STUART A. KLUGMAN · HARRY H. PANJER

GORDON E. WILLMOT

LOSS MODELS

FROM DATA TO DECISIONS

FIFTH EDITION

with website

SOCIETY OF ACTUARIES®

WILEY

LOSS MODELS

LOSS MODELS

From Data to Decisions

Fifth Edition

Stuart A. Klugman
Society of Actuaries

Harry H. Panjer
University of Waterloo

Gordon E. Willmot
University of Waterloo

SOCIETY OF
ACTUARIES®

WILEY

CONTENTS

Part II   Actuarial Models

Part III   Mathematical Statistics

Part V   Credibility

Part VI   Simulation

PREFACE

The preface to the first edition of this text explained our mission as follows:

> This textbook is organized around the principle that much of actuarial science consists of the construction and analysis of mathematical models that describe the process by which funds flow into and out of an insurance system. An analysis of the entire system is beyond the scope of a single text, so we have concentrated our efforts on the loss process, that is, the outflow of cash due to the payment of benefits.

We have not assumed that the reader has any substantial knowledge of insurance systems. Insurance terms are defined when they are first used. In fact, most of the material could be disassociated from the insurance process altogether, and this book could be just another applied statistics text. What we have done is kept the examples focused on insurance, presented the material in the language and context of insurance, and tried to avoid getting into statistical methods that are not relevant with respect to the problems being addressed.

We will not repeat the evolution of the text over the first four editions but will instead focus on the key changes in this edition. They are:

1. Since the first edition, this text has been a major resource for professional actuarial exams. When the curriculum for these exams changes it is incumbent on us to revise the book accordingly. For exams administered after July 1, 2018, the Society of Actuaries will be using a new syllabus with new learning objectives. Exam C (Construction of Actuarial Models) will be replaced by Exam STAM (Short-Term Actuarial Mathematics). As topics move in and out, it is necessary to adjust the presentation so that candidates who only want to study the topics on their exam can

do so without frequent breaks in the exposition. As has been the case, we continue to include topics not on the exam syllabus that we believe are of interest.

2. The material on nonparametric estimation, such as the Kaplan–Meier estimate, is being moved to the new Exam LTAM (Long-Term Actuarial Mathematics). Therefore, this material and the large sample approximations have been consolidated.

3. The previous editions had not assumed knowledge of mathematical statistics. Hence some of that education was woven throughout. The revised Society of Actuaries requirements now include mathematical statistics as a Validation by Educational Experience (VEE) requirement. Material that overlaps with this subject has been isolated, so exam candidates can focus on material that extends the VEE knowledge.

4. The section on score-based approaches to model selection now includes the Akaike Information Criterion in addition to the Schwarz Bayesian Criterion.

5. Examples and exercises have been added and other clarifications provided where needed.

6. The appendix on numerical optimization and solution of systems of equations has been removed. At the time the first edition was written there were limited options for numerical optimization, particularly for situations with relatively flat surfaces, such as the likelihood function. The simplex method was less well known and worth introducing to readers. Today there are many options and it is unlikely practitioners are writing their own optimization routines.

As in the previous editions, we assume that users will often be doing calculations using a spreadsheet program such as Microsoft Excel®.[1] At various places in the text we indicate how Excel® commands may help. This is not an endorsement by the authors but, rather, a recognition of the pervasiveness of this tool.

As in the first four editions, many of the exercises are taken from examinations of the Society of Actuaries. They have been reworded to fit the terminology and notation of this book and the five answer choices from the original questions are not provided. Such exercises are indicated with an asterisk (*). Of course, these questions may not be representative of those asked on examinations given in the future.

Although many of the exercises either are directly from past professional examinations or are similar to such questions, there are many other exercises meant to provide additional insight into the given subject matter. Consequently, it is recommended that readers interested in particular topics consult the exercises in the relevant sections in order to obtain a deeper understanding of the material.

Many people have helped us through the production of the five editions of this text—family, friends, colleagues, students, readers, and the staff at John Wiley & Sons. Their contributions are greatly appreciated.

S. A. Klugman, H. H. Panjer, and G. E. Willmot

Schaumburg, Illinois; Comox, British Columbia; and Waterloo, Ontario

---

[1]Microsoft® and Excel® are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

ABOUT THE COMPANION WEBSITE

This book is accompanied by a companion website:

    www.wiley.com/go/klugman/lossmodels5e

- Data files to accompany the examples and exercises in Excel and/or comma separated value formats.

PART I

INTRODUCTION

1

MODELING

1.1 The Model-Based Approach

The model-based approach should be considered in the context of the objectives of any given problem. Many problems in actuarial science involve the building of a mathematical model that can be used to forecast or predict insurance costs in the future.

A model is a simplified mathematical description that is constructed based on the knowledge and experience of the actuary combined with data from the past. The data guide the actuary in selecting the form of the model as well as in calibrating unknown quantities, usually called **parameters**. The model provides a balance between simplicity and conformity to the available data.

The simplicity is measured in terms of such things as the number of unknown parameters (the fewer the simpler); the conformity to data is measured in terms of the discrepancy between the data and the model. Model selection is based on a balance between the two criteria, namely, fit and simplicity.

1.1.1 The Modeling Process

The modeling process is illustrated in Figure 1.1, which describes the following six stages:
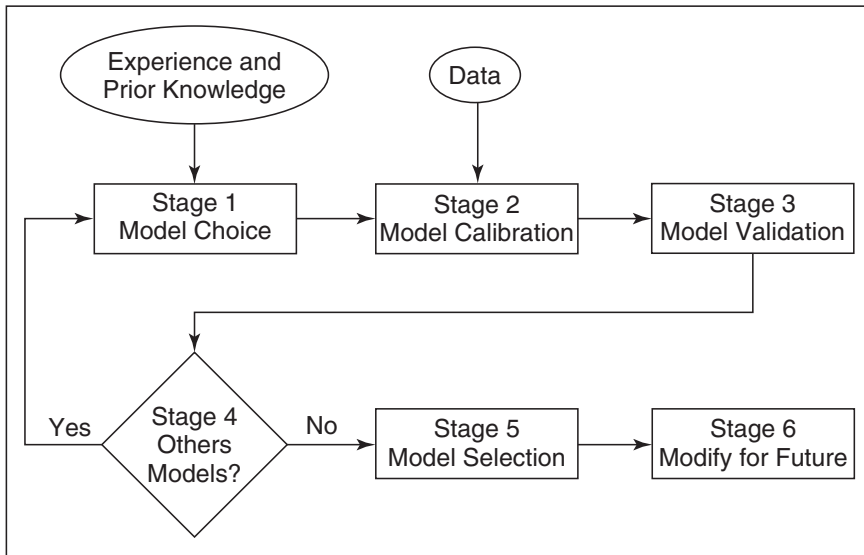
Figure 1.1   The modeling process.

Stage 1 One or more models are selected based on the analyst's prior knowledge and experience, and possibly on the nature and form of the available data. For example, in studies of mortality, models may contain covariate information such as age, sex, duration, policy type, medical information, and lifestyle variables. In studies of the size of an insurance loss, a statistical distribution (e.g. lognormal, gamma, or Weibull) may be chosen.

Stage 2 The model is calibrated based on the available data. In mortality studies, these data may be information on a set of life insurance policies. In studies of property claims, the data may be information about each of a set of actual insurance losses paid under a set of property insurance policies.

Stage 3 The fitted model is validated to determine if it adequately conforms to the data. Various diagnostic tests can be used. These may be well-known statistical tests, such as the chi-square goodness-of-fit test or the Kolmogorov–Smirnov test, or may be more qualitative in nature. The choice of test may relate directly to the ultimate purpose of the modeling exercise. In insurance-related studies, the total loss given by the fitted model is often required to equal the total loss actually experienced in the data. In insurance practice, this is often referred to as **unbiasedness** of a model.

Stage 4 An opportunity is provided to consider other possible models. This is particularly useful if Stage 3 revealed that all models were inadequate. It is also possible that more than one valid model will be under consideration at this stage.

Stage 5 All valid models considered in Stages 1–4 are compared, using some criteria to select between them. This may be done by using the test results previously obtained or it may be done by using another criterion. Once a winner is selected, the losers may be retained for sensitivity analyses.

Stage 6 Finally, the selected model is adapted for application to the future. This could involve adjustment of parameters to reflect anticipated inflation from the time the data were collected to the period of time to which the model will be applied.

As new data are collected or the environment changes, the six stages will need to be repeated to improve the model.

In recent years, actuaries have become much more involved in "big data" problems. Massive amounts of data bring with them challenges that require adaptation of the steps outlined above. Extra care must be taken to avoid building overly complex models that match the data but perform less well when used to forecast future observations. Techniques such as hold-out samples and cross-validation are employed to addresses such issues. These topics are beyond the scope of this book. There are numerous references available, among them [61].

1.1.2 The Modeling Advantage

Determination of the advantages of using models requires us to consider the alternative: decision-making based strictly upon empirical evidence. The empirical approach assumes that the future can be expected to be exactly like a sample from the past, perhaps adjusted for trends such as inflation. Consider Example 1.1.

■ **EXAMPLE 1.1**

A portfolio of group life insurance certificates consists of 1,000 employees of various ages and death benefits. Over the past five years, 14 employees died and received a total of 580,000 in benefits (adjusted for inflation because the plan relates benefits to salary). Determine the empirical estimate of next year's expected benefit payment.

The empirical estimate for next year is then 116,000 (one-fifth of the total), which would need to be further adjusted for benefit increases. The danger, of course, is that it is unlikely that the experience of the past five years will accurately reflect the future of this portfolio, as there can be considerable fluctuation in such short-term results. □

It seems much more reasonable to build a model, in this case a mortality table. This table would be based on the experience of many lives, not just the 1,000 in our group. With this model, not only can we estimate the expected payment for next year, but we can also measure the risk involved by calculating the standard deviation of payments or, perhaps, various percentiles from the distribution of payments. This is precisely the problem covered in texts such as [25] and [28].

This approach was codified by the Society of Actuaries Committee on Actuarial Principles. In the publication "Principles of Actuarial Science" [114, p. 571], Principle 3.1 states that "Actuarial risks can be stochastically modeled based on assumptions regarding the probabilities that will apply to the actuarial risk variables in the future, including assumptions regarding the future environment." The actuarial risk variables referred to are occurrence, timing, and severity – that is, the chances of a claim event, the time at which the event occurs if it does, and the cost of settling the claim.

1.2 The Organization of This Book

This text takes us through the modeling process but not in the order presented in Section 1.1. There is a difference between how models are best applied and how they are best learned. In this text, we first learn about the models and how to use them, and then we learn how to determine which model to use, because it is difficult to select models in a vacuum. Unless the analyst has a thorough knowledge of the set of available models, it is difficult to narrow the choice to the ones worth considering. With that in mind, the organization of the text is as follows:

1. Review of probability – Almost by definition, contingent events imply probability models. Chapters 2 and 3 review random variables and some of the basic calculations that may be done with such models, including moments and percentiles.

2. Understanding probability distributions – When selecting a probability model, the analyst should possess a reasonably large collection of such models. In addition, in order to make a good a priori model choice, the characteristics of these models should be available. In Chapters 4–7, various distributional models are introduced and their characteristics explored. This includes both continuous and discrete distributions.

3. Coverage modifications – Insurance contracts often do not provide full payment. For example, there may be a deductible (e.g. the insurance policy does not pay the first $250) or a limit (e.g. the insurance policy does not pay more than $10,000 for any one loss event). Such modifications alter the probability distribution and affect related calculations such as moments. Chapter 8 shows how this is done.

4. Aggregate losses – To this point, the models are either for the amount of a single payment or for the number of payments. Of interest when modeling a portfolio, line of business, or entire company is the total amount paid. A model that combines the probabilities concerning the number of payments and the amounts of each payment is called an **aggregate loss model**. Calculations for such models are covered in Chapter 9.

5. Introduction to mathematical statistics – Because most of the models being considered are probability models, techniques of mathematical statistics are needed to estimate model specifications and make choices. While Chapters 10 and 11 are not a replacement for a thorough text or course in mathematical statistics, they do contain the essential items that are needed later in this book. Chapter 12 covers estimation techniques for counting distributions, as they are of particular importance in actuarial work.

6. Bayesian methods – An alternative to the frequentist approach to estimation is presented in Chapter 13. This brief introduction introduces the basic concepts of Bayesian methods.

7. Construction of empirical models – Sometimes it is appropriate to work with the empirical distribution of the data. This may be because the volume of data is sufficient or because a good portrait of the data is needed. Chapter 14 covers empirical models for the simple case of straightforward data, adjustments for truncated and censored data, and modifications suitable for large data sets, particularly those encountered in mortality studies.

8. Selection of parametric models – With estimation methods in hand, the final step is to select an appropriate model. Graphic and analytic methods are covered in Chapter 15.

9. Adjustment of estimates – At times, further adjustment of the results is needed. When there are one or more estimates based on a small number of observations, accuracy can be improved by adding other, related observations; care must be taken if the additional data are from a different population. Credibility methods, covered in Chapters 16–18, provide a mechanism for making the appropriate adjustment when additional data are to be included.

10. Simulation – When analytic results are difficult to obtain, simulation (use of random numbers) may provide the needed answer. A brief introduction to this technique is provided in Chapter 19.

2

RANDOM VARIABLES

2.1 Introduction

An actuarial model is a representation of an uncertain stream of future payments. The uncertainty may be with respect to any or all of occurrence (is there a payment?), timing (when is the payment made?), and severity (how much is paid?). Because the most useful means of representing uncertainty is through probability, we concentrate on probability models. For now, the relevant probability distributions are assumed to be known. The determination of appropriate distributions is covered in Chapters 10 through 15. In this part, the following aspects of actuarial probability models are covered:

1. Definition of random variable and important functions, with some examples.

2. Basic calculations from probability models.

3. Specific probability distributions and their properties.

4. More advanced calculations using severity models.

5. Models incorporating the possibility of a random number of payments, each of random amount.

The commonality we seek here is that all models for random phenomena have similar elements. For each, there is a set of possible outcomes. The particular outcome that occurs will determine the success of our enterprise. Attaching probabilities to the various outcomes allows us to quantify our expectations and the risk of not meeting them. In this spirit, the underlying random variable will almost always be denoted with uppercase italic letters near the end of the alphabet, such as X or Y. The context will provide a name and some likely characteristics. Of course, there are actuarial models that do not look like those covered here. For example, in life insurance a **model office** is a list of cells containing policy type, age range, gender, and so on, along with the number of contracts with those characteristics.

To expand on this concept, consider the following definitions from "Principles Underlying Actuarial Science" [5, p. 7]:

> *Phenomena* are occurrences that can be observed. An *experiment* is an observation of a given phenomenon under specified conditions. The result of an experiment is called an *outcome*; an *event* is a set of one or more possible outcomes. A *stochastic phenomenon* is a phenomenon for which an associated experiment has more than one possible outcome. An event associated with a stochastic phenomenon is said to be *contingent*. ... *Probability* is a measure of the likelihood of the occurrence of an event, measured on a scale of increasing likelihood from zero to one. ... A *random variable* is a function that assigns a numerical value to every possible outcome.

The following list contains 12 random variables that might be encountered in actuarial work **(Model #** refers to examples introduced in the next section):

1. The age at death of a randomly selected birth. **(Model 1)**

2. The time to death from when insurance was purchased for a randomly selected insured life.

3. The time from occurrence of a disabling event to recovery or death for a randomly selected workers compensation claimant.

4. The time from the incidence of a randomly selected claim to its being reported to the insurer.

5. The time from the reporting of a randomly selected claim to its settlement.

6. The number of dollars paid on a randomly selected life insurance claim.

7. The number of dollars paid on a randomly selected automobile bodily injury claim. **(Model 2)**

8. The number of automobile bodily injury claims in one year from a randomly selected insured automobile. **(Model 3)**

9. The total dollars in medical malpractice claims paid in one year owing to events at a randomly selected hospital. (**Model 4**)

10. The time to default or prepayment on a randomly selected insured home loan that terminates early.

11. The amount of money paid at maturity on a randomly selected high-yield bond.

12. The value of a stock index on a specified future date.

Because all of these phenomena can be expressed as random variables, the machinery of probability and mathematical statistics is at our disposal both to create and to analyze models for them. The following paragraphs discuss five key functions used in describing a random variable: cumulative distribution, survival, probability density, probability mass, and hazard rate. They are illustrated with four ongoing models as identified in the preceding list plus one more to be introduced later.

2.2   Key Functions and Four Models

Definition 2.1 *The **cumulative distribution function**, also called the **distribution function** and usually denoted $F_X(x)$ or $F(x)$,[1] for a random variable X is the probability that X is less than or equal to a given number. That is, $F_X(x) = \Pr(X \leq x)$. The abbreviation **cdf** is often used.*

The distribution function must satisfy a number of requirements:[2]

- $0 \leq F(x) \leq 1$ for all x.

- $F(x)$ is nondecreasing.

- $F(x)$ is right-continuous.[3]

- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

Because it need not be left-continuous, it is possible for the distribution function to jump. When it jumps, the value is assigned to the top of the jump.

Here are possible distribution functions for each of the four models.

Model 1[4] This random variable could serve as a model for the age at death. All ages between 0 and 100 are possible. While experience suggests that there is an upper bound for human lifetime, models with no upper limit may be useful if they assign extremely low probabilities to extreme ages. This allows the modeler to avoid setting a specific maximum age:

$$F_1(x) = \begin{cases} 0, & x < 0, \\ 0.01x, & 0 \leq x < 100, \\ 1, & x \geq 100. \end{cases}$$

This cdf is illustrated in Figure 2.1. ☐

Model 2 This random variable could serve as a model for the number of dollars paid on an automobile insurance claim. All positive values are possible. As with mortality, there is

---

[1]When denoting functions associated with random variables, it is common to identify the random variable through a subscript on the function. Here, subscripts are used only when needed to distinguish one random variable from another. In addition, for the five models to be introduced shortly, rather than write the distribution function for random variable 2 as $F_{X_2}(x)$, it is simply denoted $F_2(x)$.

[2]The first point follows from the last three.

[3]Right-continuous means that at any point x_0 the limiting value of $F(x)$ as x approaches x_0 from the right is equal to $F(x_0)$. This need not be true as x approaches x_0 from the left.

[4]The five models (four introduced here and one later) are identified by the numbers 1–5. Other examples use the traditional numbering scheme as used for definitions and the like.
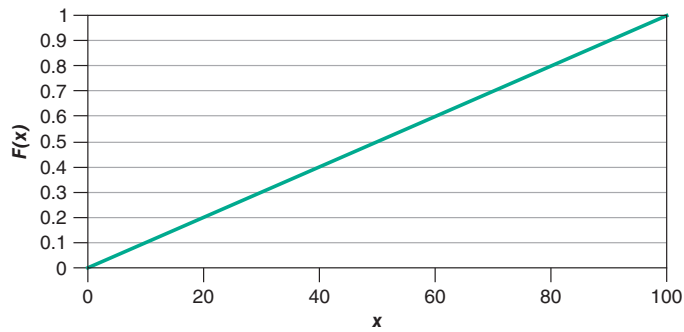
Figure 2.1 The distribution function for Model 1.
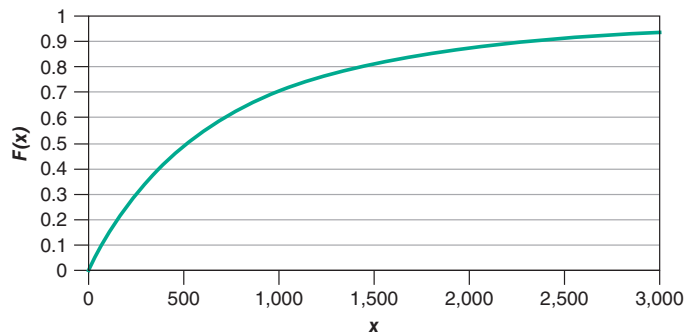
Figure 2.2 The distribution function for Model 2.

likely an upper limit (all the money in the world comes to mind), but this model illustrates that, in modeling, correspondence to reality need not be perfect:

$$F_2(x) = \begin{cases} 0, & x < 0, \\ 1 - \left(\dfrac{2,000}{x + 2,000} \right)^3, & x \geq 0. \end{cases}$$

This cdf is illustrated in Figure 2.2. □

Model 3 This random variable could serve as a model for the number of claims on one policy in one year. Probability is concentrated at the five points $(0, 1, 2, 3, 4)$ and the probability at each is given by the size of the jump in the distribution function:

$$F_3(x) = \begin{cases} 0, & x < 0, \\ 0.5, & 0 \leq x < 1, \\ 0.75, & 1 \leq x < 2, \\ 0.87, & 2 \leq x < 3, \\ 0.95, & 3 \leq x < 4, \\ 1, & x \geq 4. \end{cases}$$

While this model places a maximum on the number of claims, models with no limit (such as the Poisson distribution) could also be used. □

Model 4 This random variable could serve as a model for the total dollars paid on a medical malpractice policy in one year. Most of the probability is at zero (0.7) because in most years nothing is paid. The remaining 0.3 of probability is distributed over positive values:

$$F_4(x) = \begin{cases} 0, & x < 0, \\ 1 - 0.3e^{-0.00001x}, & x \geq 0. \end{cases}$$

□

Definition 2.2 *The **support** of a random variable is the set of numbers that are possible values of the random variable.*

Definition 2.3 *A random variable is called **discrete** if the support contains at most a countable number of values. It is called **continuous** if the distribution function is continuous and is differentiable everywhere with the possible exception of a countable number of values. It is called **mixed** if it is not discrete and is continuous everywhere with the exception of at least one value and at most a countable number of values.*

These three definitions do not exhaust all possible random variables but will cover all cases encountered in this book. The distribution function for a discrete random variable will be constant except for jumps at the values with positive probability. A mixed distribution will have at least one jump. Requiring continuous variables to be differentiable allows the variable to have a density function (defined later) at almost all values.

■ **EXAMPLE 2.1**

For each of the four models, determine the support and indicate which type of random variable it is.

The distribution function for Model 1 is continuous and is differentiable except at 0 and 100, and therefore is a continuous distribution. The support is values from 0 to 100 with it not being clear if 0 or 100 are included.[5] The distribution function for Model 2 is continuous and is differentiable except at 0, and therefore is a continuous distribution. The support is all positive real numbers and perhaps 0. The random variable for Model 3 places probability only at 0, 1, 2, 3, and 4 (the support) and thus is discrete. The distribution function for Model 4 is continuous except at 0, where it jumps. It is a mixed distribution with support on nonnegative real numbers. □

These four models illustrate the most commonly encountered forms of the distribution function. Often in the remainder of the book, when functions are presented, values outside the support are not given (most commonly where the distribution and survival functions are 0 or 1).

---

[5]The reason it is not clear is that the underlying random variable is not described. Suppose that Model 1 represents the percentage of value lost on a randomly selected house after a hurricane. Then 0 and 100 are both possible values and are included in the support. It turns out that a decision regarding including endpoints in the support of a continuous random variable is rarely needed. If there is no clear answer, an arbitrary choice can be made.

Definition 2.4 *The **survival function**, usually denoted $S_X(x)$ or $S(x)$, for a random variable X is the probability that X is greater than a given number. That is, $S_X(x) = \Pr(X > x) = 1 - F_X(x)$.*

As a result:

- $0 \le S(x) \le 1$ for all x.

- $S(x)$ is nonincreasing.

- $S(x)$ is right-continuous.

- $\lim_{x \to -\infty} S(x) = 1$ and $\lim_{x \to \infty} S(x) = 0$.

Because the survival function need not be left-continuous, it is possible for it to jump (down). When it jumps, the value is assigned to the bottom of the jump.

The survival function is the complement of the distribution function, and thus knowledge of one implies knowledge of the other. Historically, when the random variable is measuring time, the survival function is presented, while when it is measuring dollars, the distribution function is presented.

■ **EXAMPLE 2.2**

For completeness, here are the survival functions for the four models:

$$S_1(x) = 1 - 0.01x, \quad 0 \le x < 100,$$

$$S_2(x) = \left(\frac{2,000}{x + 2,000} \right)^3, \quad x \ge 0,$$

$$S_3(x) = \begin{cases} 0.5, & 0 \le x < 1, \\ 0.25, & 1 \le x < 2, \\ 0.13, & 2 \le x < 3, \\ 0.05, & 3 \le x < 4, \\ 0, & x \ge 4, \end{cases}$$

$$S_4(x) = 0.3e^{-0.00001x}, \quad x \ge 0.$$

The survival functions for Models 1 and 2 are illustrated in Figures 2.3 and 2.4.   □

Either the distribution or the survival function can be used to determine probabilities. Let $F(b-) = \lim_{x \nearrow b} F(x)$ and let $S(b-)$ be similarly defined. That is, we want the limit as x approaches b from below. We have $\Pr(a < X \le b) = F(b) - F(a) = S(a) - S(b)$ and $\Pr(X = b) = F(b) - F(b-) = S(b-) - S(b)$. When the distribution function is continuous at x, $\Pr(X = x) = 0$; otherwise, the probability is the size of the jump. The next two functions are more directly related to the probabilities. The first is for continuous distributions, the second for discrete distributions.

Definition 2.5 *The **probability density function**, also called the **density function** and usually denoted $f_X(x)$ or $f(x)$, is the derivative of the distribution function or, equivalently,*
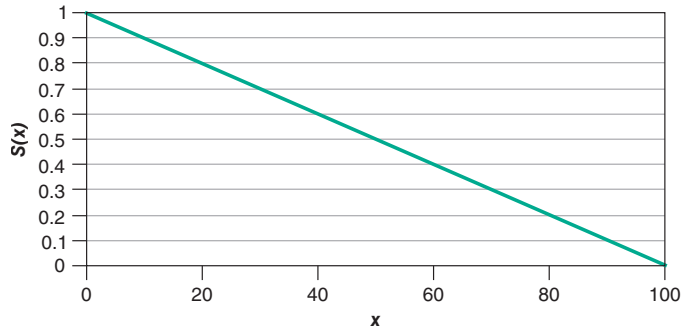
Figure 2.3   The survival function for Model 1.
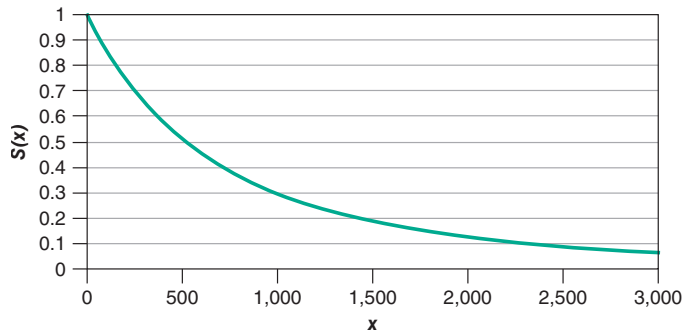
Figure 2.4   The survival function for Model 2.

*the negative of the derivative of the survival function. That is, $f(x) = F'(x) = -S'(x)$. The density function is defined only at those points where the derivative exists. The abbreviation **pdf** is often used.*

While the density function does not directly provide probabilities, it does provide relevant information. Values of the random variable in regions with higher density values are more likely to occur than those in regions with lower values. Probabilities for intervals and the distribution and survival functions can be recovered by integration. That is, when the density function is defined over the relevant interval, $\Pr(a < X \le b) = \int_a^b f(x)\,dx$, $F(b) = \int_{-\infty}^b f(x)\,dx$, and $S(b) = \int_b^\infty f(x)\,dx$.

■ **EXAMPLE 2.3**

For our models,

$$f_1(x) = 0.01, \quad 0 < x < 100,$$

$$f_2(x) = \frac{3(2{,}000)^3}{(x + 2{,}000)^4}, \quad x > 0,$$

$$f_3(x) \text{ is not defined,}$$

$$f_4(x) = 0.000003e^{-0.00001x}, \quad x > 0.$$

It should be noted that for Model 4 the density function does not completely describe the probability distribution. As a mixed distribution, there is also discrete probability at 0. The density functions for Models 1 and 2 are illustrated in Figures 2.5 and 2.6. □
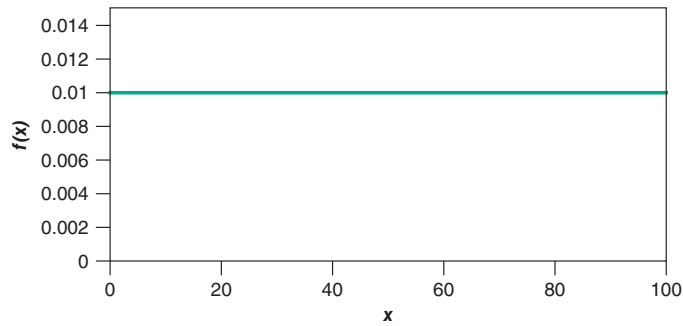
Figure 2.5 The density function for Model 1.
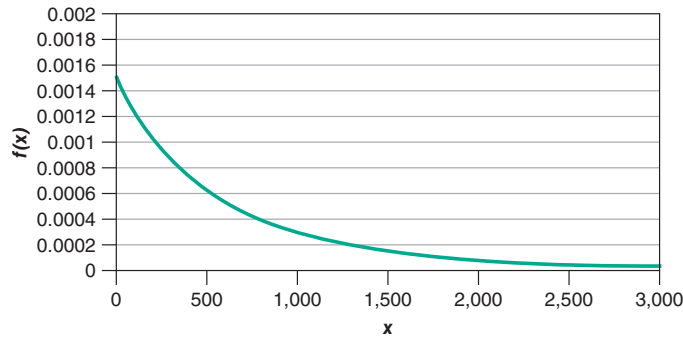
Figure 2.6 The density function for Model 2.

Definition 2.6 *The* **probability function,** *also called the* **probability mass function** *and usually denoted* $p_X(x)$ *or* $p(x)$, *describes the probability at a distinct point when it is not* 0. *The formal definition is* $p_X(x) = \Pr(X = x)$.

For discrete random variables, the distribution and survival functions can be recovered as $F(x) = \sum_{y \leq x} p(y)$ and $S(x) = \sum_{y > x} p(y)$.

■ **EXAMPLE 2.4**

For our models,

$$p_1(x) \text{ is not defined,}$$

$$p_2(x) \text{ is not defined,}$$

$$p_3(x) = \begin{cases} 0.50, & x = 0, \\ 0.25, & x = 1, \\ 0.12, & x = 2, \\ 0.08, & x = 3, \\ 0.05, & x = 4, \end{cases}$$

$$p_4(0) = 0.7.$$

It is again noted that the distribution in Model 4 is mixed, so the preceding describes only the discrete portion of that distribution. There is no easy way to present probabilities/densities for a mixed distribution. For Model 4, we would present the probability density function as

$$f_4(x) = \begin{cases} 0.7, & x = 0, \\ 0.000003e^{-0.00001x}, & x > 0, \end{cases}$$

realizing that, technically, it is not a probability density function at all. When the density function is assigned a value at a specific point, as opposed to being defined on an interval, it is understood to be a discrete probability mass.   □

Definition 2.7 *The* **hazard rate***, also known as the* **force of mortality** *and the* **failure rate** *and usually denoted* $h_X(x)$ *or* $h(x)$*, is the ratio of the density and survival functions when the density function is defined. That is,* $h_X(x) = f_X(x)/S_X(x)$*.*

When called the force of mortality, the hazard rate is often denoted $\mu(x)$, and when called the failure rate, it is often denoted $\lambda(x)$. Regardless, it may be interpreted as the probability density at x given that the argument will be at least x. We also have $h_X(x) = -S'(x)/S(x) = -d \ln S(x)/dx$. The survival function can be recovered from $S(b) = e^{-\int_0^b h(x)\,dx}$. Though not necessary, this formula implies that the support is on nonnegative numbers. In mortality terms, the force of mortality is the annualized probability that a person age x will die in the next instant, expressed as a death rate per year.[6] In this text, we always use $h(x)$ to denote the hazard rate, although one of the alternative names may be used.

■ **EXAMPLE 2.5**

For our models,

$$h_1(x) = \frac{0.01}{1 - 0.01x}, \quad 0 < x < 100,$$
$$h_2(x) = \frac{3}{x + 2{,}000}, \quad x > 0,$$
$$h_3(x) \text{ is not defined,}$$
$$h_4(x) = 0.00001, \quad x > 0.$$

Once again, note that for the mixed distribution the hazard rate is only defined over part of the random variable's support. This is different from the preceding problem where both a probability density function and a probability function are involved. Where there is a discrete probability mass, the hazard rate is not defined. The hazard rate functions for Models 1 and 2 are illustrated in Figures 2.7 and 2.8.   □

The following model illustrates a situation in which there is a point where the density and hazard rate functions are not defined.

---

[6]Note that the force of mortality is not a probability (in particular, it can be greater than 1), although it does no harm to visualize it as a probability.
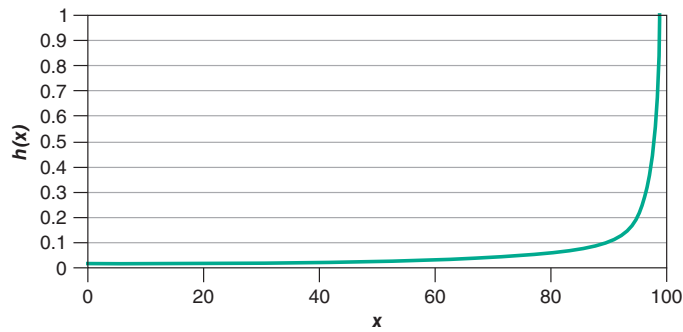
Figure 2.7   The hazard rate function for Model 1.
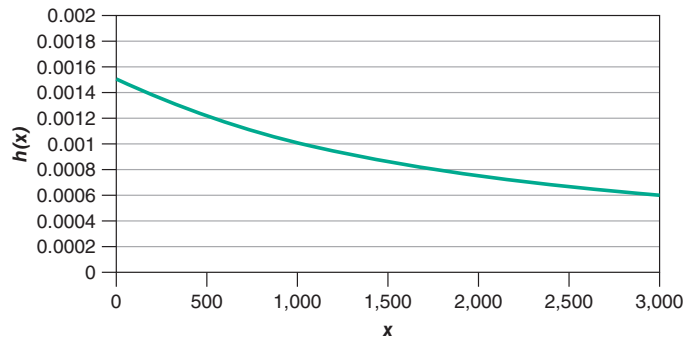
Figure 2.8   The hazard rate function for Model 2.

Model 5 An alternative to the simple lifetime distribution in Model 1 is given here. Note that it is piecewise linear and the derivative at 50 is not defined. Therefore, neither the density function nor the hazard rate function is defined at 50. Unlike the mixed model of Model 4, there is no discrete probability mass at this point. Because the probability of 50 occurring is zero, the density or hazard rate at 50 could be arbitrarily defined with no effect on subsequent calculations. In this book, such values are arbitrarily defined so that the function is right-continuous.[7] For an example, see the solution to Exercise 2.1.

$$S_5(x) = \begin{cases} 1 - 0.01x, & 0 \le x < 50, \\ 1.5 - 0.02x, & 50 \le x < 75. \end{cases}$$ ☐

A variety of commonly used continuous distributions are presented in Appendix A and many discrete distributions are presented in Appendix B.

An interesting feature of a random variable is the value that is most likely to occur.

Definition 2.8 *The **mode** of a random variable is the most likely value. For a discrete variable, it is the value with the largest probability. For a continuous variable, it is the*

---

[7]By arbitrarily defining the value of the density or hazard rate function at such a point, it is clear that using either of them to obtain the survival function will work. If there is discrete probability at this point (in which case these functions are left undefined), then the density and hazard functions are not sufficient to completely describe the probability distribution.

value for which the density function is largest. If there are local maxima, these points are also considered to be modes.

■ **EXAMPLE 2.6**

Where possible, determine the mode for Models 1–5.

   For Model 1, the density function is constant. All values from 0 to 100 could be the mode or, equivalently, it could be said that there is no mode. For Model 2, the density function is strictly decreasing and so the mode is at 0. For Model 3, the probability is highest at 0. As a mixed distribution, it is not possible to define a mode for Model 4. Model 5 has a density that is constant over two intervals, with higher values from 50 to 75. These values are all modes. □

2.2.1 Exercises

2.1 Determine the distribution, density, and hazard rate functions for Model 5.

2.2 Construct graphs of the distribution function for Models 3, 4, and 5. Also graph the density or probability function as appropriate and the hazard rate function, where it exists.

2.3 (*) A random variable X has density function $f(x) = 4x(1+x^2)^{-3}$, $x > 0$. Determine the mode of X.

2.4 (*) A nonnegative random variable has a hazard rate function of $h(x) = A+e^{2x}$, $x \geq 0$. You are also given $S(0.4) = 0.5$. Determine the value of A.

2.5 (*) X has a Pareto distribution with parameters $\alpha = 2$ and $\theta = 10,000$. Y has a Burr distribution with parameters $\alpha = 2$, $\gamma = 2$, and $\theta = \sqrt{20,000}$. Let r be the ratio of $\Pr(X > d)$ to $\Pr(Y > d)$. Determine $\lim_{d \to \infty} r$.

3

BASIC DISTRIBUTIONAL QUANTITIES

3.1 Moments

There are a variety of interesting calculations that can be done from the models described in Chapter 2. Examples are the average amount paid on a claim that is subject to a deductible or policy limit or the average remaining lifetime of a person age 40.

Definition 3.1 *The **kth raw moment** of a random variable is the expected (average) value of the kth power of the variable, provided that it exists. It is denoted by $\mathrm{E}(X^k)$ or by μ'_k. The first raw moment is called the **mean** of the random variable and is usually denoted by μ.*

Note that μ is not related to $\mu(x)$, the force of mortality from Definition 2.7. For random variables that take on only nonnegative values (i.e. $\Pr(X \geq 0) = 1$), k may be any real number. When presenting formulas for calculating this quantity, a distinction between continuous and discrete variables needs to be made. Formulas will be presented for random variables that are either everywhere continuous or everywhere discrete. For mixed models, evaluate the formula by integrating with respect to its density function wherever the random variable is continuous, and by summing with respect to its probability function wherever

the random variable is discrete and adding the results. The formula for the kth raw moment is

$$\mu'_k = \mathrm{E}(X^k) = \int_{-\infty}^{\infty} x^k f(x)\,dx \quad \text{if the random variable is continuous}$$

$$= \sum_j x_j^k p(x_j) \qquad \text{if the random variable is discrete,} \qquad (3.1)$$

where the sum is to be taken over all x_j with positive probability. Finally, note that it is possible that the integral or sum will not converge, in which case the moment is said not to exist.

■ **EXAMPLE 3.1**

Determine the first two raw moments for each of the five models.

The subscripts on the random variable X indicate which model is being used.

$$\mathrm{E}(X_1) = \int_0^{100} x(0.01)\,dx = 50,$$

$$\mathrm{E}(X_1^2) = \int_0^{100} x^2(0.01)\,dx = 3{,}333.33,$$

$$\mathrm{E}(X_2) = \int_0^{\infty} x\frac{3(2{,}000)^3}{(x+2{,}000)^4}\,dx = 1{,}000,$$

$$\mathrm{E}(X_2^2) = \int_0^{\infty} x^2\frac{3(2{,}000)^3}{(x+2{,}000)^4}\,dx = 4{,}000{,}000,$$

$$\mathrm{E}(X_3) = 0(0.5) + 1(0.25) + 2(0.12) + 3(0.08) + 4(0.05) = 0.93,$$

$$\mathrm{E}(X_3^2) = 0(0.5) + 1(0.25) + 4(0.12) + 9(0.08) + 16(0.05) = 2.25,$$

$$\mathrm{E}(X_4) = 0(0.7) + \int_0^{\infty} x(0.000003)e^{-0.00001x}\,dx = 30{,}000,$$

$$\mathrm{E}(X_4^2) = 0^2(0.7) + \int_0^{\infty} x^2(0.000003)e^{-0.00001x}\,dx = 6{,}000{,}000{,}000,$$

$$\mathrm{E}(X_5) = \int_0^{50} x(0.01)\,dx + \int_{50}^{75} x(0.02)\,dx = 43.75,$$

$$\mathrm{E}(X_5^2) = \int_0^{50} x^2(0.01)\,dx + \int_{50}^{75} x^2(0.02)\,dx = 2{,}395.83.$$

□

Definition 3.2 *The **kth central moment** of a random variable is the expected value of the kth power of the deviation of the variable from its mean. It is denoted by $\mathrm{E}[(X-\mu)^k]$ or by μ_k. The second central moment is usually called the **variance** and denoted σ^2 or $\mathrm{Var}(X)$, and its square root, σ, is called the **standard deviation**. The ratio of the standard deviation to the mean is called the **coefficient of variation**. The ratio of the third central moment to the cube of the standard deviation, $\gamma_1 = \mu_3/\sigma^3$, is called the **skewness**. The ratio of the*

fourth central moment to the fourth power of the standard deviation, $\gamma_2 = \mu_4/\sigma^4$, is called the **kurtosis**.[1]

The continuous and discrete formulas for calculating central moments are

$$\mu_k = \mathrm{E}[(X - \mu)^k]$$

$$= \int_{-\infty}^{\infty} (x - \mu)^k f(x)\, dx \quad \text{if the random variable is continuous}$$

$$= \sum_j (x_j - \mu)^k p(x_j) \qquad \text{if the random variable is discrete.} \qquad (3.2)$$

In reality, the integral needs be taken only over those x values where $f(x)$ is positive. The standard deviation is a measure of how much the probability is spread out over the random variable's possible values. It is measured in the same units as the random variable itself. The coefficient of variation measures the spread relative to the mean. The skewness is a measure of asymmetry. A symmetric distribution has a skewness of zero, while a positive skewness indicates that probabilities to the right tend to be assigned to values further from the mean than those to the left. The kurtosis measures flatness of the distribution relative to a normal distribution (which has a kurtosis of 3).[2] Kurtosis values above 3 indicate that (keeping the standard deviation constant), relative to a normal distribution, more probability tends to be at points away from the mean than at points near the mean. The coefficients of variation, skewness, and kurtosis are all dimensionless.

There is a link between raw and central moments. The following equation indicates the connection between second moments. The development uses the continuous version from (3.1) and (3.2), but the result applies to all random variables:

$$\mu_2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx = \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f(x)\, dx$$

$$= \mathrm{E}(X^2) - 2\mu \mathrm{E}(X) + \mu^2 = \mu_2' - \mu^2. \qquad (3.3)$$

■ **EXAMPLE 3.2**

The density function of the gamma distribution appears to be positively skewed. Demonstrate that this is true and illustrate with graphs.

From Appendix A, the first three raw moments of the gamma distribution are $\alpha\theta$, $\alpha(\alpha + 1)\theta^2$, and $\alpha(\alpha + 1)(\alpha + 2)\theta^3$. From (3.3) the variance is $\alpha\theta^2$, and from the solution to Exercise 3.1 the third central moment is $2\alpha\theta^3$. Therefore, the skewness is $2\alpha^{-1/2}$. Because α must be positive, the skewness is always positive. Also, as α decreases, the skewness increases.

Consider the following two gamma distributions. One has parameters $\alpha = 0.5$ and $\theta = 100$ while the other has $\alpha = 5$ and $\theta = 10$. These have the same mean, but their skewness coefficients are 2.83 and 0.89, respectively. Figure 3.1 demonstrates the difference. ☐

---

[1]It would be more accurate to call these items the "coefficient of skewness" and "coefficient of kurtosis" because there are other quantities that also measure asymmetry and flatness. The simpler expressions are used in this text.
[2]Because of this, an alternative definition of kurtosis has 3 subtracted from our definition, giving the normal distribution a kurtosis of zero, which can be used as a convenient benchmark.
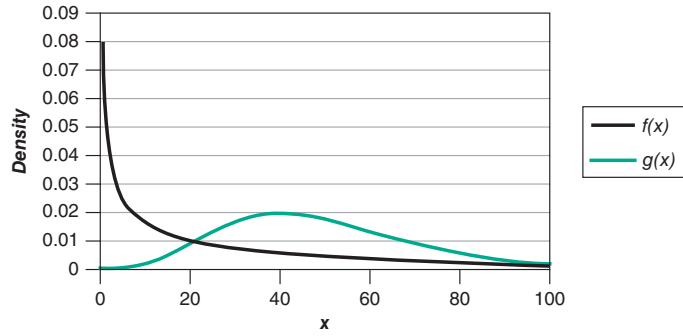
Figure 3.1   The densities of $f(x) \sim \text{gamma}(0.5, 100)$ and $g(x) \sim \text{gamma}(5, 10)$.

Finally, when calculating moments, it is possible that the integral or sum will not exist (as is the case for the third and fourth moments for Model 2). For the models that we typically encounter, the integrand and summand are nonnegative, and so failure to exist implies that the limit of the integral or sum is infinity. For an illustration, see Example 3.9.

Definition 3.3 *For a given value of d with* $\Pr(X > d) > 0$*, the **excess loss variable** is* $Y^P = X - d$*, given that* $X > d$*. Its expected value,*

$$e_X(d) = e(d) = \text{E}(Y^P) = \text{E}(X - d \mid X > d),$$

*is called the **mean excess loss function**. Other names for this expectation are **mean residual life function** and **complete expectation of life**. When the latter terminology is used, the commonly used symbol is* $\overset{\circ}{e}_d$*.*

This variable could also be called a **left truncated and shifted variable**. It is left truncated because any values of X below d are not observed. It is shifted because d is subtracted from the remaining values. When X is a payment variable, the mean excess loss is the expected amount paid, given that there has been a payment in excess of a deductible of d.[3] When X is the age at death, the mean excess loss is the expected remaining time until death, given that the person is alive at age d. The kth moment of the excess loss variable is determined from

$$e_X^k(d) = \frac{\int_d^\infty (x - d)^k f(x)\, dx}{1 - F(d)} \qquad \text{if the variable is continuous}$$

$$= \frac{\sum_{x_j > d} (x_j - d)^k p(x_j)}{1 - F(d)} \qquad \text{if the variable is discrete.} \qquad (3.4)$$

Here, $e_X^k(d)$ is defined only if the integral or sum converges. There is a particularly convenient formula for calculating the first moment. The development given below is for a continuous random variable, but the result holds for all types of random variables.

---

[3]This provides the meaning of the superscript P, indicating that this payment is per payment. It is made to distinguish this variable from Y^L, the per-loss variable to be introduced shortly. These two variables are explored in depth in Chapter 8.

The second line is based on an integration by parts, where the antiderivative of $f(x)$ is taken as $-S(x)$:

$$e_X(d) = \frac{\int_d^\infty (x-d)f(x)\,dx}{1-F(d)}$$

$$= \frac{-(x-d)S(x)|_d^\infty + \int_d^\infty S(x)\,dx}{S(d)}$$

$$= \frac{\int_d^\infty S(x)\,dx}{S(d)}. \tag{3.5}$$

Definition 3.4 *The **left censored and shifted variable** is*

$$Y^L = (X-d)_+ = \begin{cases} 0, & X \le d, \\ X-d, & X > d. \end{cases}$$

It is left censored because values below d are not ignored but are set equal to zero. There is no standard name or symbol for the moments of this variable. For dollar events, the distinction between the excess loss variable and the left censored and shifted variable is one of **per payment** versus **per loss**. In the per-payment situation, the variable exists only when a payment is made. The per-loss variable takes on the value zero whenever a loss produces no payment. The moments can be calculated from

$$\mathrm{E}[(X-d)_+^k] = \int_d^\infty (x-d)^k f(x)\,dx \quad \text{if the variable is continuous.}$$

$$= \sum_{x_j > d} (x_j - d)^k p(x_j) \quad \text{if the variable is discrete.} \tag{3.6}$$

It should be noted that
$$\mathrm{E}[(X-d)_+^k] = e^k(d)[1-F(d)]. \tag{3.7}$$

■ **EXAMPLE 3.3**

Construct graphs to illustrate the difference between the excess loss variable and the left censored and shifted variable.

The two graphs in Figures 3.2 and 3.3 plot the modified variable Y as a function of the unmodified variable X. The only difference is that for X values below 100 the variable is undefined, while for the left censored and shifted variable it is set equal to zero. □

These concepts are most easily demonstrated with a discrete random variable.

■ **EXAMPLE 3.4**

An automobile insurance policy with no coverage modifications has the following possible losses, with probabilities in parentheses: 100 (0.4), 500 (0.2), 1,000 (0.2), 2,500 (0.1), and 10,000 (0.1). Determine the probability mass functions and expected
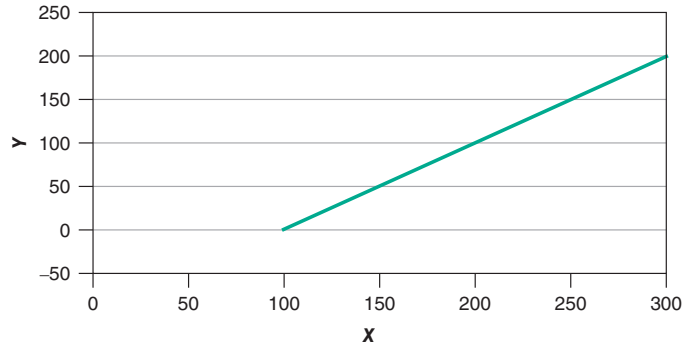
Figure 3.2 The excess loss variable.
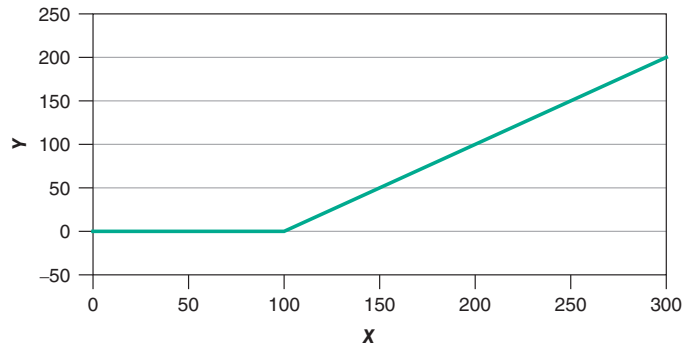
Figure 3.3 A left censored and shifted variable.

values for the excess loss and left censored and shifted variables, where the deductible is set at 750.

For the excess loss variable, 750 is subtracted from each possible loss above that value. Thus the possible values for this random variable are 250, 1,750, and 9,250. The conditional probabilities are obtained by dividing each of the three probabilities by 0.4 (the probability of exceeding the deductible). They are 0.5, 0.25, and 0.25, respectively. The expected value is $250(0.5) + 1,750(0.25) + 9,250(0.25) = 2,875$.

For the left censored and shifted variable, the probabilities that had been assigned to values below 750 are now assigned to zero. The other probabilities are unchanged, but the values they are assigned to are reduced by the deductible. The probability mass function is 0 (0.6), 250 (0.2), 1,750 (0.1), and 9,750 (0.1). The expected value is $0(0.6) + 250(0.2) + 1,750(0.1) + 9,250(0.1) = 1,150$. As noted in (3.7), the ratio of the two expected values is the probability of exceeding the deductible.

Another way to understand the difference in these expected values is to consider 10 accidents with losses conforming exactly to the above distribution. Only four of the accidents produce payments, and multiplying by the expected payment per payment gives a total of $4(2,875) = 11,500$ expected to be paid by the company. Or, consider that the 10 accidents each have an expected payment of 1,150 per loss (accident) for a total expected value of 11,500. Therefore, what is important is not the variable being used but, rather, that it be used appropriately. ☐
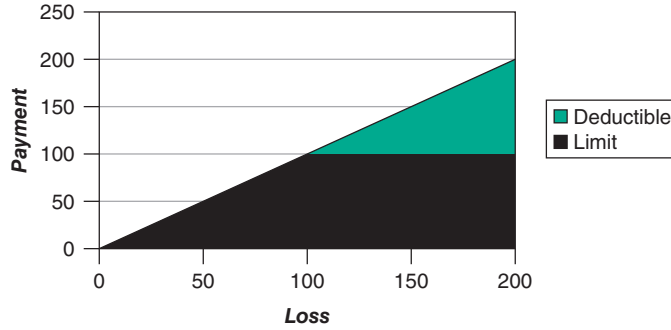
Figure 3.4 A limit of 100 plus a deductible of 100 equals full coverage.

The next definition provides a complementary variable to the excess loss variable.

Definition 3.5 *The **limited loss variable** is*

$$Y = X \wedge u = \begin{cases} X, & X < u, \\ u, & X \geq u. \end{cases}$$

*Its expected value, $E(X \wedge u)$, is called the **limited expected value**.*

This variable could also be called the **right censored variable**. It is right censored because values above u are set equal to u. An insurance phenomenon that relates to this variable is the existence of a policy limit that sets a maximum on the benefit to be paid. Note that $(X - d)_+ + (X \wedge d) = X$. That is, buying one insurance policy with a limit of d and another with a deductible of d is equivalent to buying full coverage. This is illustrated in Figure 3.4.

The most direct formulas for the kth moment of the limited loss variable are

$$E[(X \wedge u)^k] = \int_{-\infty}^{u} x^k f(x)\, dx + u^k [1 - F(u)]$$

if the random variable is continuous.

$$= \sum_{x_j \leq u} x_j^k p(x_j) + u^k [1 - F(u)]$$

if the random variable is discrete. (3.8)

Another interesting formula is derived as follows:

$$\begin{aligned} E[(X \wedge u)^k] &= \int_{-\infty}^{0} x^k f(x)\, dx + \int_{0}^{u} x^k f(x)\, dx + u^k [1 - F(u)] \\ &= x^k F(x)\big|_{-\infty}^{0} - \int_{-\infty}^{0} k x^{k-1} F(x)\, dx \\ &\quad - x^k S(x)\big|_{0}^{u} + \int_{0}^{u} k x^{k-1} S(x)\, dx + u^k S(u) \\ &= -\int_{-\infty}^{0} k x^{k-1} F(x)\, dx + \int_{0}^{u} k x^{k-1} S(x)\, dx, \end{aligned} \qquad (3.9)$$

where the second line uses integration by parts. For $k = 1$, we have

$$E(X \wedge u) = -\int_{-\infty}^{0} F(x)\,dx + \int_{0}^{u} S(x)\,dx.$$

The corresponding formula for discrete random variables is not particularly interesting. The limited expected value also represents the expected dollar saving per incident when a deductible is imposed. The kth limited moment of many common continuous distributions is presented in Appendix A. Exercise 3.8 asks you to develop a relationship between the three first moments introduced previously.

■ **EXAMPLE 3.5**

(*Example 3.4 continued*) Calculate the probability function and the expected value of the limited loss variable with a limit of 750. Then show that the sum of the expected values of the limited loss and left censored and shifted random variables is equal to the expected value of the original random variable.

All possible values at or above 750 are assigned a value of 750 and their probabilities summed. Thus the probability function is 100 (0.4), 500 (0.2), and 750 (0.4), with an expected value of $100(0.4) + 500(0.2) + 750(0.4) = 440$. The expected value of the original random variable is $100(0.4) + 500(0.2) + 1,000(0.2) + 2,500(0.1) + 10,000(0.1) = 1,590$, which is $440 + 1,150$. □

3.1.1 Exercises

3.1 Develop formulas similar to (3.3) for μ_3 and μ_4.

3.2 Calculate the standard deviation, skewness, and kurtosis for each of the five models. It may help to note that Model 2 is a Pareto distribution and the density function in the continuous part of Model 4 is an exponential distribution. Formulas that may help with calculations for these models appear in Appendix A.

3.3 (*) A random variable has a mean and a coefficient of variation of 2. The third raw moment is 136. Determine the skewness.

3.4 (*) Determine the skewness of a gamma distribution that has a coefficient of variation of 1.

3.5 Determine the mean excess loss function for Models 1–4. Compare the functions for Models 1, 2, and 4.

3.6 (*) For two random variables, X and Y, $e_Y(30) = e_X(30) + 4$. Let X have a uniform distribution on the interval from 0 to 100 and let Y have a uniform distribution on the interval from 0 to w. Determine w.

3.7 (*) A random variable has density function $f(x) = \lambda^{-1}e^{-x/\lambda}$, $x, \lambda > 0$. Determine $e(\lambda)$, the mean excess loss function evaluated at λ.

3.8 Show that the following relationship holds:

$$E(X) = e(d)S(d) + E(X \wedge d). \tag{3.10}$$

3.9 Determine the limited expected value function for Models 1–4. Do this using both (3.8) and (3.10). For Models 1 and 2, also obtain the function using (3.9).

3.10 (*) Which of the following statements are true?

   (a) The mean excess loss function for an empirical distribution is continuous.

   (b) The mean excess loss function for an exponential distribution is constant.

   (c) If it exists, the mean excess loss function for a Pareto distribution is decreasing.

3.11 (*) Losses have a Pareto distribution with $\alpha = 0.5$ and $\theta = 10,000$. Determine the mean excess loss at 10,000.

3.12 Define a right truncated variable and provide a formula for its kth moment.

3.13 (*) The severity distribution of individual claims has pdf

$$f(x) = 2.5x^{-3.5}, \quad x \geq 1.$$

Determine the coefficient of variation.

3.14 (*) Claim sizes are for 100, 200, 300, 400, or 500. The true probabilities for these values are 0.05, 0.20, 0.50, 0.20, and 0.05, respectively. Determine the skewness and kurtosis for this distribution.

3.15 (*) Losses follow a Pareto distribution with $\alpha > 1$ and θ unspecified. Determine the ratio of the mean excess loss function at $x = 2\theta$ to the mean excess loss function at $x = \theta$.

3.16 (*) A random sample of size 10 has two claims of 400, seven claims of 800, and one claim of 1,600. Determine the empirical skewness coefficient for a single claim.

3.2 Percentiles

One other value of interest that may be derived from the distribution function is the percentile function.[4] It is the inverse of the distribution function but, because this quantity is not always well defined, an arbitrary definition must be created.

Definition 3.6 *The* **100p***th* **percentile** *of a random variable is any value* π_p *such that* $F(\pi_p-) \leq p \leq F(\pi_p)$. *The* 50*th percentile,* $\pi_{0.5}$ *is called the* **median**.

   This quantity is sometimes referred to as a quantile. Generally, the quantile uses values between 0 and 1, while the percentile uses values between 0 and 100. Thus the 70th

---

[4]As will be seen from the definition it may not be a function in the mathematical sense, in that it is possible for this function to not produce unique values.
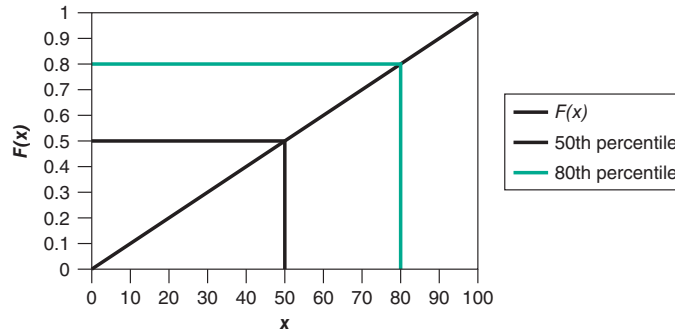
Figure 3.5   The percentiles for Model 1.
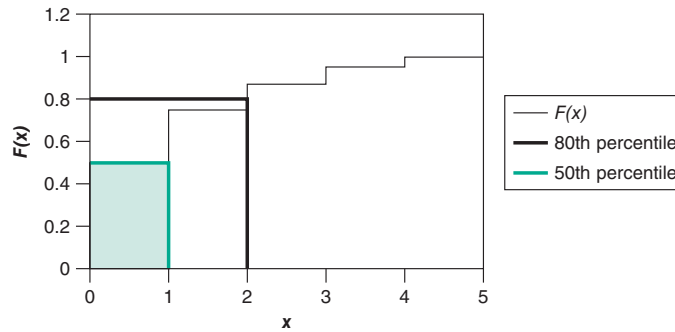
Figure 3.6   The percentiles for Model 3.

percentile and the 0.7 quantile are the same. However, regardless of the term used, we will always use decimal values when using subscripts. So it will be $\pi_{0.5}$ and never π_{50}. If the distribution function has a value of p for one and only one x value, then the percentile is uniquely defined. In addition, if the distribution function jumps from a value below p to a value above p, then the percentile is at the location of the jump. The only time the percentile is not uniquely defined is when the distribution function is constant at a value of p over a range of values of the random variable. In that case, any value in that range (including both endpoints) can be used as the percentile.

■   **EXAMPLE 3.6**

Determine the 50th and 80th percentiles for Models 1 and 3.

For Model 1, the pth percentile can be obtained from $p = F(\pi_p) = 0.01\pi_p$ and so $\pi_p = 100p$ and, in particular, the requested percentiles are 50 and 80 (see Figure 3.5). For Model 3, the distribution function equals 0.5 for all $0 \leq x < 1$, and so any value from 0 to 1 inclusive can be the 50th percentile. For the 80th percentile, note that at $x = 2$ the distribution function jumps from 0.75 to 0.87 and so $\pi_{0.8} = 2$ (see Figure 3.6).   □

3.2.1 Exercises

3.17 (*) The cdf of a random variable is $F(x) = 1 - x^{-2}$, $x \geq 1$. Determine the mean, median, and mode of this random variable.

3.18 Determine the 50th and 80th percentiles for Models 2, 4, and 5.

3.19 (*) Losses have a Pareto distribution with parameters α and θ. The 10th percentile is $\theta - k$. The 90th percentile is $5\theta - 3k$. Determine the value of α.

3.20 (*) Losses have a Weibull distribution with parameters τ and θ. The 25th percentile is 1,000 and the 75th percentile is 100,000. Determine the value of τ.

3.3 Generating Functions and Sums of Random Variables

Consider a portfolio of insurance risks covered by insurance policies issued by an insurance company. The total claims paid by the insurance company on all policies are the sum of all payments made by the insurer. Thus, it is useful to be able to determine properties of $S_k = X_1 + \cdots + X_k$. The first result is a version of the central limit theorem.

Theorem 3.7 *For a random variable S_k as previously defined, $\mathrm{E}(S_k) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_k)$. Also, if $X_1, \ldots, X_k$ are independent, $\mathrm{Var}(S_k) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_k)$. If the random variables $X_1, X_2, \ldots$ are independent and their first two moments meet certain conditions, $\lim_{k \to \infty} [S_k - \mathrm{E}(S_k)]/\sqrt{Var(S_k)}$ has a normal distribution with mean 0 and variance 1.*

When working with a sequence of random variables, there are many types of limit. The limit used in the theorem is called **convergence in distribution**. It means that for a given argument, the distribution function converges to its limiting case. Here, if we define $Z_k = [S_k - \mathrm{E}(S_k)]/\sqrt{\mathrm{Var}(S_k)}$, then, for any value of z, $\lim_{k \to \infty} F_{Z_k}(z) = F_Z(z)$, where Z has a standard normal distribution. Thus, probabilities of sums of random variables can often be approximated by those from the normal distribution.

Obtaining the exact distribution or density function of S_k is usually very difficult. However, there are a few cases where it is simple. The key to this simplicity is the generating function.

Definition 3.8 *For a random variable X, the **moment generating function** (mgf) is $M_X(z) = \mathrm{E}(e^{zX})$ for all z for which the expected value exists. The **probability generating function** (pgf) is $P_X(z) = \mathrm{E}(z^X)$ for all z for which the expectation exists.*

Note that $M_X(z) = P_X(e^z)$ and $P_X(z) = M_X(\ln z)$. Often, the mgf is used for continuous random variables and the pgf for discrete random variables. For us, the value of these functions is not so much that they generate moments or probabilities but that there is a one-to-one correspondence between a random variable's distribution function and its mgf and pgf (i.e. two random variables with different distribution functions cannot have the same mgf or pgf). The following result aids in working with sums of random variables.

Theorem 3.9 *Let $S_k = X_1 + \cdots + X_k$, where the random variables in the sum are independent. Then, $M_{S_k}(z) = \prod_{j=1}^{k} M_{X_j}(z)$ and $P_{S_k}(z) = \prod_{j=1}^{k} P_{X_j}(z)$, provided that all the component mgfs and pgfs exist.*

Proof: We use the fact that the expected product of independent random variables is the product of the individual expectations. Then,

$$M_{S_k}(z) = \mathrm{E}(e^{zS_k}) = \mathrm{E}[e^{z(X_1+\cdots+X_k)}]$$

$$= \prod_{j=1}^{k} \mathrm{E}(e^{zX_j}) = \prod_{j=1}^{k} M_{X_j}(z).$$

A similar argument can be used for the pgf. □

■ **EXAMPLE 3.7**

Show that the sum of independent gamma random variables, each with the same value of θ, has a gamma distribution.

The moment generating function of a gamma variable is

$$\mathrm{E}(e^{zX}) = \frac{\int_0^\infty e^{zx} x^{\alpha-1} e^{-x/\theta}\, dx}{\Gamma(\alpha)\theta^\alpha}$$

$$= \frac{\int_0^\infty x^{\alpha-1} e^{-x(-z+1/\theta)}\, dx}{\Gamma(\alpha)\theta^\alpha}$$

$$= \frac{\int_0^\infty y^{\alpha-1}(-z+1/\theta)^{-\alpha} e^{-y}\, dy}{\Gamma(\alpha)\theta^\alpha}$$

$$= \frac{\Gamma(\alpha)(-z+1/\theta)^{-\alpha}}{\Gamma(\alpha)\theta^\alpha} = \left(\frac{1}{1-\theta z}\right)^\alpha, \quad z < 1/\theta.$$

Now let X_j have a gamma distribution with parameters α_j and θ. Then, the moment generating function of the sum is

$$M_{S_k}(z) = \prod_{j=1}^{k} \left(\frac{1}{1-\theta z}\right)^{\alpha_j} = \left(\frac{1}{1-\theta z}\right)^{\alpha_1+\cdots+\alpha_k},$$

which is the moment generating function of a gamma distribution with parameters $\alpha_1 + \cdots + \alpha_k$ and θ. □

■ **EXAMPLE 3.8**

Obtain the mgf and pgf for the Poisson distribution.

The pgf is

$$P_X(z) = \sum_{x=0}^{\infty} z^x \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(z\lambda)^x}{x!} = e^{-\lambda} e^{z\lambda} = e^{\lambda(z-1)}.$$

Then, the mgf is $M_X(z) = P_X(e^z) = \exp[\lambda(e^z - 1)]$. □

3.3.1   Exercises

3.21 (*) A portfolio contains 16 independent risks, each with a gamma distribution with parameters $\alpha = 1$ and $\theta = 250$. Give an expression using the incomplete gamma function for the probability that the sum of the losses exceeds 6,000. Then approximate this probability using the central limit theorem.

3.22 (*) The severities of individual claims have a Pareto distribution with parameters $\alpha = 8/3$ and $\theta = 8,000$. Use the central limit theorem to approximate the probability that the sum of 100 independent claims will exceed 600,000.

3.23 (*) The severities of individual claims have a gamma distribution (see Appendix A) with parameters $\alpha = 5$ and $\theta = 1,000$. Use the central limit theorem to approximate the probability that the sum of 100 independent claims exceeds 525,000.

3.24 A sample of 1,000 health insurance contracts on adults produced a sample mean of 1,300 for the annual benefits paid with a standard deviation of 400. It is expected that 2,500 contracts will be issued next year. Use the central limit theorem to estimate the probability that benefit payments will be more than 101% of the expected amount.

3.4   Tails of Distributions

The **tail** of a distribution (more properly, the right tail) is the portion of the distribution corresponding to large values of the random variable. Understanding large possible loss values is important because these have the greatest effect on total losses. Random variables that tend to assign higher probabilities to larger values are said to be heavier tailed. Tail weight can be a relative concept (model A has a heavier tail than model B) or an absolute concept (distributions with a certain property are classified as heavy tailed). When choosing models, tail weight can help narrow the choices or can confirm a choice for a model.

3.4.1   Classification Based on Moments

Recall that in the continuous case, the kth raw moment for a random variable that takes on only positive values (like most insurance payment variables) is given by $\int_0^\infty x^k f(x)\, dx$. Depending on the density function and the value of k, this integral may not exist (i.e. it may be infinite). One way of classifying distributions is on the basis of whether all moments exist. It is generally agreed that the existence of all positive moments indicates a (relatively) light right tail, while the existence of only positive moments up to a certain value (or existence of no positive moments at all) indicates a heavy right tail.

■ **EXAMPLE 3.9**

Demonstrate that for the gamma distribution all positive moments exist but for the Pareto distribution they do not.

For the gamma distribution, the raw moments are

$$
\begin{aligned}
\mu'_k &= \int_0^\infty x^k \frac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}\,dx \\
&= \int_0^\infty (y\theta)^k \frac{(y\theta)^{\alpha-1}e^{-y}}{\Gamma(\alpha)\theta^\alpha}\theta\,dy, \quad \text{making the substitution } y = x/\theta \\
&= \frac{\theta^k}{\Gamma(\alpha)}\Gamma(\alpha+k) < \infty \text{ for all } k > 0.
\end{aligned}
$$

For the Pareto distribution, they are

$$
\begin{aligned}
\mu'_k &= \int_0^\infty x^k \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}\,dx \\
&= \int_\theta^\infty (y-\theta)^k \frac{\alpha\theta^\alpha}{y^{\alpha+1}}\,dy, \quad \text{making the substitution } y = x+\theta \\
&= \alpha\theta^\alpha \int_\theta^\infty \sum_{j=0}^k \binom{k}{j} y^{j-\alpha-1}(-\theta)^{k-j}\,dy, \quad \text{for integer values of } k.
\end{aligned}
$$

The integral exists only if all of the exponents on y in the sum are less than -1, that is, if $j - \alpha - 1 < -1$ for all j or, equivalently, if $k < \alpha$. Therefore, only some moments exist. □

By this classification, the Pareto distribution is said to have a heavy tail and the gamma distribution is said to have a light tail. A look at the moment formulas in Appendix A reveals which distributions have heavy tails and which do not, as indicated by the existence of moments.

It is instructive to note that if a distribution does not have all its positive moments, then it does not have a moment generating function (i.e. if X is the associated random variable, then $E(e^{zX}) = \infty$ for all $z > 0$). However, the converse is not true. The lognormal distribution has no moment generating function even though all its positive moments are finite.

Further comparisons of tail behavior can be made on the basis of ratios of moments (assuming they exist). In particular, heavy-tailed behavior is typically associated with large values of quantities such as the coefficient of variation, the skewness, and the kurtosis (see Definition 3.2).

3.4.2 Comparison Based on Limiting Tail Behavior

A commonly used indication that one distribution has a heavier tail than another distribution with the same mean is that the ratio of the two survival functions should diverge to infinity (with the heavier-tailed distribution in the numerator) as the argument becomes large. The divergence implies that the numerator distribution puts significantly more probability on large values. Note that it is equivalent to examine the ratio of density functions. The limit of the ratio will be the same, as can be seen by an application of L'Hôpital's rule:

$$
\lim_{x\to\infty} \frac{S_1(x)}{S_2(x)} = \lim_{x\to\infty} \frac{S'_1(x)}{S'_2(x)} = \lim_{x\to\infty} \frac{-f_1(x)}{-f_2(x)} = \lim_{x\to\infty} \frac{f_1(x)}{f_2(x)}.
$$
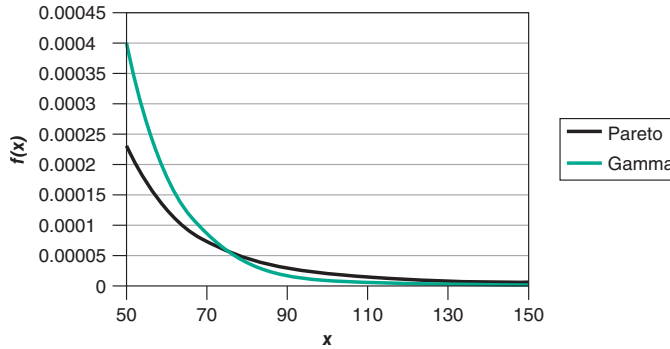
Figure 3.7   The tails of the gamma and Pareto distributions.

■ **EXAMPLE 3.10**

Demonstrate that the Pareto distribution has a heavier tail than the gamma distribution using the limit of the ratio of their density functions.

To avoid confusion, the letters τ and λ will be used for the parameters of the gamma distribution instead of the customary α and θ. Then, the required limit is

$$\lim_{x \to \infty} \frac{f_{\text{Pareto}}(x)}{f_{\text{gamma}}(x)} = \lim_{x \to \infty} \frac{\alpha \theta^{\alpha}(x + \theta)^{-\alpha-1}}{x^{\tau-1}e^{-x/\lambda}\lambda^{-\tau}\Gamma(\tau)^{-1}}$$

$$= c \lim_{x \to \infty} \frac{e^{x/\lambda}}{(x + \theta)^{\alpha+1}x^{\tau-1}}$$

$$> c \lim_{x \to \infty} \frac{e^{x/\lambda}}{(x + \theta)^{\alpha+\tau}}$$

and, either by application of L'Hôpital's rule or by remembering that exponentials go to infinity faster than polynomials, the limit is infinity. Figure 3.7 shows a portion of the density functions for a Pareto distribution with parameters $\alpha = 3$ and $\theta = 10$ and a gamma distribution with parameters $\alpha = \frac{1}{3}$ and $\theta = 15$. Both distributions have a mean of 5 and a variance of 75. The graph is consistent with the algebraic derivation.
□

3.4.3   Classification Based on the Hazard Rate Function

The hazard rate function also reveals information about the tail of the distribution. Distributions with decreasing hazard rate functions have heavy tails. Distributions with increasing hazard rate functions have light tails. In the ensuing discussion, we understand "decreasing" to mean "nonincreasing" and "increasing" to mean "nondecreasing." That is, a decreasing function can be level at times. The exponential distribution, which has a constant hazard rate, is therefore said to have both a decreasing and an increasing hazard rate. For distributions with monotone hazard rates, distributions with exponential tails divide the distributions into heavy-tailed and light-tailed distributions.

Comparisons between distributions can be made on the basis of the rate of increase or decrease of the hazard rate function. For example, a distribution has a lighter tail than

another if its hazard rate function is increasing at a faster rate. Often, these comparisons and classifications are of interest primarily in the right tail of the distribution, that is, for large functional values.

■ **EXAMPLE 3.11**

Compare the tails of the Pareto and gamma distributions by looking at their hazard rate functions.

The hazard rate function for the Pareto distribution is

$$h(x) = \frac{f(x)}{S(x)} = \frac{\alpha\theta^{\alpha}(x+\theta)^{-\alpha-1}}{\theta^{\alpha}(x+\theta)^{-\alpha}} = \frac{\alpha}{x+\theta},$$

which is decreasing. For the gamma distribution we need to be a bit more clever because there is no closed-form expression for $S(x)$. Observe that

$$\frac{1}{h(x)} = \frac{\int_x^{\infty} f(t)\,dt}{f(x)} = \frac{\int_0^{\infty} f(x+y)\,dy}{f(x)},$$

and so, if $f(x+y)/f(x)$ is an increasing function of x for any fixed y, then $1/h(x)$ will be increasing in x and thus the random variable will have a decreasing hazard rate. Now, for the gamma distribution,

$$\frac{f(x+y)}{f(x)} = \frac{(x+y)^{\alpha-1}e^{-(x+y)/\theta}}{x^{\alpha-1}e^{-x/\theta}} = \left(1+\frac{y}{x}\right)^{\alpha-1}e^{-y/\theta},$$

which is strictly increasing in x provided that $\alpha < 1$ and strictly decreasing in x if $\alpha > 1$. By this measure, some gamma distributions have a heavy tail (those with $\alpha < 1$) and some (with $\alpha > 1$) have a light tail. Note that when $\alpha = 1$, we have the exponential distribution and a constant hazard rate. Also, even though $h(x)$ is complicated in the gamma case, we know what happens for large x. Because $f(x)$ and $S(x)$ both go to 0 as $x \to \infty$, L'Hôpital's rule yields

$$\lim_{x\to\infty} h(x) = \lim_{x\to\infty} \frac{f(x)}{S(x)} = -\lim_{x\to\infty} \frac{f'(x)}{f(x)} = -\lim_{x\to\infty}\left[\frac{d}{dx}\ln f(x)\right]$$

$$= -\lim_{x\to\infty}\frac{d}{dx}\left[(\alpha-1)\ln x - \frac{x}{\theta}\right] = \lim_{x\to\infty}\left(\frac{1}{\theta} - \frac{\alpha-1}{x}\right) = \frac{1}{\theta}.$$

That is, $h(x) \to 1/\theta$ as $x \to \infty$. ☐

3.4.4 Classification Based on the Mean Excess Loss Function

The mean excess loss function also gives information about tail weight. If the mean excess loss function is increasing in d, the distribution is considered to have a heavy tail. If the mean excess loss function is decreasing in d, the distribution is considered to have a light tail. Comparisons between distributions can be made on the basis of whether the mean excess loss function is increasing or decreasing. In particular, a distribution with an increasing mean excess loss function has a heavier tail than a distribution with a decreasing mean excess loss function.

In fact, the mean excess loss function and the hazard rate are closely related in several ways. First, note that

$$\frac{S(y+d)}{S(d)} = \frac{\exp\left[-\int_0^{y+d} h(x)\,dx\right]}{\exp\left[-\int_0^{d} h(x)\,dx\right]} = \exp\left[-\int_d^{y+d} h(x)\,dx\right]$$

$$= \exp\left[-\int_0^{y} h(d+t)\,dt\right].$$

Therefore, if the hazard rate is decreasing, then for fixed y it follows that $\int_0^y h(d+t)dt$ is a decreasing function of d, and from the preceding, $S(y+d)/S(d)$ is an increasing function of d. But from (3.5) the mean excess loss function may be expressed as

$$e(d) = \frac{\int_d^\infty S(x)\,dx}{S(d)} = \int_0^\infty \frac{S(y+d)}{S(d)}\,dy.$$

Thus, if the hazard rate is a decreasing function, then the mean excess loss function $e(d)$ is an increasing function of d because the same is true of $S(y+d)/S(d)$ for fixed y. Similarly, if the hazard rate is an increasing function, then the mean excess loss function is a decreasing function. It is worth noting (and is perhaps counterintuitive), however, that the converse implication is not true. Exercise 3.29 gives an example of a distribution that has a decreasing mean excess loss function, but the hazard rate is not increasing for all values. Nevertheless, the implications just described are generally consistent with the preceding discussions of heaviness of the tail.

There is a second relationship between the mean excess loss function and the hazard rate. As $d \to \infty$, $S(d)$ and $\int_d^\infty S(x)\,dx$ go to zero. Thus, the limiting behavior of the mean excess loss function as $d \to \infty$ may be ascertained using L'Hôpital's rule because formula (3.5) holds. We have

$$\lim_{d\to\infty} e(d) = \lim_{d\to\infty} \frac{\int_d^\infty S(x)\,dx}{S(d)} = \lim_{d\to\infty} \frac{-S(d)}{-f(d)} = \lim_{d\to\infty} \frac{1}{h(d)}$$

as long as the indicated limits exist. These limiting relationships may be useful if the form of $F(x)$ is complicated.

■ **EXAMPLE 3.12**

Examine the behavior of the mean excess loss function of the gamma distribution.

Because $e(d) = \int_d^\infty S(x)\,dx/S(d)$ and $S(x)$ is complicated, $e(d)$ is complicated. But $e(0) = E(X) = \alpha\theta$, and, using Example 3.11, we have

$$\lim_{x\to\infty} e(x) = \lim_{x\to\infty} \frac{1}{h(x)} = \frac{1}{\lim\limits_{x\to\infty} h(x)} = \theta.$$

Also, from Example 3.11, $h(x)$ is strictly decreasing in x for $\alpha < 1$ and strictly increasing in x for $\alpha > 1$, implying that $e(d)$ is strictly increasing from $e(0) = \alpha\theta$ to $e(\infty) = \theta$ for $\alpha < 1$ and strictly decreasing from $e(0) = \alpha\theta$ to $e(\infty) = \theta$ for $\alpha > 1$. For $\alpha = 1$, we have the exponential distribution for which $e(d) = \theta$. □

3.4.5 Equilibrium Distributions and Tail Behavior

Further insight into the mean excess loss function and the heaviness of the tail may be obtained by introducing the **equilibrium distribution** (also called the **integrated tail distribution**). For positive random variables with $S(0) = 1$, it follows from Definition 3.3 and (3.5) with $d = 0$ that $E(X) = \int_0^\infty S(x)\,dx$ or, equivalently, $1 = \int_0^\infty [S(x)/E(X)]\,dx$, so that

$$f_e(x) = \frac{S(x)}{E(X)}, \quad x \geq 0, \tag{3.11}$$

is a probability density function. The corresponding survival function is

$$S_e(x) = \int_x^\infty f_e(t)dt = \frac{\int_x^\infty S(t)\,dt}{E(X)}, \quad x \geq 0.$$

The hazard rate corresponding to the equilibrium distribution is

$$h_e(x) = \frac{f_e(x)}{S_e(x)} = \frac{S(x)}{\int_x^\infty S(t)\,dt} = \frac{1}{e(x)}$$

using (3.5). Thus, the reciprocal of the mean excess function is itself a hazard rate, and this fact may be used to show that the mean excess function uniquely characterizes the original distribution. We have

$$f_e(x) = h_e(x)S_e(x) = h_e(x)e^{-\int_0^x h_e(t)\,dt},$$

or, equivalently,

$$S(x) = \frac{e(0)}{e(x)}e^{-\int_0^x \left[\frac{1}{e(t)}\right]dt}$$

using $e(0) = E(X)$.

The equilibrium distribution also provides further insight into the relationship between the hazard rate, the mean excess function, and the heaviness of the tail. Assuming that $S(0) = 1$, and thus $e(0) = E(X)$, we have $\int_x^\infty S(t)\,dt = e(0)S_e(x)$, and from (3.5), $\int_x^\infty S(t)\,dt = e(x)S(x)$. Equating these two expressions results in

$$\frac{e(x)}{e(0)} = \frac{S_e(x)}{S(x)}.$$

If the mean excess function is increasing (which is implied if the hazard rate is decreasing), then $e(x) \geq e(0)$, which is obviously equivalent to $S_e(x) \geq S(x)$ from the preceding equality. This, in turn, implies that

$$\int_0^\infty S_e(x)\,dx \geq \int_0^\infty S(x)\,dx.$$

But $E(X) = \int_0^\infty S(x)\,dx$ from Definition 3.3 and (3.5) if $S(0) = 1$. Also,

$$\int_0^\infty S_e(x)\,dx = \int_0^\infty x f_e(x)\,dx,$$

since both sides represent the mean of the equilibrium distribution. This may be evaluated using (3.9) with $u = \infty$, $k = 2$, and $F(0) = 0$ to give the equilibrium mean, that is,

$$\int_0^\infty S_e(x)\, dx = \int_0^\infty x f_e(x)\, dx = \frac{1}{E(X)} \int_0^\infty x S(x)\, dx = \frac{E(X^2)}{2E(X)}.$$

The inequality may thus be expressed as

$$\frac{E(X^2)}{2E(X)} \geq E(X)$$

or using $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$ as $\mathrm{Var}(X) \geq [E(X)]^2$. That is, the squared coefficient of variation, and hence the coefficient of variation itself, is at least 1 if $e(x) \geq e(0)$. Reversing the inequalities implies that the coefficient of variation is at most 1 if $e(x) \leq e(0)$, which is in turn implied if the mean excess function is decreasing or the hazard rate is increasing. These values of the coefficient of variation are consistent with the comments made here about the heaviness of the tail.

3.4.6   Exercises

3.25 Using the methods in this section (except for the mean excess loss function), compare the tail weight of the Weibull and inverse Weibull distributions.

3.26 Arguments as in Example 3.10 place the lognormal distribution between the gamma and Pareto distributions with regard to heaviness of the tail. To reinforce this conclusion, consider: a gamma distribution with parameters $\alpha = 0.2$, $\theta = 500$; a lognormal distribution with parameters $\mu = 3.709290$, $\sigma = 1.338566$; and a Pareto distribution with parameters $\alpha = 2.5$, $\theta = 150$. First demonstrate that all three distributions have the same mean and variance. Then, numerically demonstrate that there is a value such that the gamma pdf is smaller than the lognormal and Pareto pdfs for all arguments above that value, and that there is another value such that the lognormal pdf is smaller than the Pareto pdf for all arguments above that value.

3.27 For a Pareto distribution with $\alpha > 2$, compare $e(x)$ to $e(0)$ and also determine the coefficient of variation. Confirm that these results are consistent with the Pareto distribution being heavy tailed.

3.28 Let Y be a random variable that has the equilibrium density from (3.11). That is, $f_Y(y) = f_e(y) = S_X(y)/E(X)$ for some random variable X. Use integration by parts to show that

$$M_Y(z) = \frac{M_X(z) - 1}{zE(X)}$$

whenever $M_X(z)$ exists.

3.29 You are given that the random variable X has probability density function $f(x) = (1 + 2x^2)e^{-2x}$, $x \geq 0$.

  (a) Determine the survival function $S(x)$.

  (b) Determine the hazard rate $h(x)$.

(c) Determine the survival function $S_e(x)$ of the equilibrium distribution.

(d) Determine the mean excess function $e(x)$.

(e) Determine $\lim_{x \to \infty} h(x)$ and $\lim_{x \to \infty} e(x)$.

(f) Prove that $e(x)$ is strictly decreasing but $h(x)$ is not strictly increasing.

3.30 Assume that X has probability density function $f(x), x \geq 0$.

(a) Prove that
$$S_e(x) = \frac{\int_x^\infty (y - x)f(y)\, dy}{\mathrm{E}(X)}.$$

(b) Use (a) to show that
$$\int_x^\infty y f(y)\, dy = x S(x) + \mathrm{E}(X) S_e(x).$$

(c) Prove that (b) may be rewritten as
$$S(x) = \frac{\int_x^\infty y f(y)\, dy}{x + e(x)}$$
and that this, in turn, implies that
$$S(x) \leq \frac{\mathrm{E}(X)}{x + e(x)}.$$

(d) Use (c) to prove that, if $e(x) \geq e(0)$, then
$$S(x) \leq \frac{\mathrm{E}(X)}{x + \mathrm{E}(X)}$$
and thus
$$S[k\mathrm{E}(X)] \leq \frac{1}{k + 1},$$
which for $k = 1$ implies that the mean is at least as large as the (smallest) median.

(e) Prove that (b) may be rewritten as
$$S_e(x) = \frac{e(x)}{x + e(x)} \frac{\int_x^\infty y f(y)\, dy}{\mathrm{E}(X)}$$
and thus that
$$S_e(x) \leq \frac{e(x)}{x + e(x)}.$$

3.5   Measures of Risk

3.5.1   Introduction

Probability-based models provide a description of risk exposure. The level of exposure to risk is often described by one number, or perhaps a small set of numbers. These numbers are functions of the model and are often called **key risk indicators**. Such key risk indicators inform actuaries and other risk managers about the degree to which the company is subject to particular aspects of risk. In particular, Value at Risk (VaR) is a quantile of the distribution of aggregate losses. Risk managers often look at the chance of an adverse outcome. This can be expressed through the VaR at a particular probability level. The VaR can also be used in the determination of the amount of capital required to withstand such adverse outcomes. Investors, regulators, and rating agencies are particularly interested in the company's ability to withstand such events.

The VaR suffers from some undesirable properties. A more informative and more useful measure of risk is Tail Value at Risk (TVaR). It has arisen independently in a variety of areas and has been given different names, including Conditional Value at Risk (CVaR), Average Value at Risk (AVaR), Conditional Tail Expectation (CTE), and Expected Shortfall (ES). See Tasche [119] and Acerbi and Tasche [3].

While these measures have been developed in a risk management context, they are useful in assessing any random variable.

3.5.2   Risk Measures and Coherence

A risk measure is a mapping from the set of random variables representing the loss associated with the risks to the real line (the set of all real numbers). A risk measure gives a single number that is intended to quantify the risk exposure. For example, the standard deviation, or a multiple of the standard deviation of a distribution, is a measure of risk because it provides a measure of uncertainty. It is clearly appropriate when using the normal distribution. In the field of finance, the size of loss for which there is a small (e.g. 0.05%) probability of exceedence is a simple risk measure.

Risk measures are denoted by the function $\rho(X)$. It is convenient to think of $\rho(X)$ as the amount of assets required to protect against adverse outcomes of risk X. Studies of risk measures and their properties have included the behavior of risk measures when several losses are combined and treated as a single loss. Combining risks is important in the study of capital needs of an insurance company when considering the overall risk exposure of the insurance company. The insurance company may have several divisions or departments specializing in different products, for example, individual life, homeowners, automobile, group life, annuities, and health. When risk measures are applied to the individual departments, the results should be consistent in some way with the results that are obtained when the risk measure is applied to the entire company. In what follows, it is useful to think of the random variables X and Y as the loss random variables for two divisions and $X + Y$ as the loss random variable for the entity created by combining the two divisions.

The study of risk measures and their properties has been carried out by numerous authors, such as Wang [126] and [127]. Specific desirable properties of risk measures were proposed as axioms in connection with risk pricing by Wang, Young, and Panjer [128] and more generally in risk measurement by Artzner et al. [7]. The Artzner paper

introduced the concept of coherence and is considered to be the groundbreaking paper in risk measurement.

We consider the set of random variables such that if X and Y are two members of the set, then both cX and $X + Y$ are also in the set. This is not very restrictive, but it does eliminate risks that are measured as percentages, as with Model 1 of Chapter 2.

Definition 3.10 *A **coherent risk measure** is a risk measure $\rho(X)$ that has the following four properties for any two loss random variables X and Y:*

1. *Subadditivity: $\rho(X + Y) \leq \rho(X) + \rho(Y)$.*
2. *Monotonicity: if $X \leq Y$ for all possible outcomes, then $\rho(X) \leq \rho(Y)$.*
3. *Positive homogeneity: for any positive constant c, $\rho(cX) = c\rho(X)$.*
4. *Translation invariance: for any constant c, $\rho(X + c) = \rho(X) + c$.*

Subadditivity means that the risk measure (and, hence, the capital required to support it) for two risks combined will not be greater than for the risks treated separately. Subadditivity reflects the fact that there should be some diversification benefit from combining risks. In general, this is necessary at the corporate level. Otherwise, companies would find it to be an advantage to disaggregate into smaller companies. There has been some debate about the appropriateness of the subadditivity requirement. In particular, the merger of several small companies into a larger one exposes each of the small companies to the reputational risk of the others. We will continue to require subadditivity as it reflects the benefit of diversification.

Monotonicity means that if one risk always has greater losses than another risk under all circumstances,[5] the risk measure (and, hence, the capital required to support it) should always be greater. This requirement should be self-evident from an economic viewpoint.

Positive homogeneity means that the risk measure (and, hence, the capital required to support it) is independent of the currency in which the risk is measured. Equivalently, it means that, for example, doubling the exposure to a particular risk requires double the capital. This is sensible because doubling the position provides no diversification.

Translation invariance means that there is no additional risk (and, hence, capital required to support it) for an additional risk for which there is no additional uncertainty. In particular, by making X identically zero, the value of the assets required for a certain outcome is exactly the value of that outcome. Also, when a company meets the capital requirement by setting up additional risk-free capital, the act of injecting the additional capital does not, in itself, trigger a further injection (or reduction) of capital.

Risk measures satisfying these four criteria are deemed to be coherent. There are many such risk measures.

■ **EXAMPLE 3.13**

(*Standard deviation principle*) The standard deviation is a measure of the uncertainty of a distribution. Consider a loss distribution with mean μ and standard deviation σ. The quantity $\mu + k\sigma$, where k is the same fixed constant for all distributions, is a risk measure (often called the **standard deviation principle**). The coefficient k is usually

---

[5]Technically, this means that, for the joint distribution of (X, Y), $\Pr(X > Y) = 0$.

chosen to ensure that losses will exceed the risk measure for some distribution, such as the normal distribution, with some specified small probability.   □

In Exercise 3.31, you are asked to prove that the standard deviation principle is not coherent.

If X follows the normal distribution, a value of $k = 1.645$ results in an exceedence probability of $\Pr(X > \mu + k\sigma) = 5\%$ while, if $k = 2.576$, then $\Pr(X > \mu + k\sigma) = 0.5\%$. However, if the distribution is not normal, the same multiples of the standard deviation will lead to different exceedence probabilities. We can also begin with the exceedence probability, obtaining the quantile $\mu + k\sigma$ and the equivalent value of k. This is the key idea behind Value at Risk.

3.5.3 Value at Risk

In general terms, Value at Risk (VaR) is the amount of capital required to ensure, with a high degree of certainty, that the enterprise does not become technically insolvent. The degree of certainty chosen is arbitrary. In practice, it can be a high number such as 99.95% for the entire enterprise, or it can be much lower, such as 95%, for a single unit or risk class within the enterprise. This lower percentage may reflect the interunit or interrisk type diversification that exists.

Suppose that $F_X(x)$ represents the distribution function of outcomes over a fixed period of time, such as one year, of a portfolio of risks (such as a set of insurance risks or an entire insurance company). An adverse outcome is referred to as a "loss." In the notation used throughout this book, positive values of the random variable X are adverse outcomes, that is, losses. The VaR of the random variable X is the $100p$th percentile of the distribution of X, denoted by $\text{VaR}_p(X) = \pi_p$. This shows why VaR is often called a **quantile risk measure**. When the insurance company has this amount of capital available, it can absorb $100p\%$ of possible outcomes. When set at 99.95% for a one-year time period, the interpretation is that there is only a very small chance (0.05%) that the insurance company will be bankrupted by an adverse outcome over the next year.

Definition 3.11 *Let X denote a loss random variable. The **Value at Risk** of X at the $100p\%$ level, denoted $\text{VaR}_p(X)$ or π_p, is the $100p$th percentile of the distribution of X. More precisely,*

$$\text{VaR}_p(X) = \inf_{x \geq 0}[x|F_X(x) \geq p], \quad 0 < p < 1.$$

Thus, $F_X[\text{VaR}_p(X)] \geq p$, with equality holding if X is continuous at $\text{VaR}_p(X)$.

For continuous distributions, we can simply write $\text{VaR}_p(X)$ for X as the value of π_p satisfying

$$\Pr\left(X > \pi_p\right) = 1 - p.$$

It is well known that VaR does not satisfy one of the four criteria for coherence, the subadditivity requirement. The failure of VaR to be subadditive can be shown by a simple but extreme example inspired by a more complicated one from Wirch [132].

