Credibility

Credibility theory provides the basic analytical framework for pricing insurance products. The importance of combining information about the recent experience of the individuals versus the aggregate past experience has been recognized in the literature through the classical approach. Rigorous analytical treatment of the subject started with Hans Bühlmann, and much work has been accomplished by him and his students. Bühlmann's approach provides a simple solution to the Bayesian method and achieves optimality within the subset of linear predictors. In this part of the book we introduce the classical approach, the Bühlmann approach, the Bayesian method, as well as the empirical implementation of these techniques.

6

Classical credibility

Credibility models were first proposed in the beginning of the twentieth century to update predictions of insurance losses in light of recently available data of insurance claims. The oldest approach is the limited-fluctuation credibility method, also called the classical approach, which proposes to update the loss prediction as a weighted average of the prediction based purely on the recent data and the rate in the insurance manual. Full credibility is achieved if the amount of recent data is sufficient, in which case the updated prediction will be based on the recent data only. If, however, the amount of recent data is insufficient, only partial credibility is attributed to the data and the updated prediction depends on the manual rate as well.

We consider the calculation of the minimum size of the data above which full credibility is attributed to the data. For cases where the data are insufficient we derive the partial-credibility factor and the updating formula for the prediction of the loss. The classical credibility approach is applied to update the prediction of loss measures such as the frequency of claims, the severity of claims, the aggregate loss, and the pure premium of a block of insurance policies.

Learning objectives

1 Basic framework of credibility
2 The limited-fluctuation (classical) credibility approach
3 Full credibility
4 Partial credibility
5 Prediction of claim frequency, claim severity, aggregate loss, and pure premium

6.1 Framework and notations

We consider a block of insurance policies, referred to as a **risk group**. Examples of risk groups of interest are workers of a company covered under a workers accident compensation scheme, employees of a firm covered under employees health insurance, and a block of vehicle insurance policies. The risk group is covered over a period of time (say, one year) upon the payment of a premium. The premium is partially based on a rate specified in the manual, called the **manual rate** and partially on the specific risk characteristics of the group. Based upon the recent **claim experience** of the risk group, the premium for the next period will be revised. Credibility theory concerns the updating of the prediction of the claim for the next period using the recent claim experience and the manual rate. The revised prediction determines the insurance premium of the next period for the risk group.

Credibility theory may be applied to different measures of claim experience. We summarize below the key factors of interest and define our notations to be used subsequently:

Claim frequency: The number of claims in the period is denoted by N.

Aggregate loss: We denote the amount of the ith claim by X_i and the aggregate loss by S, so that $S = X_1 + X_2 + \cdots + X_N$.

Claim severity: The average claim severity is the sample mean of $X_1, \cdots, X_N$, i.e. $\bar{X} = S/N$.

Pure premium: Let E be the number of exposure units of the risk group, the pure premium P is defined as $P = S/E$.

The loss measures N, X_i, S, $\bar{X}$, and P are random variables determined by uncertain events, while the exposure E is a known constant measuring the size of the risk group. For workers compensation and employees health insurance, E may be measured as the number of workers or employees covered under the policies.

We denote generically the predicted loss based on the manual by M, and the observed value of the loss based on recent data of the experience of the risk group by D. Thus, M and D may refer to the predicted value and observed value, respectively, of N, S, $\bar{X}$, and P. The **classical credibility** approach (also called the **limited-fluctuation credibility** approach) proposes to formulate the updated prediction of the loss measure as a weighted average of D and M. The weight attached to D is called the **credibility factor**, and is denoted by Z, with $0 \leq Z \leq 1$. Thus, the updated prediction, generically denoted by U, is given by

$$U = ZD + (1 - Z)M. \tag{6.1}$$

Example 6.1 The loss per worker insured in a ship-building company was \$230 last year. If the pure premium per worker in a similar industry is \$292 and the credibility factor of the company (the risk group) is 0.46, calculate the updated predicted pure premium for the company's insurance.

Solution We have $D = 230$, $M = 292$, and $Z = 0.46$, so that from equation (6.1) we obtain

$$U = (0.46)(230) + (1 - 0.46)(292) = \$263.48,$$

which will be the pure premium charged per worker of the company next year. □

From equation (6.1) we observe that U is always between the experience measure D and the manual rate M. The closer Z is to 1, the closer the updated predicted value U will be to the observed measure D. The credibility factor Z determines the relative importance of the data in calculating the updated prediction. **Full credibility** is said to be achieved if $Z = 1$, in which case the prediction depends upon the data only but not the manual. When $Z < 1$, the data are said to have **partial credibility**. Intuitively, a larger data set would justify a larger Z.

For the classical frequentist approach in statistics with no extraneous (or prior) information, estimation and prediction are entirely based on the data available. Thus, in the frequentist statistical framework one might say that all data have full credibility. The credibility theory literature, however, attempts to use extraneous (or prior) information (as provided by insurance manuals) to obtain an improved update of the prediction.

6.2 Full credibility

The classical credibility approach determines the minimum data size required for the experience data to be given full credibility (namely, for setting $Z = 1$). The minimum data size is called the **standard for full credibility**, which depends on the loss measure of interest. In this section we derive the formulas for the computation of the full-credibility standards for loss measures such as claim frequency, claim severity, aggregate loss, and pure premium.

6.2.1 Full credibility for claim frequency

Assume that the claim frequency random variable N has mean μ_N and variance σ_N^2. To assess how likely an observed value of N is "representative" of the true

mean, we ask the following question: What is the probability of observing claim frequency within $100k\%$ of the mean? This probability is given by

$$\Pr(\mu_N - k\mu_N \le N \le \mu_N + k\mu_N) = \Pr\left(-\frac{k\mu_N}{\sigma_N} \le \frac{N-\mu_N}{\sigma_N} \le \frac{k\mu_N}{\sigma_N}\right). \tag{6.2}$$

If we further assume that N is normally distributed, then $(N-\mu_N)/\sigma_N$ follows a standard normal distribution. Thus, denoting $\Phi(\cdot)$ as the df of the standard normal random variable, expression (6.2) becomes

$$\begin{aligned}\Pr\left(-\frac{k\mu_N}{\sigma_N} \le \frac{N-\mu_N}{\sigma_N} \le \frac{k\mu_N}{\sigma_N}\right) &= \Phi\left(\frac{k\mu_N}{\sigma_N}\right) - \Phi\left(-\frac{k\mu_N}{\sigma_N}\right) \\ &= \Phi\left(\frac{k\mu_N}{\sigma_N}\right) - \left[1 - \Phi\left(\frac{k\mu_N}{\sigma_N}\right)\right] \\ &= 2\Phi\left(\frac{k\mu_N}{\sigma_N}\right) - 1. \end{aligned} \tag{6.3}$$

If

$$\frac{k\mu_N}{\sigma_N} = z_{1-\frac{\alpha}{2}}, \tag{6.4}$$

where z_β is the 100βth percentile of the standard normal, i.e. $\Phi(z_\beta) = \beta$, then the probability in (6.3) is given by $2(1-\alpha/2) - 1 = 1 - \alpha$. Thus, there is a probability of $1-\alpha$ that an observed claim frequency is within $100k\%$ of the true mean, where α satisfies equation (6.4).

Example 6.2 Suppose the claim frequency of a risk group is normally distributed with mean 420 and variance 521, find the probability that the observed number of claims is within 10% of the true mean. Find the symmetric interval about the mean which covers 90% of the observed claim frequency. Express this interval in percentage error of the mean.

Solution We first calculate

$$\frac{k\mu_N}{\sigma_N} = \frac{(0.1)(420)}{\sqrt{521}} = 1.8401.$$

As $\Phi(1.8401) = 0.9671$, the required probability is $2(0.9671) - 1 = 0.9342$. Thus, the probability of observing claim frequency within 10% of the true mean is 93.42%.

Now $z_{0.95} = 1.645$. Using equation (6.4), we have

$$\frac{k\mu_N}{\sigma_N} = 1.645,$$

so that

$$k = \frac{1.645\,\sigma_N}{\mu_N} = \frac{1.645\sqrt{521}}{420} = 0.0894,$$

and there is a probability of 90% that the observed claim frequency is within 8.94% of the true mean. □

Note that the application of equation (6.4) requires knowledge of μ_N and σ_N^2. To simplify the computation, it is often assumed that the claim frequency is distributed as a Poisson variable with mean λ_N, which is large enough so that the normal approximation applies. Due to the Poisson assumption, we have $\mu_N = \sigma_N^2 = \lambda_N$, and equation (6.4) can be written as

$$\frac{k\lambda_N}{\sqrt{\lambda_N}} = k\sqrt{\lambda_N} = z_{1-\frac{\alpha}{2}}. \tag{6.5}$$

Example 6.3 Repeat Example 6.2, assuming that the claim frequency is distributed as a Poisson with mean 850, and that the normal approximation can be used for the Poisson.

Solution From equation (6.5), we have

$$k\sqrt{\lambda_N} = 0.1\sqrt{850} = 2.9155.$$

As $\Phi(2.9155) = 0.9982$, the coverage probability within 10% of the mean is $2(0.9982) - 1 = 99.64\%$. For coverage probability of 90%, we have

$$k = \frac{1.645}{\sqrt{850}} = 0.0564,$$

so that there is a probability of 90% that the observed claim frequency is within 5.64% of the mean. □

In the classical credibility approach, full credibility is attained for claim frequency if there is a probability of at least $1 - \alpha$ that the observed number of claims is within $100k\%$ of the true mean, where α and k are some given values. From equation (6.5) we can see that, under the Poisson assumption, the above probability statement holds if

$$k\sqrt{\lambda_N} \geq z_{1-\frac{\alpha}{2}}. \tag{6.6}$$

Table 6.1. *Selected values of standard for full credibility λ_F for claim frequency*

α	Coverage probability	k 10%	5%	1%
0.20	80%	165	657	16,424
0.10	90%	271	1,083	27,056
0.05	95%	385	1,537	38,415
0.01	99%	664	2,654	66,349

Hence, subject to the following assumptions concerning the claim frequency distribution:

1 the claim frequency is distributed as a Poisson variable,
2 the mean of the Poisson variable is large enough to justify the normal approximation to the Poisson,

full credibility for claim frequency is attributed to the data if $\lambda_N \geq (z_{1-\frac{\alpha}{2}}/k)^2$. We now define

$$\lambda_F \equiv \left(\frac{z_{1-\frac{\alpha}{2}}}{k}\right)^2, \tag{6.7}$$

which is the standard for full credibility for claim frequency, i.e. full credibility is attained if $\lambda_N \geq \lambda_F$.

As the expected number of claims λ_N is unknown in practice, the implementation of the credibility model is to compare the observed value of N in the recent period against λ_F calculated using equation (6.7). Full credibility is attained if $N \geq \lambda_F$. Table 6.1 presents the values of λ_F for selected values of α and k.

Table 6.1 shows that, given the accuracy parameter k, the standard for full credibility λ_F increases with the required coverage probability $1-\alpha$. Likewise, given the required coverage probability $1-\alpha$, λ_F increases with the required accuracy (i.e. decreases with k).

Example 6.4 If an insurance company requires a coverage probability of 99% for the number of claims to be within 5% of the true expected claim frequency, how many claims in the recent period are required for full credibility? If the insurance company receives 2,890 claims this year from the risk group and the manual list of expected claim is 3,000, what is the updated expected number of claims next year? Assume the claim-frequency distribution is Poisson and the normal approximation applies.

Solution We compute λ_F using equation (6.7) to obtain

$$\lambda_F = \left(\frac{z_{0.995}}{0.05}\right)^2 = \left(\frac{2.576}{0.05}\right)^2 = 2{,}653.96.$$

Hence, 2,654 claims are required for full credibility. As the observed claim frequency of 2,890 is larger than 2,654, full credibility is attributed to the data, i.e. $Z = 1$. Thus, $1 - Z = 0$, and from equation (6.1) the updated estimate of the expected number of claims in the next period is 2,890. Note that as full credibility is attained, the updated prediction does not depend on the manual value of $M = 3{,}000$. □

Example 6.5 If an insurance company decides to assign full credibility for 800 claims or more, what is the required coverage probability for the number of claims to be within 8% of the true value? Assume the claim-frequency distribution is Poisson and the normal approximation applies.

Solution To find the coverage probability when 800 claims are sufficient to acquire full credibility to within 8% of the true mean, we apply equation (6.5) to find α, which satisfies

$$k\sqrt{\lambda_N} = 0.08\sqrt{800} = 2.2627 = z_{1-\frac{\alpha}{2}},$$

so that $\alpha = 0.0237$ and the coverage probability is $1 - 0.0237 = 97.63\%$. □

Standard for full credibility is sometimes expressed in terms of the number of exposure units. The example below illustrates an application.

Example 6.6 Recent experience of a workers compensation insurance has established the mean accident rate as 0.045 and the standard for full credibility of claims as 1,200. For a group with a similar risk profile, what is the minimum number of exposure units (i.e. number of workers in the group) required for full credibility?

Solution As the standard for full credibility has been established for claim frequency, the standard for full credibility based on exposure is

$$\frac{1{,}200}{0.045} = 26{,}667 \text{ workers.}$$

□

6.2.2 Full credibility for claim severity

We now consider the standard for full credibility when the loss measure of interest is the claim severity. Suppose there is a sample of N claims of amounts $X_1, X_2, \cdots, X_N$. We assume $\{X_i\}$ to be iid with mean μ_X and variance σ_X^2, and use the sample mean $\bar{X}$ to estimate μ_X. Full credibility is attributed to $\bar{X}$ if the

probability of $\bar{X}$ being within $100k\%$ of the true mean of claim loss μ_X is at least $1-\alpha$, for given values of k and α. We also assume that the sample size N is sufficiently large so that $\bar{X}$ is approximately normally distributed with mean μ_X and variance σ_X^2/N. Hence, the coverage probability is

$$\Pr(\mu_X - k\mu_X \le \bar{X} \le \mu_X + k\mu_X) = \Pr\left(-\frac{k\mu_X}{\frac{\sigma_X}{\sqrt{N}}} \le \frac{\bar{X}-\mu_X}{\frac{\sigma_X}{\sqrt{N}}} \le \frac{k\mu_X}{\frac{\sigma_X}{\sqrt{N}}}\right)$$
$$\simeq 2\Phi\left(\frac{k\mu_X}{\frac{\sigma_X}{\sqrt{N}}}\right) - 1. \tag{6.8}$$

For the coverage probability to be larger than $1-\alpha$, we must have

$$\frac{k\mu_X}{\frac{\sigma_X}{\sqrt{N}}} \ge z_{1-\frac{\alpha}{2}}, \tag{6.9}$$

so that

$$N \ge \left(\frac{z_{1-\frac{\alpha}{2}}}{k}\right)^2 \left(\frac{\sigma_X}{\mu_X}\right)^2, \tag{6.10}$$

which is the standard for full credibility for severity. Note that the coefficient of variation of X is $C_X = \sigma_X/\mu_X$. Using equation (6.7), expression (6.10) can be written as

$$N \ge \lambda_F C_X^2. \tag{6.11}$$

Hence, $\lambda_F C_X^2$ is the standard for full credibility for claim severity. If the experience claim frequency exceeds $\lambda_F C_X^2$, $\bar{X}$ will be the predictor for the average severity of the next period (i.e. the manual rate will not be relevant). Note that in the derivation above, N is treated as a constant rather than a random variable dependent on the claim experience. To implement the methodology in practice, μ_X and σ_X^2 have to be estimated from the sample.

Example 6.7 What is the standard for full credibility for claim severity with $\alpha = 0.01$ and $k = 0.05$, given that the mean and variance estimates of the severity are 1,000 and 2,000,000, respectively?

Solution From Table 1, we have $\lambda_F = 2{,}654$. Thus, using equation (6.11), the standard for full credibility for severity is

$$2{,}654\left[\frac{2{,}000{,}000}{(1{,}000)(1{,}000)}\right] = 5{,}308.$$

□

In this example the standard for full credibility is higher for the severity estimate than for the claim-frequency estimate. As shown in equation (6.11), the standard for full credibility for severity is higher (lower) than that for claim frequency if the coefficient of variation of X is larger (smaller) than 1.

In deriving the standard for full credibility for severity, we do not make use of the assumption of Poisson distribution for claim frequency. The number of claims, however, must be large enough to justify normal approximation for the average loss per claim $\bar{X}$.

6.2.3 Full credibility for aggregate loss

To derive the standard for full credibility for aggregate loss, we determine the minimum (expected) claim frequency such that the probability of the observed aggregate loss S being within $100k\%$ of the expected aggregate loss is at least $1-\alpha$; that is, denoting μ_S and σ_S^2 as the mean and variance of S, respectively, we need to evaluate

$$\Pr(\mu_S - k\mu_S \le S \le \mu_S + k\mu_S) = \Pr\left(-\frac{k\mu_S}{\sigma_S} \le \frac{S-\mu_S}{\sigma_S} \le \frac{k\mu_S}{\sigma_S}\right). \quad (6.12)$$

To compute μ_S and σ_S^2, we use the compound distribution formulas derived in Appendix A.12 and applied in Section 1.5.1. Specifically, if N and $X_1, X_2, \cdots, X_N$ are mutually independent, we have $\mu_S = \mu_N \mu_X$ and $\sigma_S^2 = \mu_N \sigma_X^2 + \mu_X^2 \sigma_N^2$. If we further assume that N is distributed as a Poisson variable with mean λ_N, then $\mu_N = \sigma_N^2 = \lambda_N$, and we have $\mu_S = \lambda_N \mu_X$ and $\sigma_S^2 = \lambda_N(\mu_X^2 + \sigma_X^2)$. Thus

$$\frac{\mu_S}{\sigma_S} = \frac{\lambda_N \mu_X}{\sqrt{\lambda_N(\mu_X^2+\sigma_X^2)}} = \frac{\mu_X\sqrt{\lambda_N}}{\sqrt{\mu_X^2+\sigma_X^2}}. \quad (6.13)$$

Applying normal approximation to the distribution of the aggregate loss S, equation (6.12) can be written as

$$\begin{aligned}\Pr(\mu_S - k\mu_S \le S \le \mu_S + k\mu_S) &\simeq 2\Phi\left(\frac{k\mu_S}{\sigma_S}\right) - 1 \\ &= 2\Phi\left(\frac{k\mu_X\sqrt{\lambda_N}}{\sqrt{\mu_X^2+\sigma_X^2}}\right) - 1. \quad (6.14)\end{aligned}$$

For the above probability to be at least $1-\alpha$, we must have

$$\frac{k\mu_X\sqrt{\lambda_N}}{\sqrt{\mu_X^2+\sigma_X^2}} \geq z_{1-\frac{\alpha}{2}}, \tag{6.15}$$

so that

$$\lambda_N \geq \left(\frac{z_{1-\frac{\alpha}{2}}}{k}\right)^2 \left(\frac{\mu_X^2+\sigma_X^2}{\mu_X^2}\right). \tag{6.16}$$

Thus, the standard for full credibility for aggregate loss is

$$\left(\frac{z_{1-\frac{\alpha}{2}}}{k}\right)^2 \left(\frac{\mu_X^2+\sigma_X^2}{\mu_X^2}\right) = \lambda_F\left(1+C_X^2\right). \tag{6.17}$$

From equations (6.7) and (6.17), it can be seen that the standard for full credibility for aggregate loss is always higher than that for claim frequency. This result is due to the randomness of both the claim frequency and the claim severity in impacting the aggregate loss. Indeed, as

$$\lambda_F\left(1+C_X^2\right) = \lambda_F + \lambda_F C_X^2, \tag{6.18}$$

we conclude that

Standard for full credibility for aggregate loss
= Standard for full credibility for claim frequency
+ Standard for full credibility for claim severity.

Example 6.8 A block of health insurance policies has estimated mean severity of 25 and variance of severity of 800. For $\alpha = 0.15$ and $k = 0.08$, calculate the standard for full credibility for claim frequency and aggregate loss. Assume the claim frequency follows a Poisson distribution and normal approximation can be used for the claim-frequency and aggregate-loss distributions. If the block has an expected number of claims of 400 for the next period, is full credibility attained?

Solution As $z_{0.925} = \Phi^{-1}(0.925) = 1.4395$, from equation (6.7) we have

$$\lambda_F = \left(\frac{1.4395}{0.08}\right)^2 = 323.78,$$

which is the standard for full credibility for claim frequency. The coefficient of variation of the claim severity is

$$\frac{\sqrt{800}}{25} = 1.1314.$$

Thus, using equation (6.17) the standard for full credibility for aggregate loss is

$$(323.78)[1 + (1.1314)^2] = 738.24,$$

which is 2.28 times that of the standard for full credibility for claim frequency. The expected number of claims for the next period, 400, is larger than 323.78 but smaller than 738.24. Thus, full credibility is attained for the risk group for claim frequency but not for aggregate loss. □

Example 6.9 Data for the claim experience of a risk group in the current period show the following: (a) there are 542 claims and (b) the sample mean and variance of the claim severity are, respectively, 48 and 821. For $\alpha = 0.01$ and $k = 0.1$, do the data justify full credibility for claim frequency and claim severity for the next period?

Solution From Table 1, the standard for full credibility for claim frequency at $\alpha = 0.01$ and $k = 0.1$ is 664, which is larger than the claim frequency of 542. Thus, full credibility is not attained for claim frequency. To calculate the standard for full credibility for severity, we use equation (6.11) to obtain

$$\lambda_F C_X^2 = 664 \left[\frac{821}{(48)^2} \right] = 236.61.$$

As $542 > 236.61$, full credibility is attained for claim severity. □

6.2.4 Full credibility for pure premium

Pure premium, denoted by P, is the premium charged to cover losses before taking account of expenses and profits. Repeating the arguments as before, we evaluate the following probability

$$\Pr(\mu_P - k\mu_P \leq P \leq \mu_P + k\mu_P) = 2\Phi\left(\frac{k\mu_P}{\sigma_P}\right) - 1, \qquad (6.19)$$

where μ_P and σ_P^2 are, respectively, the mean and variance of P. As $P = S/E$, where the number of exposure units E is a constant, we have $\mu_P/\sigma_P = \mu_S/\sigma_S$. Thus, the final expression in equation (6.14) can be used to calculate the

probability in equation (6.19), and we conclude that the standard for full credibility for pure premium is the same as that for aggregate loss.

Example 6.10 A block of accident insurance policies has mean claim frequency of 0.03 per policy. Claim-frequency distribution is assumed to be Poisson. If the claim-severity distribution is lognormally distributed with $\mu = 5$ and $\sigma = 1$, calculate the number of policies required to attain full credibility for pure premium, with $\alpha = 0.02$ and $k = 0.05$.

Solution The mean and the variance of the claim severity are

$$\mu_X = \exp\left(\mu + \frac{\sigma^2}{2}\right) = \exp(5.5) = 244.6919$$

and

$$\sigma_X^2 = \left[\exp\left(2\mu + \sigma^2\right)\right]\left[\exp(\sigma^2) - 1\right] = 102{,}880.6497.$$

Thus, the coefficient of variation of claim severity is

$$C_X = \frac{\sqrt{102{,}880.6497}}{244.6919} = 1.3108.$$

Now $z_{0.99} = \Phi^{-1}(0.99) = 2.3263$, so that the standard for full credibility for pure premium requires a minimum expected claim frequency of

$$\lambda_F(1 + C_X^2) = \left(\frac{2.3263}{0.05}\right)^2\left[1 + (1.3108)^2\right] = 5{,}884.2379.$$

Hence, the minimum number of policies for full credibility for pure premium is

$$\frac{5{,}884.2379}{0.03} = 196{,}142.$$

□

6.3 Partial credibility

When the risk group is not sufficiently large, full credibility cannot be attained. In this case, a value of $Z < 1$ has to be determined. Denoting generically the loss measure of interest by W, the basic assumption in deriving Z is that the probability of ZW lying within the interval $[Z\mu_W - k\mu_W, Z\mu_W + k\mu_W]$ is $1-\alpha$ for a given value of k. For the case where the loss measure of interest is the claim frequency N, we require

$$\Pr\left(Z\mu_N - k\mu_N \leq ZN \leq Z\mu_N + k\mu_N\right) = 1 - \alpha, \tag{6.20}$$

which upon standardization becomes

$$\Pr\left(\frac{-k\mu_N}{Z\sigma_N} \leq \frac{N-\mu_N}{\sigma_N} \leq \frac{k\mu_N}{Z\sigma_N}\right) = 1-\alpha. \tag{6.21}$$

Assuming Poisson claim-frequency distribution with mean λ_N and applying the normal approximation, the left-hand side of the above equation reduces to

$$2\Phi\left(\frac{k\sqrt{\lambda_N}}{Z}\right) - 1. \tag{6.22}$$

Thus, we have

$$\frac{k\sqrt{\lambda_N}}{Z} = z_{1-\frac{\alpha}{2}}, \tag{6.23}$$

so that

$$Z = \left(\frac{k}{z_{1-\frac{\alpha}{2}}}\right)\sqrt{\lambda_N} = \sqrt{\frac{\lambda_N}{\lambda_F}}. \tag{6.24}$$

Equation (6.24) is called the **square-root rule for partial credibility**. For predicting claim frequency, the rule states that the **partial-credibility factor** Z is the square root of the ratio of the expected claim frequency to the standard for full credibility for claim frequency. The principle in deriving the partial credibility factor for claim frequency can be applied to other loss measures as well, and similar results are obtained. In general, the partial credibility factor is the square root of the ratio of the size of the risk group (measured in number of exposure units, number of claims or expected number of claims) to the standard for full credibility.

The partial credibility factors for claim severity, aggregate loss, and pure premium are summarized below

$$\textbf{Claim severity: } Z = \sqrt{\frac{N}{\lambda_F C_X^2}}$$

$$\textbf{Aggregate loss/Pure premium: } Z = \sqrt{\frac{\lambda_N}{\lambda_F(1+C_X^2)}}.$$

Example 6.11 A block of insurance policies had 896 claims this period with mean loss of 45 and variance of loss of 5,067. Full credibility is based on a coverage probability of 98% for a range of within 10% deviation from the true mean. The mean frequency of claims is 0.09 per policy and the block has 18,600 policies. Calculate Z for the claim frequency, claim severity, and aggregate loss for the next period.

Solution The expected claim frequency is $\lambda_N = 18{,}600(0.09) = 1{,}674$. We have $z_{0.99} = \Phi^{-1}(0.99) = 2.3263$, so that the full-credibility standard for claim frequency is

$$\lambda_F = \left(\frac{2.3263}{0.1}\right)^2 = 541.17 < 1{,}674 = \lambda_N.$$

Thus, for claim frequency there is full credibility and $Z = 1$. The estimated coefficient of variation for claim severity is

$$C_X = \frac{\sqrt{5{,}067}}{45} = 1.5818,$$

so that the standard for full credibility for claim severity is

$$\lambda_F C_X^2 = (541.17)(1.5818)^2 = 1{,}354.13,$$

which is larger than the sample size 896. Hence, full credibility is not attained for claim severity. The partial credibility factor is

$$Z = \sqrt{\frac{896}{1{,}354.13}} = 0.8134.$$

For aggregate loss, the standard for full credibility is

$$\lambda_F(1 + C_X^2) = 1{,}895.23 > 1{,}674 = \lambda_N.$$

Thus, full credibility is not attained for aggregate loss, and the partial credibility factor is

$$Z = \sqrt{\frac{1{,}674}{1{,}895.23}} = 0.9398.$$

□

6.4 Variation of assumptions

The classical credibility approach relies heavily on the assumption of Poisson distribution for the number of claims. This assumption can be relaxed, however. We illustrate the case where credibility for claim frequency is considered under an alternative assumption.

Assume the claim frequency N is distributed as a binomial random variable with parameters E and θ, i.e. $N \sim \mathcal{BN}(E, \theta)$. Thus, θ is the probability of a claim and E is the number of exposure units. The mean and variance of N are

$\lambda_N = E\theta$ and $\sigma_N^2 = E\theta(1-\theta)$. The standard for full credibility for claim frequency, as given by equation (6.4), is

$$\frac{k\mu_N}{\sigma_N} = \frac{kE\theta}{\sqrt{E\theta(1-\theta)}} = k\sqrt{\frac{E\theta}{1-\theta}} = z_{1-\frac{\alpha}{2}}, \tag{6.25}$$

which reduces to

$$E = \left(\frac{z_{1-\frac{\alpha}{2}}}{k}\right)^2 \left(\frac{1-\theta}{\theta}\right) = \lambda_F \left(\frac{1-\theta}{\theta}\right). \tag{6.26}$$

This can also be expressed in terms of the expected number of claims, which is

$$E\theta = \lambda_F(1-\theta). \tag{6.27}$$

As $\lambda_F(1-\theta) < \lambda_F$, the standard for full credibility under the binomial assumption is less than that under the Poisson assumption. However, as θ is typically small, $1-\theta$ is close to 1 and the difference between the two models is small.

Example 6.12 Assume full credibility is based on 99% coverage of observed claim frequency within 1% of the true mean. Compare the standard for full credibility for claim frequency based on assumptions of Poisson claim frequency versus binomial claim frequency with the probability of claim per policy being 0.05.

Solution From Table 6.1, full-credibility standard for claim frequency requires an expected claim frequency of 66,349, or exposure of 66,349/0.05 = 1.327 million units, if the Poisson assumption is adopted. Under the binomial assumption, the expected claim number from equation (6.27) is $66{,}349(1-0.05) = 63{,}031.55$. In terms of exposure, we require $63{,}031.55/0.05 = 1.261$ million units. □

6.5 Summary and discussions

Table 6.2 summarizes the formulas for various cases of full- and partial-credibility factors for different loss measures.

The above results for the classical credibility approach assume the approximation of the normal distribution for the loss measure of interest. Also, if predictions of claim frequency and aggregate loss are required, the claim frequency is assumed to be Poisson. The coefficient of variation of the claim severity is assumed to be given or reliable estimates are obtainable from the data.

Table 6.2. *Summary of standards for full-credibility and partial-credibility factor Z*

Loss measure	Standard for full credibility	Partial-credibility factor Z
Claim frequency	$\lambda_F = \left(\frac{z_{1-\frac{\alpha}{2}}}{k}\right)^2$	$\sqrt{\frac{\lambda_N}{\lambda_F}}$
Claim severity	$\lambda_F C_X^2$	$\sqrt{\frac{N}{\lambda_F C_X^2}}$
Aggregate loss/Pure premium	$\lambda_F(1 + C_X^2)$	$\sqrt{\frac{\lambda_N}{\lambda_F(1 + C_X^2)}}$

The computation of the full standards depends on the given level of probability coverage $1 - \alpha$ and the accuracy parameter k. As Table 6.1 shows, the full-credibility standard has large variations over different values of α and k, and it may be difficult to determine the suitable values to adopt.

Although the classical credibility approach is easy to apply, it is not based on a well-adopted statistical principle of prediction. In particular, there are several shortcomings of the approach, such as:

1 This approach emphasizes the role of D. It does not attach any importance to the accuracy of the prior information M.
2 The full-credibility standards depend on some unknown parameter values. The approach does not address the issue of how the calibration of these parameters may affect the credibility.
3 There are some limitations in the assumptions, which are made for the purpose of obtaining tractable analytical results.

Exercises

6.1 Assume the claim severity has a mean of 256 and a standard deviation of 532. A sample of 456 claims are observed. Answer the following questions.
   (a) What is the probability that the sample mean is within 10% of the true mean?
   (b) What is the coefficient of variation of the claim-severity distribution?
   (c) What is the coefficient of variation of the sample mean of the claim severity?

(d) Within what percentage of the true mean will the sample mean be observed with a probability of 92%?

(e) What assumptions have you made in answering the above questions?

6.2 Assume the aggregate-loss distribution follows a compound distribution with the claim frequency distributed as a Poisson with mean 569, and the claim severity distributed with mean 120 and standard deviation 78.

(a) Calculate the mean and the variance of the aggregate loss.

(b) Calculate the probability that the observed aggregate loss is within 6% of the mean aggregate loss.

(c) If the mean of the claim frequency increases to 620, how might the claim-severity distribution be changed so that the probability in (b) remains unchanged (give one possible answer)?

(d) If the standard deviation of the claim-severity distribution reduces to 60, how might the claim-frequency distribution be changed so that the probability in (b) remains unchanged?

6.3 A risk group has 569 claims this period, giving a claim average of 1,290 and standard deviation of 878. Calculate the standard for full credibility for claim frequency and claim severity, where full credibility is based on deviation of up to 6% of the true mean with a coverage probability of 94%. You may assume the claim frequency to be Poisson. Is full credibility attained in each case?

6.4 Assume the variance of the claim-frequency distribution is twice its mean (the distribution is not Poisson). Find the standard for full credibility for claim frequency and aggregate loss.

6.5 Assume Poisson distribution for the claim frequency. Show that the partial-credibility factor for claim severity is $\sqrt{N/(\lambda_F C_X^2)}$. The notations are as defined in the text.

6.6 Assume Poisson distribution for the claim frequency. Show that the partial-credibility factor for aggregate loss is $\sqrt{\lambda_N/[\lambda_F(1+C_X^2)]}$. The notations are as defined in the text.

6.7 The standard for full credibility for claim frequency of a risk group is 2,156. If the standard is based on a coverage probability of 94%, what is the accuracy parameter k?

(a) If the required accuracy parameter is halved, will the standard for full credibility increase or decrease? What is its new value?

(b) For the standard defined in (a), if the standard for full credibility for claim severity is 4,278 claims, what is the standard for full credibility for aggregate loss?

6.8 Claim severity is uniformly distributed in the interval [2000, 3000]. If claim frequency is distributed as a Poisson, determine the standard for full credibility for aggregate loss.

6.9 Assume claim frequency to be Poisson. If claim severity is distributed exponentially with mean 356, find the standard for full credibility for aggregate loss. If the maximum amount of a claim is capped at 500, calculate the revised standard for full credibility for the aggregate loss.

6.10 A block of health insurance policies has 2,309 claims this year, with mean claim of \$239 and standard deviation of \$457. If full credibility is based on 95% coverage to within 5% of the true mean claim severity, and the prior mean severity is \$250, what is the updated prediction for the mean severity next year based on the limited-fluctuation approach?

6.11 Claim severity is distributed lognormally with $\mu = 5$ and $\sigma^2 = 2$. The classical credibility approach is adopted to predict mean severity, with a minimum coverage probability of 92% to within 5% of the mean severity. An average loss of 354 per claim was calculated for 6,950 claims for the current year. What is the predicted loss for each claim? If the average loss of 354 was actually calculated for 9,650 claims, what is the predicted loss for each claim? State any assumptions you have made in the calculation.

6.12 Claim severity has mean 358 and standard deviation 421. An insurance company has 85,000 insurance policies. Using the classical credibility approach with coverage probability of 96% to within 6% of the aggregate loss, determine the credibility factor Z if the average claim per policy is (a) 3%, and (b) 5%.

6.13 Claim frequency has a binomial distribution with the probability of claim per policy being 0.068. Assume full credibility is based on 98% coverage of observed claim frequency within 4% of the true mean. Determine the credibility factor if there are 68,000 policies.

6.14 Claim severity has mean 26. In Year 1, 1,200 claims were filed with mean severity of 32. Based on the limited-fluctuation approach, severity per policy for Year 2 was then revised to 29.82. In Year 2, 1,500 claims were filed with mean severity of 24.46. What is the revised mean severity prediction for Year 3, if the same actuarial assumptions are used as for the prediction for Year 2?

6.15 Claim frequency N has a Poisson distribution, and claim size X is distributed as $\mathcal{P}(6, 0.5)$, where N and X are independent. For full credibility of the pure premium, the observed pure premium is required to be within 2% of the expected pure premium 90% of the time. Determine the expected number of claims required for full credibility.

Questions adapted from SOA exams

6.16 An insurance company has 2,500 policies. The annual amount of claims for each policy follows a compound distribution. The primary distribution is $\mathcal{NB}(2, 1/1.2)$ and the secondary distribution is $\mathcal{P}(3, 1000)$. Full credibility is attained if the observed aggregate loss is within 5% of the expected aggregate loss 90% of the time. Determine the partial credibility of the annual aggregate loss of the company.

6.17 An insurance company has determined that the limited-fluctuation full credibility standard is 2,000 if (a) the total number of claims is to be within 3% of the expected value with probability $1 - \alpha$, and (b) the number of claims follows a Poisson distribution. The standard is then changed so that the total cost of claims is to be within 5% of the expected value with probability $1 - \alpha$, where claim severity is distributed as $\mathcal{U}(0, 10000)$. Determine the expected number of claims for the limited-fluctuation full credibility standard.

6.18 The number of claims is distributed as $\mathcal{NB}(r, 0.25)$, and the claim severity takes values 1, 10, and 100 with probabilities 0.4, 0.4, and 0.2, respectively. If claim frequency and claim severity are independent, determine the expected number of claims needed for the observed aggregate losses to be within 10% of the expected aggregate losses with 95% probability.

7

Bühlmann credibility

While the classical credibility theory addresses the important problem of combining claim experience and prior information to update the prediction for loss, it does not provide a very satisfactory solution. The method is based on arbitrary selection of the coverage probability and the accuracy parameter. Furthermore, for tractability some restrictive assumptions about the loss distribution have to be imposed.

Bühlmann credibility theory sets the problem in a rigorous statistical framework of optimal prediction, using the least mean squared error criterion. It is flexible enough to incorporate various distributional assumptions of loss variables. The approach is further extended to enable the claim experience of different blocks of policies with different exposures to be combined for improved forecast through the Bühlmann–Straub model.

The Bühlmann and Bühlmann–Straub models recognize the interaction of two sources of variability in the data, namely the variation due to between-group differences and variation due to within-group fluctuations. We begin this chapter with the set-up of the Bühlmann credibility model, and a review of how the variance of the loss variable is decomposed into between-group and within-group variations. We derive the Bühlmann credibility factor and updating formula as the minimum mean squared error predictor. The approach is then extended to the Bühlmann–Straub model, in which the loss random variables have different exposures.

Learning objectives

1 Basic framework of Bühlmann credibility
2 Variance decomposition
3 Expected value of the process variance
4 Variance of the hypothetical mean
5 Bühlmann credibility
6 Bühlmann–Straub credibility

7.1 Framework and notations

Consider a risk group or block of insurance policies with loss measure denoted by X, which may be claim frequency, claim severity, aggregate loss, or pure premium. We assume that the risk profiles of the group are characterized by a parameter θ, which determines the distribution of the loss measure X. We denote the conditional mean and variance of X given θ by

$$\mathrm{E}(X \mid \theta) = \mu_X(\theta), \tag{7.1}$$

and

$$\mathrm{Var}(X \mid \theta) = \sigma_X^2(\theta). \tag{7.2}$$

We assume that the insurance company has similar blocks of policies with different risk profiles. Thus, the parameter θ varies with different risk groups. We treat θ as the realization of a random variable Θ, the distribution of which is called the **prior distribution**. When θ varies over the support of Θ, the conditional mean and variance of X become random variables in Θ, and are denoted by $\mu_X(\Theta) = \mathrm{E}(X \mid \Theta)$ and $\sigma_X^2(\Theta) = \mathrm{Var}(X \mid \Theta)$, respectively.

Example 7.1 An insurance company has blocks of worker compensation policies. The claim frequency is known to be Poisson with parameter λ, where λ is 20 for the low-risk group and 50 for the high-risk group. Suppose 30% of the risk groups are low risk and 70% are high risk. What are the conditional mean and variance of the claim frequency?

Solution The parameter determining the claim frequency X is λ, which is a realization of the random variable Λ. As X is Poisson, the conditional mean and conditional variance of X are equal to λ. Thus, we have the results in Table 7.1, so that

$$\mu_X(\Lambda) = \mathrm{E}(X \mid \Lambda) = \begin{cases} 20, & \text{with probability } 0.30, \\ 50, & \text{with probability } 0.70. \end{cases}$$

Likewise, we have

$$\sigma_X^2(\Lambda) = \mathrm{Var}(X \mid \Lambda) = \begin{cases} 20, & \text{with probability } 0.30, \\ 50, & \text{with probability } 0.70. \end{cases}$$

□

Example 7.2 The claim severity X of a block of health insurance policies is normally distributed with mean θ and variance 10. If θ takes values within the interval [100, 200] and follows a uniform distribution, what are the conditional mean and conditional variance of X?

Table 7.1. *Results for Example 7.1*

λ	$\Pr(\Lambda = \lambda)$	$\mathrm{E}(X \mid \lambda)$	$\mathrm{Var}(X \mid \lambda)$
20	0.3	20	20
50	0.7	50	50

Solution The conditional variance of X is 10, irrespective of θ. Hence, we have $\sigma_X^2(\Theta) = \mathrm{Var}(X \mid \Theta) = 10$ with probability 1. The conditional mean of X is Θ, i.e. $\mu_X(\Theta) = \mathrm{E}(X \mid \Theta) = \Theta$, which is uniformly distributed in [100, 200] with pdf

$$f_\Theta(\theta) = \begin{cases} 0.01, & \text{for } \theta \in [100, 200], \\ 0, & \text{otherwise.} \end{cases}$$

□

The Bühlmann model assumes that there are n observations of losses, denoted by $\boldsymbol{X} = \{X_1, \cdots, X_n\}$. The observations may be losses recorded in n periods and they are assumed to be independently and identically distributed as X, which depends on the parameter θ. The task is to update the prediction of X for the next period, i.e. X_{n+1}, based on $\boldsymbol{X}$. In the Bühlmann approach the solution depends on the variation between the conditional means as well as the average of the conditional variances of the risk groups. In the next section we discuss the calculation of these components, after which we will derive the updating formula proposed by Bühlmann.

7.2 Variance components

The variation of the loss measure X consists of two components: the variation between risk groups and the variation within risk groups. The first component, variation between risk groups, is due to the randomness of the risk profiles of each group and is captured by the parameter Θ. The second component, variation within risk group, is measured by the conditional variance of the risk group.[1]

We first consider the calculation of the overall mean of the loss measure X. The **unconditional mean** (or **overall mean**) of X measures the overall central tendency of X, averaged over all the underlying differences in the risk groups. Applying equation (A.111) of iterative expectation, the unconditional mean

[1] Readers may refer to Appendix A.11 for a review of the calculation of conditional expectation and total variance. To economize on notations, we will use X to denote a general loss measure as well as claim severity.

of X is

$$\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X \mid \Theta)] = \mathrm{E}[\mu_X(\Theta)]. \tag{7.3}$$

Thus, the unconditional mean is the average of the conditional means taken over the distribution of Θ.[2]

For the **unconditional variance** (or **total variance**), the calculation is more involved. The total variance of X is due to the variation in Θ as well as the variance of X conditional on Θ. We use the results derived in Appendix A.11. Applying equation (A.115), we have

$$\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X \mid \Theta)] + \mathrm{Var}[\mathrm{E}(X \mid \Theta)]. \tag{7.4}$$

Note that $\mathrm{Var}(X \mid \Theta)$ measures the variance of a given risk group. It is a function of the random variable Θ and we call this the **process variance**. Thus, $\mathrm{E}[\mathrm{Var}(X \mid \Theta)]$ is the **expected value of the process variance** (**EPV**). On the other hand, $\mathrm{E}(X \mid \Theta)$ is the mean of a given risk group. We call this conditional mean the **hypothetical mean**. Thus, $\mathrm{Var}[\mathrm{E}(X \mid \Theta)]$ is the **variance of the hypothetical means** (**VHM**), as it measures the variations in the *means* of the risk groups. Verbally, equation (7.4) can be written as

$$\begin{aligned}\textbf{total variance} = {} & \textbf{expected value of process variance} \\ & + \textbf{variance of hypothetical means},\end{aligned} \tag{7.5}$$

or

$$\textbf{total variance} = \textbf{EPV} + \textbf{VHM}. \tag{7.6}$$

It can also be stated alternatively as

$$\begin{aligned}\textbf{total variance} = {} & \textbf{mean of conditional variance} \\ & + \textbf{variance of conditional mean}.\end{aligned} \tag{7.7}$$

Symbolically, we use the following notations

$$\mathrm{E}[\mathrm{Var}(X \mid \Theta)] = \mathrm{E}[\sigma_X^2(\Theta)] = \mu_{\mathrm{PV}}, \tag{7.8}$$

and

$$\mathrm{Var}[\mathrm{E}(X \mid \Theta)] = \mathrm{Var}[\mu_X(\Theta)] = \sigma_{\mathrm{HM}}^2, \tag{7.9}$$

[2] Note that the expectation operations in equation (7.3) have different meanings. The operation in $\mathrm{E}(X)$ is taken unconditionally on X. The first (outer) operation in $\mathrm{E}[\mathrm{E}(X \mid \Theta)]$ is taken over Θ, while the second (inner) operation is taken over X *conditional* on Θ. Lastly, the operation in $\mathrm{E}[\mu_X(\Theta)]$ is taken over Θ unconditionally.

so that equation (7.4) can be written as

$$\text{Var}(X) = \mu_{\text{PV}} + \sigma^2_{\text{HM}}. \tag{7.10}$$

Example 7.3 For Examples 7.1 and 7.2, calculate the unconditional mean, the expected value of the process variance, the variance of the hypothetical means, and the total variance.

Solution For Example 7.1, the unconditional mean is

$$\begin{aligned}\text{E}(X) &= \Pr(\Lambda = 20)\text{E}(X \mid \Lambda = 20) + \Pr(\Lambda = 50)\text{E}(X \mid \Lambda = 50)\\ &= (0.3)(20) + (0.7)(50)\\ &= 41.\end{aligned}$$

The expected value of the process variance, EPV, is

$$\begin{aligned}\text{E}[\text{Var}(X \mid \Lambda)] &= \Pr(\Lambda = 20)\text{Var}(X \mid \Lambda = 20) + \Pr(\Lambda = 50)\text{Var}(X \mid \Lambda = 50)\\ &= (0.3)(20) + (0.7)(50)\\ &= 41.\end{aligned}$$

As the mean of the hypothetical means (i.e. the unconditional mean) is 41, the variance of the hypothetical means, VHM, is

$$\text{Var}[\text{E}(X \mid \Lambda)] = (0.3)(20 - 41)^2 + (0.7)(50 - 41)^2 = 189.$$

Thus, the total variance of X is

$$\text{Var}(X) = \text{E}[\text{Var}(X \mid \Lambda)] + \text{Var}[\text{E}(X \mid \Lambda)] = 41 + 189 = 230.$$

For Example 7.2, as Θ is uniformly distributed in $[100,\ 200]$, the unconditional mean of X is

$$\text{E}(X) = \text{E}[\text{E}(X \mid \Theta)] = \text{E}(\Theta) = 150.$$

As X has a constant variance of 10, the expected value of the process variance is

$$\text{E}[\text{Var}(X \mid \Theta)] = \text{E}(10) = 10.$$

The variance of the hypothetical means is[3]

$$\text{Var}[\text{E}(X \mid \Theta)] = \text{Var}(\Theta) = \frac{(200 - 100)^2}{12} = 833.33,$$

[3] See Appendix A.10.3 for the variance of the uniform distribution.

and the total variance of X is

$$\text{Var}(X) = 10 + 833.33 = 843.33.$$

□

If we divide the variance of the hypothetical means by the total variance, we obtain the proportion of the variation in X that is due to the differences in the means of the risk groups. Thus, for Example 7.1, we have

$$\frac{\sigma^2_{\text{HM}}}{\text{Var}(X)} = \frac{189}{230} = 82.17\%,$$

so that 82.17% of the variation in a randomly observed X is due to the differences in the averages of the risk groups. For Example 7.2, this figure is $833.33/843.33 = 98.81\%$.

Example 7.4 The claim severity X of a block of health insurance policies is normally distributed with mean 100 and variance σ^2. If σ^2 takes values within the interval [50, 100] and follows a uniform distribution, find the conditional mean of claim severity, the expected value of the process variance, the variance of the hypothetical means, and the total variance.

Solution We denote the random variable of the variance of X by Ω . Note that the conditional mean of X does not vary with Ω, and we have $\text{E}(X \mid \Omega) = 100$, so that the unconditional mean of X is

$$\text{E}(X) = \text{E}[\text{E}(X \mid \Omega)] = \text{E}(100) = 100.$$

As Ω is uniformly distributed in [50, 100], the expected value of the process variance is

$$\text{EPV} = \mu_{\text{PV}} = \text{E}[\text{Var}(X \mid \Omega)] = \text{E}(\Omega) \ = 75.$$

For the variance of the hypothetical means, we have

$$\text{VHM} = \sigma^2_{\text{HM}} = \text{Var}[\text{E}(X \mid \Omega)] = \text{Var}(100) = 0.$$

Thus, the total variance of X is 75, which is *entirely* due to the process variance, as there is no variation in the conditional mean. □

Example 7.5 An insurance company sells workers compensation policies, each of which belongs to one of three possible risk groups. The risk groups have claim frequencies N that are Poisson distributed with parameter λ and claim severity X that are gamma distributed with parameters α and β. Claim frequency and claim

Table 7.2. *Data for Example 7.5*

Risk group	Relative frequency	Distribution of N: $\mathcal{PN}(\lambda)$	Distribution of X: $\mathcal{G}(\alpha, \beta)$
1	0.2	$\lambda = 20$	$\alpha = 5, \beta = 2$
2	0.4	$\lambda = 30$	$\alpha = 4, \beta = 3$
3	0.4	$\lambda = 40$	$\alpha = 3, \beta = 2$

severity are independently distributed given a risk group, and the aggregate loss is S. The data of the risk groups are given in Table 7.2. For each of the following loss measures: (a) claim frequency N, (b) claim severity X, and (c) aggregate loss S, calculate EPV, VHM, and the total variance.

Solution (a) Claim frequency We first calculate the conditional mean and conditional variance of N given the risk group, which is characterized by the parameter Λ. As N is Poisson, the mean and variance are equal to Λ, so that we have the results in Table 7.3.

Table 7.3. *Results for Example 7.5 (a)*

Risk group	Probability	$\mathrm{E}(N \mid \Lambda) = \mu_N(\Lambda)$	$\mathrm{Var}(N \mid \Lambda) = \sigma_N^2(\Lambda)$
1	0.2	20	20
2	0.4	30	30
3	0.4	40	40

Thus, the EPV is

$$\mu_{\mathrm{PV}} = \mathrm{E}[\mathrm{Var}(N \mid \Lambda)] = (0.2)(20) + (0.4)(30) + (0.4)(40) = 32,$$

which is also equal to the unconditional mean $\mathrm{E}[\mu_N(\Lambda)]$. For VHM, we first calculate

$$\mathrm{E}\{[\mu_N(\Lambda)]^2\} = (0.2)(20)^2 + (0.4)(30)^2 + (0.4)(40)^2 = 1{,}080,$$

so that

$$\sigma_{\mathrm{HM}}^2 = \mathrm{Var}[\mu_N(\Lambda)] = \mathrm{E}\{[\mu_N(\Lambda)]^2\} - \{\mathrm{E}[\mu_N(\Lambda)]\}^2 = 1{,}080 - (32)^2 = 56.$$

Therefore, the total variance of N is

$$\text{Var}(N) = \mu_{\text{PV}} + \sigma^2_{\text{HM}} = 32 + 56 = 88.$$

(b) Claim severity There are three claim-severity distributions, which are specific to each risk group. Note that the relative frequencies of the risk groups as well as the claim frequencies in the risk groups jointly determine the relative occurrence of each claim-severity distribution. The probabilities of occurrence of the severity distributions, as well as their conditional means and variances are given in Table 7.4, in which Γ denotes the vector random variable representing α and β.

Table 7.4. *Results for Example 7.5 (b)*

Group	Group probability	λ	Col 2 $\times$ Col 3	Probability of severity X	$\text{E}(X \mid \Gamma) = \mu_X(\Gamma)$	$\text{Var}(X \mid \Gamma) = \sigma^2_X(\Gamma)$
1	0.2	20	4	0.125	10	20
2	0.4	30	12	0.375	12	36
3	0.4	40	16	0.500	6	12

Column 4 gives the expected number of claims in each group weighted by the group probability. Column 5 gives the probability of occurrence of each type of claim-severity distribution, which is obtained by dividing the corresponding figure in Column 4 by the sum of Column 4 (e.g. 0.125 = 4/(4 + 12 + 16)). The last two columns give the conditional mean $\alpha\beta$ and conditional variance $\alpha\beta^2$ corresponding to the three different distributions of claim severity. Similar to the calculation in (a), we have

$$\text{E}(X) = \text{E}[\text{E}(X \mid \Gamma)] = (0.125)(10) + (0.375)(12) + (0.5)(6) = 8.75,$$

and

$$\mu_{\text{PV}} = (0.125)(20) + (0.375)(36) + (0.5)(12) = 22.$$

To calculate VHM, we first compute the raw second moment of the conditional mean of X, which is

$$\text{E}\{[\mu_X(\Gamma)]^2\} = (0.125)(10)^2 + (0.375)(12)^2 + (0.5)(6)^2 = 84.50.$$

Hence

$$\sigma^2_{\text{HM}} = \text{Var}[\mu_X(\Gamma)] = \text{E}\{[\mu_X(\Gamma)]^2\} - \{\text{E}[\mu_X(\Gamma)]\}^2$$
$$= 84.50 - (8.75)^2 = 7.9375.$$

Therefore, the total variance of X is

$$\text{Var}(X) = \mu_{\text{PV}} + \sigma^2_{\text{HM}} = 22 + 7.9375 = 29.9375.$$

(c) Aggregate loss The distribution of the aggregate loss S is determined jointly by Λ and Γ, which we shall denote as Θ. For the conditional mean of S, we have

$$\text{E}(S \mid \Theta) = \text{E}(N \mid \Theta)\text{E}(X \mid \Theta) = \lambda\alpha\beta.$$

For the conditional variance of S, we use the result on compound distribution with Poisson claim frequency stated in equation (A.123), and make use of the assumption of gamma severity to obtain

$$\text{Var}(S \mid \Theta) = \lambda[\sigma^2_X(\Gamma) + \mu^2_X(\Gamma)] = \lambda(\alpha\beta^2 + \alpha^2\beta^2).$$

The conditional means and conditional variances of S are summarized in Table 7.5.

Table 7.5. *Results for Example 7.5 (c)*

Group	Group probability	Parameters λ, α, β	$\text{E}(S \mid \Theta) = \mu_S(\Theta)$	$\text{Var}(S \mid \Theta) = \sigma^2_S(\Theta)$
1	0.2	20, 5, 2	200	2,400
2	0.4	30, 4, 3	360	5,400
3	0.4	40, 3, 2	240	1,920

The unconditional mean of S is

$$\text{E}(S) = \text{E}[\text{E}(S \mid \Theta)] = (0.2)(200) + (0.4)(360) + (0.4)(240) = 280,$$

and the EPV is

$$\mu_{\text{PV}} = (0.2)(2{,}400) + (0.4)(5{,}400) + (0.4)(1{,}920) = 3{,}408.$$

Also, the VHM is given by

$$\begin{aligned}\sigma^2_{\text{HM}} &= \text{Var}[\mu_S(\Theta)] \\ &= \text{E}\{[\mu_S(\Theta)]^2\} - \{\text{E}[\mu_S(\Theta)]\}^2 \\ &= \left[(0.2)(200)^2 + (0.4)(360)^2 + (0.4)(240)^2\right] - (280)^2 \\ &= 4{,}480.\end{aligned}$$

Therefore, the total variance of S is

$$\text{Var}(S) = 3{,}408 + 4{,}480 = 7{,}888.$$

□

EPV and VHM measure two different aspects of the total variance. When a risk group is homogeneous so that the loss claims are similar within the group, the conditional variance is small. If all risk groups have similar loss claims within the group, the expected value of the process variance EPV is small. On the other hand, if the risk groups have very different risk profiles across groups, their hypothetical means will differ more and thus the variance of the hypothetical means VHM will be large. In other words, it will be easier to distinguish between risk groups if the variance of the hypothetical means is large and the average of the process variance is small.

We define k as the ratio of EPV to VHM, i.e.

$$k = \frac{\mu_{\text{PV}}}{\sigma^2_{\text{HM}}} = \frac{\text{EPV}}{\text{VHM}}. \tag{7.11}$$

A small EPV or large VHM will give rise to a small k. The risk groups will be more *distinguishable* in the mean when k is smaller, in which case we may put more weight on the data in updating our revised prediction for future losses. For the cases in Example 7.5, the values of k for claim frequency, claim severity, and aggregate loss are, respectively, 0.5714, 2.7717, and 0.7607. For Example 7.4, as $\sigma^2_{\text{HM}} = 0$, k is infinite.[4]

Example 7.6 Frequency of claim per year, N, is distributed as a binomial random variable $\mathcal{BN}(10, \theta)$, and claim severity, X, is distributed as an exponential random variable with mean $c\theta$, where c is a known constant. Given θ, claim frequency and claim severity are independently distributed. Derive an expression of k for the aggregate loss per year, S, in terms of c and the moments

[4] In this case, the data contain no useful information for updating the *mean* of the risk group separately from the overall mean, although they might be used to update the specific group *variance* if required.

of Θ, and show that it does not depend on c. If Θ is 0.3 or 0.7 with equal probabilities, calculate k.

Solution We first calculate the conditional mean of S as a function of θ. Due to the independence assumption of N and X, the hypothetical mean of S is

$$\mathrm{E}(S \mid \Theta) = \mathrm{E}(N \mid \Theta)\mathrm{E}(X \mid \Theta) = (10\Theta)(c\Theta) = 10c\Theta^2.$$

Using equation (A.122), the process variance is

$$\begin{aligned}
\mathrm{Var}(S \mid \Theta) &= \mu_N(\Theta)\sigma_X^2(\Theta) + \sigma_N^2(\Theta)\mu_X^2(\Theta) \\
&= (10\Theta)(c\Theta)^2 + [10\Theta(1-\Theta)](c\Theta)^2 \\
&= 10c^2\Theta^3 + 10c^2\Theta^3(1-\Theta) \\
&= 10c^2\Theta^3(2-\Theta).
\end{aligned}$$

Hence, the unconditional mean of S is

$$\mathrm{E}(S) = \mathrm{E}[\mathrm{E}(S \mid \Theta)] = \mathrm{E}(10c\Theta^2) = 10c\mathrm{E}(\Theta^2)$$

and the variance of the hypothetical means is

$$\begin{aligned}
\sigma_{\mathrm{HM}}^2 &= \mathrm{Var}[\mathrm{E}(S \mid \Theta)] \\
&= \mathrm{Var}(10c\Theta^2) \\
&= 100c^2\mathrm{Var}(\Theta^2) \\
&= 100c^2\{\mathrm{E}(\Theta^4) - [\mathrm{E}(\Theta^2)]^2\}.
\end{aligned}$$

The expected value of the process variance is

$$\begin{aligned}
\mu_{\mathrm{PV}} &= \mathrm{E}[\mathrm{Var}(S \mid \Theta)] \\
&= \mathrm{E}[10c^2\Theta^3(2-\Theta)] \\
&= 10c^2[2\mathrm{E}(\Theta^3) - \mathrm{E}(\Theta^4)].
\end{aligned}$$

Combining the above results we conclude that

$$k = \frac{\mu_{\mathrm{PV}}}{\sigma_{\mathrm{HM}}^2} = \frac{10c^2[2\mathrm{E}(\Theta^3) - \mathrm{E}(\Theta^4)]}{100c^2\{\mathrm{E}(\Theta^4) - [\mathrm{E}(\Theta^2)]^2\}} = \frac{2\mathrm{E}(\Theta^3) - \mathrm{E}(\Theta^4)}{10\{\mathrm{E}(\Theta^4) - [\mathrm{E}(\Theta^2)]^2\}}.$$

Thus, k does not depend on c. To compute its value for the given distribution of Θ, we present the calculations in Table 7.6.

Table 7.6. *Calculations of Example 7.6*

θ	$\Pr(\Theta = \theta)$	θ^2	θ^3	θ^4
0.3	0.5	0.09	0.027	0.0081
0.7	0.5	0.49	0.343	0.2401

Thus, the required moments of Θ are

$$\begin{aligned} \mathrm{E}(\Theta) &= (0.5)(0.3) + (0.5)(0.7) = 0.5, \\ \mathrm{E}(\Theta^2) &= (0.5)(0.09) + (0.5)(0.49) = 0.29, \\ \mathrm{E}(\Theta^3) &= (0.5)(0.027) + (0.5)(0.343) = 0.185 \end{aligned}$$

and

$$\mathrm{E}(\Theta^4) = (0.5)(0.0081) + (0.5)(0.2401) = 0.1241,$$

so that

$$k = \frac{2(0.185) - 0.1241}{10\left[0.1241 - (0.29)^2\right]} = 0.6148.$$

In this example, note that both EPV and VHM depend on c. However, as the effects of c on these components are the same, the ratio of EPV to VHM is invariant to c. Also, though X and N are independent *given* θ, they are correlated *unconditionally* due to their common dependence on Θ. □

7.3 Bühlmann credibility

Bühlmann's approach of updating the predicted loss measure is based on a linear predictor using past observations. It is also called the **greatest accuracy approach** or the **least squares approach**. Recall that for the classical credibility approach, the updated prediction U is given by (see equation (6.1))

$$U = ZD + (1 - Z)M. \tag{7.12}$$

The Bühlmann credibility method has a similar basic equation, in which D is the sample mean of the data and M is the overall prior mean $\mathrm{E}(X)$. The Bühlmann

credibility factor Z depends on the sample size n and the EPV to VHM ratio k. In particular, Z varies with n and k as follows:

1 Z increases with the sample size n of the data.
2 Z increases with the *distinctiveness* of the risk groups. As argued above, the risk groups are more distinguishable when k is small. Thus, Z increases as k decreases.

We now state formally the assumptions of the Bühlmann model and derive the updating formula as the **least mean squared error (MSE) linear predictor**.

1 $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ are loss measures that are independently and identically distributed as the random variable X. The distribution of X depends on the parameter θ.
2 The parameter θ is a realization of a random variable Θ. Given θ, the conditional mean and variance of X are

$$\mathrm{E}(X \mid \theta) = \mu_X(\theta), \tag{7.13}$$

and

$$\mathrm{Var}(X \mid \theta) = \sigma_X^2(\theta). \tag{7.14}$$

3 The unconditional mean of X is $\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X \mid \Theta)] = \mu_X$. The mean of the conditional variance of X is

$$\begin{aligned} \mathrm{E}[\mathrm{Var}(X \mid \Theta)] &= \mathrm{E}[\sigma_X^2(\Theta)] \\ &= \mu_{\mathrm{PV}} \\ &= \text{Expected value of process variance} \\ &= \mathrm{EPV}, \end{aligned} \tag{7.15}$$

and the variance of the conditional mean is

$$\begin{aligned} \mathrm{Var}[\mathrm{E}(X \mid \Theta)] &= \mathrm{Var}[\mu_X(\Theta)] \\ &= \sigma_{\mathrm{HM}}^2 \\ &= \text{Variance of hypothetical means} \\ &= \mathrm{VHM}. \end{aligned} \tag{7.16}$$

The unconditional variance (or total variance) of X is

$$\begin{aligned}\text{Var}(X) &= \text{E}[\text{Var}(X \mid \Theta)] + \text{Var}[\text{E}(X \mid \Theta)] \\ &= \mu_{\text{PV}} + \sigma^2_{\text{HM}} \\ &= \text{EPV} + \text{VHM}. \end{aligned} \tag{7.17}$$

4 The Bühlmann approach formulates a predictor of X_{n+1} based on a linear function of $\boldsymbol{X}$, where X_{n+1} is assumed to have the same distribution as X. The predictor minimizes the mean squared error in predicting X_{n+1} over the joint distribution of Θ, X_{n+1}, and $\boldsymbol{X}$. Specifically, the predictor is given by

$$\hat{X}_{n+1} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n, \tag{7.18}$$

where $\beta_0, \beta_1, \cdots, \beta_n$ are chosen to minimize the mean squared error, MSE, defined as

$$\text{MSE} = \text{E}[(X_{n+1} - \hat{X}_{n+1})^2]. \tag{7.19}$$

To solve the above problem we make use of the least squares regression results in Appendix A.17. We define $\boldsymbol{W}$ as the $(n+1) \times 1$ vector $(1, \boldsymbol{X}')'$, and $\boldsymbol{\beta}$ as the $(n+1) \times 1$ vector $(\beta_0, \beta_1, \cdots, \beta_n)'$. We also write $\boldsymbol{\beta}_S$ as $(\beta_1, \cdots, \beta_n)'$. Thus, the predictor $\hat{X}_{n+1}$ can be written as

$$\hat{X}_{n+1} = \boldsymbol{\beta}'\boldsymbol{W} = \beta_0 + \boldsymbol{\beta}_S'\boldsymbol{X}. \tag{7.20}$$

The MSE is then given by

$$\begin{aligned}\text{MSE} &= \text{E}[(X_{n+1} - \hat{X}_{n+1})^2] \\ &= \text{E}[\big(X_{n+1} - \boldsymbol{\beta}'\boldsymbol{W}\big)^2] \\ &= \text{E}(X^2_{n+1} + \boldsymbol{\beta}'\boldsymbol{W}\boldsymbol{W}'\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{W}X_{n+1}) \\ &= \text{E}(X^2_{n+1}) + \boldsymbol{\beta}'\text{E}(\boldsymbol{W}\boldsymbol{W}')\boldsymbol{\beta} - 2\boldsymbol{\beta}'\text{E}(\boldsymbol{W}X_{n+1}). \end{aligned} \tag{7.21}$$

Thus, the MSE has the same form as RSS in equation (A.167), with the sample moments replaced by the population moments. Hence, the solution of $\boldsymbol{\beta}$ that minimizes MSE is, by virtue of equation (A.168)

$$\hat{\boldsymbol{\beta}} = \big[\text{E}(\boldsymbol{W}\boldsymbol{W}')\big]^{-1}\text{E}(\boldsymbol{W}X_{n+1}). \tag{7.22}$$

Following the results in equations (A.174) and (A.175), we have

$$\hat{\boldsymbol{\beta}}_S = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{pmatrix} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(X_1, X_n) & \mathrm{Cov}(X_2, X_n) & \cdots & \mathrm{Var}(X_n) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{Cov}(X_1, X_{n+1}) \\ \mathrm{Cov}(X_2, X_{n+1}) \\ \vdots \\ \mathrm{Cov}(X_n, X_{n+1}) \end{bmatrix} \tag{7.23}$$

and

$$\hat{\beta}_0 = \mathrm{E}(X_{n+1}) - \sum_{i=1}^{n} \hat{\beta}_i \mathrm{E}(X_i) = \mu_X - \mu_X \sum_{i=1}^{n} \hat{\beta}_i. \tag{7.24}$$

From equation (7.17), we have

$$\mathrm{Var}(X_i) = \mu_{\mathrm{PV}} + \sigma^2_{\mathrm{HM}}, \qquad \text{for } i = 1, \cdots, n. \tag{7.25}$$

Also, $\mathrm{Cov}(X_i, X_j)$ is given by (for $i \neq j$)

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j) \\ &= \mathrm{E}\left[\mathrm{E}(X_i X_j \mid \Theta)\right] - \mu_X^2 \\ &= \mathrm{E}\left[\mathrm{E}(X_i \mid \Theta)\mathrm{E}(X_j \mid \Theta)\right] - \mu_X^2 \\ &= \mathrm{E}\left\{[\mu_X(\Theta)]^2\right\} - \{\mathrm{E}[\mu_X(\Theta)]\}^2 \\ &= \mathrm{Var}[\mu_X(\Theta)] \\ &= \sigma^2_{\mathrm{HM}}. \end{aligned} \tag{7.26}$$

Thus, equation (7.23) can be written as

$$\hat{\boldsymbol{\beta}}_S = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{pmatrix} = \left(\mu_{\mathrm{PV}}\mathbf{I} + \sigma^2_{\mathrm{HM}}\mathbf{1}\mathbf{1}'\right)^{-1} \left(\sigma^2_{\mathrm{HM}}\mathbf{1}\right), \tag{7.27}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{1}$ is the $n \times 1$ vector of ones. We now write

$$k = \frac{\mu_{\text{PV}}}{\sigma^2_{\text{HM}}}, \tag{7.28}$$

and evaluate the inverse matrix on the right-hand side of equation (7.27) as follows

$$\begin{aligned}\left(\mu_{\text{PV}}\mathbf{I} + \sigma^2_{\text{HM}}\mathbf{1}\mathbf{1}'\right)^{-1} &= \frac{1}{\mu_{\text{PV}}}\left(\mathbf{I} + \frac{\sigma^2_{\text{HM}}}{\mu_{\text{PV}}}\mathbf{1}\mathbf{1}'\right)^{-1} \\ &= \frac{1}{\mu_{\text{PV}}}\left(\mathbf{I} + \frac{1}{k}\mathbf{1}\mathbf{1}'\right)^{-1}. \end{aligned} \tag{7.29}$$

With $\mathbf{1}'\mathbf{1} = n$, it is easy to verify that

$$\left(\mathbf{I} + \frac{1}{k}\mathbf{1}\mathbf{1}'\right)^{-1} = \mathbf{I} - \frac{1}{n+k}\mathbf{1}\mathbf{1}'. \tag{7.30}$$

Substituting equation (7.30) into equations (7.27) and (7.29), we obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}}_S &= \frac{1}{\mu_{\text{PV}}}\left(\mathbf{I} - \frac{1}{n+k}\mathbf{1}\mathbf{1}'\right)\left(\sigma^2_{\text{HM}}\mathbf{1}\right) \\ &= \frac{1}{k}\left(\mathbf{I} - \frac{1}{n+k}\mathbf{1}\mathbf{1}'\right)\mathbf{1} \\ &= \frac{1}{k}\left(\mathbf{1} - \frac{n}{n+k}\mathbf{1}\right) \\ &= \frac{1}{n+k}\mathbf{1}. \end{aligned} \tag{7.31}$$

Note that equation (7.24) can be written as

$$\hat{\beta}_0 = \mu_X - \mu_X\hat{\boldsymbol{\beta}}'_S\mathbf{1}. \tag{7.32}$$

Thus, the least MSE linear predictor of X_{n+1} is

$$\begin{aligned}\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'_S\boldsymbol{X} &= (\mu_X - \mu_X\hat{\boldsymbol{\beta}}'_S\mathbf{1}) + \hat{\boldsymbol{\beta}}'_S\boldsymbol{X} \\ &= \left(\mu_X - \frac{\mu_X}{n+k}\mathbf{1}'\mathbf{1}\right) + \frac{1}{n+k}\mathbf{1}'\boldsymbol{X} \\ &= \frac{1}{n+k}\mathbf{1}'\boldsymbol{X} + \frac{k\mu_X}{n+k}. \end{aligned} \tag{7.33}$$

Noting that $\mathbf{1}'\boldsymbol{X} = n\bar{X}$, we conclude that

$$\hat{X}_{n+1} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_S'\boldsymbol{X} = \frac{n\bar{X}}{n+k} + \frac{k\mu_X}{n+k} = Z\bar{X} + (1-Z)\mu_X, \tag{7.34}$$

where

$$Z = \frac{n}{n+k}. \tag{7.35}$$

Z defined in equation (7.35) is called the **Bühlmann credibility factor** or simply the **Bühlmann credibility**. It depends on the EPV to VHM ratio k, which is called the **Bühlmann credibility parameter**. The optimal linear forecast $\hat{X}_{n+1}$ given in equation (7.34) is also called the **Bühlmann premium**. Note that k depends only on the parameters of the model,[5] while Z is a function of k and the size n of the *data*. For predicting claim frequency N, the sample size n is the number of periods over which the number of claims is aggregated.[6] For predicting claim severity X, the sample size n is the number of claims. As aggregate loss S refers to the total loss payout per period, the sample size is the number of periods of claim experience.

Example 7.7 Refer to Example 7.5. Suppose the claim experience last year was 26 claims with an average claim size of 12. Calculate the updated prediction of (a) the claim frequency, (b) the average claim size, and (c) the aggregate loss, for next year.

Solution (a) Claim frequency From Example 7.5, we have $k = 0.5714$ and $M = \mathrm{E}(N) = 32$. Now we are given $n = 1$ and $D = 26$. Hence

$$Z = \frac{1}{1+0.5714} = 0.6364,$$

so that the updated prediction of the claim frequency of this group is

$$U = (0.6364)(26) + (1-0.6364)(32) = 28.1816.$$

(b) Claim severity We have $k = 2.7717$ and $M = \mathrm{E}(X) = 8.75$, with $n = 26$ and $D = 12$. Thus

$$Z = \frac{26}{26+2.7717} = 0.9037,$$

[5] Hence, k is fixed, but needs to be estimated in practice. We shall come back to this issue in Chapter 9.

[6] Note that N is the number of claims per period, say year, and n is the number of periods of claim-frequency experience.

so that the updated prediction of the claim severity of this group is

$$U = (0.9037)(12) + (1 - 0.9037)(8.75) = 11.6870.$$

(c) Aggregate loss With $k = 0.7607$, $M = \mathrm{E}(S) = 280$, $n = 1$ and $D = (26)(12) = 312$, we have

$$Z = \frac{1}{1 + 0.7607} = 0.5680,$$

so that the updated prediction of the aggregate loss of this group is

$$U = (0.5680)(312) + (1 - 0.5680)(280) = 298.1760.$$

□

Example 7.8 Refer to Example 7.6. Suppose the numbers of claims in the last three years were 8, 4, and 7, with the corresponding average amount of claim in each of the three years being 12, 19, and 9. Calculate the updated prediction of the aggregate loss for next year for $c = 20$ and 30.

Solution We first calculate the average aggregate loss per year in the last three years, which is

$$\frac{1}{3}[(8)(12) + (4)(19) + (7)(9)] = 78.3333.$$

As shown in Example 7.6, $k = 0.6148$, which does not vary with c. As there are three observations of S, the Bühlmann credibility factor is

$$Z = \frac{3}{3 + 0.6148} = 0.8299.$$

The unconditional mean of S is

$$\mathrm{E}[\mathrm{E}(S \mid \Theta)] = 10c\mathrm{E}(\Theta^2) = (10)(0.29)c = 2.9c.$$

Hence, using equation (7.34), the updated prediction of S is

$$(0.8299)(78.3333) + (1 - 0.8299)(2.9c),$$

which gives a predicted value of 74.8746 when $c = 20$, and 79.8075 when $c = 30$. □

We have derived the Bühlmann credibility predictor for future loss as the linear predictor (in $\boldsymbol{X}$) that minimizes the mean squared prediction error in equation (7.19). However, we can also consider the problem of a linear *estimator* (in $\boldsymbol{X}$) that minimizes the squared error in estimating the *expected* future loss, i.e. $\mathrm{E}\big[(\mu_{n+1} - \hat{\mu}_{n+1})^2\big]$, where $\mu_{n+1} = \mathrm{E}(X_{n+1}) = \mu_X$ and $\hat{\mu}_{n+1}$ is a linear estimator of μ_{n+1}. Readers are invited to show that the result is the same as the Bühlmann credibility predictor for future loss (see Exercise 7.5). Thus, we shall use the terminologies Bühlmann credibility predictor for future loss and Bühlmann credibility estimator of the expected loss interchangeably.

7.4 Bühlmann–Straub credibility

An important limitation of the Bühlmann credibility theory is that the loss observations X_i are assumed to be *identically* distributed. This assumption is violated if the data are over different periods with different exposures (the definition of exposure will be explained below). The **Bühlmann–Straub credibility model** extends the **Bühlmann theory** to cases where the loss data X_i are not identically distributed. In particular, the process variance of the loss measure is assumed to depend on the exposure. We denote the exposure by m_i, and the *loss per unit of exposure* by X_i. Note that the exposure needs not be the number of insureds, although that may often be the case. We then assume the following for the conditional variance of X_i

$$\mathrm{Var}(X_i \mid \Theta) = \frac{\sigma_X^2(\Theta)}{m_i}, \tag{7.36}$$

for a suitably defined $\sigma_X^2(\Theta)$. The following are some examples:

1 X_i is the average number of claims per insured in year i, $\sigma_X^2(\Theta)$ is the variance of the claim frequency of an insured, and the exposure m_i is the number of insureds covered in year i.[7]
2 X_i is the average aggregate loss per month of the ith block of policies, $\sigma_X^2(\Theta)$ is the variance of the aggregate loss of the block in a month, and the exposure m_i is the number of months of insurance claims for the ith block of policies.
3 X_i is the average loss per unit premium in year i, $\sigma_X^2(\Theta)$ is the variance of the claim amount of an insured per year divided by the premium per insured, and the exposure m_i is the amount of premiums received in year i. To see this, assume there are N_i insured in year i, each paying a premium P. Thus,

[7] Note that $\mathrm{Var}(X_i \mid \Theta)$ is the variance of the *average claim frequency per insured*, while $\sigma_X^2(\Theta)$ is the variance of the *claim frequency of each insured*.

$m_i = N_i P$ and

$$\begin{aligned}\mathrm{Var}(X_i \mid \Theta) &= \mathrm{Var}\left(\frac{\text{loss per insured}}{P}\right)\\ &= \frac{1}{P^2}\left[\frac{\mathrm{Var}(\text{claim amount of an insured})}{N_i}\right]\\ &= \frac{1}{m_i}\left[\frac{\mathrm{Var}(\text{claim amount of an insured})}{P}\right]\\ &= \frac{\sigma_X^2(\Theta)}{m_i}. \end{aligned} \tag{7.37}$$

In each of the examples above, the distributions of X_i are not identical. Instead, the conditional variance of X_i varies with m_i. We now formally summarize below the assumptions of the Bühlmann–Straub credibility model.

1 Let m_i be the exposure in period i and X_i be the loss per unit exposure, for $i = 1, \cdots, n$. Suppose $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ are independently (but not identically) distributed and the distribution of X_i depends on the parameter θ.
2 The parameter θ is a realization of a random variable Θ. Given θ, the conditional mean and variance of $\boldsymbol{X}$ are

$$\mathrm{E}(X_i \mid \theta) = \mu_X(\theta), \tag{7.38}$$

and

$$\mathrm{Var}(X_i \mid \theta) = \frac{\sigma_X^2(\theta)}{m_i}, \tag{7.39}$$

for $i \in \{1, \cdots, n\}$, where $\sigma_X^2(\theta)$ is suitably defined as in the examples above.
3 The unconditional mean of X_i is $\mathrm{E}(X_i) = \mathrm{E}[\mathrm{E}(X_i \mid \Theta)] = \mathrm{E}[\mu_X(\Theta)] = \mu_X$. The mean of the conditional variance of X_i is

$$\begin{aligned}\mathrm{E}[\mathrm{Var}(X_i \mid \Theta)] &= \mathrm{E}\left[\frac{\sigma_X^2(\Theta)}{m_i}\right]\\ &= \frac{\mu_{\mathrm{PV}}}{m_i}, \end{aligned} \tag{7.40}$$

for $i \in \{1, \cdots, n\}$, where $\mu_{\mathrm{PV}} = \mathrm{E}[\sigma_X^2(\Theta)]$, and the variance of its conditional mean is

$$\begin{aligned}\mathrm{Var}[\mathrm{E}(X_i \mid \Theta)] &= \mathrm{Var}[\mu_X(\Theta)]\\ &= \sigma_{\mathrm{HM}}^2. \end{aligned} \tag{7.41}$$

4 The Bühlmann–Straub predictor minimizes the MSE of all predictors of X_{n+1} that are linear in $\boldsymbol{X}$ over the joint distribution of Θ, X_{n+1} and $\boldsymbol{X}$. The predictor is given by

$$\hat{X}_{n+1} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n, \tag{7.42}$$

where $\beta_0, \beta_1, \cdots, \beta_n$ are chosen to minimize the MSE of the predictor.

The steps in the last section can be used to solve for the least MSE predictor of the Bühlmann–Straub model. In particular, equations (7.22), (7.23), and (7.24) hold. The solution differs due to the variance term in equation (7.25). First, for the total variance of X_i, we have

$$\begin{aligned} \mathrm{Var}(X_i) &= \mathrm{E}[\mathrm{Var}(X_i \mid \Theta)] + \mathrm{Var}[\mathrm{E}(X_i \mid \Theta)] \\ &= \frac{\mu_{\mathrm{PV}}}{m_i} + \sigma^2_{\mathrm{HM}}. \end{aligned} \tag{7.43}$$

Second, following the same argument as in equation (7.26), the covariance terms are

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathrm{Var}[\mu_X(\Theta)] \\ &= \sigma^2_{\mathrm{HM}}, \end{aligned} \tag{7.44}$$

for $i \neq j$. Now equation (7.27) can be written as

$$\hat{\boldsymbol{\beta}}_S = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{pmatrix} = \left(\mathbf{V} + \sigma^2_{\mathrm{HM}} \mathbf{1}\mathbf{1}'\right)^{-1} \left(\sigma^2_{\mathrm{HM}} \mathbf{1}\right), \tag{7.45}$$

where $\mathbf{V}$ is the $n \times n$ diagonal matrix

$$\mathbf{V} = \mu_{\mathrm{PV}} \begin{bmatrix} m_1^{-1} & 0 & 0 & \cdots & 0 \\ 0 & m_2^{-1} & 0 & \cdots & 0 \\ 0 & 0 & & & \\ \vdots & \vdots & & & \\ 0 & 0 & & \cdots & m_n^{-1} \end{bmatrix}. \tag{7.46}$$

It can be verified that

$$\left(\mathbf{V} + \sigma^2_{\mathrm{HM}} \mathbf{1}\mathbf{1}'\right)^{-1} = \mathbf{V}^{-1} - \frac{\sigma^2_{\mathrm{HM}} (\mathbf{V}^{-1}\mathbf{1})(\mathbf{1}'\mathbf{V}^{-1})}{1 + \sigma^2_{\mathrm{HM}} \mathbf{1}'\mathbf{V}^{-1}\mathbf{1}}. \tag{7.47}$$

Thus, denoting

$$m = \sum_{i=1}^{n} m_i \tag{7.48}$$

and $\boldsymbol{m} = (m_1, \cdots, m_n)'$, we have

$$\mathbf{V}^{-1} - \frac{\sigma_{\mathrm{HM}}^2(\mathbf{V}^{-1}\mathbf{1})(\mathbf{1}'\mathbf{V}^{-1})}{1 + \sigma_{\mathrm{HM}}^2\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}} = \mathbf{V}^{-1} - \frac{1}{\mu_{\mathrm{PV}}}\left(\frac{\sigma_{\mathrm{HM}}^2\boldsymbol{m}\boldsymbol{m}'}{\mu_{\mathrm{PV}} + m\sigma_{\mathrm{HM}}^2}\right), \tag{7.49}$$

so that from equation (7.45) we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_S &= \left[\mathbf{V}^{-1} - \frac{1}{\mu_{\mathrm{PV}}}\left(\frac{\sigma_{\mathrm{HM}}^2\boldsymbol{m}\boldsymbol{m}'}{\mu_{\mathrm{PV}} + m\sigma_{\mathrm{HM}}^2}\right)\right]\left(\sigma_{\mathrm{HM}}^2\mathbf{1}\right) \\
&= \left(\frac{\sigma_{\mathrm{HM}}^2}{\mu_{\mathrm{PV}}}\right)\boldsymbol{m} - \left(\frac{\sigma_{\mathrm{HM}}^2}{\mu_{\mathrm{PV}}}\right)\left(\frac{\sigma_{\mathrm{HM}}^2 m\boldsymbol{m}}{\mu_{\mathrm{PV}} + m\sigma_{\mathrm{HM}}^2}\right) \\
&= \left(\frac{\sigma_{\mathrm{HM}}^2}{\mu_{\mathrm{PV}}}\right)\left(1 - \frac{\sigma_{\mathrm{HM}}^2 m}{\mu_{\mathrm{PV}} + m\sigma_{\mathrm{HM}}^2}\right)\boldsymbol{m} \\
&= \frac{\sigma_{\mathrm{HM}}^2\boldsymbol{m}}{\mu_{\mathrm{PV}} + m\sigma_{\mathrm{HM}}^2}.
\end{aligned} \tag{7.50}$$

We now define

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{n} m_i X_i = \frac{1}{m}\boldsymbol{m}'\boldsymbol{X} \tag{7.51}$$

and

$$k = \frac{\mu_{\mathrm{PV}}}{\sigma_{\mathrm{HM}}^2} \tag{7.52}$$

to obtain

$$\hat{\boldsymbol{\beta}}_S'\boldsymbol{X} = \frac{\sigma_{\mathrm{HM}}^2\boldsymbol{m}'\boldsymbol{X}}{\mu_{\mathrm{PV}} + m\sigma_{\mathrm{HM}}^2} = \frac{m}{m+k}\bar{X} = Z\bar{X}, \tag{7.53}$$

where

$$Z = \frac{m}{m+k}. \tag{7.54}$$

If we replace $\boldsymbol{X}$ in equation (7.53) by $\mathbf{1}$, we have

$$\hat{\boldsymbol{\beta}}_S'\mathbf{1} = Z, \tag{7.55}$$

so that from equation (7.32) we obtain

$$\hat{\beta}_0 = \mu_X - \mu_X \hat{\boldsymbol{\beta}}_S' \mathbf{1} = (1 - Z)\mu_X . \tag{7.56}$$

Combining the results in equations (7.53) and (7.56), we conclude that

$$\hat{X}_{n+1} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_S' \boldsymbol{X} = Z\bar{X} + (1 - Z)\mu_X , \tag{7.57}$$

where Z is defined in equation (7.54).

Example 7.9 The number of accident claims incurred per year for each insured is distributed as a binomial random variable $\mathcal{BN}(2, \theta)$, and the claim incidences are independent across insureds. The probability θ of the binomial has a beta distribution with parameters $\alpha = 1$ and $\beta = 10$. The data in Table 7.7 are given for a block of policies.

Table 7.7. *Data for Example 7.9*

Year	Number of insureds	Number of claims
1	100	7
2	200	13
3	250	18
4	280	–

Calculate the Bühlmann–Straub credibility prediction of the number of claims in the fourth year.

Solution Let m_i be the number of insureds in Year i, and X_i be the number of claims per insured in Year i. Define X_{ij} as the number of claims for the jth insured in Year i, which is distributed as $\mathcal{BN}(2, \theta)$. Thus, we have

$$\mathrm{E}(X_i \mid \Theta) = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathrm{E}(X_{ij} \mid \Theta) = 2\Theta,$$

and

$$\sigma^2_{\mathrm{HM}} = \mathrm{Var}[\mathrm{E}(X_i \mid \Theta)] = \mathrm{Var}(2\Theta) = 4\mathrm{Var}(\Theta).$$

As Θ has a beta distribution with parameters $\alpha = 1$ and $\beta = 10$, we have[8]

$$\mathrm{Var}(\Theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{10}{(11)^2(12)} = 0.006887.$$

[8] See Appendix A.10.6 for the moments of the beta distribution.

For the conditional variance of X_i, we have

$$\text{Var}(X_i \mid \Theta) = \frac{2\Theta(1-\Theta)}{m_i}.$$

Thus

$$\mu_{\text{PV}} = 2\text{E}[\Theta(1-\Theta)].$$

As

$$\text{E}(\Theta) = \frac{\alpha}{\alpha+\beta} = 0.0909,$$

we have

$$\begin{aligned}\mu_{\text{PV}} &= 2[\text{E}(\Theta) - \text{E}(\Theta^2)] \\ &= 2\left\{\text{E}(\Theta) - \left(\text{Var}(\Theta) + [\text{E}(\Theta)]^2\right)\right\} \\ &= 2\{0.0909 - [0.006887 + (0.0909)^2]\} = 0.1515.\end{aligned}$$

Thus

$$k = \frac{\mu_{\text{PV}}}{\sigma^2_{\text{HM}}} = \frac{0.1515}{(4)(0.006887)} = 5.5.$$

As $m = 100 + 200 + 250 = 550$, we have

$$Z = \frac{550}{550 + 5.5} = 0.9901.$$

Now

$$\mu_X = \text{E}[\text{E}(X_i \mid \Theta)] = (2)(0.0909) = 0.1818$$

and

$$\bar{X} = \frac{7 + 13 + 18}{550} = 0.0691.$$

Thus, the predicted number of claims per insured is

$$(0.9901)(0.0691) + (1 - 0.9901)(0.1818) = 0.0702,$$

and the predicted number of claims in Year 4 is

$$(280)(0.0702) = 19.66.$$

□

Example 7.10 The number of accident claims incurred per year for each insured is a Bernoulli random variable with probability θ, which takes value 0.1 with probability 0.8 and 0.2 with probability 0.2. Each claim may be of amount 20, 30, or 40, with equal probabilities. Claim frequency and claim severity are assumed to be independent for each insured. The data for the total claim amount are given in Table 7.8.

Table 7.8. *Data for Example 7.10*

Year	Number of insureds	Total claim amount
1	100	240
2	200	380
3	250	592
4	280	–

Calculate the Bühlmann–Straub credibility prediction of the pure premium and total loss in the fourth year.

Solution Let X_{ij} be the claim amount for the jth insured in Year i, each of which is distributed as $X = NW$, where

$$N = \begin{cases} 0, & \text{with probability } 1 - \Theta, \\ 1, & \text{with probability } \Theta, \end{cases}$$

and $W = 20$, 30, and 40, with equal probabilities. We have

$$\mathrm{E}(N \mid \Theta) = \Theta,$$

and

$$\mathrm{Var}(N \mid \Theta) = \Theta(1 - \Theta).$$

We evaluate the moments of Θ to obtain

$$\mathrm{E}(\Theta) = (0.1)(0.8) + (0.2)(0.2) = 0.12,$$

and

$$\mathrm{E}(\Theta^2) = (0.1)^2(0.8) + (0.2)^2(0.2) = 0.016.$$

Also

$$\mathrm{E}(W) = \frac{20 + 30 + 40}{3} = 30,$$

and

$$\mathrm{E}(W^2) = \frac{(20)^2 + (30)^2 + (40)^2}{3} = 966.6667,$$

so that

$$\mathrm{E}(X \mid \Theta) = \mathrm{E}(N \mid \Theta)\mathrm{E}(W) = 30\Theta.$$

Using equation (A.118), we have

$$\begin{aligned}\mathrm{Var}(X \mid \Theta) &= \mathrm{E}(W^2)\mathrm{Var}(N \mid \Theta) + [\mathrm{E}(N \mid \Theta)]^2\mathrm{Var}(W) \\ &= 966.6667\Theta(1 - \Theta) + \Theta^2[966.6667 - (30)^2] \\ &= 966.6667\Theta - 900\Theta^2.\end{aligned}$$

Thus, EPV is

$$\mu_{\mathrm{PV}} = \mathrm{E}[\mathrm{Var}(X \mid \Theta)] = (966.6667)(0.12) - (900)(0.016) = 101.60,$$

and VHM is

$$\begin{aligned}\sigma^2_{\mathrm{HM}} &= \mathrm{Var}[\mathrm{E}(X \mid \Theta)] \\ &= \mathrm{Var}(30\Theta) \\ &= 900\left\{\mathrm{E}(\Theta^2) - [\mathrm{E}(\Theta)]^2\right\} \\ &= 900[0.016 - (0.12)^2] \\ &= 1.44.\end{aligned}$$

Thus

$$k = \frac{\mu_{\mathrm{PV}}}{\sigma^2_{\mathrm{HM}}} = \frac{101.60}{1.44} = 70.5556.$$

As $m = 100 + 200 + 250 = 550$

$$Z = \frac{550}{550 + 70.5556} = 0.8863.$$

Now

$$\bar{X} = \frac{240 + 380 + 592}{550} = 2.2036,$$

and

$$\mu_X = \mathrm{E}(X) = 30\mathrm{E}(\Theta) = (30)(0.12) = 3.60,$$

so that the Bühlmann–Straub prediction for the pure premium is

$$(0.8863)(2.2036) + (1 - 0.8863)(3.60) = 2.3624,$$

and the Bühlmann–Straub prediction for the total claim amount in Year 4 is

$$(280)(2.3624) = 661.4638.$$

□

7.5 Summary and discussions

The Bühlmann–Straub credibility model is a generalization of the Bühlmann credibility model, and we summarize its main results again here. Let X_i be the loss measure per unit exposure in period i, with the amount of exposure being m_i, for $i = 1, \cdots, n$. Let $m = \sum_{i=1}^{n} m_i$ and Θ be the parameter determining the distribution of X_i. Assume X_i are independently distributed.