■   **EXAMPLE 3.14**

(*Incoherence of VaR*) Let Z denote a loss random variable of the continuous type with the following cdf values:

$$F_Z(1) = 0.91,$$
$$F_Z(90) = 0.95,$$
$$F_Z(100) = 0.96.$$

The 95% quantile, $\text{VaR}_{0.95}(Z)$, is 90 because there is a 5% chance of exceeding 90.

Suppose that we now split the risk Z into two separate (but dependent) risks X and Y such that the two separate risks in total are equivalent to risk Z, that is, $X + Y = Z$. One way to define them is

$$X = \begin{cases} Z, & Z \leq 100, \\ 0, & Z > 100, \end{cases}$$

and

$$Y = \begin{cases} 0, & Z \leq 100, \\ Z, & Z > 100. \end{cases}$$

The cdf for risk X satisfies

$$F_X(1) = 0.95,$$
$$F_X(90) = 0.99,$$
$$F_X(100) = 1,$$

indicating that $\text{VaR}_{0.95}(X) = 1$.

Similarly, the cdf for risk Y satisfies $F_Y(0) = 0.96$, indicating that there is a 96% chance of no loss. Therefore, the 95% quantile cannot exceed zero, and so $\text{VaR}_{0.95}(Y) \leq 0$. Consequently, the sum of the 95% quantiles for X and Y is less than $\text{VaR}_{0.95}(Z)$, which violates subadditivity.   □

Although this example may appear to be somewhat artificial, the existence of such possibilities creates opportunities for strange or unproductive manipulation. Therefore, we turn to a risk measure that is coherent.

3.5.4   Tail Value at Risk

As a risk measure, VaR is used extensively in financial risk management of trading risk over a fixed (usually relatively short) time period. In these situations, the normal distribution is often used for describing gains or losses. If distributions of gains or losses are restricted to the normal distribution, VaR satisfies all coherency requirements. However, the normal distribution is generally not used for describing insurance losses, which are typically skewed. Consequently, the use of VaR is problematic because of the lack of subadditivity.

Definition 3.12 *Let X denote a loss random variable. The **Tail Value at Risk** of X at the $100p\%$ security level, denoted* $\text{TVaR}_p(X)$*, is the average of all* VaR *values above the security level, p, where $0 < p < 1$. That is,*

$$\text{TVaR}_p(X) = \frac{\int_p^1 \text{VaR}_u(X)\, du}{1 - p}.$$

It is well known (e.g. Artzner et al. [7]) that TVaR is a coherent risk measure. An alternative formula is (e.g. Embrechts and Wang [35])

$$\text{TVaR}_p(X) = \text{VaR}_p(X) + \frac{1 - F_X[\text{VaR}_p(X)]}{1 - p}\{E[X|X > \text{VaR}_p(X)] - \text{VaR}_p(X)\}.$$

If X is continuous at $\text{VaR}_p(X)$, then $F_X[\text{VaR}_p(X)] = p$ and the formula simplifies to

$$\text{TVaR}_p(X) = E[X|X > \text{VaR}_p(X)].$$

Furthermore, in this case,

$$\begin{aligned}
\text{TVaR}_p(X) &= E(X \mid X > \pi_p) \\
&= \pi_p + E(X - \pi_p|X > \pi_p) \\
&= \text{VaR}_p(X) + e(\pi_p),
\end{aligned} \tag{3.12}$$

where $e(\pi_p)$ is the mean excess loss function evaluated at the $100p$th percentile. Thus TVaR is larger than the corresponding VaR by the average excess of all losses that exceed VaR. Furthermore, because $\pi_p = \text{Var}_p(X)$, (3.12) expresses $\text{TVaR}_p(X)$ as a function of $\text{Var}_p(X)$, and in Exercise 3.37 the fact that $\text{TVaR}_p(X)$ is a nondecreasing function of $\text{Var}_p(X)$ is established.

Overbeck [96] also discusses VaR and TVaR as risk measures. He argues that VaR is an "all or nothing" risk measure, in that if an extreme event in excess of the VaR threshold occurs, there is no capital to cushion losses. He also argues that the VaR quantile in TVaR provides a definition of "bad times," which are those where losses exceed the VaR threshold, thereby not using up all available capital when TVaR is used to determine capital. Then, TVaR provides the average excess loss in "bad times," that is, when the VaR "bad times" threshold has been exceeded.

■ **EXAMPLE 3.15**

(*Normal distribution*) Consider a normal distribution with mean μ, standard deviation σ, and pdf

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

Let $\phi(x)$ and $\Phi(x)$ denote the pdf and the cdf of the standard normal distribution ($\mu = 0$, $\sigma = 1$). Then,

$$\text{VaR}_p(X) = \mu + \sigma\Phi^{-1}(p),$$

and, with a bit of calculus, it can be shown that

$$\text{TVaR}_p(X) = \mu + \sigma\frac{\phi\left[\Phi^{-1}(p)\right]}{1 - p}.$$

Note that, in both cases, the risk measure can be translated to the standard deviation principle with an appropriate choice of k. □

■ **EXAMPLE 3.16**

(*Exponential distribution*) Consider an exponential distribution with mean θ and pdf,

$$f(x) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right), \quad x > 0.$$

Then,

$$\text{VaR}_p(X) = -\theta \ln(1-p)$$

and

$$\text{TVaR}_p(X) = \text{VaR}_p(X) + \theta.$$

The excess of TVaR over VaR is a constant θ for all values of p because of the memoryless property of the exponential distribution. □

■ **EXAMPLE 3.17**

(*Pareto distribution*) Consider a Pareto distribution with scale parameter θ, shape parameter $\alpha > 1$, and cdf

$$F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^\alpha, \quad x > 0.$$

Then,

$$\text{VaR}_p(X) = \theta\left[(1-p)^{-1/\alpha} - 1\right]$$

and

$$\text{TVaR}_p(X) = \text{VaR}_p(X) + \frac{\text{VaR}_p(X) + \theta}{\alpha - 1}.$$

The excess of TVaR over VaR is a linear increasing function in the VaR. This means that a larger VaR results in a larger mean excess loss over the VaR indicating a dangerous distribution. □

TVaR is one of many possible coherent risk measures. However, it is particularly well suited to insurance applications where you may want to reflect the shape of the tail beyond the VaR threshold in some way. TVaR represents that shape through a single number: the mean excess loss or expected shortfall.

■ **EXAMPLE 3.18**

(*Tail comparisons*) Consider three loss distributions for an insurance company. Losses for the next year are estimated to be 100 million with standard deviation 223.607 million. You are interested in finding high quantiles of the distribution of losses. Using the normal, Pareto, and Weibull distributions, obtain VaR at the 99%, 99.9%, and 99.99% security levels.

From the mean and standard deviation, using the moment formulas in Appendix A, the distributions and their parameters (in millions) are Normal(100, 223.607), Pareto(150, 2.5), and Weibull(50, 0.5). From the formulas for the cumulative distribution functions, the quantiles $\pi_{0.90}$, $\pi_{0.99}$, and $\pi_{0.999}$ are obtained. They are listed, in millions, in Table 3.1. □

Table 3.1   The quantiles for Example 3.18.

Security level	Normal	Pareto	Weibull
0.900	386.56	226.78	265.09
0.990	620.19	796.44	1,060.38
0.999	791.00	2,227.34	2,385.85

From this example, it should be noted that the results can vary widely depending on the choice of distribution. The normal distribution has a lighter tail than the others. Therefore, the probabilities at extreme outcomes are relatively small, leading to smaller quantiles. The Pareto distribution and the Weibull distribution with $\tau < 1$ have heavy tails and thus relatively larger extreme quantiles. This book is devoted to the exercise of selecting the best distribution based on the available information. In the preceding example, knowing only the mean and standard deviation is not enough to estimate the extreme quantiles needed for determining capital requirements.

In practice, obtaining numerical values of VaR or TVaR can be done either from the data directly or from the distributional formulas, as was done in Example 3.18. When estimating VaR from data, the methods of obtaining quantiles from empirical distributions described in Section 10.4.1 can be used. Since TVaR is the expected value of the observations that are larger than a given threshold, the natural estimator is the average of the values of the observations that exceed the threshold. However, we caution against using this approach in practice unless there are a large number of observations in excess of the threshold. In cases where there are not many observations in excess of the threshold, we prefer to obtain a model for the distribution of all of the observations, or at least of all observations in excess of some relatively low threshold. The values of VaR and TVaR can then be calculated directly from the fitted distribution. This can be done easily for the continuous distributions listed in Appendix A using the relation

$$
\begin{aligned}
\mathrm{TVaR}_p(X) &= \mathrm{E}\left(X \mid X > \pi_p\right) \\
&= \pi_p + \frac{\int_{\pi_p}^{\infty} \left(x - \pi_p\right) f(x)\, dx}{1 - p} \\
&= \pi_p + \frac{\int_{-\infty}^{\infty} \left(x - \pi_p\right) f(x)\, dx - \int_{-\infty}^{\pi_p} \left(x - \pi_p\right) f(x)\, dx}{1 - p} \\
&= \pi_p + \frac{\mathrm{E}(X) - \int_{-\infty}^{\pi_p} x f(x)\, dx - \pi_p \left[1 - F(\pi_p)\right]}{1 - p} \\
&= \pi_p + \frac{\mathrm{E}(X) - \mathrm{E}\left[\min\left(X, \pi_p\right)\right]}{1 - p} \\
&= \pi_p + \frac{\mathrm{E}(X) - \mathrm{E}\left(X \wedge \pi_p\right)}{1 - p}.
\end{aligned} \tag{3.13}
$$

The notation $\mathrm{E}(X \wedge \pi_p)$ is defined in Definition 3.5.

■ **EXAMPLE 3.19**

Using (3.13), obtain TVaR at the 99.9% security level for the Pareto(150, 2.5) distribution.

For the Pareto(150, 2.5) distribution, from Example 3.17 and Appendix A,

$$
\pi_p = \text{VaR}_p(X) = \theta\left[(1-p)^{-1/\alpha} - 1\right]
$$
$$
= 150\left[(1-.999)^{-1/2.5} - 1\right] = 2{,}227.34,
$$
$$
E(X) = \frac{\theta}{\alpha - 1} = \frac{150}{1.5} = 100,
$$
$$
E\left(X \wedge \pi_p\right) = \frac{\theta}{\alpha - 1}\left[1 - \left(\frac{\theta}{\pi_p + \theta}\right)^{\alpha - 1}\right]
$$
$$
= \frac{150}{1.5}\left[1 - \left(\frac{150}{2227.34 + 150}\right)^{1.5}\right] = 98.4151,
$$
$$
\text{TVaR}_p(X) = \pi_p + \frac{E(X) - E\left(X \wedge \pi_p\right)}{1 - p}
$$
$$
= 2{,}227.34 + \frac{100 - 98.4151}{0.001} = 3{,}812.23. \qquad \square
$$

3.5.5 Exercises

3.31 Prove that the standard deviation principle satisfies coherence criteria 1, 3, and 4. To see that it does not satisfy criterion 2, consider the bivariate variable (X, Y) that takes on the value (0,4) with probability 0.25 and the value (4,4) with probability 0.75. Using $k = 1$, show that monotonicity is not satisfied.

3.32 Show that the VaR and TVaR formulas in Example 3.16 are correct.

3.33 Show that the VaR and TVaR formulas in Example 3.17 are correct.

3.34 Verify the parameters and the VaR calculation in Example 3.18

3.35 Using (3.13), obtain TVaR at the 99.9% security level for the Weibull(50, 0.5) distribution.

3.36 Consider an exponential distribution with $\theta = 500$ and a Pareto distribution with $\alpha = 3$ and $\theta = 1{,}000$. Determine VaR and TVaR at the 95% security level.

3.37 Suppose that the distribution of X is continuous on (x_0, ∞), where $-\infty < x_0 < \infty$ (this does not rule out the possibility that $\Pr(X \le x_0) > 0$ with discrete mass points at or below x_0). For $x > x_0$, let $f(x)$ be the pdf, $h(x)$ be the hazard rate function, and $e(x)$ be the mean excess loss function. Demonstrate that

$$
\frac{d}{dx}E(X|X > x) = e(x)h(x), \quad x > x_0,
$$

and hence that $E(X|X > x)$ is nondecreasing in x for $x > x_0$.

PART II

ACTUARIAL MODELS

4

CHARACTERISTICS OF ACTUARIAL MODELS

4.1 Introduction

Basic probability models for actuarial situations tend to be either continuous or discrete (i.e. not mixed). This calls for either counting something (discrete) or paying something (continuous). In both cases, it is unlikely that the model will need to accommodate negative values. The set of all possible distribution functions is too large to comprehend. Therefore, when searching for a distribution function to use as a model for a random phenomenon, it can be helpful if the field can be narrowed.

In Chapter 3, distributions were distinguished by tail weight. In Section 4.2 distributions are classified based on the complexity of the model. Then, in Chapter 5, a variety of continuous models are developed. Chapters 6 and 7 provide a similar treatment of discrete models.

4.2 The Role of Parameters

In this section, models are characterized by how much information is needed to specify the model. The number of quantities (parameters) needed to do so gives some indication of

how complex a model is, in the sense that many items are needed to describe a complex model. Arguments for a simple model include the following:

- With few items required in its specification, it is more likely that each item can be determined more accurately.

- It is more likely to be stable across time and across settings. That is, if the model does well today, it (perhaps with small changes to reflect inflation or similar phenomena) will probably do well tomorrow and will also do well in other, similar situations.

- Because data can often be irregular, a simple model may provide necessary smoothing.

  Of course, complex models also have advantages:

- With many items required in its specification, a complex model can more closely match reality.

- With many items required in its specification, it can more closely match irregularities in the data.

Another way to express the difference is that simpler models can be estimated more accurately, but the model itself may be too superficial. The principle of parsimony states that the simplest model that adequately reflects reality should be used. The definition of "adequately" will depend on the purpose for which the model is to be used.

In the following sections, we move from simpler models to more complex models. There is some difficulty in naming the various classifications because there is not universal agreement on the definitions. With the exception of parametric distributions, the other category names have been created by the authors. It should also be understood that these categories do not cover the universe of possible models, nor will every model be easy to categorize. These should be considered as qualitative descriptions.

4.2.1 Parametric and Scale Distributions

These models are simple enough to be specified by a few key numbers.

Definition 4.1 *A **parametric distribution** is a set of distribution functions each member of which is determined by specifying one or more values called **parameters**. The number of parameters is fixed and finite.*

The most familiar parametric distribution is the normal distribution with parameters μ and σ^2. When values for these two parameters are specified, the distribution function is completely known.

These are the simplest distributions in this section, because typically only a small number of values need to be specified. All of the individual distributions in Appendices A and B are parametric. Within this class, distributions with fewer parameters are simpler than those with more parameters.

For much of actuarial modeling work, it is especially convenient if the name of the distribution is unchanged when the random variable is multiplied by a constant. The most common uses for this phenomenon are to model the effect of inflation and to accommodate a change in the monetary unit.

Definition 4.2 *A parametric distribution is a **scale distribution** if, when a random variable from that set of distributions is multiplied by a positive constant, the resulting random variable is also in that set of distributions.*

■ **EXAMPLE 4.1**

Demonstrate that the exponential distribution is a scale distribution.

According to Appendix A, the distribution function is $F_X(x) = 1 - e^{-x/\theta}$. Let $Y = cX$, where $c > 0$. Then,

$$
\begin{aligned}
F_Y(y) &= \Pr(Y \le y) \\
&= \Pr(cX \le y) \\
&= \Pr\left(X \le \frac{y}{c}\right) \\
&= 1 - e^{-y/c\theta}.
\end{aligned}
$$

This is an exponential distribution with parameter $c\theta$. □

Definition 4.3 *For random variables with nonnegative support, a **scale parameter** is a parameter for a scale distribution that meets two conditions. First, when a member of the scale distribution is multiplied by a positive constant, the scale parameter is multiplied by the same constant. Second, when a member of the scale distribution is multiplied by a positive constant, all other parameters are unchanged.*

■ **EXAMPLE 4.2**

Demonstrate that the gamma distribution, as defined in Appendix A, has a scale parameter.

Let X have the gamma distribution and $Y = cX$. Then, using the incomplete gamma notation in Appendix A,

$$
\begin{aligned}
F_Y(y) &= \Pr\left(X \le \frac{y}{c}\right) \\
&= \Gamma\left(\alpha; \frac{y}{c\theta}\right),
\end{aligned}
$$

which indicates that Y has a gamma distribution with parameters α and $c\theta$. Therefore, the parameter θ is a scale parameter. □

Many textbooks write the density function for the gamma distribution as

$$
f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0.
$$

In this book, we have chosen to use the version of the density function that has a scale parameter, θ, that is equal to $1/\beta$ in the above version. The mean is proportional to θ in our version. Our version makes it easier to get ballpark estimates of this parameter, although,

for the alternative definition, all that is needed is to recall that the parameter β is inversely proportional to the mean.

It is often possible to recognize a scale parameter from looking at the distribution or density function. In particular, in the distribution function, x only appears together with θ in the form x/θ.

4.2.2 Parametric Distribution Families

A slightly more complex version of a parametric distribution is one in which the number of parameters is finite but not fixed in advance.

Definition 4.4 *A **parametric distribution family** is a set of parametric distributions that are related in some meaningful way.*

The most common type of parametric distribution family is described in the following example.

■ **EXAMPLE 4.3**

One type of parametric distribution family is based on a specified parametric distribution. Other members of the family are obtained by setting one or more parameters from the specified distribution equal to a preset value or to each other. Demonstrate that the transformed beta family as defined in Appendix A is a parametric distribution family.

The transformed beta distribution has four parameters. Each of the other named distributions in the family is a transformed beta distribution with certain parameters set equal to 1 (e.g. the Pareto distribution has $\gamma = \tau = 1$) or to each other (the paralogistic distribution has $\tau = 1$ and $\gamma = \alpha$). Note that the number of parameters (ranging from two to four) is not known in advance. There is a subtle difference in definitions. A modeler who uses the transformed beta distribution looks at all four parameters over their range of possible values. A modeler who uses the transformed beta family pays particular attention to the possibility of using special cases such as the Burr distribution. For example, if the former modeler collects some data and decides that $\tau = 1.01$, that will be the value to use. The latter modeler will note that $\tau = 1$ gives a Burr distribution and will likely use that model instead. ☐

4.2.3 Finite Mixture Distributions

By themselves, mixture distributions are no more complex, but later in this section we find a way to increase the complexity level. One motivation for mixing is that the underlying phenomenon may actually be several phenomena that occur with unknown probabilities. For example, a randomly selected dental claim may be from a checkup, from a filling, from a repair (such as a crown), or from a surgical procedure. Because of the differing modes for these possibilities, a mixture model may work well.

Definition 4.5 *Consider a set of random variables $\{X_1, X_2, \ldots, X_k\}$ with associated marginal distribution functions $\{F_{X_1}, F_{X_2}, \ldots, F_{X_k}\}$. The random variable Y is a **k-point***

mixture[1] *of the random variables* $\{X_1, X_2, \ldots, X_k\}$ *if its cdf is given by*

$$F_Y(y) = a_1 F_{X_1}(y) + a_2 F_{X_2}(y) + \cdots + a_k F_{X_k}(y), \tag{4.1}$$

where all $a_j > 0$ *and* $a_1 + a_2 + \cdots + a_k = 1$.

Note that the joint distribution of $\{X_1, X_2, \ldots, X_k\}$ is not relevant here. The distribution of Y depends only on the marginal distribution functions of these variables.

An interpretation of this model is that it assigns probability a_j to the outcome that Y is a realization of the random variable X_j. Note that, if we have 20 choices for a given random variable, a two-point mixture allows us to create over 200 new distributions.[2] This may be sufficient for most modeling situations. Nevertheless, these are still parametric distributions, though perhaps with many parameters.

■ **EXAMPLE 4.4**

For models involving general liability insurance, actuaries at the Insurance Services Office once considered a mixture of two Pareto distributions. This model contains a total of five parameters. They decided that five parameters were not necessary and selected the mixture distribution with distribution function

$$F(x) = 1 - a \left(\frac{\theta_1}{\theta_1 + x} \right)^\alpha - (1 - a) \left(\frac{\theta_2}{\theta_2 + x} \right)^{\alpha+2}.$$

Note that the shape parameters in the two Pareto distributions differ by 2. The second distribution places more probability on smaller values. This might be a model for frequent, small claims while the first distribution covers large but infrequent claims. This distribution has four parameters, bringing some parsimony to the modeling process.

□

Suppose that we do not know how many distributions should be in the mixture. Then the value of k becomes a parameter, as indicated in the following definition.

Definition 4.6 *A **variable-component mixture distribution** has a distribution function that can be written as*

$$F(x) = \sum_{j=1}^{K} a_j F_j(x), \quad \sum_{j=1}^{K} a_j = 1, \quad a_j > 0, \ j = 1, \ldots, K, \ K = 1, 2, \ldots.$$

These models have been called *semiparametric* because in complexity they are between parametric models and nonparametric models (see Section 4.2.4). This distinction becomes more important when model selection is discussed in Chapter 15. When the

---

[1] The words "mixed" and "mixture" have been used interchangeably to refer to the type of distribution described here as well as distributions that are partly discrete and partly continuous. To add to this, any distribution that is partly discrete and partly continuous can be written as a two-point mixture. This text does not attempt to resolve that confusion. The context will make clear which type of distribution is being considered.

[2] There are actually $\binom{20}{2} + 20 = 210$ choices. The extra 20 represent the cases where both distributions are of the same type but with different parameters.

number of parameters is to be estimated from data, hypothesis tests to determine the appropriate number of parameters become more difficult. When all of the components have the same parametric distribution (but different parameters), the resulting distribution is called a "variable mixture of gs" distribution, where g stands for the name of the component distribution.

■ **EXAMPLE 4.5**

Determine the distribution, density, and hazard rate functions for the variable mixture-of-exponentials distribution.

A mixture of exponential distribution functions can be written as

$$F(x) = 1 - a_1 e^{-x/\theta_1} - a_2 e^{-x/\theta_2} - \cdots - a_K e^{-x/\theta_K},$$

$$\sum_{j=1}^{K} a_j = 1, \quad a_j, \theta_j > 0, \; j = 1, \dots, K, \; K = 1, 2, \dots.$$

Then, the other functions are

$$f(x) = a_1 \theta_1^{-1} e^{-x/\theta_1} + a_2 \theta_2^{-1} e^{-x/\theta_2} + \cdots + a_K \theta_K^{-1} e^{-x/\theta_K},$$

$$h(x) = \frac{a_1 \theta_1^{-1} e^{-x/\theta_1} + a_2 \theta_2^{-1} e^{-x/\theta_2} + \cdots + a_K \theta_K^{-1} e^{-x/\theta_K}}{a_1 e^{-x/\theta_1} + a_2 e^{-x/\theta_2} + \cdots + a_K e^{-x/\theta_K}}.$$

The number of parameters is not fixed, nor is it even limited. For example, when $K = 2$ there are three parameters $(a_1, \theta_1, \theta_2)$, noting that a_2 is not a parameter because once a_1 is set, the value of a_2 is determined. However, when $K = 4$, there are seven parameters. □

The paper by Keatinge [68] presents a strong argument for using the mixture-of-exponentials distribution as an all-purpose model. An extension where one of the distributions in the mixture is not exponential is provided by Klugman and Rioux [75].

■ **EXAMPLE 4.6**

Illustrate how a two-point mixture of gamma variables can create a bimodal distribution.

Consider a 50–50 mixture of two gamma distributions. One has parameters $\alpha = 4$ and $\theta = 7$ (for a mode of 21) and the other has parameters $\alpha = 15$ and $\theta = 7$ (for a mode of 98). The density function is

$$f(x) = 0.5 \frac{x^3 e^{-x/7}}{3! \, 7^4} + 0.5 \frac{x^{14} e^{-x/7}}{14! \, 7^{15}},$$

and a graph is shown in Figure 4.1.

□

4.2.4 Data-Dependent Distributions

Models 1–5 and many of the examples rely on an associated phenomenon (the random variable) but not on observations of that phenomenon. For example, without having
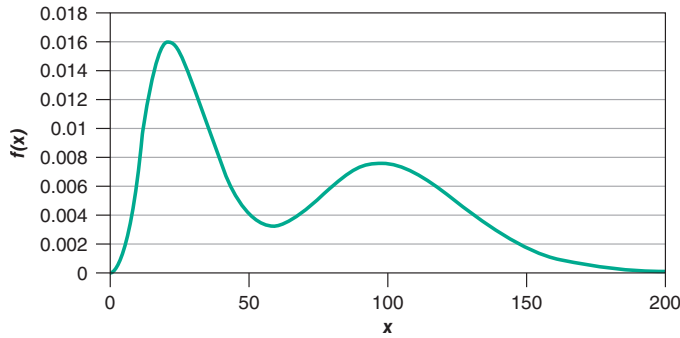
Figure 4.1   The two-point mixture-of-gammas distribution.

observed any dental claims, we could postulate a lognormal distribution with parameters $\mu = 5$ and $\sigma = 1$ or perhaps a mixture of two lognormal distributions. Our model may be a poor description of dental claims, but that is a different matter. There is another type of model that, unlike the lognormal example, requires data. These models also have parameters but are often called nonparametric.

Definition 4.7 *A **data-dependent distribution** is at least as complex as the data or knowledge that produced it, and the number of "parameters" increases as the number of data points or amount of knowledge increases.*

Essentially, these models have as many (or more) "parameters" than observations in the data set.

Definition 4.8 *The **empirical model** is a discrete distribution based on a sample of size n that assigns probability $1/n$ to each data point.*

■ **EXAMPLE 4.7**

Consider a sample of size 8 in which the observed data points were 3, 5, 6, 6, 6, 7, 7, and 10. The empirical model then has probability function

$$p(x) = \begin{cases} 0.125, & x = 3, \\ 0.125, & x = 5, \\ 0.375, & x = 6, \\ 0.25, & x = 7, \\ 0.125, & x = 10. \end{cases}$$

□

Note that many discrete models with finite support look like empirical models. Model 3 in Section 2.2 could have been the empirical model for a sample of size 100 that contained 50 zeros, 25 ones, 12 twos, 8 threes, and 5 fours. Regardless, we use the term "empirical model" only when there is an actual sample behind it.

The empirical distribution is a data-dependent distribution. Each data point contributes probability $1/n$ to the probability function, so the n parameters are the n observations in the data set that produced the empirical distribution.
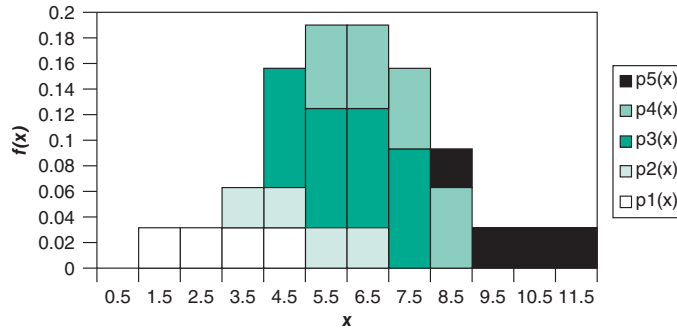
Figure 4.2   The kernel density distribution.

Another example of a data-dependent model is the kernel smoothing model, which is covered in more detail in Section 14.6. Rather than placing a probability mass of $1/n$ at each data point, a continuous density function with area $1/n$ replaces the data point. This piece is centered at the data point so that this model follows the data, but not perfectly. It provides some smoothing when compared to the empirical distribution. A simple example follows.

■ **EXAMPLE 4.8**

Construct a kernel smoothing model from the data that produced Example 4.7 using the uniform kernel and a bandwidth of 2.

The probability density function is

$$f(x) = \sum_{j=1}^{5} p_8(x_j) K_j(x),$$

$$K_j(x) = \begin{cases} 0, & |x - x_j| > 2, \\ 0.25, & |x - x_j| \le 2, \end{cases}$$

where the sum is taken over the five points where the original model has positive probability. For example, the first term of the sum is the function

$$p_8(x_1) K_1(x) = \begin{cases} 0, & x < 1, \\ 0.03125, & 1 \le x \le 5, \\ 0, & x > 5. \end{cases}$$

The complete density function is the sum of five such functions, which are illustrated in Figure 4.2.   □

Note that both the kernel smoothing model and the empirical distribution can also be written as mixture distributions. The reason why these models are classified separately is that the number of components relates to the sample size rather than to the phenomenon and its random variable.

4.2.5 Exercises

4.1 Demonstrate that the lognormal distribution as parameterized in Appendix A is a scale distribution but has no scale parameter. Display an alternative parametrization of this distribution that does have a scale parameter.

4.2 Which of Models 1–5 could be considered as members of a parametric distribution? For those that are, name or describe the distribution.

4.3 (*) Claims have a Pareto distribution with $\alpha = 2$ and θ unknown. Claims the following year experience 6% uniform inflation. Let r be the ratio of the proportion of claims that will exceed d next year to the proportion of claims that exceed d this year. Determine the limit of r as d goes to infinity.

4.4 Determine the mean and second moment of the two-point mixture distribution in Example 4.4. The solution to this exercise provides general formulas for raw moments of a mixture distribution.

4.5 Determine expressions for the mean and variance of the mixture-of-gammas distribution.

4.6 Which of Models 1–5 could be considered to be from parametric distribution families? Which could be considered to be from variable-component mixture distributions?

4.7 (*) Seventy-five percent of claims have a normal distribution with a mean of 3,000 and a variance of 1,000,000. The remaining 25% have a normal distribution with a mean of 4,000 and a variance of 1,000,000. Determine the probability that a randomly selected claim exceeds 5,000.

4.8 (*) Let X have a Burr distribution with parameters $\alpha = 1$, $\gamma = 2$, and $\theta = \sqrt{1,000}$ and let Y have a Pareto distribution with parameters $\alpha = 1$ and $\theta = 1,000$. Let Z be a mixture of X and Y with equal weight on each component. Determine the median of Z. Let $W = 1.1Z$. Demonstrate that W is also a mixture of a Burr and a Pareto distribution and determine the parameters of W.

4.9 (*) Consider three random variables: X is a mixture of a uniform distribution on the interval 0–2 and a uniform distribution on the interval 0–3; Y is the sum of two random variables, one uniform on 0–2 and the other uniform on 0–3; Z is a normal distribution that has been right censored at 1. Match these random variables with the following descriptions:

  (a) Both the distribution and density functions are continuous.
  (b) The distribution function is continuous but the density function is discontinuous.
  (c) The distribution function is discontinuous.

4.10 Demonstrate that the model in Example 4.8 is a mixture of uniform distributions.

4.11 Show that the inverse Gaussian distribution as parameterized in Appendix A is a scale family but does not have a scale parameter.

4.12 Show that the Weibull distribution has a scale parameter.

5

CONTINUOUS MODELS

5.1 Introduction

In this chapter, a variety of continuous models are introduced. The collection developed here should be sufficient for most modeling situations. The discussion begins by showing how new distributions can be created from existing ones. This bottom-up approach allows for additional parameters to be added.

5.2 Creating New Distributions

Many continuous distributions are listed in Appendix A where formulas are given for the pdf, cdf, and other quantities. In actuarial applications, we are mainly interested in distributions that have only positive support, that is, where $F(0) = 0$. Distributions with this property that are familiar to most students of statistics include the exponential, gamma, Pareto, and lognormal distributions.

For any distribution, it is possible to construct other, new distributions by making a transformation or by using some other mathematical technique. Many of the distributions in Appendix A can be obtained by applying such methods to other distributions that are

also listed in Appendix A. In this section, we illustrate such methods through numerous examples. The examples help explain some of the relationships between the distributions. In particular, Section 5.3 examines "families" of distributions where the members of the family are all special cases of a "parent" distribution.

5.2.1   Multiplication by a Constant

This transformation is equivalent to applying inflation uniformly across all loss levels and is known as a change of scale. For example, if this year's losses are given by the random variable X, then uniform inflation of 5% indicates that next year's losses can be modeled with the random variable $Y = 1.05X$.

Theorem 5.1 *Let X be a continuous random variable with pdf $f_X(x)$ and cdf $F_X(x)$. Let $Y = \theta X$ with $\theta > 0$. Then,*

$$F_Y(y) = F_X\left(\frac{y}{\theta}\right), \quad f_Y(y) = \frac{1}{\theta} f_X\left(\frac{y}{\theta}\right).$$

Proof:

$$F_Y(y) = \Pr(Y \le y) = \Pr(\theta X \le y) = \Pr\left(X \le \frac{y}{\theta}\right) = F_X\left(\frac{y}{\theta}\right)$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{\theta} f_X\left(\frac{y}{\theta}\right).$$

$\square$

Corollary 5.2 *The parameter θ is a scale parameter for the random variable Y.*

The following example illustrates this process.

■ **EXAMPLE 5.1**

Let X have pdf $f(x) = e^{-x}$, $x > 0$. Determine the cdf and pdf of $Y = \theta X$.

$$F_X(x) = 1 - e^{-x}, \quad F_Y(y) = 1 - e^{-y/\theta},$$

$$f_Y(y) = \frac{1}{\theta} e^{-y/\theta}.$$

We recognize this as the exponential distribution. $\square$

5.2.2   Raising to a Power

Theorem 5.3 *Let X be a continuous random variable with pdf $f_X(x)$ and cdf $F_X(x)$ and with $F_X(0) = 0$. Let $Y = X^{1/\tau}$. Then, if $\tau > 0$,*

$$F_Y(y) = F_X(y^\tau), \quad f_Y(y) = \tau y^{\tau-1} f_X(y^\tau), \quad y > 0$$

while, if $\tau < 0$,

$$F_Y(y) = 1 - F_X(y^\tau), \quad f_Y(y) = -\tau y^{\tau-1} f_X(y^\tau). \tag{5.1}$$

Proof: If $\tau > 0$,

$$F_Y(y) = \Pr(X \leq y^\tau) = F_X(y^\tau),$$

while, if $\tau < 0$,

$$F_Y(y) = \Pr(X \geq y^\tau) = 1 - F_X(y^\tau).$$

The pdf follows by differentiation.   □

It is more common to keep parameters positive and so, when τ is negative, create a new parameter $\tau^* = -\tau$. Then, (5.1) becomes

$$F_Y(y) = 1 - F_X(y^{-\tau^*}), \quad f_Y(y) = \tau^* y^{-\tau^*-1} f_X(y^{-\tau^*}).$$

We drop the asterisk for future use of this positive parameter.

Definition 5.4 *When raising a distribution to a power, if $\tau > 0$, the resulting distribution is called **transformed**; if $\tau = -1$, it is called **inverse**; and if $\tau < 0$ (but is not -1), it is called **inverse transformed**. To create the distributions in Appendix A and to retain θ as a scale parameter, the base distribution should be raised to a power before being multiplied by θ.*

■ **EXAMPLE 5.2**

Suppose that X has an exponential distribution. Determine the cdf of the inverse, transformed, and inverse transformed exponential distributions.

The inverse exponential distribution with no scale parameter has cdf

$$F(y) = 1 - [1 - e^{-1/y}] = e^{-1/y}.$$

With the scale parameter added, it is $F(y) = e^{-\theta/y}$.

The transformed exponential distribution with no scale parameter has cdf

$$F(y) = 1 - \exp(-y^\tau).$$

With the scale parameter added, it is $F(y) = 1 - \exp[-(y/\theta)^\tau]$. This distribution is more commonly known as the **Weibull distribution**.

The inverse transformed exponential distribution with no scale parameter has cdf

$$F(y) = 1 - [1 - \exp(-y^{-\tau})] = \exp(-y^{-\tau}).$$

With the scale parameter added, it is $F(y) = \exp[-(\theta/y)^\tau]$. This distribution is the **inverse Weibull**.   □

Another base distribution has pdf $f(x) = x^{\alpha-1} e^{-x}/\Gamma(\alpha)$. When a scale parameter is added, this becomes the **gamma distribution.** It has inverse and transformed versions that can be created using the results in this section. Unlike the distributions introduced to this point, this one does not have a closed-form cdf. The best we can do is define notation for the function.

Definition 5.5 *The **incomplete gamma function** with parameter $\alpha > 0$ is denoted and defined by*

$$\Gamma(\alpha; x) = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t} \, dt,$$

*while the **gamma function** is denoted and defined by*

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \, dt.$$

In addition, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, and for positive integer values of n, $\Gamma(n) = (n-1)!$. Appendix A provides details on numerical methods of evaluating these quantities. Furthermore, these functions are built into most spreadsheet programs and many statistical and numerical analysis programs.

5.2.3 Exponentiation

Theorem 5.6 *Let X be a continuous random variable with pdf $f_X(x)$ and cdf $F_X(x)$ and with $f_X(x) > 0$ for all real x. Let $Y = \exp(X)$. Then, for $y > 0$,*

$$F_Y(y) = F_X(\ln y), \quad f_Y(y) = \frac{1}{y} f_X(\ln y).$$

Proof: $F_Y(y) = \Pr(e^X \leq y) = \Pr(X \leq \ln y) = F_X(\ln y)$. The pdf follows by differentiation. $\qquad\square$

■ **EXAMPLE 5.3**

Let X have a normal distribution with mean μ and variance σ^2. Determine the cdf and pdf of $Y = e^X$.

$$F_Y(y) = \Phi\left(\frac{\ln y - \mu}{\sigma}\right)$$

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

$\square$

We could try to add a scale parameter by creating $W = \theta Y$, but this adds no value, as is demonstrated in Exercise 5.5. This example created the **lognormal** distribution (the name has stuck even though "expnormal" would seem more descriptive).

5.2.4 Mixing

The concept of mixing can be extended from mixing a finite number of random variables to mixing an uncountable number. In the following theorem, the pdf $f_\Lambda(\lambda)$ plays the role of the discrete probabilities a_j in the k-point mixture.

Theorem 5.7 *Let X have pdf $f_{X|\Lambda}(x|\lambda)$ and cdf $F_{X|\Lambda}(x|\lambda)$, where λ is a parameter of X. While X may have other parameters, they are not relevant. Let λ be a realization of the random variable Λ with pdf $f_\Lambda(\lambda)$. Then, the unconditional pdf of X is*

$$f_X(x) = \int f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) \, d\lambda, \tag{5.2}$$

*where the integral is taken over all values of λ with positive probability. The resulting distribution is a **mixture distribution**. The distribution function can be determined from*

$$F_X(x) = \int_{-\infty}^x \int f_{X|\Lambda}(y|\lambda) f_\Lambda(\lambda) \, d\lambda \, dy$$

$$= \int \int_{-\infty}^x f_{X|\Lambda}(y|\lambda) f_\Lambda(\lambda) \, dy \, d\lambda$$

$$= \int F_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) \, d\lambda.$$

Moments of the mixture distribution can be found from

$$\mathrm{E}(X^k) = \mathrm{E}[\mathrm{E}(X^k|\Lambda)]$$

and, in particular,

$$\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|\Lambda)] + \mathrm{Var}[\mathrm{E}(X|\Lambda)].$$

Proof: The integrand is, by definition, the joint density of X and Λ. The integral is then the marginal density. For the expected value (assuming that the order of integration can be reversed),

$$\mathrm{E}(X^k) = \int \int x^k f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) \, d\lambda \, dx$$

$$= \int \left[\int x^k f_{X|\Lambda}(x|\lambda) \, dx \right] f_\Lambda(\lambda) \, d\lambda$$

$$= \int \mathrm{E}(X^k|\lambda) f_\Lambda(\lambda) \, d\lambda$$

$$= \mathrm{E}[\mathrm{E}(X^k|\Lambda)].$$

For the variance,

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

$$= \mathrm{E}[\mathrm{E}(X^2|\Lambda)] - \{\mathrm{E}[\mathrm{E}(X|\Lambda)]\}^2$$

$$= \mathrm{E}\{\mathrm{Var}(X|\Lambda) + [\mathrm{E}(X|\Lambda)]^2\} - \{\mathrm{E}[\mathrm{E}(X|\Lambda)]\}^2$$

$$= \mathrm{E}[\mathrm{Var}(X|\Lambda)] + \mathrm{Var}[\mathrm{E}(X|\Lambda)]. \qquad \square$$

Note that if $f_\Lambda(\lambda)$ is discrete, the integrals must be replaced with sums. An alternative way to write the results is $f_X(x) = \mathrm{E}_\Lambda[f_{X|\Lambda}(x|\Lambda)]$ and $F_X(x) = \mathrm{E}_\Lambda[F_{X|\Lambda}(x|\Lambda)]$, where the subscript on E indicates that the random variable is Λ.

An interesting phenomenon is that mixture distributions tend to be heavy tailed, so this method is a good way to generate such a model. In particular, if $f_{X|\Lambda}(x|\lambda)$ has a decreasing hazard rate function for all λ, then the mixture distribution will also have a decreasing hazard rate function (for details, see Ross [107, pp. 407–409]). The following example shows how a familiar heavy-tailed distribution may be obtained by mixing.

■ **EXAMPLE 5.4**

Let $X|\Lambda$ have an exponential distribution with parameter $1/\Lambda$. Let Λ have a gamma distribution. Determine the unconditional distribution of X.

We have (note that the parameter θ in the gamma distribution has been replaced by its reciprocal)

$$
\begin{aligned}
f_X(x) &= \frac{\theta^\alpha}{\Gamma(\alpha)} \int_0^\infty \lambda e^{-\lambda x} \lambda^{\alpha-1} e^{-\theta\lambda} \, d\lambda \\
&= \frac{\theta^\alpha}{\Gamma(\alpha)} \int_0^\infty \lambda^\alpha e^{-\lambda(x+\theta)} \, d\lambda \\
&= \frac{\theta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{(x+\theta)^{\alpha+1}} \\
&= \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}.
\end{aligned}
$$

This is a Pareto distribution.   □

The following example provides an illustration that is useful in Chapter 16.

■ **EXAMPLE 5.5**

Suppose that, given $\Theta = \theta$, X is normally distributed with mean θ and variance v, so that

$$
f_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{1}{2v}(x-\theta)^2\right], \quad -\infty < x < \infty,
$$

and Θ is itself normally distributed with mean μ and variance a, that is,

$$
f_\Theta(\theta) = \frac{1}{\sqrt{2\pi a}} \exp\left[-\frac{1}{2a}(\theta-\mu)^2\right], \quad -\infty < \theta < \infty.
$$

Determine the marginal pdf of X.

The marginal pdf of X is

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{1}{2v}(x-\theta)^2\right] \frac{1}{\sqrt{2\pi a}} \exp\left[-\frac{1}{2a}(\theta-\mu)^2\right] d\theta \\
&= \frac{1}{2\pi\sqrt{va}} \int_{-\infty}^\infty \exp\left[-\frac{1}{2v}(x-\theta)^2 - \frac{1}{2a}(\theta-\mu)^2\right] d\theta.
\end{aligned}
$$

We leave as an exercise the verification of the algebraic identity

$$
\frac{(x-\theta)^2}{v} + \frac{(\theta-\mu)^2}{a} = \frac{a+v}{va}\left(\theta - \frac{ax+v\mu}{a+v}\right)^2 + \frac{(x-\mu)^2}{a+v}
$$

obtained by completion of the square in θ. Thus,

$$
f_X(x) = \frac{\exp\left[-\dfrac{(x-\mu)^2}{2(a+v)}\right]}{\sqrt{2\pi(a+v)}} \int_{-\infty}^\infty \sqrt{\frac{a+v}{2\pi va}} \exp\left[-\frac{a+v}{2va}\left(\theta - \frac{ax+v\mu}{a+v}\right)^2\right] d\theta.
$$

We recognize the integrand as the pdf (as a function of θ) of a normal distribution with mean $(ax+v\mu)/(a+v)$ and variance $(va)/(a+v)$. Thus the integral is 1 and so

$$f_X(x) = \frac{\exp\left[-\dfrac{(x-\mu)^2}{2(a+v)}\right]}{\sqrt{2\pi(a+v)}}, \quad -\infty < x < \infty;$$

that is, X is normal with mean μ and variance $a + v$. $\qquad\square$

The following example is taken from Hayne [51]. It illustrates how a mixture distribution can arise. In particular, continuous mixtures are often used to provide a model for parameter uncertainty. That is, the exact value of a parameter is not known but a probability density function can be elucidated to describe possible values of that parameter.

■ **EXAMPLE 5.6**

In the valuation of warranties on automobiles, it is important to recognize that the number of miles driven varies from driver to driver. It is also the case that for a particular driver, the number of miles varies from year to year. Suppose that the number of miles for a randomly selected driver has an inverse Weibull distribution but the year-to-year variation in the scale parameter has a transformed gamma distribution with the same value for τ. Determine the distribution for the number of miles driven in a randomly selected year by a randomly selected driver.

Using the parameterizations from Appendix A, the inverse Weibull for miles driven in a year has parameters Λ (in place of Θ) and τ, while the transformed gamma distribution for the scale parameter Λ has parameters τ, θ, and α. The marginal density is

$$\begin{aligned}
f(x) &= \int_0^\infty \frac{\tau\lambda^\tau}{x^{\tau+1}} e^{-(\lambda/x)^\tau} \frac{\tau\lambda^{\tau\alpha-1}}{\theta^{\tau\alpha}\Gamma(\alpha)} e^{-(\lambda/\theta)^\tau}\, d\lambda \\
&= \frac{\tau^2}{\theta^{\tau\alpha}\Gamma(\alpha)x^{\tau+1}} \int_0^\infty \lambda^{\tau+\tau\alpha-1} \exp[-\lambda^\tau(x^{-\tau}+\theta^{-\tau})]\, d\lambda \\
&= \frac{\tau^2}{\theta^{\tau\alpha}\Gamma(\alpha)x^{\tau+1}} \int_0^\infty [y^{1/\tau}(x^{-\tau}+\theta^{-\tau})^{-1/\tau}]^{\tau+\tau\alpha-1} e^{-y} \\
&\quad \times y^{\tau^{-1}-1} \tau^{-1}(x^{-\tau}+\theta^{-\tau})^{-1/\tau}\, dy \\
&= \frac{\tau}{\theta^{\tau\alpha}\Gamma(\alpha)x^{\tau+1}(x^{-\tau}+\theta^{-\tau})^{\alpha+1}} \int_0^\infty y^\alpha e^{-y}\, dy \\
&= \frac{\tau\Gamma(\alpha+1)}{\theta^{\tau\alpha}\Gamma(\alpha)x^{\tau+1}(x^{-\tau}+\theta^{-\tau})^{\alpha+1}} \\
&= \frac{\tau\alpha\theta^\tau x^{\tau\alpha-1}}{(x^\tau+\theta^\tau)^{\alpha+1}}.
\end{aligned}$$

The third line is obtained by the transformation $y = \lambda^\tau(x^{-\tau}+\theta^{-\tau})$. The final line uses the fact that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. The result is an inverse Burr distribution. Note that this distribution applies to a particular driver. Another driver may have a different Weibull shape parameter τ: also, that driver's Weibull scale parameter Θ may have a different distribution and, in particular, a different mean. $\qquad\square$

5.2.5 Frailty Models

An important type of mixture distribution is a **frailty model**. Although the physical motivation for this particular type of mixture is originally from the analysis of lifetime distributions in survival analysis, the resulting mathematical convenience implies that the approach may also be viewed as a useful way to generate new distributions by mixing.

We begin by introducing a **frailty** random variable $\Lambda > 0$ and define the conditional hazard rate (given $\Lambda = \lambda$) of X to be $h_{X|\Lambda}(x|\lambda) = \lambda a(x)$, where $a(x)$ is a known function of x (i.e. $a(x)$ is to be specified in a particular application). The frailty is meant to quantify uncertainty associated with the hazard rate, which by the preceding specification of the conditional hazard rate acts in a multiplicative manner.

The conditional survival function of $X|\Lambda$ is therefore

$$S_{X|\Lambda}(x|\lambda) = e^{-\int_0^x h_{X|\Lambda}(t|\lambda)dt} = e^{-\lambda A(x)},$$

where $A(x) = \int_0^x a(t)dt$. In order to specify the mixture distribution (i.e. the marginal distribution of X), we define the moment generating function of the frailty random variable Λ to be $M_\Lambda(z) = E(e^{z\Lambda})$. Then, the marginal survival function is

$$S_X(x) = E[e^{-\Lambda A(x)}] = M_\Lambda[-A(x)], \tag{5.3}$$

and, obviously, $F_X(x) = 1 - S_X(x)$.

The type of mixture to be used determines the choice of $a(x)$ and, hence, $A(x)$. The most important subclass of the frailty models is the class of exponential mixtures with $a(x) = 1$ and $A(x) = x$, so that $S_{X|\Lambda}(x|\lambda) = e^{-\lambda x}$, $x \geq 0$. Other useful mixtures include Weibull mixtures with $a(x) = \gamma x^{\gamma-1}$ and $A(x) = x^\gamma$.

Evaluation of the frailty distribution requires an expression for the moment generating function $M_\Lambda(z)$ of Λ. The most common choice is gamma frailty, but other choices such as inverse Gaussian frailty are also used.

■ **EXAMPLE 5.7**

Let Λ have a gamma distribution and let $X|\Lambda$ have a Weibull distribution with conditional survival function $S_{X|\Lambda}(x|\lambda) = e^{-\lambda x^\gamma}$. Determine the unconditional or marginal distribution of X.

In this case, it follows from Example 3.7 that the gamma moment generating function is $M_\Lambda(z) = (1 - \theta z)^{-\alpha}$, and from (5.3) it follows that X has survival function

$$S_X(x) = M_\Lambda(-x^\gamma) = (1 + \theta x^\gamma)^{-\alpha}.$$

This is a Burr distribution (see Appendix A), with the usual parameter θ replaced by $\theta^{-1/\gamma}$. Note that when $\gamma = 1$, this is an exponential mixture which is a Pareto distribution, considered previously in Example 5.4. □

As mentioned earlier, mixing tends to create heavy-tailed distributions and, in particular, a mixture of distributions that all have decreasing hazard rates also has a decreasing hazard rate. In Exercise 5.16, you are asked to prove this fact for frailty models. For further details on frailty models, see Hougaard [59].

5.2.6 Splicing

Another method for creating a new distribution is by splicing. This approach is similar to mixing in that it might be believed that two or more separate processes are responsible for generating the losses. With mixing, the various processes operate on subsets of the population. Once the subset is identified, a simple loss model suffices. For splicing, the processes differ with regard to the loss amount. That is, one model governs the behavior of losses in some interval of possible losses while other models cover the other intervals. Definition 5.8 makes this precise.

Definition 5.8 *A **k-component spliced distribution** has a density function that can be expressed as follows:*

$$f_X(x) = \begin{cases} a_1 f_1(x), & c_0 < x < c_1, \\ a_2 f_2(x), & c_1 < x < c_2, \\ \vdots & \vdots \\ a_k f_k(x), & c_{k-1} < x < c_k. \end{cases}$$

For $j = 1, \ldots, k$, each $a_j > 0$ and each $f_j(x)$ must be a legitimate density function with all probability on the interval (c_{j-1}, c_j). Also, $a_1 + \cdots + a_k = 1$.

■ **EXAMPLE 5.8**

Demonstrate that Model 5 in Section 2.2 is a two-component spliced model.

The density function is

$$f(x) = \begin{cases} 0.01, & 0 \le x < 50, \\ 0.02, & 50 \le x < 75, \end{cases}$$

and the spliced model is created by letting $f_1(x) = 0.02$, $0 \le x < 50$, which is a uniform distribution on the interval from 0 to 50, and $f_2(x) = 0.04$, $50 \le x < 75$, which is a uniform distribution on the interval from 50 to 75. The coefficients are then $a_1 = 0.5$ and $a_2 = 0.5$. □

It was not necessary to use density functions and coefficients, but this is one way to ensure that the result is a legitimate density function. When using parametric models, the motivation for splicing is that the tail behavior may be inconsistent with the behavior for small losses. For example, experience (based on knowledge beyond that available in the current, perhaps small, data set) may indicate that the tail follows the Pareto distribution, but there is a positive mode more in keeping with the lognormal or inverse Gaussian distributions. A second instance is when there is a large amount of data below some value but a limited amount of information elsewhere. We may want to use the empirical distribution (or a smoothed version of it) up to a certain point and a parametric model beyond that value. Definition 5.8 is appropriate when the break points $c_0, \ldots, c_k$ are known in advance.

Another way to construct a spliced model is to use standard distributions over the range from c_0 to c_k. Let $g_j(x)$ be the jth such density function. Then, in Definition 5.8, replace $f_j(x)$ with $g_j(x)/[G(c_j) - G(c_{j-1})]$. This formulation makes it easier to have the break points become parameters that can be estimated.
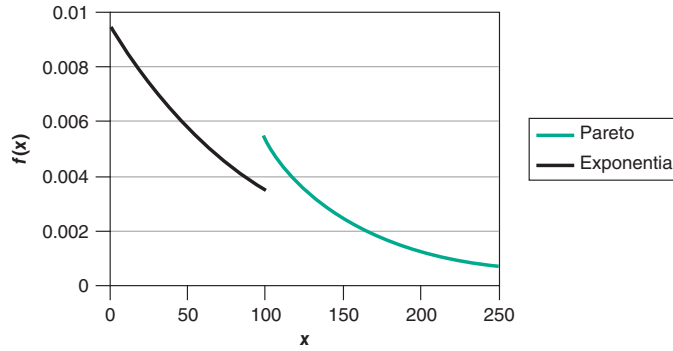
Figure 5.1   The two-component spliced density.

Neither approach to splicing ensures that the resulting density function will be continuous (i.e. the components will meet at the break points). Such a restriction could be added to the specification.

■ **EXAMPLE 5.9**

Create a two-component spliced model using an exponential distribution from 0 to c and a Pareto distribution (using γ in place of θ) from c to ∞.

The basic format is

$$f_X(x) = \begin{cases} a_1 \dfrac{\theta^{-1}e^{-x/\theta}}{1 - e^{-c/\theta}}, & 0 < x < c, \\[2ex] a_2 \dfrac{\alpha\gamma^\alpha(x+\gamma)^{-\alpha-1}}{\gamma^\alpha(c+\gamma)^{-\alpha}}, & c < x < \infty. \end{cases}$$

However, we must force the density function to integrate to 1. All that is needed is to let $a_1 = v$ and $a_2 = 1 - v$. The spliced density function becomes

$$f_X(x) = \begin{cases} v \dfrac{\theta^{-1}e^{-x/\theta}}{1 - e^{-c/\theta}}, & 0 < x < c, \\[2ex] (1-v)\dfrac{\alpha(c+\gamma)^\alpha}{(x+\gamma)^{\alpha+1}}, & c < x < \infty \end{cases}, \quad \theta, \alpha, \gamma, c > 0, \; 0 < v < 1.$$

Figure 5.1 illustrates this density function using the values $c = 100$, $v = 0.6$, $\theta = 100$, $\gamma = 200$, and $\alpha = 4$. It is clear that this density is not continuous. □

5.2.7 Exercises

5.1 Let X have cdf $F_X(x) = 1 - (1 + x)^{-\alpha}$, $x, \alpha > 0$. Determine the pdf and cdf of $Y = \theta X$.

5.2 (*) One hundred observed claims in 1995 were arranged as follows: 42 were between 0 and 300, 3 were between 300 and 350, 5 were between 350 and 400, 5 were between 400 and 450, 0 were between 450 and 500, 5 were between 500 and 600, and the remaining 40 were above 600. For the next three years, all claims are inflated by 10% per year. Based on the empirical distribution from 1995, determine a range for the probability that a claim exceeds 500 in 1998 (there is not enough information to determine the probability exactly).

5.3 Let X have a Pareto distribution. Determine the cdf of the inverse, transformed, and inverse transformed distributions. Check Appendix A to determine if any of these distributions have special names.

5.4 Let X have a loglogistic distribution. Demonstrate that the inverse distribution also has a loglogistic distribution. Therefore, there is no need to identify a separate inverse loglogistic distribution.

5.5 Let Y have a lognormal distribution with parameters μ and σ. Let $Z = \theta Y$. Show that Z also has a lognormal distribution and, therefore, the addition of a third parameter has not created a new distribution.

5.6 (*) Let X have a Pareto distribution with parameters α and θ. Let $Y = \ln(1 + X/\theta)$. Determine the name of the distribution of Y and its parameters.

5.7 Venter [124] notes that if X has a transformed gamma distribution and its scale parameter θ has an inverse transformed gamma distribution (where the parameter τ is the same in both distributions), the resulting mixture has the transformed beta distribution. Demonstrate that this is true.

5.8 (*) Let N have a Poisson distribution with mean Λ. Let Λ have a gamma distribution with mean 1 and variance 2. Determine the unconditional probability that $N = 1$.

5.9 (*) Given a value of $\Theta = \theta$, the random variable X has an exponential distribution with hazard rate function $h(x) = \theta$, a constant. The random variable Θ has a uniform distribution on the interval $(1, 11)$. Determine $S_X(0.5)$ for the unconditional distribution.

5.10 (*) Let N have a Poisson distribution with mean Λ. Let Λ have a uniform distribution on the interval $(0, 5)$. Determine the unconditional probability that $N \geq 2$.

5.11 (*) A two-point mixed distribution has, with probability p, a binomial distribution with parameters $m = 2$ and $q = 0.5$, and with probability $1 - p$, a binomial distribution with parameters $m = 4$ and $q = 0.5$. Determine, as a function of p, the probability that this random variable takes on the value 2.

5.12 Determine the probability density function and the hazard rate of the frailty distribution.

5.13 Suppose that $X|\Lambda$ has a Weibull survival function $S_{X|\Lambda}(x|\lambda) = e^{-\lambda x^{\gamma}}$, $x \geq 0$, and Λ has an exponential distribution. Demonstrate that the unconditional distribution of X is loglogistic.

5.14 Consider the exponential–inverse Gaussian frailty model with

$$a(x) = \frac{\theta}{2\sqrt{1 + \theta x}}, \quad \theta > 0.$$

(a) Verify that the conditional hazard rate $h_{X|\Lambda}(x|\lambda)$ of $X|\Lambda$ is indeed a valid hazard rate.

   (b) Determine the conditional survival function $S_{X|\Lambda}(x|\lambda)$.

   (c) If Λ has a gamma distribution with parameters $\theta = 1$ and α replaced by 2α, determine the marginal or unconditional survival function of X.

   (d) Use (c) to argue that a given frailty model may arise from more than one combination of conditional distributions of $X|\Lambda$ and frailty distributions of Λ.

5.15 Suppose that X has survival function $S_X(x) = 1 - F_X(x)$, given by (5.3). Show that $S_1(x) = F_X(x)/[E(\Lambda)A(x)]$ is again a survival function of the form given by (5.3), and identify the distribution of Λ associated with $S_1(x)$.

5.16 Fix $s \geq 0$, and define an "Esscher-transformed" frailty random variable Λ_s with probability density function (or discrete probability mass function in the discrete case) $f_{\Lambda_s}(\lambda) = e^{-s\lambda} f_\Lambda(\lambda)/M_\Lambda(-s)$, $\lambda \geq 0$.

   (a) Show that Λ_s has moment generating function

$$M_{\Lambda_s}(z) = E(e^{z\Lambda_s}) = \frac{M_\Lambda(z - s)}{M_\Lambda(-s)}.$$

   (b) Define the cumulant generating function of Λ to be

$$c_\Lambda(z) = \ln[M_\Lambda(z)],$$

   and use (a) to prove that

$$c_\Lambda'(-s) = E(\Lambda_s) \text{ and } c_\Lambda''(-s) = \text{Var}(\Lambda_s).$$

   (c) For the frailty model with survival function given by (5.3), prove that the associated hazard rate may be expressed as $h_X(x) = a(x)c_\Lambda'[-A(x)]$, where c_Λ is defined in (b).

   (d) Use (c) to show that

$$h_X'(x) = a'(x)c_\Lambda'[-A(x)] - [a(x)]^2 c_\Lambda''[-A(x)].$$

   (e) Prove using (d) that if the conditional hazard rate $h_{X|\Lambda}(x|\lambda)$ is nonincreasing in x, then $h_X(x)$ is also nonincreasing in x.

5.17 Write the density function for a two-component spliced model in which the density function is proportional to a uniform density over the interval from 0 to 1,000 and is proportional to an exponential density function from 1,000 to ∞. Ensure that the resulting density function is continuous.

5.18 Let X have pdf $f(x) = \exp(-|x/\theta|)/2\theta$ for $-\infty < x < \infty$. Let $Y = e^X$. Determine the pdf and cdf of Y.

5.19 (*) Losses in 1993 follow the density function $f(x) = 3x^{-4}$, $x \geq 1$, where x is the loss in millions of dollars. Inflation of 10% impacts all claims uniformly from 1993 to 1994. Determine the cdf of losses for 1994 and use it to determine the probability that a 1994 loss exceeds 2,200,000.

5.20 Consider the inverse Gaussian random variable X with pdf (from Appendix A)

$$f(x) = \sqrt{\frac{\theta}{2\pi x^3}} \exp\left[-\frac{\theta}{2x}\left(\frac{x-\mu}{\mu}\right)^2\right], \quad x > 0,$$

where $\theta > 0$ and $\mu > 0$ are parameters.

(a) Derive the pdf of the reciprocal inverse Gaussian random variable $1/X$.

(b) Prove that the "joint" moment generating function of X and $1/X$ is given by

$$M(z_1, z_2) = E\left(e^{z_1 X + z_2 X^{-1}}\right)$$

$$= \sqrt{\frac{\theta}{\theta - 2z_2}} \exp\left[\frac{\theta - \sqrt{(\theta - 2\mu^2 z_1)(\theta - 2z_2)}}{\mu}\right],$$

where $z_1 < \theta/(2\mu^2)$ and $z_2 < \theta/2$.

(c) Use (b) to show that the moment generating function of X is

$$M_X(z) = E\left(e^{zX}\right) = \exp\left[\frac{\theta}{\mu}\left(1 - \sqrt{1 - \frac{2\mu^2}{\theta}z}\right)\right], \quad z < \frac{\theta}{2\mu^2}.$$

(d) Use (b) to show that the reciprocal inverse Gaussian random variable $1/X$ has moment generating function

$$M_{1/X}(z) = E\left(e^{zX^{-1}}\right)$$

$$= \sqrt{\frac{\theta}{\theta - 2z}} \exp\left[\frac{\theta}{\mu}\left(1 - \sqrt{1 - \frac{2}{\theta}z}\right)\right], \quad z < \frac{\theta}{2}.$$

Hence prove that $1/X$ has the same distribution as $Z_1 + Z_2$, where Z_1 has a gamma distribution, Z_2 has an inverse Gaussian distribution, and Z_1 is independent of Z_2. Also, identify the gamma and inverse Gaussian parameters in this representation.

(e) Use (b) to show that

$$Z = \frac{1}{X}\left(\frac{X-\mu}{\mu}\right)^2$$

has a gamma distribution with parameters $\alpha = \frac{1}{2}$ and the usual parameter θ (in Appendix A) replaced by $2/\theta$.

(f) For the mgf of the inverse Gaussian random variable X in (c), prove by induction on k that, for $k = 1, 2, \ldots$, the kth derivative of the mgf is

$$M_X^{(k)}(z) = M_X(z) \sum_{n=0}^{k-1} \frac{(k+n-1)!}{(k-n-1)!n!} \left(\frac{1}{2}\right)^{\frac{k+3n}{2}} \theta^{\frac{k-n}{2}} \left(\frac{\theta}{2\mu^2} - z\right)^{-\frac{k+n}{2}}$$

and hence that the inverse Gaussian random variable has integer moments

$$E[X^k] = \sum_{n=0}^{k-1} \frac{(k+n-1)!}{(k-n-1)!n!} \frac{\mu^{n+k}}{(2\theta)^n}, \quad k = 1, 2, \ldots.$$

(g) The modified Bessel function, $K_\lambda(x)$ may be defined, for half-integer values of the index parameter λ, by $K_{-\lambda}(x) = K_\lambda(x)$, together with

$$K_{m+\frac{1}{2}}(x) = \sqrt{\frac{\pi}{2x}} e^{-x} \sum_{j=0}^{m} \frac{(m+j)!}{(m-j)!j!} \left(\frac{1}{2x}\right)^j, \quad m = 0, 1, \dots .$$

Use part (f) to prove that, for $\alpha > 0$, $\theta > 0$, and $m = 0, 1, \dots$,

$$\int_0^\infty x^{m-\frac{3}{2}} e^{-\alpha x - \frac{\theta}{2x}} \, dx = 2 \left(\frac{\theta}{2\alpha}\right)^{\frac{m}{2}-\frac{1}{4}} K_{m-\frac{1}{2}}\left(\sqrt{2\alpha\theta}\right).$$

5.3   Selected Distributions and Their Relationships

5.3.1   Introduction

There are many ways to organize distributions into groups. Families such as Pearson (12 types), Burr (12 types), Stoppa (5 types), and Dagum (11 types) are discussed in Chapter 2 of [69]. The same distribution can appear in more than one system, indicating that there are many relations among the distributions beyond those presented here. The systems presented in Section 5.3.2 are particularly useful for actuarial modeling because all the members have support on the positive real line and all tend to be skewed to the right. For a comprehensive set of continuous distributions, the two volumes by Johnson, Kotz, and Balakrishnan [63, 64] are a valuable reference. In addition, there are entire books devoted to single distributions (such as Arnold [6] for the Pareto distribution). Leemis and McQueston [78] present 76 distributions on one page, with arrows showing all the various relationships.

5.3.2   Two Parametric Families

As noted when defining parametric families, many of the distributions presented in this section and in Appendix A are special cases of others. For example, a Weibull distribution with $\tau = 1$ and θ arbitrary is an exponential distribution. Through this process, many of our distributions can be organized into groupings, as illustrated in Figures 5.2 and 5.3. The transformed beta family includes two special cases of a different nature. The paralogistic and inverse paralogistic distributions are created by setting the two nonscale parameters of the Burr and inverse Burr distributions equal to each other rather than to a specified value.

5.3.3   Limiting Distributions

The classification in Section 5.3.2 involved distributions that are special cases of other distributions. Another way to relate distributions is to see what happens as parameters go to their limiting values of zero or infinity.

■   **EXAMPLE 5.10**

Show that the transformed gamma distribution is a limiting case of the transformed beta distribution as $\theta \to \infty$, $\alpha \to \infty$, and $\theta/\alpha^{1/\gamma} \to \xi$, a constant.
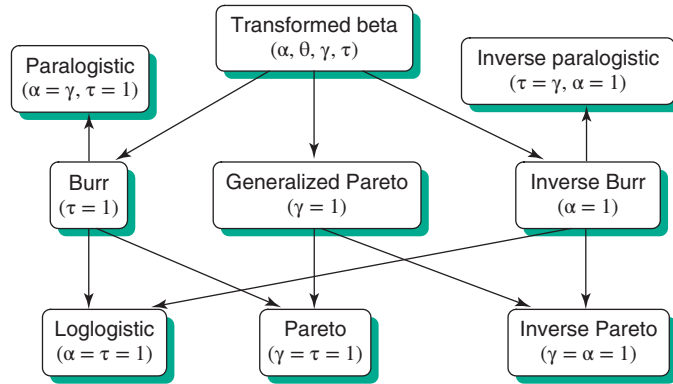
Figure 5.2   The transformed beta family.
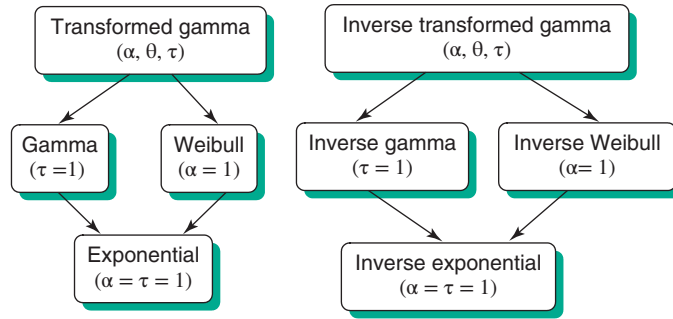
Figure 5.3   The transformed/inverse transformed gamma family.

The demonstration relies on two facts concerning limits:

$$\lim_{\alpha \to \infty} \frac{e^{-\alpha} \alpha^{\alpha-1/2} (2\pi)^{1/2}}{\Gamma(\alpha)} = 1 \tag{5.4}$$

and

$$\lim_{a \to \infty} \left(1 + \frac{x}{a}\right)^{a+b} = e^x. \tag{5.5}$$

The limit in (5.4) is known as Stirling's formula and provides an approximation for the gamma function. The limit in (5.5) is a standard result found in most calculus texts.

To ensure that the ratio $\theta/\alpha^{1/\gamma}$ goes to a constant, it is sufficient to force it to be constant as α and θ become larger and larger. This can be accomplished by substituting $\xi \alpha^{1/\gamma}$ for θ in the transformed beta pdf and then letting $\alpha \to \infty$. The first steps, which

also include using Stirling's formula to replace two of the gamma function terms, are

$$f(x) = \frac{\Gamma(\alpha + \tau)\gamma x^{\gamma\tau-1}}{\Gamma(\alpha)\Gamma(\tau)\theta^{\gamma\tau}(1 + x^\gamma \theta^{-\gamma})^{\alpha+\tau}}$$

$$= \frac{e^{-\alpha-\tau}(\alpha+\tau)^{\alpha+\tau-1/2}(2\pi)^{1/2}\gamma x^{\gamma\tau-1}}{e^{-\alpha}\alpha^{\alpha-1/2}(2\pi)^{1/2}\Gamma(\tau)(\xi\alpha^{1/\gamma})^{\gamma\tau}(1 + x^\gamma \xi^{-\gamma}\alpha^{-1})^{\alpha+\tau}}$$

$$= \frac{e^{-\tau}[(\alpha+\tau)/\alpha]^{\alpha+\tau-1/2}\gamma x^{\gamma\tau-1}}{\Gamma(\tau)\xi^{\gamma\tau}\left[1 + (x/\xi)^\gamma/\alpha\right]^{\alpha+\tau}}.$$

The two limits,

$$\lim_{\alpha\to\infty}\left(1 + \frac{\tau}{\alpha}\right)^{\alpha+\tau-1/2} = e^\tau \text{ and } \lim_{\alpha\to\infty}\left[1 + \frac{(x/\xi)^\gamma}{\alpha}\right]^{\alpha+\tau} = e^{(x/\xi)^\gamma},$$

can be substituted to yield

$$\lim_{\alpha\to\infty} f(x) = \frac{\gamma x^{\gamma\tau-1}e^{-(x/\xi)^\gamma}}{\Gamma(\tau)\xi^{\gamma\tau}},$$

which is the pdf of the transformed gamma distribution.   □

With a similar argument, the inverse transformed gamma distribution is obtained by letting τ go to infinity instead of α (see Exercise 5.23).

Because the Burr distribution is a transformed beta distribution with $\tau = 1$, its limiting case is the transformed gamma with $\tau = 1$ (using the parameterization in the previous example), which is the Weibull distribution. Similarly, the inverse Burr has the inverse Weibull as a limiting case. Finally, letting $\tau = \gamma = 1$ shows that the limiting case for the Pareto distribution is the exponential (and similarly for their inverse distributions).

As a final illustration of a limiting case, consider the transformed gamma distribution as parameterized previously. Let $\gamma^{-1}\sqrt{\xi^\gamma} \to \sigma$ and $\gamma^{-1}(\xi^\gamma\tau - 1) \to \mu$. If this is done by letting $\tau \to \infty$ (so both γ and ξ must go to zero), the limiting distribution will be lognormal.

In Figure 5.4, some of the limiting and special case relationships are shown. Other interesting facts about the various distributions are also given.[1]

5.3.4 Two Heavy-Tailed Distributions

With an increased emphasis on risk management, two distributions have received particular attention as being **extreme value distributions**. This a well-developed area of study, but here we present the results without going into the theory. The first setting for developing such distributions is to examine the maximum observation from a sample of independently and identically distributed (i.i.d.) observations. In insurance applications, knowing something about the largest claim that might be paid provides information about the risk. Note that this is not the same as measuring VaR because the quantile being examined depends on the sample size. The Fisher–Tippett theorem [40] states that in the limit the maximum (properly scaled) will have one of only three possible distributions. The only one that is interesting for actuarial applications is the **Fréchet distribution**, which is identical to

---

[1]Thanks to Dave Clark of Munich Reinsurance America, Inc. for creating this graphic.
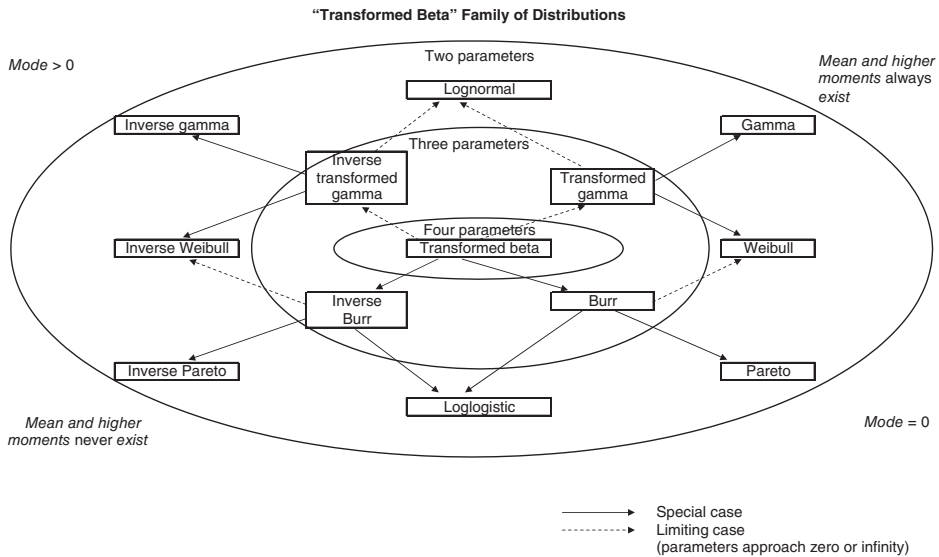
Figure 5.4   Distributional relationships and characteristics.

what we have called the inverse Weibull distribution. Thus, if the goal is to model the maximum observation from a random sample, the inverse Weibull distribution is likely to be a good candidate. Intuitively, the maximum from a random sample (particularly one from a heavy-tailed distribution) will be heavy tailed and thus the inverse Weibull is also confirmed as a good choice when a heavy-tailed model is needed.

The second setting examines the excess loss random variable as the truncation point is increased. The Balkema–de Haan–Pickands theorem (see Balkema and de Haan [11] and Pickands [102]) states that in the limit this distribution (properly scaled) converges to one of three distributions. Two are relevant for insurance observations. One is the exponential distribution, which is the limiting case for lighter-tailed distributions. The other is what we term the Pareto distribution (and is called the generalized Pareto distribution in extreme value theory, which is not the same as the distribution of that name in this book). The insurance situation of interest is when there is a high deductible, as often occurs in reinsurance contracts. If there is no interest in the distribution for moderate losses, the Pareto distribution is likely to be a good model for large losses. Again, even if this is not the situation of interest, this development provides further confirmation of the Pareto distribution as a good model for heavy-tailed losses.

More detail about extreme value distribution can be found in Section 5.6 of the third edition of this book [73] and also appears in the companion book to the fourth edition on advanced loss models [74].

5.3.5   Exercises

5.21 For a Pareto distribution, let both α and θ go to infinity with the ratio α/θ held constant. Show that the result is an exponential distribution.

5.22 Determine the limiting distribution of the generalized Pareto distribution as α and θ both go to infinity.

5.23 Show that as $\tau \to \infty$ in the transformed beta distribution, the result is the inverse transformed gamma distribution.

5.4   The Linear Exponential Family

A large parametric family that includes many of the distributions of interest to actuaries has special use in Bayesian analysis (Section 13.3) and in credibility (Section 17.7). The definition is as follows.

Definition 5.9 *A random variable X (discrete or continuous) has a distribution from the* ***linear exponential family*** *if its pdf may be parameterized in terms of a parameter θ and expressed as*

$$f(x; \theta) = \frac{p(x)e^{r(\theta)x}}{q(\theta)}. \tag{5.6}$$

The function $p(x)$ depends only on x (not on θ), and the function $q(\theta)$ is a normalizing constant. Also, the support of the random variable must not depend on θ. The parameter $r(\theta)$ is called the ***canonical parameter*** *of the distribution.*

Many standard distributions are of this form, as shown in the following examples.

■ **EXAMPLE 5.11**

Show that the normal distribution is a member of the linear exponential family.

The pdf is, letting the mean be θ,

$$f(x; \theta) = (2\pi v)^{-1/2} \exp\left[-\frac{1}{2v}(x-\theta)^2\right]$$

$$= (2\pi v)^{-1/2} \exp\left(-\frac{x^2}{2v} + \frac{\theta}{v}x - \frac{\theta^2}{2v}\right)$$

$$= \frac{\left[(2\pi v)^{-1/2} \exp\left(-\frac{x^2}{2v}\right)\right] \exp\left(\frac{\theta}{v}x\right)}{\exp\left(\frac{\theta^2}{2v}\right)},$$

which is of the form (5.6) with $p(x) = (2\pi v)^{-1/2} \exp[-x^2/(2v)]$, $r(\theta) = \theta/v$, and $q(\theta) = \exp[\theta^2/(2v)]$.   □

■ **EXAMPLE 5.12**

Show that the gamma distribution is a member of the linear exponential family.

The pdf is (from Appendix A)

$$f(x; \theta) = \frac{\theta^{-\alpha} x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha)},$$

which is of the form (5.6) with $r(\theta) = -1/\theta$, $q(\theta) = \theta^\alpha$, and $p(x) = x^{\alpha-1}/\Gamma(\alpha)$.   □

As is clear from Examples 5.11 and 5.12, there may be other parameters in addition to θ, but they will have no explicit role in any subsequent analysis involving the linear exponential family.

We now find the mean and variance of the distribution defined by (5.6). First, note that

$$\ln f(x; \theta) = \ln p(x) + r(\theta)x - \ln q(\theta).$$

Differentiate with respect to θ to obtain

$$\frac{\partial}{\partial \theta} f(x; \theta) = \left[r'(\theta)x - \frac{q'(\theta)}{q(\theta)} \right] f(x; \theta). \tag{5.7}$$

Integrate over the range of x (known not to depend on θ) to obtain

$$\int \frac{\partial}{\partial \theta} f(x; \theta)\, dx = r'(\theta) \int x f(x; \theta)\, dx - \frac{q'(\theta)}{q(\theta)} \int f(x; \theta)\, dx.$$

On the left-hand side, interchange the order of differentiation and integration to obtain

$$\frac{\partial}{\partial \theta} \left[\int f(x; \theta)\, dx \right] = r'(\theta) \int x f(x; \theta)\, dx - \frac{q'(\theta)}{q(\theta)} \int f(x; \theta)\, dx.$$

We know that $\int f(x; \theta)\, dx = 1$ and $\int x f(x; \theta)\, dx = \mathrm{E}(X)$ and thus

$$\frac{\partial}{\partial \theta}(1) = r'(\theta)\mathrm{E}(X) - \frac{q'(\theta)}{q(\theta)}.$$

In other words, the mean is

$$\mathrm{E}(X) = \mu(\theta) = \frac{q'(\theta)}{r'(\theta)q(\theta)}. \tag{5.8}$$

To obtain the variance, (5.7) may first be rewritten as

$$\frac{\partial}{\partial \theta} f(x; \theta) = r'(\theta)[x - \mu(\theta)]f(x; \theta).$$

Differentiate again with respect to θ, to obtain

$$\frac{\partial^2}{\partial \theta^2} f(x; \theta) = r''(\theta)[x - \mu(\theta)]f(x; \theta) - r'(\theta)\mu'(\theta)f(x; \theta)$$

$$+ [r'(\theta)]^2 [x - \mu(\theta)]^2 f(x; \theta).$$

Again, integrate over the range of x to obtain

$$\int \frac{\partial^2}{\partial \theta^2} f(x, \theta)\, dx = r''(\theta) \int [x - \mu(\theta)]f(x; \theta)\, dx - r'(\theta)\mu'(\theta) \int f(x; \theta)\, dx$$

$$+ [r'(\theta)]^2 \int [x - \mu(\theta)]^2 f(x; \theta)\, dx.$$

In other words, because (5.8) holds,

$$[r'(\theta)]^2 \int [x - \mu(\theta)]^2 f(x; \theta)\, dx = r'(\theta)\mu'(\theta) + \frac{\partial^2}{\partial \theta^2} \int f(x; \theta)\, dx.$$

Because $\mu(\theta)$ is the mean, the left-hand side is the variance (by definition) multiplied by $[r'(\theta)]^2$, and then, because the second term on the right-hand side is zero, we obtain

$$\text{Var}(X) = v(\theta) = \frac{\mu'(\theta)}{r'(\theta)}. \tag{5.9}$$

Although the emphasis in this section has been on continuous distributions, the linear exponential family also includes discrete distributions such as the Poisson, binomial, and negative binomial (among others). In this case, $f(x;\theta)$ in (5.6) is a probability function (pf) rather than a pdf, but the mean and variance are still given by (5.8) and (5.9), respectively (their derivation is identical, with integrals replaced by sums).

5.4.1   Exercises

5.24   Use the results of Example 5.12 and equations (5.8) and (5.9) to derive the mean and variance of the gamma distribution.

5.25   Consider the generalization of (5.6) given by

$$f(x;\theta) = \frac{p(m,x)e^{mr(\theta)x}}{[q(\theta)]^m},$$

where m is a known parameter. This distribution is said to be of the **exponential dispersion** type. Prove that the mean is still given by (5.8) but the variance is given by $v(\theta)/m$, where $v(\theta)$ is given by (5.9).

5.26   For $j = 1, 2, \ldots, m$, let X_j have a pf of the form (5.6), namely

$$f_j(x;\theta) = \frac{p_j(x)e^{r(\theta)x}}{q_j(\theta)},$$

and assume that $X_1, \ldots, X_m$ are independent.

  (a) Show that $S = X_1 + X_2 + \cdots + X_m$ also has a pdf of the form (5.6) with

$$q(\theta) = \prod_{j=1}^{m} q_j(\theta).$$

  (b) Now assume that $X_1, X_2, \ldots, X_m$ are also identically distributed. Show that $\bar{X} = S_m/m$ has a pdf of the form given in Exercise 5.25.

6

DISCRETE DISTRIBUTIONS

6.1 Introduction

The purpose of this chapter is to introduce a large class of counting distributions. Counting distributions are discrete distributions with probabilities only on the nonnegative integers, that is, probabilities are defined only at the points $0, 1, 2, 3, 4, \ldots$. In an insurance context, counting distributions can be used to describe the number of events such as losses to the insured or claims to the insurance company. An understanding of both the number of claims and the size of claims provides a deeper insight into a variety of issues surrounding insurance payments than if information is only available about total losses. The description of total losses in terms of numbers and amounts separately makes it possible to address issues of modification of an insurance contract. Another reason for separating numbers and amounts of claims is that models for the number of claims are fairly easy to obtain, and experience has shown that the commonly used distributions really do model the propensity to generate losses.

We now formalize some of the notation used for models for discrete phenomena. The **probability function** (pf) p_k denotes the probability that exactly k events (such as claims or losses) occur. Let N be a random variable representing the number of such events.

Then,

$$p_k = \Pr(N = k), \qquad k = 0, 1, 2, \ldots .$$

As a reminder, the probability generating function (pgf) of a discrete random variable N with pf p_k is

$$P(z) = P_N(z) = \mathrm{E}\left(z^N\right) = \sum_{k=0}^{\infty} p_k z^k. \tag{6.1}$$

As is true with the moment generating function, the pgf can be used to generate moments. In particular, $P'(1) = \mathrm{E}(N)$ and $P''(1) = \mathrm{E}[N(N-1)]$ (see Exercise 6.1). To see that the pgf really does generate probabilities, observe that

$$P^{(m)}(z) = \mathrm{E}\left(\frac{d^m}{dz^m} z^N\right) = \mathrm{E}[N(N-1)\cdots(N-m+1)z^{N-m}]$$

$$= \sum_{k=m}^{\infty} k(k-1)\cdots(k-m+1)z^{k-m}p_k$$

$$P^{(m)}(0) = m! p_m \text{ or } p_m = \frac{P^{(m)}(0)}{m!}.$$

6.1.1 Exercise

6.1 The moment generating function (mgf) for discrete variables is defined as

$$M_N(z) = \mathrm{E}\left(e^{zN}\right) = \sum_{k=0}^{\infty} p_k e^{zk}.$$

Demonstrate that $P_N(z) = M_N(\ln z)$. Use the fact that $\mathrm{E}(N^k) = M_N^{(k)}(0)$ to show that $P'(1) = \mathrm{E}(N)$ and $P''(1) = \mathrm{E}[N(N-1)]$.

6.2 The Poisson Distribution

The pf for the Poisson distribution is

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots .$$

The probability generating function from Example 3.8 is

$$P(z) = e^{\lambda(z-1)}, \qquad \lambda > 0.$$

The mean and variance can be computed from the probability generating function as follows:

$$\mathrm{E}(N) = P'(1) = \lambda$$

$$\mathrm{E}[N(N-1)] = P''(1) = \lambda^2$$

$$\mathrm{Var}(N) = \mathrm{E}[N(N-1)] + \mathrm{E}(N) - [\mathrm{E}(N)]^2$$

$$= \lambda^2 + \lambda - \lambda^2$$

$$= \lambda.$$

For the Poisson distribution, the variance is equal to the mean. The Poisson distribution and Poisson processes (which give rise to Poisson distributions) are discussed in many

textbooks on probability, statistics, and actuarial science, including Panjer and Willmot [100] and Ross [109].

The Poisson distribution has at least two additional useful properties. The first is given in the following theorem.

Theorem 6.1 *Let $N_1, \dots, N_n$ be independent Poisson variables with parameters $\lambda_1, \dots, \lambda_n$. Then, $N = N_1 + \cdots + N_n$ has a Poisson distribution with parameter $\lambda_1 + \cdots + \lambda_n$.*

Proof: The pgf of the sum of independent random variables is the product of the individual pgfs. For the sum of Poisson random variables, we have

$$P_N(z) = \prod_{j=1}^{n} P_{N_j}(z) = \prod_{j=1}^{n} \exp[\lambda_j(z-1)]$$

$$= \exp\left[\sum_{j=1}^{n} \lambda_j(z-1)\right]$$

$$= e^{\lambda(z-1)},$$

where $\lambda = \lambda_1 + \cdots + \lambda_n$. Just as is true with moment generating functions, the pgf is unique and, therefore, N must have a Poisson distribution with parameter λ. □

The second property is particularly useful in modeling insurance risks. Suppose that the number of claims in a fixed time period, such as one year, follows a Poisson distribution. Further suppose that the claims can be classified into m distinct types. For example, claims could be classified by size, such as those below a fixed limit and those above the limit. It turns out that, if we are interested in studying the number of claims above the limit, that distribution is also Poisson but with a new Poisson parameter.

This second property is also useful when considering removing or adding a part of an insurance coverage. Suppose that the number of claims for a complicated medical benefit coverage follows a Poisson distribution. Consider the "types" of claims to be the different medical procedures or medical benefits under the plan. If one of the benefits is removed from the plan, again it turns out that the distribution of the number of claims under the revised plan will still have a Poisson distribution, but with a new parameter.

In each of the cases mentioned in the previous paragraph, the number of claims of the different types will not only be Poisson distributed but will also be independent of each other, that is, the distributions of the number of claims above the limit and the number below the limit will be independent of each other. This is a somewhat surprising result. For example, suppose that we, the insurer, currently sell an insurance policy with a deductible of 50 and experience has indicated that a Poisson distribution with a certain parameter is a valid model for the number of payments. Further suppose that we are also comfortable with the assumption that the number of losses in a period also has the Poisson distribution but we do not know the parameter. Without additional information, it is impossible to infer the value of the Poisson parameter should the deductible be lowered or removed entirely. We now formalize these ideas in the following theorem.

Theorem 6.2 *Suppose that the number of events N is a Poisson random variable with mean λ. Further suppose that each event can be classified into one of m types with probabilities $p_1, \dots, p_m$ independent of all other events. Then, the number of events*

$N_1, \dots, N_m$ *corresponding to event types* $1, \dots, m$, *respectively, are mutually independent Poisson random variables with means* $\lambda p_1, \dots, \lambda p_m$, *respectively.*

Proof: For fixed $N = n$, the conditional joint distribution of $(N_1, \dots, N_m)$ is multinomial with parameters $(n, p_1, \dots, p_m)$. Also, for fixed $N = n$, the conditional marginal distribution of N_j is binomial with parameters (n, p_j).

The joint pf of $(N_1, \dots, N_m)$ is given by

$$
\begin{aligned}
\Pr(N_1 = n_1, \dots, N_m = n_m) &= Pr(N_1 = n_1, \dots, N_m = n_m | N = n) \\
&\quad \times \Pr(N = n) \\
&= \frac{n!}{n_1! n_2! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m} \frac{e^{-\lambda} \lambda^n}{n!} \\
&= \prod_{j=1}^{m} e^{-\lambda p_j} \frac{(\lambda p_j)^{n_j}}{n_j!},
\end{aligned}
$$

where $n = n_1 + n_2 + \cdots + n_m$. Similarly, the marginal pf of N_j is

$$
\begin{aligned}
\Pr(N_j = n_j) &= \sum_{n=n_j}^{\infty} \Pr(N_j = n_j | N = n) \Pr(N = n) \\
&= \sum_{n=n_j}^{\infty} \binom{n}{n_j} p_j^{n_j} (1 - p_j)^{n-n_j} \frac{e^{-\lambda} \lambda^n}{n!} \\
&= e^{-\lambda} \frac{(\lambda p_j)^{n_j}}{n_j!} \sum_{n=n_j}^{\infty} \frac{[\lambda(1 - p_j)]^{n-n_j}}{(n - n_j)!} \\
&= e^{-\lambda} \frac{(\lambda p_j)^{n_j}}{n_j!} e^{\lambda(1-p_j)} \\
&= e^{-\lambda p_j} \frac{(\lambda p_j)^{n_j}}{n_j!}.
\end{aligned}
$$

The joint pf is the product of the marginal pfs, establishing mutual independence.   □

■ **EXAMPLE 6.1**

In a study of medical insurance, the expected number of claims per individual policy is 2.3 and the number of claims is Poisson distributed. You are considering removing one medical procedure from the coverage under this policy. Based on historical studies, this procedure accounts for approximately 10% of the claims. Determine the new frequency distribution.

From Theorem 6.2, we know that the distribution of the number of claims expected under the revised insurance policy after removing the procedure from coverage is Poisson with mean $0.9(2.3) = 2.07$. Note that this does not imply a 10% reduction in premium, as the distribution of the amount of a claim may change now that this procedure has been removed.   □

6.3   The Negative Binomial Distribution

The negative binomial distribution has been used extensively as an alternative to the Poisson distribution. Like the Poisson distribution, it has positive probabilities on the nonnegative integers. Because it has two parameters, it has more flexibility in shape than the Poisson.

Definition 6.3 *The probability function of the **negative binomial distribution** is given by*

$$\Pr(N = k) = p_k = \binom{k + r - 1}{k} \left(\frac{1}{1 + \beta} \right)^r \left(\frac{\beta}{1 + \beta} \right)^k ,$$
$$k = 0, 1, 2, \dots , \quad r > 0, \ \beta > 0. \qquad (6.2)$$

The binomial coefficient is to be evaluated as

$$\binom{x}{k} = \frac{x(x - 1) \cdots (x - k + 1)}{k!}.$$

While k must be an integer, x may be any real number. When $x > k - 1$, it can also be written as

$$\binom{x}{k} = \frac{\Gamma(x + 1)}{\Gamma(k + 1)\Gamma(x - k + 1)},$$

which may be useful because $\ln \Gamma(x)$ is available in many spreadsheets, programming languages, and mathematics packages.

It is not difficult to show that the probability generating function for the negative binomial distribution is

$$P(z) = [1 - \beta(z - 1)]^{-r}.$$

From this, it follows that the mean and variance of the negative binomial distribution are

$$\mathrm{E}(N) = r\beta \text{ and } \mathrm{Var}(N) = r\beta(1 + \beta).$$

Because β is positive, the variance of the negative binomial distribution exceeds the mean. This relationship is in contrast to the Poisson distribution, for which the variance is equal to the mean. Thus, for a particular set of data, if the observed variance is larger than the observed mean, the negative binomial might be a better candidate than the Poisson distribution as a model to be used.

The negative binomial distribution is a generalization of the Poisson in at least two different ways, namely, as a mixed Poisson distribution with a gamma mixing distribution (demonstrated later in this section) and as a compound Poisson distribution with a logarithmic secondary distribution (see Section 7.1).

The **geometric distribution** is the special case of the negative binomial distribution when $r = 1$. The geometric distribution is, in some senses, the discrete analog of the continuous exponential distribution. Both the geometric and exponential distributions have an exponentially decaying probability function and, hence, the memoryless property, which can be interpreted in various contexts as follows. If the exponential distribution is a distribution of lifetimes, then the expected future lifetime is constant for any age. If the exponential distribution describes the size of insurance claims, then the memoryless property can be interpreted as follows: *Given that a claim exceeds a certain level d, the expected amount of the claim in excess of d is constant and so does not depend on d.* That

is, if a deductible of d is imposed, the expected payment per claim will be unchanged but, of course, the expected number of payments will decrease. If the geometric distribution describes the number of claims, then the memoryless property can be interpreted as follows: *Given that there are at least m claims, the probability distribution of the number of claims in excess of m does not depend on m.* Among continuous distributions, the exponential distribution is used to distinguish between *subexponential* distributions with heavy (or fat) tails and distributions with light (or thin) tails. Similarly for frequency distributions, distributions that decay in the tail slower than the geometric distribution are often considered to have heavy tails, whereas distributions that decay more rapidly than the geometric have light tails. The negative binomial distribution has a heavy tail (decays more slowly than the geometric distribution) when $r < 1$ and a lighter tail than the geometric distribution when $r > 1$.

As noted earlier, one way to create the negative binomial distribution is as a mixture of Poissons. Suppose that we know that a risk has a Poisson number of claims distribution when the risk parameter λ is known. Now treat λ as being the outcome of a random variable Λ. We denote the pdf/pf of Λ by $u(\lambda)$, where Λ may be continuous or discrete, and denote the cdf by $U(\lambda)$. The idea that λ is the outcome of a random variable can be justified in several ways. First, we can think of the population of risks as being heterogeneous with respect to the risk parameter Λ. In practice, this makes sense. Consider a portfolio of insurance policies with the same premium, such as a group of automobile drivers in the same rating category. Such categories are usually broad ranges, such as 0–7,500 miles driven per year, garaged in a rural area, commuting less than 50 miles per week, and so on. We know that not all drivers in the same rating category are the same, even though they may "appear" to be the same from the point of view of the insurer and are charged the same premium. The parameter λ measures the expected number of accidents for a given driver. If λ varies across the population of drivers, then we can think of the insured individual as a sample value drawn from the population of possible drivers. For a particular driver, λ is unknown to the insurer but follows some distribution, in this case $u(\lambda)$, over the population of drivers. The true value of λ is unobservable. We can only observe the number of accidents coming from the driver. There is now an additional degree of uncertainty, that is, uncertainty about the parameter.

This is the same mixing process that was discussed with regard to continuous distributions in Section 5.2.4. In some contexts, this is referred to as **parameter uncertainty**. In the Bayesian context, the distribution of Λ is called a **prior distribution** and the parameters of its distribution are sometimes called **hyperparameters**. The role of the distribution $u(\cdot)$ is very important in credibility theory, the subject of Chapter 16. When the parameter λ is unknown, the probability that exactly k claims will arise can be written as the expected value of the same probability but conditional on $\Lambda = \lambda$, where the expectation is taken with respect to the distribution of Λ. From the law of total probability, we can write

$$
\begin{aligned}
p_k &= \Pr(N = k) \\
&= \mathrm{E}[\Pr(N = k|\Lambda)] \\
&= \int_0^\infty \Pr(N = k|\Lambda = \lambda) u(\lambda)\, d\lambda \\
&= \int_0^\infty \frac{e^{-\lambda}\lambda^k}{k!} u(\lambda)\, d\lambda.
\end{aligned}
$$

Now suppose that Λ has a gamma distribution. Then,

$$p_k = \int_0^\infty \frac{e^{-\lambda}\lambda^k}{k!} \frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\theta}}}{\theta^\alpha\Gamma(\alpha)} \, d\lambda = \frac{1}{k!}\frac{1}{\theta^\alpha\Gamma(\alpha)}\int_0^\infty e^{-\lambda(1+\frac{1}{\theta})}\lambda^{k+\alpha-1}\,d\lambda.$$

From the definition of the gamma distribution in Appendix A, this expression can be evaluated as

$$p_k = \frac{\Gamma(k+\alpha)}{k!\Gamma(\alpha)}\frac{\theta^k}{(1+\theta)^{k+\alpha}}$$

$$= \binom{k+\alpha-1}{k}\left(\frac{\theta}{1+\theta}\right)^k\left(\frac{1}{1+\theta}\right)^\alpha.$$

This formula is of the same form as (6.2), demonstrating that the mixed Poisson, with a gamma mixing distribution, is the same as a negative binomial distribution.

It is worth noting that the Poisson distribution is a limiting case of the negative binomial distribution. To see this, let r go to infinity and β go to zero while keeping their product constant. Let $\lambda = r\beta$ be that constant. Substitution of $\beta = \lambda/r$ in the pgf leads to (using L'Hôpital's rule in lines 3 and 5)

$$\lim_{r\to\infty}\left[1 - \frac{\lambda(z-1)}{r}\right]^{-r} = \exp\left\{\lim_{r\to\infty}-r\ln\left[1 - \frac{\lambda(z-1)}{r}\right]\right\}$$

$$= \exp\left\{-\lim_{r\to\infty}\frac{\ln[1-\lambda(z-1)/r]}{r^{-1}}\right\}$$

$$= \exp\left\{\lim_{r\to\infty}\frac{[1-\lambda(z-1)/r]^{-1}\lambda(z-1)/r^2}{r^{-2}}\right\}$$

$$= \exp\left[\lim_{r\to\infty}\frac{r\lambda(z-1)}{r-\lambda(z-1)}\right]$$

$$= \exp\left\{\lim_{r\to\infty}[\lambda(z-1)]\right\}$$

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

which is the pgf of the Poisson distribution.

6.4   The Binomial Distribution

The **binomial distribution** is another counting distribution that arises naturally in claim number modeling. It possesses some properties different from those of the Poisson and the negative binomial that make it particularly useful. First, its variance is smaller than its mean, making it useful for data sets in which the observed sample variance is less than the sample mean. This property contrasts with the negative binomial, where the variance exceeds the mean, and it also contrasts with the Poisson distribution, where the variance is equal to the mean.

Second, it describes a physical situation in which m risks are each subject to claim or loss. We can formalize this situation as follows. Consider m independent and identical risks, each with probability q of making a claim.[1] This might apply to a life insurance

---

[1] It is more common to use p for the parameter of this distribution. Because we have used p for the probabilities and because q is the standard actuarial symbol for the probability of death, we have elected to use q as the parameter.

situation in which all the individuals under consideration are in the same mortality class, that is, they may all be male smokers at age 35 and duration 5 of an insurance policy. In that case, q is the probability that a person with those attributes will die in the next year. Then, the number of claims for a single person follows a **Bernoulli distribution**, a distribution with probability $1 - q$ at 0 and probability q at 1. The probability generating function of the number of claims per individual is then given by

$$P(z) = (1 - q)z^0 + qz^1 = 1 + q(z - 1).$$

Now, if there are m such independent individuals, then the probability generating functions can be multiplied together to give the probability generating function of the total number of claims arising from the group of m individuals. That probability generating function is

$$P(z) = [1 + q(z - 1)]^m, \qquad 0 < q < 1.$$

Then, from this it is easy to show that the probability of exactly k claims from the group is

$$p_k = \Pr(N = k) = \binom{m}{k} q^k (1 - q)^{m-k}, \quad k = 0, 1, \dots, m, \tag{6.3}$$

the pf for a binomial distribution with parameters m and q. From this Bernoulli trial framework, it is clear that at most m events (claims) can occur. Hence, the distribution only has positive probabilities on the nonnegative integers up to and including m.

Consequently, a sometimes useful attribute of the binomial distribution is that it has finite support, that is, the range of values for which there exist positive probabilities has finite length. This attribute may be useful, for instance, in modeling the number of individuals injured in an automobile accident or the number of family members covered under a health insurance policy. In each case, it is reasonable to have an upper limit on the range of possible values. It is useful also in connection with situations in which it is believed that it is unreasonable to assign positive probabilities beyond some point. For example, if we are modeling the number of accidents per automobile during a one-year period, it is probably physically impossible for there to be more than some number, say 12, of claims during the year, given the time it would take to repair the automobile between accidents. If a model with probabilities that extend beyond 12 were used, those probabilities should be very small, so that they have little impact on any decisions that are made.

The mean and variance of the binomial distribution are given by

$$E(N) = mq, \quad \text{Var}(N) = mq(1 - q).$$

6.5   The $(a, b, 0)$ Class

The following definition characterizes the members of this class of distributions.

Definition 6.4 *Let p_k be the pf of a discrete random variable. It is a member of the $(a, b, 0)$ class of distributions provided that there exist constants a and b such that*

$$p_k = \left(a + \frac{b}{k} \right) p_{k-1}, \qquad k = 1, 2, 3, \dots.$$

This recursion describes the relative size of successive probabilities in the counting distribution. The probability at zero, p_0, can be obtained from the recursive formula because the probabilities must sum to 1. The $(a, b, 0)$ class of distributions is a two-parameter class, the two parameters being a and b. The following example illustrates these ideas by demonstrating that the binomial distribution is a member of the $(a, b, 0)$ class.

■ **EXAMPLE 6.2**

Demonstrate that the binomial distribution is a member of the $(a, b, 0)$ class.

The binomial distribution with parameters m and q has probabilities

$$p_k = \binom{m}{k} q^k (1 - q)^{m-k}, \qquad k = 0, 1, \dots, m,$$

and $p_k = 0$, otherwise. The probabilities for $k = 1, 2, \dots, m$ can be rewritten as

$$
\begin{aligned}
p_k &= \frac{m!}{(m - k)!k!} q^k (1 - q)^{m-k} \\
&= \frac{m - k + 1}{k} \frac{q}{1 - q} \left\{ \frac{m}{[m - (k - 1)]!(k - 1)!} q^{k-1} (1 - q)^{m-(k-1)} \right\} \\
&= \frac{q}{1 - q} \left(-1 + \frac{m + 1}{k} \right) p_{k-1}.
\end{aligned}
$$

Hence, $p_k = (a + b/k)p_{k-1}$ holds for $k = 1, 2, \dots, m$ with $a = -q/(1 - q)$ and $b = (m + 1)q/(1 - q)$. To complete the example, we must verify that the recursion holds for $k = m + 1, m + 2, \dots$. For $k = m + 1$, we have

$$\left(a + \frac{b}{m + 1} \right) p_m = \left(-\frac{q}{1 - q} + \frac{q}{1 - q} \right) p_m = 0 = p_{m+1}.$$

For $k = m + 2, m + 3, \dots$ the recursion holds trivially, with both sides clearly being zero. This demonstrates that the binomial distribution is a member of the $(a, b, 0)$ class.

□

As in the above example, substituting in the probability function for the Poisson and negative binomial distributions on each side of the recursive formula in Definition 6.4, with the values of a and b given in Table 6.1, demonstrates that these two distributions are also members of the $(a, b, 0)$ class. In addition, Table 6.1 gives the values of p_0, the starting value for the recursion. The geometric distribution, the one-parameter special case ($r = 1$) of the negative binomial distribution, is also in the table.

It can be shown (see Panjer and Willmot [100, Chapter 6]) that these are the only possible distributions satisfying this recursive formula.

The recursive formula can be rewritten (if $p_{k-1} > 0$) as

$$k \frac{p_k}{p_{k-1}} = ak + b, \qquad k = 1, 2, 3, \dots.$$

The expression on the left-hand side is a linear function in k. Note from Table 6.1 that the slope a of the straight line is zero for the Poisson distribution, is negative for the binomial distribution, and is positive for the negative binomial distribution, including the geometric

Table 6.1   The members of the $(a, b, 0)$ class.

Distribution	a	b	p_0
Poisson	0	λ	$e^{-\lambda}$
Binomial	$-\dfrac{q}{1-q}$	$(m+1)\dfrac{q}{1-q}$	$(1-q)^m$
Negative binomial	$\dfrac{\beta}{1+\beta}$	$(r-1)\dfrac{\beta}{1+\beta}$	$(1+\beta)^{-r}$
Geometric	$\dfrac{\beta}{1+\beta}$	0	$(1+\beta)^{-1}$

special case. This relationship suggests a graphical way of indicating which of the three distributions might be selected for fitting to data. We begin by plotting

$$k\frac{\hat{p}_k}{\hat{p}_{k-1}} = k\frac{n_k}{n_{k-1}}$$

against k. The observed values should form approximately a straight line if one of these models is to be selected, and the value of the slope should be an indication of which of the models should be selected. Note that this cannot be done if any of the n_k are zero. Hence this procedure is less useful for a small number of observations.

■ **EXAMPLE 6.3**

Consider the accident data in Table 6.2, which is taken from Thyrion [120]. For the 9,461 automobile insurance policies studied, the number of accidents under the policy is recorded in the table. Also recorded in the table is the observed value of the quantity that should be linear.

Figure 6.1 plots the value of the quantity of interest against k, the number of accidents. It can be seen from the graph that the quantity of interest looks approximately linear except for the point at $k = 6$. The reliability of the quantities as k increases diminishes because the number of observations becomes small and the variability of the results grows, illustrating a weakness of this *ad hoc* procedure. Visually, all the points appear to have equal value. However, the points on the left are more reliable than the points on the right due to the larger number of observations. From the graph, it can be seen that the slope is positive and the data appear approximately linear, suggesting that the negative binomial distribution is an appropriate model. Whether or not the slope is significantly different from zero is also not easily judged from the graph. By rescaling the vertical axis of the graph, the slope can be made to look steeper and, hence, the slope could be made to appear to be significantly different from zero. Graphically, it is difficult to distinguish between the Poisson and the negative binomial distribution because the Poisson requires a slope of zero. However, we can say that the binomial distribution is probably not a good choice, since there is no evidence of a negative slope. In this case, it is advisable to fit both the Poisson and negative binomial distributions and conduct a more formal test to choose between them.   □

It is also possible to compare the appropriateness of the distributions by looking at the relationship of the variance to the mean. For this data set, the mean number of claims per policy is 0.2144. The variance is 0.2889. Because the variance exceeds the mean, the

Table 6.2   The accident profile from Thyrion [120].

Number of accidents, k	Number of policies, n_k	$k\dfrac{n_k}{n_{k-1}}$
0	7,840	
1	1,317	0.17
2	239	0.36
3	42	0.53
4	14	1.33
5	4	1.43
6	4	6.00
7	1	1.75
8+	0	
Total	9,461	
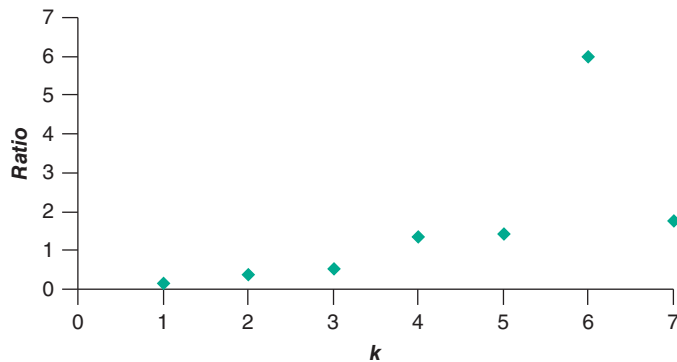
Figure 6.1   The plot of the ratio kn_k/n_{k-1} against k.

negative binomial should be considered as an alternative to the Poisson. Again, this is a qualitative comment because we have, at this point, no formal way of determining whether the variance is sufficiently larger than the mean to warrant use of the negative binomial. To do some formal analysis, Table 6.3 gives the results of maximum likelihood estimation (discussed in Chapters 11 and 12) of the parameters of the Poisson and negative binomial distributions and the negative loglikelihood in each case. In Chapter 15, formal selection methods are presented. They would indicate that the negative binomial is superior to the Poisson as a model for this data set. However, those methods also indicate that the negative binomial is not a particularly good model and, thus, some of the distributions yet to be introduced should be considered.

   In subsequent sections, we will expand the class of the distributions beyond the three discussed in this section by constructing more general models related to the Poisson, binomial, and negative binomial distributions.

6.5.1   Exercises

6.2 For each of the data sets in Exercises 12.3 and 12.5 in Section 12.7, calculate values similar to those in Table 6.2. For each, determine the most appropriate model from the $(a, b, 0)$ class.

Table 6.3   Comparison between Poisson and negative binomial models.

Distribution	Parameter estimates	−Loglikelihood
Poisson	$\hat{\lambda} = 0.2143537$	5,490.78
Negative binomial	$\hat{\beta} = 0.3055594$	5,348.04
	$\hat{r} = 0.7015122$	

6.3 Use your knowledge of the permissible ranges for the parameters of the Poisson, negative binomial, and binomial to determine all possible values of a and b for these members of the $(a, b, 0)$ class. Because these are the only members of the class, all other pairs must not lead to a legitimate probability distribution (nonnegative values that sum to 1). Show that the pair $a = -1$ and $b = 1.5$ (which is not on the list of possible values) does not lead to a legitimate distribution.

6.6   Truncation and Modification at Zero

At times, the distributions discussed previously do not adequately describe the characteristics of some data sets encountered in practice. This may be because the tail of the negative binomial is not heavy enough or because the distributions in the $(a, b, 0)$ class cannot capture the shape of the data set in some other part of the distribution.

In this section, we address the problem of a poor fit at the left-hand end of the distribution, in particular, the probability at zero.

For insurance count data, the probability at zero is the probability that no claims occur during the period under study. For applications in insurance where the probability of occurrence of a loss is very small, the probability at zero has the largest value. Thus, it is important to pay special attention to the fit at this point.

There are also situations that naturally occur which generate unusually large probabilities at zero. Consider the case of group dental insurance. If, in a family, both husband and wife have coverage with their respective employer-sponsored plans and both group insurance contracts provide coverage for all family members, the claims will be made to the insurer of the plan that provides the better benefits, and no claims may be made under the other contract. Then, in conducting studies for a specific insurer, we may find a higher than expected number of individuals who made no claim.

Similarly, it is possible to have situations in which there is less than the expected number, or even zero, occurrences at zero. For example, if we are counting the number of claims from accidents resulting in a claim, the minimum observed value is 1.

An adjustment of the probability at zero is easily handled for the Poisson, binomial, and negative binomial distributions.

Definition 6.5 *Let p_k be the pf of a discrete random variable. It is a member of the* $(a, b, 1)$ *class of distributions provided that there exist constants a and b such that*

$$p_k = \left(a + \frac{b}{k} \right) p_{k-1}, \qquad k = 2, 3, 4, \dots .$$

Note that the only difference from the $(a, b, 0)$ class is that the recursion begins at p_1

rather than p_0. The distribution from $k = 1$ to $k = \infty$ has the same shape as the $(a, b, 0)$ class in the sense that the probabilities are the same up to a constant of proportionality, because $\sum_{k=1}^{\infty} p_k$ can be set to any number in the interval $[0, 1]$. The remaining probability is at $k = 0$.

We distinguish between the situations in which $p_0 = 0$ and those where $p_0 > 0$. The first subclass is called the **truncated** (more specifically, **zero-truncated**) distributions. The members are the zero-truncated Poisson, zero-truncated binomial, and zero-truncated negative binomial distributions (and the special case of the latter, the zero-truncated geometric distribution).

The second subclass is referred to as the **zero-modified** distributions because the probability is modified from that for the $(a, b, 0)$ class. These distributions can be viewed as a mixture of an $(a, b, 0)$ distribution and a degenerate distribution with all the probability at zero. Alternatively, they can be called **truncated with zeros** distributions because the distribution can be viewed as a mixture of a truncated distribution and a degenerate distribution with all the probability at zero. We now show this equivalence more formally. Note that all zero-truncated distributions can be considered as zero-modified distributions, with the particular modification being to set $p_0 = 0$. The abbreviations ZT and ZM will be used at times. For example, a reference to the ZT Poisson distribution.

With three types of distributions, the notation can become confusing. When writing about discrete distributions in general, we continue to let $p_k = \Pr(N = k)$. When referring to a zero-truncated distribution, we use p_k^T, and when referring to a zero-modified distribution, we use p_k^M. Once again, it is possible for a zero-modified distribution to be a zero-truncated distribution.

Let $P(z) = \sum_{k=0}^{\infty} p_k z^k$ denote the pgf of a member of the $(a, b, 0)$ class. Let $P^M(z) = \sum_{k=0}^{\infty} p_k^M z^k$ denote the pgf of the corresponding member of the $(a, b, 1)$ class, that is,

$$p_k^M = cp_k, \qquad k = 1, 2, 3, \dots ,$$

and p_0^M is an arbitrary number. Then,

$$P^M(z) = p_0^M + \sum_{k=1}^{\infty} p_k^M z^k$$

$$= p_0^M + c \sum_{k=1}^{\infty} p_k z^k$$

$$= p_0^M + c[P(z) - p_0].$$

Because $P^M(1) = P(1) = 1$,

$$1 = p_0^M + c(1 - p_0),$$

resulting in

$$c = \frac{1 - p_0^M}{1 - p_0} \text{ or } p_0^M = 1 - c(1 - p_0).$$

This relationship is necessary to ensure that the p_k^M sum to 1. We then have

$$P^M(z) = p_0^M + \frac{1 - p_0^M}{1 - p_0}[P(z) - p_0]$$

$$= \left(1 - \frac{1 - p_0^M}{1 - p_0}\right)1 + \frac{1 - p_0^M}{1 - p_0}P(z). \tag{6.4}$$

This is a weighted average of the pgfs of the degenerate distribution and the corresponding $(a, b, 0)$ member. Furthermore,

$$p_k^M = \frac{1 - p_0^M}{1 - p_0}p_k, \qquad k = 1, 2, \dots . \tag{6.5}$$

Let $P^T(z)$ denote the pgf of the zero-truncated distribution corresponding to an $(a, b, 0)$ pgf $P(z)$. Then, by setting $p_0^M = 0$ in (6.4) and (6.5),

$$P^T(z) = \frac{P(z) - p_0}{1 - p_0}$$

and

$$p_k^T = \frac{p_k}{1 - p_0}, \qquad k = 1, 2, \dots . \tag{6.6}$$

Then, from (6.5),

$$p_k^M = (1 - p_0^M)p_k^T, \qquad k = 1, 2, \dots , \tag{6.7}$$

and

$$P^M(z) = p_0^M(1) + (1 - p_0^M)P^T(z). \tag{6.8}$$

Then, the zero-modified distribution is also the weighted average of a degenerate distribution and the zero-truncated member of the $(a, b, 0)$ class. The following example illustrates these relationships.

■ **EXAMPLE 6.4**

Consider a negative binomial random variable with parameters $\beta = 0.5$ and $r = 2.5$. Determine the first four probabilities for this random variable. Then determine the corresponding probabilities for the zero-truncated and zero-modified (with $p_0^M = 0.6$) versions.

From Table 6.4 we have, for the negative binomial distribution,

$$p_0 = (1 + 0.5)^{-2.5} = 0.362887,$$
$$a = \frac{0.5}{1.5} = \frac{1}{3},$$
$$b = \frac{(2.5 - 1)(0.5)}{1.5} = \frac{1}{2}.$$

The first three recursions are

$$p_1 = 0.362887 \left(\tfrac{1}{3} + \tfrac{1}{2}\tfrac{1}{1} \right) = 0.302406,$$

$$p_2 = 0.302406 \left(\tfrac{1}{3} + \tfrac{1}{2}\tfrac{1}{2} \right) = 0.176404,$$

$$p_3 = 0.176404 \left(\tfrac{1}{3} + \tfrac{1}{2}\tfrac{1}{3} \right) = 0.088202.$$

For the zero-truncated random variable, $p_0^T = 0$ by definition. The recursions start with [from (6.6)] $p_1^T = 0.302406/(1 - 0.362887) = 0.474651$. Then,

$$p_2^T = 0.474651 \left(\tfrac{1}{3} + \tfrac{1}{2}\tfrac{1}{2} \right) = 0.276880,$$

$$p_3^T = 0.276880 \left(\tfrac{1}{3} + \tfrac{1}{2}\tfrac{1}{3} \right) = 0.138440.$$

If the original values were all available, then the zero-truncated probabilities could have all been obtained by multiplying the original values by $1/(1 - 0.362887) = 1.569580$.

For the zero-modified random variable, $p_0^M = 0.6$ arbitrarily. From (6.5), $p_1^M = (1 - 0.6)(0.302406)/(1 - 0.362887) = 0.189860$. Then,

$$p_2^M = 0.189860 \left(\tfrac{1}{3} + \tfrac{1}{2}\tfrac{1}{2} \right) = 0.110752,$$

$$p_3^M = 0.110752 \left(\tfrac{1}{3} + \tfrac{1}{2}\tfrac{1}{3} \right) = 0.055376.$$

In this case, each original negative binomial probability has been multiplied by $(1 - 0.6)/(1 - 0.362887) = 0.627832$. Also note that, for $j \geq 1$, $p_j^M = 0.4 p_j^T$.   □

A special case of the zero-modified distributions is called **zero-inflated**. The only difference is that for such distributions it is required that $p_0^M > p_0$. It is shown in Frees [41] that for the zero-inflated Poisson distribution the variance is always larger than the mean. This provides an alternative to the negative binomial model when that property is desired.

Although we have only discussed the zero-modified distributions of the $(a, b, 0)$ class, the $(a, b, 1)$ class admits additional distributions. The (a, b) parameter space can be expanded to admit an extension of the negative binomial distribution to include cases where $-1 < r < 0$. For the $(a, b, 0)$ class, $r > 0$ is required. By adding the additional region to the sample space, the "extended" truncated negative binomial (ETNB) distribution has parameter restrictions $\beta > 0, r > -1, r \neq 0$.

To show that the recursive equation

$$p_k = p_{k-1} \left(a + \frac{b}{k} \right), \qquad k = 2, 3, \ldots, \tag{6.9}$$

with $p_0 = 0$ defines a proper distribution, it is sufficient to show that for any value of p_1, the successive values of p_k obtained recursively are each positive and that $\sum_{k=1}^{\infty} p_k < \infty$. For the ETNB, this must be done for the parameter space

$$a = \frac{\beta}{1 + \beta}, \quad \beta > 0, \quad \text{and} \quad b = (r - 1)\frac{\beta}{1 + \beta}, \quad r > -1, r \neq 0$$

(see Exercise 6.4).

When $r \to 0$, the limiting case of the ETNB is the **logarithmic distribution** with

$$p_k^T = \frac{[\beta/(1+\beta)]^k}{k \ln(1+\beta)}, \quad k = 1, 2, 3, \ldots \tag{6.10}$$

(see Exercise 6.5). The pgf of the logarithmic distribution is

$$P^T(z) = 1 - \frac{\ln[1 - \beta(z-1)]}{\ln(1+\beta)} \tag{6.11}$$

(see Exercise 6.6). The zero-modified logarithmic distribution is created by assigning an arbitrary probability at zero and reducing the remaining probabilities.

It is also interesting that the special extreme case with $-1 < r < 0$ and $\beta \to \infty$ is a proper distribution and is sometimes called the **Sibuya distribution**. It has pgf $P(z) = 1 - (1-z)^{-r}$ and no moments exist (see Exercise 6.7). Distributions with no moments are not particularly interesting for modeling claim numbers (unless the right tail is subsequently modified), because then an infinite number of claims is expected. An insurance policy covering such a case might be difficult to price!

■ **EXAMPLE 6.5**

Determine the probabilities for an ETNB distribution with $r = -0.5$ and $\beta = 1$. Do this both for the truncated version and for the modified version, with $p_0^M = 0.6$ set arbitrarily.

We have $a = 1/(1+1) = 0.5$ and $b = (-0.5 - 1)(1)/(1+1) = -0.75$. From Appendix B, we also have $p_1^T = -0.5(1)/[(1+1)^{0.5} - (1+1)] = 0.853553$. Subsequent values are

$$p_2^T = \left(0.5 - \frac{0.75}{2}\right)(0.853553) = 0.106694,$$

$$p_3^T = \left(0.5 - \frac{0.75}{3}\right)(0.106694) = 0.026674.$$

For the modified probabilities, the truncated probabilities need to be multiplied by 0.4 to produce $p_1^M = 0.341421$, $p_2^M = 0.042678$, and $p_3^M = 0.010670$. □

It is reasonable to ask if there is a "natural" member of the ETNB distribution for the example, that is, one for which the recursion would begin with p_1 rather than p_2. The natural value of p_0 would have to satisfy $p_1 = (0.5 - 0.75/1)p_0 = -0.25p_0$. This would force one of the two probabilities to be negative and so there is no acceptable solution. It is easy to show that this occurs for any $r < 0$.

There are no other members of the $(a, b, 1)$ class beyond the ones just discussed. A summary is given in Table 6.4.

6.6.1   Exercises

6.4 Show that for the extended truncated negative binomial distribution with any $\beta > 0$ and $r > -1$, but $r \neq 0$, the successive values of p_k given by (6.9) are, for any p_1, positive and $\sum_{k=1}^{\infty} p_k < \infty$.

6.5 Show that when, in the zero-truncated negative binomial distribution, $r \to 0$, the pf is as given in (6.10).

6.6 Show that the pgf of the logarithmic distribution is as given in (6.11).

6.7 Show that for the Sibuya distribution, which is the ETNB distribution with $-1 < r < 0$ and $\beta \to \infty$, the mean does not exist (i.e. the sum that defines the mean does not converge). Because this random variable takes on nonnegative values, this also shows that no other positive moments exist.

6.8 If $p_k = \Pr(N = k)$, and $\{p_0, p_1, p_2, \dots\}$ is a member of the $(a, b, 1)$ class, demonstrate that for $k = 1, 2, \dots$, and $a \neq 1$,

$$E(N \mid N > k) = \frac{a + b}{1 - a} + \frac{a(k + 1) + b}{1 - a} \frac{p_k}{\bar{P}_k},$$

where $\bar{P}_k = p_{k+1} + p_{k+2} + \dots$.

Table 6.4   The members of the $(a, b, 1)$ class.

Distribution[a]	p_0	a	b	Parameter space
Poisson	$e^{-\lambda}$	0	λ	$\lambda > 0$
ZT Poisson	0	0	λ	$\lambda > 0$
ZM Poisson	Arbitrary	0	λ	$\lambda > 0$
Binomial	$(1 - q)^m$	$-\dfrac{q}{1 - q}$	$(m + 1)\dfrac{q}{1 - q}$	$0 < q < 1$
ZT binomial	0	$-\dfrac{q}{1 - q}$	$(m + 1)\dfrac{q}{1 - q}$	$0 < q < 1$
ZM binomial	Arbitrary	$-\dfrac{q}{1 - q}$	$(m + 1)\dfrac{q}{1 - q}$	$0 < q < 1$
Negative binomial	$(1 + \beta)^{-r}$	$\dfrac{\beta}{1 + \beta}$	$(r - 1)\dfrac{\beta}{1 + \beta}$	$r > 0, \ \beta > 0$
ETNB	0	$\dfrac{\beta}{1 + \beta}$	$(r - 1)\dfrac{\beta}{1 + \beta}$	$r > -1,^b \ \beta > 0$
ZM ETNB	Arbitrary	$\dfrac{\beta}{1 + \beta}$	$(r - 1)\dfrac{\beta}{1 + \beta}$	$r > -1,^b \ \beta > 0$
Geometric	$(1 + \beta)^{-1}$	$\dfrac{\beta}{1 + \beta}$	0	$\beta > 0$
ZT geometric	0	$\dfrac{\beta}{1 + \beta}$	0	$\beta > 0$
ZM geometric	Arbitrary	$\dfrac{\beta}{1 + \beta}$	0	$\beta > 0$
Logarithmic	0	$\dfrac{\beta}{1 + \beta}$	$-\dfrac{\beta}{1 + \beta}$	$\beta > 0$
ZM logarithmic	Arbitrary	$\dfrac{\beta}{1 + \beta}$	$-\dfrac{\beta}{1 + \beta}$	$\beta > 0$

[a]ZT = zero truncated, ZM = zero modified.

[b]Excluding $r = 0$, which is the logarithmic distribution.

7

ADVANCED DISCRETE DISTRIBUTIONS

7.1 Compound Frequency Distributions

A larger class of distributions can be created by the processes of compounding any two discrete distributions. The term *compounding* reflects the idea that the pgf of the new distribution, $P_S(z)$, is written as

$$P_S(z) = P_N[P_M(z)], \qquad (7.1)$$

where $P_N(z)$ and $P_M(z)$ are called the *primary* and *secondary* distributions, respectively.

The compound distributions arise naturally as follows. Let N be a counting random variable with pgf $P_N(z)$. Let $M_1, M_2, \ldots$ be i.i.d. counting random variables each with pgf $P_M(z)$. Assuming that the M_js do not depend on N, the pgf of the random sum

$S = M_1 + M_2 + \cdots + M_N$ (where $N = 0$ implies that $S = 0$) is $P_S(z) = P_N[P_M(z)]$. This is shown as follows:

$$P_S(z) = \sum_{k=0}^{\infty} \Pr(S = k)z^k = \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \Pr(S = k | N = n) \Pr(N = n)z^k$$

$$= \sum_{n=0}^{\infty} \Pr(N = n) \sum_{k=0}^{\infty} \Pr(M_1 + \cdots + M_n = k | N = n)z^k$$

$$= \sum_{n=0}^{\infty} \Pr(N = n)[P_M(z)]^n$$

$$= P_N[P_M(z)].$$

In insurance contexts, this distribution can arise naturally. If N represents the number of accidents arising in a portfolio of risks and $\{M_k \,:\, k = 1, 2, \ldots, N\}$ represents the number of claims (injuries, number of cars, etc.) from the accidents, then S represents the total number of claims from the portfolio. This kind of interpretation is not necessary to justify the use of a compound distribution. If a compound distribution fits data well, that may be enough justification itself. Also, there are other motivations for these distributions, as presented in Section 7.5.

■ **EXAMPLE 7.1**

Demonstrate that any zero-modified distribution is a compound distribution.

Consider a primary Bernoulli distribution. It has pgf $P_N(z) = 1 - q + qz$. Then consider an arbitrary secondary distribution with pgf $P_M(z)$. Then, from (7.1), we obtain
$$P_S(z) = P_N[P_M(z)] = 1 - q + qP_M(z).$$
From (6.4) this is the pgf of a ZM distribution with

$$q = \frac{1 - p_0^M}{1 - p_0}.$$

That is, the ZM distribution has assigned arbitrary probability p_0^M at zero, while p_0 is the probability assigned at zero by the secondary distribution. □

■ **EXAMPLE 7.2**

Consider the case where both M and N have a Poisson distribution. Determine the pgf of this distribution.

This distribution is called the Poisson–Poisson or Neyman Type A distribution. Let $P_N(z) = e^{\lambda_1(z-1)}$ and $P_M(z) = e^{\lambda_2(z-1)}$. Then,

$$P_S(z) = e^{\lambda_1[e^{\lambda_2(z-1)}-1]}.$$

When λ_2 is a lot larger than λ_1, for example, $\lambda_1 = 0.1$ and $\lambda_2 = 10$, the resulting distribution will have two local modes. □

The probability of exactly k claims can be written as

$$\Pr(S = k) = \sum_{n=0}^{\infty} \Pr(S = k|N = n)\Pr(N = n)$$

$$= \sum_{n=0}^{\infty} \Pr(M_1 + \cdots + M_N = k|N = n)\Pr(N = n)$$

$$= \sum_{n=0}^{\infty} \Pr(M_1 + \cdots + M_n = k)\Pr(N = n). \tag{7.2}$$

Letting $g_n = \Pr(S = n)$, $p_n = \Pr(N = n)$, and $f_n = \Pr(M = n)$, this is rewritten as

$$g_k = \sum_{n=0}^{\infty} p_n f_k^{*n}, \tag{7.3}$$

where f_k^{*n}, $k = 0, 1, \dots$, is the "n-fold convolution" of the function f_k, $k = 0, 1, \dots$, that is, the probability that the sum of n random variables which are each i.i.d. with probability function f_k will take on value k.

When $P_N(z)$ is chosen to be a member of the $(a, b, 0)$ class,

$$p_k = \left(a + \frac{b}{k}\right) p_{k-1}, \qquad k = 1, 2, \dots, \tag{7.4}$$

then a simple recursive formula can be used. This formula avoids the use of convolutions and thus reduces the computations considerably.

Theorem 7.1 *If the primary distribution is a member of the $(a, b, 0)$ class, the recursive formula is*

$$g_k = \frac{1}{1 - af_0} \sum_{j=1}^{k} \left(a + \frac{bj}{k}\right) f_j g_{k-j}, \qquad k = 1, 2, 3, \dots. \tag{7.5}$$

Proof: From (7.4),
$$np_n = a(n - 1)p_{n-1} + (a + b)p_{n-1}.$$

Multiplying each side by $[P_M(z)]^{n-1} P'_M(z)$ and summing over n yields

$$\sum_{n=1}^{\infty} np_n[P_M(z)]^{n-1} P'_M(z) = a \sum_{n=1}^{\infty} (n - 1)p_{n-1}[P_M(z)]^{n-1} P'_M(z)$$

$$+ (a + b) \sum_{n=1}^{\infty} p_{n-1}[P_M(z)]^{n-1} P'_M(z).$$

Because $P_S(z) = \sum_{n=0}^{\infty} p_n[P_M(z)]^n$, the previous equation is

$$P'_S(z) = a \sum_{n=0}^{\infty} np_n[P_M(z)]^n P'_M(z) + (a + b) \sum_{n=0}^{\infty} p_n[P_M(z)]^n P'_M(z).$$

Therefore,

$$P'_S(z) = aP'_S(z)P_M(z) + (a+b)P_S(z)P'_M(z).$$

Each side can be expanded in powers of z. The coefficients of z^{k-1} in such an expansion must be the same on both sides of the equation. Hence, for $k = 1, 2, \ldots$, we have

$$kg_k = a\sum_{j=0}^{k}(k-j)f_j g_{k-j} + (a+b)\sum_{j=0}^{k} jf_j g_{k-j}$$

$$= akf_0 g_k + a\sum_{j=1}^{k}(k-j)f_j g_{k-j} + (a+b)\sum_{j=1}^{k} jf_j g_{k-j}$$

$$= akf_0 g_k + ak\sum_{j=1}^{k} f_j g_{k-j} + b\sum_{j=1}^{k} jf_j g_{k-j}.$$

Therefore,

$$g_k = af_0 g_k + \sum_{j=1}^{k}\left(a + \frac{bj}{k}\right)f_j g_{k-j}.$$

Rearrangement yields (7.5). □

In order to use (7.5), the starting value g_0 is required and is given in Theorem 7.3. If the primary distribution is a member of the $(a, b, 1)$ class, the proof must be modified to reflect the fact that the recursion for the primary distribution begins at $k = 2$. The result is the following.

Theorem 7.2 *If the primary distribution is a member of the $(a, b, 1)$ class, the recursive formula is*

$$g_k = \frac{[p_1 - (a+b)p_0]f_k + \sum_{j=1}^{k}(a+bj/k)f_j g_{k-j}}{1 - af_0}, \quad k = 1, 2, 3, \ldots. \tag{7.6}$$

Proof: It is similar to the proof of Theorem 7.1 and is omitted. □

■ **EXAMPLE 7.3**

Develop the recursive formula for the case where the primary distribution is Poisson.

In this case, $a = 0$ and $b = \lambda$, yielding the recursive form

$$g_k = \frac{\lambda}{k}\sum_{j=1}^{k} jf_j g_{k-j}.$$

The starting value is, from (7.1),

$$\begin{aligned}
g_0 &= \Pr(S = 0) = P(0)\\
&= P_N[P_M(0)] = P_N(f_0)\\
&= e^{-\lambda(1-f_0)}.
\end{aligned} \tag{7.7}$$

Distributions of this type are called **compound Poisson**. When the secondary distribution is specified, the compound distribution is called Poisson–X, where X is the name of the secondary distribution. ☐

The method used to obtain g_0 applies to any compound distribution.

Theorem 7.3 *For any compound distribution, $g_0 = P_N(f_0)$, where $P_N(z)$ is the pgf of the primary distribution and f_0 is the probability that the secondary distribution takes on the value zero.*

Proof: See the second line of (7.7) ☐

Note that the secondary distribution is not required to be in any special form. However, to keep the number of distributions manageable, secondary distributions are selected from the $(a, b, 0)$ or the $(a, b, 1)$ class.

■ **EXAMPLE 7.4**

Calculate the probabilities for the Poisson–ETNB distribution, where $\lambda = 3$ for the Poisson distribution and the ETNB distribution has $r = -0.5$ and $\beta = 1$.

From Example 6.5, the secondary probabilities are $f_0 = 0$, $f_1 = 0.853553$, $f_2 = 0.106694$, and $f_3 = 0.026674$. From (7.7), $g_0 = \exp[-3(1 - 0)] = 0.049787$. For the Poisson primary distribution, $a = 0$ and $b = 3$. The recursive formula in (7.5) becomes

$$g_k = \frac{\sum_{j=1}^{k}(3j/k)f_j g_{k-j}}{1 - 0(0)} = \sum_{j=1}^{k} \frac{3j}{k} f_j g_{k-j}.$$

Then,

$$g_1 = \frac{3(1)}{1}0.853553(0.049787) = 0.127488,$$

$$g_2 = \frac{3(1)}{2}0.853553(0.127488) + \frac{3(2)}{2}0.106694(0.049787) = 0.179163,$$

$$g_3 = \frac{3(1)}{3}0.853553(0.179163) + \frac{3(2)}{3}0.106694(0.127488)$$

$$+ \frac{3(3)}{3}0.026674(0.049787) = 0.184114.$$ ☐

■ **EXAMPLE 7.5**

Demonstrate that the Poisson–logarithmic distribution is a negative binomial distribution.

The negative binomial distribution has pgf

$$P(z) = [1 - \beta(z - 1)]^{-r}.$$

Suppose that $P_N(z)$ is Poisson(λ) and $P_M(z)$ is logarithmic(β). Then,

$$P_N[P_M(z)] = \exp\{\lambda[P_M(z) - 1]\}$$

$$= \exp\left\{\lambda\left[1 - \frac{\ln[1 - \beta(z - 1)]}{\ln(1 + \beta)} - 1\right]\right\}$$

$$= \exp\left\{\frac{-\lambda}{\ln(1 + \beta)}\ln[1 - \beta(z - 1)]\right\}$$

$$= [1 - \beta(z - 1)]^{-\lambda/\ln(1+\beta)}$$

$$= [1 - \beta(z - 1)]^{-r},$$

where $r = \lambda/\ln(1 + \beta)$. This shows that the negative binomial distribution can be written as a compound Poisson distribution with a logarithmic secondary distribution.

□

Example 7.5 shows that the Poisson–logarithmic distribution does not create a new distribution beyond the $(a, b, 0)$ and $(a, b, 1)$ classes. As a result, this combination of distributions is not useful to us. Another combination that does not create a new distribution beyond the $(a, b, 1)$ class is the compound geometric distribution, where both the primary and secondary distributions are geometric. The resulting distribution is a zero-modified geometric distribution, as shown in Exercise 7.2. The following theorem shows that certain other combinations are also of no use in expanding the class of distributions through compounding. Suppose that $P_S(z) = P_N[P_M(z)]$ as before. Now, $P_M(z)$ can always be written as

$$P_M(z) = f_0 + (1 - f_0)P_M^*(z), \tag{7.8}$$

where $P_M^*(z)$ is the pgf of the conditional distribution over the positive range (in other words, the zero-truncated version).

Theorem 7.4 *Suppose that the pgf $P_N(z; \theta)$ satisfies*

$$P_N(z; \theta) = B[\theta(z - 1)]$$

for some parameter θ and some function $B(z)$ that is independent of θ. That is, the parameter θ and the argument z only appear in the pgf as $\theta(z - 1)$. There may be other parameters as well, and they may appear anywhere in the pgf. Then, $P_S(z) = P_N[P_M(z); \theta]$ can be rewritten as

$$P_S(z) = P_N[P_M^T(z); \theta(1 - f_0)].$$

Proof:

$$P_S(z) = P_N[P_M(z); \theta]$$

$$= P_N[f_0 + (1 - f_0)P_M^T(z); \theta]$$

$$= B\{\theta[f_0 + (1 - f_0)P_M^T(z) - 1]\}$$

$$= B\{\theta(1 - f_0)[P_M^T(z) - 1]\}$$

$$= P_N[P_M^T(z); \theta(1 - f_0)].$$

□

This shows that adding, deleting, or modifying the probability at zero in the secondary distribution does not add a new distribution because it is equivalent to modifying the parameter θ of the primary distribution. Thus, for example, a Poisson primary distribution with a Poisson, zero-truncated Poisson, or zero-modified Poisson secondary distribution will still lead to a Neyman Type A (Poisson–Poisson) distribution.

■ **EXAMPLE 7.6**

Determine the probabilities for a Poisson–zero-modified ETNB distribution where the parameters are $\lambda = 7.5$, $p_0^M = 0.6$, $r = -0.5$, and $\beta = 1$.

From Example 6.5, the secondary probabilities are $f_0 = 0.6$, $f_1 = 0.341421$, $f_2 = 0.042678$, and $f_3 = 0.010670$. From (7.7), $g_0 = \exp[-7.5(1-0.6)] = 0.049787$. For the Poisson primary distribution, $a = 0$ and $b = 7.5$. The recursive formula in (7.5) becomes

$$g_k = \frac{\sum_{j=1}^{k}(7.5j/k)f_j g_{k-j}}{1 - 0(0.6)} = \sum_{j=1}^{k}\frac{7.5j}{k}f_j g_{k-j}.$$

Then,

$$g_1 = \frac{7.5(1)}{1}0.341421(0.049787) = 0.127487,$$

$$g_2 = \frac{7.5(1)}{2}0.341421(0.127487) + \frac{7.5(2)}{2}0.042678(0.049787) = 0.179161,$$

$$g_3 = \frac{7.5(1)}{3}0.341421(0.179161) + \frac{7.5(2)}{3}0.042678(0.127487)$$

$$+ \frac{7.5(3)}{3}0.010670(0.049787) = 0.184112.$$

Except for slight rounding differences, these probabilities are the same as those obtained in Example 7.4.   □

7.1.1 Exercises

7.1 Do all the members of the $(a, b, 0)$ class satisfy the condition of Theorem 7.4? For those that do, identify the parameter (or function of its parameters) that plays the role of θ in the theorem.

7.2 Show that the following three distributions are identical: (1) geometric–geometric, (2) Bernoulli–geometric, and (3) zero-modified geometric. That is, for any one of the distributions with arbitrary parameters, show that there is a member of the other two distribution types that has the same pf or pgf.

7.3 Show that the binomial–geometric and negative binomial–geometric (with negative binomial parameter r a positive integer) distributions are identical.

7.2 Further Properties of the Compound Poisson Class

Of central importance within the class of compound frequency models is the class of compound Poisson frequency distributions. Physical motivation for this model arises

from the fact that the Poisson distribution is often a good model to describe the number of claim-causing accidents, and the number of claims from an accident is often itself a random variable. There are numerous convenient mathematical properties enjoyed by the compound Poisson class. In particular, those involving recursive evaluation of the probabilities were also discussed in Section 7.1. In addition, there is a close connection between the compound Poisson distributions and the mixed Poisson frequency distributions that is discussed in Section 7.3.2. Here, we consider some other properties of these distributions. The compound Poisson pgf may be expressed as

$$P(z) = \exp\{\lambda[Q(z) - 1]\}, \tag{7.9}$$

where $Q(z)$ is the pgf of the secondary distribution.

■ **EXAMPLE 7.7**

Obtain the pgf for the Poisson–ETNB distribution and show that it looks like the pgf of a Poisson–negative binomial distribution.

The ETNB distribution has pgf

$$Q(z) = \frac{[1 - \beta(z - 1)]^{-r} - (1 + \beta)^{-r}}{1 - (1 + \beta)^{-r}}$$

for $\beta > 0$, $r > -1$, and $r \neq 0$. Then, the Poisson–ETNB distribution has, as the logarithm of its pgf,

$$\ln P(z) = \lambda \left\{ \frac{[1 - \beta(z - 1)]^{-r} - (1 + \beta)^{-r}}{1 - (1 + \beta)^{-r}} - 1 \right\}$$

$$= \lambda \left\{ \frac{[1 - \beta(z - 1)]^{-r} - 1}{1 - (1 + \beta)^{-r}} \right\}$$

$$= \mu\{[1 - \beta(z - 1)]^{-r} - 1\},$$

where $\mu = \lambda/[1 - (1 + \beta)^{-r}]$. This defines a compound Poisson distribution with primary mean μ and secondary pgf $[1 - \beta(z - 1)]^{-r}$, which is the pgf of a negative binomial random variable, as long as r and, hence, μ, are positive. This observation illustrates that the probability at zero in the secondary distribution has no impact on the compound Poisson form. Also, the preceding calculation demonstrates that the Poisson–ETNB pgf $P(z)$, with $\ln P(z) = \mu\{[1 - \beta(z - 1)]^{-r} - 1\}$, has parameter space $\{\beta > 0, r > -1, \mu r > 0\}$, a useful observation with respect to estimation and analysis of the parameters.   □

We can compare the skewness (third moment) of these distributions to develop an appreciation of the amount by which the skewness and, hence, the tails of these distributions can vary even when the mean and variance are fixed. From (7.9) (see Exercise 7.5) and Definition 3.2, the mean and second and third central moments of the compound Poisson distribution are

$$\mu'_1 = \mu = \lambda m'_1,$$
$$\mu_2 = \sigma^2 = \lambda m'_2, \tag{7.10}$$
$$\mu_3 = \lambda m'_3,$$

where m'_j is the jth raw moment of the secondary distribution. The coefficient of skewness is

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{m'_3}{\lambda^{1/2}(m'_2)^{3/2}}.$$

For the Poisson–binomial distribution, with a bit of algebra (see Exercise 7.6), we obtain

$$\mu = \lambda mq,$$
$$\sigma^2 = \mu[1 + (m-1)q], \tag{7.11}$$
$$\mu_3 = 3\sigma^2 - 2\mu + \frac{m-2}{m-1}\frac{(\sigma^2 - \mu)^2}{\mu}.$$

Carrying out similar exercises for the negative binomial, Polya–Aeppli, Neyman Type A, and Poisson–ETNB distributions yields

$$\text{Negative binomial: } \mu_3 = 3\sigma^2 - 2\mu + 2\frac{(\sigma^2 - \mu)^2}{\mu}$$

$$\text{Polya–Aeppli: } \mu_3 = 3\sigma^2 - 2\mu + \frac{3}{2}\frac{(\sigma^2 - \mu)^2}{\mu}$$

$$\text{Neyman Type A: } \mu_3 = 3\sigma^2 - 2\mu + \frac{(\sigma^2 - \mu)^2}{\mu}$$

$$\text{Poisson–ETNB: } \mu_3 = 3\sigma^2 - 2\mu + \frac{r+2}{r+1}\frac{(\sigma^2 - \mu)^2}{\mu}$$

For the Poisson–ETNB distribution, the range of r is $-1 < r < \infty$, $r \neq 0$. The other three distributions are special cases. Letting $r \to 0$, the secondary distribution is logarithmic, resulting in the negative binomial distribution. Setting $r = 1$ defines the Polya–Aeppli distribution. Letting $r \to \infty$, the secondary distribution is Poisson, resulting in the Neyman Type A distribution.

Note that for fixed mean and variance, the third moment only changes through the coefficient in the last term for each of the five distributions. For the Poisson distribution, $\mu_3 = \lambda = 3\sigma^2 - 2\mu$, and so the third term for each expression for μ_3 represents the change from the Poisson distribution. For the Poisson–binomial distribution, if $m = 1$, the distribution is Poisson because it is equivalent to a Poisson–zero-truncated binomial as truncation at zero leaves only probability at 1. Another view is that from (7.11) we have

$$\mu_3 = 3\sigma^2 - 2\mu + \frac{m-2}{m-1}\frac{(m-1)^2 q^4 \lambda^2 m^2}{\lambda mq}$$
$$= 3\sigma^2 - 2\mu + (m-2)(m-1)q^3 \lambda m,$$

which reduces to the Poisson value for μ_3 when $m = 1$. Hence, it is necessary that $m \geq 2$ for non-Poisson distributions to be created. Then, the coefficient satisfies

$$0 \leq \frac{m-2}{m-1} < 1.$$

For the Poisson–ETNB, because $r > -1$, the coefficient satisfies

$$1 < \frac{r+2}{r+1} < \infty,$$

noting that when $r = 0$ this refers to the negative binomial distribution. For the Neyman Type A distribution, the coefficient is exactly 1. Hence, these three distributions provide any desired degree of skewness greater than that of the Poisson distribution.

Table 7.1   The data of Hossack et al. [58].

Number of claims	Observed frequency
0	565,664
1	68,714
2	5,177
3	365
4	24
5	6
6+	0

■ **EXAMPLE 7.8**

The data in Table 7.1 are taken from Hossack et al. [58] and give the distribution of
the number of claims on automobile insurance policies in Australia. Determine an
appropriate frequency model based on the skewness results of this section.

The mean, variance, and third central moment are 0.1254614, 0.1299599, and
0.1401737, respectively. For these numbers,

$$\frac{\mu_3 - 3\sigma^2 + 2\mu}{(\sigma^2 - \mu)^2/\mu} = 7.543865.$$

From among the Poisson–binomial, negative binomial, Polya–Aeppli, Neyman Type
A, and Poisson–ETNB distributions, only the latter is appropriate. For this distribution,
an estimate of r can be obtained from

$$7.543865 = \frac{r+2}{r+1},$$

resulting in $r = -0.8471851$. In Example 15.12, a more formal estimation and
selection procedure is applied, but the conclusion is the same.   □

A very useful property of the compound Poisson class of probability distributions is the
fact that it is closed under convolution. We have the following theorem.

Theorem 7.5 *Suppose that S_i has a compound Poisson distribution with Poisson param-
eter λ_i and secondary distribution $\{q_n(i) : n = 0, 1, 2, \dots\}$ for $i = 1, 2, 3, \dots, k$. Suppose
also that $S_1, S_2, \dots, S_k$ are independent random variables. Then, $S = S_1 + S_2 + \cdots + S_k$
also has a compound Poisson distribution with Poisson parameter $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_k$
and secondary distribution $\{q_n : n = 0, 1, 2, \dots\}$, where $q_n = [\lambda_1 q_n(1) + \lambda_2 q_n(2) + \cdots +
\lambda_k q_n(k)]/\lambda$.*

Proof: Let $Q_i(z) = \sum_{n=0}^{\infty} q_n(i)z^n$ for $i = 1, 2, \dots, k$. Then, S_i has pgf $P_{S_i}(z) = E(z^{S_i}) =$

$\exp\{\lambda_i[Q_i(z) - 1]\}$. Because the S_is are independent, S has pgf

$$P_S(z) = \prod_{i=1}^{k} P_{S_i}(z)$$

$$= \prod_{i=1}^{k} \exp\{\lambda_i[Q_i(z) - 1]\}$$

$$= \exp\left[\sum_{i=1}^{k} \lambda_i Q_i(z) - \sum_{i=1}^{k} \lambda_i\right]$$

$$= \exp\{\lambda[Q(z) - 1]\},$$

where $\lambda = \sum_{i=1}^{k} \lambda_i$ and $Q(z) = \sum_{i=1}^{k} \lambda_i Q_i(z)/\lambda$. The result follows by the uniqueness of the generating function. □

One main advantage of this result is computational. If we are interested in the sum of independent compound Poisson random variables, then we do not need to compute the distribution of each compound Poisson random variable separately (i.e. recursively using Example 7.3), because Theorem 7.5 implies that a single application of the compound Poisson recursive formula in Example 7.3 will suffice. The following example illustrates this idea.

■ **EXAMPLE 7.9**

Suppose that $k = 2$ and S_1 has a compound Poisson distribution with $\lambda_1 = 2$ and secondary distribution $q_1(1) = 0.2, q_2(1) = 0.7$, and $q_3(1) = 0.1$. Also, S_2 (independent of S_1) has a compound Poisson distribution with $\lambda_2 = 3$ and secondary distribution $q_2(2) = 0.25, q_3(2) = 0.6$, and $q_4(2) = 0.15$. Determine the distribution of $S = S_1 + S_2$.

We have $\lambda = \lambda_1 + \lambda_2 = 2 + 3 = 5$. Then,

$$q_1 = 0.4(0.2) + 0.6(0) = 0.08,$$
$$q_2 = 0.4(0.7) + 0.6(0.25) = 0.43,$$
$$q_3 = 0.4(0.1) + 0.6(0.6) = 0.40,$$
$$q_4 = 0.4(0) + 0.6(0.15) = 0.09.$$

Thus, S has a compound Poisson distribution with Poisson parameter $\lambda = 5$ and secondary distribution $q_1 = 0.08, q_2 = 0.43, q_3 = 0.40$, and $q_4 = 0.09$. Numerical values of the distribution of S may be obtained using the recursive formula

$$\Pr(S = x) = \frac{5}{x} \sum_{n=1}^{x} n q_n \Pr(S = x - n), \quad x = 1, 2, \ldots,$$

beginning with $\Pr(S = 0) = e^{-5}$. □

In various situations, the convolution of negative binomial distributions is of interest. The following example indicates how this distribution may be evaluated.

■ **EXAMPLE 7.10**

(*Convolutions of negative binomial distributions*) Suppose that N_i has a negative binomial distribution with parameters r_i and β_i for $i = 1, 2, \dots, k$ and that $N_1, N_2, \dots, N_k$ are independent. Determine the distribution of $N = N_1 + N_2 + \cdots + N_k$.

The pgf of N_i is $P_{N_i}(z) = [1 - \beta_i(z - 1)]^{-r_i}$ and that of N is

$$P_N(z) = \prod_{i=1}^{k} P_{N_i}(z) = \prod_{i=1}^{k} [1 - \beta_i(z - 1)]^{-r_i}.$$

If $\beta_i = \beta$ for $i = 1, 2, \dots, k$, then

$$P_N(z) = [1 - \beta(z - 1)]^{-(r_1 + r_2 + \cdots + r_k)},$$

and N has a negative binomial distribution with parameters $r = r_1 + r_2 + \cdots + r_k$ and β.

If not all the β_is are identical, however, we may proceed as follows. From Example 7.5,

$$P_{N_i}(z) = [1 - \beta_i(z - 1)]^{-r_i} = e^{\lambda_i[Q_i(z)-1]},$$

where $\lambda_i = r_i \ln(1 + \beta_i)$, and

$$Q_i(z) = 1 - \frac{\ln[1 - \beta_i(z - 1)]}{\ln(1 + \beta_i)} = \sum_{n=1}^{\infty} q_n(i)z^n$$

with

$$q_n(i) = \frac{[\beta_i/(1 + \beta_i)]^n}{n \ln(1 + \beta_i)}, \quad n = 1, 2, \dots .$$

But Theorem 7.5 implies that $N = N_1 + N_2 + \cdots + N_k$ has a compound Poisson distribution with Poisson parameter

$$\lambda = \sum_{i=1}^{k} r_i \ln(1 + \beta_i)$$

and secondary distribution

$$q_n = \sum_{i=1}^{k} \frac{\lambda_i}{\lambda} q_n(i)$$

$$= \frac{\sum_{i=1}^{k} r_i[\beta_i/(1 + \beta_i)]^n}{n \sum_{i=1}^{k} r_i \ln(1 + \beta_i)}, \quad n = 1, 2, 3, \dots .$$

The distribution of N may be computed recursively using the formula

$$\Pr(N = n) = \frac{\lambda}{n} \sum_{k=1}^{n} k q_k \Pr(N = n - k), \quad n = 1, 2, \dots ,$$

beginning with $\Pr(N = 0) = e^{-\lambda} = \prod_{i=1}^{k} (1 + \beta_i)^{-r_i}$ and with λ and q_n as given previously.   □

It is not hard to see that Theorem 7.5 is a generalization of Theorem 6.1, which may be recovered with $q_1(i) = 1$ for $i = 1, 2, \ldots, k$. Similarly, the decomposition result of Theorem 6.2 may also be extended to compound Poisson random variables, where the decomposition is on the basis of the region of support of the secondary distribution. For further details, see Panjer and Willmot [100, Sec. 6.4] or Karlin and Taylor [67, Sec. 16.9].

7.2.1 Exercises

7.4 For $i = 1, \ldots, n$ let S_i have independent compound Poisson frequency distributions with Poisson parameter λ_i and a secondary distribution with pgf $P_2(z)$. Note that all n of the variables have the same secondary distribution. Determine the distribution of $S = S_1 + \cdots + S_n$.

7.5 Show that, for any pgf, $P^{(k)}(1) = \mathrm{E}[N(N-1)\cdots(N-k+1)]$ provided that the expectation exists. Here, $P^{(k)}(z)$ indicates the kth derivative. Use this result to confirm the three moments as given in (7.10).

7.6 Verify the three moments as given in (7.11).

7.3 Mixed-Frequency Distributions

7.3.1 The General Mixed-Frequency Distribution

Many compound distributions can arise in a way that is very different from compounding. In this section, we examine mixture distributions by treating one or more parameters as being "random" in some sense. This section expands on the ideas discussed in Section 6.3 in connection with the gamma mixture of the Poisson distribution being negative binomial.

We assume that the parameter is distributed over the population under consideration and that the sampling scheme that generates our data has two stages. First, a value of the parameter is selected. Then, given that parameter value, an observation is generated using that parameter value.

In automobile insurance, for example, classification schemes attempt to put individuals into (relatively) homogeneous groups for the purpose of pricing. Variables used to develop the classification scheme might include age, experience, a history of violations, accident history, and other variables. Because there will always be some residual variation in accident risk within each class, mixed distributions provide a framework for modeling this heterogeneity.

Let $P(z|\theta)$ denote the pgf of the number of events (e.g. claims) if the risk parameter is known to be θ. The parameter, θ, might be the Poisson mean, for example, in which case the measurement of risk is the expected number of events in a fixed time period.

Let $U(\theta) = \Pr(\Theta \le \theta)$ be the cdf of Θ, where Θ is the risk parameter, which is viewed as a random variable. Then, $U(\theta)$ represents the probability that, when a value of Θ is selected (e.g. a driver is included in our sample), the value of the risk parameter does not exceed θ. Let $u(\theta)$ be the pf or pdf of Θ. Then,

$$P(z) = \int P(z|\theta)u(\theta)\,d\theta \quad \text{or} \quad P(z) = \sum P(z|\theta_j)u(\theta_j) \qquad (7.12)$$

is the unconditional pgf of the number of events (where the formula selected depends on whether Θ is discrete or continuous). The corresponding probabilities are denoted by

$$p_k = \int p_k(\theta)u(\theta)\,d\theta \quad \text{or} \quad p_k = \sum p_k(\theta_j)u(\theta_j). \tag{7.13}$$

The **mixing distribution** denoted by $U(\theta)$ may be of the discrete or continuous type or even a combination of discrete and continuous types. **Discrete mixtures** are mixtures of distributions when the mixing function is of the discrete type; similarly for **continuous mixtures**. This phenomenon was introduced for continuous mixtures of severity distributions in Section 5.2.4 and for finite discrete mixtures in Section 4.2.3.

It should be noted that the mixing distribution is unobservable because the data are drawn from the mixed distribution.

■ **EXAMPLE 7.11**

Demonstrate that the zero-modified distributions may be created by using a two-point mixture.

Suppose that
$$P(z) = p(1) + (1 - p)P_2(z).$$

This is a (discrete) two-point mixture of a degenerate distribution that places all probability at zero and a distribution with pgf $P_2(z)$. From (7.8), this is also a compound Bernoulli distribution. □

Many mixed models can be constructed beginning with a simple distribution. Two examples are given here.

■ **EXAMPLE 7.12**

Determine the pf for a mixed binomial with a beta mixing distribution. This distribution is called binomial–beta, negative hypergeometric, or Polya–Eggenberger.

The beta distribution has pdf

$$u(q) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}q^{a-1}(1 - q)^{b-1}, \quad a > 0,\ b > 0.$$

Then, the mixed distribution has probabilities

$$\begin{aligned}
p_k &= \int_0^1 \binom{m}{k}q^k(1 - q)^{m-k}\frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}q^{a-1}(1 - q)^{b-1}\,dq \\[2mm]
&= \frac{\Gamma(a + b)\Gamma(m + 1)\Gamma(a + k)\Gamma(b + m - k)}{\Gamma(a)\Gamma(b)\Gamma(k + 1)\Gamma(m - k + 1)\Gamma(a + b + m)} \\[2mm]
&= \frac{\binom{-a}{k}\binom{-b}{m - k}}{\binom{-a - b}{m}}, \quad k = 0, 1, 2, \dots.
\end{aligned}$$

□

■ **EXAMPLE 7.13**

Determine the pf for a mixed negative binomial distribution with mixing on the parameter $p = (1 + \beta)^{-1}$. Let p have a beta distribution. The mixed distribution is called the **generalized Waring distribution**.

Arguing as in Example 7.12, we have

$$
\begin{aligned}
p_k &= \frac{\Gamma(r + k)}{\Gamma(r)\Gamma(k + 1)} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a+r-1}(1 - p)^{b+k-1} \, dp \\
&= \frac{\Gamma(r + k)}{\Gamma(r)\Gamma(k + 1)} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a + r)\Gamma(b + k)}{\Gamma(a + r + b + k)}, \quad k = 0, 1, 2, \dots.
\end{aligned}
$$

When $b = 1$, this distribution is called the **Waring distribution**. When $r = b = 1$, it is termed the **Yule distribution**. □

7.3.2 Mixed Poisson Distributions

If we let $p_k(\theta)$ in (7.13) have the Poisson distribution, this leads to a class of distributions with useful properties. A simple example of a Poisson mixture is the two-point mixture.

■ **EXAMPLE 7.14**

Suppose that drivers can be classified as "good drivers" and "bad drivers," each group with its own Poisson distribution. Determine the pf for this model and fit it to the data from Example 12.5. This model and its application to the data set are from Tröbliger [121].

From (7.13) the pf is

$$
p_k = p \frac{e^{-\lambda_1} \lambda_1^k}{k!} + (1 - p) \frac{e^{-\lambda_2} \lambda_2^k}{k!}.
$$

The maximum likelihood estimates[1] were calculated by Tröbliger to be $\hat{p} = 0.94$, $\hat{\lambda}_1 = 0.11$, and $\hat{\lambda}_2 = 0.70$. This means that about 94% of drivers were "good" with a risk of $\lambda_1 = 0.11$ expected accidents per year and 6% were "bad" with a risk of $\lambda_2 = 0.70$ expected accidents per year. Note that it is not possible to return to the data set and identify which were the bad drivers. □

This example illustrates two important points about finite mixtures. First, the model is probably oversimplified in the sense that risks (e.g. drivers) probably exhibit a continuum of risk levels rather than just two. The second point is that finite mixture models have a lot of parameters to be estimated. The simple two-point Poisson mixture has three parameters. Increasing the number of distributions in the mixture to r will then involve $r - 1$ mixing parameters in addition to the total number of parameters in the r component distributions. Consequently, continuous mixtures are frequently preferred.

The class of mixed Poisson distributions has some interesting properties that are developed here. Let $P(z)$ be the pgf of a mixed Poisson distribution with arbitrary mixing

---

[1]Maximum likelihood estimation is discussed in Chapter 11.

distribution $U(\theta)$. Then (with formulas given for the continuous case), by introducing a scale parameter λ, we have

$$P(z) = \int e^{\lambda\theta(z-1)}u(\theta)\,d\theta = \int \left[e^{\lambda(z-1)}\right]^{\theta} u(\theta)\,d\theta$$

$$= E\left\{\left[e^{\lambda(z-1)}\right]^{\theta}\right\} = M_{\Theta}\left[\lambda(z-1)\right], \tag{7.14}$$

where $M_{\Theta}(z)$ is the mgf of the mixing distribution.

Therefore, $P'(z) = \lambda M'_{\Theta}[\lambda(z-1)]$ and with $z = 1$ we obtain $E(N) = \lambda E(\Theta)$, where N has the mixed Poisson distribution. Also, $P''(z) = \lambda^2 M''_{\Theta}[\lambda(z-1)]$, implying that $E[N(N-1)] = \lambda^2 E(\Theta^2)$ and, therefore,

$$\begin{aligned}
\mathrm{Var}(N) &= E[N(N-1)] + E(N) - [E(N)]^2 \\
&= \lambda^2 E(\Theta^2) + E(N) - \lambda^2 [E(\Theta)]^2 \\
&= \lambda^2 \,\mathrm{Var}(\Theta) + E(N) \\
&> E(N),
\end{aligned}$$

and thus for mixed Poisson distributions the variance is always greater than the mean.

Most continuous distributions in this book involve a scale parameter. This means that scale changes to distributions do not cause a change in the form of the distribution, but only in the value of its scale parameter. For the mixed Poisson distribution, with pgf (7.14), any change in λ is equivalent to a change in the scale parameter of the mixing distribution. Hence, it may be convenient to simply set $\lambda = 1$ where a mixing distribution with a scale parameter is used.

Douglas [29] proves that for any mixed Poisson distribution, the mixing distribution is unique. This means that two different mixing distributions cannot lead to the same mixed Poisson distribution and this allows us to identify the mixing distribution in some cases.

There is also an important connection between mixed Poisson distributions and compound Poisson distributions.

Definition 7.6 *A distribution is said to be **infinitely divisible** if for all values of $n = 1, 2, 3, \ldots$ its characteristic function $\varphi(z)$ can be written as*

$$\varphi(z) = [\varphi_n(z)]^n,$$

where $\varphi_n(z)$ is the characteristic function of some random variable.

In other words, taking the $(1/n)$th power of the characteristic function still results in a characteristic function. The characteristic function is defined as follows.

Definition 7.7 *The **characteristic function** of a random variable X is*

$$\varphi_X(z) = E(e^{izX}) = E(\cos zX + i\sin zX),$$

where $i = \sqrt{-1}$.

In Definition 7.6, "characteristic function" could have been replaced by "moment generating function" or "probability generating function," or some other transform. That

is, if the definition is satisfied for one of these transforms, it will be satisfied for all others that exist for the particular random variable. We choose the characteristic function because it exists for all distributions, while the moment generating function does not exist for some distributions with heavy tails. Because many earlier results involved probability generating functions, it is useful to note the relationship between it and the characteristic function.

Theorem 7.8 *If the probability generating function exists for a random variable X, then $P_X(z) = \varphi(-i \ln z)$ and $\varphi_X(z) = P(e^{iz})$.*

Proof:

$$P_X(z) = \mathrm{E}(z^X) = \mathrm{E}(e^{X \ln z}) = \mathrm{E}[e^{-i(i \ln z)X}] = \varphi_X(-i \ln z)$$

and

$$\varphi_X(z) = \mathrm{E}(e^{izX}) = \mathrm{E}[(e^{iz})^X] = P_X(e^{iz}). \qquad \square$$

The following distributions, among others, are infinitely divisible: normal, gamma, Poisson, and negative binomial. The binomial distribution is not infinitely divisible because the exponent m in its pgf must take on integer values. Dividing m by $n = 1, 2, 3, \ldots$ will result in nonintegral values. In fact, no distributions with a finite range of support (the range over which positive probabilities exist) can be infinitely divisible. Now to the important result.

Theorem 7.9 *Suppose that $P(z)$ is a mixed Poisson pgf with an infinitely divisible mixing distribution. Then, $P(z)$ is also a compound Poisson pgf and may be expressed as*

$$P(z) = e^{\lambda[P_2(z)-1]},$$

where $P_2(z)$ is a pgf. If we insist that $P_2(0) = 0$, then $P_2(z)$ is unique.

A proof can be found in Feller [37, Chapter 12]. If we choose any infinitely divisible mixing distribution, the corresponding mixed Poisson distribution can be equivalently described as a compound Poisson distribution. For some distributions, this is a distinct advantage when carrying out numerical work, because the recursive formula (7.5) can be used in evaluating the probabilities once the secondary distribution is identified. For most cases, this identification is easily carried out. A second advantage is that, because the same distribution can be motivated in two different ways, a specific explanation is not required in order to use it. Conversely, the fact that one of these models fits well does not imply that it is the result of mixing or compounding. For example, the fact that claims follow a negative binomial distribution does not necessarily imply that individuals have the Poisson distribution and the Poisson parameter has a gamma distribution.

To obtain further insight into these results, we remark that if a counting distribution with pgf $P(z) = \sum_{n=0}^{\infty} p_n z^n$ is known to be of compound Poisson form (or, equivalently, is an infinitely divisible pgf), then the quantities λ and $P_2(z)$ in Theorem 7.9 may be expressed in terms of $P(z)$. Because $P_2(0) = 0$, it follows that $P(0) = p_0 = e^{-\lambda}$ or, equivalently,

$$\lambda = -\ln P(0). \tag{7.15}$$

Thus, using (7.15),

$$P_2(z) = 1 + \frac{1}{\lambda} \ln P(z) = \frac{\ln P(0) - \ln P(z)}{\ln P(0)}. \tag{7.16}$$

The following examples illustrate the use of these ideas.

■ **EXAMPLE 7.15**

Use the preceding results and (7.14) to express the negative binomial distribution in both mixed Poisson and compound Poisson form.

The moment generating function of the gamma distribution with pdf denoted by $u(\theta)$ is (from Example 3.7 with α replaced by r and θ replaced by β)

$$M_\Theta(t) = (1 - \beta t)^{-r} = \int_0^\infty e^{t\theta} u(\theta)\, d\theta, \quad t < 1/\beta.$$

This is clearly infinitely divisible because $\left[M_\Theta(t)\right]^{1/n}$ is the mgf of another gamma distribution with r replaced by r/n. Thus, using (7.14) with $\lambda = 1$ yields the negative binomial pgf

$$P(z) = M_\Theta(z - 1) = \int_0^\infty e^{\theta(z-1)} u(\theta)\, d\theta = [1 - \beta(z - 1)]^{-r}.$$

Because the gamma mixing distribution is infinitely divisible, Theorem 7.9 guarantees that the negative binomial distribution is also of compound Poisson form, in agreement with Example 7.5. The identification of the Poisson parameter λ and the secondary distribution in Example 7.5, although algebraically correct, does not provide as much insight as in the present discussion. In particular, from (7.15) we find directly that

$$\lambda = r \ln(1 + \beta)$$

and, from (7.16),

$$P_2(z) = \frac{-r \ln(1 + \beta) + r \ln[1 - \beta(z - 1)]}{-r \ln(1 + \beta)}$$

$$= \frac{\ln\left(\frac{1+\beta-\beta z}{1+\beta}\right)}{\ln(1 + \beta)^{-1}}$$

$$= \frac{\ln\left(1 - \frac{\beta}{1+\beta} z\right)}{\ln\left(1 - \frac{\beta}{1+\beta}\right)},$$

the logarithmic series pdf as before.   □

■ **EXAMPLE 7.16**

Show that a mixed Poisson with an inverse Gaussian mixing distribution is the same as a Poisson–ETNB distribution with $r = -0.5$.

The inverse Gaussian distribution is described in Appendix A. It has pdf

$$f(x) = \left(\frac{\theta}{2\pi x^3}\right)^{1/2} \exp\left[-\frac{\theta}{2x}\left(\frac{x - \mu}{\mu}\right)^2\right], \quad x > 0,$$

and mgf

$$M(t) = \int_0^\infty e^{tx} f(x)\, dx = \exp\left[\frac{\theta}{\mu}\left(1 - \sqrt{1 - \frac{2\mu^2}{\theta}t}\right)\right],$$

where $\theta > 0$ and $\mu > 0$ are parameters. Note that

$$[M(t)]^{1/n} = \exp\left[\frac{\theta}{n\mu}\left(1 - \sqrt{1 - \frac{2\mu^2}{\theta}t}\right)\right]$$

$$= \exp\left\{\frac{\theta/n^2}{\mu/n}\left[1 - \sqrt{1 - \frac{2(\mu/n)^2}{(\theta/n^2)}t}\right]\right\}.$$

This is the mgf of an inverse Gaussian distribution with θ replaced by θ/n^2 and μ by μ/n, and thus the inverse Gaussian distribution is infinitely divisible.

Hence, by Theorem 7.9, the Poisson mixed over the inverse Gaussian distribution is also compound Poisson. Its pgf is then, from (7.14) with $\lambda = 1$,

$$P(z) = M(z-1) = \exp\left\{\frac{\theta}{\mu}\left[1 - \sqrt{1 - \frac{2\mu^2}{\theta}(z-1)}\right]\right\},$$

which may be represented, using (7.15) and (7.16) in the compound Poisson form of Theorem 7.9 with

$$\lambda = -\ln P(0) = \frac{\theta}{\mu}\left(\sqrt{1 + \frac{2\mu^2}{\theta}} - 1\right),$$

and

$$P_2(z) = \frac{\frac{\theta}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2}{\theta}}\right) + \frac{\theta}{\mu}\left[\sqrt{1 - \frac{2\mu^2}{\theta}(z-1)} - 1\right]}{\frac{\theta}{\mu}\left(1 - \sqrt{1 + \frac{2\mu^2}{\theta}}\right)}$$

$$= \frac{\sqrt{1 - \frac{2\mu^2}{\theta}(z-1)} - \sqrt{1 + \frac{2\mu^2}{\theta}}}{1 - \sqrt{1 + \frac{2\mu^2}{\theta}}}.$$

We recognize that $P_2(z)$ is the pgf of an extended truncated negative binomial distribution with $r = -1/2$ and $\beta = 2\mu^2/\theta$. Unlike the negative binomial distribution, which is itself a member of the $(a, b, 0)$ class, the compound Poisson representation is of more use for computational purposes than the original mixed Poisson formulation. $\square$

It is not difficult to see that, if $u(\theta)$ is the pf for any discrete random variable with pgf $P_\Theta(z)$, then the pgf of the mixed Poisson distribution is $P_\Theta\left[\exp^{\lambda(z-1)}\right]$, a compound distribution with a Poisson secondary distribution.

Table 7.2   Pairs of compound and mixed Poisson distributions.

Name	Compound secondary distribution	Mixing distribution
Negative binomial	Logarithmic	Gamma
Neyman–Type A	Poisson	Poisson
Poisson–inverse Gaussian	ETNB ($r = -0.5$)	Inverse Gaussian

■ **EXAMPLE 7.17**

Demonstrate that the Neyman Type A distribution can be obtained by mixing.

If in (7.14) the mixing distribution has pgf

$$P_\Theta(z) = e^{\mu(z-1)},$$

then the mixed Poisson distribution has pgf

$$P(z) = \exp\{\mu[e^{\lambda(z-1)} - 1]\},$$

the pgf of a compound Poisson with a Poisson secondary distribution, that is, the Neyman Type A distribution. □

A further interesting result obtained by Holgate [57] is that, if a mixing distribution is absolutely continuous and unimodal, then the resulting mixed Poisson distribution is also unimodal. Multimodality can occur when discrete mixing functions are used. For example, the Neyman Type A distribution can have more than one mode. You should try this calculation for various combinations of the two parameters. The relationships between mixed and compound Poisson distributions are given in Table 7.2.

In this chapter, we focus on distributions that are easily handled computationally. Although many other discrete distributions are available, we believe that those discussed form a sufficiently rich class for most problems.

7.3.3 Exercises

7.7 Show that the negative binomial–Poisson compound distribution is the same as a mixed Poisson distribution with a negative binomial mixing distribution.

7.8 For $i = 1, \dots, n$ let N_i have a mixed Poisson distribution with parameter λ. Let the mixing distribution for N_i have pgf $P_i(z)$. Show that $N = N_1 + \cdots + N_n$ has a mixed Poisson distribution and determine the pgf of the mixing distribution.

7.9 Let N have a Poisson distribution with (given that $\Theta = \theta$) parameter $\lambda\theta$. Let the distribution of the random variable Θ have a scale parameter. Show that the mixed distribution does not depend on the value of λ.

7.10 Let N have a Poisson distribution with (given that $\Theta = \theta$) parameter θ. Let the distribution of the random variable Θ have pdf $u(\theta) = \alpha^2(\alpha + 1)^{-1}(\theta + 1)e^{-\alpha\theta}$, $\theta > 0$. Determine the pf of the mixed distribution. In addition, show that the mixed distribution is also a compound distribution.

7.11 Consider the mixed Poisson distribution

$$p_n = \int_0^\infty \frac{(\lambda\theta)^n e^{-\lambda\theta}}{n!} u(\theta)\, d\theta, \quad n = 0, 1, \ldots ,$$

where the pdf $u(\theta)$ is that of the positive stable distribution (see, e.g. Feller [38, pp. 448, 583]) given by

$$u(\theta) = \frac{1}{\pi} \sum_{k=1}^\infty \frac{\Gamma(k\alpha + 1)}{k!} (-1)^{k-1} \theta^{-k\alpha-1} \sin(k\alpha\pi), \quad \theta > 0,$$

where $0 < \alpha < 1$. The Laplace transform is $\int_0^\infty e^{-s\theta} u(\theta)\, d\theta = \exp(-s^\alpha)$, $s \geq 0$. Prove that $\{p_n; \ n = 0, 1, \ldots\}$ is a compound Poisson distribution with Sibuya secondary distribution (this mixed Poisson distribution is sometimes called a discrete stable distribution).

7.12 Consider the Poisson–inverse Gaussian distribution (Example 7.16) with pgf

$$P(z) = \sum_{n=0}^\infty p_n z^n = \exp\left\{ \frac{\mu}{\beta} \left[1 - \sqrt{1 + 2\beta(1 - z)} \right] \right\}.$$

Use the results of Exercise 5.20(g) to prove that

$$p_0 = \exp\left[\frac{\mu}{\beta} \left(1 - \sqrt{1 + 2\beta} \right) \right]$$

and, for $n = 1, 2, \ldots ,$

$$p_n = p_0 \frac{\mu^n}{n!} \sum_{j=0}^{n-1} \frac{(n + j - 1)!}{(n - j - 1)! j!} \left(\frac{\beta}{2\mu} \right)^j (1 + 2\beta)^{-\frac{n+j}{2}}.$$

7.13 Let N have a Poisson distribution with mean Λ, where Λ has the reciprocal inverse Gaussian distribution with mgf [from Exercise 5.20(d)] given by

$$M(t) = \sqrt{\frac{\theta}{\theta - 2t}} \exp\left[\frac{\theta}{\mu} \left(1 - \sqrt{1 - \frac{2}{\theta} t} \right) \right].$$

(a) Derive the pgf of N and explain in words what type of distribution has this pgf.

(b) Express the pgf in (a) in compound Poisson form, identifying the Poisson parameter and the secondary distribution.

7.14 Suppose the following:

- Λ is a random variable with support on the positive real numbers.

- $N|\Lambda = \lambda$ is a Poisson random variable with mean λ and N has pgf $P_N(z)$.

- $\Lambda_* = \Lambda + \mu$, where $\mu > 0$ is constant.

- $N_*|\Lambda_* = \lambda$ is a Poisson random variable with mean λ.

(a) Prove that the pgf of N_* is

$$P_{N_*}(x) = e^{\mu(z-1)} P_N(z)$$

and explain in words what type of distribution has this pgf.

(b) Assume Λ has an inverse Gaussian distribution. Describe carefully how you could recursively calculate the probabilities of the compound distribution with pgf $P_{N_*}[Q(z)]$, where $Q(z)$ is itself a pgf.

7.4   The Effect of Exposure on Frequency

Assume that the current portfolio consists of n entities, each of which could produce claims. Let N_j be the number of claims produced by the jth entity. Then, $N = N_1 + \cdots + N_n$. If we assume that the N_js are independent and identically distributed, then

$$P_N(z) = [P_{N_1}(z)]^n.$$

Now suppose that the portfolio is expected to expand to n^* entities with frequency N^*. Then,

$$P_{N^*}(z) = [P_{N_1}(z)]^{n^*} = [P_N(z)]^{n^*/n}.$$

Thus, if N is infinitely divisible, the distribution of N^* will have the same form as that of N, but with modified parameters.

■ **EXAMPLE 7.18**

It has been determined from past studies that the number of workers compensation claims for a group of 300 employees in a certain occupation class has a negative binomial distribution with $\beta = 0.3$ and $r = 10$. Determine the frequency distribution for a group of 500 such individuals.

The pgf of N^* is

$$P_{N^*}(z) = [P_N(z)]^{500/300} = \{[1 - 0.3(z-1)]^{-10}\}^{500/300}$$
$$= [1 - 0.3(z-1)]^{-16.67},$$

which is negative binomial with $\beta = 0.3$ and $r = 16.67$.   □

For the $(a, b, 0)$ class, all members except the binomial have this property. For the $(a, b, 1)$ class, none of the members do. For compound distributions, it is the primary distribution that must be infinitely divisible. In particular, compound Poisson and compound negative binomial (including the geometric) distributions will be preserved under an increase in exposure. Earlier, some reasons were given to support the use of zero-modified distributions. If exposure adjustments are anticipated, it may be better to choose a compound model, even if the fit is not quite as good. It should be noted that compound models have the ability to place large amounts of probability at zero.

Table 7.3   Relationships among discrete distributions.

Distribution	Is a special case of	Is a limiting case of
Poisson	ZM Poisson	Negative binomial
		Poisson–binomial
		Poisson–inverse Gaussian
		Polya–Aeppli[a]
		Neyman–Type A[b]
ZT Poisson	ZM Poisson	ZT negative binomial
ZM Poisson		ZM negative binomial
Geometric	Negative binomial,	Geometric–Poisson
	ZM geometric	
ZT geometric	ZT negative binomial	
ZM geometric	ZM negative binomial	
Logarithmic		ZT negative binomial
ZM logarithmic		ZM negative binomial
Binomial	ZM binomial	
Negative binomial	ZM negative binomial,	
	Poisson–ETNB	
Poisson–inverse Gaussian	Poisson–ETNB	
Polya–Aeppli	Poisson–ETNB	
Neyman–Type A		Poisson–ETNB

[a] Also called Poisson–geometric.

[b] Also called Poisson–Poisson.

7.5 An Inventory of Discrete Distributions

We have introduced the simple $(a, b, 0)$ class, generalized to the $(a, b, 1)$ class, and then used compounding and mixing to create a larger class of distributions. Calculation of the probabilities of these distributions can be carried out by using simple recursive procedures. In this section, we note that there are relationships among the various distributions similar to those of Section 5.3.2. The specific relationships are given in Table 7.3.

It is clear from earlier developments that members of the $(a, b, 0)$ class are special cases of members of the $(a, b, 1)$ class and that zero-truncated distributions are special cases of zero-modified distributions. The limiting cases are best discovered through the probability generating function, as was done in Section 6.3 where the Poisson distribution is shown to be a limiting case of the negative binomial distribution.

We have not listed compound distributions where the primary distribution is one of the two-parameter models, such as the negative binomial or Poisson–inverse Gaussian. They are excluded because these distributions are often themselves compound Poisson distributions and, as such, are generalizations of distributions already presented. This collection forms a particularly rich set of distributions in terms of shape. However, many other distributions are also possible and are discussed in Johnson, Kotz, and Kemp [65], Douglas [29], and Panjer and Willmot [100].

7.5.1 Exercises

7.15 Calculate $\Pr(N = 0)$, $\Pr(N = 1)$, and $\Pr(N = 2)$ for each of the following distributions:

(a) Poisson($\lambda = 4$)

(b) Geometric($\beta = 4$)

(c) Negative binomial($r = 2, \beta = 2$)

(d) Binomial($m = 8, q = 0.5$)

(e) Logarithmic($\beta = 4$)

(f) ETNB($r = -0.5, \beta = 4$)

(g) Poisson–inverse Gaussian($\lambda = 2, \beta = 4$)

(h) Zero-modified geometric($p_0^M = 0.5, \beta = 4$)

(i) Poisson–Poisson(Neyman Type A)($\lambda_{\text{primary}} = 4, \lambda_{\text{secondary}} = 1$)

(j) Poisson–ETNB($\lambda = 4, r = 2, \beta = 0.5$)

(k) Poisson–zero-modified geometric ($\lambda = 8, p_0^M = 0.5, r = 2, \beta = 0.5$)

7.16 A frequency model that has not been mentioned up to this point is the **zeta distribution**. It is a zero-truncated distribution with $p_k^T = k^{-(\rho+1)}/\zeta(\rho + 1)$, $k = 1, 2, \ldots, \rho > 0$. The denominator is the zeta function, which must be evaluated numerically as $\zeta(\rho + 1) = \sum_{k=1}^{\infty} k^{-(\rho+1)}$. The zero-modified zeta distribution can be formed in the usual way. More information can be found in Luong and Doray [84]. Verify that the zeta distribution is not a member of the $(a, b, 1)$ class.

7.17 For the discrete counting random variable N with probabilities $p_n = \Pr(N = n)$; $n = 0, 1, 2, \ldots$, let $a_n = \Pr(N > n) = \sum_{k=n+1}^{\infty} p_k$; $n = 0, 1, 2, \ldots$.

(a) Demonstrate that $\mathrm{E}(N) = \sum_{n=0}^{\infty} a_n$.

(b) Demonstrate that $A(z) = \sum_{n=0}^{\infty} a_n z^n$ and $P(z) = \sum_{n=0}^{\infty} p_n z^n$ are related by $A(z) = [1 - P(z)] / (1 - z)$. What happens as $z \to 1$?

(c) Suppose that N has the negative binomial distribution

$$p_n = \binom{n + r - 1}{n} \left(\frac{1}{1 + \beta} \right)^r \left(\frac{\beta}{1 + \beta} \right)^n, \quad n = 0, 1, 2, \ldots,$$

where r is a positive integer. Prove that

$$a_n = \beta \sum_{k=1}^{r} \binom{n + k - 1}{n} \left(\frac{1}{1 + \beta} \right)^k \left(\frac{\beta}{1 + \beta} \right)^n, \quad n = 0, 1, 2, \ldots.$$

(d) Suppose that N has the Sibuya distribution with pgf

$$P(z) = 1 - (1 - z)^{-r}, \quad -1 < r < 0.$$

Prove that

$$p_n = \frac{(-r)\Gamma(n + r)}{n!\Gamma(1 + r)}, \quad n = 1, 2, 3, \ldots,$$

and that

$$a_n = \binom{n+r}{n}, \; n = 0, 1, 2, \dots.$$

(e) Suppose that N has the mixed Poisson distribution with

$$p_n = \int_0^\infty \frac{(\lambda\theta)^n e^{-\lambda\theta}}{n!} dU(\theta), \; n = 0, 1, 2, \dots,$$

where $U(\theta)$ is a cumulative distribution function. Prove that

$$a_n = \lambda \int_0^\infty \frac{(\lambda\theta)^n e^{-\lambda\theta}}{n!} [1 - U(\theta)] d\theta, \; n = 0, 1, 2, \dots.$$

8

FREQUENCY AND SEVERITY WITH COVERAGE MODIFICATIONS

8.1 Introduction

We have seen a variety of examples that involve functions of random variables. In this chapter, we relate those functions to insurance applications. Throughout this chapter, we assume that all random variables have support on all or a subset of the nonnegative real numbers. At times in this chapter and later in the text, we need to distinguish between a random variable that measures the payment per loss (so zero is a possibility, taking place when there is a loss without a payment) and a variable that measures the payment per payment (the random variable is not defined when there is no payment). For notation, a per-loss variable is denoted Y^L and a per-payment variable is denoted Y^P. When the distinction is not material (e.g. setting a maximum payment does not create a difference), the superscript is omitted.

8.2   Deductibles

Insurance policies are often sold with a per-loss deductible of d. When the loss, x, is at or below d, the insurance pays nothing. When the loss is above d, the insurance pays $x - d$. In the language of Chapter 3, such a deductible can be defined as follows.

Definition 8.1 *An **ordinary deductible** modifies a random variable into either the excess loss or left censored and shifted variable (see Definition 3.3). The difference depends on whether the result of applying the deductible is to be per payment or per loss, respectively.*

This concept has already been introduced along with formulas for determining its moments. The per-payment variable is

$$Y^P = \begin{cases} \text{undefined}, & X \le d, \\ X - d, & X > d, \end{cases}$$

while the per-loss variable is

$$Y^L = \begin{cases} 0, & X \le d, \\ X - d, & X > d. \end{cases}$$

Note that the per-payment variable $Y^P = Y^L | Y^L > 0$. That is, the per-payment variable is the per-loss variable conditioned on the loss being positive. For the excess loss/per-payment variable, the density function is

$$f_{Y^P}(y) = \frac{f_X(y + d)}{S_X(d)}, \quad y > 0, \tag{8.1}$$

noting that for a discrete distribution, the density function need only be replaced by the probability function. Other key functions are

$$S_{Y^P}(y) = \frac{S_X(y + d)}{S_X(d)},$$

$$F_{Y^P}(y) = \frac{F_X(y + d) - F_X(d)}{1 - F_X(d)},$$

$$h_{Y^P}(y) = \frac{f_X(y + d)}{S_X(y + d)} = h_X(y + d).$$

Note that as a per-payment variable, the excess loss variable places no probability at zero.

The left censored and shifted variable has discrete probability at zero of $F_X(d)$, representing the probability that a payment of zero is made because the loss did not exceed d. Above zero, the density function is

$$f_{Y^L}(y) = f_X(y + d), \quad y > 0, \tag{8.2}$$

while the other key functions are[1] (for $y \geq 0$)

$$S_{Y^L}(y) = S_X(y + d),$$
$$F_{Y^L}(y) = F_X(y + d).$$

It is important to recognize that when counting claims on a per-payment basis, changing the deductible will change the frequency with which payments are made (while the frequency of losses will be unchanged). The nature of these changes is discussed in Section 8.6.

■ **EXAMPLE 8.1**

Determine the previously discussed functions for a Pareto distribution with $\alpha = 3$ and $\theta = 2{,}000$ for an ordinary deductible of 500.

Using the preceding formulas, for the excess loss variable,

$$f_{Y^P}(y) = \frac{3(2{,}000)^3(2{,}000 + y + 500)^{-4}}{(2{,}000)^3(2{,}000 + 500)^{-3}} = \frac{3(2{,}500)^3}{(2{,}500 + y)^4},$$

$$S_{Y^P}(y) = \left(\frac{2{,}500}{2{,}500 + y} \right)^3,$$

$$F_{Y^P}(y) = 1 - \left(\frac{2{,}500}{2{,}500 + y} \right)^3,$$

$$h_{Y^P}(y) = \frac{3}{2{,}500 + y}.$$

Note that this is a Pareto distribution with $\alpha = 3$ and $\theta = 2{,}500$. For the left censored and shifted variable,

$$f_{Y^L}(y) = \begin{cases} 0.488, & y = 0, \\ \dfrac{3(2{,}000)^3}{(2{,}500 + y)^4}, & y > 0, \end{cases}$$

$$S_{Y^L}(y) = \begin{cases} 0.512, & y = 0, \\ \dfrac{(2{,}000)^3}{(2{,}500 + y)^3}, & y > 0, \end{cases}$$

$$F_{Y^L}(y) = \begin{cases} 0.488, & y = 0, \\ 1 - \dfrac{(2{,}000)^3}{(2{,}500 + y)^3}, & y > 0, \end{cases}$$

$$h_{Y^L}(y) = \begin{cases} \text{undefined}, & y = 0, \\ \dfrac{3}{2{,}500 + y}, & y > 0. \end{cases}$$

---

[1] The hazard rate function is not presented because it is not defined at zero, making it of limited value. Note that for the excess loss variable, the hazard rate function is simply shifted.
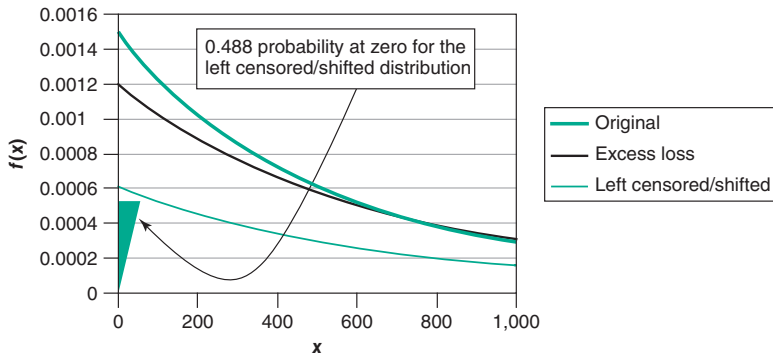
Figure 8.1   The densities for Example 8.1.