Let

$$\mu_{\mathrm{PV}} = m_i \, \mathrm{E}[\mathrm{Var}(X_i \mid \Theta)], \tag{7.58}$$

and

$$\sigma^2_{\mathrm{HM}} = \mathrm{Var}[\mathrm{E}(X_i \mid \Theta)], \tag{7.59}$$

then the Bühlmann–Straub prediction of X_{n+1} is

$$\hat{X}_{n+1} = Z\bar{X} + (1 - Z)\mu_X, \tag{7.60}$$

where $\mu_X = \mathrm{E}(X_i)$, for $i = 1, \cdots, n$, and

$$\bar{X} = \frac{1}{m} \sum_{i=1}^{n} m_i X_i, \tag{7.61}$$

and

$$Z = \frac{m}{m + k}, \tag{7.62}$$

with

$$k = \frac{\mu_{\text{PV}}}{\sigma^2_{\text{HM}}}. \tag{7.63}$$

In the special case where the exposures of all periods are the same, say $m_i = \bar{m}$ for $i = 1, \cdots, n$, then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{7.64}$$

and

$$Z = \frac{n\bar{m}}{n\bar{m} + \dfrac{\mu_{\text{PV}}}{\sigma^2_{\text{HM}}}} = \frac{n}{n + \dfrac{\mu_{\text{PV}}}{\bar{m}\sigma^2_{\text{HM}}}} = \frac{n}{n + \dfrac{\text{E}[\text{Var}(X_i \mid \Theta)]}{\sigma^2_{\text{HM}}}}. \tag{7.65}$$

Thus, the Bühlmann–Straub credibility predictor can be specialized to the Bühlmann predictor.

For the examples given in this chapter we assume that the variance components are known. In practice, they have to be estimated from the data. In Chapter 9 we shall consider the empirical implementation of the Bühlmann and Bühlmann–Straub credibility models when EPV and VHM are unknown. While we have proved the optimality of the Bühlmann predictor in the class of *linear* predictors, it turns out that its optimality may be more general. We shall see the details of this in the next chapter.

Exercises

7.1 Refer to Example 7.6. Calculate the unconditional covariance between X and N, $\text{Cov}(X, N)$.

7.2 Refer to Example 7.6. Find the Bühlmann credibility parameters for claim frequency N and claim severity X.

7.3 Refer to Example 7.6. If the claim-severity distribution X is gamma with parameters $\alpha = c\,\theta$ and $\beta = 1/c$, derive an expression of the Bühlmann credibility parameter k for the aggregate loss per year S in terms of c.

7.4 Refer to Example 7.8. Calculate the updated prediction for claim frequency and claim severity for next year, for $c = 20$ and 30.

7.5 Following the set-up in Section 7.3, let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample of losses that are independently and identically distributed as the random variable X, the distribution of which depends on a risk parameter θ. Denote $\mu_{n+1}(\Theta) = \text{E}(X_{n+1}|\Theta)$, and consider the

problem of estimating $\mu_{n+1}(\Theta)$ by a linear function of $\boldsymbol{X}$, denoted by $\hat{\mu}_{n+1}$, that minimizes the mean squared error $\mathrm{E}\big[(\mu_{n+1}(\Theta) - \hat{\mu}_{n+1})^2\big]$. Show that $\hat{\mu}_{n+1}$ is the same as the Bühlmann credibility predictor for future loss given in equations (7.34) and (7.35).

Questions adapted from SOA exams

7.6 The Bühlmann credibility assigned for estimating X_5 based on $X_1, \cdots, X_4$ is $Z = 0.4$. If the expected value of the process variance is 8, calculate $\mathrm{Cov}(X_i, X_j)$ for $i \neq j$.

7.7 The annual number of claims of a policy is distributed as $\mathcal{GM}(1/(1+\theta))$. If Θ follows the $\mathcal{P}(\alpha, 1)$ distribution, where $\alpha > 2$, and a randomly selected policy has x claims in Year 1, derive the Bühlmann credibility estimate of the expected number of claims of the policy in Year 2.

7.8 For a portfolio of insurance policies the annual claim amount X of a policy has the following pdf

$$f_X(x \mid \theta) = \frac{2x}{\theta^2}, \qquad 0 < x < \theta.$$

The prior distribution of Θ has the following pdf

$$f_\Theta(\theta) = 4\theta^3, \qquad 0 < \theta < 1.$$

A randomly selected policy has claim amount 0.1 in Year 1. Determine the Bühlmann credibility estimate of the expected claim amount of the selected policy in Year 2.

7.9 The number of claims in a year of a selected risk group follows the $\mathcal{PN}(\lambda)$ distribution. Claim severity follows the $\mathcal{E}(1/\theta)$ distribution and is independent of the claim frequency. If $\Lambda \sim \mathcal{E}(1)$ and $\Theta \sim \mathcal{PN}(1)$, and Λ and Θ are independent, determine the Bühlmann credibility parameter k for the estimation of the expected annual aggregate loss.

7.10 An insurance company sells two types of policies with the following characteristics:

Type of policy	Proportion of policies	Annual claim frequency, Poisson
1	θ	$\lambda = 0.5$
2	$1 - \theta$	$\lambda = 1.5$

A randomly selected policyholder has one claim in Year 1. Determine the Bühlmann credibility factor Z of this policyholder.

7.11 Claim frequency follows a Poisson distribution with mean λ. Claim size follows an exponential distribution with mean 10λ and is independent of claim frequency. If the distribution of Λ has pdf

$$f_\Lambda(\lambda) = \frac{5}{\lambda^6}, \qquad \lambda > 1,$$

calculate the Bühlmann credibility parameter k for aggregate losses.

7.12 Two risks have the following severity distributions:

Claim amount	Probability of claim amount for Risk 1	Probability of claim amount for Risk 2
250	0.5	0.7
2,500	0.3	0.2
60,000	0.2	0.1

If Risk 1 is twice as likely to be observed as Risk 2 and a claim of 250 is observed, determine the Bühlmann credibility estimate of the expected second claim amount from the same risk.

7.13 Claim frequency in a month is distributed as $\mathcal{PN}(\lambda)$, and the distribution of Λ is $\mathcal{G}(6, 0.01)$. The following data are available:

Month	Number of insureds	Number of claims
1	100	6
2	150	8
3	200	11
4	300	–

Calculate the Bühlmann–Straub credibility estimate of the expected number of claims in Month 4.

7.14 The number of claims made by an individual insured in a year is distributed as $\mathcal{PN}(\lambda)$, where Λ is distributed as $\mathcal{G}(1, 1.2)$. If three claims are observed in Year 1 and no claim is observed in Year 2, calculate the Bühlmann credibility estimate of the expected number of claims in Year 3.

7.15 Annual claim frequency of an individual policyholder has mean λ, which is distributed as $\mathcal{U}(0.5, 1.5)$, and variance σ^2, which is

distributed as exponential with mean 1.25. A policyholder is selected randomly and found to have no claim in Year 1. Using Bühlmann credibility, estimate the expected number of claims in Year 2 for the selected policyholder.

7.16 You are given the following joint distribution of X and Θ:

	Θ	
X	0	1
0	0.4	0.1
1	0.1	0.2
2	0.1	0.1

For a given (but unknown) value of Θ and a sample of ten observations of X with a total of 10, determine the Bühlmann credibility premium.

7.17 There are four classes of insureds, each of whom may have zero or one claim, with the following probabilities:

	Number of claims	
Class	0	1
A	0.9	0.1
B	0.8	0.2
C	0.5	0.5
D	0.1	0.9

A class is selected randomly, with probability of one-fourth, and four insureds are selected at random from the class. The total number of claims is two. If five insureds are selected at random from the same class, determine the Bühlmann–Straub credibility estimate of the expected total number of claims.

7.18 An insurance company has a large portfolio of insurance policies. Each insured may file a maximum of one claim per year, and the probability of a claim for each insured is constant over time. A randomly selected insured has a probability 0.1 of filing a claim in a year, and the variance of the claim probability of individual insured is 0.01. A randomly selected individual is found to have filed no claim over the past ten years. Determine the Bühlmann credibility estimate for the expected number of claims the selected insured will file over the next five years.

7.19 A portfolio of insurance policies comprises of 100 insureds. The aggregate loss of each insured in a year follows a compound distribution, where the primary distribution is $\mathcal{NB}(r, 1/1.2)$ and the secondary distribution is $\mathcal{P}(3, 1000)$. If the distribution of r is exponential with mean 2, determine the Bühlmann credibility factor Z of the portfolio.

7.20 An insurance company has a large portfolio of employee compensation policies. The losses of each employee are independently and identically distributed. The overall average loss of each employee is 20, the variance of the hypothetical means is 40, and the expected value of the process variance is 8,000. The following data are available in the last three years for a randomly selected policyholder:

Year	Average loss per employee	Number of employees
1	15	800
2	10	600
3	5	400

Determine the Bühlmann–Straub credibility premium per employee for this policyholder.

7.21 Claim severity has mean μ and variance 500, where the distribution of μ has a mean of 1,000 and a variance of 50. The following three claims were observed: 750, 1,075 and 2,000. Calculate the Bühlmann estimate of the expected claim severity of the next claim.

7.22 Annual losses are distributed as $\mathcal{G}(\alpha, \beta)$, where β does not vary with policyholders. The distribution of α has a mean of 50, and the Bühlmann credibility factor based on two years of experience is 0.25. Calculate the variance of the distribution of α.

7.23 The aggregate losses per year per exposure of a portfolio of insurance risks follow a normal distribution with mean μ and standard deviation 1,000. You are given that μ varies by class of risk as follows:

Class	μ	Probability of class
A	2,000	0.6
B	3,000	0.3
C	4,000	0.1

A randomly selected risk has the following experience over three years:

Year	Number of exposures	Aggregate losses
1	24	24,000
2	30	36,000
3	26	28,000

Calculate the Bühlmann–Straub estimate of the expected aggregate loss per exposure in Year 4 for this risk.

7.24 The annual loss of an individual policy is distributed as $\mathcal{G}(4, \beta)$, where the mean of the distribution of β is 600. A randomly selected policy had losses of 1,400 in Year 1 and 1,900 in Year 2. Loss data for Year 3 were misfiled and the Bühlmann credibility estimate of the expected loss for the selected policy in Year 4 based on the data for Years 1 and 2 was 1,800. The loss for the selected policy in Year 3, however, was found later to be 2,763. Determine the Bühlmann credibility estimate of the expected loss for the selected policy in Year 4 based on the data of Years 1, 2, and 3.

7.25 Claim frequency in a month is distributed as $\mathcal{PN}(\lambda)$, where λ is distributed as $\mathcal{W}(2, 0.1)$. You are given the following data:

Month	Number of insureds	Number of claims
1	100	10
2	150	11
3	250	14

Calculate the Bühlmann–Straub credibility estimate of the expected number of claims in the next 12 months for 300 insureds. [*Hint:* You may use the Excel function GAMMALN to compute the natural logarithm of the gamma function.]

8

Bayesian approach

In this chapter we consider the Bayesian approach in updating the prediction for future losses. We consider the derivation of the posterior distribution of the risk parameters based on the prior distribution of the risk parameters and the likelihood function of the data. The Bayesian estimate of the risk parameter under the squared-error loss function is the mean of the posterior distribution. Likewise, the Bayesian estimate of the mean of the random loss is the posterior mean of the loss conditional on the data.

In general, the Bayesian estimates are difficult to compute, as the posterior distribution may be quite complicated and intractable. There are, however, situations where the computation may be straightforward, as in the case of conjugate distributions. We define conjugate distributions and provide some examples for cases that are of relevance in analyzing loss measures. Under specific classes of conjugate distributions, the Bayesian predictor is the same as the Bühlmann predictor. Specifically, when the likelihood belongs to the linear exponential family and the prior distribution is the natural conjugate, the Bühlmann credibility estimate is equal to the Bayesian estimate. This result provides additional justification for the use of the Bühlmann approach.

Learning objectives

1 Bayesian inference and estimation
2 Prior and posterior pdf
3 Bayesian credibility
4 Conjugate prior distribution
5 Linear exponential distribution
6 Bühlmann credibility versus Bayesian credibility

8.1 Bayesian inference and estimation

The classical and Bühlmann credibility models update the prediction for future losses based on recent claim experience and existing prior information. In these models, the random loss variable X has a distribution that varies with different risk groups. Based on a sample of n observations of random losses, the predicted value of the loss for the next period is updated. The predictor is a weighted average of the sample mean of X and the prior mean, where the weights depend on the distribution of X across different risk groups.

We formulate the aforementioned as a statistical problem suitable for the Bayesian approach of statistical inference and estimation. The set-up is summarized as follows:[1]

1 Let X denote the random loss variable (such as claim frequency, claim severity, and aggregate loss) of a risk group. The distribution of X is dependent on a parameter θ, which varies with different risk groups and is hence treated as the realization of a random variable Θ.
2 Θ has a statistical distribution called the **prior distribution**. The **prior pdf** of Θ is denoted by $f_{\Theta}(\theta \mid \gamma)$ (or simply $f_{\Theta}(\theta)$), which depends on the parameter γ, called the **hyperparameter**.
3 The conditional pdf of X given the parameter θ is denoted by $f_{X \mid \Theta}(x \mid \theta)$. Suppose $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ is a random sample of X, and $\boldsymbol{x} = (x_1, \cdots, x_n)$ is a realization of $\boldsymbol{X}$. The conditional pdf of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} f_{X \mid \Theta}(x_i \mid \theta). \tag{8.1}$$

We call $f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta)$ the **likelihood function**.
4 Based on the sample data $\boldsymbol{x}$, the distribution of Θ is updated. The conditional pdf of Θ given $\boldsymbol{x}$ is called the **posterior pdf**, and is denoted by $f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})$.
5 An estimate of the mean of the random loss, which is a function of Θ, is computed using the posterior pdf of Θ. This estimate, called the **Bayes estimate**, is also the predictor of future losses.

Bayesian inference differs from classical statistical inference in its treatment of the prior distribution of the parameter θ. Under classical statistical inference, θ is assumed to be *fixed* and *unknown*, and the relevant entity for inference is the likelihood function. For Bayesian inference, the prior distribution has an important role. The likelihood function and the prior pdf jointly determine the posterior pdf, which is then used for statistical inference.

[1] For convenience of exposition, we assume all distributions (both the prior and the likelihood) are continuous. Thus, we use the terminology "pdf" and compute the marginal pdf using integration. If the distribution is discrete, we need to replace "pdf" by "pf", and use summation instead of integration. A brief introduction to Bayesian inference can be found in Appendix A.15.

We now discuss the derivation of the posterior pdf and the Bayesian approach of estimating Θ.

8.1.1 Posterior distribution of parameter

Given the prior pdf of Θ and the likelihood function of $\boldsymbol{X}$, the joint pdf of Θ and X can be obtained as follows

$$f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x}) = f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_{\Theta}(\theta). \tag{8.2}$$

Integrating out θ from the joint pdf of Θ and $\boldsymbol{X}$, we obtain the marginal pdf of $\boldsymbol{X}$ as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{\theta \in \Omega_{\Theta}} f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_{\Theta}(\theta)\, d\theta, \tag{8.3}$$

where Ω_{Θ} is the support of Θ.

Now we can turn the question around and consider the conditional pdf of Θ given the data $\boldsymbol{x}$, i.e. $f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})$. Combining equations (8.2) and (8.3), we have

$$\begin{aligned} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) &= \frac{f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_{\Theta}(\theta)}{\int_{\theta \in \Omega_{\Theta}} f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_{\Theta}(\theta)\, d\theta}. \end{aligned} \tag{8.4}$$

The posterior pdf describes the distribution of Θ based on prior information about Θ and the sample data $\boldsymbol{x}$. Bayesian inference about the population as described by the risk parameter Θ is then based on the posterior pdf.

Example 8.1 Let X be the Bernoulli random variable which takes value 1 with probability θ and 0 with probability $1 - \theta$. If Θ follows the beta distribution with parameters α and β, i.e. $\Theta \sim \mathcal{B}(\alpha, \beta)$, calculate the posterior pdf of Θ given X.[2]

Solution As X is Bernoulli, the likelihood function of X is

$$f_{X \mid \Theta}(x \mid \theta) = \theta^{x}(1 - \theta)^{1-x}, \qquad \text{for } x = 0, 1.$$

Since Θ is assumed to follow the beta distribution with hyperparameters α and β, the prior pdf of Θ is

$$f_{\Theta}(\theta) = \frac{\theta^{\alpha-1}(1 - \theta)^{\beta-1}}{B(\alpha, \beta)}, \qquad \text{for } \theta \in (0, 1),$$

[2] See Appendix A.10.6 for some properties of the $\mathcal{B}(\alpha, \beta)$ distribution. The support of $\mathcal{B}(\alpha, \beta)$ is the interval (0, 1), so that the distribution is suitable for modeling probability as a random variable.

where $B(\alpha, \beta)$ is the beta function defined in equation (A.102). Thus, the joint pf–pdf of Θ and X is

$$f_{\Theta X}(\theta, x) = f_{X \mid \Theta}(x \mid \theta) f_{\Theta}(\theta) = \frac{\theta^{\alpha+x-1}(1-\theta)^{(\beta-x+1)-1}}{B(\alpha, \beta)},$$

from which we compute the marginal pf of X by integration to obtain

$$\begin{aligned} f_X(x) &= \int_0^1 \frac{\theta^{\alpha+x-1}(1-\theta)^{(\beta-x+1)-1}}{B(\alpha, \beta)} \, d\theta \\ &= \frac{B(\alpha + x, \beta - x + 1)}{B(\alpha, \beta)}. \end{aligned}$$

Thus, we conclude

$$\begin{aligned} f_{\Theta \mid X}(\theta \mid x) &= \frac{f_{\Theta X}(\theta, x)}{f_X(x)} \\ &= \frac{\theta^{\alpha+x-1}(1-\theta)^{(\beta-x+1)-1}}{B(\alpha + x, \beta - x + 1)}, \end{aligned}$$

which is the pdf of a beta distribution with parameters $\alpha + x$ and $\beta - x + 1$. □

Example 8.2 In Example 8.1, if there is a sample of n observations of X denoted by $\boldsymbol{X} = \{X_1, \cdots, X_n\}$, compute the posterior pdf of Θ.

Solution We first compute the likelihood of $\boldsymbol{X}$ as follows

$$\begin{aligned} f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{\sum_{i=1}^n (1-x_i)}, \end{aligned}$$

and the joint pf–pdf is

$$\begin{aligned} f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x}) &= f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_{\Theta}(\theta) \\ &= \left[\theta^{\sum_{i=1}^n x_i}(1-\theta)^{\sum_{i=1}^n (1-x_i)}\right] \left[\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}\right] \\ &= \frac{\theta^{(\alpha+n\bar{x})-1}(1-\theta)^{(\beta+n-n\bar{x})-1}}{B(\alpha, \beta)}. \end{aligned}$$

As

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \int_0^1 f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x})\, d\theta \\ &= \int_0^1 \frac{\theta^{(\alpha+n\bar{x})-1}(1-\theta)^{(\beta+n-n\bar{x})-1}}{B(\alpha,\beta)}\, d\theta \\ &= \frac{B(\alpha+n\bar{x}, \beta+n-n\bar{x})}{B(\alpha,\beta)}, \end{aligned}$$

we conclude that

$$\begin{aligned} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) &= \frac{f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{\theta^{(\alpha+n\bar{x})-1}(1-\theta)^{(\beta+n-n\bar{x})-1}}{B(\alpha+n\bar{x}, \beta+n-n\bar{x})}, \end{aligned}$$

and the posterior pdf of Θ follows a beta distribution with parameters $\alpha + n\bar{x}$ and $\beta + n - n\bar{x}$. □

Note that the denominator in equation (8.4) is a function of $\boldsymbol{x}$ but not θ. Denoting

$$K(\boldsymbol{x}) = \frac{1}{\int_{\theta \in \Omega_\Theta} f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta)\, d\theta}, \tag{8.5}$$

we can rewrite the posterior pdf of Θ as

$$\begin{aligned} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) &= K(\boldsymbol{x}) f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta) \\ &\propto f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta). \end{aligned} \tag{8.6}$$

$K(\boldsymbol{x})$ is free of θ and is a **constant of proportionality**. It scales the posterior pdf so that it integrates to 1. The expression $f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta)$ enables us to identify the functional form of the posterior pdf in terms of θ without computing the marginal pdf of $\boldsymbol{X}$.

Example 8.3 Let $X \sim \mathcal{BN}(m, \theta)$, and $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample of X. If $\Theta \sim \mathcal{B}(\alpha, \beta)$, what is the posterior distribution of Θ?

Solution From equation (8.6), we have

$$\begin{aligned} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) &\propto f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta) \\ &\propto \left[\theta^{n\bar{x}}(1-\theta)^{\sum_{i=1}^n (m-x_i)}\right]\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] \\ &\propto \theta^{(\alpha+n\bar{x})-1}(1-\theta)^{(\beta+mn-n\bar{x})-1}. \end{aligned}$$

Comparing the above equation with equation (A.101), we conclude that the posterior pdf belongs to the class of beta distributions. We can further conclude that the hyperparameters of the beta posterior pdf are $\alpha + n\bar{x}$ and $\beta + mn - n\bar{x}$. Note that this is done without computing the expression for the constant of proportionality $K(\boldsymbol{x})$ nor the marginal pdf of $\boldsymbol{X}$. □

8.1.2 Loss function and Bayesian estimation

We now consider the problem of estimating $\mu_X(\Theta) = \mathrm{E}(X \mid \Theta)$ given the observed data $\boldsymbol{x}$. The Bayesian approach of estimation views the estimator as a decision rule, which assigns a value to $\mu_X(\Theta)$ based on the data. Thus, let $w(\boldsymbol{x})$ be an estimator of $\mu_X(\Theta)$. A nonnegative function $L[\mu_X(\Theta), w(\boldsymbol{x})]$, called the **loss function**, is then defined to reflect the penalty in making a wrong decision about $\mu_X(\Theta)$. Typically, the larger the difference between $\mu_X(\Theta)$ and $w(\boldsymbol{x})$, the larger the loss $L[\mu_X(\Theta), w(\boldsymbol{x})]$. A commonly used loss function is the **squared-error loss function** (or **quadratic loss function**) defined by

$$L[\mu_X(\Theta), w(\boldsymbol{x})] = [\mu_X(\Theta) - w(\boldsymbol{x})]^2. \tag{8.7}$$

Other popularly used loss functions include the **absolute-error loss function** and the **zero-one loss function**.[3] We assume, however, that the squared-error loss function is adopted in Bayesian inference.

Given the decision rule and the data, the expected loss in the estimation of $\mu_X(\Theta)$ is

$$\mathrm{E}\{L[\mu_X(\Theta), w(\boldsymbol{x})] \mid \boldsymbol{x}\} = \int_{\theta \in \Omega_\Theta} L[\mu_X(\Theta), w(\boldsymbol{x})] f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, d\theta. \tag{8.8}$$

It is naturally desirable to have a decision rule that gives as small an expected loss as possible. Thus, for any given $\boldsymbol{x}$, if the decision rule $w(\boldsymbol{x})$ assigns a value to $\mu_X(\Theta)$ that minimizes the expected loss, then the decision rule $w(\boldsymbol{x})$ is called the **Bayes estimator** of $\mu_X(\Theta)$ with respect to the chosen loss function. In other words, the Bayes estimator, denoted by $w^*(\boldsymbol{x})$, satisfies

$$\mathrm{E}\{L[\mu_X(\Theta), w^*(\boldsymbol{x})] \mid \boldsymbol{x}\} = \min_{w(\cdot)} \mathrm{E}\{L[\mu_X(\Theta), w(\boldsymbol{x})] \mid \boldsymbol{x}\}, \tag{8.9}$$

for any given $\boldsymbol{x}$. For the squared-error loss function, the decision rule (estimator) that minimizes the expected loss $\mathrm{E}\{[\mu_X(\Theta) - w(\boldsymbol{x})]^2 \mid \boldsymbol{x}\}$ is[4]

$$w^*(\boldsymbol{x}) = \mathrm{E}[\mu_X(\Theta) \mid \boldsymbol{x}]. \tag{8.10}$$

[3] For estimating θ with the estimator $\hat{\theta}$, the absolute-error loss function is defined by $L[\theta, \hat{\theta}] = |\theta - \hat{\theta}|$ and the zero-one loss function is defined by $L[\theta, \hat{\theta}] = 0$ if $\hat{\theta} = \theta$ and 1 otherwise.

[4] See DeGroot and Schervish (2002, p. 348) for a proof of this result.

Thus, for the squared-error loss function, the Bayes estimator of $\mu_X(\Theta)$ is the posterior mean, denoted by $\hat{\mu}_X(\boldsymbol{x})$, so that[5]

$$\hat{\mu}_X(\boldsymbol{x}) = \mathrm{E}[\mu_X(\Theta) \mid \boldsymbol{x}] = \int_{\theta \in \Omega_\Theta} \mu_X(\theta) f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, d\theta. \tag{8.11}$$

In the credibility literature (where X is a loss random variable), $\hat{\mu}_X(x)$ is called the **Bayesian premium**.

An alternative way to interpret the Bayesian premium is to consider the prediction of the loss in the next period, namely, X_{n+1}, given the data $\boldsymbol{x}$. To this effect, we first calculate the conditional pdf of X_{n+1} given $\boldsymbol{x}$, which is

$$\begin{aligned} f_{X_{n+1} \mid \boldsymbol{X}}(x_{n+1} \mid \boldsymbol{x}) &= \frac{f_{X_{n+1}\boldsymbol{X}}(x_{n+1}, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{\int_{\theta \in \Omega_\Theta} f_{X_{n+1}\boldsymbol{X} \mid \Theta}(x_{n+1}, \boldsymbol{x} \mid \theta) f_\Theta(\theta)\, d\theta}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{\int_{\theta \in \Omega_\Theta} \left[\prod_{i=1}^{n+1} f_{X_i \mid \Theta}(x_i \mid \theta)\right] f_\Theta(\theta)\, d\theta}{f_{\boldsymbol{X}}(\boldsymbol{x})}. \end{aligned} \tag{8.12}$$

As the posterior pdf of Θ given $\boldsymbol{X}$ is

$$f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) = \frac{f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{\left[\prod_{i=1}^{n} f_{X_i \mid \Theta}(x_i \mid \theta)\right] f_\Theta(\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})}, \tag{8.13}$$

we conclude

$$\left[\prod_{i=1}^{n} f_{X_i \mid \Theta}(x_i \mid \theta)\right] f_\Theta(\theta) = f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}). \tag{8.14}$$

Substituting (8.14) into (8.12), we obtain

$$f_{X_{n+1} \mid \boldsymbol{X}}(x_{n+1} \mid \boldsymbol{x}) = \int_{\theta \in \Omega_\Theta} f_{X_{n+1} \mid \boldsymbol{\Theta}}(x_{n+1} \mid \theta) f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, d\theta. \tag{8.15}$$

Equation (8.15) shows that the conditional pdf of X_{n+1} given $\boldsymbol{X}$ can be interpreted as a mixture of the conditional pdf of X_{n+1}, where the mixing density is the posterior pdf of Θ.

[5] The Bayes estimator based on the absolute-error loss function is the posterior median, and the Bayes estimator based on the zero-one loss function is the posterior mode. See DeGroot and Schervish (2002, p. 349) for more discussions.

We now consider the prediction of X_{n+1} given $\boldsymbol{X}$. A natural predictor is the conditional expected value of X_{n+1} given $\boldsymbol{X}$, i.e. $\mathrm{E}(X_{n+1} \mid \boldsymbol{x})$, which is given by

$$\mathrm{E}(X_{n+1} \mid \boldsymbol{x}) = \int_0^\infty x_{n+1} f_{X_{n+1} \mid \boldsymbol{X}}(x_{n+1} \mid \boldsymbol{x})\, dx_{n+1}. \tag{8.16}$$

Using equation (8.15), we have

$$\begin{aligned}
\mathrm{E}(X_{n+1} \mid \boldsymbol{x}) &= \int_0^\infty x_{n+1} \left[\int_{\theta \in \Omega_\Theta} f_{X_{n+1} \mid \boldsymbol{\Theta}}(x_{n+1} \mid \theta) f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, d\theta \right] dx_{n+1} \\
&= \int_{\theta \in \Omega_\Theta} \left[\int_0^\infty x_{n+1} f_{X_{n+1} \mid \boldsymbol{\Theta}}(x_{n+1} \mid \theta)\, dx_{n+1} \right] f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, d\theta \\
&= \int_{\theta \in \Omega_\Theta} \mathrm{E}(X_{n+1} \mid \theta) f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, d\theta \\
&= \int_{\theta \in \Omega_\Theta} \mu_X(\theta) f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, d\theta \\
&= \mathrm{E}[\mu_X(\Theta) \mid \boldsymbol{x}].
\end{aligned} \tag{8.17}$$

Thus, the Bayesian premium can also be interpreted as the conditional expectation of X_{n+1} given $\boldsymbol{X}$.

In summary, the Bayes estimate of the mean of the random loss X, called the Bayesian premium, is the posterior mean of X conditional on the data $\boldsymbol{x}$, as given in equation (8.11). It is also equal to the conditional expectation of future loss given the data $\boldsymbol{x}$, as shown in equation (8.17). Thus, we shall use the terminologies Bayesian estimate of expected loss and Bayesian predictor of future loss interchangeably.

8.1.3 Some examples of Bayesian credibility

We now re-visit Examples 8.2 and 8.3 to illustrate the calculation of the Bayesian estimate of the expected loss.

Example 8.4 Let X be the Bernoulli random variable which takes value 1 with probability θ and 0 with probability $1-\theta$, and $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample of X. If $\Theta \sim \mathcal{B}(\alpha, \beta)$, calculate the posterior mean of $\mu_X(\Theta)$ and the expected value of a future observation X_{n+1} given the sample data.

Solution From Example 8.2, we know that the posterior distribution of Θ given $\boldsymbol{x}$ is beta with parameters $\alpha^* = \alpha + n\bar{x}$ and $\beta^* = \beta + n - n\bar{x}$. As X is a Bernoulli random variable, $\mu_X(\Theta) = \mathrm{E}(X \mid \Theta) = \Theta$. Hence, the posterior mean of $\mu_X(\Theta)$

is $\mathrm{E}(\Theta \,|\, \boldsymbol{x}) = \alpha^*/(\alpha^* + \beta^*)$. Now the conditional pdf of X_{n+1} given Θ is

$$f_{X_{n+1}\,|\,\Theta}(x_{n+1} \,|\, \theta) = \begin{cases} \theta, & \text{for } x_{n+1} = 1, \\ 1 - \theta, & \text{for } x_{n+1} = 0. \end{cases}$$

From equation (8.15), the conditional pdf of X_{n+1} given $\boldsymbol{x}$ is

$$f_{X_{n+1}\,|\,\boldsymbol{X}}(x_{n+1} \,|\, \boldsymbol{x}) = \begin{cases} \mathrm{E}(\Theta \,|\, \boldsymbol{x}) = \dfrac{\alpha^*}{\alpha^* + \beta^*}, & \text{for } x_{n+1} = 1, \\ 1 - \mathrm{E}(\Theta \,|\, \boldsymbol{x}) = 1 - \dfrac{\alpha^*}{\alpha^* + \beta^*}, & \text{for } x_{n+1} = 0. \end{cases}$$

Now we apply the above results to equation (8.16) to obtain the conditional mean of X_{n+1} given $\boldsymbol{x}$ as[6]

$$\begin{aligned} \mathrm{E}(X_{n+1} \,|\, \boldsymbol{x}) &= (1)\left[f_{X_{n+1}\,|\,\boldsymbol{X}}(1 \,|\, \boldsymbol{x})\right] + (0)\left[f_{X_{n+1}\,|\,\boldsymbol{X}}(0 \,|\, \boldsymbol{x})\right] \\ &= \frac{\alpha^*}{\alpha^* + \beta^*}, \end{aligned}$$

which is equal to the posterior mean of Θ, $\mathrm{E}(\Theta \,|\, \boldsymbol{x})$. □

Example 8.5 Let $X \sim \mathcal{BN}(2, \theta)$, and $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample of X. If $\Theta \sim \mathcal{B}(\alpha, \beta)$, calculate the posterior mean of $\mu_X(\Theta)$ and the expected value of a future observation X_{n+1} given the sample data.

Solution From Example 8.3, we know that the posterior distribution of Θ is beta with parameters $\alpha^* = \alpha + n\bar{x}$ and $\beta^* = \beta + 2n - n\bar{x}$. As X is a binomial random variable, $\mu_X(\Theta) = \mathrm{E}(X \,|\, \Theta) = 2\Theta$. Hence, the posterior mean of $\mu_X(\Theta)$ is $\mathrm{E}(2\Theta \,|\, \boldsymbol{x}) = 2\alpha^*/(\alpha^* + \beta^*)$. Now the conditional pdf of X_{n+1} given Θ is

$$f_{X_{n+1}\,|\,\Theta}(x_{n+1} \,|\, \theta) = \binom{2}{x_{n+1}} \theta^{x_{n+1}} (1 - \theta)^{2 - x_{n+1}}, \qquad x_{n+1} \in \{0, 1, 2\}.$$

From equation (8.15), the conditional pdf of X_{n+1} given $\boldsymbol{x}$ is

$$f_{X_{n+1}\,|\,\boldsymbol{X}}(x_{n+1} \,|\, \boldsymbol{x}) = \begin{cases} \mathrm{E}[(1 - \Theta)^2 \,|\, \boldsymbol{x}], & \text{for } x_{n+1} = 0, \\ 2\mathrm{E}[\Theta(1 - \Theta) \,|\, \boldsymbol{x}], & \text{for } x_{n+1} = 1, \\ \mathrm{E}[\Theta^2 \,|\, \boldsymbol{x}], & \text{for } x_{n+1} = 2. \end{cases}$$

[6] Note that X_{n+1} is a discrete random variable and we have to replace the integration in equation (8.16) by summation.

Now we apply the above results to equation (8.16) to obtain the conditional mean of X_{n+1} given $\boldsymbol{x}$ as

$$\begin{aligned}
\mathrm{E}(X_{n+1} \mid \boldsymbol{x}) &= (1)\left[f_{X_{n+1} \mid \boldsymbol{X}}(1 \mid \boldsymbol{x})\right] + (2)\left[f_{X_{n+1} \mid \boldsymbol{X}}(2 \mid \boldsymbol{x})\right] \\
&= 2\mathrm{E}[\Theta(1-\Theta) \mid \boldsymbol{x}] + 2\mathrm{E}[\Theta^2 \mid \boldsymbol{x}] \\
&= 2\mathrm{E}[\Theta \mid \boldsymbol{x}] \\
&= \frac{2\alpha^*}{\alpha^* + \beta^*},
\end{aligned}$$

which is equal to the posterior mean of $\mu_X(\Theta)$. □

Examples 8.4 and 8.5 illustrate the equivalence of equations (8.11) and (8.16). The results can be generalized to the case when X is a binomial random variable with parameters m and θ, where m is any positive integer. Readers may wish to prove this result as an exercise (see Exercise 8.1).

Example 8.6 X is the claim-severity random variable that can take values 10, 20, or 30. The distribution of X depends on the risk group defined by parameter Θ, which are labeled 1, 2, and 3. The relative frequencies of risk groups with Θ equal to 1, 2, and 3 are, respectively, 0.4, 0.4, and 0.2. The conditional distribution of X given the risk parameter Θ is given in Table 8.1.

Table 8.1. *Data for Example 8.6*

		$\Pr(X = x \mid \theta)$		
θ	$\Pr(\Theta = \theta)$	$x = 10$	$x = 20$	$x = 30$
1	0.4	0.2	0.3	0.5
2	0.4	0.4	0.4	0.2
3	0.2	0.5	0.5	0.0

A sample of three claims with $\boldsymbol{x} = (20, 20, 30)$ is observed. Calculate the posterior mean of X. Compute the conditional pdf of X_4 given $\boldsymbol{x}$, and calculate the expected value of X_4 given $\boldsymbol{x}$.

Solution We first calculate the conditional probability of $\boldsymbol{x}$ given Θ as follows

$$\begin{aligned}
f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid 1) &= (0.3)(0.3)(0.5) = 0.045, \\
f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid 2) &= (0.4)(0.4)(0.2) = 0.032,
\end{aligned}$$

and

$$f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid 3) = (0.5)(0.5)(0) = 0.$$

Thus, the joint pf of $\boldsymbol{x}$ and Θ is

$$f_{\Theta \boldsymbol{X}}(1, \boldsymbol{x}) = f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid 1) f_{\Theta}(1) = (0.045)(0.4) = 0.018,$$
$$f_{\Theta \boldsymbol{X}}(2, \boldsymbol{x}) = f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid 2) f_{\Theta}(2) = (0.032)(0.4) = 0.0128,$$

and

$$f_{\Theta \boldsymbol{X}}(3, \boldsymbol{x}) = f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid 3) f_{\Theta}(3) = 0(0.2) = 0.$$

Thus, we obtain

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = 0.018 + 0.0128 + 0 = 0.0308,$$

so that the posterior distribution of Θ is

$$f_{\Theta \mid \boldsymbol{X}}(1 \mid \boldsymbol{x}) = \frac{f_{\Theta \boldsymbol{X}}(1, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{0.018}{0.0308} = 0.5844,$$
$$f_{\Theta \mid \boldsymbol{X}}(2 \mid \boldsymbol{x}) = \frac{f_{\Theta \boldsymbol{X}}(2, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{0.0128}{0.0308} = 0.4156,$$

and $f_{\Theta \mid \boldsymbol{X}}(3 \mid \boldsymbol{x}) = 0$. The conditional means of X are

$$\mathrm{E}(X \mid \Theta = 1) = (10)(0.2) + (20)(0.3) + (30)(0.5) = 23,$$
$$\mathrm{E}(X \mid \Theta = 2) = (10)(0.4) + (20)(0.4) + (30)(0.2) = 18,$$

and

$$\mathrm{E}(X \mid \Theta = 3) = (10)(0.5) + (20)(0.5) + (30)(0) = 15.$$

Thus, the posterior mean of X is

$$\begin{aligned} \mathrm{E}[\mathrm{E}(X \mid \Theta) \mid \boldsymbol{x}] &= \sum_{\theta=1}^{3} [\mathrm{E}(X \mid \theta)] f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) \\ &= (23)(0.5844) + (18)(0.4156) + (15)(0) = 20.92. \end{aligned}$$

Now we compute the conditional distribution of X_4 given $\boldsymbol{x}$. We note that

$$f_{X_4}(x_4 \mid \boldsymbol{x}) = \sum_{\theta=1}^{3} f_{X_4 \mid \Theta}(x_4 \mid \theta) f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}).$$

As $f_{\Theta \mid \boldsymbol{X}}(3 \mid \boldsymbol{x}) = 0$, we have

$$f_{X_4}(10 \mid \boldsymbol{x}) = (0.2)(0.5844) + (0.4)(0.4156) = 0.2831,$$
$$f_{X_4}(20 \mid \boldsymbol{x}) = (0.3)(0.5844) + (0.4)(0.4156) = 0.3416,$$

and

$$f_{X_4}(30 \mid \boldsymbol{x}) = (0.5)(0.5844) + (0.2)(0.4156) = 0.3753.$$

Thus, the conditional mean of X_4 given $\boldsymbol{x}$ is

$$\mathrm{E}(X_4 \mid \boldsymbol{x}) = (10)(0.2831) + (20)(0.3416) + (30)(0.3753) = 20.92,$$

and the result

$$\mathrm{E}[\mu_X(\Theta) \mid \boldsymbol{x}] = \mathrm{E}(X_4 \mid \boldsymbol{x})$$

is verified. □

8.2 Conjugate distributions

A difficulty in applying the Bayes approach of statistical inference is the computation of the posterior pdf, which requires the computation of the marginal pdf of the data. However, as the Bayes estimate under squared-error loss is the mean of the posterior distribution, the estimate cannot be calculated unless the posterior pdf is known.

It turns out that there are classes of prior pdfs, which, together with specific likelihood functions, give rise to posterior pdfs that belong to the same class as the prior pdf. Such prior pdf and likelihood are said to be a **conjugate** pair. In Example 8.1, we see that if the prior pdf is beta and the likelihood is Bernoulli, the posterior pdf also follows a beta distribution, albeit with hyperparameters different from those of the prior pdf. In Example 8.3, we see that if the prior pdf is beta and the likelihood is binomial, then the posterior pdf is also beta, though with hyperparameters different from those of the prior. Thus, in these cases, the observed data $\boldsymbol{x}$ do not change the class of the prior, they only change the parameters of the prior.

A formal definition of **conjugate prior distribution** is as follows. Let the prior pdf of Θ be $f_\Theta(\theta \mid \gamma)$, where γ is the hyperparameter. The prior pdf $f_\Theta(\theta \mid \gamma)$ is conjugate to the likelihood function $f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta)$ if the posterior pdf is equal to $f_\Theta(\theta \mid \gamma^*)$, which has the same functional form as the prior pdf but, generally, a different hyperparameter γ^*. In other words, the prior and posterior belong to the same family of distributions.

We adopt the convention of "prior–likelihood" to describe the conjugate distribution. Thus, as shown in Examples 8.1 and 8.3, beta–Bernoulli and beta–binomial are conjugate distributions. We now present further examples of conjugate distributions which may be relevant for analyzing random losses. More conjugate distributions can be found in Table A.3 in the Appendix.

8.2.1 The gamma–Poisson conjugate distribution

Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ be iid $\mathcal{PN}(\lambda)$. We assume $\Lambda \sim \mathcal{G}(\alpha, \beta)$. As shown in Appendix A.16.3, the posterior distribution of Λ is $\mathcal{G}(\alpha^*, \beta^*)$, where

$$\alpha^* = \alpha + n\bar{x} \tag{8.18}$$

and

$$\beta^* = \left[n + \frac{1}{\beta}\right]^{-1} = \frac{\beta}{n\beta + 1}. \tag{8.19}$$

Hence, the gamma prior pdf is conjugate to the Poisson likelihood.

8.2.2 The beta–geometric conjugate distribution

Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ be iid $\mathcal{GM}(\theta)$. If the prior pdf of Θ is $\mathcal{B}(\alpha, \beta)$, then, as shown in Appendix A.16.4, the posterior distribution of Θ is $\mathcal{B}(\alpha^*, \beta^*)$, with

$$\alpha^* = \alpha + n \tag{8.20}$$

and

$$\beta^* = \beta + n\bar{x}, \tag{8.21}$$

so that the beta prior is conjugate to the geometric likelihood.

8.2.3 The gamma–exponential conjugate distribution

Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ be iid $\mathcal{E}(\lambda)$. If the prior distribution of Λ is $\mathcal{G}(\alpha, \beta)$, then, as shown in Appendix A.16.5, the posterior distribution of Λ is $\mathcal{G}(\alpha^*, \beta^*)$, with

$$\alpha^* = \alpha + n \tag{8.22}$$

and

$$\beta^* = \left[\frac{1}{\beta} + n\bar{x}\right]^{-1} = \frac{\beta}{1 + \beta n\bar{x}}. \tag{8.23}$$

Thus, the gamma prior is conjugate to the exponential likelihood.

8.3 Bayesian versus Bühlmann credibility

If the prior distribution is conjugate to the likelihood, the Bayes estimate is easy to obtain. It turns out that for the conjugate distributions discussed in the last

section, the Bühlmann credibility estimate is equal to the Bayes estimate. The examples below give the details of these results.

Example 8.7 (gamma–Poisson case) The claim-frequency random variable X is assumed to be distributed as $\mathcal{PN}(\lambda)$, and the prior distribution of Λ is $\mathcal{G}(\alpha, \beta)$. If a random sample of n observations of $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ is available, derive the Bühlmann credibility estimate of the future claim frequency, and show that this is the same as the Bayes estimate.

Solution As $X_i \sim$ iid $\mathcal{P}(\lambda)$, we have

$$\mu_{\text{PV}} = \text{E}[\sigma_X^2(\Lambda)] = \text{E}(\Lambda).$$

Since $\Lambda \sim \mathcal{G}(\alpha, \beta)$, we conclude that $\mu_{\text{PV}} = \alpha\beta$. Also, $\mu_X(\Lambda) = \text{E}(X \mid \Lambda) = \Lambda$, so that

$$\sigma_{\text{HM}}^2 = \text{Var}[\mu_X(\Lambda)] = \text{Var}(\Lambda) = \alpha\beta^2.$$

Thus

$$k = \frac{\mu_{\text{PV}}}{\sigma_{\text{HM}}^2} = \frac{1}{\beta},$$

and the Bühlmann credibility factor is

$$Z = \frac{n}{n+k} = \frac{n\beta}{n\beta+1}.$$

The prior mean of the claim frequency is

$$M = \text{E}[\text{E}(X \mid \Lambda)] = \text{E}(\Lambda) = \alpha\beta.$$

Hence, we obtain the Bühlmann credibility estimate of future claim frequency as

$$\begin{aligned} U &= Z\bar{X} + (1-Z)M \\ &= \frac{n\beta\bar{X}}{n\beta+1} + \frac{\alpha\beta}{n\beta+1} \\ &= \frac{\beta(n\bar{X}+\alpha)}{n\beta+1}. \end{aligned}$$

The Bayes estimate of the expected claim frequency is the posterior mean of Λ. From Section 8.2.1, the posterior distribution of Λ is $\mathcal{G}(\alpha^*, \beta^*)$, where α^* and β^* are given in equations (8.18) and (8.19), respectively. Thus, the Bayes estimate of the expected claim frequency is

$$\begin{aligned}
\mathrm{E}(X_{n+1} \,|\, \boldsymbol{x}) &= \mathrm{E}[\mathrm{E}(X_{n+1} \,|\, \Lambda) \,|\, \boldsymbol{x}] \\
&= \mathrm{E}(\Lambda \,|\, \boldsymbol{x}) \\
&= \alpha^* \beta^* \\
&= (\alpha + n\bar{X}) \left[\frac{\beta}{n\beta + 1} \right] \\
&= U,
\end{aligned}$$

which is the Bühlmann credibility estimate. □

Example 8.8 (beta–geometric case) The claim-frequency random variable X is assumed to be distributed as $\mathcal{GM}(\theta)$, and the prior distribution of Θ is $\mathcal{B}(\alpha, \beta)$, where $\alpha > 2$. If a random sample of n observations of $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ is available, derive the Bühlmann credibility estimate of the future claim frequency, and show that this is the same as the Bayes estimate.

Solution As $X_i \sim$ iid $\mathcal{G}(\theta)$, we have

$$\mu_X(\Theta) = \mathrm{E}(X \mid \Theta) = \frac{1 - \Theta}{\Theta},$$

and

$$\sigma_X^2(\Theta) = \mathrm{Var}(X \mid \Theta) = \frac{1 - \Theta}{\Theta^2}.$$

Assuming $\Theta \sim \mathcal{B}(\alpha, \beta)$, we first compute the following moments

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{\Theta}\right) &= \int_0^1 \frac{1}{\theta} \left[\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \right] d\theta \\
&= \frac{B(\alpha - 1, \beta)}{B(\alpha, \beta)} \\
&= \frac{\alpha + \beta - 1}{\alpha - 1},
\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}\left(\frac{1}{\Theta^2}\right) &= \int_0^1 \frac{1}{\theta^2}\left[\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\right] d\theta \\ &= \frac{B(\alpha-2,\beta)}{B(\alpha,\beta)} \\ &= \frac{(\alpha+\beta-1)(\alpha+\beta-2)}{(\alpha-1)(\alpha-2)}.\end{aligned}$$

Hence, the expected value of the process variance is

$$\begin{aligned}\mu_{\mathrm{PV}} &= \mathrm{E}[\sigma_X^2(\Theta)] \\ &= \mathrm{E}\left(\frac{1-\Theta}{\Theta^2}\right) \\ &= \mathrm{E}\left(\frac{1}{\Theta^2}\right) - \mathrm{E}\left(\frac{1}{\Theta}\right) \\ &= \frac{(\alpha+\beta-1)(\alpha+\beta-2)}{(\alpha-1)(\alpha-2)} - \frac{\alpha+\beta-1}{\alpha-1} \\ &= \frac{(\alpha+\beta-1)\beta}{(\alpha-1)(\alpha-2)},\end{aligned}$$

and the variance of the hypothetical means is

$$\begin{aligned}\sigma_{\mathrm{HM}}^2 &= \mathrm{Var}[\mu_X(\Theta)] \\ &= \mathrm{Var}\left(\frac{1-\Theta}{\Theta}\right) \\ &= \mathrm{Var}\left(\frac{1}{\Theta}\right) \\ &= \mathrm{E}\left(\frac{1}{\Theta^2}\right) - \left[\mathrm{E}\left(\frac{1}{\Theta}\right)\right]^2 \\ &= \frac{(\alpha+\beta-1)(\alpha+\beta-2)}{(\alpha-1)(\alpha-2)} - \left(\frac{\alpha+\beta-1}{\alpha-1}\right)^2 \\ &= \frac{(\alpha+\beta-1)\beta}{(\alpha-1)^2(\alpha-2)}.\end{aligned}$$

Thus, the ratio of μ_{PV} to σ_{HM}^2 is

$$k = \frac{\mu_{\mathrm{PV}}}{\sigma_{\mathrm{HM}}^2} = \alpha - 1,$$

and the Bühlmann credibility factor is

$$Z = \frac{n}{n+k} = \frac{n}{n+\alpha-1}.$$

As the prior mean of X is

$$M = \mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X \mid \Theta)] = \mathrm{E}\left(\frac{1-\Theta}{\Theta}\right) = \frac{\alpha+\beta-1}{\alpha-1} - 1 = \frac{\beta}{\alpha-1},$$

the Bühlmann credibility prediction of future claim frequency is

$$\begin{aligned} U &= Z\bar{X} + (1-Z)M \\ &= \frac{n\bar{X}}{n+\alpha-1} + \frac{\alpha-1}{n+\alpha-1}\left(\frac{\beta}{\alpha-1}\right) \\ &= \frac{n\bar{X}+\beta}{n+\alpha-1}. \end{aligned}$$

To compute the Bayes estimate of future claim frequency we note, from Section 8.2.2, that the posterior distribution of Θ is $\mathcal{B}(\alpha^*, \beta^*)$, where α^* and β^* are given in equations (8.20) and (8.21), respectively. Thus, we have

$$\begin{aligned} \mathrm{E}(X_{n+1} \mid \boldsymbol{x}) &= \mathrm{E}[\mathrm{E}(X_{n+1} \mid \Theta) \mid \boldsymbol{x}] \\ &= \mathrm{E}\left(\frac{1-\Theta}{\Theta} \mid \boldsymbol{x}\right) \\ &= \frac{\alpha^*+\beta^*-1}{\alpha^*-1} - 1 \\ &= \frac{\beta^*}{\alpha^*-1} \\ &= \frac{n\bar{X}+\beta}{n+\alpha-1}, \end{aligned}$$

which is the same as the Bühlmann credibility estimate. □

Example 8.9 (gamma–exponential case) The claim-severity random variable X is assumed to be distributed as $\mathcal{E}(\lambda)$, and the prior distribution of Λ is $\mathcal{G}(\alpha, \beta)$, where $\alpha > 2$. If a random sample of n observations of $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ is available, derive the Bühlmann credibility estimate of the future claim severity, and show that this is the same as the Bayes estimate.

Solution As $X_i \sim$ iid $\mathcal{E}(\lambda)$, we have

$$\mu_X(\Lambda) = \mathrm{E}(X \mid \Lambda) = \frac{1}{\Lambda},$$

and

$$\sigma_X^2(\Lambda) = \text{Var}(X \mid \Lambda) = \frac{1}{\Lambda^2}.$$

Since $\Lambda \sim \mathcal{G}(\alpha, \beta)$, the expected value of the process variance is

$$\begin{aligned}
\mu_{\text{PV}} &= \text{E}[\sigma_X^2(\Lambda)] \\
&= \text{E}\left(\frac{1}{\Lambda^2}\right) \\
&= \int_0^\infty \frac{1}{\lambda^2}\left[\frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)\beta^\alpha}\right] d\lambda \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty \lambda^{\alpha-3}e^{-\frac{\lambda}{\beta}}\, d\lambda \\
&= \frac{\Gamma(\alpha-2)\beta^{\alpha-2}}{\Gamma(\alpha)\beta^\alpha} \\
&= \frac{1}{(\alpha-1)(\alpha-2)\beta^2}.
\end{aligned}$$

The variance of the hypothetical means is

$$\sigma_{\text{HM}}^2 = \text{Var}[\mu_X(\Lambda)] = \text{Var}\left(\frac{1}{\Lambda}\right) = \text{E}\left(\frac{1}{\Lambda^2}\right) - \left[\text{E}\left(\frac{1}{\Lambda}\right)\right]^2.$$

Now

$$\begin{aligned}
\text{E}\left(\frac{1}{\Lambda}\right) &= \int_0^\infty \frac{1}{\lambda}\left[\frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)\beta^\alpha}\right] d\lambda \\
&= \frac{\Gamma(\alpha-1)\beta^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \\
&= \frac{1}{(\alpha-1)\beta},
\end{aligned}$$

so that

$$\begin{aligned}
\sigma_{\text{HM}}^2 &= \frac{1}{(\alpha-1)(\alpha-2)\beta^2} - \left[\frac{1}{(\alpha-1)\beta}\right]^2 \\
&= \frac{1}{(\alpha-1)^2(\alpha-2)\beta^2}.
\end{aligned}$$

Thus, we have

$$k = \frac{\mu_{\text{PV}}}{\sigma^2_{\text{HM}}} = \alpha - 1,$$

and the Bühlmann credibility factor is

$$Z = \frac{n}{n+k} = \frac{n}{n+\alpha-1}.$$

The prior mean of X is

$$M = \mathrm{E}\left[\mathrm{E}(X \mid \Lambda)\right] = \mathrm{E}\left(\frac{1}{\Lambda}\right) = \frac{1}{(\alpha-1)\beta}.$$

Hence we obtain the Bühlmann credibility estimate as

$$\begin{aligned} U &= Z\bar{X} + (1-Z)M \\ &= \frac{n\bar{X}}{n+\alpha-1} + \frac{\alpha-1}{n+\alpha-1}\left[\frac{1}{(\alpha-1)\beta}\right] \\ &= \frac{\beta n\bar{X}+1}{(n+\alpha-1)\beta}. \end{aligned}$$

To calculate the Bayes estimate, we note, from Section 8.2.3, that the posterior pdf of Λ is $\mathcal{G}(\alpha^*, \beta^*)$, where α^* and β^* are given in equations (8.22) and (8.23), respectively. Thus, the Bayes estimate of the expected claim severity is

$$\begin{aligned} \mathrm{E}(X_{n+1} \mid \boldsymbol{x}) &= \mathrm{E}\left(\frac{1}{\Lambda} \mid \boldsymbol{x}\right) \\ &= \frac{1}{(\alpha^*-1)\beta^*} \\ &= \frac{1+\beta n\bar{X}}{(\alpha+n-1)\beta} \\ &= U, \end{aligned}$$

and the equality of the Bühlmann estimate and the Bayes estimate is proven. □

We have shown that if the conjugate distributions discussed in the last section are used to model loss variables, where the distribution of the loss variable follows the likelihood function and the distribution of the risk parameters follows the conjugate prior, then the Bühlmann credibility estimate of the expected loss is equal to the Bayes estimate. In such cases, the Bühlmann

credibility estimate is said to have **exact credibility**. Indeed, there are other conjugate distributions for which the Bühlmann credibility is *exact*. For example, the Bühlmann estimate for the case of normal–normal conjugate has exact credibility. In the next section, we discuss a general result for which the Bühlmann credibility estimate is exact.

8.4 Linear exponential family and exact credibility

Consider a random variable X with pdf or pf $f_{X\,|\,\Theta}(x\,|\,\theta)$, where θ is the parameter of the distribution. X is said to have a **linear exponential distribution** if $f_{X\,|\,\Theta}(x\,|\,\theta)$ can be written as

$$f_{X\,|\,\Theta}(x\,|\,\theta) = \exp\left[A(\theta)x + B(\theta) + C(x)\right], \tag{8.24}$$

for some functions $A(\theta)$, $B(\theta)$, and $C(x)$. By identifying these respective functions it is easy to show that some of the commonly used distributions in the actuarial science literature belong to the linear exponential family. Table 8.2 summarizes some of these distributions.[7]

Table 8.2. *Some linear exponential distributions*

Distribution	$\log f_{X\,\|\,\Theta}(x\,\|\,\theta)$	$A(\theta)$	$B(\theta)$	$C(x)$
Binomial, $\mathcal{BN}(m,\theta)$	$\log(C_x^m) + x\log\theta$ $+(m-x)\log(1-\theta)$	$\log\theta - \log(1-\theta)$	$m\log(1-\theta)$	$\log(C_x^m)$
Geometric, $\mathcal{GM}(\theta)$	$\log\theta + x\log(1-\theta)$	$\log(1-\theta)$	$\log\theta$	0
Poisson, $\mathcal{PN}(\theta)$	$x\log\theta - \theta - \log(x!)$	$\log\theta$	$-\theta$	$-\log(x!)$
Exponential, $\mathcal{E}(\theta)$	$-\theta x + \log\theta$	$-\theta$	$\log\theta$	0

If the likelihood function belongs to the linear exponential family, we can identify the prior distribution that is conjugate to the likelihood. Suppose the distribution of Θ has two hyperparameters, denoted by α and β. The **natural conjugate** of the likelihood given in equation (8.24) is

$$f_{\Theta}(\theta\,|\,\alpha,\beta) = \exp\left[A(\theta)a(\alpha,\beta) + B(\theta)b(\alpha,\beta) + D(\alpha,\beta)\right], \tag{8.25}$$

for some functions $a(\alpha,\beta)$, $b(\alpha,\beta)$, and $D(\alpha,\beta)$. To see this, we combine equations (8.24) and (8.25) to obtain the posterior pdf of Θ conditional on the

[7] For convenience we use θ to denote generically the parameters of the distributions instead of the usual notations (e.g. θ replaces λ for the parameter of the Poisson and exponential distributions). C_x^m in Table 8.2 denotes the combinatorial function.

sample $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ as

$$f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) \propto \exp\left\{A(\theta)\left[a(\alpha,\beta)+\sum_{i=1}^{n} x_i\right] + B(\theta)\left[b(\alpha,\beta)+n\right]\right\}. \quad (8.26)$$

Hence, the posterior pdf belongs to the same family as the prior pdf with parameters α^* and β^*, assuming they can be solved uniquely from the following equations

$$a(\alpha^*, \beta^*) = a(\alpha, \beta) + n\bar{x}, \quad (8.27)$$

and

$$b(\alpha^*, \beta^*) = b(\alpha, \beta) + n. \quad (8.28)$$

For the Poisson, geometric, and exponential likelihoods, we identify the functions $a(\alpha, \beta)$ and $b(\alpha, \beta)$, and summarize them in Table 8.3. The natural conjugate priors are then obtainable using equation (8.25).[8]

Table 8.3. *Examples of natural conjugate priors*

Conjugate distribution	$\log f_\Theta(\theta \mid \alpha, \beta)$	$a(\alpha, \beta)$	$b(\alpha, \beta)$
gamma–Poisson	$(\alpha - 1)\log\theta - \dfrac{\theta}{\beta} - \log[\Gamma(\alpha)\beta^\alpha]$	$\alpha - 1$	$\dfrac{1}{\beta}$
beta–geometric	$(\alpha - 1)\log\theta + (\beta - 1)\log(1-\theta) - \log B(\alpha, \beta)$	$\beta - 1$	$\alpha - 1$
gamma–exponential	$(\alpha - 1)\log\theta - \theta\beta - \log[\Gamma(\alpha)\beta^{-\alpha}]$	$\dfrac{1}{\beta}$	$\alpha - 1$

The results above enable us to compute the prior pdf (up to a scaling factor) using equation (8.25). To illustrate the calculation of the posterior pdf, we consider the gamma–Poisson conjugate distribution. As $a(\alpha, \beta) = \alpha - 1$, we have, from equation (8.27)

$$\alpha^* - 1 = \alpha - 1 + n\bar{x}, \quad (8.29)$$

which implies

$$\alpha^* = \alpha + n\bar{x}. \quad (8.30)$$

[8] In Table 8.3, the function $D(\alpha, \beta)$, which defines the scaling factor, is ignored. The case of the binomial likelihood and its natural conjugate is left as an exercise (see Exercise 8.2).

Also, from equation (8.28), we have

$$\frac{1}{\beta^*} = \frac{1}{\beta} + n, \tag{8.31}$$

which implies

$$\beta^* = \left[\frac{1}{\beta} + n\right]^{-1}. \tag{8.32}$$

Thus, the results are the same as in equations (8.18) and (8.19).

Having seen some examples of linear exponential likelihoods and their natural conjugates, we now state a theorem which relates the Bühlmann credibility estimate to the Bayes estimate.

Theorem 8.1 *Let X be a random loss variable. If the likelihood of X belongs to the linear exponential family with parameter θ, and the prior distribution of Θ is the natural conjugate of the likelihood of X, then the Bühlmann credibility estimate of the mean of X is the same as the Bayes estimate.*

Proof See Klugman *et al.* (2004, Section 16.4.6), or Jewell (1974), for a proof of this theorem.[9] □

When the conditions of Theorem 8.1 hold, the Bühlmann credibility estimate is the same as the Bayes estimate and is said to have exact credibility. When the conditions of the theorem do not hold, the Bühlmann credibility estimator generally has a larger mean squared error than the Bayes estimator. The Bühlmann credibility estimator, however, still has the minimum mean squared error in the class of linear estimators based on the sample. The example below illustrates a comparison between the two methods, as well as the sample mean.

Example 8.10 Assume the claim frequency X over different periods are iid as $\mathcal{PN}(\lambda)$, and the prior pf of Λ is

$$\Lambda = \begin{cases} 1, & \text{with probability 0.5,} \\ 2, & \text{with probability 0.5.} \end{cases}$$

A random sample of $n = 6$ observations of X is available. Calculate the Bühlmann credibility estimate and the Bayes estimate of the expected claim frequency. Compare the mean squared errors of these estimates as well as that of the sample mean.

[9] Note that the proofs in these references require a transformation of the parameters of the prior pdf, so that the prior pdf is expressed in a form different from equation (8.24).

Solution The expected claim frequency is $\mathrm{E}(X) = \Lambda$. Thus, the mean squared error of the sample mean as an estimate of the expected claim frequency is

$$\begin{aligned}
\mathrm{E}\left[(\bar{X} - \Lambda)^2\right] &= \mathrm{E}\left\{\mathrm{E}\left[(\bar{X} - \Lambda)^2 \mid \Lambda\right]\right\} \\
&= \mathrm{E}\left\{\left[\mathrm{Var}(\bar{X} \mid \Lambda)\right]\right\} \\
&= \mathrm{E}\left[\frac{\mathrm{Var}(X \mid \Lambda)}{n}\right] \\
&= \frac{\mathrm{E}(\Lambda)}{n} \\
&= \frac{1.5}{6} \\
&= 0.25.
\end{aligned}$$

We now derive the Bühlmann credibility estimator. As $\mu_X(\Lambda) = \mathrm{E}(X \mid \Lambda) = \Lambda$ and $\sigma_X^2(\Lambda) = \mathrm{Var}(X \mid \Lambda) = \Lambda$, we have

$$\mu_{\mathrm{PV}} = \mathrm{E}[\sigma_X^2(\Lambda)] = \mathrm{E}(\Lambda) = 1.5,$$

and

$$\sigma_{\mathrm{HM}}^2 = \mathrm{Var}[\mu_X(\Lambda)] = \mathrm{Var}(\Lambda) = (0.5)(1 - 1.5)^2 + (0.5)(2 - 1.5)^2 = 0.25.$$

Thus, we have

$$k = \frac{\mu_{\mathrm{PV}}}{\sigma_{\mathrm{HM}}^2} = \frac{1.5}{0.25} = 6,$$

and the Bühlmann credibility factor is

$$Z = \frac{n}{n+6} = \frac{6}{6+6} = 0.5.$$

As the prior mean of X is

$$M = \mathrm{E}[\mathrm{E}(X \mid \Lambda)] = \mathrm{E}(\Lambda) = 1.5,$$

the Bühlmann credibility estimator is

$$U = Z\bar{X} + (1 - Z)M = 0.5\bar{X} + (0.5)(1.5) = 0.5\bar{X} + 0.75.$$

Given $\Lambda = \lambda$, the expected values of the sample mean and the Bühlmann credibility estimator are, respectively, λ and $0.5\lambda + 0.75$. Thus, the sample

mean is an unbiased estimator of λ, while the Bühlmann credibility estimator is generally not. However, when λ varies as a random variable the expected value of the Bühlmann credibility estimator is equal to 1.5, which is the prior mean of X, and so is the expected value of the sample mean.

The mean squared error of the Bühlmann credibility estimate of the expected value of X is computed as follows

$$\begin{aligned}
\mathrm{E}\left\{[U - \mathrm{E}(X)]^2\right\} &= \mathrm{E}\left[(0.5\bar{X} + 0.75 - \Lambda)^2\right] \\
&= \mathrm{E}\left\{\mathrm{E}\left[(0.5\bar{X} + 0.75 - \Lambda)^2 \mid \Lambda\right]\right\} \\
&= \mathrm{E}\Big\{\mathrm{E}\Big[0.25\bar{X}^2 + (0.75)^2 + \Lambda^2 + 0.75\bar{X} \\
&\quad - 1.5\Lambda - \Lambda\bar{X} \mid \Lambda\Big]\Big\} \\
&= \mathrm{E}\Big[0.25(\mathrm{Var}(\bar{X} \mid \Lambda) + \left[\mathrm{E}(\bar{X} \mid \Lambda)\right]^2) \\
&\quad + \mathrm{E}\left\{(0.75)^2 + \Lambda^2 + 0.75\bar{X} - 1.5\Lambda - \Lambda\bar{X} \mid \Lambda\right\}\Big] \\
&= \mathrm{E}\left[0.25\left(\frac{\Lambda}{6} + \Lambda^2\right) + (0.75)^2 + \Lambda^2 + 0.75\Lambda\right. \\
&\quad \left. - 1.5\Lambda - \Lambda^2\right] \\
&= \mathrm{E}\left[0.25\left(\frac{\Lambda}{6} + \Lambda^2\right) + (0.75)^2 - 0.75\Lambda\right] \\
&= \mathrm{E}\left(0.25\Lambda^2 - 0.7083\Lambda + 0.5625\right) \\
&= 0.25\left[(1)(0.5) + (2)^2(0.5)\right] - (0.7083)(1.5) + 0.5625 \\
&= 0.1251.
\end{aligned}$$

Hence, the mean squared error of the Bühlmann credibility estimator is about half of that of the sample mean.

As the Bayes estimate is the posterior mean, we first derive the posterior pf of Λ. The marginal pf of $\boldsymbol{X}$ is

$$\begin{aligned}
f_{\boldsymbol{X}}(\boldsymbol{x}) &= \sum_{\lambda \in \{1,2\}} f_{\boldsymbol{X} \mid \Lambda}(\boldsymbol{x} \mid \lambda) \Pr(\Lambda = \lambda) \\
&= 0.5\left[\sum_{\lambda \in \{1,2\}} \left(\prod_{i=1}^{6} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right)\right]
\end{aligned}$$

$$= 0.5\left[\left(\frac{1}{e^6}\prod_{i=1}^{6}\frac{1}{x_i!}\right)+\left(\frac{1}{e^{12}}\prod_{i=1}^{6}\frac{2^{x_i}}{x_i!}\right)\right]$$
$$= K, \qquad \text{say.}$$

Thus, the posterior pdf of Λ is

$$f_{\Lambda\,|\,\boldsymbol{X}}(\lambda\,|\,\boldsymbol{x}) = \begin{cases} \dfrac{0.5}{e^6 K}\left(\displaystyle\prod_{i=1}^{6}\frac{1}{x_i!}\right), & \text{for } \lambda = 1, \\ \dfrac{0.5}{e^{12} K}\left(\displaystyle\prod_{i=1}^{6}\frac{2^{x_i}}{x_i!}\right), & \text{for } \lambda = 2. \end{cases}$$

The posterior mean of Λ is

$$\mathrm{E}(\Lambda\,|\,\boldsymbol{x}) = \frac{0.5}{e^6 K}\left(\prod_{i=1}^{6}\frac{1}{x_i!}\right)+\left(\frac{1}{e^{12} K}\prod_{i=1}^{6}\frac{2^{x_i}}{x_i!}\right).$$

Thus, the Bayes estimate is a highly nonlinear function of the data, and the computation of its mean squared error is intractable. We estimate the mean squared error using simulation as follows:[10]

1 Generate λ with value of 1 or 2 with probability of 0.5 each.
2 Using the value of λ generated in Step 1, generate six observations of X, $x_1, \cdots, x_6$, from the distribution $\mathcal{PN}(\lambda)$.
3 Compute the posterior mean of Λ of this sample using the expression

$$\frac{0.5}{e^6 K}\left(\prod_{i=1}^{6}\frac{1}{x_i!}\right)+\left(\frac{1}{e^{12} K}\prod_{i=1}^{6}\frac{2^{x_i}}{x_i!}\right).$$

4 Repeat Steps 1 through 3 m times. Denote the values of λ generated in Step 1 by $\lambda_1, \cdots, \lambda_m$, and the corresponding Bayes estimates computed in Step 3 by $\hat{\lambda}_1, \cdots, \hat{\lambda}_m$. The estimated mean squared error of the Bayes estimate is

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{\lambda}_i - \lambda_i)^2.$$

We perform a simulation with $m = 100{,}000$ runs. The estimated mean squared error is 0.1103. Thus, the mean squared error of the Bayes estimate is lower than that of the Bühlmann credibility estimate (0.1251), which is in turn lower than that of the sample mean (0.25). □

[10] The methodology of simulation is covered in detail in Chapters 14 and 15.

8.5 Summary and discussions

The prediction of future random losses can be usefully formulated under the Bayesian framework. Suppose the random loss variable X has a mean $\mathrm{E}(X \mid \Theta) = \mu_X(\Theta)$, and a random sample of $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ is available. The Bayesian premium is equal to $\mathrm{E}[\mu_X(\Theta) \mid \boldsymbol{x}]$, which is also equal to $\mathrm{E}[X_{n+1} \mid \boldsymbol{x}]$. The former is the Bayes estimate of the expected loss, and the latter is the Bayes predictor of future loss. The Bayes estimate (prediction) is the posterior mean of the expected loss (posterior mean of the future loss), and it has the minimum mean squared error among all estimators of the expected loss.

The Bühlmann credibility estimate is the minimum mean squared error estimate in the class of estimators that are linear in $\boldsymbol{X}$. When the likelihood belongs to the linear exponential family and the prior distribution is the natural conjugate, the Bühlmann credibility estimate is equal to the Bayes estimate. However, in other situations the performance of the Bühlmann credibility estimate is generally inferior to the Bayes estimate.

While the Bayes estimate has optimal properties, its computation is generally complicated. In practical applications, the posterior mean of the expected loss may not be analytically available and has to be computed numerically.

Exercises

8.1 Let $X \sim \mathcal{BN}(m, \theta)$ and $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample of X. Assume Θ follows a beta distribution with hyperparameters α and β.
(a) Calculate the posterior mean of $\mu_X(\Theta) = \mathrm{E}(X \mid \Theta)$.
(b) What is the conditional pf of X_{n+1} given the sample data $\boldsymbol{x}$?

8.2 Let $X \sim \mathcal{BN}(m, \theta)$ and $\Theta \sim \mathcal{B}(\alpha, \beta)$. Given the functions $A(\theta)$ and $B(\theta)$ for the $\mathcal{BN}(m, \theta)$ distribution in Table 8.2, identify the functions $a(\alpha, \beta)$ and $b(\alpha, \beta)$ as defined in equation (8.25) so that the $\mathcal{B}(\alpha, \beta)$ distribution is a natural conjugate of $\mathcal{BN}(m, \theta)$. Hence, derive the hyperparameters of the posterior distribution of Θ using equations (8.27) and (8.28).

8.3 In Example 8.10, the mean squared error of the expected loss is analytically derived for the sample mean and the Bühlmann credibility estimate, and numerically estimated for the Bayes estimate. Derive the mean squared errors of the sample mean and the Bühlmann premium as predictors for future loss X_{n+1}. Suggest a simulation procedure for the estimation of the mean squared error of the Bayes predictor for future loss.

8.4 Given $\Theta = \theta, X \sim \mathcal{NB}(r, \theta)$. If the prior distribution of Θ is $\mathcal{B}(\alpha, \beta)$, determine the unconditional pf of X.

8.5 Show that the negative binomial distribution $\mathcal{NB}(r,\theta)$ belongs to the linear exponential family, where r is known and θ is the unknown parameter. Identify the functions $A(\theta)$, $B(\theta)$, and $C(x)$ in equation (8.24).

8.6 Given N, X is distributed as $\mathcal{BN}(N,\theta)$. Derive the unconditional distribution of X assuming N is distributed as (a) $\mathcal{PN}(\lambda)$, and (b) $\mathcal{BN}(m,\beta)$, where m is known.

Questions adapted from SOA exams

8.7 The pf of the annual number of claims N of a particular insurance policy is: $f_N(0) = 2\theta, f_N(1) = \theta$, and $f_N(2) = 1 - 3\theta$. Over different policies, the pf of Θ is: $f_\Theta(0.1) = 0.8$ and $f_\Theta(0.3) = 0.2$. If there is one claim in Year 1, calculate the Bayes estimate of the expected number of claims in Year 2.

8.8 In a portfolio of insurance policies, the number of claims for each policyholder in each year, denoted by N, may be 0, 1, or 2, with the following pf: $f_N(0) = 0.1, f_N(1) = 0.9 - \theta$, and $f_N(2) = \theta$. The prior pdf of Θ is

$$f_\Theta(\theta) = \frac{\theta^2}{0.039}, \qquad 0.2 < \theta < 0.5.$$

A randomly selected policyholder has two claims in Year 1 and two claims in Year 2. Determine the Bayes estimate of the expected number of claims in Year 3 of this policyholder.

8.9 The number of claims N of each policy is distributed as $\mathcal{BN}(8,\theta)$, and the prior distribution of Θ is $\mathcal{B}(\alpha, 9)$. A randomly selected policyholder is found to have made two claims in Year 1 and k claims in Year 2. The Bayesian credibility estimate of the expected number of claims in Year 2 based on the experience of Year 1 is 2.54545, and the Bayesian credibility estimate of the expected number of claims in Year 3 based on the experience of Years 1 and 2 is 3.73333. Determine k.

8.10 Claim severity is distributed as $\mathcal{E}(1/\theta)$. The prior distribution of Θ is inverse gamma with pdf

$$f_\Theta(\theta) = \frac{c^2}{\theta^3}\exp\left(-\frac{c}{\theta}\right), \qquad 0 < \theta < \infty,\ 0 < c.$$

Given an observed loss is x, calculate the mean of the posterior distribution of Θ.

8.11 Consider two random variables D and G, where

$$\Pr(D = d \mid G = g) = g^{1-d}(1-g)^d, \qquad d = 0, 1,$$

and

$$\Pr\left(G = \frac{1}{5}\right) = \frac{3}{5} \qquad \text{and} \qquad \Pr\left(G = \frac{1}{3}\right) = \frac{2}{5}.$$

Calculate

$$\Pr\left(G = \frac{1}{3} \mid D = 0\right).$$

8.12 A portfolio has 100 independently and identically distributed risks. The number of claims of each risk follows a $\mathcal{PN}(\lambda)$ distribution. The prior pdf of Λ is $\mathcal{G}(4, 0.02)$. In Year 1, the following loss experience is observed

Number of claims	Number of risks
0	90
1	7
2	2
3	1
Total	100

Determine the Bayesian expected number of claims of the portfolio in Year 2.

8.13 Claim severity X is distributed as $\mathcal{E}(1/\theta)$. It is known that 80% of the policies have $\theta = 8$ and the other 20% have $\theta = 2$. A randomly selected policy has one claim of size 5. Calculate the Bayes expected size of the next claim of this policy.

8.14 The claim frequency N in a period is distributed as $\mathcal{PN}(\lambda)$, where the prior distribution of Λ is $\mathcal{E}(1)$. If a policyholder makes no claim in a period, determine the posterior pdf of Λ for this policyholder.

8.15 Annual claim frequencies follow a Poisson distribution with mean λ. The prior distribution of Λ has pdf

$$f_\Lambda(\lambda) = 0.4\left(\frac{1}{6}e^{-\frac{\lambda}{6}}\right) + 0.6\left(\frac{1}{12}e^{-\frac{\lambda}{12}},\right), \qquad \lambda > 0.$$

Ten claims are observed for an insured in Year 1. Determine the Bayesian expected number of claims for the insured in Year 2.

8.16 The annual number of claims for a policyholder follows a Poisson distribution with mean λ. The prior distribution of Λ is $\mathcal{G}(5, 0.5)$.

A randomly selected insured has five claims in Year 1 and three claims in Year 2. Determine the posterior mean of Λ.

8.17 The annual number of claims of a given policy is distributed as $\mathcal{GM}(\theta)$. One third of the policies have $\theta = 1/3$ and the remaining two-thirds have $\theta = 1/6$. A randomly selected policy had two claims in Year 1. Calculate the Bayes expected number of claims for the selected policy in Year 2.

8.18 An insurance company sells three types of policies with the following characteristics

Type of policy	Proportion of total policies	Distribution of annual claim frequency
A	5%	$\mathcal{PN}(0.25)$
B	20%	$\mathcal{PN}(0.50)$
C	75%	$\mathcal{PN}(1.00)$

A randomly selected policy is observed to have one claim in each of Years 1 through 4. Determine the Bayes estimate of the expected number of claims of this policyholder in Year 5.

8.19 The annual number of claims for a policyholder is distributed as $\mathcal{BN}(2, \theta)$. The prior distribution of Θ has pdf $f_\Theta(\theta) = 4\theta^3$ for $0 < \theta < 1$. This policyholder had one claim in each of Years 1 and 2. Determine the Bayes estimate of the expected number of claims in Year 3.

8.20 Claim sizes follow the $\mathcal{P}(1, \gamma)$ distribution. Half of the policies have $\gamma = 1$, while the other half have $\gamma = 3$. For a randomly selected policy, the claim in Year 1 was 5. Determine the posterior probability that the claim amount of the policy in Year 2 will exceed 8.

8.21 The probability that an insured will have at least one loss during any year is θ. The prior distribution of Θ is $\mathcal{U}(0, 0.5)$. An insured had at least one loss every year in the last eight years. Determine the posterior probability that the insured will have at least one loss in Year 9.

8.22 The probability that an insured will have exactly one claim is θ. The prior distribution of Θ has pdf

$$f_\Theta(\theta) = \frac{3\sqrt{\theta}}{2}, \qquad 0 < \theta < 1.$$

A randomly selected insured is found to have exactly one claim. Determine the posterior probability of $\theta > 0.6$.

8.23 For a group of insureds, the claim size is distributed as $\mathcal{U}(0, \theta)$, where $\theta > 0$. The prior distribution of Θ has pdf

$$f_\Theta(\theta) = \frac{500}{\theta^2}, \qquad \theta > 500.$$

If two independent claims of amounts 400 and 600 are observed, calculate the probability that the next claim will exceed 550.

8.24 The annual number of claims of each policyholder is distributed as $\mathcal{PN}(\lambda)$. The prior distribution of Λ is $\mathcal{G}(2, 1)$. If a randomly selected policyholder had at least one claim last year, determine the posterior probability that this policyholder will have at least one claim this year.

9

Empirical implementation of credibility

We have discussed the limited-fluctuation credibility method, the Bühlmann and Bühlmann–Straub credibility methods, as well as the Bayesian method for future loss prediction. The implementation of these methods requires the knowledge or assumptions of some unknown parameters of the model. For the limited-fluctuation credibility method, Poisson distribution is usually assumed for claim frequency. In addition, we need to know the coefficient of variation of claim severity if predictions of claim severity or aggregate loss/pure premium are required. For the Bühlmann and Bühlmann–Straub methods, the key quantities required are the expected value of the process variance, μ_{PV}, and the variance of the hypothetical means, σ^2_{HM}. These quantities depend on the assumptions of the prior distribution of the risk parameters and the conditional distribution of the random loss variable. For the Bayesian method, the predicted loss can be obtained relatively easily if the prior distribution is conjugate to the likelihood. Yet the posterior mean, which is the Bayesian predictor of the future loss, depends on the hyperparameters of the posterior distribution. Thus, for the empirical implementation of the Bayesian method, the hyperparameters have to be estimated.

In this chapter, we discuss the estimation of the required parameters for the implementation of the credibility estimates. We introduce the empirical Bayes method, which may be nonparametric, semiparametric, or parametric, depending on the assumptions concerning the prior distribution and the likelihood. Our main focus is on the Bühlmann and Bühlmann–Straub credibility models, the nonparametric implementation of which is relatively straightforward.

Learning objectives

1 Empirical Bayes method
2 Nonparametric estimation

3 Semiparametric estimation
4 Parametric estimation

9.1 Empirical Bayes method

Implementation of the credibility estimates requires the knowledge of some unknown parameters in the model. For the limited-fluctuation method, depending on the loss variable of interest, the mean and/or the variance of the loss variable are required. For example, to determine whether full credibility is attained for the prediction of claim frequency, we need to know λ_N, which can be estimated by the sample mean of the claim frequency.[1] For predicting claim severity and aggregate loss/pure premium, the coefficient of variation of the loss variable, C_X, is also required, which may be estimated by

$$\hat{C}_X = \frac{s_X}{\bar{X}}, \tag{9.1}$$

where s_X and $\bar{X}$ are the sample standard deviation and sample mean of X, respectively.

In the Bühlmann and Bühlmann–Straub frameworks, the key quantities of interest are the expected value of the process variance, μ_{PV}, and the variance of the hypothetical means, σ^2_{HM}. These quantities can be derived from the Bayesian framework and depend on both the prior distribution and the likelihood. In a strictly Bayesian approach, the prior distribution is given and inference is drawn based on the given prior. For practical applications when researchers are not in a position to state the prior, empirical methods may be applied to estimate the hyperparameters. This is called the **empirical Bayes method**. Depending on the assumptions about the prior distribution and the likelihood, empirical Bayes estimation may adopt one of the following approaches:[2]

1 **Nonparametric approach:** In this approach, no assumptions are made about the particular forms of the prior density of the risk parameters $f_\Theta(\theta)$ and the conditional density of the loss variable $f_{X\,|\,\Theta}(x\,|\,\theta)$. The method is very general and applies to a wide range of models.
2 **Semiparametric approach:** In some practical applications, prior experience may suggest a particular distribution for the loss variable X, while the specification of the prior distribution remains elusive. In such cases, parametric assumptions concerning $f_{X\,|\,\Theta}(x\,|\,\theta)$ may be made, while the prior distribution of the risk parameters $f_\Theta(\theta)$ remains unspecified.

[1] Refer to the assumptions for full credibility and its derivation in the limited-fluctuation approach in Section 6.2.1. Recall that λ_N is the mean of the claim frequency, which is assumed to be Poisson.
[2] Further discussions of parametric versus nonparametric estimation can be found in Chapter 10.

3 **Parametric approach:** When the researcher makes specific assumptions about $f_{X \mid \Theta}(x \mid \theta)$ and $f_{\Theta}(\theta)$, the estimation of the parameters in the model may be carried out using the maximum likelihood estimation (MLE) method. The properties of these estimators follow the classical results of MLE, as discussed in Appendix A.19 and Chapter 12. While in some cases the MLE can be derived analytically, in many situations they have to be computed numerically.

9.2 Nonparametric estimation

To implement the limited-fluctuation credibility prediction for claim severity and aggregate loss/pure premium, an estimate of the coefficient of variation C_X is required. $\hat{C}_X$ as defined in equation (9.1) is an example of a nonparametric estimator. Note that under the assumption of a random sample, s_X and $\bar{X}$ are consistent estimators for the population standard deviation and the population mean, respectively, irrespective of the actual distribution of the random loss variable X. Thus, $\hat{C}_X$ is a consistent estimator for C_X, although it is generally not unbiased.[3]

For the implementation of the Bühlmann and Bühlmann–Straub credibility models, the key quantities required are the expected value of the process variance, μ_{PV}, and the variance of the hypothetical means, σ^2_{HM}, which together determine the Bühlmann credibility parameter k. We present below unbiased estimates of these quantities. To the extent that the unbiasedness holds under the mild assumption that the loss observations are statistically independent, and that no specific assumption is made about the likelihood of the loss random variables and the prior distribution of the risk parameters, the estimates are nonparametric.

In Section 7.4 we set up the Bühlmann–Straub credibility model with a sample of loss observations from a risk group. We shall extend this set-up to consider multiple risk groups, each with multiple samples of loss observations over possibly different periods. The results in this set-up will then be specialized to derive results for the situations discussed in Chapter 7. We now formally state the assumptions of the extended set-up as follows:

1 Let X_{ij} denote the loss per unit of exposure and m_{ij} denote the amount of exposure. The index i denotes the ith risk group, for $i = 1, \cdots, r$, with $r > 1$. Given i, the index j denotes the jth loss observation in the ith group, for $j = 1, \cdots, n_i$, where $n_i > 1$ for $i = 1, \cdots, r$. The number of loss observations n_i in each risk group may differ. We may think of j as indexing

[3] Properties of estimators will be discussed in Chapter 10.

an individual within the risk group or a period of the risk group. Thus, for the ith risk group we have loss observations of n_i individuals or periods.

2 X_{ij} are assumed to be independently distributed. The risk parameter of the ith group is denoted by θ_i, which is a realization of the random variable Θ_i. We assume Θ_i to be independently and identically distributed as Θ.

3 The following assumptions are made for the hypothetical means and the process variance

$$\mathrm{E}(X_{ij} \mid \Theta = \theta_i) = \mu_X(\theta_i), \qquad \text{for } i = 1, \cdots, r;\ j = 1, \cdots, n_i, \tag{9.2}$$

and

$$\mathrm{Var}(X_{ij} \mid \theta_i) = \frac{\sigma_X^2(\theta_i)}{m_{ij}}, \qquad \text{for } i = 1, \cdots, r;\ j = 1, \cdots, n_i. \tag{9.3}$$

We define the overall mean of the loss variable as

$$\mu_X = \mathrm{E}[\mu_X(\Theta_i)] = \mathrm{E}[\mu_X(\Theta)], \tag{9.4}$$

the mean of the process variance as

$$\mu_{\mathrm{PV}} = \mathrm{E}[\sigma_X^2(\Theta_i)] = \mathrm{E}[\sigma_X^2(\Theta)], \tag{9.5}$$

and the variance of the hypothetical means as

$$\sigma_{\mathrm{HM}}^2 = \mathrm{Var}[\mu_X(\Theta_i)] = \mathrm{Var}[\mu_X(\Theta)]. \tag{9.6}$$

For future reference, we also define the following quantities

$$m_i = \sum_{j=1}^{n_i} m_{ij}, \qquad \text{for } i = 1, \cdots, r, \tag{9.7}$$

which is the total exposure for the ith risk group; and

$$m = \sum_{i=1}^{r} m_i, \tag{9.8}$$

which is the total exposure over all risk groups. Also, we define

$$\bar{X}_i = \frac{1}{m_i} \sum_{j=1}^{n_i} m_{ij} X_{ij}, \qquad \text{for } i = 1, \cdots, r, \tag{9.9}$$

as the exposure-weighted mean of the ith risk group; and

$$\bar{X} = \frac{1}{m} \sum_{i=1}^{r} m_i \bar{X}_i \tag{9.10}$$

as the overall weighted mean.

The Bühlmann–Straub credibility predictor of the loss in the next period or a random individual of the ith risk group is

$$Z_i \bar{X}_i + (1 - Z_i)\mu_X, \tag{9.11}$$

where

$$Z_i = \frac{m_i}{m_i + k}, \tag{9.12}$$

with

$$k = \frac{\mu_{\text{PV}}}{\sigma^2_{\text{HM}}}. \tag{9.13}$$

To implement the credibility prediction, we need to estimate μ_X, μ_{PV}, and σ^2_{HM}. It is natural to estimate μ_X by $\bar{X}$. To show that $\bar{X}$ is an unbiased estimator of μ_X, we first note that $\text{E}(X_{ij}) = \text{E}[\text{E}(X_{ij} \mid \Theta_i)] = \text{E}[\mu_X(\Theta_i)] = \mu_X$, so that

$$\begin{aligned} \text{E}(\bar{X}_i) &= \frac{1}{m_i} \sum_{j=1}^{n_i} m_{ij} \text{E}(X_{ij}) \\ &= \frac{1}{m_i} \sum_{j=1}^{n_i} m_{ij} \mu_X \\ &= \mu_X, \qquad \text{for } i = 1, \cdots, r. \end{aligned} \tag{9.14}$$

Thus, we have

$$\begin{aligned} \text{E}(\bar{X}) &= \frac{1}{m} \sum_{i=1}^{r} m_i \text{E}(\bar{X}_i) \\ &= \frac{1}{m} \sum_{i=1}^{r} m_i \mu_X \\ &= \mu_X, \end{aligned} \tag{9.15}$$

so that $\bar{X}$ is an unbiased estimator of μ_X.

We now present an unbiased estimator of μ_{PV} in the following theorem.

Theorem 9.1 *The following quantity is an unbiased estimator of* μ_{PV}

$$\hat{\mu}_{\text{PV}} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} m_{ij} (X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^{r} (n_i - 1)}. \tag{9.16}$$

Proof We re-arrange the inner summation term in the numerator of equation (9.16) to obtain

$$\begin{aligned}\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2 &= \sum_{j=1}^{n_i} m_{ij}\left\{[X_{ij}-\mu_X(\theta_i)]-[\bar{X}_i-\mu_X(\theta_i)]\right\}^2 \\ &= \sum_{j=1}^{n_i} m_{ij}[X_{ij}-\mu_X(\theta_i)]^2+\sum_{j=1}^{n_i} m_{ij}[\bar{X}_i-\mu_X(\theta_i)]^2 \\ &\quad -2\sum_{j=1}^{n_i} m_{ij}[X_{ij}-\mu_X(\theta_i)][\bar{X}_i-\mu_X(\theta_i)]. \qquad (9.17)\end{aligned}$$

Simplifying the last two terms on the right-hand side of the above equation, we have

$$\begin{aligned}&\sum_{j=1}^{n_i} m_{ij}[\bar{X}_i-\mu_X(\theta_i)]^2-2\sum_{j=1}^{n_i} m_{ij}[X_{ij}-\mu_X(\theta_i)][\bar{X}_i-\mu_X(\theta_i)] \\ &= m_i[\bar{X}_i-\mu_X(\theta_i)]^2-2[\bar{X}_i-\mu_X(\theta_i)]\sum_{j=1}^{n_i} m_{ij}[X_{ij}-\mu_X(\theta_i)] \\ &= m_i[\bar{X}_i-\mu_X(\theta_i)]^2-2m_i[\bar{X}_i-\mu_X(\theta_i)]^2 \\ &= -m_i[\bar{X}_i-\mu_X(\theta_i)]^2. \qquad (9.18)\end{aligned}$$

Combining equations (9.17) and (9.18), we obtain

$$\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2=\left[\sum_{j=1}^{n_i} m_{ij}[X_{ij}-\mu_X(\theta_i)]^2\right]-m_i[\bar{X}_i-\mu_X(\theta_i)]^2. \qquad (9.19)$$

We now take expectations of the two terms on the right-hand side of the above. First, we have

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{n_i} m_{ij}[X_{ij} - \mu_X(\Theta_i)]^2\right] &= \mathrm{E}\left[\mathrm{E}\left(\sum_{j=1}^{n_i} m_{ij}[X_{ij} - \mu_X(\Theta_i)]^2 \mid \Theta_i\right)\right] \\
&= \mathrm{E}\left[\sum_{j=1}^{n_i} m_{ij}\mathrm{Var}(X_{ij} \mid \Theta_i)\right] \\
&= \mathrm{E}\left[\sum_{j=1}^{n_i} m_{ij}\left[\frac{\sigma_X^2(\Theta_i)}{m_{ij}}\right]\right] \\
&= \sum_{j=1}^{n_i} \mathrm{E}[\sigma_X^2(\Theta_i)] \\
&= n_i\,\mu_{\mathrm{PV}}, \qquad (9.20)
\end{aligned}$$

and, noting that $\mathrm{E}(\bar{X}_i \mid \Theta_i) = \mu_X(\Theta_i)$, we have

$$\begin{aligned}
\mathrm{E}\left\{m_i[\bar{X}_i - \mu_X(\Theta_i)]^2\right\} &= m_i\,\mathrm{E}\left[\mathrm{E}\left\{[\bar{X}_i - \mu_X(\Theta_i)]^2 \mid \Theta_i\right\}\right] \\
&= m_i\,\mathrm{E}\left[\mathrm{Var}(\bar{X}_i \mid \Theta_i)\right] \\
&= m_i\,\mathrm{E}\left[\mathrm{Var}\left(\frac{1}{m_i}\sum_{j=1}^{n_i} m_{ij}X_{ij} \mid \Theta_i\right)\right] \\
&= m_i\,\mathrm{E}\left[\frac{1}{m_i^2}\sum_{j=1}^{n_i} m_{ij}^2\,\mathrm{Var}(X_{ij} \mid \Theta_i)\right] \\
&= m_i\,\mathrm{E}\left[\frac{1}{m_i^2}\sum_{j=1}^{n_i} m_{ij}^2\left(\frac{\sigma_X^2(\Theta_i)}{m_{ij}}\right)\right] \\
&= \frac{1}{m_i}\sum_{j=1}^{n_i} m_{ij}\,\mathrm{E}[\sigma_X^2(\Theta_i)] \\
&= \mathrm{E}\left[\sigma_X^2(\Theta_i)\right] \\
&= \mu_{\mathrm{PV}}. \qquad (9.21)
\end{aligned}$$

Combining equations (9.19), (9.20), and (9.21), we conclude that

$$\mathrm{E}\left[\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2\right] = n_i\mu_{\mathrm{PV}} - \mu_{\mathrm{PV}} = (n_i-1)\mu_{\mathrm{PV}}. \tag{9.22}$$

Thus, taking expectation of equation (9.16), we have

$$\begin{aligned}\mathrm{E}\left(\hat{\mu}_{\mathrm{PV}}\right) &= \frac{\sum_{i=1}^{r}\mathrm{E}\left[\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2\right]}{\sum_{i=1}^{r}(n_i-1)} \\ &= \frac{\sum_{i=1}^{r}(n_i-1)\mu_{\mathrm{PV}}}{\sum_{i=1}^{r}(n_i-1)} \\ &= \mu_{\mathrm{PV}},\end{aligned} \tag{9.23}$$

so that $\hat{\mu}_{\mathrm{PV}}$ is an unbiased estimator of μ_{PV}. □

Note that equation (9.22) shows that

$$\hat{\sigma}_i^2 = \frac{1}{(n_i-1)}\left[\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2\right] \tag{9.24}$$

is also an unbiased estimator of μ_{PV}, for $i = 1,\cdots,r$. These estimators, however, make use of data in the ith risk group only, and are thus not as efficient as $\hat{\mu}_{\mathrm{PV}}$. In contrast, $\hat{\mu}_{\mathrm{PV}}$ is a weighted average of $\hat{\sigma}_i^2$, as it can be written as

$$\hat{\mu}_{\mathrm{PV}} = \sum_{i=1}^{r} w_i\,\hat{\sigma}_i^2, \tag{9.25}$$

where

$$w_i = \frac{n_i-1}{\sum_{i=1}^{r}(n_i-1)}, \tag{9.26}$$

so that the weights are proportional to the degrees of freedom of the risk groups.

We now turn to the estimation of σ^2_{HM} and present an unbiased estimator of σ^2_{HM} in the following theorem.

Theorem 9.2 *The following quantity is an unbiased estimator of* σ^2_{HM}

$$\hat{\sigma}^2_{\mathrm{HM}} = \frac{\left[\sum_{i=1}^{r} m_i(\bar{X}_i-\bar{X})^2\right] - (r-1)\hat{\mu}_{\mathrm{PV}}}{m - \frac{1}{m}\sum_{i=1}^{r} m_i^2}, \tag{9.27}$$

where $\hat{\mu}_{\text{PV}}$ *is defined in equation* (9.16).

Proof We begin our proof by expanding the term $\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2$ in the numerator of equation (9.27) as follows

$$\begin{aligned}
\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2 &= \sum_{i=1}^{r} m_i \left[(\bar{X}_i - \mu_X) - (\bar{X} - \mu_X)\right]^2 \\
&= \sum_{i=1}^{r} m_i(\bar{X}_i - \mu_X)^2 + \sum_{i=1}^{r} m_i(\bar{X} - \mu_X)^2 \\
&\quad - 2\sum_{i=1}^{r} m_i(\bar{X}_i - \mu_X)(\bar{X} - \mu_X) \\
&= \left[\sum_{i=1}^{r} m_i(\bar{X}_i - \mu_X)^2\right] + m(\bar{X} - \mu_X)^2 \\
&\quad - 2(\bar{X} - \mu_X)\sum_{i=1}^{r} m_i(\bar{X}_i - \mu_X) \\
&= \left[\sum_{i=1}^{r} m_i(\bar{X}_i - \mu_X)^2\right] + m(\bar{X} - \mu_X)^2 \\
&\quad - 2m(\bar{X} - \mu_X)^2 \\
&= \left[\sum_{i=1}^{r} m_i(\bar{X}_i - \mu_X)^2\right] - m(\bar{X} - \mu_X)^2. \qquad (9.28)
\end{aligned}$$

We then take expectations on both sides of equation (9.28) to obtain

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2\right] &= \left[\sum_{i=1}^{r} m_i \,\mathrm{E}\left[(\bar{X}_i - \mu_X)^2\right]\right] - m\,\mathrm{E}\left[(\bar{X} - \mu_X)^2\right] \\
&= \left[\sum_{i=1}^{r} m_i \mathrm{Var}(\bar{X}_i)\right] - m\mathrm{Var}(\bar{X}). \qquad (9.29)
\end{aligned}$$

Applying the result in equation (A.115) to $\mathrm{Var}(\bar{X}_i)$, we have

$$\mathrm{Var}(\bar{X}_i) = \mathrm{Var}\left[\mathrm{E}(\bar{X}_i \mid \Theta_i)\right] + \mathrm{E}\left[\mathrm{Var}(\bar{X}_i \mid \Theta_i)\right]. \qquad (9.30)$$

From equation (9.21) we conclude

$$\mathrm{Var}(\bar{X}_i \mid \Theta_i) = \frac{\sigma_X^2(\Theta_i)}{m_i}. \qquad (9.31)$$

Also, as $\mathrm{E}(\bar{X}_i \mid \Theta_i) = \mu_X(\Theta_i)$, equation (9.30) becomes

$$\mathrm{Var}(\bar{X}_i) = \mathrm{Var}\left[\mu_X(\Theta_i)\right] + \frac{\mathrm{E}\left[\sigma_X^2(\Theta_i)\right]}{m_i} = \sigma_{\mathrm{HM}}^2 + \frac{\mu_{\mathrm{PV}}}{m_i}. \tag{9.32}$$

Next, for $\mathrm{Var}(\bar{X})$ in equation (9.29), we have

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{r} m_i \bar{X}_i\right) \\
&= \frac{1}{m^2}\sum_{i=1}^{r} m_i^2 \mathrm{Var}(\bar{X}_i) \\
&= \frac{1}{m^2}\sum_{i=1}^{r} m_i^2 \left(\sigma_{\mathrm{HM}}^2 + \frac{\mu_{\mathrm{PV}}}{m_i}\right) \\
&= \left[\sum_{i=1}^{r} \frac{m_i^2}{m^2}\right]\sigma_{\mathrm{HM}}^2 + \frac{\mu_{\mathrm{PV}}}{m}.
\end{aligned} \tag{9.33}$$

Substituting equations (9.32) and (9.33) into (9.29), we obtain

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2\right] &= \left[\sum_{i=1}^{r} m_i\left(\sigma_{\mathrm{HM}}^2 + \frac{\mu_{\mathrm{PV}}}{m_i}\right)\right] \\
&\quad - \left[\left(\sum_{i=1}^{r} \frac{m_i^2}{m}\right)\sigma_{\mathrm{HM}}^2 + \mu_{\mathrm{PV}}\right] \\
&= \left[m - \frac{1}{m}\sum_{i=1}^{r} m_i^2\right]\sigma_{\mathrm{HM}}^2 + (r-1)\mu_{\mathrm{PV}}.
\end{aligned} \tag{9.34}$$

Thus, taking expectation of $\hat{\sigma}_{\mathrm{HM}}^2$, we can see that

$$\begin{aligned}
\mathrm{E}\left(\hat{\sigma}_{\mathrm{HM}}^2\right) &= \frac{\mathrm{E}\left[\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2\right] - (r-1)\mathrm{E}(\hat{\mu}_{\mathrm{PV}})}{m - \frac{1}{m}\sum_{i=1}^{r} m_i^2} \\
&= \frac{\left[m - \frac{1}{m}\sum_{i=1}^{r} m_i^2\right]\sigma_{\mathrm{HM}}^2 + (r-1)\mu_{\mathrm{PV}} - (r-1)\mu_{\mathrm{PV}}}{m - \frac{1}{m}\sum_{i=1}^{r} m_i^2} \\
&= \sigma_{\mathrm{HM}}^2.
\end{aligned} \tag{9.35}$$

□

From equation (9.16), we can see that an unbiased estimate of μ_{PV} can be obtained with a single risk group, i.e. $r = 1$. However, as can be seen from equation (9.27), $\hat{\sigma}^2_{\text{HM}}$ cannot be computed unless $r > 1$. This is due to the fact that σ^2_{HM} measures the variations in the hypothetical means and requires at least two risk groups for a well-defined estimate.

As the Bühlmann model is a special case of the Bühlmann–Straub model, the results in Theorems 9.1 and 9.2 can be used to derive unbiased estimators of μ_{PV} and σ^2_{HM} for the Bühlmann model. This is summarized in the following corollary.

Corollary 9.1 *In the Bühlmann model with r risk groups, denote the loss observations by X_{ij}, for $i = 1, \cdots, r$, with $r > 1$, and $j = 1, \cdots, n_i$. Let the exposures of X_{ij} be the same, so that without loss of generality we set $m_{ij} \equiv 1$. The following quantity is then an unbiased estimator of μ_{PV}*

$$\tilde{\mu}_{\text{PV}} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^{r} (n_i - 1)}, \tag{9.36}$$

and the following quantity is an unbiased estimator of σ^2_{HM}

$$\tilde{\sigma}^2_{\text{HM}} = \frac{\left[\sum_{i=1}^{r} n_i (\bar{X}_i - \bar{X})^2\right] - (r-1)\tilde{\mu}_{\text{PV}}}{n - \frac{1}{n}\sum_{i=1}^{r} n_i^2}, \tag{9.37}$$

where $n = \sum_{i=1}^{r} n_i$. In particular, if all risk groups have the same sample size, so that $n_i = n^*$ for $i = 1, \cdots, r$, then we have

$$\begin{aligned} \tilde{\mu}_{\text{PV}} &= \frac{1}{r(n^* - 1)} \left[\sum_{i=1}^{r} \sum_{j=1}^{n^*} (X_{ij} - \bar{X}_i)^2 \right] \\ &= \frac{1}{r} \left[\sum_{i=1}^{r} s_i^2 \right], \end{aligned} \tag{9.38}$$

*where s_i^2 is the sample variance of the losses of the i*th *group, and*

$$\begin{aligned} \tilde{\sigma}^2_{\text{HM}} &= \frac{1}{r-1} \left[\sum_{i=1}^{r} (\bar{X}_i - \bar{X})^2 \right] - \frac{\tilde{\mu}_{PV}}{n^*} \\ &= S^2 - \frac{\tilde{\mu}_{PV}}{n^*}, \end{aligned} \tag{9.39}$$

where S^2 is the between-group sample variance.

Proof The proof is a straightforward application of Theorems 9.1 and 9.2, and is left as an exercise. □

With estimated values of the model parameters, the Bühlmann–Straub credibility predictor of the ith risk group can be calculated as

$$\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i)\bar{X}, \tag{9.40}$$

where

$$\hat{Z}_i = \frac{m_i}{m_i + \hat{k}}, \tag{9.41}$$

with

$$\hat{k} = \frac{\hat{\mu}_{\text{PV}}}{\hat{\sigma}^2_{\text{HM}}}. \tag{9.42}$$

For the Bühlmann credibility predictor, equations (9.41) and (9.42) are replaced by

$$\tilde{Z}_i = \frac{n_i}{n_i + \tilde{k}} \tag{9.43}$$

and

$$\tilde{k} = \frac{\tilde{\mu}_{\text{PV}}}{\tilde{\sigma}^2_{\text{HM}}}. \tag{9.44}$$

While $\hat{\mu}_{\text{PV}}$ and $\hat{\sigma}^2_{\text{HM}}$ are unbiased estimators of μ_{PV} and σ^2_{HM}, respectively, $\hat{k}$ is not unbiased for k, due to the fact that k is a nonlinear function of μ_{PV} and σ^2_{HM}. Likewise, $\tilde{k}$ is not unbiased for k.

Note that $\hat{\sigma}^2_{\text{HM}}$ and $\tilde{\sigma}^2_{\text{HM}}$ may be negative in empirical applications. In such circumstances, they may be set to zero, which implies that $\hat{k}$ and $\tilde{k}$ will be infinite, and that $\hat{Z}_i$ and $\tilde{Z}_i$ will be zero for all risk groups. Indeed, if the hypothetical means have no variation, the risk groups are homogeneous and there should be no differential weighting. In sum, the predicted loss is the overall average.

From equation (9.10), the total loss experienced is $m\bar{X} = \sum_{i=1}^{r} m_i \bar{X}_i$. Now if future losses are predicted according to equation (9.40), the total loss predicted will in general be different from the total loss experienced. If it is desired to equate the total loss predicted to the total loss experienced, some re-adjustment is needed. This may be done by using an alternative estimate of the average loss, denoted by $\hat{\mu}_X$, in place of $\bar{X}$ in equation (9.40). Now the total loss predicted becomes

$$\sum_{i=1}^{r} m_i[\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i)\hat{\mu}_X] = \sum_{i=1}^{r} m_i \left\{[1 - (1 - \hat{Z}_i)]\bar{X}_i + (1 - \hat{Z}_i)\hat{\mu}_X \right\}. \tag{9.45}$$

As $m_i(1-\hat{Z}_i) = \hat{Z}_i\hat{k}$, the above equation can be written as

$$\begin{aligned}\text{Total loss predicted} &= \sum_{i=1}^{r} m_i \left\{[1-(1-\hat{Z}_i)]\bar{X}_i + (1-\hat{Z}_i)\hat{\mu}_X\right\} \\ &= \sum_{i=1}^{r} m_i\bar{X}_i + \sum_{i=1}^{r} m_i(1-\hat{Z}_i)(\hat{\mu}_X - \bar{X}_i) \\ &= \text{Total loss experienced} \\ &\quad + \hat{k}\sum_{i=1}^{r}\hat{Z}_i(\hat{\mu}_X - \bar{X}_i). \end{aligned} \tag{9.46}$$

Thus, to balance the total loss predicted and the total loss experienced, we must have $\hat{k}\sum_{i=1}^{r}\hat{Z}_i\left(\hat{\mu}_X - \bar{X}_i\right) = 0$, which implies

$$\hat{\mu}_X = \frac{\sum_{i=1}^{r}\hat{Z}_i\bar{X}_i}{\sum_{i=1}^{r}\hat{Z}_i}, \tag{9.47}$$

and the loss predicted for the ith group is $\hat{Z}_i\bar{X}_i + (1-\hat{Z}_i)\hat{\mu}_X$.

Example 9.1 An analyst has data of the claim frequencies of workers' compensations of three insured companies. Table 9.1 gives the data of company A for the last three years and companies B and C for the last four years. The numbers of workers (in hundreds) and the numbers of claims each year per hundred workers are given.

Table 9.1. *Data for Example* 9.1

		Years			
Company		1	2	3	4
A	Claims per hundred workers	–	1.2	0.9	1.8
	Workers (in hundreds)	–	10	11	12
B	Claims per hundred workers	0.6	0.8	1.2	1.0
	Workers (in hundreds)	5	5	6	6
C	Claims per hundred workers	0.7	0.9	1.3	1.1
	Workers (in hundreds)	8	8	9	10

Calculate the Bühlmann–Straub credibility predictions of the numbers of claims per hundred workers for the three companies next year, without and with corrections for balancing the total loss with the predicted loss.

Solution The total exposures of each company are

$$m_A = 10 + 11 + 12 = 33,$$

$$m_B = 5 + 5 + 6 + 6 = 22,$$

and

$$m_C = 8 + 8 + 9 + 10 = 35,$$

which give the total exposures of all companies as $m = 33 + 22 + 35 = 90$. The exposure-weighted means of the claim frequency of the companies are

$$\bar{X}_A = \frac{(10)(1.2) + (11)(0.9) + (12)(1.8)}{33} = 1.3182,$$

$$\bar{X}_B = \frac{(5)(0.6) + (5)(0.8) + (6)(1.2) + (6)(1.0)}{22} = 0.9182,$$

and

$$\bar{X}_C = \frac{(8)(0.7) + (8)(0.9) + (9)(1.3) + (10)(1.1)}{35} = 1.0143.$$

The numerator of $\hat{\mu}_{\text{PV}}$ in equation (9.16) is

$$\begin{aligned}
&(10)(1.2 - 1.3182)^2 + (11)(0.9 - 1.3182)^2 + (12)(1.8 - 1.3182)^2 \\
&+ (5)(0.6 - 0.9182)^2 + (5)(0.8 - 0.9182)^2 + (6)(1.2 - 0.9182)^2 \\
&+ (6)(1.0 - 0.9182)^2 + (8)(0.7 - 1.0143)^2 + (8)(0.9 - 1.0143)^2 \\
&+ (9)(1.3 - 1.0143)^2 + (10)(1.1 - 1.0143)^2 \\
&= 7.6448.
\end{aligned}$$

Hence, we have

$$\hat{\mu}_{\text{PV}} = \frac{7.6448}{2 + 3 + 3} = 0.9556.$$

The overall mean is

$$\bar{X} = \frac{(1.3182)(33) + (0.9182)(22) + (1.0143)(35)}{90} = 1.1022.$$

The first term in the numerator of $\hat{\sigma}^2_{\text{HM}}$ in equation (9.27) is

$$(33)(1.3182 - 1.1022)^2 + (22)(0.9182 - 1.1022)^2$$

$$+ (35)(1.0143 - 1.1022)^2 = 2.5549,$$

and the denominator is

$$90 - \frac{1}{90}[(33)^2 + (22)^2 + (35)^2] = 58.9111,$$

so that

$$\hat{\sigma}^2_{\text{HM}} = \frac{2.5549 - (2)(0.9556)}{58.9111} = \frac{0.6437}{58.9111} = 0.0109.$$

Thus, the Bühlmann–Straub credibility parameter estimate is

$$\hat{k} = \frac{\hat{\mu}_{\text{PV}}}{\hat{\sigma}^2_{\text{HM}}} = \frac{0.9556}{0.0109} = 87.6697,$$

and the Bühlmann–Straub credibility factor estimates of the companies are

$$\hat{Z}_A = \frac{33}{33 + 87.6697} = 0.2735,$$

$$\hat{Z}_B = \frac{22}{22 + 87.6697} = 0.2006,$$

and

$$\hat{Z}_C = \frac{35}{35 + 87.6697} = 0.2853.$$

We then compute the Bühlmann–Straub credibility predictors of the claim frequencies per hundred workers for company A as

$$(0.2735)(1.3182) + (1 - 0.2735)(1.1022) = 1.1613,$$

for company B as

$$(0.2006)(0.9182) + (1 - 0.2006)(1.1022) = 1.0653,$$

and for company C as

$$(0.2853)(1.0143) + (1 - 0.2853)(1.1022) = 1.0771.$$

Note that the total claim frequency predicted based on the historical exposure is

$$(33)(1.1613) + (22)(1.0653) + (35)(1.0771) = 99.4580,$$

which is not equal to the total recorded claim frequency of $(90)(1.1022) = 99.20$. To balance the two figures, we use equation (9.47) to obtain

$$\hat{\mu}_X = \frac{(0.2735)(1.3182) + (0.2006)(0.9182) + (0.2853)(1.0143)}{0.2735 + 0.2006 + 0.2853} = 1.0984.$$

Using this as the credibility complement, we obtain the updated predictors as

$$\text{A}: \ (0.2735)(1.3182) + (1 - 0.2735)(1.0984) = 1.1585,$$

$$\text{B}: \ (0.2006)(0.9182) + (1 - 0.2006)(1.0984) = 1.0623,$$

$$\text{C}: \ (0.2853)(1.0143) + (1 - 0.2853)(1.0984) = 1.0744.$$

It can be checked that the total claim frequency predicted based on the historical exposure is

$$(33)(1.1585) + (22)(1.0623) + (35)(1.0744) = 99.20,$$

which balances with the total claim frequency recorded. □

Example 9.2 An insurer sold health policies to three companies. The claim experience of these companies in the last period is summarized in Table 9.2.

Table 9.2. *Data for Example 9.2*

Company	Number of employees	Mean claim amount per employee	Standard deviation of claim amount
A	350	467.20	116.48
B	673	328.45	137.80
C	979	390.23	86.50

Suppose company A has 380 employees in the new period, calculate the Bühlmann credibility predictor of its aggregate claim amount.

Solution Assuming the claim amounts of the employees within each company are independently and identically distributed, we employ the Bühlmann model. From equation (9.36), we have

$$\begin{aligned}\tilde{\mu}_{\text{PV}} &= \frac{(349)(116.48)^2 + (672)(137.80)^2 + (978)(86.50)^2}{349 + 672 + 978} \\ &= \frac{24{,}813{,}230.04}{1{,}999} \\ &= 12{,}412.84.\end{aligned}$$

The overall mean of the claim amounts is

$$\begin{aligned}\bar{X} &= \frac{(350)(467.20) + (673)(328.45) + (979)(390.23)}{350 + 673 + 979} \\ &= \frac{766{,}602.02}{2{,}002} \\ &= 382.92.\end{aligned}$$

We compute the denominator of equation (9.37) to obtain

$$2{,}002 - \frac{1}{2{,}002}[(350)^2 + (673)^2 + (979)^2] = 1{,}235.83.$$

Thus, from equation (9.37), we have

$$\begin{aligned}\tilde{\sigma}^2_{\text{HM}} &= [(350)(467.20-382.92)^2+(673)(328.45-382.92)^2 \\ &\quad + (979)(390.23-382.92)^2-(2)(12{,}412.84)]/1{,}235.83 \\ &= 3{,}649.66,\end{aligned}$$

and the Bühlmann credibility parameter estimate is

$$\tilde{k}=\frac{12{,}412.84}{3{,}649.66}=3.4011.$$

For company A, its Bühlmann credibility factor is

$$\tilde{Z}_A=\frac{350}{350+3.4011}=0.99,$$

so that the Bühlmann credibility predictor for the claim amount of the current period is

$$(380)\,[(0.99)(467.20)+(1-0.99)(382.92)]=177{,}215.36. \qquad \square$$

Example 9.3 An insurer insures two rental car companies with similar sizes and operations. The aggregate-loss (in thousand dollars) experience in the last three years is summarized in Table 9.3. Assume the companies have stable business and operations in this period, calculate the predicted aggregate loss of company B next year.

Table 9.3. *Data for Example 9.3*

Company	Mean annual aggregate loss over 3 years	Standard deviation of annual aggregate loss
A	235.35	48.42
B	354.52	76.34

Solution In this problem, the numbers of observations of each risk group are $n^*=3$. We calculate $\tilde{\mu}_{\text{PV}}$ using equation (9.38) to obtain

$$\tilde{\mu}_{\text{PV}}=\frac{(48.42)^2+(76.34)^2}{2}=4{,}086.1460.$$

As the overall mean is

$$\bar{X}=\frac{235.35+354.52}{2}=294.94,$$

using equation (9.39) we obtain $\tilde{\sigma}^2_{\text{HM}}$ as

$$\begin{aligned}\tilde{\sigma}^2_{\text{HM}} &= (235.35 - 294.94)^2 + (354.52 - 294.94)^2 - \frac{4{,}086.1460}{3} \\ &= 5{,}738.6960.\end{aligned}$$

Thus, the Bühlmann credibility parameter estimate is

$$\tilde{k} = \frac{4{,}086.1460}{5{,}738.6960} = 0.7120,$$

so that the estimate of the Bühlmann credibility factor of company B is

$$\tilde{Z}_B = \frac{3}{3 + 0.7120} = 0.8082.$$

Therefore, the Bühlmann credibility prediction of the aggregate loss of company B next year is

$$(0.8082)(354.52) + (1 - 0.8082)(294.94) = 343.09. \qquad \square$$

9.3 Semiparametric estimation

The unbiasedness of $\hat{\mu}_{\text{PV}}$ and $\hat{\sigma}^2_{\text{HM}}$ holds under very mild conditions that the loss random variables X_{ij} are statistically independent of each other and are identically distributed within each risk group (under the same risk parameters). Other than this, no particular assumptions are necessary for the prior distribution of the risk parameters and the conditional distribution of the loss variables. In some applications, however, researchers may have information about the possible conditional distribution $f_{X_{ij} \mid \Theta_i}(x \mid \theta_i)$ of the loss variables. For example, claim frequency per exposure may be assumed to be Poisson distributed. In contrast, the prior distribution of the risk parameters, which are not observable, are usually best assumed to be unknown. Under such circumstances, estimates of the parameters of the Bühlmann–Straub model can be estimated using the semiparametric method.

Suppose X_{ij} are the claim frequencies per exposure and $X_{ij} \sim \mathcal{P}(\lambda_i)$, for $i = 1, \cdots, r$ and $j = 1, \cdots, n_i$. As $\sigma^2_X(\lambda_i) = \lambda_i$, we have

$$\mu_{\text{PV}} = \text{E}[\sigma^2_X(\Lambda_i)] = \text{E}(\Lambda_i) = \text{E}[\text{E}(X \mid \Lambda_i)] = \text{E}(X). \tag{9.48}$$

Thus, μ_{PV} can be estimated using the overall sample mean of X, $\bar{X}$. From (9.27) an alternative estimate of σ^2_{HM} can then be obtained by substituting $\hat{\mu}_{\text{PV}}$ with $\bar{X}$.

Example 9.4 In Example 9.1, if the claim frequencies are assumed to be Poisson distributed, estimate the Bühlmann–Straub credibility parameter k using the semiparametric method.

Solution We estimate μ_{PV} using $\bar{X} = 1.1022$. Thus, the estimate of σ^2_{HM} is

$$\hat{\sigma}^2_{\text{HM}} = \frac{2.5549 - (2)(1.1022)}{58.9111} = 0.005950,$$

so that the semiparametric estimate of the Bühlmann–Straub credibility parameter k is

$$\hat{k} = \frac{1.1022}{0.005950} = 185.24. \qquad \square$$

9.4 Parametric estimation

If the prior distribution of Θ and the conditional distribution of X_{ij} given Θ_i, for $i = 1, \cdots, r$ and $j = 1, \cdots, n_i$ are of known functional forms, then the hyperparameter of Θ, γ, can be estimated using the maximum likelihood estimation (MLE) method.[4] The quantities μ_{PV} and σ^2_{HM} are functions of γ, and we denote them by $\mu_{\text{PV}} = \mu_{\text{PV}}(\gamma)$ and $\sigma^2_{\text{HM}} = \sigma^2_{\text{HM}}(\gamma)$. As k is a function of μ_{PV} and σ^2_{HM}, the MLE of k can be obtained by replacing γ in $\mu_{\text{PV}} = \mu_{\text{PV}}(\gamma)$ and $\sigma^2_{\text{HM}} = \sigma^2_{\text{HM}}(\gamma)$ by the MLE of γ, $\hat{\gamma}$. Specifically, the MLE of k is

$$\hat{k} = \frac{\mu_{\text{PV}}(\hat{\gamma})}{\sigma^2_{\text{HM}}(\hat{\gamma})}. \tag{9.49}$$

We now consider the estimation of γ. For simplicity, we assume $m_{ij} \equiv 1$. The marginal pdf of X_{ij} is given by

$$f_{X_{ij}}(x_{ij} \mid \gamma) = \int_{\theta_i \in \Omega_\Theta} f_{X_{ij} \mid \Theta_i}(x_{ij} \mid \theta_i) f_{\Theta_i}(\theta_i \mid \gamma)\, d\theta_i. \tag{9.50}$$

Given the data X_{ij}, for $i = 1, \cdots, r$ and $j = 1, \cdots, n_i$, the likelihood function $L(\gamma)$ is

$$L(\gamma) = \prod_{i=1}^{r} \prod_{j=1}^{n_i} f_{X_{ij}}(x_{ij} \mid \gamma), \tag{9.51}$$

and the log-likelihood function is

$$\log[L(\gamma)] = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \log f_{X_{ij}}(x_{ij} \mid \gamma). \tag{9.52}$$

[4] See Appendix A.19 for a review of the maximum likelihood estimation method. The result in equation (9.49) is justified by the invariance principle of the MLE (see Section 12.3).

The MLE of γ, $\hat{\gamma}$, is obtained by maximizing $L(\gamma)$ in equation (9.51) or $\log[L(\gamma)]$ in equation (9.52) with respect to γ.

Example 9.5 The claim frequencies X_{ij} are assumed to be Poisson distributed with parameter λ_i, i.e. $X_{ij} \sim \mathcal{PN}(\lambda_i)$. The prior distribution of Λ_i is gamma with hyperparameters α and β, where α is a known constant. Derive the MLE of β and k.

Solution As α is a known constant, the only hyperparameter of the prior is β. The marginal pf of X_{ij} is

$$\begin{aligned} f_{X_{ij}}(x_{ij} \mid \beta) &= \int_0^\infty \left[\frac{\lambda_i^{x_{ij}} \exp(-\lambda_i)}{x_{ij}!}\right]\left[\frac{\lambda_i^{\alpha-1} \exp\left(-\frac{\lambda_i}{\beta}\right)}{\Gamma(\alpha)\beta^\alpha}\right] d\lambda_i \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha x_{ij}!} \int_0^\infty \lambda_i^{x_{ij}+\alpha-1} \exp\left[-\lambda_i\left(\frac{1}{\beta}+1\right)\right] d\lambda_i \\ &= \frac{\Gamma(x_{ij}+\alpha)}{\Gamma(\alpha)\beta^\alpha x_{ij}!}\left(\frac{1}{\beta}+1\right)^{-(x_{ij}+\alpha)} \\ &= \frac{c_{ij}\beta^{x_{ij}}}{(1+\beta)^{x_{ij}+\alpha}}, \end{aligned}$$

where c_{ij} does not involve β. Thus, the likelihood function is

$$L(\beta) = \prod_{i=1}^{r}\prod_{j=1}^{n_i} \frac{c_{ij}\beta^{x_{ij}}}{(1+\beta)^{x_{ij}+\alpha}},$$

and ignoring the term that does not involve β, the log-likelihood function is

$$\log[L(\beta)] = (\log\beta)\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}\right) - [\log(1+\beta)]\left[n\alpha + \sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}\right],$$

where $n = \sum_{i=1}^{r} n_i$. The derivative of $\log[L(\beta)]$ with respect to β is

$$\frac{n\bar{x}}{\beta} - \frac{n(\alpha+\bar{x})}{1+\beta},$$

where

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}\right).$$

The MLE of β, $\hat{\beta}$, is obtained by solving for β when the first derivative of $\log[L(\beta)]$ is set to zero. Hence, we obtain[5]

$$\hat{\beta} = \frac{\bar{x}}{\alpha}.$$

As $X_{ij} \sim \mathcal{PN}(\lambda_i)$ and $\Lambda_i \sim \mathcal{G}(\alpha, \beta)$, $\mu_{\text{PV}} = \text{E}[\sigma_X^2(\Lambda_i)] = \text{E}(\Lambda_i) = \alpha\beta$. Also, $\sigma_{\text{HM}}^2 = \text{Var}[\mu_X(\Lambda_i)] = \text{Var}(\Lambda_i) = \alpha\beta^2$, so that

$$k = \frac{\alpha\beta}{\alpha\beta^2} = \frac{1}{\beta}.$$

Thus, the MLE of k is

$$\hat{k} = \frac{1}{\hat{\beta}} = \frac{\alpha}{\bar{x}}.$$ □

9.5 Summary and discussions

While the hyperparameters of the prior distribution are assumed to be known values in the Bayesian model, they are typically unknown in practical applications. The empirical Bayes method adopts the Bayesian approach of analysis, but treats the hyperparameters as quantities to be obtained from the data. In this chapter, we discuss some empirical Bayes approaches for the estimation of the quantities necessary for the implementation of the credibility prediction.

The nonparametric approach makes no assumption about the prior pdf (or pf) of the risk parameters and the conditional pdf (or pf) of the loss variables. For the Bühlmann–Straub and Bühlmann models, these estimates are easy to calculate. The semiparametric approach assumes knowledge of the conditional pdf (or pf) of the loss variable but not the prior distribution of the risk parameters. We illustrate an application of this approach when the loss variable is distributed as Poisson. When assumptions are made for both the prior and conditional distributions, the likelihood function of the hyperparameters can be derived, at least in principle. We can then estimate the hyperparameters using the MLE method, which may require numerical methods.

[5] It can be verified that the second derivative of $\log[L(\beta)]$ evaluated at $\hat{\beta}$ is negative, so that $\log[L(\beta)]$ is maximized. Note that α is a known constant in the computation of $\hat{\beta}$.

Exercises

9.1 The claim experience of three policyholders in three years is given as follows:

Policyholder		Year 1	Year 2	Year 3
1	Total claims	–	2,200	2,700
	Number in group	–	100	110
2	Total claims	2,100	2,000	1,900
	Number in group	90	80	85
3	Total claims	2,400	2,800	3,000
	Number in group	120	130	140

Determine the Bühlmann–Straub credibility premium for each group in Year 4.

9.2 An actuary is making credibility estimates for rating factors using the Bühlmann–Straub nonparametric empirical Bayes method. Let X_{it} denote the rating for Group i in Year t, for $i = 1$, 2, and 3, and $t = 1, \cdots, T$, and m_{it} denote the exposure. Define $m_i = \sum_{t=1}^{T} m_{it}$ and $\bar{X}_i = (\sum_{t=1}^{T} m_{it} X_{it})/m_i$. The following data are available:

Group	m_i	$\bar{X}_i$
1	50	15.02
2	300	13.26
3	150	11.63

The actuary computed the empirical Bayes estimate of the Bühlmann–Straub credibility factor of Group 1 to be 0.6791.

(a) What are the Bühlmann–Straub credibility estimates of the rating factors for the three groups using the overall mean of X_{it} as the manual rating?

(b) If it is desired to set the aggregate estimated credibility rating equal to the aggregate experienced rating, estimate the rating factors of the three groups.

Questions adapted from SOA exams

9.3 You are given the following data:

	Year 1	Year 2
Total losses	12,000	14,000
Number of policyholders	25	30

If the estimate of the variance of the hypothetical means is 254, determine the credibility factor for Year 3 using the nonparametric empirical Bayes method.

9.4 The claim experience of three territories in the region in a year is as follows:

Territory	Number of insureds	Number of claims
A	10	4
B	20	5
C	30	3

The numbers of claims for each insured each year are independently Poisson distributed. Each insured in a territory has the same number of expected claim frequencies, and the number of insured is constant over time for each territory. Determine the empirical Bayes estimate of the Bühlmann–Straub credibility factor for Territory A.

9.5 You are given the following data:

Group		Year 1	Year 2	Year 3	Total
A	Total claims		10,000	15,000	25,000
	Number in group		50	60	110
	Average		200	250	227.27
B	Total claims	16,000	18,000		34,000
	Number in group	100	90		190
	Average	160	200		178.95
	Total claims				59,000
	Number in group				300
	Average				196.67

If the estimate of the variance of the hypothetical means is 651.03, determine the nonparametric empirical Bayes estimate of the Bühlmann–Straub credibility factor of Group A.

9.6 During a two-year period, 100 policies had the following claims experience:

Total claims in Years 1 and 2	Number of policies
0	50
1	30
2	15
3	4
4	1

You are given that the number of claims per year follows a Poisson distribution, and that each policyholder was insured for the entire two-year period. A randomly selected policyholder had one claim over the two-year period. Using the semiparametric empirical Bayes method, estimate the number of claims in Year 3 for the selected policyholder.

9.7 The number of claims of each policyholder in a block of auto-insurance policies is Poisson distributed. In Year 1, the following data are observed for 8,000 policyholders:

Number of claims	Number of policyholders
0	5,000
1	2,100
2	750
3	100
4	50
5 or more	0

A randomly selected policyholder had one claim in Year 1. Determine the semiparametric Bayes estimate of the number of claims in Year 2 of this policyholder.

9.8 Three policyholders have the following claims experience over three months:

Policyholder	Month 1	Month 2	Month 3	Mean	Variance
A	4	6	5	5	1
B	8	11	8	9	3
C	5	7	6	6	1

Calculate the nonparametric empirical Bayes estimate of the credibility factor for Month 4.

9.9 Over a three-year period, the following claims experience was observed for two insureds who own delivery vans:

Insured		Year 1	Year 2	Year 3
A	Number of vehicles	2	2	1
	Number of claims	1	1	0
B	Number of vehicles	–	3	2
	Number of claims	–	2	3

The number of claims of each insured each year follows a Poisson distribution. Determine the semiparametric empirical Bayes estimate of the claim frequency per vehicle for Insured A in Year 4.

9.10 Three individual policyholders have the following claim amounts over four years:

Policyholder	Year 1	Year 2	Year 3	Year 4
A	2	3	3	4
B	5	5	4	6
C	5	5	3	3

Using the nonparametric empirical Bayes method, calculate the estimated variance of the hypothetical means.

9.11 Two policyholders A and B had the following claim experience in the last four years:

Policyholder	Year 1	Year 2	Year 3	Year 4
A	730	800	650	700
B	655	650	625	750

Determine the credibility premium for Policyholder B using the nonparametric empirical Bayes method.

9.12 The number of claims of each driver is Poisson distributed. The experience of 100 drivers in a year is as follows:

Number of claims	Number of drivers
0	54
1	33
2	10
3	2
4	1

Determine the credibility factor of a single driver using the semiparametric empirical Bayes method.

9.13 During a five-year period, 100 policies had the following claim experience:

Number of claims in Years 1 through 5	Number of policies
0	46
1	34
2	13
3	5
4	2

The number of claims of each policyholder each year follows a Poisson distribution, and each policyholder was insured for the entire period. A randomly selected policyholder had three claims over the five-year period. Using the semiparametric empirical Bayes method, determine the estimate for the number of claims in Year 6 for the same policyholder.

9.14 Denoting X_{ij} as the loss of the ith policyholder in Year j, the following data of four policyholders in seven years are known

$$\sum_{i=1}^{4}\sum_{j=1}^{7}(X_{ij}-\bar{X}_i)^2 = 33.60, \qquad \sum_{i=1}^{4}(\bar{X}_i-\bar{X})^2 = 3.30.$$

Using nonparametric empirical Bayes method, calculate the credibility factor for an individual policyholder.

Part IV

Model construction and evaluation

Model construction and evaluation are two important aspects of the empirical implementation of loss models. To construct a parametric model of loss distributions, the parameters of the distribution have to be estimated based on observed data. Alternatively, we may consider the estimation of the distribution function or density function without specifying their functional forms, in which case nonparametric methods are used. We discuss the estimation techniques for both failure-time data and loss data. Competing models are selected and evaluated based on model selection criteria, including goodness-of-fit tests.

Computer simulation using random numbers is an important tool in analyzing complex problems for which analytical answers are difficult to obtain. We discuss methods of generating random numbers suitable for various continuous and discrete distributions. We also consider the use of simulation for the estimation of the mean squared error of an estimator and the p-value of a hypothesis test, as well as the generation of asset-price paths.

10

Model estimation and types of data

Given the assumption that a loss random variable has a certain parametric distribution, the empirical analysis of the properties of the loss requires the parameters to be estimated. In this chapter we review the theory of parametric estimation, including the properties of an estimator and the concepts of point estimation, interval estimation, unbiasedness, consistency, and efficiency. Apart from the parametric approach, we may also estimate the distribution functions and the probability (density) functions of the loss random variables directly, without assuming a certain parametric form. This approach is called nonparametric estimation. The purpose of this chapter is to provide a brief review of the theory of estimation, with the discussion of specific estimation methods postponed to the next two chapters.

Although the focus of this book is on nonlife actuarial risks, the estimation methods discussed are also applicable to life-contingency models. Specifically, the estimation methods may be used for failure-time data (life risks) as well as loss data (nonlife risks). In many practical applications, only incomplete data observations are available. These observations may be left truncated or right censored. We define the notations to be used in subsequent chapters with respect to left truncation, right censoring, and the risk set. Furthermore, in certain set-ups individual observations may not be available and we may have to work with grouped data. Different estimation methods are required, depending on whether the data are complete or incomplete, and whether they are individual or grouped.

Learning objectives

1 Parametric versus nonparametric estimation
2 Point estimate and interval estimate
3 Unbiasedness, consistency, and efficiency
4 Failure-time data and loss data

5 Complete versus incomplete data, left truncation, and right censoring
6 Individual versus grouped data

10.1 Estimation

We review in this section the basic concepts of estimation, distinguishing between parametric and nonparametric estimation. Properties of parametric estimators will be discussed. Methods of estimation, however, will be postponed to the next two chapters.

10.1.1 Parametric and nonparametric estimation

Loss distributions may be estimated using the parametric or nonparametric approach. In the parametric approach, the distribution is determined by a finite number of parameters. Let the df and pdf (or pf) of the loss random variable X be $F(x;\theta)$ and $f(x;\theta)$, respectively, where θ is the parameter of the df and pdf (pf). To economize notations we shall suppress the suffix X in the functions. This convention will be adopted subsequently unless there is a need to be more specific. Furthermore, θ may be a scalar or vector, although for convenience of exposition we shall treat it as a scalar. When θ is known, the distribution of X is completely specified and various quantities (such as the mean and variance) may be computed. In practical situations θ is unknown and has to be estimated using observed data. Let $\{X_1, \cdots, X_n\}$ be a random sample of n observations of X. We denote $\hat{\theta}$ as an estimator of θ using the random sample.[1] Thus, $F(x;\hat{\theta})$ and $f(x;\hat{\theta})$ are the parametric estimates of the df and pdf (pf), respectively. Parametric estimation of loss distributions will be covered in Chapter 12.

On the other hand, $F(x)$ and $f(x)$ may be estimated directly for all values of x without assuming specific parametric forms, resulting in nonparametric estimates of these functions. For example, the histogram of the sample data is a nonparametric estimate of $f(\cdot)$. Nonparametric estimation of loss distributions will be covered in Chapter 11. Unlike the parametric approach, the nonparametric approach has the benefit of requiring few assumptions about the loss distribution.

10.1.2 Point and interval estimation

As $\hat{\theta}$ assigns a specific value to θ based on the sample, it is called a **point estimator**. In contrast, an **interval estimator** of an unknown parameter is a

[1] An estimator may be thought of as a rule of assigning a point value to the unknown parameter value using the sample observations. In exposition we shall use the terms estimator (the rule) and estimate (the assigned value) interchangeably.

random interval constructed from the sample data, which covers the true value of θ with a certain probability. Specifically, let $\hat{\theta}_L$ and $\hat{\theta}_U$ be functions of the sample data $\{X_1, \cdots, X_n\}$, with $\hat{\theta}_L < \hat{\theta}_U$. The interval $(\hat{\theta}_L, \hat{\theta}_U)$ is said to be a $100(1-\alpha)\%$ **confidence interval** of θ if

$$\Pr(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha. \tag{10.1}$$

For example, let the mean and variance of $\hat{\theta}$ be θ and $\sigma^2_{\hat{\theta}}$, respectively. Suppose $\hat{\theta}$ is normally distributed so that $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2_{\hat{\theta}})$, then a $100(1-\alpha)\%$ confidence interval of θ is

$$(\hat{\theta} - z_{1-\frac{\alpha}{2}}\sigma_{\hat{\theta}},\ \hat{\theta} + z_{1-\frac{\alpha}{2}}\sigma_{\hat{\theta}}). \tag{10.2}$$

When $\sigma_{\hat{\theta}}$ is unknown, it has to be estimated and an alternative quantile may be needed to replace $z_{1-\frac{\alpha}{2}}$ in equation (10.2).

10.1.3 Properties of estimators

As there are possibly many different estimators for the same parameter, an intelligent choice among them is important. We naturally desire the estimate to be *close* to the true parameter value *on average*, leading to the unbiasedness criterion as follows.

Definition 10.1 (Unbiasedness) An estimator of θ, $\hat{\theta}$, is said to be unbiased if and only if $\mathrm{E}(\hat{\theta}) = \theta$.

The unbiasedness of $\hat{\theta}$ for θ requires the condition $\mathrm{E}(\hat{\theta}) = \theta$ to hold for samples of all sizes. In some applications, although $\mathrm{E}(\hat{\theta})$ may not be equal to θ in finite samples, it may approach to θ arbitrarily closely in large samples. We say $\hat{\theta}$ is **asymptotically unbiased** for θ if[2]

$$\lim_{n\to\infty} \mathrm{E}(\hat{\theta}) = \theta. \tag{10.3}$$

As we may have more than one unbiased estimator, the unbiasedness criterion alone may not be sufficient to provide guidance for choosing a good estimator. If we have two unbiased estimators, the closeness requirement suggests that the one with the smaller variance should be preferred. This leads us to the following definition:

Definition 10.2 (Minimum variance unbiased estimator) Suppose $\hat{\theta}$ and $\tilde{\theta}$ are two unbiased estimators of θ, $\hat{\theta}$ is more efficient than $\tilde{\theta}$ if $\mathrm{Var}(\hat{\theta}) < \mathrm{Var}(\tilde{\theta})$. In

[2] Note that $\mathrm{E}(\hat{\theta})$ generally depends on the sample size n.

particular, if the variance of $\hat{\theta}$ is smaller than the variance of any other unbiased estimator of θ, then $\hat{\theta}$ is the minimum variance unbiased estimator of θ.

While asymptotic unbiasedness requires the *mean* of $\hat{\theta}$ to approach θ arbitrarily closely in large samples, a stronger condition is to require $\hat{\theta}$ itself to approach θ arbitrarily closely in large samples. This leads us to the property of consistency.

Definition 10.3 (Consistency) [3] $\hat{\theta}$ is a consistent estimator of θ if it **converges in probability** to θ, which means that for any $\delta > 0$

$$\lim_{n\to\infty} \Pr(|\hat{\theta} - \theta| < \delta) = 1. \tag{10.4}$$

Note that unbiasedness is a property that refers to samples of all sizes, large or small. In contrast, consistency is a property that refers to large samples only.

The following theorem may be useful in identifying consistent estimators:

Theorem 10.1 *$\hat{\theta}$ is a consistent estimator of θ if it is asymptotically unbiased and $Var(\hat{\theta}) \to 0$ when $n \to \infty$.*

Proof See the proof of a similar result in DeGroot and Schervish (2002, p. 234). □

Biased estimators are not necessarily inferior if their average deviation from the true parameter value is small. Hence, we may use the **mean squared error** as a criterion for selecting estimators. The mean squared error of $\hat{\theta}$ as an estimator of θ, denoted by $\text{MSE}(\hat{\theta})$, is defined as

$$\text{MSE}(\hat{\theta}) = \text{E}[(\hat{\theta} - \theta)^2]. \tag{10.5}$$

We note that

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \text{E}[(\hat{\theta} - \theta)^2] \\
&= \text{E}[\{(\hat{\theta} - \text{E}(\hat{\theta})) + (\text{E}(\hat{\theta}) - \theta)\}^2] \\
&= \text{E}[\{\hat{\theta} - \text{E}(\hat{\theta})\}^2] + [\text{E}(\hat{\theta}) - \theta]^2 + 2[\text{E}(\hat{\theta}) - \theta]\text{E}[\hat{\theta} - \text{E}(\hat{\theta})] \\
&= \text{Var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2,
\end{aligned} \tag{10.6}$$

where

$$\text{bias}(\hat{\theta}) = \text{E}(\hat{\theta}) - \theta \tag{10.7}$$

[3] The consistency concept defined here is also called *weak consistency*, and is the only consistency property considered in this book. For further discussions of convergence in probability, readers may refer to DeGroot and Schervish (2002, p. 233).

is the bias of $\hat{\theta}$ as an estimator of θ. Thus, MSE($\hat{\theta}$) is the sum of the variance of $\hat{\theta}$ and the squared bias. A small bias in $\hat{\theta}$ may be tolerated, if the variance of $\hat{\theta}$ is small so that the overall MSE is low.

Example 10.1 Let $\{X_1, \cdots, X_n\}$ be a random sample of X with mean μ and variance σ^2. Prove that the sample mean $\bar{X}$ is a consistent estimator of μ.

Solution First, $\bar{X}$ is unbiased for μ as $\mathrm{E}(\bar{X}) = \mu$. Thus, $\bar{X}$ is asymptotically unbiased. Second, the variance of $\bar{X}$ is

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n},$$

which tends to 0 when n tends to ∞. Hence, by Theorem 10.1, $\bar{X}$ is consistent for μ. □

Example 10.2 Let $\{X_1, \cdots, X_n\}$ be a random sample of X which is distributed as $\mathcal{U}(0, \theta)$. Define $Y = \max\{X_1, \cdots, X_n\}$, which is used as an estimator of θ. Calculate the mean, variance, and mean squared error of Y. Is Y a consistent estimator of θ?

Solution We first determine the distribution of Y. The df of Y is

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) \\ &= \Pr(X_1 \le y, \cdots, X_n \le y) \\ &= [\Pr(X \le y)]^n \\ &= \left(\frac{y}{\theta}\right)^n. \end{aligned}$$

Thus, the pdf of Y is

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{ny^{n-1}}{\theta^n}.$$

Hence, the first two raw moments of Y are

$$\mathrm{E}(Y) = \frac{n}{\theta^n}\int_0^\theta y^n\, dy = \frac{n\theta}{n+1},$$

and

$$\mathrm{E}(Y^2) = \frac{n}{\theta^n}\int_0^\theta y^{n+1}\, dy = \frac{n\theta^2}{n+2}.$$

From equation (10.7), the bias of Y is

$$\mathrm{bias}(Y) = \mathrm{E}(Y) - \theta = \frac{n\theta}{n+1} - \theta = -\frac{\theta}{n+1},$$

so that Y is downward biased for θ. However, as bias(Y) tends to 0 when n tends to ∞, Y is asymptotically unbiased for θ. The variance of Y is

$$\begin{aligned}\text{Var}(Y) &= \text{E}(Y^2) - [\text{E}(Y)]^2 \\ &= \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 \\ &= \frac{n\theta^2}{(n+2)(n+1)^2},\end{aligned}$$

which tends to 0 when n tends to ∞. Thus, by Theorem 10.1, Y is a consistent estimator of θ. Finally, the MSE of Y is

$$\begin{aligned}\text{MSE}(Y) &= \text{Var}(Y) + [\text{bias}(Y)]^2 \\ &= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} \\ &= \frac{2\theta^2}{(n+2)(n+1)},\end{aligned}$$

which also tends to 0 when n tends to ∞. □

10.2 Types of data

The estimation methods to be used for analyzing loss distributions depend crucially on the type of data available. In this section we discuss different data formats and define the notations and terminologies. While our main interests are in loss distributions, the estimation methods discussed are also applicable to data for life contingency. Indeed the adopted terminologies for the estimation methods in the literature have strong connotations of life-contingent data. Thus, we shall introduce the terminologies and estimation methods using examples of life-contingent as well as loss data.

10.2.1 Duration data and loss data

Suppose we are interested in modeling the age-at-death of the individuals in a population. The key variable of interest is then the time each individual has lived since birth, called the age-at-death variable, which involves length-of-time data or **duration data**. There are similar problems in which duration is the key variable of interest. Examples are: (a) the duration of unemployment of an individual in the labor force, (b) the duration of stay of a patient in a hospital, and (c) the survival time of a patient after a major operation. There may be other cases where estimation methods of duration distributions are applicable.

Depending on the specific problem of interest, the methodology may be applied to **failure-time data**, **age-at-death data, survival-time data**, or any duration data in general.

In nonlife actuarial risks, a key variable of interest is the claim-severity or loss distribution. Examples of applications are: (a) the distribution of medical-cost claims in a health insurance policy, (b) the distribution of car insurance claims, and (c) the distribution of compensation for work accidents. These cases involve analysis of **loss data**.

Depending on how the data are collected, information about individuals in the data set may or may not be complete. For example, in a post-operation survival-time study, the researcher may have knowledge about the survival of patients up to a certain time point (end of the study), but does not have information beyond that point. Such incompleteness of information has important implications for the estimation methods used. We shall define the data set-up and notations prior to discussing the estimation methods.

10.2.2 Complete individual data

We begin with the case in which the researcher has complete knowledge about the relevant duration or loss data of the individuals. Let X denote the variable of interest (duration or loss), and $X_1, \cdots, X_n$ denote the values of X for n individuals. We denote the observed sample values by $x_1, \cdots, x_n$. However, there may be duplications of values in the sample, and we assume there are m distinct values arranged in the order $0 < y_1 < \cdots < y_m$, with $m \leq n$. Furthermore, we assume y_j occurs w_j times in the sample, for $j = 1, \cdots, m$. Thus, $\sum_{j=1}^{m} w_j = n$. In the case of age-at-death data, w_j individuals die at age y_j. If all individuals are observed from birth until they die, we have a **complete individual** data set. We define r_j as the **risk set** at *time* y_j, which is the number of individuals in the sample exposed to the possibility of death at time y_j (prior to observing the deaths at y_j).[4] For example, $r_1 = n$, as all individuals in the sample are exposed to the risk of death just prior to time y_1. Similarly, we can see that $r_j = \sum_{i=j}^{m} w_i$, which is the number of individuals who are surviving just prior to time y_j.

The following example illustrates this set-up.

Example 10.3 Let $x_1, \cdots, x_{16}$ be a sample of failure times of a machine part. The values of x_i, arranged in increasing order, are as follows:

$$2, 3, 5, 5, 5, 6, 6, 8, 8, 8, 12, 14, 18, 18, 24, 24.$$

Summarize the data in terms of the set-up above.

[4] The terminology "risk set" refers to the set of individuals as well as the number of individuals exposed to the risk.

Solution There are nine distinct values of failure time in this data set, so that $m = 9$. Table 10.1 summarizes the data in the notations described above.

Table 10.1. *Failure-time data in Example 10.3*

j	y_j	w_j	r_j
1	2	1	16
2	3	1	15
3	5	3	14
4	6	2	11
5	8	3	9
6	12	1	6
7	14	1	5
8	18	2	4
9	24	2	2

From the table it is obvious that $r_{j+1} = r_j - w_j$ for $j = 1, \cdots, m-1$. □

To understand the terminology introduced, let us assume that all individuals in the data were born at the same time $t = 0$. Then all 16 of them are exposed to the risk of death up to time 2, so that $r_1 = 16$. Upon the death of an individual at time $t = 2$, 15 individuals are exposed to death up to time $t = 3$ when another individual dies, so that $r_2 = 15$. This argument is repeated to complete the sequence of risk sets, r_j.

We present another example below based on complete individual data of losses. Although the life-contingent connotation does not apply, the same terminology is used.

Example 10.4 Let $x_1, \cdots, x_{20}$ be a sample of claims of a group medical insurance policy. The values of x_i, arranged in increasing order, are as follows:

$$15, 16, 16, 16, 20, 21, 24, 24, 24, 28, 28, 34, 35, 36, 36, 36, 40, 40, 48, 50.$$

There are no deductible and policy limit. Summarize the data in terms of the set-up above.

Solution There are 12 distinct values of claim costs in this data set, so that $m = 12$. As there are no deductible and policy limits, the observations are ground-up losses with no censoring or truncation. Thus, we have a complete individual data set. Table 10.2 summarizes the data in the notations above.

The risk set r_j at y_j may be interpreted as the number of claims in the data set with claim size larger than or equal to y_j, having observed claims of amount up to but not including y_j. Thus, 20 claims have sizes larger than or equal to 15,

Table 10.2. *Medical claims data in Example 10.4*

j	y_j	w_j	r_j
1	15	1	20
2	16	3	19
3	20	1	16
4	21	1	15
5	24	3	14
6	28	2	11
7	34	1	9
8	35	1	8
9	36	3	7
10	40	2	4
11	48	1	2
12	50	1	1

having observed that no claim of amount less than 15 exists. When one claim of amount 15 is known, there are 19 claims remaining in the data set. Thus, prior to knowing the number of claims of amount 16, there are possibly 19 claims of amount greater than or equal to 16. Again when three claims of amount 16 are known, there are 16 claims remaining in the sample. As the next higher claim is of amount 20, prior to observing claims of amount 20, there are possibly 16 claims of amount greater than or equal to 20. This argument is repeatedly applied to interpret the meaning of the risk set for other values of y_j. □

10.2.3 Incomplete individual data

In certain studies the researcher may not have complete information about each individual observed in the sample. To illustrate this problem, we consider a study of the survival time of patients after a surgical operation. When the study begins it includes data of patients who have recently received an operation. New patients who are operated on during the study are included in the sample as well. All patients are observed until the end of the study, and their survival times are recorded.

If a patient received an operation some time before the study began, the researcher has the information about how long this patient has survived after the operation and the future survival time is conditional on this information. Other patients who received operations at the same time as this individual but did not live up till the study began would not be in the sample. Thus, this individual is observed from a population which has been **left truncated,** i.e. information is not available for patients who do not survive till the beginning of the study. On the other hand, if an individual survives until the end of the study,

the researcher knows the survival time of the patient up to that time, but has no information about when the patient dies. Thus, the observation pertaining to this individual is **right censored**, i.e. the researcher has the partial information that this individual's survival time goes beyond the study but does not know its exact value. While there may be studies with left censoring or right truncation, we shall not consider such cases.

We now define further notations for analyzing incomplete data. Using survival-time studies for exposition, we use d_i to denote the left-truncation status of individual i in the sample. Specifically, $d_i = 0$ if there is no left truncation (the operation was done during the study period), and $d_i > 0$ if there is left truncation (the operation was done d_i periods before the study began). Let x_i denote the survival time (time till death after operation) of the ith individual. If an individual i survives until the end of the study, x_i is not observed and we denote the survival time up to that time by u_i. Thus, for each individual i, there is a x_i value or u_i value (but not both) associated with it. The example below illustrates the construction of the variables introduced.

Example 10.5 A sample of ten patients receiving a major operation is available. The data are collected over 12 weeks and are summarized in Table 10.3. Column 2 gives the time when the individual was first observed, with a value of zero indicating that the individual was first observed when the study began. A nonzero value gives the time when the operation was done, which is also the time when the individual was first observed. For cases in which the operation was done prior to the beginning of the study Column 3 gives the duration from the operation to the beginning of the study. Column 4 presents the time when the observation ceased, either due to death of patient (D in Column 5) or end of study (S in Column 5).

Table 10.3. *Survival time after a surgical operation*

Ind i	Time ind i first obs	Time since operation when ind i first obs	Time when ind i ends	Status when ind i ends
1	0	2	7	D
2	0	4	4	D
3	2	0	9	D
4	4	0	10	D
5	5	0	12	S
6	7	0	12	S
7	0	2	12	S
8	0	6	12	S
9	8	0	12	S
10	9	0	11	D

Determine the d_i, x_i, and u_i values of each individual.

Solution The data are reconstructed in Table 10.4.

Table 10.4. *Reconstruction of Table 10.3*

i	d_i	x_i	u_i
1	2	9	–
2	4	8	–
3	0	7	–
4	0	6	–
5	0	–	7
6	0	–	5
7	2	–	14
8	6	–	18
9	0	–	4
10	0	2	–

Note that d_i is Column 3 of Table 10.3. x_i is defined when there is a 'D' in Column 5 of Table 10.3, while u_i is defined when there is a 'S' in the column. x_i is Column 4 minus Column 2, or Column 4 plus Column 3, depending on when the individual was first observed. u_i is computed in a similar way. □

As in the case of a complete data set, we assume that there are m distinct failure-time numbers x_i in the sample, arranged in increasing order, as $0 < y_1 < \cdots < y_m$, with $m \leq n$. Assume y_j occurs w_j times in the sample, for $j = 1, \cdots, m$. Again, we denote r_j as the risk set at y_j, which is the number of individuals in the sample exposed to the possibility of death at time y_j (prior to observing the deaths at y_j). To update the risk set r_j after knowing the number of deaths at time y_{j-1}, we use the following formula

$$r_j = r_{j-1} - w_{j-1} + \text{number of observations with } y_{j-1} \leq d_i < y_j$$
$$- \text{number of observations with } y_{j-1} \leq u_i < y_j, \quad j = 2, \cdots, m. \quad (10.8)$$

Note that upon w_{j-1} deaths at failure-time y_{j-1}, the risk set is reduced to $r_{j-1} - w_{j-1}$. This number is supplemented by the number of individuals with $y_{j-1} \leq d_i < y_j$, who are now exposed to risk at failure-time y_j, but are formerly not in the risk set due to left truncation. Note that if an individual has a d value that ties with y_j, this individual is *not included* in the risk set r_j. Furthermore, the risk set is reduced by the number of individuals with $y_{j-1} \leq u_i < y_j$, i.e. those whose failure times are not observed due to right censoring. If an individual

has a u value that ties with y_j, this individual is *not excluded* from the risk set r_j

Equation (10.8) can also be computed equivalently using the following formula

$$\begin{aligned} r_j = &\text{ number of observations with } d_i < y_j \\ &- \text{number of observations with } x_i < y_j \\ &- \text{number of observations with } u_i < y_j, \quad j = 1, \cdots, m. \end{aligned} \tag{10.9}$$

Note that the number of observations with $d_i < y_j$ is the total number of individuals who are potentially facing the risk of death at failure-time y_j. However, individuals with $x_i < y_j$ or $u_i < y_j$ are removed from this risk set as they have either died prior to time y_j (when $x_i < y_j$) or have been censored from the study (when $u_i < y_j$). Indeed, to compute r_j using equation (10.8), we need to calculate r_1 using equation (10.9) to begin the recursion.

Example 10.6 Using the data in Example 10.5, determine the risk set at each failure time in the sample.

Solution The results are summarized in Table 10.5. Columns 5 and 6 describe the computation of the risk sets using equations (10.8) and (10.9), respectively.

Table 10.5. *Ordered death times and risk sets of Example 10.6*

j	y_j	w_j	r_j	Eq (10.8)	Eq (10.9)
1	2	1	6	–	$6-0-0$
2	6	1	6	$6-1+3-2$	$9-1-2$
3	7	1	6	$6-1+1-0$	$10-2-2$
4	8	1	4	$6-1+0-1$	$10-3-3$
5	9	1	3	$4-1+0-0$	$10-4-3$

It can be seen that equations (10.8) and (10.9) give the same answers. □

The methods for computing the risk sets of duration data can also be applied to compute the risk sets of loss data. Left truncation in loss data occurs when there is a deductible in the policy. Likewise, right censoring occurs when there is a policy limit. When the loss data come from insurance policies with the same deductible, say d, and the same maximum covered loss, say u, these values are applied to all observations. The example below illustrates a case where there are several policies in the data with different deductibles and maximum covered losses.

Example 10.7 Table 10.6 summarizes the loss claims of 20 insurance policies, numbered by i, with d_i = deductible, x_i = ground-up loss, and u_i^* = maximum covered loss. For policies with losses larger than u_i^*, only the u_i^* value is recorded. The right-censoring variable is denoted by u_i. Determine the risk set r_j of each distinct loss value y_j.

Table 10.6. *Insurance claims data of Example 10.7*

i	d_i	x_i	u_i^*	u_i	i	d_i	x_i	u_i^*	u_i
1	0	12	15	–	11	3	14	15	–
2	0	10	15	–	12	3	–	15	15
3	0	8	12	–	13	3	12	18	–
4	0	–	12	12	14	4	15	18	–
5	0	–	15	15	15	4	–	18	18
6	2	13	15	–	16	4	8	18	–
7	2	10	12	–	17	4	–	15	15
8	2	9	15	–	18	5	–	20	20
9	2	–	18	18	19	5	18	20	–
10	3	6	12	–	20	5	8	20	–

Solution The distinct values of x_i, arranged in order, are

$$6, 8, 9, 10, 12, 13, 14, 15, 18,$$

so that $m = 9$. The results are summarized in Table 10.7. As in Table 10.5 of Example 10.6, Columns 5 and 6 describe the computation of the risk sets using equations (10.8) and (10.9), respectively.

Table 10.7. *Ordered claim losses and risk sets of Example 10.7*

j	y_j	w_j	r_j	Eq (10.8)	Eq (10.9)
1	6	1	20	–	$20-0-0$
2	8	3	19	$20-1+0-0$	$20-1-0$
3	9	1	16	$19-3+0-0$	$20-4-0$
4	10	2	15	$16-1+0-0$	$20-5-0$
5	12	2	13	$15-2+0-0$	$20-7-0$
6	13	1	10	$13-2+0-1$	$20-9-1$
7	14	1	9	$10-1+0-0$	$20-10-1$
8	15	1	8	$9-1+0-0$	$20-11-1$
9	18	1	4	$8-1+0-3$	$20-12-4$

Thus, just like the case of duration data, the risk sets of loss data can be computed using equations (10.8) and (10.9). □

Example 10.8 A sample of insurance policies with a deductible of 4 and maximum covered loss of 20 has the following ground-up loss amounts

$$5, 7, 8, 10, 10, 16, 17, 17, 17, 19, 20, 20^+, 20^+, 20^+, 20^+,$$

where 20^+ denotes the right-censored loss amount of 20. Determine the risk sets of the distinct loss values.

Solution Each of the observations has $d_i = 4$, and there are four observations with $u_i = 20$. The eight distinct loss values y_j, with their associated w_j and r_j values, are given in Table 10.8.

Table 10.8. *Results of Example 10.8*

j	y_j	w_j	r_j
1	5	1	15
2	7	1	14
3	8	1	13
4	10	2	12
5	16	1	10
6	17	3	9
7	19	1	6
8	20	1	5

Note that all observations from this sample are from a left truncated population. Hence, analysis of the ground-up loss distribution is not possible and only the conditional distribution (given loss larger than 4) can be analyzed. □

10.2.4 Grouped data

Sometimes we work with grouped observations rather than individual observations. This situation is especially common when the number of individual observations is large and there is no significant loss of information in working with grouped data.

Let the values of the failure-time or loss data be divided into k intervals: $(c_0, c_1]$, $(c_1, c_2], \cdots, (c_{k-1}, c_k]$, where $0 \le c_0 < c_1 < \cdots < c_k$. The observations are classified into the interval groups according to the values of x_i (failure time or loss). We first consider complete data. Let there be n

observations of x_i in the sample, with n_j observations of x_i in interval $(c_{j-1}, c_j]$, so that $\sum_{j=1}^{k} n_j = n$. As there is no left truncation; all observations in the sample are in the risk set for the first interval. Thus, the risk set in interval $(c_0, c_1]$ is n. The number of deaths in each interval is then subtracted from the risk set to obtain the risk set for the next interval. This is due to the fact that there is no new addition of risk (no left truncation) and no extra reduction in risk (no right censoring). Thus, the risk set in interval $(c_1, c_2]$ is $n - n_1$. In general, the risk set in interval $(c_{j-1}, c_j]$ is $n - \sum_{i=1}^{j-1} n_i = \sum_{i=j}^{k} n_i$.

When the data are incomplete, with possible left truncation and/or right censoring, approximations may be required to compute the risk sets. For this purpose, we first define the following quantities based on the attributes of individual observations:

$$
\begin{aligned}
D_j &= \text{number of observations with } c_{j-1} \le d_i < c_j, \text{ for } j = 1, \cdots, k, \\
U_j &= \text{number of observations with } c_{j-1} < u_i \le c_j, \text{ for } j = 1, \cdots, k, \\
V_j &= \text{number of observations with } c_{j-1} < x_i \le c_j, \text{ for } j = 1, \cdots, k.
\end{aligned}
$$

Thus, D_j is the number of new additions to the risk set in the interval $(c_{j-1}, c_j]$, U_j is the number of right-censored observations that exit the sample in the interval $(c_{j-1}, c_j]$, and V_j is the number of deaths or loss values in $(c_{j-1}, c_j]$. Note the difference in the intervals for defining D_j versus U_j and V_j. A tie of a d value with c_{j-1} is a new addition of risk to the interval $(c_{j-1}, c_j]$. On the other hand, a tie of a death or a right-censored observation with c_j is a loss in the risk set from the interval $(c_{j-1}, c_j]$. This is due to the fact that D_j determines entry into $(c_{j-1}, c_j]$, while U_j and V_j record exits from it.

We now define R_j as the risk set for the interval $(c_{j-1}, c_j]$, which is the total number of observations in the sample exposed to the risk of failure or loss in $(c_{j-1}, c_j]$. We assume that any entry in an interval contributes to the risk group in the whole interval, while any exit only reduces the risk group in the next interval. Thus, for the first interval $(c_0, c_1]$, we have $R_1 = D_1$. Subsequent updating of the risk set is computed as

$$
R_j = R_{j-1} - V_{j-1} + D_j - U_{j-1}, \qquad j = 2, \cdots, k. \tag{10.10}
$$

An alternative formula for R_j is

$$
R_j = \sum_{i=1}^{j} D_i - \sum_{i=1}^{j-1} (V_i + U_i), \qquad j = 2, \cdots, k. \tag{10.11}
$$

Note that equations (10.10) and (10.11) can be compared against equations (10.8) and (10.9), respectively.

In the case of complete data, $D_1 = n$ and $D_j = 0$ for $j = 2, \cdots, k$. Also, $U_j = 0$ and $V_j = n_j$ for $j = 1, \cdots, k$. Thus, $R_1 = D_1 = n$, and from equation (10.10) we have

$$\begin{aligned} R_j &= R_{j-1} - n_{j-1} \\ &= n - (n_1 + \cdots + n_{j-1}) \\ &= n_j + \cdots + n_k, \end{aligned} \tag{10.12}$$

using recursive substitution. This result has been obtained directly from the properties of complete data.

Example 10.9 For the data in Table 10.6, the observations are grouped into the intervals: (0, 4], (4, 8], (8, 12], (12, 16], and (16, 20]. Determine the risk set in each interval.

Solution We tabulate the results as shown in Table 10.9.

Table 10.9. *Results of Example 10.9*

Group j	D_j	U_j	V_j	R_j
(0, 4]	13	0	0	13
(4, 8]	7	0	4	20
(8, 12]	0	1	5	16
(12, 16]	0	3	3	10
(16, 20]	0	3	1	4

D_j and U_j are obtained from the d and u values in Table 10.6, respectively. Likewise, V_j are obtained by accumulating the w values in Table 10.7. Note that the sum of D_j equals the total number of observations; the sum of U_j equals the number of right-censored observations (i.e., 7) and the sum of V_j equals the number of uncensored loss observations (i.e., 13). The risk sets R_j are calculated using equation (10.10). □

10.3 Summary and discussions

We have reviewed the theories of model estimation and briefly surveyed the use of the parametric and nonparametric estimation approaches. As the estimation methods used are applicable to both duration and loss data, we use examples from both literature. While the terminologies follow the interpretation of age-at-death data, they are also used for loss data. The purpose of this chapter is to define the notations used for complete and incomplete data, as preparation for the discussion of various estimation methods in the next two chapters. In

particular, we explain the concept of the risk set, and show how this can be computed using individual and grouped data. For each observation we have to keep track of a set of values representing left truncation, right censoring, and age at death. The computation of the risk sets requires special care when the observations are incomplete, due to left truncation and/or right censoring. In grouped data some assumptions regarding entry and exit of risks are required. We have discussed one simple approach, although other methods are available.

Exercises

10.1 Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be a random sample of n observations of X, which has mean μ and variance σ^2. Prove that the sample mean $\bar{x}$ and the sample variance s^2 of $\boldsymbol{x}$ are unbiased estimators of μ and σ^2, respectively.

10.2 Let $\hat{\theta}_1$ be an estimator of θ, which is known to lie in the interval (a, b). Define an estimator of θ, $\hat{\theta}_2$, by $\hat{\theta}_2 = \hat{\theta}_1$ if $\hat{\theta}_1 \in (a, b)$, $\hat{\theta}_2 = a$ if $\hat{\theta}_1 \leq a$, and $\hat{\theta}_2 = b$ if $\hat{\theta}_1 \geq b$. Prove that $\text{MSE}(\hat{\theta}_2) \leq \text{MSE}(\hat{\theta}_1)$. If $\hat{\theta}_1$ is an unbiased estimator of θ, also show that $\text{Var}(\hat{\theta}_2) \leq \text{Var}(\hat{\theta}_1)$.

10.3 Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be a random sample of n observations of a continuous distribution X with pdf $f_X(\cdot)$ and df $F_X(\cdot)$. Denote $x_{(r)}$ as the rth-order statistic, such that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Show that the joint pdf of $X_{(r)}$ and $X_{(s)}$, $r < s$, is

$$\frac{n!}{(r-1)!(s-r-1)!(n-s)!} \left[F_X(x_{(r)})\right]^{r-1} \times \left[F_X(x_{(s)}) - F_X(x_{(r)})\right]^{s-r-1} \left[1 - F_X(x_{(s)})\right]^{n-s} f_X(x_{(r)}) f_X(x_{(s)}).$$

If $X \sim \mathcal{U}(0, \theta)$, show that the joint pdf of $X_{(n-1)}$ and $X_{(n)}$ is

$$\frac{n(n-1)x_{(n-1)}^{n-2}}{\theta^n}.$$

For $X \sim \mathcal{U}(0, \theta)$, compute $\text{E}(X_{(n-1)})$ and $\text{Var}(X_{(n-1)})$. What is the MSE of $X_{(n-1)}$ as an estimator of θ? Compare this against the MSE of $X_{(n)}$. What is the covariance of $X_{(n-1)}$ and $X_{(n)}$?

10.4 Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be a random sample of n observations of a continuous distribution X, with pdf $f_X(x) = \exp[-(x - \delta)]$ for $x > \delta$ and 0 otherwise.

(a) Is the sample mean $\bar{X}$ an unbiased estimator of δ? Is it a consistent estimator of δ?

(b) Is the first-order statistic $X_{(1)}$ an unbiased estimator of δ? Is it a consistent estimator of δ?

10.5 If $X \sim \mathcal{BN}(n, \theta)$, show that the sample proportion $p = X/n$ is an unbiased estimator of θ. Is $np(1-p)$ an unbiased estimator of the variance of X?

10.6 Suppose the moments of X exist up to order k, and $\boldsymbol{x} = (x_1, \cdots, x_n)$ is a random sample of X. Show that $\hat{\mu}'_r = (\sum_{i=1}^{n} x^r)/n$ is an unbiased estimate of $\mathrm{E}(X^r)$ for $r = 1, \cdots, k$. Can you find an unbiased estimate of $[\mathrm{E}(X)]^2$?

10.7 Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be a random sample of n observations of $\mathcal{N}(\mu, \sigma^2)$. Construct a $100(1-\alpha)\%$ confidence interval of σ^2.

10.8 Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be a random sample of n observations of $\mathcal{E}(\lambda)$. Construct a $100(1-\alpha)\%$ confidence interval of λ.

10.9 Applicants for a job were given a task to perform and the time they took to finish the task was recorded. The table below summarizes the time applicant i started the task (B_i) and the time it was finished (E_i):

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
B_i	2	4	4	5	6	7	8	8	8	9	10	11	11	12	15	18	18	20
E_i	7	6	9	12	14	17	14	13	20	21	21	23	20	19	18	24	25	24

Summarize the data, giving the durations of the task, the distinct observed durations and their frequencies, and the risk set at each observed duration.

10.10 In a graduate employment survey the starting monthly salary of each respondent, to the nearest hundred dollars, was recorded. The following data are collected:

$$23, 23, 25, 27, 27, 27, 28, 30, 30, 31, 33, 38, 42, 45, 45, 45.$$

What is the risk set at each observed salary?

10.11 The Labor Bureau conducted a survey on unemployment in a small town. The data below give the time B_i the unemployed person i reported out of job ($B_i = -k$ if the person was first unemployed k weeks before the survey started) and the time E_i the unemployed person found a job. The survey spanned a period of 20 weeks, and people who remained unemployed at the end of the survey are

recorded as U:

i	B_i	E_i/U
1	−4	12
2	−2	9
3	1	6
4	3	16
5	4	12
6	4	18
7	4	19
8	5	U
9	6	19
10	6	U

Compute the risk set at each observed unemployment duration.

10.12 Insurance policies have deductibles d and ground-up losses x. Twenty losses are summarized below:

i	d_i	x_i	i	d_i	x_i
1	0	8	11	4	3
2	0	6	12	4	5
3	0	6	13	6	7
4	2	3	14	6	5
5	2	9	15	6	8
6	2	7	16	6	10
7	2	9	17	6	7
8	2	5	18	6	4
9	4	6	19	6	8
10	4	8	20	6	9

Compute the risk set of each observed ground-up loss amount.

10.13 Ground-up losses of 25 insurance policies with a deductible of 5 and maximum covered loss of 18 are summarized as follows:

2, 4, 4, 5, 6, 6, 6, 8, 8, 12, 13, 13, 14, 15, 15, 16, 16, 17, 17, 18, 18, 19, 20, 23, 28.

Determine the risk sets of the reported distinct ground-up loss values.

10.14 Insurance policies have deductibles d, maximum covered losses 15, and ground-up losses x. Twenty losses are recorded below:

i	d_i	x_i	i	d_i	x_i
1	0	12	11	4	13
2	0	16	12	4	18
3	0	4	13	4	7
4	0	12	14	4	12
5	0	15	15	4	18
6	2	14	16	5	10
7	2	13	17	5	9
8	2	17	18	5	8
9	2	9	19	5	18
10	2	8	20	5	13

(a) Compute the risk set of each observed ground-up loss amount.
(b) If the loss observations are grouped into the intervals $(0, 5]$, $(5, 10]$, and $(10, 15]$, determine the risk set in each interval.

11

Nonparametric model estimation

The main focus of this chapter is the estimation of the distribution function and probability (density) function of duration and loss variables. The methods used depend on whether the data are for individual or grouped observations, and whether the observations are complete or incomplete.

For complete individual observations, the relative frequency distribution of the sample observations defines a discrete distribution called the empirical distribution. Moments and df of the true distribution can be estimated using the empirical distribution. Smoothing refinements can be applied to the empirical df to improve its performance. We also discuss kernel-based estimation methods for the estimation of the df and pdf.

When the sample observations are incomplete, with left truncation and/or right censoring, the Kaplan–Meier (product-limit) estimator and the Nelson–Aalen estimator can be used to estimate the survival function. These estimators compute the conditional survival probabilities using observations arranged in increasing order. They make use of the data set-up discussed in the last chapter, in particular the risk set at each observed data point. We also discuss the estimation of their variance, the Greenwood formula, and interval estimation.

For grouped data, smoothing techniques are used to estimate the moments, the quantiles, and the df. The Kaplan–Meier and Nelson–Aalen estimators can also be applied to grouped incomplete data.

Learning objectives

1 Empirical distribution
2 Moments and df of the empirical distribution
3 Kernel estimates of df and pdf
4 Kaplan–Meier (product-limit) estimator and Nelson–Aalen estimator
5 Greenwood formula
6 Estimation based on grouped observations

11.1 Estimation with complete individual data

We first consider the case where we have complete individual observations. We discuss methods of estimating the df, the quantiles, and moments of the duration and loss variables, as well as censored moments when these variables are subject to censoring.

11.1.1 Empirical distribution

We continue to use the notations introduced in the last chapter. Thus, we have a sample of n observations of failure times or losses X, denoted by $x_1, \cdots, x_n$. The distinct values of the observations are arranged in increasing order and are denoted by $0 < y_1 < \cdots < y_m$, where $m \leq n$. The value of y_j is repeated w_j times, so that $\sum_{j=1}^{m} w_j = n$. We also denote g_j as the partial sum of the number of observations not more than y_j, i.e. $g_j = \sum_{h=1}^{j} w_h$.

The **empirical distribution** of the data is defined as the discrete distribution which can take values $y_1, \cdots, y_m$ with probabilities $w_1/n, \cdots, w_m/n$, respectively. Alternatively, it is a discrete distribution for which the values $x_1, \cdots, x_n$ (with possible repetitions) occur with equal probabilities. Denoting $\hat{f}(\cdot)$ and $\hat{F}(\cdot)$ as the pf and df of the empirical distribution, respectively, these functions are given by

$$\hat{f}(y) = \begin{cases} \dfrac{w_j}{n}, & \text{if } y = y_j \text{ for some } j, \\ 0, & \text{otherwise,} \end{cases} \tag{11.1}$$

and

$$\hat{F}(y) = \begin{cases} 0, & \text{for } y < y_1, \\ \dfrac{g_j}{n}, & \text{for } y_j \leq y < y_{j+1}, \, j = 1, \cdots, m-1, \\ 1, & \text{for } y_m \leq y. \end{cases} \tag{11.2}$$

Thus, the **mean of the empirical distribution** is

$$\sum_{j=1}^{m} \frac{w_j}{n} y_j = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{11.3}$$

which is the sample mean of $x_1, \cdots, x_n$, i.e. $\bar{x}$. The **variance of the empirical distribution** is

$$\sum_{j=1}^{m} \frac{w_j}{n} (y_j - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{11.4}$$

which is not equal to the sample variance of $x_1, \cdots, x_n$, and is biased for the variance of X.

Estimates of the moments of X can be computed from their sample analogues. In particular, censored moments can be estimated from the censored sample. For example, for a policy with policy limit u, the censored kth moment $\mathrm{E}\left[(X \wedge u)^k\right]$ can be estimated by

$$\sum_{j=1}^{r} \frac{w_j}{n} y_j^k + \frac{n - g_r}{n} u^k, \qquad \text{where } y_r \le u < y_{r+1} \text{ for some } r. \tag{11.5}$$

$\hat{F}(y)$ defined in equation (11.2) is also called the **empirical distribution function** of X. Likewise, the **empirical survival function** of X is $\hat{S}(y) = 1 - \hat{F}(y)$, which is an estimate of $\Pr(X > y)$.

To compute an estimate of the df for a value of y not in the set $y_1, \cdots, y_m$, we may *smooth* the empirical df to obtain $\tilde{F}(y)$ as follows

$$\tilde{F}(y) = \frac{y - y_j}{y_{j+1} - y_j} \hat{F}(y_{j+1}) + \frac{y_{j+1} - y}{y_{j+1} - y_j} \hat{F}(y_j), \tag{11.6}$$

where $y_j \le y < y_{j+1}$ for some $j = 1, \cdots, m-1$. Thus, $\tilde{F}(y)$ is the linear interpolation of $\hat{F}(y_{j+1})$ and $\hat{F}(y_j)$, called the **smoothed empirical distribution function**. At values of $y = y_j$, $\tilde{F}(y) = \hat{F}(y)$.

To estimate the quantiles of the distribution, we also use interpolation. Recall that the quantile x_δ is defined as $F^{-1}(\delta)$. We use y_j as an estimate of the $(g_j/(n+1))$-quantile (or the $(100g_j/(n+1))$th percentile) of X.[1] The δ-quantile of X, denoted by $\hat{x}_\delta$, may be computed as

$$\hat{x}_\delta = \left[\frac{(n+1)\delta - g_j}{w_{j+1}}\right] y_{j+1} + \left[\frac{g_{j+1} - (n+1)\delta}{w_{j+1}}\right] y_j, \tag{11.7}$$

where

$$\frac{g_j}{n+1} \le \delta < \frac{g_{j+1}}{n+1}, \qquad \text{for some } j. \tag{11.8}$$

Thus, $\hat{x}_\delta$ is a smoothed estimate of the sample quantiles, and is obtained by linearly interpolating y_j and y_{j+1}. It can be verified that if $\delta = g_j/(n+1)$,

[1] Hence, the largest observation in the sample, y_m, is the $(100n/(n+1))$th percentile, and *not* the 100th percentile. Recall from Example 10.2 that, if X is distributed as $\mathcal{U}(0, \theta)$, the mean of the largest sample observation is $n\theta/(n+1)$. Likewise, if y_1 has no tie, it is an estimate of the $(100/(n+1))$th percentile. Note that this is not the only way to define the sample quantile. We may alternatively use y_j as an estimate of the $((g_j - 0.5)/n)$-quantile. See Hyndman and Fan (1996) for more details. Unless otherwise stated, we shall use equations (11.7) through (11.10) to compute the quantile $\hat{x}_\delta$.

$\hat{x}_\delta = y_j$. Furthermore, when there are no ties in the observations, $w_j = 1$ and $g_j = j$ for $j = 1, \cdots, n$. Equation (11.7) then reduces to

$$\hat{x}_\delta = [(n+1)\delta - j]\, y_{j+1} + [j + 1 - (n+1)\delta]\, y_j, \tag{11.9}$$

where

$$\frac{j}{n+1} \le \delta < \frac{j+1}{n+1}, \qquad \text{for some } j. \tag{11.10}$$

Example 11.1 A sample of losses has the following ten observations:

$$2, 4, 5, 8, 8, 9, 11, 12, 12, 16.$$

Plot the empirical distribution function, the smoothed empirical distribution function, and the smoothed quantile function. Determine the estimates $\tilde{F}(7.2)$ and $\hat{x}_{0.75}$. Also, estimate the censored variance $\text{Var}[(X \wedge 11.5)]$.

Solution The plots of various functions are given in Figure 11.1. The empirical distribution function is a step function represented by the solid lines. The dashed line represents the smoothed empirical df, and the dotted line gives the (inverse) of the quantile function.

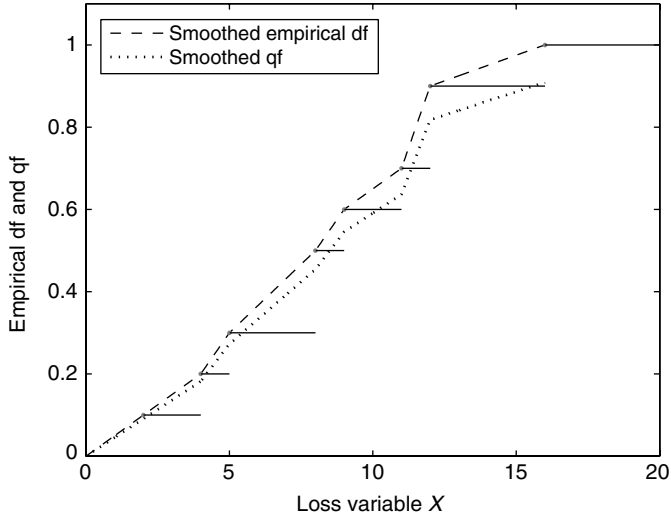

Figure 11.1 Empirical df and qf of Example 11.1

For $\tilde{F}(7.2)$, we first note that $\hat{F}(5) = 0.3$ and $\hat{F}(8) = 0.5$. Thus, using equation (11.6) we have

$$\begin{aligned}\tilde{F}(7.2) &= \left[\frac{7.2-5}{8-5}\right]\hat{F}(8) + \left[\frac{8-7.2}{8-5}\right]\hat{F}(5) \\ &= \left[\frac{2.2}{3}\right](0.5) + \left[\frac{0.8}{3}\right](0.3) \\ &= 0.4467.\end{aligned}$$

For $\hat{x}_{0.75}$, we first note that $g_6 = 7$ and $g_7 = 9$ (note that $y_6 = 11$ and $y_7 = 12$). With $n = 10$, we have

$$\frac{7}{11} \leq 0.75 < \frac{9}{11},$$

so that j defined in equation (11.8) is 6. Hence, using equation (11.7), we compute the smoothed quantile as

$$\hat{x}_{0.75} = \left[\frac{(11)(0.75)-7}{2}\right](12) + \left[\frac{9-(11)(0.75)}{2}\right](11) = 11.625.$$

We estimate the first moment of the censored loss $\mathrm{E}[(X \wedge 11.5)]$ by

$$\begin{aligned}&(0.1)(2) + (0.1)(4) + (0.1)(5) + (0.2)(8) + (0.1)(9) \\ &\quad + (0.1)(11) + (0.3)(11.5) = 8.15,\end{aligned}$$

and the second raw moment of the censored loss $\mathrm{E}\left[(X \wedge 11.5)^2\right]$ by

$$\begin{aligned}&(0.1)(2)^2 + (0.1)(4)^2 + (0.1)(5)^2 + (0.2)(8)^2 + (0.1)(9)^2 \\ &\quad + (0.1)(11)^2 + (0.3)(11.5)^2 = 77.175.\end{aligned}$$

Hence, the estimated variance of the censored loss is

$$77.175 - (8.15)^2 = 10.7525. \qquad \square$$

Note that $\hat{F}(y)$ can also be written as

$$\hat{F}(y) = \frac{Y}{n}, \tag{11.11}$$

where Y is the number of observations less than or equal to y, so that $Y \sim \mathcal{BN}(n, F(y))$. Using results in binomial distribution, we conclude that

$$\mathrm{E}[\hat{F}(y)] = \frac{\mathrm{E}(Y)}{n} = \frac{nF(y)}{n} = F(y), \tag{11.12}$$

and

$$\text{Var}[\hat{F}(y)] = \frac{F(y)[1-F(y)]}{n}. \tag{11.13}$$

Thus, $\hat{F}(y)$ is an unbiased estimator of $F(y)$ and its variance tends to 0 as n tends to infinity, implying $\hat{F}(y)$ is a consistent estimator of $F(y)$. As $F(y)$ is unknown, we may estimate $\text{Var}[\hat{F}(y)]$ by replacing $F(y)$ in equation (11.13) by $\hat{F}(y)$. Then in large samples an approximate $100(1-\alpha)\%$ confidence interval estimate of $F(y)$ may be computed as

$$\hat{F}(y) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{F}(y)[1-\hat{F}(y)]}{n}}. \tag{11.14}$$

A drawback of (11.14) in the estimation of the confidence interval of $F(y)$ is that it may fall outside the interval $(0, 1)$. We will discuss a remedy for this problem later.