Figure 8.1 contains a plot of the densities. The modified densities are created as follows. For the excess loss variable, take the portion of the original density from 500 and above. Then, shift it to start at zero and multiply it by a constant so that the area under it is still 1. The left censored and shifted variable also takes the original density function above 500 and shifts it to the origin, but then leaves it alone. The remaining probability is concentrated at zero, rather than spread out. □

An alternative to the ordinary deductible is the franchise deductible. This deductible differs from the ordinary deductible in that, when the loss exceeds the deductible, the loss is paid in full. One example is in disability insurance where, for example, if a disability lasts seven or fewer days, no benefits are paid. However, if the disability lasts more than seven days, daily benefits are paid retroactively to the onset of the disability.

Definition 8.2 *A **franchise deductible** modifies the ordinary deductible by adding the deductible when there is a positive amount paid.*

The terms *left censored and shifted* and *excess loss* are not used here. Because this modification is unique to insurance applications, we use per-payment and per-loss terminology. The per-loss variable is

$$Y^L = \begin{cases} 0, & X \leq d, \\ X, & X > d, \end{cases}$$

while the per-payment variable is

$$Y^P = \begin{cases} \text{undefined}, & X \leq d, \\ X, & X > d. \end{cases}$$

Note that, as usual, the per-payment variable is a conditional random variable. The related

functions are now

$$f_{Y^L}(y) = \begin{cases} F_X(d), & y = 0, \\ f_X(y), & y > d, \end{cases}$$

$$S_{Y^L}(y) = \begin{cases} S_X(d), & 0 \le y \le d, \\ S_X(y), & y > d, \end{cases}$$

$$F_{Y^L}(y) = \begin{cases} F_X(d), & 0 \le y \le d, \\ F_X(y), & y > d, \end{cases}$$

$$h_{Y^L}(y) = \begin{cases} 0, & 0 < y < d, \\ h_X(y), & y > d, \end{cases}$$

for the per-loss variable and

$$f_{Y^P}(y) = \frac{f_X(y)}{S_X(d)}, \quad y > d,$$

$$S_{Y^P}(y) = \begin{cases} 1, & 0 \le y \le d, \\ \dfrac{S_X(y)}{S_X(d)}, & y > d, \end{cases}$$

$$F_{Y^P}(y) = \begin{cases} 0, & 0 \le y \le d, \\ \dfrac{F_X(y) - F_X(d)}{1 - F_X(d)}, & y > d, \end{cases}$$

$$h_{Y^P}(y) = \begin{cases} 0, & 0 < y < d, \\ h_X(y), & y > d, \end{cases}$$

for the per-payment variable.

■ **EXAMPLE 8.2**

Repeat Example 8.1 for a franchise deductible.

Using the preceding formulas for the per-payment variable, for $y > 500$,

$$f_{Y^P}(y) = \frac{3(2,000)^3(2,000 + y)^{-4}}{(2,000)^3(2,000 + 500)^{-3}} = \frac{3(2,500)^3}{(2,000 + y)^4},$$

$$S_{Y^P}(y) = \left(\frac{2,500}{2,000 + y} \right)^3,$$

$$F_{Y^P}(y) = 1 - \left(\frac{2,500}{2,000 + y} \right)^3,$$

$$h_{Y^P}(y) = \frac{3}{2,000 + y}.$$

For the per-loss variable,

$$f_{Y^L}(y) = \begin{cases} 0.488, & y = 0, \\ \dfrac{3(2{,}000)^3}{(2{,}000 + y)^4}, & y > 500, \end{cases}$$

$$S_{Y^L}(y) = \begin{cases} 0.512, & 0 \le y \le 500, \\ \dfrac{(2{,}000)^3}{(2{,}000 + y)^3}, & y > 500, \end{cases}$$

$$F_{Y^L}(y) = \begin{cases} 0.488, & 0 \le y \le 500, \\ 1 - \dfrac{(2{,}000)^3}{(2{,}000 + y)^3}, & y > 500, \end{cases}$$

$$h_{Y^L}(y) = \begin{cases} 0, & 0 < y < 500, \\ \dfrac{3}{2{,}000 + y}, & y > 500. \end{cases} \qquad \square$$

Expected costs for the two types of deductible may also be calculated.

Theorem 8.3 *For an ordinary deductible, the expected cost per loss is*

$$E(X) - E(X \wedge d)$$

and the expected cost per payment is

$$\frac{E(X) - E(X \wedge d)}{1 - F(d)}.$$

For a franchise deductible the expected cost per loss is

$$E(X) - E(X \wedge d) + d[1 - F(d)]$$

and the expected cost per payment is

$$\frac{E(X) - E(X \wedge d)}{1 - F(d)} + d.$$

Proof: For the per-loss expectation with an ordinary deductible, we have, from (3.7) and (3.10), that the expectation is $E(X) - E(X \wedge d)$. From (8.1) and (8.2) we see that, to change to a per-payment basis, division by $1 - F(d)$ is required. The adjustments for the franchise deductible come from the fact that when there is a payment, it will exceed that for the ordinary deductible by d. $\square$

■   **EXAMPLE 8.3**

Determine the four expectations for the Pareto distribution from Examples 8.1 and 8.2, using a deductible of 500.

Expectations could be derived directly from the density functions obtained in Examples 8.1 and 8.2. Using Theorem 8.3 and recognizing that we have a Pareto distribution, we can also look up the required values (the formulas are in Appendix A). That is,

$$F(500) = 1 - \left(\frac{2{,}000}{2{,}000 + 500} \right)^3 = 0.488,$$

$$E(X \wedge 500) = \frac{2{,}000}{2} \left[1 - \left(\frac{2{,}000}{2{,}000 + 500} \right)^2 \right] = 360.$$

With $E(X) = 1{,}000$, for the ordinary deductible, the expected cost per loss is $1{,}000 - 360 = 640$, while the expected cost per payment is $640/0.512 = 1{,}250$. For the franchise deductible, the expectations are $640 + 500(1 - 0.488) = 896$ and $1{,}250 + 500 = 1{,}750$. $\qquad\qquad\square$

8.2.1 Exercises

8.1 Perform the calculations in Example 8.1 for the following distribution, using an ordinary deductible of 5,000:

$$F_4(x) = \begin{cases} 0, & x < 0, \\ 1 - 0.3e^{-0.00001x}, & x \geq 0. \end{cases}$$

8.2 Repeat Exercise 8.1 for a franchise deductible.

8.3 Repeat Example 8.3 for the model in Exercise 8.1 and a 5,000 deductible.

8.4 (*) Risk 1 has a Pareto distribution with parameters $\alpha > 2$ and θ. Risk 2 has a Pareto distribution with parameters 0.8α and θ. Each risk is covered by a separate policy, each with an ordinary deductible of k. Determine the expected cost per loss for risk 1. Determine the limit as k goes to infinity of the ratio of the expected cost per loss for risk 2 to the expected cost per loss for risk 1.

8.5 (*) Losses (prior to any deductibles being applied) have a distribution as reflected in Table 8.1. There is a per-loss ordinary deductible of 10,000. The deductible is then raised so that half the number of losses exceed the new deductible as exceeded the old deductible. Determine the percentage change in the expected cost per payment when the deductible is raised.

Table 8.1 The data for Exercise 8.5.

x	$F(x)$	$E(X \wedge x)$
10,000	0.60	6,000
15,000	0.70	7,700
22,500	0.80	9,500
32,500	0.90	11,000
∞	1.00	20,000

8.3   The Loss Elimination Ratio and the Effect of Inflation for Ordinary Deductibles

A ratio that can be meaningful in evaluating the effect of a deductible is the loss elimination ratio (LER).

Definition 8.4 *The **loss elimination ratio** is the ratio of the decrease in the expected payment with an ordinary deductible to the expected payment without the deductible.*

While many types of coverage modifications can decrease the expected payment, the term loss elimination ratio is reserved for the effect of changing the deductible. Without the deductible, the expected payment is $E(X)$. With the deductible, the expected payment (from Theorem 8.3) is $E(X) - E(X \wedge d)$. Therefore, the loss elimination ratio is

$$\frac{E(X) - [E(X) - E(X \wedge d)]}{E(X)} = \frac{E(X \wedge d)}{E(X)}$$

provided that $E(X)$ exists.

■   **EXAMPLE 8.4**

Determine the loss elimination ratio for the Pareto distribution with $\alpha = 3$ and $\theta = 2{,}000$ with an ordinary deductible of 500.

From Example 8.3, we have a loss elimination ratio of $360/1{,}000 = 0.36$. Thus 36% of losses can be eliminated by introducing an ordinary deductible of 500.   □

Inflation increases costs, but it turns out that when there is a deductible, the effect of inflation is magnified. First, some events that formerly produced losses below the deductible will now lead to payments. Second, the relative effect of inflation is magnified because the deductible is subtracted after inflation. For example, suppose that an event formerly produced a loss of 600. With a 500 deductible, the payment is 100. Inflation at 10% will increase the loss to 660 and the payment to 160, a 60% increase in the cost to the insurer.

Theorem 8.5 *For an ordinary deductible of d after uniform inflation of $1 + r$, the expected cost per loss is*

$$(1 + r)\{E(X) - E[X \wedge d/(1 + r)]\}.$$

If $F[d/(1 + r)] < 1$, then the expected cost per payment is obtained by dividing by $1 - F[d/(1 + r)]$.

Proof: After inflation, losses are given by the random variable $Y = (1 + r)X$. From

Theorem 5.1, $f_Y(y) = f_X[y/(1+r)]/(1+r)$ and $F_Y(y) = F_X[y/(1+r)]$. Using (3.8),

$$
\begin{aligned}
E(Y \wedge d) &= \int_0^d y f_Y(y)\, dy + d[1 - F_Y(d)] \\
&= \int_0^d \frac{y f_X[y/(1+r)]}{1+r}\, dy + d\left[1 - F_X\left(\frac{d}{1+r}\right)\right] \\
&= \int_0^{d/(1+r)} (1+r)x f_X(x)\, dx + d\left[1 - F_X\left(\frac{d}{1+r}\right)\right] \\
&= (1+r)\left\{\int_0^{d/(1+r)} x f_X(x)\, dx + \frac{d}{1+r}\left[1 - F_X\left(\frac{d}{1+r}\right)\right]\right\} \\
&= (1+r)E\left(X \wedge \frac{d}{1+r}\right),
\end{aligned}
$$

where the third line follows from the substitution $x = y/(1+r)$. Then, $E(Y) = (1+r)E(X)$ completes the first statement of the theorem. The per-payment result follows from the relationship between the distribution functions of Y and X. ☐

■ EXAMPLE 8.5

Determine the effect of inflation at 10% on an ordinary deductible of 500 applied to a Pareto distribution with $\alpha = 3$ and $\theta = 2{,}000$.

From Example 8.3 the expected costs are 640 and 1,250 per loss and per payment, respectively. With 10% inflation, we need

$$
E\left(X \wedge \frac{500}{1.1}\right) = E(X \wedge 454.55)
$$

$$
= \frac{2{,}000}{2}\left[1 - \left(\frac{2{,}000}{2{,}000 + 454.55}\right)^2\right] = 336.08.
$$

The expected cost per loss after inflation is $1.1(1{,}000 - 336.08) = 730.32$, an increase of 14.11%. On a per-payment basis, we need

$$
F_Y(500) = F_X(454.55)
$$

$$
= 1 - \left(\frac{2{,}000}{2{,}000 + 454.55}\right)^3
$$

$$
= 0.459.
$$

The expected cost per payment is $730.32/(1 - 0.459) = 1{,}350$, an increase of 8%. ☐

8.3.1 Exercises

8.6 Determine the loss elimination ratio for the distribution given here, with an ordinary deductible of 5,000. This is the same model as used in Exercise 8.1.

$$
F_4(x) = \begin{cases} 0, & x < 0, \\ 1 - 0.3e^{-0.00001x}, & x \geq 0. \end{cases}
$$

Table 8.2   The data for Exercise 8.11.

x	$F(x)$	$E(X \wedge x)$
10,000	0.60	6,000
15,000	0.70	7,700
22,500	0.80	9,500
∞	1.00	20,000

8.7 Determine the effect of inflation at 10% on an ordinary deductible of 5,000 applied to the distribution in Exercise 8.6.

8.8 (*) Losses have a lognormal distribution with $\mu = 7$ and $\sigma = 2$. There is a deductible of 2,000, and 10 losses are expected each year. Determine the loss elimination ratio. If there is uniform inflation of 20% but the deductible remains at 2,000, how many payments will be expected?

8.9 (*) Losses have a Pareto distribution with $\alpha = 2$ and $\theta = k$. There is an ordinary deductible of $2k$. Determine the loss elimination ratio before and after 100% inflation.

8.10 (*) Losses have an exponential distribution with a mean of 1,000. There is a deductible of 500. Determine the amount by which the deductible would have to be raised to double the loss elimination ratio.

8.11 (*) The values in Table 8.2 are available for a random variable X. There is a deductible of 15,000 per loss and no policy limit. Determine the expected cost per payment using X and then assuming 50% inflation (with the deductible remaining at 15,000).

8.12 (*) Losses have a lognormal distribution with $\mu = 6.9078$ and $\sigma = 1.5174$. Determine the ratio of the loss elimination ratio at 10,000 to the loss elimination ratio at 1,000. Then determine the percentage increase in the number of losses that exceed 1,000 if all losses are increased by 10%.

8.13 (*) Losses have a mean of 2,000. With a deductible of 1,000, the loss elimination ratio is 0.3. The probability of a loss being greater than 1,000 is 0.4. Determine the average size of a loss, given that it is less than or equal to 1,000.

8.4 Policy Limits

The opposite of a deductible is a **policy limit**. The typical policy limit arises in a contract where for losses below u the insurance pays the full loss but for losses above u the insurance pays only u. The effect of the limit is to produce a right censored random variable. It will have a mixed distribution with distribution and density function given by (where Y is the random variable after the limit has been imposed)

$$F_Y(y) = \begin{cases} F_X(y), & y < u, \\ 1, & y \geq u \end{cases}$$

and

$$f_Y(y) = \begin{cases} f_X(y), & y < u, \\ 1 - F_X(u), & y = u. \end{cases}$$

The effect of inflation can be calculated as follows.

Theorem 8.6 *For a policy limit of u, after uniform inflation of $1 + r$, the expected cost is $(1 + r)\mathrm{E}[X \wedge u/(1 + r)]$.*

Proof: The expected cost is $\mathrm{E}(Y \wedge u)$. The proof of Theorem 8.5 shows that this equals the expression given in this theorem.   □

For policy limits, the concepts of per payment and per loss are not relevant. All losses that produced payments prior to imposing the limit will produce payments after the limit is imposed.

■ **EXAMPLE 8.6**

Impose a limit of 3,000 on a Pareto distribution with $\alpha = 3$ and $\theta = 2,000$. Determine the expected cost per loss with the limit as well as the proportional reduction in expected cost. Repeat these calculations after 10% uniform inflation is applied.

For this Pareto distribution, the expected cost is

$$\mathrm{E}(X \wedge 3,000) = \frac{2,000}{2}\left[1 - \left(\frac{2,000}{2,000 + 3,000}\right)^2\right] = 840$$

and the proportional reduction is $(1,000 - 840)/1,000 = 0.16$. After inflation, the expected cost is

$$1.1\mathrm{E}(X \wedge 3,000/1.1) = 1.1\frac{2,000}{2}\left[1 - \left(\frac{2,000}{2,000 + 3,000/1.1}\right)^2\right] = 903.11$$

for a proportional reduction of $(1,100 - 903.11)/1,100 = 0.179$. Note also that after inflation the expected cost has increased by 7.51%, less than the general inflation rate. The effect is the opposite of the deductible – inflation is tempered, not exacerbated.

Figure 8.2 shows the density function for the right censored random variable. From 0 to 3,000, it matches the original Pareto distribution. The probability of exceeding 3,000, $\Pr(X > 3,000) = (2,000/5,000)^3 = 0.064$, is concentrated at 3,000.   □

A policy limit and an ordinary deductible go together in the sense that, whichever applies to the insurance company's payments, the other applies to the policyholder's payments. For example, when the policy has a deductible of 500, the cost per loss to the policyholder is a random variable that is right censored at 500. When the policy has a limit of 3,000, the policyholder's payments are a variable that is left truncated and shifted (as in an ordinary deductible). The opposite of the franchise deductible is a coverage that right truncates any losses (see Exercise 3.12). This coverage is rarely, if ever, sold. (Would you buy a policy that pays you nothing if your loss exceeds u?)
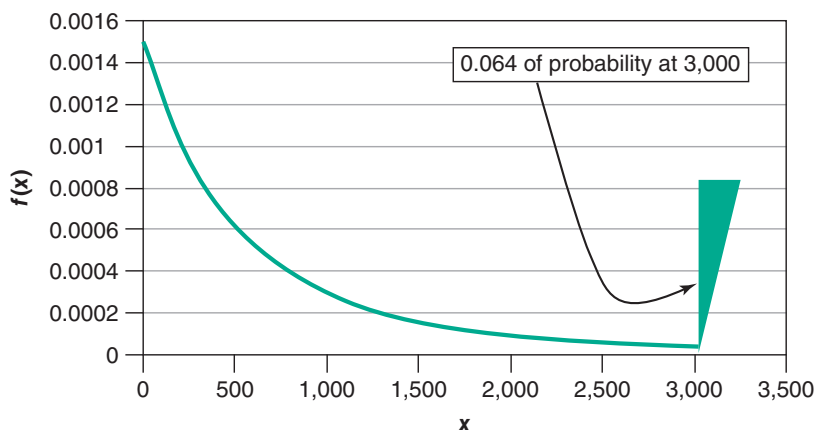
Figure 8.2   The density function for Example 8.6.

8.4.1   Exercises

8.14 Determine the effect of 10% inflation on a policy limit of 150,000 on the following distribution. This is the same distribution as used in Exercises 8.1 and 8.6.

$$F_4(x) = \begin{cases} 0, & x < 0, \\ 1 - 0.3e^{-0.00001x}, & x \ge 0. \end{cases}$$

8.15 (*) Let X have a Pareto distribution with $\alpha = 2$ and $\theta = 100$. Determine the range of the mean excess loss function $e(d)$ as d ranges over all positive numbers. Then, let $Y = 1.1X$. Determine the range of the ratio $e_Y(d)/e_X(d)$ as d ranges over all positive numbers. Finally, let Z be X right censored at 500 (i.e. a limit of 500 is applied to X). Determine the range of $e_Z(d)$ as d ranges over the interval 0 to 500.

8.5   Coinsurance, Deductibles, and Limits

The final common coverage modification is coinsurance. In this case, the insurance company pays a proportion, α, of the loss and the policyholder pays the remaining fraction. If coinsurance is the only modification, this changes the loss variable X to the payment variable, $Y = \alpha X$. The effect of multiplication has already been covered. When all four items covered in this chapter are present (ordinary deductible, limit, coinsurance, and inflation), we create the following per-loss random variable:

$$Y^L = \begin{cases} 0, & X < \dfrac{d}{1+r}, \\ \alpha[(1+r)X - d], & \dfrac{d}{1+r} \le X < \dfrac{u}{1+r}, \\ \alpha(u - d), & X \ge \dfrac{u}{1+r}. \end{cases}$$

For this definition, the quantities are applied in a particular order. In particular, the coinsurance is applied last. For the illustrated contract, the policy limit is $\alpha(u - d)$, the

maximum amount payable. In this definition, u is the loss above which no additional benefits are paid and is called the **maximum covered loss**. For the per-payment variable, Y^P is undefined for $X < d/(1 + r)$.

Previous results can be combined to produce the following theorem, presented without proof.

Theorem 8.7 *For the per-loss variable,*

$$\text{E}(Y^L) = \alpha(1 + r)\left[\text{E}\left(X \wedge \frac{u}{1 + r}\right) - \text{E}\left(X \wedge \frac{d}{1 + r}\right)\right].$$

The expected value of the per-payment variable is obtained as

$$\text{E}(Y^P) = \frac{\text{E}(Y^L)}{1 - F_X\left(\frac{d}{1+r}\right)}.$$

Higher moments are more difficult. Theorem 8.8 gives the formula for the second moment. The variance can then be obtained by subtracting the square of the mean.

Theorem 8.8 *For the per-loss variable,*

$$\text{E}\left[(Y^L)^2\right] = \alpha^2(1 + r)^2 \{\text{E}[(X \wedge u^*)^2] - \text{E}[(X \wedge d^*)^2]$$
$$- 2d^*\text{E}(X \wedge u^*) + 2d^*\text{E}(X \wedge d^*)\},$$

where $u^* = u/(1 + r)$ and $d^* = d/(1 + r)$. For the second moment of the per-payment variable, divide this expression by $1 - F_X(d^*)$.

Proof: From the definition of Y^L,

$$Y^L = \alpha(1 + r)[(X \wedge u^*) - (X \wedge d^*)]$$

and, therefore,

$$\frac{(Y^L)^2}{[\alpha(1 + r)]^2} = [(X \wedge u^*) - (X \wedge d^*)]^2$$
$$= (X \wedge u^*)^2 + (X \wedge d^*)^2 - 2(X \wedge u^*)(X \wedge d^*)$$
$$= (X \wedge u^*)^2 - (X \wedge d^*)^2 - 2(X \wedge d^*)[(X \wedge u^*) - (X \wedge d^*)].$$

The final term on the right-hand side can be written as

$$2(X \wedge d^*)[(X \wedge u^*) - (X \wedge d^*)] = 2d^*[(X \wedge u^*) - (X \wedge d^*)].$$

To see this, note that when $X < d^*$, both sides equal zero; when $d^* \leq X < u^*$, both sides equal $2d^*(X - d^*)$; and when $X \geq u^*$, both sides equal $2d^*(u^* - d^*)$. Make this substitution and take expectations on each side to complete the proof.[2]   □

---

[2]Thanks to Ken Burton for providing this improved proof.

■ **EXAMPLE 8.7**

Determine the mean and standard deviation per loss for a Pareto distribution with $\alpha = 3$ and $\theta = 2,000$ with a deductible of 500 and a policy limit of 2,500. Note that the maximum covered loss is $u = 3,000$.

From earlier examples, $E(X \wedge 500) = 360$ and $E(X \wedge 3,000) = 840$. The second limited moment is

$$
\begin{aligned}
E[(X \wedge u)^2] &= \int_0^u x^2 \frac{3(2,000)^3}{(x+2,000)^4} \, dx + u^2 \left(\frac{2,000}{u+2,000} \right)^3 \\
&= 3(2,000)^3 \int_{2,000}^{u+2,000} (y-2,000)^2 y^{-4} dy + u^2 \left(\frac{2,000}{u+2,000} \right)^3 \\
&= 3(2,000)^3 \left(-y^{-1} + 2,000y^{-2} - \frac{2,000^2}{3} y^{-3} \Big|_{2,000}^{u+2,000} \right) \\
&\quad + u^2 \left(\frac{2,000}{u+2,000} \right)^3 \\
&= 3(2,000)^3 \left[-\frac{1}{u+2,000} + \frac{2,000}{(u+2,000)^2} - \frac{2,000^2}{3(u+2,000)^3} \right] \\
&\quad + 3(2,000)^3 \left[\frac{1}{2,000} - \frac{2,000}{2,000^2} + \frac{2,000^2}{3(2,000)^3} \right] \\
&\quad + u^2 \left(\frac{2,000}{u+2,000} \right)^3 \\
&= (2,000)^2 - \left(\frac{2,000}{u+2,000} \right)^3 (2u+2,000)(u+2,000).
\end{aligned}
$$

Then, $E[(X \wedge 500)^2] = 160,000$ and $E[(X \wedge 3,000)^2] = 1,440,000$, and so

$$E(Y) = 840 - 360 = 480,$$
$$E(Y^2) = 1,440,000 - 160,000 - 2(500)(840) + 2(500)(360) = 800,000$$

for a variance of $800,000 - 480^2 = 569,600$ and a standard deviation of 754.72. □

8.5.1 Exercises

8.16 (*) You are given that $e(0) = 25$, $S(x) = 1 - x/w$, $0 \le x \le w$, and Y^P is the excess loss variable for $d = 10$. Determine the variance of Y^P.

8.17 (*) The loss ratio (R) is defined as total losses (L) divided by earned premiums (P). An agent will receive a bonus (B) if the loss ratio on his business is less than 0.7. The bonus is given as $B = P(0.7 - R)/3$ if this quantity is positive; otherwise, it is zero. Let $P = 500,000$ and L have a Pareto distribution with parameters $\alpha = 3$ and $\theta = 600,000$. Determine the expected value of the bonus.

8.18 (*) Losses this year have a distribution such that $E(X \wedge d) = -0.025d^2 + 1.475d - 2.25$ for $d = 10, 11, 12, \dots, 26$. Next year, losses will be uniformly higher by 10%. An insurance policy reimburses 100% of losses subject to a deductible of 11 up to a maximum reimbursement of 11. Determine the ratio of next year's reimbursements to this year's reimbursements.

8.19 (*) Losses have an exponential distribution with a mean of 1,000. An insurance company will pay the amount of each claim in excess of a deductible of 100. Determine the variance of the amount paid by the insurance company for one claim, including the possibility that the amount paid is zero.

8.20 (*) Total claims for a health plan have a Pareto distribution with $\alpha = 2$ and $\theta = 500$. The health plan implements an incentive to physicians that will pay a bonus of 50% of the amount by which total claims are less than 500; otherwise, no bonus is paid. It is anticipated that with the incentive plan the claim distribution will change to become Pareto with $\alpha = 2$ and $\theta = K$. With the new distribution, it turns out that expected claims plus the expected bonus is equal to expected claims prior to the bonus system. Determine the value of K.

8.21 (*) In year a, total expected losses are 10,000,000. Individual losses in year a have a Pareto distribution with $\alpha = 2$ and $\theta = 2,000$. A reinsurer pays the excess of each individual loss over 3,000. For this, the reinsurer is paid a premium equal to 110% of expected covered losses. In year b, losses will experience 5% inflation over year a, but the frequency of losses will not change. Determine the ratio of the premium in year b to the premium in year a.

8.22 (*) Losses have a uniform distribution from 0 to 50,000. There is a per-loss deductible of 5,000 and a policy limit of 20,000 (meaning that the maximum covered loss is 25,000). Determine the expected payment given that a payment has been made.

8.23 (*) Losses have a lognormal distribution with $\mu = 10$ and $\sigma = 1$. For losses below 50,000, no payment is made. For losses between 50,000 and 100,000, the full amount of the loss is paid. For losses in excess of 100,000, the limit of 100,000 is paid. Determine the expected cost per loss.

8.24 (*) The loss severity random variable X has an exponential distribution with mean 10,000. Determine the coefficient of variation of the variables Y^P and Y^L based on $d = 30,000$.

8.25 (*) The claim size distribution is uniform over the intervals $(0, 50), (50, 100), (100, 200)$, and $(200, 400)$. Of the total probability, 30% is in the first interval, 36% in the second interval, 18% in the third interval, and 16% in the fourth interval. Determine $E[(X \wedge 350)^2]$.

8.26 (*) Losses follow a two-parameter Pareto distribution with $\alpha = 2$ and $\theta = 5,000$. An insurance policy pays the following for each loss. There is no insurance payment for the first 1,000. For losses between 1,000 and 6,000, the insurance pays 80%. Losses above 6,000 are paid by the insured until the insured has made a total payment of 10,000. For

any remaining part of the loss, the insurance pays 90%. Determine the expected insurance payment per loss.

8.27 (*) The amount of a loss has a Poisson distribution with mean $\lambda = 3$. Consider two insurance contracts. One has an ordinary deductible of 2. The second one has no deductible and a coinsurance in which the insurance company pays α of the loss. Determine the value of α so that the expected cost of the two contracts is the same.

8.28 (*) The amount of a loss has cdf $F(x) = (x/100)^2$, $0 \le x \le 100$. An insurance policy pays 80% of the amount of a loss in excess of an ordinary deductible of 20. The maximum payment is 60 per loss. Determine the expected payment, given that a payment has been made.

8.6   The Impact of Deductibles on Claim Frequency

An important component in analyzing the effect of policy modifications pertains to the change in the frequency distribution of payments when the deductible (ordinary or franchise) is imposed or changed. When a deductible is imposed or increased, there will be fewer payments per period, while if a deductible is lowered, there will be more payments.

We can quantify this process if it can be assumed that the imposition of coverage modifications does not affect the process that produces losses or the type of individual who will purchase insurance. For example, those who buy a 250 deductible on an automobile property damage coverage may (correctly) view themselves as less likely to be involved in an accident than those who buy full coverage. Similarly, an employer may find that the rate of permanent disability declines when reduced benefits are provided to employees in the first few years of employment.

To begin, suppose that X_j, the severity, represents the ground-up loss on the jth such loss and there are no coverage modifications. Let N^L denote the number of losses. Now consider a coverage modification such that v is the probability that a loss will result in a payment. For example, if there is a deductible of d, $v = \Pr(X > d)$. Next, define the indicator random variable I_j by $I_j = 1$ if the jth loss results in a payment and $I_j = 0$ otherwise. Then, I_j has a Bernoulli distribution with parameter v and the pgf of I_j is $P_{I_j}(z) = 1 - v + vz$. Then, $N^P = I_1 + \cdots + I_{N^L}$ represents the number of payments. If $I_1, I_2, \ldots$ are mutually independent and are also independent of N^L, then N^P has a compound distribution with N^L as the primary distribution and a Bernoulli secondary distribution. Thus

$$P_{N^P}(z) = P_{N^L}[P_{I_j}(z)] = P_{N^L}[1 + v(z - 1)].$$

In the important special case where the distribution of N^L depends on a parameter θ such that

$$P_{N^L}(z) = P_{N^L}(z; \theta) = B[\theta(1 - z)],$$

where $B(z)$ is functionally independent of θ (as in Theorem 7.4), then

$$
\begin{aligned}
P_{N^P}(z) &= B[\theta(1 - 1 - vz + v)] \\
&= B[v\theta(1 - z)] \\
&= P_{N^L}(z; v\theta).
\end{aligned}
$$

This result implies that N^L and N^P are both from the same parametric family and only the parameter θ need be changed.

■ **EXAMPLE 8.8**

Demonstrate that the preceding result applies to the negative binomial distribution, and illustrate the effect when a deductible of 250 is imposed on a negative binomial distribution with $r = 2$ and $\beta = 3$. Assume that losses have a Pareto distribution with $\alpha = 3$ and $\theta = 1{,}000$.

The negative binomial pgf is $P_{N^L}(z) = [1 - \beta(z - 1)]^{-r}$. Here, β takes on the role of θ in the result and $B(z) = (1 + z)^{-r}$. Then, N^P must have a negative binomial distribution with $r^* = r$ and $\beta^* = \upsilon\beta$. For the particular situation described,

$$\upsilon = 1 - F(250) = \left(\frac{1{,}000}{1{,}000 + 250} \right)^3 = 0.512,$$

and so $r^* = 2$ and $\beta^* = 3(0.512) = 1.536$.   □

This result may be generalized for zero-modified and zero-truncated distributions. Suppose that N^L depends on parameters θ and α such that

$$P_{N^L}(z) = P_{N^L}(z; \theta, \alpha) = \alpha + (1 - \alpha)\frac{\phi[\theta(1 - z)] - \phi(\theta)}{1 - \phi(\theta)}. \tag{8.3}$$

Note that $\alpha = P_{N^L}(0) = \Pr(N^L = 0)$ and so is the modified probability at zero. It is also the case that, if $\phi[\theta(1 - z)]$ is itself a pgf, then the pgf given in (8.3) is that for the corresponding zero-modified distribution. However, it is not necessary for $\phi[\theta(1 - z)]$ to be a pgf in order for $P_{N^L}(z)$ as given in (8.3) to be a pgf. In particular, $\phi(z) = 1 + \ln(1 + z)$ yields the ZM logarithmic distribution, even though there is no distribution with $\phi(z)$ as its pgf. Similarly, $\phi(z) = (1 + z)^{-r}$ for $-1 < r < 0$ yields the ETNB distribution. A few algebraic steps reveal (see Exercise 8.35) that for (8.3)

$$P_{N^P}(z) = P_{N^L}(z; \upsilon\theta, \alpha^*),$$

where $\alpha^* = \Pr(N^P = 0) = P_{N^P}(0) = P_{N^L}(1 - \upsilon; \theta, \alpha)$. It is expected that the imposition of a deductible will increase the value of α because periods with no payments will become more likely. In particular, if N^L is zero truncated, N^P will be zero modified.

■ **EXAMPLE 8.9**

Repeat Example 8.8, only now let the frequency distribution be zero-modified negative binomial with $r = 2$, $\beta = 3$, and $p_0^M = 0.4$.

The pgf is

$$P_{N^L}(z) = p_0^M + (1 - p_0^M)\frac{[1 - \beta(z - 1)]^{-r} - (1 + \beta)^{-r}}{1 - (1 + \beta)^{-r}}.$$

Then, $\alpha = p_0^M$ and $\phi(z) = (1 + z)^{-r}$. We then have $r^* = r$, $\beta^* = \upsilon\beta$, and

$$\alpha^* = p_0^{M*} = p_0^M + (1 - p_0^M)\frac{[1 + \upsilon\beta)]^{-r} - (1 + \beta)^{-r}}{1 - (1 + \beta)^{-r}}$$

$$= \frac{p_0^M - (1 + \beta)^{-r} + (1 + \upsilon\beta)^{-r} - p_0^M(1 + \upsilon\beta)^{-r}}{1 - (1 + \beta)^{-r}}.$$

For the particular distribution given, the new parameters are $r^* = 2$, $\beta^* = 3(0.512) = 1.536$, and

$$p_0^{M*} = \frac{0.4 - 4^{-2} + 2.536^{-2} - 0.4(2.536)^{-2}}{1 - 4^{-2}} = 0.4595.$$ ☐

In applications, it may be the case that we want to determine the distribution of N^L from that of N^P. For example, data may have been collected on the number of payments in the presence of a deductible and from that data the parameters of N^P can be estimated. We may then want to know the distribution of payments if the deductible is removed. Arguing as before,

$$P_{N^L}(z) = P_{N^P}(1 - v^{-1} + zv^{-1}).$$

This result implies that the formulas derived previously hold with v replaced by $1/v$. However, it is possible that the resulting pgf for N^L is not valid. In this case, one of the modeling assumptions is invalid (e.g. the assumption that changing the deductible does not change claim-related behavior). It is instructive to note that the resulting pgf for N^L is guaranteed to be a valid pgf [regardless of the value of $v \in (0, 1)$] if and only if the pgf for N^P is of mixed Poisson form (for a proof of this fact, see Grandell [47, pp. 25–26]). This provides yet another strong motivation for the use of mixed Poisson models for insurance claim counts. The following two examples illustrate this point.

■ **EXAMPLE 8.10**

Suppose that payments on a policy with a deductible of 250 have a zero-modified negative binomial distribution with $r^* = 2$, $\beta^* = 1.536$, and $p_0^{M*} = 0.4595$. Losses have a Pareto distribution with $\alpha = 3$ and $\theta = 1,000$. Determine the distribution of the number of payments when the deductible is removed. Repeat this calculation assuming that $p_0^{M*} = 0.002$.

In the first case, the formulas use $v = 1/0.512 = 1.953125$, and so $r = 2$ and $\beta = 1.953125(1.536) = 3$. Also,

$$p_0^{M*} = \frac{0.4595 - 2.536^{-2} + 4^{-2} - 0.4595(4)^{-2}}{1 - 2.536^{-2}} = 0.4$$

as expected. For the second case,

$$p_0^{M*} = \frac{0.002 - 2.536^{-2} + 4^{-2} - 0.002(4)^{-2}}{1 - 2.536^{-2}} = -0.1079,$$

which is not a legitimate probability. ☐

All members of the $(a, b, 0)$ and $(a, b, 1)$ classes meet the conditions of this section. Table 8.3 indicates how the parameters change when moving from N^L to N^P. If N^L has a compound distribution, then we can write $P_{N^L}(z) = P_1[P_2(z)]$ and therefore

$$P_{N^P}(z) = P_{N^L}[1 + v(z - 1)] = P_1\{P_2[1 + v(z - 1)]\}.$$

Thus N^P will also have a compound distribution with the secondary distribution modified as indicated. If the secondary distribution has an $(a, b, 0)$ distribution, then it can modified as in Table 8.3. The following example indicates the adjustment to be made if the secondary distribution has an $(a, b, 1)$ distribution.

Table 8.3   Frequency adjustments.

N^L	Parameters for N^P
Poisson	$\lambda^* = v\lambda$
ZM Poisson	$p_0^{M*} = \dfrac{p_0^M - e^{-\lambda} + e^{-v\lambda} - p_0^M e^{-v\lambda}}{1 - e^{-\lambda}}, \ \ \lambda^* = v\lambda$
Binomial	$q^* = vq$
ZM binomial	$p_0^{M*} = \dfrac{p_0^M - (1-q)^m + (1-vq)^m - p_0^M(1-vq)^m}{1-(1-q)^m}$
	$q^* = vq$
Negative binomial	$\beta^* = v\beta, \ \ r^* = r$
ZM negative binomial	$p_0^{M*} = \dfrac{p_0^M - (1+\beta)^{-r} + (1+v\beta)^{-r} - p_0^M(1+v\beta)^{-r}}{1-(1+\beta)^{-r}}$
	$\beta^* = v\beta, \ \ r^* = r$
ZM logarithmic	$p_0^{M*} = 1 - (1 - p_0^M)\ln(1+v\beta)/\ln(1+\beta)$
	$\beta^* = v\beta$

■   **EXAMPLE 8.11**

Suppose that N^L is Poisson–ETNB with $\lambda = 5$, $\beta = 0.3$, and $r = 4$. If $v = 0.5$, determine the distribution of N^P.

From the preceding discussion, N^P is compound Poisson with $\lambda^* = 5$, but the secondary distribution is a zero-modified negative binomial with (from Table 8.3) $\beta^* = 0.5(0.3) = 0.15$,

$$p_0^{M*} = \frac{0 - 1.3^{-4} + 1.15^{-4} - 0(1.15)^{-4}}{1 - 1.3^{-4}} = 0.34103,$$

and $r^* = 4$. This would be sufficient, except that we have acquired the habit of using the ETNB as the secondary distribution. From Theorem 7.4, a compound Poisson distribution with a zero-modified secondary distribution is equivalent to a compound Poisson distribution with a zero-truncated secondary distribution. The Poisson parameter must be changed to $(1 - p_0^{M*})\lambda^*$. Therefore, N^P has a Poisson–ETNB distribution with $\lambda^* = (1 - 0.34103)5 = 3.29485$, $\beta^* = 0.15$, and $r^* = 4$.   □

The results can be further generalized to an increase or decrease in the deductible. Let N^d be the frequency when the deductible is d and let N^{d^*} be the frequency when the deductible is d^*. Let $v = [1 - F_X(d^*)]/[1 - F_X(d)]$, and then Table 8.3 can be used to move from the parameters of N^d to the parameters of N^{d^*}. As long as $d^* > d$, we will have $v < 1$ and the formulas will lead to a legitimate distribution for N^{d^*}. This includes the special case of $d = 0$ that was used at the start of this section. If $d^* < d$, then $v > 1$ and there is no assurance that a legitimate distribution will result. This includes the special case $d^* = 0$ (removal of a deductible) covered earlier.

It should be noted that policy limits have no effect on the frequency distribution. Imposing, removing, or changing a limit will not change the number of payments made.

Finally, it is important to restate that all the results in this section depend on an assumption that may not hold. The assumption is that changing the deductible (or other

coverage modifications) does not change the number of loss-producing incidents or the amount of losses.

8.6.1   Exercises

8.29 A group life insurance policy has an accidental death rider. For ordinary deaths, the benefit is 10,000; however, for accidental deaths, the benefit is 20,000. The insureds are approximately the same age, so it is reasonable to assume that they all have the same claim probabilities. Let them be 0.97 for no claim, 0.01 for an ordinary death claim, and 0.02 for an accidental death claim. A reinsurer has been asked to bid on providing an excess reinsurance that will pay 10,000 for each accidental death.

(a) The claim process can be modeled with a frequency component that has the Bernoulli distribution (the event is claim/no claim) and a two-point severity component (the probabilities are associated with the two claim levels, given that a claim occurred). Specify the probability distributions for the frequency and severity random variables.

(b) Suppose that the reinsurer wants to retain the same frequency distribution. Determine the modified severity distribution that will reflect the reinsurer's payments.

(c) Determine the reinsurer's frequency and severity distributions when the severity distribution is to be conditional on a reinsurance payment being made.

8.30 Individual losses have a Pareto distribution with $\alpha = 2$ and $\theta = 1,000$. With a deductible of 500, the frequency distribution for the number of payments is Poisson–inverse Gaussian with $\lambda = 3$ and $\beta = 2$. If the deductible is raised to 1,000, determine the distribution for the number of payments. Also, determine the pdf of the severity distribution (per payment) when the new deductible is in place.

8.31 Losses have a Pareto distribution with $\alpha = 2$ and $\theta = 1,000$. The frequency distribution for a deductible of 500 is zero-truncated logarithmic with $\beta = 4$. Determine a model for the number of payments when the deductible is reduced to 0.

8.32 Suppose that the number of losses N^L has the Sibuya distribution (see Exercise 6.7) with pgf $P_{N^L}(z) = 1 - (1 - z)^{-r}$, where $-1 < r < 0$. Demonstrate that the number of payments has a zero-modified Sibuya distribution.

8.33 (*) The frequency distribution for the number of losses when there is no deductible is negative binomial with $r = 3$ and $\beta = 5$. Loss amounts have a Weibull distribution with $\tau = 0.3$ and $\theta = 1,000$. Determine the expected number of payments when a deductible of 200 is applied.

8.34 Consider the situation in Exercise 8.28. Suppose that with the deductible in place, the number of payments has a Poisson distribution with $\lambda = 4$. Determine the expected number of payments if the deductible is removed.

8.35 Suppose that the number of losses has pgf

$$P(z) = \alpha + (1 - \alpha)\frac{\phi[\beta(1 - z)] - \phi(\beta)}{1 - \phi(\beta)}$$

where $\phi(x)$ is a function and $\beta > 0$. Also, suppose that the probability that a loss leads to a payment is v.

(a) Show that the number of payments has pgf

$$P_1(z) = q_0^* + (1 - q_0^*)\frac{\phi[\beta^*(1 - z)] - \phi(\beta^*)}{1 - \phi(\beta^*)},$$

where

$$\beta^* = v\beta \text{ and } q_0^* = (1 - \alpha)\frac{\phi(v\beta) - \phi(\beta)}{1 - \phi(\beta)}.$$

(b) Assume that $\alpha = 0$ and $P(z)$ is the pgf of a zero-truncated distribution. Using (a), explain why $P_1(z)$ is the pgf of a zero-modified distribution. Then, further assume that $P(z)$ is the pgf of a zero-truncated member of the $(a, b, 0)$ class. Explain why $P_1(z)$ is the pgf of a member of the $(a, b, 1)$ class.

(c) Show that if $\alpha = 0$ and $\phi(x) = 1 + \ln(1 + x)$, then $P(z)$ is the logarithmic series pgf.

9

AGGREGATE LOSS MODELS

9.1 Introduction

An insurance enterprise exists because of its ability to pool risks. By insuring many people, the individual risks are combined into an aggregate risk that is manageable and can be priced at a level that will attract customers. Consider the following simple example.

■ **EXAMPLE 9.1**

An insurable event has a 10% probability of occurring and when it occurs results in a loss of 5,000. Market research has indicated that consumers will pay at most 550 to purchase insurance against this event. How many policies must a company sell in order to have a 95% chance of making money (ignoring expenses)?

Let n be the number of policies sold. A reasonable model for the number of claims, C, is a binomial distribution with $m = n$ and $q = 0.1$, and the total paid will be

$5,000C$. To achieve the desired outcome,

$$0.95 \leq \Pr(5,000C < 550n)$$
$$= \Pr(C < 0.11n)$$
$$\doteq \Pr\left[Z < \frac{0.11n - 0.1n}{\sqrt{0.1(0.9)n}}\right],$$

where the approximation uses the central limit theorem. With the normal distribution

$$\frac{0.11n - 0.1n}{\sqrt{0.1(0.9)n}} = 1.645,$$

which gives the answer $n = 2{,}435.42$, and so at least 2,436 policies must be sold. □

The goal of this chapter is to build a model for the total payments by an insurance system (which may be the entire company, a line of business, those covered by a group insurance contract, or even a single policy). The building blocks are random variables that describe the number of claims and the amounts of those claims, subjects covered in the previous chapters.

There are two ways to build a model for the amount paid on all claims occurring in a fixed time period on a defined set of insurance contracts. The first is to record the payments as they are made and then add them up. In that case, we can represent the aggregate losses as a sum, S, of a random number, N, of individual payment amounts $(X_1, X_2, \ldots, X_N)$. Hence,

$$S = X_1 + X_2 + \cdots + X_N, \quad N = 0, 1, 2, \ldots , \tag{9.1}$$

where $S = 0$ when $N = 0$.

Definition 9.1 *The **collective risk model** has the representation in (9.1) with the X_js being independent and identically distributed (i.i.d.) random variables, unless otherwise specified. More formally, the independence assumptions are:*

1. *Conditional on $N = n$, the random variables $X_1, X_2, \ldots, X_n$ are i.i.d. random variables.*

2. *Conditional on $N = n$, the common distribution of the random variables $X_1, X_2, \ldots, X_n$ does not depend on n.*

3. *The distribution of N does not depend in any way on the values of $X_1, X_2, \ldots$.*

The second model, the one used in Example 9.1, assigns a random variable to each contract.

Definition 9.2 *The **individual risk model** represents the aggregate loss as a sum, $S = X_1 + \cdots + X_n$, of a fixed number, n, of insurance contracts. The loss amounts for the n contracts are $(X_1, X_2, \ldots, X_n)$, where the X_js are assumed to be independent but are not assumed to be identically distributed. The distribution of the X_js usually has a probability mass at zero, corresponding to the probability of no loss or payment on that contract.*

The individual risk model is used to add together the losses or payments from a fixed number of insurance contracts or sets of insurance contracts. It is used in modeling the

losses of a group life or health insurance policy that covers a group of n employees. Each employee can have different coverage (life insurance benefit as a multiple of salary) and different levels of loss probabilities (different ages and health statuses).

In the special case where the X_js are identically distributed, the individual risk model becomes a special case of the collective risk model, with the distribution of N being the degenerate distribution with all of the probability at $N = n$, that is, $\Pr(N = n) = 1$.

The distribution of S in (9.1) is obtained from the distribution of N and the common distribution of the X_js. Using this approach, the frequency and the severity of claims are modeled separately. The information about these distributions is used to obtain information about S. An alternative to this approach is to simply gather information about S (e.g. total losses each month for a period of months) and to use some model from the earlier chapters to model the distribution of S. Modeling the distribution of N and the distribution of the X_js separately has seven distinct advantages:

1. The expected number of claims changes as the number of insured policies changes. Growth in the volume of business needs to be accounted for in forecasting the number of claims in future years based on past years' data.

2. The effects of general economic inflation and additional claims inflation are reflected in the losses incurred by insured parties and the claims paid by insurance companies. Such effects are often masked when insurance policies have deductibles and policy limits that do not depend on inflation and aggregate results are used.

3. The impact of changing individual deductibles and policy limits is easily implemented by changing the specification of the severity distribution.

4. The impact on claims frequencies of changing deductibles is better understood.

5. Data that are heterogeneous in terms of deductibles and limits can be combined to obtain the hypothetical loss size distribution. This approach is useful when data from several years in which policy provisions were changing are combined.

6. Models developed for noncovered losses to insureds, claim costs to insurers, and claim costs to reinsurers can be mutually consistent. This feature is useful for a direct insurer when studying the consequence of shifting losses to a reinsurer.

7. The shape of the distribution of S depends on the shapes of both distributions of N and X. The understanding of the relative shapes is useful when modifying policy details. For example, if the severity distribution has a much heavier tail than the frequency distribution, the shape of the tail of the distribution of aggregate claims or losses will be determined by the severity distribution and will be relatively insensitive to the choice of frequency distribution.

In summary, a more accurate and flexible model can be constructed by examining frequency and severity separately.

In constructing the model (9.1) for S, if N represents the actual number of losses to the insured, then the X_js can represent (i) the losses to the insured, (ii) the claim payments of the insurer, (iii) the claim payments of a reinsurer, or (iv) the deductibles (self-insurance) paid by the insured. In each case, the interpretation of S is different and the severity distribution can be constructed in a consistent manner.

Because the random variables N, $X_1, X_2, \ldots$, and S provide much of the focus for this chapter, we want to be especially careful when referring to them. To that end, we refer to N as the **claim count random variable** and refer to its distribution as the **claim count distribution**. The expression **number of claims** is also used and, occasionally, just **claims**. Another term commonly used is **frequency distribution**. The X_js are the **individual** or **single-loss random variables**. The modifier *individual* or *single* is dropped when the reference is clear. In Chapter 8, a distinction is made between losses and payments. Strictly speaking, the X_js are payments because they represent a real cash transaction. However, the term **loss** is more customary, and we continue with it. Another common term for the X_js is **severity**. Finally, S is the **aggregate loss random variable** or the **total loss random variable**.

■ **EXAMPLE 9.2**

Describe how a collective risk model could be used for the total payments made in one year on an automobile physical damage policy with a deductible of 250.

There are two ways to do this. First, let N^L be the number of accidents, including those that do not exceed the deductible. The individual loss variables are the Y^L variables from Chapter 8. The other way is to let N^P count the number of payments. In this case, the individual loss variable is Y^P.   □

9.1.1 Exercises

9.1 Show how the model in Example 9.1 could be written as a collective risk model.

9.2 For each of the following situations, determine which model (individual or collective) is more likely to provide a better description of aggregate losses.

(a) A group life insurance contract where each employee has a different age, gender, and death benefit.

(b) A reinsurance contract that pays when the annual total medical malpractice costs at a certain hospital exceeds a given amount.

(c) A dental policy on an individual pays for at most two checkups per year per family member. A single contract covers any size family at the same price.

9.2 Model Choices

In many cases of fitting frequency or severity distributions to data, several distributions may be good candidates for models. However, some distributions may be preferable for a variety of practical reasons.

In general, it is useful for the severity distribution to be a scale distribution (see Definition 4.2) because the choice of currency (e.g. US dollars or British pounds) should not affect the result. Also, scale families are easy to adjust for inflationary effects over time (this is, in effect, a change in currency; e.g. from 1994 US dollars to 1995 US dollars). When forecasting the costs for a future year, the anticipated rate of inflation can be factored in easily by adjusting the parameters.

A similar consideration applies to frequency distributions. As an insurance company's portfolio of contracts grows, the number of claims can be expected to grow, all other things being equal. Models that have probability generating functions of the form

$$P_N(z; \alpha) = Q(z)^\alpha \qquad (9.2)$$

for some parameter α have the expected number of claims proportional to α. Increasing the volume of business by $100r\%$ results in expected claims being proportional to $\alpha^* = (1 + r)\alpha$. This approach is discussed in Section 7.4. Because r is any value satisfying $r > -1$, the distributions satisfying (9.2) should allow α to take on any positive value. Such distributions can be shown to be infinitely divisible (see Definition 7.6).

A related consideration, the concept of invariance over the time period of the study, also supports using frequency distributions that are infinitely divisible. Ideally, the model selected should not depend on the length of the time period used in the study of claims frequency. In particular, the expected frequency should be proportional to the length of the time period after any adjustment for growth in business. In this case, a study conducted over a period of 10 years can be used to develop claims frequency distributions for periods of one month, one year, or any other period. Furthermore, the form of the distribution for a one-year period is the same as for a one-month period with a change of parameter. The parameter α corresponds to the length of a time period. For example, if $\alpha = 1.7$ in (9.2) for a one-month period, then the identical model with $\alpha = 20.4$ is an appropriate model for a one-year period.

Distributions that have a modification at zero are not of the form (9.2). However, it may still be desirable to use a zero-modified distribution if the physical situation suggests it. For example, if a certain proportion of policies never make a claim, due to duplication of coverage or other reason, it may be appropriate to use this same proportion in future periods for a policy selected at random.

9.2.1 Exercises

9.3 For pgfs satisfying (9.2), show that the mean is proportional to α.

9.4 Which of the distributions in Appendix B satisfy (9.2) for any positive value of α?

9.3   The Compound Model for Aggregate Claims

Let S denote aggregate losses associated with a set of N observed claims $X_1, X_2, \ldots, X_N$ satisfying the independence assumptions following (9.1). The approach in this chapter involves the following three steps:

1. Develop a model for the distribution of N based on data.

2. Develop a model for the common distribution of the X_js based on data.

3. Using these two models, carry out necessary calculations to obtain the distribution of S.

Completion of the first two steps follows the ideas developed elsewhere in this text. We now presume that these two models are developed and that we only need to carry out numerical work in obtaining solutions to problems associated with the distribution of S.

9.3.1   Probabilities and Moments

The random sum

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

(where N has a counting distribution) has distribution function

$$
\begin{aligned}
F_S(x) &= \Pr(S \leq x) \\
&= \sum_{n=0}^{\infty} p_n \Pr(S \leq x | N = n) \\
&= \sum_{n=0}^{\infty} p_n F_X^{*n}(x),
\end{aligned}
\tag{9.3}
$$

where $F_X(x) = \Pr(X \leq x)$ is the common distribution function of the X_js and $p_n = \Pr(N = n)$. The distribution of S is called a **compound distribution** In (9.3), $F_X^{*n}(x)$ is the "n-fold convolution" of the cdf of X. It can be obtained as

$$
F_X^{*0}(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0, \end{cases}
$$

and

$$
F_X^{*k}(x) = \int_{-\infty}^{\infty} F_X^{*(k-1)}(x - y)\, dF_X(y), \quad \text{for } k = 1, 2, \ldots.
\tag{9.4}
$$

The tail may then be written, for all $x \geq 0$, as

$$
1 - F_S(x) = \sum_{n=1}^{\infty} p_n [1 - F_X^{*n}(x)].
\tag{9.5}
$$

If X is a continuous random variable with probability zero on nonpositive values, (9.4) reduces to

$$
F_X^{*k}(x) = \int_0^x F_X^{*(k-1)}(x - y) f_X(y)\, dy, \quad \text{for } k = 2, 3, \ldots.
$$

For $k = 1$, this equation reduces to $F_X^{*1}(x) = F_X(x)$. By differentiating, the pdf is

$$
f_X^{*k}(x) = \int_0^x f_X^{*(k-1)}(x - y) f_X(y)\, dy \text{ for } k = 2, 3, \ldots.
$$

Therefore, if X is continuous, then S has a pdf, which, for $x > 0$, is given by

$$
f_S(x) = \sum_{n=1}^{\infty} p_n f_X^{*n}(x),
\tag{9.6}
$$

and a discrete mass point, $\Pr(S = 0) = p_0$ at $x = 0$. Note that $\Pr(S = 0) \neq f_S(0) = \lim_{x \to 0_+} f_S(x)$.

If X has a discrete counting distribution, with probabilities at $0, 1, 2, \ldots$, (9.4) reduces to

$$
F_X^{*k}(x) = \sum_{y=0}^{x} F_X^{*(k-1)}(x - y) f_X(y), \quad \text{for } x = 0, 1, \ldots,\ k = 2, 3, \ldots.
$$

The corresponding pf is

$$f_X^{*k}(x) = \sum_{y=0}^{x} f_X^{*(k-1)}(x-y) f_X(y), \quad \text{for } x = 0, 1, \dots, \ k = 2, 3, \dots.$$

For notational purposes, let $f_X^{*0}(0) = 1$ and $f_X^{*0}(x) = 0$ for $x \neq 0$. Then, in this case, S has a discrete distribution with pf

$$f_S(x) = \Pr(S = x) = \sum_{n=0}^{\infty} p_n f_X^{*n}(x), \quad x = 0, 1, \dots. \tag{9.7}$$

Arguing as in Section 7.1, the pgf of S is

$$\begin{aligned}
P_S(z) &= E[z^S] \\
&= E[z^0] \Pr(N = 0) + \sum_{n=1}^{\infty} E[z^{X_1 + X_2 + \cdots + X_n} | N = n] \Pr(N = n) \\
&= \Pr(N = 0) + \sum_{n=1}^{\infty} E\left[\prod_{j=1}^{n} z^{X_j} \right] \Pr(N = n) \\
&= \sum_{n=0}^{\infty} \Pr(N = n)[P_X(z)]^n \\
&= E[P_X(z)^N] = P_N[P_X(z)] \tag{9.8}
\end{aligned}$$

due to the independence of $X_1, \dots, X_n$ for fixed n. The pgf is typically used when S is discrete. With regard to the moment generating function, we have

$$M_S(z) = P_N[M_X(z)].$$

The pgf of compound distributions is discussed in Section 7.1 where the "secondary" distribution plays the role of the claim size distribution in this chapter. In that section, the claim size distribution is always discrete.

In the case where $P_N(z) = P_1[P_2(z)]$ (i.e. N is itself a compound distribution), $P_S(z) = P_1\{P_2[P_X(z)]\}$, which in itself produces no additional difficulties.

From (9.8), the moments of S can be obtained in terms of the moments of N and the X_js. The first three moments are

$$\begin{aligned}
E(S) &= \mu'_{S1} = \mu'_{N1} \mu'_{X1} = E(N)E(X), \\
\text{Var}(S) &= \mu_{S2} = \mu'_{N1} \mu_{X2} + \mu_{N2}(\mu'_{X1})^2 = E(N)\text{Var}(X) + \text{Var}(N)[E(X)]^2, \\
E\{[S - E(S)]^3\} &= \mu_{S3} = \mu'_{N1} \mu_{X3} + 3\mu_{N2} \mu'_{X1} \mu_{X2} + \mu_{N3}(\mu'_{X1})^3. \tag{9.9}
\end{aligned}$$

Here, the first subscript indicates the appropriate random variable, the second subscript indicates the order of the moment, and the superscript is a prime ($'$) for raw moments (moments about the origin) and is unprimed for central moments (moments about the mean). The moments can be used on their own to provide approximations for probabilities of aggregate claims by matching the first few model and sample moments.

■ **EXAMPLE 9.3**

The observed mean (and standard deviation) of the number of claims and the individual losses over the past 10 months are 6.7 (2.3) and 179,247 (52,141), respectively. Determine the mean and standard deviation of aggregate claims per month.

$$E(S) = 6.7(179{,}247) = 1{,}200{,}955,$$
$$\text{Var}(S) = 6.7(52{,}141)^2 + (2.3)^2(179{,}247)^2$$
$$= 1.88180 \times 10^{11}.$$

Hence, the mean and standard deviation of aggregate claims are 1,200,955 and 433,797, respectively. □

■ **EXAMPLE 9.4**

(*Example 9.3 continued*) Using normal and lognormal distributions as approximating distributions for aggregate claims, calculate the probability that claims will exceed 140% of expected costs. That is,

$$\Pr(S > 1.40 \times 1{,}200{,}955) = \Pr(S > 1{,}681{,}337).$$

For the normal distribution

$$\Pr(S > 1{,}681{,}337) = \Pr\left(\frac{S - E(S)}{\sqrt{\text{Var}(S)}} > \frac{1{,}681{,}337 - 1{,}200{,}955}{433{,}797}\right)$$
$$= \Pr(Z > 1.107) = 1 - \Phi(1.107) = 0.134.$$

For the lognormal distribution, from Appendix A, the mean and second raw moment of the lognormal distribution are

$$E(S) = \exp(\mu + \tfrac{1}{2}\sigma^2) \quad \text{and} \quad E(S^2) = \exp(2\mu + 2\sigma^2).$$

Equating these to 1.200955×10^6 and $1.88180 \times 10^{11} + (1.200955 \times 10^6)^2 = 1.63047 \times 10^{12}$ and taking logarithms results in the following two equations in two unknowns:

$$\mu + \tfrac{1}{2}\sigma^2 = 13.99863, \quad 2\mu + 2\sigma^2 = 28.11989.$$

From this, $\mu = 13.93731$ and $\sigma^2 = 0.1226361$. Then,

$$\Pr(S > 1{,}681{,}337) = 1 - \Phi\left[\frac{\ln 1{,}681{,}337 - 13.93731}{(0.1226361)^{0.5}}\right]$$
$$= 1 - \Phi(1.135913) = 0.128.$$

The normal distribution provides a good approximation when $E(N)$ is large. In particular, if N has the Poisson, binomial, or negative binomial distribution, a version of the central limit theorem indicates that, as λ, m, or r, respectively, goes to infinity, the distribution of S becomes normal. In this example, $E(N)$ is small, so the distribution of S is likely to be skewed. In this case, the lognormal distribution may provide a good approximation, although there is no theory to support this choice. □

Table 9.1   The loss distribution for Example 9.5.

x	$f_X(x)$
1	0.150
2	0.200
3	0.250
4	0.125
5	0.075
6	0.050
7	0.050
8	0.050
9	0.025
10	0.025

Table 9.2   The frequency distribution for Example 9.5.

n	p_n
0	0.05
1	0.10
2	0.15
3	0.20
4	0.25
5	0.15
6	0.06
7	0.03
8	0.01

■ **EXAMPLE 9.5**

(*Group dental insurance*) Under a group dental insurance plan covering employees and their families, the premium for each married employee is the same regardless of the number of family members. The insurance company has compiled statistics showing that the annual cost of dental care per person for the benefits provided by the plan has the distribution in Table 9.1 (given in units of 25).

Furthermore, the distribution of the number of persons per insurance certificate (i.e. per employee) receiving dental care in any year has the distribution given in Table 9.2.

The insurer is now in a position to calculate the distribution of the cost per year per married employee in the group. The cost per married employee is

$$f_S(x) = \sum_{n=0}^{8} p_n f_X^{*n}(x).$$

Determine the pf of S up to 525. Determine the mean and standard deviation of total payments per employee.

The distribution up to amounts of 525 is given in Table 9.3. For example, $f_X^{*3}(4) = f_X^{*1}(1)f_X^{*2}(3) + f_X^{*1}(2)f_X^{*2}(2)$. In general, pick two columns whose superscripts sum

Table 9.3 The aggregate probabilities for Example 9.5.

x	f_X^{*0}	f_X^{*1}	f_X^{*2}	f_X^{*3}	f_X^{*4}	f_X^{*5}	f_X^{*6}	f_X^{*7}	f_X^{*8}	$f_S(x)$
0	1	0	0	0	0	0	0	0	0	0.05000
1	0	0.150	0	0	0	0	0	0	0	0.01500
2	0	0.200	0.02250	0	0	0	0	0	0	0.02338
3	0	0.250	0.06000	0.00338	0	0	0	0	0	0.03468
4	0	0.125	0.11500	0.01350	0.00051	0	0	0	0	0.03258
5	0	0.075	0.13750	0.03488	0.00270	0.00008	0	0	0	0.03579
6	0	0.050	0.13500	0.06144	0.00878	0.00051	0.00001	0	0	0.03981
7	0	0.050	0.10750	0.08569	0.01999	0.00198	0.00009	0.00000	0	0.04356
8	0	0.050	0.08813	0.09750	0.03580	0.00549	0.00042	0.00002	0.00000	0.04752
9	0	0.025	0.07875	0.09841	0.05266	0.01194	0.00136	0.00008	0.00000	0.04903
10	0	0.025	0.07063	0.09338	0.06682	0.02138	0.00345	0.00031	0.00002	0.05190
11	0	0	0.06250	0.08813	0.07597	0.03282	0.00726	0.00091	0.00007	0.05138
12	0	0	0.04500	0.08370	0.08068	0.04450	0.01305	0.00218	0.00022	0.05119
13	0	0	0.03125	0.07673	0.08266	0.05486	0.02062	0.00448	0.00060	0.05030
14	0	0	0.01750	0.06689	0.08278	0.06314	0.02930	0.00808	0.00138	0.04818
15	0	0	0.01125	0.05377	0.08081	0.06934	0.03826	0.01304	0.00279	0.04576
16	0	0	0.00750	0.04125	0.07584	0.07361	0.04677	0.01919	0.00505	0.04281
17	0	0	0.00500	0.03052	0.06811	0.07578	0.05438	0.02616	0.00829	0.03938
18	0	0	0.00313	0.02267	0.05854	0.07552	0.06080	0.03352	0.01254	0.03575
19	0	0	0.00125	0.01673	0.04878	0.07263	0.06573	0.04083	0.01768	0.03197
20	0	0	0.00063	0.01186	0.03977	0.06747	0.06882	0.04775	0.02351	0.02832
21	0	0	0	0.00800	0.03187	0.06079	0.06982	0.05389	0.02977	0.02479
p_n	0.05	0.10	0.15	0.20	0.25	0.15	0.06	0.03	0.01	

to the superscript of the desired column (in this case, $1 + 2 = 3$). Then, add all combinations from these columns where the arguments sum to the desired argument (in this case, $1 + 3 = 4$ and $2 + 2 = 4$). To obtain $f_S(x)$, each row of the matrix of convolutions of $f_X(x)$ is multiplied by the probabilities from the row below the table and the products are summed. For example, $f_S(2) = 0.05(0) + 0.10(0.2) + 0.15(0.225) = 0.02338$.

You may wish to verify, using (9.9), that the first two moments of the distribution $f_S(x)$ are

$$E(S) = 12.58, \quad \text{Var}(S) = 58.7464.$$

Hence the annual cost of the dental plan has mean $12.58 \times 25 = 314.50$ and standard deviation 191.6155. (Why can't the calculations be done from Table 9.3?)   □

9.3.2 Stop-Loss Insurance

It is common for insurance to be offered in which a deductible is applied to the aggregate losses for the period. When the losses occur to a policyholder, it is called **insurance coverage**, and when the losses occur to an insurance company, it is called **reinsurance coverage**. The latter version is a common method for an insurance company to protect itself against an adverse year (as opposed to protecting against a single, very large claim). More formally, we present the following definition.

Definition 9.3 *Insurance on the aggregate losses, subject to a deductible, is called **stop-loss insurance**. The expected cost of this insurance is called the **net stop-loss premium** and can be computed as $E[(S - d)_+]$, where d is the deductible and the notation $(\cdot)_+$ means to use the value in parentheses if it is positive and to use zero otherwise.*

For any aggregate distribution,

$$E[(S - d)_+] = \int_d^\infty [1 - F_S(x)]\, dx.$$

If the distribution is continuous for $x > d$, the net stop-loss premium can be computed directly from the definition as

$$E[(S - d)_+] = \int_d^\infty (x - d) f_S(x)\, dx.$$

Similarly, for discrete random variables,

$$E[(S - d)_+] = \sum_{x>d} (x - d) f_S(x).$$

Any time there is an interval with no aggregate probability, the following result may simplify calculations.

Theorem 9.4 *Suppose that $\Pr(a < S < b) = 0$. Then, for $a \leq d \leq b$,*

$$E[(S - d)_+] = \frac{b - d}{b - a} E[(S - a)_+] + \frac{d - a}{b - a} E[(S - b)_+].$$

That is, the net stop-loss premium can be calculated via linear interpolation.

Proof: From the assumption, $F_S(x) = F_S(a)$, $a \leq x < b$. Then,

$$
\begin{aligned}
E[(S-d)_+] &= \int_d^\infty [1 - F_S(x)]\, dx \\
&= \int_a^\infty [1 - F_S(x)]\, dx - \int_a^d [1 - F_S(x)]\, dx \\
&= E[(S-a)_+] - \int_a^d [1 - F_S(a)]\, dx \\
&= E[(S-a)_+] - (d-a)[1 - F_S(a)]. \qquad (9.10)
\end{aligned}
$$

Then, by setting $d = b$ in (9.10),

$$
E[(S-b)_+] = E[(S-a)_+] - (b-a)[1 - F_S(a)]
$$

and, therefore,

$$
1 - F_S(a) = \frac{E[(S-a)_+] - E[(S-b)_+]}{b-a}.
$$

The substitution of this formula in (9.10) produces the desired result. □

Further simplification is available in the discrete case provided that S places probability at equally spaced values.

Theorem 9.5 *Assume that* $\Pr(S = kh) = f_k \geq 0$ *for some fixed* $h > 0$ *and* $k = 0, 1, \ldots$ *and* $\Pr(S = x) = 0$ *for all other* x. *Then, provided that* $d = jh$, *with* j *a nonnegative integer,*

$$
E[(S-d)_+] = h \sum_{m=0}^\infty \{1 - F_S[(m+j)h]\}.
$$

Proof:

$$
\begin{aligned}
E[(S-d)_+] &= \sum_{x>d} (x-d) f_S(x) \\
&= \sum_{k=j}^\infty (kh - jh) f_k \\
&= h \sum_{k=j}^\infty \sum_{m=0}^{k-j-1} f_k \\
&= h \sum_{m=0}^\infty \sum_{k=m+j+1}^\infty f_k \\
&= h \sum_{m=0}^\infty \{1 - F_S[(m+j)h]\}.
\end{aligned}
$$

□

In the discrete case with probability at equally spaced values, a simple recursion holds.

Corollary 9.6 *Under the conditions of Theorem* 9.5,

$$E\{[S - (j + 1)h]_+\} = E[(S - jh)_+] - h[1 - F_S(jh)].$$

This result is easy to use because, when $d = 0$, $E[(S - 0)_+] = E(S) = E(N)E(X)$, which can be obtained directly from the frequency and severity distributions.

■ **EXAMPLE 9.6**

(*Example 9.5 continued*) The insurer is examining the effect of imposing an aggregate deductible per employee. Determine the reduction in the net premium as a result of imposing deductibles of 25, 30, 50, and 100 dollars.

From Table 9.3, the cdf at 0, 25, 50, and 75 dollars has values 0.05, 0.065, 0.08838, and 0.12306. With $E(S) = 25(12.58) = 314.5$, we have

$$E[(S - 25)_+] = 314.5 - 25(1 - 0.05) = 290.75,$$
$$E[(S - 50)_+] = 290.75 - 25(1 - 0.065) = 267.375,$$
$$E[(S - 75)_+] = 267.375 - 25(1 - 0.08838) = 244.5845,$$
$$E[(S - 100)_+] = 244.5845 - 25(1 - 0.12306) = 222.661.$$

From Theorem 9.4, $E[(S - 30)_+] = \frac{20}{25}290.75 + \frac{5}{25}267.375 = 286.07$. When compared to the original premium of 314.5, the reductions are 23.75, 28.43, 47.125, and 91.839 for the four deductibles. □

9.3.3 The Tweedie Distribution

The **Tweedie distribution** [123] brings together two concepts. First, for certain parameter values it is a compound Poisson distribution with a gamma severity distribution. Hence it may be a useful model for aggregate claims. Second, it is a member of the linear exponential family as discussed in Section 5.4. As such, it can be a useful distributional model when constructing generalized linear models to relate claims to policyholder characteristics.

We begin by looking at this compound Poisson distribution. Let N have a Poisson distribution with mean λ and let X have a gamma distribution with parameters α and γ (which is used in place of θ as that letter is used in the definition of the linear exponential family). We then have, for the compound distribution $S = X_1 + X_2 + \cdots + X_N$, that

$$\Pr(S = 0) = \Pr(N = 0) = e^{-\lambda},$$
$$f_S(s) = \sum_{n=1}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \frac{s^{n\alpha-1} e^{-s/\gamma}}{\Gamma(n\alpha)\gamma^{n\alpha}},$$
$$E(S) = E(N)E(X) = \lambda\alpha\gamma,$$
$$\text{Var}(S) = E(N)\text{Var}(X) + \text{Var}(N)[E(X)]^2 = \lambda\alpha\gamma^2 + \lambda(\alpha\gamma)^2 = \lambda\gamma^2\alpha(\alpha + 1).$$

The second equation arises from the fact that the n-fold convolution of a gamma distribution is also gamma, with the shape parameter (α) multiplied by n.

The Tweedie distribution is often reparameterized through the relations

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \text{and} \quad \gamma = \phi(p-1)\mu^{p-1}$$

where $1 < p < 2$, $\mu > 0$, and $\phi > 0$. Substitution of the above three formulas into the moment equations yields

$$\mathrm{E}(S) = \mu \quad \text{and} \quad \mathrm{Var}(S) = \phi\mu^p. \tag{9.11}$$

This provides a convenient relationship between the variance and the mean that can help in deciding if this is an appropriate model for a given problem.

As mentioned, the Tweedie distribution is a member of the linear exponential family. Definition 5.9 states that members of this family have the form

$$f(x;\theta) = \frac{p(x)e^{r(\theta)x}}{q(\theta)}$$

where $p(x)$ may include parameters other than θ. For this discussion, we write the general form of the linear exponential distribution as

$$f(x;\theta) = p(x)\exp\left[\frac{\theta y - a(\theta)}{\phi}\right].$$

The two differences are that $r(\theta)$ is replaced by θ and the parameter ϕ is introduced. The replacement is just a reparameterization. The additional parameter is called the **dispersion parameter**. As can be seen in (9.11), this parameter allow for additional flexibility with respect to how the variance relates to the mean. For a demonstration that the Tweedie distribution is a member of the linear exponential family, see Clark and Thayer [24].

The Tweedie distribution exists for other values of p. All nonnegative values other than $0 < p < 1$ are possible. Some familiar distributions that are special cases are normal ($p = 0$), Poisson ($p = 1$), gamma ($p = 2$), and inverse Gaussian ($p = 3$). Note that for $p = 1$ we have, from (9.11), that $\mathrm{Var}(X) = \phi\mu$. Hence, to obtain the Poisson distribution as a special case, we must have $\phi = 1$. When ϕ takes on values larger than one, the distribution is called the **overdispersed Poisson** distribution. It will not be discussed further in this text, but is often used when constructing generalized linear models for count data.

9.3.4 Exercises

9.5 From (9.8), show that the relationships between the moments in (9.9) hold.

9.6 (*) When an individual is admitted to the hospital, the hospital charges have the following characteristics:

1.

Charges	Mean	Standard deviation
Room	1,000	500
Other	500	300

2. The covariance between an individual's room charges and other charges is 100,000.

An insurer issues a policy that reimburses 100% for room charges and 80% for other charges. The number of hospital admissions has a Poisson distribution with parameter 4. Determine the mean and standard deviation of the insurer's payout for the policy.

Table 9.4   The data for Exercise 9.8.

Class	Proportion of population	λ
1	0.25	5
2	0.25	3
3	0.50	2

Table 9.5   The distributions for Exercise 9.9.

x	$f_1(x)$	$f_2(x)$	$f_3(x)$
0	0.90	0.50	0.25
1	0.10	0.30	0.25
2	0.00	0.20	0.25
3	0.00	0.00	0.25

Table 9.6   The distributions for Exercise 9.10.

x	$f_1(x)$	$f_2(x)$	$f_3(x)$
0	p	0.6	0.25
1	$1 - p$	0.2	0.25
2	0	0.1	0.25
3	0	0.1	0.25

9.7 Aggregate claims have been modeled by a compound negative binomial distribution with parameters $r = 15$ and $\beta = 5$. The claim amounts are uniformly distributed on the interval $(0, 10)$. Using the normal approximation, determine the premium such that the probability that claims will exceed premium is 0.05.

9.8 Automobile drivers can be divided into three homogeneous classes. The number of claims for each driver follows a Poisson distribution with parameter λ. Determine the variance of the number of claims for a randomly selected driver, using the data in Table 9.4.

9.9 (*) Assume that X_1, X_2, and X_3 are mutually independent loss random variables with probability functions as given in Table 9.5. Determine the pf of $S = X_1 + X_2 + X_3$.

9.10 (*) Assume that X_1, X_2, and X_3 are mutually independent random variables with probability functions as given in Table 9.6. If $S = X_1 + X_2 + X_3$ and $f_S(5) = 0.06$, determine p.

9.11 (*) Consider the following information about AIDS patients:

1. The conditional distribution of an individual's medical care costs given that the individual does not have AIDS has mean 1,000 and variance 250,000.

2. The conditional distribution of an individual's medical care costs given that the individual does have AIDS has mean 70,000 and variance 1,600,000.

Table 9.7   The data for Exercise 9.12.

	Male	Female
Mean	6	3
Variance	64	31

3. The number of individuals with AIDS in a group of m randomly selected adults has a binomial distribution with parameters m and $q = 0.01$.

An insurance company determines premiums for a group as the mean plus 10% of the standard deviation of the group's aggregate claims distribution. The premium for a group of 10 independent lives for which all individuals have been proven *not* to have AIDS is P. The premium for a group of 10 randomly selected adults is Q. Determine P/Q.

9.12 (*) You have been asked by a city planner to analyze office cigarette smoking patterns. The planner has provided the information in Table 9.7 about the distribution of the number of cigarettes smoked during a workday.

The number of male employees in a randomly selected office of N employees has a binomial distribution with parameters N and 0.4. Determine the mean plus the standard deviation of the number of cigarettes smoked during a workday in a randomly selected office of eight employees.

9.13 (*) For a certain group, aggregate claims are uniformly distributed over (0, 10). Insurer A proposes stop-loss coverage with a deductible of 6 for a premium equal to the expected stop-loss claims. Insurer B proposes group coverage with a premium of 7 and a dividend (a premium refund) equal to the excess, if any, of $7k$ over claims. Calculate k such that the expected cost to the group is equal under both proposals.

9.14 (*) For a group health contract, aggregate claims are assumed to have an exponential distribution, where the mean of the distribution is estimated by the group underwriter. Aggregate stop-loss insurance for total claims in excess of 125% of the expected claims is provided by a gross premium that is twice the expected stop-loss claims. You have discovered an error in the underwriter's method of calculating expected claims. The underwriter's estimate is 90% of the correct estimate. Determine the actual percentage loading in the premium.

9.15 (*) A random loss, X, has the probability function given in Table 9.8. You are given that $E(X) = 4$ and $E[(X - d)_+] = 2$. Determine d.

9.16 (*) A reinsurer pays aggregate claim amounts in excess of d, and in return it receives a stop-loss premium $E[(S - d)_+]$. You are given $E[(S - 100)_+] = 15$, $E[(S - 120)_+] = 10$, and the probability that the aggregate claim amounts are greater than 80 and less than or equal to 120 is zero. Determine the probability that the aggregate claim amounts are less than or equal to 80.

Table 9.8   The data for Exercise 9.15.

x	f(x)
0	0.05
1	0.06
2	0.25
3	0.22
4	0.10
5	0.05
6	0.05
7	0.05
8	0.05
9	0.12

9.17 (*) A loss random variable X has pdf $f(x) = \frac{1}{100}, 0 < x < 100$. Two policies can be purchased to alleviate the financial impact of the loss:

$$A = \begin{cases} 0, & x < 50k, \\ \dfrac{x}{k} - 50, & x \geq 50k, \end{cases}$$

and

$$B = kx, \quad 0 < x < 100,$$

where A and B are the amounts paid when the loss is x. Both policies have the same net premium, that is, $E(A) = E(B)$. Determine k.

9.18 (*) For a nursing home insurance policy, you are given that the average length of stay is 440 days and 30% of the stays are terminated in the first 30 days. These terminations are distributed uniformly during that period. The policy pays 20 per day for the first 30 days and 100 per day thereafter. Determine the expected benefits payable for a single stay.

9.19 (*) An insurance portfolio produces N claims, where

n	$\Pr(N = n)$
0	0.5
1	0.4
3	0.1

Individual claim amounts have the following distribution:

x	$f_X(x)$
1	0.9
10	0.1

Individual claim amounts and N are mutually independent. Calculate the probability that the ratio of aggregate claims to expected claims will exceed 3.0.

Benefit amount	Number covered	Probability of claim
Table 9.9	The data for Exercise 9.21.	
Benefit amount	Number covered	Probability of claim
1	8,000	0.025
2	3,500	0.025
4	4,500	0.025

9.20 (*) A company sells group travel-accident life insurance, with m payable in the event of a covered individual's death in a travel accident. The gross premium for a group is set equal to the expected value plus the standard deviation of the group's aggregate claims. The standard premium is based on the following two assumptions:

1. All individual claims within the group are mutually independent.

2. $m^2q(1-q) = 2,500$, where q is the probability of death by travel accident for an individual.

In a certain group of 100 lives, the independence assumption fails because three specific individuals always travel together. If one dies in an accident, all three are assumed to die. Determine the difference between this group's premium and the standard premium.

9.21 (*) A life insurance company covers 16,000 lives for one-year term life insurance, as in Table 9.9.

All claims are mutually independent. The insurance company's retention limit is two units per life. Reinsurance is purchased for 0.03 per unit. The probability that the insurance company's retained claims, S, plus the cost of reinsurance will exceed 1,000 is

$$\Pr\left[\frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}} > K\right].$$

Determine K using a normal approximation.

9.22 (*) The probability density function of individual losses Y is

$$f(y) = \begin{cases} 0.02\left(1 - \dfrac{y}{100}\right), & 0 < y < 100, \\ 0, & \text{elsewhere.} \end{cases}$$

The amount paid, Z, is 80% of that portion of the loss that exceeds a deductible of 10. Determine $\mathrm{E}(Z)$.

9.23 (*) An individual loss distribution is normal with $\mu = 100$ and $\sigma^2 = 9$. The distribution for the number of claims, N, is given in Table 9.10. Determine the probability that aggregate claims exceed 100.

Table 9.10   The distribution for Exercise 9.23.

n	$\Pr(N = n)$
0	0.5
1	0.2
2	0.2
3	0.1

Table 9.11   The distribution for Exercise 9.24.

n	$f(n)$
0	1/16
1	1/4
2	3/8
3	1/4
4	1/16

9.24 (*) An employer self-insures a life insurance program with the following two characteristics:

1. Given that a claim has occurred, the claim amount is 2,000 with probability 0.4 and 3,000 with probability 0.6.

2. The number of claims has the distribution given in Table 9.11.

The employer purchases aggregate stop-loss coverage that limits the employer's annual claims cost to 5,000. The aggregate stop-loss coverage costs 1,472. Determine the employer's expected annual cost of the program, including the cost of stop-loss coverage.

9.25 (*) The probability that an individual admitted to the hospital will stay k days or less is $1 - 0.8^k$ for $k = 0, 1, 2, \ldots$. A hospital indemnity policy provides a fixed amount per day for the 4th day through the 10th day (i.e. for a maximum of 7 days). Determine the percentage increase in the expected cost per admission if the maximum number of days paid is increased from 7 to 14.

9.26 (*) The probability density function of aggregate claims, S, is given by $f_S(x) = 3x^{-4}$, $x \geq 1$. The relative loading θ and the value λ are selected so that

$$\Pr[S \leq (1 + \theta)\mathrm{E}(S)] = \Pr\left[S \leq \mathrm{E}(S) + \lambda\sqrt{\mathrm{Var}(S)}\right] = 0.90.$$

Calculate λ and θ.

9.27 (*) An insurance policy reimburses aggregate incurred expenses at the rate of 80% of the first 1,000 in excess of 100, 90% of the next 1,000, and 100% thereafter. Express the expected cost of this coverage in terms of $R_d = \mathrm{E}[(S - d)_+]$ for different values of d.

9.28 (*) The number of accidents incurred by an insured driver in a single year has a Poisson distribution with parameter $\lambda = 2$. If an accident occurs, the probability that the damage amount exceeds the deductible is 0.25. The number of claims and the damage amounts are independent. What is the probability that there will be no damages exceeding the deductible in a single year?

9.29 (*) The aggregate loss distribution is modeled by an insurance company using an exponential distribution. However, the mean is uncertain. The company uses a uniform distribution $(2,000,000, \ 4,000,000)$ to express its view of what the mean should be. Determine the expected aggregate losses.

9.30 (*) A group hospital indemnity policy provides benefits at a continuous rate of 100 per day of hospital confinement for a maximum of 30 days. Benefits for partial days of confinement are prorated. The length of hospital confinement in days, T, has the following continuance (survival) function for $0 \leq t \leq 30$:

$$\Pr(T \geq t) = \begin{cases} 1 - 0.04t, & 0 \leq t \leq 10, \\ 0.95 - 0.035t, & 10 < t \leq 20, \\ 0.65 - 0.02t, & 20 < t \leq 30. \end{cases}$$

For a policy period, each member's probability of a single hospital admission is 0.1 and that of more than one admission is zero. Determine the pure premium per member, ignoring the time value of money.

9.31 (*) Medical and dental claims are assumed to be independent with compound Poisson distributions as follows:

Claim type	Claim amount distribution	λ
Medical claims	Uniform (0, 1,000)	2
Dental claims	Uniform (0, 200)	3

Let X be the amount of a given claim under a policy that covers both medical and dental claims. Determine $E[(X - 100)_+]$, the expected cost (in excess of 100) of any given claim.

9.32 (*) For a certain insured, the distribution of aggregate claims is binomial with parameters $m = 12$ and $q = 0.25$. The insurer will pay a dividend, D, equal to the excess of 80% of the premium over claims, if positive. The premium is 5. Determine $E[D]$.

9.33 (*) The number of claims in one year has a geometric distribution with mean 1.5. Each claim has a loss of 100. An insurance policy pays zero for the first three claims in one year and then pays 100 for each subsequent claim. Determine the expected insurance payment per year.

9.34 (*) A compound Poisson distribution has $\lambda = 5$ and claim amount distribution $p(100) = 0.80$, $p(500) = 0.16$, and $p(1,000) = 0.04$. Determine the probability that aggregate claims will be exactly 600.

9.35 (*) Aggregate payments have a compound distribution. The frequency distribution is negative binomial with $r = 16$ and $\beta = 6$, and the severity distribution is uniform on the interval $(0, 8)$. Use the normal approximation to determine the premium such that the probability is 5% that aggregate payments will exceed the premium.

9.36 (*) The number of losses is Poisson with $\lambda = 3$. Loss amounts have a Burr distribution with $\alpha = 3$, $\theta = 2$, and $\gamma = 1$. Determine the variance of aggregate losses.

9.4 Analytic Results

For most choices of distributions of N and the X_js, the compound distributional values can only be obtained numerically. Subsequent sections in this chapter are devoted to such numerical procedures.

However, for certain combinations of choices, simple analytic results are available, thus reducing the computational problems considerably.

■ **EXAMPLE 9.7**

(*Compound negative binomial–exponential*) Determine the distribution of S when the frequency distribution is negative binomial with an integer value for the parameter r and the severity distribution is exponential.

The mgf of S is

$$
\begin{aligned}
M_S(z) &= P_N[M_X(z)] \\
&= P_N[(1 - \theta z)^{-1}] \\
&= \{1 - \beta[(1 - \theta z)^{-1} - 1]\}^{-r}.
\end{aligned}
$$

With a bit of algebra, this can be rewritten as

$$
M_S(z) = \left(1 + \frac{\beta}{1 + \beta}\{[1 - \theta(1 + \beta)z]^{-1} - 1\}\right)^r,
$$

which is of the form

$$
M_S(z) = P_N^*[M_X^*(z)],
$$

where

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

the pgf of the binomial distribution with parameters r and $\beta/(1 + \beta)$, and $M_X^*(z)$ is the mgf of the exponential distribution with mean $\theta(1 + \beta)$.

This transformation reduces the computation of the distribution function to the finite sum, that is,

$$
F_S(x) = 1 - \sum_{n=1}^{r} \binom{r}{n} \left(\frac{\beta}{1 + \beta}\right)^n \left(\frac{1}{1 + \beta}\right)^{r-n}
$$

$$
\times \sum_{j=0}^{n-1} \frac{[x\theta^{-1}(1 + \beta)^{-1}]^j e^{-x\theta^{-1}(1+\beta)^{-1}}}{j!}.
$$

When $r = 1$, S has a compound geometric distribution, and in this case the preceding formula reduces to

$$F_S(x) = 1 - \frac{\beta}{1 + \beta} \exp\left[-\frac{x}{\theta(1 + \beta)}\right], \quad x \geq 0.$$

Hence, $\Pr(S = 0) = F_S(0) = (1 + \beta)^{-1}$, and because $F_S(x)$ is differentiable, it has pdf $f_S(x) = F'_S(x)$, for $x > 0$. That is, for $x > 0$, S has pdf

$$f_S(x) = \frac{\beta}{\theta(1 + \beta)^2} \exp\left[-\frac{x}{\theta(1 + \beta)}\right].$$

To summarize, if $r = 1$, S has a point mass of $(1 + \beta)^{-1}$ at zero and an exponentially decaying density over the positive axis. $\qquad\qquad\square$

As is clear from Example 9.7, useful formulas may result with exponential claim sizes. The following example considers this case in more detail.

■ **EXAMPLE 9.8**

(*Exponential severities*) Determine the cdf of S for any compound distribution with exponential severities.

The mgf of the sum of n independent exponential random variables each with mean θ is

$$M_{X_1 + X_2 + \cdots + X_n}(z) = (1 - \theta z)^{-n},$$

which is the mgf of the gamma distribution with cdf

$$F_X^{*n}(x) = \Gamma\left(n; \frac{x}{\theta}\right)$$

(see Appendix A). For integer values of α, the values of $\Gamma(\alpha; x)$ can be calculated exactly (see Appendix A for the derivation) as

$$\Gamma(n; x) = 1 - \sum_{j=0}^{n-1} \frac{x^j e^{-x}}{j!}, \quad n = 1, 2, 3, \ldots. \tag{9.12}$$

From (9.3)

$$F_S(x) = p_0 + \sum_{n=1}^{\infty} p_n \Gamma\left(n; \frac{x}{\theta}\right). \tag{9.13}$$

The density function can be obtained by differentiation,

$$f_S(x) = \sum_{n=1}^{\infty} p_n \frac{x^{n-1} e^{-x/\theta}}{\theta^n \Gamma(n)}. \tag{9.14}$$

Returning to the distribution function, the substitution of (9.12) in (9.13) yields

$$F_S(x) = 1 - \sum_{n=1}^{\infty} p_n \sum_{j=0}^{n-1} \frac{(x/\theta)^j e^{-x/\theta}}{j!}, \quad x \geq 0. \tag{9.15}$$

Interchanging the order of summation yields

$$F_S(x) = 1 - e^{-x/\theta} \sum_{j=0}^{\infty} \frac{(x/\theta)^j}{j!} \sum_{n=j+1}^{\infty} p_n$$

$$= 1 - e^{-x/\theta} \sum_{j=0}^{\infty} \bar{P}_j \frac{(x/\theta)^j}{j!}, \quad x \geq 0, \tag{9.16}$$

where $\bar{P}_j = \sum_{n=j+1}^{\infty} p_n$ for $j = 0, 1, \ldots$. $\qquad\square$

For frequency distributions that assign positive probability to all nonnegative integers, (9.15) can be evaluated by taking sufficient terms in the first summation. For distributions for which $\Pr(N > n^*) = 0$, the first summation becomes finite. For example, for the binomial frequency distribution, (9.15) becomes

$$F_S(x) = 1 - \sum_{n=1}^{m} \binom{m}{n} q^n (1-q)^{m-n} \sum_{j=0}^{n-1} \frac{(x/\theta)^j e^{-x/\theta}}{j!}. \tag{9.17}$$

The following theorem provides a shortcut when adding independent compound Poisson random variables. This may arise, for example, in a group insurance contract in which each member has a compound Poisson model for aggregate losses and we are interested in computing the distribution of total aggregate losses. Similarly, we may want to evaluate the combined losses from several independent lines of business. The theorem implies that it is not necessary to compute the distribution for each line or group member and then determine the distribution of the sum. Instead, a weighted average of the loss severity distributions of each component may be used as a severity distribution and the Poisson parameters added to obtain the frequency distribution. Then, a single aggregate loss distribution calculation is sufficient.

Theorem 9.7 *Suppose that S_j has a compound Poisson distribution with Poisson parameter λ_j and a severity distribution with cdf $F_j(x)$ for $j = 1, 2, \ldots, n$. Suppose also that $S_1, S_2, \ldots, S_n$ are independent. Then, $S = S_1 + \cdots + S_n$ has a compound Poisson distribution with Poisson parameter $\lambda = \lambda_1 + \cdots + \lambda_n$ and a severity distribution with cdf*

$$F(x) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda} F_j(x).$$

Proof: Let $M_j(t)$ be the mgf of $F_j(x)$ for $j = 1, 2, \ldots, n$. Then, S_j has mgf

$$M_{S_j}(t) = \mathrm{E}(e^{tS_j}) = \exp\{\lambda_j[M_j(t) - 1]\},$$

and, by the independence of the S_js, S has mgf

$$M_S(t) = \prod_{j=1}^{n} M_{S_j}(t) = \prod_{j=1}^{n} \exp\{\lambda_j[M_j(t) - 1]\}$$

$$= \exp\left\{ \left[\sum_{j=1}^{n} \lambda_j M_j(t) \right] - \lambda \right\}$$

$$= \exp\left(\lambda \left\{ \left[\sum_{j=1}^{n} \frac{\lambda_j}{\lambda} M_j(t) \right] - 1 \right\} \right).$$

Because $\sum_{j=1}^{n} \frac{\lambda_j}{\lambda} M_j(t)$ is the mgf of $F(x) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda} F_j(x)$, $M_S(t)$ is a compound Poisson mgf and the result follows. $\qquad\square$

■ **EXAMPLE 9.9**

Policy A has a compound Poisson distribution with $\lambda = 2$ and severity probabilities 0.6 on a payment of 1 and 0.4 on a payment of 2. Policy B has a compound Poisson distribution with $\lambda = 1$ and probabilities 0.7 on a payment of 1 and 0.3 on a payment of 3. Determine the probability that the total payment on the two policies will be 2.

This problem is simple enough that Theorem 9.7 is not needed (the probability can be directly obtained by enumerating the ways in which a total of 2 can be achieved). Also, a faster method of computing compound Poisson probabilities is developed in Section 9.6.

Using Theorem 9.7, the total payment has a compound Poisson distribution with $\lambda = 2 + 1 = 3$ and a severity distribution

$$f_X(1) = (2/3)(0.6) + (1/3)(0.7) = 0.63333,$$
$$f_X(2) = (2/3)(0.4) + (1/3)(0.0) = 0.26667,$$
$$f_X(3) = (2/3)(0.0) + (1/3)(0.3) = 0.1.$$

Then,

$$\begin{aligned} \Pr(S = 2) &= \Pr(N = 1, X_1 = 2) + \Pr(N = 2, X_1 = 1, X_2 = 1) \\ &= e^{-3}(3/1)(0.26667) + e^{-3}(3^2/2)(0.63333)(0.63333) \\ &= 2.605e^{-3} = 0.12970. \end{aligned}$$

Treating each policy separately proceeds as follows. For policy A,

$$\Pr(S_A = 0) = e^{-2},$$
$$\Pr(S_A = 1) = e^{-2}(2/1)(0.6) = 1.2e^{-2},$$
$$\Pr(S_A = 2) = e^{-2}(2/1)(0.4) + e^{-2}(2^2/2)(0.6)^2 = 1.52e^{-2}.$$

For policy B,

$$\Pr(S_B = 0) = e^{-1},$$
$$\Pr(S_B = 1) = e^{-1}(0.7),$$
$$\Pr(S_B = 2) = e^{-1}(1/2)(0.7)^2 = 0.245e^{-1}.$$

Then, $\Pr(S_A + S_B = 2) = e^{-2}(0.245e^{-1}) + 1.2e^{-2}(0.7e^{-1}) + 1.52e^{-2}(e^{-1}) = 2.605e^{-3} = 0.12970.$ $\qquad\square$

9.4.1 Exercises

9.37 A compound negative binomial distribution has parameters $\beta = 1$, $r = 2$, and severity distribution $f_X(x)$, $x = 0, 1, 2, \dots$. How do the parameters of the distribution change if the severity distribution is $g_X(x) = f_X(x)/[1 - f_X(0)]$, $x = 1, 2, \dots$ but the aggregate claims distribution remains unchanged?