11.1.2 Kernel estimation of probability density function

The empirical pf summarizes the data as a discrete distribution. However, if the variable of interest (loss or failure time) is continuous, it is desirable to estimate a pdf. This can be done using the **kernel density estimation method**.

Consider the observation x_i in the sample. The empirical pf assigns a probability mass of $1/n$ to the point x_i. Given that X is continuous, we may wish to *distribute* the probability mass to a neighborhood of x_i rather than assigning it completely to point x_i. Let us assume that we wish to distribute the mass *evenly* in the interval $[x_i - b, x_i + b]$ for a given value of $b > 0$, called the **bandwidth**. To do this, we define a function $f_i(x)$ as follows[2]

$$f_i(x) = \begin{cases} \dfrac{0.5}{b}, & \text{for } x_i - b \le x \le x_i + b, \\ 0, & \text{otherwise.} \end{cases} \tag{11.15}$$

This function is rectangular in shape, with a base of length $2b$ and height of $0.5/b$, so that its area is 1. It may be interpreted as the pdf contributed by the observation x_i. Note that $f_i(x)$ is also the pdf of a $\mathcal{U}(x_i - b, x_i + b)$ variable. Thus, only values of x in the interval $[x_i - b, x_i + b]$ receive contributions from x_i. As

[2] Note that the suffix i in $f_i(x)$ highlights that the function depends on x_i. Also, $2b$ is sometimes called the **window width**.

each x_i contributes a probability mass of $1/n$, the pdf of X may be estimated as

$$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x). \tag{11.16}$$

We now rewrite $f_i(x)$ in equation (11.15) as

$$f_i(x) = \begin{cases} \dfrac{0.5}{b}, & \text{for } -1 \le \dfrac{x - x_i}{b} \le 1, \\ 0, & \text{otherwise,} \end{cases} \tag{11.17}$$

and define

$$K_R(\psi) = \begin{cases} 0.5, & \text{for } -1 \le \psi \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{11.18}$$

Then it can be seen that

$$f_i(x) = \frac{1}{b} K_R(\psi_i), \tag{11.19}$$

where

$$\psi_i = \frac{x - x_i}{b}. \tag{11.20}$$

Note that $K_R(\psi)$ as defined in (11.18) does not depend on the data. However, $K_R(\psi_i)$ with ψ_i defined in equation (11.20) is a function of x (x is the argument of $K_R(\psi_i)$, and x_i and b are the *parameters*). Using equation (11.19), we rewrite equation (11.16) as

$$\tilde{f}(x) = \frac{1}{nb}\sum_{i=1}^{n} K_R(\psi_i). \tag{11.21}$$

$K_R(\psi)$ as defined in equation (11.18) is called the **rectangular** (or **box**, **uniform**) **kernel function**. $\tilde{f}(x)$ defined in equation (11.21) is the estimate of the pdf of X using the rectangular kernel.

It can be seen that $K_R(\psi)$ satisfies the following properties

$$K_R(\psi) \ge 0, \qquad \text{for } -\infty < \psi < \infty, \tag{11.22}$$

and

$$\int_{-\infty}^{\infty} K_R(\psi)\, d\psi = 1. \tag{11.23}$$

Hence, $K_R(\psi)$ is itself the pdf of a random variable taking values over the real line. Indeed any function $K(\psi)$ satisfying equations (11.22) and (11.23) may be called a **kernel function**. Furthermore, the expression in equation (11.21), with $K(\psi)$ replacing $K_R(\psi)$ and ψ_i defined in equation (11.20), is called the **kernel estimate** of the pdf. Apart from the rectangular kernel, two other commonly used kernels are the **triangular kernel**, denoted by $K_T(\psi)$, and the **Gaussian kernel**, denoted by $K_G(\psi)$. The triangular kernel is defined as

$$K_T(\psi) = \begin{cases} 1 - |\psi|, & \text{for } -1 \leq \psi \leq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{11.24}$$

and the Gaussian kernel is given by

$$K_G(\psi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\psi^2}{2}\right), \qquad \text{for } -\infty < \psi < \infty, \tag{11.25}$$

which is just the standard normal density function.

Figure 11.2 presents the plots of the rectangular, triangular, and Gaussian kernels. For the rectangular and triangular kernels, the width of the neighborhood to which the mass of each observation x_i is distributed is $2b$. In contrast, for the Gaussian kernel, the neighborhood is infinite. Regardless of the kernel used, the larger the value of b, the smoother the estimated pdf is. While the rectangular kernel distributes the mass uniformly in the neighborhood, the triangular and Gaussian kernels diminish gradually towards the two ends of the neighborhood. We also note that all three kernel functions are even, such that $K(-\psi) = K(\psi)$. Hence, these kernels are symmetric.

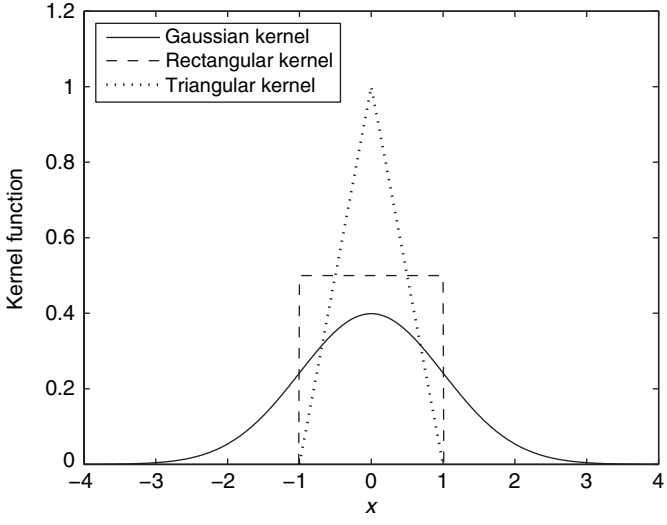

Figure 11.2 Some commonly used kernel functions

Example 11.2 A sample of losses has the following ten observations:

$$5, 6, 6, 7, 8, 8, 10, 12, 13, 15.$$

Determine the kernel estimate of the pdf of the losses using the rectangular kernel for $x = 8.5$ and 11.5 with a bandwidth of 3.

Solution For $x = 8.5$ with $b = 3$, there are six observations within the interval [5.5, 11.5]. From equation (11.21) we have

$$\tilde{f}(8.5) = \frac{1}{(10)(3)}(6)(0.5) = \frac{1}{10}.$$

Similarly, there are three observations in the interval [8.5, 14.5], so that

$$\tilde{f}(11.5) = \frac{1}{(10)(3)}(3)(0.5) = \frac{1}{20}. \qquad \square$$

Example 11.3 A sample of losses has the following 40 observations:

$$15, 16, 16, 17, 19, 19, 19, 23, 24, 27, 28, 28, 28, 28, 31, 34, 34, 34, 36, 36$$
$$37, 40, 41, 41, 43, 46, 46, 46, 47, 47, 49, 50, 50, 53, 54, 54, 59, 61, 63, 64.$$

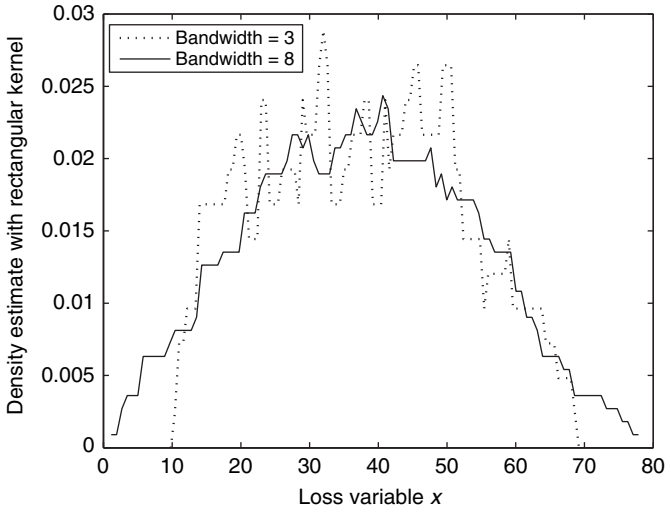

Figure 11.3 Estimated pdf of Example 11.3 using rectangular kernel

Estimate the pdf of the loss distribution using the rectangular kernel and Gaussian kernel, with bandwidth of 3 and 8.

Solution The rectangular kernel estimates are plotted in Figure 11.3, and the Gaussian kernel estimates are plotted in Figure 11.4. It can be seen that the Gaussian kernel estimates are smoother than the rectangular kernel estimates. Also, the kernels with bandwidth of 8 provide smoother pdfs than kernels with bandwidths of 3. For the Gaussian kernel with bandwidth of 8, the estimated pdf extends to values of negative losses. This is undesirable and is due to the fact that the Gaussian kernel has infinite support and the bandwidth is wide.

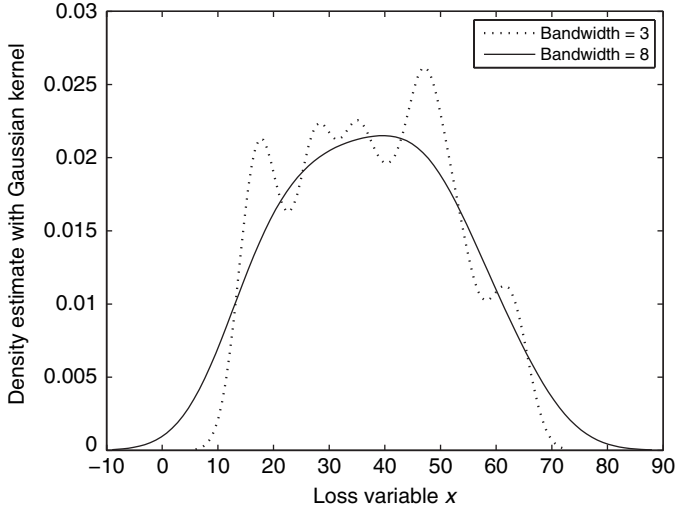

Figure 11.4 Estimated pdf of Example 11.3 using Gaussian kernel

□

The estimated kernel df, denoted by $\tilde{F}(x)$, can be computed by integrating the estimated kernel pdf. From equation (11.21), we have

$$\begin{aligned}\tilde{F}(x) &= \int_0^x \tilde{f}(x)\,dx \\ &= \frac{1}{nb}\sum_{i=1}^{n}\int_0^x K\left(\frac{x-x_i}{b}\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-x_i/b}^{(x-x_i)/b} K(\psi)\,d\psi, \end{aligned} \tag{11.26}$$

where $K(\psi)$ is any well-defined kernel function.

11.2 Estimation with incomplete individual data

We now consider samples with incomplete data, i.e. observations with left truncation or right censoring. Our focus is on the nonparametric estimation of the survival function $S(x)$. Suppose the data consist of n observations with m distinct values $0 < y_1 < \cdots < y_m$, with corresponding risk sets $r_1, \cdots, r_m$, and counts of repetition $w_1, \cdots, w_m$ (numbers of times $y_1, \cdots, y_m$ are in the sample). We let $y_0 \equiv 0$, so that $S(y_0) = S(0) = 1$. We shall provide heuristic derivations of the Kaplan–Meier (product-limit) estimator and the Nelson–Aalen estimator of the sf.[3]

11.2.1 Kaplan–Meier (product-limit) estimator

We consider the estimation of $S(y_j) = \Pr(X > y_j)$, for $j = 1, \cdots, m$. Using the rule of conditional probability, we have

$$\begin{aligned} S(y_j) &= \Pr(X > y_1) \Pr(X > y_2 \,|\, X > y_1) \cdots \Pr(X > y_j \,|\, X > y_{j-1}) \\ &= \Pr(X > y_1) \prod_{h=2}^{j} \Pr(X > y_h \,|\, X > y_{h-1}). \end{aligned} \tag{11.27}$$

As the risk set for y_1 is r_1, and w_1 observations are found to have value y_1, $\Pr(X > y_1)$ can be estimated by

$$\widehat{\Pr}(X > y_1) = 1 - \frac{w_1}{r_1}. \tag{11.28}$$

Likewise, the risk set for y_h is r_h, and w_h individuals are observed to have value y_h. Thus, $\Pr(X > y_h \,|\, X > y_{h-1})$ can be estimated by

$$\widehat{\Pr}(X > y_h \,|\, X > y_{h-1}) = 1 - \frac{w_h}{r_h}, \quad \text{for } h = 2, \cdots, m. \tag{11.29}$$

Hence, we may estimate $S(y_j)$ by

$$\begin{aligned} \hat{S}(y_j) &= \widehat{\Pr}(X > y_1) \prod_{h=2}^{j} \widehat{\Pr}(X > y_h \,|\, X > y_{h-1}) \\ &= \prod_{h=1}^{j} \left(1 - \frac{w_h}{r_h}\right). \end{aligned} \tag{11.30}$$

[3] Rigorous derivations of these estimators are beyond the scope of this book. Interested readers may refer to London (1988) and the references therein.

For values of y with $y_j \le y < y_{j+1}$, we have $\hat{S}(y) = \hat{S}(y_j)$, as no observation is found in the sample in the interval $(y_j, y]$.

We now summarize the above arguments and define the Kaplan–Meier estimator, denoted by $\hat{S}_K(y)$, as follows

$$\hat{S}_K(y) = \begin{cases} 1, & \text{for } 0 < y < y_1, \\ \prod_{h=1}^{j} \left(1 - \frac{w_h}{r_h}\right), & \text{for } y_j \le y < y_{j+1}, j = 1, \cdots, m-1, \\ \prod_{h=1}^{m} \left(1 - \frac{w_h}{r_h}\right), & \text{for } y_m \le y. \end{cases} \tag{11.31}$$

Note that if $w_m = r_m$, then $\hat{S}_K(y) = 0$ for $y_m \le y$. However, if $w_m < r_m$ (i.e. the largest observation is a censored observation and not a failure time), then $\hat{S}_K(y_m) > 0$. There are several ways to compute $\hat{S}_K(y)$ for $y > y_m$. First, we may adopt the definition in equation (11.31). This method, however, is inconsistent with the property of the sf that $S(y) \to 0$ as $y \to \infty$. Second, we may let $\hat{S}_K(y) = 0$ for $y > y_m$. Third, we may allow $\hat{S}_K(y)$ to decay geometrically to 0 by defining

$$\hat{S}_K(y) = \hat{S}_K(y_m)^{\frac{y}{y_m}}, \quad \text{for } y > y_m. \tag{11.32}$$

Example 11.4 Refer to the loss claims in Example 10.7. Determine the Kaplan–Meier estimate of the sf.

Solution Using the data compiled in Table 10.7, we present the calculation of the Kaplan–Meier estimates in Table 11.1. For $y > 18$, the sf may be alternatively computed as 0 or $(0.2888)^{\frac{y}{18}}$. □

Example 11.5 Refer to the loss claims in Example 10.8. Determine the Kaplan–Meier estimate of the sf.

Solution As all policies are with a deductible of 4, we can only estimate the conditional sf $S(y \mid y > 4)$. Also, as there is a maximum covered loss of 20 for all policies, we can only estimate the conditional sf up to $S(20 \mid y > 4)$. Using the data compiled in Table 10.8, the Kaplan–Meier estimates are summarized in Table 11.2. □

For complete data, $d_i = 0$ for $i = 1, \cdots, n$, and there is no u_i value. Thus, $r_1 = n$ and $r_j = r_{j-1} - w_{j-1}$, for $j = 2, \cdots, m$ (see equation (10.8)). Therefore, when the Kaplan–Meier estimate is applied to complete data, we have, for

Table 11.1. *Kaplan–Meier estimates of Example 11.4*

Interval containing y	$\hat{S}_K(y)$
$(0, 6)$	1
$[6, 8)$	$1 - \frac{1}{20} = 0.95$
$[8, 9)$	$0.95\left[1 - \frac{3}{19}\right] = 0.8$
$[9, 10)$	$0.8\left[1 - \frac{1}{16}\right] = 0.75$
$[10, 12)$	$0.75\left[1 - \frac{2}{15}\right] = 0.65$
$[12, 13)$	$0.65\left[1 - \frac{2}{13}\right] = 0.55$
$[13, 14)$	$0.55\left[1 - \frac{1}{10}\right] = 0.495$
$[14, 15)$	$0.495\left[1 - \frac{1}{9}\right] = 0.44$
$[15, 18)$	$0.44\left[1 - \frac{1}{8}\right] = 0.385$
$[18, \infty)$	$0.385\left[1 - \frac{1}{4}\right] = 0.2888$

$j = 1, \cdots, m-1$

$$\hat{S}_K(y_j) = \prod_{h=1}^{j}\left(1 - \frac{w_h}{r_h}\right) = \prod_{h=1}^{j}\frac{r_{h+1}}{r_h} = \frac{r_{j+1}}{r_1}. \tag{11.33}$$

As $r_{j+1} = r_1 - \sum_{h=1}^{j} w_h = r_1 - g_j$, we conclude from equation (11.2) that

$$\hat{S}_K(y_j) = \frac{r_1 - g_j}{r_1} = 1 - \frac{g_j}{n} = 1 - \hat{F}(y_j). \tag{11.34}$$

Table 11.2. *Kaplan–Meier estimates of Example 11.5*

Interval containing y	$\hat{S}_K(y \mid y > 4)$
(4, 5)	1
[5, 7)	0.9333
[7, 8)	0.8667
[8, 10)	0.8000
[10, 16)	0.6667
[16, 17)	0.6000
[17, 19)	0.4000
[19, 20)	0.3333
20	0.2667

Furthermore, $\hat{S}_K(y_m) = 0 = 1 - \hat{F}(y_m)$. Thus, the Kaplan–Meier estimate gives the same result as the empirical df when all observations are complete.

We now consider the mean and the variance of the Kaplan–Meier estimator. Our derivation is heuristic and is based on the assumption that the risk sets r_j and the values of the observed loss or failure-time y_j are known. Specifically, we denote $\mathcal{C}$ as the information set $\{y_1, \cdots, y_m, r_1, \cdots, r_m\}$. We also denote W_j as the random variable representing w_j, for $j = 1, \cdots, m$. Now given $\mathcal{C}$, the conditional probability of observing a loss or failure time of value y_j is

$$\Pr(X \leq y_j \mid X > y_{j-1}) = \frac{S(y_{j-1}) - S(y_j)}{S(y_{j-1})} = 1 - S_j, \qquad \text{for } j = 1, \cdots, m, \tag{11.35}$$

where S_j is defined as

$$S_j = \frac{S(y_j)}{S(y_{j-1})}. \tag{11.36}$$

Hence, given $\mathcal{C}$, $W_1, \cdots, W_m$ are distributed as binomial random variables. Specifically

$$W_j \mid \mathcal{C} \sim \mathcal{BN}(r_j, 1 - S_j), \qquad \text{for } j = 1, \cdots, m, \tag{11.37}$$

and

$$r_j - W_j \mid \mathcal{C} \sim \mathcal{BN}(r_j, S_j), \qquad \text{for } j = 1, \cdots, m. \tag{11.38}$$

Thus, we have the following results

$$\mathrm{E}\left[\frac{r_j - W_j}{r_j} \mid \mathcal{C}\right] = S_j \tag{11.39}$$

and

$$\begin{aligned}\mathrm{E}\left[\left(\frac{r_j - W_j}{r_j}\right)^2 \mid \mathcal{C}\right] &= \frac{S_j(1-S_j)}{r_j} + S_j^2 \\ &= S_j^2\left[\frac{1-S_j}{S_j r_j} + 1\right].\end{aligned} \tag{11.40}$$

From equation (11.31), assuming the independence of $W_1, \cdots, W_m$ given $\mathcal{C}$, the mean of the Kaplan–Meier estimator is

$$\begin{aligned}\mathrm{E}\left[\hat{S}_K(y_j) \mid \mathcal{C}\right] &= \mathrm{E}\left[\prod_{h=1}^{j}\left(1 - \frac{W_h}{r_h}\right) \mid \mathcal{C}\right] \\ &= \prod_{h=1}^{j}\mathrm{E}\left[\left(1 - \frac{W_h}{r_h}\right) \mid \mathcal{C}\right] \\ &= \prod_{h=1}^{j} S_h \\ &= \prod_{h=1}^{j}\frac{S(y_h)}{S(y_{h-1})} \\ &= S(y_j),\end{aligned} \tag{11.41}$$

so that $\hat{S}_K(y_j)$ is unbiased for $S(y_j)$. The variance of $\hat{S}_K(y_j)$ is

$$\begin{aligned}\mathrm{Var}\left[\hat{S}_K(y_j) \mid \mathcal{C}\right] &= \mathrm{E}\left[\left\{\prod_{h=1}^{j}\left(1 - \frac{W_h}{r_h}\right)\right\}^2 \mid \mathcal{C}\right] - \left\{\mathrm{E}\left[\prod_{h=1}^{j}\left(1 - \frac{W_h}{r_h}\right) \mid \mathcal{C}\right]\right\}^2 \\ &= \prod_{h=1}^{j}\mathrm{E}\left[\left(1 - \frac{W_h}{r_h}\right)^2 \mid \mathcal{C}\right] - \left\{\prod_{h=1}^{j}\mathrm{E}\left[\left(1 - \frac{W_h}{r_h}\right) \mid \mathcal{C}\right]\right\}^2\end{aligned}$$

$$= \prod_{h=1}^{j} S_h^2 \left[\frac{1 - S_h}{S_h r_h} + 1 \right] - \left[\prod_{h=1}^{j} S_h \right]^2$$

$$= \left[\prod_{h=1}^{j} S_h \right]^2 \left\{ \prod_{h=1}^{j} \left[\frac{1 - S_h}{S_h r_h} + 1 \right] - 1 \right\}$$

$$= \left[S(y_j) \right]^2 \left\{ \prod_{h=1}^{j} \left[\frac{1 - S_h}{S_h r_h} + 1 \right] - 1 \right\}. \tag{11.42}$$

Now using the approximation

$$\prod_{h=1}^{j} \left[\frac{1 - S_h}{S_h r_h} + 1 \right] \simeq 1 + \sum_{h=1}^{j} \frac{1 - S_h}{S_h r_h}, \tag{11.43}$$

equation (11.42) can be written as

$$\operatorname{Var}\left[\hat{S}_K(y_j) \mid \mathcal{C} \right] \simeq \left[S(y_j) \right]^2 \left(\sum_{h=1}^{j} \frac{1 - S_h}{S_h r_h} \right). \tag{11.44}$$

Estimating $S(y_j)$ by $\hat{S}_K(y_j)$ and S_h by $(r_h - w_h)/r_h$, the variance estimate of the Kaplan–Meier estimator can be computed as

$$\widehat{\operatorname{Var}}\left[\hat{S}_K(y_j) \mid \mathcal{C} \right] \simeq \left[\hat{S}_K(y_j) \right]^2 \left(\sum_{h=1}^{j} \frac{w_h}{r_h(r_h - w_h)} \right), \tag{11.45}$$

which is called the **Greenwood approximation** for the variance of the Kaplan–Meier estimator.

In the special case where all observations are complete, the Greenwood approximation becomes

$$\widehat{\operatorname{Var}}\left[\hat{S}_K(y_j) \mid \mathcal{C} \right] = \left[\frac{n - g_j}{n} \right]^2 \left[\sum_{h=1}^{j} \left(\frac{1}{r_h - w_h} - \frac{1}{r_h} \right) \right]$$

$$= \left[\frac{n - g_j}{n} \right]^2 \left[\frac{1}{r_j - w_j} - \frac{1}{r_1} \right]$$

$$= \left[\frac{n-g_j}{n}\right]^2 \left[\frac{g_j}{n(n-g_j)}\right]$$
$$= \frac{g_j(n-g_j)}{n^3}$$
$$= \frac{\hat{F}(y_j)\left[1-\hat{F}(y_j)\right]}{n}, \tag{11.46}$$

which is the usual estimate of the variance of the empirical df (see equation (11.13)).

Example 11.6 Refer to the loss claims in Examples 10.7 and 11.4. Determine the approximate variance of $\hat{S}_K(10.5)$ and the 95% confidence interval of $S_K(10.5)$.

Solution From Table 11.1, we can see that Kaplan–Meier estimate of $S_K(10.5)$ is 0.65. The Greenwood approximate for the variance of $\hat{S}_K(10.5)$ is

$$(0.65)^2\left[\frac{1}{(20)(19)}+\frac{3}{(19)(16)}+\frac{1}{(16)(15)}+\frac{2}{(15)(13)}\right]=0.0114.$$

Thus, the estimate of the standard deviation of $\hat{S}_K(10.5)$ is $\sqrt{0.0114}=0.1067$, and, assuming the normality of $\hat{S}_K(10.5)$, the 95% confidence interval of $S_K(10.5)$ is

$$0.65 \pm (1.96)(0.1067) = (0.4410,\ 0.8590). \qquad \square$$

The above example uses the normal approximation for the distribution of $\hat{S}_K(y_j)$ to compute the confidence interval of $S(y_j)$. This is sometimes called the **linear confidence interval**. A disadvantage of this estimate is that the computed confidence interval may fall outside the range $(0, 1)$. This drawback can be remedied by considering a transformation of the survival function. We first define the transformation $\zeta(\cdot)$ by

$$\zeta(x) = \log\left[-\log(x)\right], \tag{11.47}$$

and let

$$\hat{\zeta} = \zeta(\hat{S}(y)) = \log[-\log(\hat{S}(y))], \tag{11.48}$$

where $\hat{S}(y)$ is an estimate of the sf $S(y)$ for a given y. Using the delta method (see Appendix A.19), the variance of $\hat{\zeta}$ can be approximated by

$$\widehat{\text{Var}}(\hat{\zeta}) = [\zeta'(\hat{S}(y))]^2\, \widehat{\text{Var}}[\hat{S}(y)], \tag{11.49}$$

where $\zeta'(\hat{S}(y))$ is the first derivative of $\zeta(\cdot)$ evaluated at $\hat{S}(y)$. Now

$$\frac{d\zeta(x)}{dx} = \left[\frac{1}{-\log x}\right]\left[-\frac{1}{x}\right] = \frac{1}{x\log x}, \tag{11.50}$$

so that we can estimate $\text{Var}(\hat{\zeta})$ by

$$\widehat{\text{Var}}(\hat{\zeta}) = \frac{\widehat{\text{Var}}[\hat{S}(y)]}{[\hat{S}(y)\log\hat{S}(y)]^2} \equiv \hat{V}(y). \tag{11.51}$$

A $100(1-\alpha)\%$ confidence interval of $\zeta(S(y))$ can be computed as

$$\zeta(\hat{S}(y)) \pm z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(y)}. \tag{11.52}$$

Taking the inverse transform of Equation (11.47), we have

$$S(y) = \exp\{-\exp[\zeta(S(y))]\}. \tag{11.53}$$

Thus, a $100(1-\alpha)\%$ confidence interval of $S(y)$ can be computed by taking the inverse transform of equation (11.52) to obtain[4]

$$\left(\exp\left[-\exp\left\{\zeta(\hat{S}(y)) + z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(y)}\right\}\right],\right.$$
$$\left.\exp\left[-\exp\left\{\zeta(\hat{S}(y)) \quad -z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(y)}\right\}\right]\right), \tag{11.54}$$

which reduces to

$$\left(\hat{S}(y)^U, \hat{S}(y)^{\frac{1}{U}}\right), \tag{11.55}$$

where

$$U = \exp\left[z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(y)}\right]. \tag{11.56}$$

This is known as the **logarithmic transformation method**.

Example 11.7 Refer to the loss claims in Examples 10.8 and 11.5. Determine the approximate variance of $\hat{S}_K(7)$ and the 95% confidence interval of $S(7)$.

Solution From Table 11.2, we have $\hat{S}_K(7) = 0.8667$. The Greenwood approximate variance of $\hat{S}_K(7)$ is

$$(0.8667)^2\left[\frac{1}{(15)(14)} + \frac{1}{(14)(13)}\right] = 0.0077.$$

[4] As $\zeta(\cdot)$ is a monotonic *decreasing* function, the upper and lower limits of $\zeta(S(y))$ in equation (11.52) are reversed to obtain the limits in equation (11.54).

Using normal approximation to the distribution of $\hat{S}_K(7)$, the 95% confidence interval of $S(7)$ is

$$0.8667 \pm 1.96\sqrt{0.0077} = (0.6947,\ 1.0387).$$

Thus, the upper limit exceeds 1, which is undesirable. To apply the logarithmic transformation method, we compute $\hat{V}(7)$ in equation (11.51) to obtain

$$\hat{V}(7) = \frac{0.0077}{[0.8667 \log(0.8667)]^2} = 0.5011,$$

so that U in equation (11.56) is

$$\exp\left[(1.96)\sqrt{0.5011}\right] = 4.0048.$$

From (11.55), the 95% confidence interval of $S(7)$ is

$$\{(0.8667)^{4.0048}, (0.8667)^{\frac{1}{4.0048}}\} = (0.5639,\ 0.9649),$$

which is within the range $(0, 1)$.

We finally remark that as all policies in this example have a deductible of 4. The sf of interest is conditional on the loss exceeding 4. □

11.2.2 Nelson–Aalen estimator

Denoting the cumulative hazard function defined in equation (2.8) by $H(y)$, so that

$$H(y) = \int_0^y h(y)\, dy, \tag{11.57}$$

where $h(y)$ is the hazard function, we have

$$S(y) = \exp\left[-H(y)\right]. \tag{11.58}$$

Thus

$$H(y) = -\log\left[S(y)\right]. \tag{11.59}$$

If we use $\hat{S}_K(y)$ to estimate $S(y)$ for $y_j \le y < y_{j+1}$, an estimate of the cumulative hazard function can be computed as

$$\begin{aligned}\hat{H}(y) &= -\log\left[\hat{S}_K(y)\right] \\ &= -\log\left[\prod_{h=1}^{j}\left(1-\frac{w_h}{r_h}\right)\right] \\ &= -\sum_{h=1}^{j}\log\left(1-\frac{w_h}{r_h}\right). \end{aligned} \tag{11.60}$$

Using the approximation

$$-\log\left(1-\frac{w_h}{r_h}\right) \simeq \frac{w_h}{r_h}, \tag{11.61}$$

we obtain $\hat{H}(y)$ as

$$\hat{H}(y) = \sum_{h=1}^{j}\frac{w_h}{r_h}, \tag{11.62}$$

which is the **Nelson–Aalen estimate of the cumulative hazard function**. We complete its formula as follows

$$\hat{H}(y) = \begin{cases} 0, & \text{for } 0 < y < y_1, \\ \displaystyle\sum_{h=1}^{j}\frac{w_h}{r_h}, & \text{for } y_j \le y < y_{j+1}, j = 1, \cdots, m-1, \\ \displaystyle\sum_{h=1}^{m}\frac{w_h}{r_h}, & \text{for } y_m \le y. \end{cases} \tag{11.63}$$

We now construct an alternative estimator of $S(y)$ using equation (11.58). This is called the **Nelson–Aalen estimator of the survival function**, denoted by $\hat{S}_N(y)$. The formula of $\hat{S}_N(y)$ is

$$\hat{S}_N(y) = \begin{cases} 1, & \text{for } 0 < y < y_1, \\ \exp\left(-\displaystyle\sum_{h=1}^{j}\frac{w_h}{r_h}\right), & \text{for } y_j \le y < y_{j+1}, j = 1, \cdots, m-1, \\ \exp\left(-\displaystyle\sum_{h=1}^{m}\frac{w_h}{r_h}\right), & \text{for } y_m \le y. \end{cases} \tag{11.64}$$

For $y > y_m$, we may also compute $\hat{S}_N(y)$ as 0 or $[\hat{S}_N(y_m)]^{\frac{y}{y_m}}$.

The Nelson–Aalen estimator is easy to compute. In the case of complete data, with one observation at each point y_j, we have $w_h = 1$ and $r_h = n - h + 1$ for $h = 1, \cdots, n$, so that

$$\hat{S}_N(y_j) = \exp\left(-\sum_{h=1}^{j} \frac{1}{n-h+1}\right). \tag{11.65}$$

Example 11.8 Refer to the loss claims in Examples 11.6 and 11.7. Compute the Nelson–Aalen estimates of the sf.

Solution For the data in Example 11.6, the Nelson–Aalen estimate of $S(10.5)$ is

$$\hat{S}_N(10.5) = \exp\left(-\frac{1}{20} - \frac{3}{19} - \frac{1}{16} - \frac{2}{15}\right) = \exp(-0.4037) = 0.6678.$$

This may be compared against $\hat{S}_K(10.5) = 0.65$ from Example 11.5. Likewise, the Nelson–Aalen estimate of $S(7)$ in Example 11.7 is

$$\hat{S}_N(7) = \exp\left(-\frac{1}{15} - \frac{1}{14}\right) = \exp(-0.1381) = 0.8710,$$

which may be compared against $\hat{S}_K(7) = 0.8667$. □

To derive an approximate formula for the variance of $\hat{H}(y)$, we assume the conditional distribution of W_h, given the information set $\mathcal{C}$, to be Poisson. Thus, we estimate $\text{Var}(W_h)$ by w_h. An estimate of $\text{Var}[\hat{H}(y_j)]$ can then be computed as

$$\widehat{\text{Var}}[\hat{H}(y_j)] = \widehat{\text{Var}}\left(\sum_{h=1}^{j} \frac{W_h}{r_h}\right) = \sum_{h=1}^{j} \frac{\widehat{\text{Var}}(W_h)}{r_h^2} = \sum_{h=1}^{j} \frac{w_h}{r_h^2}. \tag{11.66}$$

Hence, a $100(1-\alpha)\%$ confidence interval of $H(y_j)$, assuming normal approximation, is given by

$$\hat{H}(y_j) \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}[\hat{H}(y_j)]}. \tag{11.67}$$

To ensure the lower limit of the confidence interval of $H(y_j)$ to be positive,[5] we consider the transformation

$$\zeta(x) = \log x, \tag{11.68}$$

[5] Note that the cumulative hazard function takes values in the range $(0, \infty)$.

and define

$$\hat{\zeta} = \zeta(\hat{H}(y_j)) = \log[\hat{H}(y_j)]. \tag{11.69}$$

Thus, using the delta method, an approximate variance of $\hat{\zeta}$ is

$$\widehat{\text{Var}}(\hat{\zeta}) = \frac{\widehat{\text{Var}}[\hat{H}(y_j)]}{[\hat{H}(y_j)]^2}, \tag{11.70}$$

and a $100(1-\alpha)\%$ approximate confidence interval of $\zeta(H(y_j))$ can be obtained as

$$\zeta(\hat{H}(y_j)) \pm z_{1-\frac{\alpha}{2}} \frac{\sqrt{\widehat{\text{Var}}[\hat{H}(y_j)]}}{\hat{H}(y_j)}. \tag{11.71}$$

Taking the inverse transformation of $\zeta(\cdot)$, a $100(1-\alpha)\%$ approximate confidence interval of $H(y_j)$ is

$$\left(\hat{H}(y_j)\left(\frac{1}{U}\right), \hat{H}(y_j)U\right), \tag{11.72}$$

where

$$U = \exp\left[z_{1-\frac{\alpha}{2}} \frac{\sqrt{\widehat{\text{Var}}[\hat{H}(y_j)]}}{\hat{H}(y_j)}\right]. \tag{11.73}$$

Example 11.9 Refer to Examples 11.6 and 11.7. Compute the 95% confidence intervals of the Nelson–Aalen estimates of the cumulative hazard function.

Solution For the data in Example 11.6, the Nelson–Aalen estimate of $H(10.5)$ is 0.4037 (see Example 11.8). From equation (11.66), the estimated variance of $\hat{H}(10.5)$ is

$$\frac{1}{(20)^2} + \frac{3}{(19)^2} + \frac{1}{(16)^2} + \frac{2}{(15)^2} = 0.0236.$$

Thus, the 95% linear confidence interval of $H(10.5)$ is

$$0.4037 \pm 1.96\sqrt{0.0236} = (0.1026, 0.7048).$$

Using the logarithmic transformation, U in equation (11.73) is

$$\exp\left[(1.96)\frac{\sqrt{0.0236}}{0.4037}\right] = 2.1084,$$

and the 95% confidence interval of $H(10.5)$ using the logarithmic transformation method is

$$\left\{\frac{0.4037}{2.1084},\ (0.4037)(2.1084)\right\} = (0.1914,\ 0.8512)\,.$$

For Example 11.7, the estimated variance of $\hat{H}(7)$ is

$$\frac{1}{(15)^2} + \frac{1}{(14)^2} = 0.0095.$$

The 95% linear confidence interval of $H(7)$ is

$$0.1381 \pm 1.96\sqrt{0.0095} = (-0.0534,\ 0.3296)\,.$$

Thus, the lower limit falls below zero. Using the logarithmic transformation method, we have $U = 4.0015$, so that the 95% confidence interval of $H(7)$ is $(0.0345, 0.5526)$. □

11.3 Estimation with grouped data

We now consider data that are grouped into intervals. Following the notations developed in Section 10.2.4, we assume that the values of the failure-time or loss data x_i are grouped into k intervals: $(c_0, c_1], (c_1, c_2], \cdots, (c_{k-1}, c_k]$, where $0 \le c_0 < c_1 < \cdots < c_k$. We first consider the case where the data are complete, with no truncation or censoring. Let there be n observations of x in the sample, with n_j observations in the interval $(c_{j-1}, c_j]$, so that $\sum_{j=1}^{k} n_j = n$. Assuming the observations within each interval are uniformly distributed, the empirical pdf of the failure-time or loss variable X can be written as

$$\hat{f}(x) = \sum_{j=1}^{k} p_j f_j(x), \tag{11.74}$$

where

$$p_j = \frac{n_j}{n} \tag{11.75}$$

and

$$f_j(x) = \begin{cases} \dfrac{1}{c_j - c_{j-1}}, & \text{for } c_{j-1} < x \le c_j, \\ 0, & \text{otherwise.} \end{cases} \tag{11.76}$$

Thus, $\hat{f}(x)$ is the pdf of a mixture distribution. To compute the moments of X we note that

$$\int_0^\infty f_j(x)x^r\,dx = \frac{1}{c_j - c_{j-1}}\int_{c_{j-1}}^{c_j} x^r\,dx = \frac{c_j^{r+1} - c_{j-1}^{r+1}}{(r+1)(c_j - c_{j-1})}. \tag{11.77}$$

Hence, the mean of the empirical pdf is

$$\mathrm{E}(X) = \sum_{j=1}^{k} p_j\left[\frac{c_j^2 - c_{j-1}^2}{2(c_j - c_{j-1})}\right] = \sum_{j=1}^{k}\frac{n_j}{n}\left[\frac{c_j + c_{j-1}}{2}\right], \tag{11.78}$$

and its rth raw moment is

$$\mathrm{E}(X^r) = \sum_{j=1}^{k}\frac{n_j}{n}\left[\frac{c_j^{r+1} - c_{j-1}^{r+1}}{(r+1)(c_j - c_{j-1})}\right]. \tag{11.79}$$

The censored moments are more complex. Suppose it is desired to compute $\mathrm{E}[(X \wedge u)^r]$. First, we consider the case where $u = c_h$ for some $h = 1, \cdots, k-1$, i.e. u is the end point of an interval. Then the rth raw moment is

$$E\left[(X \wedge c_h)^r\right] = \sum_{j=1}^{h}\frac{n_j}{n}\left[\frac{c_j^{r+1} - c_{j-1}^{r+1}}{(r+1)(c_j - c_{j-1})}\right] + c_h^r\sum_{j=h+1}^{k}\frac{n_j}{n}. \tag{11.80}$$

However, if $c_{h-1} < u < c_h$, for some $h = 1, \cdots, k$, then we have

$$\begin{aligned} E\left[(X \wedge u)^r\right] = &\sum_{j=1}^{h-1}\frac{n_j}{n}\left[\frac{c_j^{r+1} - c_{j-1}^{r+1}}{(r+1)(c_j - c_{j-1})}\right] + u^r\sum_{j=h+1}^{k}\frac{n_j}{n} \\ &+ \frac{n_h}{n(c_h - c_{h-1})}\left[\frac{u^{r+1} - c_{h-1}^{r+1}}{r+1} + u^r(c_h - u)\right]. \end{aligned} \tag{11.81}$$

The last term in the above equation arises from the assumption that $(c_h - u)/(c_h - c_{h-1})$ of the observations in the interval (c_{h-1}, c_h) are larger than or equal to u.

The empirical df at the upper end of each interval is easy to compute. Specifically, we have

$$\hat{F}(c_j) = \frac{1}{n}\sum_{h=1}^{j} n_h, \quad \text{for } j = 1, \cdots, k. \tag{11.82}$$

For other values of x, we use the interpolation formula given in equation (11.6), i.e.

$$\hat{F}(x) = \frac{x - c_j}{c_{j+1} - c_j} \hat{F}(c_{j+1}) + \frac{c_{j+1} - x}{c_{j+1} - c_j} \hat{F}(c_j), \tag{11.83}$$

where $c_j \le x < c_{j+1}$, for some $j = 0, 1, \cdots, k-1$, with $\hat{F}(c_0) = 0$. $\hat{F}(x)$ is also called the **ogive**. The quantiles of X can then be determined by taking the inverse of the smoothed empirical df $\hat{F}(x)$.

When the observations are incomplete, we may use the Kaplan–Meier and Nelson–Aalen methods to estimate the sf. Using equations (10.10) or (10.11), we calculate the risk sets R_j and the number of failures or losses V_j in the interval $(c_{j-1}, c_j]$. These numbers are taken as the risk sets and observed failures or losses at points c_j. $\hat{S}_K(c_j)$ and $\hat{S}_N(c_j)$ may then be computed using equations (11.31) and (11.64), respectively, with R_h replacing r_h and V_h replacing w_h. Finally, we use the interpolation method as in equation (11.83) to estimate other values of the sf $S(x)$ for x not at the end points c_j. The df is then estimated as $\hat{F}(x) = 1 - \hat{S}(x)$.

Example 11.10 Refer to Example 10.9. Compute the Kaplan–Meier estimates of the sf and df.

Solution The computations are presented in Table 11.3. The last two columns summarize the estimated sf and df, respectively, at points c_j. The interpolated estimate of the df is plotted in Figure 11.5. For $x > 20$, the geometric decay

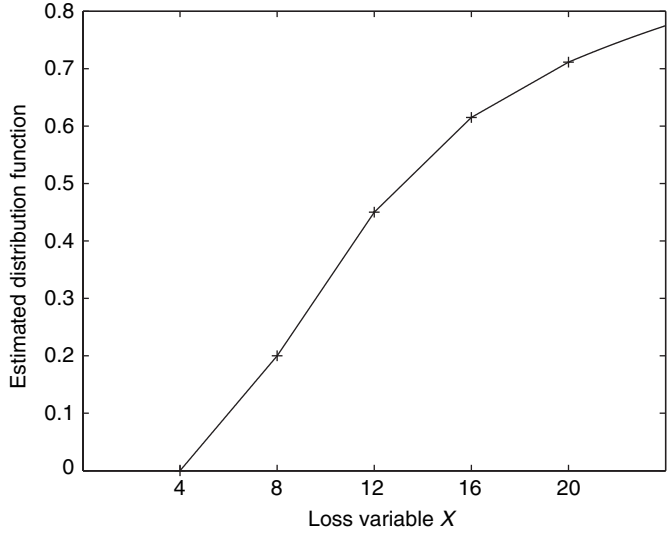

Figure 11.5 Estimated df of Example 11.10 for grouped incomplete data

Table 11.3. *Results of Example 11.10*

Interval	c_j	V_j	R_j	$\hat{S}_K(c_j)$	$\hat{F}_K(c_j)$
(0, 4]	4	0	13	1	0
(4, 8]	8	4	20	$1\left(\frac{16}{20}\right) = 0.8$	0.2
(8, 12]	12	5	16	$0.8\left(\frac{11}{16}\right) = 0.55$	0.45
(12, 16]	16	3	10	$0.55\left(\frac{7}{10}\right) = 0.385$	0.615
(16, 20]	20	1	4	$0.385\left(\frac{3}{4}\right) = 0.2888$	0.7113

method is used to compute the estimated sf (see equation (11.32)) and then the estimated df. □

11.4 Excel computation notes

Excel has functions for the computation of higher-order moments, such as `SKEW` for the computation of skewness and `KURT` for the computation of kurtosis. These functions, however, apply corrections for the degrees of freedom and do not produce the standardized skewness and kurtosis of the empirical distribution. Furthermore, they work only for ungrouped individual observations.

Excel also provides the function `PERCENTILE` for the computation of the precentile of a data set of individual observations. The method used, however, is different from that given in equations (11.7) through (11.10).

For the computation of the kernel estimates of the pdf and the Kaplan–Meier estimates of the survival function, more advanced statistical computational tools (e.g., Matlab) are required.

11.5 Summary and discussions

We have discussed methods of estimating the df and sf of failure-time and loss data, as well as their moments. Cases of complete versus incomplete observations (with left truncation and right censoring), and individual versus

grouped data are discussed. The focus is on nonparametric methods in which no parametric functional forms of the distributions are assumed. For incomplete data, the Kaplan–Meier and Nelson–Aalen estimates are two convenient methods to estimate the sf, both for individual as well as grouped data. We have discussed methods of estimating the variance of the estimated sf and df. The confidence intervals of the sf and df based on the linear confidence interval method may give rise to undesirable results with estimated values lying outside the theoretical range. This shortcoming may be remedied using suitable logarithmic transformations.

Exercises

11.1 A sample from a distribution X has the following 16 observations:

$$6, 8, 8, 11, 13, 14, 14, 20, 21, 26, 27, 27, 27, 30, 32, 33.$$

(a) Determine the smoothed empirical distribution functions $\tilde{F}(15)$ and $\tilde{F}(27)$.
(b) Determine the smoothed quantiles $\hat{x}_{0.25}$ and $\hat{x}_{0.75}$, and hence the interquartile range $\hat{x}_{0.75} - \hat{x}_{0.25}$.
(c) Compute the mean, the variance, and the skewness (see equation (A.28)) of the empirical distribution.
(d) Compute $\text{Var}[(X \wedge 27)]$ and $\text{Var}[(X \wedge 31.5)]$ of the empirical distribution.
(e) If $\tilde{X} = X \mid X > 10$, estimate $\Pr(\tilde{X} \leq 25)$ and $\text{E}[(\tilde{X} \wedge 20.5)]$.

11.2 A sample of ground-up losses X has the following observations:

$$80, 85, 99, 120, 125, 155, 160, 166, 176, 198.$$

(a) If the policies have a deductible of 90, estimate the expected loss and variance of loss in a loss event, as well as the expected loss and variance of loss in a payment event. Use the sample variance as the variance estimate.
(b) Estimate the probability of X not exceeding 160, and estimate the variance of this estimated probability.

11.3 A researcher has collected data on the paired variables (X, Y). However, she has lost the pairing of the data and only managed to keep the unpaired observations, with the values of x being: 2, 7, 4, 6, 8, 7, and the values of y being: 3, 2, 7, 6, 5, 2. What are the maximum and minimum possible sample correlation coefficients of X and Y (see equation (A.50)) based on the empirical data.

11.4 A sample of losses X has the following observations:

$$4, 6, 6, 9, 10, 14, 18, 20, 22, 25.$$

(a) Determine the median $\hat{x}_{0.5}$.
(b) If $Y = X \mid X > 5$, determine the median $\hat{y}_{0.5}$.

11.5 A sample of losses has the following observations:

$$5, 8, 8, 12, 14, 17, 21, 22, 26, 28.$$

(a) Compute the kernel density estimates $\tilde{f}(10)$ and $\tilde{f}(15)$ using the rectangular kernel with bandwidth of 4 and 6.
(b) Compute the kernel density estimates $\tilde{f}(10)$ and $\tilde{f}(15)$ using the triangular kernel with bandwidth of 4 and 6.

11.6 The following grouped data of losses X are given:

$(c_{j-1}, c_j]$	n_j
$(0, 10]$	22
$(10, 20]$	29
$(20, 30]$	38
$(30, 40]$	21
$(40, 50]$	16
$(50, 60]$	8

(a) Compute the mean and the standard deviation of the empirical distribution of X.
(b) Determine $\mathrm{E}[(X \wedge 40)]$ and $\mathrm{E}[(X \wedge 45)]$ of the empirical distribution of X.
(c) Determine $\mathrm{Var}[(X \wedge 40)]$ and $\mathrm{Var}[(X \wedge 45)]$ of the empirical distribution of X.
(d) If there is a deductible of 20, compute the expected loss in a loss event using the empirical distribution.
(e) Compute the empirical distribution functions $\hat{F}(40)$ and $\hat{F}(48)$.
(f) What is the 30th percentile of the empirical distribution?

11.7 You are given the following grouped loss data:

$(c_{j-1}, c_j]$	n_j
$(0, 100]$	18
$(100, 200]$	29
$(200, 300]$	32
$(300, 400]$	21
$(400, 500]$	12

(a) If the policies have a deductible of 200 and a maximum covered loss of 400, determine the mean and the standard deviation of the loss in a loss event.

(b) If the policies have a maximum covered loss of 280 and no deductible, compute the mean and the standard deviation of the loss in a loss event.

11.8 The following grouped data of losses X are given:

$(c_{j-1}, c_j]$	n_j	Total losses	Total squared losses
(0, 20]	15	164	2,208
(20, 40]	24	628	23,683
(40, 60]	26	1,284	68,320
(60, 80]	14	1,042	81,230
(80, 100]	6	620	52,863

(a) Compute the mean and the standard deviation of the empirical loss distribution.

(b) If there is a maximum loss limit of 60, determine the mean and the standard deviation of the loss in a loss event.

11.9 Refer to the duration data in Exercise 10.9.

(a) Estimate the probability of the duration exceeding 8 using the Kaplan–Meier method. Compute the 95% linear confidence interval of this probability.

(b) Estimate the probability of the duration exceeding 3 and less than 6 using the Kaplan–Meier method. What is the estimated variance of this probability?

(c) Estimate the probability of the duration exceeding 8, given that it exceeds 4, using the Kaplan–Meier method. What is the estimated variance of this probability?

(d) Estimate the cumulative hazard function at 10.5, $H(10.5)$, using the Nelson–Aalen method. Compute the 95% linear confidence interval of $H(10.5)$.

11.10 Refer to the graduate employment survey data in Exercise 10.10, and denote X as the starting monthly salary in hundred dollars.

(a) Estimate $\Pr(25 \leq X \leq 28)$ using the Kaplan–Meier estimator.

(b) Estimate $\Pr(X > 30 \,|\, X > 27)$ using the Nelson–Aalen estimator.

11.11 Refer to the unemployment survey data in Exercise 10.11.

(a) Estimate the mean unemployment duration using the Kaplan–Meier estimator.

(b) Compute the cumulative hazard function at unemployment duration 12 using the Kaplan–Meier estimator and the Nelson–Aalen estimator.

11.12 Refer to the loss data in Exercise 10.12.

(a) Estimate the probability of loss exceeding 9.5, $S(9.5)$, using the Kaplan–Meier estimator. Compute an approximate variance of this estimate.

(b) Compute the 95% linear confidence interval of $S(9.5)$, and compare this against the confidence interval using the logarithmic transformation method.

(c) Estimate the cumulative hazard function at 5, $H(5)$, using the Nelson–Aalen estimator. Compute an approximate variance of this estimate.

(d) Compute the 95% linear confidence interval of $H(5)$, and compare this against the confidence interval using the logarithmic transformation method.

11.13 Refer to the loss data in Exercise 10.13. Estimate the mean of the loss payment in a payment event using the Kaplan–Meier and Nelson–Aalen methods. Assume geometrical decay for the survival function beyond the maximum loss observation.

11.14 Refer to the loss data in Exercise 10.14.

(a) Compute the 90% logarithmic transformed confidence interval of the probability of loss less than 11 using the Kaplan–Meier method.

(b) Compute the 95% logarithmic transformed confidence interval of the cumulative hazard function at 8 using the Nelson–Aalen method.

11.15 Refer to the grouped loss data for the loss variable X in Exercise 10.14 (b).

(a) Estimate $\Pr(X \leq 10)$.

(b) Estimate $\Pr(X \leq 8)$.

(c) Estimate $\Pr(X > 12 \,|\, X > 4)$.

Questions adapted from SOA exams

11.16 The 95% linear confidence interval of the cumulative hazard function $H(x_0)$ is (1.54, 2.36). Compute the 95% logarithmic transformed confidence interval of $H(x_0)$.

11.17 In a sample of 40 insurance policy losses without deductible, there are five losses of amount 4, four losses of amount 8, and three losses of amount 12. In addition, there are two censored losses of amount 9 and one censored loss of amount 10. Compute the 90% logarithmic transformed confidence interval of the cumulative hazard function at 12, $H(12)$.

11.18 In a mortality study with right-censored data, the cumulative hazard function $H(t)$ is estimated using the Nelson–Aalen estimator. It is known that no death occurs between times t_i and t_{i+1}, that the 95% linear confidence interval of $H(t_i)$ is (0.07125, 0.22875), and that the 95% linear confidence interval of $H(t_{i+1})$ is (0.15607, 0.38635). What is the number of deaths at time t_{i+1}?

11.19 Fifteen cancer patients were observed from the time of diagnosis until death or 36 months from diagnosis, whichever comes first. Time (in months) since diagnosis, T, and number of deaths, n, are recorded as follows (d is unknown):

T	15	20	24	30	34	36
n	2	3	2	d	2	1

The Nelson–Aalen estimate of the cumulative hazard function at time 35, $\hat{H}(35)$, is 1.5641. Compute the Nelson–Aalen estimate of the variance of $\hat{H}(35)$.

11.20 You are given the following information about losses grouped by interval:

$(c_{j-1}, c_j]$	Number of losses V_j	Risk set R_j
(0, 20]	13	90
(20, 40]	27	88
(40, 60]	38	55
(60, 100]	21	36
(100, 200]	15	18

Estimate the probability of loss larger than 50 using the Nelson–Aalen method.

11.21 The times to death in a study of five lives from the onset of a disease to death are: 2, 3, 3, 3, and 7. Using a triangular kernel with bandwidth 2, estimate the density function at 2.5.

11.22 In a study of claim-payment times, the data were not truncated or censored, and at most one claim was paid at any one time. The Nelson–Aalen estimate of the cumulative hazard function $H(t)$ immediately following the second paid claim was 23/132. Determine the Nelson–Aalen estimate of $H(t)$ immediately after the fourth paid claim.

11.23 Suppose the 95% log-transformed confidence interval of $S(t_0)$ using the product-limit estimator is (0.695, 0.843). Determine $\hat{S}_K(t_0)$.

11.24 The claim payments of a sample of ten policies are ($^+$ indicates a right-censored payment): 2, 3, 3, 5, 5^+, 6, 7, 7^+, 9, and 10^+. Using the product-limit estimator, calculate the probability that the loss of a policy exceeds 8.

11.25 You are given the following data for a mortality study:

	Country A		Country B	
t_i	w_i	r_i	w_i	r_i
1	20	200	15	100
2	54	180	20	85
3	14	126	20	65
4	22	112	10	45

Let $\hat{S}_1(t)$ be the product-limit estimate of $S(t)$ based on the data for all countries and $\hat{S}_2(t)$ be the product-limit estimate of $S(t)$ based on the data for Country B. Determine $|\hat{S}_1(4) - \hat{S}_2(4)|$.

11.26 In a sample of 200 accident claims, t denotes the time (in months) a claim is submitted after the accident. There are no right-censored observations. $\hat{S}(t)$ is the Kaplan–Meier estimator and

$$c^2(t) = \frac{\widehat{\text{Var}}[\hat{S}(t)]}{[\hat{S}(t)]^2},$$

where $\widehat{\text{Var}}[\hat{S}(t)]$ is computed using Greenwood's approximation. If $\hat{S}(8) = 0.22$, $\hat{S}(9) = 0.16$, $c^2(9) = 0.02625$, and $c^2(10) = 0.04045$, determine the number of claims that were submitted ten months after an accident.

11.27 Eight people joined an exercise program on the same day. They stayed in the program until they reached their weight loss goal or switched to a diet program. Their experience is shown below:

	Time at which ...	
Member	Reach weight loss goal	Switch to diet program
1		4
2		8
3	8	
4	12	
5		12
6	12	
7	22	
8	36	

Let t be the time to reach the weight loss goal, and $H(t)$ be the cumulative hazard function at t. Using the Nelson–Aalen estimator, compute the upper limit of the symmetric 90% linear confidence interval of $H(12)$.

11.28 All members of a mortality study are observed from birth. Some leave the study by means other than death. You are given the following: $w_4 = 3$, $\hat{S}_K(y_3) = 0.65$, $\hat{S}_K(y_4) = 0.50$, and $\hat{S}_K(y_5) = 0.25$. Furthermore, between times y_4 and y_5, six observations are censored. Assuming no observations are censored at the times of death, determine w_5.

11.29 You are given the following data:

150, 150, 150, 362, 366, 452, 500, 500, 601, 693, 750, 750.

Let $\hat{H}_1(700)$ be the Nelson–Aalen estimate of the cumulative hazard function computed under the assumption that all observations are uncensored, and $\hat{H}_2(700)$ be the Nelson–Aalen estimate of the cumulative hazard function computed under the assumption that all occurrences of the values 150, 500, and 750 are right censored, while the remaining values are uncensored. Determine $|\hat{H}_1(700) - \hat{H}_2(700)|$.

11.30 You are given the following ages at time of death for ten individuals:

25, 30, 35, 35, 37, 39, 45, 47, 49, 55.

Using a uniform kernel with bandwidth 10, determine the kernel density estimate of the probability of survival to age 40.

11.31 For a mortality study with right-censored data, you are given the following:

t_i	w_i	r_i
3	1	50
5	3	49
6	5	k
10	7	21

If the Nelson–Aalen estimate of the survival function at time 10 is 0.575, determine k.

12

Parametric model estimation

Some models assume that the failure-time or loss variables follow a certain family of distributions, specified up to a number of unknown parameters. To compute quantities such as the average loss or VaR, the parameters of the distributions have to be estimated. This chapter discusses various methods of estimating the parameters of a failure-time or loss distribution.

Matching moments and percentiles to the data are two simple methods of parametric estimation. These methods, however, are subject to the decisions of the set of moments or percentiles to be used. The most important estimation method in the classical statistics literature is perhaps the maximum likelihood estimation method. It is applicable to a wide class of problems: variables that are discrete or continuous, and data observations that are complete or incomplete. On the other hand, the Bayesian approach provides an alternative perspective to parametric estimation, and has been made easier to adopt due to the advances in computational techniques.

Parametric models can be extended to allow the distributions to vary with some attributes of the objects, such as the years of driving experience of the insured in predicting vehicle accident claims. This gives rise to the use of models with covariates. A very important model in the parametric estimation of models with covariates is Cox's proportional hazards model.

We also include in this chapter a brief introduction to the use of copula in modeling the joint distributions of several variables. This approach is flexible in maintaining the assumptions about the marginal distributions, while analyzing the joint distribution through the copula function.

Learning objectives

1 Methods of moments and percentile matching
2 Maximum likelihood estimation
3 Bayesian estimation

4 Cox's proportional hazards model
5 Modeling joint distributions using copula

12.1 Methods of moments and percentile matching

Let $f(\cdot;\theta)$ be the pdf or pf of a failure-time or loss variable X, where $\theta = (\theta_1, \cdots, \theta_k)'$ is a k-element parameter vector. We denote μ'_r as the rth raw moment of X. In general, μ'_r are functions of the parameter θ, and we assume the functional dependence of μ'_r on θ is known so that we can write the raw moments as $\mu'_r(\theta)$. Given a random sample $\boldsymbol{x} = (x_1, \cdots, x_n)$ of X, the sample analogues of $\mu'_r(\theta)$ (i.e. the sample moments) are straightforward to compute. We denote these by $\hat{\mu}'_r$, so that

$$\hat{\mu}'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r. \tag{12.1}$$

12.1.1 Method of moments

The **method-of-moments** estimate $\hat{\theta}$ is the solution of θ in the equations

$$\mu'_r(\theta) = \hat{\mu}'_r, \quad \text{for } r = 1, \cdots, k. \tag{12.2}$$

Thus, we have a set of k equations involving k unknowns $\theta_1, \cdots, \theta_k$. We assume that a solution to the equations in (12.2) exists. However, there may be multiple solutions to the equations, in which case the method-of-moments estimate is not unique.

Example 12.1 Let X be the claim-frequency random variable. Determine the method-of-moments estimates of the parameter of the distribution of X, if X is distributed as (a) $\mathcal{PN}(\lambda)$, (b) $\mathcal{GM}(\theta)$, and (c) $\mathcal{BN}(m, \theta)$, where m is a known constant.

Solution All the distributions in this example are discrete with a single parameter in the pf. Hence, $k = 1$ and we need to match only the population mean $\mathrm{E}(X)$ to the sample mean $\bar{x}$. For (a), $\mathrm{E}(X) = \lambda$. Hence, $\hat{\lambda} = \bar{x}$. For (b), we have

$$\mathrm{E}(X) = \frac{1-\theta}{\theta} = \bar{x},$$

so that

$$\hat{\theta} = \frac{1}{1+\bar{x}}.$$

For (c), we equate $\mathrm{E}(X) = m\theta$ to $\bar{x}$ and obtain

$$\hat{\theta} = \frac{\bar{x}}{m},$$

which is the sample proportion. □

Example 12.2 Let X be the claim-severity random variable. Determine the method-of-moments estimates of the parameters of the distribution of X, if X is distributed as (a) $\mathcal{G}(\alpha, \beta)$, (b) $\mathcal{P}(\alpha, \gamma)$, and (c) $\mathcal{U}(a, b)$.

Solution All the distributions in this example are continuous with two parameters in the pdf. Thus, $k = 2$, and we need to match the first two population moments μ'_1 and μ'_2 to the sample moments $\hat{\mu}'_1$ and $\hat{\mu}'_2$. For (a), we have

$$\mu'_1 = \alpha\beta = \hat{\mu}'_1 \quad \text{and} \quad \mu'_2 = \alpha\beta^2 + \alpha^2\beta^2 = \hat{\mu}'_2,$$

from which we obtain

$$\beta\mu'_1 + \mu'^2_1 = \mu'_2.$$

Hence, the method-of-moments estimates are[1]

$$\hat{\beta} = \frac{\hat{\mu}'_2 - \hat{\mu}'^2_1}{\hat{\mu}'_1}$$

and

$$\hat{\alpha} = \frac{\hat{\mu}'_1}{\hat{\beta}} = \frac{\hat{\mu}'^2_1}{\hat{\mu}'_2 - \hat{\mu}'^2_1}.$$

For (b), the population moments are

$$\mu'_1 = \frac{\gamma}{\alpha - 1} \quad \text{and} \quad \mu'_2 = \frac{2\gamma^2}{(\alpha - 1)(\alpha - 2)},$$

from which we obtain

$$\mu'_2 = \frac{2\mu'^2_1(\alpha - 1)}{\alpha - 2}.$$

[1] Due to the Cauchy–Schwarz inequality, which states that for any $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_n\}$, $\left(\sum_{i=1}^n x_i y_i\right)^2 \le \left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i^2\right)$, we conclude that $\left(\sum_{i=1}^n x_i\right)^2 \le n\left(\sum_{i=1}^n x_i^2\right)$ so that $\hat{\mu}'_2 - \hat{\mu}'^2_1 \ge 0$, with equality only when all x_i are equal. Thus, $\hat{\alpha}$ and $\hat{\beta}$ are positive.

Hence

$$\hat{\alpha} = \frac{2(\hat{\mu}'_2 - \hat{\mu}'^2_1)}{\hat{\mu}'_2 - 2\hat{\mu}'^2_1}$$

and

$$\hat{\gamma} = (\hat{\alpha} - 1)\hat{\mu}'_1.$$

Note that if $\hat{\mu}'_2 - 2\hat{\mu}'^2_1 < 0$, then $\hat{\alpha} < 0$ and the model $\mathcal{P}(\hat{\alpha}, \hat{\gamma})$ is not well defined.[2]

For (c), the population moments are

$$\mu'_1 = \frac{a+b}{2} \quad \text{and} \quad \mu'_2 = \frac{(b-a)^2}{12} + \mu'^2_1.$$

Solving for a and b, and evaluating the solutions at $\hat{\mu}'_1$ and $\hat{\mu}'_2$, we obtain

$$\hat{a} = \hat{\mu}'_1 - \sqrt{3(\hat{\mu}'_2 - \hat{\mu}'^2_1)} \quad \text{and} \quad \hat{b} = \hat{\mu}'_1 + \sqrt{3(\hat{\mu}'_2 - \hat{\mu}'^2_1)}.$$

However, if $\min\{x_1, \cdots, x_n\} < \hat{a}$, or $\max\{x_1, \cdots, x_n\} > \hat{b}$, the model $\mathcal{U}(\hat{a}, \hat{b})$ is incompatible with the claim-severity data. □

Although the method of moments is generally easy to apply, the results may not be always satisfactory. As can be seen from Example 12.2, the estimates may be incompatible with the model assumption. However, as the sample moments are consistent estimates of the population moments, provided the parameters of the distribution can be solved uniquely from the population moments, the method-of-moments estimates are *consistent* for the model parameters. Of course, incompatibility may still exist when the sample size is small.

We have assumed that the matching of moments is based on the raw moments. An alternative is to consider central moments. For example, for a two-parameter distribution, we may match the sample mean to the population mean, and the sample variance to the population variance. This approach would result in different estimates from matching the raw moments, due to the degrees-of-freedom correction in the computation of the sample variance. In large samples, however, the difference is immaterial.

The method of moments can also be applied to censored or truncated distributions, as illustrated in the following examples.

[2] Existence of variance for the Pareto distribution requires α to be larger than 2. Hence, an estimate of $\hat{\alpha} \leq 2$ suggests a misfit of the model, if the variance is assumed to exist.

Example 12.3 A random sample of 15 ground-up losses, X, with a policy limit of 15 has the following observations:

$$2, 3, 4, 5, 8, 8, 9, 10, 11, 11, 12, 12, 15, 15, 15.$$

If X is distributed as $\mathcal{U}(0, b)$, determine the method-of-moments estimate of b.

Solution To estimate b we match the sample mean of the loss payments to the mean of the censored uniform distribution. The mean of the sample of 15 observations is 9.3333. As

$$\mathrm{E}[(X \wedge u)] = \int_0^u [1 - F(x)]\,dx = \int_0^u \frac{b-x}{b}\,dx = u - \frac{u^2}{2b},$$

and $u = 15$, we have

$$15 - \frac{(15)^2}{2\hat{b}} = 9.3333,$$

so that $\hat{b} = 19.8528$. □

Example 12.4 A random sample of ten insurance claims with a deductible of 5 has the following ground-up losses:

$$12, 13, 14, 16, 17, 19, 23, 27, 74, 97.$$

If the ground-up loss is distributed as $\mathcal{P}(\alpha, \gamma)$, determine the method-of-moments estimates of α and γ.

Solution From Example 3.7, we know that if the ground-up loss is distributed as $\mathcal{P}(\alpha, \gamma)$ and there is a deductible of d, then the distribution of the modified losses in a payment event is $\mathcal{P}(\alpha, \gamma + d)$. Hence, if μ_1' and μ_2' are the first two raw moments of the modified loss payments, from Example 12.2 the method-of-moments estimates of α and γ are

$$\hat{\alpha} = \frac{2(\hat{\mu}_2' - \hat{\mu}_1'^2)}{\hat{\mu}_2' - 2\hat{\mu}_1'^2}$$

and

$$\hat{\gamma} = (\hat{\alpha} - 1)\hat{\mu}_1' - d.$$

Now the modified claim amounts are:

$$7, 8, 9, 11, 12, 14, 18, 22, 69, 92,$$

so that $\hat{\mu}'_1 = 26.2$ and $\hat{\mu}'_2 = 1{,}468.8$. Hence

$$\hat{\alpha} = \frac{2[1{,}468.8 - (26.2)(26.2)]}{1{,}468.8 - 2(26.2)(26.2)} = 16.3128,$$

and

$$\hat{\gamma} = (16.3128 - 1)(26.2) - 5 = 396.1943.$$

We end this example by commenting that the Pareto distribution cannot be adopted if $\hat{\mu}'_2 - 2\hat{\mu}'^2_1 < 0$. Indeed, the estimation of the parameters of the Pareto distribution using the method of moments is generally numerically unstable.□

The classical method of moments can be made more flexible. First, instead of using the raw moments, we may consider a p-element vector-valued function $h(\boldsymbol{w};\theta)$, where $\boldsymbol{w}$ includes the loss variable of interest (such as the claim amount) as well as some covariates (such as some attributes of the insured), and θ is a k-element vector of parameters. The function $h(\boldsymbol{w};\theta)$ is defined in such a way that $\mathrm{E}[h(\boldsymbol{w};\theta_0)] = 0$ at the true parameter value θ_0, and $h(\boldsymbol{w};\theta)$ is called the **orthogonality condition**. For example, when the method of moments is applied to match the first two raw moments, $h(\boldsymbol{w};\theta)$ is defined as

$$h(\boldsymbol{w};\theta) = \begin{bmatrix} X - \mu'_1(\theta) \\ X^2 - \mu'_2(\theta) \end{bmatrix}. \tag{12.3}$$

Having defined $h(\boldsymbol{w};\theta)$, we estimate its expected value using the sample data, and denote this by $\hat{h}(\theta)$. Thus, the sample estimate of $\mathrm{E}[h(\boldsymbol{w};\theta)]$ with $h(\boldsymbol{w};\theta)$ given in equation (12.3) is

$$\hat{h}(\theta) = \begin{bmatrix} \hat{\mu}'_1 - \mu'_1(\theta) \\ \hat{\mu}'_2 - \mu'_2(\theta) \end{bmatrix}. \tag{12.4}$$

The classical method-of-moments estimation solves for $\hat{\theta}$ such that the sample estimate $\hat{h}(\theta)$ evaluated at $\hat{\theta}$ is zero. This solution, however, may not exist in general, especially when $p > k$ (we have more orthogonality conditions than the number of parameters). Hence, we modify the objective to finding the value $\hat{\theta}$ such that the sum of squares of the components of $\hat{h}(\hat{\theta})$, i.e. $\hat{h}(\hat{\theta})' \hat{h}(\hat{\theta})$, is minimized. Alternatively, we may consider minimizing a weighted sum of squares $\hat{h}(\hat{\theta})' \Omega \hat{h}(\hat{\theta})$, where Ω is a suitably defined weighting matrix dependent on the data. These extensions have been commonly used in the econometrics literature, in which the method is given the name **generalized method of moments**. In the statistics literature, this approach has been developed as the **estimation-function method**.

12.1.2 Method of percentile matching

It is well known that for some statistical distributions with thick tails (such as the **Cauchy** distribution and some members of the **stable distribution family**), moments of *any* order do not exist. For such distributions, the method of moments breaks down. On the other hand, as quantiles or percentiles of a distribution always exist, we may estimate the model parameters by matching the population percentiles (as functions of the parameters) to the sample percentiles. This approach is called **the method of percentile or quantile matching**.

Consider k quantities $0 < \delta_1, \cdots, \delta_k < 1$, and let $\delta_i = F(x_{\delta_i}; \theta)$ so that $x_{\delta_i} = F^{-1}(\delta_i; \theta)$, where θ is a k-element vector of the parameters of the df. Thus, x_{δ_i} is the δ_i-quantile of the loss variable X, which we write as $x_{\delta_i}(\theta)$, emphasizing its dependence on θ. Let $\hat{x}_{\delta_i}$ be the δ_i-quantile computed from the sample, as given in equations (11.7) and (11.9). The quantile-matching method solves for the value of $\hat{\theta}$, so that

$$x_{\delta_i}(\hat{\theta}) = \hat{x}_{\delta_i}, \quad \text{for } i = 1, \cdots, k. \tag{12.5}$$

Again we assume that a solution of $\hat{\theta}$ exists for the above equations, and it is called the **percentile- or quantile-matching estimate**.

Example 12.5 Let X be distributed as $\mathcal{W}(\alpha, \lambda)$. Determine the quantile-matching estimates of α and λ.

Solution Let $0 < \delta_1, \delta_2 < 1$. From equation (2.36), we have

$$\delta_i = 1 - \exp\left[-\left(\frac{x_{\delta_i}}{\lambda}\right)^{\alpha}\right], \quad i = 1, 2,$$

so that

$$-\left(\frac{x_{\delta_i}}{\lambda}\right)^{\alpha} = \log(1 - \delta_i), \quad i = 1, 2.$$

We take the ratio of the case of $i = 1$ to $i = 2$ to obtain

$$\left(\frac{x_{\delta_1}}{x_{\delta_2}}\right)^{\alpha} = \frac{\log(1 - \delta_1)}{\log(1 - \delta_2)},$$

which implies

$$\hat{\alpha} = \frac{\log\left[\dfrac{\log(1 - \delta_1)}{\log(1 - \delta_2)}\right]}{\log\left(\dfrac{\hat{x}_{\delta_1}}{\hat{x}_{\delta_2}}\right)},$$

where $\hat{x}_{\delta_1}$ and $\hat{x}_{\delta_2}$ are sample quantiles. Given $\hat{\alpha}$, we further solve for $\hat{\lambda}$ to obtain

$$\hat{\lambda} = \frac{\hat{x}_{\delta_1}}{[-\log(1-\delta_1)]^{\frac{1}{\hat{\alpha}}}} = \frac{\hat{x}_{\delta_2}}{[-\log(1-\delta_2)]^{\frac{1}{\hat{\alpha}}}}.$$

Thus, analytical solutions of $\hat{\alpha}$ and $\hat{\lambda}$ are obtainable. □

Example 12.6 Let X be distributed as $\mathcal{P}(\alpha, \gamma)$. Determine the quantile-matching estimates of α and γ.

Solution Let $0 < \delta_1, \delta_2 < 1$. From equation (2.38), we have

$$\delta_i = 1 - \left(\frac{\gamma}{x_{\delta_i} + \gamma}\right)^{\alpha}, \quad i = 1, 2,$$

so that

$$\alpha \log\left(\frac{\gamma}{x_{\delta_i} + \gamma}\right) = \log(1 - \delta_i), \quad i = 1, 2.$$

We take the ratio of the case of $i = 1$ to $i = 2$ to eliminate α. Evaluating the ratio at the sample quantiles, we obtain

$$\frac{\log\left(\frac{\gamma}{\hat{x}_{\delta_1} + \gamma}\right)}{\log\left(\frac{\gamma}{\hat{x}_{\delta_2} + \gamma}\right)} = \frac{\log(1-\delta_1)}{\log(1-\delta_2)}.$$

This equation involves only the unknown parameter γ. However, it cannot be solved analytically, and the solution $\hat{\gamma}$ has to be computed numerically. Given $\hat{\gamma}$, we can calculate $\hat{\alpha}$ as

$$\hat{\alpha} = \frac{\log(1-\delta_1)}{\log\left(\frac{\hat{\gamma}}{\hat{x}_{\delta_1} + \hat{\gamma}}\right)} = \frac{\log(1-\delta_2)}{\log\left(\frac{\hat{\gamma}}{\hat{x}_{\delta_2} + \hat{\gamma}}\right)}.$$ □

The above examples show that the quantile-matching method may be straightforward for some models, but numerically involving for others. One question that remains to be answered is the choice of the quantile set δ_i. Given the same data, the use of different quantiles may result in very different estimates of the parameter θ. Nonetheless, for models with analytically difficult pdf and nonexistence of moments, the quantile-matching method may be useful.[3]

[3] See Adler *et al.* (1998) for the use of quantile-based methods in estimating the stable distributions.

12.2 Bayesian estimation method

The Bayesian approach to parametric model estimation has been introduced in Section 8.1, in which the emphasis has been on the estimation of the expected value of the claim amount or claim frequency. For the purpose of estimating the unknown parameter θ in a model, the Bayesian approach views θ as the realization of a random variable Θ. The Bayesian estimator of Θ is a decision rule of assigning a value to Θ based on the observed data. The consequence of making a wrong decision about Θ is reflected in a loss function. Given a loss function, the decision rule is chosen to give as small an expected loss as possible. In particular, if the squared-error loss (or quadratic loss) function is adopted, the Bayesian estimator (the decision rule) is the mean of the posterior distribution (given the data) of Θ.

To compute the Bayes estimate of Θ, we need to obtain the posterior pdf of Θ, denoted by $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$. In general, the computation of $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ is quite complex, as it requires the knowledge of the marginal pdf of the data $\boldsymbol{X}$. We have surveyed in Chapter 8 some conjugate distributions that make the computation of the posterior mean of Θ particularly easy. If the prior pdf is conjugate to the likelihood, the posterior pdf belongs to the same family as the prior. This facilitates the computation of the posterior mean tremendously. Otherwise, numerical computational algorithms have to be adopted to determine the posterior mean. The required computational techniques have been advancing favorably to make the Bayesian approach more viable and easier to adopt.

Example 12.7 Let X be the loss random variable. Consider the following assumptions about the distribution of X and the prior distribution:

(a) $X \sim \mathcal{PN}(\lambda)$ and $\Lambda \sim \mathcal{G}(\alpha, \beta)$,
(b) $X \sim \mathcal{GM}(\theta)$ and $\Theta \sim \mathcal{B}(\alpha, \beta)$,
(c) $X \sim \mathcal{E}(\lambda)$ and $\Lambda \sim \mathcal{G}(\alpha, \beta)$.

In each case we have a random sample of n observations $x_1, \cdots, x_n$ of X. Determine the Bayes estimate of λ in (a) and (c), and θ in (b).

Solution For (a) we know from Section 8.2.1 that the posterior distribution of Λ is $\mathcal{G}(\alpha^*, \beta^*)$, where α^* and β^* are given in equations (8.19) and (8.20), respectively. The Bayes estimate of λ is the posterior mean of Λ, i.e.

$$\alpha^*\beta^* = \frac{\beta(\alpha + n\bar{x})}{n\beta + 1}.$$

Note that this result is the same as the posterior mean of X_{n+1} as derived in Example 8.7. This is due to the fact that the mean of X_{n+1} is Λ.

For (b) we know from Section 8.2.2 that the posterior distribution of Θ is $\mathcal{B}(\alpha^*, \beta^*)$ where α^* and β^* are given in equations (8.21) and (8.22), respectively. The Bayes estimate of θ is the posterior mean of Θ, which, from equation (A.103), is

$$\frac{\alpha^*}{\alpha^* + \beta^*} = \frac{\alpha + n}{\alpha + \beta + n + n\bar{x}}.$$

Note that this problem is different from the one in Example 8.8. In the latter example, the interest was in the posterior mean of X_{n+1}, which is equal to the posterior mean of $(1 - \Theta)/\Theta$. In the current problem, our interest is in the posterior mean of Θ.

For (c) we know from Section 8.2.3 that the posterior distribution of Λ is $\mathcal{G}(\alpha^*, \beta^*)$, where α^* and β^* are given in equations (8.23) and (8.24), respectively. The Bayes estimate of λ is the posterior mean of Λ, i.e.

$$\alpha^*\beta^* = \frac{\beta(\alpha + n)}{1 + n\beta\bar{x}}.$$

Again, this problem is different from that in Example 8.9, in which we were interested in the posterior mean of X_{n+1}, which is equal to the posterior mean of $1/\Lambda$. □

12.3 Maximum likelihood estimation method

Suppose we have a random sample of n observations of X, denoted by $\boldsymbol{x} = (x_1, \cdots, x_n)$. Given the pdf or pf of X, $f(\cdot;\theta)$, we define the **likelihood function** of the sample as the product of $f(x_i;\theta)$, denoted by $L(\theta;\boldsymbol{x})$. Thus, we have

$$L(\theta;\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta), \tag{12.6}$$

which is taken as a function of θ given $\boldsymbol{x}$. As the observations are independent, $L(\theta;\boldsymbol{x})$ is the joint probability or joint density of the observations. We further define the **log-likelihood function** as the logarithm of $L(\theta;\boldsymbol{x})$, i.e.

$$\log L(\theta;\boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i;\theta). \tag{12.7}$$

The value of θ, denoted by $\hat{\theta}$, that maximizes the likelihood function is called the **maximum likelihood estimator (MLE)** of θ. As the logarithm is a monotonic nondecreasing function, $\hat{\theta}$ also maximizes the log-likelihood function. Indeed,

maximization of the log-likelihood function is often easier than maximization of the likelihood function, as the former is the *sum* of n terms involving θ while the latter is a product.

We now discuss the asymptotic properties of the MLE and its applications. We first consider the case where $\boldsymbol{X}$ are independently and identically distributed. This is the case where we have complete individual loss observations. We then extend the discussion to the case where $\boldsymbol{X}$ are not identically distributed, such as for grouped or incomplete data. The properties of the MLE are well established in the statistics literature and their validity depends on some technical conditions, referred to as **regularity conditions**. We will not elaborate the regularity conditions for the properties of the MLE to hold. Instead, we will only state the standard asymptotic properties of the MLE and discuss its applications to problems concerning actuarial data.[4]

Appendix A.18 discusses some properties of the likelihood function. We shall summarize some of these results here, with the details deferred to the Appendix. We first consider the case where θ is a scalar. The **Fisher information in a single observation**, denoted by $I(\theta)$, is defined as

$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right], \tag{12.8}$$

which is also equal to

$$\mathrm{E}\left[-\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right]. \tag{12.9}$$

In addition, the **Fisher information** in a **random sample** $\boldsymbol{X}$, denoted by $I_n(\theta)$, is defined as

$$I_n(\theta) = \mathrm{E}\left[\left(\frac{\partial \log L(\theta;\boldsymbol{X})}{\partial \theta}\right)^2\right], \tag{12.10}$$

which is n times the Fisher information in a single observation, i.e.

$$I_n(\theta) = nI(\theta). \tag{12.11}$$

Also, $I_n(\theta)$ can be computed as

$$I_n(\theta) = \mathrm{E}\left[-\frac{\partial^2 \log L(\theta;\boldsymbol{X})}{\partial \theta^2}\right]. \tag{12.12}$$

[4] Readers may refer to Hogg and Craig (1995, Chapter 8) for the proof of the case of random samples. For more advanced results, see Amemiya (1985). An important regularity condition for the asymptotic properties of the MLE to hold is that the support of the distribution does not depend on the unknown parameter. For example, the assumption $X \sim \mathcal{U}(0,\theta)$ violates this condition, as X cannot take a value exceeding θ.

For any unbiased estimator $\tilde{\theta}$ of θ, the **Cramér–Rao inequality** states that

$$\text{Var}(\tilde{\theta}) \geq \frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)}, \tag{12.13}$$

and an unbiased estimator is said to be **efficient** if it attains the **Cramér–Rao lower bound**.

The MLE $\hat{\theta}$ is formally defined as

$$\hat{\theta} = \max_{\theta} \{L(\theta; \boldsymbol{x})\} = \max_{\theta} \{\log L(\theta; \boldsymbol{x})\}, \tag{12.14}$$

which can be computed by solving the first-order condition

$$\frac{\partial \log L(\theta; \boldsymbol{x})}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log f(x_i; \theta)}{\partial \theta} = 0. \tag{12.15}$$

We now state the asymptotic properties of the MLE for the random-sample case as follows:

Theorem 12.1 *Under certain regularity conditions, the distribution of* $\sqrt{n}(\hat{\theta} - \theta)$ *converges to the normal distribution with mean* 0 *and variance* $1/I(\theta)$, i.e.

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right), \tag{12.16}$$

where $\xrightarrow{D}$ *denotes convergence in distribution.*[5]

The above theorem has several important implications. First, $\hat{\theta}$ is asymptotically unbiased and consistent. Second, in large samples $\hat{\theta}$ is approximately normally distributed with mean θ and variance $1/I_n(\theta)$. Third, since the variance of $\hat{\theta}$ converges to the Cramér–Rao lower bound, $\hat{\theta}$ is *asymptotically* efficient.

We now generalize the results to the case where $\theta = (\theta_1, \cdots, \theta_k)'$ is a k-element vector. The **Fisher information matrix** in an observation is now defined as the $k \times k$ matrix

$$I(\theta) = \text{E}\left[\frac{\partial \log f(X; \theta)}{\partial \theta} \frac{\partial \log f(X; \theta)}{\partial \theta'}\right], \tag{12.17}$$

which is also equal to

$$\text{E}\left[-\frac{\partial^2 \log f(X; \theta)}{\partial \theta \, \partial \theta'}\right]. \tag{12.18}$$

[5] See DeGroot and Schervish (2002, p. 288) for discussions of convergence in distribution.

The Fisher information matrix in a random sample of n observations is $I_n(\theta) = nI(\theta)$. Let $\tilde{\theta}$ be any unbiased estimator of θ. We denote the variance matrix of $\tilde{\theta}$ by $\text{Var}(\tilde{\theta})$. Hence, the ith diagonal element of $\text{Var}(\tilde{\theta})$ is $\text{Var}(\tilde{\theta}_i)$, and its (i,j)th element is $\text{Cov}(\tilde{\theta}_i, \tilde{\theta}_j)$. Denoting $I_n^{-1}(\theta)$ as the inverse of $I_n(\theta)$, the multivariate version of the Cramér–Rao inequality states that

$$\text{Var}(\tilde{\theta}) - I_n^{-1}(\theta) \tag{12.19}$$

is a nonnegative definite matrix. As a property of nonnegative definite matrices, the diagonal elements of $\text{Var}(\tilde{\theta}) - I_n^{-1}(\theta)$ are nonnegative, i.e. the lower bound of $\text{Var}(\tilde{\theta}_i)$ is the ith diagonal element of $I_n^{-1}(\theta)$. An unbiased estimator is said to be efficient if it attains the Cramér–Rao lower bound $I_n^{-1}(\theta)$.

To compute the MLE $\hat{\theta}$, we solve the first-order condition in equation (12.15). The multivariate version of Theorem 12.1 is stated as follows:

Theorem 12.2 *Under certain regularity conditions, the distribution of* $\sqrt{n}(\hat{\theta} - \theta)$ *converges to the* ***multivariate normal distribution*** *with mean vector* 0 *and variance matrix* $I^{-1}(\theta)$*, i.e.*[6]

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}\left(0, I^{-1}(\theta)\right). \tag{12.20}$$

Again, this theorem says that the MLE is asymptotically unbiased, consistent, asymptotically normal, and efficient. Thus, its properties are very desirable, which explains its popularity.

The MLE has the convenient property that it satisfies the **invariance principle**. Suppose $g(\cdot)$ is a one-to-one function and $\hat{\theta}$ is the MLE of θ, then the invariance principle states that $g(\hat{\theta})$ is the MLE of $g(\theta)$.[7]

12.3.1 Complete individual data

Complete individual observations form a random sample for which the likelihood and log-likelihood functions are given in equations (12.6) and (12.7), respectively. Maximization through equation (12.15) then applies.

Example 12.8 Determine the MLE of the following models with a random sample of n observations: (a) $\mathcal{PN}(\lambda)$, (b) $\mathcal{GM}(\theta)$, (c) $\mathcal{E}(\lambda)$, and (d) $\mathcal{U}(0,\theta)$.

Solution Note that (a) and (b) are discrete models, while (c) and (d) are continuous. The same method, however, applies. For (a) the log-likelihood

[6] We use $\mathcal{N}$ to denote a multivariate normal distribution as well as a univariate normal. See Appendix A.19 for some properties of the multivariate normal distribution. Also, 0 denotes a vector of zeros.

[7] The invariance principle is not restricted to one-to-one functions. See DeGroot and Schervish (2002, p. 365) for an extension of the result.

function is[8]

$$\log L(\lambda; \boldsymbol{x}) = n\bar{x} \log \lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!),$$

and the first-order condition is

$$\frac{\partial \log L(\lambda; \boldsymbol{x})}{\partial \lambda} = \frac{n\bar{x}}{\lambda} - n = 0.$$

Thus, the MLE of λ is

$$\hat{\lambda} = \bar{x},$$

which is equal to the method-of-moments estimate derived in Example 12.1.
For (b), the log-likelihood function is

$$\log L(\theta; \boldsymbol{x}) = n \log \theta + [\log(1 - \theta)] \sum_{i=1}^{n} x_i,$$

and the first-order condition is

$$\frac{\partial \log L(\theta; \boldsymbol{x})}{\partial \theta} = \frac{n}{\theta} - \frac{n\bar{x}}{1 - \theta} = 0.$$

Solving for the above, we obtain

$$\hat{\theta} = \frac{1}{1 + \bar{x}},$$

which is also the method-of-moments estimate derived in Example 12.1.

For (c), the log-likelihood function is

$$\log L(\lambda; \boldsymbol{x}) = n \log \lambda - n\lambda\bar{x},$$

with the first-order condition being

$$\frac{\partial \log L(\lambda; \boldsymbol{x})}{\partial \lambda} = \frac{n}{\lambda} - n\bar{x} = 0.$$

Thus, the MLE of λ is

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$

[8] Note that the last term of the equation does not involve λ and can be ignored for the purpose of finding the MLE.

For (d), it is more convenient to consider the likelihood function, which is

$$L(\theta;\boldsymbol{x}) = \left(\frac{1}{\theta}\right)^n,$$

for $0 < x_1, \cdots, x_n \leq \theta$, and 0 otherwise. Thus, the value of θ that maximizes the above expression is $\hat{\theta} = \max\{x_1, \cdots, x_n\}$. Note that in this case the MLE is not solved from equation (12.15).

A remark for the $\mathcal{U}(0,\theta)$ case is of interest. Note that from Theorem 12.1, we conclude that $\text{Var}(\sqrt{n}\hat{\theta})$ converges to a positive constant when n tends to infinity, where $\hat{\theta}$ is the MLE. From Example 10.2, however, we learn that the variance of $\max\{x_1, \cdots, x_n\}$ is

$$\frac{n\theta^2}{(n+2)(n+1)^2},$$

so that $\text{Var}(n \max\{x_1, \cdots, x_n\})$ converges to a positive constant when n tends to infinity. Hence, Theorem 12.1 breaks down. This is due to the violation of the regularity conditions for this model, as discussed in Footnote 4. □

Example 12.9 Determine the MLE of the following models with a random sample of n observations: (a) $\mathcal{G}(\alpha,\beta)$, and (b) $\mathcal{P}(\alpha,\gamma)$.