9.38 Consider the compound logarithmic distribution with exponential severity distribution.

(a) Show that the density of aggregate losses may be expressed as

$$f_S(x) = \frac{1}{\ln(1+\beta)} \sum_{n=1}^{\infty} \frac{1}{n!} \left[\frac{\beta}{\theta(1+\beta)} \right]^n x^{n-1} e^{-x/\theta}.$$

(b) Reduce this to

$$f_S(x) = \frac{\exp\{-x/[\theta(1+\beta)]\} - \exp(-x/\theta)}{x \ln(1+\beta)}.$$

9.39 For a compound distribution, N has a binomial distribution with parameters $m = 3$ and $q = 0.4$ and X has an exponential distribution with a mean of 100. Calculate $\Pr(S \leq 300)$.

9.40 A company sells three policies. For policy A, all claim payments are 10,000 and a single policy has a Poisson number of claims with mean 0.01. For policy B, all claim payments are 20,000 and a single policy has a Poisson number of claims with mean 0.02. For policy C, all claim payments are 40,000 and a single policy has a Poisson number of claims with mean 0.03. All policies are independent. For the coming year, there are 5,000, 3,000, and 1,000 of policies A, B, and C, respectively. Calculate the expected total payment, the standard deviation of total payment, and the probability that total payments will exceed 30,000.

9.5   Computing the Aggregate Claims Distribution

The computation of the compound distribution function

$$F_S(x) = \sum_{n=0}^{\infty} p_n F_X^{*n}(x) \tag{9.18}$$

or the corresponding probability (density) function is generally not an easy task, even in the simplest of cases. In this section, we discuss several approaches to numerical evaluation of (9.18) for specific choices of the frequency and severity distributions as well as for arbitrary choices of one or both distributions.

One approach is to use an **approximating distribution** to avoid direct calculation of (9.18). This approach is used in Example 9.4, where the method of moments is used to estimate the parameters of the approximating distribution. The advantage of this method is that it is simple and easy to apply. However, the disadvantages are significant. First, there is no way of knowing how good the approximation is. Choosing different approximating distributions can result in very different results, particularly in the right-hand tail of the distribution. Of course, the approximation should improve as more moments are used, but after four moments, we quickly run out of distributions!

The approximating distribution may also fail to accommodate special features of the true distribution. For example, when the loss distribution is of the continuous type and there is a maximum possible claim (e.g. when there is a policy limit), the severity distribution may have a point mass ("atom" or "spike") at the maximum. The true aggregate

claims distribution is of the mixed type, with spikes at integral multiples of the maximum corresponding to $1, 2, 3, \ldots$ claims at the maximum. These spikes, if large, can have a significant effect on the probabilities near such multiples. These jumps in the aggregate claims distribution function cannot be replicated by a smooth approximating distribution.

The second method to evaluate (9.18) or the corresponding pdf is **direct calculation**. The most difficult (or computer-intensive) part is the evaluation of the n-fold convolutions of the severity distribution for $n = 2, 3, 4, \ldots$. The convolutions need to be evaluated numerically using

$$F_X^{*k}(x) = \int_{-\infty}^{\infty} F_X^{*(k-1)}(x - y) \, dF_X(y). \tag{9.19}$$

When the losses are limited to nonnegative values (as is usually the case), the range of integration becomes finite, reducing (9.19) to

$$F_X^{*k}(x) = \int_{0-}^{x} F_X^{*(k-1)}(x - y) \, dF_X(y). \tag{9.20}$$

These integrals are written in Lebesgue–Stieltjes form because of possible jumps in the cdf $F_X(x)$ at zero and at other points.[1] Evaluation of (9.20) usually requires numerical integration methods. Because of the first term inside the integral, (9.20) needs to be evaluated for all possible values of x. This approach quickly becomes technically overpowering.

As seen in Example 9.5, when the severity distribution is discrete, the calculations reduce to numerous multiplications and additions. For continuous severities, a simple way to avoid these technical problems is to replace the severity distribution by a discrete distribution defined at multiples $0, 1, 2 \ldots$ of some convenient monetary unit such as 1,000.

In practice, the monetary unit can be made sufficiently small to accommodate spikes at maximum insurance amounts. The spike must be a multiple of the monetary unit to have it located at exactly the right point. As the monetary unit of measurement becomes small, the discrete distribution function needs to approach the true distribution function. The simplest approach is to round all amounts to the nearest multiple of the monetary unit; for example, round all losses or claims to the nearest 1,000. More sophisticated methods are discussed later in this chapter.

When the severity distribution is defined on nonnegative integers $0, 1, 2, \ldots$, calculating $f_X^{*k}(x)$ for integral x requires $x + 1$ multiplications. Then, carrying out these calculations for all possible values of k and x up to n requires a number of multiplications that are of order n^3, written as $O(n^3)$, to obtain the distribution of (9.18) for $x = 0$ to $x = n$. When the maximum value, n, for which the aggregate claims distribution is calculated is large, the number of computations quickly becomes prohibitive, even for fast computers. For example, in real applications n can easily be as large as 1,000 and requires about 10^9 multiplications. Further, if $\Pr(X = 0) > 0$ and the frequency distribution is unbounded, an infinite number of calculations is required to obtain any single probability. This is because $F_X^{*n}(x) > 0$ for all n and all x, and so the sum in (9.18) contains an infinite number of terms. When $\Pr(X = 0) = 0$, we have $F_X^{*n}(x) = 0$ for $n > x$ and so (9.18) will have no more than $x + 1$ positive terms. Table 9.3 provides an example of this latter case.

An alternative method to more quickly evaluate the aggregate claims distribution is discussed in Section 9.6. This method, the *recursive method*, reduces the number of

---

[1] Without going into the formal definition of the Lebesgue–Stieltjes integral, it suffices to interpret $\int g(y) \, dF_X(y)$ as to be evaluated by integrating $g(y) f_X(y)$ over those y values for which X has a continuous distribution and then adding $g(y_i) \Pr(X = y_i)$ over those points where $\Pr(X = y_i) > 0$. This formulation allows for a single notation to be used for continuous, discrete, and mixed random variables.

computations discussed previously to $O(n^2)$, which is a considerable savings in computer time, a reduction of about 99.9% when $n = 1,000$ compared to direct calculation. However, the method is limited to certain frequency distributions. Fortunately, it includes all of the frequency distributions discussed in Chapter 6 and Appendix B.

9.6   The Recursive Method

Suppose that the severity distribution $f_X(x)$ is defined on $x = 0, 1, 2, \ldots, m$ representing multiples of some convenient monetary unit. The number m represents the largest possible payment and could be infinite. Further, suppose that the frequency distribution, p_k, is a member of the $(a, b, 1)$ class and therefore satisfies

$$p_k = \left(a + \frac{b}{k} \right) p_{k-1}, \quad k = 2, 3, 4, \ldots .$$

Then, the following result holds.

Theorem 9.8 *For the $(a, b, 1)$ class,*

$$f_S(x) = \frac{[p_1 - (a + b)p_0]f_X(x) + \sum_{y=1}^{x \wedge m}(a + by/x)f_X(y)f_S(x - y)}{1 - af_X(0)}, \tag{9.21}$$

noting that $x \wedge m$ is notation for $\min(x, m)$.

Proof: This result is identical to Theorem 7.2 with appropriate substitution of notation and recognition that the argument of $f_X(y)$ cannot exceed m. □

Corollary 9.9 *For the $(a, b, 0)$ class, the result* (9.21) *reduces to*

$$f_S(x) = \frac{\sum_{y=1}^{x \wedge m} (a + by/x)\, f_X(y)f_S(x - y)}{1 - af_X(0)}. \tag{9.22}$$

Note that when the severity distribution has no probability at zero, the denominator of (9.21) and (9.22) equals 1. Further, in the case of the Poisson distribution, (9.22) reduces to

$$f_S(x) = \frac{\lambda}{x} \sum_{y=1}^{x \wedge m} y f_X(y)f_S(x - y), \quad x = 1, 2, \ldots . \tag{9.23}$$

The starting value of the recursive schemes (9.21) and (9.22) is $f_S(0) = P_N[f_X(0)]$ following Theorem 7.3 with an appropriate change of notation. In the case of the Poisson distribution, we have

$$f_S(0) = e^{-\lambda[1 - f_X(0)]}.$$

Starting values for other frequency distributions are found in Appendix D.

■ **EXAMPLE 9.10**

(*Example 9.9 continued*) Calculate the probabilities at 0, 1, 2, 3, and 4 using the recursive formula.

To use the formula, $f_X(0) = 0$, $f_X(1) = 0.63333$, $f_X(2) = 0.26667$, and $f_X(3) = 0.1$. Then, $f_S(0) = e^{-3(1-0)} = e^{-3}$. The recursive formula is then (with $\lambda = 3$)

$$
f_S(x) = \frac{3}{x} \sum_{y=1}^{3} y f_X(y) f_S(x - y)
$$

$$
= \frac{3}{x} [1(0.63333) f_S(x - 1) + 2(0.26667) f_S(x - 2) + 3(0.1) f_S(x - 3)]
$$

noting that when the argument of $f_S(x)$ is negative, the value is zero. Then,

$$
f_S(1) = 3(0.63333)e^{-3} = 1.9e^{-3},
$$
$$
f_S(2) = (3/2)[0.63333(1.9e^{-3}) + 0.53333e^{-3} = 2.605e^{-3},
$$
$$
f_S(3) = (3/3)[0.63333(2.605e^{-3}) + 0.53333(1.9e^{-3}) + 0.3e^{-3} = 2.96315e^{-3},
$$
$$
f_S(4) = (3/4)[0.63333(2.96315e^{-3}) + 0.53333(2.605e^{-3}) + 0.3(1.9e^{-3})
$$
$$
= 3.83598e^{-3}. \qquad\qquad \square
$$

■ **EXAMPLE 9.11**

A compound distribution has a zero-modified binomial distribution with $m = 3$, $q = 0.3$, and $p_0^M = 0.4$. Individual payments are 0, 50, and 150, with probabilities 0.3, 0.5, and 0.2, respectively. Use the recursive formula to determine the probability distribution of S.

Because payments must be equally spaced, the appropriate unit to use is 50. Then, $f_X(0) = 0.3$, $f_X(1) = 0.5$, $f_X(2) = 0$, and $f_X(3) = 0.2$. The starting value is

$$
f_X(0) = 0.4 + 0.6 \frac{[1 + 0.3(-0.7)]^3 - 0.7^3}{1 - 0.7^3} = 0.53702.
$$

Key values are:

$$
a = -0.3/0.7 = -3/7,
$$
$$
b = (3 + 1)(0.3)/0.7 = 12/7, \text{ and}
$$
$$
p_1 = 3(0.3)(0.7)^2(0.6)/(1 - 0.7^3) = 0.40274.
$$

The recursive formula is

$$
f_S(s) = \frac{[p_1 - (a + b)p_0] f_X(x) + \sum_{y=1}^{3} \left(-\frac{3}{7} + \frac{12y}{7x}\right) f_X(y) f_S(x - y)}{1 - (-3/7)(0.3)}
$$

$$
= \frac{70}{79} \left[\left(0.40274 - \frac{3.6}{7}\right) f_X(x) + \left(-\frac{3}{7} + \frac{12}{7x}\right) 0.5 f_S(x - 1) \right.
$$

$$
\left. + \left(-\frac{3}{7} + \frac{36}{7x}\right) 0.2 f_S(x - 3) \right].
$$

The first few values are

$$f_S(1) = \frac{70}{79}\left[-0.11155(0.5) + \left(-\frac{3}{7} + \frac{12}{7}\right)(0.5)(0.53702)\right] = 0.25648,$$

$$f_S(2) = \frac{70}{79}\left(-\frac{3}{7} + \frac{12}{14}\right)(0.5)(0.25648) = 0.04870,$$

$$f_S(3) = \frac{70}{79}\left[-0.11155(0.2) + \left(-\frac{3}{7} + \frac{12}{21}\right)(0.5)(0.04870)\right.$$
$$\left. + \left(-\frac{3}{7} + \frac{36}{21}\right)(0.2)(0.53702)\right] = 0.10567,$$

$$f_S(4) = \frac{70}{79}\left[\left(-\frac{3}{7} + \frac{12}{28}\right)(0.5)(0.10567) + \left(-\frac{3}{7} + \frac{36}{28}\right)(0.2)(0.25648)\right]$$
$$= 0.03896.$$

The remaining values use a formula similar to that for $f_S(4)$.   □

9.6.1 Applications to Compound Frequency Models

When the frequency distribution can be represented as a compound distribution (e.g. Neyman Type A, Poisson–inverse Gaussian) involving only distributions from the $(a, b, 0)$ or $(a, b, 1)$ classes, the recursive formula (9.21) can be used two or more times to obtain the aggregate claims distribution. If the frequency distribution can be written as

$$P_N(z) = P_1[P_2(z)],$$

then the aggregate claims distribution has pgf

$$P_S(z) = P_N[P_X(z)]$$
$$= P_1\{P_2[P_X(z)]\},$$

which can be rewritten as

$$P_S(z) = P_1[P_{S_1}(z)], \tag{9.24}$$

where

$$P_{S_1}(z) = P_2[P_X(z)]. \tag{9.25}$$

Now (9.25) has the same form as an aggregate claims distribution. Thus, if $P_2(z)$ is in the $(a, b, 0)$ or $(a, b, 1)$ class, the distribution of S_1 can be calculated using (9.21). The resulting distribution is the "severity" distribution in (9.25). Thus, a second application of (9.21) to (9.24) results in the distribution of S.

The following example illustrates the use of this algorithm.

■ **EXAMPLE 9.12**

The number of claims has a Poisson–ETNB distribution with Poisson parameter $\lambda = 2$ and ETNB parameters $\beta = 3$ and $r = 0.2$. The claim size distribution has probabilities 0.3, 0.5, and 0.2 at 0, 10, and 20, respectively. Determine the total claims distribution recursively.

In the preceding terminology, N has pgf $P_N(z) = P_1[P_2(z)]$, where $P_1(z)$ and $P_2(z)$ are the Poisson and ETNB pgfs, respectively. Then, the total dollars of claims has pgf $P_S(z) = P_1[P_{S_1}(z)]$, where $P_{S_1}(z) = P_2[P_X(z)]$ is a compound ETNB pgf.

We will first compute the distribution of S_1. We have (in monetary units of 10) $f_X(0) = 0.3$, $f_X(1) = 0.5$, and $f_X(2) = 0.2$. To use the compound ETNB recursion, we start with

$$f_{S_1}(0) = P_2 \left[f_X(0) \right]$$
$$= \frac{\left\{ 1 + \beta \left[1 - f_X(0) \right] \right\}^{-r} - (1+\beta)^{-r}}{1 - (1+\beta)^{-r}}$$
$$= \frac{\{1 + 3(1 - 0.3)\}^{-0.2} - (1 + 3)^{-0.2}}{1 - (1 + 3)^{-0.2}}$$
$$= 0.16369.$$

The remaining values of $f_{S_1}(x)$ may be obtained from (9.21) with S replaced by S_1. In this case, we have

$$a = \frac{3}{1+3} = 0.75, \quad b = (0.2 - 1)a = -0.6,$$
$$p_0 = 0, \quad p_1 = \frac{0.2(3)}{(1+3)^{0.2+1} - (1+3)} = 0.46947.$$

Then, (9.21) becomes

$$f_{S_1}(x) = \frac{[0.46947 - (0.75 - 0.6)(0)] f_X(x) + \sum_{y=1}^{x} (0.75 - 0.6y/x) f_X(y) f_{S_1}(x - y)}{1 - (0.75)(0.3)}$$
$$= 0.60577 f_X(x) + 1.29032 \sum_{y=1}^{x} \left(0.75 - 0.6 \frac{y}{x} \right) f_X(y) f_{S_1}(x - y).$$

The first few probabilities are

$$f_{S_1}(1) = 0.60577(0.5) + 1.29032 \left[0.75 - 0.6 \left(\tfrac{1}{1} \right) \right] (0.5)(0.16369)$$
$$= 0.31873,$$
$$f_{S_1}(2) = 0.60577(0.2) + 1.29032 \left\{ \left[0.75 - 0.6 \left(\tfrac{1}{2} \right) \right] (0.5)(0.31873) \right.$$
$$\left. + \left[0.75 - 0.6 \left(\tfrac{2}{2} \right) \right] (0.2)(0.16369) \right\} = 0.22002,$$
$$f_{S_1}(3) = 1.29032 \left\{ \left[0.75 - 0.6 \left(\tfrac{1}{3} \right) \right] (0.5)(0.22002) \right.$$
$$\left. + \left[0.75 - 0.6 \left(\tfrac{2}{3} \right) \right] (0.2)(0.31873) \right\} = 0.10686,$$
$$f_{S_1}(4) = 1.29032 \left\{ \left[0.75 - 0.6 \left(\tfrac{1}{4} \right) \right] (0.5)(0.10686) \right.$$
$$\left. + \left[0.75 - 0.6 \left(\tfrac{2}{4} \right) \right] (0.2)(0.22002) \right\} = 0.06692.$$

We now turn to evaluation of the distribution of S with compound Poisson pgf

$$P_S(z) = P_1 \left[P_{S_1}(z) \right] = e^{\lambda \left[P_{S_1}(z) - 1 \right]}.$$

Thus the distribution

$$f_{S_1}(x), \quad x = 0, 1, 2, \dots$$

becomes the "secondary" or "claim size" distribution in an application of the compound Poisson recursive formula. Therefore,

$$f_S(0) = P_S(0) = e^{\lambda\left[P_{S_1}(0)-1\right]} = e^{\lambda\left[f_{S_1}(0)-1\right]} = e^{2(0.16369-1)} = 0.18775.$$

The remaining probabilities may be found from the recursive formula

$$f_S(x) = \frac{2}{x}\sum_{y=1}^{x} y f_{S_1}(y) f_S(x-y), \quad x = 1, 2, \dots .$$

The first few probabilities are

$$f_S(1) = 2\left(\tfrac{1}{1}\right)(0.31873)(0.18775) = 0.11968,$$

$$f_S(2) = 2\left(\tfrac{1}{2}\right)(0.31873)(0.11968) + 2\left(\tfrac{2}{2}\right)(0.22002)(0.18775) = 0.12076,$$

$$f_S(3) = 2\left(\tfrac{1}{3}\right)(0.31873)(0.12076) + 2\left(\tfrac{2}{3}\right)(0.22002)(0.11968)$$

$$+ 2\left(\tfrac{3}{3}\right)(0.10686)(0.18775) = 0.10090,$$

$$f_S(4) = 2\left(\tfrac{1}{4}\right)(0.31873)(0.10090) + 2\left(\tfrac{2}{4}\right)(0.22002)(0.12076)$$

$$+ 2\left(\tfrac{3}{4}\right)(0.10686)(0.11968) + 2\left(\tfrac{4}{4}\right)(0.06692)(0.18775)$$

$$= 0.08696. \qquad\qquad \square$$

When the severity distribution has a maximum possible value at m, the computations are speeded up even more because the sum in (9.21) will be restricted to at most m nonzero terms. In this case, then, the computations can be considered to be of order $O(x)$.

9.6.2 Underflow/Overflow Problems

The recursion (9.21) starts with the calculated value of $P(S = 0) = P_N[f_X(0)]$. For large insurance portfolios, this probability is very small, sometimes smaller than the smallest number that can be represented on the computer. When this happens, it is stored in the computer as zero and the recursion (9.21) fails. This problem can be overcome in several different ways (see Panjer and Willmot [99]). One of the easiest ways is to start with an arbitrary set of values for $f_S(0), f_S(1), \dots, f_S(k)$ such as $(0, 0, 0, \dots, 0, 1)$, where k is sufficiently far to the left in the distribution so that the true value of $F_S(k)$ is still negligible. Setting k to a point that lies six standard deviations to the left of the mean is usually sufficient. Recursion (9.21) is used to generate values of the distribution with this set of starting values until the values are consistently less than $f_S(k)$. The "probabilities" are then summed and divided by the sum so that the "true" probabilities add to 1. Trial and error will dictate how small k should be for a particular problem.

Another method to obtain probabilities when the starting value is too small is to carry out the calculations for a subset of the portfolio. For example, for the Poisson distribution

with mean λ, find a value of $\lambda^* = \lambda/2^n$ so that the probability at zero is representable on the computer when λ^* is used as the Poisson mean. Equation (9.21) is now used to obtain the aggregate claims distribution when λ^* is used as the Poisson mean. If $P_*(z)$ is the pgf of the aggregate claims using Poisson mean λ^*, then $P_S(z) = [P_*(z)]^{2^n}$. Hence we can obtain successively the distributions with pgfs $[P_*(z)]^2$, $[P_*(z)]^4$, $[P_*(z)]^8$, ..., $[P_*(z)]^{2^n}$ by convoluting the result at each stage with itself. This approach requires an additional n convolutions in carrying out the calculations but involves no approximations. It can be carried out for any frequency distributions that are closed under convolution. For the negative binomial distribution, the analogous procedure starts with $r^* = r/2^n$. For the binomial distribution, the parameter m must be integer valued. A slight modification can be used. Let $m^* = \lfloor m/2^n \rfloor$ when $\lfloor \cdot \rfloor$ indicates the **integer part of** the function. When the n convolutions are carried out, we still need to carry out the calculations using (9.21) for parameter $m - m^*2^n$. This result is then convoluted with the result of the n convolutions. For compound frequency distributions, only the primary distribution needs to be closed under convolution.

9.6.3 Numerical Stability

Any recursive formula requires accurate computation of values because each such value will be used in computing subsequent values. Recursive schemes suffer the risk of errors propagating through all subsequent values and potentially blowing up. In the recursive formula (9.21), errors are introduced through rounding at each stage because computers represent numbers with a finite number of significant digits. The question about stability is: *How fast do the errors in the calculations grow as the computed values are used in successive computations?* This work has been done by Panjer and Wang [98]. The analysis is quite complicated and well beyond the scope of this book. However, we can draw some general conclusions.

Errors are introduced in subsequent values through the summation

$$\sum_{y=1}^{x} \left(a + \frac{by}{x} \right) f_X(y) f_S(x - y)$$

in recursion (9.21). In the extreme right-hand tail of the distribution of S, this sum is positive (or at least nonnegative), and subsequent values of the sum will be decreasing. The sum will stay positive, even with rounding errors, when each of the three factors in each term in the sum is positive. In this case, the recursive formula is stable, producing relative errors that do not grow fast. For the Poisson and negative binomial-based distributions, the factors in each term are always positive.

However, for the binomial distribution, the sum can have negative terms because a is negative, b is positive, and y/x is a positive function not exceeding 1. In this case, the negative terms can cause the successive values to blow up with alternating signs. When this occurs, the nonsensical results are immediately obvious. Although it does not happen frequently in practice, you should be aware of this possibility in models based on the binomial distribution.

9.6.4 Continuous Severity

The recursive method as presented here requires a discrete severity distribution, while it is customary to use a continuous distribution for severity. In the case of continuous

severities, the analog of the recursion (9.21) is an integral equation, the solution of which is the aggregate claims distribution.

Theorem 9.10 *For the $(a, b, 1)$ class of frequency distributions and any continuous severity distribution with probability on the positive real line, the following integral equation holds:*

$$f_S(x) = p_1 f_X(x) + \int_0^x \left(a + \frac{by}{x} \right) f_X(y) f_S(x - y) \, dy. \tag{9.26}$$

The proof of this result is beyond the scope of this book. For a detailed proof, see Theorems 6.14.1 and 6.16.1 of Panjer and Willmot [100], along with the associated corollaries. They consider the more general (a, b, m) class of distributions, which allow for arbitrary modification of m initial values of the distribution. Note that the initial term is $p_1 f_X(x)$, not $[p_1 - (a + b)p_0] f_X(x)$ as in (9.21). Also, (9.26) holds for members of the $(a, b, 0)$ class as well.

Integral equations of the form (9.26) are Volterra integral equations of the second kind. Numerical solution of this type of integral equation has been studied in the text by Baker [10]. Instead, we consider an alternative approach for continuous severity distributions. It is to use a discrete approximation of the severity distribution in order to use the recursive method (9.21) and avoid the more complicated methods of Baker [10].

9.6.5 Constructing Arithmetic Distributions

The easiest approach to constructing a discrete severity distribution from a continuous one is to place the discrete probabilities on multiples of a convenient unit of measurement h, the **span**. Such a distribution is called **arithmetic** because it is defined on the nonnegative integers. In order to arithmetize a distribution, it is important to preserve the properties of the original distribution both locally through the range of the distribution and globally – that is, for the entire distribution. This should preserve the general shape of the distribution and at the same time preserve global quantities such as moments.

The methods suggested here apply to the discretization (arithmetization) of continuous, mixed, and nonarithmetic discrete distributions.

9.6.5.1 The Method of Rounding (Mass Dispersal) Let f_j denote the probability placed at jh, $j = 0, 1, 2, \dots$. Then, set[2]

$$f_0 = \Pr\left(X < \frac{h}{2} \right) = F_X\left(\frac{h}{2} - 0 \right),$$

$$f_j = \Pr\left(jh - \frac{h}{2} \le X < jh + \frac{h}{2} \right)$$

$$= F_X\left(jh + \frac{h}{2} - 0 \right) - F_X\left(jh - \frac{h}{2} - 0 \right), \quad j = 1, 2, \dots .$$

This method concentrates all the probability one-half span on either side of jh and places it at jh. There is an exception for the probability assigned to zero. This, in effect, rounds all amounts to the nearest convenient monetary unit, h, the span of the distribution. When

---

[2]The notation $F_X(x-0)$ indicates that discrete probability at x should not be included. For continuous distributions, this will make no difference. Another way to look at this is that when there is discrete probability at one of the boundary points, it should be assigned to the value one-half span above that point.

the continuous severity distribution is unbounded, it is reasonable to halt the discretization process at some point once most of the probability has been accounted for. If the index for this last point is m, then $f_m = 1 - F_X[(m - 0.5)h - 0]$. With this method, the discrete probabilities are never negative and sum to 1, ensuring that the resulting distribution is legitimate.

9.6.5.2 *The Method of Local Moment Matching* In this method, we construct an arithmetic distribution that matches p moments of the arithmetic and the true severity distributions. Consider an arbitrary interval of length ph, denoted by $[x_k, x_k + ph)$. We locate point masses $m_0^k, m_1^k, \ldots, m_p^k$ at points $x_k, x_k + h, \ldots, x_k + ph$ so that the first p moments are preserved. The system of $p + 1$ equations reflecting these conditions is

$$\sum_{j=0}^{p}(x_k + jh)^r m_j^k = \int_{x_k-0}^{x_k+ph-0} x^r dF_X(x), \quad r = 0, 1, 2, \ldots, p, \tag{9.27}$$

where the notation "-0" at the limits of the integral indicates that discrete probability at x_k is to be included but discrete probability at $x_k + ph$ is to be excluded.

Arrange the intervals so that $x_{k+1} = x_k + ph$ and so that the endpoints coincide. Then, the point masses at the endpoints are added together. With $x_0 = 0$, the resulting discrete distribution has successive probabilities:

$$\begin{aligned} f_0 &= m_0^0, & f_1 &= m_1^0, & f_2 &= m_2^0, \ldots, \\ f_p &= m_p^0 + m_0^1, & f_{p+1} &= m_1^1, & f_{p+2} &= m_2^1, \ldots. \end{aligned} \tag{9.28}$$

By summing (9.27) for all possible values of k, with $x_0 = 0$, it is clear that the first p moments are preserved for the entire distribution and that the probabilities add to 1 exactly. It only remains to solve the system of equations (9.27).

Theorem 9.11 *The solution of (9.27) is*

$$m_j^k = \int_{x_k-0}^{x_k+ph-0} \prod_{i\neq j} \frac{x - x_k - ih}{(j - i)h} dF_X(x), \quad j = 0, 1, \ldots, p. \tag{9.29}$$

Proof: The Lagrange formula for collocation of a polynomial $f(y)$ at points $y_0, y_1, \ldots, y_n$ is

$$f(y) = \sum_{j=0}^{n} f(y_j) \prod_{i\neq j} \frac{y - y_i}{y_j - y_i}.$$

Applying this formula to the polynomial $f(y) = y^r$ over the points $x_k, x_k + h, \ldots, x_k + ph$ yields

$$x^r = \sum_{j=0}^{p}(x_k + jh)^r \prod_{i\neq j} \frac{x - x_k - ih}{(j - i)h}, \quad r = 0, 1, \ldots, p.$$

Integrating over the interval $[x_k, x_k + ph)$ with respect to the severity distribution results in

$$\int_{x_k-0}^{x_k+ph-0} x^r dF_X(x) = \sum_{j=0}^{p}(x_k + jh)^r m_j^k,$$

where m_j^k is given by (9.29). Hence, the solution (9.29) preserves the first p moments, as required. □

Table 9.12   The discretization of the exponential distribution by two methods.

j	f_j rounding	f_j matching
0	0.09516	0.09365
1	0.16402	0.16429
2	0.13429	0.13451
3	0.10995	0.11013
4	0.09002	0.09017
5	0.07370	0.07382
6	0.06034	0.06044
7	0.04940	0.04948
8	0.04045	0.04051
9	0.03311	0.03317
10	0.02711	0.02716

■ **EXAMPLE 9.13**

Suppose that X has an exponential distribution with pdf $f(x) = 0.1e^{-0.1x}$. Use a span of $h = 2$ to discretize this distribution by the method of rounding and by matching the first moment.

For the method of rounding, the general formulas are

$$f_0 = F(1) = 1 - e^{-0.1(1)} = 0.09516,$$
$$f_j = F(2j + 1) - F(2j - 1) = e^{-0.1(2j-1)} - e^{-0.1(2j+1)}.$$

The first few values are given in Table 9.12.

For matching the first moment we have $p = 1$ and $x_k = 2k$. The key equations become

$$m_0^k = \int_{2k}^{2k+2} \frac{x - 2k - 2}{-2}(0.1)e^{-0.1x}\, dx = 5e^{-0.1(2k+2)} - 4e^{-0.1(2k)},$$
$$m_1^k = \int_{2k}^{2k+2} \frac{x - 2k}{2}(0.1)e^{-0.1x}\, dx = -6e^{-0.1(2k+2)} + 5e^{-0.1(2k)},$$

and then

$$f_0 = m_0^0 = 5e^{-0.2} - 4 = 0.09365,$$
$$f_j = m_1^{j-1} + m_0^j = 5e^{-0.1(2j-2)} - 10e^{-0.1(2j)} + 5e^{-0.1(2j+2)}.$$

The first few values are also given in Table 9.12. A more direct solution for matching the first moment is provided in Exercise 9.41. □

This method of local moment matching was introduced by Gerber and Jones [44] and Gerber [43], and further studied by Panjer and Lutek [97] for a variety of empirical and analytic severity distributions. In assessing the impact of errors on aggregate stop-loss net premiums (aggregate excess-of-loss pure premiums), Panjer and Lutek [97] found that two moments were usually sufficient and that adding a third moment requirement adds

Table 9.13   The data for Exercise 9.42.

Type of hit	Probability of hit per time at bat	Compensation per hit
Single	0.14	x
Double	0.05	$2x$
Triple	0.02	$3x$
Home run	0.03	$4x$

only marginally to the accuracy. Furthermore, the rounding method and the first-moment method ($p = 1$) had similar errors, while the second-moment method ($p = 2$) provided significant improvement. The specific formulas for the method of rounding and the method of matching the first moment are given in Appendix E. A reason to favor matching zero or one moment is that the resulting probabilities will always be nonnegative. When matching two or more moments, this cannot be guaranteed.

The methods described here are qualitatively similar to numerical methods used to solve Volterra integral equations such as (9.26) developed in numerical analysis (see, e.g. Baker [10]).

9.6.6   Exercises

9.41 Show that the method of local moment matching with $k = 1$ (matching total probability and the mean) using (9.28) and (9.29) results in

$$f_0 = 1 - \frac{\mathrm{E}[X \wedge h]}{h},$$

$$f_i = \frac{2\mathrm{E}[X \wedge ih] - \mathrm{E}[X \wedge (i-1)h] - \mathrm{E}[X \wedge (i+1)h]}{h}, \quad i = 1, 2, \dots,$$

and that $\{f_i;\ i = 0, 1, 2, \dots\}$ forms a valid distribution with the same mean as the original severity distribution. Using the formula given here, verify the formula given in Example 9.13.

9.42 You are the agent for a baseball player who desires an incentive contract that will pay the amounts given in Table 9.13. The number of times at bat has a Poisson distribution with $\lambda = 200$. The parameter x is determined so that the probability of the player earning at least 4,000,000 is 0.95. Determine the player's expected compensation.

9.43 A weighted average of two Poisson distributions

$$p_k = w\frac{e^{-\lambda_1}\lambda_1^k}{k!} + (1 - w)\frac{e^{-\lambda_2}\lambda_2^k}{k!}$$

has been used by some authors (e.g. Tröbliger [121]) to treat drivers as either "good" or "bad."

(a) Find the pgf $P_N(z)$ of the number of losses in terms of the two pgfs $P_1(z)$ and $P_2(z)$ of the number of losses of the two types of drivers.

Table 9.14   The data for Exercise 9.45.

x	$f_S(x)$
3	0.0132
4	0.0215
5	0.0271
6	$f_S(6)$
7	0.0410

(b) Let $f_X(x)$ denote a severity distribution defined on the nonnegative integers. How can (9.23) be used to compute the distribution of aggregate claims for the entire group?

(c) Can your approach from (b) be extended to other frequency distributions?

9.44 (*) A compound Poisson aggregate loss model has five expected claims per year. The severity distribution is defined on positive multiples of 1,000. Given that $f_S(1) = e^{-5}$ and $f_S(2) = \frac{5}{2}e^{-5}$, determine $f_X(2)$.

9.45 (*) For a compound Poisson distribution, $\lambda = 6$ and individual losses have pf $f_X(1) = f_X(2) = f_X(4) = \frac{1}{3}$. Some of the pf values for the aggregate distribution S are given in Table 9.14. Determine $f_S(6)$.

9.46 Consider the $(a, b, 0)$ class of frequency distributions and any severity distribution defined on the positive integers $\{1, 2, \dots, M < \infty\}$, where M is the maximum possible single loss.

(a) Show that, for the compound distribution, the following backward recursion holds:

$$f_S(x) = \frac{f_S(x + M) - \sum_{y=1}^{M-1}\left(a + b\dfrac{M - y}{x + M}\right) f_X(M - y)f_S(x + y)}{\left(a + b\dfrac{M}{x + M}\right) f_X(M)}.$$

(b) For the binomial (m, q) frequency distribution, how can the preceding formula be used to obtain the distribution of aggregate losses?

9.47 (*) Aggregate claims are compound Poisson with $\lambda = 2$, $f_X(1) = \frac{1}{4}$, and $f_X(2) = \frac{3}{4}$. For a premium of 6, an insurer covers aggregate claims and agrees to pay a dividend (a refund of premium) equal to the excess, if any, of 75% of the premium over 100% of the claims. Determine the excess of premium over expected claims and dividends.

9.48 On a given day, a physician provides medical care to N_A adults and N_C children. Assume that N_A and N_C have Poisson distributions with parameters 3 and 2, respectively. The distributions of length of care per patient are as follows:

	Adult	Child
1 hour	0.4	0.9
2 hour	0.6	0.1

Let N_A, N_C, and the lengths of care for all individuals be independent. The physician charges 200 per hour of patient care. Determine the probability that the office income on a given day is less than or equal to 800.

9.49 (*) A group policyholder's aggregate claims, S, has a compound Poisson distribution with $\lambda = 1$ and all claim amounts equal to 2. The insurer pays the group the following dividend:

$$D = \begin{cases} 6 - S, & S < 6, \\ 0, & S \geq 6. \end{cases}$$

Determine $E[D]$.

9.50 You are given two independent compound Poisson random variables, S_1 and S_2, where $f_j(x)$, $j = 1, 2$, are the two single-claim size distributions. You are given $\lambda_1 = \lambda_2 = 1$, $f_1(1) = 1$, and $f_2(1) = f_2(2) = 0.5$. Let $F_X(x)$ be the single-claim size distribution function associated with the compound distribution $S = S_1 + S_2$. Calculate $F_X^{*4}(6)$.

9.51 (*) The variable S has a compound Poisson claims distribution with the following characteristics:

1. Individual claim amounts equal to 1, 2, or 3.

2. $E(S) = 56$.

3. $\text{Var}(S) = 126$.

4. $\lambda = 29$.

Determine the expected number of claims of size 2.

9.52 (*) For a compound Poisson distribution with positive integer claim amounts, the probability function follows:

$$f_S(x) = \frac{1}{x}[0.16 f_S(x-1) + k f_S(x-2) + 0.72 f_S(x-3)], \quad x = 1, 2, 3, \ldots.$$

The expected value of aggregate claims is 1.68. Determine the expected number of claims.

9.53 (*) For a portfolio of policies, you are given the following:

1. The number of claims has a Poisson distribution.

2. Claim amounts can be 1, 2, or 3.

3. A stop-loss reinsurance contract has net premiums for various deductibles as given in Table 9.15.

Determine the probability that aggregate claims will be either 5 or 6.

Table 9.15   The data for Exercise 9.53.

Deductible	Net premium
4	0.20
5	0.10
6	0.04
7	0.02

9.54 (*) For group disability income insurance, the expected number of disabilities per year is 1 per 100 lives covered. The continuance (survival) function for the length of a disability in days, Y, is

$$\Pr(Y > y) = 1 - \frac{y}{10}, \quad y = 0, 1, \dots, 10.$$

The benefit is 20 per day following a five-day waiting period. Using a compound Poisson distribution, determine the variance of aggregate claims for a group of 1,500 independent lives.

9.55 A population has two classes of drivers. The number of accidents per individual driver has a geometric distribution. For a driver selected at random from Class I, the geometric distribution parameter has a uniform distribution over the interval $(0, 1)$. Twenty-five percent of the drivers are in Class I. All drivers in Class II have expected number of claims 0.25. For a driver selected at random from this population, determine the probability of exactly two accidents.

9.56 (*) A compound Poisson claim distribution has $\lambda = 5$ and individual claim amount distribution $p(5) = 0.6$ and $p(k) = 0.4$ where $k > 5$. The expected cost of an aggregate stop-loss insurance with a deductible of 5 is 28.03. Determine the value of k.

9.57 (*) Aggregate losses have a compound Poisson claim distribution with $\lambda = 3$ and individual claim amount distribution $p(1) = 0.4$, $p(2) = 0.3$, $p(3) = 0.2$, and $p(4) = 0.1$. Determine the probability that aggregate losses do not exceed 3.

9.58 Repeat Exercise 9.57 with a negative binomial frequency distribution with $r = 6$ and $\beta = 0.5$.
  Note: Exercises 9.59 and 9.60 require the use of a computer.

9.59 A policy covers physical damage incurred by the trucks in a company's fleet. The number of losses in a year has a Poisson distribution with $\lambda = 5$. The amount of a single loss has a gamma distribution with $\alpha = 0.5$ and $\theta = 2,500$. The insurance contract pays a maximum annual benefit of 20,000. Determine the probability that the maximum benefit will be paid. Use a span of 100 and the method of rounding.

9.60 An individual has purchased health insurance, for which he pays 10 for each physician visit and 5 for each prescription. The probability that a payment will be 10 is 0.25, and the probability that it will be 5 is 0.75. The total number of payments per year has the Poisson–Poisson (Neyman Type A) distribution with $\lambda_1 = 10$ and $\lambda_2 = 4$. Determine the probability that total payments in one year will exceed 400. Compare your answer to a normal approximation.

9.61 Demonstrate that if the exponential distribution is discretized by the method of rounding, the resulting discrete distribution is a ZM geometric distribution.

9.7 The Impact of Individual Policy Modifications on Aggregate Payments

In Section 8.6 the manner in which individual deductibles (both ordinary and franchise) affect both the individual loss amounts and the claim frequency distribution is discussed. In this section, we consider the impact on aggregate losses. It is worth noting that both individual coinsurance and individual policy limits have an impact on the individual losses but not on the frequency of such losses, so in what follows we focus primarily on the deductible issues. We continue to assume that the presence of policy modifications does not have an underwriting impact on the individual loss distribution through an effect on the risk characteristics of the insured population, an issue discussed in Section 8.6. That is, the **ground-up** distribution of the individual loss amount X is assumed to be unaffected by the policy modifications, and only the payments themselves are affected.

From the standpoint of the aggregate losses, the relevant facts are now described. Regardless of whether the deductible is of the ordinary or franchise type, we shall assume that an individual loss results in a payment with probability v. The individual ground-up loss random variable X has policy modifications (including deductibles) applied, so that a payment is then made. Individual payments may then be viewed on a *per-loss* basis, where the amount of such payment, denoted by Y^L, will be zero if the loss results in no payment. Thus, on a per-loss basis, the payment amount is determined on each and every loss. Alternatively, individual payments may also be viewed on a *per-payment* basis. In this case, the amount of payment is denoted by Y^P, and on this basis payment amounts are only determined on losses that actually result in a nonzero payment being made. Therefore, by definition, $\Pr(Y^P = 0) = 0$, and the distribution of Y^P is the conditional distribution of Y^L given that $Y^L > 0$. Notationally, we write $Y^P = Y^L | Y^L > 0$. Therefore, the cumulative distribution functions are related by

$$F_{Y^L}(y) = (1 - v) + v F_{Y^P}(y), \quad y \geq 0,$$

because $1 - v = \Pr(Y^L = 0) = F_{Y^L}(0)$ (recall that Y^L has a discrete probability mass point $1 - v$ at zero, even if X and hence Y^P and Y^L have continuous probability density functions for $y > 0$). The moment generating functions of Y^L and Y^P are thus related by

$$M_{Y^L}(z) = (1 - v) + v M_{Y^P}(z), \tag{9.30}$$

which may be restated in terms of expectations as

$$\mathrm{E}(e^{zY^L}) = \mathrm{E}(e^{zY^L} | Y^L = 0) \Pr\left(Y^L = 0\right) + \mathrm{E}(e^{zY^L} | Y^L > 0) \Pr\left(Y^L > 0\right).$$

It follows from Section 8.6 that the number of losses N^L and the number of payments N^P are related through their probability generating functions by

$$P_{N^P}(z) = P_{N^L}(1 - v + vz), \tag{9.31}$$

where $P_{N^P}(z) = \mathrm{E}\left(z^{N^P}\right)$ and $P_{N^L}(z) = \mathrm{E}\left(z^{N^L}\right)$.

We now turn to the analysis of the aggregate payments. On a per-loss basis, the total payments may be expressed as $S = Y_1^L + Y_2^L + \cdots + Y_{N^L}^L$, with $S = 0$ if $N^L = 0$,

and where Y_j^L is the payment amount on the jth loss. Alternatively, ignoring losses on which no payment is made, we may express the total payments on a per-payment basis as $S = Y_1^P + Y_2^P + \cdots + Y_{N^P}^P$, with $S = 0$ if $N^P = 0$, and Y_j^P is the payment amount on the jth loss, which results in a nonzero payment. Clearly, S may be represented in two distinct ways on an aggregate basis. Of course, the moment generating function of S on a per-loss basis is

$$M_S(z) = \mathrm{E}\left(e^{zS}\right) = P_{N^L}\left[M_{Y^L}(z)\right], \tag{9.32}$$

whereas on a per-payment basis we have

$$M_S(z) = \mathrm{E}\left(e^{zS}\right) = P_{N^P}\left[M_{Y^P}(z)\right]. \tag{9.33}$$

Obviously, (9.32) and (9.33) are equal, as may be seen from (9.30) and (9.31). That is,

$$P_{N^L}\left[M_{Y^L}(z)\right] = P_{N^L}\left[1 - \upsilon + \upsilon M_{Y^P}(z)\right] = P_{N^P}\left[M_{Y^P}(z)\right].$$

Consequently, any analysis of the aggregate payments S may be done on either a per-loss basis (with compound representation (9.32) for the moment generating function) or on a per-payment basis (with (9.33) as the compound moment generating function). The basis selected should be determined by whatever is more suitable for the particular situation at hand. While by no means a hard-and-fast rule, we have found it more convenient to use the per-loss basis to evaluate moments of S. In particular, the formulas given in Section 8.5 for the individual mean and variance are on a per-loss basis, and the mean and variance of the aggregate payments S may be computed using these and (9.9), but with N replaced by N^L and X by Y^L.

If the (approximated) distribution of S is of more interest than the moments, then a per-payment basis is normally to be preferred. The reason for this choice is that on a per-loss basis, underflow problems may result if $\mathrm{E}(N^L)$ is large, and computer storage problems may occur due to the presence of a large number of zero probabilities in the distribution of Y^L, particularly if a franchise deductible is employed. Also, for convenience, we normally elect to apply policy modifications to the individual loss distribution first and then discretize (if necessary), rather than discretizing and then applying policy modifications to the discretized distributions. This issue is only relevant, however, if the deductible and policy limit are not integer multiples of the discretization span. The following example illustrates these ideas.

■ **EXAMPLE 9.14**

The number of ground-up losses is Poisson distributed with mean $\lambda = 3$. The individual loss distribution is Pareto with parameters $\alpha = 4$ and $\theta = 10$. An individual ordinary deductible of 6, coinsurance of 75%, and an individual loss limit of 24 (before application of the deductible and coinsurance) are all applied. Determine the mean, variance, and distribution of aggregate payments.

We first compute the mean and variance on a per-loss basis. The mean number of losses is $\mathrm{E}(N^L) = 3$, and the mean individual payment on a per-loss basis is (using Theorem 8.7 with $r = 0$ and the Pareto distribution)

$$\mathrm{E}(Y^L) = 0.75\left[\mathrm{E}(X \wedge 24) - \mathrm{E}(X \wedge 6)\right] = 0.75(3.2485 - 2.5195) = 0.54675.$$

The mean of the aggregate payments is thus

$$E(S) = E(N^L)E(Y^L) = (3)(0.54675) = 1.64.$$

The second moment of the individual payments on a per-loss basis is, using Theorem 8.8 with $r = 0$ and the Pareto distribution,

$$
\begin{aligned}
E\left[(Y^L)^2\right] &= (0.75)^2\{E\left[(X \wedge 24)^2\right] - E\left[(X \wedge 6)^2\right] \\
&\quad - 2(6)E(X \wedge 24) + 2(6)E(X \wedge 6)\} \\
&= (0.75)^2 \left[26.3790 - 10.5469 - 12(3.2485) + 12(2.5195)\right] \\
&= 3.98481.
\end{aligned}
$$

To compute the variance of aggregate payments, we do not need to explicitly determine $\text{Var}(Y^L)$ because S is compound Poisson distributed, which implies (using, for example (7.10)) that

$$\text{Var}(S) = \lambda E\left[(Y^L)^2\right] = 3(3.98481) = 11.9544 = (3.46)^2.$$

To compute the (approximate) distribution of S, we use the per-payment basis. First note that $v = Pr(X > 6) = [10/(10 + 6)]^4 = 0.15259$, and the number of payments N^P is Poisson distributed with mean $E(N^P) = \lambda v = 3(0.15259) = 0.45776$. Let $Z = X - 6|X > 6$, so that Z is the individual payment random variable with only a deductible of 6. Then,

$$Pr(Z > z) = \frac{Pr(X > z + 6)}{Pr(X > 6)}.$$

With coinsurance of 75%, $Y^P = 0.75Z$ has cumulative distribution function

$$F_{Y^P}(y) = 1 - Pr(0.75Z > y) = 1 - \frac{Pr(X > 6 + y/0.75)}{Pr(X > 6)}.$$

That is, for y less than the maximum payment of $(0.75)(24 - 6) = 13.5$,

$$F_{Y^P}(y) = \frac{Pr(X > 6) - Pr(X > 6 + y/0.75)}{Pr(X > 6)}, \quad y < 13.5,$$

and $F_{Y^P}(y) = 1$ for $y \geq 13.5$. We then discretize the distribution of Y^P (we thus apply the policy modifications first and then discretize) using a span of 2.25 and the method of rounding. This approach yields $f_0 = F_{Y^P}(1.125) = 0.30124$, $f_1 = F_{Y^P}(3.375) - F_{Y^P}(1.125) = 0.32768$, and so on. In this situation, care must be exercised in the evaluation of f_6, and we have $f_6 = F_{Y^P}(14.625) - F_{Y^P}(12.375) = 1 - 0.94126 = 0.05874$. Then, $f_n = 1 - 1 = 0$ for $n = 7, 8, \ldots$. The approximate distribution of S may then be computed using the compound Poisson recursive formula, namely $f_S(0) = e^{-0.45776(1-0.30124)} = 0.72625$, and

$$f_S(x) = \frac{0.45776}{x} \sum_{y=1}^{x \wedge 6} y f_y f_S(x - y), \quad x = 1, 2, 3, \ldots.$$

Thus, $f_S(1) = (0.45776)(1)(0.32768)(0.72625) = 0.10894$, for example. $\quad\square$

9.7.1 Exercises

9.62 Suppose that the number of ground-up losses N^L has probability generating function $P_{N^L}(z) = B[\theta(z-1)]$, where θ is a parameter and B is functionally independent of θ. The individual ground-up loss distribution is exponential with cumulative distribution function $F_X(x) = 1 - e^{-\mu x}$, $x \geq 0$. Individual losses are subject to an ordinary deductible of d and coinsurance of α. Demonstrate that the aggregate payments, on a per-payment basis, have a compound moment generating function given by (9.33), where N^P has the same distribution as N^L but with θ replaced by $\theta e^{-\mu d}$, and Y^P has the same distribution as X, but with μ replaced by μ/α.

9.63 A ground-up model of individual losses has a gamma distribution with parameters $\alpha = 2$ and $\theta = 100$. The number of losses has a negative binomial distribution with $r = 2$ and $\beta = 1.5$. An ordinary deductible of 50 and a loss limit of 175 (before imposition of the deductible) are applied to each individual loss.

(a) Determine the mean and variance of the aggregate payments on a per-loss basis.

(b) Determine the distribution of the number of payments.

(c) Determine the cumulative distribution function of the amount Y^P of a payment, given that a payment is made.

(d) Discretize the severity distribution from (c) using the method of rounding and a span of 40.

(e) Use the recursive formula to calculate the discretized distribution of aggregate payments up to a discretized amount paid of 120.

9.8 The Individual Risk Model

9.8.1 The Model

The **individual risk model** represents the aggregate loss as a fixed sum of independent (but not necessarily identically distributed) random variables:

$$S = X_1 + X_2 + \cdots + X_n.$$

This formula is usually thought of as the sum of the losses from n insurance contracts, for example, n persons covered under a group insurance policy.

The individual risk model was originally developed for life insurance, in which the probability of death within a year is q_j and the fixed benefit paid for the death of the jth person is b_j. In this case, the distribution of the loss to the insurer for the jth policy is

$$f_{X_j}(x) = \begin{cases} 1 - q_j, & x = 0, \\ q_j, & x = b_j. \end{cases}$$

The mean and variance of aggregate losses are

$$E(S) = \sum_{j=1}^{n} b_j q_j$$

and

$$\text{Var}(S) = \sum_{j=1}^{n} b_j^2 q_j (1 - q_j)$$

because the X_js are assumed to be independent. Then, the pgf of aggregate losses is

$$P_S(z) = \prod_{j=1}^{n} (1 - q_j + q_j z^{b_j}). \tag{9.34}$$

In the special case where all the risks are identical with $q_j = q$ and $b_j = 1$, the pgf reduces to

$$P_S(z) = [1 + q(z - 1)]^n,$$

and in this case S has a binomial distribution.

The individual risk model can be generalized as follows. Let $X_j = I_j B_j$, where $I_1, \ldots, I_n$, $B_1, \ldots, B_n$ are independent. The random variable I_j is an indicator variable that takes on the value 1 with probability q_j and the value 0 with probability $1 - q_j$. This variable indicates whether the jth policy produced a payment. The random variable B_j can have any distribution and represents the amount of the payment in respect of the jth policy given that a payment was made. In the life insurance case, B_j is degenerate, with all probability on the value b_j.

The mgf corresponding to (9.34) is

$$M_S(z) = \prod_{j=1}^{n} [1 - q_j + q_j M_{B_j}(z)]. \tag{9.35}$$

If we let $\mu_j = E(B_j)$ and $\sigma_j^2 = \text{Var}(B_j)$, then

$$E(S) = \sum_{j=1}^{n} q_j \mu_j \tag{9.36}$$

and

$$\text{Var}(S) = \sum_{j=1}^{n} [q_j \sigma_j^2 + q_j (1 - q_j) \mu_j^2]. \tag{9.37}$$

You are asked to verify these formulas in Exercise 9.64. The following example is a simple version of this situation.

■ **EXAMPLE 9.15**

Consider a group life insurance contract with an accidental death benefit. Assume that for all members the probability of death in the next year is 0.01 and that 30% of deaths are accidental. For 50 employees, the benefit for an ordinary death is 50,000 and for an accidental death it is 100,000. For the remaining 25 employees, the benefits are

75,000 and 150,000, respectively. Develop an individual risk model and determine its mean and variance.

For all 75 employees, $q_j = 0.01$. For 50 employees, B_j takes on the value 50,000 with probability 0.7 and 100,000 with probability 0.3. For them, $\mu_j = 65,000$ and $\sigma_j^2 = 525,000,000$. For the remaining 25 employees, B_j takes on the value 75,000 with probability 0.7 and 150,000 with probability 0.3. For them, $\mu_j = 97,500$ and $\sigma_j^2 = 1,181,250,000$. Then,

$$\begin{aligned} E(S) &= 50(0.01)(65,000) + 25(0.01)(97,500) \\ &= 56,875 \end{aligned}$$

and

$$\begin{aligned} \text{Var}(S) &= 50(0.01)(525,000,000) + 50(0.01)(0.99)(65,000)^2 \\ &\quad + 25(0.01)(1,181,250,000) + 25(0.01)(0.99)(97,500)^2 \\ &= 5,001,984,375. \end{aligned}$$ $\square$

With regard to calculating the probabilities, there are at least three options. One is to do an exact calculation, which involves numerous convolutions and almost always requires more excessive computing time. Recursive formulas have been developed, but they are cumbersome and are not presented here. For one such method, see De Pril [27]. One alternative is a parametric approximation as discussed for the collective risk model. Another alternative is to replace the individual risk model with a similar collective risk model and then do the calculations with that model. These two approaches are presented here.

9.8.2   Parametric Approximation

A normal, gamma, lognormal, or any other distribution can be used to approximate the distribution, usually done by matching the first few moments. Because the normal, gamma, and lognormal distributions each have two parameters, the mean and variance are sufficient.

■ **EXAMPLE 9.16**

(*Group life insurance*) A small manufacturing business has a group life insurance contract on its 14 permanent employees. The actuary for the insurer has selected a mortality table to represent the mortality of the group. Each employee is insured for the amount of his or her salary rounded up to the next 1,000. The group's data are given in Table 9.16.

If the insurer adds a 45% relative loading to the net (pure) premium, what are the chances that it will lose money in a given year? Use the normal and lognormal approximations.

The mean and variance of the aggregate losses for the group are

$$E(S) = \sum_{j=1}^{14} b_j q_j = 2,054.41$$

Table 9.16 The employee data for Example 9.16.

Employee, j	Age (years)	Sex	Benefit, b_j	Mortality rate, q_j
1	20	M	15,000	0.00149
2	23	M	16,000	0.00142
3	27	M	20,000	0.00128
4	30	M	28,000	0.00122
5	31	M	31,000	0.00123
6	46	M	18,000	0.00353
7	47	M	26,000	0.00394
8	49	M	24,000	0.00484
9	64	M	60,000	0.02182
10	17	F	14,000	0.00050
11	22	F	17,000	0.00050
12	26	F	19,000	0.00054
13	37	F	30,000	0.00103
14	55	F	55,000	0.00479
Total			373,000	

and

$$\text{Var}(S) = \sum_{j=1}^{14} b_j^2 q_j (1 - q_j) = 1.02534 \times 10^8.$$

The premium being charged is $1.45 \times 2,054.41 = 2{,}978.89$. For the normal approximation (in units of 1,000), the mean is 2.05441 and the variance is 102.534. Then, the probability of a loss is

$$\Pr(S > 2.97889) = \Pr\left[Z > \frac{2.97889 - 2.05441}{(102.534)^{1/2}}\right]$$
$$\doteq \Pr(Z > 0.0913)$$
$$= 0.46 \text{ or } 46\%.$$

For the lognormal approximation (as in Example 9.4),

$$\mu + \tfrac{1}{2}\sigma^2 = \ln 2.05441 = 0.719989$$

and

$$2\mu + 2\sigma^2 = \ln(102.534 + 2.05441^2) = 4.670533.$$

From this, $\mu = -0.895289$ and $\sigma^2 = 3.230555$. Then,

$$\Pr(S > 2.97889) = 1 - \Phi\left[\frac{\ln 2.97889 + 0.895289}{(3.230555)^{1/2}}\right]$$
$$= 1 - \Phi(1.105)$$
$$= 0.13, \text{ or } 13\%.$$

$\square$

9.8.3 Compound Poisson Approximation

Because of the computational complexity of calculating the distribution of total claims for a portfolio of n risks using the individual risk model, it has been popular to attempt to approximate the distribution by using the compound Poisson distribution. As seen in Section 9.6, use of the compound Poisson allows calculation of the total claims distribution by using a very simple recursive procedure.

To proceed, note that the indicator random variable I_j has pgf $P_{I_j}(z) = 1 - q_j + q_j z$, and thus (9.35) may be expressed as

$$M_S(z) = \prod_{j=1}^{n} P_{I_j}[M_{B_j}(z)]. \tag{9.38}$$

Note that I_j has a binomial distribution with parameters $m = 1$ and $q = q_j$. To obtain the compound Poisson approximation, assume that I_j has a Poisson distribution with mean λ_j. If $\lambda_j = q_j$, then the Poisson mean is the same as the binomial mean, which should provide a good approximation if q_j is close to zero. An alternative to equating the mean is to equate the probability of no loss. For the binomial distribution, that probability is $1 - q_j$, and for the Poisson distribution, it is $\exp(-\lambda_j)$. Equating these two probabilities gives the alternative approximation $\lambda_j = -\ln(1 - q_j) > q_j$. This second approximation is appropriate in the context of a group life insurance contract where a life is "replaced" upon death, leaving the Poisson intensity unchanged by the death. Naturally the expected number of losses is greater than $\sum_{j=1}^{n} q_j$. An alternative choice is proposed by Kornya [76]. It uses $\lambda_j = q_j/(1 - q_j)$ and results in an expected number of losses that exceeds that using the method that equates the no-loss probabilities (see Exercise 9.65).

Regardless of the approximation used, Theorem 9.7 yields, from (9.38) using $P_{I_j}(x) = \exp[\lambda_j(z - 1)]$,

$$M_S(z) = \prod_{j=1}^{n} \exp\{\lambda_j[M_{B_j}(z) - 1]\} = \exp\{\lambda[M_X(z) - 1]\},$$

where

$$\lambda = \sum_{j=1}^{n} \lambda_j,$$

$$M_X(z) = \lambda^{-1} \sum_{j=1}^{n} \lambda_j M_{B_j}(z),$$

and so X has pf or pdf

$$f_X(x) = \lambda^{-1} \sum_{j=1}^{n} \lambda_j f_{B_j}(x), \tag{9.39}$$

which is a weighted average of the n individual severity densities.

If $\Pr(B_j = b_j) = 1$ as in life insurance, then (9.39) becomes

$$f_X(x) = \Pr(X = x) = \lambda^{-1} \sum_{\{j : b_j = x\}} \lambda_j. \tag{9.40}$$

The numerator sums all probabilities associated with amount b_j.

Table 9.17   The aggregate distribution for Example 9.17.

x	$F_S(x)$	x	$F_S(x)$	x	$F_S(x)$	x	$F_S(x)$
0	0.9530099	20	0.9618348	40	0.9735771	60	0.9990974
1	0.9530099	21	0.9618348	41	0.9735850	61	0.9990986
2	0.9530099	22	0.9618348	42	0.9736072	62	0.9990994
3	0.9530099	23	0.9618348	43	0.9736133	63	0.9990995
4	0.9530099	24	0.9664473	44	0.9736346	64	0.9990995
5	0.9530099	25	0.9664473	45	0.9736393	65	0.9990996
6	0.9530099	26	0.9702022	46	0.9736513	66	0.9990997
7	0.9530099	27	0.9702022	47	0.9736541	67	0.9990997
8	0.9530099	28	0.9713650	48	0.9736708	68	0.9990998
9	0.9530099	29	0.9713657	49	0.9736755	69	0.9991022
10	0.9530099	30	0.9723490	50	0.9736956	70	0.9991091
11	0.9530099	31	0.9735235	51	0.9736971	71	0.9991156
12	0.9530099	32	0.9735268	52	0.9737101	72	0.9991179
13	0.9530099	33	0.9735328	53	0.9737102	73	0.9991341
14	0.9534864	34	0.9735391	54	0.9737195	74	0.9991470
15	0.9549064	35	0.9735433	55	0.9782901	75	0.9991839
16	0.9562597	36	0.9735512	56	0.9782947	76	0.9992135
17	0.9567362	37	0.9735536	57	0.9782994	77	0.9992239
18	0.9601003	38	0.9735604	58	0.9783006	78	0.9992973
19	0.9606149	39	0.9735679	59	0.9783021	79	0.9993307

■ **EXAMPLE 9.17**

(*Example 9.16 continued*) Consider the group life case of Example 9.16. Derive a compound Poisson approximation with the means matched.

Using the compound Poisson approximation of this section with Poisson parameter $\lambda = \sum q_j = 0.04813$, the distribution function given in Table 9.17 is obtained. When these values are compared to the exact solution (not presented here), the maximum error of 0.0002708 occurs at $x = 0$.   □

■ **EXAMPLE 9.18**

(*Example 9.15 continued*) Develop compound Poisson approximations using all three methods suggested here. Compute the mean and variance for each approximation and compare them to the exact value.

Using the method that matches the mean, we have $\lambda = 50(0.01)+25(0.01) = 0.75$.

The severity distribution is

$$f_X(50{,}000) = \frac{50(0.01)(0.7)}{0.75} = 0.4667,$$

$$f_X(75{,}000) = \frac{25(0.01)(0.7)}{0.75} = 0.2333,$$

$$f_X(100{,}000) = \frac{50(0.01)(0.3)}{0.75} = 0.2000,$$

$$f_X(150{,}000) = \frac{25(0.01)(0.3)}{0.75} = 0.1000.$$

The mean is $\lambda E(X) = 0.75(75{,}833.33) = 56{,}875$, which matches the exact value, and the variance is $\lambda E(X^2) = 0.75(6{,}729{,}166{,}667) = 5{,}046{,}875{,}000$, which exceeds the exact value.

For the method that preserves the probability of no losses, $\lambda = -75 \ln(0.99)$ $= 0.753775$. For this method, the severity distribution turns out to be exactly the same as before (because all individuals have the same value of q_j). Thus the mean is 57,161 and the variance is 5,072,278,876, both of which exceed the previous approximate values.

Using Kornya's method, $\lambda = 75(0.01)/0.99 = 0.757576$ and again the severity distribution is unchanged. The mean is 57,449 and the variance is 5,097,853,535, which are the largest values of all. $\square$

9.8.4   Exercises

9.64 Derive (9.36) and (9.37).

9.65 Demonstrate that the compound Poisson model given by $\lambda_j = q_j$ and (9.40) produces a model with the same mean as the exact distribution but with a larger variance. Then show that the one using $\lambda_j = -\ln(1 - q_j)$ must produce a larger mean and an even larger variance, and, finally, show that the one using $\lambda_j = q_j/(1 - q_j)$ must produce the largest mean and variance of all.

9.66 (*) Individual members of an insured group have independent claims. Aggregate payment amounts for males have mean 2 and variance 4, while females have mean 4 and variance 10. The premium for a group with future claims S is the mean of S plus 2 times the standard deviation of S. If the genders of the members of a group of m members are not known, the number of males is assumed to have a binomial distribution with parameters m and $q = 0.4$. Let A be the premium for a group of 100 for which the genders of the members are not known and let B be the premium for a group of 40 males and 60 females. Determine A/B.

9.67 (*) An insurance company assumes that the claim probability for smokers is 0.02, while for nonsmokers it is 0.01. A group of mutually independent lives has coverage of 1,000 per life. The company assumes that 20% of the lives are smokers. Based on this assumption, the premium is set equal to 110% of expected claims. If 30% of the lives are smokers, the probability that claims will exceed the premium is less than 0.20. Using the normal approximation, determine the minimum number of lives that must be in the group.

Table 9.18   The distribution for Exercise 9.68.

x	$F_S(x)$
0	0.40
100	0.58
200	0.64
300	0.69
400	0.70
500	0.78
600	0.96
700	1.00

Table 9.19   The data for Exercise 9.69.

Age group	Number in age group	Probability of claim per life	Mean of the exponential distribution of claim amounts
18–35	400	0.03	5
36–50	300	0.07	3
51–65	200	0.10	2

Table 9.20   The data for Exercise 9.70.

Service	Probability of claim	Distribution of annual charges given that a claim occurs	
		Mean	Variance
Office visits	0.7	160	4,900
Surgery	0.2	600	20,000
Other services	0.5	240	8,100

9.68 (*) Based on the individual risk model with independent claims, the cumulative distribution function of aggregate claims for a portfolio of life insurance policies is as shown in Table 9.18. One policy with face amount 100 and probability of claim 0.20 is increased in face amount to 200. Determine the probability that aggregate claims for the revised portfolio will not exceed 500.

9.69 (*) A group life insurance contract covering independent lives is rated in the three age groupings as given in Table 9.19. The insurer prices the contract so that the probability that claims will exceed the premium is 0.05. Using the normal approximation, determine the premium that the insurer will charge.

9.70 (*) The probability model for the distribution of annual claims per member in a health plan is shown in Table 9.20. Independence of costs and occurrences among services and members is assumed. Using the normal approximation, determine the minimum number of members that a plan must have such that the probability that actual charges will exceed 115% of the expected charges is less than 0.10.

Table 9.21   The data for Exercise 9.71.

Class	Probability of claim	Benefit	Number of risks
Standard	0.2	k	3,500
Substandard	0.6	αk	2,000

Table 9.22   The data for Exercise 9.72.

Class	Number in class	Benefit amount	Probability of a claim
1	500	x	0.01
2	500	$2x$	0.02

Table 9.23   The data for Exercise 9.73.

Class	Benefit amount	Probability of death	Number of policies
1	100,000	0.10	500
2	200,000	0.02	500
3	300,000	0.02	500
4	200,000	0.10	300
5	200,000	0.10	500

9.71 (*) An insurer has a portfolio of independent risks as given in Table 9.21. The insurer sets α and k such that aggregate claims have expected value 100,000 and minimum variance. Determine α.

9.72 (*) An insurance company has a portfolio of independent one-year term life policies as given in Table 9.22. The actuary approximates the distribution of claims in the individual model using the compound Poisson model, in which the expected number of claims is the same as in the individual model. Determine the maximum value of x such that the variance of the compound Poisson approximation is less than 4,500.

9.73 (*) An insurance company sold one-year term life insurance on a group of 2,300 independent lives as given in Table 9.23. The insurance company reinsures amounts in excess of 100,000 on each life. The reinsurer wishes to charge a premium that is sufficient to guarantee that it will lose money 5% of the time on such groups. Obtain the appropriate premium by each of the following ways:

(a) Using a normal approximation to the aggregate claims distribution.
(b) Using a lognormal approximation.
(c) Using a gamma approximation.
(d) Using the compound Poisson approximation that matches the means.

9.74 A group insurance contract covers 1,000 employees. An employee can have at most one claim per year. For 500 employees, there is a 0.02 probability of a claim, and when there is a claim, the amount has an exponential distribution with mean 500. For 250 other employees, there is a 0.03 probability of a claim and amounts are exponential with mean 750. For the remaining 250 employees, the probability is 0.04 and the mean is 1,000. Determine the exact mean and variance of total claims payments. Next, construct a compound Poisson model with the same mean and determine the variance of this model.

MATHEMATICAL STATISTICS

10

INTRODUCTION TO MATHEMATICAL STATISTICS

10.1   Introduction and Four Data Sets

Before studying empirical models and then parametric models, we review some concepts from mathematical statistics. Mathematical statistics is a broad subject that includes many topics not covered in this chapter. For those topics that are covered, it is assumed that you have had some prior exposure. The topics of greatest importance for constructing actuarial models are estimation and hypothesis testing. Because the Bayesian approach to statistical inference is often either ignored or treated lightly in introductory mathematical statistics texts and courses, it receives more in-depth coverage in this text, in Chapter 13. Bayesian methodology also provides the basis for the credibility methods covered in Chapter 16.

To see the need for methods of statistical inference, consider the case where your supervisor needs a model for basic dental payments. One option is to simply announce the model. You proclaim that it is the lognormal distribution with $\mu = 5.1239$ and $\sigma = 1.0345$. (The many decimal places are designed to give your proclamation an aura of precision.) When your supervisor, a regulator, or an attorney who has put you on the witness stand, asks you how you know that to be so, it will likely not be sufficient to answer that "I just know these things," "trust me, I am a trained statistician," "it is too complicated, you wouldn't understand," or "my friend at Gamma Dental uses that model."

Table 10.1 Data Set A.

Number of accidents	Number of drivers
0	81,714
1	11,306
2	1,618
3	250
4	40
5 or more	7

An alternative is to collect some data and use it to formulate a model. Most distributional models have two components. The first is a name, such as "Pareto." The second is the set of parameter values that complete the specification. Matters would be simpler if modeling could be done in that order. Most of the time, we need to fix the parameters that go with a named model before we can decide if we want to use that model.

Because the parameter estimates are based on a sample from the population and not the entire population, the results will not be the true values. It is important to have an idea of the potential error. One way to express this error is with an interval estimate. That is, rather than announcing a particular value, a range of plausible values is presented.

When named parametric distributions are used, the parameterizations used are those from Appendices A and B.

Alternatively, you may want to construct a nonparametric model (also called an empirical model), where the goal is to determine a model that essentially reproduces the data. Such models are discussed in Chapter 14.

At this point we present four data sets, referred to as Data Sets A, B, C, and D. They will be used several times, both in this chapter and in later chapters.

Data Set A This data set is well known in the casualty actuarial literature. It was first analyzed in the paper [30] by Dropkin in 1959. From 1956 to 1958, he collected data on the number of accidents by one driver in one year. The results for 94,935 drivers are shown in Table 10.1.

Data Set B These numbers (and those in the next two data sets) are artificial. They represent the amounts paid on workers compensation medical benefits but are not related to any particular policy or set of policyholders. These payments are the full amount of the loss. A random sample of 20 payments is given in Table 10.2.

Data Set C These observations represent payments on 227 claims from a general liability insurance policy. The data are shown in Table 10.3.

Data Set D This data set is from the experience of five-year term insurance policies. The study period is a fixed time period. The columns are interpreted as follows: (1) i is the policy number, 1–40; and (2) d_i is the time since issue to when the insured was first observed. Thus, policies 1–30 were observed from when the policy was sold. The remaining policies

Table 10.2 Data Set B.

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1,193	1,340	1,884	2,558	15,743

Table 10.3   Data Set C.

Payment range	Number of payments
0–7,500	99
7,500–17,500	42
17,500–32,500	29
32,500–67,500	28
67,500–125,000	17
125,000–300,000	9
Over 300,000	3

Table 10.4   Data Set D.

i	d_i	x_i	u_i	i	d_i	x_i	u_i
1	0	—	0.1	16	0	4.8	—
2	0	—	0.5	17	0	—	4.8
3	0	—	0.8	18	0	—	4.8
4	0	0.8	—	19–30	0	—	5.0
5	0	—	1.8	31	0.3	—	5.0
6	0	—	1.8	32	0.7	—	5.0
7	0	—	2.1	33	1.0	4.1	—
8	0	—	2.5	34	1.8	3.1	—
9	0	—	2.8	35	2.1	—	3.9
10	0	2.9	—	36	2.9	—	5.0
11	0	2.9	—	37	2.9	—	4.8
12	0	—	3.9	38	3.2	4.0	—
13	0	4.0	—	39	3.4	—	5.0
14	0	—	4.0	40	3.9	—	5.0
15	0	—	4.1				

were issued prior to the start of the observation period and were known to be alive at that duration. (3) x_i is the time since issue to when the insured was observed to die. Those who were not observed to die during the five years have "—" in that column. (4) u_i is the latest time since issue at which those who were not observed to die were observed. That could be because they surrendered their policy before the five years elapsed, reached the end of the five-year term, or the study ended while the policy was still in force. The data are shown in Table 10.4.

10.2   Point Estimation

10.2.1   Introduction

Regardless of how a model is estimated, it is extremely unlikely that the estimated model will exactly match the true distribution. Ideally, we would like to be able to measure the error we will be making when using the estimated model. But doing so is clearly impossible! If we knew the amount of error we had made, we could adjust our estimate

by that amount and then have no error at all. The best we can do is discover how much error is inherent in repeated use of the *procedure*, as opposed to how much error we made with our current estimate. Therefore, we are concerned about the quality of the ensemble of answers produced from the procedure, not about the quality of a particular answer.

This is a critical point with regard to actuarial practice. What is important is that an appropriate procedure be used, with everyone understanding that even the best procedure can lead to a poor result once the random future outcome has been revealed. This point is stated nicely in a Society of Actuaries principles draft [115, pp. 779–780] regarding the level of adequacy of a provision for a portfolio of life insurance risk obligations (i.e. the probability that the company will have enough money to meet its contractual obligations):

> The indicated level of adequacy is prospective, but the actuarial model is generally validated against past experience. It is incorrect to conclude on the basis of subsequent experience that the actuarial assumptions were inappropriate or that the indicated level of adequacy was overstated or understated.

When constructing models, there are several types of error. Some, such as model error (choosing the wrong model) and sampling frame error (trying to draw inferences about a population that differs from the one sampled), are not covered here. An example of model error is selecting a Pareto distribution when the true distribution is, or is close to, Weibull. An example of sampling frame error is sampling claims from insurance policies that were sold by independent agents to price policies that are to be sold over the internet.

The type of error that we can measure is that resulting from using a sample from the population to make inferences about the entire population. Errors occur when the items sampled do not represent the population. As noted earlier, we cannot know if the particular items sampled today do or do not represent the population. We can, however, estimate the extent to which estimators are affected by the possibility of a nonrepresentative sample.

The approach taken in this chapter is to consider all the samples that might be taken from the population. Each such sample leads to an estimated quantity (e.g. a probability, a parameter value, or a moment). We do not expect the estimated quantities to always match the true value. For a sensible estimation procedure, we do expect that for some samples the quantity will match the true value, for many it will be close, and for only a few it will be quite different. If we can construct a measure of how well the set of potential estimates matches the true value, we have a handle on the quality of our estimation procedure. The approach outlined here is often called the **classical** or **frequentist** approach to estimation.

Finally, we need a word about the difference between **estimate** and **estimator**. The former refers to the specific value obtained when applying an estimation procedure to a set of numbers. The latter refers to a rule or formula that produces the estimate. An estimate is a number or function, while an estimator is a random variable or a random function. Usually, both the words and the context will make the reference clear.

10.2.2   Measures of Quality

10.2.2.1   Introduction   There are a variety of ways to measure the quality of an estimator. Three of them are discussed here. Two examples are used throughout to illustrate them.

■   **EXAMPLE 10.1**

A population contains the values 1, 3, 5, and 9. We want to estimate the population mean by taking a sample of size 2 with replacement.   □

■ EXAMPLE 10.2

A population has an exponential distribution with a mean of θ. We want to estimate the population mean by taking a sample of size 3 with replacement.   □

Both examples are clearly artificial in that we know the answers prior to sampling (4.5 and θ). However, that knowledge will make apparent the error in the procedure we select. For practical applications, we need to be able to estimate the error when we do not know the true value of the quantity being estimated.

10.2.2.2 Unbiasedness   When constructing an estimator, it would be good if, on average, the errors we make were to cancel each other out. More formally, let θ be the quantity we want to estimate. Let $\hat{\theta}$ be the random variable that represents the estimator and let $E(\hat{\theta}|\theta)$ be the expected value of the estimator $\hat{\theta}$ when θ is the true parameter value.

Definition 10.1 *An estimator, $\hat{\theta}$, is **unbiased** if $E(\hat{\theta}|\theta) = \theta$ for all θ. The **bias** is* $\text{bias}_{\hat{\theta}}(\theta) = E(\hat{\theta}|\theta) - \theta$.

The bias depends on the estimator being used and may also depend on the particular value of θ.

■ EXAMPLE 10.3

For Example 10.1, determine the bias of the sample mean as an estimator of the population mean.

The population mean is $\theta = 4.5$. The sample mean is the average of the two observations. In all cases, we assume that sampling is random. In other words, every sample of size n has the same chance of being drawn. Such sampling also implies that any member of the population has the same chance of being observed as any other member. For this example, there are 16 equally likely ways in which the sample could have turned out:

1,1	1,3	1,5	1,9	3,1	3,3	3,5	3,9
5,1	5,3	5,5	5,9	9,1	9,3	9,5	9,9

These samples lead to the following 16 equally likely values for the sample mean:

1	2	3	5	2	3	4	6
3	4	5	7	5	6	7	9

Combining the common values, the sample mean, usually denoted $\bar{X}$, has the following probability distribution:

x	1	2	3	4	5	6	7	9
$p_{\bar{X}}(x)$	1/16	2/16	3/16	2/16	3/16	2/16	2/16	1/16

The expected value of the estimator is

$$E(\bar{X}) = [1(1) + 2(2) + 3(3) + 4(2) + 5(3) + 6(2) + 7(2) + 9(1)]/16 = 4.5,$$

and so the sample mean is an unbiased estimator of the population mean for this example.   □

■ **EXAMPLE 10.4**

For Example 10.2 determine the bias of the sample mean and the sample median as estimators of the population mean.

The sample mean is $\bar{X} = (X_1 + X_2 + X_3)/3$, where each X_j represents one of the observations from the exponential population. Its expected value is

$$E(\bar{X}) = E\left[\frac{1}{3}\left(X_1 + X_2 + X_3\right)\right] = \frac{1}{3}\left[E(X_1) + E(X_2) + E(X_3)\right]$$
$$= \frac{1}{3}(\theta + \theta + \theta) = \theta$$

and, therefore, the sample mean is an unbiased estimator of the population mean.

Investigating the sample median is a bit more difficult. The distribution function of the middle of three observations can be found as follows, using Y as the random variable of interest and X_j as the random variable for the jth observation from the population:

$$F_Y(y) = \Pr(Y \le y) = \Pr(X_1, X_2, X_3 \le y) + \Pr(X_1, X_2 \le y, X_3 > y)$$
$$+ \Pr(X_1, X_3 \le y, X_2 > y) + \Pr(X_2, X_3 \le y, X_1 > y)$$
$$= F_X(y)^3 + 3F_X(y)^2[1 - F_X(y)]$$
$$= [1 - e^{-y/\theta}]^3 + 3[1 - e^{-y/\theta}]^2 e^{-y/\theta}.$$

The first two lines follow because for the median to be less than or equal to y, either all three observations or exactly two of them must be less than or equal to y. The density function is

$$f_Y(y) = F_Y'(y) = \frac{6}{\theta}\left(e^{-2y/\theta} - e^{-3y/\theta}\right).$$

The expected value of this estimator is

$$E(Y|\theta) = \int_0^\infty y\frac{6}{\theta}\left(e^{-2y/\theta} - e^{-3y/\theta}\right) dy$$
$$= \frac{5\theta}{6}.$$

This estimator is clearly biased,[1] with

$$\text{bias}_Y(\theta) = 5\theta/6 - \theta = -\theta/6.$$

On average, this estimator underestimates the true value. It is also easy to see that the sample median can be turned into an unbiased estimator by multiplying it by 1.2. □

---

[1]The sample median is unlikely to be selected as an estimator of the population mean. This example studies it for comparison purposes. Because the population median is $\theta \ln 2$, the sample median is also biased for the population median.

For Example 10.2, we have two estimators (the sample mean and 1.2 times the sample median) that are both unbiased. We will need additional criteria to decide which one we prefer.

Some estimators exhibit a small amount of bias, which vanishes as the sample size goes to infinity.

Definition 10.2 *Let $\hat{\theta}_n$ be an estimator of θ based on a sample size of n. The estimator is* **asymptotically unbiased** *if*

$$\lim_{n\to\infty} \mathrm{E}(\hat{\theta}_n|\theta) = \theta$$

for all θ.

■ **EXAMPLE 10.5**

Suppose that a random variable has a uniform distribution on the interval $(0, \theta)$. Consider the estimator $\hat{\theta}_n = \max(X_1, \dots, X_n)$. Show that this estimator is asymptotically unbiased.

Let Y_n be the maximum from a sample of size n. Then,

$$F_{Y_n}(y) = \Pr(Y_n \le y) = \Pr(X_1 \le y, \dots, X_n \le y)$$
$$= [F_X(y)]^n$$
$$= (y/\theta)^n,$$
$$f_{Y_n}(y) = \frac{ny^{n-1}}{\theta^n}, \quad 0 < y < \theta.$$

The expected value is

$$\mathrm{E}(Y_n|\theta) = \int_0^\theta y(ny^{n-1}\theta^{-n})\,dy = \frac{n}{n+1}y^{n+1}\theta^{-n}\Big|_0^\theta = \frac{n\theta}{n+1}.$$

As $n \to \infty$, the limit is θ, showing that this estimator is asymptotically unbiased. □

A drawback to unbiasedness as a measure of the quality of an estimator is that an unbiased estimator may often not be very close to the parameter, as would be the case if the estimator has a large variance. We will now demonstrate that there is a limit to the accuracy of an unbiased estimator in general, in the sense that there is a lower bound (called the Cramèr–Rao lower bound) on its variance.

In what follows, suppose that $\mathbf{X} = (X_1, X_2, \dots, X_n)$ has joint pf or pdf $g(\mathbf{x}; \theta)$, where $\mathbf{x} = (x_1, x_2, \dots, x_n)$. In the i.i.d. special case $g(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta)$, where $f(x; \theta)$ is the common pf or pdf of the X_is. Of central importance in many discussions of parameter estimation is the **score function**, $U = \partial \ln g(\mathbf{X}; \theta)/\partial\theta$. We assume regularity conditions on g that will be discussed in detail later, but at this point we assume that g is twice differentiable with respect to θ and that the order of differentiation and expectation may be interchanged. In particular, this excludes situations in which an end point of the distribution depends on θ.

■ **EXAMPLE 10.6**

Determine the score function for the i.i.d. exponential case.

Let $f(x; \theta) = \theta^{-1} e^{-x/\theta}$. Then,

$$g(\mathbf{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = \theta^{-n} \exp\left(-\sum_{i=1}^{n} x_i/\theta\right),$$

and thus,

$$U = \frac{\partial}{\partial \theta}\left(-n \ln \theta - \frac{1}{\theta}\sum_{i=1}^{n} X_i\right) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i.$$

☐

As is clear from the above example, U is a random function of θ (i.e. U is a random variable and a function of θ).

In the i.i.d. special case, let $W_i = \partial \ln f(x_i; \theta)/\partial \theta$ for $i = 1, 2, \ldots, n$, implying that $W_1, W_2, \ldots, W_n$ are i.i.d. Then, $U = \sum_{i=1}^{n} W_i$.

We now turn to the evaluation of the mean of the score function. In the discrete case (the continuous case is similar),

$$\begin{aligned}
\mathrm{E}(U) &= \sum_{\text{all } \mathbf{x}} \left[\frac{\partial}{\partial \theta} \ln g(\mathbf{x}; \theta)\right] g(\mathbf{x}; \theta) \\
&= \sum_{\text{all } \mathbf{x}} \frac{\frac{\partial}{\partial \theta} g(\mathbf{x}; \theta)}{g(\mathbf{x}; \theta)} g(\mathbf{x}; \theta) \\
&= \sum_{\text{all } \mathbf{x}} \frac{\partial}{\partial \theta} g(\mathbf{x}; \theta) \\
&= \frac{\partial}{\partial \theta} \sum_{\text{all } \mathbf{x}} g(\mathbf{x}; \theta) \\
&= \frac{\partial}{\partial \theta} 1 = 0.
\end{aligned}$$

The last step follows because the sum of the probabilities over all possible values must be 1.

Also,

$$\frac{\partial^2}{\partial \theta^2} \ln g(\mathbf{x}; \theta) = \frac{\partial}{\partial \theta}\left[\frac{\partial}{\partial \theta} \ln g(\mathbf{x}; \theta)\right] = \frac{\partial}{\partial \theta}\left[\frac{\frac{\partial}{\partial \theta} g(\mathbf{x}; \theta)}{g(\mathbf{x}; \theta)}\right],$$

and so, by the quotient rule for differentiation,

$$\begin{aligned}
\frac{\partial^2}{\partial \theta^2} \ln g(\mathbf{x}; \theta) &= \frac{\frac{\partial^2}{\partial \theta^2} g(\mathbf{x}; \theta)}{g(\mathbf{x}; \theta)} - \left[\frac{\frac{\partial}{\partial \theta} g(\mathbf{x}; \theta)}{g(\mathbf{x}; \theta)}\right]^2 \\
&= \frac{\frac{\partial^2}{\partial \theta^2} g(\mathbf{x}; \theta)}{g(\mathbf{x}; \theta)} - \left[\frac{\partial}{\partial \theta} \ln g(\mathbf{x}; \theta)\right]^2.
\end{aligned}$$

Taking expectations yields

$$
\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln g(\mathbf{X};\theta)\right] = \mathrm{E}\left[\frac{\frac{\partial^2}{\partial\theta^2}g(\mathbf{X};\theta)}{g(\mathbf{X};\theta)}\right] - \mathrm{E}(U^2)
$$

$$
= \sum_{\text{all }\mathbf{x}} \frac{\frac{\partial^2}{\partial\theta^2}g(\mathbf{x};\theta)}{g(\mathbf{x};\theta)} g(\mathbf{x};\theta) - \mathrm{E}(U^2)
$$

$$
= \sum_{\text{all }\mathbf{x}} \frac{\partial^2}{\partial\theta^2}g(\mathbf{x};\theta) - \mathrm{E}(U^2)
$$

$$
= \frac{\partial^2}{\partial\theta^2}\left[\sum_{\text{all }\mathbf{x}} g(\mathbf{x};\theta)\right] - \mathrm{E}(U^2).
$$

The first term is on the right-hand side is zero and therefore

$$
\mathrm{E}(U^2) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln g(\mathbf{X};\theta)\right].
$$

Alternatively, using the definition of U, we have

$$
\mathrm{E}(U^2) = \mathrm{E}\left\{\left[\frac{\partial}{\partial\theta}\ln g(\mathbf{X};\theta)\right]^2\right\}.
$$

Recall that $\mathrm{E}(U) = 0$. Then,

$$
\mathrm{Var}(U) = \mathrm{E}(U^2) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln g(\mathbf{X};\theta)\right] = \mathrm{E}\left\{\left[\frac{\partial}{\partial\theta}\ln g(\mathbf{X};\theta)\right]^2\right\}.
$$

Before proceeding, we digress to note that, for any two random variables Z_1 and Z_2,

$$
\mathrm{Cov}(Z_1, Z_2) \le \sqrt{\mathrm{Var}(Z_1)}\sqrt{\mathrm{Var}(Z_2)}.
$$

To see that this is true, let $\sigma_1^2 = \mathrm{Var}(Z_1)$, $\sigma_2^2 = \mathrm{Var}(Z_2)$, $\sigma_{12} = \mathrm{Cov}(Z_1, Z_2)$, and $\rho = \sigma_{12}/(\sigma_1\sigma_2)$. Then,

$$
0 \le \mathrm{Var}\left(\frac{Z_1}{\sigma_1} - \rho\frac{Z_2}{\sigma_2}\right)
$$

$$
= \frac{1}{\sigma_1^2}\mathrm{Var}(Z_1) + \left(\frac{\rho}{\sigma_2}\right)^2 \mathrm{Var}(Z_2) - \frac{2\rho}{\sigma_1\sigma_2}\mathrm{Cov}(Z_1, Z_2)
$$

$$
= 1 + \rho^2 - 2\rho^2 = 1 - \rho^2.
$$

Note that this development also proves that $-1 \le \rho \le 1$.

Now let $T = T(\mathbf{X})$ be an unbiased estimator of θ. Then, by the definition of unbiasedness,

$$
\theta = \mathrm{E}(T) = \sum_{\text{all }\mathbf{x}} T(\mathbf{x})g(\mathbf{x};\theta)
$$

and differentiating with respect to θ yields (recalling our assumption that the order of differentiation and summation/integration may be interchanged)

$$
\begin{aligned}
1 &= \sum_{\text{all } \mathbf{x}} T(\mathbf{x}) \frac{\partial}{\partial \theta} g(\mathbf{x}; \theta) \\
&= \sum_{\text{all } \mathbf{x}} T(\mathbf{x}) \left[\frac{\partial}{\partial \theta} \ln g(\mathbf{x}; \theta) \right] g(\mathbf{x}; \theta) \\
&= \mathrm{E}(TU).
\end{aligned}
$$

Then,

$$
\mathrm{Cov}(T, U) = \mathrm{E}(TU) - \mathrm{E}(T)\mathrm{E}(U) = 1 - (\theta)(0) = 1.
$$

We next have

$$
1 = \mathrm{Cov}(T, U) \le \sqrt{\mathrm{Var}(T)} \sqrt{\mathrm{Var}(U)}.
$$

This implies that

$$
\mathrm{Var}(T) \ge \frac{1}{\mathrm{Var}(U)} = \frac{1}{-\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \ln g(\mathbf{X}; \theta) \right]} = \frac{1}{\mathrm{E}\left\{ \left[\frac{\partial}{\partial \theta} \ln g(\mathbf{X}; \theta) \right]^2 \right\}}. \tag{10.1}
$$

In the i.i.d. case, $\mathrm{Var}(U) = \sum_{i=1}^{n} \mathrm{Var}(W_i) = n\mathrm{Var}(W)$, where $W = \frac{\partial}{\partial \theta} \ln f(X; \theta)$ and X is a generic version of the X_is. Then, (10.1) becomes

$$
\mathrm{Var}(T) \ge \frac{1}{\mathrm{Var}(U)} = \frac{1}{-n\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \ln f(X; \theta) \right]} = \frac{1}{n\mathrm{E}\left\{ \left[\frac{\partial}{\partial \theta} \ln f(X; \theta) \right]^2 \right\}}. \tag{10.2}
$$

Generally, the version using second partial derivatives (rather than the square of the first derivative) is easier to calculate.

The lower bounds (10.1) and (10.2) are often referred to as **Cramèr–Rao lower bounds** for the variance of unbiased estimators. This is extremely valuable for maximum likelihood and other estimation procedures. The denominators in each case are referred to as the **Fisher** or **expected** information.

■ **EXAMPLE 10.7**

Determine the Cramèr–Rao lower bound for an unbiased estimator of the sample mean from a population with an exponential distribution. Use both formulas from (10.2).

For the exponential distribution,

$$f(X; \theta) = \theta^{-1} e^{-X/\theta},$$
$$\ln f(X : \theta) = -\ln \theta - X/\theta,$$
$$\frac{\partial}{\partial \theta} \ln f(X : \theta) = -\theta^{-1} + X\theta^{-2},$$
$$\frac{\partial^2}{\partial \theta^2} \ln f(X : \theta) = \theta^{-2} - 2X\theta^{-3},$$
$$\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \ln f(X; \theta)\right] = \mathrm{E}\left(\theta^{-2} - 2X\theta^{-3}\right) = \theta^{-2} - 2\theta^{-2} = -\theta^{-2},$$
$$\mathrm{E}\left\{\left[\frac{\partial}{\partial \theta} \ln f(X; \theta)\right]^2\right\} = \mathrm{E}\left[\left(-\theta^{-1} + X\theta^{-2}\right)^2\right] = \mathrm{E}\left(\theta^{-2} - 2X\theta^{-3} + X^2\theta^{-4}\right)$$
$$= \theta^{-2} - 2\theta^{-2} + 2\theta^{-2} = \theta^{-2}.$$

The lower bound is then $1/(n\theta^{-2}) = \theta^2/n$. $\qquad\square$

10.2.2.3 *Consistency* Another desirable property of an estimator is that it works well for extremely large samples. Slightly more formally, as the sample size goes to infinity, the probability that the estimator is in error by more than a small amount goes to zero. A formal definition follows.

Definition 10.3 *An estimator is **consistent** (often called, in this context, **weakly consistent**) if, for all $\delta > 0$ and any θ,*

$$\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| > \delta) = 0.$$

A sufficient (although not necessary) condition for weak consistency is that the estimator be asymptotically unbiased and $\mathrm{Var}(\hat{\theta}_n) \to 0$ [equivalently, from (10.3), the mean squared error goes to zero as $n \to \infty$].

■ **EXAMPLE 10.8**

Prove that, if the variance of a random variable is finite, the sample mean is a consistent estimator of the population mean.

From Exercise 10.2, the sample mean is unbiased. In addition,

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)$$
$$= \frac{1}{n^2}\sum_{j=1}^{n} \mathrm{Var}(X_j)$$
$$= \frac{\mathrm{Var}(X)}{n} \to 0.$$

The second step follows from assuming that the observations are independent. $\qquad\square$

■ **EXAMPLE 10.9**

Show that the maximum observation from a uniform distribution on the interval $(0, \theta)$ is a consistent estimator of θ.