Solution These models are continuous with two parameters. For (a) the log-likelihood function is

$$\log L(\alpha,\beta;\boldsymbol{x}) = (\alpha-1)\left[\sum_{i=1}^{n}\log x_i\right] - \frac{n\bar{x}}{\beta} - n\log[\Gamma(\alpha)] - n\alpha\log\beta.$$

This is a complex expression and there is no analytic solution for the maximum. Differentiation of the expression is not straightforward as it involves the gamma function $\Gamma(\alpha)$. On the other hand, if α is known, then there is only one unknown parameter β and the first-order condition becomes

$$\frac{\partial \log L(\beta;\alpha,\boldsymbol{x})}{\partial\beta} = \frac{n\bar{x}}{\beta^2} - \frac{n\alpha}{\beta} = 0,$$

so that the MLE of β is

$$\hat{\beta} = \frac{\bar{x}}{\alpha}.$$

For (b) the log-likelihood function is

$$\log L(\alpha,\gamma;\boldsymbol{x}) = n\log\alpha + n\alpha\log\gamma - (\alpha+1)\sum_{i=1}^{n}\log(x_i+\gamma).$$

The first-order conditions are

$$\frac{\partial \log L(\alpha, \gamma; \boldsymbol{x})}{\partial \alpha} = \frac{n}{\alpha} + n \log \gamma - \sum_{i=1}^{n} \log(x_i + \gamma) = 0,$$

and

$$\frac{\partial \log L(\alpha, \gamma; \boldsymbol{x})}{\partial \gamma} = \frac{n\alpha}{\gamma} - (\alpha + 1) \sum_{i=1}^{n} \frac{1}{x_i + \gamma} = 0.$$

Again, an analytic solution is not possible and the MLE have to be solved by numerical methods. If γ is known, then only the first equation above is required, and its solution is

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log(x_i + \gamma) - n \log \gamma}.$$

On the other hand, if α is known, we solve for γ from the second first-order condition above. However, an analytic solution is still not possible and the problem can only be solved numerically. □

Theorems 12.1 and 12.2 can be used to derive the asymptotic variance of the MLE and hence the confidence interval estimates of the parameters. The example below illustrates this application.

Example 12.10 Determine the asymptotic distribution of the MLE of the following models with a random sample of n observations: (a) $\mathcal{PN}(\lambda)$ and (b) $\mathcal{GM}(\theta)$. Hence, derive $100(1 - \alpha)\%$ confidence interval estimates for the parameters of the models.

Solution For (a) the second derivative of the log-likelihood of an observation is

$$\frac{\partial^2 \log f(x; \lambda)}{\partial \lambda^2} = -\frac{x}{\lambda^2}.$$

Thus

$$I(\lambda) = \mathrm{E}\left[-\frac{\partial^2 \log f(X; \lambda)}{\partial \lambda^2}\right] = \frac{1}{\lambda},$$

so that

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{D} \mathcal{N}(0, \lambda).$$

As in Example 12.8, $\hat{\lambda} = \bar{x}$, which is also the estimate for the variance. Hence, in large samples $\bar{x}$ is approximately normally distributed with mean λ and

variance λ/n (estimated by $\bar{x}/n$). A $100(1-\alpha)\%$ confidence interval of λ is computed as

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{x}}{n}}.$$

Note that we can also estimate the $100(1-\alpha)\%$ confidence interval of λ by

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}},$$

where s^2 is the sample variance. This estimate, however, will not be as efficient if X is Poisson.

For (b) the second derivative of the log-likelihood of an observation is

$$\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} = -\frac{1}{\theta^2} - \frac{x}{(1-\theta)^2}.$$

As $\mathrm{E}(X) = (1-\theta)/\theta$, we have

$$I(\theta) = \mathrm{E}\left[-\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right] = \frac{1}{\theta^2} + \frac{1}{\theta(1-\theta)} = \frac{1}{\theta^2(1-\theta)}.$$

Thus

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}\left(0, \theta^2(1-\theta)\right),$$

where, from Example 12.8

$$\hat{\theta} = \frac{1}{1+\bar{x}}.$$

A $100(1-\alpha)\%$ confidence interval of θ can be computed as

$$\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\theta}^2(1-\hat{\theta})}{n}}.$$

Note that the asymptotic variance of $\hat{\theta}$ can also be derived using the delta method, together with the result for the variance of the sample mean $\bar{x}$. Readers are invited to show the derivation (see Exercise 12.11). □

12.3.2 Grouped and incomplete data

When the sample data are grouped and/or incomplete, the observations are no longer iid. Nonetheless, we can still formulate the likelihood function and

compute the MLE. The first step is to write down the likelihood function or log-likelihood function of the sample that is appropriate for the way the observations are sampled.

We first consider the case where we have complete observations that are grouped into k intervals: $(c_0, c_1]$, $(c_1, c_2], \cdots, (c_{k-1}, c_k]$, where $0 \le c_0 < c_1 < \cdots < c_k = \infty$. Let the number of observations in the interval $(c_{j-1}, c_j]$ be n_j so that $\sum_{j=1}^{k} n_j = n$. Given a parametric df $F(\cdot;\theta)$, the probability of a single observation falling inside the interval $(c_{j-1}, c_j]$ is $F(c_j;\theta) - F(c_{j-1};\theta)$. Assuming the *individual* observations are iid, the likelihood of having n_j observations in the interval $(c_{j-1}, c_j]$, for $j = 1, \cdots, k$, is

$$L(\theta; \boldsymbol{n}) = \prod_{j=1}^{k} \left[F(c_j;\theta) - F(c_{j-1};\theta)\right]^{n_j}, \tag{12.21}$$

where $\boldsymbol{n} = (n_1, \cdots, n_k)$. The log-likelihood function of the sample is

$$\log L(\theta; \boldsymbol{n}) = \sum_{j=1}^{k} n_j \log \left[F(c_j;\theta) - F(c_{j-1};\theta)\right]. \tag{12.22}$$

Now we consider the case where we have individual observations that are right censored. If the ground-up loss is continuous, the claim amount will have a distribution of the mixed type, described by a pf–pdf. Specifically, if there is a policy limit of u, only claims of amounts in the interval $(0, u]$ are observable. Losses of amounts exceeding u are censored, so that the probability of a claim of amount u is $1 - F(u;\theta)$. Thus, if the claim data consist of $\boldsymbol{x} = (x_1, \cdots, x_{n_1})$, where $0 < x_1, \cdots, x_{n_1} < u$, and n_2 claims of amount u, with $n = n_1 + n_2$, then the likelihood function is given by

$$L(\theta; \boldsymbol{x}, n_2) = \left[\prod_{i=1}^{n_1} f(x_i;\theta)\right] [1 - F(u;\theta)]^{n_2}. \tag{12.23}$$

The log-likelihood function is

$$\log L(\theta; \boldsymbol{x}, n_2) = n_2 \log [1 - F(u;\theta)] + \sum_{i=1}^{n_1} \log f(x_i;\theta). \tag{12.24}$$

If the insurance policy has a deductible of d, the data of claim payments are sampled from a population with truncation, i.e. only losses with amounts exceeding d are sampled. Thus, the pdf of the ground-up loss *observed* is

$$\frac{f(x;\theta)}{1 - F(d;\theta)}, \quad \text{for } d < x. \tag{12.25}$$

If we have a sample of claim data $\boldsymbol{x} = (x_1, \cdots, x_n)$, then the likelihood function is given by

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} \frac{f(x_i;\theta)}{1 - F(d;\theta)} = \frac{1}{[1 - F(d;\theta)]^n} \prod_{i=1}^{n} f(x_i;\theta), \quad \text{where } d < x_1, \cdots, x_n. \tag{12.26}$$

Thus, the log-likelihood function is

$$\log L(\theta; \boldsymbol{x}) = -n \log [1 - F(d;\theta)] + \sum_{i=1}^{n} \log f(x_i;\theta). \tag{12.27}$$

We denote y_i as the modified loss amount, such that $y_i = x_i - d$. Let $\boldsymbol{y} = (y_1, \cdots, y_n)$. Suppose we wish to model the distribution of the payment in a payment event, and denote the pdf of this distribution by $\tilde{f}(\cdot;\theta^*)$, then the likelihood function of $\boldsymbol{y}$ is

$$L(\theta^*; \boldsymbol{y}) = \prod_{i=1}^{n} \tilde{f}(y_i;\theta^*), \quad \text{for } 0 < y_1, \cdots, y_n. \tag{12.28}$$

This model is called the **shifted model**. It captures the distribution of the loss in a payment event and may be different from the model of the ground-up loss distribution, i.e. $\tilde{f}(\cdot)$ may differ from $f(\cdot)$.

The above models may be extended and combined in various ways. For example, the data may have policies with different policy limits and/or deductibles. Policies may also have deductibles as well as maximum covered losses, etc. The principles illustrated above should be applied to handle different variations.

As the observations in general may not be iid, Theorems 12.1 and 12.2 may not apply. The asymptotic properties of the MLE beyond the iid assumption are summarized in the theorem below, which applies to a broad class of models.

Theorem 12.3 *Let $\hat{\theta}$ denote the MLE of the k-element parameter θ of the likelihood function $L(\theta; \boldsymbol{x})$. Under certain regularity conditions, the distribution of $\sqrt{n}(\hat{\theta} - \theta)$ converges to the multivariate normal distribution with mean vector 0 and variance matrix $\mathcal{I}^{-1}(\theta)$, i.e.*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}\left(0, \mathcal{I}^{-1}(\theta)\right), \tag{12.29}$$

where

$$\mathcal{I}(\theta) = \lim_{n\to\infty} E\left[-\frac{1}{n} \frac{\partial^2 \log L(\theta; \boldsymbol{x})}{\partial\theta\, \partial\theta'}\right]. \tag{12.30}$$

Note that $\mathcal{I}(\theta)$ requires the evaluation of an expectation and depends on the unknown parameter θ. In practical applications it may be estimated by its sample counterpart. Once $\mathcal{I}(\theta)$ is estimated, confidence intervals of θ may be computed.

Example 12.11 Let the ground-up loss X be distributed as $\mathcal{E}(\lambda)$. Consider the following cases:

(a) Claims are grouped into k intervals: $(0, c_1]$, $(c_1, c_2], \cdots, (c_{k-1}, \infty]$, with no deductible nor policy limit. Let $\boldsymbol{n} = (n_1, \cdots, n_k)$ denote the numbers of observations in the intervals.
(b) There is a policy limit of u. n_1 uncensored claims with ground-up losses $\boldsymbol{x} = (x_1, \cdots, x_{n_1})$ are available, and n_2 claims have a censored amount u.
(c) There is a deductible of d, and n claims with ground-up losses $\boldsymbol{x} = (x_1, \cdots, x_n)$ are available.
(d) Policy has a deductible of d and maximum covered loss of u. n_1 uncensored claims with ground-up losses $\boldsymbol{x} = (x_1, \cdots, x_{n_1})$ are available, and n_2 claims have a censored claim amount $u - d$. Denote $n = n_1 + n_2$.
(e) Similar to (d), but there are two blocks of policies with deductibles of d_1 and d_2 for Block 1 and Block 2, respectively. The maximum covered losses are u_1 and u_2 for Block 1 and Block 2, respectively. In Block 1 there are n_{11} uncensored claim observations and n_{12} censored claims of amount $u_1 - d_1$. In Block 2 there are n_{21} uncensored claim observations and n_{22} censored claims of amount $u_2 - d_2$.

Determine the MLE of λ in each case.

Solution The df of $\mathcal{E}(\lambda)$ is $F(x; \lambda) = 1 - e^{-\lambda x}$. For (a), using equation (12.21), the likelihood function is (with $c_0 = 0$)

$$L(\lambda; \boldsymbol{n}) = \left[\prod_{j=1}^{k-1} \left(e^{-c_{j-1}\lambda} - e^{-c_j \lambda} \right)^{n_j} \right] \left(e^{-c_{k-1}\lambda} \right)^{n_k},$$

so that the log-likelihood function is

$$\log L(\lambda; \boldsymbol{n}) = -c_{k-1} n_k \lambda + \sum_{j=1}^{k-1} n_j \log \left(e^{-c_{j-1}\lambda} - e^{-c_j \lambda} \right).$$

The MLE is solved by maximizing the above expression with respect to λ, for which numerical method is required.

For (b) the likelihood function is

$$L(\lambda;\boldsymbol{x}) = \left[\prod_{i=1}^{n_1} \lambda e^{-\lambda x_i}\right] e^{-\lambda u n_2},$$

and the log-likelihood function is

$$\log L(\lambda;\boldsymbol{x}) = -\lambda u n_2 - \lambda n_1 \bar{x} + n_1 \log \lambda.$$

The first-order condition is

$$\frac{\partial \log L(\lambda;\boldsymbol{x})}{\partial \lambda} = -u n_2 - n_1 \bar{x} + \frac{n_1}{\lambda} = 0,$$

which produces the MLE

$$\hat{\lambda} = \frac{n_1}{n_1 \bar{x} + n_2 u}.$$

For (c) the likelihood function is

$$L(\lambda;\boldsymbol{x}) = \frac{1}{e^{-\lambda d n}} \left[\prod_{i=1}^{n} \lambda e^{-\lambda x_i}\right],$$

and the log-likelihood function is

$$\log L(\lambda;\boldsymbol{x}) = \lambda d n - \lambda n \bar{x} + n \log \lambda.$$

The first-order condition is

$$\frac{\partial \log L(\lambda;\boldsymbol{x})}{\partial \lambda} = n d - n \bar{x} + \frac{n}{\lambda} = 0,$$

so that the MLE is

$$\hat{\lambda} = \frac{1}{\bar{x} - d}.$$

For (d) the likelihood function is

$$L(\lambda;\boldsymbol{x}) = \frac{1}{e^{-\lambda d n}} \left[\prod_{i=1}^{n_1} \lambda e^{-\lambda x_i}\right] e^{-\lambda u n_2},$$

with log-likelihood

$$\log L(\lambda;\boldsymbol{x}) = \lambda d n - \lambda n_1 \bar{x} + n_1 \log \lambda - \lambda u n_2,$$

and first-order condition

$$\frac{\partial \log L(\lambda; \boldsymbol{x})}{\partial \lambda} = nd - n_1 \bar{x} + \frac{n_1}{\lambda} - un_2 = 0.$$

The MLE is

$$\hat{\lambda} = \frac{n_1}{n_1(\bar{x} - d) + n_2(u - d)}.$$

For (e) the log-likelihood is the sum of the two blocks of log-likelihoods given in (d). Solving for the first-order condition, we obtain the MLE as

$$\begin{aligned}\hat{\lambda} &= \frac{n_{11} + n_{21}}{n_{11}(\bar{x}_1 - d_1) + n_{21}(\bar{x}_2 - d_2) + n_{12}(u_1 - d_1) + n_{22}(u_2 - d_2)} \\ &= \frac{\sum_{i=1}^{2} n_{i1}}{\sum_{i=1}^{2} [n_{i1}(\bar{x}_i - d_i) + n_{i2}(u_i - d_i)]}.\end{aligned}$$

□

Example 12.12 A sample of the ground-up loss X of two blocks of policies has the following observations (d_i = deductible, u_i = maximum covered loss)

Block 1 ($d_1 = 1.4$, $u_1 = \infty$):	0.13, 2.54, 2.16, 4.72, 1.88, 4.03, 1.39, 4.03, 3.23, 1.79,
Block 2 ($d_2 = 0$, $u_2 = 3$):	3.16, 2.64, 2.88, 4.38, 1.81, 2.29, 1.11, 1.78, 0.52, 3.69.

Assuming X is distributed as $\mathcal{W}(\alpha, \lambda)$, compute the MLE of α and λ, and estimate the 95% confidence intervals of these parameters.

Solution For Block 1, two losses (i.e., 0.13 and 1.39) are below the deductible, and are not observed in practice. There are eight claims, with ground-up losses denoted by x_{1i}, for $i = 1, \cdots, 8$. The likelihood function of this block of losses is

$$\prod_{i=1}^{8} \frac{f(x_{1i}; \alpha, \lambda)}{1 - F(d_1; \alpha, \lambda)}.$$

Using equations (2.34) and (2.36), the log-likelihood function can be written as

$$(\alpha - 1) \sum_{i=1}^{8} \log x_{1i} + 8 [\log \alpha - \alpha \log \lambda] - \sum_{i=1}^{8} \left(\frac{x_{1i}}{\lambda}\right)^{\alpha} + 8 \left(\frac{d_1}{\lambda}\right)^{\alpha}.$$

For Block 2, 3 claims (i.e. 3.16, 4.38, and 3.69) are right censored at $u_2 = 3$. We denote the uncensored losses by x_{2i}, for $i = 1, \cdots, 7$. As there is no deductible,

the likelihood function is

$$\left[\prod_{i=1}^{7} f(x_{2i};\alpha,\lambda)\right][1-F(u_2;\alpha,\lambda)]^3 .$$

Thus, the log-likelihood function of Block 2 is

$$(\alpha-1)\sum_{i=1}^{7}\log x_{2i}+7\,[\log\alpha-\alpha\log\lambda]-\sum_{i=1}^{7}\left(\frac{x_{2i}}{\lambda}\right)^{\alpha}-3\left(\frac{u_2}{\lambda}\right)^{\alpha}.$$

The log-likelihood of the whole sample is equal to the sum of the log-likelihoods of the two blocks. Maximizing this numerically with respect to α and λ, we obtain the MLE as (the standard errors of the estimates are in parentheses)[9]

$$\hat{\alpha} = 2.1669\ (0.5332),$$
$$\hat{\lambda} = 2.9064\ (0.3716).$$

Thus, the 95% confidence interval estimates of α and λ are, respectively

$$2.1669 \pm (1.96)(0.5332) = (1.1218,\ 3.2120)$$

and

$$2.9064 \pm (1.96)(0.3716) = (2.1781,\ 3.6347).$$ □

Example 12.13 Let the $\mathcal{W}(\alpha,\lambda)$ assumption be adopted for both the ground-up loss distribution and the payment distribution in a payment event (the shifted model) for the data of the policies of Block 1 in Example 12.12. Determine the MLE of these models.

Solution For the ground-up loss distribution, the log-likelihood of Block 1 in Example 12.12 applies. Maximizing this function (the log-likelihood of Block 2 is not required) we obtain the following estimates

$$\hat{\alpha} = 2.6040, \quad \hat{\lambda} = 3.1800.$$

Using equation (2.35), we derive the estimate of the mean of the ground-up loss as 2.8248.

For the shifted model, the log-likelihood function is

$$(\alpha-1)\sum_{i=1}^{8}\log(x_{1i}-d_1)+8\,[\log\alpha-\alpha\log\lambda]-\sum_{i=1}^{8}\left[\frac{x_{1i}-d_1}{\lambda}\right]^{\alpha},$$

[9] The standard error is the square root of the estimate of the corresponding diagonal element of $\mathcal{I}^{-1}(\theta)$ (see equation (12.30)) divided by $\sqrt{n}$. It is estimated using the sample analogue.

where $d_1 = 1.4$. The MLE (denoted with tildes) are

$$\tilde{\alpha} = 1.5979, \quad \tilde{\lambda} = 1.8412,$$

from which we obtain the estimate of the mean of the shifted loss as 1.6509. □

12.4 Models with covariates

We have so far assumed that the failure-time or loss distributions are *homogeneous*, in the sense that the same distribution applies to all insured objects, regardless of any attributes of the object that might be relevant. In practice, however, the future lifetime of smokers and non-smokers might differ. The accident rates of teenage drivers and middle-aged drivers might differ, etc. We now discuss some approaches in modeling the failure-time and loss distributions in which some attributes (called the **covariates**) of the objects affect the distributions.

Let $S(x;\theta)$ denote the survival function of interest, called the **baseline survival function**, which applies to the distribution independent of the object's attributes. Now suppose for the ith insured object, there is a vector of k attributes, denoted by $z_i = (z_{i1}, \cdots, z_{ik})'$, which affects the survival function. We denote the survival function of the ith object by $S(x;\theta,z_i)$. There are several ways in which the differences in the distribution can be formulated. We shall start with the popular **Cox's proportional hazards model**.

12.4.1 Proportional hazards model

Given the survival function $S(x;\theta)$, the hazard function $h(x;\theta)$ is defined as

$$h(x;\theta) = -\frac{d \log S(x;\theta)}{dx}, \tag{12.31}$$

from which we have

$$S(x;\theta) = \exp\left(-\int_0^x h(x;\theta)\,dx\right). \tag{12.32}$$

We now allow the hazard function to vary with the individuals and denote it by $h(x;\theta,z_i)$. In contrast, $h(x;\theta)$, which does not vary with i, is called the **baseline hazard function**. A simple model can be constructed by assuming that there exists a function $m(\cdot)$, such that if we denote $m_i = m(z_i)$, then

$$h(x;\theta,z_i) = m_i\, h(x;\theta). \tag{12.33}$$

This is called the proportional hazards model, which postulates that the hazard function of the ith individual is a multiple of the baseline hazard function, and the multiple depends on the covariate z_i.

An important implication of the proportional hazards model is that the survival function of the ith individual is given by

$$\begin{aligned} S(x;\theta,z_i) &= \exp\left(-\int_0^x h(x;\theta,z_i)\,dx\right) \\ &= \exp\left(-\int_0^x m_i\,h(x;\theta)\,dx\right) \\ &= \left[\exp\left(-\int_0^x h(x;\theta)\,dx\right)\right]^{m_i} \\ &= [S(x;\theta)]^{m_i}. \end{aligned} \tag{12.34}$$

For equation (12.33) to provide a well-defined hazard function, m_i must be positive for all z_i. Thus, the choice of the function $m(\cdot)$ is important. A popular assumption which satisfies this requirement is

$$m_i = \exp(\beta' z_i), \tag{12.35}$$

where $\beta = (\beta_1, \cdots, \beta_k)'$ is a vector of parameters. Based on this assumption, an individual has the baseline hazard function if $z_i = 0$.

The pdf of the ith individual can be written as

$$\begin{aligned} f(x;\theta,z_i) &= -\frac{dS(x;\theta,z_i)}{dx} \\ &= -\frac{d\,[S(x;\theta)]^{m_i}}{dx} \\ &= m_i\,[S(x;\theta)]^{m_i-1} f(x;\theta), \end{aligned} \tag{12.36}$$

where $f(x;\theta) = -dS(x;\theta)/dx$ is the **baseline pdf**. From this equation the likelihood of a sample can be obtained, which depends on the parameters θ and β. Given the functional form of the baseline pdf (or sf), the MLE of θ and β can be computed.

Example 12.14 There are two blocks of policies such that $z_i = 1$ if the insured is from Block 1, and $z_i = 0$ if the insured is from Block 2. We adopt the model in equation (12.35) so that $m_i = e^{\beta}$ if $z_i = 1$, and $m_i = 1$ if $z_i = 0$. Let the baseline loss distribution be $\mathcal{E}(\lambda)$. Suppose there are n_1 losses in Block 1 and n_2 losses in Block 2, with $n = n_1 + n_2$. What is the log-likelihood function of the sample?

Solution The baseline distribution applies to losses in Block 2. Note that $m(1) = e^{\beta}$ and $m(0) = 1$. As $S(x;\lambda) = e^{-\lambda x}$ and $f(x;\lambda) = \lambda e^{-\lambda x}$, from equation (12.36), the pdf of losses in Block 1 is

$$f(x;\lambda,1) = e^{\beta}\left(e^{-\lambda x}\right)^{e^{\beta}-1}(\lambda e^{-\lambda x}) = \lambda e^{\beta-\lambda x e^{\beta}}.$$

The pdf of losses in Block 2 is the pdf of $\mathcal{E}(\lambda)$, i.e.

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

If the Block 1 losses are denoted by $x_{11},\cdots,x_{1n_1}$ and the Block 2 losses are denoted by $x_{21},\cdots,x_{2n_2}$, the log-likelihood function of the sample is

$$n\log\lambda + n_1\beta - \lambda e^{\beta}\sum_{j=1}^{n_1} x_{1j} - \lambda\sum_{j=1}^{n_2} x_{2j}.$$

□

The above example shows that the MLE of the full model may be quite complicated even for a simple baseline model such as the exponential. Furthermore, it may be desirable to separate the estimation of the parameters in the proportional hazards function, i.e. β, versus the estimation of the baseline hazard function. For example, a researcher may only be interested in the *relative* effect of smoking (versus non-smoking) on future lifetime. Indeed, the estimation can be done in two stages. The first stage involves estimating β using the **partial likelihood method**, and the second stage involves estimating the baseline hazard function using a nonparametric method, such as the Kaplan–Meier or Nelson–Aalen estimators.

The partial likelihood method can be used to estimate β in the proportional hazards function. We now explain this method using the failure-time data terminology. Recall the notations introduced in Chapter 10 that the observed distinct failure times in the data are arranged in the order $0 < y_1 < \cdots < y_m$, where $m \le n$. There are w_j failures at time y_j and the risk set at time y_j is r_j. Suppose object i fails at time y_j, the partial likelihood of object i, denoted by $L_i(\beta)$, is defined as the probability of object i failing at time y_j given that some objects fail at time y_j. Thus, we have

$$\begin{aligned} L_i(\beta) &= \Pr(\text{object } i \text{ fails at time } y_j \mid \text{some objects fail at time } y_j) \\ &= \frac{\Pr(\text{object } i \text{ fails at time } y_j)}{\Pr(\text{some objects fail at time } y_j)} \\ &= \frac{h(y_j;\theta,z_i)}{\sum_{i'\in r_j} h(y_j;\theta,z_{i'})} \end{aligned}$$

$$= \frac{m_i\, h(y_j;\theta)}{\sum_{i' \in r_j} m_{i'}\, h(y_j;\theta)}$$
$$= \frac{m_i}{\sum_{i' \in r_j} m_{i'}}$$
$$= \frac{\exp(\beta' z_i)}{\sum_{i' \in r_j} \exp(\beta' z_{i'})}, \quad \text{for } i = 1, \cdots, n. \tag{12.37}$$

The third line in the above equation is due to the definition of a hazard function; the last line is due to the assumption in equation (12.35). Note that $L_i(\beta)$ does not depend on the baseline hazard function. It is also not dependent on the value of y_j.

The partial likelihood of the sample, denoted by $L(\beta)$, is defined as

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta). \tag{12.38}$$

Note that only β appears in the partial likelihood function, which can be maximized to obtain the estimate of β without any assumptions about the baseline hazard function and its estimates.

Example 12.15 A proportional hazards model has two covariates $z = (z_1, z_2)'$, each taking possible values 0 and 1. We denote $z_{(1)} = (0,0)'$, $z_{(2)} = (1,0)'$, $z_{(3)} = (0,1)'$, and $z_{(4)} = (1,1)'$. The failure times observed are

2 (1), 3 (2), 4 (3), 4 (4), 5 (1), 7 (3), 8 (1), 8 (4), 9 (2), 11 (2), 11 (2), 12 (3),

where the index i of the covariate vector $z_{(i)}$ of the observed failures are given in parentheses.[10] Also, an object with covariate vector $z_{(2)}$ is censored at time 6, and another object with covariate vector $z_{(4)}$ is censored at time 8. Compute the partial likelihood estimate of β.

Solution As there are two covariates, we let $\beta = (\beta_1, \beta_2)'$. Next we compute the multiples of the baseline hazard function. Thus, $m_{(1)} = \exp(\beta' z_{(1)}) = 1$, $m_{(2)} = \exp(\beta' z_{(2)}) = \exp(\beta_1)$, $m_{(3)} = \exp(\beta' z_{(3)}) = \exp(\beta_2)$, and $m_{(4)} = \exp(\beta' z_{(4)}) = \exp(\beta_1 + \beta_2)$. We tabulate the data and the computation of the partial likelihood in Table 12.1.

If two objects, i and i', have the same failure time y_j, their partial likelihoods have the same denominator (see equation (12.37)). With a slight abuse of notation, we denote $L_j(\beta)$ as the partial likelihood of the object (or the product of the partial likelihoods of the objects) with failure time y_j. Then the partial

[10] For instance, the first failure occurred at time 2 and the covariate vector of the failed object is $z_{(1)} = (0,0)'$.

Table 12.1. *Computation of the partial likelihood for Example 12.15*

		Covariate	r_j of covariate $z_{(i)}$				$L_j(\beta) = \text{num}_j/\text{den}_j$	
j	y_j	vector	(1)	(2)	(3)	(4)	num_j	den_j
1	2	$z_{(1)}$	3	5	3	3	$m_{(1)}$	$3m_{(1)} + 5m_{(2)} + 3m_{(3)} + 3m_{(4)}$
2	3	$z_{(2)}$	2	5	3	3	$m_{(2)}$	$2m_{(1)} + 5m_{(2)} + 3m_{(3)} + 3m_{(4)}$
3	4	$z_{(3)}, z_{(4)}$	2	4	3	3	$m_{(3)}m_{(4)}$	$\left[2m_{(1)} + 4m_{(2)} + 3m_{(3)} + 3m_{(4)}\right]^2$
4	5	$z_{(1)}$	2	4	2	2	$m_{(1)}$	$2m_{(1)} + 4m_{(2)} + 2m_{(3)} + 2m_{(4)}$
5	7	$z_{(3)}$	1	3	2	2	$m_{(3)}$	$m_{(1)} + 3m_{(2)} + 2m_{(3)} + 2m_{(4)}$
6	8	$z_{(1)}, z_{(4)}$	1	3	1	2	$m_{(1)}m_{(4)}$	$\left[m_{(1)} + 3m_{(2)} + m_{(3)} + 2m_{(4)}\right]^2$
7	9	$z_{(2)}$	0	3	1	0	$m_{(2)}$	$3m_{(2)} + m_{(3)}$
8	11	$z_{(2)}, z_{(2)}$	0	2	1	0	$m_{(2)}^2$	$\left[2m_{(2)} + m_{(3)}\right]^2$
9	12	$z_{(3)}$	0	0	1	0	$m_{(3)}$	$m_{(3)}$

likelihood of the sample is equal to

$$L(\beta) = \prod_{i=1}^{12} L_i(\beta) = \prod_{j=1}^{9} L_j(\beta) = \prod_{j=1}^{9} \frac{\text{num}_j}{\text{den}_j},$$

where num_j and den_j are given in the last two columns of Table 12.1. Maximizing $L(\beta)$ with respect to β, we obtain $\hat{\beta}_1 = -0.6999$ and $\hat{\beta}_2 = -0.5518$. These results imply $\hat{m}_{(1)} = 1$, $\hat{m}_{(2)} = 0.4966$, $\hat{m}_{(3)} = 0.5759$, and $\hat{m}_{(4)} = 0.2860$. □

Having estimated the parameter β in the proportional hazards model, we can continue to estimate the baseline hazard function nonparametrically using the Nelson–Aalen method. Recall that the cumulative hazard function $H(y;\theta)$, defined by

$$H(y;\theta) = \int_0^y h(y;\theta)\, dy, \tag{12.39}$$

can be estimated by the Nelson–Aalen method using the formula

$$\hat{H}(y;\theta) = \sum_{\ell=1}^{j} \frac{w_\ell}{r_\ell}, \qquad \text{for } y_j \le y < y_{j+1}. \tag{12.40}$$

Now the cumulative hazard function with covariate z_i is given by

$$\begin{aligned} H(y;\theta,z_i) &= \int_0^y h(y;\theta,z_i)\, dy \\ &= m_i \int_0^y h(y;\theta)\, dy \\ &= m_i\, H(y;\theta). \end{aligned} \tag{12.41}$$

This suggests that equation (12.40) may be modified as follows

$$\hat{H}(y;\theta) = \sum_{\ell=1}^{j} \frac{w_\ell}{r_\ell^*}, \qquad \text{for } y_j \le y < y_{j+1}, \tag{12.42}$$

where r_ℓ^* is the modified risk set defined by

$$r_\ell^* = \sum_{i' \in r_\ell} m_{i'}. \tag{12.43}$$

Instead of deriving formula (12.42), we now show that the method works for two special cases. First, note that if there are no covariates, all objects have the baseline hazard function so that $m_{i'} = 1$ for all $i' \in r_\ell$. Then $r_\ell^* = r_\ell$, and equation (12.42) gives us the basic Nelson–Aalen formula.[11] On the other hand, if all objects have the same covariate z^* and $m(z^*) = m^*$, then $r_\ell^* = \sum_{i' \in r_\ell} m^* = m^* r_\ell$. Thus, we have

$$\sum_{\ell=1}^{j} \frac{w_\ell}{r_\ell^*} = \sum_{\ell=1}^{j} \frac{w_\ell}{m^* r_\ell} = \frac{1}{m^*} \sum_{\ell=1}^{j} \frac{w_\ell}{r_\ell} = \frac{\hat{H}(y;\theta,z^*)}{m^*} = \hat{H}(y;\theta), \tag{12.44}$$

which is the result in equation (12.42).

Example 12.16 For the data in Example 12.15, compute the Nelson–Aalen estimate of the baseline hazard function and the baseline survival function. Estimate the survival functions $S(3.5; z_{(2)})$ and $S(8.9; z_{(4)})$.

Solution The results are summarized in Table 12.2. Note that r_ℓ^* in Column 4 are taken from the last Column of Table 12.1 (ignore the square, if any) evaluated at $\hat{\beta}$.

We can now compute the survival functions for given covariates. In particular, we have

$$\hat{S}(3.5; z_{(2)}) = (0.7669)^{0.4966} = 0.8765,$$

and

$$\hat{S}(8.9; z_{(4)}) = (0.2161)^{0.2860} = 0.6452.$$

The values of $\hat{m}_{(2)} = 0.4966$ and $\hat{m}_{(4)} = 0.2860$ are taken from Example 12.15.

[11] Recall that r_ℓ stands for the collection of items in the risk set as well as the number of items in the risk set.

Table 12.2. *Nelson–Aalen estimates for Example 12.16*

ℓ	y_ℓ	w_ℓ	r^*_ℓ	$\frac{w_\ell}{r^*_\ell}$	$\hat{H}(y)=\sum_{j=1}^{\ell}\frac{w_j}{r^*_j}$	$\hat{S}(y)=\exp\left[-\hat{H}(y)\right]$
1	2	1	8.0689	0.1239	0.1239	0.8834
2	3	1	7.0689	0.1414	0.2654	0.7669
3	4	2	6.5723	0.3043	0.5697	0.5656
4	5	1	5.7104	0.1751	0.7448	0.4748
5	7	1	4.2137	0.2373	0.9821	0.3745
6	8	2	3.6378	0.5497	1.5319	0.2161
7	9	1	2.0658	0.4840	2.0159	0.1331
8	11	2	1.5691	1.2745	3.2905	0.0372
9	12	1	0.5759	1.7363	5.0269	0.0065

□

12.4.2 Generalized linear model

While the mean of the loss may or may not be directly a parameter of the pdf, it is often a quantity of paramount importance. If some covariates are expected to affect the loss distribution, it may be natural to assume that they determine the mean of the loss. Thus, a modeling strategy is to assume that the mean of the loss variable X, denoted by μ, is a function of the covariate z. To ensure the mean loss is positive, we may adopt the following model

$$\mathrm{E}(X)=\mu=\exp(\beta' z). \tag{12.45}$$

This model is called the **generalized linear model**. The exponential function used in the above equation is called the **link function**, which relates the mean loss to the covariate. The exponential link function is appropriate for modeling loss distributions as it ensures that the expected loss is positive.

For illustration, if $X\sim\mathcal{PN}(\lambda)$, a generalized linear model may assume that the mean of X, λ, is given by $\lambda=\exp(\beta' z)$, so that the log-likelihood function of a random sample of n observations is (after dropping the irrelevant *constant*)

$$\log L(\beta;\boldsymbol{x})=\sum_{i=1}^{n}x_i(\beta' z_i)-\sum_{i=1}^{n}\exp(\beta' z_i). \tag{12.46}$$

Example 12.17 If $X\sim\mathcal{G}(\alpha,\beta)$, where the covariate z determines the mean of X, construct a generalized linear model for the distribution of X.

Solution We assume $\mathrm{E}(X)=\mu=\exp(\delta' z)$.[12] As $\mathrm{E}(X)=\alpha\beta$, we write $\beta=\mu/\alpha=\exp(\delta' z)/\alpha$, and re-parameterize the pdf of X by α and δ, with z

[12] The switch of the notation from β to δ is to avoid confusion.

as the covariate. Thus, the pdf of X is

$$f(x;\alpha,\delta,z) = \frac{\alpha^\alpha}{\Gamma(\alpha)\mu^\alpha}\, x^{\alpha-1} e^{-\frac{\alpha x}{\mu}} = \frac{\alpha^\alpha}{\Gamma(\alpha)\exp(\alpha\delta' z)}\, x^{\alpha-1} \exp\left(-\alpha x e^{-\delta' z}\right).$$

□

12.4.3 Accelerated failure-time model

In the accelerated failure-time model, the survival function of object i with covariate z_i, $S(x;\theta,z_i)$, is related to the baseline (i.e. $z = 0$) survival function as follows

$$S(x;\theta,z_i) = S(m_i x;\theta,0), \tag{12.47}$$

where $m_i = m(z_i)$ for an appropriate function $m(\cdot)$. Again, a convenient assumption is $m(z_i) = \exp(\beta' z_i)$. We now denote $X(z_i)$ as the failure-time random variable for an object with covariate z_i. The expected lifetime (at birth) is

$$\begin{aligned}
\mathrm{E}[X(z_i)] &= \int_0^\infty S(x;\theta,z_i)\,dx \\
&= \int_0^\infty S(m_i x;\theta,0)\,dx \\
&= \frac{1}{m_i}\int_0^\infty S(x;\theta,0)\,dx \\
&= \frac{1}{m_i}\mathrm{E}[X(0)]. \qquad (12.48)
\end{aligned}$$

The third line in the above equation is due to the change of variable in the integration. Hence, the expected lifetime at birth of an object with covariate z_i is $1/m_i$ times the expected lifetime at birth of a *baseline* object. This result can be further extended. We define the random variable $T(x;z_i)$ as the future lifetime of an object with covariate z_i aged x. Thus

$$T(x;z_i) = X(z_i) - x \,|\, X(z_i) > x. \tag{12.49}$$

It can be shown that

$$\mathrm{E}[T(x;z_i)] = \frac{1}{m_i}\mathrm{E}[T(m_i x;0)]. \tag{12.50}$$

Thus, the expected future lifetime of an object with covariate z_i aged x is $1/m_i$ times the expected future lifetime of an object with covariate 0 aged $m_i x$. Readers are invited to prove this result (see Exercise 12.24).

Example 12.18 The baseline distribution of the age-at-death random variable X is $\mathcal{E}(\lambda)$. Let z_i be the covariate of object i and assume the accelerated failure-time model with $m_i = \exp(\beta' z_i)$. Derive the log-likelihood function of a sample of n lifes.

Solution The baseline survival function is $S(x; \lambda, 0) = \exp(-\lambda x)$, and we have

$$S(x; \lambda, z_i) = S(m_i x; \lambda, 0) = \exp(-\lambda x e^{\beta' z_i}).$$

Thus, the pdf of $X(z_i)$ is

$$f(x; \lambda, z_i) = \lambda e^{\beta' z_i} \exp(-\lambda x e^{\beta' z_i}),$$

and the log-likelihood of the sample is

$$\log L(\lambda, \beta; \boldsymbol{x}) = n \log \lambda + \sum_{i=1}^{n} \beta' z_i - \lambda \sum_{i=1}^{n} x_i e^{\beta' z_i}.$$

□

12.5 Modeling joint distribution using copula

In many practical applications researchers are often required to analyze multiple risks of the same group, similar risks from different groups, or different aspects of a risk group. Thus, techniques for modeling multivariate distributions are required. While modeling the joint distribution directly may be an answer, this approach may face some difficulties in practice. First, there may not be an appropriate standard multivariate distribution that is suitable for the problem at hand. Second, the standard multivariate distributions usually have marginals from the same family, and this may put severe constraints on the solution. Third, researchers may have an accepted marginal distribution that suits the data, and would like to maintain the marginal model while extending the analysis to model the joint distribution.

The use of **copula** provides a flexible approach to modeling multivariate distributions. Indeed, the literature on this area is growing fast, and many applications are seen in finance and actuarial science. In this section we provide a brief introduction to the technique. For simplicity of exposition, we shall consider only bivariate distributions, and assume that we have continuous distributions for the variables of interest.

Definition 12.1 A bivariate copula $C(u_1, u_2)$ is a mapping from the unit square $[0, 1]^2$ to the unit interval $[0, 1]$. It is increasing in each component and satisfies the following conditions:

1 $C(1, u_2) = u_2$ and $C(u_1, 1) = u_1$, for $0 \le u_1, u_2 \le 1$,

2 For any $0 \leq a_1 \leq b_1 \leq 1$ and $0 \leq a_2 \leq b_2 \leq 1$, $C(b_1, b_2) - C(a_1, b_2) - C(b_1, a_2) + C(a_1, a_2) \geq 0$.

A bivariate copula is in fact a joint df on $[0, 1]^2$ with standard uniform marginals, i.e. $C(u_1, u_2) = \Pr(U_1 \leq u_1, U_2 \leq u_2)$, where U_1 and U_2 are uniformly distributed on $[0, 1]$. The first condition above implies that the marginal distribution of each component of the copula is uniform. The second condition is called the **rectangle inequality**. It ensures that $\Pr(a_1 \leq U_1 \leq b_1, a_2 \leq U_2 \leq b_2)$ is nonnegative.

Let $F_{X_1X_2}(\cdot, \cdot)$ be the joint df of X_1 and X_2, with marginal df $F_{X_1}(\cdot)$ and $F_{X_2}(\cdot)$. The theorem below, called the **Sklar Theorem**, states the representation of the joint df using a copula. It also shows how a joint distribution can be created via a copula.

Theorem 12.4 *Given the joint and marginal df of X_1 and X_2, there exists a unique copula $C(\cdot, \cdot)$, such that*

$$F_{X_1X_2}(x_1, x_2) = C(F_{X_1}(x_1), F_{X_2}(x_2)). \tag{12.51}$$

Conversely, if $C(\cdot, \cdot)$ is a copula, and $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ are univariate df of X_1 and X_2, respectively, then $C(F_{X_1}(x_1), F_{X_2}(x_2))$ is a bivariate df with marginal df $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$.

Proof See McNeil *et al.* (2005, p. 187). □

If the inverse functions $F_{X_1}^{-1}(\cdot)$ and $F_{X_2}^{-1}(\cdot)$ exist, the copula satisfying equation (12.51) is given by

$$C(u_1, u_2) = F_{X_1X_2}(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)). \tag{12.52}$$

The second part of Theorem 12.4 enables us to construct a bivariate distribution with given marginals. With a well-defined copula satisfying Definition 12.1, $C(F_{X_1}(x_1), F_{X_2}(x_2))$ establishes a bivariate distribution with the known marginals. This can be described as a *bottom-up* approach in creating a bivariate distribution.

The theorem below, called the **Fréchet bounds** for copulas, establishes the maximum and minimum of a copula.

Theorem 12.5 *The following bounds apply to any bivariate copula*

$$\max\{0, u_1 + u_2 - 1\} \leq C(u_1, u_2) \leq \min\{u_1, u_2\}. \tag{12.53}$$

Proof See McNeil *et al.* (2005, p. 188). □

The likelihood function of a bivariate distribution created by a copula can be computed using the following theorem.

Theorem 12.6 *Let X_1 and X_2 be two continuous distributions with pdf $f_{X_1}(\cdot)$ and $f_{X_2}(\cdot)$, respectively. If the joint df of X_1 and X_2 is given by equation (12.51), their joint pdf can be written as*

$$f_{X_1X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) c(F_{X_1}(x_1), F_{X_2}(x_2)), \tag{12.54}$$

where

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \, \partial u_2} \tag{12.55}$$

is called the ***copula density***.

Proof This can be obtained by differentiating equation (12.51). □

From Theorem 12.6, we can conclude that the log-likelihood of a bivariate random variable with df given by equation (12.51) is

$$\log\left[f_{X_1X_2}(x_1, x_2)\right] = \log\left[f_{X_1}(x_1)\right] + \log\left[f_{X_2}(x_2)\right] + \log\left[c(F_{X_1}(x_1), F_{X_2}(x_2))\right], \tag{12.56}$$

which is the log-likelihood of two *independent* observations of X_1 and X_2, plus a term which measures the dependence.

We now introduce some simple bivariate copulas. **Clayton's copula**, denoted by $C_C(u_1, u_2)$, is defined as

$$C_C(u_1, u_2) = \left(u_1^{-\alpha} + u_2^{-\alpha} - 1\right)^{-\frac{1}{\alpha}}, \quad \alpha > 0. \tag{12.57}$$

The Clayton copula density is given by

$$c_C(u_1, u_2) = \frac{1 + \alpha}{(u_1 u_2)^{1+\alpha}} \left(u_1^{-\alpha} + u_2^{-\alpha} - 1\right)^{-2-\frac{1}{\alpha}}. \tag{12.58}$$

Frank's copula, denoted by $C_F(u_1, u_2)$, is defined as

$$C_F(u_1, u_2) = -\frac{1}{\alpha} \log\left[1 + \frac{\left(e^{-\alpha u_1} - 1\right)\left(e^{-\alpha u_2} - 1\right)}{e^{-\alpha} - 1}\right], \quad \alpha \neq 0, \tag{12.59}$$

which has the following copula density

$$c_F(u_1, u_2) = \frac{\alpha e^{-\alpha(u_1+u_2)}\left(1 - e^{-\alpha}\right)}{\left[e^{-\alpha(u_1+u_2)} - e^{-\alpha u_1} - e^{-\alpha u_2} + e^{-\alpha}\right]^2}. \tag{12.60}$$

Another popular copula is the **Gaussian copula** defined by

$$C_G(u_1, u_2) = \Psi_\alpha(\Phi^{-1}(u_1), \Phi^{-1}(u_2)), \quad -1 < \alpha < 1, \tag{12.61}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal df and $\Psi_\alpha(\cdot, \cdot)$ is the df of a standard bivariate normal variate with correlation coefficient α. The Gaussian copula density is

$$c_G(u_1, u_2) = \frac{1}{\sqrt{1-\alpha^2}} \exp\left[-\frac{\eta_1^2 - 2\alpha\eta_1\eta_2 + \eta_2^2}{2(1-\alpha^2)}\right] \exp\left[\frac{\eta_1^2 + \eta_2^2}{2}\right], \tag{12.62}$$

where $\eta_i = \Phi^{-1}(u_i)$, for $i = 1, 2$.

Example 12.19 Let $X_1 \sim \mathcal{W}(0.5, 2)$ and $X_2 \sim \mathcal{G}(3, 2)$, and assume that Clayton's copula with parameter α fits the bivariate distribution of X_1 and X_2. Determine the probability $p = \Pr(X_1 \leq \mathrm{E}(X_1), X_2 \leq \mathrm{E}(X_2))$ for $\alpha = 0.001$, 1, 2, 3, and 10.

Solution The means of X_1 and X_2 are

$$\mathrm{E}(X_1) = 2\Gamma(3) = 4 \quad \text{and} \quad \mathrm{E}(X_2) = (2)(3) = 6.$$

Let $u_1 = F_{X_1}(4) = 0.7569$ and $u_2 = F_{X_2}(6) = 0.5768$, so that

$$\begin{aligned} p &= \Pr(X_1 \leq 4, X_2 \leq 6) \\ &= C_C(0.7569, 0.5768) \\ &= \left[(0.7569)^{-\alpha} + (0.5768)^{-\alpha} - 1\right]^{-\frac{1}{\alpha}}. \end{aligned}$$

The computed values of p are

α	0.001	1	2	3	10
p	0.4366	0.4867	0.5163	0.5354	0.5734

Note that when X_1 and X_2 are independent, $p = (0.7569)(0.5768) = 0.4366$, which corresponds to the case where α approaches 0. The dependence between X_1 and X_2 increases with α, as can be seen from the numerical results. □

12.6 Excel computation notes

The Excel Solver can solve for the root of a nonlinear equation as well as the maximum or minimum of a nonlinear function. It is useful for computing the parameter estimates that require the solution of a nonlinear equation. The

example below illustrates the solution of the parameter estimates as discussed in Example 12.6.

Example 12.20 Refer to Example 12.6, in which the parameters of the $\mathcal{P}(\alpha, \gamma)$ distribution are estimated using the quantile-matching method. Suppose the 25th and 75th percentiles are used and the sample estimates are $\hat{x}_{0.25} = 0.52$ and $\hat{x}_{0.75} = 2.80$. Estimate the values of α and γ.

Solution Using the results in Example 12.6, with $\delta_1 = 0.25$ and $\delta_2 = 0.75$, the quantile-matching estimate of γ is the solution of the following equation

$$\frac{\log\left(\dfrac{\gamma}{0.52+\gamma}\right)}{\log\left(\dfrac{\gamma}{2.80+\gamma}\right)} - \frac{\log(1-0.25)}{\log(1-0.75)} = 0.$$

Figure 12.1 illustrates the solution of γ in this equation. First, a guess value is entered in cell A1. Second, the expression on the left-hand side of the above equation is entered in cell A2. The Solver is then called up, with target cell set to A2 for a value of zero by changing the value in cell A1. The solution is found to be 8.9321. Hence, the estimate of α is computed as

$$\hat{\alpha} = \frac{\log(0.75)}{\log\left(\dfrac{8.9321}{0.52+8.9321}\right)} = 5.0840.$$

□

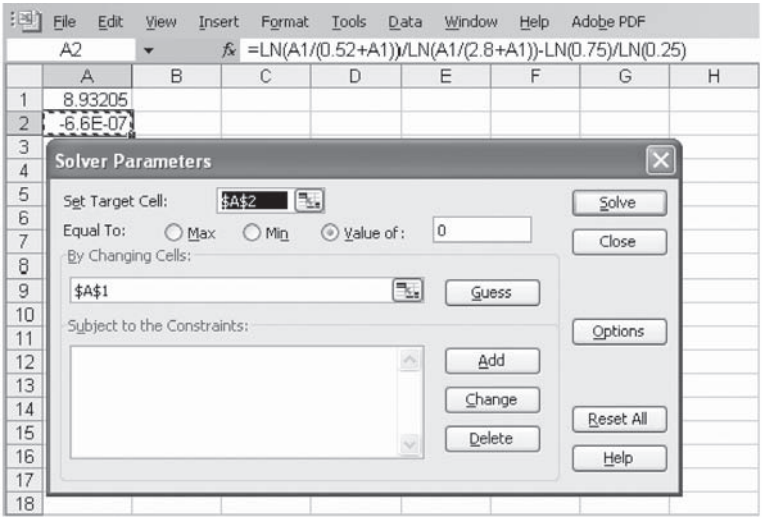

Figure 12.1 Excel computation of Example 12.20

For computing the maximum likelihood estimates, the Excel Solver may be used to solve for the root of the first-order condition as stated in equation (12.15). It may also be used to compute the maximum of the log-likelihood function directly. Furthermore, the Solver can be applied to multivariate functions. The example below illustrates the maximization of the log-likelihood function as discussed in Example 12.13.

Example 12.21 Refer to Example 12.13, in which the parameters of the $\mathcal{W}(\alpha, \lambda)$ distribution are estimated using the maximum likelihood method based on the data of the Block 1 business in Example 12.12.

Solution As discussed in Example 12.12, there are eight observable ground-up losses and the log-likelihood function is

$$(\alpha - 1) \sum_{i=1}^{8} \log x_i + 8 (\log \alpha - \alpha \log \lambda) - \sum_{i=1}^{8} \left(\frac{x_i}{\lambda}\right)^{\alpha} + 8 \left(\frac{1.4}{\lambda}\right)^{\alpha},$$

The computation is illustrated in Figure 12.2. The eight observations are entered in cells A1 through A8. Cells B1 through B8 compute the logarithm of the corresponding cells in A1 through A8, with the total computed in cell B9. The (initial) values of α and λ are entered in cells A10 and A11, respectively. Cells C1 through C8 compute $(x_i/\lambda)^{\alpha}$, with their sum given in cell C9. Finally, the log-likelihood function defined by the above equation is computed in cell C11. The Solver is called up, with the target cell set to C11, which is to be maximized by changing the values in cells A10 and A11. The answers are shown to be $\hat{\alpha} = 2.6040$ and $\hat{\lambda} = 3.1800$. □

12.7 Summary and discussions

We have discussed various methods of estimating the parameters of a failure-time or loss distribution. The maximum likelihood estimation method is by far the most popular, as it is asymptotically normal and efficient for many standard cases. Furthermore, it can be applied to data which are complete or incomplete, provided the likelihood function is appropriately defined. The asymptotic standard errors of the MLE can be computed in most standard estimation packages, from which interval estimates can be obtained. For models with covariates, we have surveyed the use of the proportional hazards model, the generalized linear model, and the accelerated failure-time model. The partial likelihood method provides a convenient way to estimate the regression coefficients of the proportional hazards model without estimating the baseline model. Finally, we have introduced the use of copula in modeling multivariate data. This method provides a simple approach to extend the adopted univariate models to fit the multivariate observations.

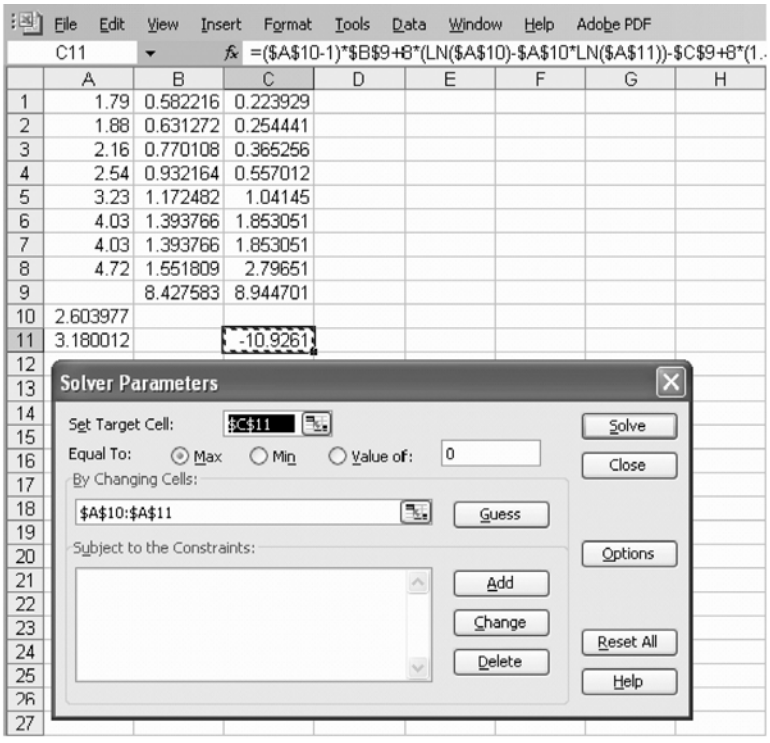

Figure 12.2 Excel computation of Example 12.21

Exercises

12.1 A random sample of n insurance claims with a deductible of d has a ground-up loss distribution following $\mathcal{P}(\alpha, \gamma)$. If the value of α is known, derive the method-of-moments estimate of γ by matching the first-order moment.

12.2 The 25th and 75th percentiles of a sample of losses are 6 and 15, respectively. What is $\text{VaR}_{0.99}$ of the loss, assuming the loss distribution is (a) Weibull, and (b) lognormal?

12.3 The median and the 90th percentile of a sample from a Weibull distribution are 25,000 and 260,000, respectively. Compute the mean and the standard deviation of the distribution.

12.4 The following observations are obtained for a sample of losses:

$$20, 23, 30, 31, 39, 43, 50, 55, 67, 72.$$

Estimate the parameters of the distribution using the method of moments if the losses are distributed as (a) $\mathcal{G}(\alpha, \beta)$ and (b) $\mathcal{U}(a, b)$.

12.5 A random sample of ten observations of X from a $\mathcal{L}(\mu, \sigma^2)$ distribution is as follows:

$$45, 49, 55, 56, 63, 71, 78, 79, 82, 86.$$

(a) Estimate μ and σ^2 by matching the moments of X, and hence estimate the probability of X exceeding 80.
(b) Estimate μ and σ^2 by matching the moments of $\log X$, and hence estimate the probability of X exceeding 80.

12.6 The following observations are obtained from a right-censored loss distribution with maximum covered loss of 18 (18^+ denotes a right-censored observation):

$$4, 4, 7, 10, 14, 15, 16, 18^+, 18^+, 18^+.$$

Estimate the parameters of the loss distribution using the method of moments if the losses are distributed as (a) $\mathcal{U}(0, b)$, (b) $\mathcal{E}(\lambda)$, and (c) $\mathcal{P}(2, \gamma)$.

12.7 Let X be a mixture distribution with 40% chance of $\mathcal{E}(\lambda_1)$ and 60% chance of $\mathcal{E}(\lambda_2)$. If the mean and the variance of X are estimated to be 4 and 22, respectively, estimate λ_1 and λ_2 using the method of moments.

12.8 The following observations are obtained from a loss distribution:

$$11, 15, 16, 23, 28, 29, 31, 35, 40, 42, 44, 48, 52, 53, 55, 59, 60, 65, 65, 71.$$

(a) Estimate $\text{VaR}_{0.95}$, assuming the loss follows a lognormal distribution, with parameters estimated by matching the 25th and 75th percentiles.
(b) Estimate $\text{VaR}_{0.90}$, assuming the loss follows a Weibull distribution, with parameters estimated by matching the 30th and 70th percentiles.

12.9 Let X be distributed with pdf $f_X(x)$ given by

$$f_X(x) = \frac{2(\theta - x)}{\theta^2}, \qquad \text{for } 0 < x < \theta,$$

and 0 elsewhere. Estimate θ using the method of moments.

12.10 Let $\{X_1, \cdots, X_n\}$ and $\{Y_1, \cdots, Y_m\}$ be independent random samples from normal populations with means μ_X and μ_Y, respectively, and common variance σ^2. Determine the maximum likelihood estimators of μ_X, μ_Y, and σ^2.

12.11 For a loss variable following the $\mathcal{GM}(\theta)$ distribution, the maximum likelihood estimator of θ based on a random sample of n observations

is $\hat{\theta} = 1/(1 + \bar{x})$. What is the variance of $\bar{x}$? Derive the asymptotic variance of $\hat{\theta}$ using the delta method.

12.12 Let $\{X_1, \cdots, X_n\}$ be a random sample from $\mathcal{U}(\theta - 0.5, \theta + 0.5)$. Determine the maximum likelihood estimator of θ and show that it is not unique. If $X_{(i)}$ denotes the ith-order statistic of the sample, determine whether the following estimators are maximum likelihood estimators of θ: (a) $(X_{(1)} + X_{(n)})/2$ and (b) $(X_{(1)} + 2X_{(n)})/3$.

12.13 The following grouped observations are available for a loss random variable X:

Interval	$(0, 2]$	$(2, 4]$	$(4, 6]$	$(6, 8]$	$(8, \infty)$
No of obs	4	7	10	6	3

Determine the log-likelihood of the sample if X is distributed as (a) $\mathcal{E}(\lambda)$ and (b) $\mathcal{P}(\alpha, \gamma)$.

12.14 You are given the following loss observations, where a policy limit of 28 has been imposed:

8, 10, 13, 14, 16, 16, 19, 24, 26, 28, 28, 28.

(a) If the losses are distributed as $\mathcal{U}(0, b)$, determine the maximum likelihood estimate of b assuming all payments of 28 have losses exceeding the policy limit. How would your result be changed if only two of these payments have losses exceeding the policy limit?

(b) If the losses are distributed as $\mathcal{E}(\lambda)$, determine the maximum likelihood estimate of λ assuming all payments of 28 have losses exceeding the policy limit.

12.15 Refer to Example 12.13. Using the Excel Solver, verify that the maximum likelihood estimates of the shifted model are $\tilde{\alpha} = 1.5979$ and $\tilde{\lambda} = 1.8412$.

12.16 An insurance policy has a deductible of 4 and maximum covered loss of 14. The following five observations of payment are given, including two censored payments: 3, 6, 8, 10, and 10.

(a) If the ground-up losses are distributed as $\mathcal{E}(\lambda)$, compute the maximum likelihood estimate of λ.

(b) If the payment in a payment event is distributed as $\mathcal{E}(\lambda^*)$, compute the maximum likelihood estimate of λ^*.

(c) Compute the maximum likelihood estimates of λ and λ^* if the ground-up loss and the shifted loss distributions are, respectively, $\mathcal{W}(3, \lambda)$ and $\mathcal{W}(3, \lambda^*)$.

12.17 The following observations are obtained from the distribution $X \sim \mathcal{E}(\lambda)$: 0.47, 0.49, 0.91, 1.00, 2.47, 5.03, and 16.09. Compute the maximum likelihood estimate of $x_{0.25}$.

12.18 A sample of five loss observations are: 158, 168, 171, 210, and 350. The df $F_X(x)$ of the loss variable X is

$$F_X(x) = 1 - \left(\frac{100}{x}\right)^{\alpha}, \qquad \text{for } x > 100,$$

where $\alpha > 0$. Compute the maximum likelihood estimate of α.

12.19 Suppose $\{X_1, \cdots, X_n\}$ is a random sample from the $\mathcal{N}(\mu, \sigma^2)$ distribution.

(a) If σ^2 is known, what is the maximum likelihood estimator of μ? Using Theorem 12.1, derive the asymptotic distribution of the maximum likelihood estimator of μ.

(b) If μ is known, what is the maximum likelihood estimator of σ^2? Using Theorem 12.1, derive the asymptotic distribution of the maximum likelihood estimator of σ^2.

(c) If μ and σ^2 are both unknown, what are the maximum likelihood estimators of μ and σ^2? Using Theorem 12.2, derive the joint asymptotic distribution of the maximum likelihood estimators of μ and σ^2.

12.20 Suppose $\{X_1, \cdots, X_n\}$ is a random sample from the $\mathcal{N}(\mu, \sigma^2)$ distribution. Determine the asymptotic variance of $\bar{X}^3$.

12.21 Suppose $\{X_1, \cdots, X_n\}$ is a random sample from the $\mathcal{N}(0, \sigma^2)$ distribution. Determine the asymptotic variance of $n/\left(\sum_{i=1}^{n} X_i^2\right)$.

12.22 Suppose $X \sim \mathcal{E}(\lambda)$ and a random sample of X has the following observations:

$$14, 17, 23, 25, 25, 28, 30, 34, 40, 41, 45, 49.$$

(a) Compute the maximum likelihood estimate of λ.

(b) Estimate the variance of the maximum likelihood estimate of λ.

(c) Compute an approximate 95% confidence interval of λ. Do you think this confidence interval is reliable?

(d) Estimate the standard deviation of the variance estimate in (b).

(e) Estimate $\Pr(X > 25)$ and determine a 95% linear confidence interval of this probability.

12.23 Refer to the loss distribution X in Example 12.14. You are given the following losses in Block 1: 3, 5, 8, 12; and the following losses in Block 2: 2, 3, 5, 5, 7. The proportional hazards model is assumed.

(a) Compute the maximum likelihood estimates of β and λ, assuming the baseline distribution is exponential.

(b) Compute the estimate of β using the partial likelihood method, hence estimate the baseline survival function at the observed loss values. Estimate $\Pr(X > 6)$ for a Block 1 risk.

12.24 Express the survival functions of $T(x; z_i)$ and $T(x; 0)$ as defined in equation (12.49) in terms of $S(\cdot; \theta, z_i)$. Hence, prove equation (12.50).

12.25 Suppose $X \sim \mathcal{PN}(\lambda)$ and a generalized linear model has $\lambda = \exp(\beta z)$ where z is a single index. The following data are given:

x_i	4	6	3	7	5	9	4	10
z_i	1	3	1	2	2	3	1	3

Determine the log-likelihood function of the sample and the maximum likelihood estimate of β.

12.26 Refer to Example 12.18. Suppose the covariate is a scalar index z and the following data are given:

x_i	10	26	13	17	25	29	34	40	21	12
z_i	2	2	2	2	1	1	1	1	2	2

(a) Determine the log-likelihood function of the sample if the baseline distribution is $\mathcal{E}(\lambda)$, and estimate the parameters of the model.

(b) Determine the log-likelihood function of the sample if the baseline distribution is $\mathcal{P}(3, \gamma)$, and estimate the parameters of the model.

12.27 Derive equations (12.58) and (12.60).

12.28 Suppose $X_1 \sim \mathcal{E}(\lambda_1)$, $X_2 \sim \mathcal{E}(\lambda_2)$, and Clayton's copula applies to the joint df of X_1 and X_2. You are given the following data:

x_{1i}	2.5	2.6	1.3	2.7	1.4	2.9	4.5	5.3
x_{2i}	3.5	4.2	1.9	4.8	2.6	5.6	6.8	7.8

(a) Compute the maximum likelihood estimates of the model.

(b) Estimate $\Pr(X_1 \leq 3, X_2 \leq 4)$.

Questions adapted from SOA exams

12.29 There were 100 insurance claims in Year 1, with an average size of 10,000. In Year 2, there were 200 claims, with an average claim size of 12,500. Inflation is found to be 10% per year, and the claim-size distribution follows a $\mathcal{P}(3, \gamma)$ distribution. Estimate γ for Year 3 using the method of moments.

12.30 A random sample of three losses from the $\mathcal{P}(\alpha, 150)$ distribution has the values: 225, 525, and 950. Compute the maximum likelihood estimate of α.

12.31 Suppose $\{X_1, \cdots, X_n\}$ is a random sample from the $\mathcal{L}(\mu, \sigma^2)$ distribution. The maximum likelihood estimates of the parameters are $\hat{\mu} = 4.2150$ and $\hat{\sigma}^2 = 1.1946$. The estimated variance of $\hat{\mu}$ is 0.1195 and that of $\hat{\sigma}^2$ is 0.2853, with estimated covariance between $\hat{\mu}$ and $\hat{\sigma}^2$ being zero. The mean of $\mathcal{L}(\mu, \sigma^2)$ is $\mu_X = \exp(\mu + \sigma^2/2)$. Estimate the variance of the maximum likelihood estimate of μ_X. [*Hint*: see equation (A.200) in the Appendix.]

12.32 The pdf of X is

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

where $\theta > 0$. If the maximum likelihood estimate of θ based on a sample of 40 observations is $\hat{\theta} = 2$, estimate the mean squared error of $\hat{\theta}$.

12.33 The duration of unemployment is assumed to follow the proportional hazards model, where the baseline distribution is exponential and the covariate z is years of education. When $z = 10$, the mean duration of unemployment is 0.2060 years. When $z = 25$, the median duration of unemployment is 0.0411 years. If $z = 5$, what is the probability that the unemployment duration exceeds one year?

12.34 A proportional hazards model has covariate $z_1 = 1$ for males and $z_1 = 0$ for females, and $z_2 = 1$ for adults and $z_2 = 0$ for children. The maximum likelihood estimates of the coefficients are $\hat{\beta}_1 = 0.25$ and $\hat{\beta}_2 = -0.45$. The covariance matrix of the estimators is

$$\text{Var}(\hat{\beta}_1, \hat{\beta}_2) = \begin{pmatrix} 0.36 & 0.10 \\ 0.10 & 0.20 \end{pmatrix}.$$

Determine a 95% confidence interval of $\beta_1 - \beta_2$.

12.35 The ground-up losses of dental claims follow the $\mathcal{E}(\lambda)$ distribution. There is a deductible of 50 and a policy limit of 350. A random sample of claim payments has the following observations: 50, 150, 200, 350^+, and 350^+, where $^+$ indicates that the original loss exceeds 400. Determine the likelihood function.

12.36 A random sample of observations is taken from the shifted exponential distribution with the following pdf

$$f(x) = \frac{1}{\theta} e^{-\frac{x-\delta}{\theta}}, \quad \delta < x < \infty.$$

The sample mean and median are 300 and 240, respectively. Estimate δ by matching these two sample quantities to the corresponding population quantities.

12.37 The distribution of the number of claims per policy during a one-year period for a block of 3,000 insurance policies is:

Number of claims per policy	Number of policies
0	1,000
1	1,200
2	600
3	200
≥ 4	0

A Poisson model is fitted to the number of claims per policy using the maximum likelihood estimation method. Determine the lower end point of the large-sample 90% linear confidence interval of the mean of the distribution.

12.38 A Cox proportional hazards model was used to study the losses on two groups of policies. A single covariate z was used with $z = 0$ for a policy in Group 1 and $z = 1$ for a policy in Group 2. The following losses were observed:

Group 1: 275, 325, 520,
Group 2: 215, 250, 300.

The baseline survival function is

$$S(x) = \left(\frac{200}{x}\right)^{\alpha}, \quad x > 200, \ \alpha > 0.$$

Compute the maximum likelihood estimate of the coefficient β of the proportional hazards model.

12.39 Losses follow the $\mathcal{W}(\alpha, \lambda)$ distribution, and the following observations are given:

54, 70, 75, 81, 84, 88, 97, 105, 109, 114, 122, 125, 128, 139, 146, 153.

The model is estimated by percentile matching using the 20th and 70th smoothed empirical percentiles. Calculate the estimate of λ.

12.40 An insurance company records the following ground-up loss amounts from a policy with a deductible of 100 (losses less than 100 are not reported):

120, 180, 200, 270, 300, 1000, 2500.

Assuming the losses follow the $\mathcal{P}(\alpha, 400)$ distribution, use the maximum likelihood estimate of α to estimate the expected loss with no deductible.

12.41 Losses are assumed to follow an inverse exponential distribution with df

$$F(x) = e^{-\frac{\theta}{x}}, \quad x > 0.$$

Three losses of amounts: 186, 91, and 66 are observed and seven other amounts are known to be less than or equal to 60. Calculate the maximum likelihood estimate of the population mode.

12.42 At time 4, there are five working light bulbs. The five bulbs are observed for another duration p. Three light bulbs burn out at times 5, 9, and 13, while the remaining light bulbs are still working at time $4 + p$. If the distribution of failure time is $\mathcal{U}(0, \omega)$ and the maximum likelihood estimate of ω is 29, determine p.

12.43 A Cox proportional hazards model was used to compare the fuel efficiency of traditional and hybrid cars. A single covariate z was used with $z = 0$ for a traditional car and $z = 1$ for a hybrid car. The following observed miles per gallon are given:

Traditional: 22, 25, 28, 33, 39,
Hybrid: 27, 31, 35, 42, 45.

The partial likelihood estimate of the coefficient β of the proportional hazards model is -1. Calculate the estimate of the baseline cumulative hazard function $H(32)$ using an analogue of the Nelson–Aalen estimator.

12.44 Losses have the following df

$$F(x) = 1 - \frac{\theta}{x}, \quad \theta < x < \infty.$$

A sample of 20 losses resulted in the following grouped distribution:

Interval	Number of losses
$x \leq 10$	9
$10 < x \leq 25$	6
$x > 25$	5

Calculate the maximum likelihood estimate of θ.

13

Model evaluation and selection

After a model has been estimated, we have to evaluate it to ascertain that the assumptions applied are acceptable and supported by the data. This should be done prior to using the model for prediction and pricing. Model evaluation can be done using graphical methods, as well as formal misspecification tests and diagnostic checks.

Nonparametric methods have the advantage of using minimal assumptions and allowing the data to determine the model. However, they are more difficult to analyze theoretically. On the other hand, parametric methods are able to summarize the model in a small number of parameters, albeit with the danger of imposing the wrong structure and oversimplification. Using graphical comparison of the estimated df and pdf, we can often detect if the estimated parametric model has any abnormal deviation from the data.

Formal misspecification tests can be conducted to compare the estimated model (parametric or nonparametric) against a hypothesized model. When the key interest is the comparison of the df, we may use the Kolmogorov–Smirnov test and Anderson–Darling test. The chi-square goodness-of-fit test is an alternative for testing distributional assumptions, by comparing the observed frequencies against the theoretical frequencies. The likelihood ratio test is applicable to testing the validity of restrictions on a model, and can be used to decide if a model can be simplified.

When several estimated models pass most of the diagnostics, the adoption of a particular model may be decided using some information criteria, such as the Akaike information criterion or the Schwarz information criterion.

Learning objectives

1 Graphical presentation and comparison
2 Misspecification tests and diagnostic checks
3 Kolmogorov–Smirnov test and Anderson–Darling test

4 Likelihood ratio test and chi-square goodness-of-fit test
5 Model selection and information criteria

13.1 Graphical methods

For complete individual observations the empirical df in equation (11.2) provides a consistent estimate of the df without imposing any parametric assumption. When parametric models are assumed, the parameters may be estimated by MLE or other methods. However, for the estimates to be consistent for the true values, the pdf assumed has to be correct. One way to assess if the assumption concerning the distribution is correct is to plot the estimated parametric df against the empirical df. If the distributional assumption is incorrect, we would expect the two plotted graphs to differ.

For exposition purpose, we denote $\hat{F}(\cdot)$ as an estimated df using the nonparametric method, such as the empirical df and the Kaplan–Meier estimate, and $F^*(\cdot)$ as a hypothesized df or parametrically estimated df. Thus, $\hat{F}(\cdot)$ is assumed to be an estimate which is entirely data based, and $F^*(\cdot)$ (and the assumption underlying it) is assessed against $\hat{F}(\cdot)$.

Example 13.1 A sample of 20 loss observations are as follows:

0.003, 0.012, 0.180, 0.253, 0.394, 0.430, 0.491, 0.743, 1.066, 1.126,
1.303, 1.508, 1.740, 4.757, 5.376, 5.557, 7.236, 7.465, 8.054, 14.938.

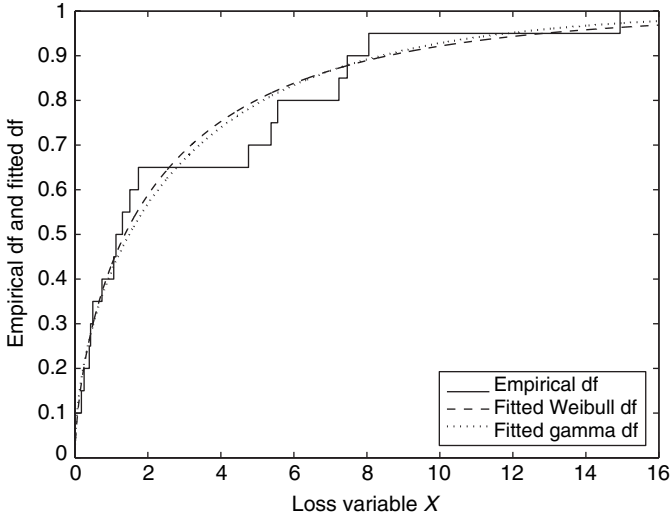

Figure 13.1 Empirical df and estimated parametric df of Example 13.1

Two parametric models are fitted to the data using the MLE, assuming that the underlying distribution is (a) $\mathcal{W}(\alpha, \lambda)$ and (b) $\mathcal{G}(\alpha, \beta)$. The fitted models are $\mathcal{W}(0.6548, 2.3989)$ and $\mathcal{G}(0.5257, 5.9569)$. Compare the empirical df against the df of the two estimated parametric models.

Solution The plots of the empirical df and the estimated parametric df are given in Figure 13.1. It can be seen that both estimated parametric models fit the data quite well. Thus, from the df plots it is difficult to ascertain which is the preferred model. □

Another useful graphical device is the *p-p* plot. Suppose the sample observations $x = (x_1, \cdots, x_n)$ are arranged in increasing order $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, so that $x_{(i)}$ is the ith-order statistic. We approximate the probability of having an observation less than or equal to $x_{(i)}$ using the sample data by the sample proportion $p_i = i/(n+1)$.[1] Now the hypothesized or parametrically estimated probability of an observation less than or equal to $x_{(i)}$ is $F^*(x_{(i)})$. A plot of $F^*(x_{(i)})$ against p_i is called the *p-p* plot. If $F^*(\cdot)$ fits the data well, the *p-p* plot should approximately follow the 45-degree line.

Example 13.2 For the data and the fitted Weibull model in Example 13.1, assess the model using the *p-p* plot.

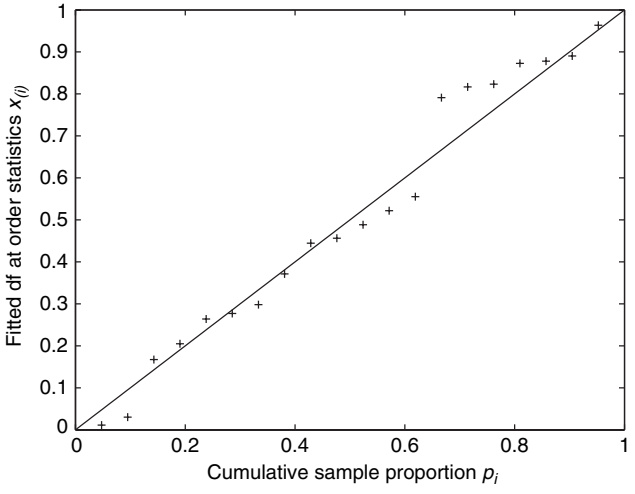

Figure 13.2 *p-p* plot of the estimated Weibull distribution in Example 13.2

Solution The *p-p* plot is presented in Figure 13.2. It can be seen that most points lie closely to the 45-degree line, apart from some deviations around $p_i = 0.7$. □

[1] Another definition of sample proportion can be found in Footnote 1 of Chapter 11.

Another graphical method equivalent to the *p-p* plot is the *q-q* plot. In a *q-q* plot, $F^{*-1}(p_i)$ is plotted against $x_{(i)}$. If $F^*(\cdot)$ fits the data well, the *q-q* plot should approximately follow a straight line.

When the data include incomplete observations that are right censored, the empirical df is an inappropriate estimate of the df. We replace it by the Kaplan–Meier estimate. Parametric estimate of the df can be computed using MLE, with the log-likelihood function suitably defined, as discussed in Section 12.3. The estimated df based on the MLE may then be compared against the Kaplan–Meier estimate.

Example 13.3 For the data in Example 13.1, assume that observations larger than 7.4 are censored. Compare the estimated df based on the MLE under the Weibull assumption against the Kaplan–Meier estimate.

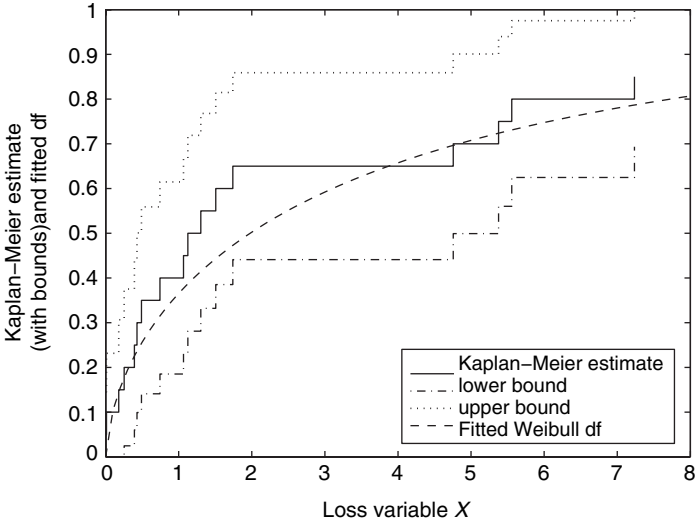

Figure 13.3 Estimated Kaplan–Meier df and estimated Weibull df of Example 13.3

Solution In the data set three observations are larger than 7.4 and are censored. The plots of the Kaplan–Meier estimate and the estimated df using the MLE of the Weibull model are given in Figure 13.3. For the Kaplan–Meier estimate we also plot the lower and upper bounds of the 95% confidence interval estimates of the df. It can be seen that the estimated parametric df falls inside the band of the estimated Kaplan–Meier estimate. □

For grouped data, the raw sample data may be summarized using a histogram, which may be compared against the pdf of a hypothesized or parametrically

fitted model. For estimation using MLE the log-likelihood given by equation (12.22) should be used. If the method of moments is adopted, the sample moments are those of the empirical distribution, as given by equation (11.80). These moments are then equated to the population moments and the model parameter estimates are computed.

Example 13.4 A sample of grouped observations of losses are summarized as follows, with notations given in Section 11.3:

$(c_{j-1}, c_j]$	(0, 6]	(6, 12]	(12, 18]	(18, 24]	(24, 30]	(30, 50]
n_j	3	9	9	4	3	2

Assuming the losses are distributed as $\mathcal{G}(\alpha, \beta)$, estimate α and β using the MLE and the method of moments. Compare the pdf of the estimated parametric models against the sample histogram.

Solution To compute the method-of-moments estimates, we first calculate the sample moments. We have

$$\hat{\mu}_1 = \sum_{j=1}^{6} \frac{n_j}{n} \left[\frac{c_j + c_{j-1}}{2} \right] = 15.6667$$

and

$$\hat{\mu}_2 = \sum_{j=1}^{6} \frac{n_j}{n} \left[\frac{c_j^3 - c_{j-1}^3}{3(c_j - c_{j-1})} \right] = 336.0889.$$

Thus, from Example 12.2, we have

$$\tilde{\beta} = \frac{336.0889 - (15.6667)^2}{15.6667} = 5.7857$$

and

$$\tilde{\alpha} = \frac{15.6667}{5.7857} = 2.7078.$$

The variance of this distribution is 90.6418. On the other hand, using the log-likelihood function of the grouped data as given in equation (12.22), the MLE of α and β are computed as[2]

$$\hat{\alpha} = 3.2141 \qquad \text{and} \qquad \hat{\beta} = 4.8134.$$

[2] Note that c_6 is now taken as ∞ for the computation of the log-likelihood.

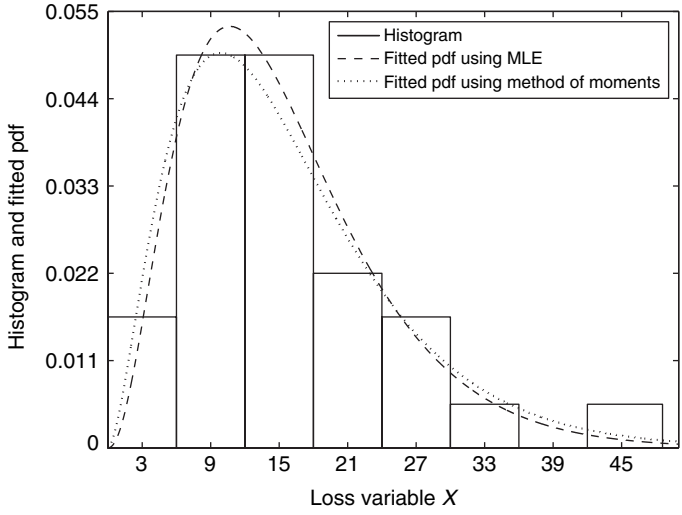

Figure 13.4 Estimated pdf versus sample histogram of Example 13.4

The mean and variance of this distribution are 15.4707 and 74.4669, respectively. Thus, the mean of the two estimated distributions are similar, while the variance of the distribution estimated by the MLE is much lower.

The plots of the estimated pdf versus the sample **histogram** are given in Figure 13.4.[3] It appears that the fitted pdf based on the MLE performs better than that based on the method of moments. □

13.2 Misspecification tests and diagnostic checks

While graphical methods are able to provide visual aids to assess whether a model fits the data, they do not provide *quantitative measures* about any possible deviations that are not commensurate with the hypothesized model. A formal **significance test** has the advantage of providing a probabilistic assessment of whether a decision concerning the model is likely to be wrong.

A formal significance test may be set up by establishing a **null hypothesis** about the distribution of the losses. This hypothesis is tested against the data using a **test statistic**. Given an assigned **level of significance**, we can determine a **critical region** such that if the test statistic falls inside the critical region, we conclude that the data do not support the null hypothesis. Alternatively, we

[3] The histogram is plotted with additional information about the two right-hand tail observations not given in the question.

can compute the probability of obtaining a test statistic more *extreme* than the computed value if the null hypothesis is correct, and call this value the ***p*-value**. The smaller the *p*-value, the more unlikely the null hypothesis is correct. We call statistical significance tests that aim at examining the model's distributional assumptions **misspecification tests**. They are also **diagnostic checks** for the model assumption before we use the model for pricing or other analysis.

In this section we discuss several misspecification tests for the model assumptions of loss distributions.

13.2.1 Kolmogorov–Smirnov test

We specify a null hypothesis about the df of a continuous loss variable, and denote it by $F^*(\cdot)$. To examine if the data support the null hypothesis, we compare $F^*(\cdot)$ against the empirical df $\hat{F}(\cdot)$ and consider the statistic

$$\max_{x_{(1)} \le x \le x_{(n)}} |\, \hat{F}(x) - F^*(x) \,|, \tag{13.1}$$

where $x_{(1)}$ and $x_{(n)}$ are the minimum and maximum of the observations, respectively. However, as $\hat{F}(\cdot)$ is a right-continuous increasing step function and $F^*(\cdot)$ is also increasing we only need to compare the differences at the observed data points, namely at the order statistics $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. Furthermore, the maximum may only occur at a jump point $x_{(i)}$ or immediately to the left of it. We now denote the statistic in expression (13.1) by D, which is called the **Kolmogorov–Smirnov statistic** and can be written as[4]

$$D = \max_{i \in \{1,\cdots,n\}} \left\{ \max \left\{ |\, \hat{F}(x_{(i-1)}) - F^*(x_{(i)}) \,|, |\, \hat{F}(x_{(i)}) - F^*(x_{(i)}) \,| \right\} \right\}, \tag{13.2}$$

where $\hat{F}(x_{(0)}) \equiv 0$. When we have complete individual observations

$$\hat{F}(x_{(i)}) = \frac{i}{n}, \tag{13.3}$$

and D can be written as

$$D = \max_{i \in \{1,\cdots,n\}} \left\{ \max \left\{ \left| \frac{i-1}{n} - F^*(x_{(i)}) \right|, \left| \frac{i}{n} - F^*(x_{(i)}) \right| \right\} \right\}. \tag{13.4}$$

If the true df of the data is not $F^*(\cdot)$, we would expect D to be large. Given a level of significance α, the null hypothesis that the data have df $F^*(\cdot)$ is rejected

[4] The Kolmogorov–Smirnov statistic was proposed by Kolmogorov in 1933 and developed by Smirnov in 1939.

if D is larger than the critical value. When $F^*(\cdot)$ is completely specified (as in the case where there is no unknown parameter in the df) the critical values of D for some selected values of α are given as follows

Level of significance α	0.10	0.05	0.01
Critical value	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Note that the critical values above apply to all df, as long as they are completely specified. Any unknown parameters in the df, however, have to be estimated for the computation of $F^*(x_{(i)})$. Then the critical values above will not apply. There are, however, some critical values in the literature that are estimated using Monte Carlo methods for different null distributions with unknown parameter values.[5]

When the data are left truncated at point d and right censored at point u, the statistic in (13.1) is modified to

$$\max_{d \le x \le u} |\hat{F}(x) - F^*(x)|. \tag{13.5}$$

Equation (13.2) still applies, and the order statistics are within the range (d, u). If some parameters have to be estimated, one of the methods in Chapter 12 can be used. When the observations are not complete, the critical values should be revised downwards. The actual value used, however, has to be estimated by Monte Carlo methods.

Example 13.5 Compute the Kolmogorov–Smirnov statistics for the data in Example 13.1, with the estimated Weibull and gamma models as the hypothesized distributions.

Solution We denote $F_1^*(x_{(j)})$ as the df of $\mathcal{W}(0.6548, 2.3989)$ evaluated at $x_{(j)}$, and $F_2^*(x_{(j)})$ as the df of $\mathcal{G}(0.5257, 5.9569)$ evaluated at $x_{(j)}$, which are the df of the estimated Weibull and gamma distributions in Example 13.1. We further denote

$$D_{ij} = \max\left\{\left|\hat{F}(x_{(j-1)}) - F_i^*(x_{(j)})\right|, \left|\hat{F}(x_{(j)}) - F_i^*(x_{(j)})\right|\right\},$$

for $i = 1, 2$, and $j = 1, \cdots, 20$. Note that

$$\hat{F}(x_{(j)}) = \frac{j}{20},$$

as we have complete individual data. The results are summarized in Table 13.1.

[5] See Stephens (1974) for some critical values estimated by Monte Carlo methods.

Table 13.1. *Results for Example 13.5*

$\hat{F}(x_{(j)})$	$F_1^*(x_{(j)})$	D_{1j}	$F_2^*(x_{(j)})$	D_{2j}
0.0500	0.0112	0.0388	0.0191	0.0309
0.1000	0.0301	0.0699	0.0425	0.0575
0.1500	0.1673	0.0673	0.1770	0.0770
0.2000	0.2050	0.0550	0.2112	0.0612
0.2500	0.2639	0.0639	0.2643	0.0643
0.3000	0.2771	0.0271	0.2761	0.0261
0.3500	0.2981	0.0519	0.2951	0.0549
0.4000	0.3714	0.0286	0.3617	0.0383
0.4500	0.4445	0.0445	0.4294	0.0294
0.5000	0.4563	0.0437	0.4405	0.0595
0.5500	0.4885	0.0615	0.4710	0.0790
0.6000	0.5218	0.0782	0.5030	0.0970
0.6500	0.5553	0.0947	0.5357	0.1143
0.7000	0.7910	**0.1410**	0.7813	**0.1313**
0.7500	0.8166	0.1166	0.8097	0.1097
0.8000	0.8233	0.0733	0.8172	0.0672
0.8500	0.8726	0.0726	0.8728	0.0728
0.9000	0.8779	0.0279	0.8789	0.0289
0.9500	0.8903	0.0597	0.8929	0.0571
1.0000	0.9636	0.0364	0.9728	0.0272

The Kolmogorov–Smirnov statistics D for the Weibull and gamma distributions are, 0.1410 and 0.1313, respectively, both occurring at $x_{(14)}$. The critical value of D at the level of significance of 10% is

$$\frac{1.22}{\sqrt{20}} = 0.2728,$$

which is larger than the computed D for both models. However, as the hypothesized df are estimated, the critical value has to be adjusted. Monte Carlo methods can be used to estimate the p-value of the tests, which will be discussed in Chapter 15. □

13.2.2 Anderson–Darling test

Similar to the Kolmogorov–Smirnov test, the **Anderson–Darling test** can be used to test for the null hypothesis that the variable of interest has the df $F^*(\cdot)$.[6] Assuming we have complete and individual observations arranged in the order: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, the Anderson–Darling statistic, denoted by A^2,

[6] This test was introduced by Anderson and Darling in 1952.

is defined as

$$A^2 = -\frac{1}{n}\left[\sum_{j=1}^{n}(2j-1)\left\{\log\left[F^*(x_{(j)})\right] + \log\left[1 - F^*(x_{(n+1-j)})\right]\right\}\right] - n. \tag{13.6}$$

If $F^*(\cdot)$ is fully specified with no unknown parameters, the critical values of A^2 are 1.933, 2.492, and 3.857 for levels of significance $\alpha = 0.10$, 0.05, and 0.01, respectively. Also, critical values are available for certain distributions with unknown parameters. Otherwise, they may be estimated using Monte Carlo methods.