From Example 10.5, the maximum is asymptotically unbiased. The second moment is

$$E(Y_n^2) = \int_0^\theta y^2 (ny^{n-1}\theta^{-n})\, dy = \left. \frac{n}{n+2} y^{n+2}\theta^{-n} \right|_0^\theta = \frac{n\theta^2}{n+2},$$

and then

$$\text{Var}(Y_n) = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1} \right)^2 = \frac{n\theta^2}{(n+2)(n+1)^2} \to 0.$$

☐

10.2.2.4 Mean Squared Error While consistency is nice, most estimators have this property. What would be truly impressive is an estimator that is not only correct on average but comes very close most of the time and, in particular, comes closer than rival estimators. One measure for a finite sample is motivated by the definition of consistency. The quality of an estimator could be measured by the probability that it gets within δ of the true value – that is, by measuring $\Pr(|\hat{\theta}_n - \theta| < \delta)$. But the choice of δ is arbitrary, and we prefer measures that cannot be altered to suit the investigator's whim. Then we might consider $E(|\hat{\theta}_n - \theta|)$, the average absolute error. But we know that working with absolute values often presents unpleasant mathematical challenges, and so the following has become widely accepted as a measure of accuracy.

Definition 10.4 *The **mean squared error** (**MSE**) of an estimator is*

$$\text{MSE}_{\hat{\theta}}(\theta) = E[(\hat{\theta} - \theta)^2 | \theta].$$

Note that the MSE is a function of the true value of the parameter. An estimator may perform extremely well for some values of the parameter but poorly for others.

■ **EXAMPLE 10.10**

Consider the estimator $\hat{\theta} = 5$ of an unknown parameter θ. The MSE is $(5 - \theta)^2$, which is very small when θ is near 5 but becomes poor for other values. Of course, this estimate is both biased and inconsistent unless θ is exactly equal to 5. ☐

A result that follows directly from the various definitions is

$$\text{MSE}_{\hat{\theta}}(\theta) = E\{[\hat{\theta} - E(\hat{\theta}|\theta) + E(\hat{\theta}|\theta) - \theta]^2 | \theta\} = \text{Var}(\hat{\theta}|\theta) + [\text{bias}_{\hat{\theta}}(\theta)]^2. \qquad (10.3)$$

If we restrict attention to only unbiased estimators, the best such estimator could be defined as follows.

Definition 10.5 *An estimator $\hat{\theta}$ is called a **uniformly minimum variance unbiased estimator** (**UMVUE**) if it is unbiased and, for any true value of θ, there is no other unbiased estimator that has a smaller variance.*

Because we are looking only at unbiased estimators, it would have been equally effective to formulate the definition in terms of MSE. We could also generalize the definition by looking for estimators that are uniformly best with regard to MSE, but the previous example indicates why that is not feasible. There are some results that can often assist with the determination of UMVUEs (e.g. Hogg et al. [56, ch. 7]). However, such estimators are often difficult to determine. Nevertheless, MSE is still a useful criterion for comparing two alternative estimators.

■ **EXAMPLE 10.11**

For Example 10.2, compare the MSEs of the sample mean and 1.2 times the sample median. Demonstrate that for the exponential distribution, the sample mean is a UMVUE.

The sample mean has variance

$$\frac{\text{Var}(X)}{3} = \frac{\theta^2}{3}.$$

When multiplied by 1.2, the sample median has second moment

$$E[(1.2Y)^2] = 1.44 \int_0^\infty y^2 \frac{6}{\theta} \left(e^{-2y/\theta} - e^{-3y/\theta}\right) dy$$

$$= 1.44 \frac{6}{\theta} \left[y^2 \left(\frac{-\theta}{2} e^{-2y/\theta} + \frac{\theta}{3} e^{-3y/\theta} \right) \right.$$

$$- 2y \left(\frac{\theta^2}{4} e^{-2y/\theta} - \frac{\theta^2}{9} e^{-3y/\theta} \right)$$

$$\left. + 2 \left(\frac{-\theta^3}{8} e^{-2y/\theta} + \frac{\theta^3}{27} e^{-3y/\theta} \right) \right] \Big|_0^\infty$$

$$= \frac{8.64}{\theta} \left(\frac{2\theta^3}{8} - \frac{2\theta^3}{27} \right) = \frac{38\theta^2}{25}$$

for a variance of

$$\frac{38\theta^2}{25} - \theta^2 = \frac{13\theta^2}{25} > \frac{\theta^2}{3}.$$

The sample mean has the smaller MSE regardless of the true value of θ. Therefore, for this problem, it is a superior estimator of θ.

From Example 10.7, we see that the minimum possible variance for an unbiased estimator is, for a sample of size 3, $\theta^2/3$. This matches the variance of the sample mean and therefore no other unbiased estimator can have a smaller variance. Hence the sample mean is a UMVUE. □

■ **EXAMPLE 10.12**

For the uniform distribution on the interval $(0, \theta)$, compare the MSE of the estimators $2\bar{X}$ and $[(n + 1)/n] \max(X_1, \ldots, X_n)$. Also evaluate the MSE of $\max(X_1, \ldots, X_n)$.

The first two estimators are unbiased, so it is sufficient to compare their variances. For twice the sample mean,

$$\text{Var}(2\bar{X}) = \frac{4}{n} \text{Var}(X) = \frac{4\theta^2}{12n} = \frac{\theta^2}{3n}.$$

For the adjusted maximum, the second moment is

$$E\left[\left(\frac{n+1}{n}Y_n\right)^2\right] = \frac{(n+1)^2}{n^2}\frac{n\theta^2}{n+2} = \frac{(n+1)^2\theta^2}{(n+2)n}$$

for a variance of

$$\frac{(n+1)^2\theta^2}{(n+2)n} - \theta^2 = \frac{\theta^2}{n(n+2)}.$$

Except for the case $n = 1$ (in which the two estimators are identical), the one based on the maximum has the smaller MSE. The third estimator is biased. The MSE of the third estimator is

$$\frac{n\theta^2}{(n+2)(n+1)^2} + \left(\frac{n\theta}{n+1} - \theta\right)^2 = \frac{2\theta^2}{(n+1)(n+2)},$$

which is also larger than that for the adjusted maximum. □

For this example, the regularity conditions underlying the derivation of the Cramèr–Rao lower bound do not hold and so (10.2) cannot be used to set a minimum possible value.

10.2.3   Exercises

10.1 For Example 10.1, show that the mean of three observations drawn without replacement is an unbiased estimator of the population mean, while the median of three observations drawn without replacement is a biased estimator of the population mean.

10.2 Prove that, for random samples, the sample mean is always an unbiased estimator of the population mean.

10.3 Let X have the uniform distribution over the range $(\theta - 2, \theta + 2)$. That is, $f_X(x) = 0.25$, $\theta - 2 < x < \theta + 2$. Show that the median from a sample of size 3 is an unbiased estimator of θ.

10.4 Explain why the sample mean may not be a consistent estimator of the population mean for a Pareto distribution.

10.5 For the sample of size 3 in Exercise 10.3, compare the MSE of the sample mean and median as estimates of θ.

10.6 (*) You are given two independent estimators of an unknown quantity θ. For estimator A, $E(\hat{\theta}_A) = 1{,}000$ and $Var(\hat{\theta}_A) = 160{,}000$, while for estimator B, $E(\hat{\theta}_B) = 1{,}200$ and $Var(\hat{\theta}_B) = 40{,}000$. Estimator C is a weighted average, $\hat{\theta}_C = w\hat{\theta}_A + (1 - w)\hat{\theta}_B$. Determine the value of w that minimizes $Var(\hat{\theta}_C)$.

10.7 (*) A population of losses has a Pareto distribution (see Appendix A) with $\theta = 6{,}000$ and α unknown. Simulation of the results from maximum likelihood estimation based on samples of size 10 has indicated that $E(\hat{\alpha}) = 2.2$ and $MSE(\hat{\alpha}) = 1$. Determine $Var(\hat{\alpha})$ if it is known that $\alpha = 2$.

10.8 (*) Two instruments are available for measuring a particular nonzero distance. The random variable X represents a measurement with the first instrument and the random variable Y one with the second instrument. Assume that X and Y are independent with $E(X) = 0.8m$, $E(Y) = m$, $Var(X) = m^2$, and $Var(Y) = 1.5m^2$, where m is the true distance. Consider estimators of m that are of the form $Z = \alpha X + \beta Y$. Determine the values of α and β that make Z a UMVUE within the class of estimators of this form.

10.9 A population contains six members, with values 1, 1, 2, 3, 5, and 10. A random sample of size 3 is drawn without replacement. In each case, the objective is to estimate the population mean. *Note:* The use of a spreadsheet with an optimization routine may be the best way to solve this problem.

    (a) Determine the bias, variance, and MSE of the sample mean.

    (b) Determine the bias, variance, and MSE of the sample median.

    (c) Determine the bias, variance, and MSE of the sample midrange (the average of the largest and smallest observations).

    (d) Consider an arbitrary estimator of the form $aX_{(1)} + bX_{(2)} + cX_{(3)}$, where $X_{(1)} \leq X_{(2)} \leq X_{(3)}$ are the sample order statistics.

        i. Determine a restriction on the values of a, b, and c that will assure that the estimator is unbiased.

        ii. Determine the values of a, b, and c that will produce the unbiased estimator with the smallest variance.

        iii. Determine the values of a, b, and c that will produce the (possibly biased) estimator with the smallest MSE.

10.10 (*) Two different estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, are being considered. To test their performance, 75 trials have been simulated, each with the true value set at $\theta = 2$. The following totals have been obtained:

$$\sum_{j=1}^{75} \hat{\theta}_{1j} = 165, \quad \sum_{j=1}^{75} \hat{\theta}_{1j}^2 = 375, \quad \sum_{j=1}^{75} \hat{\theta}_{2j} = 147, \quad \sum_{j=1}^{75} \hat{\theta}_{2j}^2 = 312,$$

where $\hat{\theta}_{ij}$ is the estimate based on the jth simulation using estimator $\hat{\theta}_i$. Estimate the MSE for each estimator and determine the **relative efficiency** (the ratio of the MSEs).

10.11 Consider an i.i.d. random sample $X_1, X_2, \ldots, X_n$ from the distribution with cdf $F(x) = 1 - (\theta/x)^3$, for $x > \theta$ and zero otherwise.

    (a) Let $\hat{\theta}_1 = \min\{X_1, X_2, \ldots, X_n\}$ be an estimator of θ. Demonstrate that

$$E(\hat{\theta}_1) = \frac{3n}{3n - 1}\theta.$$

    (b) Determine the mean squared error of $\hat{\theta}_1$, and demonstrate that it may be expressed in the form θ^2/k_n, where k_n is a positive integer. Also show that $\hat{\theta}_1$ is a consistent estimator of θ.

    (c) Let $\hat{\theta}_2 = \frac{3n-1}{3n}\hat{\theta}_1$. Prove that $\hat{\theta}_2$ is an unbiased and consistent estimator of θ.

    (d) Let $\hat{\theta}_3 = (2/3)\bar{X}$, where $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$. Derive the expected value and mean squared error of $\hat{\theta}_3$, and explain why $\hat{\theta}_2$ would generally be preferred to both $\hat{\theta}_1$ and $\hat{\theta}_3$ as an estimator of θ.

10.3   Interval Estimation

All of the estimators discussed to this point have been **point estimators**. That is, the estimation process produces a single value that represents our best attempt to determine the value of the unknown population quantity. While that value may be a good one, we do not expect it to match the true value exactly. A more useful statement is often provided by an **interval estimator**. Instead of a single value, the result of the estimation process is a range of possible numbers, any of which is likely to be the true value. A specific type of interval estimator is the confidence interval.

Definition 10.6 *A* $100(1 - \alpha)\%$ ***confidence interval*** *for a parameter θ is a pair of random values, L and U, computed from a random sample such that* $\Pr(L \le \theta \le U) \ge 1 - \alpha$ *for all θ.*

    Note that this definition does not uniquely specify the interval. Because the definition is a probability statement and must hold for all θ, it says nothing about whether or not a particular interval encloses the true value of θ from a particular population. Instead, the **level of confidence**, $1 - \alpha$, is a property of the method used to obtain L and U and not of the particular values obtained. The proper interpretation is that, if we use a particular interval estimator over and over on a variety of samples, at least $100(1 - \alpha)\%$ of the time our interval will enclose the true value. Keep in mind that it is the interval end points that are random.

    The construction of confidence intervals is usually very difficult. For example, we know that, if a population has a normal distribution with unknown mean and variance, a $100(1 - \alpha)\%$ confidence interval for the mean uses

$$L = \bar{X} - t_{\alpha/2,n-1}s/\sqrt{n}, \quad U = \bar{X} + t_{\alpha/2,n-1}s/\sqrt{n}, \tag{10.4}$$

where $s = \sqrt{\sum_{j=1}^{n}(X_j - \bar{X})^2/(n - 1)}$ and $t_{\alpha/2,b}$ is the $100(1 - \alpha/2)$th percentile of the t distribution with b degrees of freedom. But it takes a great deal of effort to verify that (10.4) is correct (see, e.g. Hogg et al. [56, p. 186]).

    However, there is a method for constructing approximate confidence intervals that is often accessible. Suppose that we have a point estimator $\hat{\theta}$ of parameter θ such that $E(\hat{\theta}) \doteq \theta$, $Var(\hat{\theta}) \doteq v(\theta)$, and $\hat{\theta}$ has approximately a normal distribution. Theorem 11.4 shows that these three properties are often the case. With all these approximations, we have that, approximately,

$$1 - \alpha \doteq \Pr\left(-z_{\alpha/2} \le \frac{\hat{\theta} - \theta}{\sqrt{v(\theta)}} \le z_{\alpha/2}\right), \tag{10.5}$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$th percentile of the standard normal distribution. Solving for θ produces the desired interval. It is sometimes difficult to obtain the solution (due to

the appearance of θ in the denominator) and so, if necessary, replace $v(\theta)$ in (10.5) with $v(\hat{\theta})$ to obtain a further approximation:

$$1 - \alpha \doteq \Pr\left(\hat{\theta} - z_{\alpha/2}\sqrt{v(\hat{\theta})} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sqrt{v(\hat{\theta})}\right). \tag{10.6}$$

■ **EXAMPLE 10.13**

Use (10.6) to construct an approximate 95% confidence interval for the mean of a normal population with unknown variance.

Use $\hat{\theta} = \bar{X}$ and then note that $E(\hat{\theta}) = \theta$, $Var(\hat{\theta}) = \sigma^2/n$, and $\hat{\theta}$ does have a normal distribution. The confidence interval is then $\bar{X} \pm 1.96s/\sqrt{n}$. Because $t_{0.025,n-1} > 1.96$, this approximate interval must be narrower than the exact interval given by (10.4), which implies that our level of confidence is something less than 95%. □

■ **EXAMPLE 10.14**

Use (10.5) and (10.6) to construct approximate 95% confidence intervals for the mean of a Poisson distribution. Obtain intervals for the particular case where $n = 25$ and $\bar{x} = 0.12$.

Here, $\theta = \lambda$, the mean of the Poisson distribution. Let $\hat{\theta} = \bar{X}$, the sample mean. For the Poisson distribution, $E(\hat{\theta}) = E(X) = \theta$ and $v(\theta) = Var(\bar{X}) = Var(X)/n = \theta/n$. For the first interval,

$$0.95 \doteq \Pr\left(-1.96 \leq \frac{\bar{X} - \theta}{\sqrt{\theta/n}} \leq 1.96\right)$$

is true if and only if

$$|\bar{X} - \theta| \leq 1.96\sqrt{\frac{\theta}{n}},$$

which is equivalent to

$$(\bar{X} - \theta)^2 \leq \frac{3.8416\theta}{n}$$

or

$$\theta^2 - \theta\left(2\bar{X} + \frac{3.8416}{n}\right) + \bar{X}^2 \leq 0.$$

Solving the quadratic produces the interval

$$\bar{X} + \frac{1.9208}{n} \pm \frac{1}{2}\sqrt{\frac{15.3664\bar{X} + 3.8416^2/n}{n}},$$

and for this problem the interval is 0.197 ± 0.156.

For the second approximation, the interval is $\bar{X} \pm 1.96\sqrt{\bar{X}/n}$, and for the example, it is 0.12 ± 0.136. This interval extends below zero (which is not possible for the true value of θ) because (10.6) is too crude an approximation in this case. □

10.3.1   Exercises

10.12 Let $x_1, \ldots, x_n$ be a random sample from a population with pdf $f(x) = \theta^{-1} e^{-x/\theta}$, $x > 0$. This exponential distribution has a mean of θ and a variance of θ^2. Consider the sample mean, $\bar{X}$, as an estimator of θ. It turns out that $\bar{X}/\theta$ has a gamma distribution with $\alpha = n$ and $\theta = 1/n$, where in the second expression the "θ" on the left is the parameter of the gamma distribution. For a sample of size 50 and a sample mean of 275, develop 95% confidence intervals by each of the following methods. In each case, if the formula requires the true value of θ, substitute the estimated value.

> (a) Use the gamma distribution to determine an exact interval.
> (b) Use a normal approximation, estimating the variance prior to solving the inequal-ities as in (10.6).
> (c) Use a normal approximation, estimating θ after solving the inequalities as in Example 10.14.

10.13 (*) A sample of 2,000 policies had 1,600 with no claims and 400 with one or more claims. Using the normal approximation, determine the symmetric 95% confidence interval for the probability that a single policy has one or more claims.

10.4   The Construction of Parametric Estimators

In previous sections, we developed methods for assessing the quality of an estimator. In all the examples, the estimators being evaluated were arbitrary, though reasonable. This section reviews two methods for constructing estimators. A third is covered in Chapter 11. In this section, we assume that n independent observations from the same parametric distribution have been collected. There are two, essentially incompatible, approaches to estimating parameters. This section and Chapter 11 cover the frequentist approach to estimation introduced in Section 10.2. An alternative estimation approach, known as Bayesian estimation, is covered in Chapter 13.

    The methods introduced in Section 10.4.1 are relatively easy to implement but tend to give poor results. Chapter 11 covers maximum likelihood estimation. This method is more difficult to use but has superior statistical properties and is considerably more flexible.

10.4.1   The Method of Moments and Percentile Matching

Let the distribution function for an individual observation be given by

$$F(x|\theta), \quad \theta^T = (\theta_1, \theta_2, \ldots, \theta_p),$$

where θ^T is the transpose of θ. That is, θ is a column vector containing the p parameters to be estimated. Furthermore, let $\mu'_k(\theta) = E(X^k|\theta)$ be the kth raw moment, and let $\pi_g(\theta)$ be the $100g$th percentile of the random variable. That is, $F[\pi_g(\theta)|\theta] = g$. If the distribution function is continuous, there will be at least one solution to that equation.

    For a sample of n independent observations from this random variable, let $\hat{\mu}'_k = \frac{1}{n}\sum_{j=1}^n x_j^k$ be the empirical estimate of the kth moment and let $\hat{\pi}_g$ be the empirical estimate of the $100g$th percentile.

Definition 10.7 *A **method of moments estimate** of* θ *is any solution of the p equations*

$$\mu_k'(\theta) = \hat{\mu}_k', \ k = 1, 2, \ldots, p.$$

The motivation for this estimator is that it produces a model that has the same first p raw moments as the data (as represented by the empirical distribution). The traditional definition of the method of moments uses positive integers for the moments. Arbitrary negative or fractional moments could also be used. In particular, when estimating parameters for inverse distributions, the matching of negative moments may be a superior approach.[2]

■ **EXAMPLE 10.15**

Use the method of moments to estimate parameters for the exponential, gamma, and Pareto distributions for Data Set B.

The first two sample moments are

$$\hat{\mu}_1' = \tfrac{1}{20}(27 + \cdots + 15{,}743) = 1{,}424.4,$$
$$\hat{\mu}_2' = \tfrac{1}{20}(27^2 + \cdots + 15{,}743^2) = 13{,}238{,}441.9.$$

For the exponential distribution, the equation is

$$\theta = 1{,}424.4,$$

with the obvious solution $\hat{\theta} = 1{,}424.4$.

For the gamma distribution, the two equations are

$$E(X) = \alpha\theta = 1{,}424.4,$$
$$E(X^2) = \alpha(\alpha + 1)\theta^2 = 13{,}238{,}441.9.$$

Dividing the second equation by the square of the first equation yields

$$\frac{\alpha + 1}{\alpha} = 6.52489, \quad 1 = 5.52489\alpha,$$

and so $\hat{\alpha} = 1/5.52489 = 0.18100$ and $\hat{\theta} = 1{,}424.4/0.18100 = 7{,}869.61$.

For the Pareto distribution, the two equations are

$$E(X) = \frac{\theta}{\alpha - 1} = 1{,}424.4,$$
$$E(X^2) = \frac{2\theta^2}{(\alpha - 1)(\alpha - 2)} = 13{,}238{,}441.9.$$

Dividing the second equation by the square of the first equation yields

$$\frac{2(\alpha - 1)}{\alpha - 2} = 6.52489,$$

[2]One advantage is that, with appropriate moments selected, the equations may have a solution within the range of allowable parameter values.

with a solution of $\hat{\alpha} = 2.442$ and then $\hat{\theta} = 1{,}424.4(1.442) = 2{,}053.985$. □

There is no guarantee that the equations will have a solution or, if there is a solution, that it will be unique.

Definition 10.8 *A **percentile-matching estimate** of θ is any solution of the p equations*

$$\pi_{g_k}(\theta) = \hat{\pi}_{g_k}, \ k = 1, 2, \dots, p,$$

where $g_1, g_2, \dots, g_p$ are p arbitrarily chosen percentiles. From the definition of percentile, the equations can also be written as

$$F(\hat{\pi}_{g_k}|\theta) = g_k, \ k = 1, 2, \dots, p.$$

The motivation for this estimator is that it produces a model with p percentiles that match the data (as represented by the empirical distribution). As with the method of moments, there is no guarantee that the equations will have a solution or, if there is a solution, that it will be unique. One problem with this definition is that percentiles for discrete random variables (such as the empirical distribution) are not always well defined. For example, Data Set B has 20 observations. Any number between 384 and 457 has 10 observations below and 10 above, and so could serve as the median. The convention is to use the midpoint. However, for other percentiles, there is no "official" interpolation scheme.[3] The following definition is used here.

Definition 10.9 *The **smoothed empirical estimate** of a percentile is calculated as*

$$\hat{\pi}_g = (1 - h)x_{(j)} + hx_{(j+1)},$$

where

$$j = \lfloor (n+1)g \rfloor \quad and \quad h = (n+1)g - j.$$

Here, $\lfloor \cdot \rfloor$ indicates the greatest integer function and $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the order statistics from the sample.

Unless there are two or more data points with the same value, no two percentiles will have the same value. One feature of this definition is that $\hat{\pi}_g$ cannot be obtained for $g < 1/(n+1)$ or $g > n/(n+1)$. This seems reasonable, as we should not expect to be able to infer the value of very large or small percentiles from small samples. We use the smoothed version whenever an empirical percentile estimate is needed.

■ **EXAMPLE 10.16**

Use percentile matching to estimate parameters for the exponential and Pareto distributions for Data Set B.

---

[3]Hyndman and Fan [60] present nine different methods. They recommend a slight modification of the one presented here, using $j = \lfloor g(n + \frac{1}{3}) + \frac{1}{3} \rfloor$ and $h = g(n + \frac{1}{3}) + \frac{1}{3} - j$.

For the exponential distribution, select the 50th percentile. The empirical estimate is the traditional median of $\hat{\pi}_{0.5} = (384 + 457)/2 = 420.5$ and the equation to solve is

$$0.5 = F(420.5|\theta) = 1 - e^{-420.5/\theta},$$

$$\ln 0.5 = \frac{-420.5}{\theta},$$

$$\hat{\theta} = \frac{-420.5}{\ln 0.5} = 606.65.$$

For the Pareto distribution, select the 30th and 80th percentiles. The smoothed empirical estimates are found as follows:

$$
\begin{aligned}
\text{30th:} \quad j &= \lfloor 21(0.3) \rfloor = \lfloor 6.3 \rfloor = 6, \, h = 6.3 - 6 = 0.3, \\
\hat{\pi}_{0.3} &= 0.7(161) + 0.3(243) = 185.6, \\
\text{80th:} \quad j &= \lfloor 21(0.8) \rfloor = \lfloor 16.8 \rfloor = 16, \, h = 16.8 - 16 = 0.8, \\
\hat{\pi}_{0.8} &= 0.2(1{,}193) + 0.8(1{,}340) = 1{,}310.6.
\end{aligned}
$$

The equations to solve are

$$0.3 = F(185.6) = 1 - \left(\frac{\theta}{185.6 + \theta} \right)^{\alpha},$$

$$0.8 = F(1{,}310.6) = 1 - \left(\frac{\theta}{1{,}310.6 + \theta} \right)^{\alpha},$$

$$\ln 0.7 = -0.356675 = \alpha \ln \left(\frac{\theta}{185.6 + \theta} \right),$$

$$\ln 0.2 = -1.609438 = \alpha \ln \left(\frac{\theta}{1{,}310.6 + \theta} \right),$$

$$\frac{-1.609438}{-0.356675} = 4.512338 = \frac{\ln \left(\dfrac{\theta}{1{,}310.6 + \theta} \right)}{\ln \left(\dfrac{\theta}{185.6 + \theta} \right)}.$$

Numerical methods are needed to solve this equation for $\hat{\theta} = 715.03$. Then, from the first equation,

$$0.3 = 1 - \left(\frac{715.03}{185.6 + 715.03} \right)^{\alpha},$$

which yields $\hat{\alpha} = 1.54559$.   □

The estimates are much different from those obtained in Example 10.15, which is one indication that these methods may not be particularly reliable.

10.4.2 Exercises

10.14 Determine the method of moments estimate for a lognormal model for Data Set B.

10.15 (*) The 20th and 80th percentiles from a sample are 5 and 12, respectively. Using the percentile-matching method, estimate $S(8)$ assuming that the population has a Weibull distribution.

10.16 (*) From a sample, you are given that the mean is 35,000, the standard deviation is 75,000, the median is 10,000, and the 90th percentile is 100,000. Using the percentile-matching method, estimate the parameters of a Weibull distribution.

10.17 (*) A sample of size 5 has produced the values 4, 5, 21, 99, and 421. You fit a Pareto distribution using the method of moments. Determine the 95th percentile of the fitted distribution.

10.18 (*) In year 1 there are 100 claims with an average size of 10,000 and in year 2 there are 200 claims with an average size of 12,500. Inflation increases the size of all claims by 10% per year. A Pareto distribution with $\alpha = 3$ and θ unknown is used to model the claim size distribution. Estimate θ for year 3 using the method of moments.

10.19 (*) From a random sample, the 20th percentile is 18.25 and the 80th percentile is 35.8. Estimate the parameters of a lognormal distribution using percentile matching and then use these estimates to estimate the probability of observing a value in excess of 30.

10.20 (*) A claim process is a mixture of two random variables A and B, where A has an exponential distribution with a mean of 1 and B has an exponential distribution with a mean of 10. A weight of p is assigned to distribution A and $1 - p$ to distribution B. The standard deviation of the mixture is 2. Estimate p by the method of moments.

10.21 (*) A random sample of 20 observations has been ordered as follows:

$$12 \quad 16 \quad 20 \quad 23 \quad 26 \quad 28 \quad 30 \quad 32 \quad 33 \quad 35$$
$$36 \quad 38 \quad 39 \quad 40 \quad 41 \quad 43 \quad 45 \quad 47 \quad 50 \quad 57$$

Determine the 60th sample percentile using the smoothed empirical estimate.

10.22 (*) The following 20 wind losses (in millions of dollars) were recorded in one year:

$$1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 3 \quad 3 \quad 4$$
$$6 \quad 6 \quad 8 \quad 10 \quad 13 \quad 14 \quad 15 \quad 18 \quad 22 \quad 25$$

Determine the sample 75th percentile using the smoothed empirical estimate.

10.23 (*) The observations 1,000, 850, 750, 1,100, 1,250, and 900 were obtained as a random sample from a gamma distribution with unknown parameters α and θ. Estimate these parameters by the method of moments.

10.24 (*) A random sample of claims has been drawn from a loglogistic distribution. In the sample, 80% of the claims exceed 100 and 20% exceed 400. Estimate the loglogistic parameters by percentile matching.

10.25 (*) Let $x_1, \ldots, x_n$ be a random sample from a population with cdf $F(x) = x^p$, $0 < x < 1$. Determine the method of moments estimate of p.

Table 10.5   The data for Exercise 10.30.

Number of claims	Number of policies
0	9,048
1	905
2	45
3	2
4+	0

10.26 (*) A random sample of 10 claims obtained from a gamma distribution is given as follows:

$$1,500 \quad 6,000 \quad 3,500 \quad 3,800 \quad 1,800 \quad 5,500 \quad 4,800 \quad 4,200 \quad 3,900 \quad 3,000$$

Estimate α and θ by the method of moments.

10.27 (*) A random sample of five claims from a lognormal distribution is given as follows:

$$500 \quad 1,000 \quad 1,500 \quad 2,500 \quad 4,500$$

Estimate μ and σ by the method of moments. Estimate the probability that a loss will exceed 4,500.

10.28 (*) The random variable X has pdf $f(x) = \beta^{-2}x \exp(-0.5x^2/\beta^2)$, $x, \beta > 0$. For this random variable, $E(X) = (\beta/2)\sqrt{2\pi}$ and $\text{Var}(X) = 2\beta^2 - \pi\beta^2/2$. You are given the following five observations:

$$4.9 \quad 1.8 \quad 3.4 \quad 6.9 \quad 4.0$$

Determine the method of moments estimate of β.

10.29 The random variable X has pdf $f(x) = \alpha\lambda^\alpha(\lambda + x)^{-\alpha-1}$, $x, \alpha, \lambda > 0$. It is known that $\lambda = 1,000$. You are given the following five observations:

$$43 \quad 145 \quad 233 \quad 396 \quad 775$$

Determine the method of moments estimate of α.

10.30 Use the data in Table 10.5 to determine the method of moments estimate of the parameters of the negative binomial model.

10.31 Use the data in Table 10.6 to determine the method of moments estimate of the parameters of the negative binomial model.

10.32 (*) Losses have a Burr distribution with $\alpha = 2$. A random sample of 15 losses is 195, 255, 270, 280, 350, 360, 365, 380, 415, 450, 490, 550, 575, 590, and 615. Use the smoothed empirical estimates of the 30th and 65th percentiles and percentile matching to estimate the parameters γ and θ.

Table 10.6   The data for Exercise 10.31.

Number of claims	Number of policies
0	861
1	121
2	13
3	3
4	1
5	0
6	1
7+	0

10.33 (*) Losses have a Weibull distribution. A random sample of 16 losses is 54, 70, 75, 81, 84, 88, 97, 105, 109, 114, 122, 125, 128, 139, 146, and 153. Use the smoothed empirical estimates of the 20th and 70th percentiles and percentile matching to estimate the parameters τ and θ.

10.34 (*) Losses follow a distribution with pdf $f(x) = \theta^{-1} \exp[-(x - \delta)/\theta]$, $x > \delta$. The sample mean is 300 and the sample median is 240. Estimate δ and θ by matching these two quantities.

10.5   Tests of Hypotheses

Hypothesis testing is covered in detail in most mathematical statistics texts. This review is fairly straightforward and does not address philosophical issues or consider alternative approaches. A hypothesis test begins with two hypotheses, one called the **null** and one called the **alternative**. The traditional notation is H_0 for the null hypothesis and H_1 for the alternative hypothesis. The two hypotheses are not treated symmetrically. Reversing them may alter the results. To illustrate this process, a simple example is used.

■   **EXAMPLE 10.17**

Your company has been basing its premiums on an assumption that the average claim is 1,200. You want to raise the premium, and a regulator has insisted that you provide evidence that the average now exceeds 1,200. Let Data Set B be a sample of 20 claims. What are the hypotheses for this problem?

Let μ be the population mean. One hypothesis (the one you claim is true) is that $\mu > 1,200$. Because hypothesis tests must present an either/or situation, the other hypothesis must be $\mu \leq 1,200$. The only remaining task is to decide which of them is the null hypothesis. Whenever the universe of continuous possibilities is divided in two, there is likely to be a boundary that needs to be assigned to one hypothesis or the other. The hypothesis that includes the boundary must be the null hypothesis.

Therefore, the problem can be succinctly stated as:

$$H_0 : \mu \leq 1{,}200,$$
$$H_1 : \mu > 1{,}200.$$

☐

The decision is made by calculating a quantity called a **test statistic**. It is a function of the observations and is treated as a random variable. That is, in designing the test procedure, we are concerned with the samples that might have been obtained and not with the particular sample that was obtained. The test specification is completed by constructing a **rejection region**. It is a subset of the possible values of the test statistic. If the value of the test statistic for the observed sample is in the rejection region, the null hypothesis is rejected and the alternative hypothesis is announced as the result that is supported by the data. Otherwise, the null hypothesis is not rejected (more on this later). The boundaries of the rejection region (other than plus or minus infinity) are called the **critical values**.

■ EXAMPLE 10.18

(*Example 10.17 continued*) Complete the test using the test statistic and rejection region that is promoted in most statistics books. Assume that the population has a normal distribution with standard deviation 3,435.

The traditional test statistic for this problem (normal population and standard deviation known) is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1{,}424.4 - 1{,}200}{3{,}435/\sqrt{20}} = 0.292,$$

where μ_0 is the value that separates the null and alternative hypotheses. The null hypothesis is rejected if $z > 1.645$. Because 0.292 is less than 1.645, the null hypothesis is not rejected. The data do not support the assertion that the average claim exceeds 1,200.

☐

The test in the previous example was constructed to meet certain objectives. The first objective is to control what is called the **Type I error**. It is the error made when the test rejects the null hypothesis in a situation in which it happens to be true. In the example, the null hypothesis can be true in more than one way. As a result, a measure of the propensity of a test to make a Type I error must be carefully defined.

Definition 10.10 *The **significance level** of a hypothesis test is the probability of making a Type I error given that the null hypothesis is true. If it can be true in more than one way, the level of significance is the maximum of such probabilities. The significance level is usually denoted by α.*

This is a conservative definition in that it looks at the worst case. It is typically a case that is on the boundary between the two hypotheses.

■ EXAMPLE 10.19

Determine the level of significance for the test in Example 10.18.

Begin by computing the probability of making a Type I error when the null hypothesis is true with $\mu = 1,200$. Then,

$$\Pr(Z > 1.645 | \mu = 1,200) = 0.05$$

because the assumptions imply that Z has a standard normal distribution.

Now suppose that μ has a value below 1,200. Then,

$$\Pr\left(\frac{\bar{X} - 1,200}{3,435/\sqrt{20}} > 1.645\right) = \Pr\left(\frac{\bar{X} - \mu + \mu - 1,200}{3,435/\sqrt{20}} > 1.645\right)$$

$$= \Pr\left(\frac{\bar{X} - \mu}{3,435/\sqrt{20}} > 1.645 - \frac{\mu - 1,200}{3,435/\sqrt{20}}\right).$$

The random variable on the left has a standard normal distribution. Because μ is known to be less than 1,200, the right-hand side is always greater than 1.645. Therefore the probability is less than 0.05, and so the significance level is 0.05. ☐

The significance level is usually set in advance and is often between 1% and 10%. The second objective is to keep the **Type II error** (not rejecting the null hypothesis when the alternative is true) probability small. Generally, attempts to reduce the probability of one type of error increase the probability of the other. The best we can do once the significance level has been set is to make the Type II error as small as possible, though there is no assurance that the probability will be a small number. The best test is one that meets the following requirement.

Definition 10.11 *A hypothesis test is **uniformly most powerful** if no other test exists that has the same or lower significance level and, for a particular value within the alternative hypothesis, has a smaller probability of making a Type II error.*

■ **EXAMPLE 10.20**

(*Example 10.19 continued*) Determine the probability of making a Type II error when the alternative hypothesis is true with $\mu = 2,000$.

$$\Pr\left(\frac{\bar{X} - 1,200}{3,435/\sqrt{20}} < 1.645 | \mu = 2,000\right)$$

$$= \Pr(\bar{X} - 1,200 < 1,263.51 | \mu = 2,000)$$

$$= \Pr(\bar{X} < 2,463.51 | \mu = 2,000)$$

$$= \Pr\left(\frac{\bar{X} - 2,000}{3,435/\sqrt{20}} < \frac{2,463.51 - 2,000}{3,435/\sqrt{20}} = 0.6035\right) = 0.7269.$$

For this value of μ, the test is not very powerful, having a more than 70% chance of making a Type II error. Nevertheless (though this is not easy to prove), the test used is the most powerful test for this problem. ☐

Because the Type II error probability can be high, it is customary to not make a strong statement when the null hypothesis is not rejected. Rather than saying that we choose or

accept the null hypothesis, we say that we fail to reject it. That is, there was not enough evidence in the sample to make a strong argument in favor of the alternative hypothesis, so we take no stand at all.

A common criticism of this approach to hypothesis testing is that the choice of the significance level is arbitrary. In fact, by changing the significance level, any result can be obtained.

■ **EXAMPLE 10.21**

(*Example 10.20 continued*) Complete the test using a significance level of $\alpha = 0.45$. Then determine the range of significance levels for which the null hypothesis is rejected and for which it is not rejected.

Because $\Pr(Z > 0.1257) = 0.45$, the null hypothesis is rejected when

$$\frac{\bar{X} - 1{,}200}{3{,}435/\sqrt{20}} > 0.1257.$$

In this example, the test statistic is 0.292, which is in the rejection region, and thus the null hypothesis is rejected. Of course, few people would place confidence in the results of a test that was designed to make errors 45% of the time. Because $\Pr(Z > 0.292) = 0.3851$, the null hypothesis is rejected for those who select a significance level that is greater than 38.51% and is not rejected by those who use a significance level that is less than 38.51%.   □

Few people are willing to make errors 38.51% of the time. Announcing this figure is more persuasive than the earlier conclusion based on a 5% significance level. When a significance level is used, those interpreting the output are left to wonder what the outcome would have been with other significance levels. The value of 38.51% is called a *p*-value. A working definition follows.

Definition 10.12 *For a hypothesis test, the **p-value** is the probability that the test statistic takes on a value that is less in agreement with the null hypothesis than the value obtained from the sample. Tests conducted at a significance level that is greater than the p-value will lead to a rejection of the null hypothesis, while tests conducted at a significance level that is smaller than the p-value will lead to a failure to reject the null hypothesis.*

Also, because the *p*-value must be between 0 and 1, it is on a scale that carries some meaning. The closer to zero the value is, the more support the data give to the alternative hypothesis. Common practice is that values above 10% indicate that the data provide no evidence in support of the alternative hypothesis, while values below 1% indicate strong support for the alternative hypothesis. Values in between indicate uncertainty as to the appropriate conclusion, and may call for more data or a more careful look at the data or the experiment that produced it.

This approach to hypothesis testing has some consequences that can create difficulties when answering actuarial questions. The following example illustrates these problems.

■ **EXAMPLE 10.22**

You believe that the lognormal model is appropriate for the problem you are investigating. You have collected some data and would like to test this hypothesis. What are the null and alternative hypotheses and what will you learn after completing the test?

Methods for conducting this test are presented in Section 15.4. One hypothesis is that the population has the lognormal distribution and the other is that it does not. The first one is the statement of equality and so must be the null hypothesis. The problem is that while the data can confirm that the population is *not* lognormal, the method does not allow you to assert that the population *is* lognormal. A second problem is that often the null hypothesis is known to be false. In this case, we know that the population is unlikely to be exactly lognormal. If our sample size is large enough, the hypothesis test will discover this, and it is likely that all models will be rejected.   □

It is important to keep in mind that hypothesis testing was invented for situations in which collecting data was either expensive or inconvenient. For example, in deciding if a new drug cures a disease, it is important to confirm this fact with the smallest possible sample so that, if the results are favorable, the drug can be approved and made available. Or, consider testing a new crop fertilizer. Every test acre planted costs time and money. In contrast, in many types of actuarial problems, a large amount of data is available from historical records. In this case, unless the data follow a parametric model extremely closely, almost any model can be rejected by using a sufficiently large set of data.

10.5.1 Exercise

10.35 (*Exercise 10.12 continued*) Test $H_0 : \theta \geq 325$ versus $H_1 : \theta < 325$ using a significance level of 5% and the sample mean as the test statistic. Also, compute the p-value. Do this twice, using: (i) the exact distribution of the test statistic and (ii) a normal approximation.

11

MAXIMUM LIKELIHOOD ESTIMATION

11.1   Introduction

Estimation by the method of moments and percentile matching is often easy to do, but these estimators tend to perform poorly, mainly because they use only a few features of the data, rather than the entire set of observations. It is particularly important to use as much information as possible when the population has a heavy right tail. For example, when estimating parameters for the normal distribution, the sample mean and variance are sufficient.[1] However, when estimating parameters for a Pareto distribution, it is important to know all the extreme observations in order to successfully estimate α. Another drawback of these methods is that they require that all the observations are from the same random variable. Otherwise, it is not clear what to use for the population moments or percentiles. For example, if half the observations have a deductible of 50 and half have a deductible of

---

[1]This applies both in the formal statistical definition of sufficiency (not covered here) and in the conventional sense. If the population has a normal distribution, the sample mean and variance convey as much information as the original observations.

100, it is not clear to what the sample mean should be equated.[2] Finally, these methods allow the analyst to make arbitrary decisions regarding the moments or percentiles to use.

There are a variety of estimators that use the individual data points. All of them are implemented by setting an objective function and then determining the parameter values that optimize that function. Of the many possibilities, the only one presented here is the maximum likelihood estimator.

To define the maximum likelihood estimator, let the data set be the n events $A_1, \ldots, A_n$, where A_j is whatever was observed for the jth observation. For example, A_j may consist of a single point or an interval. The latter arises, for example, with grouped data. Further assume that the event A_j results from observing the random variable X_j. The random variables $X_1, \ldots, X_n$ need not have the same probability distribution, but their distributions must depend on the same parameter vector, θ. In addition, the random variables are assumed to be independent.

Definition 11.1 *The **likelihood function** is*

$$L(\theta) = \prod_{j=1}^{n} \Pr(X_j \in A_j | \theta)$$

*and the **maximum likelihood estimate** of θ is the vector that maximizes the likelihood function.*[3]

In the definition, if A_j is a single point and the distribution is continuous, then $\Pr(X_j \in A_j | \theta)$ is interpreted as the probability density function evaluated at that point. In all other cases, it is the probability of that event.

There is no guarantee that the function has a maximum at eligible parameter values. It is possible that as various parameters become zero or infinite, the likelihood function will continue to increase. Care must be taken when maximizing this function because there may be local maxima in addition to the global maximum. Often, it is not possible to analytically maximize the likelihood function (by setting partial derivatives equal to zero). Numerical approaches will usually be needed.

Because the observations are assumed to be independent, the product in the definition represents the joint probability $\Pr(X_1 \in A_1, \ldots, X_n \in A_n | \theta)$, that is, the likelihood function is the probability of obtaining the sample results that were obtained, given a particular parameter value. The estimate is then the parameter value that produces the model under which the actual observations are most likely to be observed. One of the major attractions of this estimator is that it is almost always available. That is, if you can write an expression for the desired probabilities, you can execute this method. If you cannot write and evaluate an expression for probabilities using your model, there is no point in postulating that model in the first place, because you will not be able to use it to solve the larger actuarial problem.

---

[2]One way to rectify that drawback is to first determine a data-dependent model such as the Kaplan–Meier estimate introduced in Section 14.3. Then use percentiles or moments from that model.
[3]Some authors write the likelihood function as $L(\theta|\mathbf{x})$, where the vector $\mathbf{x}$ represents the observed data. Because observed data can take many forms, the dependence of the likelihood function on the data is suppressed in the notation.

Table 11.1	Data Set B.								
27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1,193	1,340	1,884	2,558	15,743

■ **EXAMPLE 11.1**

Suppose that the data in Data Set B, introduced in Chapter 10 and reproduced here as Table 11.1 were such that the exact value of all observations above 250 was unknown. All that is known is that the value was greater than 250. Determine the maximum likelihood estimate of θ for an exponential distribution.

For the first seven values, the set A_j contains the single point equal to the observation x_j. Thus the first seven terms of the product are

$$f(27)f(82)\cdots f(243) = \theta^{-1}e^{-27/\theta}\theta^{-1}e^{-82/\theta}\cdots\theta^{-1}e^{-243/\theta} = \theta^{-7}e^{-909/\theta}.$$

For each of the final 13 terms, the set A_j is the interval from 250 to infinity and, therefore, $\Pr(X_j \in A_j) = \Pr(X_j > 250) = e^{-250/\theta}$. There are 13 such factors, making the likelihood function

$$L(\theta) = \theta^{-7}e^{-909/\theta}(e^{-250/\theta})^{13} = \theta^{-7}e^{-4,159/\theta}.$$

It is easier to maximize the logarithm of the likelihood function. Because it occurs so often, we denote the **loglikelihood function** as $l(\theta) = \ln L(\theta)$. Then,

$$l(\theta) = -7\ln\theta - 4{,}159\theta^{-1},$$
$$l'(\theta) = -7\theta^{-1} + 4{,}159\theta^{-2} = 0,$$
$$\hat{\theta} = \frac{4{,}159}{7} = 594.14.$$

In this case, the calculus technique of setting the first derivative equal to zero is easy to do. Also, evaluating the second derivative at this solution produces a negative number, verifying that this solution is a maximum.   □

11.2 Individual Data

Consider the special case where the value of each observation is recorded. It is easy to write the loglikelihood function:

$$L(\theta) = \prod_{j=1}^{n} f_{X_j}(x_j|\theta), \quad l(\theta) = \sum_{j=1}^{n} \ln f_{X_j}(x_j|\theta).$$

The notation indicates that it is not necessary for each observation to come from the same distribution, but we continue to assume that the parameter vector is common to each distribution.

■ **EXAMPLE 11.2**

Using Data Set B, determine the maximum likelihood estimates for an exponential distribution, for a gamma distribution where α is known to equal 2, and for a gamma distribution where both parameters are unknown.

For the exponential distribution, the general solution is

$$l(\theta) = \sum_{j=1}^{n} \left(-\ln \theta - x_j \theta^{-1} \right) = -n \ln \theta - n\bar{x}\theta^{-1},$$

$$l'(\theta) = -n\theta^{-1} + n\bar{x}\theta^{-2} = 0,$$

$$n\theta = n\bar{x},$$

$$\hat{\theta} = \bar{x}.$$

For Data Set B, $\hat{\theta} = \bar{x} = 1{,}424.4$. The value of the loglikelihood function is -165.23. For this situation, the method of moments and maximum likelihood estimates are identical. For further insight, see Exercise 11.16.

For the gamma distribution with $\alpha = 2$,

$$f(x|\theta) = \frac{x^{2-1}e^{-x/\theta}}{\Gamma(2)\theta^2} = x\theta^{-2}e^{-x/\theta},$$

$$\ln f(x|\theta) = \ln x - 2\ln \theta - x\theta^{-1},$$

$$l(\theta) = \sum_{j=1}^{n} \ln x_j - 2n\ln \theta - n\bar{x}\theta^{-1},$$

$$l'(\theta) = -2n\theta^{-1} + n\bar{x}\theta^{-2} = 0,$$

$$\hat{\theta} = \tfrac{1}{2}\bar{x}.$$

For Data Set B, $\hat{\theta} = 1{,}424.4/2 = 712.2$ and the value of the loglikelihood function is -179.98. Again, this estimate is the same as the method of moments estimate.

For the gamma distribution with unknown parameters, the function is not as simple:

$$f(x|\alpha, \theta) = \frac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha},$$

$$\ln f(x|\alpha, \theta) = (\alpha - 1)\ln x - x\theta^{-1} - \ln \Gamma(\alpha) - \alpha \ln \theta.$$

The partial derivative with respect to α requires the derivative of the gamma function. The resulting equation cannot be solved analytically. Using numerical methods, the estimates are $\hat{\alpha} = 0.55616$ and $\hat{\theta} = 2{,}561.1$ and the value of the loglikelihood function is -162.29. These do not match the method of moments estimates.   □

11.2.1 Exercises

11.1 Repeat Example 11.2 using the inverse exponential, inverse gamma with $\alpha = 2$, and inverse gamma distributions. Compare your estimates with the method of moments estimates.

11.2 (*) You are given the five observations 521, 658, 702, 819, and 1,217. Your model is the single-parameter Pareto distribution with distribution function

$$F(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500, \alpha > 0.$$

Determine the maximum likelihood estimate of α.

11.3 (*) You have observed the following five claim severities: 11.0, 15.2, 18.0, 21.0, and 25.8. Determine the maximum likelihood estimate of μ for the following model (which is the reciprocal inverse Gaussian distribution; see Exercise 5.20a of [74]):

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

11.4 (*) The following values were calculated from a random sample of 10 losses:

$$\sum_{j=1}^{10} x_j^{-2} = 0.00033674, \quad \sum_{j=1}^{10} x_j^{-1} = 0.023999,$$

$$\sum_{j=1}^{10} x_j^{-0.5} = 0.34445, \quad \sum_{j=1}^{10} x_j^{0.5} = 488.97$$

$$\sum_{j=1}^{10} x_j = 31{,}939, \quad \sum_{j=1}^{10} x_j^2 = 211{,}498{,}983.$$

Losses come from a Weibull distribution with $\tau = 0.5$ [so $F(x) = 1 - e^{-(x/\theta)^{0.5}}$]. Determine the maximum likelihood estimate of θ.

11.5 (*) A sample of n independent observations $x_1, \ldots, x_n$ came from a distribution with a pdf of $f(x) = 2\theta x \exp(-\theta x^2), \ x > 0$. Determine the maximum likelihood estimator (mle) of θ.

11.6 (*) Let $x_1, \ldots, x_n$ be a random sample from a population with cdf $F(x) = x^p, \ 0 < x < 1$. Determine the mle of p.

11.7 A random sample of 10 claims obtained from a gamma distribution is given as follows:

  1,500   6,000   3,500   3,800   1,800   5,500   4,800   4,200   3,900   3,000

(a) (*) Suppose it is known that $\alpha = 12$. Determine the maximum likelihood estimate of θ.

(b) Determine the maximum likelihood estimates of α and θ.

11.8 A random sample of five claims from a lognormal distribution is given as follows:

  500   1,000   1,500   2,500   4,500

Estimate μ and σ by maximum likelihood. Estimate the probability that a loss will exceed 4,500.

11.9 (*) Let $x_1, \ldots, x_n$ be a random sample from a random variable with pdf $f(x) = \theta^{-1}e^{-x/\theta}$, $x > 0$. Determine the mle of θ.

11.10 (*) The random variable X has pdf $f(x) = \beta^{-2}x\exp(-0.5x^2/\beta^2)$, $x, \beta > 0$. For this random variable, $E(X) = (\beta/2)\sqrt{2\pi}$ and $\text{Var}(X) = 2\beta^2 - \pi\beta^2/2$. You are given the following five observations:

$$4.9 \quad 1.8 \quad 3.4 \quad 6.9 \quad 4.0$$

Determine the maximum likelihood estimate of β.

11.11 (*) Let $x_1, \ldots, x_n$ be a random sample from a random variable with cdf $F(x) = 1 - x^{-\alpha}$, $x > 1$, $\alpha > 0$. Determine the mle of α.

11.12 (*) The random variable X has pdf $f(x) = \alpha\lambda^{\alpha}(\lambda + x)^{-\alpha-1}$, $x, \alpha, \lambda > 0$. It is known that $\lambda = 1,000$. You are given the following five observations:

$$43 \quad 145 \quad 233 \quad 396 \quad 775$$

Determine the maximum likelihood estimate of α.

11.13 The following 20 observations were collected. It is desired to estimate $\text{Pr}(X > 200)$. When a parametric model is called for, use the single-parameter Pareto distribution for which $F(x) = 1 - (100/x)^{\alpha}$, $x > 100$, $\alpha > 0$.

$$132 \quad 149 \quad 476 \quad 147 \quad 135 \quad 110 \quad 176 \quad 107 \quad 147 \quad 165$$
$$135 \quad 117 \quad 110 \quad 111 \quad 226 \quad 108 \quad 102 \quad 108 \quad 227 \quad 102$$

(a) Determine the empirical estimate of $\text{Pr}(X > 200)$.

(b) Determine the method of moments estimate of the single-parameter Pareto parameter α and use it to estimate $\text{Pr}(X > 200)$.

(c) Determine the maximum likelihood estimate of the single-parameter Pareto parameter α and use it to estimate $\text{Pr}(X > 200)$.

11.14 Consider the inverse Gaussian distribution with density given by

$$f_X(x) = \left(\frac{\theta}{2\pi x^3}\right)^{1/2}\exp\left[-\frac{\theta}{2x}\left(\frac{x - \mu}{\mu}\right)^2\right], \quad x > 0.$$

(a) Show that

$$\sum_{j=1}^{n}\frac{(x_j - \mu)^2}{x_j} = \mu^2\sum_{j=1}^{n}\left(\frac{1}{x_j} - \frac{1}{\bar{x}}\right) + \frac{n}{\bar{x}}(\bar{x} - \mu)^2,$$

where $\bar{x} = (1/n)\sum_{j=1}^{n}x_j$.

(b) For a sample $(x_1, \ldots, x_n)$, show that the maximum likelihood estimates of μ and θ are

$$\hat{\mu} = \bar{x}$$

and

$$\hat{\theta} = \frac{n}{\sum\limits_{j=1}^{n} \left(\dfrac{1}{x_j} - \dfrac{1}{\bar{x}} \right)}.$$

11.15 Suppose that $X_1, \ldots, X_n$ are independent and normally distributed with mean $E(X_j) = \mu$ and $\mathrm{Var}(X_j) = (\theta m_j)^{-1}$, where $m_j > 0$ is a known constant. Prove that the maximum likelihood estimates of μ and θ are

$$\hat{\mu} = \bar{X}$$

and

$$\hat{\theta} = n \left[\sum_{j=1}^{n} m_j (X_j - \bar{X})^2 \right]^{-1},$$

where $\bar{X} = (1/m) \sum_{j=1}^{n} m_j X_j$ and $m = \sum_{j=1}^{n} m_j$.

11.16 Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution (5.6). Prove that the maximum likelihood estimate of the mean is the sample mean. In other words, if $\hat{\theta}$ is the mle of θ, prove that

$$\widehat{\mu(\theta)} = \mu(\hat{\theta}) = \bar{X}.$$

11.3 Grouped Data

When data are grouped and the groups span the range of possible observations, the observations may be summarized as follows. Begin with a set of numbers $c_0 < c_1 < \cdots < c_k$, where c_0 is the smallest possible observation (often zero) and c_k is the largest possible observation (often infinity). From the sample, let n_j be the number of observations in the interval $(c_{j-1}, c_j]$. For such data, the likelihood function is

$$L(\theta) = \prod_{j=1}^{k} [F(c_j|\theta) - F(c_{j-1}|\theta)]^{n_j},$$

and its logarithm is

$$l(\theta) = \sum_{j=1}^{k} n_j \ln[F(c_j|\theta) - F(c_{j-1}|\theta)].$$

■ **EXAMPLE 11.3**

Using Data Set C from Chapter 10, reproduced here as Table 11.2, determine the maximum likelihood estimate for an exponential distribution.

The loglikelihood function is

$$l(\theta) = 99 \ln[F(7{,}500) - F(0)] + 42 \ln[F(17{,}500) - F(7{,}500)] + \cdots$$
$$+ 3 \ln[1 - F(300{,}000)]$$
$$= 99 \ln(1 - e^{-7{,}500/\theta}) + 42 \ln(e^{-7{,}500/\theta} - e^{-17{,}500/\theta}) + \cdots$$
$$+ 3 \ln e^{-300{,}000/\theta}.$$

Table 11.2 Data Set C.

Payment range	Number of payments
0–7,500	99
7,500–17,500	42
17,500–32,500	29
32,500–67,500	28
67,500–125,000	17
125,000–300,000	9
Over 300,000	3

Table 11.3 The data for Exercise 11.19.

Loss	Number of observations	Loss	Number of observations
0–25	5	350–500	17
25–50	37	500–750	13
50–75	28	750–1000	12
75–100	31	1,000–1,500	3
100–125	23	1,500–2,500	5
125–150	9	2,500–5,000	5
150–200	22	5,000–10,000	3
200–250	17	10,000–25,000	3
250–350	15	25,000–	2

A numerical routine is needed to produce $\hat{\theta} = 29{,}721$, and the value of the loglikelihood function is -406.03. □

11.3.1 Exercises

11.17 From Data Set C, determine the maximum likelihood estimates for gamma, inverse exponential, and inverse gamma distributions.

11.18 (*) Losses follow a distribution with cdf $F(x) = 1 - \theta/x$, $x > \theta$. A sample of 20 losses contained 9 below 10, 6 between 10 and 25, and 5 in excess of 25. Determine the maximum likelihood estimate of θ.

11.19 Table 11.3 presents the results of a sample of 250 losses.

Consider the inverse exponential distribution with cdf $F(x) = e^{-\theta/x}$, $x > 0$, $\theta > 0$. Determine the maximum likelihood estimate of θ.

11.4 Truncated or Censored Data

The definition of right censoring is as follows.

Definition 11.2 *An observation is **censored from above** (also called **right censored**) at u if when it is at or above u it is recorded as being equal to u, but when it is below u it is recorded at its observed value.*

A similar definition applies to left censoring, which is uncommon in insurance settings.

Data Set D in Chapter 10 illustrates how censoring can occur in mortality data. If observation of an insured ends before death, all we know is that death occurs sometime after the time of the last observation. Another common situation is a policy limit where, if the actual loss exceeds the limit, all that is known is that the limit was exceeded.

When data are censored, there is no additional complication. Right censoring simply creates an interval running from the censoring point to infinity. Data below the censoring point are individual data, and so the likelihood function contains both density and distribution function terms.

The definition of truncation is as follows.

Definition 11.3 *An observation is **truncated from below** (also called **left truncated**) at d if when it is at or below d it is not recorded, but when it is above d it is recorded at its observed value.*

A similar definition applies to right truncation, which is uncommon in insurance settings.

Data Set D also illustrates left truncation. For policies sold before the observation period begins, some insured lives will die while others will be alive to enter observation. Not only will their death times be unrecorded, but we will not even know how many there were. Another common situation is a deductible. Losses below the deductible are not recorded and there is no count of how many losses were below the deductible.

Truncated data present more of a challenge. There are two ways to proceed. One is to shift the data by subtracting the truncation point from each observation. The other is to accept the fact that there is no information about values below the truncation point but then attempt to fit a model for the original population.

■ **EXAMPLE 11.4**

Assume that the values in Data Set B had been truncated from below at 200. Using both methods, estimate the value of α for a Pareto distribution with $\theta = 800$ known. Then use the model to estimate the cost per payment with deductibles of 0, 200, and 400.

Using the shifting approach, the values become 43, 94, 140, 184, 257, 480, 655, 677, 774, 993, 1,140, 1,684, 2,358, and 15,543. The likelihood function is

$$L(\alpha) = \prod_{j=1}^{14} \frac{\alpha(800^\alpha)}{(800 + x_j)^{\alpha+1}},$$

$$l(\alpha) = \sum_{j=1}^{14} [\ln \alpha + \alpha \ln 800 - (\alpha + 1) \ln(x_j + 800)]$$

$$= 14 \ln \alpha + 93.5846\alpha - 103.969(\alpha + 1)$$

$$= 14 \ln \alpha - 103.969 - 10.384\alpha,$$

$$l'(\alpha) = 14\alpha^{-1} - 10.384,$$

$$\hat{\alpha} = \frac{14}{10.384} = 1.3482.$$

Because the data have been shifted, it is not possible to estimate the cost with no deductible. With a deductible of 200, the expected cost is the expected value of the estimated Pareto distribution, $800/0.3482 = 2,298$. Raising the deductible to 400 is equivalent to imposing a deductible of 200 on the modeled distribution. From Theorem 8.7, the expected cost per payment is

$$\frac{E(X) - E(X \wedge 200)}{1 - F(200)} = \frac{\frac{800}{0.3482}\left(\frac{800}{200 + 800}\right)^{0.3482}}{\left(\frac{800}{200 + 800}\right)^{1.3482}} = \frac{1,000}{0.3482} = 2,872.$$

For the unshifted approach, we need to ask the key question required when constructing the likelihood function. That is, what is the probability of observing each value knowing that values under 200 are omitted from the data set? This becomes a conditional probability and therefore the likelihood function is (where the x_j values are now the original values)

$$L(\alpha) = \prod_{j=1}^{14} \frac{f(x_j|\alpha)}{1 - F(200|\alpha)} = \prod_{j=1}^{14}\left[\frac{\alpha(800^\alpha)}{(800 + x_j)^{\alpha+1}} \Big/ \left(\frac{800}{800 + 200}\right)^\alpha\right]$$

$$= \prod_{j=1}^{14} \frac{\alpha(1,000^\alpha)}{(800 + x_j)^{\alpha+1}},$$

$$l(\alpha) = 14 \ln \alpha + 14\alpha \ln 1,000 - (\alpha + 1)\sum_{j=1}^{14}\ln(800 + x_j),$$

$$= 14 \ln \alpha + 96.709\alpha - (\alpha + 1)105.810,$$

$$l'(\alpha) = 14\alpha^{-1} - 9.101,$$

$$\hat{\alpha} = 1.5383.$$

This model is for losses with no deductible, and therefore the expected payment without a deductible is $800/0.5383 = 1,486$. Imposing deductibles of 200 and 400 produces the following results:

$$\frac{E(X) - E(X \wedge 200)}{1 - F(200)} = \frac{1,000}{0.5383} = 1,858,$$

$$\frac{E(X) - E(X \wedge 400)}{1 - F(400)} = \frac{1,200}{0.5383} = 2,229. \qquad \square$$

It should now be clear that the contribution to the likelihood function can be written for most any type of observation. In general, the likelihood is always (proportional to) the probability of observing the data under the model, with the understanding for this purpose that pdfs for continuous variables should be viewed as probabilities. The following two steps summarize the process:

1. For the numerator, use $f(x)$ if the exact value, x, of the observation is known. If it is only known that the observation is between y and z, use $F(z) - F(y)$.

2. For the denominator, if there is no truncation, the denominator is 1. Else, let d be the truncation point, in which case the denominator is $1 - F(d)$.

Table 11.4   The likelihood function for Example 11.5.

Obs.	x, y	d	L	Obs.	x, y	d	L
1	$y = 0.1$	0	$1 - F(0.1)$	16	$x = 4.8$	0	$f(4.8)$
2	$y = 0.5$	0	$1 - F(0.5)$	17	$y = 4.8$	0	$1 - F(4.8)$
3	$y = 0.8$	0	$1 - F(0.8)$	18	$y = 4.8$	0	$1 - F(4.8)$
4	$x = 0.8$	0	$f(0.8)$	19–30	$y = 5.0$	0	$1 - F(5.0)$
5	$y = 1.8$	0	$1 - F(1.8)$	31	$y = 5.0$	0.3	$\frac{1-F(5.0)}{1-F(0.3)}$
6	$y = 1.8$	0	$1 - F(1.8)$	32	$y = 5.0$	0.7	$\frac{1-F(5.0)}{1-F(0.7)}$
7	$y = 2.1$	0	$1 - F(2.1)$	33	$x = 4.1$	1.0	$\frac{f(4.1)}{1-F(1.0)}$
8	$y = 2.5$	0	$1 - F(2.5)$	34	$x = 3.1$	1.8	$\frac{f(3.1)}{1-F(1.8)}$
9	$y = 2.8$	0	$1 - F(2.8)$	35	$y = 3.9$	2.1	$\frac{1-F(3.9)}{1-F(2.1)}$
10	$x = 2.9$	0	$f(2.9)$	36	$y = 5.0$	2.9	$\frac{1-F(5.0)}{1-F(2.9)}$
11	$x = 2.9$	0	$f(2.9)$	37	$y = 4.8$	2.9	$\frac{1-F(4.8)}{1-F(2.9)}$
12	$y = 3.9$	0	$1 - F(3.9)$	38	$x = 4.0$	3.2	$\frac{f(4.0)}{1-F(3.2)}$
13	$x = 4.0$	0	$f(4.0)$	39	$y = 5.0$	3.4	$\frac{1-F(5.0)}{1-F(3.4)}$
14	$y = 4.0$	0	$1 - F(4.0)$	40	$y = 5.0$	3.9	$\frac{1-F(5.0)}{1-F(3.9)}$
15	$y = 4.1$	0	$1 - F(4.1)$				

■ **EXAMPLE 11.5**

Determine Pareto and gamma models for the time to death for Data Set D, introduced in Chapter 10.

Table 11.4 shows how the likelihood function is constructed for these observations. For observed deaths, the time is known, and so the exact value of x is available. For surrenders or those reaching time 5, the observation is censored, and therefore death is known to be some time in the interval from the surrender time, y, to infinity. In the table, $z = \infty$ is not noted because all interval observations end at infinity. The likelihood function must be maximized numerically. For the Pareto distribution, there is no solution. The likelihood function keeps getting larger as α and θ get larger.[4] For the gamma distribution, the maximum is at $\hat{\alpha} = 2.617$ and $\hat{\theta} = 3.311$.   □

Discrete data present no additional problems.

---

[4]For a Pareto distribution, the limit as the parameters α and θ become infinite with the ratio being held constant is an exponential distribution. Thus, for this example, the exponential distribution is a better model (as measured by the likelihood function) than any Pareto model.

Table 11.5   Data Set A.

Number of accidents	Number of drivers
0	81,714
1	11,306
2	1,618
3	250
4	40
5 or more	7

■ **EXAMPLE 11.6**

For Data Set A, which was introduced in Chapter 10 and is reproduced here as Table 11.5, assume that the seven drivers with five or more accidents all had exactly five accidents. Determine the maximum likelihood estimate for a Poisson distribution and for a binomial distribution with $m = 8$.

In general, for a discrete distribution with complete data, the likelihood function is

$$L(\theta) = \prod_{j=1}^{\infty} [p(x_j|\theta)]^{n_j},$$

where x_j is one of the observed values, $p(x_j|\theta)$ is the probability of observing x_j, and n_x is the number of times x was observed in the sample. For the Poisson distribution,

$$L(\lambda) = \prod_{x=0}^{\infty} \left(\frac{e^{-\lambda} \lambda^x}{x!} \right)^{n_x} = \prod_{x=0}^{\infty} \frac{e^{-n_x \lambda} \lambda^{xn_x}}{(x!)^{n_x}},$$

$$l(\lambda) = \sum_{x=0}^{\infty} (-n_x \lambda + xn_x \ln \lambda - n_x \ln x!) = -n\lambda + n\bar{x} \ln \lambda - \sum_{x=0}^{\infty} n_x \ln x!,$$

$$l'(\lambda) = -n + \frac{n\bar{x}}{\lambda} = 0,$$

$$\hat{\lambda} = \bar{x}.$$

For the binomial distribution,

$$L(q) = \prod_{x=0}^{m} \left[\binom{m}{x} q^x (1-q)^{m-x} \right]^{n_x} = \prod_{x=0}^{m} \frac{m!^{n_x} q^{xn_x} (1-q)^{(m-x)n_x}}{(x!)^{n_x} [(m-x)!]^{n_x}},$$

$$l(q) = \sum_{x=0}^{m} [n_x \ln m! + xn_x \ln q + (m-x)n_x \ln(1-q)]$$

$$- \sum_{x=0}^{m} [n_x \ln x! + n_x \ln(m-x)!],$$

$$l'(q) = \sum_{x=0}^{m} \frac{xn_x}{q} - \frac{(m-x)n_x}{1-q} = \frac{n\bar{x}}{q} - \frac{mn - n\bar{x}}{1-q} = 0,$$

$$\hat{q} = \frac{\bar{x}}{m}.$$

For this problem, $\bar{x} = [81{,}714(0) + 11{,}306(1) + 1{,}618(2) + 250(3) + 40(4) + 7(5)]/94{,}935 = 0.16313$. Therefore, for the Poisson distribution, $\hat{\lambda} = 0.16313$, and for the binomial distribution, $\hat{q} = 0.16313/8 = 0.02039$. □

In Exercise 11.23, you are asked to estimate the Poisson parameter when the actual values for those with five or more accidents are not known.

11.4.1 Exercises

11.20 Determine maximum likelihood estimates for Data Set B using the inverse exponential, gamma, and inverse gamma distributions. Assume that the data have been censored at 250 and then compare your answers to those obtained in Example 11.2 and Exercise 11.1.

11.21 Repeat Example 11.4 using a Pareto distribution with both parameters unknown.

11.22 Repeat Example 11.5, this time finding the distribution of the time to surrender.

11.23 Repeat Example 11.6, but this time assume that the actual values for the seven drivers who have five or more accidents are unknown. Note that this is a case of censoring.

11.24 (*) Five hundred losses are observed. Five of the losses are 1,100, 3,200, 3,300, 3,500, and 3,900. All that is known about the other 495 losses is that they exceed 4,000. Determine the maximum likelihood estimate of the mean of an exponential model.

11.25 (*) Ten claims were observed. The values of seven of them (in thousands) were 3, 7, 8, 12, 12, 13, and 14. The remaining three claims were all censored at 15. The proposed model has a hazard rate function given by

$$
h(t) = \begin{cases} \lambda_1, & 0 < t < 5, \\ \lambda_2, & 5 \le t < 10, \\ \lambda_3, & t \ge 10. \end{cases}
$$

Determine the maximum likelihood estimates of the three parameters.

11.26 (*) A random sample of size 5 is taken from a Weibull distribution with $\tau = 2$. Two of the sample observations are known to exceed 50 and the three remaining observations are 20, 30, and 45. Determine the maximum likelihood estimate of θ.

11.27 (*) Phil and Sylvia are competitors in the light bulb business. Sylvia advertises that her light bulbs burn twice as long as Phil's. You were able to test 20 of Phil's bulbs and 10 of Sylvia's. You assumed that both of their bulbs have an exponential distribution, with time measured in hours. You have separately estimated the parameters as $\hat{\theta}_P = 1{,}000$ and $\hat{\theta}_S = 1{,}500$ for Phil and Sylvia, respectively, using maximum likelihood. Using all 30 observations, determine $\hat{\theta}^*$, the maximum likelihood estimate of θ_P restricted by Sylvia's claim that $\theta_S = 2\theta_P$.

11.28 (*) A sample of 100 losses revealed that 62 were below 1,000 and 38 were above 1,000. An exponential distribution with mean θ is considered. Using only the given information, determine the maximum likelihood estimate of θ. Now suppose you are also

given that the 62 losses that were below 1,000 totaled 28,140, while the total for the 38 above 1,000 remains unknown. Using this additional information, determine the maximum likelihood estimate of θ.

11.29 (*) For claims reported in 1997, the number settled in 1997 (year 0) was unknown, the number settled in 1998 (year 1) was three, and the number settled in 1999 (year 2) was one. The number settled after 1999 is unknown. For claims reported in 1998, there were five settled in year 0, two settled in year 1, and the number settled after year 1 is unknown. For claims reported in 1999, there were four settled in year 0 and the number settled after year 0 is unknown. Let N be the year in which a randomly selected claim is settled and assume that it has probability function $\Pr(N = n) = p_n = (1 - p)p^n$, $n = 0, 1, 2, \dots$. Determine the maximum likelihood estimate of p.

11.30 (*) Losses have a uniform distribution on the interval $(0, w)$. Five losses are observed, all with a deductible of 4. Three losses are observed with values of 5, 9, and 13. The other two losses are censored at a value of $4 + p$. The maximum likelihood estimate of w is 29. Determine the value of p.

11.31 (*) Three losses are observed with values 66, 91, and 186. Seven other losses are known to be less than or equal to 60. Losses have an inverse exponential distribution with cdf $F(x) = e^{-\theta/x}$, $x > 0$. Determine the maximum likelihood estimate of the population mode.

11.32 (*) Policies have a deductible of 100. Seven losses are observed, with values 120, 180, 200, 270, 300, 1,000, and 2,500. Ground-up losses have a Pareto distribution with $\theta = 400$ and α unknown. Determine the maximum likelihood estimate of α.

11.5   Variance and Interval Estimation for Maximum Likelihood Estimators

In general, it is not easy to determine the variance of complicated estimators such as the mle. However, it is possible to approximate the variance. The key is a theorem that can be found in most mathematical statistics books. The particular version stated here and its multiparameter generalization are taken from [106] and stated without proof. Recall that $L(\theta)$ is the likelihood function and $l(\theta)$ its logarithm. All of the results assume that the population has a distribution that is a member of the chosen parametric family.

Theorem 11.4 *Assume that the pdf (pf in the discrete case)* $f(x; \theta)$ *satisfies the following for* θ *in an interval containing the true value (replace integrals by sums for discrete variables):*

(i) $\ln f(x; \theta)$ *is three times differentiable with respect to* θ.

(ii) $\int \dfrac{\partial}{\partial \theta} f(x; \theta)\, dx = 0$. *This formula implies that the derivative may be taken outside the integral and so we are just differentiating the constant* 1.[5]

---

[5]The integrals in (ii) and (iii) are to be evaluated over the range of x values for which $f(x; \theta) > 0$.

(iii) $\int \dfrac{\partial^2}{\partial\theta^2} f(x;\theta)\,dx = 0.$ *This formula is the same concept for the second derivative.*

(iv) $-\infty < \int f(x;\theta)\dfrac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\,dx < 0.$ *This inequality establishes that the indicated integral exists and that the location where the derivative is zero is a maximum.*

(v) There exists a function $H(x)$ such that

$$\int H(x)f(x;\theta)\,dx < \infty \quad \text{with} \quad \left|\dfrac{\partial^3}{\partial\theta^3}\ln f(x;\theta)\right| < H(x).$$

This inequality makes sure that the population is not overpopulated with regard to extreme values.

Then the following results hold:

(a) As $n \to \infty$, the probability that the likelihood equation $[L'(\theta) = 0]$ has a solution goes to 1.

(b) As $n \to \infty$, the distribution of the mle $\hat{\theta}_n$ converges to a normal distribution with mean θ and variance such that $I(\theta)\,\mathrm{Var}(\hat{\theta}_n) \to 1$, where

$$I(\theta) = -n\mathrm{E}\left[\dfrac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right] = -n\int f(x;\theta)\dfrac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\,dx$$

$$= n\mathrm{E}\left[\left(\dfrac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right] = n\int f(x;\theta)\left(\dfrac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2 dx.$$

For any z, part (b) is to be interpreted as

$$\lim_{n\to\infty}\Pr\left(\dfrac{\hat{\theta}_n - \theta}{[I(\theta)]^{-1/2}} < z\right) = \Phi(z),$$

and therefore $[I(\theta)]^{-1}$ is a useful approximation for $\mathrm{Var}(\hat{\theta}_n)$. The quantity $I(\theta)$ is called the **information** (sometimes, more specifically, **Fisher information**). It follows from this result that the mle is asymptotically unbiased and consistent. The conditions in statements *(i)–(v)* are often referred to as "mild regularity conditions." A skeptic would translate this statement as "conditions that are almost always true but are often difficult to establish, so we'll just assume they hold in our case." Their purpose is to ensure that the density function is fairly smooth with regard to changes in the parameter and that there is nothing unusual about the density itself.[6]

The preceding results assume that the sample consists of i.i.d. random observations. A more general version of the result uses the logarithm of the likelihood function:

$$I(\theta) = -\mathrm{E}\left[\dfrac{\partial^2}{\partial\theta^2}l(\theta)\right] = \mathrm{E}\left[\left(\dfrac{\partial}{\partial\theta}l(\theta)\right)^2\right].$$

The only requirement here is that the same parameter value apply to each observation.

---

[6]For an example of a situation in which these conditions do not hold, see Exercise 11.34.

If there is more than one parameter, the only change is that the vector of maximum likelihood estimates now has an asymptotic multivariate normal distribution. The covariance matrix[7] of this distribution is obtained from the inverse of the matrix with (r, s)th element,

$$\mathbf{I}(\theta)_{rs} = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_s\,\partial\theta_r}l(\theta)\right] = -n\mathrm{E}\left[\frac{\partial^2}{\partial\theta_s\,\partial\theta_r}\ln f(X;\theta)\right]$$

$$= \mathrm{E}\left[\frac{\partial}{\partial\theta_r}l(\theta)\frac{\partial}{\partial\theta_s}l(\theta)\right] = n\mathrm{E}\left[\frac{\partial}{\partial\theta_r}\ln f(X;\theta)\frac{\partial}{\partial\theta_s}\ln f(X;\theta)\right].$$

The first expression on each line is always correct. The second expression assumes that the likelihood is the product of n identical densities. This matrix is often called the **information matrix**. The information matrix also forms the Cramér–Rao lower bound, as developed in Section 10.2.2.2. That is, under the usual conditions, no unbiased estimator has a smaller variance than that given by the inverse of the information. Therefore, at least asymptotically, no unbiased estimator is more accurate than the mle.

■ **EXAMPLE 11.7**

Estimate the covariance matrix of the mle for the lognormal distribution. Then apply this result to Data Set B.

The likelihood function and its logarithm are

$$L(\mu,\sigma) = \prod_{j=1}^n \frac{1}{x_j\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln x_j - \mu)^2}{2\sigma^2}\right],$$

$$l(\mu,\sigma) = \sum_{j=1}^n\left[-\ln x_j - \ln\sigma - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\left(\frac{\ln x_j - \mu}{\sigma}\right)^2\right].$$

The first partial derivatives are

$$\frac{\partial l}{\partial\mu} = \sum_{j=1}^n\frac{\ln x_j - \mu}{\sigma^2} \text{ and } \frac{\partial l}{\partial\sigma} = -\frac{n}{\sigma} + \sum_{j=1}^n\frac{(\ln x_j - \mu)^2}{\sigma^3}.$$

The second partial derivatives are

$$\frac{\partial^2 l}{\partial\mu^2} = -\frac{n}{\sigma^2},$$

$$\frac{\partial^2 l}{\partial\sigma\,\partial\mu} = -2\sum_{j=1}^n\frac{\ln x_j - \mu}{\sigma^3},$$

$$\frac{\partial^2 l}{\partial\sigma^2} = \frac{n}{\sigma^2} - 3\sum_{j=1}^n\frac{(\ln x_j - \mu)^2}{\sigma^4}.$$

[7]For any multivariate random variable, the covariance matrix has the variances of the individual random variables on the main diagonal and covariances in the off-diagonal positions.

The expected values are ($\ln X_j$ has a normal distribution with mean μ and standard deviation σ)

$$E\left(\frac{\partial^2 l}{\partial \mu^2}\right) = -\frac{n}{\sigma^2},$$

$$E\left(\frac{\partial^2 l}{\partial \mu\, \partial \sigma}\right) = 0,$$

$$E\left(\frac{\partial^2 l}{\partial \sigma^2}\right) = -\frac{2n}{\sigma^2}.$$

Changing the signs and inverting produces an estimate of the covariance matrix (it is an estimate because Theorem 11.4 only provides the covariance matrix in the limit). It is

$$\begin{bmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{\sigma^2}{2n} \end{bmatrix}.$$

For the lognormal distribution, the maximum likelihood estimates are the solutions to the two equations

$$\sum_{j=1}^{n} \frac{\ln x_j - \mu}{\sigma^2} = 0 \quad \text{and} \quad -\frac{n}{\sigma} + \sum_{j=1}^{n} \frac{(\ln x_j - \mu)^2}{\sigma^3} = 0.$$

From the first equation, $\hat{\mu} = (1/n)\sum_{j=1}^{n} \ln x_j$, and from the second equation, $\hat{\sigma}^2 = (1/n)\sum_{j=1}^{n}(\ln x_j - \hat{\mu})^2$. For Data Set B, the values are $\hat{\mu} = 6.1379$ and $\hat{\sigma}^2 = 1.9305$, or $\hat{\sigma} = 1.3894$. With regard to the covariance matrix, the true values are needed. The best we can do is substitute the estimated values to obtain

$$\widehat{\text{Var}}(\hat{\mu}, \hat{\sigma}) = \begin{bmatrix} 0.0965 & 0 \\ 0 & 0.0483 \end{bmatrix}. \tag{11.1}$$

The multiple "hats" in the expression indicate that this is an estimate of the variance of the estimators. □

The zeros off the diagonal indicate that the two parameter estimators are asymptotically uncorrelated. For the particular case of the lognormal distribution, the estimators are uncorrelated for any sample size. One thing we could do with this information is construct approximate 95% confidence intervals for the true parameter values. These would be 1.96 standard deviations on either side of the estimate:

$$\mu: \quad 6.1379 \pm 1.96(0.0965)^{1/2} = 6.1379 \pm 0.6089,$$

$$\sigma: \quad 1.3894 \pm 1.96(0.0483)^{1/2} = 1.3894 \pm 0.4308.$$

To obtain the information matrix, it is necessary to take both derivatives and expected values, which is not always easy to do. A way to avoid this problem is to simply not take the expected value. Rather than working with the number that results from the expectation, use the observed data points. The result is called the **observed information**.

■ **EXAMPLE 11.8**

Estimate the covariance in Example 11.7 using the observed information.

Substitution of the observations into the second derivatives produces

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} = -\frac{20}{\sigma^2},$$

$$\frac{\partial^2 l}{\partial \sigma \, \partial \mu} = -2 \sum_{j=1}^{n} \frac{\ln x_j - \mu}{\sigma^3} = -2 \frac{122.7576 - 20\mu}{\sigma^3},$$

$$\frac{\partial^2 l}{\partial \sigma^2} = \frac{n}{\sigma^2} - 3 \sum_{j=1}^{n} \frac{(\ln x_j - \mu)^2}{\sigma^4} = \frac{20}{\sigma^2} - 3 \frac{792.0801 - 245.5152\mu + 20\mu^2}{\sigma^4}.$$

Inserting the parameter estimates produces the negatives of the entries of the observed information,

$$\frac{\partial^2 l}{\partial \mu^2} = -10.3600, \quad \frac{\partial^2 l}{\partial \sigma \, \partial \mu} = 0, \quad \frac{\partial^2 l}{\partial \sigma^2} = -20.7190.$$

Changing the signs and inverting produces the same values as in (11.1). This is a feature of the lognormal distribution that need not hold for other models. □

Sometimes it is not even possible to take the derivative. In that case, an approximate second derivative can be used. A reasonable approximation is

$$\frac{\partial^2 f(\theta)}{\partial \theta_i \, \partial \theta_j} \doteq \frac{1}{h_i h_j} [f(\theta + \tfrac{1}{2} h_i \mathbf{e}_i + \tfrac{1}{2} h_j \mathbf{e}_j) - f(\theta + \tfrac{1}{2} h_i \mathbf{e}_i - \tfrac{1}{2} h_j \mathbf{e}_j)$$

$$- f(\theta - \tfrac{1}{2} h_i \mathbf{e}_i + \tfrac{1}{2} h_j \mathbf{e}_j) + f(\theta - \tfrac{1}{2} h_i \mathbf{e}_i - \tfrac{1}{2} h_j \mathbf{e}_j)],$$

where $\mathbf{e}_i$ is a vector with all zeros except for a 1 in the ith position and $h_i = \theta_i / 10^v$, where v is one-third the number of significant digits used in calculations.

■ **EXAMPLE 11.9**

Repeat Example 11.8 using approximate derivatives.

Assume that there are 15 significant digits being used. Then, $h_1 = 6.1379/10^5$ and $h_2 = 1.3894/10^5$. Reasonably close values are 0.00006 and 0.00001. The first approximation is

$$\frac{\partial^2 l}{\partial \mu^2} \doteq \frac{l(6.13796, 1.3894) - 2l(6.1379, 1.3894) + l(6.13784, 1.3894)}{(0.00006)^2}$$

$$= \frac{-157.71389308198 - 2(-157.71389304968) + (-157.71389305468)}{(0.00006)^2}$$

$$= -10.3604.$$

The other two approximations are

$$\frac{\partial^2 l}{\partial \sigma \, \partial \mu} \doteq 0.0003 \quad \text{and} \quad \frac{\partial^2 l}{\partial \sigma^2} \doteq -20.7208.$$

We see that here the approximation works very well. □

11.5.1 Exercises

11.33 Determine 95% confidence intervals for the parameters of exponential and gamma models for Data Set B. The likelihood function and maximum likelihood estimates were determined in Example 11.2.

11.34 Let X have a uniform distribution on the interval from 0 to θ. Show that the maximum likelihood estimator is $\hat{\theta} = \max(X_1, \ldots, X_n)$. Use Examples 10.5 and 10.9 to show that this estimator is asymptotically unbiased and to obtain its variance. Show that Theorem 11.4 yields a negative estimate of the variance and that item *(ii)* in the conditions does not hold.

11.35 (*) A distribution has two parameters, α and β. A sample of size 10 produced the following loglikelihood function:

$$l(\alpha, \beta) = -2.5\alpha^2 - 3\alpha\beta - \beta^2 + 50\alpha + 2\beta + k,$$

where k is a constant. Estimate the covariance matrix of the mle $(\hat{\alpha}, \hat{\beta})$.

11.36 (*) A sample of size 40 has been taken from a population with pdf

$$f(x) = (2\pi\theta)^{-1/2} e^{-x^2/(2\theta)}, \quad -\infty < x < \infty, \ \theta > 0.$$

The mle of θ is $\hat{\theta} = 2$. Approximate the MSE of $\hat{\theta}$.

11.37 Four observations were made from a random variable having the density function $f(x) = 2\lambda x e^{-\lambda x^2}$, $x, \lambda > 0$. Exactly one of the four observations was less than 2.

    (a) (*) Determine the mle of λ.

    (b) Approximate the variance of the mle of λ.

11.38 Consider a random sample of size n from a Weibull distribution. For this exercise, write the Weibull survival function as

$$S(x) = \exp\left\{-\left[\frac{\Gamma(1 + \tau^{-1})x}{\mu}\right]^\tau\right\}.$$

For this exercise, assume that τ is known and that only μ is to be estimated.

    (a) Show that $E(X) = \mu$.

    (b) Show that the maximum likelihood estimate of μ is

$$\hat{\mu} = \Gamma(1 + \tau^{-1})\left(\frac{1}{n}\sum_{j=1}^n x_j^\tau\right)^{1/\tau}.$$

    (c) Show that using the observed information produces the variance estimate

$$V\hat{a}r(\hat{\mu}) = \frac{\hat{\mu}^2}{n\tau^2},$$

where μ is replaced by $\hat{\mu}$.

(d) Show that using the information (again replacing μ with $\hat{\mu}$) produces the same variance estimate as in part (c).

(e) Show that $\hat{\mu}$ has a transformed gamma distribution with $\alpha = n, \theta = \mu n^{-1/\tau}$, and $\tau = \tau$. Use this result to obtain the exact variance of $\hat{\mu}$ (as a function of μ).

Hint: The variable X^τ has an exponential distribution, and so the variable $\sum_{j=1}^{n} X_j^\tau$ has a gamma distribution with first parameter equal to n and second parameter equal to the mean of the exponential distribution.

11.6 Functions of Asymptotically Normal Estimators

We are often more interested in a quantity that is a function of the parameters. For example, we might be interested in the lognormal mean as an estimate of the population mean. That is, we want to use $\exp(\hat{\mu} + \hat{\sigma}^2/2)$ as an estimate of the population mean, where the maximum likelihood estimates of the parameters are used. It not trivial to evaluate the mean and variance of this random variable, because it is a function of two variables with different distributions. The following theorem (from [104]) provides an approximate solution. The method is often called the **delta method**.

Theorem 11.5 *Let* $\mathbf{X}_n = (X_{1n}, \ldots, X_{kn})^T$ *be a multivariate random variable of dimension k based on a sample of size n. Assume that* $\mathbf{X}$ *is asymptotically normal with mean* θ *and covariance matrix* $\boldsymbol{\Sigma}/n$, *where neither* θ *nor* $\boldsymbol{\Sigma}$ *depend on n. Let g be a function of k variables that is totally differentiable. Let* $G_n = g(X_{1n}, \ldots, X_{kn})$. *Then* G_n *is asymptotically normal with mean* $g(\theta)$ *and variance* $(\mathbf{A}^T \boldsymbol{\Sigma} \mathbf{A})/n$, *where* $\mathbf{A}$ *is the vector of first derivatives of g, that is,* $\mathbf{A} = (\partial g/\partial \theta_1, \ldots, \partial g/\partial \theta_k)^T$ *and it is to be evaluated at* θ, *the true parameters of the original random variable.*

The statement of the theorem is hard to decipher. The Xs are the estimators and g is the function of the parameters that are being estimated. For a model with one parameter, the theorem reduces to the following statement. Let $\hat{\theta}$ be an estimator of θ that has an asymptotic normal distribution with mean θ and variance $\text{Var}(\hat{\theta})$. Then, $g(\hat{\theta})$ has an asymptotic normal distribution with mean $g(\theta)$ and asymptotic variance $[g'(\theta)]\text{Var}(\hat{\theta})[g'(\theta)] = g'(\theta)^2\text{Var}(\hat{\theta})$. To obtain a useful numerical result, any parameters that appear will usually be replaced by an estimate.

Note that the theorem does not specify the type of estimator used. All that is required is that the estimator have an asymptotic multivariate normal distribution. For maximum likelihood estimators, under the usual regularity conditions, this holds and the information matrix can be used to estimate the asymptotic variance.

■ **EXAMPLE 11.10**

Use the delta method to approximate the variance of the mle of the probability that an observation from an exponential distribution exceeds 200. Apply this result to Data Set B.

From Example 11.2, we know that the maximum likelihood estimate of the exponential parameter is the sample mean. We are asked to estimate $p = \text{Pr}(X > 200) = \exp(-200/\theta)$. The maximum likelihood estimate is $\hat{p} = \exp(-200/\hat{\theta}) = \exp(-200/\bar{x})$. It is not easy to determine the mean and variance of this quantity, but

we do know that $\text{Var}(\bar{X}) = \text{Var}(X)/n = \theta^2/n$. Furthermore,

$$g(\theta) = e^{-200/\theta}, \quad g'(\theta) = 200\theta^{-2}e^{-200/\theta},$$

and therefore the delta method gives

$$\text{Var}(\hat{p}) \doteq \frac{(200\theta^{-2}e^{-200/\theta})^2\theta^2}{n} = \frac{40{,}000\theta^{-2}e^{-400/\theta}}{n}.$$

For Data Set B,

$$\bar{x} = 1{,}424.4,$$

$$\hat{p} = \exp\left(-\frac{200}{1{,}424.4}\right) = 0.86900,$$

$$\widehat{\text{Var}}(\hat{p}) = \frac{40{,}000(1{,}424.4)^{-2}\exp(-400/1{,}424.4)}{20} = 0.0007444.$$

A 95% confidence interval for p is $0.869 \pm 1.96\sqrt{0.0007444}$, or 0.869 ± 0.053.   □

■ EXAMPLE 11.11

Construct a 95% confidence interval for the mean of a lognormal population using Data Set B. Compare this to the more traditional confidence interval based on the sample mean.

From Example 11.7, we have $\hat{\mu} = 6.1379$, $\hat{\sigma} = 1.3894$, and an estimated covariance matrix of

$$\frac{\hat{\Sigma}}{n} = \begin{bmatrix} 0.0965 & 0 \\ 0 & 0.0483 \end{bmatrix}.$$

The function is $g(\mu, \sigma) = \exp(\mu + \sigma^2/2)$. The partial derivatives are

$$\frac{\partial g}{\partial \mu} = \exp\left(\mu + \tfrac{1}{2}\sigma^2\right),$$

$$\frac{\partial g}{\partial \sigma} = \sigma\exp\left(\mu + \tfrac{1}{2}\sigma^2\right),$$

and the estimates of these quantities are 1,215.75 and 1,689.16, respectively. The delta method produces the following approximation:

$$\widehat{\text{Var}}[g(\hat{\mu}, \hat{\sigma})] = \begin{bmatrix} 1{,}215.75 & 1{,}689.16 \end{bmatrix}\begin{bmatrix} 0.0965 & 0 \\ 0 & 0.0483 \end{bmatrix}\begin{bmatrix} 1{,}215.75 \\ 1{,}689.16 \end{bmatrix}$$

$$= 280{,}444.$$

The confidence interval is $1{,}215.75 \pm 1.96\sqrt{280{,}444}$, or $1{,}215.75 \pm 1{,}037.96$.

The customary confidence interval for a population mean is $\bar{x} \pm 1.96s/\sqrt{n}$, where s^2 is the sample variance. For Data Set B, the interval is $1{,}424.4 \pm 1.96(3{,}435.04)/\sqrt{20}$, or $1{,}424.4 \pm 1{,}505.47$. It is not surprising that this is a wider interval because we know that (for a lognormal population) the mle is asymptotically UMVUE.   □

11.6.1 Exercises

11.39 Use the delta method to construct a 95% confidence interval for the mean of a gamma distribution using Data Set B. Preliminary calculations are in Exercise 11.33.

11.40 (*) For a lognormal distribution with parameters μ and σ, you are given that the maximum likelihood estimates are $\hat{\mu} = 4.215$ and $\hat{\sigma} = 1.093$. The estimated covariance matrix of $(\hat{\mu}, \hat{\sigma})$ is

$$\begin{bmatrix} 0.1195 & 0 \\ 0 & 0.0597 \end{bmatrix}.$$

The mean of a lognormal distribution is given by $\exp(\mu + \sigma^2/2)$. Estimate the variance of the maximum likelihood estimator of the mean of this lognormal distribution using the delta method.

11.41 This is a continuation of Exercise 11.6. Let $x_1, \ldots, x_n$ be a random sample from a population with cdf $F(x) = x^p$, $0 < x < 1$.

    (a) Determine the asymptotic variance of the maximum likelihood estimate of p.

    (b) Use your answer to obtain a general formula for a 95% confidence interval for p.

    (c) Determine the mle of $E(X)$ and obtain its asymptotic variance and a formula for a 95% confidence interval.

11.42 This is a continuation of Exercise 11.9. Let $x_1, \ldots, x_n$ be a random sample from a population with pdf $f(x) = \theta^{-1} e^{-x/\theta}$, $x > 0$.

    (a) Determine the asymptotic variance of the mle of θ.

    (b) (*) Use your answer to obtain a general formula for a 95% confidence interval for θ.

    (c) Determine the mle of $\text{Var}(X)$ and obtain its asymptotic variance and a formula for a 95% confidence interval.

11.43 (*) Losses have an exponential distribution. Five observations from this distribution are 100, 200, 400, 800, 1,400, and 3,100. Use the delta method to approximate the variance of the mle of $S(1,500)$. Then construct a symmetric 95% confidence interval for the true value.

11.44 Estimate the covariance matrix of the mles for the data in Exercise 11.7 with both α and θ unknown. Do so by computing approximate derivatives of the loglikelihood. Then construct a 95% confidence interval for the mean.

11.45 Estimate the variance of the mle for Exercise 11.12 and use it to construct a 95% confidence interval for $E(X \wedge 500)$.

11.7   Nonnormal Confidence Intervals

Section 11.5 created confidence intervals based on two assumptions. The first was that the normal distribution is a reasonable approximation of the true distribution of the maximum likelihood estimator. We know that this assumption is asymptotically true but may not hold for small or even moderate samples. Second, it was assumed that when there is more than one parameter, separate confidence intervals should be constructed for each parameter. Separate intervals can be used in cases such as the lognormal distribution where the parameter estimates are independent, but in most cases that is not true. When there is high correlation, it is better to postulate a confidence region, which could be done using the asymptotic covariances and a multivariate normal distribution. However, there is an easier method that does not require a normal distribution assumption (though it is still based on asymptotic results).

One way to motivate a confidence region is to consider the meaning of the likelihood function. The parameter value that maximizes this function is our best choice. It is reasonable that values of the parameter which produce likelihood function values close to the maximum are good alternative choices for the true parameter value. Thus, for some choice of c, a confidence region for the parameter might be

$$\{\theta : l(\theta) \geq c\},$$

the set of all parameters for which the loglikelihood exceeds c. The discussion of the likelihood ratio test in Section 15.4.4 confirms that the loglikelihood is the correct function to use and also indicates how c should be selected to produce a $100(1 - \alpha)\%$ confidence region. The value is

$$c = l(\hat{\theta}) - 0.5\chi_\alpha^2,$$

where the first term is the loglikelihood value at the maximum likelihood estimate and the second term is the $1 - \alpha$ percentile from the chi-square distribution, with degrees of freedom equal to the number of estimated parameters.

■   **EXAMPLE 11.12**

Use this method to construct a 95% confidence interval for the parameter of an exponential distribution. Compare the answer to the normal approximation using Data Set B.

We know that $\hat{\theta} = \bar{x}$ and for a sample of size n, $l(\bar{x}) = -n - n \ln \bar{x}$. With one degree of freedom, the 95th percentile of the chi-square distribution is 3.84. The confidence region is

$$\left\{\theta : -\frac{n\bar{x}}{\theta} - n \ln \theta \geq -n - n \ln \bar{x} - 1.92\right\},$$

which must be evaluated numerically. For Data Set B, the equation is

$$-\frac{20(1{,}424.4)}{\theta} - 20 \ln \theta \geq -20 - 20 \ln(1{,}424.4) - 1.92,$$

$$-\frac{28{,}488}{\theta} - 20 \ln \theta \geq -167.15,$$

and the solution is $946.85 \leq \theta \leq 2{,}285.05$.

For the normal approximation, the asymptotic variance of the maximum likelihood estimate is θ^2/n, which happens to be the true variance. Inserting sample values, the normal confidence interval is

$$1{,}424.4 \pm 1.96\sqrt{1{,}424.4^2/20},$$

$$1{,}424.4 \pm 624.27,$$

which is $800.14 \le \theta \le 2{,}048.76$. Note that the widths of the two intervals are similar, but the first one is not symmetric about the sample mean. This asymmetry is reasonable in that a sample of size 20 is unlikely to be large enough to have the sample mean remove the skewness of the underlying exponential distribution. □

The extension to two parameters is similar, as illustrated in Example 11.13.

■ **EXAMPLE 11.13**

In Example 11.2, the maximum likelihood estimates for a gamma model for Data Set B were $\hat\alpha = 0.55616$ and $\hat\theta = 2{,}561.1$. Determine a 95% confidence region for the true values.

The region consists of all pairs (α, θ) that satisfy

$$(\alpha - 1) \sum_{j=1}^{20} \ln x_j - \frac{1}{\theta} \sum_{j=1}^{20} x_j - 20 \ln \Gamma(\alpha) - 20\alpha \ln \theta$$

$$\ge (0.55616 - 1) \sum_{j=1}^{20} \ln x_j - \frac{1}{2{,}561.1} \sum_{j=1}^{20} x_j - 20 \ln \Gamma(0.55616)$$

$$- 20(0.55616) \ln 2{,}561.1 - 2.996 = -165.289,$$

where 2.996 is one-half of the 95th percentile of a chi-square distribution with two degrees of freedom. Figure 11.1 shows the resulting confidence region. If the normal approximation were appropriate, this region would be elliptical in shape. □

For functions of parameters, the same method can be applied as illustrated in Example 11.14.

■ **EXAMPLE 11.14**

Determine a 95% confidence interval for the mean of the gamma distribution in Example 11.13.

First, reparameterize the gamma density so that the mean is a parameter, which can be done by setting $\mu = \alpha\theta$ and leaving α unchanged. The density function is now

$$f(x) = \frac{x^{\alpha-1} e^{-x\alpha/\mu}}{\Gamma(\alpha)(\mu/\alpha)^\alpha}.$$

Due to the invariance of mles, we have $\hat\mu = \hat\alpha\hat\theta = 1{,}424.4$. We then look for alternative μ values that produce loglikelihood values that are within 1.92 of the
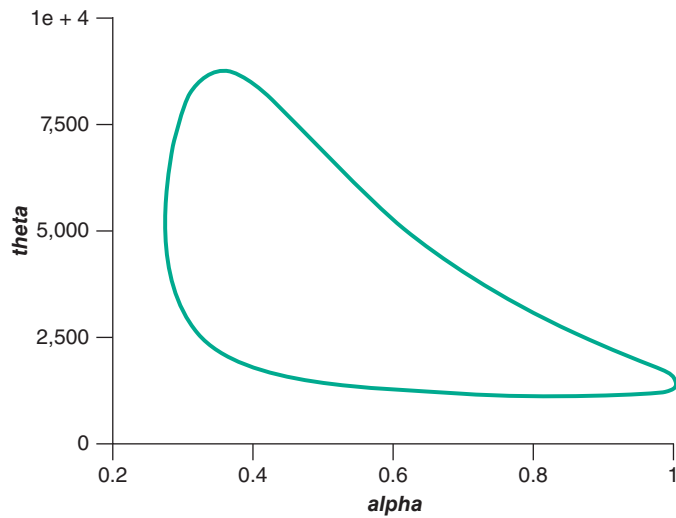
Figure 11.1 The 95% confidence region for the gamma parameters.

maximum (there is only one degree of freedom because there is only one parameter, the mean, being evaluated). When we try a μ-value, to give it the best chance to be accepted, the accompanying α-value should be the one that maximizes the likelihood given μ. Numerical maximizations and trial and error reveal the confidence interval $811 \leq \mu \leq 2{,}846$. ☐

11.7.1 Exercise

11.46 Using an exponential model, Data Set B, and the method of this section, determine a 95% confidence interval for the probability that an observation exceeds 200. Compare your answer to that from Example 11.10.

12

FREQUENTIST ESTIMATION FOR DISCRETE DISTRIBUTIONS

There are special considerations that apply when estimating parameters for discrete distributions. They are discussed in this chapter.

12.1 The Poisson Distribution

The principles of estimation discussed in Chapters 10 and 11 can be applied to frequency distributions. We now illustrate the methods of estimation by fitting a Poisson model.

■ **EXAMPLE 12.1**

A hospital liability policy has experienced the number of claims over a 10-year period given in Table 12.1. Estimate the Poisson parameter using the method of moments and the method of maximum likelihood.

These data can be summarized in a different way. We can count the number of years in which exactly zero claims occurred, one claim occurred, and so on, as in Table 12.2.

Table 12.1   The number of hospital liability claims by year.

Year	Number of claims
1985	6
1986	2
1987	3
1988	0
1989	2
1990	1
1991	2
1992	5
1993	1
1994	3

Table 12.2   Hospital liability claims by frequency.

Frequency (k)	Number of observations (n_k)
0	1
1	2
2	3
3	2
4	0
5	1
6	1
7+	0

The total number of claims for the period 1985–1994 is 25. Hence, the average number of claims per year is 2.5. The average can also be computed from Table 12.2. Let n_k denote the number of years in which a frequency of exactly k claims occurred. The expected frequency (sample mean) is

$$\bar{x} = \frac{\sum_{k=0}^{\infty} k n_k}{\sum_{k=0}^{\infty} n_k},$$

where n_k represents the number of observed values at frequency k. Hence the method of moments estimate of the Poisson parameter is $\hat{\lambda} = 2.5$.

Maximum likelihood estimation can easily be carried out on these data. The likelihood contribution of an observation of k is p_k. Then, the likelihood for the entire set of observations is

$$L = \prod_{k=0}^{\infty} p_k^{n_k}$$

and the loglikelihood is

$$l = \sum_{k=0}^{\infty} n_k \ln p_k.$$

The likelihood and loglikelihood functions are considered to be functions of the unknown parameters. In the case of the Poisson distribution, there is only one parameter, making the maximization easy.

For the Poisson distribution,

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}$$

and

$$\ln p_k = -\lambda + k \ln \lambda - \ln k!.$$

The loglikelihood is

$$l = \sum_{k=0}^{\infty} n_k(-\lambda + k \ln \lambda - \ln k!)$$

$$= -\lambda n + \sum_{k=0}^{\infty} k\, n_k \ln \lambda - \sum_{k=0}^{\infty} n_k \ln k!,$$

where $n = \sum_{k=0}^{\infty} n_k$ is the sample size. Differentiating the loglikelihood with respect to λ, we obtain

$$\frac{dl}{d\lambda} = -n + \sum_{k=0}^{\infty} k\, n_k \frac{1}{\lambda}.$$

By setting the derivative of the loglikelihood to zero, the maximum likelihood estimate is obtained as the solution of the resulting equation. The estimator is then

$$\hat{\lambda} = \frac{\sum_{k=0}^{\infty} kn_k}{n} = \bar{x}$$

and, thus, for the Poisson distribution, the maximum likelihood and the method of moments estimators are identical.

If N has a Poisson distribution with mean λ, then

$$E(\hat{\lambda}) = E(N) = \lambda$$

and

$$\text{Var}(\hat{\lambda}) = \frac{\text{Var}(N)}{n} = \frac{\lambda}{n}.$$

Hence, $\hat{\lambda}$ is unbiased and consistent. From Theorem 11.4, the mle is asymptotically normally distributed with mean λ and variance

$$\text{Var}(\hat{\lambda}) = \left\{ -n\text{E}\left[\frac{d^2}{d\lambda^2} \ln p_N \right] \right\}^{-1}$$

$$= \left\{ -n\text{E}\left[\frac{d^2}{d\lambda^2}(-\lambda + N \ln \lambda - \ln N!) \right] \right\}^{-1}$$

$$= \left[n\text{E}(N/\lambda^2) \right]^{-1}$$

$$= \left(n\lambda^{-1} \right)^{-1} = \frac{\lambda}{n}.$$

Table 12.3   The data for Example 12.2.

Number of claims/day	Observed number of days
0	47
1	97
2	109
3	62
4	25
5	16
6+	9

In this case, the asymptotic approximation to the variance is equal to its true value. From this information, we can construct an approximate 95% confidence interval for the true value of the parameter. The interval is $\hat{\lambda} \pm 1.96(\hat{\lambda}/n)^{1/2}$. For this example, the interval becomes $(1.52, 3.48)$. This confidence interval is only an approximation because it relies on large-sample theory. The sample size is very small and such a confidence interval should be used with caution. $\square$

The formulas presented so far have assumed that the counts at each observed frequency are known. Occasionally, data are collected so that all these counts are not given. The most common example is to have a final entry given as $k+$, where the count is the number of times k or more claims were observed. If n_{k+} is that number of times, the contribution to the likelihood function is

$$(p_k + p_{k+1} + \cdots)^{n_{k+}} = (1 - p_0 - \cdots - p_{k-1})^{n_{k+}}.$$

The same adjustments apply to grouped frequency data of any kind. Suppose that there were five observations at frequencies 3–5. The contribution to the likelihood function would be

$$(p_3 + p_4 + p_5)^5.$$

■ **EXAMPLE 12.2**

For the data in Table 12.3, determine the maximum likelihood estimate for the Poisson distribution.

The likelihood function is

$$L = p_0^{47} p_1^{97} p_2^{109} p_3^{62} p_4^{25} p_5^{16}(1 - p_0 - p_1 - p_2 - p_3 - p_4 - p_5)^9,$$

and when written as a function of λ, it becomes somewhat complicated. While the derivative can be taken, solving the equation when it is set equal to zero requires numerical methods. It may be just as easy to use a numerical method to directly maximize the function. A reasonable starting value can be obtained by assuming that all nine observations were exactly at 6 and then using the sample mean. Of course, this assumption will understate the true maximum likelihood estimate, but should be a good place to start. For this particular example, the maximum likelihood estimate is $\hat{\lambda} = 2.0226$, which is very close to the value obtained when all the counts were recorded. $\square$

12.2   The Negative Binomial Distribution

The moment equations are

$$r\beta = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \bar{x} \tag{12.1}$$

and

$$r\beta(1 + \beta) = \frac{\sum_{k=0}^{\infty} k^2 n_k}{n} - \left(\frac{\sum_{k=0}^{\infty} k n_k}{n} \right)^2 = s^2, \tag{12.2}$$

with solutions $\hat{\beta} = (s^2/\bar{x}) - 1$ and $\hat{r} = \bar{x}/\hat{\beta}$. Note that this variance estimate is obtained by dividing by n, not $n - 1$. This is a common, though not required, approach when using the method of moments. Also note that, if $s^2 < \bar{x}$, the estimate of β will be negative, an inadmissible value.

■ **EXAMPLE 12.3**

(*Example 12.1 continued*) Estimate the negative binomial parameters by the method of moments.

The sample mean and the sample variance are 2.5 and 3.05 (verify this), respectively, and the estimates of the parameters are $\hat{r} = 11.364$ and $\hat{\beta} = 0.22$. □

When compared to the Poisson distribution with the same mean, it can be seen that β is a measure of "extra-Poisson" variation. A value of $\beta = 0$ means no extra-Poisson variation, while a value of $\beta = 0.22$ implies a 22% increase in the variance when compared to the Poisson distribution with the same mean.

We now examine maximum likelihood estimation. The loglikelihood for the negative binomial distribution is

$$
\begin{aligned}
l &= \sum_{k=0}^{\infty} n_k \ln p_k \\
&= \sum_{k=0}^{\infty} n_k \left[\ln \binom{r + k - 1}{k} - r \ln(1 + \beta) + k \ln \beta - k \ln(1 + \beta) \right].
\end{aligned}
$$

The loglikelihood is a function of the two parameters β and r. To find the maximum of the loglikelihood, we differentiate with respect to each of the parameters, set the derivatives equal to zero, and solve for the parameters. The derivatives of the loglikelihood are

$$\frac{\partial l}{\partial \beta} = \sum_{k=0}^{\infty} n_k \left(\frac{k}{\beta} - \frac{r + k}{1 + \beta} \right) \tag{12.3}$$

and

$$\frac{\partial l}{\partial r} = -\sum_{k=0}^{\infty} n_k \ln(1 + \beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \ln \frac{(r + k - 1) \cdots r}{k!}$$

$$= -n \ln(1 + \beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \ln \prod_{m=0}^{k-1} (r + m)$$

$$= -n \ln(1 + \beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \sum_{m=0}^{k-1} \ln(r + m)$$

$$= -n \ln(1 + \beta) + \sum_{k=1}^{\infty} n_k \sum_{m=0}^{k-1} \frac{1}{r + m}. \tag{12.4}$$

Setting these equations to zero yields

$$\hat{\mu} = \hat{r}\hat{\beta} = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \bar{x} \tag{12.5}$$

and

$$n \ln(1 + \hat{\beta}) = \sum_{k=1}^{\infty} n_k \left(\sum_{m=0}^{k-1} \frac{1}{\hat{r} + m} \right). \tag{12.6}$$

Note that the mle of the mean is the sample mean (as, by definition, in the method of moments). Equations (12.5) and (12.6) can be solved numerically. Replacing $\hat{\beta}$ in (12.6) by $\hat{\mu}/\hat{r} = \bar{x}/\hat{r}$ yields the equation

$$H(\hat{r}) = n \ln \left(1 + \frac{\bar{x}}{\hat{r}} \right) - \sum_{k=1}^{\infty} n_k \left(\sum_{m=0}^{k-1} \frac{1}{\hat{r} + m} \right) = 0. \tag{12.7}$$

If the right-hand side of (12.2) is greater than the right-hand side of (12.1), it can be shown that there is a unique solution of (12.7). If not, then the negative binomial model is probably not a good model to use, because the sample variance does not exceed the sample mean.[1]

Equation (12.7) can be solved numerically for $\hat{r}$ using the Newton–Raphson method. The required equation for the kth iteration is

$$r_k = r_{k-1} - \frac{H(r_{k-1})}{H'(r_{k-1})}.$$

A useful starting value for r_0 is the moment-based estimator of r. Of course, any numerical root-finding method (e.g. bisection, secant) may be used.

The loglikelihood is a function of two variables and can be maximized numerically. For the case of the negative binomial distribution with complete data, because we know that the estimator of the mean must be the sample mean, setting $\beta = \bar{x}/r$ reduces this to a one-dimensional problem.

---

[1] In other words, when the sample variance is less than or equal to the mean, the loglikelihood function will not have a maximum. The function will keep increasing as r goes to infinity and β goes to zero with their product remaining constant. This effectively says that the negative binomial distribution that best matches the data is the Poisson distribution that is a limiting case.

Table 12.4 Two models for the automobile claims frequency.

Number of claims/year	Number of drivers	Poisson expected	Negative binomial expected
0	20,592	20,420.9	20,596.8
1	2,651	2,945.1	2,631.0
2	297	212.4	318.4
3	41	10.2	37.8
4	7	0.4	4.4
5	0	0.0	0.5
6	1	0.0	0.1
7+	0	0.0	0.0
Parameters		$\lambda = 0.144220$	$r = 1.11790$
			$\beta = 0.129010$
Loglikelihood		$-10,297.84$	$-10,223.42$

■ **EXAMPLE 12.4**

Determine the maximum likelihood estimates of the negative binomial parameters for the data in Example 12.1.

The maximum occurs at $\hat{r} = 10.9650$ and $\hat{\beta} = 0.227998$. □

■ **EXAMPLE 12.5**

Tröbliger [121] studied the driving habits of 23,589 automobile drivers in a class of automobile insurance by counting the number of accidents per driver in a one-year time period. The data, as well as fitted Poisson and negative binomial distributions, are given in Table 12.4. Based on the information presented, which distribution appears to provide a better model?

The expected counts are found by multiplying the sample size (23,589) by the probability assigned by the model. It is clear that the negative binomial probabilities produce expected counts that are much closer to those that were observed. In addition, the loglikelihood function is maximized at a significantly higher value. Formal procedures for model selection (including what it means to be *significantly higher*) are discussed in Chapter 15. However, in this case, the superiority of the negative binomial model is apparent. □

12.3 The Binomial Distribution

The binomial distribution has two parameters, m and q. Frequently, the value of m is known and fixed. In this case, only one parameter, q, needs to be estimated. In many insurance situations, q is interpreted as the probability of some event such as death or disability. In such cases, the value of q is usually estimated as

$$\hat{q} = \frac{\text{number of observed events}}{\text{maximum number of possible events}},$$

which is the method of moments estimator when m is known.

In situations where frequency data are in the form of the previous examples in this chapter, the value of the parameter m, the largest possible observation, may be known and fixed or unknown. In any case, m must be no smaller than the largest observation. The loglikelihood is

$$l = \sum_{k=0}^{m} n_k \ln p_k$$
$$= \sum_{k=0}^{m} n_k \left[\ln \binom{m}{k} + k \ln q + (m - k) \ln(1 - q) \right].$$

When m is known and fixed, we need only maximize l with respect to q:

$$\frac{\partial l}{\partial q} = \frac{1}{q} \sum_{k=0}^{m} k n_k - \frac{1}{1 - q} \sum_{k=0}^{m} (m - k) n_k.$$

Setting this expression equal to zero yields

$$\hat{q} = \frac{1}{m} \frac{\sum_{k=0}^{m} k n_k}{\sum_{k=0}^{m} n_k},$$

which is the sample proportion of observed events. For the method of moments, with m fixed, the estimator of q is the same as the mle because the moment equation is

$$mq = \frac{\sum_{k=0}^{m} k n_k}{\sum_{k=0}^{m} n_k}.$$

When m is unknown, the maximum likelihood estimator of q is

$$\hat{q} = \frac{1}{\hat{m}} \frac{\sum_{k=0}^{\infty} k n_k}{\sum_{k=0}^{\infty} n_k}, \tag{12.8}$$

where $\hat{m}$ is the maximum likelihood estimate of m. An easy way to approach the maximum likelihood estimation of m and q is to create a *likelihood profile* for various possible values of m as follows:

Step 1: Start with $\hat{m}$ equal to the largest observation.

Step 2: Obtain $\hat{q}$ using (12.8).

Step 3: Calculate the loglikelihood at these values.

Step 4: Increase $\hat{m}$ by 1.

Step 5: Repeat steps 2–4 until a maximum is found.

As with the negative binomial, there need not be a pair of parameters that maximizes the likelihood function. In particular, if the sample mean is less than or equal to the sample variance, this procedure will lead to ever-increasing loglikelihood values as the value of $\hat{m}$ is increased. Once again, the trend is toward a Poisson model. This phenomenon can be checked out using the data from Example 12.1.

Table 12.5   The number of claims per policy.

Number of claims/policy	Number of policies
0	5,367
1	5,893
2	2,870
3	842
4	163
5	23
6	1
7	1
8+	0

Table 12.6   The binomial likelihood profile.

$\hat{m}$	$\hat{q}$	$-$Loglikelihood
7	0.140775	19,273.56
8	0.123178	19,265.37
9	0.109491	19,262.02
10	0.098542	19,260.98
11	0.089584	19,261.11
12	0.082119	19,261.84

■ **EXAMPLE 12.6**

The numbers of claims per policy during a one-year period for a portfolio of 15,160 insurance policies are given in Table 12.5. Obtain moment-based and maximum likelihood estimators.

The sample mean and variance are 0.985422 and 0.890355, respectively. The variance is smaller than the mean, suggesting the binomial as a reasonable distribution to try. The method of moments leads to

$$mq = 0.985422$$

and

$$mq(1 - q) = 0.890355.$$

Hence, $\hat{q} = 0.096474$ and $\hat{m} = 10.21440$. However, m can only take on integer values. We choose $\hat{m} = 10$ by rounding. Then, we adjust the estimate of $\hat{q}$ to 0.0985422 from the first moment equation. Doing so will result in a model variance that differs from the sample variance because $10(0.0985422)(1 - 0.0985422) = 0.888316$. This difference shows one of the pitfalls of using the method of moments with integer-valued parameters.

We now turn to maximum likelihood estimation. From the data, $m \geq 7$. If m is known, then only q needs to be estimated. If m is unknown, then we can produce a likelihood profile by maximizing the likelihood for fixed values of m starting at 7 and increasing until a maximum is found. The results are shown in Table 12.6.

The largest loglikelihood value occurs at $m = 10$. If, a priori, the value of m is unknown, then the maximum likelihood estimates of the parameters are $\hat{m} = 10$ and $\hat{q} = 0.0985422$. This result is the same as the adjusted moment estimates, but it is not necessarily the case for all data sets. □

12.4 The $(a, b, 1)$ Class

Estimation of the parameters for the $(a, b, 1)$ class follows the same general principles used in connection with the $(a, b, 0)$ class.

Assuming that the data are in the same form as in the previous examples, the likelihood is, using (6.7),

$$L = \left(p_0^M\right)^{n_0} \prod_{k=1}^{\infty} (p_k^M)^{n_k} = \left(p_0^M\right)^{n_0} \prod_{k=1}^{\infty} \left[(1 - p_0^M)p_k^T\right]^{n_k}.$$

The loglikelihood is

$$l = n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k[\ln(1 - p_0^M) + \ln p_k^T]$$

$$= n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k \ln(1 - p_0^M) + \sum_{k=1}^{\infty} n_k[\ln p_k - \ln(1 - p_0)],$$

where the last statement follows from $p_k^T = p_k/(1 - p_0)$. The three parameters of the $(a, b, 1)$ class are p_0^M, a, and b, where a and b determine $p_1, p_2, \dots$.

Then, it can be seen that

$$l = l_0 + l_1,$$

with

$$l_0 = n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k \ln(1 - p_0^M),$$

$$l_1 = \sum_{k=1}^{\infty} n_k[\ln p_k - \ln(1 - p_0)],$$

where l_0 depends only on the parameter p_0^M and l_1 is independent of p_0^M, depending only on a and b. This separation simplifies the maximization because

$$\frac{\partial l}{\partial p_0^M} = \frac{\partial l_0}{\partial p_0^M} = \frac{n_0}{p_0^M} - \sum_{k=1}^{\infty} \frac{n_k}{1 - p_0^M} = \frac{n_0}{p_0^M} - \frac{n - n_0}{1 - p_0^M},$$

resulting in

$$\hat{p}_0^M = \frac{n_0}{n},$$

the proportion of observations at zero. This is the natural estimator because p_0^M represents the probability of an observation of zero.

Similarly, because the likelihood factors conveniently, the estimation of a and b is independent of p_0^M. Note that although a and b are parameters, maximization should

not be done with respect to them because not all values of a and b produce admissible probability distributions.[2] For the zero-modified Poisson distribution, the relevant part of the loglikelihood is

$$l_1 = \sum_{k=1}^{\infty} n_k \left[\ln \frac{e^{-\lambda} \lambda^k}{k!} - \ln(1 - e^{-\lambda}) \right]$$

$$= -(n - n_0)\lambda + \left(\sum_{k=1}^{\infty} k \, n_k \right) \ln \lambda - (n - n_0) \ln(1 - e^{-\lambda}) + c$$

$$= -(n - n_0)[\lambda + \ln(1 - e^{-\lambda})] + n\bar{x} \ln \lambda + c,$$

where $\bar{x} = \frac{1}{n} \sum_{k=0}^{\infty} k n_k$ is the sample mean, $n = \sum_{k=0}^{\infty} n_k$, and c is independent of λ. Hence,

$$\frac{\partial l_1}{\partial \lambda} = -(n - n_0) - (n - n_0)\frac{e^{-\lambda}}{1 - e^{-\lambda}} + n\frac{\bar{x}}{\lambda}$$

$$= -\frac{n - n_0}{1 - e^{-\lambda}} + \frac{n\bar{x}}{\lambda}.$$

Setting this expression equal to zero yields

$$\bar{x}(1 - e^{-\lambda}) = \frac{n - n_0}{n} \lambda. \tag{12.9}$$

By graphing each side as a function of λ, it is clear that, if $n_0 > 0$, there exist exactly two roots: one is $\lambda = 0$ and the other is $\lambda > 0$. Equation (12.9) can be solved numerically to obtain $\hat{\lambda}$. Note that, because $\hat{p}_0^M = n_0/n$ and $p_0 = e^{-\lambda}$, (12.9) can be rewritten as

$$\bar{x} = \frac{1 - \hat{p}_0^M}{1 - p_0} \lambda. \tag{12.10}$$

Because the right-hand side of (12.10) is the theoretical mean of the zero-modified Poisson distribution (when $\hat{p}_0^M$ is replaced with p_0^M), (12.10) is a moment equation. Hence, an alternative estimation method yielding the same results as the maximum likelihood method is to equate p_0^M to the sample proportion at zero and the theoretical mean to the sample mean. This approach suggests that, by fixing the zero probability to the observed proportion at zero and equating the low-order moments, a modified moment method can be used to get starting values for numerical maximization of the likelihood function. Because the maximum likelihood method has better asymptotic properties, it is preferable to use the modified moment method only to obtain starting values.

For the purpose of obtaining estimates of the asymptotic variance of the mle of λ, it is easy to obtain

$$\frac{\partial^2 l_1}{\partial \lambda^2} = (n - n_0)\frac{e^{-\lambda}}{(1 - e^{-\lambda})^2} - \frac{n\bar{x}}{\lambda^2},$$

and the expected value is obtained by observing that $\mathrm{E}(\bar{x}) = (1 - p_0^M)\lambda/(1 - e^{-\lambda})$. Finally, p_0^M may be replaced by its estimator, n_0/n. The variance of $\hat{p}_0^M$ is obtained by observing that the numerator, n_0, has a binomial distribution and, therefore, the variance is $p_0^M(1 - p_0^M)/n$.

---

[2]Maximization can be done with respect to any parameterization because maximum likelihood estimation is invariant under parameter transformations. However, it is more difficult to maximize over bounded regions, because numerical methods are difficult to constrain and analytic methods will fail due to a lack of differentiability. Therefore, estimation is usually done with respect to particular class members, such as the Poisson.

For the zero-modified binomial distribution,

$$
\begin{aligned}
l_1 &= \sum_{k=1}^{m} n_k \left\{ \ln\left[\binom{m}{k} q^k (1-q)^{m-k} \right] - \ln[1-(1-q)^m] \right\} \\
&= \left(\sum_{k=1}^{m} k n_k \right) \ln q + \sum_{k=1}^{m} (m-k) n_k \ln(1-q) \\
&\quad - \sum_{k=1}^{m} n_k \ln[1-(1-q)^m] + c \\
&= n\bar{x} \ln q + m(n-n_0)\ln(1-q) - n\bar{x}\ln(1-q) \\
&\quad - (n-n_0)\ln[1-(1-q)^m] + c,
\end{aligned}
$$

where c does not depend on q and

$$
\frac{\partial l_1}{\partial q} = \frac{n\bar{x}}{q} - \frac{m(n-n_0)}{1-q} + \frac{n\bar{x}}{1-q} - \frac{(n-n_0)m(1-q)^{m-1}}{1-(1-q)^m}.
$$

Setting this expression equal to zero yields

$$
\bar{x} = \frac{1-\hat{p}_0^M}{1-p_0} mq, \tag{12.11}
$$

where we recall that $p_0 = (1-q)^m$. This equation matches the theoretical mean with the sample mean.

If m is known and fixed, the mle of p_0^M is still

$$
\hat{p}_0^M = \frac{n_0}{n}.
$$

However, even with m known, (12.11) must be solved numerically for q. When m is unknown and also needs to be estimated, this procedure can be followed for different values of m until the maximum of the likelihood function is obtained.

The zero-modified negative binomial (or extended truncated negative binomial) distribution is a bit more complicated because three parameters need to be estimated. Of course, the mle of p_0^M is $\hat{p}_0^M = n_0/n$ as before, reducing the problem to the estimation of r and β. The part of the loglikelihood relevant to r and β is

$$
l_1 = \sum_{k=1}^{\infty} n_k \ln p_k - (n-n_0)\ln(1-p_0). \tag{12.12}
$$

Hence,

$$
\begin{aligned}
l_1 &= \sum_{k=1}^{\infty} n_k \ln \left[\binom{k+r-1}{k} \left(\frac{1}{1+\beta} \right)^r \left(\frac{\beta}{1+\beta} \right)^k \right] \\
&\quad - (n-n_0)\ln\left[1 - \left(\frac{1}{1+\beta} \right)^r \right]. \tag{12.13}
\end{aligned}
$$

This function needs to be maximized over the (r, β) plane to obtain the mles, which can be done numerically. Starting values can be obtained by the modified moment method by

setting $\hat{p}_0^M = n_0/n$ and equating the first two moments of the distribution to the first two sample moments. It is generally easier to use raw moments (moments about the origin) than central moments for this purpose. In practice, it may be more convenient to maximize (12.12) rather than (12.13) because one can take advantage of the recursive scheme

$$p_k = p_{k-1}\left(a + \frac{b}{k}\right)$$

in evaluating (12.12). This approach makes computer programming a bit easier.

For zero-truncated distributions, there is no need to estimate the probability at zero because it is known to be zero. The remaining parameters are estimated using the same formulas developed for the zero-modified distributions.

■ **EXAMPLE 12.7**

The data set in Table 12.7 comes from Beard et al. [13]. Determine a model that adequately describes the data.   □

When a Poisson distribution is fitted to it, the resulting fit is very poor. There is too much probability for one accident and two little at subsequent values. The geometric distribution is tried as a one-parameter alternative. It has loglikelihood

$$l = -n\ln(1+\beta) + \sum_{k=1}^{\infty} n_k \ln\left(\frac{\beta}{1+\beta}\right)^k$$

$$= -n\ln(1+\beta) + \sum_{k=1}^{\infty} kn_k[\ln\beta - \ln(1+\beta)]$$

$$= -n\ln(1+\beta) + n\bar{x}[\ln\beta - \ln(1+\beta)]$$

$$= -(n+n\bar{x})\ln(1+\beta) + n\bar{x}\ln\beta,$$

where $\bar{x} = \sum_{k=1}^{\infty} k\,n_k/n$ and $n = \sum_{k=0}^{\infty} n_k$.

Differentiation reveals that the loglikelihood has a maximum at

$$\hat{\beta} = \bar{x}.$$

Table 12.7   The results of fitting distributions to the Beard et al. data.

Accidents	Observed	Poisson	Geometric	ZM Poisson	ZM geom.
0	370,412	369,246.9	372,206.5	370,412.0	370,412.0
1	46,545	48,643.6	43,325.8	46,432.1	46,555.2
2	3,935	3,204.1	5,043.2	4,138.6	3,913.6
3	317	140.7	587.0	245.9	329.0
4	28	4.6	68.3	11.0	27.7
5	3	0.1	8.0	0.4	2.3
6+	0	0.0	1.0	0.0	0.2
Parameters		λ: 0.13174	β: 0.13174	p_0^M: 0.87934	p_0^M: 0.87934
				λ: 0.17827	β: 0.091780
Loglikelihood		−171,373	−171,479	−171,160	−171,133

A qualitative look at the numbers indicates that the zero-modified geometric distribution matches the data better than the other three models considered. A formal analysis is done in Example 15.14.

12.5 Compound Models

For the method of moments, the first few moments can be matched with the sample moments. The system of equations can be solved to obtain the moment-based estimators. Note that the number of parameters in the compound model is the sum of the number of parameters in the primary and secondary distributions. The first two theoretical moments for compound distributions are

$$E(S) = E(N)E(M),$$

$$Var(S) = E(N) Var(M) + E(M)^2 Var(N).$$

These results were developed in Chapter 9. The first three moments for the compound Poisson distribution are given in (7.10).

Maximum likelihood estimation is also carried out as before. The loglikelihood to be maximized is

$$l = \sum_{k=0}^{\infty} n_k \ln g_k.$$

When g_k is the probability of a compound distribution, the loglikelihood can be maximized numerically. The first and second derivatives of the loglikelihood can be obtained by using approximate differentiation methods as applied directly to the loglikelihood function at the maximum value.

■ **EXAMPLE 12.8**

Determine various properties of the Poisson–zero-truncated geometric distribution. This distribution is also called the Polya–Aeppli distribution.

For the zero-truncated geometric distribution, the pgf is

$$P_2(z) = \frac{[1 - \beta(z - 1)]^{-1} - (1 + \beta)^{-1}}{1 - (1 + \beta)^{-1}},$$

and therefore the pgf of the Polya–Aeppli distribution is

$$P(z) = P_1[P_2(z)] = \exp\left(\lambda\left\{\frac{[1 - \beta(z - 1)]^{-1} - (1 + \beta)^{-1}}{1 - (1 + \beta)^{-1}} - 1\right\}\right)$$

$$= \exp\left\{\lambda\frac{[1 - \beta(z - 1)]^{-1} - 1}{1 - (1 + \beta)^{-1}}\right\}.$$

The mean is

$$P'(1) = \lambda(1 + \beta)$$

and the variance is

$$P''(1) + P'(1) - [P'(1)]^2 = \lambda(1 + \beta)(1 + 2\beta).$$

Alternatively, $E(N) = \text{Var}(N) = \lambda$, $E(M) = 1 + \beta$, and $\text{Var}(M) = \beta(1 + \beta)$. Then,

$$E(S) = \lambda(1 + \beta),$$
$$\text{Var}(S) = \lambda\beta(1 + \beta) + \lambda(1 + \beta)^2 = \lambda(1 + \beta)(1 + 2\beta).$$

From Theorem 7.3, the probability at zero is

$$g_0 = P_1(0) = e^{-\lambda}.$$

The successive values of g_k are computed easily using the compound Poisson recursion

$$g_k = \frac{\lambda}{k} \sum_{j=1}^{k} j f_j g_{k-j}, \qquad k = 1, 2, 3, \ldots , \tag{12.14}$$

where $f_j = \beta^{j-1}/(1 + \beta)^j$, $j = 1, 2, \ldots$. For any values of λ and β, the loglikelihood function can easily be evaluated. □

Example 15.15 provides a data set for which the Polya–Aeppli distribution is a good choice.

   Another useful compound Poisson distribution is the Poisson–extended truncated negative binomial (Poisson–ETNB) distribution. Although it does not matter if the secondary distribution is modified or truncated, we prefer the truncated version here, so that the parameter r may be extended.[3] Special cases are $r = 1$, which is the Poisson–geometric (also called Polya–Aeppli); $r \to 0$, which is the Poisson–logarithmic (negative binomial); and $r = -0.5$, which is called the Poisson–inverse Gaussian. This name is not consistent with the others. Here, the inverse Gaussian distribution is a mixing distribution (see Section 7.3). Example 15.16 provides a data set for which the Poisson–inverse Gaussian distribution is a good choice.

12.6   The Effect of Exposure on Maximum Likelihood Estimation

In Section 7.4, the effect of exposure on discrete distributions is discussed. When aggregate data from a large group of insureds is obtained, maximum likelihood estimation is still possible. The following example illustrates this fact for the Poisson distribution.

■   **EXAMPLE 12.9**

Determine the maximum likelihood estimate of the Poisson parameter for the data in Table 12.8.

   Let λ be the Poisson parameter for a single exposure. If year k has e_k exposures, then the number of claims has a Poisson distribution with parameter λe_k. If n_k is the number of claims in year k, the likelihood function is

$$L = \prod_{k=1}^{6} \frac{e^{-\lambda e_k}(\lambda e_k)^{n_k}}{n_k!}.$$

---

[3]This preference does not contradict Theorem 7.4. When $-1 < r < 0$, it is still the case that changing the probability at zero will not produce new distributions. What is true is that there is no probability at zero that will lead to an ordinary $(a, b, 0)$ negative binomial secondary distribution.

<p style="text-align:center">Table 12.8   Automobile claims by year.</p>

Year	Number of Exposures	Number of Claims
1986	2,145	207
1987	2,452	227
1988	3,112	341
1989	3,458	335
1990	3,698	362
1991	3,872	359

The maximum likelihood estimate is found by

$$l = \ln L = \sum_{k=1}^{6}[-\lambda e_k + n_k \ln(\lambda e_k) - \ln(n_k!)],$$

$$\frac{\partial l}{\partial \lambda} = \sum_{k=1}^{6}\left(-e_k + n_k \lambda^{-1}\right) = 0,$$

$$\hat{\lambda} = \frac{\sum_{k=1}^{6} n_k}{\sum_{k=1}^{6} e_k} = \frac{1{,}831}{18{,}737} = 0.09772.$$

In this example, the answer is what we expected it to be: the average number of claims per exposure. This technique will work for any distribution in the $(a, b, 0)^4$ and compound classes. But care must be taken in the interpretation of the model. For example, if we use a negative binomial distribution, we are assuming that each exposure unit produces a number of claims according to a negative binomial distribution. This is different from assuming that the total number of claims has a negative binomial distribution because they arise from individuals who each have a Poisson distribution but with different parameters.

12.7 Exercises

12.1 Assume that the binomial parameter m is known. Consider the mle of q.

(a) Show that the mle is unbiased.

(b) Determine the variance of the mle.

(c) Show that the asymptotic variance as given in Theorem 11.4 is the same as that developed in part (b).

(d) Determine a simple formula for a confidence interval using (10.6) in Section 10.3 that is based on replacing q with $\hat{q}$ in the variance term.

(e) Determine a more complicated formula for a confidence interval using (10.5) that is not based on such a replacement.

---

[4]For the binomial distribution, the usual problem that m must be an integer remains.

Table 12.9   The data for Exercise 12.3.

Number of claims	Number of policies
0	9,048
1	905
2	45
3	2
4+	0

Table 12.10   The data for Exercise 12.4.

Number of claims	Underinsured	Uninsured
0	901	947
1	92	50
2	5	2
3	1	1
4	1	0
5+	0	0

12.2 Use (12.5) to determine the mle of β for the geometric distribution. In addition, determine the variance of the mle and verify that it matches the asymptotic variance as given in Theorem 11.4.

12.3 A portfolio of 10,000 risks produced the claim counts in Table 12.9.

(a) Determine the mle of λ for a Poisson model and then determine a 95% confidence interval for λ.

(b) Determine the mle of β for a geometric model and then determine a 95% confidence interval for β.

(c) Determine the mle of r and β for a negative binomial model.

(d) Assume that $m = 4$. Determine the mle of q of the binomial model.

(e) Construct 95% confidence intervals for q using the methods developed in parts (d) and (e) of Exercise 12.1.

(f) Determine the mle of m and q by constructing a likelihood profile.

12.4 An automobile insurance policy provides benefits for accidents caused by both underinsured and uninsured motorists. Data on 1,000 policies reveal the information in Table 12.10.

(a) Determine the mle of λ for a Poisson model for each of the variables $N_1 = $ number of underinsured claims and $N_2 = $ number of uninsured claims.

(b) Assume that N_1 and N_2 are independent. Use Theorem 6.1 in Section 6.2 to determine a model for $N = N_1 + N_2$.

Table 12.11   The data for Exercise 12.5.

Number of claims	Number of policies
0	861
1	121
2	13
3	3
4	1
5	0
6	1
7+	0

Table 12.12   The data for Exercise 12.6.

Number of prescriptions	Frequency	Number of prescriptions	Frequency
0	82	16–20	40
1–3	49	21–25	38
4–6	47	26–35	52
7–10	47	36–	91
11–15	57		

12.5 An alternative method of obtaining a model for N in Exercise 12.4 would be to record the total number of underinsured and uninsured claims for each of the 1,000 policies. Suppose that this was done and the results were as shown in Table 12.11.

(a) Determine the mle of λ for a Poisson model.

(b) The answer to part (a) matches the answer to part (c) of the previous exercise. Demonstrate that this must always be so.

(c) Determine the mle of β for a geometric model.

(d) Determine the mle of r and β for a negative binomial model.

(e) Assume that $m = 7$. Determine the mle of q of the binomial model.

(f) Determine the mles of m and q by constructing a likelihood profile.

12.6 The data in Table 12.12 represent the number of prescriptions filled in one year for a group of elderly members of a group insurance plan.

(a) Determine the mle of λ for a Poisson model.

(b) Determine the mle of β for a geometric model and then determine a 95% confidence interval for β.

(c) Determine the mle of r and β for a negative binomial model.

12.7 (*) A sample of 3,000 policies contains 1,000 with no claims, 1,200 with one claim, 600 with two claims, and 200 with three claims. Use maximum likelihood estimation and a normal approximation to construct a 90% confidence interval for the mean of a Poisson model.

12.8 (*) A sample of size 10 from a Poisson distribution contains the values 10, 2, 4, 0, 6, 2, 4, 5, 4, and 2. Estimate the coefficient of variation of the mle of the Poisson parameter.

12.9 Consider the data $\{n_1, n_2, \ldots\}$, where n_k is the number of observations equal to k. It is assumed that the data follow a zero-truncated negative binomial distribution.

(a) Assume that r is known. Prove that $\hat{\beta}$, the maximum likelihood estimator of β, is equal to the method of moments estimator. That is, $\hat{\beta}$ satisfies

$$\bar{x} = \frac{r\hat{\beta}}{1 - (1 + \hat{\beta})^{-r}}.$$

Note that (a proof is not expected) if $\bar{x} > 1$, there is a unique value of $\hat{\beta} > 0$ that satisfies this equation. Further note that $\bar{x} \leq 1$ if and only if all the observed values are equal to one.

(b) Suppose that $-1 < r < 0$ and r is rational. That is, $r = -k/m$, where $k < m$ and k and m are positive integers. Prove that

$$\sum_{j=0}^{k-1}(1 + \hat{\beta})^{j/m} = \frac{(1 + \hat{\beta})^{k/m} - 1}{(1 + \hat{\beta})^{1/m} - 1},$$

and hence $\hat{\beta}$ satisfies the polynomial-type equation

$$k \sum_{j=k}^{m-1} \left[(1 + \hat{\beta})^{1/m}\right]^j - (m\bar{x} - k) \sum_{j=0}^{k-1} \left[(1 + \hat{\beta})^{1/m}\right]^j = 0.$$

(c) Prove each of the following three statements:

i. If $r = -1/2$, then $\hat{\beta} = 4\bar{x}(\bar{x} - 1)$.

ii. If $r = -1/3$, then $\hat{\beta} = \left[\frac{\sqrt{3(4\bar{x}-1)}-1}{2}\right]^3 - 1$.

iii. If $r = -2/3$, then $\hat{\beta} = \left[\frac{3\bar{x}-2+\sqrt{3(3\bar{x}-2)(\bar{x}+2)}}{4}\right]^3 - 1$.

(d) Suppose that $r > 0$ and r is rational. That is, $r = k/m$, where k and m are positive integers. Prove, as in (b), that $\hat{\beta}$ satisfies the polynomial-type equation

$$k \sum_{j=k}^{m+k-1} \left[(1 + \hat{\beta})^{1/m}\right]^j - m\bar{x} \sum_{j=0}^{k-1} \left[(1 + \hat{\beta})^{1/m}\right]^j = 0.$$

(e) Prove each of the following three statements:

i. If $r = 1$, then $\hat{\beta} = \bar{x} - 1$.

ii. If $r = 1/2$, then $\hat{\beta} = \left(\frac{\sqrt{8\bar{x}+1}-1}{2}\right)^2 - 1$.

iii. If $r = 2$, then $\hat{\beta} = \frac{\sqrt{\bar{x}(\bar{x}+8)}+\bar{x}-4}{4}$.

13

BAYESIAN ESTIMATION

All of the previous discussion on estimation has assumed a frequentist approach. That is, the population distribution has been fixed but unknown, and our decisions have been concerned not only with the sample we obtained from the population, but also with the possibilities attached to other samples that might have been obtained. The Bayesian approach assumes that only the data actually observed are relevant and it is the population distribution that is variable. For parameter estimation, the following definitions describe the process and then Bayes' theorem provides the solution.

13.1 Definitions and Bayes' Theorem

Definition 13.1 *The **prior distribution** is a probability distribution over the space of possible parameter values. It is denoted $\pi(\theta)$ and represents our opinion concerning the relative chances that various values of θ are the true value of the parameter.*

As before, the parameter θ may be scalar or vector valued. Determination of the prior distribution has always been one of the barriers to the widespread acceptance of Bayesian methods. It is almost certainly the case that your experience has provided some insights

about possible parameter values before the first data point has been observed. (If you have no such opinions, perhaps the wisdom of the person who assigned this task to you should be questioned.) The difficulty is translating this knowledge into a probability distribution. An excellent discussion about prior distributions and the foundations of Bayesian analysis can be found in Lindley [80], and for a discussion about issues surrounding the choice of Bayesian versus frequentist methods, see Efron [33]. The book by Klugman [72] contains more detail on the Bayesian approach, along with several actuarial applications. More recent papers applying Bayesian methods to actuarial problems include deAlba [26], Fellingham, Kottas, and Hartman [39], Meyers [88, 89], Mildenhall [90], Ntzoufras and Dellaportas [95], Scollnik [111], Verrall [125], and Wuthrich [133]. General applications of actuarial interest can be found in Hartman [49] and Hartman, Richardson, and Bateman [50]. For a thorough mathematical treatment of Bayesian methods, a good source is the text by Berger [14]. In recent years, many advancements in Bayesian calculations have taken place. Two good reviews are Carlin and Louis [22] and Gelman et al. [42].

Due to the difficulty of finding a prior distribution that is convincing (you will have to convince others that your prior opinions are valid) and the possibility that you may really have no prior opinion, the definition of prior distribution can be loosened.

Definition 13.2 *An **improper prior distribution** is one for which the probabilities (or pdf) are nonnegative but their sum (or integral) is infinite.*

A great deal of research has gone into the determination of a so-called **noninformative** or **vague** prior. Its purpose is to reflect minimal knowledge. Universal agreement on the best way to construct a vague prior does not exist. However, there is agreement that the appropriate noninformative prior for a scale parameter is $\pi(\theta) = 1/\theta, \theta > 0$. Note that this is an improper prior.

For a Bayesian analysis, the model is no different than before. In our development, we will use *pdf* to represent discrete and mixed distributions in addition to those that are continuous. In the formulas, integrals should be replaced by sums as appropriate.

Definition 13.3 *The **model distribution** is the probability distribution for the data as collected given a particular value for the parameter. Note that this matches Definition 11.1 for the likelihood function. However, consistent with Bayesian notion, the model pdf is denoted $f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)$, where vector notation for $\mathbf{x}$ is used to remind us that all the data appear here.*

Thus, as with maximum likelihood estimation, a necessary step is the ability to write the likelihood function for the given situation. Data that have been truncated or censored can thus be analyzed by Bayesian methods. We use concepts from multivariate statistics to obtain two more definitions.

Definition 13.4 *The **joint distribution** has pdf*

$$f_{\mathbf{X},\Theta}(\mathbf{x}, \theta) = f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi(\theta).$$

Definition 13.5 *The **marginal distribution** of $\mathbf{x}$ has pdf*

$$f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi(\theta)\, d\theta.$$

Note that if there is more than one parameter, this equation will be a multiple integral. Compare this definition to that of a mixture distribution given by (5.2) in Section 5.2.4. The final two quantities of interest are the posterior and predictive distributions.

Definition 13.6 *The **posterior distribution** is the conditional probability distribution of the parameters given the observed data. It is denoted $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$.*

Definition 13.7 *The **predictive distribution** is the conditional probability distribution of a new observation y given the data* $\mathbf{x}$*. It is denoted $f_{Y|\mathbf{X}}(y|\mathbf{x})$.*[1]

These last two items are the key output of a Bayesian analysis. The posterior distribution tells us how our opinion about the parameter has changed once we have observed the data. The predictive distribution tells us what the next observation might look like given the information contained in the data (as well as, implicitly, our prior opinion). Bayes' theorem tells us how to compute the posterior distribution.

Theorem 13.8 *The posterior distribution can be computed as*

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi(\theta)}{\displaystyle\int f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi(\theta)\,d\theta}, \tag{13.1}$$

while the predictive distribution can be computed as

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \int f_{Y|\Theta}(y|\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\,d\theta, \tag{13.2}$$

where $f_{Y|\Theta}(y|\theta)$ *is the pdf of the new observation given the parameter value.*

The predictive distribution can be interpreted as a mixture distribution, where the mixing is with respect to the posterior distribution. Example 13.1 illustrates the preceding definitions and results. The setting is taken from Meyers [87], though the data are not.

■ **EXAMPLE 13.1**

The following amounts were paid on a hospital liability policy:

<div align="center">125   132   141   107   133   319   126   104   145   223</div>

The amount of a single payment has the single-parameter Pareto distribution with $\theta = 100$ and α unknown. The prior distribution has the gamma distribution with $\alpha = 2$ and $\theta = 1$. Determine all of the relevant Bayesian quantities.

The prior has a gamma distribution and the pdf is

$$\pi_A(\alpha) = \alpha e^{-\alpha}, \quad \alpha > 0,$$

---

[1]In this section and in any subsequent Bayesian discussions, we reserve $f(\cdot)$ for distributions concerning observations (such as the model and predictive distributions) and $\pi(\cdot)$ for distributions concerning parameters (such as the prior and posterior distributions). The arguments will usually make it clear which particular distribution is being used. To make matters explicit, we also employ subscripts to enable us to keep track of the random variables.

while the model is (evaluated at the data points)

$$f_{\mathbf{X}|A}(\mathbf{x}|\alpha) = \frac{\alpha^{10}(100)^{10\alpha}}{\left(\prod_{j=1}^{10} x_j^{\alpha+1}\right)} = \alpha^{10} e^{-3.801121\alpha - 49.852823}.$$

The joint pdf of $\mathbf{x}$ and A is (again evaluated at the data points)

$$f_{\mathbf{X},A}(\mathbf{x}, \alpha) = \alpha^{11} e^{-4.801121\alpha - 49.852823}.$$

The posterior pdf of α is

$$\pi_{A|\mathbf{X}}(\alpha|\mathbf{x}) = \frac{\alpha^{11} e^{-4.801121\alpha - 49.852823}}{\int_0^\infty \alpha^{11} e^{-4.801121\alpha - 49.852823} \, d\alpha} = \frac{\alpha^{11} e^{-4.801121\alpha}}{(11!)(1/4.801121)^{12}}. \tag{13.3}$$

There is no need to evaluate the integral in the denominator. Because we know that the result must be a probability distribution, the denominator is just the appropriate normalizing constant. A look at the numerator reveals that we have a gamma distribution with $\alpha = 12$ and $\theta = 1/4.801121$.

The predictive distribution is

$$\begin{aligned}
f_{Y|\mathbf{X}}(y|\mathbf{x}) &= \int_0^\infty \frac{\alpha 100^\alpha}{y^{\alpha+1}} \frac{\alpha^{11} e^{-4.801121\alpha}}{(11!)(1/4.801121)^{12}} \, d\alpha \\
&= \frac{1}{y(11!)(1/4.801121)^{12}} \int_0^\infty \alpha^{12} e^{-(0.195951 + \ln y)\alpha} \, d\alpha \\
&= \frac{1}{y(11!)(1/4.801121)^{12}} \frac{(12!)}{(0.195951 + \ln y)^{13}} \\
&= \frac{12(4.801121)^{12}}{y(0.195951 + \ln y)^{13}}, \quad y > 100. \tag{13.4}
\end{aligned}$$

While this density function may not look familiar, you are asked to show in Exercise 13.1 that $\ln Y - \ln 100$ has a Pareto distribution. $\square$

■ **EXAMPLE 13.2**

(*Example 13.1 continued*) Repeat the example, assuming that the data have been censored at 200.

The model is (evaluated at the data points)

$$f_{\mathbf{X}|A}(\mathbf{x}|\alpha) = \frac{\alpha^8 (100)^{8\alpha}}{\left(\prod_{x_j < 200} x_j^{\alpha+1}\right)} \left(\frac{100}{200}\right)^{2\alpha} = \alpha^8 e^{-3.225393\alpha - 38.680460}.$$

The joint density of $\mathbf{x}$ and A is (again evaluated at the data points)

$$f_{\mathbf{X},A}(\mathbf{x}, \alpha) = \alpha^9 e^{-4.225393\alpha - 38.680460}.$$

The posterior distribution of α is

$$\pi_{A|\mathbf{X}}(\alpha|\mathbf{x}) = \frac{\alpha^9 e^{-4.225393\alpha - 38.680460}}{\int_0^\infty \alpha^9 e^{-4.225393\alpha - 38.680460} \, d\alpha} = \frac{\alpha^9 e^{-4.225393\alpha}}{(9!)(1/4.225393)^{10}}.$$

There is no need to evaluate the integral in the denominator. Because we know that the result must be a probability distribution, the denominator is just the appropriate normalizing constant. A look at the numerator reveals that we have a gamma distribution with $\alpha = 10$ and $\theta = 1/4.225393$.

The predictive distribution is $f_{Y|\mathbf{X}}(y|\mathbf{x}) = \dfrac{10(4.225393)^{11}}{y(-0.379777 + \ln y)^{11}}, \quad y > 100.$ $\square$

13.2 Inference and Prediction

In one sense, the analysis is complete. We begin with a distribution that quantifies our knowledge about the parameter and/or the next observation, and we end with a revised distribution. But we suspect that your boss may not be satisfied if you produce a distribution in response to his or her request. No doubt a specific number, perhaps with a margin for error, is what is desired. The usual Bayesian solution is to pose a loss function.

Definition 13.9 *A **loss function** $l_j(\hat{\theta}_j, \theta_j)$ describes the penalty paid by the investigator when $\hat{\theta}_j$ is the estimate and θ_j is the true value of the jth parameter.*

It is possible to have a multidimensional loss function $l(\hat{\theta}, \theta)$ that allows the loss to depend simultaneously on the errors in the various parameter estimates.

Definition 13.10 *The **Bayes estimate** for a given loss function is the one that minimizes the expected loss given the posterior distribution of the parameter in question.*

The three most commonly used loss functions are defined as follows.

Definition 13.11 *For **squared-error loss**, the loss function is (all subscripts are dropped for convenience) $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. For **absolute loss**, it is $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$. For **zero–one loss** it is $l(\hat{\theta}, \theta) = 0$ if $\hat{\theta} = \theta$ and is 1 otherwise.*

The following theorem indicates the Bayes estimates for these three common loss functions.

Theorem 13.12 *For squared-error loss, the Bayes estimate is the mean of the posterior distribution; for absolute loss, it is a median; and for zero–one loss, it is a mode.*

Note that there is no guarantee that the posterior mean exists or that the posterior median or mode will be unique. Further note that if the improper prior $\pi(\theta) = 1$ is used and the estimate is the posterior mode, then the estimate will match the maximum likelihood estimate. When not otherwise specified, the term **Bayes estimate** refers to the posterior mean.

■ EXAMPLE 13.3

(*Examples 13.1 and 13.2 continued*) Determine the three Bayes estimates of α for both examples.

For Example 13.1, the mean of the posterior gamma distribution is $\alpha\theta = 12/4.801121 = 2.499416$. The median of 2.430342 must be determined numerically, while the mode

is $(\alpha - 1)\theta = 11/4.801121 = 2.291132$. Note that the α used here is the parameter of the posterior gamma distribution, not the α for the single-parameter Pareto distribution that we are trying to estimate. For Example 13.2, the corresponding values are $10/4.225393 = 2.366644$, 2.288257, and $9/4.225393 = 2.129979$. $\square$

For forecasting purposes, the expected value of the predictive distribution is often of interest. It can be thought of as providing a point estimate of the $(n + 1)$th observation given the first n observations and the prior distribution. It is

$$
\begin{aligned}
\mathrm{E}(Y|\mathbf{x}) &= \int y f_{Y|\mathbf{X}}(y|\mathbf{x})\, dy \\
&= \int y \int f_{Y|\Theta}(y|\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\, dy \\
&= \int \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \int y f_{Y|\Theta}(y|\theta)\, dy\, d\theta \\
&= \int \mathrm{E}(Y|\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta. \tag{13.5}
\end{aligned}
$$

Equation (13.5) can be interpreted as a weighted average, using the posterior distribution as the weights.

■ **EXAMPLE 13.4**

(*Example 13.1 continued*) Determine the expected value of the 11th observation given the first 10.

For the single-parameter Pareto distribution, $\mathrm{E}(Y|\alpha) = 100\alpha/(\alpha - 1)$ for $\alpha > 1$. Because the posterior distribution assigns positive probability to values of $\alpha \leq 1$, the expected value of the predictive distribution is not defined. $\square$

The Bayesian equivalent of a confidence interval is easy to construct. The following definition suffices.

Definition 13.13 *The points $a < b$ define a $100(1-\alpha)\%$ **credibility interval** for θ_j provided that* $\Pr(a \leq \Theta_j \leq b\,|\,\mathbf{x}) \geq 1 - \alpha$.

The use of the term *credibility* has no relationship to its use in actuarial analyses as developed in Chapter 16. The inequality is present for the case where the posterior distribution of θ_j is discrete. Then it may not be possible for the probability to be exactly $1 - \alpha$. This definition does not produce a unique solution. The following theorem indicates one way to produce a unique interval.

Theorem 13.14 *If the posterior random variable $\theta_j|\mathbf{x}$ is continuous and unimodal, then the $100(1 - \alpha)\%$ credibility interval with smallest width $b - a$ is the unique solution to*

$$
\int_a^b \pi_{\Theta_j|\mathbf{X}}(\theta_j|\mathbf{x})\, d\theta_j = 1 - \alpha,
$$

$$
\pi_{\Theta|\mathbf{X}}(a|\mathbf{x}) = \pi_{\Theta|\mathbf{X}}(b|\mathbf{x}).
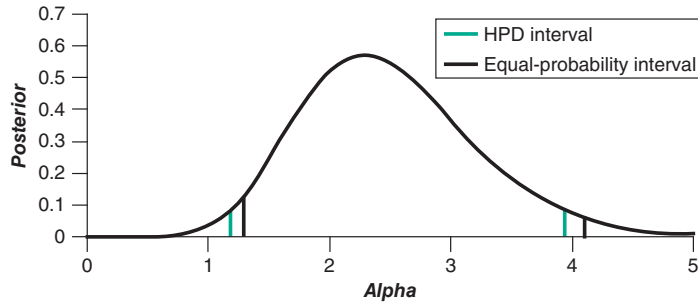$$

Figure 13.1   Two Bayesian credibility intervals.

This interval is a special case of a highest posterior density (*HPD*) *credibility set.*

The following example may clarify the theorem.

■ **EXAMPLE 13.5**

(*Example 13.1 continued*) Determine the shortest 95% credibility interval for the parameter α. Also determine the interval that places 2.5% probability at each end.

The two equations from Theorem 13.14 are

$$\Pr(a \le A \le b \,|\, \mathbf{x}) = \Gamma(12; 4.801121b) - \Gamma(12; 4.801121a) = 0.95,$$
$$a^{11}e^{-4.801121a} = b^{11}e^{-4.801121b}.$$

Numerical methods can be used to find the solution $a = 1.1832$ and $b = 3.9384$. The width of this interval is 2.7552.

Placing 2.5% probability at each end yields the two equations

$$\Gamma(12; 4.801121b) = 0.975 \quad \text{and} \quad \Gamma(12; 4.801121a) = 0.025.$$

This solution requires either access to the inverse of the incomplete gamma function or the use of root-finding techniques with the incomplete gamma function itself. The solution is $a = 1.2915$ and $b = 4.0995$. The width is 2.8080, wider than the first interval. Figure 13.1 shows the difference in the two intervals. The blue vertical bars represent the HPD interval. The total area to the left and right of these bars is 0.05. Any other 95% interval must also have this probability. To create the interval with 0.025 probability on each side, both bars must be moved to the right. To subtract the same probability on the right end that is added on the left end, the right limit must be moved a greater distance, because the posterior density is lower over that interval than it is on the left end. These adjustments must lead to a wider interval. □

The following definition provides the equivalent result for any posterior distribution.

Definition 13.15 *For any posterior distribution, the* $100(1 - \alpha)$% **HPD credibility set** *is the set of parameter values* **C** *such that*

$$\Pr(\theta_j \in C) \ge 1 - \alpha \tag{13.6}$$

and

$$C = \{\theta_j \, : \, \pi_{\Theta_j|\mathbf{X}}(\theta_j|\mathbf{x}) \geq c\} \, for \, some \, c,$$

where c is the largest value for which the inequality (13.6) holds.

This set may be the union of several intervals (which can happen with a multimodal posterior distribution). This definition produces the set of minimum total width that has the required posterior probability. Construction of the set is done by starting with a high value of c and then lowering it. As it decreases, the set C gets larger, as does the probability. The process continues until the probability reaches $1 - \alpha$. It should be obvious to see how the definition can be extended to the construction of a simultaneous credibility region for a vector of parameters, θ.

Sometimes it is the case that, while computing posterior probabilities is difficult, computing posterior moments may be easy. We can then use the Bayesian central limit theorem. The following theorem is paraphrased from Berger [14, p. 224].

Theorem 13.16 *If $\pi(\theta)$ and $f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)$ are both twice differentiable in the elements of θ and other commonly satisfied assumptions hold, then the posterior distribution of Θ given $\mathbf{X} = \mathbf{x}$ is asymptotically normal.*

The "commonly satisfied assumptions" are like those in Theorem 11.4. As in that theorem, it is possible to do further approximations. In particular, the asymptotic normal distribution also results if the posterior mode is substituted for the posterior mean and/or if the posterior covariance matrix is estimated by inverting the matrix of second partial derivatives of the negative logarithm of the posterior density.

■ **EXAMPLE 13.6**

(*Example 13.1 continued*) Construct a 95% credibility interval for α using the Bayesian central limit theorem.

The posterior distribution has mean 2.499416 and variance $\alpha\theta^2 = 0.520590$. Using the normal approximation, the credibility interval is $2.499416 \pm 1.96(0.520590)^{1/2}$, which produces $a = 1.0852$ and $b = 3.9136$. This interval (with regard to the normal approximation) is HPD due to the symmetry of the normal distribution.

The approximation is centered at the posterior mode of 2.291132 (see Example 13.3). The second derivative of the negative logarithm of the posterior density [from (13.3)] is

$$-\frac{d^2}{d\alpha^2} \ln\left[\frac{\alpha^{11}e^{-4.801121\alpha}}{(11!)(1/4.801121)^{12}}\right] = \frac{11}{\alpha^2}.$$

The variance estimate is the reciprocal. Evaluated at the modal estimate of α, we get $(2.291132)^2/11 = 0.477208$ for a credibility interval of $2.29113 \pm 1.96(0.477208)^{1/2}$, which produces $a = 0.9372$ and $b = 3.6451$.   □

The same concepts can apply to the predictive distribution. However, the Bayesian central limit theorem does not help here because the predictive sample has only one member. The only potential use for it is that, for a large original sample size, we can replace the true posterior distribution in (13.2) with a multivariate normal distribution.

■ **EXAMPLE 13.7**

(*Example 13.1 continued*) Construct a 95% highest density prediction interval for the next observation.

It is easy to see that the predictive density function (13.4) is strictly decreasing. Therefore, the region with highest density runs from $a = 100$ to b. The value of b is determined from

$$
\begin{aligned}
0.95 &= \int_{100}^{b} \frac{12(4.801121)^{12}}{y(0.195951 + \ln y)^{13}} \, dy \\
&= \int_{0}^{\ln(b/100)} \frac{12(4.801121)^{12}}{(4.801121 + x)^{13}} \, dx \\
&= 1 - \left[\frac{4.801121}{4.801121 + \ln(b/100)} \right]^{12},
\end{aligned}
$$

and the solution is $b = 390.1840$. It is interesting to note that the mode of the predictive distribution is 100 (because the pdf is strictly decreasing), while the mean is infinite (with $b = \infty$ and an additional y in the integrand, after the transformation, the integrand is like $e^x x^{-13}$, which goes to infinity as x goes to infinity). □

The following example revisits a calculation done in Section 6.3. There, the negative binomial distribution was derived as a gamma mixture of Poisson variables. Example 13.8 shows how the same calculations arise in a Bayesian context.

■ **EXAMPLE 13.8**

The number of claims in one year on a given policy is known to have a Poisson distribution. The parameter is not known, but the prior distribution has a gamma distribution with parameters α and θ. Suppose that in the past year the policy had x claims. Use Bayesian methods to estimate the number of claims in the next year. Then repeat these calculations assuming claim counts for the past n years, $x_1, \ldots, x_n$.

The key distributions are (where $x = 0, 1, \ldots$; $\lambda, \alpha, \theta > 0$):

$$
\text{Prior:} \quad \pi(\lambda) = \frac{\lambda^{\alpha-1} e^{-\lambda/\theta}}{\Gamma(\alpha)\theta^{\alpha}}.
$$

$$
\text{Model:} \quad p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.
$$

$$
\text{Joint:} \quad p(x, \lambda) = \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}}{x!\Gamma(\alpha)\theta^{\alpha}}.
$$

$$
\begin{aligned}
\text{Marginal:} \quad p(x) &= \int_{0}^{\infty} \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}}{x!\Gamma(\alpha)\theta^{\alpha}} \, d\lambda \\
&= \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)\theta^{\alpha}(1+1/\theta)^{x+\alpha}} \\
&= \binom{x+\alpha-1}{x} \left(\frac{1}{1+\theta} \right)^{\alpha} \left(\frac{\theta}{1+\theta} \right)^{x}.
\end{aligned}
$$

$$\text{Posterior:} \quad \pi(\lambda|x) = \frac{\lambda^{x+\alpha-1}e^{-(1+1/\theta)\lambda}}{x!\Gamma(\alpha)\theta^\alpha} \Bigg/ \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)\theta^\alpha(1+1/\theta)^{x+\alpha}}$$

$$= \frac{\lambda^{x+\alpha-1}e^{-(1+1/\theta)\lambda}(1+1/\theta)^{x+\alpha}}{\Gamma(x+\alpha)}.$$

The marginal distribution is negative binomial with $r = \alpha$ and $\beta = \theta$. The posterior distribution is gamma with shape parameter "α" equal to $x + \alpha$ and scale parameter "θ" equal to $(1 + 1/\theta)^{-1} = \theta/(1 + \theta)$. The Bayes estimate of the Poisson parameter is the posterior mean, $(x + \alpha)\theta/(1 + \theta)$. For the predictive distribution, (13.2) gives

$$p(y|x) = \int_0^\infty \frac{\lambda^y e^{-\lambda}}{y!} \frac{\lambda^{x+\alpha-1}e^{-(1+1/\theta)\lambda}(1+1/\theta)^{x+\alpha}}{\Gamma(x+\alpha)} d\lambda$$

$$= \frac{(1+1/\theta)^{x+\alpha}}{y!\Gamma(x+\alpha)} \int_0^\infty \lambda^{y+x+\alpha-1}e^{-(2+1/\theta)\lambda} d\lambda$$

$$= \frac{(1+1/\theta)^{x+\alpha}\Gamma(y+x+\alpha)}{y!\Gamma(x+\alpha)(2+1/\theta)^{y+x+\alpha}}, \quad y = 0, 1, \dots ,$$

and some rearranging shows this to be a negative binomial distribution with $r = x + \alpha$ and $\beta = \theta/(1+\theta)$. The expected number of claims for the next year is $(x+\alpha)\theta/(1+\theta)$. Alternatively, from (13.5),

$$\mathrm{E}(Y|x) = \int_0^\infty \lambda \frac{\lambda^{x+\alpha-1}e^{-(1+1/\theta)\lambda}(1+1/\theta)^{x+\alpha}}{\Gamma(x+\alpha)} d\lambda = \frac{(x+\alpha)\theta}{1+\theta}.$$

For a sample of size n, the key change is that the model distribution is now

$$p(\mathbf{x}|\lambda) = \frac{\lambda^{x_1+\cdots+x_n}e^{-n\lambda}}{x_1! \cdots x_n!}.$$

Using the same development as for a single observation, the posterior distribution is still gamma, now with shape parameter $x_1 + \cdots + x_n + \alpha = n\bar{x} + \alpha$ and scale parameter $\theta/(1 + n\theta)$. The predictive distribution is still negative binomial, now with $r = n\bar{x} + \alpha$ and $\beta = \theta/(1 + n\theta)$. $\square$

When only moments are needed, the double-expectation formulas can be very useful. Provided that the moments exist, for any random variables X and Y,

$$\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y|X)], \tag{13.7}$$

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)]. \tag{13.8}$$

For the predictive distribution,

$$\mathrm{E}(Y|\mathbf{x}) = \mathrm{E}_{\Theta|\mathbf{x}}[\mathrm{E}(Y|\Theta, \mathbf{x})]$$

$$= \mathrm{E}_{\Theta|\mathbf{x}}[\mathrm{E}(Y|\Theta)]$$

and

$$\mathrm{Var}(Y|\mathbf{x}) = \mathrm{E}_{\Theta|\mathbf{x}}[\mathrm{Var}(Y|\Theta, \mathbf{x})] + \mathrm{Var}_{\Theta|\mathbf{x}}[\mathrm{E}(Y|\Theta, \mathbf{x})]$$

$$= \mathrm{E}_{\Theta|\mathbf{x}}[\mathrm{Var}(Y|\Theta)] + \mathrm{Var}_{\Theta|\mathbf{x}}[\mathrm{E}(Y|\Theta)].$$

The simplification on the inner expected value and variance results from the fact that, if Θ is known, the value of $\mathbf{x}$ provides no additional information about the distribution of Y. This is simply a restatement of (13.5).

■ **EXAMPLE 13.9**

Apply these formulas to obtain the predictive mean and variance for Example 13.8. Then anticipate the credibility formulas of Chapters 16 and 17.

The predictive mean uses $E(Y|\lambda) = \lambda$. Then,

$$E(Y|\mathbf{x}) = E(\lambda|\mathbf{x}) = \frac{(n\bar{x} + \alpha)\theta}{1 + n\theta}.$$

The predictive variance uses $\text{Var}(Y|\lambda) = \lambda$, and then

$$\text{Var}(Y|\mathbf{x}) = E(\lambda|\mathbf{x}) + \text{Var}(\lambda|\mathbf{x})$$

$$= \frac{(n\bar{x} + \alpha)\theta}{1 + n\theta} + \frac{(n\bar{x} + \alpha)\theta^2}{(1 + n\theta)^2}$$

$$= (n\bar{x} + \alpha)\frac{\theta}{1 + n\theta}\left(1 + \frac{\theta}{1 + n\theta}\right).$$

These agree with the mean and variance of the known negative binomial distribution for y. However, these quantities were obtained from moments of the model (Poisson) and posterior (gamma) distributions. The predictive mean can be written as

$$\frac{n\theta}{1 + n\theta}\bar{x} + \frac{1}{1 + n\theta}\alpha\theta,$$

which is a weighted average of the mean of the data and the mean of the prior distribution. Note that, as the sample size increases, more weight is placed on the data and less on the prior opinion. The variance of the prior distribution can be increased by letting θ become large. As it should, this also increases the weight placed on the data. The credibility formulas in Chapter 16 generally consist of weighted averages of an estimate from the data and a prior opinion.   □

13.2.1 Exercises

13.1 Show that, if Y is the predictive distribution in Example 13.1, then $\ln Y - \ln 100$ has a Pareto distribution.

13.2 Determine the posterior distribution of α in Example 13.1 if the prior distribution is an arbitrary gamma distribution. To avoid confusion, denote the first parameter of this gamma distribution by γ. Next, determine a particular combination of gamma parameters so that the posterior mean is the maximum likelihood estimate of α regardless of the specific values of $x_1, \ldots, x_n$. Is this prior improper?

13.3 Let $x_1, \ldots, x_n$ be a random sample from a lognormal distribution with unknown parameters μ and σ. Let the prior density be $\pi(\mu, \sigma) = \sigma^{-1}$.

(a) Write the posterior pdf of μ and σ up to a constant of proportionality.

(b) Determine Bayesian estimators of μ and σ by using the posterior mode.

(c) Fix σ at the posterior mode as determined in part (b) and then determine the exact (conditional) pdf of μ. Then use it to determine a 95% HPD credibility interval for μ.

13.4 A random sample of size 100 has been taken from a gamma distribution with α known to be 2, but θ unknown. For this sample, $\sum_{j=1}^{100} x_j = 30{,}000$. The prior distribution for θ is inverse gamma, with β taking the role of α and λ taking the role of θ.

    (a) Determine the exact posterior distribution of θ. At this point, the values of β and λ have yet to be specified.

    (b) The population mean is 2θ. Determine the posterior mean of 2θ using the prior distribution first with $\beta = \lambda = 0$ [this is equivalent to $\pi(\theta) = \theta^{-1}$] and then with $\beta = 2$ and $\lambda = 250$ (which is a prior mean of 250). Then, in each case, determine a 95% credibility interval with 2.5% probability on each side.

    (c) Determine the posterior variance of 2θ and use the Bayesian central limit theorem to construct a 95% credibility interval for 2θ using each of the two prior distributions given in part (b).

    (d) Determine the maximum likelihood estimate of θ and then use the estimated variance to construct a 95% confidence interval for 2θ.

13.5 Suppose that, given $\Theta = \theta$, the random variables $X_1, \ldots, X_n$ are independent and binomially distributed with pf

$$f_{X_j|\Theta}(x_j|\theta) = \binom{K_j}{x_j}\theta^{x_j}(1-\theta)^{K_j-x_j}, \quad x_j = 0, 1, \ldots, K_j,$$

and Θ itself is beta distributed with parameters a and b and pdf

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1.$$

    (a) Verify that the marginal pf of X_j is

$$f_{X_j}(x_j) = \frac{\binom{-a}{x_j}\binom{-b}{K_j - x_j}}{\binom{-a-b}{K_j}}, \quad x_j = 0, 1, \ldots, K_j,$$

and $\mathrm{E}(X_j) = aK_j/(a+b)$. This distribution is termed the **binomial–beta** or **negative hypergeometric** distribution.

    (b) Determine the posterior pdf $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ and the posterior mean $\mathrm{E}(\Theta|\mathbf{x})$.

13.6 Suppose that, given $\Theta = \theta$, the random variables $X_1, \ldots, X_n$ are independent and identically exponentially distributed with pdf

$$f_{X_j|\Theta}(x_j|\theta) = \theta e^{-\theta x_j}, \quad x_j > 0,$$

and Θ is itself gamma distributed with parameters $\alpha > 1$ and $\beta > 0$,

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}, \quad \theta > 0.$$

(a) Verify that the marginal pdf of X_j is

$$f_{X_j}(x_j) = \alpha \beta^{-\alpha}(\beta^{-1} + x_j)^{-\alpha-1}, \quad x_j > 0,$$

and

$$E(X_j) = \frac{1}{\beta(\alpha - 1)}.$$

This distribution is one form of the Pareto distribution.

(b) Determine the posterior pdf $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ and the posterior mean $E(\Theta|\mathbf{x})$.

13.7 Suppose that, given $\Theta = \theta$, the random variables $X_1, \ldots, X_n$ are independent and identically negative binomially distributed with parameters r and θ with pf

$$f_{X_j|\Theta}(x_j|\theta) = \binom{r + x_j - 1}{x_j} \theta^r (1 - \theta)^{x_j}, \quad x_j = 0, 1, 2, \ldots,$$

and Θ itself is beta distributed with parameters a and b and pdf

$$\pi(\theta) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1}, \quad 0 < \theta < 1.$$

(a) Verify that the marginal pf of X_j is

$$f_{X_j}(x_j) = \frac{\Gamma(r + x_j)}{\Gamma(r)x_j!} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a + r)\Gamma(b + x_j)}{\Gamma(a + r + b + x_j)}, \quad x_j = 0, 1, 2, \ldots,$$

and

$$E(X_j) = \frac{rb}{a - 1}.$$

This distribution is termed the **generalized Waring distribution**. The special case where $b = 1$ is the **Waring distribution** and is the **Yule distribution** if $r = 1$ and $b = 1$.

(b) Determine the posterior pdf $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ and the posterior mean $E(\Theta|\mathbf{x})$.

13.8 Suppose that, given $\Theta = \theta$, the random variables $X_1, \ldots, X_n$ are independent and identically normally distributed with mean μ and variance θ^{-1}, and Θ is gamma distributed with parameters α and (θ replaced by) $1/\beta$.

(a) Verify that the marginal pdf of X_j is

$$f_{X_j}(x_j) = \frac{\Gamma(\alpha + \frac{1}{2})}{\sqrt{2\pi\beta}\,\Gamma(\alpha)} \left[1 + \frac{1}{2\beta}(x_j - \mu)^2\right]^{-\alpha-1/2}, \quad -\infty < x_j < \infty,$$

which is a form of the t-distribution.

(b) Determine the posterior pdf $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ and the posterior mean $E(\theta|\mathbf{x})$.

13.9 Suppose that, for $j = 1, 2, \ldots, n$, the random variable Y_{1j} has (conditional on $\Theta = \theta$) the Poisson pf

$$f_{Y_{1j}|\Theta}(y_{1j}|\theta) = \frac{\theta^{y_{1j}} e^{-\theta}}{(y_{1j})!}, \quad y_{1j} = 0, 1, \ldots$$

and Y_{2j} has (conditional on $\Theta = \theta$) the binomial pf

$$f_{Y_{2j}|\Theta}(y_{2j}|\theta) = \binom{N}{y_{2j}} \left(\frac{\theta}{1+\theta}\right)^{y_{2j}} \left(\frac{1}{1+\theta}\right)^{N-y_{2j}}, \quad y_{2j} = 0, 1, \ldots, N,$$

with $\theta > 0$ and N a known positive integer. Further assume that all random variables are independent (conditional on $\Theta = \theta$). Let $X_j = Y_{1j} + Y_{2j}$ for $j = 1, 2, \ldots, n$.

(a) Show that X_j has (conditional on $\Theta = \theta$) the Poisson–binomial pf

$$f_{X_j|\Theta}(x_j|\theta) = \frac{p(x_j)\theta^{x_j}}{q(\theta)}, \quad x_j = 0, 1, \ldots,$$

where

$$p(x) = \sum_{y=0}^{x} \binom{N}{y} \frac{1}{(x-y)!} \quad \text{and} \quad q(\theta) = e^{\theta}(1+\theta)^N.$$

(b) If $X_1, X_2, \ldots, X_n$ have the pf in (a), demonstrate that the conjugate prior for this situation is

$$\pi(\theta) = \frac{\theta^{\mu k-1}(1+\theta)^{-Nk} e^{-k\theta}}{c(\mu, k)}, \quad \theta > 0,$$

where $\mu > 0$ and $k > 0$. Show further that

$$c(\mu, k) = \Gamma(\mu k)\Psi(\mu k, \mu k - Nk + 1, k)$$

where $\Psi(a, b, z)$ is the confluent hypergeometric function of the second kind, which can be expressed as

$$\Psi(a, b, z) = \frac{1}{\Gamma(a)} \int_0^{\infty} t^{a-1}(1+t)^{b-a-1} e^{-zt} dt.$$

13.10 Suppose that, given N, the random variable X is binomially distributed with parameters N and p.

(a) Show that, if N is Poisson distributed, so is X (unconditionally) and identify the parameters.

(b) Show that, if N is binomially distributed, so is X (unconditionally) and identify the parameters.

(c) Show that, if N is negative binomially distributed, so is X (unconditionally) and identify the parameters.

13.11 (*) A die is selected at random from an urn that contains two six-sided dice. Die number 1 has three faces with the number 2, while one face each has the numbers 1, 3, and 4. Die number 2 has three faces with the number 4, while one face each has the numbers

1, 2, and 3. The first five rolls of the die yielded the numbers 2, 3, 4, 1, and 4, in that order. Determine the probability that the selected die was die number 2.

13.12 (*) The number of claims in a year, Y, has a distribution that depends on a parameter θ. As a random variable, Θ has the uniform distribution on the interval $(0, 1)$. The unconditional probability that Y is 0 is greater than 0.35. For each of the following conditional pfs, determine if it is possible that it is the true conditional pf of Y:

(a) $\Pr(Y = y|\theta) = e^{-\theta}\theta^y/y!$.

(b) $\Pr(Y = y|\theta) = (y + 1)\theta^2(1 - \theta)^y$.

(c) $\Pr(Y = y|\theta) = \binom{2}{y}\theta^y(1 - \theta)^{2-y}$.

13.13 (*) Your prior distribution concerning the unknown value of H is $\Pr(H = \frac{1}{4}) = \frac{4}{5}$ and $\Pr(H = \frac{1}{2}) = \frac{1}{5}$. The observation from a single experiment has distribution $\Pr(D = d|H = h) = h^d(1 - h)^{1-d}$ for $d = 0, 1$. The result of a single experiment is $d = 1$. Determine the posterior distribution of H.

13.14 (*) The number of claims for an individual in one year has a Poisson distribution with parameter λ. The prior distribution for λ has a gamma distribution with mean 0.14 and variance 0.0004. During the past two years, a total of 110 claims has been observed. In each year, there were 310 policies in force. Determine the expected value and variance of the posterior distribution of λ.

13.15 (*) An individual risk has exactly one claim each year. The amount of the single claim has an exponential distribution with pdf $f(x) = te^{-tx}$, $x > 0$. The parameter t has a prior distribution with pdf $\pi(t) = te^{-t}$. A claim of 5 has been observed. Determine the posterior pdf of t.

13.16 (*) Given $Q = q$, $X_1, \ldots, X_m$ are i.i.d. Bernoulli random variables with parameter q. Let $S_m = X_1 + \cdots + X_m$. The prior distribution of Q is beta with $a = 1$, $b = 99$, and $\theta = 1$. Determine the smallest value of m such that the mean of the marginal distribution of S_m is greater than or equal to 50.

13.17 (*) Given β, a loss X has the exponential pdf $f(x) = \beta^{-1}e^{-x/\beta}$, $x > 0$. The prior distribution is $\pi(\beta) = 100\beta^{-3}e^{-10/\beta}$, $\beta > 0$, an inverse gamma distribution. A single loss of x has been observed. Determine the mean of the posterior distribution as a function of x.

13.18 In Exercise 11.24, 500 losses are observed. Five of the losses are 1,100, 3,200, 3,300, 3,500, and 3,900. All that is known about the other 495 losses is that they exceed 4,000. Determine the Bayes estimate of the mean of an exponential model using the improper prior $\pi(\theta) = 1/\theta$ and compare your answer to the maximum likelihood estimate.

13.19 Suppose that, given $\Theta_1 = \theta_1$ and $\Theta_2 = \theta_2$, the random variables $X_1, \ldots, X_n$ are independent and identically normally distributed with mean θ_1 and variance θ_2^{-1}. Suppose also that the conditional distribution of Θ_1 given $\Theta_2 = \theta_2$ is a normal distribution with mean μ and variance σ^2/θ_2 and Θ_2 is gamma distributed with parameters α and $\theta = 1/\beta$.

(a) Show that the posterior conditional distribution of Θ_1 given $\Theta_2 = \theta_2$ is normally distributed with mean

$$\mu_* = \frac{1}{1 + n\sigma^2}\mu + \frac{n\sigma^2}{1 + n\sigma^2}\bar{x}$$

and variance

$$\sigma_*^2 = \frac{\sigma^2}{\theta_2(1 + n\sigma^2)},$$

and the posterior marginal distribution of Θ_2 is gamma distributed with parameters

$$\alpha_* = \alpha + \frac{n}{2}$$

and

$$\beta_* = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n\left(\bar{x} - \mu\right)^2}{2\left(1 + n\sigma^2\right)}.$$

(b) Find the posterior marginal means $E(\Theta_1|\mathbf{x})$ and $E(\Theta_2|\mathbf{x})$.

13.3   Conjugate Prior Distributions and the Linear Exponential Family

The linear exponential family introduced in Section 5.4 is also useful in connection with Bayesian analysis, as is demonstrated in this section.

In Example 13.8, it turned out the posterior distribution was of the same type as the prior distribution (gamma). A definition of this concept follows.

Definition 13.17 *A prior distribution is said to be a **conjugate prior distribution** for a given model if the resulting posterior distribution is from the same family as the prior (but perhaps with different parameters).*

The following theorem shows that, if the model is a member of the linear exponential family, a conjugate prior distribution is easy to find.

Theorem 13.18 *Suppose that given $\Theta = \theta$, the random variables $X_1, \ldots, X_n$ are i.i.d. with pf*

$$f_{X_j|\Theta}(x_j|\theta) = \frac{p(x_j)e^{r(\theta)x_j}}{q(\theta)},$$

where Θ has pdf

$$\pi(\theta) = \frac{[q(\theta)]^{-k}e^{\mu kr(\theta)}r'(\theta)}{c(\mu, k)},$$

where k and μ are parameters of the distribution and $c(\mu, k)$ is the normalizing constant. Then, the posterior pf $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ is of the same form as $\pi(\theta)$.

Proof: The posterior distribution is

$$\pi(\theta|\mathbf{x}) \propto \frac{\left[\prod_{j=1}^n p(x_j)\right] e^{r(\theta)\Sigma x_j}}{[q(\theta)]^n} \frac{[q(\theta)]^{-k} e^{\mu k r(\theta)} r'(\theta)}{c(\mu, k)}$$

$$\propto [q(\theta)]^{-(k+n)} \exp\left[\left(r(\theta)\frac{\mu k + \sum x_j}{k+n}\right)(k+n)\right] r'(\theta)$$

$$= [q(\theta)]^{-k^*} \exp[r(\theta)\mu^* k^*] r'(\theta),$$

which is of the same form as $\pi(\theta)$ with parameters

$$k^* = k + n,$$

$$\mu^* = \frac{\mu k + \sum x_j}{k+n} = \frac{k}{k+n}\mu + \frac{n}{k+n}\bar{x}. \qquad \square$$

■ **EXAMPLE 13.10**

Show that, for the Poisson model, the conjugate prior as given in Theorem 13.18 is the gamma distribution.

The Poisson pf can be written

$$p(x) = \frac{\theta^x e^{-\theta}}{x!} = \frac{(x!)^{-1} e^{x \ln \theta}}{e^\theta}.$$

Thus we have that $q(\theta) = e^\theta$ and $r(\theta) = \ln \theta$. The prior as given by the theorem is

$$\pi(\theta) \propto e^{-k\theta} e^{\mu k \ln \theta} \theta^{-1},$$

which can be rewritten

$$\pi(\theta) \propto \theta^{\mu k - 1} e^{-k\theta},$$

which is the kernel of a gamma distribution with $\alpha = \mu k$ and scale parameter $1/k$. $\square$

Other well-known examples of linear exponential family members include the binomial and negative binomial distributions both with beta conjugate prior (see Exercises 13.5 and 13.7, respectively). Similarly, for the exponential distribution, the gamma distribution is the conjugate prior (see Exercise 13.6).

13.3.1 Exercises

13.20 Let $X_1, \ldots, X_n$ be i.i.d. random variables, conditional on Θ, with pf

$$f_{X_j|\Theta}(x_j|\theta) = \frac{p(x_j)e^{r(\theta)x_j}}{q(\theta)}.$$

Let $S = X_1 + \cdots + X_n$. Use Exercise 5.26(a) to prove that the posterior distribution $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ is the same as the (conditional) distribution of $\Theta|S$,

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{S|\Theta}(s|\theta)\pi(\theta)}{f_S(s)},$$

where $\pi(\theta)$ is the pf of Θ and $f_S(s)$ is the marginal pf of S.

13.21 (*) The number of claims in one year, Y, has a Poisson distribution with parameter θ. The parameter θ has a exponential distribution with pdf $\pi(\theta) = e^{-\theta}$. A particular insured had no claims in one year. Determine the posterior distribution of θ for this insured.

13.22 (*) The number of claims in one year, Y, has a Poisson distribution with parameter θ. The prior distribution has a gamma distribution with pdf $\pi(\theta) = \theta e^{-\theta}$. There was one claim in one year. Determine the posterior pdf of θ.

13.23 (*) Each individual car's claim count has a Poisson distribution with parameter λ. All individual cars have the same parameter. The prior distribution is gamma with parameters $\alpha = 50$ and $\theta = 1/500$. In a two-year period, the insurer covers 750 and 1,100 cars in years 1 and 2, respectively. There were 65 and 112 claims in years 1 and 2, respectively. Determine the coefficient of variation of the posterior gamma distribution.

13.24 (*) The number of claims, r, made by an individual in one year has a binomial distribution with pf $f(r) = \binom{3}{r}\theta^r(1-\theta)^{3-r}$. The prior distribution for θ has pdf $\pi(\theta) = 6(\theta - \theta^2)$. There was one claim in a one-year period. Determine the posterior pdf of θ.

13.25 (*) The number of claims for an individual in one year has a Poisson distribution with parameter λ. The prior distribution for λ is exponential with an expected value of 2. There were three claims in the first year. Determine the posterior distribution of λ.

13.26 (*) The number of claims in one year has a binomial distribution with $n = 3$ and θ unknown. The prior distribution for θ is beta with pdf $\pi(\theta) = 280\theta^3(1-\theta)^4$, $\quad 0 < \theta < 1$. Two claims were observed. Determine each of the following:

(a) The posterior distribution of θ.

(b) The expected value of θ from the posterior distribution.

13.27 (*) The number of claims is binomial, with $m = 4$ and q unknown. The prior distribution is $\pi(q) = 6q(1-q)$, $0 < q < 1$. A single observation has a value of 2. Determine the mean and mode of the posterior distribution of q.

13.4 Computational Issues

It should be obvious by now that all Bayesian analyses proceed by taking integrals or sums. So, at least conceptually, it is always possible to do a Bayesian analysis. However, only in rare cases are the integrals or sums easy to do, and that means most Bayesian analyses will require numerical integration. While one-dimensional integrations are easy to do to a high degree of accuracy, multidimensional integrals are much more difficult to approximate. A great deal of effort has been expended with regard to solving this problem. A number of ingenious methods have been developed. Some of them are summarized in Klugman [72]. However, the one that is widely used today is called Markov Chain Monte Carlo simulation. A good discussion of this method can be found in Gelman et al. [42].

CONSTRUCTION OF MODELS

14

CONSTRUCTION OF EMPIRICAL MODELS

14.1   The Empirical Distribution

The material presented here has traditionally been presented under the heading of "survival models," with the accompanying notion that the techniques are useful only when studying lifetime distributions. Standard texts on the subject, such as Klein and Moeschberger [70] and Lawless [77], contain examples that are exclusively oriented in that direction. However, the same problems that occur when modeling lifetime occur when modeling payment amounts. The examples we present are of both types. However, the latter sections focus on special considerations when constructing decrement models. Only a handful of references are presented, most of the results being well developed in the survival models literature. If you want more detail and proofs, consult a text dedicated to the subject, such as the ones just mentioned.

In Chapter 4, models were divided into two types – data-dependent and parametric. The definitions are repeated here.

Definition 14.1 *A **data-dependent distribution** is at least as complex as the data or knowledge that produced it, and the number of "parameters" increases as the number of data points or amount of knowledge increases.*

Table 14.1   Data Set B.

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1,193	1,340	1,884	2,558	15,743

Table 14.2   Data Set C.

Payment range	Number of payments
0–7,500	99
7,500–17,500	42
17,500–32,500	29
32,500–67,500	28
67,500–125,000	17
125,000–300,000	9
Over 300,000	3

Definition 14.2 *A **parametric distribution** is a set of distribution functions, each member of which is determined by specifying one or more values called **parameters**. The number of parameters is fixed and finite.*

This chapter will focus on data-dependent distributions as models, making few, if any, assumptions about the underlying distribution. To fix the most important concepts, we begin by assuming that we have a sample of n observations that are an independent and identically distributed sample from the same (unspecified) continuous distribution. This is referred to as a **complete data** situation. In that context, we have the following definition.

Definition 14.3 *The **empirical distribution** is obtained by assigning probability $1/n$ to each data point.*

When observations are collected from a probability distribution, the ideal situation is to have the (essentially) exact[1] value of each observation. This case is referred to as **complete, individual data** and applies to Data Set B, introduced in Chapter 10 and reproduced here as Table 14.1. There are two reasons why exact data may not be available. One is grouping, in which all that is recorded is the range of values in which the observation belongs. Grouping applies to Data Set C and to Data Set A for those with five or more accidents. There data sets were introduced in Chapter 10 and are reproduced here as Tables 14.2 and 14.3, respectively.