Example 13.6 Compute the Anderson–Darling statistics for the data in Example 13.1, with the estimated Weibull and gamma models as the hypothesized distributions.

Solution The statistics computed using equation (13.6) are 0.3514 and 0.3233, respectively, for the fitted Weibull and gamma distributions. These values are much lower than the critical value of 1.933 at $\alpha = 0.10$. However, as parameters are estimated for the hypothesized distributions, the true critical values would be lower. The p-values of the tests can be estimated using Monte Carlo methods. □

13.2.3 Chi-square goodness-of-fit test

The chi-square goodness-of-fit test is applicable to grouped data. Suppose the sample observations are classified into the intervals $(0, c_1], (c_1, c_2], \cdots, (c_{k-1}, \infty)$, with n_j observations in $(c_{j-1}, c_j]$ such that $\sum_{j=1}^{k} n_j = n$. The expected number of observations in $(c_{j-1}, c_j]$ based on $F^*(\cdot)$ is

$$e_j = n\left[F^*(c_j) - F^*(c_{j-1})\right]. \tag{13.7}$$

To test the null hypothesis that the df $F^*(\cdot)$ fits the data, we define the chi-square goodness-of-fit statistic X^2 as

$$X^2 = \sum_{j=1}^{k}\frac{(e_j - n_j)^2}{e_j} = \left(\sum_{j=1}^{k}\frac{n_j^2}{e_j}\right) - n. \tag{13.8}$$

If $F^*(\cdot)$ is fully specified with no unknown parameters, X^2 is approximately distributed as a χ^2_{k-1} (chi-square distribution with $k-1$ degrees of freedom) under the null hypothesis when n is large. For the test to work well, we also

require the expected number of observations in each class to be not smaller than 5. If the parameters of $F^*(\cdot)$ are estimated using the **multinomial MLE (MMLE)**, then the asymptotic distribution of X^2 is χ^2_{k-r-1}, where r is the number of parameters estimated. To compute the multinomial MLE, we use the log-likelihood function

$$\log L(\theta; \boldsymbol{n}) = \sum_{j=1}^{k} n_j \log \left[F^*(\theta; c_j) - F^*(\theta; c_{j-1})\right], \tag{13.9}$$

where θ is the r-element parameter vector. The log-likelihood $\log L(\theta; \boldsymbol{n})$ is maximized with respect to θ to obtain $\hat{\theta}$.[7] The expected frequency e_j in interval $(c_{j-1}, c_j]$ is then given by

$$e_j = n\,[F^*(\hat{\theta}; c_j) - F^*(\hat{\theta}; c_{j-1})]. \tag{13.10}$$

Example 13.7 For the data in Example 13.4, compute the chi-square goodness-of-fit statistic assuming the loss distribution is (a) $\mathcal{G}(\alpha, \beta)$ and (b) $\mathcal{W}(\alpha, \lambda)$. In each case, estimate the parameters using the multinomial MLE.

Solution The multinomial MLE for the $\mathcal{G}(\alpha, \beta)$ assumption can be found in Example 13.4. We estimate the Weibull case using the multinomial MLE method to obtain the distribution $\mathcal{W}(1.9176, 17.3222)$. For each of the fitted distributions, the expected frequencies are computed and compared alongside the observed frequencies in each interval. The results are given in Table 13.2.

Table 13.2. *Results for Example 13.7*

$(c_{j-1}, c_j]$	(0, 6]	(6, 12]	(12, 18]	(18, 24]	(24, 30]	(30, ∞)
n_j	3	9	9	4	3	2
gamma	3.06	9.02	8.27	5.10	2.58	1.97
Weibull	3.68	8.02	8.07	5.60	2.92	1.71

Thus, for the gamma distribution, the chi-square statistic is

$$X^2 = \frac{(3)^2}{3.06} + \frac{(9)^2}{9.02} + \frac{(9)^2}{8.27} + \frac{(4)^2}{5.10} + \frac{(3)^2}{2.58} + \frac{(2)^2}{1.97} - 30 = 0.3694.$$

Similarly, the X^2 statistic for the fitted Weibull distribution is 0.8595. The degrees of freedom of the test statistics is $6 - 2 - 1 = 3$ for both fitted

[7] Note that if the parameters are estimated by MLE using individual observations, X^2 is *not* asymptotically distributed as a χ^2. This was pointed out by Chernoff and Lehmann in 1954. Specifically, X^2 is bounded between the χ^2_{k-1} and χ^2_{k-r-1} distributions. See DeGroot and Schervish (2002, p. 547) for more details.

distributions, and the critical value of the test statistic at the 5% level of significance is $\chi^2_{3,0.95} = 7.815$. Thus, both the gamma and Weibull assumptions cannot be rejected for the loss distribution. □

13.2.4 Likelihood ratio test

The likelihood ratio test compares two hypotheses, one of which is a special case of the other in the sense that the parameters of the model are stated to satisfy some constraints. The unconstrained model is the **alternative hypothesis,** and the model under the parametric constraints is the **null hypothesis**. Thus, the null hypothesis is said to be **nested** within the alternative hypothesis. The constraints imposed on the parameter vector θ can be zero restrictions (i.e. some of the parameters in θ are zero), linear restrictions, or nonlinear restrictions. Let $\hat{\theta}_U$ denote the unrestricted MLE under the alternative hypothesis, $\hat{\theta}_R$ denote the restricted MLE under the null hypothesis, and r denote the number of restrictions. The unrestricted and restricted maximized likelihoods are denoted by $L(\hat{\theta}_U, \boldsymbol{x})$ and $L(\hat{\theta}_R, \boldsymbol{x})$, respectively. The **likelihood ratio statistic**, denoted by ℓ, is defined as

$$\ell = 2\log\left[\frac{L(\hat{\theta}_U, \boldsymbol{x})}{L(\hat{\theta}_R, \boldsymbol{x})}\right] = 2\left[\log L(\hat{\theta}_U, \boldsymbol{x}) - \log L(\hat{\theta}_R, \boldsymbol{x})\right]. \tag{13.11}$$

As $\hat{\theta}_U$ and $\hat{\theta}_R$ are the unrestricted and restricted maxima, $L(\hat{\theta}_U, \boldsymbol{x}) \geq L(\hat{\theta}_R, \boldsymbol{x})$. If the restrictions under the null are true, we would expect the difference between $L(\hat{\theta}_U, \boldsymbol{x})$ and $L(\hat{\theta}_R, \boldsymbol{x})$ to be small. Thus, a large discrepancy between $L(\hat{\theta}_U, \boldsymbol{x})$ and $L(\hat{\theta}_R, \boldsymbol{x})$ is an indication that the null hypothesis is incorrect. Hence, we should reject the null hypothesis for large values of ℓ. Furthermore, when the null is true, ℓ converges to a χ^2_r distribution as the sample size n tends to ∞. Thus, the decision rule of the test is to reject the null hypothesis (i.e. conclude the restrictions do not hold) if $\ell > \chi^2_{r,1-\alpha}$ at level of significance α, where $\chi^2_{r,1-\alpha}$ is the $100(1-\alpha)$-percentile of the χ^2_r distribution.

Example 13.8 For the data in Example 13.1, estimate the loss distribution assuming it is exponential. Test the exponential assumption against the gamma assumption using the likelihood ratio test.

Solution The exponential distribution is a special case of the $\mathcal{G}(\alpha, \beta)$ distribution with $\alpha = 1$. For the alternative hypothesis of a gamma distribution where α is not restricted, the fitted distribution is $\mathcal{G}(0.5257, 5.9569)$ and the log-likelihood is

$$\log L(\hat{\theta}_U, \boldsymbol{x}) = -39.2017.$$

The MLE of λ for the $\mathcal{E}(\lambda)$ distribution is $1/\bar{x}$ (or the estimate of β in $\mathcal{G}(\alpha, \beta)$ with $\alpha = 1$ is $\bar{x}$). Now $\bar{x} = 3.1315$ and the maximized restricted log-likelihood is

$$\log L(\hat{\theta}_R, \boldsymbol{x}) = -42.8305.$$

Thus, the likelihood ratio statistic is

$$\ell = 2(42.8305 - 39.2017) = 7.2576.$$

As $\chi^2_{1,0.95} = 3.841 < 7.2576$, the null hypothesis of $\alpha = 1$ is rejected at the 5% level of significance. Thus, the exponential distribution is not supported by the data. □

Example 13.9 For the data in Example 13.1, estimate the loss distribution assuming it is $\mathcal{W}(0.5, \lambda)$. Test this assumption against the unrestricted $\mathcal{W}(\alpha, \lambda)$ alternative using the likelihood ratio test.

Solution The unrestricted Weibull model is computed in Example 13.1, for which the fitted model is $\mathcal{W}(0.6548, 2.3989)$. The log-likelihood of this model is -39.5315. Under the restriction of $\alpha = 0.5$, λ is estimated to be 2.0586, and the log-likelihood is -40.5091. Thus, the likelihood ratio statistic is

$$\ell = 2(40.5091 - 39.5315) = 1.9553,$$

which is smaller than $\chi^2_{1,0.95} = 3.841$. Hence, at the 5% level of significance there is no evidence to reject $\alpha = 0.5$ against the alternative hypothesis that the loss distribution is $\mathcal{W}(\alpha, \lambda)$. □

We have discussed four diagnostic checks for modeling. The likelihood ratio test is a general testing approach for restrictions on a model. We may use it to test if the model can be simplified to have a smaller number of unknown parameters. This approach, however, adopts the alternative hypothesis as the maintained model, which is itself not tested. The other three tests, namely the Kolmogorov–Smirnov test, the Anderson–Darling test, and the chi-square goodness-of-fit test, are for testing the fit of a model to the data.

If there are two or more possible models that are non-nested, we may begin by determining if each of these models can be simplified. This follows the **principle of parsimony**, and can be done by testing for parametric restrictions using the likelihood ratio test. The smaller models that are not rejected may then be tested using one of the three misspecification tests. If the final outcome is that only one of the models is not rejected, this model will be adopted. In circumstances when more than one model are not rejected, decision would

not be straightforward. Factors such as model parsimony, prior information and simplicity of theoretical analysis have to be considered. We may also use information criteria for model selection, which helps to identify the model to be adopted.

13.3 Information criteria for model selection

When two non-nested models are compared, the larger model with more parameters have the advantage of being able to fit the in-sample data with a more flexible function and thus possibly a larger log-likelihood. To compare models on more equal terms, **penalized log-likelihood** may be adopted. The **Akaike information criterion**, denoted by AIC, proposes to penalize large models by subtracting the number of parameters in the model from its log-likelihood. Thus, AIC is defined as

$$\text{AIC} = \log L(\hat{\theta}; \boldsymbol{x}) - p, \tag{13.12}$$

where $\log L(\hat{\theta}; \boldsymbol{x})$ is the log-likelihood evaluated at the MLE $\hat{\theta}$ and p is the number of estimated parameters in the model. Based on this approach the model with the largest AIC is selected.

Although intuitively easy to understand, the AIC has an undesirable property. Consider two models $\mathcal{M}_1$ and $\mathcal{M}_2$, so that $\mathcal{M}_1 \subset \mathcal{M}_2$; that is, $\mathcal{M}_1$ is a smaller model and is nested by $\mathcal{M}_2$. Then, using AIC, the probability of choosing $\mathcal{M}_1$ when it is true converges to a number that is strictly less than 1 when the sample size tends to infinity. In this sense, we say that the Akaike information criterion is *inconsistent*. On the other hand, if the true model belongs to $\mathcal{M}_2 - \mathcal{M}_1$ (i.e. $\mathcal{M}_1$ is not correct but the true model lies in $\mathcal{M}_2$), then the probability of rejecting $\mathcal{M}_1$ under AIC converges to 1. Hence, the problem of not being able to identify the true model even when we have very large samples occurs when the smaller model is correct.

The above problem of the AIC can be corrected by imposing a different penalty on the log-likelihood. The **Schwarz information criterion**, also called the **Bayesian information criterion**, denoted by BIC, is defined as

$$\text{BIC} = \log L(\hat{\theta}; \boldsymbol{x}) - \frac{p}{2} \log n. \tag{13.13}$$

Thus, compared to the AIC, heavier penalty is placed on larger models when $\log n > 2$, i.e. $n > 8$. Unlike the AIC, the BIC is consistent in the sense that the probability it will choose the smaller model when it is true converges to 1 when the sample size tends to infinity. Also, as for the case of the AIC, if the true model belongs to $\mathcal{M}_2 - \mathcal{M}_1$, the probability of rejecting $\mathcal{M}_1$ under BIC also converges to 1.

Example 13.10 For the data in Example 13.1, consider the following models: (a) $\mathcal{W}(\alpha, \lambda)$, (b) $\mathcal{W}(0.5, \lambda)$, (c) $\mathcal{G}(\alpha, \beta)$, and (d) $\mathcal{G}(1, \beta)$. Compare these models using AIC and BIC, and comment on your choice of model.

Solution The MLE of the four models, as well as their maximized log-likelihood, appear in previous examples. Table 13.3 summarizes the results and the values of the AIC and BIC.

Table 13.3. *Results for Example 13.10*

Model	$\log L(\hat{\theta}; \boldsymbol{x})$	AIC	BIC
$\mathcal{W}(\alpha, \lambda)$	-39.5315	-41.5315	-42.5272
$\mathcal{W}(0.5, \lambda)$	-40.5091	-41.5091	-42.0070
$\mathcal{G}(\alpha, \beta)$	-39.2017	-41.2017	-42.1974
$\mathcal{G}(1, \beta)$	-42.8305	-43.8305	-45.3284

AIC is maximized for the $\mathcal{G}(\alpha, \beta)$ model, giving a value of -41.2017. BIC is maximized for the $\mathcal{W}(0.5, \lambda)$ model, giving a value of -42.0070. Based on the BIC, $\mathcal{W}(0.5, \lambda)$ is the preferred model. □

13.4 Summary and discussions

We have reviewed some methods for evaluating a fitted model for loss distributions. Graphical tools such as a plot of the estimated df against the empirical df, the *p-p* plot, the *q-q* plot, and a plot of the estimated pdf against the sample histogram can be used to assess the fit of the model. The key point is to identify if there is any systematic deviation of the parametrically fitted model from the sample data.

The likelihood ratio test is applicable to test for parametric restrictions. The purpose is to examine if a bigger and more general model can be reduced to a smaller model in which some parameters take certain specific values. Misspecification tests for the fitness of a hypothesized distribution, whether the parameters are fully specified or subject to estimation, can be performed using the Kolmogorov–Smirnov test, the Anderson–Darling test, and the chi-square goodness-of-fit test. When more than one model pass the diagnostic checks, model selection may require additional criteria, such as the principle of parsimony and other considerations. Information selection criteria such as the Akaike and Schwarz information criteria provide quantitative rules for model selection. The Schwarz information criterion has the advantage that it is consistent, while the Akaike information criterion is not.

Exercises

13.1 You are given the following loss observations, which are assumed to follow an exponential distribution:

$$2, 5, 6, 9, 14, 18, 23.$$

The model is fitted using the maximum likelihood method.

(a) A *p-p* plot is constructed, with the sample proportion p_i of the ith-order statistic computed as $i/(n+1)$, where n is the sample size. What are the co-ordinates of the *p-p* plot?

(b) Repeat (a) above if p_i is computed as $(i-0.5)/n$.

(c) Repeat (a) above if the observation "5" is corrected to "6", and the repeated observation is treated as the 2nd- and 3rd-order statistics.

(d) Repeat (b) above if the observation "5" is corrected to "6", and the repeated observation is treated as the 2nd- and 3rd-order statistics.

13.2 You are given the following loss observations, which are assumed to come from a $\mathcal{G}(5, 6)$ distribution:

$$12, 15, 18, 21, 23, 28, 32, 38, 45, 58.$$

(a) A *q-q* plot is constructed, with the sample proportion p_i of the ith-order statistic computed as $i/(n+1)$, where n is the sample size. What are the co-ordinates of the *q-q* plot?

(b) Repeat (a) above if the sample proportion is computed as $(i-0.5)/n$.

13.3 Suppose X follows an exponential distribution. A point on the *p-p* plot of a random sample of X is (0.2, 0.24), and its corresponding co-ordinates in the *q-q* plot is $(x, 23.6)$. Determine the value of x.

13.4 Suppose X follows a $\mathcal{P}(5, \gamma)$ distribution. A point on the *p-p* plot of a random sample of X is $(0.246, p)$, and its corresponding co-ordinates in the *q-q* plot is (45.88, 54.62). Determine the value of p.

13.5 Develop an Excel spreadsheet to compute the Kolmogorov–Smirnov statistic of a sample of observations. You are given the following observations from a $\mathcal{G}(\alpha, \beta)$ distribution:

$$0.4, 1.8, 2.6, 3.5, 5.4, 6.7, 8.9, 15.5, 18.9, 19.5, 24.6.$$

Estimate the distribution using the method of moments and compute the Kolmogorov–Smirnov statistic of the sample. What is your conclusion regarding the assumption of the gamma distribution?

13.6 Use the data in Exercise 13.5, now assuming the observations are distributed as $\mathcal{W}(\alpha, \lambda)$. Estimate the parameters of the Weibull distribution by matching the 25th and 75th percentiles. Compute

the Kolmogorov–Smirnov statistic of the fitted data. What is your conclusion regarding the assumption of the Weibull distribution?

13.7 Compute the Anderson–Darling statistic for the $\mathcal{G}(\alpha, \beta)$ distribution fitted to the data in Exercise 13.5. What is your conclusion regarding the assumption of the gamma distribution?

13.8 Compute the Anderson–Darling statistic for the fitted $\mathcal{W}(\alpha, \lambda)$ distribution in Exercise 13.6. What is your conclusion regarding the assumption of the Weibull distribution?

Questions adapted from SOA exams

13.9 A particular line of business has three types of claims. The historical probability and the number of claims for each type in the current year are:

Type	Historical probability	Number of claims in current year
A	0.2744	112
B	0.3512	180
C	0.3744	138

You test the null hypothesis that the probability of each type of claim in the current year is the same as the historical probability. Determine the chi-square goodness-of-fit statistic.

13.10 You fit a $\mathcal{P}(\alpha, \gamma)$ distribution to a sample of 200 claim amounts and use the likelihood ratio test to test the hypothesis that $\alpha = 1.5$ and $\gamma = 7.8$. The maximum likelihood estimates are $\hat{\alpha} = 1.4$ and $\hat{\gamma} = 7.6$. The log-likelihood function evaluated at the maximum likelihood estimates is –817.92, and $\sum_{i=1}^{200} \log(x_i + 7.8) = 607.64$. Determine the results of the test.

13.11 A sample of claim payments is: 29, 64, 90, 135, and 182. Claim sizes are assumed to follow an exponential distribution, and the mean of the exponential distribution is estimated using the method of moments. Compute the Kolmogorov–Smirnov statistic.

13.12 You are given the following observed claim-frequency data collected over a period of 365 days:

Number of claims per day	Observed number of days
0	50
1	122
2	101
3	92
≥ 4	0

The Poisson model is fitted using the maximum likelihood method, and the data are re-grouped into four groups: 0, 1, 2, and ≥ 3. Determine the results of the chi-square goodness-of-fit test for the Poisson assumption.

13.13 A computer program simulated 1,000 $\mathcal{U}(0,1)$ variates, which were then grouped into 20 groups of equal length. If the sum of the squares of the observed frequencies in each group is 51,850, determine the results of the chi-square goodness-of-fit test for the $\mathcal{U}(0,1)$ hypothesis.

13.14 Twenty claim amounts are randomly selected from the $\mathcal{P}(2,\gamma)$ distribution. The maximum likelihood estimate of γ is 7. If $\sum_{i=1}^{20} \log(x_i+7) = 49.01$ and $\sum_{i=1}^{20} \log(x_i+3.1) = 39.30$, determine the results of the likelihood ratio test for the hypothesis that $\gamma = 3.1$.

13.15 A uniform kernel density estimator with bandwidth 50 is used to smooth the following workers compensation loss payments: 82, 126, 161, 294, and 384. If $F^*(x)$ denotes the kernel estimate and $\hat{F}(x)$ denotes the empirical distribution function, determine $|F^*(150) - \hat{F}(150)|$.

13.16 The Kolmogorov–Smirnov test is used to assess the fit of the logarithmic loss to a distribution with distribution function $F^*(\cdot)$. There are $n = 200$ observations and the maximum value of $|\hat{F}(x) - F^*(x)|$, where $\hat{F}(\cdot)$ is the empirical distribution function, occurs for x between 4.26 and 4.42. You are given the following:

x	$F^*(x)$	$\hat{F}(x)$
4.26	0.584	0.510
4.30	0.599	0.515
4.35	0.613	0.520
4.36	0.621	0.525
4.39	0.636	0.530
4.42	0.638	0.535

Also, $\hat{F}(4.26^-)$, which is the empirical distribution function immediately to the left of 4.26, is 0.505. Determine the results of the test.

13.17 Five models are fitted to a sample of 260 observations with the results in the table. Which model will be selected based on the Schwarz Bayesian criterion? Which will be selected based on the Akaike Information criterion?

Model	Number of parameters	Log-likelihood
1	1	–414
2	2	–412
3	3	–411
4	4	–409
5	6	–409

13.18 A random sample of five observations is: 0.2, 0.7, 0.9, 1.1, and 1.3. The Kolmogorov–Smirnov test is used to test the null hypothesis that the probability density function of the population is

$$f(x) = \frac{4}{(1+x)^5}, \quad x > 0.$$

Determine the results of the test.

13.19 You test the hypothesis that a given set of data comes from a known distribution with distribution function $F(x)$. The following data were collected (x_i is the upper end point of the interval):

Interval	$F(x_i)$	Number of observations
$x < 2$	0.035	5
$2 \le x < 5$	0.130	42
$5 \le x < 7$	0.630	137
$7 \le x < 8$	0.830	66
$8 \le x$	1.000	50

Determine the results of the chi-square goodness-of-fit test.

13.20 One thousand workers insured under a workers compensation policy were observed for one year. The number of work days missed is given below:

Number of days missed	Number of workers
0	818
1	153
2	25
≥ 3	4

The total number of days missed is 230 and a Poisson distribution is fitted, with the parameter estimated by the average number of days

missed. Determine the results of the chi-square goodness-of-fit test for the model.

13.21 A random sample of losses from a $\mathcal{W}(\alpha, \lambda)$ distribution is: 595, 700, 789, 799, and 1109. The maximized log-likelihood function is –33.05, and the maximum likelihood estimate of λ assuming $\alpha = 2$ is 816.7. Determine the results of the likelihood ratio test for the hypothesis that $\alpha = 2$.

13.22 A sample of claim amounts is: 400, 1000, 1600, 3000, 5000, 5400, and 6200. The data are hypothesized to come from the $\mathcal{E}(1/3300)$ distribution. Let (s, t) be the co-ordinates of the *p-p* plot for claim amount 3000, and let $D(x)$ be the empirical distribution function at x minus the hypothesized distribution function at x. Compute (s, t) and $D(3000)$.

13.23 You are given the following distribution of 500 claims:

Claim size	Number of claims
[0, 500)	200
[500, 1000)	110
[1000, 2000)	x
[2000, 5000)	y
[5000, 10000)	–
[10000, 25000)	–
[25000, ∞)	–

Let $\hat{F}(\cdot)$ be the ogive computed for the sample. If $\hat{F}(1500) = 0.689$ and $\hat{F}(3500) = 0.839$, determine y.

13.24 Claim size has the following probability density function

$$f(x) = \frac{\theta e^{-\frac{\theta}{x}}}{x^2}, \quad x > 0.$$

A random sample of claims is: 1, 2, 3, 5, and 13. For $\theta = 2$, compute the Kolmogorov–Smirnov statistic.

14

Basic Monte Carlo methods

Some problems arising from loss modeling may be analytically intractable. Many of these problems, however, can be formulated in a stochastic framework, with a solution that can be estimated empirically. This approach is called Monte Carlo simulation. It involves drawing samples of observations randomly according to the distribution required, in a manner determined by the analytic problem.

To solve the stochastic problem, samples of the specified distribution have to be generated, invariably using computational algorithms. The basic random number generators required in Monte Carlo methods are for generating observations from the uniform distribution. Building upon uniform random number generators, we can generate observations from other distributions by constructing appropriate random number generators, using methods such as inverse transformation and acceptance–rejection. We survey specific random number generators for some commonly used distributions, some of which are substantially more efficient than standard methods. An alternative method of generating numbers resembling a uniformly distributed sample of observations is the quasi-random number generator or the low-discrepancy sequence.

The accuracy of the Monte Carlo estimates depends on the variance of the estimator. To speed up the convergence of the Monte Carlo estimator to the deterministic solution, we consider designs of Monte Carlo sampling schemes and estimation methods that will produce smaller variances. Methods involving the use of antithetic variable, control variable, and importance sampling are discussed.

Learning objectives

1 Generation of uniform random numbers, mixed congruential method
2 Low-discrepancy sequence
3 Inversion transformation and acceptance–rejection methods

4 Generation of specific discrete and continuous random variates
5 Generation of correlated normal random variables
6 Variance reduction techniques
7 Antithetic variable, control variable, and importance sampling

14.1 Monte Carlo simulation

Suppose $h(\cdot)$ is a smooth integrable function over the interval $[0, 1]$, and it is desired to compute the integral

$$\int_0^1 h(x)\, dx. \tag{14.1}$$

Now consider a random variable U distributed uniformly in the interval $[0, 1]$, i.e. $U \sim \mathcal{U}(0, 1)$. Then the integral in (14.1) is equal to $\mathrm{E}[h(U)]$. If the solution of (14.1) is difficult to obtain analytically, we may consider the stochastic solution of it as the mean of $h(U)$. The stochastic solution can be estimated by drawing a random sample of n observations $(u_1, \cdots, u_n)$ from U, and the computed estimate is given by

$$\hat{\mathrm{E}}\,[h(U)] = \frac{1}{n}\sum_{i=1}^{n} h(u_i). \tag{14.2}$$

By the law of large numbers $\hat{\mathrm{E}}[\,h(U)]$ converges to $\mathrm{E}[\,h(U)]$ when n tends to ∞. Thus, provided the sample size is sufficiently large, the stochastic solution can be made very close to the deterministic solution, at least in the probabilistic sense.

There are many deterministic problems that can be formulated in a stochastic framework, such as the solution of differential equations and eigenvalue systems. Von Neumann and Ulam coined the use of the term **Monte Carlo method** to describe this technique, which requires the generation of observations as random numbers produced by a computer. This technique can also be extended to study the solution of any simulated stochastic process (not necessarily with a deterministic counterpart), called **statistical simulation**. While some authors require the term simulation to describe a process evolving over time, we would not make this distinction and will treat Monte Carlo method and simulation as synonymous.[1]

Equation (14.2) requires samples of uniform random numbers. Indeed, the generation of uniform random numbers is a main component of a Monte

[1] See Kennedy and Gentle (1980) and Herzog and Lord (2002) for further discussions of the historical developments of Monte Carlo simulation methods.

Carlo study. We shall review methods of generating uniform random numbers using congruential algorithms. However, there are problems for which we require random numbers from other distributions. General methods to generate random numbers for an arbitrary distribution using uniform random numbers will be discussed. These include the inverse transformation method and the acceptance–rejection method. As the Monte Carlo method provides an estimated answer to the solution, we shall study its accuracy, which obviously depends on the Monte Carlo sample size. We will also discuss methods to improve the accuracy of the Monte Carlo estimate, or reduce its variance. Techniques such as antithetic variable, control variable, and importance sampling will be discussed.

14.2 Uniform random number generators

Independent random variates from the $\mathcal{U}(0, 1)$ distribution can be generated in the computer by dividing random integers in the interval $[0, m)$ by m, where m is a large number. An important method for generating sequences of random integers is the use of the congruential algorithm. We first define the expression

$$y \equiv z \pmod{m}, \tag{14.3}$$

where m is an integer, and y and z are integer-valued expressions, to mean that there exists an integer k, such that

$$z = mk + y. \tag{14.4}$$

This also means that y is the remainder when z is divided by m. The **mixed-congruential method** of generating a sequence of random integers x_i is defined by the equation

$$x_{i+1} \equiv (ax_i + c) \pmod{m}, \qquad \text{for } i = 0, 1, 2, \cdots, \tag{14.5}$$

where a is the **multiplier**, c is the **increment**, and m is the **modulus**. The mixed-congruential method requires the restrictions: $m > 0$, $0 < a < m$, and $0 \leq c < m$. When $c = 0$, the method is said to be **multiplicative-congruential**.

The integers produced from equation (14.5) are in the interval $[0, m)$. To generate numbers in the interval $[0, 1)$, we divide them by m. To start the sequence of x_i, we need a **seed** x_0. Given the seed x_0, the sequence of numbers x_i are completely determined. However, for appropriately chosen parameters of the congruential algorithm, the sequence will *appear* to be *random*. Indeed, random numbers generated by computer algorithms usually follow deterministic sequences, and are called **pseudo-random numbers**. Such

sequences, however, pass stringent tests for randomness and may be regarded as random for practical purposes, i.e.

$$\frac{x_i}{m} \sim \text{iid}\, \mathcal{U}(0,1), \qquad \text{for } i = 1, 2, \cdots . \tag{14.6}$$

Given a seed x_0, when the algorithm produces a value $x_k = x_h$ for certain integers h and k, such that $k > h \geq 0$, the sequence will start to repeat itself. We define the **period of the seed** as the shortest subsequence of numbers, which, by repeating itself, forms the complete sequence generated. Note that given an algorithm, different seeds generally have different periods. The **period of the generator** is the *largest* period among all seeds. Naturally, it is desirable to have pseudo-random number generators with long periods.

Example 14.1 Consider the following mixed-congruential generator

$$x_{i+1} \equiv (3x_i + 1) \pmod 8.$$

What is the period of (a) $x_0 = 2$ and (b) $x_0 = 4$? What is the period of the generator?

Solution For (a) it is easy to show that the sequence of numbers generated are:

$$2, 7, 6, 3, 2, \cdots$$

and this sequence repeats itself. Thus, the period of $x_0 = 2$ is 4. For (b), we have the sequence:

$$4, 5, 0, 1, 4, \cdots$$

Hence, the period of $x_0 = 4$ is again 4. To summarize, for given seed values x_0, the values of x_1 are given as follows:

x_0	0	1	2	3	4	5	6	7
x_1	1	4	7	2	5	0	3	6

All seeds have period 4, and the generated sequences belong to one of the two sequences above. Thus, the period of the generator is 4. □

Congruential algorithms provide efficient methods to generate random sequences. For example, the Super–Duper algorithm has the following formula

$$x_{i+1} \equiv (69{,}069x_i + 1) \pmod{2^{32}}. \tag{14.7}$$

Another simple multiplicative-congruential generator is RANDU, which is defined by the equation[2]

$$x_{i+1} \equiv 65{,}539 x_i \pmod{2^{31}}. \tag{14.8}$$

In estimating the integral over the [0, 1] interval, the use of n random points in the interval [0, 1] produces an estimate with a standard error of order $1/\sqrt{n}$. Thus, we describe the error as $O_p(1/\sqrt{n})$, meaning that for large enough n, there is a negligibly small probability of the absolute error exceeding a given bound divided by $\sqrt{n}$. Now instead of using n randomly selected numbers we may use n *equally spaced* numbers in the interval [0, 1]. The use of these equally spaced numbers in equation (14.2) produces a solution for the definite integral by summing the areas of the rectangles approximating the area under the curve $f(\cdot)$. It can be shown that the error of this **numerical integration** is bounded above by a constant times $1/n$, i.e. the error is of order $O(1/n)$.[3] Indeed, if we use the **trapezoidal rule** for the numerical integration, the error is of order $O(1/n^2)$. Hence, the Monte Carlo method is not as efficient as numerical integration for integrals of one dimension.

The scenario, however, changes when we consider multi-dimensional integrals. For a two-dimensional integral, the error of the Monte Carlo method remains of order $O_p(1/\sqrt{n})$. Using n equally spaced points over the unit square produces distances of order $1/\sqrt{n}$ between the points. Thus, the error in the numerical integration using approximating rectangles is of order $O(1/\sqrt{n})$. In general, when the dimension of the integral is of order d and the trapezoidal rule is used for approximation, the error in numerical integration using evenly spaced points is of order $O(n^{-2/d})$. Hence, for higher-dimensional integrals, numerical integration using evenly spaced points becomes unattractive.

Instead of using evenly spaced points over the d-dimensional unit hypercube, we may use **low-discrepancy sequences**. These sequences may be generated using various algorithms, many of which are based on number-theoretic results. They are also called **quasi-random numbers** or **quasi-random sequences**. The numerical integration methods using such sequences are called **quasi-Monte Carlo methods**. Like pseudo-random numbers, the quasi-random sequences are deterministic, and so are the error bounds for the numerical integrals using such sequences. The objective of the quasi-Monte Carlo method is to approximate the integral as accurately as possible, and this is done by trying to avoid using repeated choices of sequences. Thus, achieving randomness is not the basic underpinning of this approach. For further details of this method,

[2] See McLeish (2005) for other congruential generators, as well as other techniques such as shuffling of generators and linear combinations of outputs of generators.

[3] Note that there is no subscript p for this error, as its bound is deterministic.

readers may refer to Boyle *et al.* (1997), Tan and Boyle (2000), and McLeish (2005).

14.3 General random number generators

In many practical applications, we may be required to generate random numbers from distributions other than $\mathcal{U}(0,1)$. It turns out that the generation of random numbers following an arbitrary distribution can be done using uniform random numbers via the inversion transformation. We first define the important **probability integral transform**, which is basically the transformation of a random variable using its distribution function.

Definition 14.1 Let X be a random variable with df $F(\cdot)$. The probability integral transform Y of X is a random variable defined by $Y = F(X)$.

Thus, the probability integral transform is just a df, where the argument is a random variable rather than a fixed number. It turns out that through the probability integral transform, we can obtain a random variable that is distributed as $\mathcal{U}(0,1)$.

Theorem 14.1 ***(a) Probability integral transform theorem*** *If X is a random variable with continuous df $F(\cdot)$, then the random variable $Y = F(X)$ is distributed as $\mathcal{U}(0,1)$.* ***(b) Quantile function theorem*** *Let $F(\cdot)$ be a df, and define $F^{-1}(\cdot)$ as $F^{-1}(y) = \inf\{x : F(x) \geq y\}$, for $0 < y < 1$. If $U \sim \mathcal{U}(0,1)$, then the df of $X = F^{-1}(U)$ is $F(\cdot)$.*

Proof For Part (a), if $F(\cdot)$ is strictly increasing, the proof is quite straightforward. In this case, for $0 < y < 1$, there exists a unique x such that $F(x) = y$. Furthermore, $Y \leq y$ if and only if $X \leq x$. Thus, if $G(\cdot)$ is the df of Y, then

$$G(y) = \Pr(Y \leq y) = \Pr(X \leq x) = F(x) = y.$$

Hence, $G(y) = y$, which implies $Y \sim \mathcal{U}(0,1)$. For a general proof requiring $F(\cdot)$ to be continuous only, see Angus (1994).

For Part (b), we note that $X \leq x$ if and only if $U \leq F(x)$. Thus, we conclude

$$\Pr(X \leq x) = \Pr(U \leq F(x)) = F(x).$$

The last equality above is due to the fact that $U \sim \mathcal{U}(0,1)$. Hence, the df of X is $F(\cdot)$, as required by the theorem. □

14.3.1 Inversion method

Part (b) of Theorem 14.1 provides a convenient method to generate a random number with a known df from a uniform random number generator. Provided we can invert the function $F(\cdot)$ to obtain $F^{-1}(\cdot)$, $F^{-1}(U)$ will be a random variable with df $F(\cdot)$. This is called the **inversion method** for generating a random number for an arbitrary distribution.

Example 14.2 Let X have pdf $3x^2$ for $x \in [0, 1]$. Derive an algorithm to generate X. If two random numbers distributed as $\mathcal{U}(0, 1)$ are generated as 0.4521 and 0.8747, what are the values of X generated?

Solution We first derive the df of X, which is

$$F(x) = x^3, \qquad \text{for } 0 \le x \le 1.$$

Thus, inverting $F(\cdot)$, X may be generated as

$$X = U^{\frac{1}{3}}.$$

For the given values of U generated, the results for X are 0.7675 and 0.9564. □

Example 14.3 Derive algorithms to generate random numbers from the following distributions: (a) $\mathcal{W}(\alpha, \lambda)$, and (b) $\mathcal{P}(\alpha, \gamma)$.

Solution For (a), from equation (2.36), the df of $\mathcal{W}(\alpha, \lambda)$ is

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\lambda}\right)^{\alpha}\right].$$

Inverting the df, we generate X using the formula

$$X = \lambda\,[-\log(1 - U)]^{\frac{1}{\alpha}}.$$

As $1 - U$ is also distributed as $\mathcal{U}(0, 1)$, we can use the simplified formula

$$X = \lambda\,[-\log U]^{\frac{1}{\alpha}}$$

to generate $\mathcal{W}(\alpha, \lambda)$.

For (b), from equation (2.38), the df of $\mathcal{P}(\alpha, \gamma)$ is

$$F(x) = 1 - \left(\frac{\gamma}{x + \gamma}\right)^{\alpha}.$$

Thus, random numbers from $\mathcal{P}(\alpha, \gamma)$ may be generated using the equation

$$X = \gamma(U^{-\frac{1}{\alpha}} - 1).$$

□

The above examples illustrate the use of the inverse transform of the df to generate continuous random numbers. The inversion method can also be used to generate discrete or mixed-type variables. The example below provides an illustration.

Example 14.4 The ground-up loss X of an insurance policy is distributed as $\mathcal{W}(0.5, 5)$. There is a deductible of $d = 1$ and maximum covered loss of $u = 8$. Derive an algorithm to generate the loss in a loss event variable X_L using a $\mathcal{U}(0, 1)$ variate U. What are the values of X_L generated when $U = 0.8, 0.25$, and 0.5?

Solution $X_L = 0$ when $X \leq 1$. Thus

$$F_{X_L}(0) = \Pr(X \leq 1) = \Pr(\mathcal{W}(0.5, 5) \leq 1) = 1 - \exp\left[-\left(\frac{1}{5}\right)^{0.5}\right]$$

$$= 0.3606.$$

X_L is also right censored at point 7, with

$$\Pr(X_L = 7) = \Pr(X \geq 8) = \exp\left[-\left(\frac{8}{5}\right)^{0.5}\right] = 0.2823.$$

Hence, $\Pr(X_L < 7) = 1 - 0.2823 = 0.7177$, and the df of X_L is

$$F_{X_L}(x) = \begin{cases} 0.3606, & \text{for } x = 0, \\ 1 - \exp\left[-\left(\dfrac{x+1}{5}\right)^{0.5}\right], & \text{for } 0 < x < 7, \\ 1, & \text{for } x \geq 7. \end{cases} \tag{14.9}$$

Thus, X_L is a mixed-type random variable, and its df is plotted in Figure 14.1. We may invert $F_{X_L}(x)$ as follows to generate a random variate of X_L given a $\mathcal{U}(0, 1)$ variate U

$$X_L = \begin{cases} 0, & \text{for } 0 \leq U < 0.3606, \\ 5\left[-\log(1-U)\right]^2 - 1, & \text{for } 0.3606 \leq U < 0.7177, \\ 7, & \text{for } 0.7177 \leq U < 1. \end{cases} \tag{14.10}$$

When $U = 0.8$, $X_L = 7$. When $U = 0.25$, $X_L = 0$. Finally, when $U = 0.5$, X_L is computed as

$$X_L = 5\left[-\log(1 - 0.5)\right]^2 - 1 = 1.4023.$$

Note that X_L can also be generated by left-truncating and right-censoring a Weibull variate computed using the inversion method. □

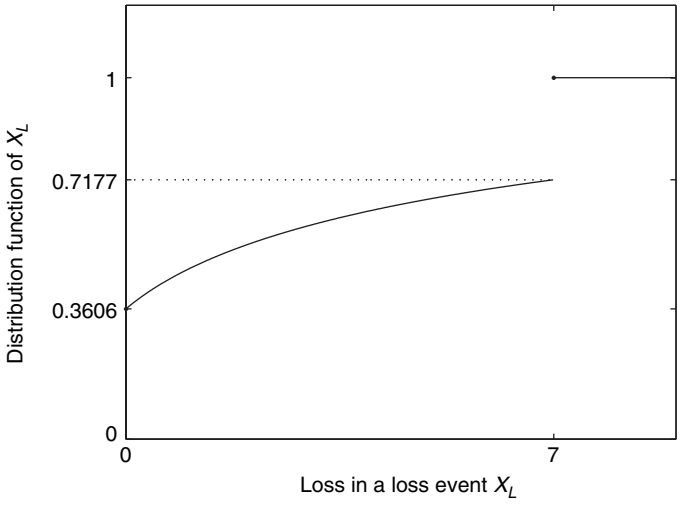

Figure 14.1 Distribution function of X_L in Example 14.4

While the inversion method is straightforward and easy to understand, there are situations when the df cannot be inverted analytically. When this happens alternative methods may be required to provide an efficient generator of the random numbers.

14.3.2 Acceptance–rejection method

The acceptance–rejection method can be used for cases where the df of a random variable has no analytic form or its analytic df cannot be easily inverted (e.g., the normal distribution and the gamma distribution). Let $f(\cdot)$ be the pdf of a random variable X, the df of which cannot be easily inverted, and let Y be another random variable with pdf $q(\cdot)$ for which an easy and efficient generator is available. Assume X and Y have the same support $[a, b]$, and there exists a constant c such that $M(x) \equiv cq(x) \geq f(x)$ for $x \in [a, b]$. We now state the acceptance–rejection procedure for generating random numbers of X, followed by a proof of the validity of the procedure. The steps of the **acceptance–rejection procedure** are as follows:

1 Generate a number x from the distribution with pdf $q(\cdot)$.
2 Generate a number u independently from the $\mathcal{U}(0, 1)$ distribution.
3 If $u \leq f(x)/M(x)$, assign $z = x$, otherwise return to Step 1.

It turns out that the sequence of numbers z obtained from the above procedure have pdf $f(\cdot)$. To prove this statement, we consider the df of the random variable

Z generated, which is given by

$$\begin{aligned}
\Pr(Z \le z) &= \Pr\left(Y \le z \mid U \le \frac{f(Y)}{M(Y)}\right) \\
&= \frac{\int_a^z \int_0^{\frac{f(x)}{M(x)}} q(x)\, du\, dx}{\int_a^b \int_0^{\frac{f(x)}{M(x)}} q(x)\, du\, dx} \\
&= \frac{\int_a^z q(x) \left(\int_0^{\frac{f(x)}{M(x)}} du\right) dx}{\int_a^b q(x) \left(\int_0^{\frac{f(x)}{M(x)}} du\right) dx} \\
&= \frac{\int_a^z q(x) \dfrac{f(x)}{M(x)}\, dx}{\int_a^b q(x) \dfrac{f(x)}{M(x)}\, dx} \\
&= \frac{\int_a^z f(x)\, dx}{\int_a^b f(x)\, dx} \\
&= \int_a^z f(x)\, dx. \qquad (14.11)
\end{aligned}$$

Note that the second to last equation above is due to the fact that $q(x)/M(x) = 1/c$ for $z \in [a, b]$, and is thus canceled out in the ratio. Finally, differentiating $\Pr(Z \le z)$ with respect to z, we obtain the pdf of Z, which is $f(\cdot)$.

The pdf $q(\cdot)$ is called the **majorizing density**, and the function $M(x) = cq(x)$ is called the **majorizing function**. The principle is to find a majorizing function that *envelopes* the pdf $f(\cdot)$ as closely as possible. For a given majorizing density $q(\cdot)$, c should be chosen to tighten the enveloping of $M(x)$ over $f(x)$, i.e. the optimum c should be

$$c = \inf\{r : rq(x) \ge f(x) \text{ for } x \in [a, b]\}. \qquad (14.12)$$

However, even if the optimum c is not used, the acceptance–rejection procedure stated above remains valid, albeit there is loss in efficiency.

Example 14.5 Let the pdf of X be

$$f(x) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad \text{for } x \ge 0.$$

Suppose the majorizing density is selected to be

$$q(x) = e^{-x}, \qquad \text{for } x \ge 0.$$

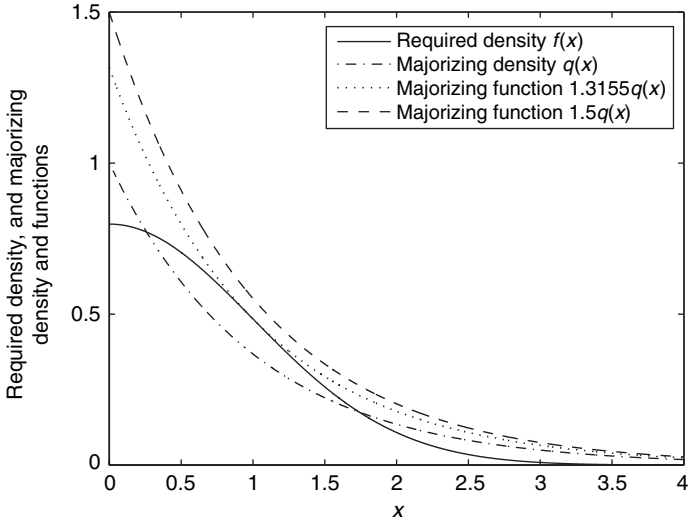

Figure 14.2 Required pdf and majorizing pdf in Example 14.5

Discuss the use of the acceptance–rejection procedure for the generation of random numbers of X.

Solution X is obtained as the absolute value of the standard normal random variable. The inverse transformation method is intractable for this distribution. Figure 14.2 plots the pdf $f(\cdot)$ and $q(\cdot)$. The two functions cross each other. To create the optimum $cq(\cdot)$ the value of c is $\sqrt{2e/\pi} = 1.3155$.[4] However, any value of $c \geq 1.3155$ may be used to compute the majorizing function and appropriate random numbers will be produced. Figure 14.2 also shows the majorizing function with $c = 1.5$, which is not optimum.

The acceptance–rejection procedure for generating X is summarized as follows:

1 Generate a number x with pdf e^{-x}. This can be done by computing $x = -\log v$, where v is a random number from $\mathcal{U}(0, 1)$.
2 Generate a number u independently from $\mathcal{U}(0, 1)$.
3 For a selected value of $c \geq 1.3155$, if

$$u \leq \frac{\frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)}{c \exp(-x)} = \frac{1}{c}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2} + x\right) \equiv R(x),$$

assign $Z = x$. Otherwise, return to Step 1.

[4] We skip the technical derivation of this result, which can be proved using equation (14.12).

Table 14.1 shows a sample of values of random numbers generated using $c = 1.5$. The last row of values Z are the random numbers having pdf $f(\cdot)$.

Table 14.1. *Illustrative results for Example 14.5*

i	1	2	3	4
u	0.3489	0.9236	0.5619	0.4581
v	0.4891	0.0910	0.5047	0.9057
x	0.7152	2.3969	0.6838	0.0990
$R(x)$	0.8421	0.3306	0.8342	0.5844
Z	0.7152	reject	0.6838	0.0990

□

The probability of acceptance in Step 3 of the acceptance–rejection procedure is given by

$$\begin{aligned}
\Pr\left(U \le \frac{f(X)}{M(X)}\right) &= \int_a^b \int_0^{\frac{f(x)}{M(x)}} q(x)\, du\, dx \\
&= \int_a^b q(x) \left(\int_0^{\frac{f(x)}{M(x)}} du \right) dx \\
&= \int_a^b q(x) \frac{f(x)}{M(x)}\, dx \\
&= \frac{1}{c} \int_a^b f(x) dx \\
&= \frac{1}{c}. \qquad (14.13)
\end{aligned}$$

Thus, we may use $1/c$ as a measure of the efficiency of the procedure.

14.3.3 Generation of correlated random variables

In some Monte Carlo studies we are required to generate observations that are correlated. For example, in estimating the VaR of a portfolio of assets, we may need to generate returns of the components of the portfolio, which are invariably correlated. In estimating aggregate losses of a block of insurance policies, we may need to generate loss observations that are correlated.

The correlation structure of the random numbers required depends on specific problems. We shall discuss the problem of generating samples of normal random variables that are correlated. This problem is of particular interest as returns are often assumed to be normal in financial applications (i.e. asset prices are lognormally distributed). Also, the multivariate normal distribution has the nice property that its distribution is completely determined by its mean and variance.

We shall discuss the main properties of a multivariate normal distribution, followed by methods of generating correlated multivariate normal variates.

Let $\boldsymbol{X} = (X_1, \cdots, X_k)'$ be a k-element random variable. If $\boldsymbol{X}$ has a multivariate normal distribution, its joint df is completely determined by its mean vector $\mu = (\mu_1, \cdots, \mu_k)'$ and its variance matrix

$$\Omega = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & & \\ \sigma_{1k} & & \cdots & \sigma_k^2 \end{bmatrix}, \tag{14.14}$$

where

$$\mu_i = \mathrm{E}(X_i) \quad \text{and} \quad \sigma_i^2 = \mathrm{Var}(X_i), \qquad \text{for } i = 1, \cdots, k, \tag{14.15}$$

and

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j), \qquad \text{for } i, j = 1, \cdots, k. \tag{14.16}$$

We will then write

$$\boldsymbol{X} \sim \mathcal{N}(\mu, \Omega). \tag{14.17}$$

If $\boldsymbol{X}$ has a nondegenerate distribution, there exists a lower triangular $k \times k$ matrix C (i.e. the elements in the upper triangle of the matrix are all zero), denoted by

$$C = \begin{bmatrix} c_{11} & 0 & 0 & \cdots & 0 \\ c_{21} & c_{22} & 0 & \cdots & 0 \\ \vdots & & & & \\ c_{k1} & & & \cdots & c_{kk} \end{bmatrix}, \tag{14.18}$$

such that

$$\Omega = CC'. \tag{14.19}$$

The equation above is called the **Choleski decomposition** of Ω. The lower triangular matrix C is obtainable in many statistical packages.

The multivariate normal distribution has some very convenient properties. Let A be a $m \times k$ ($m \leq k$) constant matrix and b be a $m \times 1$ constant vector.

Then

$$A\boldsymbol{X} + b \sim \mathcal{N}(A\mu + b, A\Omega A'). \tag{14.20}$$

If $\boldsymbol{Y} = (Y_1, \cdots, Y_k)'$ has a multivariate normal distribution with mean vector $\mu = (0, \cdots, 0)' = \mathbf{0}$ and variance matrix $\Omega = \mathbf{I}$ (i.e. the $k \times k$ identity matrix), we write

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{14.21}$$

$Y_1, \cdots, Y_k$ are iid standard normal variates. Furthermore, if we define

$$\boldsymbol{X} = C\boldsymbol{Y} + \mu, \tag{14.22}$$

then from equation (14.20) we conclude

$$\boldsymbol{X} \sim \mathcal{N}(C\mathbf{0} + \mu, C\mathbf{I}C') \equiv \mathcal{N}(\mu, \Omega). \tag{14.23}$$

Thus, to generate random numbers of the multivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\mu, \Omega)$, we first generate k iid standard normal variates $\boldsymbol{Y} = (Y_1, \cdots, Y_k)'$. Then using equation (14.22), we obtain the required random numbers for $\boldsymbol{X}$. The generation of standard normal variates will be discussed in Section 14.4.1.

Example 14.6 Let X_1 and X_2 be jointly normally distributed with means μ_1 and μ_2, respectively, variances σ_1^2 and σ_2^2, respectively, and covariance σ_{12}. How would you generate random numbers of X_1 and X_2 given independent random numbers of the standard normal distribution?

Solution We first solve for the Choleski decomposition of

$$\Omega = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

It can be easily checked that

$$C = \begin{bmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{bmatrix}$$

where ρ is the correlation coefficient, i.e.

$$\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}.$$

Hence, if Z_1 and Z_2 are independently distributed $\mathcal{N}(0,1)$ variates, then we can generate X_1 and X_2 from the equation

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$
$$= \begin{bmatrix} \sigma_1 Z_1 + \mu_1 \\ \rho\sigma_2 Z_1 + \sigma_2\sqrt{1-\rho^2}\, Z_2 + \mu_2 \end{bmatrix}.$$

It is easy to verify that $\mathrm{E}(X_1) = \mu_1$, $\mathrm{E}(X_2) = \mu_2$, $\mathrm{Var}(X_1) = \sigma_1^2$,

$$\mathrm{Var}(X_2) = \rho^2\sigma_2^2 + \sigma_2^2\left(1-\rho^2\right) = \sigma_2^2,$$

and

$$\mathrm{Cov}(X_1, X_2) = \rho\sigma_1\sigma_2 = \sigma_{12}.$$

14.4 Specific random number generators

Many commonly used distributions cannot be generated by the standard methods reviewed so far, while others can be generated using some specific but more efficient methods. We now discuss the generation of some of these distributions.

14.4.1 Some continuous distributions

Normal distribution

Normal distribution is perhaps one of the most commonly used distributions in statistics. Yet as the df of a normal distribution does not have an analytic expression, the inversion method does not work. A crude method is to invoke the law of large numbers and approximate a normal random variable using the sum of a large number of uniform random numbers. This method, however, is only approximate and is obviously very inefficient in execution.

The **Box–Muller method** generates pairs of standard normal variates from pairs of uniform variates. Let U_1 and U_2 be two independent $\mathcal{U}(0,1)$ variates. The transformations defined by

$$X_1 = \cos(2\pi U_1)\sqrt{-2\log U_2},$$
$$X_2 = \sin(2\pi U_1)\sqrt{-2\log U_2}, \tag{14.24}$$

produce a pair of independent $\mathcal{N}(0,1)$ variates X_1 and X_2.

Another algorithm, called the **Marsaglia–Bray method**, uses a mixture distribution together with the acceptance–rejection method. This method, however, requires the generation of more uniform variates. More algorithms for generating normal random variates can be found in Kennedy and Gentle (1980).

Gamma distribution

The gamma distribution $\mathcal{G}(\alpha, \beta)$ covers several standard distributions as special cases, including the exponential distribution $\mathcal{E}(1/\beta)$ (when $\alpha = 1$), the Erlang distribution (when $\beta = 1$ and α is a positive integer), and the chi-square distribution χ^2_r (when $\beta = 2$ and $\alpha = r/2$). As the df of a general $\mathcal{G}(\alpha, \beta)$ variable cannot be inverted easily, the inversion method for generating random numbers of $\mathcal{G}(\alpha, \beta)$ does not work.

A gamma random variable with $\beta = 1$ is said to have a standard gamma distribution. As a $\mathcal{G}(\alpha, \beta)$ variate has the same distribution as a $\mathcal{G}(\alpha, 1)/\beta$ variate, we only need to consider standard gamma distributions.

From Section 2.2.2, we see that the sum of n independent $\mathcal{E}(1)$ variates is distributed as $\mathcal{G}(n, 1)$. Furthermore, if $U \sim \mathcal{U}(0, 1)$, $-\log U \sim \mathcal{E}(1)$. Hence, if α is an integer (Erlang distribution), $X \sim \mathcal{G}(\alpha, 1)$ can be generated by the equation

$$X = -\sum_{i=1}^{\alpha} \log U_i, \tag{14.25}$$

where $U_i \sim \text{iid}\, \mathcal{U}(0, 1)$, for $i = 1, \cdots, \alpha$.

If $X_i \sim \mathcal{G}(\alpha_i, 1)$, for $i = 1, 2$, are independently distributed, then $X_1 + X_2 \sim \mathcal{G}(\alpha_1 + \alpha_2, 1)$. Hence, to generate a standard gamma variate $\mathcal{G}(\alpha, 1)$ we may split α into the sum of its largest integer part and a term that is between 0 and 1. Thus, $\mathcal{G}(\alpha, 1)$ is the sum of two gamma variates, one of which has an Erlang distribution and the other has a standard gamma distribution with a parameter in the interval $(0, 1)$. We now consider the case of generating a $\mathcal{G}(\alpha, 1)$ variate with $\alpha \in (0, 1)$.

The **Ahrens method** provides an efficient procedure to generate a $\mathcal{G}(\alpha, 1)$ variate with $\alpha \in (0, 1)$ using the acceptance–rejection approach. The required pdf is

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \qquad \text{for } \alpha \in (0, 1),\ x \geq 0. \tag{14.26}$$

The majorizing frequency consists of two segments defined as follows[5]

$$q(x) = \begin{cases} \dfrac{e}{\alpha + e}\, \alpha x^{\alpha - 1}, & \text{for } 0 \le x \le 1, \\ \dfrac{\alpha}{\alpha + e}\, e^{1-x}, & \text{for } 1 < x. \end{cases} \tag{14.27}$$

The df of this density, denoted by $Q(\cdot)$, is

$$Q(x) = \begin{cases} \dfrac{e}{\alpha + e}\, x^{\alpha}, & \text{for } 0 \le x \le 1, \\ 1 - \dfrac{\alpha}{\alpha + e}\, e^{1-x}, & \text{for } 1 < x. \end{cases} \tag{14.28}$$

Using the inverse transformation, we can generate a random number X with df $Q(\cdot)$ from a $\mathcal{U}(0, 1)$ variate U as follows

$$X = \begin{cases} \left[\dfrac{(\alpha + e)\, U}{e}\right]^{\frac{1}{\alpha}}, & \text{for } 0 \le U \le \dfrac{e}{\alpha + e}, \\ 1 - \log\left[\dfrac{(1 - U)(\alpha + e)}{\alpha}\right], & \text{for } \dfrac{e}{\alpha + e} < U < 1. \end{cases} \tag{14.29}$$

To envelope the pdf $f(\cdot)$ we use the majorizing function $M(x) = cq(x)$, where c is given by

$$c = \frac{\alpha + e}{\Gamma(\alpha)\alpha e}. \tag{14.30}$$

Thus, the majorizing function is

$$M(x) = cq(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)}\, x^{\alpha - 1}, & \text{for } 0 \le x \le 1, \\ \dfrac{1}{\Gamma(\alpha)}\, e^{-x}, & \text{for } 1 < x. \end{cases} \tag{14.31}$$

We further note that

$$\frac{f(x)}{M(x)} = \begin{cases} e^{-x}, & \text{for } 0 \le x \le 1, \\ x^{\alpha - 1}, & \text{for } 1 < x. \end{cases} \tag{14.32}$$

[5] It can be checked that $\int_0^{\infty} q(x)\, dx = 1$.

Thus, it can be easily checked that $M(x) \geq f(x)$ for $x \geq 0$. We are now ready to state the Ahrens method for the generation of the random variable X with pdf $f(\cdot)$ as follows:

1 Generate a random number u_1 from $\mathcal{U}(0,1)$. If $u_1 > e/(\alpha + e)$, go to Step 3. Otherwise, continue with Step 2.
2 Set $z = [(\alpha + e)\, u_1/e]^{\frac{1}{\alpha}}$ and generate independently another random number u_2 from $\mathcal{U}(0,1)$. If $u_2 > e^{-z}$, go to Step 1, otherwise assign $X = z$.
3 Set $z = 1 - \log[(1-u_1)(\alpha + e)/\alpha]$ and generate independently another random number u_2 from $\mathcal{U}(0,1)$. If $u_2 > z^{\alpha-1}$, go to Step 1, otherwise assign $X = z$.

14.4.2 Some discrete distributions

We now discuss the generation of two discrete distributions, namely the binomial and Poisson distributions. Note that while the binomial distribution has a finite support, the support of the Poisson distribution is infinite.

Binomial distribution

The pf of the binomial distribution $\mathcal{BN}(n,\theta)$ is

$$f(i) = \binom{n}{i}\theta^i(1-\theta)^{n-i}, \qquad \text{for } i = 0, 1, \cdots, n. \tag{14.33}$$

As the $\mathcal{BN}(n,\theta)$ distribution has a finite support, we can use a simple **table look-up method** to generate the random numbers. We first compute the df of $\mathcal{BN}(n,\theta)$ as

$$F(i) = \sum_{j=0}^{i}\binom{n}{j}\theta^j(1-\theta)^{n-j}, \qquad \text{for } i = 0, 1, \cdots, n. \tag{14.34}$$

Now a random number X from $\mathcal{BN}(n,\theta)$ may be generated as follows. For a number u generated from $\mathcal{U}(0,1)$, we set $X = 0$ if $u \leq F(0)$, and set $X = r+1$ if $F(r) < u \leq F(r+1)$, for $r = 0, \cdots, n-1$.

Alternatively X can be generated by exploiting the fact that it may be interpreted as the number of successes in n independent trials, where the probability of success in each trial is θ. Thus, we generate n random variates U_i, $i = 1, \cdots, n$, from $\mathcal{U}(0,1)$ and compute X as the number of U_i that are less than θ.

Poisson distribution

As the Poisson distribution has an infinite support, the table look-up method does not work. However, we may make use of the relationship between the exponential distribution and the Poisson distribution to derive an algorithm.

Let $X \sim \mathcal{PN}(\lambda)$ be the number of arrivals of a certain event in a unit time interval. Then the inter-arrival time Y of the events follows the exponential distribution $\mathcal{E}(\lambda)$, i.e. an exponential distribution with mean waiting time $1/\lambda$. Thus, we can generate Y_i from $\mathcal{E}(\lambda)$ and accumulate them to obtain the *total* waiting time. We then set X to be the largest number of Y_i accumulated such that their total is less than 1, i.e.

$$X = \min\left\{n : \sum_{i=1}^{n+1} Y_i > 1\right\}. \tag{14.35}$$

As Y_i can be generated by $(-\log U_i)/\lambda$, where $U_i \sim \mathcal{U}(0,1)$, we re-write the above as

$$\begin{aligned} X &= \min\left\{n : \sum_{i=1}^{n+1} \frac{1}{\lambda}(-\log U_i) > 1\right\} \\ &= \min\left\{n : \sum_{i=1}^{n+1} \log U_i < -\lambda\right\} \\ &= \min\left\{n : \prod_{i=1}^{n+1} U_i < e^{-\lambda}\right\}. \end{aligned} \tag{14.36}$$

Hence, X can be generated by multiplying uniformly distributed variates. The number of such variates required to generate a single value of X increases with λ.

14.5 Accuracy and Monte Carlo sample size

Apart from providing an estimate of the required deterministic solution, Monte Carlo samples can also be used to provide an assessment of the accuracy of the estimate. We may use the Monte Carlo sample to estimate the standard error of the estimated solution and hence obtain a confidence interval for the solution. The standard error may also be used to estimate the required sample size to produce a solution within a required accuracy given a certain probability level.

Example 14.7 The specific damages X covered by a liability insurance policy are distributed as $\mathcal{G}(4,3)$. The total damages, inclusive of punitive damages, are given by cX, where $c > 1$. The policy has a maximum covered loss of u.

Using Monte Carlo methods or otherwise, determine the expected loss of the insured and the probability that the total damages do not exceed u. Discuss the accuracy of your solutions. Consider the case of $c = 1.1$, and $u = 20$.

Solution The pdf of X is

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, \qquad x \geq 0.$$

We denote the df of $X \sim \mathcal{G}(\alpha, \beta)$ by

$$\gamma_x(\alpha, \beta) = \Pr(X \leq x),$$

and note that

$$\gamma_x(\alpha, \beta) = \gamma_{\frac{x}{\beta}}(\alpha, 1) \equiv \gamma_{\frac{x}{\beta}}(\alpha).$$

The function $\gamma_x(\alpha)$ is also called the (lower) **incomplete gamma function**.

The expected loss is

$$\begin{aligned}
\mathrm{E}[(cx) \wedge u] &= \int_0^{\frac{u}{c}} cx \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} dx + u \int_{\frac{u}{c}}^{\infty} \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} dx \\
&= \frac{c\Gamma(\alpha+1)\beta}{\Gamma(\alpha)} \int_0^{\frac{u}{c}} \frac{x^{(\alpha+1)-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha+1)\beta^{\alpha+1}} dx + u\left[1 - \gamma_{\frac{u}{c\beta}}(\alpha)\right] \\
&= c\alpha\beta\gamma_{\frac{u}{c\beta}}(\alpha+1) + u\left[1 - \gamma_{\frac{u}{c\beta}}(\alpha)\right].
\end{aligned}$$

Thus, the expected loss can be computed using the incomplete gamma function.

Similarly, we can derive the second moment of the loss as

$$\mathrm{E}\left[((cx) \wedge u)^2\right] = c^2(\alpha+1)\alpha\beta^2\gamma_{\frac{u}{c\beta}}(\alpha+2) + u^2\left[1 - \gamma_{\frac{u}{c\beta}}(\alpha)\right],$$

from which we can compute $\mathrm{Var}[(cx) \wedge u]$.

Now with the given values of $c = 1.1$ and $u = 20$ we obtain

$$\mathrm{E}[(cx) \wedge u] = 12.4608 \qquad \text{and} \qquad \mathrm{Var}[(cx) \wedge u] = 25.9197.$$

Using a Monte Carlo sample of 10,000 observations, we obtain estimates of the mean and variance of the loss as (these are the sample mean and the sample variance of the simulated losses)

$$\hat{\mathrm{E}}[(cx) \wedge u] = 12.5466 \qquad \text{and} \qquad \widehat{\mathrm{Var}}[(cx) \wedge u] = 25.7545.$$

The Monte Carlo standard error of $\hat{\mathrm{E}}[(cx) \wedge u]$ is

$$\sqrt{\frac{\widehat{\mathrm{Var}}\,[(cx) \wedge u]}{10{,}000}} = \sqrt{\frac{25.7545}{10{,}000}} = 0.0507.$$

Thus, using normal approximation, the Monte Carlo estimate of the 95% confidence interval of the expected loss is

$$12.5466 \pm (1.96)(0.0507) = (12.4471,\ 12.6461),$$

which covers the true value of 12.4608.

The probability of the total damages not exceeding u is

$$\Pr(cX \le u) = \int_0^{\frac{u}{c}} \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}\,dx = \gamma_{\frac{u}{c\beta}}(\alpha),$$

and we have

$$\gamma_{\frac{20}{(1.1)(3)}}(4) = 0.8541.$$

The Monte Carlo estimate of this probability is the sample proportion of $1.1X \le 20$, which was found to be 0.8543. The 95% confidence interval of the true probability is

$$0.8543 \pm 1.96\sqrt{\frac{(0.8543)(1-0.8543)}{10{,}000}} = (0.8474,\ 0.8612),$$

which again covers the true probability. □

Example 14.8 Continue with Example 14.7 and modify the total damages to $bX^2 + cX$. Using Monte Carlo methods or otherwise, determine the expected loss of the insured and the probability that the total damages do not exceed u. Discuss the accuracy of your solutions. Consider the case of $b = 0.1$, $c = 1.1$, and $u = 30$.

Solution The mean loss is now given by

$$\mathrm{E}\left[(bx^2 + cx) \wedge u\right] = \int_0^r \left(bx^2 + cx\right) \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}\,dx + u\int_r^{\infty} \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}\,dx,$$

where

$$br^2 + cr = u.$$

We solve for r to obtain

$$r = \frac{\sqrt{c^2 + 4bu} - c}{2b}.$$

Thus, the expected loss is

$$\mathrm{E}\left[(bx^2 + cx) \wedge u\right] = b(\alpha+1)\alpha\beta^2\gamma_{\frac{r}{\beta}}(\alpha+2) + c\alpha\beta\gamma_{\frac{r}{\beta}}(\alpha+1) + u\left[1 - \gamma_{\frac{r}{\beta}}(\alpha)\right].$$

We will not derive the variance analytically, but will estimate it using the Monte Carlo method. Using the values $b = 0.1$, $c = 1.1$, and $u = 30$, we obtain

$$\mathrm{E}\left[(bx^2 + cx) \wedge u\right] = 21.7192.$$

A Monte Carlo sample of 10,000 observations produced the following estimates

$$\hat{\mathrm{E}}\left[(bx^2 + cx) \wedge u\right] = 21.7520 \quad \text{and} \quad \sqrt{\frac{\widehat{\mathrm{Var}}\left[(bx^2 + cx) \wedge u\right]}{10{,}000}} = 0.0867.$$

Thus, a 95% confidence interval estimate of the expected loss is

$$21.7520 \pm (1.96)(0.0867) = (21.5821,\ 21.9219),$$

which covers the true value.

The probability of the total damages not exceeding u is

$$\Pr(bX^2 + cX \le u) = \Pr(X \le r) = \gamma_{\frac{r}{\beta}}(\alpha) = 0.6091.$$

This is estimated by the proportion of Monte Carlo observations of total damages not exceeding 30, and was found to be 0.6076. A 95% confidence interval of the probability of the total damages not exceeding 30 is

$$0.6076 \pm 1.96\sqrt{\frac{(0.6076)(1 - 0.6076)}{10{,}000}} = (0.5980,\ 0.6172),$$

which again covers the true probability. □

14.6 Variance reduction techniques

Consider a deterministic problem with solution equal to $\mathrm{E}[h(X)]$, where X is a random variable (not necessarily uniformly distributed) and $h(\cdot)$ is an integrable

function over the support of X. The **crude Monte Carlo** estimate of the solution is

$$\hat{\mathrm{E}}[h(X)] = \frac{1}{n}\sum_{i=1}^{n} h(x_i), \tag{14.37}$$

where $(x_1, \cdots, x_n)$ are sample values of X. The accuracy of this stochastic solution depends on the variance of the estimator. We now discuss some methods of sampling and estimation that aim at reducing the variance of the Monte Carlo estimator.

14.6.1 Antithetic variable

The **antithetic variable method** attempts to reduce the variance of the Monte Carlo estimate through the use of the random numbers generated. To illustrate the idea, consider a Monte Carlo sample of two observations X_1 and X_2. If X_1 and X_2 are iid, The variance of the Monte Carlo estimator is

$$\mathrm{Var}\left(\hat{\mathrm{E}}[h(X)]\right) = \frac{\mathrm{Var}[h(X)]}{2}. \tag{14.38}$$

However, if X_1 and X_2 are identically distributed as X, but *not independent*, then the variance of the Monte Carlo estimator is

$$\mathrm{Var}\left(\hat{\mathrm{E}}[h(X)]\right) = \frac{\mathrm{Var}[h(X)] + \mathrm{Cov}(h(X_1), h(X_2))}{2}. \tag{14.39}$$

Now if $\mathrm{Cov}(h(X_1), h(X_2)) < 0$, the variance of the Monte Carlo estimator is reduced. Random numbers generated in such a way that the functional evaluations at these numbers are negatively correlated are said to be antithetic variables.

If $X_1 \sim \mathcal{U}(0,1)$, then $X_2 = 1 - X_1 \sim \mathcal{U}(0,1)$ and is negatively correlated with X_1. It should be noted, however, that for the antithetic variable technique to work well, it is the negative correlation between $h(X_1)$ and $h(X_2)$ that is required. Negative correlation between X_1 and X_2 in itself does not guarantee reduction in the variance of the Monte Carlo estimator.

The above discussion can be generalized to a sample of n observations, in which reduction in the variance may be obtained if the pairwise correlations of $h(X_i)$ are negative. Also, this technique can be extended to the case of a vector random variable.

Example 14.9 Consider the distribution of the loss in a loss event variable X_L in Example 14.4. Estimate $\mathrm{E}(X_L)$ using a crude Monte Carlo method and a Monte Carlo simulation with antithetic variable.

Solution To estimate $\mathrm{E}(X_L)$ using the crude Monte Carlo method, we generate a random sample of n variates U_i from $\mathcal{U}(0,1)$. For each U_i we compute $X_L(U_i)$ using equation (14.10). $\mathrm{E}(X_L)$ is then estimated by

$$\hat{\mathrm{E}}_{\mathrm{CR}}(X_L) = \frac{1}{n}\sum_{i=1}^{n} X_L(U_i).$$

To use the antithetic variable technique, we generate $n/2$ (n being even) $\mathcal{U}(0,1)$ variates U_i, $i = 1, \cdots, n/2$, and augment this set of numbers by another $n/2$ $\mathcal{U}(0,1)$ variates by taking $U_{i+\frac{n}{2}} = 1 - U_i$. The sample mean of X_L computed from this set of U_i is the Monte Carlo estimate with antithetic variables, denoted by $\hat{\mathrm{E}}_{\mathrm{AT}}(X_L)$.

We performed a Monte Carlo simulation with $n = 10{,}000$. The sample means and sample variances are

$$\hat{\mathrm{E}}_{\mathrm{CR}}(X_L) = 2.8766, \qquad s^2_{\mathrm{CR}} = 9.2508,$$

and

$$\hat{\mathrm{E}}_{\mathrm{AT}}(X_L) = 2.8555, \qquad s^2_{\mathrm{AT}} = 9.1719.$$

We can see that the results of the two Monte Carlo estimates of $\mathrm{E}(X_L)$ are very similar. However, there is only a small gain in using antithetic variables in this estimation. If we refer to Figure 14.1, we can see that the probability contents at the two ends of the distribution (at values 0 and 7) are similar. Thus, little negative correlation is induced by taking the complements of the $\mathcal{U}(0,1)$ variates. □

14.6.2 Control variable

To estimate $\mathrm{E}[h(X)]$ using control variable, we consider an auxiliary function $g(\cdot)$ and the associated expectation $\mathrm{E}[g(X)]$. We select the function $g(\cdot)$ so that it is close to $h(\cdot)$ and yet $\mathrm{E}[g(X)]$ is *known*. Now a Monte Carlo estimate of $\mathrm{E}[h(X)]$ can be computed as

$$\hat{\mathrm{E}}_{\mathrm{CV}}[h(X)] = \mathrm{E}[g(X)] + \frac{1}{n}\sum_{i=1}^{n}\left[h(X_i) - g(X_i)\right]. \tag{14.40}$$

It is easy to check that $\hat{\mathrm{E}}_{\mathrm{CV}}[h(X)]$ is an unbiased estimate of $\mathrm{E}[h(X)]$. The variance of this estimator is

$$\mathrm{Var}(\hat{\mathrm{E}}_{\mathrm{CV}}[h(X)]) = \frac{\mathrm{Var}\left[h(X) - g(X)\right]}{n}, \tag{14.41}$$

which is smaller than the variance of $\hat{\mathrm{E}}_{\mathrm{CR}}[h(X)]$ if

$$\mathrm{Var}[h(X) - g(X)] < \mathrm{Var}[h(X)]. \tag{14.42}$$

Example 14.10 Consider the distribution of the loss in a loss event variable X_L in Example 14.4. Estimate $\mathrm{E}(X_L)$ using a Monte Carlo simulation with control variable.

Solution To estimate $\mathrm{E}(X_L)$ using control variable, we consider a random variable $\tilde{X}_L$ with the following df

$$F_{\tilde{X}_L}(x) = \begin{cases} 0.3606, & \text{for } x = 0, \\ 0.3606 + 0.0510x, & \text{for } 0 < x < 7, \\ 1, & \text{for } x \geq 7, \end{cases}$$

where

$$0.0510 = \frac{0.7177 - 0.3606}{7} = \frac{0.3571}{7}$$

is the slope of the line joining the points $(0, 0.3606)$ and $(7, 0.7177)$. Comparing the above with equation (14.9), we can see that the df of X_L in the interval $[0.3606, 0.7177)$ is now *linearized*. The mean of $\tilde{X}_L$ is

$$\mathrm{E}(\tilde{X}_L) = (0.3571)(3.5) + (0.2823)(7) = 3.2260.$$

Given a $\mathcal{U}(0, 1)$ variate U_i, X_L can be generated using equation (14.10), and we denote this by $X_L(U_i)$. Now the inverse transformation of $\tilde{X}_L$ is

$$\tilde{X}_L = \begin{cases} 0, & \text{for } 0 \leq U < 0.3606, \\ (U - 0.3606)/0.0510, & \text{for } 0.3606 \leq U < 0.7177, \\ 7, & \text{for } 0.7177 \leq U < 1. \end{cases}$$

Hence, the Monte Carlo estimate of $\mathrm{E}(X_L)$ using the control variable $\tilde{X}_L$ is computed as

$$3.2260 + \frac{1}{n}\sum_{i=1}^{n}\left[X_L(U_i) - \tilde{X}_L(U_i)\right].$$

Note that $X_L(U_i) - \tilde{X}_L(U_i)$ is nonzero only when $U_i \in [0.3606, 0.7177)$, in which case we have

$$X_L(U_i) - \tilde{X}_L(U_i) = 5[-\log(1 - U_i)]^2 - 1 - \frac{U_i - 0.3606}{0.0510}.$$

We performed a Monte Carlo simulation with $n = 10{,}000$. The sample mean and sample variance are

$$\hat{\mathrm{E}}_{\mathrm{CV}}(X_L) = 2.8650, \qquad s^2_{\mathrm{CV}} = 0.3150.$$

Thus, there is a substantial increase in efficiency versus the crude Monte Carlo and Monte Carlo with antithetic variable. □

14.6.3 Importance sampling

Consider the following integral of a smooth integrable function $h(\cdot)$ over the interval $[a, b]$

$$\int_a^b h(x)\,dx, \tag{14.43}$$

which can be re-written as

$$\int_a^b [(b-a)h(x)]\,\frac{1}{b-a}\,dx. \tag{14.44}$$

Thus, the integral can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}(b-a)h(X_i), \tag{14.45}$$

where X_i are iid $\mathcal{U}(a, b)$. In general, if $\tilde{X}$ is a random variable with support $[a, b]$ and pdf $q(\cdot)$, the integral in equation (14.43) can be written as

$$\int_a^b h(x)\,dx = \int_a^b \left[\frac{h(x)}{q(x)}\right] q(x)\,dx, \tag{14.46}$$

which can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{h(\tilde{X}_i)}{q(\tilde{X}_i)}, \tag{14.47}$$

where $\tilde{X}_i$ are iid as $\tilde{X}$. The estimator in equation (14.47) has a smaller variance than the estimator in equation (14.45) if

$$\mathrm{Var}\left[\frac{h(\tilde{X}_i)}{q(\tilde{X}_i)}\right] < \mathrm{Var}\left[(b-a)h(X_i)\right]. \tag{14.48}$$

The technique of using equation (14.47), with a change of the pdf of the random variable to be generated and the function to be integrated, is called **importance sampling**. From equation (14.48) we can see that the advantage of importance sampling is likely to be large if the variation in the ratio $h(\cdot)/q(\cdot)$ is small over the interval $[a, b]$ (i.e. the two functions are close to each other).

Example 14.11 Consider the distribution of the loss in a loss event variable X_L in Example 14.4. Estimate $\mathrm{E}(X_L)$ using a Monte Carlo simulation with importance sampling.

Solution Defining $h(U)$ as $X_L(U)$ in equation (14.10), we have

$$\begin{aligned}\mathrm{E}(X_L) &= \int_0^1 h(x)\,dx \\ &= \int_{0.3606}^{0.7177} \left(5\,[-\log(1-x)]^2 - 1\right) dx + \int_{0.7177}^{1} 7\,dx \\ &= \int_{0.3606}^{0.7177} \left(5\,[-\log(1-x)]^2 - 1\right) dx + 1.9761.\end{aligned}$$

The integral above is the expected value of

$$(0.7177 - 0.3606)(5[-\log(1-\tilde{U})]^2 - 1),$$

where $\tilde{U} \sim \mathcal{U}(0.3606,\ 0.7177)$. Thus, we estimate $\mathrm{E}(X_L)$ by

$$1.9761 + \frac{0.3571}{n}\sum_{i=1}^{n}(5[-\log(1-\tilde{U}_i)]^2 - 1),$$

where

$$\tilde{U}_i = 0.3606 + 0.3571 U_i,$$

with $U_i \sim \text{iid}\,\mathcal{U}(0,1)$.

We performed a Monte Carlo simulation with 10,000 observations, and obtained $\hat{\mathrm{E}}(X_L) = 2.8654$, with a sample variance of 0.4937. Thus, the importance sampling method produced a big gain in efficiency over the crude Monte Carlo method, although the gain is not as much as that from the control variable method in Example 14.10. □

14.7 Excel computation notes

The RAND function in Excel uses a combination of multiplicative-congruential generators to generate random variates of $\mathcal{U}(0,1)$. It can be used to generate

$\mathcal{U}(a, b)$ variates using the transformation: `RAND()*(b-a)+a`. For example, a random variate of $\mathcal{U}(2, 12)$ can be generated using the code `RAND()*10+2`. More details of the algorithm used by `RAND` can be found in Exercise 14.4.

The following random number generators can be called in the Excel drop-down menu (using Tools → Data Analysis → Random Number Generation → Distribution): Uniform, Normal, Bernoulli, Binomial, Poisson, and Discrete. Users are required to input the parameters desired for the specific distribution selected. Figure 14.3 illustrates the generation of an arbitrary discrete distribution using the Random Number Generation menu. The desired Discrete distribution takes values 1, 2, and 3 with probabilities 0.1, 0.3, and 0.6, respectively. The possible values of the distribution and their associated probabilities are entered in cells A1 to B3 of the spreadsheet, and the Random Number Generation menu is then called up. The Distribution is selected (Discrete), and the Required Number of Random Numbers (30) with their

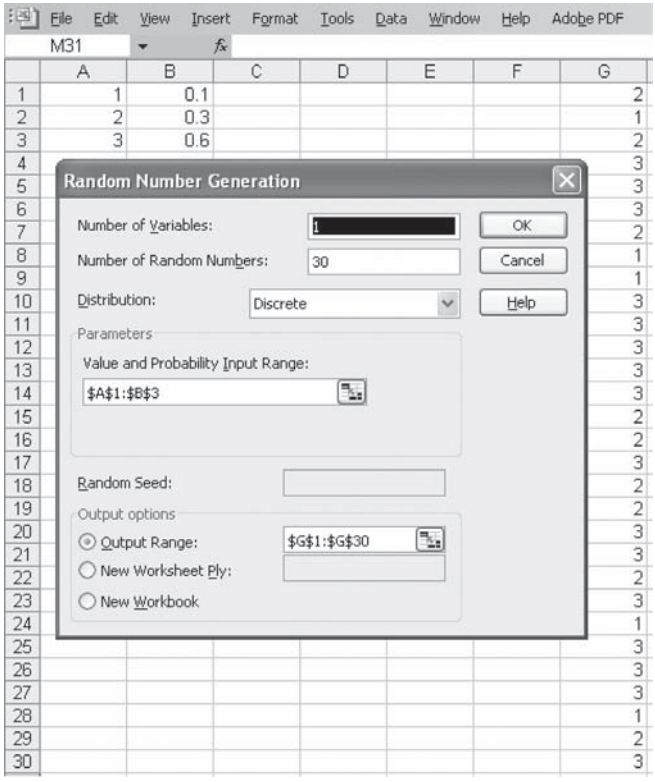

Figure 14.3 Generation of a discrete distribution in Excel

required Parameters (cells A1 to B3) and desired Output Range (cells G1 to G30) are entered. The generated results are exhibited in the figure.

14.8 Summary and discussions

Many deterministic problems can be formulated as a stochastic problem, the solution of which can be estimated using Monte Carlo methods. In a Monte Carlo simulation, random numbers are generated to resemble the sampling of observations. To this effect, efficient algorithms for the generation of uniformly distributed random numbers form a basic component of the method. We have surveyed methods for generating random numbers that are uniformly distributed, from which random numbers following other distributions can be generated using methods such as inversion transformation and acceptance–rejection. For certain commonly used distributions, specific generators are available, which provide efficient algorithms to facilitate the computation. As the solution of the Monte Carlo simulation is a stochastic estimator, its accuracy depends on the variance of the estimator. We discuss methods to improve the efficiency, or reduce the variance. Antithetic variable, control variable, and importance sampling are some common techniques. The performance of these techniques, however, depends on the actual problems considered and may vary considerably.

Exercises

14.1 You are given the following multiplicative-congruential generator

$$x_{i+1} \equiv (16807\, x_i) \pmod{2^{31} - 1}.$$

Using the Excel `MOD` function and the seed $x_1 = 401$, compute the numbers $x_2, \cdots, x_5$ and the corresponding $\mathcal{U}(0, 1)$ variates $u_2, \cdots, u_5$. If the seed is changed to $x_1 = 245987$, determine the new sequence of x_i and u_i.

14.2 You are given the following mixed-congruential generator

$$x_{i+1} \equiv (69069\, x_i + 1) \pmod{2^{32}}.$$

Using the Excel `MOD` function and the seed $x_1 = 747$, compute the numbers $x_2, \cdots, x_5$ and the corresponding $\mathcal{U}(0, 1)$ variates $u_2, \cdots, u_5$. If the seed is changed to $x_1 = 380$, determine the new sequence of x_i and u_i.

14.3 You are given that $x_1 = 7, x_2 = 5$ and

$$x_{i+1} \equiv (5\,x_i + c) \pmod{11}, \qquad \text{for } i = 0, 1, 2, \cdots .$$

Compute x_{101}.

14.4 The RAND function in Excel uses a combination of three multiplicative-congruential generators. The following generators are defined

$$\begin{aligned} x_{i+1} &\equiv (171\,x_i) \pmod{30269}, \\ y_{i+1} &\equiv (172\,y_i) \pmod{30307}, \\ z_{i+1} &\equiv (170\,z_i) \pmod{30323}. \end{aligned}$$

The $\mathcal{U}(0, 1)$ variates generated by each of the above generators are summed up and the fractional part is taken as the output of the RAND function. This algorithm provides $\mathcal{U}(0, 1)$ variates with a cycle length exceeding 2.78×10^{13}. Using the Excel MOD function with seeds $x_1 = 320$, $y_1 = 777$ and $z_1 = 380$, compute the next five $\mathcal{U}(0, 1)$ variates. [*Hint:* How do you obtain the fractional part of a number using the MOD function?]

14.5 The inverse Pareto distribution X has the following pdf

$$f(x) = \frac{\alpha\theta x^{\alpha-1}}{(x+\theta)^{\alpha+1}},$$

with

$$\mathrm{E}(X^{-1}) = \frac{1}{\theta(\alpha-1)} \quad \text{and} \quad \mathrm{E}(X^{-2}) = \frac{2}{\theta^2(\alpha-1)(\alpha-2)}.$$

Suppose $\alpha = \theta = 4$.

(a) How would you generate X variates using the inverse transformation method?

(b) Use Excel to simulate 50 random variates of X, and estimate $\mathrm{E}(X^{-1})$ and $\mathrm{E}(X^{-2})$ using the Monte Carlo sample. Compute a 95% confidence interval of $\mathrm{E}(X^{-1})$ using your Monte Carlo sample. Does this interval cover the true value?

(c) Use your Monte Carlo sample in (b) to estimate the Monte Carlo sample size required if you wish to estimate $\mathrm{E}(X^{-1})$ up to 1% error with a probability of 95%.

14.6 The inverse Weibull distribution X has the following pdf

$$f(x) = \frac{\alpha \left(\frac{\theta}{x}\right)^{\alpha} e^{-\left(\frac{\theta}{x}\right)^{\alpha}}}{x},$$

with

$$\mathrm{E}(X) = \theta\, \Gamma\left(1 - \frac{1}{\alpha}\right) \quad \text{and} \quad \mathrm{E}(X^2) = \theta^2\, \Gamma\left(1 - \frac{2}{\alpha}\right).$$

Suppose $\alpha = \theta = 4$.

(a) How would you generate X variates using the inverse transformation method?

(b) Use Excel to simulate 50 random variates of X, and estimate $\mathrm{E}(X)$ and $\mathrm{E}(X^2)$ using the Monte Carlo sample. Compute a 95% confidence interval of $\mathrm{E}(X)$ using your Monte Carlo sample. Does this interval cover the true value?

(c) Use your Monte Carlo sample in (b) to estimate the Monte Carlo sample size required if you wish to estimate $\mathrm{E}(X)$ up to 1% error with a probability of 95%.

14.7 Suppose X is distributed with pdf

$$f(x) = 12x^2(1 - x), \quad 0 < x < 1.$$

Using the majorizing density $q(x) = 1$ for $0 < x < 1$, suggest an acceptance–rejection algorithm to generate random variates of X. What is the efficiency measure of your algorithm? Why would you use the acceptance–rejection method rather than the inverse transformation method?

14.8 Let $X = U^2$ and $Y = (1 - U)^2$, where $U \sim \mathcal{U}(0, 1)$.

(a) Compute $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, $\mathrm{Cov}(X, Y)$, and $\mathrm{Var}[(X + Y)/2]$.

(b) Generate 100 random numbers of U and estimate the integral $\int_0^1 u^2\, du$ by simulation. Repeat the computation using the random numbers $1 - U$. Compare the means and the variances of the two simulation samples. How would you use the random numbers to improve the precision of your estimates?

14.9 Consider the following integral

$$\int_0^1 \frac{e^u - 1}{e - 1}\, du.$$

(a) Estimate the integral using a crude Monte Carlo method with 100 random numbers of $\mathcal{U}(0, 1)$.

(b) Estimate the integral using a Monte Carlo method with antithetic variates. This is done by first generating 50 random numbers $u_i \sim$ iid $\mathcal{U}(0,1)$ as in (a) above and then using the numbers $1-u_i$ for a total of 100 random numbers.

(c) Estimate the integral using a Monte Carlo simulation with control variable, with $g(u) = u/(e-1)$ as the auxiliary function. Use 100 random numbers of $\mathcal{U}(0,1)$ for your estimation.

Compare the Monte Carlo estimates of the methods above, as well as their standard deviations. Which method provides the best efficiency?

14.10 Suppose X_1 and X_2 are jointly distributed as bivariate standard normal variates with correlation of 0.6. Use Excel to generate 100 observations of (X_1, X_2), and estimate from the sample $\Pr(X_1 \leq 1.2, X_2 \leq 2.1)$. Construct a 95% confidence interval of the probability. If you wish to obtain an estimate with error less than 0.01 with a probability of 95%, what Monte Carlo sample size would you require? [Hint: $\Pr(X_1 \leq 1.2, X_2 \leq 2.1) = 0.8783$.]

14.11 Suppose $h(\cdot)$ is an integrable function over $[a,b]$ with $0 \leq h(x) \leq c$ for $a \leq x \leq b$. Let

$$\theta = \int_a^b h(x)\,dx.$$

(a) Show that $\theta = (b-a)\mathrm{E}[h(X)]$, where $X \sim \mathcal{U}(a,b)$.

(b) Show that $\mathrm{E}[h(X)] = c\Pr(Y \leq h(X))$, where $Y \sim \mathcal{U}(0,c)$ and is independent of X.

(c) In a Monte Carlo simulation, n independent pairs of X and Y are drawn and W is the number of cases such that $y \leq h(x)$. Define $\tilde{\theta} = c(b-a)W/n$, which is called the **hit-or-miss estimator** of θ. Show that $\mathrm{E}(\tilde{\theta}) = \theta$ and $\mathrm{Var}(\tilde{\theta}) = \theta[c(b-a)-\theta]/n$.

(d) The crude Monte Carlo estimator $\hat{\theta}$ is defined as $\hat{\theta} = (b-a)\left[\sum_{i=1}^n h(X_i)\right]/n$, where X_i are iid $\mathcal{U}(a,b)$. Show that $\hat{\theta}$ is unbiased for θ and that $\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\tilde{\theta})$.

14.12 Let $F(\cdot)$ be the distribution function of $\mathcal{N}(\mu,\sigma^2)$. Show that $F^{-1}(1-u) = 2\mu - F^{-1}(u)$ for $0 < u < 1$. Suppose X is a variate generated from $\mathcal{N}(\mu,\sigma^2)$, how would you compute an antithetic variate of X?

Questions adapted from SOA exams

14.13 A mixture variable Y is distributed as $\mathcal{E}(2)$ with probability 0.3 and $\mathcal{U}(-3,3)$ with probability 0.7. Y is simulated where low values of $\mathcal{U}(0,1)$ correspond to the exponential distribution. Then the selected component is simulated where low random variates of $\mathcal{U}(0,1)$

correspond to low values of Y, and Y is computed using the inverse transformation method. The $\mathcal{U}(0,1)$ variates generated, in order, are 0.25 and 0.69. Determine the value of Y.

14.14 N is the number of accidents in a period, and is distributed with pf

$$\Pr(N = n) = 0.9(0.1)^{n-1}, \quad n = 1, 2, \cdots .$$

X_i is the claim amount of the ith accident, and are iid with pdf

$$f(x) = 0.01\, e^{-0.01x}, \quad x > 0.$$

Suppose U and $V_1, V_2, \cdots$ are independent $\mathcal{U}(0,1)$ variates. U is used to simulate N and V_i are used to simulate the required number of X_i, all using the inverse transformation method, where U and V_i correspond to small values of N and X_i, respectively. The following values are generated: $u = 0.05$, $v_1 = 0.3$, $v_2 = 0.22$, $v_3 = 0.52$, and $v_4 = 0.46$. Determine the total amount of claims simulated for the period.

14.15 The return of an asset is zero with probability 0.8 and uniformly distributed on [1000, 5000] with probability 0.2. The inverse transformation method is used to simulate the outcome, where large values of $\mathcal{U}(0,1)$ generated correspond to large returns. If the first two $\mathcal{U}(0,1)$ variates generated are 0.75 and 0.85, compute the average of these two outcomes.

14.16 The claims of a workers compensation policy follow the $\mathcal{P}(2.8, 36)$ distribution. The df of the frequency of claim N is:

n	$\Pr(N \le n)$
0	0.5556
1	0.8025
2	0.9122
3	0.9610
4	0.9827
5	0.9923
6	1.0000

Each claim is subject to a deductible of 5 and a maximum payment of 30. A $\mathcal{U}(0,1)$ variate is generated and is used to simulate the number of claims using the inverse transformation method, where small value of $\mathcal{U}(0,1)$ corresponds to small number of claims. The $\mathcal{U}(0,1)$ variate generated is 0.981. Then further $\mathcal{U}(0,1)$ variates are generated to simulate the claim amount, again using the inverse transformation method, where small value of $\mathcal{U}(0,1)$ corresponds to small amount

of claim. The following $\mathcal{U}(0, 1)$ variates are generated: 0.571, 0.932, 0.303, 0.471, and 0.878. Using as many of these numbers as necessary (in the given order), compute the aggregate simulated claim payments.

14.17 The df of the number of losses for a policy in a year, N, is:

n	$\Pr(N \leq n)$
0	0.125
1	0.312
2	0.500
3	0.656
4	0.773
5	0.855
⋮	⋮

The amount of each loss is distributed as $\mathcal{W}(2, 200)$. There is a deductible of 150 for each claim and an annual maximum out-of-pocket of 500 per policy. The inverse transformation method is used to simulate the number of losses and loss amounts, where small $\mathcal{U}(0, 1)$ variates correspond to a small number of loss amounts. For the number of losses, the $\mathcal{U}(0, 1)$ variate generated is 0.7654. For the loss amounts, the following random numbers are used (in order and as needed): 0.2738, 0.5152, 0.7537, 0.6481, and 0.3153. Determine the aggregate payments by the insurer.

14.18 A dental benefit has a deductible of 100 applied to the annual charges. Reimbursement is 80% of the remaining charges subject to an annual maximum reimbursement of 1,000. Suppose the annual dental charges are distributed exponentially with mean 1,000. Charges are simulated using the inverse transformation method, where small values of $\mathcal{U}(0, 1)$ variates correspond to low charges. The following random numbers are generated: 0.30, 0.92, 0.70, and 0.08. Compute the average annual reimbursement.

14.19 Total losses are simulated using the aggregate loss model and the inverse transformation method, where small values of $\mathcal{U}(0, 1)$ variates correspond to low values of claim frequency and claim size. Suppose claim frequency is distributed as $\mathcal{PN}(4)$, and the $\mathcal{U}(0, 1)$ variate simulated is 0.13; the claim amount is distributed as $\mathcal{E}(0.001)$, and the $\mathcal{U}(0, 1)$ variates generated are: 0.05, 0.95, and 0.1 (in that order). Determine the simulated total losses.

14.20 Losses are distributed as $\mathcal{L}(5.6, 0.75^2)$. The claim payments with deductibles are estimated using simulation, where losses are computed

from $\mathcal{U}(0,1)$ variates using the inverse transformation method, with small $\mathcal{U}(0,1)$ variates corresponding to small losses. The following $\mathcal{U}(0,1)$ variates are generated: 0.6217, 0.9941, 0.8686, and 0.0485. Using these random numbers, compute the average payment per loss for a contract with a deductible of 100.

15

Applications of Monte Carlo methods

In this chapter we discuss some applications of Monte Carlo methods to the analysis of actuarial and financial data. We first re-visit the tests of model misspecification introduced in Chapter 13. For an asymptotic test, Monte Carlo simulation can be used to improve the performance of the test when the sample size is small, in terms of getting more accurate critical values or p-values. When the asymptotic distribution of the test is unknown, as for the case of the Kolmogorov–Smirnov test when the hypothesized distribution has some unknown parameters, Monte Carlo simulation may be the only way to estimate the critical values or p-values.

The Monte Carlo estimation of critical values is generally not viable when the null hypothesis has some nuisance parameters, i.e. parameters that are not specified and not tested under the null. For such problems, the use of bootstrap may be applied to estimate the p-values. Indeed, bootstrap is one of the most powerful and exciting techniques in statistical inference and analysis. We shall discuss the use of bootstrap in model testing, as well as the estimation of the bias and mean squared error of an estimator.

The last part of this chapter is devoted to the discussion of the simulation of asset-price processes. In particular, we consider both pure diffusion processes that generate lognormally distributed asset prices, as well as jump–diffusion processes that allow for discrete jumps as a random event with a random magnitude. The simulated price paths can be used to evaluate the prices of contingent claims such as options and guaranteed payoffs.

Learning objectives

1 Monte Carlo estimation of critical values and p-values
2 Bootstrap estimation of p-values
3 Bootstrap estimation of bias and mean squared error.
4 Simulation of lognormally distributed asset prices
5 Simulation of asset prices with discrete jumps

15.1 Monte Carlo simulation for hypothesis test

In Chapter 13 we discuss some hypothesis tests for model misspecification, including the Kolmogorov–Smirnov test, the Anderson–Darling test, and the chi-square goodness-of-fit test. All these tests require large samples to justify the use of an asymptotic distribution for the test statistic under the null. In small samples, the critical values based on the asymptotic distributions may not be accurate. Furthermore, some of these tests require the null distribution to be completely specified, with no unknown parameters. When the parameters are estimated, the distribution of the test statistic is unknown even in large samples.

15.1.1 Kolmogorov–Smirnov test

We now consider the use of Monte Carlo simulation to estimate the critical values of the Kolmogorov–Smirnov D test when the parameters of the null distribution are unknown. If the null distribution is normal, even if the parameters have to be estimated for the computation of D, the critical values of the D statistic can be estimated using Monte Carlo simulation. This is due to a result in David and Johnson (1948), which states that if the parameters estimated for the null distribution are parameters of *scale* or *location*, and the estimators satisfy certain general conditions, then the joint distribution of the probability-integral transformed observations of the sample will not depend on the *true* parameter values.

The David–Johnson result is important for several reasons. First, the Kolmogorov–Smirnov test is based on the distribution function under the null, which is the probability integral transform (see Definition 14.1) of the sample observations. Second, many commonly used distributions involve parameters of scale and location. For example, for the $\mathcal{N}(\mu, \sigma^2)$ distribution, μ is the location parameter and σ is the scale parameter. The parameter λ in the $\mathcal{E}(\lambda)$ distribution is a location-and-scale parameter. In these cases the exact distributions of the D statistics under the null do not depend on the true parameter values, as long as the null distribution functions are computed using the MLE, which satisfy the conditions required by the David–Johnson result.

As the null distribution of the D statistic for the normal distribution does not depend on the true parameter values, we may assume *any* convenient values of the parameters without affecting the null distribution. This gives rise to the following Monte Carlo procedure to estimate the critical value of D for a given sample size n:

1 Generate a random sample of n (call this the **estimation sample size**) standard normal variates $x_1, \cdots, x_n$. Calculate the sample mean $\bar{x}$ and sample variance

s^2, and use these values to compute the estimated distribution function $F^*(x_i)$, where $F^*(\cdot)$ is the df of $\mathcal{N}(\bar{x}, s^2)$. Then use equation (13.4) to compute D.

2 Repeat Step 1 m times (call this the **Monte Carlo sample size**) to obtain m values of D_j, for $j = 1, \cdots, m$.

3 At the level of significance α, the critical value of the Kolmogorov–Smirnov D statistic is computed as the $(1 - \alpha)$-quantile of the sample of m values of D, estimated using the method in equations (11.9) and (11.10).

Example 15.1 Estimate the critical values of the Kolmogorov–Smirnov D statistic when the null hypothesis is that the observations are distributed normally, with unspecified mean and variance. Let the estimation sample size n be (a) 20 and (b) 30. Assume levels of significance of $\alpha = 0.10$, 0.05, and 0.01.

Solution We use the above procedure to estimate the critical values from Monte Carlo samples of size m = 10,000. The results are summarized in Table 15.1.

Table 15.1. *Results of Example 15.1*

α	$n = 20$	$n = 30$
0.10	0.176	0.146
0.05	0.191	0.158
0.01	0.224	0.185

These values are very close to those in Lilliefors (1967), where Monte Carlo samples of 1,000 were used. Compared to the critical values in Section 13.2.1, we can see that when the parameters are estimated the critical values are lower. Thus, if the estimation is not taken into account and the critical values for the completely specified null are used, the true probability of rejection of the correct model is lower than the stated level of significance α. The following critical values are proposed by Lilliefors (1967) for testing normal distributions with unknown mean and variance

Level of significance α	0.10	0.05	0.01
Critical value	$\frac{0.805}{\sqrt{n}}$	$\frac{0.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$.

□

If the null hypothesis is that the sample observations are distributed as $\mathcal{E}(\lambda)$, where λ is not specified, to estimate the critical values of the

Kolmogorov–Smirnov statistic, the following procedure can be used:

1 Generate a random sample of n variates $x_1, \cdots, x_n$ distributed as $\mathcal{E}(1)$. Calculate the sample mean $\bar{x}$ and compute the estimated distribution function $F^*(x_i)$, where $F^*(\cdot)$ is the df of $\mathcal{E}(1/\bar{x})$. Then use equation (13.4) to compute D.
2 Repeat Step 1 m times to obtain m values of D_j, for $j = 1, \cdots, m$.
3 At the level of significance α, the critical value of the Kolmogorov–Smirnov D statistic is computed as the $(1 - \alpha)$-quantile of the sample of m values of D, estimated using the method in equations (11.9) and (11.10).

Example 15.2 Estimate the critical values of the Kolmogorov–Smirnov D statistic when the null hypothesis is that the observations are distributed as $\mathcal{E}(\lambda)$, where λ is not specified. The estimation sample size n is (a) 20 and (b) 30. Assume levels of significance of $\alpha = 0.10$, 0.05, and 0.01.

Solution We use the above procedure to estimate the critical values from Monte Carlo samples of size m = 10,000. The results are summarized in Table 15.2.

Table 15.2. *Results of Example 15.2*

α	$n = 20$	$n = 30$
0.10	0.214	0.176
0.05	0.236	0.194
0.01	0.279	0.231

These values are very close to those in Lilliefors (1969), where Monte Carlo samples of 1,000 were used. Again we can see that when the parameters are estimated the critical values are lower. The following critical values are proposed by Lilliefors (1969) for testing exponential distributions with unknown mean

Level of significance α	0.10	0.05	0.01
Critical value	$\frac{0.96}{\sqrt{n}}$	$\frac{1.06}{\sqrt{n}}$	$\frac{1.25}{\sqrt{n}}$.

□

15.1.2 Chi-square goodness-of-fit test

As discussed in Section 13.2.3, the asymptotic distribution of the X^2 statistic for the goodness-of-fit test is χ^2_{k-r-1}, where k is the number of groups and r is the number of parameters estimated using the MMLE method. This result

holds asymptotically for any null distribution. Yet Monte Carlo simulation can be used to investigate the performance of the test and improve the estimates of the critical values in small samples if required.

Example 15.3 Estimate the critical values of the chi-square goodness-of-fit statistic X^2 using Monte Carlo simulation when the null hypothesis is that the observations are distributed as $\mathcal{E}(\lambda)$, where λ is unknown. Compute the X^2 statistics based on the MLE using individual observations as well as the MMLE using grouped data.

Solution We group the data into intervals $(c_{i-1}, c_i]$, and use the following four intervals: $(0, 0.4]$, $(0.4, 1]$, $(1, 1.5]$, and $(1.5, \infty)$. The MLE of λ using the complete individual data is $1/\bar{x}$. Let $\boldsymbol{n} = (n_1, \cdots, n_4)$, where n_i is the number of observations in the ith interval. Using grouped data, the MMLE is solved by maximizing the log-likelihood function

$$\log L(\lambda; \boldsymbol{n}) = \sum_{i=1}^{4} n_i \log\left[\exp(-\lambda c_{i-1}) - \exp(-\lambda c_i)\right]$$

with respect to λ. The X^2 statistic is then computed using equation (13.8). Using a Monte Carlo simulation with 10,000 samples, we obtain the estimated critical values of the X^2 statistic summarized in Table 15.3.

Table 15.3. *Results of Example 15.3*

	$n = 50$		$n = 100$		$n = 200$		$n = 300$		
α	MLE	MMLE	MLE	MMLE	MLE	MMLE	MLE	MMLE	$\chi^2_{2,1-\alpha}$
0.10	4.95	4.70	4.93	4.70	4.91	4.61	4.91	4.66	4.61
0.05	6.31	6.07	6.30	6.07	6.38	6.05	6.30	6.04	5.99
0.01	9.45	9.25	9.48	9.39	9.60	9.37	9.41	9.14	9.21

The asymptotic critical values $\chi^2_{2,1-\alpha}$ are shown in the last column. Two points can be observed from the Monte Carlo results. First, the asymptotic results are very reliable even for samples of size 50, if the correct MMLE is used to compute X^2. Second, if MLE is used to compute X^2, the use of $\chi^2_{2,1-\alpha}$ as the critical value will *over-reject* the null hypothesis. □

15.2 Bootstrap estimation of *p*-value

We have discussed test procedures for which there is a unique distribution of the test statistic under the null hypothesis. This enables us to estimate the critical

values or p-values using Monte Carlo simulation. There are situations, however, for which the distribution of the test statistic under the null hypothesis depends on some nuisance parameters not specified under the null. For such problems, tabulation of the critical values is not viable. As an alternative, we may use **bootstrap method** to estimate the p-value of the test statistic.

Consider a sample of n observations $\boldsymbol{x} = (x_1, \cdots, x_n)$ and a test statistic $T(\boldsymbol{x})$ for testing a null hypothesis H_0. Let the computed value of the test statistic for the sample $\boldsymbol{x}$ be t. Suppose the decision rule of the test is to reject H_0 when t is too large (i.e. on the right-hand extreme tail). Furthermore, assume H_0 contains a nuisance parameter θ, which is not specified. We now consider the estimation of the p-value of the test statistic, which is the probability that $T(\boldsymbol{x})$ is larger than t if the null hypothesis is true, i.e.

$$p = \Pr(T(\boldsymbol{x}) > t \,|\, H_0). \tag{15.1}$$

As H_0 contains the nuisance parameter θ, we replace the above problem by

$$p = \Pr(T(\boldsymbol{x}) > t \,|\, H_0(\hat{\theta})), \tag{15.2}$$

where $\hat{\theta}$ is an estimator of θ. The bootstrap estimate of p can be computed as follows:

1 Let the computed value of $T(\boldsymbol{x})$ based on the sample $\boldsymbol{x}$ be t, and let the estimated value of θ be $\hat{\theta}$, which may be any appropriate estimator, such as the MLE.
2 Generate a sample of observations from the distributional assumption of $H_0(\hat{\theta})$, call this $\boldsymbol{x}^*$. Compute the test statistic using data $\boldsymbol{x}^*$ and call this t^*.
3 Repeat Step 2 m times, which is the bootstrap sample size, to obtain m values of the test statistic t_j^*, for $j = 1, \cdots, m$.
4 The estimated p-value of t is computed as

$$\frac{1 + \text{number of } \{t_j^* \geq t\}}{m + 1}. \tag{15.3}$$

The above is a **parametric bootstrap** procedure, in which the samples $\boldsymbol{x}^*$ are generated from a parametric distribution. At level of significance α, the null hypothesis is rejected if the estimated p-value is less than α.

Example 15.4 You are given the following 20 observations of losses:

0.114, 0.147, 0.203, 0.378, 0.410, 0.488, 0.576, 0.868, 0.901, 0.983,
1.049, 1.555, 2.060, 2.274, 4.235, 5.400, 5.513, 5.817, 8.901, 12.699.

(a) Compute the Kolmogorov–Smirnov D statistic, assuming the data are distributed as $\mathcal{P}(\alpha, 5)$. Estimate the p-value of the test statistic using bootstrap.
(b) Repeat (a), assuming the null distribution is $\mathcal{P}(\alpha, 40)$.
(c) Repeat (a), assuming the null distribution is $\mathcal{E}(\lambda)$.

Solution For (a), we estimate α using MLE, which, from Example 12.9, is given by

$$\hat{\alpha} = \frac{20}{\sum_{i=1}^{20} \log(x_i + 5) - 20 \log(5)},$$

and we obtain $\hat{\alpha} = 2.7447$. The computed D statistic, with the null distribution being $\mathcal{P}(2.7447, 5)$, is computed using equation (13.4) to obtain 0.1424. To estimate the p-value, we generate 10,000 bootstrap samples of size 20 each from $\mathcal{P}(2.7447, 5)$, estimate α, and compute the D statistic for each sample. The generation of the Pareto random numbers is done using the inversion method discussed in Example 14.3. The proportion of the D values larger than 0.1424 calculated using equation (15.3) is 0.5775, which is the estimated p-value. Thus, the $\mathcal{P}(\alpha, 5)$ assumption cannot be rejected at any conventional level of significance.

For (b), the MLE of α is

$$\hat{\alpha} = \frac{20}{\sum_{i=1}^{20} \log(x_i + 40) - 20 \log(40)} = 15.8233.$$

The computed D statistic is 0.2138. We generate 10,000 samples of size 20 each from the $\mathcal{P}(15.8233, 40)$ distribution and compute the D statistic of each sample. The estimated p-value is 0.0996. Thus, at the level of significance of 10%, the null hypothesis $\mathcal{P}(\alpha, 40)$ is rejected, but not at the level of significance of 5%.

For (c), the MLE of λ is

$$\hat{\lambda} = \frac{1}{\bar{x}} = 0.3665,$$

and the computed D value is 0.2307. We generate 10,000 samples of size 20 each from the $\mathcal{E}(0.3665)$ distribution using the inversion method. The estimated p-value of the D statistic is 0.0603. Thus, the assumption of $\mathcal{E}(\lambda)$ is rejected at the 10% level, but not at the 5% level.

To conclude, the Kolmogorov–Smirnov test supports the $\mathcal{P}(\alpha, 5)$ distribution assumption for the loss data, but not the $\mathcal{P}(\alpha, 40)$ and $\mathcal{E}(\lambda)$ distributions. □

15.3 Bootstrap estimation of bias and mean squared error

The bootstrap method can also be used to estimate the bias and mean squared error of the parameter estimates of a distribution. Let us consider the estimation of the parameter θ (or a function of the parameter $g(\theta)$) of a distribution using an estimator $\hat{\theta}$ (or $g(\hat{\theta})$), given a random sample of n observations $\boldsymbol{x} = (x_1, \cdots, x_n)$ of X. In situations where theoretical results about the bias and mean squared error of $\hat{\theta}$ (or $g(\hat{\theta})$) are intractable, we may use bootstrap method to estimate these quantities.

When no additional assumption about the distribution of X is made, we may use the empirical distribution defined by $\boldsymbol{x}$ (see Section 11.1.1) as the assumed distribution. We generate a sample of n observations $\boldsymbol{x}^* = (x_1^*, \cdots, x_n^*)$ by re-sampling from $\boldsymbol{x}$ with replacement, and compute the estimate $\hat{\theta}^*$ (or $g(\hat{\theta}^*)$) based on $\boldsymbol{x}^*$. We do this m times to obtain m estimates $\hat{\theta}_j^*$ (or $g(\hat{\theta}_j^*)$), for $j = 1, \cdots, m$. Based on these bootstrap estimates we can compute the bias and the mean squared error of the estimator $\hat{\theta}$ (or $g(\hat{\theta})$). As $\boldsymbol{x}^*$ are generated from the empirical distribution defined by $\boldsymbol{x}$, we call this method **nonparametric bootstrap**.

To illustrate the idea, we consider the use of the sample mean and the sample variance as estimates of the population mean μ and population variance σ^2 of X, respectively. Let μ_{E} and σ_{E}^2 be the mean and the variance, respectively, of the empirical distribution defined by $\boldsymbol{x}$. We note that $\mu_{\mathrm{E}} = \bar{x}$ and $\sigma_{\mathrm{E}}^2 = (n-1)s^2/n$, where $\bar{x}$ and s^2 are the sample mean and the sample variance of $\boldsymbol{x}$, respectively. To use the bootstrap method to estimate the bias and the mean squared error of $\bar{x}$ and s^2, we adopt the following procedure:

1 Generate a random sample of n observations by re-sampling with replacement from $\boldsymbol{x}$, call this $\boldsymbol{x}^* = (x_1^*, \cdots, x_n^*)$. Compute the mean $\bar{x}^*$ and variance s^{*2} of $\boldsymbol{x}^*$.
2 Repeat Step 1 m times to obtain values $\bar{x}_j^*$ and s_j^{*2}, for $j = 1, \cdots, m$.
3 The bias and the mean squared error of $\bar{x}$ are estimated, respectively, by

$$\frac{1}{m}\sum_{j=1}^{m}(\bar{x}_j^* - \mu_{\mathrm{E}}) \qquad \text{and} \qquad \frac{1}{m}\sum_{j=1}^{m}(\bar{x}_j^* - \mu_{\mathrm{E}})^2. \tag{15.4}$$

4 The bias and the mean squared error of s^2 are estimated, respectively, by

$$\frac{1}{m}\sum_{j=1}^{m}(s_j^{*2} - \sigma_{\mathrm{E}}^2) \qquad \text{and} \qquad \frac{1}{m}\sum_{j=1}^{m}(s_j^{*2} - \sigma_{\mathrm{E}}^2)^2. \tag{15.5}$$

It is theoretically known that $\bar{x}$ and s^2 are unbiased for μ and σ^2, respectively. Furthermore, the expected value of $\bar{x}_j^*$ is μ_{E} and the expected value of s_j^{*2} is

σ_{E}^2, so that the bootstrap estimate of the biases should converge to zero when m is large.

The mean squared error of $\bar{x}$ is

$$\mathrm{MSE}(\bar{x}) = \mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}, \tag{15.6}$$

which is unknown (as σ^2 is unknown). On the other hand, the bootstrap estimate of the MSE of $\bar{x}$ in equation (15.4) converges in probability to σ_{E}^2/n, which is known given $\boldsymbol{x}$. However, when $\boldsymbol{x}$ varies $\mathrm{E}(\sigma_{\mathrm{E}}^2/n) = (n-1)\sigma^2/n^2 \neq \mathrm{MSE}(\bar{x})$.

We will not pursue the analytical derivation of $\mathrm{MSE}(s^2)$, which depends on the higher-order moments of X.

Example 15.5 You are given a sample of 20 loss observations of X as in Example 15.4.

(a) Compute the bootstrap estimates of the bias and mean squared error of the sample mean $\bar{x}$.
(b) Compute the bootstrap estimates of the bias and mean squared error of the sample variance s^2.
(c) Compute the bootstrap estimates of the bias and mean squared error of the sample proportion p in estimating $\Pr(X \leq \bar{x})$.

Solution From the data we compute the sample mean (which is also the mean of the empirical distribution μ_{E}) as 2.7286, the sample variance as 11.5442, and the variance of the empirical distribution σ_{E}^2 as 10.9670. For (a) and (b), the bootstrap procedure described above is used. Note that the bootstrap estimate of the mean squared error of $\bar{x}$ should converge to

$$\frac{\sigma_{\mathrm{E}}^2}{n} = \frac{10.9670}{20} = 0.5484.$$

For (c), we compute, for the bootstrap sample j, the proportion of observations not exceeding 2.7286, and call this p_j^*. As 14 out of 20 observations in the sample do not exceed 2.7286, the proportion in the empirical distribution is 0.7. The bias and the mean squared error of the sample proportion p are estimated by

$$\frac{1}{m}\sum_{j=1}^{m}(p_j^* - 0.7) \qquad \text{and} \qquad \frac{1}{m}\sum_{j=1}^{m}(p_j^* - 0.7)^2,$$

respectively. Note that p is unbiased for the proportion in the empirical distribution, and

$$\mathrm{Var}(p) = \frac{(0.7)(1-0.7)}{20} = 0.0105.$$

Using a bootstrap sample of 10,000 runs, we obtain the results in Table 15.4.

Table 15.4. *Results of Example 15.5*

Statistic	Bias	MSE
$\bar{x}$	0.0072	0.5583
s^2	–0.0060	24.2714
p	–0.0017	0.0105

It can be observed that the estimated mean squared error of $\bar{x}$ is very close to its theoretical limit of 0.5484, and so is the estimated mean squared error of p to its theoretical value of 0.0105. Also, the empirical results agree with the theory that the statistics are unbiased. □

We now consider a parametric loss distribution with df $F(\theta)$. The bias and mean squared error of an estimator $\hat{\theta}$ of θ can be estimated using the **parametric bootstrap** method as follows:

1 Compute the estimate $\hat{\theta}$ using sample data $\boldsymbol{x}$.
2 Generate a random sample of n observations from the distribution $F(\hat{\theta})$, and estimate θ, called $\hat{\theta}^*$. Repeat this m times to obtain the estimates $\hat{\theta}_j^*$, for $j = 1, \cdots, m$.
3 Estimates of the bias and mean squared error of $\hat{\theta}$ are computed as

$$\frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_j^* - \hat{\theta}) \qquad \text{and} \qquad \frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_j^* - \hat{\theta})^2. \tag{15.7}$$

Note that the bias and mean squared error of $\hat{\theta}$ can also be estimated using the nonparametric bootstrap, in which case the random samples in Step 2 above are generated by re-sampling with replacement from the data $\boldsymbol{x}$.

Example 15.6 Refer to the data in Example 15.4. Assume the loss data are distributed as $\mathcal{P}(\alpha, 5)$.

(a) Compute the parametric bootstrap estimates of the bias and mean squared error of the MLE of α.
(b) Compute the nonparametric bootstrap estimates of the bias and mean squared error of the MLE of α.

Solution As discussed in Example 15.4, the MLE of α assuming the $\mathcal{P}(\alpha, 5)$ distribution is $\hat{\alpha} = 2.7477$. For parametric bootstrap, we generate the bootstrap samples from the $\mathcal{P}(2.7477, 5)$ distribution and compute the

MLE of α in each bootstrap sample. We simulate 10,000 such samples to compute the bias and the mean squared error using equation (15.7). For the nonparametric bootstrap, the procedure is similar, except that the bootstrap samples are generated by re-sampling with replacement from the original data. The results are summarized in Table 15.5.

Table 15.5. *Results of Example 15.6*

Bootstrap method	Bias of $\hat{\alpha}$	MSE of $\hat{\alpha}$
Parametric	0.1417	0.4792
Nonparametric	0.1559	0.5401

We can see that the results of the parametric and nonparametric bootstraps are comparable. We also note that the MLE is upward biased (at the given sample size). □

15.4 A general framework of bootstrap

Bootstrap is a very versatile statistical method with many important applications. In this section we attempt to provide a framework of the theoretical underpinning of the applications we have discussed. Readers may refer to Davison and Hinkley (1997) for further details of the method.

Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be independently and identically distributed as X with df $F(\cdot)$, which may depend on a parameter θ. Suppose $\xi = \xi(F)$ is a quantity of the distribution (e.g., mean, median, a quantile, or a population proportion) and $\hat{\xi} = \hat{\xi}(\boldsymbol{X})$ is an estimate of ξ based on $\boldsymbol{X}$. We define

$$\eta(\boldsymbol{X}; F) = \hat{\xi}(\boldsymbol{X}) - \xi(F), \tag{15.8}$$

which is the error in estimating ξ using $\hat{\xi}$. Denoting E_F as the expectation taken using the df F, the bias of $\hat{\xi}$ is

$$\mathrm{E}_F[\eta(\boldsymbol{X}; F)] = \mathrm{E}_F[\hat{\xi}(\boldsymbol{X}) - \xi(F)] \tag{15.9}$$

and the mean squared error of $\hat{\xi}$ is

$$\mathrm{E}_F[\eta(\boldsymbol{X}; F)^2] = \mathrm{E}_F[(\hat{\xi}(\boldsymbol{X}) - \xi(F))^2]. \tag{15.10}$$

For another application, let $T(\boldsymbol{X})$ be a test statistic for a hypothesis H_0 and its value computed based on a specific sample $\boldsymbol{x} = (x_1, \cdots, x_n)$ be $t = T(\boldsymbol{x})$.

We now define

$$\eta(\boldsymbol{X}; F) = T(\boldsymbol{X}) - t. \tag{15.11}$$

If H_0 is rejected when t is too large, the p-value of the test is[1]

$$\Pr(T(\boldsymbol{X}) - t > 0 \,|\, F) = \Pr(\eta(\boldsymbol{X}; F) > 0 \,|\, F). \tag{15.12}$$

In the above cases, we are interested in the expectation or the population proportion of a suitably defined function $\eta(\boldsymbol{X}; F)$. This set-up includes the evaluation of bias and mean squared error of an estimator and the p-value of a test, as well as many other applications.

As F is unknown in practice, the quantities in equations (15.9), (15.10), and (15.12) cannot be evaluated. However, we may replace F by a known df F^* and consider instead the quantities

$$\mathrm{E}_{F^*}[\eta(\boldsymbol{X}; F^*)] = \mathrm{E}_{F^*}[\hat{\xi}(\boldsymbol{X}) - \xi(F^*)], \tag{15.13}$$

$$\mathrm{E}_{F^*}[\eta(\boldsymbol{X}; F^*)^2] = \mathrm{E}_{F^*}[(\hat{\xi}(\boldsymbol{X}) - \xi(F^*))^2], \tag{15.14}$$

and

$$\Pr(T(\boldsymbol{X}) - t > 0 \,|\, F^*) = \Pr(\eta(\boldsymbol{X}; F^*) > 0 \,|\, F^*). \tag{15.15}$$

The above quantities are called the bootstrap approximations. The reliability of these approximations depend on how good F^* is as an approximation to F. If F^* is taken as the empirical distribution defined by $\boldsymbol{x}$, we have a nonparametric bootstrap. If, however, F^* is taken as $F(\hat{\theta})$ for a suitable estimator $\hat{\theta}$ computed from the sample $\boldsymbol{x}$, then we have a parametric bootstrap.

As $\hat{\xi}(\boldsymbol{X})$ and $T(\boldsymbol{X})$ may be rather complex functions of $\boldsymbol{X}$, the evaluation of equations (15.13), (15.14), and (15.15) may remain elusive even with known or given F^*. In the case where the sample size n is small and the empirical distribution is used for F^*, we may evaluate these quantities by exhausting all possible samples of $\boldsymbol{X}$. This approach, however, will not be feasible when n is large or when a parametric df $F(\hat{\theta})$ is used. In such situations the quantities may be estimated using Monte Carlo methods, and we call the solution the Monte Carlo estimate of the bootstrap approximate, or simply the bootstrap estimate. Specifically, for nonparametric bootstrap, we re-sample with replacement from $\boldsymbol{x}$ and then compute the sample mean or the sample proportion as required. For parametric bootstrap, however, the samples are generated from the distribution with df $F(\hat{\theta})$.

[1] Tests which are two sided or with critical regions on the left-hand tail can also be defined appropriately.

15.5 Monte Carlo simulation of asset prices

We now consider continuous-time models of prices of assets, such as stocks, bonds, and commodities. We shall review briefly the basic Wiener process, and then consider its extension to the generalized Wiener process and diffusion process. These processes are important for the pricing of derivative securities, the payoffs of which are contingent on the prices of the underlying assets.

15.5.1 Wiener process and generalized Wiener process

Let W_t be a stochastic process over time t with the following properties:

1 Over a small time interval Δt, the change in W_t, denoted by $\Delta W_t = W_{t+\Delta t} - W_t$, satisfies the property

$$\Delta W_t = \epsilon\sqrt{\Delta t}, \tag{15.16}$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

2 If ΔW_{t_1} and ΔW_{t_2} are changes in the process W_t over two nonoverlapping intervals, then ΔW_{t_1} and ΔW_{t_2} are independent.

A continuous-time stochastic process satisfying the above two properties is called a **Wiener process** or **standard Brownian motion**. From the first of these two properties, we can conclude that

$$\mathrm{E}(\Delta W_t) = 0, \tag{15.17}$$

and

$$\mathrm{Var}(\Delta W_t) = \Delta t. \tag{15.18}$$

Furthermore, for the change over a finite interval $[0, T]$, we can partition the interval into N nonoverlapping small segments of length Δt each, such that

$$T = N(\Delta t) \tag{15.19}$$

and

$$W_T - W_0 = \sum_{i=0}^{N-1} \Delta W_{i(\Delta t)} = \sum_{i=1}^{N} \epsilon_i\sqrt{\Delta t}, \tag{15.20}$$

where ϵ_i, for $i = 1, \cdots, n$, are iid $\mathcal{N}(0, 1)$. Thus, given the information at time 0 we have

$$\mathrm{E}(W_T) = W_0 + \sum_{i=1}^{N} \mathrm{E}(\epsilon_i)\sqrt{\Delta t} = W_0, \tag{15.21}$$

and

$$\text{Var}(W_T) = \sum_{i=1}^{N} \text{Var}(\epsilon_i)\Delta t = \sum_{i=1}^{N} \Delta t = T. \tag{15.22}$$

From (15.20) we can see that W_T is the sum of W_0 and N iid normal variates, which implies, given W_0

$$W_T \sim \mathcal{N}(W_0, T). \tag{15.23}$$

Hence, W_T is normally distributed, with its variance increasing linearly with time T.

The Wiener process can be extended to allow for a *drift* in the process and a constant volatility parameter. Thus, we consider a **generalized Wiener process** or **Brownian motion** X_t, which has the following change over a small time interval Δt

$$\Delta X_t = a\,\Delta t + b\,\Delta W_t, \tag{15.24}$$

where a is the drift rate and b is the volatility rate (a and b are constants), and W_t is a Wiener process. It can be verified that, given X_0, we have

$$X_T \sim \mathcal{N}(X_0 + aT, b^2T), \tag{15.25}$$

for any finite T.

The Wiener and generalized Wiener processes are continuous-time processes when $\Delta t \to 0$ in equations (15.16) and (15.24), respectively. Thus, we shall write the differentials of these processes as dW_t and

$$dX_t = a\,dt + b\,dW_t, \tag{15.26}$$

respectively.

15.5.2 Diffusion process and lognormal distribution

A further extension of equation (15.26) is to allow the drift and volatility rates to depend on time t and the process value X_t. Thus, we consider the process

$$dX_t = a(X_t, t)\,dt + b(X_t, t)\,dW_t, \tag{15.27}$$

which is called an **Ito process** or **diffusion process**. The terms $a(X_t, t)$ and $b(X_t, t)$ are called the **drift rate** and the **diffusion coefficient**, respectively.

The flexibility of allowing the drift and diffusion to depend on the process value and time introduces much complexity into the problem. First, X_t is in general no longer normally distributed. Second, for many practical problems, the distribution of X_t is unknown. We now consider a specific member of diffusion processes, called the **geometric Brownian motion**, which has been commonly used to model asset prices.

Let S_t be the price of an asset at time t. S_t is said to follow a geometric Brownian motion if

$$dS_t = \mu S_t\, dt + \sigma S_t\, dW_t, \tag{15.28}$$

where μ, called the **instantaneous rate of return**, and σ, called the **volatility rate**, are constants. The above equation can also be written as

$$\frac{1}{S_t}\, dS_t = \mu\, dt + \sigma\, dW_t. \tag{15.29}$$

Further analysis shows that[2]

$$d \log S_t = \left(\mu - \frac{\sigma^2}{2}\right) dt + \sigma\, dW_t, \tag{15.30}$$

so that $\log S_t$ follows a generalized Wiener process and hence is normally distributed. Thus, following equation (15.25), we conclude

$$\log S_t \sim \mathcal{N}\left(\log S_0 + \left(\mu - \frac{\sigma^2}{2}\right) t,\, \sigma^2 t\right), \tag{15.31}$$

so that S_t is lognormally distributed with mean (see Appendix A.10.2)

$$\mathrm{E}(S_t) = \exp\left[\log S_0 + \left(\mu - \frac{\sigma^2}{2}\right) t + \frac{\sigma^2 t}{2}\right] = S_0 \exp(\mu t). \tag{15.32}$$

Equation (15.31) can also be written as

$$\log S_t - \log S_0 = \log\left(\frac{S_t}{S_0}\right) \sim \mathcal{N}\left(\left(\mu - \frac{\sigma^2}{2}\right) t,\, \sigma^2 t\right), \tag{15.33}$$

so that

$$\log\left(\frac{S_t}{S_0}\right) = \left(\mu - \frac{\sigma^2}{2}\right) t + \sigma\sqrt{t} Z, \tag{15.34}$$

[2] Note that it is erroneous to conclude from equation (15.29) that its left-hand side is equal to $d \log S_t$ so that $d \log S_t = \mu\, dt + \sigma\, dW_t$. The equation $d \log S_t / dS_t = 1/S_t$ is not valid as S_t is a *stochastic variable*, not a *mathematical variable*. The result in equation (15.30) can be proved using Ito's lemma (see McDonald, 2006, Chapter 20).

where

$$Z \sim \mathcal{N}(0, 1). \tag{15.35}$$

Note that

$$R \equiv \frac{1}{t} \log\left(\frac{S_t}{S_0}\right) \tag{15.36}$$

is the **continuously compounded rate of return** over the interval $[0, t]$.[3] Thus, from equation (15.34), the expected continuously compounded rate of return over the finite interval $[0, t]$ is

$$\mathrm{E}\left[\frac{1}{t} \log\left(\frac{S_t}{S_0}\right)\right] = \mu - \frac{\sigma^2}{2}, \tag{15.37}$$

which is less than the instantaneous rate of return μ.

The total return of an asset consists of two components: capital gain and dividend yield. As S_t is the price of the asset, μ as defined in equation (15.28) captures the instantaneous capital gain only. If the dividend yield is assumed to be continuous at the rate δ, then the total instantaneous return, denoted by μ^*, is given by

$$\mu^* = \mu + \delta. \tag{15.38}$$

Hence, expressed in terms of the total return and the dividend yield, the expected continuously compounded rate of capital gain (asset-price appreciation) is

$$\mathrm{E}\left[\frac{1}{t} \log\left(\frac{S_t}{S_0}\right)\right] = \mu - \frac{\sigma^2}{2} = \mu^* - \delta - \frac{\sigma^2}{2}. \tag{15.39}$$

We now consider the simulation of asset prices that follow the geometric Brownian motion given in equation (15.28), in which the parameter μ captures the return due to asset-price appreciation. From equation (15.34), we obtain

$$S_t = S_0 \exp\left[\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma\sqrt{t} Z\right], \tag{15.40}$$

which can be used to simulate price paths of the asset. In practical applications, we need the values of the parameters σ and

$$\mu_R = \mu - \frac{\sigma^2}{2}. \tag{15.41}$$

[3] Note that equation (15.36) can be rewritten as $S_t = S_0\, e^{Rt}$, which shows that R is the continuously compounded rate of return over the interval $[0, t]$.

If we desire to simulate price paths in the *real world* or under a *physical probability measure*, we would estimate the parameters of the price process using historical return data.[4] Suppose we sample return data of the asset over intervals of length h. Let there be n return observations (computed as differences of logarithmic asset prices) with mean $\bar{x}_R$ and sample variance s_R^2. The required parameter estimates are then given by

$$\hat{\mu}_R = \frac{\bar{x}_R}{h} \qquad \text{and} \qquad \hat{\sigma} = \frac{s_R}{\sqrt{h}}. \tag{15.42}$$

The asset prices at intervals of h can be simulated recursively using the equation[5]

$$S_{t+(i+1)h} = S_{t+ih} \exp\left[\bar{x}_R + s_R Z_i\right], \qquad \text{for } i = 0, 1, 2, \cdots, \tag{15.43}$$

where Z_i are iid $\mathcal{N}(0, 1)$.

For illustration, we use end-of-day S&P500 index values for the period January 3, 2007, through December 28, 2007. There are in total 250 index values and we compute 249 daily returns, which are the logarithmic price differences. If the stock price index indeed follows a geometric Brownian motion, we would expect the returns to be normally distributed. The price index graph and the return graph are plotted in Figure 15.1. We also plot the histogram of the return observations as well as the normal probability plot of the return data.[6] It can be observed that there is clear deviation of the return data from the normality assumption. We shall, however, ignore this deviation, and continue with the simulation of the price path using the geometric Brownian motion assumption.

We estimate the parameters of the price process and obtain $\bar{x}_R = 0.0068\%$ and $s_R = 1.0476\%$. Note that these values are in percent per day. If we take $h = 1/250$, the annualized estimate of σ is $1.0476\sqrt{250}\% = 16.5640\%$ per annum. The estimate of μ is

$$\hat{\mu} = \hat{\mu}_R + \frac{\hat{\sigma}^2}{2} = \frac{\bar{x}_R}{h} + \frac{s_R^2}{2h} = 3.0718\% \text{ per annum.} \tag{15.44}$$

We use these values to simulate the price paths, an example of which is presented in Figure 15.2.

[4] If the price process under the *risk-neutral world* is desired, the parameters required will be different. Simulation of risk-neutral processes is important for the pricing of derivative securities. See McDonald (2006, Chapter 20) for more details.

[5] Note that S_{t+ih} stands for the ith simulated price value, which is the value at chronological time $t + ih$.

[6] The normal probability plot plots the proportion of values in the sample that are less than or equal to the order statistics of the data against the order statistics. The vertical axis is scaled in such a way that if the data come from a normal distribution, the points should fall approximately on a straight line.

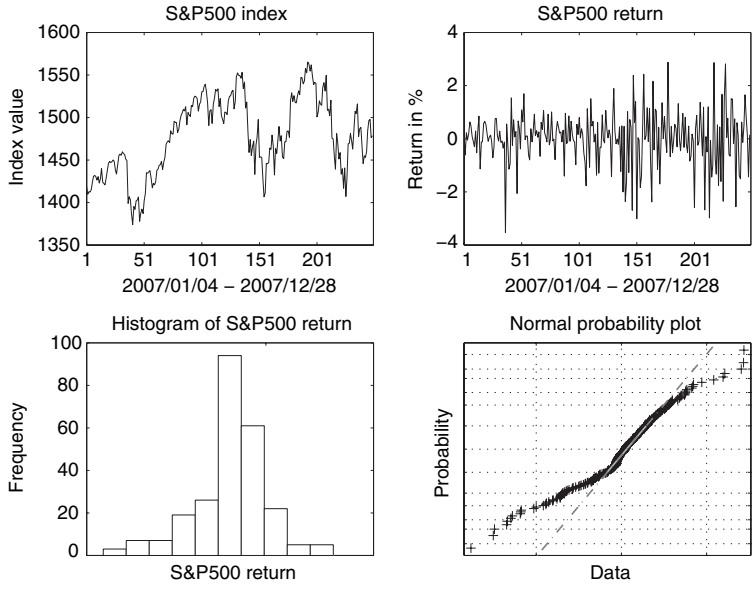

Figure 15.1 S&P500 price index and logarithmic difference

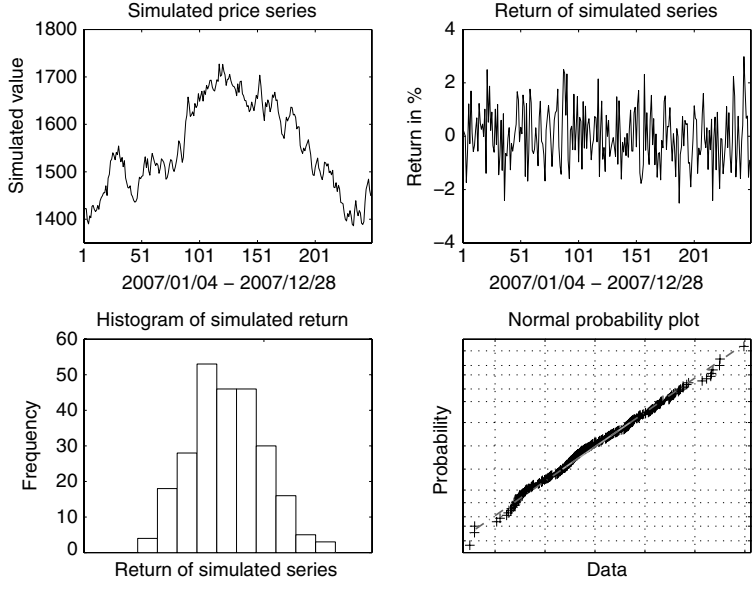

Figure 15.2 Simulated price index and logarithmic difference

A notable difference between the simulated price process versus the actual S&P500 series is that the normal probability plot of the simulated series agrees quite closely with the hypothesis that the returns are normally distributed. This is due to the fact that the price processes are actually simulated as a geometric Brownian motion, so that the prices are indeed lognormally distributed. Also, the volatility of the S&P500 series appears to be higher in the second half of the year, while that of the simulated series appears to be quite even. In other words, there is some *volatility clustering* in the real series, but not in the simulated series.

15.5.3 Jump–diffusion process

Asset prices following a diffusion process are characterized by paths that are *continuous* in time. Anecdotal evidence, however, often suggests that stock prices are more *jumpy* than what would be expected of a diffusion process. To allow for discrete jumps in asset prices, we introduce a jump component into the diffusion process and consider asset prices following a **jump–diffusion process**. In particular, we consider augmenting the geometric Brownian motion with a jump component. To this effect, we assume the occurrence of a jump in an interval has a Poisson distribution, and when a jump occurs, the jump size is distributed normally. Thus, the aggregate jump within a finite interval resembles the aggregate loss as discussed in Chapter 3.

We define N_t as a Poisson process with intensity (mean per unit time) λ.[7] Thus, over a small time interval Δt, the probability of occurrence of a jump event is $\lambda \Delta t + o(\Delta t)$, while the probability of no jump is $1 - \lambda \Delta t + o(\Delta t)$ and that of more than one jump is $o(\Delta t)$, where $o(\Delta t)$ is a term that tends to zero faster than Δt. ΔN_t is the number of jump events occurring in the interval $(t, t + \Delta t]$. We use the notation dN_t when $\Delta t \to 0$.

We now augment the geometric Brownian motion in equation (15.30) with a jump component as follows

$$d \log S_t = \left(\mu - \frac{\sigma^2}{2} \right) dt + \sigma \, dW_t + J_t \, dN_t - \lambda \mu_J \, dt, \tag{15.45}$$

where $J_t \sim \mathcal{N}(\mu_J, \sigma_J^2)$ and is distributed independently of N_t. Note that the mean of $J_t \, dN_t$ is $\lambda \mu_J \, dt$, so that the mean of the augmented component $J_t \, dN_t - \lambda \mu_J \, dt$ is zero. In other words, the addition of the term $-\lambda \mu_J \, dt$ is to center the jump component so that its mean is equal to zero. This property is of important significance because jumps are often assumed to be idiosyncratic and do not affect the expected return of the stock. Also, as the variance of dW_t and $J_t \, dN_t$

[7] See Section 5.3 for the definition of Poisson process.

are dt and $\lambda(\mu_J^2 + \sigma_J^2)dt$, respectively, the variance rate of the jump diffusion process is $\sigma^2 + \lambda(\mu_J^2 + \sigma_J^2)$.

We now re-write equation (15.45) as

$$d \log S_t = \left(\mu - \lambda\mu_J - \frac{\sigma^2}{2}\right) dt + \sigma\, dW_t + J_t\, dN_t, \tag{15.46}$$

from which we can see that the first component is a geometric Brownian motion with an adjusted drift rate of $\mu - \lambda\mu_J - \sigma^2/2$. If $\mu_J > 0$, the jump component induces price appreciation on average, and the diffusion part of the price will have a drift term adjusted downwards. On the other hand, if $\mu_J < 0$, investors will be compensated by a higher drift rate to produce the same expected return.

To simulate the jump–diffusion process defined in equation (15.46), we first consider the jump component. Suppose the time interval of the prices simulated is h, to simulate the jump component $J_t\, dN_t$ we generate a number m from the $\mathcal{PN}(\lambda h)$ distribution, and then simulate m independent variates Z_i from the $\mathcal{N}(0, 1)$ distribution. The jump component is then given by

$$m\mu_J + \sigma_J \sum_{i=1}^{m} Z_i. \tag{15.47}$$

The diffusion component is computed as

$$\left(\mu - \lambda\mu_J - \frac{\sigma^2}{2}\right) h + \sigma\sqrt{h}Z, \tag{15.48}$$

where Z is a standard normal variate independent of Z_i. To generate the value of $S_{t+(i+1)h}$ given S_{t+ih} we use the equation

$$\begin{aligned} S_{t+(i+1)h} = S_{t+ih} \exp\left[\left(\mu - \lambda\mu_J - \frac{\sigma^2}{2}\right) h + \sigma\sqrt{h}Z\right] \\ \times \exp\left[m\mu_J + \sigma_J \sum_{i=1}^{m} Z_i\right]. \end{aligned} \tag{15.49}$$

The above computation can be repeated recursively to simulate a price path of S_t.

For illustration we simulate a jump–diffusion process using the following parameters: $\mu = 3.0718\%$, $\sigma = 16.5640\%$, $\lambda = 3$, $\mu_J = -2\%$, and $\sigma_J = 3\%$ (the first three quantities are per annum). Thus, the jumps occur on average three times per year, and each jump is normally distributed with mean jump size of 2% down and standard deviation of 3%. To observe more jumps in the simulated

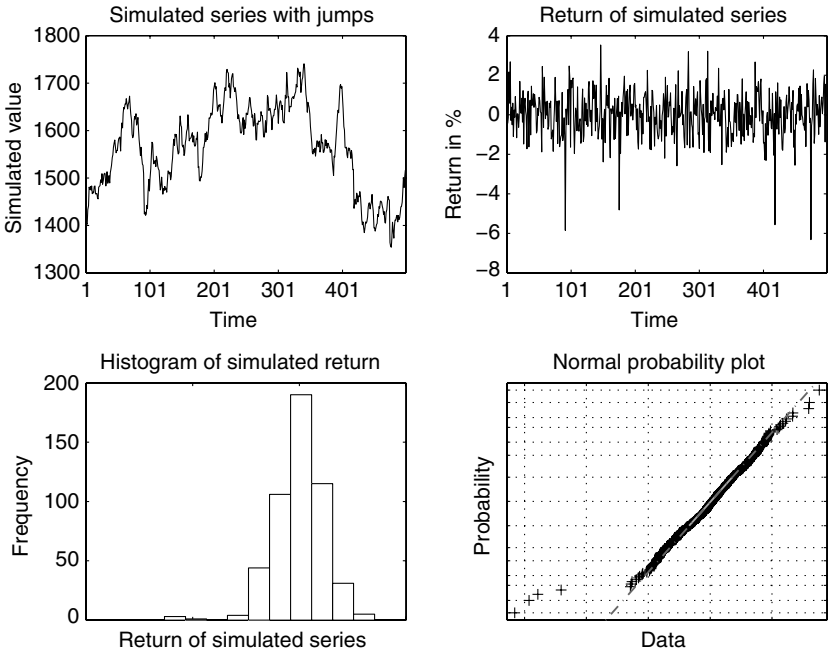

Figure 15.3 Simulated jump–diffusion and logarithmic difference

process, we simulate 500 daily observations (about 2 years) and an example is presented in Figure 15.3. From the plot of the return series, we can see some obvious jumps downwards. The histogram shows a negative skewness, due to the jumps in the process. In the normal probability plot, most of the points lie closely to the straight line, apart from a few that lie on the extreme lower tail. This is due to the aggregation of the lognormal component and the jump component to form the price series.

15.6 Summary and discussions

We have discussed some applications of Monte Carlo methods for the analysis of actuarial and financial data. Using simulated samples we can estimate the critical values and p-values of the test statistics when their exact values are unknown. These estimates are viable, however, only when there are no nuisance parameters determining the distribution of the test statistic. In cases where nuisance parameters are present, the p-values of the tests can be estimated using the bootstrap method. We discuss parametric and nonparametric bootstrap methods. They prove to be very valuable tools for model testing.

For pricing derivative securities we often require numerical solutions based on the simulation of the prices of the underlying assets. We discuss the simulation of price paths of geometric Brownian motions as well as jump–diffusion processes that incorporate both the geometric Brownian motion and a jump component. Our illustrations show distinct features of normal returns and discrete jumps in the simulated data.

Exercises

15.1 Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be a random sample of n observations of X.
(a) How would you estimate the bias and the mean squared error of the sample median as an estimator of the median of X using nonparametric bootstrap?
(b) How would you estimate the bootstrap sample size required if it is desired to estimate the bias of the sample median to within 0.05 with probability of 0.95?
(c) If X is assumed to be exponentially distributed, how would you perform a parametric bootstrap for the above problems?

15.2 Let θ denote the interquartile range of X, which is defined as the 75th percentile minus the 25th percentile. You are required to estimate the 95% confidence interval of θ based on the distribution of $\hat{\theta}/\theta$, where $\hat{\theta}$ is the difference between the 75th percentile and the 25th percentile of the sample $\boldsymbol{x} = (x_1, \cdots, x_n)$. How would you estimate the 95% confidence interval of θ using nonparametric bootstrap?

15.3 Let (X, Y) be a bivariate random variable with correlation coefficient ρ. You have a random sample of n observations (x_i, y_i), for $i = 1, \cdots, n$ with sample correlation coefficient $\hat{\rho}$.
(a) How would you estimate the bias of $\hat{\rho}$ as an estimator of ρ using nonparametric bootstrap?
(b) If (X, Y) follows a bivariate normal distribution, how would you estimate the bias of $\hat{\rho}$ as an estimator of ρ using parametric bootstrap? [*Hint:* It can be shown that the distribution of $\hat{\rho}$ depends only on ρ, not on the means and standard deviations of X and Y.]

15.4 Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be a random sample of n observations of $X \sim \mathcal{L}(\mu, \sigma^2)$. The coefficient of variation of X is $\theta = [\mathrm{Var}(X)]^{\frac{1}{2}} / \mathrm{E}(X) = (e^{\sigma^2} - 1)^{\frac{1}{2}}$. The maximum likelihood estimate of θ is $\hat{\theta} = (e^{\hat{\sigma}^2} - 1)^{\frac{1}{2}}$, where $\hat{\sigma}^2$ is the maximum likelihood estimate of σ^2. How would you estimate the bias of $\hat{\theta}$ as an estimator of θ using parametric bootstrap?

Questions adapted from SOA exams

15.5 For an insurance policy covering both fire and wind losses, a sample of fire losses were found to be 3 and 4, and wind losses in the same period were 0 and 3. Fire and wind losses are independent, but do not have identical distributions. Based on the sample, you estimate that adding a policy deductible of 2 per wind claim will eliminate 20% of the insured loss. Determine the bootstrap approximation to the mean squared error of the estimate.

15.6 Three observed values of the random variable X are: 1, 1, and 4. You estimate the third moment of X using the estimator $\hat{\mu}'_3 = \left[\sum_{i=1}^{3}(X_i - \bar{X})^3\right]/3$. Determine the bootstrap estimate of the mean squared error of $\hat{\mu}'_3$.

15.7 You are given a sample of two values of the distribution X, 5 and 9, and you estimate σ^2, the variance of X, using the estimator $\hat{\sigma}^2 = \left[\sum_{i=1}^{2}(X_i - \bar{X})^2\right]/2$. Determine the bootstrap approximation to the mean squared error of $\hat{\sigma}^2$.

15.8 The price of a stock at time t, S_t, is to be forecasted using simulation. It is assumed that

$$\log\left(\frac{S_t}{S_0}\right) \sim \mathcal{N}\left(\left[\alpha - \frac{\sigma^2}{2}\right]t,\, \sigma^2 t\right),$$

with $S_0 = 50, \alpha = 0.15$, and $\sigma = 0.3$. Three prices for S_2 are simulated using $\mathcal{U}(0, 1)$ variates and the inverse transformation method, where small values of $\mathcal{U}(0, 1)$ correspond to small stock prices. The following $\mathcal{U}(0, 1)$ variates are generated: 0.9830, 0.0384, and 0.7794. Calculate the mean of S_2 generated.

15.9 The prices of a stock taken at the end of each month for seven consecutive months are:

Month	Price
1	54
2	56
3	48
4	55
5	60
6	58
7	62

Estimate the annualized expected instantaneous rate of return of the stock assuming it does not pay any dividend.

Appendix

Review of statistics

This Appendix provides a review of the statistical tools and literature required for this book. It summarizes background material found in introductory probability textbooks as well as develops required results for use in the main text. Readers who require a quick revision may study this Appendix prior to reading the main text. Otherwise, this Appendix may be used for reference only. For the purpose of being self-contained some of the results developed in the main text are recapped here.

Students who wish to go deeper in the statistics literature will find the following texts useful: DeGroot and Schervish (2002), Hogg and Craig (1995), and Ross (2006).

A.1 Distribution function, probability density function, probability function, and survival function

If X is a random variable, the **distribution function (df)** of X evaluated at x, denoted by $F_X(x)$, is defined as

$$F_X(x) = \Pr(X \leq x). \tag{A.1}$$

X is a continuous random variable if its df $F_X(x)$ is continuous. In addition, if $F_X(x)$ is differentiable, the **probability density function (pdf)** of X, denoted by $f_X(x)$, is defined as

$$f_X(x) = \frac{dF_X(x)}{dx}. \tag{A.2}$$

If X is discrete and takes possible countable values x_i for $i = 1, \cdots, n$, where n may be finite or infinite, then the **probability function (pf)** of X is

$$f_X(x_i) = \Pr(X = x_i). \tag{A.3}$$

Thus, $f_X(x) = \Pr(X = x_i)$ if $x = x_i$ for some i and zero otherwise. We denote $\Omega_X = \{x_1, x_2, \cdots\}$ as the set of discrete values X can take. For a random variable X, whether continuous or discrete, the set of all possible values X can take (countable if X is discrete and uncountable if X is continuous) is called the **support** of X.

The **survival function (sf)** (also called the **decumulative distribution function** or the **survival distribution function**) of a random variable X, denoted by $S_X(x)$, is

$$S_X(x) = 1 - F_X(x) = \Pr(X > x). \tag{A.4}$$

The df $F_X(x)$ is monotonic nondecreasing, the pdf $f_X(x)$ is nonnegative, and the sf $S_X(x)$ is monotonic nonincreasing. Also, we have $F_X(-\infty) = S_X(\infty) = 0$ and $F_X(\infty) = S_X(-\infty) = 1$. If X is positive, then $F_X(0) = 0$ and $S_X(0) = 1$. The following equations express the df in terms of the pdf

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx, \quad \text{for continuous } X, \tag{A.5}$$

and

$$F_X(x) = \sum_{x_i \le x} f_X(x_i), \quad \text{for discrete } X. \tag{A.6}$$

A.2 Random variables of the mixed type and Stieltjes integral

Some random variables may have a mix of discrete and continuous components. A random variable X is said to be of the **mixed type** if its df $F_X(x)$ is continuous and differentiable in the support apart from at the points belonging to a countable set, say, Ω_X^D. Thus, there exists a function $f_X(x)$ such that[1]

$$F_X(x) = \Pr(X \le x) = \int_{-\infty}^{x} f_X(x)\, dx + \sum_{x_i \in \Omega_X^D, x_i \le x} \Pr(X = x_i). \tag{A.7}$$

We use the **differential** $dF_X(x)$ to mean the probability of X in the infinitesimal interval $[x, x + dx)$, i.e.

$$dF_X(x) = \Pr\{X \in [x, x + dx)\}. \tag{A.8}$$

[1] Note that $f_X(x)$ is the derivative of $F_X(x)$ at the points where $F_X(x)$ is continuous and differentiable, but it is not the pdf of X. In particular, $\int_{-\infty}^{\infty} f_X(x)\, dx \neq 1$.

If $F_X(x)$ has a jump at a point x, i.e. there is a **probability mass** at x, then

$$dF_X(x) = \Pr(X = x). \tag{A.9}$$

On the other hand, if $F_X(x)$ has a derivative $f_X(x)$ at point x, we have

$$dF_X(x) = f_X(x)\,dx. \tag{A.10}$$

We use the convenient notation of the **Stieltjes integral** to state that[2]

$$\Pr(a \le X \le b) = \int_a^b dF_X(x), \tag{A.11}$$

for any interval $[a, b]$ in the support of X. This expression incorporates continuous, discrete, and mixed random variables, where the df $F_X(x)$ may be any one of (A.5), (A.6), or (A.7).

A.3 Expected value

Let $g(x)$ be a function of x, the **expected value** of $g(X)$, denoted by $\mathrm{E}[g(X)]$, is defined as the Stieltjes integral

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\,dF_X(x), \tag{A.12}$$

which is equal to

$$\int_{-\infty}^{\infty} g(x) f_X(x)\,dx, \quad \text{if } X \text{ is continuous,} \tag{A.13}$$

$$\sum_{x_i \in \Omega_X} g(x_i)\ \Pr(X = x_i), \quad \text{if } X \text{ is discrete,} \tag{A.14}$$

and

$$\int_{-\infty}^{\infty} g(x) f_X(x)\,dx + \sum_{x_i \in \Omega_X^D} g(x_i)\ \Pr(X = x_i), \quad \text{if } X \text{ is mixed.} \tag{A.15}$$

Thus, the use of the Stieltjes integral conveniently simplifies the notations. If X is continuous and nonnegative, and $g(\cdot)$ is a nonnegative, monotonic, and differentiable function, the following result holds

$$\mathrm{E}[g(X)] = \int_0^{\infty} g(x)\,dF_X(x) = g(0) + \int_0^{\infty} g'(x)[1 - F_X(x)]\,dx, \tag{A.16}$$

[2] For the definition of Stieltjes integral, see Ross (2006, p. 404).

where $g'(x)$ is the derivative of $g(x)$ with respect to x. If X is discrete and nonnegative, taking values $0, 1, \cdots$, we have

$$\mathrm{E}[g(X)] = g(0) + \sum_{x=0}^{\infty} [1 - F_X(x)]\, \Delta g(x), \tag{A.17}$$

where $\Delta g(x) = g(x+1) - g(x)$.

To prove equation (A.16), we note that, using integration by parts, we have

$$\begin{aligned} \int_0^t g(x)\, dF_X(x) &= -\int_0^t g(x)\, d\,[1 - F_X(x)] \\ &= -g(t)\,[1 - F_X(t)] + g(0) + \int_0^t g'(x)[1 - F_X(x)]\, dx. \end{aligned} \tag{A.18}$$

It is thus sufficient to show that

$$\lim_{t\to\infty} g(t)\,[1 - F_X(t)] = 0. \tag{A.19}$$

The above equation obviously holds if $g(\cdot)$ is nonincreasing. If $g(\cdot)$ is nondecreasing, we have

$$g(t)\,[1 - F_X(t)] = g(t) \int_t^{\infty} f_X(x)\, dx \le \int_t^{\infty} g(x) f_X(x)\, dx. \tag{A.20}$$

As $\mathrm{E}[g(X)]$ exists, the last expression above tends to 0 as $t \to \infty$, which completes the proof.

A.4 Mean, variance, and other moments

The **mean** of X is

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, dF_X(x). \tag{A.21}$$

If X is continuous and nonnegative, we apply equation (A.16) to obtain[3]

$$\mathrm{E}(X) = \int_0^{\infty} [1 - F_X(x)]\, dx = \int_0^{\infty} S_X(x)\, dx. \tag{A.22}$$

[3] We need to replace the integral by a summation when X is discrete and nonnegative.

The **variance** of X, denoted by $\mathrm{Var}(X)$, is defined as

$$\mathrm{Var}(X) = \mathrm{E}\left\{[X - \mathrm{E}(X)]^2\right\} = \int_{-\infty}^{\infty} [x - \mathrm{E}(X)]^2 \, dF_X(x). \tag{A.23}$$

The kth **moment about zero**, also called the kth raw moment, of X (for $k \geq 1$), denoted by μ_k', is defined as

$$\mu_k' = \mathrm{E}(X^k) = \int_{-\infty}^{\infty} x^k \, dF_X(x). \tag{A.24}$$

Thus, $\mu_1' = \mathrm{E}(X)$. The kth **moment about the mean**, also called the kth **central moment**, of X (for $k > 1$), denoted by μ_k, is defined as

$$\mu_k = \mathrm{E}[(X - \mu_1')^k] = \int_{-\infty}^{\infty} (x - \mu_1')^k \, dF_X(x). \tag{A.25}$$

We have the relationship

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2, \tag{A.26}$$

i.e.

$$\mu_2 = \mu_2' - (\mu_1')^2. \tag{A.27}$$

If X is symmetric about the mean μ_1', the third central moment μ_3 is zero. The standardized measure of **skewness** is defined as

$$\text{skewness} = \frac{\mu_3}{\sigma^3}. \tag{A.28}$$

The standardized measure of the fourth moment is called the **kurtosis**, which is defined as

$$\text{kurtosis} = \frac{\mu_4}{\sigma^4}. \tag{A.29}$$

The kurtosis measures the *thickness* of the tail distribution, with a value of 3 for the normal distribution. The **coefficient of variation** of X is defined as

$$\frac{\sqrt{\mathrm{Var}(X)}}{\mathrm{E}(X)} = \frac{\sqrt{\mu_2}}{\mu_1'}. \tag{A.30}$$

A.5 Conditional probability and Bayes' theorem

If A and B are nonnull events in a sample space S, then

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \tag{A.31}$$

which can also be written as

$$\Pr(A \cap B) = \Pr(A \mid B) \Pr(B). \tag{A.32}$$

This is called the **multiplication rule** of probability.

If $B_1, B_2, \cdots, B_n$ are **mutually exclusive and exhaustive events** of S, i.e.

$$\bigcup_{i=1}^{n} B_i = S \quad \text{and} \quad B_i \cap B_j = \emptyset \ \text{ for } i \neq j, \tag{A.33}$$

then extending the multiplication rule, we have

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A \mid B_i) \Pr(B_i). \tag{A.34}$$

Now applying the multiplication rule to $\Pr(B_i \mid A)$ for any $i \in \{1, 2, \cdots, n\}$, we have

$$\begin{aligned} \Pr(B_i \mid A) &= \frac{\Pr(B_i \cap A)}{\Pr(A)} \\ &= \frac{\Pr(A \mid B_i) \Pr(B_i)}{\sum_{i=1}^{n} \Pr(A \mid B_i) \Pr(B_i)}. \end{aligned} \tag{A.35}$$

This result is called Bayes' Theorem.

A.6 Bivariate random variable

The **joint distribution function (joint df)** of the bivariate random variable (X, Y), denoted by $F_{XY}(x, y)$, is defined as

$$F_{XY}(x, y) = \Pr(X \leq x, Y \leq y). \tag{A.36}$$

If $F_{XY}(x, y)$ is continuous and differentiable with respect to x and y, the **joint probability density function (joint pdf)** of X and Y, denoted by $f_{XY}(x, y)$, is defined as

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \, \partial y}. \tag{A.37}$$

The pdf of X and Y are, respectively

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy, \tag{A.38}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx, \tag{A.39}$$

which are called the **marginal pdf**. The **marginal df** of X and Y, denoted by $F_X(x)$ and $F_Y(y)$, can be obtained from the marginal pdf using equation (A.5).

If X and Y are random variables with marginal densities $f_X(x)$ and $f_Y(y)$, respectively, and the joint pdf of X and Y is $f_{XY}(x, y)$, then the **conditional pdf** of X given Y, denoted by $f_{X \mid Y}(x \mid y)$, is

$$f_{X \mid Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}. \tag{A.40}$$

The above equation can also be used to compute the joint pdf from the conditional pdf and marginal pdf, i.e.

$$f_{XY}(x, y) = f_{X \mid Y}(x \mid y) f_Y(y). \tag{A.41}$$

Let dx and dy be small changes in x and y, respectively. If we multiply $f_{X \mid Y}(x \mid y)$ by dx we have

$$f_{X \mid Y}(x \mid y)\, dx = \Pr(x \le X \le x + dx \mid y \le Y \le y + dy). \tag{A.42}$$

Substituting equation (A.39) for $f_Y(y)$ into (A.40), we obtain

$$f_{X \mid Y}(x \mid y) = \frac{f_{XY}(x, y)}{\int_{-\infty}^{\infty} f_{XY}(x, y)\, dx}. \tag{A.43}$$

X and Y are **independent** if and only if

$$f_{XY}(x, y) = f_X(x) f_Y(y) \tag{A.44}$$

for *all* (x, y) in the support of (X, Y). Using equation (A.40) we can see that equation (A.44) is equivalent to

$$f_{X \mid Y}(x \mid y) = f_X(x), \qquad \text{for all } x \text{ and } y. \tag{A.45}$$

If X and Y are discrete random variables, the **joint probability function (joint pf)** of X and Y, also denoted by $f_{XY}(x, y)$, is defined as

$$f_{XY}(x, y) = \Pr(X = x, Y = y). \tag{A.46}$$

The **marginal probability function (marginal pf)** of X and Y are analogously defined as equations (A.38) and (A.39).[4]

We now consider the moments of a bivariate distribution. For exposition, we assume X and Y are continuous. The expected value of $g(X, Y)$, denoted by $\mathrm{E}[g(X, Y)]$, is defined as

$$\mathrm{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y)\, dx\, dy, \tag{A.47}$$

if the integral exists. The covariance of X and Y, denoted by $\mathrm{Cov}(X, Y)$, is defined as $\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$, where μ_X and μ_Y are the means of X and Y, respectively. Thus

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X - \mu_X) \times (Y - \mu_Y) f_{XY}(x, y)\, dx\, dy. \tag{A.48}$$

For convenience, we also use the notation σ_{XY} for $\mathrm{Cov}(X, Y)$. From equation (A.48) we can show that

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mu_X \mu_Y. \tag{A.49}$$

The correlation coefficient of X and Y, denoted by $\rho(X, Y)$, is defined as

$$\rho(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{A.50}$$

where σ_X and σ_Y are the **standard deviation** (i.e. the square root of the variance) of X and Y, respectively.

If two random variables X and Y are independent, and $g(x, y) = g_X(x) g_Y(y)$ for some functions $g_X(\cdot)$ and $g_Y(\cdot)$, then

$$\mathrm{E}\left[g(x, y)\right] = \mathrm{E}\left[g_X(x) g_Y(y)\right] = \mathrm{E}\left[g_X(x)\right] \mathrm{E}\left[g_Y(y)\right], \tag{A.51}$$

where the first two expectations are taken over the joint distribution of X and Y, and the last two expectations are taken over their marginal distributions. For notational simplicity we do not specify the distribution over which the expectation is taken, and let the content of the function of the random variable determine the required expectation.

From equation (A.51), if X and Y are independent, then $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y) = \mu_X \mu_Y$, so that $\sigma_{XY} = 0$ and $\rho(X, Y) = 0$. Thus, independence implies uncorrelatedness. The converse, however, does not stand.

[4] The case when one random variable is discrete and the other is continuous can be defined similarly.

A.7 Mean and variance of sum of random variables

Consider a set of random variables $X_1, X_2, \cdots, X_n$ with means $\mathrm{E}(X_i) = \mu_i$ and variances $\mathrm{Var}(X_i) = \sigma_i^2$, for $i = 1, 2, \cdots, n$. Let the covariance of X_i and X_j be $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, the correlation coefficient of X_i and X_j be $\rho(X_i, X_j) = \rho_{ij}$, and $w_1, w_2, \cdots, w_n$ be a set of constants. Then

$$\mathrm{E}\left(\sum_{i=1}^{n} w_i X_i\right) = \sum_{i=1}^{n} w_i \mu_i, \tag{A.52}$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} w_i X_i\right) = \sum_{i=1}^{n} w_i^2 \sigma_i^2 + \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{n}}_{i \neq j} w_i w_j \sigma_{ij}. \tag{A.53}$$

For $n = 2$, we have

$$\begin{aligned} \mathrm{Var}(w_1 X_1 \pm w_2 X_2) &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 \pm 2 w_1 w_2 \sigma_{12} \\ &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 \pm 2 w_1 w_2 \rho_{12} \sigma_1 \sigma_2. \end{aligned} \tag{A.54}$$

A.8 Moment generating function and probability generating function

The **moment generating function (mgf)** of a random variable X, denoted by $M_X(t)$, is a function of t defined by[5]

$$M_X(t) = \mathrm{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}\, dF_X(x). \tag{A.55}$$

Given the mgf of a random variable X, the moments of X, if they exist, can be obtained by successively differentiating the mgf with respect to t and evaluating the result at $t = 0$. We observe

$$M_X'(t) = \frac{dM_X(t)}{dt} = \frac{d}{dt}\mathrm{E}(e^{tX}) = \mathrm{E}\left[\frac{d}{dt}(e^{tX})\right] = \mathrm{E}(Xe^{tX}). \tag{A.56}$$

Thus

$$M_X'(0) = \mathrm{E}(X) = \mu_1'. \tag{A.57}$$

[5] If the integral in equation (A.55) does not converge, the mgf does not exist. Some random variables do not have a mgf.

Extending the above, we can see that, for any integer r

$$M_X^r(t) = \frac{d^r M_X(t)}{dt^r} = \frac{d^r}{dt^r}\mathrm{E}(e^{tX}) = \mathrm{E}\left[\frac{d^r}{dt^r}(e^{tX})\right] = \mathrm{E}(X^r e^{tX}), \quad \text{(A.58)}$$

so that

$$M_X^r(0) = \mathrm{E}(X^r) = \mu_r'. \quad \text{(A.59)}$$

If $X_1, X_2, \cdots, X_n$ are independently distributed random variables with mgf $M_1(\cdot), M_2(\cdot), \cdots, M_n(\cdot)$, respectively, and $X = X_1 + \cdots + X_n$, then the mgf of X is

$$\begin{aligned} M_X(t) &= \mathrm{E}(e^{tX}) = \mathrm{E}(e^{tX_1 + \cdots + tX_n}) = \mathrm{E}\left(\prod_{i=1}^{n} e^{tX_i}\right) \\ &= \prod_{i=1}^{n} \mathrm{E}(e^{tX_i}) = \prod_{i=1}^{n} M_i(t). \end{aligned} \quad \text{(A.60)}$$

If $X_1, X_2, \cdots, X_n$ are **independently and identically distributed** (**iid**) with mgf $M(t)$, i.e. $M_i(t) = M(t)$ for $i = 1, 2, \cdots, n$, then we have

$$M_{X_1 + \cdots + X_n}(t) = [M(t)]^n. \quad \text{(A.61)}$$

The following are two important properties of a mgf:[6]

(1) If the mgf of a random variable X exists for t in an open interval around the point $t = 0$, then all moments of X exist.
(2) If the mgf of two random variables X_1 and X_2 are identical for t in an open interval around the point $t = 0$, then the distributions of X_1 and X_2 are identical. Also, if two distributions are identical, they must have the same mgf.

Another important tool for statistical distributions is the **probability generating function (pgf)**. The pgf of a nonnegative discrete random variable X, denoted by $P_X(t)$, is defined as

$$P_X(t) = \mathrm{E}(t^X), \quad \text{(A.62)}$$

if the expectation exists. Suppose $\Omega_X = \{0, 1, \cdots\}$, with $\Pr(X = i) = p_i$ for $i = 0, 1, \cdots$, the pgf of X is

$$P_X(t) = \sum_{i=0}^{\infty} t^i p_i. \quad \text{(A.63)}$$

[6] See DeGroot and Schervish (2002, pp. 205–208) for the details.

The rth-order derivative of $P_X(t)$ is

$$P_X^r(t) = \frac{d^r}{dt^r}\left(\sum_{i=0}^{\infty} t^i p_i\right) = \sum_{i=r}^{\infty} i(i-1)\cdots(i-r+1)t^{i-r}p_i. \quad \text{(A.64)}$$

If we evaluate $P_X^r(t)$ at $t = 0$, all terms in the summation vanish except for $i = r$, which is $r!p_r$. Hence, we have

$$P_X^r(0) = r!p_r, \quad \text{(A.65)}$$

so that given the pgf, we can obtain the pf as

$$p_r = \frac{P_X^r(0)}{r!}, \quad \text{(A.66)}$$

which explains the terminology pgf.

A.9 Some discrete distributions

In this section we present some commonly used discrete distributions, namely the binomial, Poisson, geometric, negative binomial, and hypergeometric distributions.

A.9.1 Binomial distribution

Let X be the number of successes in a sequence of n independent Bernoulli trials each with probability of success θ. Then X follows a binomial distribution with parameters n and θ, denoted by $\mathcal{BN}(n, \theta)$, with pf

$$f_X(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad \text{for } x = 0, 1, \cdots, n, \quad \text{(A.67)}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}. \quad \text{(A.68)}$$

The mean and variance of X are

$$\mathrm{E}(X) = n\theta \quad \text{and} \quad \mathrm{Var}(X) = n\theta(1-\theta). \quad \text{(A.69)}$$

The mgf of X is

$$M_X(t) = (\theta e^t + 1 - \theta)^n. \quad \text{(A.70)}$$

When n is large, X is approximately normally distributed.

A.9.2 Poisson distribution

A discrete random variable X is said to have a Poisson distribution with parameter $\lambda(>0)$, denoted by $\mathcal{PN}(\lambda)$, if its pf is

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{for } x = 0, 1, \cdots . \tag{A.71}$$

The mean and variance of X are

$$\mathrm{E}(X) = \mathrm{Var}(X) = \lambda. \tag{A.72}$$

The mgf of X is

$$M_X(t) = \exp\left[\lambda(e^t - 1)\right]. \tag{A.73}$$

When λ is large, X is approximately normally distributed.

A.9.3 Geometric distribution

Suppose independent Bernoulli trials, each with probability of success θ, are performed until a success occurs. Let X be the number of failures prior to the first success. Then X has a geometric distribution with parameter θ, denoted by $\mathcal{GM}(\theta)$, and its pf is

$$f_X(x) = \theta(1-\theta)^x, \qquad \text{for } x = 0, 1, \cdots . \tag{A.74}$$

The mean and variance of X are

$$\mathrm{E}(X) = \frac{1-\theta}{\theta} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{1-\theta}{\theta^2}. \tag{A.75}$$

The mgf of X is

$$M_X(t) = \frac{\theta}{1-(1-\theta)e^t}. \tag{A.76}$$

A.9.4 Negative binomial distribution

Suppose independent Bernoulli trials, each with probability of success θ, are performed until r successes occur. Let X be the number of failures prior to the rth success. Then X has a negative binomial distribution with parameters r and θ, denoted by $\mathcal{NB}(r, \theta)$, and its pf is

$$f_X(x) = \binom{x+r-1}{r-1}\theta^r(1-\theta)^x, \qquad \text{for } x = 0, 1, \cdots . \tag{A.77}$$

The mean and variance of X are

$$\mathrm{E}(X) = \frac{r(1-\theta)}{\theta} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{r(1-\theta)}{\theta^2}. \tag{A.78}$$

These results can be easily obtained by making use of the results in Section A.9.3 and recognizing that X is the sum of r iid geometric random variables with parameter θ. The mgf of X is

$$M_X(t) = \left[\frac{\theta}{1-(1-\theta)e^t}\right]^r. \tag{A.79}$$

A.9.5 Hypergeometric distribution

Consider the probability of getting x blue balls in a random draw of m balls without replacement from an urn consisting of n_1 blue balls and n_2 red balls. The random variable X of the number of blue balls defined by this experiment has the pf

$$f_X(x) = \frac{\binom{n_1}{x}\binom{n_2}{m-x}}{\binom{n_1+n_2}{m}}, \qquad \text{for } x = 0, 1, \cdots, n_1;\ x \le m;\ m-x \le n_2, \tag{A.80}$$

and is said to have a hypergeometric distribution with parameters m, n_1, and n_2, denoted by $\mathcal{HG}(m, n_1, n_2)$. The mean and variance of X are

$$\mathrm{E}(X) = \frac{mn_1}{n_1+n_2} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{mn_1n_2(n_1+n_2-m)}{(n_1+n_2)^2(n_1+n_2-1)}. \tag{A.81}$$

Due to the complexity of the mgf of X, it is not given here.[7]

A.9.6 Summary of some discrete distributions

Table A.1 summarizes the pf, mgf, mean, and variance of the discrete distributions discussed in this section.

A.10 Some continuous distributions

In this section we present some commonly used continuous distributions, namely the normal, lognormal, uniform, exponential, gamma, beta, Pareto, and Weibull distributions.

[7] See Johnson and Kotz (1969, p. 144) for the details.

Table A.1. *Some discrete distributions*

Distribution, parameters, notation and support	pf $f_X(x)$	mgf $M_X(t)$	Mean	Variance
Binomial $\mathcal{BN}(n,\theta)$ $x \in \{0,1,\cdots,n\}$	$\binom{n}{x}\theta^x(1-\theta)^{n-x}$	$(\theta e^t + 1-\theta)^n$	$n\theta$	$n\theta(1-\theta)$
Poisson $\mathcal{PN}(\lambda)$ $x \in \{0,1,\cdots\}$	$\dfrac{\lambda^x e^{-\lambda}}{x!}$	$\exp\left[\lambda(e^t-1)\right]$	λ	λ
Geometric $\mathcal{GM}(\theta)$ $x \in \{0,1,\cdots\}$	$\theta(1-\theta)^x$	$\dfrac{\theta}{1-(1-\theta)e^t}$	$\dfrac{1-\theta}{\theta}$	$\dfrac{1-\theta}{\theta^2}$
Negative binomial $\mathcal{NB}(r,\theta)$ $x \in \{0,1,\cdots\}$	$\binom{x+r-1}{r-1}\theta^r(1-\theta)^x$	$\left[\dfrac{\theta}{1-(1-\theta)e^t}\right]^r$	$\dfrac{r(1-\theta)}{\theta}$	$\dfrac{r(1-\theta)}{\theta^2}$
Hypergeometric $\mathcal{HG}(m,n_1,n_2)$ $x \in \{0,1,\cdots,n_1\}$, $x \le m, m-x \le n_2$, Denote $n = n_1 + n_2$	$\dfrac{\binom{n_1}{x}\binom{n_2}{m-x}}{\binom{n}{m}}$	Not presented	$\dfrac{mn_1}{n}$	$\dfrac{mn_1n_2(n-m)}{n^2(n-1)}$

A.10.1 Normal distribution

Let X be a continuous random variable which can take values on the real line. X is said to follow a normal distribution with mean μ and variance σ^2, denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$, if the pdf of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{A.82}$$

X is said to be a standard normal random variable if it is normally distributed with mean 0 and variance 1. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1). \tag{A.83}$$

The mgf of $X \sim \mathcal{N}(\mu, \sigma^2)$ is

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right). \tag{A.84}$$

If X_1 and X_2 are independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2$, and w_1 and w_2 are constants, then $w_1X_1 + w_2X_2$ is normally distributed with mean $w_1\mu_1 + w_2\mu_2$ and variance $w_1^2\sigma_1^2 + w_2^2\sigma_2^2$.

Linear combinations of normally distributed random variables (not necessarily independent) are normally distributed.

A.10.2 Lognormal distribution

Suppose X is a continuous positive random variable. If $\log X$ follows a normal distribution with mean μ and variance σ^2, then X follows a lognormal distribution with parameters μ and σ^2, denoted by $\mathcal{L}(\mu, \sigma^2)$. The mean and variance of X are

$$\mathrm{E}(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{A.85}$$

$$\mathrm{Var}(X) = \left[\exp\left(2\mu + \sigma^2\right)\right]\left[\exp(\sigma^2) - 1\right]. \tag{A.86}$$

If X_1 and X_2 are independently distributed lognormal random variables with $\log X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2$, and w_1 and w_2 are constants, then $Y = X_1^{w_1} X_2^{w_2}$ is lognormally distributed with parameters $w_1\mu_1 + w_2\mu_2$

and $w_1^2\sigma_1^2 + w_2^2\sigma_2^2$. This result holds as

$$\begin{aligned}\log Y &= \log(X_1^{w_1} X_2^{w_2}) \\ &= w_1 \log X_1 + w_2 \log X_2 \\ &\sim \mathcal{N}(w_1\mu_1 + w_2\mu_2, w_1^2\sigma_1^2 + w_2^2\sigma_2^2). \end{aligned} \tag{A.87}$$

Products of powers of lognormally distributed random variables (not necessarily independent) are lognormally distributed.

The lognormal distribution has the peculiar property that even though the moments of all finite orders exist, the mgf is infinite for any $t > 0$. The pdf of the lognormal distribution with parameters μ and σ^2 is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right]. \tag{A.88}$$

A.10.3 Uniform distribution

A continuous random variable X is uniformly distributed in the interval $[a, b]$, denoted by $\mathcal{U}(a, b)$, if its pdf is

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{for } x \in [a, b], \\ 0, & \text{otherwise.} \end{cases} \tag{A.89}$$

The mean and variance of X are

$$\mathrm{E}(X) = \frac{a+b}{2} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}, \tag{A.90}$$

and its mgf is

$$M_X(t) = \frac{e^{bt} - e^{at}}{(b-a)t}. \tag{A.91}$$

A.10.4 Exponential distribution

A random variable X has an exponential distribution with parameter λ (> 0), denoted by $\mathcal{E}(\lambda)$, if its pdf is

$$f_X(x) = \lambda e^{-\lambda x}, \quad \text{for } x \geq 0. \tag{A.92}$$

The mean and variance of X are

$$\mathrm{E}(X) = \frac{1}{\lambda} \quad \text{and} \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}. \tag{A.93}$$

The mgf of X is

$$M_X(t) = \frac{\lambda}{\lambda - t}. \tag{A.94}$$

A.10.5 Gamma distribution

The integral

$$\int_0^\infty y^{\alpha-1} e^{-y}\, dy \tag{A.95}$$

exists for $\alpha > 0$ and is called the gamma function, denoted by $\Gamma(\alpha)$. Using integration by parts, it can be shown that, for $\alpha > 1$

$$\Gamma(\alpha) = (\alpha - 1) \int_0^\infty y^{\alpha-2} e^{-y}\, dy = (\alpha - 1)\Gamma(\alpha - 1). \tag{A.96}$$

In addition, if α is an integer, we have

$$\Gamma(\alpha) = (\alpha - 1)!. \tag{A.97}$$

X is said to have a gamma distribution with parameters α and β ($\alpha > 0$ and $\beta > 0$), denoted by $\mathcal{G}(\alpha, \beta)$, if its pdf is

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad \text{for } x \geq 0. \tag{A.98}$$

The mean and variance of X are

$$\mathrm{E}(X) = \alpha\beta \quad \text{and} \quad \mathrm{Var}(X) = \alpha\beta^2, \tag{A.99}$$

and its mgf is

$$M_X(t) = \frac{1}{(1 - \beta t)^\alpha}, \quad \text{for } t < \frac{1}{\beta}. \tag{A.100}$$

The special case of $\alpha = r/2$, where r is a positive integer and $\beta = 2$, is called the chi-square distribution with r degrees of freedom, denoted by χ^2_r.

A.10.6 Beta distribution

A random variable X is said to have a beta distribution with parameters α and β ($\alpha > 0, \beta > 0$), denoted by $\mathcal{B}(\alpha, \beta)$, if its pdf is given by

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \quad \text{for } 0 \leq x \leq 1, \tag{A.101}$$

where $B(\alpha, \beta)$ is the beta function defined by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{A.102}$$

The mean and variance of the beta distribution are

$$\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \tag{A.103}$$

and its mgf can be expressed as a confluent hypergeometric function. The details will not be provided here.[8]

A.10.7 Pareto distribution

A random variable X has a Pareto distribution with parameters α and γ ($\alpha > 0, \gamma > 0$), denoted by $\mathcal{P}(\alpha, \gamma)$, if its pdf is

$$f_X(x) = \frac{\alpha\gamma^\alpha}{(x+\gamma)^{\alpha+1}}, \quad \text{for } x \geq 0. \tag{A.104}$$

The df of X is

$$F_X(x) = 1 - \left(\frac{\gamma}{x+\gamma}\right)^\alpha, \quad \text{for } x \geq 0. \tag{A.105}$$

The rth moment of X exists for $r < \alpha$. For $\alpha > 2$, the mean and variance of X are

$$\mathrm{E}(X) = \frac{\gamma}{\alpha-1} \quad \text{and} \quad \mathrm{Var}(X) = \frac{\alpha\gamma^2}{(\alpha-1)^2(\alpha-2)}. \tag{A.106}$$

The Pareto distribution does not have a mgf.

A.10.8 Weibull distribution

A random variable X has a 2-parameter Weibull distribution if its pdf is

$$f_X(x) = \left(\frac{\alpha}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\alpha-1} \exp\left[-\left(\frac{x}{\lambda}\right)^\alpha\right], \quad \text{for } x \geq 0, \tag{A.107}$$

[8] See Johnson and Kotz (1970, p. 40), for the details.

Table A.2. *Some continuous distributions*

Distribution, parameters, notation, and support	pdf $f_X(x)$	mgf $M_X(t)$	Mean	Variance
Normal $\mathcal{N}(\mu, \sigma^2)$ $x \in (-\infty, \infty)$	$\dfrac{\exp\left[-\dfrac{(x-\mu)^2}{2\sigma^2}\right]}{\sqrt{2\pi}\sigma}$	$\exp\left(\mu t + \dfrac{\sigma^2 t^2}{2}\right)$	μ	σ^2
Lognormal $\mathcal{L}(\mu, \sigma^2)$ $x \in (0, \infty)$	$\dfrac{\exp\left[-\dfrac{(\log x-\mu)^2}{2\sigma^2}\right]}{\sqrt{2\pi}\sigma x}$	Does not exist	$e^{\mu+\frac{\sigma^2}{2}}$	$\left(e^{2\mu+\sigma^2}\right)\left(e^{\sigma^2}-1\right)$
Uniform $\mathcal{U}(a, b)$ $x \in [a, b]$	$\dfrac{1}{b-a}$	$\dfrac{e^{bt}-e^{at}}{(b-a)t}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Exponential $\mathcal{E}(\lambda)$ $x \in [0, \infty)$	$\lambda e^{-\lambda x}$	$\dfrac{\lambda}{\lambda - t}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$

Gamma $\mathcal{G}(\alpha, \beta)$ $x \in [0, \infty)$	$\frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$	$\frac{1}{(1-\beta t)^{\alpha}}$	$\alpha\beta$	$\alpha\beta^2$
Beta $\mathcal{B}(\alpha, \beta)$ $x \in [0, 1]$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$	Not presented	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Pareto $\mathcal{P}(\alpha, \gamma)$ $x \in [0, \infty)$	$\frac{\alpha\gamma^{\alpha}}{(x+\gamma)^{\alpha+1}}$	Does not exist	$\frac{\gamma}{\alpha-1}$	$\frac{\alpha\gamma^2}{(\alpha-1)^2(\alpha-2)}$
Weibull $\mathcal{W}(\alpha, \lambda)$ $x \in [0, \infty)$	$\left(\frac{\alpha}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{x}{\lambda}\right)^{\alpha}}$	Not presented	$\mu = \lambda\,\Gamma\left(1+\frac{1}{\alpha}\right)$	$\lambda^2\,\Gamma\left(1+\frac{2}{\alpha}\right) - \mu^2$

where α is the shape parameter and λ is the scale parameter ($\alpha > 0, \lambda > 0$). We denote the distribution by $\mathcal{W}(\alpha, \lambda)$. The mean and variance of X are

$$\mathrm{E}(X) = \mu = \lambda\,\Gamma\left(1 + \frac{1}{\alpha}\right) \quad \text{and} \quad \mathrm{Var}(X) = \lambda^2\,\Gamma\left(1 + \frac{2}{\alpha}\right) - \mu^2. \tag{A.108}$$

Due to its complexity, the mgf of the Weibull distribution is not presented here.

A.10.9 Summary of some continuous distributions

Table A.2 summarizes the results of the pdf, mgf, mean, and variance of the continuous distributions presented in this section.

A.11 Conditional expectation, conditional mean, and conditional variance

Given two random variables X and Y, and a function $g(x, y)$, the expectation $\mathrm{E}[g(X, Y)]$ defined in equation (A.47) can be evaluated by iterative expectations. First, we define the conditional expectation of $g(X, Y)$ given $Y = y$, denoted by $\mathrm{E}[g(X, Y) \mid y]$, as[9]

$$\mathrm{E}[g(X, Y) \mid y] = \int_{-\infty}^{\infty} g(x, y) f_{X \mid Y}(x \mid y)\, dx. \tag{A.109}$$

Note that $\mathrm{E}[g(X, Y) \mid y]$ is a function of y (but not x). If we allow y to vary over the support of Y, we treat $\mathrm{E}[g(X, Y) \mid y]$ as a function of the random variable Y, and denote it by $\mathrm{E}[g(X, Y) \mid Y]$. Taking the expectation of this function over the random variable Y, we have[10]

$$\begin{aligned}
\mathrm{E}\{\mathrm{E}[g(X, Y) \mid Y]\} &= \int_{-\infty}^{\infty} \mathrm{E}[g(X, Y) \mid y] f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} g(x, y) f_{X \mid Y}(x \mid y)\, dx\right] f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X \mid Y}(x \mid y) f_Y(y)\, dx\, dy
\end{aligned}$$

[9] An alternative notation for the conditional expectation is $\mathrm{E}_{X \mid Y}[g(X, Y) \mid y]$, in which the suffix of E explicitly denotes that it is a conditional expectation. We adopt the simpler notation where the suffix is dropped. Also, for simplicity of exposition, we assume X and Y are continuous. The results in this section apply to discrete and mixed random variables as well.

[10] The first expectation on the left-hand side of equation (A.110) is taken over Y and the second expectation is taken over X conditional on $Y = y$.

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y)\, dx\, dy$$
$$= \mathrm{E}[g(X, Y)]. \tag{A.110}$$

Thus, unconditional expectation can be calculated using iterative expectations. This result implies

$$\mathrm{E}[\mathrm{E}(X \mid Y)] = \mathrm{E}(X). \tag{A.111}$$

The conditional variance of X given $Y = y$ is a function of the random variable Y if y is allowed to vary over the support of Y. We denote this conditional variance by $\mathrm{Var}(X \mid Y)$, which is defined as $v(Y)$, where

$$v(y) = \mathrm{Var}(X \mid y) = \mathrm{E}\{[X - \mathrm{E}(X \mid y)]^2 \mid y\} = \mathrm{E}(X^2 \mid y) - [\mathrm{E}(X \mid y)]^2. \tag{A.112}$$

Thus, we have

$$\mathrm{Var}(X \mid Y) = \mathrm{E}(X^2 \mid Y) - [\mathrm{E}(X \mid Y)]^2, \tag{A.113}$$

which implies

$$\mathrm{E}(X^2 \mid Y) = \mathrm{Var}(X \mid Y) + [\mathrm{E}(X \mid Y)]^2. \tag{A.114}$$

Now from equations (A.26) and (A.114), we have

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \\
&= \mathrm{E}[\mathrm{E}(X^2) \mid Y] - [\mathrm{E}(X)]^2 \\
&= \mathrm{E}\{\mathrm{Var}(X \mid Y) + [\mathrm{E}(X \mid Y)]^2\} - [\mathrm{E}(X)]^2 \\
&= \mathrm{E}[\mathrm{Var}(X \mid Y)] + \mathrm{E}\{[\mathrm{E}(X \mid Y)]^2\} - [\mathrm{E}(X)]^2 \\
&= \mathrm{E}[\mathrm{Var}(X \mid Y)] + \mathrm{E}\{[\mathrm{E}(X \mid Y)]^2\} - \{\mathrm{E}[\mathrm{E}(X \mid Y)]\}^2 \\
&= \mathrm{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}[\mathrm{E}(X \mid Y)].
\end{aligned} \tag{A.115}$$

Verbally, the above equation says that the unconditional variance of X is equal to the mean of its conditional (upon Y) variance plus the variance of its conditional (upon Y) mean.

If X and Y are independent, we conclude from equation (A.51) that

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y). \tag{A.116}$$

To compute Var(XY), we use equation (A.115) and apply conditioning of XY on Y to obtain

$$\begin{aligned}\text{Var}(XY) &= \text{E}[\text{Var}(XY \mid Y)] + \text{Var}[\text{E}(XY \mid Y)] \\ &= \text{E}[Y^2\text{Var}(X \mid Y)] + \text{Var}[Y\text{E}(X \mid Y)]. \end{aligned} \tag{A.117}$$

Since X and Y are independent, $\text{Var}(X \mid Y) = \text{Var}(X)$ and $\text{E}(X \mid Y) = \text{E}(X)$. Thus, we have

$$\begin{aligned}\text{Var}(XY) &= \text{E}[Y^2\text{Var}(X)] + \text{Var}[Y\text{E}(X)] \\ &= \text{E}(Y^2)\text{Var}(X) + [\text{E}(X)]^2\text{Var}(Y). \end{aligned} \tag{A.118}$$

Note that if we use equation (A.115) and apply conditioning of XY on X, we obtain

$$\text{Var}(XY) = [\text{E}(Y)]^2\text{Var}(X) + \text{E}(X^2)\text{Var}(Y). \tag{A.119}$$

The equivalence of equations (A.118) and (A.119) can be proved using equation (A.26).

A.12 Compound distribution

Let N be a discrete random variable that takes nonnegative integer values. We consider a sequence of iid random variables $\{X_1, X_2, \cdots\}$, where $X_i \sim X$ for $i \in \{1, 2, \cdots\}$. Let

$$S = X_1 + \cdots + X_N, \tag{A.120}$$

which is the sum of N iid random variables, each distributed as X, with the number of summation terms N being random.[11] S is said to have a **compound distribution**, with N being the **primary distribution** and X the **secondary distribution**. We denote $\text{E}(N) = \mu_N$ and $\text{Var}(N) = \sigma_N^2$, and likewise $\text{E}(X) = \mu_X$ and $\text{Var}(X) = \sigma_X^2$. Using the results on conditional expectations, we have

$$\text{E}(S) = \text{E}[\text{E}(S \mid N)] = \text{E}[\text{E}(X_1 + \cdots + X_N \mid N)] = \text{E}(N\mu_X) = \mu_N\mu_X. \tag{A.121}$$

Also, using equation (A.115), we have

$$\begin{aligned}\text{Var}(S) &= \text{E}[\text{Var}(S \mid N)] + \text{Var}[\text{E}(S \mid N)] \\ &= \text{E}[N\sigma_X^2] + \text{Var}(N\mu_X) \\ &= \mu_N\sigma_X^2 + \sigma_N^2\mu_X^2. \end{aligned} \tag{A.122}$$

[11] S is defined as 0 if $N = 0$.

If N has a Poisson distribution with mean λ, so that $\mu_N = \sigma_N^2 = \lambda$, S is said to have a **compound Poisson distribution**, with

$$\text{Var}(S) = \lambda(\sigma_X^2 + \mu_X^2). \tag{A.123}$$

A.13 Convolution

Suppose X_1 and X_2 are independent discrete random variables with common support Ω, and pf $f_{X_1}(\cdot)$ and $f_{X_2}(\cdot)$, respectively, and let $X = X_1 + X_2$. The pf $f_X(\cdot)$ of X is given by

$$f_X(x) = \sum_{x_2,\, x-x_2 \in \Omega} f_{X_1}(x - x_2) f_{X_2}(x_2) = \sum_{x_1,\, x-x_1 \in \Omega} f_{X_2}(x - x_1) f_{X_1}(x_1). \tag{A.124}$$

The pf $f_X(x)$ evaluated by either expression in equation (A.124) is called the **convolution** of the pf $f_{X_1}(\cdot)$ and $f_{X_2}(\cdot)$, which may also be written as

$$f_X(x) = (f_{X_2} * f_{X_1})(x) = (f_{X_1} * f_{X_2})(x). \tag{A.125}$$

Hence, convolutions are *commutative*. If X_1 and X_2 are nonnegative, then equation (A.124) becomes

$$f_X(x) = \sum_{x_2=0}^{x} f_{X_1}(x - x_2) f_{X_2}(x_2) = \sum_{x_1=0}^{x} f_{X_2}(x - x_1) f_{X_1}(x_1), \quad \text{for } x = 0, 1, \cdots. \tag{A.126}$$

If X_1 and X_2 are continuous, equations (A.124) and (A.126) are replaced, respectively, by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X_1}(x - x_2) f_{X_2}(x_2)\, dx_2 = \int_{-\infty}^{\infty} f_{X_2}(x - x_1) f_{X_1}(x_1)\, dx_1, \tag{A.127}$$

and

$$f_X(x) = \int_0^x f_{X_1}(x - x_2) f_{X_2}(x_2)\, dx_2 = \int_0^x f_{X_2}(x - x_1) f_{X_1}(x_1)\, dx_1. \tag{A.128}$$

Convolutions may be applied to sums of more than two random variables. Thus, if X_1, X_2, and X_3 are independently distributed, we have

$$f_{X_1+X_2+X_3}(x) = (f_{X_1} * f_{X_2} * f_{X_3})(x). \tag{A.129}$$

Furthermore, if X_1, X_2, and X_3 are identically distributed with the same pf or pdf $f(\cdot)$, we write equation (A.129) as

$$f_{X_1+X_2+X_3}(x) = (f * f * f)(x) = f^{*3}(x). \tag{A.130}$$

Thus, for a sum of n iid random variables each with pf or pdf $f(\cdot)$, its pf or pdf is $f^{*n}(\cdot)$.

A.14 Mixture distribution

Let $X_1, \cdots, X_n$ be random variables with corresponding pf (if X_i are discrete) or pdf (if X_i are continuous) $f_{X_1}(\cdot), \cdots, f_{X_n}(\cdot)$. X_i are assumed to have the common support Ω. A new random variable X may be created with pf or pdf $f_X(\cdot)$ given by

$$f_X(x) = p_1 f_{X_1}(x) + \cdots + p_n f_{X_n}(x), \qquad x \in \Omega, \tag{A.131}$$

where $p_i \geq 0$ for $i = 1, \cdots, n$ and $\sum_{i=1}^{n} p_i = 1$, so that $\{p_i\}$ form a well-defined probability distribution. We call X a **mixture distribution** with a **discrete mixing distribution**.

Now consider a nonnegative random variable (discrete or continuous) with pf or pdf $f(x \mid \theta)$, which depends on the parameter θ. Let $h(\cdot)$ be a function such that $h(\theta) > 0$ for $\theta > 0$, and

$$\int_0^\infty h(\theta)\, d\theta = 1. \tag{A.132}$$

A new random variable X may be created with pf or pdf $f_X(x)$ given by

$$f_X(x) = \int_0^\infty f(x \mid \theta) h(\theta)\, d\theta. \tag{A.133}$$

Note that $h(\theta)$ is a well-defined pdf that defines the **continuous mixing distribution**, and X is a mixture distribution with continuous mixing. Furthermore, if we allow $h(\theta)$ to depend on a parameter γ, we may re-write equation (A.133) as

$$f_X(x \mid \gamma) = \int_0^\infty f(x \mid \theta) h(\theta \mid \gamma)\, d\theta. \tag{A.134}$$

Thus, the mixture distribution depends on the parameter γ, which determines the distribution of θ.

A.15 Bayesian approach of statistical inference

Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ denote a sample of iid random variables each distributed as X, and $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ be a realization of the sample. Classical statistical inference assumes that X depends on an unknown *fixed* parameter θ (which may be multidimensional). After $\boldsymbol{X}$ is observed, **statistical inference**, including **estimation** and **hypothesis testing**, concerning the parameter θ is made.

The Bayesian approach of statistical inference assumes that the parameter θ determining the distribution of X is unknown and uncertain. Thus, θ is treated as a random variable, denoted by Θ with support Ω_Θ and pdf $f_\Theta(\theta)$, which is called the **prior distribution** of Θ. Once the data $\boldsymbol{X}$ are observed, the distribution of Θ is revised and the resultant pdf of Θ is called the **posterior distribution** of Θ, with pdf denoted by $f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})$. Denoting $f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x})$ as the joint pdf of Θ and $\boldsymbol{X}$, and using the result in equation (A.40), we have

$$\begin{aligned} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) &= \frac{f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\ &= \frac{f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta)}{\int_{\theta \in \Omega_\Theta} f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta)\, d\theta}. \end{aligned} \tag{A.135}$$

The conditional pdf of $\boldsymbol{X}$, $f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta)$, is called the **likelihood function**. Multiplying the likelihood with the prior pdf of Θ gives the joint pdf of $\boldsymbol{X}$ and Θ. In classical statistical inference, only the likelihood function matters; the prior pdf does not have a role to play.

Note that the denominator of equation (A.135) is a function of $\boldsymbol{x}$ but not θ. Denoting

$$K(\boldsymbol{x}) = \frac{1}{\int_{\theta \in \Omega_\Theta} f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta)\, d\theta}, \tag{A.136}$$

we can rewrite the posterior pdf of Θ as

$$\begin{aligned} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) &= K(\boldsymbol{x}) f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta) \\ &\propto f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta). \end{aligned} \tag{A.137}$$

$K(\boldsymbol{x})$ is a **constant of proportionality** and is free of θ. It scales the posterior pdf so that it integrates to 1.

The Bayesian approach of estimating θ involves minimizing a **loss function** over the posterior distribution of Θ. The loss function measures the penalty in making a wrong decision with respect to the true value of θ. Thus, if the estimate of θ is $\hat{\theta}$, the loss function is denoted by $L(\theta, \hat{\theta})$. A

popular loss function is the **squared-error loss function** defined by

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2. \tag{A.138}$$

The squared-error loss function is symmetric, as an over-estimation and under-estimation of the same amount incur the same loss. It can be shown that the value of $\hat{\theta}$ that minimizes $(\theta - \hat{\theta})^2$ over the posterior distribution is the **posterior mean**, which is given by

$$\hat{\theta} = \mathrm{E}(\Theta \mid \boldsymbol{x}) = \int_{\theta \in \Omega_\Theta} \theta f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x})\, d\theta. \tag{A.139}$$

This is also called the **Bayes estimate** of θ (with respect to the squared-error loss function).

A.16 Conjugate distribution

A difficult step in applying the Bayesian approach of statistical inference is the computation of the posterior distribution. As equation (A.135) shows, the posterior pdf is in general difficult to evaluate, as it is the ratio of two terms where the denominator involves an integral. The evaluation of the Bayes estimate is difficult if the posterior cannot be easily computed. It turns out, however, that there are classes of prior pdf which are **conjugate** to some particular likelihood functions, in the sense that the resulting posterior pdf belongs to the same class of pdf as the prior. Thus, for conjugate priors, the observed data $\boldsymbol{X}$ do not change the class of the prior, they only change the *parameters* of the prior.

The formal definition of **conjugate prior distribution** is as follows. Let the prior pdf of Θ be $f_\Theta(\theta \mid \gamma)$, where γ is the parameter of the prior pdf, called the **hyperparameter**. The prior pdf $f_\Theta(\theta \mid \gamma)$ is conjugate to the likelihood function $f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta)$ if the posterior pdf is equal to $f_\Theta(\theta \mid \gamma^*)$, which has the same functional form as the prior pdf but, generally, a different hyperparameter. In other words, the prior and posterior distributions belong to the same family of distributions.

We now present some conjugate distributions that are commonly used in Bayesian inference.[12]

[12] We adopt the convention of "prior–likelihood" to describe the conjugate distribution. Thus, the beta–Bernoulli conjugate distribution has a beta prior and a Bernoulli likelihood.

A.16.1 The beta–Bernoulli conjugate distribution

Let X be the Bernoulli random variable which takes value 1 with probability θ and 0 with probability $1-\theta$. Thus, the likelihood of X is

$$f_{X\,|\,\Theta}(x\,|\,\theta) = \theta^x(1-\theta)^{1-x}, \quad \text{for } x = 0, 1. \tag{A.140}$$

Θ is assumed to follow the beta distribution with pdf given in equation (A.101), where the hyperparameters are α and β, i.e.

$$f_\Theta(\theta;\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \quad \text{for } \theta \in (0,1). \tag{A.141}$$

Thus, the joint pdf of Θ and X is

$$f_{\Theta X}(\theta,x) = f_{X\,|\,\Theta}(x\,|\,\theta) f_\Theta(\theta;\alpha,\beta) = \frac{\theta^{\alpha+x-1}(1-\theta)^{(\beta-x+1)-1}}{B(\alpha,\beta)}, \tag{A.142}$$

from which we obtain the marginal pdf of X as

$$\begin{aligned} f_X(x) &= \int_0^1 \frac{\theta^{\alpha+x-1}(1-\theta)^{(\beta-x+1)-1}}{B(\alpha,\beta)}\, d\theta \\ &= \frac{B(\alpha+x,\beta-x+1)}{B(\alpha,\beta)}. \end{aligned} \tag{A.143}$$

Substituting equations (A.142) and (A.143) into (A.135), we obtain

$$f_{\Theta\,|\,X}(\theta\,|\,x) = \frac{\theta^{\alpha+x-1}(1-\theta)^{(\beta-x+1)-1}}{B(\alpha+x,\beta-x+1)}, \tag{A.144}$$

which is a beta pdf with parameters $\alpha+x$ and $\beta-x+1$. Hence, the posterior and prior distributions belong to the same family, and the beta distribution is said to be conjugate to the Bernoulli distribution.

Consider now n observations of X denoted by $\boldsymbol{X} = \{X_1, \cdots, X_n\}$. Repeating the derivation above, we obtain

$$\begin{aligned} f_{\boldsymbol{X}\,|\,\Theta}(\boldsymbol{x}\,|\,\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}, \end{aligned} \tag{A.145}$$

and

$$\begin{aligned} f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x}) &= f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_{\Theta}(\theta; \alpha, \beta) \\ &= \left[\theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}\right]\left[\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\right] \\ &= \frac{\theta^{(\alpha+n\bar{x})-1}(1-\theta)^{(\beta+n-n\bar{x})-1}}{B(\alpha,\beta)}. \end{aligned} \tag{A.146}$$

As

$$\begin{aligned} \int_0^1 f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x})\, d\theta &= \int_0^1 \frac{\theta^{(\alpha+n\bar{x})-1}(1-\theta)^{(\beta+n-n\bar{x})-1}}{B(\alpha,\beta)}\, d\theta \\ &= \frac{B(\alpha+n\bar{x}, \beta+n-n\bar{x})}{B(\alpha,\beta)}, \end{aligned} \tag{A.147}$$

we conclude that

$$\begin{aligned} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) &= \frac{f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x})}{\int_0^1 f_{\Theta \boldsymbol{X}}(\theta, \boldsymbol{x})\, d\theta} \\ &= \frac{\theta^{(\alpha+n\bar{x})-1}(1-\theta)^{(\beta+n-n\bar{x})-1}}{B(\alpha+n\bar{x}, \beta+n-n\bar{x})}, \end{aligned} \tag{A.148}$$

and the posterior distribution of Θ follows a beta distribution with parameters $\alpha + n\bar{x}$ and $\beta + n - n\bar{x}$.

In the above computation we derive the posterior pdf and show that it has the same functional form as the prior pdf, apart from the differences in the parameters. However, following equation (A.137) we could have concluded that

$$\begin{aligned} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) &\propto f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_{\Theta}(\theta) \\ &\propto \left[\theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}\right]\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] \\ &\propto \theta^{(\alpha+n\bar{x})-1}(1-\theta)^{(\beta+n-n\bar{x})-1}, \end{aligned} \tag{A.149}$$

so that the posterior distribution belongs to the same class of distribution as the prior. This is done without having to compute the expression for the constant of proportionality $K(\boldsymbol{x})$. We shall adopt this simpler approach in subsequent discussions.

A.16.2 The beta–binomial conjugate distribution

Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ be a sample of binomial random variables with parameters m_i and θ, such that $X_i \sim \mathcal{BN}(m_i, \theta)$ independently. Thus, the

likelihood of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}\mid\Theta}(\boldsymbol{x}\mid\theta) = \prod_{i=1}^{n}\binom{m_i}{x_i}\theta^{x_i}(1-\theta)^{m_i-x_i}$$
$$= \left[\prod_{i=1}^{n}\binom{m_i}{x_i}\right]\left[\theta^{n\bar{x}}(1-\theta)^{\sum_{i=1}^{n}(m_i-x_i)}\right]. \tag{A.150}$$

If Θ follows the beta distribution with hyperparameters α and β, and we define $m = \sum_{i=1}^{n} m_i$, the posterior pdf of Θ satisfies

$$f_{\Theta\mid\boldsymbol{X}}(\theta\mid\boldsymbol{x}) \propto f_{\boldsymbol{X}\mid\Theta}(\boldsymbol{x}\mid\theta)f_{\Theta}(\theta;\alpha,\beta)$$
$$\propto \left[\theta^{n\bar{x}}(1-\theta)^{m-n\bar{x}}\right]\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right]$$
$$\propto \theta^{(\alpha+n\bar{x})-1}(1-\theta)^{(\beta+m-n\bar{x})-1}. \tag{A.151}$$

Comparing this against equation (A.141), we conclude that the posterior distribution of Θ is beta with parameters $\alpha + n\bar{x}$ and $\beta + m - n\bar{x}$. Thus, the beta prior distribution is conjugate to the binomial likelihood.

A.16.3 The gamma–Poisson conjugate distribution

Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ be iid Poisson random variables with parameter λ. The random variable Λ of the parameter λ is assumed to follow a gamma distribution with hyperparameters α and β, i.e. the prior pdf of Λ is

$$f_{\Lambda}(\lambda;\alpha,\beta) = \frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}, \tag{A.152}$$

and the likelihood of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}\mid\Lambda}(\boldsymbol{x}\mid\lambda) = \prod_{i=1}^{n}\frac{\lambda^{x_i}e^{-\lambda}}{x_i!}$$
$$= \frac{\lambda^{n\bar{x}}e^{-n\lambda}}{\prod_{i=1}^{n}x_i!}. \tag{A.153}$$

Thus, the posterior pdf of Λ satisfies

$$f_{\Lambda\mid\boldsymbol{X}}(\lambda\mid\boldsymbol{x}) \propto f_{\boldsymbol{X}\mid\Lambda}(\boldsymbol{x}\mid\lambda)f_{\Lambda}(\lambda;\alpha,\beta)$$
$$\propto \lambda^{\alpha+n\bar{x}-1}e^{-\lambda\left(n+\frac{1}{\beta}\right)}. \tag{A.154}$$

Comparing equations (A.154) and (A.152), we conclude that the posterior pdf of Λ is $f_\Lambda(\lambda; \alpha^*, \beta^*)$, where

$$\alpha^* = \alpha + n\bar{x} \tag{A.155}$$

and

$$\beta^* = \left[n + \frac{1}{\beta}\right]^{-1} = \frac{\beta}{n\beta + 1}. \tag{A.156}$$

Hence, the gamma prior pdf is conjugate to the Poisson likelihood.

A.16.4 The beta–geometric conjugate distribution

Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ be iid geometric random variables with parameter θ so that the likelihood of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i} = \theta^n (1-\theta)^{n\bar{x}}. \tag{A.157}$$

If the prior distribution of Θ is beta with hyperparameters α and β, then the posterior pdf of Θ satisfies

$$\begin{aligned} f_{\Theta \mid \boldsymbol{X}}(\theta \mid \boldsymbol{x}) &\propto f_{\boldsymbol{X} \mid \Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta; \alpha, \beta) \\ &\propto \theta^{\alpha+n-1}(1-\theta)^{\beta+n\bar{x}-1}. \end{aligned} \tag{A.158}$$

Thus, comparing equations (A.158) and (A.141), we conclude that the posterior distribution of Θ is beta with parameters

$$\alpha^* = \alpha + n \tag{A.159}$$

and

$$\beta^* = \beta + n\bar{x}, \tag{A.160}$$

so that the beta prior is conjugate to the geometric likelihood.

A.16.5 The gamma–exponential conjugate distribution

Let $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ be iid exponential random variables with parameter λ so that the likelihood of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X} \mid \Lambda}(\boldsymbol{x} \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda n\bar{x}}. \tag{A.161}$$

Table A.3. *Some conjugate distributions*

Prior pdf and hyperparameters	Likelihood of $\boldsymbol{X}$	Hyperparameters of posterior pdf
$\mathcal{B}(\alpha, \beta)$	Bernoulli	$\alpha + n\bar{x},\ \beta + n - n\bar{x}$
$\mathcal{B}(\alpha, \beta)$	$\mathcal{BN}(m_i, \theta)$	$\alpha + n\bar{x},\ \beta + \sum_{i=1}^{n}(m_i - x_i)$
$\mathcal{G}(\alpha, \beta)$	$\mathcal{PN}(\lambda)$	$\alpha + n\bar{x},\ \dfrac{\beta}{n\beta + 1}$
$\mathcal{B}(\alpha, \beta)$	$\mathcal{GM}(\theta)$	$\alpha + n,\ \beta + n\bar{x}$
$\mathcal{G}(\alpha, \beta)$	$\mathcal{E}(\lambda)$	$\alpha + n,\ \dfrac{\beta}{1 + \beta n\bar{x}}$

If the prior distribution of Λ is gamma with hyperparameters α and β, then the posterior pdf of Λ satisfies

$$\begin{aligned} f_{\Lambda \mid \boldsymbol{X}}(\lambda \mid \boldsymbol{x}) &\propto f_{\boldsymbol{X} \mid \Lambda}(\boldsymbol{x} \mid \lambda) f_{\Lambda}(\lambda; \alpha, \beta) \\ &\propto \lambda^{\alpha + n - 1} e^{-\lambda\left(\frac{1}{\beta} + n\bar{x}\right)}. \end{aligned} \tag{A.162}$$

Comparing equations (A.162) and (A.152), we conclude that the posterior distribution of Λ is gamma with parameters

$$\alpha^* = \alpha + n \tag{A.163}$$

and

$$\beta^* = \left[\frac{1}{\beta} + n\bar{x}\right]^{-1} = \frac{\beta}{1 + \beta n\bar{x}}. \tag{A.164}$$

Thus, the gamma prior is conjugate to the exponential likelihood.

A.16.6 Summary of conjugate distributions

Table A.3 summarizes the conjugate distributions discussed in this section.

A.17 Least squares estimation

Consider a regression model with n observations, where the $n \times 1$ vector of the observations of the dependent variable is denoted by $\boldsymbol{y} = (y_1, \cdots, y_n)'$

and the $n \times (k+1)$ matrix of the observations of the regressors is denoted by

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdot & \cdot & X_{1k} \\ 1 & X_{21} & X_{22} & \cdot & \cdot & X_{2k} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_{n1} & X_{n2} & \cdot & \cdot & X_{nk} \end{bmatrix}. \tag{A.165}$$

Thus, the first column of $\mathbf{X}$ is the vector $\mathbf{1} = (1, \cdots, 1)'$, such that the regression has a constant term, and X_{ij} is the ith observation of the jth explanatory variable. We write the regression model in matrix form as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{A.166}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_k)'$ is the vector of regression coefficients and $\boldsymbol{\varepsilon}$ is the vector of residuals. The regression coefficient $\boldsymbol{\beta}$ can be estimated using the **least squares method**, which is obtained by minimizing the residual sum of squares (RSS), defined as

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2 \\ &= (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{X}'\boldsymbol{y}, \end{aligned} \tag{A.167}$$

with respect to $\boldsymbol{\beta}$. The value of $\boldsymbol{\beta}$ which minimizes RSS is called the **least squares estimate** of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, and is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}. \tag{A.168}$$

$\mathbf{X}'\mathbf{X}$ is the $(k+1) \times (k+1)$ **raw sum-of-cross-products** matrix of the regressors, i.e.

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{h=1}^{n} X_{h1} & \sum_{h=1}^{n} X_{h2} & \cdots & \sum_{h=1}^{n} X_{hk} \\ \sum_{h=1}^{n} X_{h1} & \sum_{h=1}^{n} X_{h1}^2 & \sum_{h=1}^{n} X_{h1}X_{h2} & \cdots & \sum_{h=1}^{n} X_{h1}X_{hk} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{h=1}^{n} X_{hk} & \sum_{h=1}^{n} X_{hk}X_{h1} & \sum_{h=1}^{n} X_{hk}X_{h2} & \cdots & \sum_{h=1}^{n} X_{hk}^2 \end{bmatrix}. \tag{A.169}$$

$\mathbf{X}'\boldsymbol{y}$ is the $(k+1)\times 1$ raw sum-of-cross-products vector of the regressor and the dependent variable, i.e.,

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{h=1}^{n} y_h \\ \sum_{h=1}^{n} X_{h1}y_h \\ \sum_{h=1}^{n} X_{h2}y_h \\ \vdots \\ \sum_{h=1}^{n} X_{hk}y_h \end{bmatrix}. \tag{A.170}$$

We now denote the sum-of-cross-products matrix of the **deviation-from-mean** of the k regressors by $\mathbf{M}_{\mathbf{X}'\mathbf{X}}$, so that the (i,j)th element of the $k\times k$ matrix $\mathbf{M}_{\mathbf{X}'\mathbf{X}}$ is, for $i,j = 1,\cdots,k$

$$\sum_{h=1}^{n}(X_{hi}-\bar{X}_i)(X_{hj}-\bar{X}_j), \tag{A.171}$$

where

$$\bar{X}_i = \frac{1}{n}\sum_{h=1}^{n} X_{hi}. \tag{A.172}$$

Likewise, we define the $k\times 1$ vector $\mathbf{M}_{\mathbf{X}'\boldsymbol{y}}$ with the ith element given by

$$\sum_{h=1}^{n}(X_{hi}-\bar{X}_i)(y_h-\bar{y}), \tag{A.173}$$

where $\bar{y}$ is the sample mean of $\boldsymbol{y}$.

Then the least squares estimate of the slope coefficients of the regression model is[13]

$$\begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \mathbf{M}_{\mathbf{X}'\mathbf{X}}^{-1}\mathbf{M}_{\mathbf{X}'\boldsymbol{y}} \tag{A.174}$$

[13] Readers may refer to Johnston and DiNardo (1997, Section 3.1.3) for a proof of this result.

and the least squares estimate of the constant term is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}_1 - \cdots - \hat{\beta}_k \bar{X}_k. \tag{A.175}$$

A.18 Fisher information and Cramér–Rao inequality

Let X be a random variable with pdf or pf $f(x;\theta)$, where θ is a parameter. To economize on notations we drop the suffix X in the pdf or pf. For simplicity of exposition, we assume X is continuous, although the results below also hold for discrete distributions. We assume that the differentiation of a definite integral can be executed inside the integral. This assumption, together with others (such as the existence of derivatives of the pdf), are collectively known as the **regularity conditions**, which, although not elaborated here, will be assumed to hold.

As

$$\int_{-\infty}^{\infty} f(x;\theta)\, dx = 1, \tag{A.176}$$

differentiating the equation on both sides, and assuming that we can move the differentiation operation inside the integral, we have

$$\frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(x;\theta)\, dx = \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial \theta}\, dx = 0, \tag{A.177}$$

which can also be written as

$$\begin{aligned} \int_{-\infty}^{\infty} \left[\frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta}\right] f(x;\theta)\, dx &= \int_{-\infty}^{\infty} \left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right] f(x;\theta)\, dx \\ &= \mathrm{E}\left[\frac{\partial \log f(X;\theta)}{\partial \theta}\right] \\ &= 0. \end{aligned} \tag{A.178}$$

If we differentiate the above equation again, we obtain

$$\int_{-\infty}^{\infty} \left[\left(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right) f(x;\theta) + \frac{\partial \log f(x;\theta)}{\partial \theta} \frac{\partial f(x;\theta)}{\partial \theta}\right] dx = 0. \tag{A.179}$$

Now the second part of the integral above can be expressed as

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\partial \log f(x;\theta)}{\partial \theta} \frac{\partial f(x;\theta)}{\partial \theta}\, dx &= \int_{-\infty}^{\infty} \frac{\partial \log f(x;\theta)}{\partial \theta} \left[\frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)\right] dx \\ &= \int_{-\infty}^{\infty} \left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right]^2 f(x;\theta)\, dx \\ &= \mathrm{E}\left[\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right] \\ &\equiv I(\theta) \\ &> 0, \end{aligned} \tag{A.180}$$

which is called the **Fisher information** in an observation. From equations (A.179) and (A.180), we conclude that

$$I(\theta) = -\int_{-\infty}^{\infty} \left(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right) f(x;\theta)\; dx = \mathrm{E}\left[-\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right]. \tag{A.181}$$

Suppose we have a random sample of observations $\boldsymbol{x} = (x_1, \cdots, x_n)$. We define the **likelihood function** of the sample as

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta), \tag{A.182}$$

which is taken as a function of θ given $\boldsymbol{x}$. Then

$$\log L(\theta; \boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i;\theta). \tag{A.183}$$

Analogous to equation (A.180), we define the Fisher information in the random sample as

$$I_n(\theta) = \mathrm{E}\left[\left(\frac{\partial \log L(\theta; \boldsymbol{X})}{\partial \theta}\right)^2\right], \tag{A.184}$$

so that from equation (A.183), we have

$$I_n(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f(X_i;\theta)\right)^2\right]$$
$$= \mathrm{E}\left[\left(\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta}\right)^2\right]. \tag{A.185}$$

As the observation x_i are pairwise independent, the expectations of the cross product terms above are zero. Thus, we conclude

$$I_n(\theta) = \sum_{i=1}^{n}\mathrm{E}\left[\left(\frac{\partial \log f(X_i;\theta)}{\partial\theta}\right)^2\right] = nI(\theta). \tag{A.186}$$

Let $u(\boldsymbol{x})$ be a statistic of the sample, such that $\mathrm{E}[u(\boldsymbol{x})] = k(\theta)$, i.e. $u(\boldsymbol{x})$ is an unbiased estimator of $k(\theta)$. Then we have

$$\mathrm{Var}\,[u(\boldsymbol{x})] \geq \frac{\left[k'(\theta)\right]^2}{I_n(\theta)} = \frac{\left[k'(\theta)\right]^2}{nI(\theta)}, \tag{A.187}$$

which is called the **Cramér–Rao** inequality.[14] In the special case $k(\theta) = \theta$, we denote $u(\boldsymbol{x}) = \hat{\theta}$ and conclude

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}, \tag{A.188}$$

for any unbiased estimator of θ. An unbiased estimator is said to be **efficient** if it attains the **Cramér–Rao** lower bound, i.e. the right-hand side of equation (A.188).

In the case where θ is a k-element vector, the above results can be generalized as follows. First, $\partial \log[f(x;\theta)]/\partial\theta$ is a $k \times 1$ vector. Equation (A.178) applies, with the result being a $k \times 1$ vector. Second, the **Fisher information matrix** in an observation is now defined as the $k \times k$ matrix

$$I(\theta) = \mathrm{E}\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\frac{\partial \log f(X;\theta)}{\partial\theta'}\right], \tag{A.189}$$

and the result in equation (A.181) is replaced by

$$\mathrm{E}\left[-\frac{\partial^2 \log f(X;\theta)}{\partial\theta\,\partial\theta'}\right], \tag{A.190}$$

[14] See DeGroot and Shervish (2002, p.439) for a proof of this inequality.

so that

$$I(\theta) = \mathrm{E}\left[\frac{\partial \log f(X;\theta)}{\partial \theta}\frac{\partial \log f(X;\theta)}{\partial \theta'}\right] = \mathrm{E}\left[-\frac{\partial^2 \log f(X;\theta)}{\partial \theta\, \partial \theta'}\right]. \tag{A.191}$$

The Fisher information matrix in a random sample of n observations is $I_n(\theta) = nI(\theta)$. Third, let $\hat{\theta}$ be an unbiased estimator of θ. We denote the variance matrix of $\hat{\theta}$ by $\mathrm{Var}(\hat{\theta})$. Hence, the ith diagonal element of $\mathrm{Var}(\hat{\theta})$ is $\mathrm{Var}(\hat{\theta}_i)$, and its (i,j)th element is $\mathrm{Cov}(\hat{\theta}_i, \hat{\theta}_j)$. Denoting $I_n^{-1}(\theta)$ as the inverse of $I_n(\theta)$, the **Cramér–Rao** inequality states that

$$\mathrm{Var}(\hat{\theta}) - I_n^{-1}(\theta) \tag{A.192}$$

is a nonnegative definite matrix. As a property of nonnegative definite matrices, the diagonal elements of $\mathrm{Var}(\hat{\theta}) - I_n^{-1}(\theta)$ are nonnegative, i.e. the lower bound of $\mathrm{Var}(\hat{\theta}_i)$ is the ith diagonal element of $I_n^{-1}(\theta)$.

If the sample observations are not iid, the Fisher information matrix has to be computed differently. For the general case of a vector parameter θ, the Fisher information matrix in the sample is

$$I_n(\theta) = \mathrm{E}\left[\frac{\partial \log L(\theta;\boldsymbol{X})}{\partial \theta}\frac{\partial \log L(\theta;\boldsymbol{X})}{\partial \theta'}\right] = \mathrm{E}\left[-\frac{\partial^2 \log L(\theta;\boldsymbol{X})}{\partial \theta\, \partial \theta'}\right], \tag{A.193}$$

where the likelihood function $L(\theta;\boldsymbol{x})$ has to be established based on specific model assumptions, incorporating possibly the dependence structure and the specific pdf or pf of each observation. For the case of a scalar θ, equation (A.193) can be specialized easily.

A.19 Maximum likelihood estimation

We continue to use the notations introduced in Section A.18. Suppose there exists a value that maximizes $L(\theta;\boldsymbol{x})$ or, equivalently, $\log L(\theta;\boldsymbol{x})$, this value is called the **maximum likelihood estimate (MLE)** of θ. The MLE of θ, denoted by $\hat{\theta}$, is formally defined as

$$\hat{\theta} = \max_{\theta}\{L(\theta;\boldsymbol{x})\} = \max_{\theta}\{\log L(\theta;\boldsymbol{x})\}. \tag{A.194}$$

The MLE can be computed by solving the equation

$$\frac{\partial \log L(\theta;\boldsymbol{x})}{\partial \theta} = \sum_{i=1}^{n}\frac{\partial \log f(x_i;\theta)}{\partial \theta} = 0, \tag{A.195}$$

called the **first-order condition.** Under some *regularity conditions*, the distribution of $\sqrt{n}(\hat{\theta} - \theta)$ converges to a normal distribution with mean 0 and variance $1/I(\theta)$, i.e. in large samples $\hat{\theta}$ is approximately normally distributed with mean θ and variance $1/I_n(\theta)$. Thus, $\hat{\theta}$ is asymptotically unbiased. Also, as $I_n(\theta) = nI(\theta) \to \infty$ when $n \to \infty$, the variance of $\hat{\theta}$ converges to 0. Hence, by Theorem 10.1, $\hat{\theta}$ is consistent for θ. Furthermore, as the variance of $\hat{\theta}$ converges to the Cramér–Rao lower bound, $\hat{\theta}$ is **asymptotically efficient**.

Suppose $\tau = g(\theta)$ is a one-to-one transformation. Then the likelihood function $L(\theta; \boldsymbol{x})$ can also be expressed in terms τ, i.e. $L(\tau; \boldsymbol{x})$, and the MLE of τ, denoted by $\hat{\tau}$, can be computed accordingly. It turns out that $\hat{\tau} = g(\hat{\theta})$, so that the MLE of τ is $g(\cdot)$ evaluated at the MLE of θ.[15]

Given a function $\tau = g(\theta)$ (not necessarily a one-to-one transformation), using Taylor's expansion, we have[16]

$$\hat{\tau} = g(\hat{\theta}) \simeq g(\theta) + (\hat{\theta} - \theta)g'(\theta). \tag{A.196}$$

Thus,

$$\sqrt{n}(\hat{\tau} - \tau) \simeq \sqrt{n}(\hat{\theta} - \theta)g'(\theta), \tag{A.197}$$

so that the asymptotic distribution of $\sqrt{n}(\hat{\tau} - \tau)$ is normal with mean 0 and variance

$$\frac{\left[g'(\theta)\right]^2}{I(\theta)}. \tag{A.198}$$

In general, if the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta)$ for any consistent estimator $\hat{\theta}$ is $\sigma^2_{\hat{\theta}}$,[17] then the asymptotic variance of $\sqrt{n}(g(\hat{\theta}) - g(\theta))$ is

$$\left[g'(\theta)\right]^2 \sigma^2_{\hat{\theta}}. \tag{A.199}$$

This is known as the **delta method** for the computation of the asymptotic variance of $g(\hat{\theta})$.

The above results can be generalized to the case where θ is a vector of k elements. The MLE solved from equation (A.194) then requires the solution of a system of k equations. Furthermore, the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ follows a **multivariate normal distribution** with mean vector 0 and asymptotic variance matrix $I^{-1}(\theta)$. The multivariate normal

[15] See DeGroot and Shervish (2002, p.365) for a proof of this result.
[16] $g(\cdot)$ is assumed to be a smooth function satisfying some differentiability conditions.
[17] Here $\hat{\theta}$ is any consistent estimator of θ. It needs not be an MLE, and its asymptotic distribution needs not be normal.

distribution is a generalization of the univariate normal distribution. It has some important and convenient properties. First, the marginal distribution of each component of a multivariate normal distribution is normal. Second, any linear combination of the elements of a multivariate normal distribution is normally distributed.

Finally, we state the multivariate version of the delta method. If the asymptotic variance of a consistent estimator $\hat{\theta}$ of θ is $\Omega(\theta)$ (a $k \times k$ matrix) and $g(\cdot)$ is a smooth scalar function with a k-element argument, then the asymptotic variance of $\sqrt{n}(g(\hat{\theta}) - g(\theta))$ is

$$\frac{\partial g(\theta)}{\partial \theta'} \Omega(\theta) \frac{\partial g(\theta)}{\partial \theta}. \tag{A.200}$$

Answers to exercises

Chapter 1

1.4 $P_{X^*}(t) = \dfrac{P_X(t) - f_X(0)}{1 - f_X(0)}$

1.5 $f_X(x) = \binom{6}{x}(0.4)^x(0.6)^{6-x},\ x = 0, \cdots, 6$

1.6 (a) $\{0, \cdots, \infty\}$
(b) $\{0, \cdots, \infty\}$
(c) $\{0, \cdots, \infty\}$
(d) $\{0, \cdots, mn\}$

1.7 0.6904

1.8 0.4897

1.9 (a) $\mathrm{E}(W) = \mathrm{E}(Y)$
(b) $\mathrm{Var}(W) = \sigma_X^2 \sum_{i=1}^n p_i \geq \sigma_X^2 \sum_{i=1}^n p_i^2 = \mathrm{Var}(Y)$

1.10 (a) $\exp\left[\lambda_1\left(e^{\lambda_2(e^t-1)} - 1\right)\right]$
(b) $\mathrm{E}(S) = \lambda_1\lambda_2, \quad \mathrm{Var}(S) = \lambda_1\lambda_2(1 + \lambda_2)$
(c) $f_S(0) = \exp\left[\lambda_1(e^{-\lambda_2} - 1)\right], \quad f_S(1) = f_S(0)\lambda_1\lambda_2 e^{-\lambda_2}$

1.11 (a) $1 - \dfrac{\log[1 - \beta(t-1)]}{\log(1+\beta)}$
(b) $[1 - \beta(t-1)]^{-\frac{\lambda}{\log(1+\beta)}}$, i.e. pgf of $\mathcal{NB}(r, \theta)$, where $r = \dfrac{\lambda}{\log(1+\beta)}$ and $\theta = \dfrac{1}{(1+\beta)}$

1.12 $\Pr(X \geq 4) = 0.1188, \quad \mathrm{E}(X) = 1.2256, \quad \mathrm{Var}(X) = 2.7875$

1.13 $\Pr(S = 0) = 0.0067, \quad \Pr(S = 1) = 0.0135, \quad \Pr(S = 2) = 0.0337$

1.14 (a) $\Pr(S = 0) = 0.0907, \quad \Pr(S = 1) = 0.0435, \quad \Pr(S = 2) = 0.0453$

(b) $\Pr(S = 0) = 0.1353$, $\Pr(S = 1) = 0.0866$, $\Pr(S = 2) = 0.0840$

1.15 (a) $\exp\left[\sum_{i=1}^{n} \lambda_i (t^{x_i} - 1)\right]$

(b) $\exp\left[-\sum_{i=1}^{n} \lambda_i\right]$

1.16 $\mathrm{E}(S_1) = 8$, $\mathrm{Var}(S_1) = 48$, $\mathrm{E}(S_2) = 8$, $\mathrm{Var}(S_2) = 40$

1.17 $P_X(t) = 0.64\, e^{t-1} + 0.32\, e^{2(t-1)} + 0.04\, e^{3(t-1)}$, $\mathrm{E}(X) = 1.4$, $\mathrm{Var}(X) = 1.72$

1.18 0.0979

1.19 (a) 0.0111

(b) 0.0422

1.20

x	0	1	2
$\Pr(X_1 = x)$	0.3679	0.3679	0.1839
$\Pr(X_2 = x)$	0.1353	0.2707	0.2707

$\Pr(X_1 + X_2 \leq 2) = 0.4232$, note that $X_1 + X_2 \sim \mathcal{PN}(3)$

1.21 primary: $\mathcal{BN}(1, c)$, i.e. Bernoulli with parameter c; secondary: X

X^* is a mixture of a degenerate distribution at 1 and X, with weights $1 - c$ and c, respectively

1.24 $f_S(0) = 0.3012$, $f_S(1) = 0.1807$, $f_S(s) = [0.6 f_S(s-1) + 1.2 f_S(s-2)]/s$ for $s \geq 2$

1.25 (a) $P_S(t) = \left[\dfrac{\theta}{1 - (1-\theta)e^{\lambda(t-1)}}\right]^r$

1.26 $f_X(x) = 0.2 f_X(x-1)$ with $f_X(0) = 0.8$

$f_X^M(x) = 0.2 f_X^M(x-1)$ with $f_X^M(0) = 0.4$

mean = 0.75, variance = 0.5625

1.27 $1 - c + cP_S(t)$, where $c = 1/(1 - e^{-\lambda})$ and $P_S(t) = \left[\theta e^{\lambda(t-1)} + 1 - \theta\right]^n$, assuming primary distribution $\mathcal{BN}(n, \theta)$ and secondary distribution $\mathcal{PN}(\lambda)$

1.28 $f_S(0) = 0.5128$, $f_S(1) = 0.0393$, $f_S(2) = 0.0619$

1.29 (a) 0.2652

(b) 0.4582

(c) 0.1536

1.30 (a) $M_S(t) = \dfrac{\theta_N}{1 - (1 - \theta_N)(\theta_X e^t + 1 - \theta_X)^n}$,

$P_S(t) = \dfrac{\theta_N}{1 - (1 - \theta_N)(\theta_X t + 1 - \theta_X)^n}$

(b) $M_S(t) = \left[\theta e^{\lambda(e^t - 1)} + 1 - \theta\right]^n$, $P_S(t) = \left[\theta e^{\lambda(t-1)} + 1 - \theta\right]^n$

(c) $M_S(t) = \left[\dfrac{\theta}{1 - (1-\theta)e^{\lambda(e^t - 1)}}\right]^r$, $P_S(t) = \left[\dfrac{\theta}{1 - (1-\theta)e^{\lambda(t-1)}}\right]^r$

Chapter 2

2.2 (a) $F_X(x) = e^{-\frac{\theta}{x}}$, $S_X(x) = 1 - e^{-\frac{\theta}{x}}$, $h_X(x) = \dfrac{\theta}{\left(e^{\frac{\theta}{x}} - 1\right)x^2}$, for $0 < x < \infty$

(b) median $= \dfrac{\theta}{\log(2)}$, mode $= \dfrac{\theta}{2}$

2.3 (a) $S_X(x) = 1 - e^{-\left(\frac{\theta}{x}\right)^{\tau}}$, $f_X(x) = \dfrac{\tau}{x}\left(\dfrac{\theta}{x}\right)^{\tau} e^{-\left(\frac{\theta}{x}\right)^{\tau}}$,

$h_X(x) = \dfrac{\tau\left(\dfrac{\theta}{x}\right)^{\tau} e^{-\left(\frac{\theta}{x}\right)^{\tau}}}{x\left[1 - e^{-\left(\frac{\theta}{x}\right)^{\tau}}\right]}$, for $0 < x < \infty$

(b) median $= \dfrac{\theta}{[\log(2)]^{\frac{1}{\tau}}}$, mode $= \theta\left(\dfrac{\tau}{\tau+1}\right)^{\frac{1}{\tau}}$

2.4 (a) $S_X(x) = 1 - 0.01x$, $F_X(x) = 0.01x$, $f_X(x) = 0.01$, for $x \in [0, 100]$

(b) $\mathrm{E}(X) = 50$, $\mathrm{Var}(X) = \dfrac{(100)^2}{12}$

(c) median = 50, mode = any value in [0,100]

(d) 45

2.5 (a) $S_X(x) = 1 - \dfrac{3x^2}{400} + \dfrac{x^3}{4{,}000}$, $F_X(x) = \dfrac{3x^2}{400} - \dfrac{x^3}{4{,}000}$,

$h_X(x) = \dfrac{60x - 3x^2}{4{,}000 - 30x^2 + x^3}$

(b) $\mathrm{E}(X) = 10$, $\mathrm{Var}(X) = 20$

(c) median = mode = 10

(d) 1

2.6 2.2705

2.7 mean = 1.25, median = 0.5611

2.8 e^{-y} for $0 < y < \infty$

2.9 $\dfrac{1}{\pi\left(1 + y^2\right)}$, for $-\infty < y < \infty$

2.10 (a) $F_X(x) = \begin{cases} 0, & x < 0 \\ 0.2, & x = 0 \\ 0.2 + 0.08\left(x - \dfrac{x^2}{40}\right), & 0 < x \le 20 \\ 1, & 20 < x \end{cases}$

(b) $\mathrm{E}(X) = 5.3333$, $\mathrm{Var}(X) = 24.8889$

(c) 10

2.11 (a) $F_X(x) = \begin{cases} 0, & x < 0 \\ 0.4, & x = 0 \\ 0.4 + 0.6x^4, & 0 < x \le 1 \\ 1, & 1 < x \end{cases}$

(b) $\mathrm{E}(X) = 0.48, \quad \mathrm{Var}(X) = 0.1696$

(c) 0.9036

2.12 $\dfrac{2}{\sqrt{\alpha}}$

2.13 $f_X(x) = \begin{cases} 1.7759e^{-2x}, & 0 \le x < 0.8 \\ 2.2199xe^{-2x}, & 0.8 \le x < \infty \end{cases}$

2.14 (a) $\mathrm{E}(X) = \dfrac{\log(5)}{4}, \quad \mathrm{Var}(X) = \dfrac{2}{5} - \left[\dfrac{\log(5)}{4}\right]^2$

(b) $\dfrac{1}{4x}\left[e^{-x}\left(1 + \dfrac{1}{x}\right) - e^{-5x}\left(5 + \dfrac{1}{x}\right)\right]$

2.15 $\mathrm{E}(X) = 7, \quad \mathrm{Var}(X) = 24$

2.16 $x_{0.9} = 4.6697, \ \mathrm{CTE}_{0.9} = 8.2262$

2.17 (a) $F_{X_L}(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda d}, & x = 0 \\ 1 - e^{-\lambda(x+d)}, & 0 < x < \infty \end{cases}$

(b) $F_{X_P}(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & 0 \le x < \infty \end{cases}$

(c) $f_{X_P}(x) = \lambda e^{-\lambda x}$, for $x \ge 0$ and 0 otherwise, $\mathrm{E}(X_P) = \dfrac{1}{\lambda}$

2.18 (a) 6

(b) 0.8333

(c) 0.1667

2.20 (a) $1 - pe^{-\lambda_1 x} - (1-p)e^{-\lambda_2 x}$, for $0 < x < \infty$

(b) $\dfrac{pe^{-\lambda_1 d}}{\lambda_1} + \dfrac{(1-p)e^{-\lambda_2 d}}{\lambda_2}$

2.21 3.4286

2.22 $\mathrm{E}(X_P) = 37.5, \quad \mathrm{Var}(X_P) = 393.75$

2.23 12.1371

2.24 $\mathrm{E}(X_L) = 92.3116, \quad \mathrm{Var}(X_L) = 9{,}940.8890$

2.25 without inflation: 54.6716, with inflation: 56.9275

2.29 $\alpha e^{-\alpha y}$ for $0 \le y < \infty$

2.30 6.4745

2.31 3.7119

Chapter 3

3.1 3.0888

3.2 $\theta = 0.2268, \quad \beta = 0.0971$

3.3 $\lambda_1 = 1.3333, \quad \lambda_2 = 1.5$

3.4 mean = 800, variance = 40,000

3.5 $\left[\dfrac{\theta}{(1-\beta t)^{\alpha}} + 1 - \theta\right]^n$

3.6 (a) 0.6561, (b) 0.0820

3.7 (a) 0.0020, (b) 0.0017

3.8 $\left[\dfrac{1}{3}\left(1 + \dfrac{0.05}{0.05-t} + \left(\dfrac{0.05}{0.05-t}\right)^2\right)\right]^5$

3.9 mean = 1.05, variance = 2.4475

3.10 0.9939

3.11 $\left[0.6(0.5t + 0.3t^2 + 0.2t^3) + 0.4\right]^2$

3.12 (a) & (b): 0.9532

3.13 $$F_{X_1+X_2}(x) = \begin{cases} 0, & x < 0 \\ 0.25, & x = 0 \\ 0.25 + 0.25x + 0.03125x^2, & 0 < x \le 2 \\ 0.5 + 0.25x - 0.03125x^2, & 2 < x \le 4 \\ 1, & 4 < x \end{cases}$$

3.14 0.031

3.15 15.25

3.16 18

3.17 mean = 19.20, variance = 101.03

3.18 $\mathrm{E}(S) = 70$, $\mathrm{Var}(S) = 4{,}433.33$, $\mathrm{E}(\tilde{S}) = 44.80$, $\mathrm{Var}(\tilde{S}) = 2{,}909.02$

3.19 compound Poisson distribution with $\lambda = 3$, and $f_X(-2) = 1/3$ and $f_X(1) = 2/3$

3.20 mean = 30, variance = 72

3.21 0.74

3.22 (a) mean = 415, variance = 19,282.50

(b) using compound Poisson (other distributions may also be used), mean = 415, variance = 20,900

3.23 0.7246

3.24 1.6667

3.25 0.52

3.26 0.2883

3.27 0.0233

Chapter 4

4.6 $\dfrac{\alpha\beta}{1-\beta\rho}$

4.7 $\mathrm{VaR}_\delta(X) = x_\delta = \lambda\,[-\log(1-\delta)]^{\frac{1}{\alpha}}$, PH premium $= \lambda\rho^{\frac{1}{\alpha}}\,\Gamma\left(1+\dfrac{1}{\alpha}\right)$

4.8 (b) $1-\dfrac{c}{2}$

(c) $\mathrm{VaR}_\delta(X) = \begin{cases} \dfrac{\delta}{c}, & \text{for } 0 < \delta < c \\ 1, & \text{for } c \le \delta < 1 \end{cases}$

(d) $\mathrm{CTE}_\delta(X) = \dfrac{1}{1-\delta}\left(\dfrac{c^2-\delta^2}{2c} + 1 - c\right)$, for $\delta \in (0, c)$

4.9 $\mathrm{VaR}_{0.90} = 40$, $\mathrm{VaR}_{0.95} = 50$, $\mathrm{VaR}_{0.99} = 60$, $\mathrm{CTE}_{0.90} = 52$, $\mathrm{CTE}_{0.95} = 58$

4.10 (a) $\dfrac{2b\rho}{\rho+1}$, (b) $b\rho$, (c) $\dfrac{b\rho}{2-\rho}$; the PH premium of the Pareto loss distribution is the most sensitive to the risk-aversion parameter

4.11 (a) pure premium $= \dfrac{b}{2}$, expected-value premium $= (1+\theta)\dfrac{b}{2}$

(b) variance premium $= \dfrac{b}{2} + \dfrac{\alpha b^2}{12}$, standard-deviation premium $= \dfrac{b}{2} + \dfrac{\alpha b}{\sqrt{12}}$

(c) $\mathrm{VaR}_\delta = b\delta$, $\mathrm{CTE}_\delta = \dfrac{b(1+\delta)}{2}$

(e) $\mathrm{CVaR}_\delta = \dfrac{b(1-\delta)}{2}$, $\mathrm{TVaR}_\delta = \dfrac{b(1+\delta)}{2}$

4.12 (a) pure premium $= \dfrac{1}{\lambda}$, expected-value premium $= (1+\theta)\dfrac{1}{\lambda}$

(b) variance premium $= \dfrac{1}{\lambda} + \dfrac{\alpha}{\lambda^2}$, standard-deviation premium $= \dfrac{1+\alpha}{\lambda}$

(c) $\mathrm{VaR}_\delta = -\dfrac{\log(1-\delta)}{\lambda}$, $\mathrm{CTE}_\delta = \dfrac{1-\log(1-\delta)}{\lambda}$

(e) $\mathrm{CVaR}_\delta = \dfrac{1}{\lambda}$, $\mathrm{TVaR}_\delta = \dfrac{1-\log(1-\delta)}{\lambda}$

4.13 expected-value premium $= 1.2\lambda\alpha\beta$, loading $= \dfrac{0.2}{\beta(1+\alpha)}$

4.14 $\dfrac{\lambda\alpha\beta}{(1-\beta\rho)^{1+\alpha}}$

4.16 (a) $\mathrm{VaR}_{0.95}(P_1) = \mathrm{VaR}_{0.95}(P_2) = 0$

(b)

$P_1 + P_2$	0	100	200
prob	0.9216	0.0768	0.0016

(c) $\mathrm{VaR}_{0.95}(P_1 + P_2) = 100 > \mathrm{VaR}_{0.95}(P_1) + \mathrm{VaR}_{0.95}(P_2) = 0$, hence not sub-additive

4.17 $\text{VaR}_{0.9} = 400,\ \text{CTE}_{0.9} = 480,\ \text{CVaR}_{0.9} = 80,\ \text{TVaR}_{0.9} = 480$

4.20 PH premium of $U = 4^{1-\frac{1}{\rho}}$, PH premium of $V = \dfrac{2\rho}{3-\rho}$

Chapter 5

5.2 3.2299
5.3 0.5312
5.4 0.0237
5.5 adj coeff = 0.4055, max prob = 0.2963
5.6 adj coeff = 0.6755, max prob = 0.2590
5.7 adj coeff = 0.8109, max prob = 0.1975
5.8 adj coeff = 0.1853, max prob = 0.6903
5.9 adj coeff = 0.3333, max prob = 0.2636
5.10 adj coeff = 0.2918, max prob = 0.4167
5.11 adj coeff = 0.3749, max prob = 0.4725
5.12 adj coeff = 0.3023, max prob = 0.5463
5.13 max prob = 0.5799, loading = 12.26%, initial surplus = 4.7744
5.14 7.7778
5.15 0.4
5.16 0.36
5.17 0.199

Chapter 6

6.1 (a) 0.6958
(b) 2.0781
(c) 0.0973
(d) 17.04%
(e) $\bar{X}$ approximately normally distributed

6.2 (a) mean = 68,280, variance = 11,655,396
(b) 0.7699
(c) $\mu_X = 110.1290,\ \sigma_X = 81.6739$
(d) 500

6.3 for claim frequency: 983, for claim severity: 456, full credibility attained for claim severity but not claim frequency

6.4 for claim frequency: $2\lambda_F$, for aggregate loss: $\lambda_F(2 + C_X^2)$

6.7 $k = 4.05\%$
(a) 8,624
(b) 12,902

6.8 $1.0133\,\lambda_F$
6.9 $2\lambda_F$, $1.4394\,\lambda_F$
6.10 242.9490
6.11 356.8685, 354
6.12 (a) 0.9557
(b) 1
6.13 1
6.14 24.9038
6.15 16,910
6.16 0.47
6.17 960
6.18 2,469

Chapter 7

7.1 $0.4\,c$
7.2 for N: 0.525, for X: 7.25
7.3 $0.1523 + \dfrac{0.725}{c}$
7.4 for N: 6.1348, for X, $c = 20$: 11.7142, for X, $c = 30$: 13.0952
7.6 1.3333
7.7 $\dfrac{x+1}{\alpha}$
7.8 0.4283
7.9 1.3333
7.10 $\dfrac{\theta - \theta^2}{1.5 - \theta^2}$
7.11 2.25
7.12 10,622
7.13 16.91
7.14 1.41
7.15 0.9375
7.16 0.8565
7.17 2.40
7.18 0.2222
7.19 0.905
7.20 12
7.21 1,063
7.22 8.3333
7.23 1,138
7.24 2,075
7.25 257.11

Chapter 8

8.1 (a) $\dfrac{m\,(\alpha + n\bar{x})}{\alpha + \beta + mn}$

(b) $\dbinom{m}{x_{n+1}} \dfrac{B(\alpha + n\bar{x} + x_{n+1}, \beta + mn + m - n\bar{x} - x_{n+1})}{B(\alpha + n\bar{x}, \beta + mn - n\bar{x})}$

8.2 $a(\alpha, \beta) = \alpha - 1$, $b(\alpha, \beta) = \dfrac{(\beta - 1) + (\alpha - 1)}{m}$, $\alpha^* = \alpha + n\bar{x}$, $\beta^* = \beta + mn - n\bar{x}$

8.3 for sample mean: 1.75, for Bühlmann premium: 1.6251

8.4 $f_X(x) = \dbinom{x + r - 1}{r - 1} \dfrac{B(\alpha + r, \beta + x)}{B(\alpha, \beta)}$, for $x = 0, 1, \cdots$

8.5 $A(\theta) = \log(1 - \theta)$, $B(\theta) = r \log(\theta)$, $C(x) = \log[(x + r - 1)!] - \log(x!) - \log[(r - 1)!]$

8.6 (a) $\mathcal{PN}(\theta\lambda)$

(b) $\mathcal{BN}(m, \theta\beta)$

8.7 1.0714

8.8 1.3193

8.9 7

8.10 $\dfrac{x + c}{2}$

8.11 $\dfrac{10}{19}$

8.12 12

8.13 7.2022

8.14 $\mathcal{E}(0.5)$

8.15 9.8848

8.16 3.25

8.17 3.8293

8.18 0.9420

8.19 $\dfrac{4}{3}$

8.20 0.2126

8.21 0.45

8.22 0.7211

8.23 0.3125

8.24 0.8148

Chapter 9

9.1 Policyholder 1: 23.02, Policyholder 2: 23.20, Group 3: 21.29

9.2 (a) Group 1: 14.355, Group 2: 13.237, Group 3: 11.809

(b) Group 1: 14.428, Group 2: 13.254, Group 3: 11.840

9.3 0.852
9.4 0.323
9.5 0.499
9.6 0.393
9.7 0.575
9.8 0.872
9.9 0.633
9.10 0.778
9.11 687.38
9.12 0.074
9.13 0.221
9.14 0.818

Chapter 10

10.3 $\mathrm{E}(X_{(n-1)}) = \dfrac{(n-1)\theta}{n+1}$, $\mathrm{Var}(X_{(n-1)}) = \dfrac{2(n-1)\theta^2}{(n+1)^2(n+2)}$,
$\mathrm{MSE}(X_{(n-1)}) = \dfrac{6\theta^2}{(n+1)(n+2)} > \mathrm{MSE}(X_{(n)}) = \dfrac{2\theta^2}{(n+1)(n+2)}$,
$\mathrm{Cov}(X_{(n-1)}, X_{(n)}) = \dfrac{(n-1)\theta^2}{(n+1)^2(n+2)}$

10.4 (a) $\bar{X}$ is biased and inconsistent
(b) $X_{(1)}$ is biased but consistent

10.5 no, $np(1-p)$ is biased for $\mathrm{Var}(X)$

10.6 $\hat{\mu}_2' - s^2$, where s^2 is the sample variance, is unbiased for $[\mathrm{E}(X)]^2$

10.7 $\left(\dfrac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \dfrac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \right)$, where $\chi^2_{r,\alpha}$ is the 100α-percentile of the χ^2_r distribution

10.8 $\left(\dfrac{\gamma_{\frac{\alpha}{2}}(n)}{n\bar{x}}, \dfrac{\gamma_{1-\frac{\alpha}{2}}(n)}{n\bar{x}} \right)$, where $\gamma_\alpha(n)$ is the 100α-percentile of $\mathcal{G}(n,1)$

10.9

y_j	2	3	4	5	6	7	8	9	10	11	12
w_j	1	1	1	3	2	3	1	1	1	1	3
r_j	18	17	16	15	12	10	7	6	5	4	3

10.10

y_j	23	25	27	28	30	31	33	38	42	45
w_j	2	1	3	1	2	1	1	1	1	3
r_j	16	14	13	10	9	7	6	5	4	3

10.11

y_j	w_j	r_j
5	1	10
8	1	9
11	1	8
13	2	7
14	1	5
15	1	3
16	1	1

10.12 Note that the observations of $i = 11$, 14 and 18 are not observable (as $x_i < d_i$) and are removed from the data. Thus, we have (there are no censored data):

y_j	w_j	r_j	Eq. (10.8)	Eq. (10.9)
3	1	8	—	$8-0$
5	2	10	$8-1+3$	$11-1$
6	3	8	$10-2+0$	$11-3$
7	3	11	$8-3+6$	$17-6$
8	4	8	$11-3+0$	$17-9$
9	3	4	$8-4+0$	$17-13$
10	1	1	$4-3+0$	$17-16$

10.13

y_j	w_j	r_j
6	3	21
8	2	18
12	1	16
13	2	15
14	1	13
15	2	12
16	2	10
17	2	8
18	2	6

10.14 (a)

y_j	w_j	r_j	Eq. (10.8)	Eq. (10.9)
4	1	10	—	$10-0-0$
7	1	19	$10-1+10-0$	$20-1-0$
8	2	18	$19-1+0-0$	$20-2-0$
9	2	16	$18-2+0-0$	$20-4-0$
10	1	14	$16-2+0-0$	$20-6-0$
12	3	13	$14-1+0-0$	$20-7-0$
13	3	10	$13-3+0-0$	$20-10-0$
14	1	7	$10-3+0-0$	$20-13-0$
15	1	6	$7-1+0-0$	$20-14-0$

(b)

Group j	D_j	U_j	V_j	R_j
$(0,5]$	15	0	1	15
$(5,10]$	5	0	6	19
$(10,15]$	0	5	8	13

Chapter 11

11.1 (a) $\tilde{F}(15) = 0.4479$, $\tilde{F}(27) = 0.8125$
(b) $\hat{x}_{0.25} = 11.5$, $\hat{x}_{0.75} = 26.9167$, $\hat{x}_{0.75} - \hat{x}_{0.25} = 15.4167$
(c) mean = 19.8125, variance = 80.1523, skewness = –0.0707
(d) $\text{Var}[(X \wedge 27)] = 62.4336$, $\text{Var}[(X \wedge 31.5)] = 77.0586$
(e) $\widehat{\Pr}(\tilde{X} \leq 25) = 0.5231$, $\hat{\text{E}}[(\tilde{X} \wedge 20.5)] = 18.6538$

11.2 (a) loss event: mean = 47.9, variance = 1,469.2111
payment event: mean = 59.8750, variance = 1,069.5536
(b) prob = 0.7, variance = 0.21

11.3 minimum correlation = –0.9424, maximum correlation = 0.8870

11.4 (a) 12
(b) 14

11.5 (a) for $b = 4$: $\tilde{f}(10) = 0.05$, $\tilde{f}(15) = 0.0375$
for $b = 6$: $\tilde{f}(10) = 0.0417$, $\tilde{f}(15) = 0.0333$
(b) for $b = 4$: $\tilde{f}(10) = 0.0375$, $\tilde{f}(15) = 0.0375$
for $b = 6$: $\tilde{f}(10) = 0.0417$, $\tilde{f}(15) = 0.0333$

11.6 (a) mean = 25.2985, standard deviation = 14.5850
(b) $\text{E}[(X \wedge 40)] = 23.8060$, $\text{E}[(X \wedge 45)] = 24.5522$
(c) $\text{Var}[(X \wedge 40)] = 148.6972$, $\text{Var}[(X \wedge 45)] = 175.7964$
(d) 8.8433

(e) $\hat{F}(40) = 0.8209$, $\hat{F}(48) = 0.9164$
(f) 16.2754

11.7 (a) mean = 63.8393, standard deviation = 74.4153
(b) mean = 200.2321, standard deviation = 85.7443

11.8 (a) mean = 43.9765, standard deviation = 27.4226
(b) mean = 38.5412, standard deviation = 21.6795

11.9 (a) prob = 0.3333, confidence interval = (0.1155, 0.5511)
(b) prob = 0.2222, variance = 0.009602
(c) prob = 0.4, variance = 0.0160
(d) $\hat{H}(10.5) = 1.3531$, confidence interval = (0.5793, 2.1269)

11.10 (a) 0.3125
(b) 0.7246

11.11 (a) 12.6667
(b) $\hat{H}_K(12) = 0.3567$, $\hat{H}_N(12) = 0.3361$

11.12 (a) prob = 0.0398, variance = 0.001625
(b) linear: (–0.0392, 0.1188), lower limit < 0, undesirable
logarithmic: (0.0026, 0.1752)
(c) $\hat{H}(5) = 0.325$, variance = 0.0356
(d) linear: (–0.0449, 0.6949), lower limit < 0, undesirable
logarithmic: (0.1041, 1.0145)

11.13 Kaplan–Meier: 10.3030, Nelson–Aalen: 11.0322

11.14 (a) (0.2005, 0.5658)
(b) (0.0948, 0.7335)

11.15 (a) 0.3614
(b) 0.2435
(c) 0.5085

11.16 (1.5802, 2.4063)

11.17 (0.2144, 0.5598)

11.18 8

11.19 0.2341

11.20 0.4780

11.21 0.3

11.22 0.3854

11.23 0.7794

11.24 0.36

11.25 0.0667

11.26 10

11.27 1.0641

11.28 2

11.29 0.5833

11.30 0.485
11.31 36

Chapter 12

12.1 $(\alpha - 1)\hat{\mu}'_1$
12.2 (a) 30.1924
(b) 46.0658
12.3 mean = 97,696.51, standard deviation = 211,005.60
12.4 (a) $\mathcal{G}(6.4470, 6.6698)$
(b) $\mathcal{U}(13.6674, 72.3326)$
12.5 (a) $\hat{\mu} = 4.1740$, $\hat{\sigma}^2 = 0.0433$, $\widehat{\Pr}(X > 80) = 0.1588$
(b) $\hat{\mu} = 4.1724$, $\hat{\sigma}^2 = 0.0478$, $\widehat{\Pr}(X > 80) = 0.1687$
12.6 (a) $\hat{b} = 28.9286$
(b) $\hat{\lambda} = 0.0443$
(c) $\hat{\gamma} = 39.8571$
12.7 $\hat{\lambda}_1 = 0.1634$, $\hat{\lambda}_2 = 0.3867$
12.8 (a) 97.3093
(b) 75.2438
12.9 $3\bar{x}$
12.10 $\hat{\mu}_X = \bar{x}$, $\hat{\mu}_Y = \bar{y}$, $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{n+m}$
12.11 $\mathrm{Var}(\bar{x}) = \dfrac{1-\theta}{n\theta^2}$, $\mathrm{Var}(\hat{\theta}) = \dfrac{\theta^2(1-\theta)}{n}$
12.12 $X_{(n)} - 0.5 \le \hat{\theta} \le X_{(1)} + 0.5$
(a) yes
(b) no
12.13 (a) $4\log(1 - e^{-2\lambda}) + 7\log(e^{-2\lambda} - e^{-4\lambda}) + 10\log(e^{-4\lambda} - e^{-6\lambda}) + 6\log(e^{-6\lambda} - e^{-8\lambda}) - 24\lambda$
(b) $4\log\left[1 - \left(\dfrac{\gamma}{2+\gamma}\right)^{\alpha}\right] + 7\log\left[\left(\dfrac{\gamma}{2+\gamma}\right)^{\alpha} - \left(\dfrac{\gamma}{4+\gamma}\right)^{\alpha}\right] + 10\log\left[\left(\dfrac{\gamma}{4+\gamma}\right)^{\alpha} - \left(\dfrac{\gamma}{6+\gamma}\right)^{\alpha}\right] + 6\log\left[\left(\dfrac{\gamma}{6+\gamma}\right)^{\alpha} - \left(\dfrac{\gamma}{8+\gamma}\right)^{\alpha}\right] + 3\alpha\log\left(\dfrac{\gamma}{8+\gamma}\right)$
12.14 (a) 37.33 (3 losses exceeding 28), 33.60 (2 losses exceeding 28)
(b) 0.0391
12.16 (a) 0.0811
(b) 0.0811
(c) $\hat{\lambda} = 14.0040$, $\hat{\lambda}^* = 9.7200$
12.17 1.0874

12.18 1.4256

12.19 (a) $\hat{\mu} = \bar{x}, \quad \sqrt{n}\left(\hat{\mu} - \mu\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$

(b) $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}, \quad \sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) \xrightarrow{D} \mathcal{N}(0, 2\sigma^4)$

(c) $\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n},$

$$\sqrt{n}\begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{D} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right)$$

12.20 $\dfrac{9\mu^4\sigma^2}{n}$

12.21 $\dfrac{2}{n\sigma^4}$

12.22 (a) 0.0323

(b) 0.00008718

(c) (0.01404, 0.05065), not reliable as sample size is too small

(d) 0.00005034

(e) 0.4455, (0.2417, 0.6493)

12.23 (a) $\hat{\lambda} = 0.2273, \quad \hat{\beta} = -0.4643$

(b) $\hat{\beta} = -0.9971$

y	2	3	5	7	8	12
$\hat{S}(y)$	0.8569	0.5947	0.2865	0.1611	0.0416	0.0028

$\hat{S}(6; z_{(1)}) = 0.6305$

12.25 $\log L(\beta; \boldsymbol{x}, \boldsymbol{z}) = -\sum_{i=1}^{n} e^{\beta z_i} + \beta \sum_{i=1}^{n} x_i z_i, \quad \hat{\beta} = 0.7511$

12.26 (a) $\log L(\lambda, \beta; \boldsymbol{x}, \boldsymbol{z}) = n \log \lambda + \beta \sum_{i=1}^{10} z_i - \lambda \sum_{i=1}^{10} x_i e^{\beta z_i}$
$\hat{\lambda} = 0.0161, \quad \hat{\beta} = 0.6624$

(b) $\log L(\alpha, \gamma; \boldsymbol{x}, \boldsymbol{z}) = n(\log \alpha + \alpha \log \gamma) + \beta \sum_{i=1}^{n} z_i - (\alpha + 1) \sum_{i=1}^{n} \log\left(x_i e^{\beta z_i} + \gamma\right)$, where $\alpha = 3$;
$\hat{\gamma} = 188.59, \quad \hat{\beta} = 0.6829$

12.28 (a) $\hat{\lambda}_1 = 0.3117, \quad \hat{\lambda}_2 = 0.1925, \quad \hat{\alpha} = 21.2511$

(b) 0.5352

12.29 26,400

12.30 0.6798

12.31 2,887.66

12.32 0.2

12.33 0.0406

12.34 (–0.476, 1.876)

12.35 $\lambda^3 e^{-1100\lambda}$

12.36 104.4665

12.37 0.9700

12.38 0.9204

12.39 118.3197
12.40 471.3091
12.41 10.1226
12.42 15
12.43 1.0358
12.44 5.5

Chapter 13

13.1 (a) The (x_i, y_i) co-ordinates are

x_i	y_i
0.125	0.166
0.250	0.365
0.375	0.420
0.500	0.559
0.625	0.720
0.750	0.805
0.875	0.876

(b) the x_i values are: 0.071, 0.214, 0.357, 0.500, 0.643, 0.786, 0.929
the y_i values are the same as in (a)
(c) the x_i values are the same as in (a)
the y_i values are: 0.164, 0.416, 0.416, 0.554, 0.715, 0.801, 0.873
(d) the x_i values are the same as in (b)
the y_i values are the same as in (c)

13.2 (a) & (b): the x_i values are the same for both (a) and (b), but y_i are different

x_i	$y_i(a)$	$y_i(b)$
12	14.16	11.82
15	17.90	16.71
18	20.94	20.21
21	23.77	23.35
23	26.58	26.44
28	29.52	29.68
32	32.77	33.29
38	36.56	37.65
45	41.43	43.60
58	48.95	54.92

13.3 29.0249
13.4 0.2120
13.5 $D = 0.1723 < 0.3678$, the critical value at the 10% level of significance, but this critical value should be adjusted downwards due to parametric estimation
13.6 $D = 0.1811 < 0.3678$, the critical value at the 10% level of significance, but this critical value should be adjusted downwards due to parametric estimation
13.7 $A^2 = 0.4699 < 1.933$, the critical value at the 10% level of significance, but this critical value should be adjusted downwards due to parametric estimation
13.8 $A^2 = 0.3633 < 1.933$, the critical value at the 10% level of significance, but this critical value should be adjusted downwards due to parametric estimation
13.9 9.1507
13.10 $X^2 = 7.70$, $\chi^2_{2,0.975} = 7.38$ and $\chi^2_{2,0.99} = 9.21$; reject null at 2.5% level but not at 1% level
13.11 0.2727
13.12 $X^2 = 7.546$, $\chi^2_{2,0.975} = 7.38$ and $\chi^2_{2,0.99} = 9.21$; reject null at 2.5% level but not at 1% level
13.13 $X^2 = 37$ and $\chi^2_{19,0.99} = 36.191$; reject null at 1% level
13.14 $\ell = 6.901$ and $\chi^2_{1,0.99} = 6.635$; reject null at 1% level
13.15 0.026
13.16 $D = 0.111$, reject null at 5% level but not at 1% level
13.17 select Model 1 based on BIC, select Model 4 based on AIC
13.18 $D = 0.6803$, reject null at 5% level but not at 1% level
13.19 $X^2 = 11.022$, $\chi^2_{4,0.95} = 9.488$ and $\chi^2_{4,0.975} = 11.143$; reject null at 5% level but not at 2.5% level
13.20 $X^2 = 9.36$ and $\chi^2_{2,0.99} = 9.21$; reject null at 1% level
13.21 $\ell = 4.46$, $\chi^2_{1,0.95} = 3.84$ and $\chi^2_{1,0.975} = 5.02$; reject null at 5% level but not at 2.5% level
13.22 $(s, t) = (0.5, 0.5971)$, $D(3000) = -0.0257$
13.23 81
13.24 0.1679

Chapter 14

14.1 for seed 401, x_i are: 6.7396×10^6, 1.6034×10^9, 2.1426×10^9, 1.9729×10^9; u_i are: 0.0031, 0.7467, 0.9977, 0.9187

for seed 245987, x_i are: 1.9868×10^9, 1.2582×10^9, 1.8552×10^8, 2.0831×10^9; u_i are: 0.9252, 0.5859, 0.0864, 0.9700

14.2 for seed 747, x_i are: 5.1595×10^7, 3.0557×10^9, 1.7529×10^9, 7.9814×10^8; u_i are: 0.0120, 0.7115, 0.4081, 0.1858
for seed 380, x_i are: 2.6246×10^7, 3.2404×10^8, 4.2934×10^9, 7.9153×10^8; u_i are: 0.0061, 0.0754, 0.9996, 0.1843

14.3 7

14.4 0.3479, 0.7634, 0.8436, 0.6110, 0.1399

14.7 Generate u and v independently from $\mathcal{U}(0,1)$. If $16u/9 > f(v)$, generate another (u, v) again. If $16u/9 \leq f(v)$ output v as value of X. Inverse transformation has no explicit solution.

14.8 (a) $\text{Var}(X) = \text{Var}(Y) = \dfrac{4}{45}$, $\text{Cov}(X, Y) = -\dfrac{7}{90}$, $\text{Var}\left(\dfrac{X+Y}{2}\right) = \dfrac{1}{180}$

14.9 control variable method most efficient, antithetic variate method marginal improvement over crude Monte Carlo method

14.10 4,107

14.12 $2\mu - X$

14.13 0.5856

14.14 35.6675

14.15 1,000

14.16 41.8971

14.17 224.44

14.18 522.13

14.19 3,047.03

14.20 614.42

Chapter 15

15.1 (a) Denote the population median by θ and the median of $\boldsymbol{x}$ by $\hat{\theta}$. Draw a sample of n observations from $\boldsymbol{x}$ with replacement, and call this sample $\boldsymbol{x}^*$, with median denoted by $\hat{\theta}^*$. Perform the sampling m times to obtain m values of $\hat{\theta}_j^*$, for $j = 1, \cdots, m$. The bias of the sample median is estimated by

$$\frac{1}{m}\left[\sum_{j=1}^{m} \hat{\theta}_j^*\right] - \hat{\theta}.$$

The mean squared error of the sample median is estimated by

$$\frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_j^* - \hat{\theta})^2.$$

(b) Compute the estimate of the variance of $\hat{\theta}^*$ by

$$s^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_j^* - \bar{\theta}^*)^2,$$

where $\bar{\theta}^*$ is the sample mean of $\hat{\theta}_j^*$. The required sample size is $(1.96s/0.05)^2$.

(c) The samples $\boldsymbol{x}^*$ are generated from the $\mathcal{E}(1/\bar{x})$ distribution, where $\bar{x}$ is the sample mean of $\boldsymbol{x}$.

15.2 Draw a sample of n observations from $\boldsymbol{x}$ with replacement, and compute the interquartile range $\hat{\theta}^*$. Perform the sampling m times to obtain m values of $\hat{\theta}_j^*$, for $j = 1, \cdots, m$. Let a and b be the 0.025- and 0.975-quantiles of $\hat{\theta}_j^*/\hat{\theta}$, respectively. The 95% confidence interval of θ is estimated by $(\hat{\theta}/b, \hat{\theta}/a)$.

15.3 (a) Draw a sample of n pairs of observations (x_i^*, y_i^*) from (x_i, y_i) with replacement and compute the sample correlation coefficient $\hat{\rho}^*$. Repeat this m times to obtain $\hat{\rho}_j^*$, for $j = 1, \cdots, m$. The bias of $\hat{\rho}$ is estimated by

$$\frac{1}{m}\left[\sum_{j=1}^{m}\hat{\rho}_j^*\right] - \hat{\rho}.$$

(b) The observations (x_i^*, y_i^*) are simulated from the bivariate standard normal distribution (i.e. $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$) with correlation coefficient $\hat{\rho}$. The correlation coefficient $\hat{\rho}^*$ is computed for each sample, and the formula above is used to compute the bias.

15.4 Compute the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of μ and σ^2. These are, respectively, the sample mean and $(n-1)/n$ times the sample variance of $\log x_i$, for $i = 1, \cdots, n$. Then simulate n observations x_i^* from $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ and compute $(n-1)/n$ times the sample variance, and call this $\hat{\sigma}^{*2}$. Do this m times to obtain $\hat{\sigma}_j^{*2}$ and hence $\hat{\theta}_j^* = (e^{\hat{\sigma}_j^{*2}} - 1)^{\frac{1}{2}}$, for $j = 1, \cdots, m$. The bias of $\hat{\theta}$ is then estimated by taking the sample mean of $\hat{\theta}_j^* - \hat{\theta}$. An alternative

procedure is to note that $n\hat{\sigma}^{*2}$ is distributed as a $\hat{\sigma}^2\chi^2_{n-1}$ variable. Thus, each $\hat{\sigma}^{*2}_i$ may be directly simulated as $\hat{\sigma}^2/n$ times a χ^2_{n-1} variate.

15.5 0.0131

15.6 $\frac{44}{9}$

15.7 8

15.8 88.74

15.9 34.06%

References

Adler, R. J., Feldman, R. E., and Taqqu, M. S. (1998), *A Practical Guide to Heavy Tails*, Birkhäuser.

Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press.

Angus, J. E. (1994), "The probability integral transform and related results," *SIAM Review*, 36, 652–654.

Artzner, P. (1999), "Application of coherent risk measures to capital requirements in insurance," *North American Actuarial Journal*, 3 (2), 11–25.

Artzner, P., Delbaen, F., Eber, J., and Heath, D. (1999), "Coherent measures of risk," *Mathematical Finance*, 9, 203–228.

Bowers, N. L. Jr., Gerber, H. U., Hickman, J. C., Jones, D. A., and Nesbitt, C. J. (1997), *Actuarial Mathematics*, 2nd edition, Society of Actuaries.

Boyle, P., Broadie, M., and Glasserman, P. (1997), "Monte Carlo methods for security pricing," *Journal of Economic Dynamics and Control*, 21, 1267–1321.

David, F. N. and Johnson, N. L. (1948), "The probability integral transformation when parameters are estimated from the sample," *Biometrika*, 35, 182–190.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press.

DeGroot, M. H. and Schervish, M. J. (2002), *Probability and Statistics*, 3rd edition, Addison Wesley.

Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. (2005), *Actuarial Theory for Dependent Risks*: *Measures, Orders and Models*, John Wiley.

De Pril, N. (1985), "Recursions for convolutions of arithmetic distributions," *ASTIN Bulletin*, 15, 135–139.

De Pril, N. (1986), "On the exact computation of the aggregate claims distribution in the individual life model," *ASTIN Bulletin*, 16, 109–112.

Dickson, D. C. M. (2005), *Insurance Risk and Ruin*, Cambridge University Press.

Dowd, K. and Blake, D. (2006), "After VaR: the theory, estimation, and insurance applications of quantile-based risk measures," *Journal of Risk and Insurance*, 73, 193–228.

Hardy, M. (2003), *Investment Guarantees*, John Wiley.

Herzog, T. N. and Lord, G. (2002), *Applications of Monte Carlo Methods to Finance and Insurance*, ACTEX Publications.

Hogg, R. V. and Craig, A. T. (1995), *Introduction to Mathematical Statistics*, 5th edition, Prentice Hall.

Hyndman, R. J. and Fan, Y. (1996), "Sample quantiles in statistical packages," *The American Statistician*, 50, 361–365.

Jewell, W. S. (1974), "Credible means are exact Bayesian for exponential families," *ASTIN Bulletin*, 7, 237–269.

Johnson, N. L. and Kotz, S. (1969), *Distributions in Statistics*: *Discrete Distributions*, John Wiley.

Johnson, N. L. and Kotz, S. (1970), *Distributions in Statistics*: *Continuous Univariate Distributions-I*, John Wiley.

Johnston, J. and DiNardo, J. (1997), *Econometric Methods*, 4th edition, McGraw-Hill.

Jones, B. L., Puri, M. L. and Zitikis, R. (2006), "Testing hypotheses about the equality of several risk measure values with applications in insurance," *Insurance*: *Mathematics and Economics*, 38, 253–270.

Jones, B. L. and Zitikis, R. (2007), "Risk measures, distortion parameters, and their empirical estimation," *Insurance*: *Mathematics and Economics*, 41, 279–297.

Kennedy, W. J. Jr. and Gentle, J. E. (1980), *Statistical Computing*, Marcel Dekker.

Klugman, S. A., Panjer, H. H. and Willmot, G. E. (2004), *Loss Models*: *From Data to Decisions*, 2nd edition, John Wiley.

Lam, J. (2003), *Enterprise Risk Management*: *From Incentives to Controls*, John Wiley.

Lilliefors, H. W. (1967), "On the Kolmogorov–Smirnov test for normality with mean and variance unknown," *Journal of the American Statistical association*, 62, 399–402.

Lilliefors, H. W. (1969), "On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown," *Journal of the American Statistical association*, 64, 387–389.

London, D. (1988), *Survival Models and Their Estimation*, 2nd edition, ACTEX Publications.

McDonald, R. L. (2006), *Derivatives Markets*, 2nd edition, Addison Wesley.

McLeish, D. L. (2005), *Monte Carlo Simulation and Finance*, John Wiley.

McNeil, A. J., Frey, R., and Embrechts, P. (2005), *Quantitative Risk Management*: *Concepts, Techniques, Tools*, Princeton University Press.

Panjer, H. H. (1981), "Recursive evaluation of a family of compound distributions," *ASTIN Bulletin*, 12, 21–26.

Ross, S. (2006), *A First Course in Probability*, 7th edition, Pearson Prentice Hall.

Stephens, M. A. (1974), "EDF statistics for goodness of fit and some comparison," *Journal of the American Statistical Association*, 69, 730–737.

Tan, K. S. and Boyle, P. (2000), "Applications of randomized low discrepancy sequences to the valuation of complex securities," *Journal of Economic Dynamics and Control*, 24, 1747–1782.

Wang, S. S. (2000), "A class of distortion operators for pricing financial and insurance risks," *The Journal of Risk and Insurance*, 67, 15–36.

Wirch, J. and Hardy, M. R. (1999), "A synthesis of risk measures for capital adequacy," *Insurance*: *Mathematics and Economics*, 25, 337–347.

Index