A second reason that exact values may not be available is the presence of censoring or truncation. When data are censored from below, observations below a given value are known to be below that value but the exact value is unknown. When data are censored from above, observations above a given value are known to be above that value but the exact value is unknown. Note that censoring effectively creates grouped data. When the data are grouped in the first place, censoring has no effect. For example, the data in Data Set C may have been censored from above at 300,000, but we cannot know for sure from

---

[1]Some measurements are never exact. Ages may be rounded to the nearest whole number, monetary amounts to the nearest dollar, car mileage to the nearest mile, and so on. This text is not concerned with such rounding errors. Rounded values will be treated as if they are exact.

Table 14.3   Data Set A.

Number of accidents	Number of drivers
0	81,714
1	11,306
2	1,618
3	250
4	40
5 or more	7

the data set and that knowledge has no effect on how we treat the data. In contrast, were Data Set B to be censored at 1,000, we would have 15 individual observations and then five grouped observations in the interval from 1,000 to infinity.

In insurance settings, censoring from above is fairly common. For example, if a policy pays no more than 100,000 for an accident, any time the loss is above 100,000 the actual amount will be unknown, but we will know that it happened. Note that Data Set A has been censored from above at 5. This is more common language than saying that Data Set A has some individual data and some grouped data. When studying mortality or other decrements, the study period may end with some individuals still alive. They are censored from above in that we know the death will occur sometime after their age when the study ends.

When data are truncated from below, observations below a given value are not recorded. Truncation from above implies that observations above a given value are not recorded. In insurance settings, truncation from below is fairly common. If an automobile physical damage policy has a per-claim deductible of 250, any losses below 250 will not come to the attention of the insurance company and so will not appear in any data sets. Data sets may have truncation forced on them. For example, if Data Set B were to be truncated from below at 250, the first seven observations would disappear and the remaining 13 would be unchanged. In decrement studies it is unusual to observe individuals from birth. If someone is first observed at, say, age 20, that person is from a population where anyone who died before age 20 would not have been observed and thus is truncated from below.

As noted in Definition 14.3, the empirical distribution assigns probability $1/n$ to each data point. That definition works well when the value of each data point is recorded. An alternative definition follows.

Definition 14.4 *The **empirical distribution function** is*

$$F_n(x) = \frac{\textit{number of observations } \leq x}{n},$$

where n is the total number of observations.

■ **EXAMPLE 14.1**

Provide the empirical probability functions for the data in Data Sets A and B. For Data Set A, also provide the empirical distribution function. For Data Set A, assume that all seven drivers who had five or more accidents had exactly five accidents.

For notation, a subscript of the sample size (or of n if the sample size is not known) is used to indicate an empirical function. Without the subscript, the function represents the true function for the underlying random variable. For Data Set A, the empirical probability function is

$$
p_{94,935}(x) = \begin{cases}
81{,}714/94{,}935 = 0.860736, & x = 0, \\
11{,}306/94{,}935 = 0.119092, & x = 1, \\
1{,}618/94{,}935 = 0.017043, & x = 2, \\
250/94{,}935 = 0.002633, & x = 3, \\
40/94{,}935 = 0.000421, & x = 4, \\
7/94{,}935 = 0.000074, & x = 5,
\end{cases}
$$

where the values add to 0.999999 due to rounding. The empirical distribution function is a step function with jumps at each data point.

$$
F_{94,935}(x) = \begin{cases}
0/94{,}935 = 0.000000, & x < 0, \\
81{,}714/94{,}935 = 0.860736, & 0 \le x < 1, \\
93{,}020/94{,}935 = 0.979828, & 1 \le x < 2, \\
94{,}638/94{,}935 = 0.996872, & 2 \le x < 3, \\
94{,}888/94{,}935 = 0.999505, & 3 \le x < 4, \\
94{,}928/94{,}935 = 0.999926, & 4 \le x < 5, \\
94{,}935/94{,}935 = 1.000000, & x \ge 5.
\end{cases}
$$

For Data Set B,

$$
p_{20}(x) = \begin{cases}
0.05, & x = 27, \\
0.05, & x = 82, \\
0.05, & x = 115, \\
\vdots & \vdots \\
0.05, & x = 15{,}743.
\end{cases}
$$

□

In the following example, not all values are distinct.

■ **EXAMPLE 14.2**

Consider a data set containing the numbers 1.0, 1.3, 1.5, 1.5, 2.1, 2.1, 2.1, and 2.8. Determine the empirical distribution function.

The empirical distribution function is a step function with the following values:

$$F_8(x) = \begin{cases} 0, & x < 1.0, \\ 1 - \frac{7}{8} = 0.125, & 1.0 \le x < 1.3, \\ 1 - \frac{6}{8} = 0.250, & 1.3 \le x < 1.5, \\ 1 - \frac{4}{8} = 0.500, & 1.5 \le x < 2.1, \\ 1 - \frac{1}{8} = 0.875, & 2.1 \le x < 2.8, \\ 1, & x \ge 2.8. \end{cases}$$

□

To assess the quality of the estimate, we examine statistical properties, in particular, the mean and variance. Working with the empirical estimate of the distribution function is straightforward. To see that with complete data the empirical estimator of the survival function is unbiased and consistent, recall that the empirical estimate of $F(x)$ is $F_n(x) = Y/n$, where Y is the number of observations in the sample that are less than or equal to x. Then Y must have a binomial distribution with parameters n and $F(x)$ and

$$E[F_n(x)] = E\left(\frac{Y}{n}\right) = \frac{nF(x)}{n} = F(x).$$

demonstrating that the estimator is unbiased. The variance is

$$\mathrm{Var}[F_n(x)] = \mathrm{Var}\left(\frac{Y}{n}\right) = \frac{F(x)[1 - F(x)]}{n},$$

which has a limit of zero, thus verifying consistency.

To make use of the result, the best we can do for the variance is estimate it. It is unlikely that we will know the value of $F(x)$, because that is the quantity we are trying to estimate. The estimated variance is given by

$$\widehat{\mathrm{Var}}[F_n(x)] = \frac{F_n(x)[1 - F_n(x)]}{n}. \tag{14.1}$$

The same results hold for empirically estimated probabilities. Let $p = \Pr(a < X \le b)$. The empirical estimate of p is $\hat{p} = F_n(b) - F_n(a)$. Arguments similar to those used for $F_n(x)$ verify that $\hat{p}$ is unbiased and consistent, with $\mathrm{Var}(\hat{p}) = p(1 - p)/n$.

■ **EXAMPLE 14.3**

For the data in Example 14.2, estimate the variance of $F_8(1.4)$.

From the example, we have $F_8(1.4) = 0.25$. From (14.1),

$$\widehat{\mathrm{Var}}[F_8(1.4)] = \frac{F_8(1.4)[1 - F_8(1.4)]}{8} = \frac{0.25(1 - 0.25)}{8} = 0.02344.$$

□

14.2 Empirical Distributions for Grouped Data

For grouped data as in Data Set C, construction of the empirical distribution as defined previously is not possible. However, it is possible to approximate the empirical distribution. The strategy is to obtain values of the empirical distribution function wherever possible and then connect those values in some reasonable way. For grouped data, the distribution function is usually approximated by connecting the points with straight lines. For notation, let the group boundaries be $c_0 < c_1 < \cdots < c_k$, where often $c_0 = 0$ and $c_k = \infty$. The number of observations falling between c_{j-1} and c_j is denoted n_j, with $\sum_{j=1}^{k} n_j = n$. For such data, we are able to determine the empirical distribution at each group boundary. That is, $F_n(c_j) = (1/n) \sum_{i=1}^{j} n_i$. Note that no rule is proposed for observations that fall on a group boundary. There is no correct approach, but whatever approach is used, consistency in assignment of observations to groups should be used. Note that in Data Set C it is not possible to tell how the assignments were made. If we had that knowledge, it would not affect any subsequent calculations.[2]

Definition 14.5 *For grouped data, the distribution function obtained by connecting the values of the empirical distribution function at the group boundaries with straight lines is called the **ogive**. The formula is as follows:*

$$F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j), \ c_{j-1} \le x \le c_j.$$

This function is differentiable at all values except group boundaries. Therefore the density function can be obtained. To completely specify the density function, it is arbitrarily made right continuous.

Definition 14.6 *For grouped data, the empirical density function can be obtained by differentiating the ogive. The resulting function is called a **histogram**. The formula is as follows:*

$$f_n(x) = \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} = \frac{n_j}{n(c_j - c_{j-1})}, \ c_{j-1} \le x < c_j.$$

Many computer programs that produce histograms actually create a bar chart with bar heights proportional to n_j/n. A bar chart is acceptable if the groups have equal width, but if not, then the preceding formula is needed. The advantage of this approach is that the histogram is indeed a density function and, among other things, areas under the histogram can be used to obtain empirical probabilities.

■ **EXAMPLE 14.4**

Construct the ogive and histogram for Data Set C.

---

[2]Technically, for the interval from c_{j-1} to c_j, $x = c_j$ should be included and $x = c_{j-1}$ excluded in order for $F_n(c_j)$ to be the empirical distribution function.
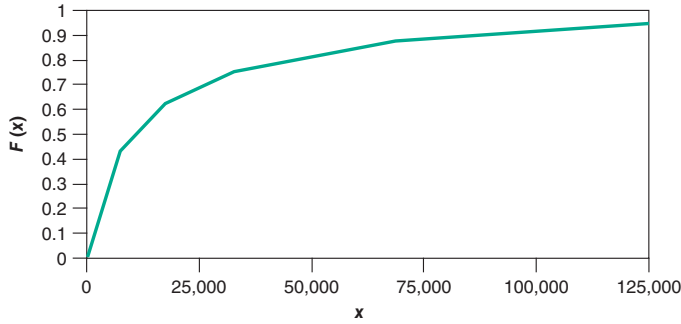
Figure 14.1   The ogive for general liability losses.

The distribution function is as follows:

$$
F_{227}(x) = \begin{cases}
0.000058150x, & 0 \le x \le 7{,}500, \\
0.29736 + 0.000018502x, & 7{,}500 \le x \le 17{,}500, \\
0.47210 + 0.000008517x, & 17{,}500 \le x \le 32{,}500, \\
0.63436 + 0.000003524x, & 32{,}500 \le x \le 67{,}500, \\
0.78433 + 0.000001302x, & 67{,}500 \le x \le 125{,}000, \\
0.91882 + 0.000000227x, & 125{,}000 \le x \le 300{,}000, \\
\text{undefined}, & x > 300{,}000,
\end{cases}
$$

where, for example, for the range $32{,}500 \le x \le 67{,}500$, the calculation is

$$
F_{227}(x) = \frac{67{,}500 - x}{67{,}500 - 32{,}500}\frac{170}{227} + \frac{x - 32{,}500}{67{,}500 - 32{,}500}\frac{198}{227}.
$$

The value is undefined above 300,000 because the last interval has a width of infinity. A graph of the ogive for values up to 125,000 appears in Figure 14.1. The derivative is simply a step function with the following values:

$$
f_{227}(x) = \begin{cases}
0.000058150, & 0 \le x < 7{,}500, \\
0.000018502, & 7{,}500 \le x < 17{,}500, \\
0.000008517, & 17{,}500 \le x < 32{,}500, \\
0.000003524, & 32{,}500 \le x < 67{,}500, \\
0.000001302, & 67{,}500 \le x < 125{,}000, \\
0.000000227, & 125{,}000 \le x < 300{,}000, \\
\text{undefined}, & x \ge 300{,}000.
\end{cases}
$$

A graph of the function up to 125,000 appears in Figure 14.2.   □

14.2.1   Exercises

14.1 Construct the ogive and histogram for the data in Table 14.4.
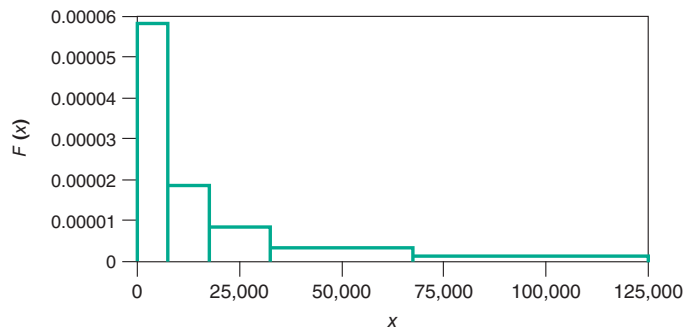
Figure 14.2   The histogram of general liability losses.

Table 14.4   The data for Exercise 14.1.

Payment range	Number of payments
0–25	6
25–50	24
50–75	30
75–100	31
100–150	57
150–250	80
250–500	85
500–1,000	54
1,000–2,000	15
2,000–4,000	10
Over 4,000	0

14.2 (*) The following 20 windstorm losses (in millions of dollars) were recorded in one year:

$$1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 3 \quad 3 \quad 4$$
$$6 \quad 6 \quad 8 \quad 10 \quad 13 \quad 14 \quad 15 \quad 18 \quad 22 \quad 25$$

(a) Construct an ogive based on using class boundaries at 0.5, 2.5, 8.5, 15.5, and 29.5.

(b) Construct a histogram using the same boundaries as in part (a).

14.3 The data in Table 14.5 are from Herzog and Laverty [53]. A certain class of 15-year mortgages was followed from issue until December 31, 1993. The issues were split into those that were refinances of existing mortgages and those that were original issues. Each entry in the table provides the number of issues and the percentage of them that were still in effect after the indicated number of years. Draw as much of the two ogives (on the same graph) as is possible from the data. Does it appear from the ogives that the lifetime variable (the time to mortgage termination) has a different distribution for refinanced versus original issues?

Table 14.5   The data for Exercise 14.3.

	Refinances		Original	
Years	Number issued	Survived	Number issued	Survived
1.5	42,300	99.97	12,813	99.88
2.5	9,756	99.82	18,787	99.43
3.5	1,550	99.03	22,513	98.81
4.5	1,256	98.41	21,420	98.26
5.5	1,619	97.78	26,790	97.45

Table 14.6   The data for Exercise 14.4.

Loss	Number of observations
0–2	25
2–10	10
10–100	10
100–1,000	5

14.4 (*) The data in Table 14.6 were collected (the units are millions of dollars). Construct the histogram.

14.5 (*) Forty losses have been observed. Sixteen losses are between 1 and $\frac{4}{3}$ (in millions), and the sum of the 16 losses is 20. Ten losses are between $\frac{4}{3}$ and 2, with a total of 15. Ten more are between 2 and 4, with a total of 35. The remaining four losses are greater than 4. Using the empirical model based on these observations, determine $E(X \wedge 2)$.

14.6 (*) A sample of size 2,000 contains 1,700 observations that are no greater than 6,000, 30 that are greater than 6,000 but no greater than 7,000, and 270 that are greater than 7,000. The total amount of the 30 observations that are between 6,000 and 7,000 is 200,000. The value of $E(X \wedge 6,000)$ for the empirical distribution associated with these observations is 1,810. Determine $E(X \wedge 7,000)$ for the empirical distribution.

14.7 (*) A random sample of unknown size produced 36 observations between 0 and 50; x between 50 and 150; y between 150 and 250; 84 between 250 and 500; 80 between 500 and 1,000; and none above 1,000. Two values of the ogive constructed from these observations are $F_n(90) = 0.21$ and $F_n(210) = 0.51$. Determine the value of x.

14.8 The data in Table 14.7 are from Hogg and Klugman [55, p. 128]. They represent the total damage done by 35 hurricanes between the years 1949 and 1980. The losses have been adjusted for inflation (using the Residential Construction Index) to be in 1981 dollars. The entries represent all hurricanes for which the trended loss was in excess of 5,000,000.

The federal government is considering funding a program that would provide 100% payment for all damages for any hurricane causing damage in excess of 5,000,000. You have been asked to make some preliminary estimates.

Table 14.7   Trended hurricane losses.

Year	Loss (10^3)	Year	Loss (10^3)	Year	Loss (10^3)
1964	6,766	1964	40,596	1975	192,013
1968	7,123	1949	41,409	1972	198,446
1971	10,562	1959	47,905	1964	227,338
1956	14,474	1950	49,397	1960	329,511
1961	15,351	1954	52,600	1961	361,200
1966	16,983	1973	59,917	1969	421,680
1955	18,383	1980	63,123	1954	513,586
1958	19,030	1964	77,809	1954	545,778
1974	25,304	1955	102,942	1970	750,389
1959	29,112	1967	103,217	1979	863,881
1971	30,146	1957	123,680	1965	1,638,000
1976	33,727	1979	140,136		

(a) Estimate the mean, standard deviation, coefficient of variation, and skewness for the population of hurricane losses.

(b) Estimate the first and second limited moments at 500,000,000.

14.9 (*) There have been 30 claims recorded in a random sampling of claims. There were 2 claims for 2,000, 6 for 4,000, 12 for 6,000, and 10 for 8,000. Determine the empirical skewness coefficient.

14.3   Empirical Estimation with Right Censored Data

In this section, we generalize the empirical approach of the previous section to situations in which the data are not complete. In particular, we assume that individual observations may be right censored. We have the following definition.

Definition 14.7 *An observation is **censored from above** (also called **right censored**) at u if when it is at or above u it is recorded as being equal to u, but when it is below u it is recorded at its observed value.*

In insurance claims data, the presence of a policy limit may give rise to right censored observations. When the amount of the loss equals or exceeds the limit u, benefits beyond that value are not paid, and so the exact value is typically not recorded. However, it is known that a loss of at least u has occurred.

When carrying out a study of the mortality of humans, if a person is alive when the study ends, right censoring has occurred. The person's age at death is not known, but it is known that it is at least as large as the age when the study ended. Right censoring also affects those who exit the study prior to its end due to surrender or lapse. Note that this discussion could have been about other decrements, such as disability, policy surrender, or retirement.

For this section and the next two, we assume that the underlying random variable has a continuous distribution. While data from discrete random variables can also be right

censored (Data Set A is an example), the use of empirical estimators is rare and thus the development of analogous formulas is unlikely to be worth the effort.

We now make specific assumptions regarding how the data are collected and recorded. It is assumed that we have a random sample for which some (but not all) of the data are right censored. For the uncensored (i.e. completely known) observations, we will denote their k unique values by $y_1 < y_2 < \cdots < y_k$. We let s_i denote the number of times that y_i appears in the sample. We also set y_0 as the minimum possible value for an observation and assume that $y_0 < y_1$. Often, $y_0 = 0$. Similarly, set y_{max} as the largest observation in the data, censored or uncensored. Hence, $y_{max} \geq y_k$. Our goal is to create an empirical (data-dependent) distribution that places probability at the values $y_1 < y_2 < \cdots < y_k$.

We often possess the specific value at which an observation was censored. However, for both the derivation of the estimator and its implementation, it is only necessary to know between which y-values it occurred. Thus, the only input needed is b_i, the number of right censored observations in the interval $[y_i, y_{i+1})$ for $i = 1, 2, \ldots, k - 1$. We make the assumption that if an observation is censored at y_i, then the observation is censored at $y_i + 0$ (i.e. in the lifetime situation, immediately after the death). It is possible to have censored observations at values between y_0 and y_1. However, because we are placing probability only at the uncensored values, these observations provide no information about those probabilities and so can be dropped. When referring to the sample size, n will denote the number of observations after these have been dropped. Observations censored at y_k or above cannot be ignored. Let b_k be the number of observations right censored at $y_k + 0$ or later. Note that if $b_k = 0$, then $y_{max} = y_k$.

The final important quantity is r_i, referred to as the number "at risk" at y_i. When thinking in terms of a mortality study, the risk set comprises the individuals who are under observation at that age. Included are all who die at that age or later and all who are censored at that age or later. Formally, we have the following definition.

Definition 14.8 *The **number at risk**, r_i at observation y_i is*

$$r_i = \begin{cases} n, & i = 1 \\ r_{i-1} - s_{i-1} - b_{i-1}, & i = 2, 3, \ldots, k + 1. \end{cases}$$

This formula reflects the fact that the number at risk at y_i is that at y_{i-1} less the s_{i-1} exact observations at y_{i-1} and the b_{i-1} censored observations in $[y_{i-1}, y_i)$. Note that $b_k = r_k - s_k$ and hence $r_{k+1} = 0$.

The following numerical example illustrates these ideas.

■ **EXAMPLE 14.5**

Determine the numbers at risk for the data in Table 14.8. Observations marked with an asterisk (*) were censored at that value.

The first step is to summarize the data using previously defined notation. The summarized values and risk set calculations appear in Table 14.9. Addition of the values in the s and b columns confirms that there are $n = 20$ observations. Thus, $r_1 = n = 20$. Had only this table been available, the fact that $b_7 = 1$ would have implied that there is one censored value larger than 12, the largest observed value. The value $y_{max} = 15$ indicates that this observation was censored at 15.

Table 14.8   The raw data for Example 14.5.

1	2	3*	4	4	4*	4*	5	7*	8
8	8	9	9	9	9	10*	12	12	15*

Table 14.9   The data for Example 14.5.

i	y_i	s_i	b_i	r_i
0	0	—	—	—
1	1	1	0	20
2	2	1	1	$20 - 1 - 0 = 19$
3	4	2	2	$19 - 1 - 1 = 17$
4	5	1	1	$17 - 2 - 2 = 13$
5	8	3	0	$13 - 1 - 1 = 11$
6	9	4	1	$11 - 3 - 0 = 8$
7	12	2	1	$8 - 4 - 1 = 3$
max	15	—	—	$3 - 2 - 1 = 0$

As noted earlier, calculation of the risk set does not require the exact values of the censored observations, only the interval in which they fall. For example, had the censored observation of 7 been at 6, the resulting numbers at risk would not change (because b_4 would still equal one). ☐

It should be noted that if there is no censoring, so that $b_i = 0$ for all i, then the data are complete and the techniques of Section 14.1 may be used. As such, the approach of this section may be viewed as a generalization.

We shall now present a heuristic derivation of a well-known generalization of the empirical distribution function. This estimator is referred to as either the **Kaplan–Meier** or the **product limit** estimator.

To proceed, we first present some basic facts regarding the distribution of a discrete random variable Y, say, with support on the points $y_1 < y_2 < \cdots < y_k$. Let $p(y_j) = \Pr(Y = y_j)$, and then the survival function is (where $i|y_i > y$ means to take the sum or product over all values of i where $y_i > y$)

$$S(y) = \Pr(Y > y) = \sum_{i|y_i > y} p(y_i).$$

Setting $y = y_j$ for $j < k$, we have

$$S(y_j) = \sum_{i=j+1}^{k} p(y_i),$$

and $S(y_k) = 0$. We also have $S(y_0) = 1$ from the definition of y_0.

Definition 14.9 *The **discrete failure rate** function is*

$$h(y_j) = \Pr(Y = y_j | Y \geq y_j) = p(y_j)/S(y_{j-1}), \quad j = 1, 2, \ldots, k.$$

Thus,

$$h(y_j) = \frac{S(y_{j-1}) - S(y_j)}{S(y_{j-1})} = 1 - \frac{S(y_j)}{S(y_{j-1})},$$

implying that $S(y_j)/S(y_{j-1}) = 1 - h(y_j)$. Hence,

$$S(y_j) = \frac{S(y_j)}{S(y_0)} = \prod_{i=1}^{j} \frac{S(y_i)}{S(y_{i-1})} = \prod_{i=1}^{j} [1 - h(y_i)].$$

Also, $p(y_1) = h(y_1)$, and for $j = 2, 3, \ldots, k$,

$$p(y_j) = h(y_j) S(y_{j-1}) = h(y_j) \prod_{i=1}^{j-1} [1 - h(y_i)].$$

The heuristic derivation proceeds by viewing $\lambda_j = h(y_j)$ for $j = 1, 2, \ldots, k$ as unknown parameters, and estimating them by a nonparametric "maximum likelihood" bsed argument.[3]   For a more detailed discussion, see Lawless [77].   For the present data, the s_j uncensored observations at y_j each contribute $p(y_j)$ to the likelihood where $p(y_1) = \lambda_1$ and

$$p(y_j) = \lambda_j \prod_{i=1}^{j-1} (1 - \lambda_i), \quad j = 2, 3, \ldots, k.$$

Each of the b_j censored observations contributes

$$S(y_j) = \prod_{i=1}^{j} (1 - \lambda_i), \quad j = 1, 2, \ldots, k-1,$$

to the likelihood (recall that $S(y) = S(y_j)$ for $y_j \leq y < y_{j+1}$), and the b_k censored observations at or above y_k each contribute $S(y_k) = \prod_{i=1}^{k} (1 - \lambda_i)$.

The likelihood is formed by taking products over all contributions (assuming independence of all data points), namely

$$L(\lambda_1, \lambda_2, \ldots, \lambda_k) = \prod_{j=1}^{k} \left\{ [p(y_j)]^{s_j} [S(y_j)]^{b_j} \right\},$$

which, in terms of the λ_js, becomes

$$L(\lambda_1, \lambda_2, \ldots, \lambda_k) = \lambda_1^{s_1} \left\{ \prod_{j=2}^{k} \left[\lambda_j \prod_{i=1}^{j-1} (1 - \lambda_i) \right]^{s_j} \right\} \prod_{j=1}^{k} \left[\prod_{i=1}^{j} (1 - \lambda_i) \right]^{b_j}$$

$$= \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) \left[\prod_{j=2}^{k} \prod_{i=1}^{j-1} (1 - \lambda_i)^{s_j} \right] \prod_{j=1}^{k} \prod_{i=1}^{j} (1 - \lambda_i)^{b_j}$$

$$= \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) \left[\prod_{i=1}^{k-1} \prod_{j=i+1}^{k} (1 - \lambda_i)^{s_j} \right] \prod_{i=1}^{k} \prod_{j=i}^{k} (1 - \lambda_i)^{b_j},$$

---

[3]Maximum likelihood estimation is covered in Chapter 11. Candidates preparing for the Society of Actuaries Long-Term Actuarial Mathematics exam will not be tested on the method itself, but only the estimators presented in this chapter.

where the last line follows by interchanging the order of multiplication in each of the two double products. Thus,

$$L(\lambda_1, \lambda_2, \ldots, \lambda_k) = \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) \left(1 - \lambda_k \right)^{b_k} \prod_{i=1}^{k-1} (1 - \lambda_i)^{b_i + \sum_{m=i+1}^{k} (s_m + b_m)}$$

$$= \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) \left(1 - \lambda_k \right)^{b_k} \prod_{i=1}^{k-1} (1 - \lambda_i)^{b_i + \sum_{m=i+1}^{k} (r_m - r_{m+1})}$$

$$= \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) \left(1 - \lambda_k \right)^{r_k - s_k} \prod_{i=1}^{k-1} \left(1 - \lambda_i \right)^{b_i + r_{i+1} - r_{k+1}}.$$

Observe that $r_{k+1} = 0$ and $b_i + r_{i+1} = r_i - s_i$. Hence,

$$L(\lambda_1, \lambda_2, \ldots, \lambda_k) = \left(\prod_{j=1}^{k} \lambda_j^{s_j} \right) \left(1 - \lambda_k \right)^{r_k - s_k} \prod_{i=1}^{k-1} (1 - \lambda_i)^{r_i - s_i}$$

$$= \prod_{j=1}^{k} \lambda_j^{s_j} (1 - \lambda_j)^{r_j - s_j}.$$

This likelihood has the appearance of a product of binomial likelihoods. That is, this is the same likelihood as if $s_1, s_2, \ldots, s_k$ were realizations of k independent binomial observations with parameters $m = r_j$ and $q = \lambda_j$. The "maximum likelihood estimate" $\hat{\lambda}_j$ of λ_j is obtained by taking logarithms, namely

$$l(\lambda_1, \lambda_2, \ldots, \lambda_k) = \ln L(\lambda_1, \lambda_2, \ldots, \lambda_k) = \sum_{j=1}^{k} \left[s_j \ln \lambda_j + (r_j - s_j) \ln(1 - \lambda_j) \right],$$

implying that

$$\frac{\partial l}{\partial \lambda_j} = \frac{s_j}{\lambda_j} - \frac{r_j - s_j}{1 - \lambda_j}, \quad j = 1, 2, \ldots, k.$$

Equating this latter expression to zero yields $\hat{\lambda}_j = s_j / r_j$.

For $y = y_k$, the Kaplan–Meier [66] estimate $S_n(y)$ of $S(y)$ is obtained by replacing λ_j by $\hat{\lambda}_j = s_j / r_j$ wherever it appears. Noting that $S(y) = S(y_j)$ for $y_j \leq y < y_{j+1}$, it follows that

$$S_n(y) = \begin{cases} 1, & y < y_1, \\ \prod_{i=1}^{j} \left(1 - \hat{\lambda}_i \right) = \prod_{i=1}^{j} \left(1 - \frac{s_i}{r_i} \right), & y_j \leq y < y_{j+1}, \quad j = 1, 2, \ldots, k - 1, \\ \prod_{i=1}^{k} \left(1 - \hat{\lambda}_i \right) = \prod_{i=1}^{k} \left(1 - \frac{s_i}{r_i} \right), & y_k \leq y < y_{\max}. \end{cases}$$

This may be written more succinctly as $S_n(y) = \prod_{i | y_i \leq y} (1 - \hat{\lambda}_i)$ for $y < y_{\max}$. When $y_{\max} = y_k$, you should interpret $y_k \leq y < y_{\max}$ as $y = y_k$.

Table 14.10   The Kaplan–Meier estimates for Example 14.6.

i	y_i	s_i	r_i	$\hat{S}_n(y_i)$
1	1	1	20	$1 - 1/20 = 0.950$
2	2	1	19	$0.95(1 - 1/19) = 0.900$
3	4	2	17	$0.9(1 - 2/17) = 0.794$
4	5	1	13	$0.794(1 - 1/13) = 0.733$
5	8	3	11	$0.733(1 - 3/11) = 0.533$
6	9	4	8	$0.533(1 - 4/8) = 0.267$
7	12	2	3	$0.267(1 - 2/3) = 0.089$

■ **EXAMPLE 14.6**

Construct the Kaplan–Meier estimate of $S(y)$ using the data in Example 14.5. Indicate how the answer would change if $s_7 = 3$ and $b_7 = 0$.

The calculations appear in Table 14.10. The estimated survival function is

$$S_{20}(y) = \begin{cases} = 1, & y < 1, \\ = 0.950, & 1 \le y < 2, \\ = 0.900, & 2 \le y < 4, \\ = 0.794, & 4 \le y < 5, \\ = 0.737, & 5 \le y < 8, \\ = 0.533, & 8 \le y < 9, \\ = 0.167, & 9 \le y < 12, \\ = 0.089, & 12 \le y < 15. \end{cases}$$

With the change in values, we have $y_{\max} = y_7 = 12$ and $S_{20}(y) = 0.267(1-3/3) = 0$ for $y = 12$. □

We now discuss estimation for $y \ge y_{\max}$. First, note that if $s_k = r_k$ (no censored observations at y_k), then $S_n(y_k) = 0$ and $S_n(y) = 0$ for $y \ge y_k$ is clearly the (only) obvious choice. However, if $S_n(y_k) > 0$, as in the previous example, there are no empirical data to estimate $S(y)$ for $y \ge y_{\max}$, and tail estimates for $y \ge y_{\max}$ (often called tail corrections) are needed. There are three popular extrapolations:

- Efron's tail correction [31] assumes that $S_n(y) = 0$ for $y \ge y_{\max}$.

- Klein and Moeschberger [70, p. 118] assume that $S_n(y) = S_n(y_k)$ for $y_k \le y < \gamma$ and $S_n(y) = 0$ for $y \ge \gamma$, where $\gamma > y_{\max}$ is a plausible upper limit for the underlying random variable. For example, in a study of human mortality, the limit might be 120 years.

- Brown, Hollander, and Korwar's exponential tail correction [18] assumes that $S_n(y_{\max}) = S_n(y_k)$ and that $S_n(y) = e^{-\hat{\beta}y}$ for $y \ge y_{\max}$. With $y = y_{\max}$, $\hat{\beta} = -\ln S_n(y_k)/y_{\max}$, and thus

$$S_n(y) = e^{y[\ln S_n(y_k)]/y_{\max}} = \left[S_n(y_k)\right]^{y/y_{\max}}, \quad y \ge y_{\max}.$$

■ **EXAMPLE 14.7**

Apply all three tail correction methods to the data used in Example 14.6. Assume that $\gamma = 22$.

Efron's method has $S_{20}(y) = 0$, $y \geq 15$. With $\gamma = 22$, Klein and Moeschberger's method has $S_{20}(y) = 0.089$, $15 \leq y < 22$. and $S_{20}(y) = 0$, $y \geq 22$. The exponential tail correction has $S_{20}(y) = (0.089)^{y/15}$, $y \geq 15$.   □

Note that if there is no censoring ($b_i = 0$ for all i), then $r_{i+1} = r_i - s_i$, and for $y_j \leq y < y_{j+1}$

$$S_n(y) = \prod_{i=1}^{j} \left(\frac{r_i - s_i}{r_i} \right) = \prod_{i=1}^{j} \frac{r_{i+1}}{r_i} = \frac{r_{j+1}}{r_1}.$$

In this case, r_{j+1} is the number of observations exceeding y and $r_1 = n$. Thus, with no censoring, the Kaplan–Meier estimate reduces to the empirical estimate of the previous section.

An alternative to the Kaplan–Meier estimator, called the **Nelson–Åalen estimator** [1], [93], is sometimes used. To motivate the estimator, note that if $S(y)$ is the survival function of a continuous distribution with failure rate $h(y)$, then $-\ln S(y) = H(y) = \int_0^y h(t)dt$ is called the cumulative hazard rate function. The discrete analog is, in the present context, given by $\sum_{i|y_i \leq y} \lambda_i$, which can intuitively be estimated by replacing λ_i by its estimate $\hat{\lambda}_i = s_i/r_i$. The Nelson–Åalen estimator of $H(y)$ is thus defined for $y < y_{\max}$ to be

$$\hat{H}(y) = \begin{cases} 0, & y < y_1, \\ \sum_{i=1}^{j} \hat{\lambda}_i = \sum_{i=1}^{j} \frac{s_i}{r_i}, & y_j \leq y < y_{j+1}, \quad j = 1, 2, \ldots, k-1, \\ \sum_{i=1}^{k} \hat{\lambda}_i = \sum_{i=1}^{k} \frac{s_i}{r_i}, & y_k \leq y < y_{\max}. \end{cases}$$

That is, $\hat{H}(y) = \sum_{i|y_i \leq y} \hat{\lambda}_i$ for $y < y_{\max}$, and the Nelson–Åalen estimator of the survival function is $\hat{S}(y) = \exp(-\hat{H}(y))$. The notation under the summation sign indicates that values of $\hat{\lambda}_i$ should be included only if $y_i \leq y$. For $y \geq y_{\max}$, the situation is similar to that involving the Kaplan–Meier estimate in the sense that a tail correction of the type discussed earlier needs to be employed. Note that, unlike the Kaplan–Meier estimate, $\hat{S}(y_k) > 0$, so that a tail correction is always needed.

■ **EXAMPLE 14.8**

Determine the Nelson–Åalen estimates for the data in Example 14.6. Continue to assume $\gamma = 22$.

The estimates of the cumulative hazard function are given in Table 14.11. The estimated survival function is

Table 14.11   The Nelson–Åalen estimates for Example 14.8.

i	y_i	s_i	r_i	$\hat{H}_n(y_i)$
1	1	1	20	$1/20 = 0.050$
2	2	1	19	$0.05 + 1/19 = 0.103$
3	4	2	17	$0.103 + 2/17 = 0.220$
4	5	1	13	$0.220 + 1/13 = 0.297$
5	8	3	11	$0.297 + 3/11 = 0.570$
6	9	4	8	$0.570 + 4/8 = 1.070$
7	12	2	3	$1.070 + 2/3 = 1.737$

$$\hat{S}_{20}(y) = \begin{cases} = 1, & y < 1, \\ = e^{-0.050} = 0.951, & 1 \le y < 2, \\ = e^{-0.103} = 0.902, & 2 \le y < 4, \\ = e^{-0.220} = 0.803, & 4 \le y < 5, \\ = e^{-0.297} = 0.743, & 5 \le y < 8, \\ = e^{-0.570} = 0.566, & 8 \le y < 9, \\ = e^{-1.070} = 0.343, & 9 \le y < 12, \\ = e^{-1.737} = 0.176, & 12 \le y < 15. \end{cases}$$

With regard to tail correction, Efron's method has $S_{20}(y) = 0$, $y \ge 15$. Klein and Moeschberger's method has $S_{20}(y) = 0.176$, $15 \le y < 22$, and $S_{20}(y) = 0$, $y \ge 22$. The exponential tail correction has $S_{20}(y) = (0.176)^{y/15}$, $y \ge 15$.   □

To assess the quality of the two estimators, we will now consider estimation of the variance. Recall that for $y < y_{\max}$, the Kaplan–Meier estimator may be expressed as

$$S_n(y) = \prod_{i|y_i \le y} \left(1 - \hat{\lambda}_i\right),$$

which is a function of the $\hat{\lambda}_i$s. Thus, to estimate the variance of $S_n(y)$, we first need the covariance matrix of $(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k)$. We estimate this from the "likelihood," using standard likelihood methods. Recall that

$$L\left(\lambda_1, \lambda_2, \ldots, \lambda_k\right) = \prod_{j=1}^k \lambda_j^{s_j}(1 - \lambda_j)^{r_j - s_j},$$

and thus $l = \ln L$ satisfies

$$l\left(\lambda_1, \lambda_2, \ldots, \lambda_k\right) = \sum_{j=1}^k \left[s_j \ln \lambda_j + (r_j - s_j)\ln(1 - \lambda_j)\right].$$

Thus,

$$\frac{\partial l}{\partial \lambda_i} = \frac{s_i}{\lambda_i} - \frac{r_i - s_i}{1 - \lambda_i}, \quad i = 1, 2, \ldots, k,$$

and

$$\frac{\partial^2 l}{\partial \lambda_i^2} = -\frac{s_i}{\lambda_i^2} - \frac{r_i - s_i}{\left(1 - \lambda_i\right)^2},$$

which, with λ_i replaced by $\hat{\lambda}_i = s_i/r_i$, becomes

$$\frac{\partial^2 l}{\partial \lambda_i^2}\Big|_{\lambda_i = \hat{\lambda}_i} = -\frac{s_i}{\hat{\lambda}_i^2} - \frac{r_i - s_i}{(1 - \hat{\lambda}_i)^2} = -\frac{r_i^3}{s_i(r_i - s_i)}.$$

For $i \neq m$,

$$\frac{\partial^2 l}{\partial \lambda_i \partial \lambda_m} = 0.$$

The observed information, evaluated at the maximum likelihood estimate, is thus a diagonal matrix, which when inverted yields the estimates

$$\mathrm{Var}(\hat{\lambda}_i) \doteq \frac{s_i(r_i - s_i)}{r_i^3} = \frac{\hat{\lambda}_i(1 - \hat{\lambda}_i)}{r_i}, \quad i = 1, 2, \dots, k,$$

and

$$\mathrm{Cov}(\hat{\lambda}_i, \hat{\lambda}_m) \doteq 0, \quad i \neq m.$$

These results also follow directly from the binomial form of the likelihood.

Returning to the problem at hand, the delta method[4] gives the approximate variance of $f(\hat{\theta})$ as $\mathrm{Var}[f(\hat{\theta})] \doteq [f'(\hat{\theta})]^2 \mathrm{Var}(\hat{\theta})$, for an estimator $\hat{\theta}$.

To proceed, note that

$$\ln S_n(y) = \sum_{i|y_i \leq y} \ln(1 - \hat{\lambda}_i),$$

and since the $\hat{\lambda}_i$s are assumed to be approximately uncorrelated,

$$\mathrm{Var}[\ln S_n(y)] \doteq \sum_{i|y_i \leq y} \mathrm{Var}[\ln(1 - \hat{\lambda}_i)].$$

The choice $f(x) = \ln(1 - x)$ yields

$$\mathrm{Var}[\ln(1 - \hat{\lambda}_i)] \doteq \frac{\mathrm{Var}(\hat{\lambda}_i)}{(1 - \hat{\lambda}_i)^2} \doteq \frac{\hat{\lambda}_i}{r_i(1 - \hat{\lambda}_i)} = \frac{s_i}{r_i(r_i - s_i)},$$

implying that

$$\mathrm{Var}[\ln S_n(y)] \doteq \sum_{i|y_i \leq y} \frac{s_i}{r_i(r_i - s_i)}.$$

Because $S_n(y) = \exp[\ln S_n(y)]$, the delta method with $f(x) = \exp(x)$ yields

$$\mathrm{Var}[S_n(y)] \doteq \{\exp[\ln S_n(y)]\}^2 \mathrm{Var}[\ln S_n(y)] = [S_n(y)]^2 \mathrm{Var}[\ln S_n(y)],$$

---

[4]The delta method is presented in Section 11.6. Candidates for the Society of Actuaries Long-Term Actuarial Mathematics exam are not expected to be able to apply this method to given problems. In this chapter, it is important to know that many of the variance formulas are approximate due to being derived by the delta method.

This yields the final version of the estimate,

$$\widehat{\mathrm{Var}}[S_n(y)] = [S_n(y)]^2 \sum_{i|y_i \le y} \frac{s_i}{r_i(r_i - s_i)}. \tag{14.2}$$

Equation (14.2) holds for $y < y_{\max}$ in all cases. However, if $y_{\max} = y_k$ (that is, there are no censored observations after the last uncensored observation), then it holds for $y \le y_{\max} = y_k$. Hence the formula always holds for $y \le y_k$.

Formula (14.2) is known as **Greenwood's approximation** to the variance of $S_n(y)$, and is known to often understate the true variance.

If there is no censoring, and we take $y_j \le y < y_{j+1}$, then Greenwood's approximation yields

$$\widehat{\mathrm{Var}}[S_n(y)] = \widehat{\mathrm{Var}}[S_n(y_j)] = \left[S_n(y_j)\right]^2 \sum_{i=1}^{j} \frac{s_i}{r_i(r_i - s_i)},$$

which may be expressed (using $r_{i+1} = r_i - s_i$ due to no censoring) as

$$\widehat{\mathrm{Var}}[S_n(y)] = [S_n(y_j)]^2 \sum_{i=1}^{j} \left(\frac{1}{r_i - s_i} - \frac{1}{r_i} \right)$$

$$= [S_n(y_j)]^2 \sum_{i=1}^{j} \left(\frac{1}{r_{i+1}} - \frac{1}{r_i} \right).$$

Because $r_1 = n$, this sum telescopes to give

$$\widehat{\mathrm{Var}}[S_n(y)] = [S_n(y_j)]^2 \left(\frac{1}{r_{j+1}} - \frac{1}{r_1} \right)$$

$$= [S_n(y_j)]^2 \left[\frac{1}{nS_n(y_j)} - \frac{1}{n} \right]$$

$$= \frac{S_n(y_j)[1 - S_n(y_j)]}{n},$$

which is the same estimate as obtained in Section 14.1, but derived without use of the delta method.

We remark that in the case with $r_k = s_k$ (i.e. $S_n(y_k) = 0$), Greenwood's approximation cannot be used to estimate the variance of $S_n(y_k)$. In this case, $r_k - s_k$ is often replaced by r_k in the denominator.

Turning now to the Nelson–Åalen estimator, we note that

$$\hat{H}(y) = \sum_{i|y_i \le y} \hat{\lambda}_i,$$

and the same reasoning used for Kaplan–Meier implies that $\mathrm{Var}[\hat{H}(y)] \doteq \sum_{i|y_i \le y} \mathrm{Var}(\hat{\lambda}_i)$, yielding the estimate

$$\widehat{\mathrm{Var}}[\hat{H}(y)] = \sum_{i|y_i \le y} \frac{s_i(r_i - s_i)}{r_i^3}, \tag{14.3}$$

which is referred to as **Klein's estimate**. A commonly used alternative estimate due to Åalen is obtained by replacing $r_i - s_i$ with r_i in the numerator.

We are typically more interested in $S(t)$ than $H(t)$. Because $\hat{S}(y) = \exp[-\hat{H}(y)]$, the delta method with $f(x) = e^{-x}$ yields Klein's survival function estimate

$$\text{Var}[\hat{S}(y)] \doteq [\exp(-\hat{H}(y))]^2 \widehat{\text{Var}}[\hat{H}(y)],$$

that is, the estimated variance is

$$\widehat{\text{Var}}[\hat{S}(y)] = [\hat{S}(y)]^2 \sum_{i|y_i \le y} \frac{s_i(r_i - s_i)}{r_i^3}, \quad y < y_{\max}.$$

■ **EXAMPLE 14.9**

For the data of Example 14.5, estimate the variance of the Kaplan–Meier estimators of $S(2)$ and $S(9)$, and the Nelson–Åalen estimator of $S(2)$.

For the Kaplan–Meier estimators,

$$\widehat{\text{Var}}[S_{20}(2)] = \left[S_{20}(2)\right]^2 \left[\frac{1}{20(19)} + \frac{1}{19(18)}\right]$$
$$= (0.90)^2(0.00556)$$
$$= 0.0045,$$

$$\widehat{\text{Var}}[S_{20}(9)] = [S_{20}(9)]^2 \left[\frac{1}{20(19)} + \frac{1}{19(18)} + \frac{2}{17(15)} + \frac{1}{13(12)} + \frac{3}{11(8)} + \frac{4}{8(4)}\right]$$
$$= (0.26656)^2(0.17890)$$
$$= 0.01271,$$

and for the Nelson–Åalen estimator,

$$\widehat{\text{Var}}[\hat{S}(2) = [\hat{S}(2)]^2 \left[\frac{1(19)}{(20)^3} + \frac{1(18)}{(19)^3}\right]$$
$$= (0.90246)^2(0.00500)$$
$$= 0.00407.$$

□

Variance estimates for $y \ge y_{\max}$ depend on the tail correction used. Efron's method gives an estimate of 0, which is not of interest in the present context. For the exponential tail correction in the Kaplan–Meier case, we have for $y \ge y_{\max}$, $S_n(y) = S_n(y_k)^{y/y_{\max}}$, and the delta method with $f(x) = x^{y/y_{\max}}$ yields

$$\widehat{\text{Var}}[S_n(y) = \left[\frac{y}{y_{\max}} S_n\left(y_k\right)^{\frac{y}{y_{\max}}-1}\right]^2 \widehat{\text{Var}}[S_n(y_k)]$$
$$= \left(\frac{y}{y_{\max}}\right)^2 [S_n(y)]^2 \sum_{i=1}^{k} \frac{s_i}{r_i(r_i - s_i)}$$
$$= \left(\frac{y}{y_{\max}}\right)^2 \left[\frac{S_n(y)}{S_n(y_k)}\right]^2 \widehat{\text{Var}}[S_n(y_k)].$$

Likelihood methods typically result in approximate asymptotic normality of the estimates, and this is true for Kaplan–Meier and Nelson–Åalen estimates as well. Using the results of Example 14.9, an approximate 95% confidence interval for $S(9)$ is given by

$$S_{20}(9) \pm 1.96\sqrt{\{\mathrm{Var}[S_{20}(9)]\}} = 0.26656 \pm 1.96\sqrt{0.01271} = (0.04557, 0.48755).$$

For $S(2)$, the Nelson–Åalen estimate gives a confidence interval of

$$0.90246 \pm 1.96\sqrt{0.00407} = (0.77740, 1.02753),$$

whereas that based on the Kaplan–Meier estimate is

$$0.90 \pm 1.96\sqrt{0.0045} = (0.76852, 1.03148).$$

Clearly, both confidence intervals for $S(2)$ are unsatisfactory, both including values greater than 1.

An alternative approach can be constructed as follows, using the Kaplan–Meier estimate as an example.

Let $Y = \ln[-\ln S_n(y)]$. Using the delta method, the variance of Y can be approximated as follows. The function of interest is $f(x) = \ln(-\ln x)$. Its derivative is

$$f'(x) = \frac{1}{-\ln x}\frac{-1}{x} = \frac{1}{x\ln x}.$$

According to the delta method,

$$\widehat{\mathrm{Var}}(Y) = \{f'[S_n(y)]\}^2\mathrm{Var}[S_n(y)] = \frac{\mathrm{Var}[S_n(y)]}{[S_n(y)\ln S_n(y)]^2}.$$

Then, an approximate 95% confidence interval for $\theta = \ln[-\ln S(y)]$ is

$$\ln[-\ln S_n(y)] \pm 1.96\frac{\sqrt{\widehat{\mathrm{Var}}[S_n(y)]}}{S_n(y)\ln S_n(y)}.$$

Because $S(y) = \exp(-e^Y)$, evaluating each endpoint of this formula provides a confidence interval for $S(y)$. For the upper limit, we have (where $\hat{v} = \widehat{\mathrm{Var}}[S_n(y)]$)

$$\exp\left\{-e^{\ln[-\ln S_n(y)]+1.96\sqrt{\hat{v}}/[S_n(y)\ln S_n(y)]}\right\} = \exp\left\{[\ln S_n(y)]e^{1.96\sqrt{\hat{v}}/[S_n(y)\ln S_n(y)]}\right\}$$

$$= S_n(y)^U, \quad U = \exp\left[\frac{1.96\sqrt{\hat{v}}}{S_n(y)\ln S_n(y)}\right].$$

Similarly, the lower limit is $S_n(y)^{1/U}$. This interval will always be inside the range 0–1 and is referred to as a **log-transformed confidence interval**.

■ **EXAMPLE 14.10**

Construct a 95% log-transformed confidence interval for $S(2)$ in Example 14.9 based on the Kaplan–Meier estimator.

In this case, $S_{20}(2) = 0.9$, and $U = \exp\left[\frac{1.96(0.06708)}{0.90\ln 0.90}\right] = 0.24994$. The lower limit is $(0.90)^{1/U} = 0.65604$ and the upper limit is $(0.90)^U = 0.97401$. $\qquad\square$

For the Nelson–Åalen estimator, a similar log-transformed confidence interval for $H(y)$ has endpoints $\hat{H}(y)U$, where $U = \exp[\pm 1.96\sqrt{\widehat{\mathrm{Var}}[\hat{H}(y)]}/\hat{H}(y)]$. Exponentiation of the negative of these endpoints yields a corresponding interval for $S(y)$.

■ **EXAMPLE 14.11**

Construct 95% linear and log-transformed confidence intervals for $H(2)$ in Example 14.9 based on the Nelson–Åalen estimate. Then construct a 95% log-transformed confidence interval for $S(2)$.

From (14.3), a 95% linear confidence interval for $H(2)$ is

$$0.10263 \pm 1.96\sqrt{\frac{1(19)}{(20)^3} + \frac{1(18)}{(19)^3}} = 0.10263 \pm 1.96\sqrt{0.00500}$$

$$= (-0.03595, 0.24121)$$

which includes negative values. For a log-transformed confidence interval, we have

$$U = \exp[\pm 1.96\sqrt{0.00500}/0.10263]$$
$$= 0.25916 \text{ and } 3.8586.$$

The interval is from

$$(0.10263)(0.25916) = 0.02660 \text{ to } (0.10263)(3.85865) = 0.39601.$$

The corresponding interval for $S(2)$ is from

$$\exp(-0.39601) = 0.67300 \text{ to } \exp(-0.02660) = 0.97375. \qquad\square$$

14.3.1 Exercises

14.10 (*) You are given the following times of first claim for five randomly selected automobile insurance policies: 1, 2, 3, 4, and 5. You are later told that one of the five times given is actually the time of policy lapse, but you are not told which one. The smallest product-limit estimate of $S(4)$, the probability that the first claim occurs after time 4, would result if which of the given times arose from the lapsed policy?

14.11 (*) For a mortality study with right censored data, you are given the information in Table 14.12. Calculate the estimate of the survival function at time 12 using the Nelson–Åalen estimate.

14.12 (*) Let n be the number of lives observed from birth. None were censored and no two lives died at the same age. At the time of the ninth death, the Nelson–Åalen estimate of the cumulative hazard rate is 0.511, and at the time of the tenth death it is 0.588. Estimate the value of the survival function at the time of the third death.

Table 14.12   The data for Exercise 14.11.

Time t_j	Number of deaths s_j	Number at risk r_j
5	2	15
7	1	12
10	1	10
12	2	6

14.13 (*) All members of a study joined at birth; however, some may exit the study by means other than death. At the time of the third death, there was one death (i.e. $s_3 = 1$); at the time of the fourth death, there were two deaths; and at the time of the fifth death, there was one death. The following product-limit estimates were obtained: $S_n(y_3) = 0.72$, $S_n(y_4) = 0.60$, and $S_n(y_5) = 0.50$. Determine the number of censored observations between times y_4 and y_5. Assume that no observations were censored at the death times.

14.14 (*) A mortality study has right censored data and no left truncated data. Uncensored observations occurred at ages 3, 5, 6, and 10. The risk sets at these ages were 50, 49, k, and 21, respectively, while the number of deaths observed at these ages were 1, 3, 5, and 7, respectively. The Nelson–Åalen estimate of the survival function at time 10 is 0.575. Determine k.

14.15 (*) Consider the observations 2,500, 2,500, 2,500, 3,617, 3,662, 4,517, 5,000, 5,000, 6,010, 6,932, 7,500, and 7,500. No truncation is possible. First, determine the Nelson–Åalen estimate of the cumulative hazard rate function at 7,000, assuming that all the observations are uncensored. Second, determine the same estimate, assuming that the observations at 2,500, 5,000, and 7,500 were right censored.

14.16 (*) No observations in a data set are truncated. Some are right censored. You are given $s_3 = 1$, $s_4 = 3$, and the Kaplan–Meier estimates $S_n(y_3) = 0.65$, $S_n(y_4) = 0.50$, and $S_n(y_5) = 0.25$. Also, between the observations y_4 and y_5 there are six right censored observations and no observations were right censored at the same value as an uncensored observation. Determine s_5.

14.17 For Data Set A, determine the empirical estimate of the probability of having two or more accidents and estimate its variance.

14.18 (*) Ten individuals were observed from birth. All were observed until death. Table 14.13 gives the death ages. Let V_1 denote the estimated conditional variance of $_3\hat{q}_7$ if calculated without any distribution assumption. Let V_2 denote the conditional variance of $_3\hat{q}_7$ if calculated knowing that the survival function is $S(t) = 1 - t/15$. Determine $V_1 - V_2$.

14.19 (*) Observations can be censored, but there is no truncation. Let y_j and y_{j+1} be consecutive death ages. A 95% linear confidence interval for $H(y_j)$ using the Klein estimator is $(0.07125, 0.22875)$, while a similar interval for $H(y_{j+1})$ is $(0.15985, 0.38257)$. Determine s_{j+1}.

Table 14.13   The data for Exercise 14.18.

Age	Number of deaths
2	1
3	1
5	1
7	2
10	1
12	2
13	1
14	1

14.20 (*) A mortality study is conducted on 50 lives, all observed from age 0. At age 15 there were two deaths; at age 17 there were three censored observations; at age 25 there were four deaths; at age 30 there were c censored observations; at age 32 there were eight deaths; and at age 40 there were two deaths. Let S be the product-limit estimate of $S(35)$ and let V be the Greenwood approximation of this estimator's variance. You are given $V/S^2 = 0.011467$. Determine the value of c.

14.21 (*) Fifteen cancer patients were observed from the time of diagnosis until the earlier of death or 36 months from diagnosis. Deaths occurred as follows: at 15 months there were two deaths; at 20 months there were three deaths; at 24 months there were two deaths; at 30 months there were d deaths; at 34 months there were two deaths; and at 36 months there was one death. The Nelson–Åalen estimate of $H(35)$ is 1.5641. Determine Klein's estimate of the variance of this estimator.

14.22 (*) Ten payments were recorded as follows: 4, 4, 5, 5, 5, 8, *10*, *10*, 12, and 15, with the italicized values representing payments at a policy limit. There were no deductibles. Determine the product-limit estimate of $S(11)$ and Greenwood's approximation of its variance.

14.23 (*) All observations begin on day 0. Eight observations were *4*, 8, 8, 12, *12*, 12, 22, and 36, with the italicized values representing right censored observations. Determine the Nelson–Åalen estimate of $H(12)$ and then determine a 90% linear confidence interval for the true value using Klein's variance estimate.

14.24 You are given the data in Table 14.14, based on 40 observations. Dashes indicate missing observations that must be deduced.

Table 14.14   The data for Exercise 14.24.

i	y_i	s_i	b_i	r_i
1	4	3	—	40
2	6	—	3	31
3	9	6	4	23
4	13	4	—	—
5	15	2	4	6

(a) Compute the Kaplan–Meier estimate $S_{40}(y_i)$ for $i = 1, 2, \ldots, 5$.

(b) Compute the Nelson–Åalen estimate $\hat{H}(y_i)$ for $i = 1, 2, \ldots, 5$.

(c) Compute $S_{40}(24)$ using the method of Brown, Hollander, and Korwar.

(d) Compute Greenwood's approximation, $\widehat{\text{Var}}[S_{40}(15)]$.

(e) Compute a 95% linear confidence interval for $S(15)$ using the Kaplan–Meier estimate.

(f) Compute a 95% log-transformed confidence interval for $S(15)$ using the Kaplan–Meier estimate.

14.25 You are given the data in Table 14.15, based on 50 observations.

(a) Compute the Kaplan–Meier estimate $S_{50}(y_i)$ for $i = 1, 2, \ldots, 6$.

(b) Compute the Nelson–Åalen estimate $\hat{S}(y_i)$ for $i = 1, 2, \ldots, 6$.

(c) Compute $S_{50}(40)$ using Efron's tail correction, and also using the exponential tail correction of Brown, Hollander, and Korwar.

(d) Compute Klein's survival function estimate of the variance of $\hat{S}(20)$.

(e) Compute a 95% log-transformed confidence interval for $H(20)$ based on the Nelson–Åalen estimate.

(f) Using the exponential tail correction method of Brown, Hollander, and Korwar, estimate the variance of $\hat{S}(40)$.

14.26 Consider the estimator

$$\tilde{S}(y) = \prod_{i|y_i \le y} \phi\left(\hat{\lambda}_i\right)$$

where ϕ is differentiable.

(a) Show that $\tilde{S}(y)$ becomes the Kaplan–Meier estimator $S_n(y)$ when $\phi(x) = 1 - x$, and $\tilde{S}(y)$ becomes the Nelson–Åalen estimator $\hat{S}(y)$ when $\phi(x) = e^{-x}$.

(b) Derive the variance estimate

$$\text{Var}\left[\tilde{S}(y)\right] \approx \left[\tilde{S}(y)\right]^2 \sum_{i|y_i \le y} \left[\frac{\phi'\left(\hat{\lambda}_i\right)}{\phi\left(\hat{\lambda}_i\right)}\right]^2 \frac{s_i(r_i - s_i)}{r_i^3}.$$

Table 14.15 The data for Exercise 14.25.

i	y_i	s_i	b_i	r_i
1	3	3	6	50
2	5	7	4	41
3	7	5	2	30
4	11	5	3	23
5	16	6	4	15
6	20	2	3	5

(c) Consider

$$\tilde{S}_m(y) = \prod_{i|y_i \le y} \phi_m\left(\hat{\lambda}_i\right) \text{ with } \phi_m(x) = \sum_{j=0}^{m}(-x)^j/j! \text{ for } m = 0, 1, 2, \dots.$$

Prove that $\hat{S}(y) \le \tilde{S}_{2m}(y)$ if $m = 0, 1, 2, \dots$, and $\hat{S}(y) \ge \tilde{S}_{2m+1}(y)$ if $m = 0, 1, 2, \dots$, and thus that $\hat{S}(y) \ge S_n(y)$ in particular.

Hint: Prove by induction on m the identity $(-1)^m \left[\phi_m(x) - e^{-x}\right] \ge 0$ for $x \ge 0$ and $m = 0, 1, 2, \dots$.

14.4   Empirical Estimation of Moments

In the previous section, we focused on estimation of the survival function $S(y)$ or, equivalently, the cumulative distribution function $F(y) = 1 - S(y)$, of a random variable Y. In many actuarial applications, other quantities such as raw moments are of interest. Of central importance in this context is the mean, particularly for premium calculation in a loss modelling context.

For estimation of the mean with complete data $Y_1, Y_2, \dots, Y_n$, an obvious (unbiased) estimation is $\hat{\mu} = (Y_1 + Y_2 + \cdots Y_n)/n$, but for incomplete data such as that of the previous section involving right censoring, other methods are needed. We continue to assume that we have the setting described in the previous section, and we will capitalize on the results obtained for $S(y)$ there. To do so, we recall that, for random variables that take on only nonnegative values, the mean satisfies

$$\mu = \mathrm{E}(Y) = \int_0^\infty S(y)\,dy,$$

and empirical estimation of μ may be done by replacing $S(y)$ with an estimator such as the Kaplan–Meier estimator $S_n(y)$ or the Nelson–Åalen estimator $\hat{S}(y)$. To unify the approach, we will assume that $S(y)$ is estimated for $y < y_{\max}$ by the estimator given in Exercise 14.26 of Section 14.3, namely

$$\tilde{S}(y) = \prod_{i|y_i \le y} \phi(\hat{\lambda}_i),$$

where $\phi(x) = 1 - x$ for the Kaplan–Meier estimator and $\phi(x) = e^{-x}$ for the Nelson–Åalen estimator. The mean is obtained by replacing $S(y)$ with $\tilde{S}(y)$ in the integrand. This yields the estimator

$$\tilde{\mu} = \int_0^\infty \tilde{S}(y)\,dy.$$

It is convenient to write

$$\tilde{\mu} = \int_0^{y_{\max}} \tilde{S}(y)\,dy + \tilde{\tau}(y_{\max}),$$

where

$$\tau(x) = \int_x^\infty S(y)\,dy, \quad x \ge 0$$

and

$$\tilde{\tau}(x) = \int_x^\infty \tilde{S}(y)\,dy, \quad x \ge 0.$$

Anticipating what follows, we wish to evaluate $\tau(y_m)$ for $m = 1, 2, \ldots, k$. For $m = k$, we have that $S(y) = S(y_k) = \prod_{i=1}^{k} \phi(\lambda_i)$ for $y_k \leq y < y_{\max}$. Thus

$$\tau(y_k) = \tau(y_{\max}) + (y_{\max} - y_k) \prod_{i=1}^{k} \phi(\lambda_i).$$

To evaluate $\tau(y_m)$ for $m = 1, 2, \ldots, k - 1$, recall that $S(y) = 1$ for $0 \leq y < y_1$ and for $y_j \leq y < y_{j+1}$, $S(y) = S(y_j) = \prod_{i=1}^{j} \phi(\lambda_i)$. Thus,

$$\tau(y_m) = \tau(y_k) + \int_{y_m}^{y_k} S(y)\,dy$$

$$= \tau(y_k) + \sum_{j=m+1}^{k} \int_{y_{j-1}}^{y_j} S(y)\,dy$$

$$= \tau(y_k) + \sum_{j=m+1}^{k} \int_{y_{j-1}}^{y_j} S(y_{j-1})\,dy$$

$$= \tau(y_k) + \sum_{j=m+1}^{k} (y_j - y_{j-1}) \prod_{i=1}^{j-1} \phi(\lambda_i).$$

For evaluation of μ, note that

$$\mu = \int_0^{y_1} S(y)\,dy + \tau(y_1) = y_1 + \tau(y_1),$$

and also that for $m = 1, 2, \ldots, k - 1$,

$$\tau(y_m) = \tau(y_{m+1}) + (y_{m+1} - y_m)S(y_m),$$

a recursive formula, beginning with $\tau(y_k)$.

For the estimates themselves, $\tilde{S}(y_j) = \prod_{i=1}^{j} \phi(\hat{\lambda}_i)$, and the above formulas continue to hold when $\tau(y_m)$ is replaced by $\tilde{\tau}(y_m)$, $S(y)$ by $\tilde{S}(y)$, and μ by $\tilde{\mu}$.

The estimate of the mean μ clearly depends on $\tau(y_{\max})$, which in turn depends on the tail correction, that is, on $S(y)$ for $y \geq y_{\max}$. If $S(y) = 0$ for $y \geq y_{\max}$ (as, for example, under Efron's tail correction), then $\tau(y_{\max}) = 0$. Under Klein and Moeschberger's method, with $S(y) = S(y_k)$ for $y_k \leq y < \gamma$, and $S(y) = 0$ for $y \geq \gamma$ where $\gamma > y_{\max}$,

$$\tau(y_{\max}) = \int_{y_{\max}}^{\gamma} S(y)\,dy + \int_{\gamma}^{\infty} S(y)\,dy = (\gamma - y_{\max})S(y_k).$$

For the exponential tail correction of Brown, Hollander, and Korwar, $S(y) = e^{-\beta y}$ for $y \geq y_{\max}$ with $\beta = -\ln S(y_k)/y_{\max}$. Thus

$$\tau(y_{\max}) = \int_{y_{\max}}^{\infty} e^{-\beta y}\,dy = \frac{1}{\beta} e^{-\beta y_{\max}} = \frac{y_{\max} S(y_k)}{-\ln S(y_k)}.$$

The following example illustrates the calculation of $\tilde{\mu}$, where all empirical quantities are obtained by substitution of estimates.

■ **EXAMPLE 14.12**

Calculate the estimates of the mean for the Nelson–Åalen estimate of Example 14.8 under the three tail corrections. Continue to assume that $\gamma = 22$.

We have, in terms of $\tilde{\tau}(y_{\max}) = \tilde{\tau}(15)$,

$$\tilde{\tau}(y_7) = \tilde{\tau}(12) = (15 - 12)\hat{S}(12) + \tilde{\tau}(15) = 3(0.176) + \tilde{\tau}(15) = 0.528 + \tilde{\tau}(15),$$

$$\tilde{\tau}(y_6) = \tilde{\tau}(9) = (12 - 9)\hat{S}(9) + \tilde{\tau}(12) = 3(0.343) + \tilde{\tau}(12) = 1.557 + \tilde{\tau}(15),$$

$$\tilde{\tau}(y_5) = \tilde{\tau}(8) = (9 - 8)\hat{S}(8) + \tilde{\tau}(9) = 1(0.566) + \tilde{\tau}(9) = 2.123 + \tilde{\tau}(15),$$

$$\tilde{\tau}(y_4) = \tilde{\tau}(5) = (8 - 5)\hat{S}(5) + \tilde{\tau}(8) = 3(0.743) + \tilde{\tau}(8) = 4.352 + \tilde{\tau}(15),$$

$$\tilde{\tau}(y_3) = \tilde{\tau}(4) = (5 - 4)\hat{S}(4) + \tilde{\tau}(5) = 1(0.803) + \tilde{\tau}(5) = 5.155 + \tilde{\tau}(15),$$

$$\tilde{\tau}(y_2) = \tilde{\tau}(2) = (4 - 2)\hat{S}(2) + \tilde{\tau}(4) = 2(0.902) + \tilde{\tau}(4) = 6.959 + \tilde{\tau}(15),$$

$$\tilde{\tau}(y_1) = \tilde{\tau}(1) = (2 - 1)\hat{S}(1) + \tilde{\tau}(2) = 1(0.951) + \tilde{\tau}(2) = 7.910 + \tilde{\tau}(15).$$

Then, $\tilde{\mu} = y_1 + \tilde{\tau}(y_1) = 1 + 7.910 + \tilde{\tau}(15) = 8.910 + \tilde{\tau}(15)$. Under Efron's method, $\tilde{\tau}(15) = 0$ and $\tilde{\mu} = 8.910$; while under Klein and Moeschberger's method, $\tilde{\tau}(15) = (22 - 15)(0.176) = 1.232$ and $\tilde{\mu} = 8.910 + 1.232 = 10.142$. Finally, for the exponential tail correction, $\tilde{\tau}(15) = (15)(0.176)/(-\ln 0.176) = 1.520$, and therefore $\tilde{\mu} = 8.910 + 1.520 = 10.43$. □

To estimate the variance of $\tilde{\mu}$, we note that $\tilde{\mu}$ is a function of $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k)$, for which we have an estimate of the variance matrix from the previous section. In particular, $\mathrm{Var}(\hat{\lambda})$ is a $k \times k$ diagonal matrix (i.e. all off-diagonal elements are 0). Thus, by the multivariate delta method, with the $1 \times k$ matrix A with jth entry $\partial \mu / \partial \lambda_j$, the estimated variance of $\tilde{\mu}$ is

$$\mathrm{Var}(\tilde{\mu}) \doteq \sum_{m=1}^{k} \left(\frac{\partial \mu}{\partial \lambda_m} \right)^2 \mathrm{Var}(\hat{\lambda}_m),$$

and it remains to identify $\partial \mu / \partial \lambda_m$ for $m = 1, 2, \ldots, k$.

To begin, first note that μ depends on $\tau(y_k)$, which in turn depends on $S(y_k) = \prod_{i=1}^{k} \phi(\lambda_i)$ but also on the tail correction employed. As such, we will express the formulas in terms of $\partial \tau(y_k) / \partial \lambda_m$ for $m = 1, 2, \ldots, k$ for the moment. We first consider $\partial \mu / \partial \lambda_m$ for $m = 1, 2, \ldots, k - 1$. Then,

$$\mu = y_1 + \tau(y_1) = y_1 + \tau(y_k) + \sum_{j=2}^{k} (y_j - y_{j-1}) \prod_{i=1}^{j-1} \phi(\lambda_i).$$

In the above expression, λ_m does not appear in the first $m - 1$ terms of the summation, that is, for $j \leq m$. Thus,

$$\frac{\partial \mu}{\partial \lambda_m} = \frac{\partial \tau(y_k)}{\partial \lambda_m} + \frac{\partial}{\partial \lambda_m} \left[\sum_{j=m+1}^{k} (y_j - y_{j-1}) \prod_{i=1}^{j-1} \phi(\lambda_i) \right]$$

$$= \frac{\partial \tau(y_k)}{\partial \lambda_m} + \frac{\phi'(\lambda_m)}{\phi(\lambda_m)} \sum_{j=m+1}^{k} (y_j - y_{j-1}) \prod_{i=1}^{j-1} \phi(\lambda_i),$$

and in terms of $\tau(y_m)$, this may be expressed as

$$\frac{\partial \mu}{\partial \lambda_m} = \frac{\partial \tau(y_k)}{\partial \lambda_m} + \frac{\phi'(\lambda_m)}{\phi(\lambda_m)} \left[\tau(y_m) - \tau(y_k) \right], \quad m = 1, 2, \ldots, k - 1.$$

It is also useful to note that $\int_0^{y_k} S(y) dy$ does not involve λ_k and thus $\partial \mu / \partial \lambda_k = \partial \tau(y_k)/\partial \lambda_k$. The general variance formula thus may be written as

$$\mathrm{Var}(\tilde{\mu}) \doteq \sum_{m=1}^{k} \left\{ [\tau(y_m) - \tau(y_k)] \frac{\phi'(\lambda_m)}{\phi(\lambda_m)} + \frac{\partial \tau(y_k)}{\partial \lambda_m} \right\}^2 \mathrm{Var}(\hat{\lambda}_m).$$

But

$$\frac{\partial \tau(y_k)}{\partial \lambda_m} = (y_{\max} - y_k) \frac{\phi'(\lambda_m)}{\phi(\lambda_m)} \left[\prod_{i=1}^{k} \phi(\lambda_i) \right] + \frac{\partial \tau(y_{\max})}{\partial \lambda_m},$$

and thus,

$$\frac{\partial \tau(y_k)}{\partial \lambda_m} = \frac{\phi'(\lambda_m)}{\phi(\lambda_m)} [\tau(y_k) - \tau(y_{\max})] + \frac{\partial \tau(y_{\max})}{\partial \lambda_m},$$

in turn implying that

$$\mathrm{Var}(\tilde{\mu}) \doteq \sum_{m=1}^{k} \left\{ [\tau(y_m) - \tau(y_{\max})] \frac{\phi'(\lambda_m)}{\phi(\lambda_m)} + \frac{\partial \tau(y_{\max})}{\partial \lambda_m} \right\}^2 \mathrm{Var}(\hat{\lambda}_m).$$

The variance is estimated by replacing parameters with their estimates in the above formula. This yields

$$\widehat{\mathrm{Var}}(\tilde{\mu}) = \sum_{m=1}^{k} \left\{ [\tilde{\tau}(y_m) - \tilde{\tau}(y_{\max})] \frac{\phi'(\hat{\lambda}_m)}{\phi(\hat{\lambda}_m)} + \frac{\partial \tilde{\tau}(y_{\max})}{\partial \hat{\lambda}_m} \right\}^2 \frac{s_m(r_m - s_m)}{r_m^3},$$

where we understand $\partial \tilde{\tau}(y_{\max})/\partial \hat{\lambda}_m$ to mean $\partial \tau(y_{\max})/\partial \lambda_m$ with $\tau(y_{\max})$ replaced by $\tilde{\tau}(y_{\max})$ and λ_m by $\hat{\lambda}_m$.

If $\tilde{\tau}(y_{\max}) = 0$, then

$$\widehat{\mathrm{Var}}(\tilde{\mu}) = \sum_{m=1}^{k} [\tilde{\tau}^2(y_m)]^2 \left[\frac{\phi'(\hat{\lambda}_m)}{\phi(\hat{\lambda}_m)} \right]^2 \frac{s_m(r_m - s_m)}{r_m^3},$$

a formula that further simplifies, under the Kaplan–Meier assumption $\phi(x) = 1-x$ (recalling that $\hat{\lambda}_m = s_m/r_m$), to

$$\widehat{\mathrm{Var}}(\tilde{\mu}) = \sum_{m=1}^{k} [\tilde{\tau}(y_m)]^2 \frac{s_m}{r_m(r_m - s_m)}.$$

We note that $\tilde{\tau}(y_{\max}) = 0$ if no tail correction is necessary, because $S(y_k) = 0$ (in which case $\tilde{\tau}(y_k) = 0$ as well and the upper limit of the summation is $k - 1$), or under Efron's approximation.

For Klein and Moeschberger's method,

$$\tau(y_{\max}) = (\gamma - y_{\max}) S(y_k) = (\gamma - y_{\max}) \prod_{i=1}^{k} \phi(\lambda_i),$$

implying that

$$\frac{\partial \tau(y_{\max})}{\partial \lambda_m} = \frac{\phi'(\lambda_m)}{\phi(\lambda_m)} \tau(y_{\max}),$$

resulting in the same variance formula as under Efron's method [but $\tilde{\tau}(y_m)$ is increased by $\tilde{\tau}(y_{\max})$ for this latter approximation].

Turning now to the exponential tail correction with $\tau(y_{max}) = -y_{\max} S(y_k)/\ln S(y_k)$, recall that $S(y_k) = \prod_{i=1}^{k} \phi(\lambda_i)$ and $\ln S(y_k) = \sum_{i=1}^{k} \ln \phi(\lambda_i)$. Thus

$$\begin{aligned}
\frac{\partial \tau(y_{max})}{\partial \lambda_m} &= -\frac{y_{max}}{\ln S(y_k)}\left[\frac{\partial}{\partial \lambda_m}S(y_k)\right] + \frac{y_{max}S(y_k)}{[\ln S(y_k)]^2}\left[\frac{\partial}{\partial \lambda_m}\ln S(y_k)\right] \\
&= -\frac{y_{max}}{\ln S(y_k)}\left[\frac{\phi'(\lambda_m)}{\phi(\lambda_m)}S(y_k)\right] + \frac{y_{max}S(y_k)}{[-\ln S(y_k)]^2}\left[\frac{\phi'(\lambda_m)}{\phi(\lambda_m)}\right] \\
&= \tau(y_{max})\frac{\phi'(\lambda_m)}{\phi(\lambda_m)} + \frac{[\tau(y_{max})]^2}{y_{max}S(y_k)}\frac{\phi'(\lambda_m)}{\phi(\lambda_m)}.
\end{aligned}$$

Therefore, under the exponential tail correction, the general variance estimate becomes

$$\widehat{\mathrm{Var}}(\tilde{\mu}) = \sum_{m=1}^{k}\left\{\tilde{\tau}(y_m) + \frac{[\tilde{\tau}(y_{\max})]^2}{y_{\max}\tilde{S}(y_k)}\right\}^2 \left[\frac{\phi'(\hat{\lambda}_m)}{\phi(\hat{\lambda}_m)}\right]^2 \frac{s_m(r_m - s_m)}{r_m^3}.$$

In the Nelson–Åalen case with $\phi(x) = e^{-x}$, the term $[\phi'(\hat{\lambda}_m)/\phi(\hat{\lambda}_m)]^2$ may obviously be omitted.

■ **EXAMPLE 14.13**

Estimate the variance of the means calculated in Example 14.12.

For both Efron's and Klein and Moeschberger's approaches, the formula becomes

$$\begin{aligned}
\widehat{\mathrm{Var}}(\tilde{\mu}) &= \sum_{m=1}^{7}[\tilde{\tau}(y_m)]^2\frac{s_m(r_m - s_m)}{r_m^3} \\
&= [7.910 + \tilde{\tau}(15)]^2\frac{(1)(19)}{(20)^3} + [6.959 + \tilde{\tau}(15)]^2\frac{(1)(18)}{(19)^3} \\
&\quad + [5.155 + \tilde{\tau}(15)]^2\frac{(2)(15)}{(17)^3} + [4.352 + \tilde{\tau}(15)]^2\frac{(1)(12)}{(13)^3} \\
&\quad + [2.123 + \tilde{\tau}(15)]^2\frac{(3)(8)}{(11)^3} + [1.557 + \tilde{\tau}(15)]^2\frac{(4)(4)}{(8)^3} \\
&\quad + [0.528 + \tilde{\tau}(15)]^2\frac{(2)(1)}{(3)^3}.
\end{aligned}$$

Under Efron's approach, $\tilde{\tau}(15) = 0$, yielding the variance estimate of 0.719, whereas under Klein and Moeschberger's method, $\tilde{\tau}(15) = 1.232$, yielding the variance estimate of 1.469. For the exponential tail correction, we have $\tilde{\tau}(15) = 1.520$ and

$$\frac{[\tilde{\tau}(y_{\max})]^2}{y_{\max}\hat{S}(y_k)} = \frac{[\tilde{\tau}(15)]^2}{15\hat{S}(12)} = \frac{(1.520)^2}{15(0.176)} = 0.875,$$

and the variance estimate is thus

$$\widehat{\text{Var}}(\tilde{\mu}) = \sum_{m=1}^{7} [\tilde{\tau}(y_m) + 0.875]^2 \frac{s_m(r_m - s_m)}{r_m^3}.$$

That is,

$$\widehat{\text{Var}}(\tilde{\mu}) = (7.910 + 1.520 + 0.875)^2 \frac{(1)(19)}{(20)^3} + (6.959 + 1.520 + 0.875)^2 \frac{(1)(18)}{(19)^3}$$

$$+ (5.155 + 1.520 + 0.875)^2 \frac{(2)(15)}{(17)^3} + (4.352 + 1.520 + 0.875)^2 \frac{(1)(12)}{(13)^3}$$

$$+ (2.123 + 1.520 + 0.875)^2 \frac{(3)(8)}{(11)^3} + (1.557 + 1.520 + 0.875)^2 \frac{(4)(4)}{(8)^3}$$

$$+ (0.528 + 1.520 + 0.875)^2 \frac{(2)(1)}{(3)^3}$$

$$= 2.568. \qquad\qquad \square$$

For higher moments, a similar approach may be used. We have, for the αth moment,

$$E(Y^\alpha) = \alpha \int_0^\infty y^{\alpha-1} S(y)\, dy,$$

which may be estimated (using $y_0 = 0$ without loss of generality) by

$$\tilde{\mu}_\alpha = \alpha \int_0^\infty y^{\alpha-1} \tilde{S}(y)\, dy$$

$$= \alpha \left[\sum_{j=1}^{k} \int_{y_{j-1}}^{y_j} y^{\alpha-1} \tilde{S}(y)\, dy \right] + \alpha \int_{y_k}^{y_{\max}} y^{\alpha-1} \tilde{S}(y)\, dy + \alpha \int_{y_{\max}}^{\infty} y^{\alpha-1} \tilde{S}(y)\, dy$$

$$= \alpha \left[\sum_{j=1}^{k} \int_{y_{j-1}}^{y_j} y^{\alpha-1} \tilde{S}(y_{j-1})\, dy \right] + \alpha \int_{y_k}^{y_{\max}} y^{\alpha-1} \tilde{S}(y_k)\, dy + \alpha \int_{y_{\max}}^{\infty} y^{\alpha-1} \tilde{S}(y)\, dy$$

$$= \alpha \left[\sum_{j=1}^{k} \tilde{S}(y_{j-1}) \int_{y_{j-1}}^{y_j} y^{\alpha-1}\, dy \right] + \alpha \tilde{S}(y_k) \int_{y_k}^{y_{\max}} y^{\alpha-1}\, dy + \alpha \int_{y_{\max}}^{\infty} y^{\alpha-1} \tilde{S}(y)\, dy$$

$$= \left[\sum_{j=1}^{k} \left(y_j^\alpha - y_{j-1}^\alpha \right) \tilde{S}(y_{j-1}) \right] + (y_{\max}^\alpha - y_k^\alpha)\tilde{S}(y_k) + \alpha \int_{y_{\max}}^{\infty} y^{\alpha-1} \tilde{S}(y)\, dy$$

$$= y_1^\alpha + \left[\sum_{j=2}^{k} (y_j^\alpha - y_{j-1}^\alpha) \prod_{i=1}^{j-1} \phi(\hat{\lambda}_i) \right] + \left[(y_{\max}^\alpha - y_k^\alpha) \prod_{i=1}^{k} \phi(\hat{\lambda}_i) \right] + \alpha \int_{y_{\max}}^{\infty} y^{\alpha-1} \tilde{S}(y)\, dy.$$

Again, the final integral on the right-hand side depends on the tail correction, and is 0 if $\tilde{S}(y_k) = 0$ or under Efron's tail correction. It is useful to note that under the exponential tail correction, $\tilde{S}(y) = e^{-\tilde{\beta}y}$ for $y \geq y_{\max}$ with $\tilde{\beta} = -\ln \tilde{S}(y_k)/y_{\max}$, and if $\alpha = 1, 2, \ldots$,

$$\alpha \int_{y_{\max}}^{\infty} y^{\alpha-1} e^{-\tilde{\beta}y} dy = \frac{\alpha!}{\tilde{\beta}^{\alpha}} \int_{y_{\max}}^{\infty} \frac{\tilde{\beta}^{\alpha} y^{\alpha-1} e^{-\tilde{\beta}y}}{(\alpha-1)!} dy$$

$$= \frac{\alpha!}{\tilde{\beta}^{\alpha}} \sum_{j=0}^{\alpha-1} \frac{(\tilde{\beta}y_{\max})^{j} e^{-\tilde{\beta}y_{\max}}}{j!},$$

using the tail function representation of the gamma distribution. That is, under the exponential tail correction,

$$\alpha \int_{y_{\max}}^{\infty} y^{\alpha-1} \tilde{S}(y) \, dy = \alpha(y_{\max})^{\alpha} \tilde{S}(y_k) \sum_{j=0}^{\alpha-1} \frac{1}{j!} \left[-\ln \tilde{S}(y_k)\right]^{j-\alpha}, \quad \alpha = 1, 2, 3, \ldots.$$

In particular, for the second moment ($\alpha = 2$),

$$2 \int_{y_{\max}}^{\infty} y\tilde{S}(y) dy = 2(y_{\max})^2 \tilde{S}(y_k) \left\{ [-\ln \tilde{S}(y_k)]^{-1} + [-\ln \tilde{S}(y_k)]^{-2} \right\}.$$

Variance estimation for $\tilde{\mu}_{\alpha}$ may be done in a similar manner as for the mean, if desired.

14.4.1 Exercises

14.27 For the data of Exercise 14.24 and using the Kaplan–Meier estimate:

(a) Compute the mean survival time estimate assuming Efron's tail correction.

(b) Compute the mean survival time estimate using the exponential tail correction of Brown, Hollander, and Korwar.

(c) Estimate the variance of the estimate in (a).

14.28 For the data of Exercise 14.25, using the Nelson–Ãalen estimate and the exponential tail correction of Brown, Hollander, and Korwar:

(a) Estimate the mean $\tilde{\mu}$.

(b) Estimate the variance of $\tilde{\mu}$ in (b).

14.29 For the data in Example 14.5 and subsequent examples, using the Nelson–Ãalen estimate with the exponential tail correction of Brown, Hollander, and Korwar, estimate the variance of Y.

14.5   Empirical Estimation with Left Truncated Data

The results of Section 14.3 apply in situations in which the data are (right) censored. In this section, we discuss the situation in which the data may also be (left) truncated. We have the following definitions.

Definition 14.10 *An observation is **truncated from below** (also called **left truncated**) at d if when it is at or below d it is not recorded, but when it is above d it is recorded at its observed value.*

*An observation is **truncated from above** (also called **right truncated**) at u if when it is at or above u it is not recorded, but when it is below u it is recorded at its observed value.*

In insurance survival data and claim data, the most common occurrences are left truncation and right censoring. Left truncation occurs when an ordinary deductible of d is applied. When a policyholder has a loss below d, he or she realizes no benefits will be paid and so does not inform the insurer. When the loss is above d, the amount of the loss is assumed to be reported.[5]   A policy limit leads to an example of right censoring. When the amount of the loss equals or exceeds u, benefits beyond that value are not paid, and so the exact value is not recorded. However, it is known that a loss of at least u has occurred.

For decrement studies, such as of human mortality, it is impractical to follow people from birth to death. It is more common to follow a group of people of varying ages for a few years during the study period. When a person joins a study, he or she is alive at that time. This person's age at death must be at least as great as the age at entry to the study and thus has been left truncated. If the person is alive when the study ends, right censoring has occurred. The person's age at death is not known, but it is known that it is at least as large as the age when the study ended. Right censoring also affects those who exit the study prior to its end due to surrender. Note that this discussion could have been about other decrements, such as disability, policy surrender, or retirement.

Because left truncation and right censoring are the most common occurrences in actuarial work, they are the only cases that are covered in this section. To save words, **truncated** always means truncated from below and **censored** always means censored from above.

When trying to construct an empirical distribution from truncated or censored data, the first task is to create notation to represent the data. For individual (as opposed to grouped) data, the following facts are needed. The first is the truncation point for that observation. Let that value be d_j for the jth observation. If there was no truncation, $d_j = 0$.[6]   Next, record the observation itself. The notation used depends on whether or not that observation was censored. If it was not censored, let its value be x_j. If it was censored, let its value be u_j. When this subject is presented more formally, a distinction is made between the case where the censoring point is known in advance and where it is not. For example, a liability insurance policy with a policy limit usually has a censoring point that is known prior to

---

[5]In some cases, an insured may elect to not report a loss that is slightly above the deductible if it is likely that the next premium will be increased.
[6]Throughout, we assume that negative values are not possible.

the receipt of any claims. By comparison, in a mortality study of insured lives, those that surrender their policy do so at an age that was not known when the policy was sold. In this chapter, no distinction is made between the two cases.

To construct the estimate, the raw data must be summarized in a useful manner. The most interesting values are the uncensored observations. As in Section 14.3, let $y_1 < y_2 < \cdots < y_k$ be the k unique values of the x_js that appear in the sample, where k must be less than or equal to the number of uncensored observations. We also continue to let s_j be the number of times the uncensored observation y_j appears in the sample. Again, an important quantity is r_j, the number "at risk" at y_j. In a decrement study, r_j represents the number under observation and subject to the decrement at that time. To be under observation at y_j, an individual must (1) either be censored or have an observation that is on or after y_j and (2) not have a truncation value that is on or after y_j. That is,

$$r_j = (\text{number of } x_i\text{s} \geq y_j) + (\text{number of } u_i\text{s} \geq y_j) - (\text{number of } d_i\text{s} \geq y_j).$$

Alternatively, because the total number of d_is is equal to the total number of x_is and u_is, we also have

$$r_j = (\text{number of } d_i\text{s} < y_j) - (\text{number of } x_i\text{s} < y_j) - (\text{number of } u_i\text{s} < y_j). \quad (14.4)$$

This latter version is a bit easier to conceptualize because it includes all who have entered the study prior to the given age less those who have already left. The key point is that the number at risk is the number of people observed alive at age y_j. If the data are loss amounts, the risk set is the number of policies with observed loss amounts (either the actual amount or the maximum amount due to a policy limit) greater than or equal to y_j less those with deductibles greater than or equal to y_j. These relationships lead to a recursive version of the formula,

$$\begin{aligned} r_j = r_{j-1} &+ (\text{number of } d_i\text{s between } y_{j-1} \text{ and } y_j) \\ &- (\text{number of } x_i\text{s equal to } y_{j-1}) \\ &- (\text{number of } u_i\text{s between } y_{j-1} \text{ and } y_j), \quad (14.5) \end{aligned}$$

where *between* is interpreted to mean greater than or equal to y_{j-1} and less than y_j, and r_0 is set equal to zero.

A consequence of the above definitions is that if a censoring or truncation time equals that of a death, the death is assumed to have happened first. That is, the censored observation is considered to be at risk while the truncated observation is not.

The definition of r_j presented here is consistent with that in Section 14.3. That is, if $d_j = 0$ for all observations, the formulas presented here reduce match those presented earlier. The following example illustrates calculating the number at risk when there is truncation.

■ **EXAMPLE 14.14**

Using Data Set D, introduced in Chapter 10 and reproduced here as Table 14.16, calculate the r_j values using both (14.4) and (14.5). To provide some context and explain the entries in the table, think of this as a study of mortality by duration for

Table 14.16   The values for Example 14.14.

i	d_i	x_i	u_i	i	d_i	x_i	u_i
1	0	—	0.1	16	0	4.8	—
2	0	—	0.5	17	0	—	4.8
3	0	—	0.8	18	0	—	4.8
4	0	0.8	—	19–30	0	—	5.0
5	0	—	1.8	31	0.3	—	5.0
6	0	—	1.8	32	0.7	—	5.0
7	0	—	2.1	33	1.0	4.1	—
8	0	—	2.5	34	1.8	3.1	—
9	0	—	2.8	35	2.1	—	3.9
10	0	2.9	—	36	2.9	—	5.0
11	0	2.9	—	37	2.9	—	4.8
12	0	—	3.9	38	3.2	4.0	—
13	0	4.0	—	39	3.4	—	5.0
14	0	—	4.0	40	3.9	—	5.0
15	0	—	4.1				

a five-year term insurance policy. The study period is a fixed time period. Thus, policies 31–40 were already insured when first observed. There are two ways in which censoring occurs. For some (1, 2, 3, 5, etc.), the study either ended while they were still alive or they terminated their policies prior to the end of the five-year term. For others (19–32, 36, etc.), the five-year term ended with them still alive. Note that the eight observed deaths are at six distinct times and that $y_6 = 4.8$. Because the highest censoring time is 5.0, $y_{\max} = 5.0$.

The calculations appear in Table 14.17.   □

The approach to developing an empirical estimator of the survival function is to use the formulas developed in Section 14.3, but with this more general definition of r_j. A theoretical treatment that incorporates left truncation is considerably more complex (for details, see Lawless [77]).

Table 14.17   The risk set calculations for Example 14.14.

j	y_j	s_j	r_j
1	0.8	1	$32 - 0 - 2 = 30$ or $0 + 32 - 0 - 2 = 30$
2	2.9	2	$35 - 1 - 8 = 26$ or $30 + 3 - 1 - 6 = 26$
3	3.1	1	$37 - 3 - 8 = 26$ or $26 + 2 - 2 - 0 = 26$
4	4.0	2	$40 - 4 - 10 = 26$ or $26 + 3 - 1 - 2 = 26$
5	4.1	1	$40 - 6 - 11 = 23$ or $26 + 0 - 2 - 1 = 23$
6	4.8	1	$40 - 7 - 12 = 21$ or $23 + 0 - 1 - 1 = 21$

The formula for the Kaplan–Meier estimate is the same as presented earlier, namely

$$
S_n(y) = \begin{cases}
1, & y < y_1, \\
\prod_{i=1}^{j}\left(1 - \hat{\lambda}_i\right) = \prod_{i=1}^{j}\left(1 - \frac{s_i}{r_i}\right), & y_j \le y < y_{j+1}, \quad j = 1, 2, \ldots, k-1, \\
\prod_{i=1}^{k}\left(1 - \hat{\lambda}_i\right) = \prod_{i=1}^{k}\left(1 - \frac{s_i}{r_i}\right), & y_k \le y < y_{\max}.
\end{cases}
$$

The same tail corrections developed in Section 14.3 can be used for $y \ge y_{\max}$ in cases where $S_n(y_k) > 0$.

■ **EXAMPLE 14.15**

Determine the Kaplan–Meier estimate for the data in Example 14.14.

Based on the previous example, we have

$$
S_{40}(y) = \begin{cases}
1, & 0 \le y < 0.8, \\
\frac{30-1}{30} = 0.9667, & 0.8 \le y < 2.9, \\
0.9667\frac{26-2}{26} = 0.8923, & 2.9 \le y < 3.1, \\
0.8923\frac{26-1}{26} = 0.8580, & 3.1 \le y < 4.0, \\
0.8580\frac{26-2}{26} = 0.7920, & 4.0 \le y < 4.1, \\
0.7920\frac{23-1}{23} = 0.7576, & 4.1 \le y < 4.8, \\
0.7576\frac{21-1}{21} = 0.7215, & 4.8 \le y < 5.0.
\end{cases}
$$

□

In this example, a tail correction is not needed because an estimate of survival beyond the five-year term is of no value when analyzing these policyholders.

The same analogy holds for the Nelson–Åalen estimator, where the formula for the cumulative hazard function remains

$$
\hat{H}(y) = \begin{cases}
0, & y < y_1, \\
\sum_{i=1}^{j}\frac{s_i}{r_i}, & y_j \le y < y_{j+1}, \quad j = 1, 2, \ldots, k-1, \\
\sum_{i=1}^{k}\frac{s_i}{r_i}, & y_k \le y < y_{\max}.
\end{cases}
$$

As before, $\hat{S}(y) = \exp[-\hat{H}(y)]$ for $y < y_{\max}$ and for $y \ge y_{\max}$ the same tail corrections can be used.

■ **EXAMPLE 14.16**

Determine the Nelson–Åalen estimate of the survival function for the data in Example 14.14.

The estimated functions are

$$\hat{H}(y) = \begin{cases} 0, & 0 \le y < 0.8, \\ \frac{1}{30} = 0.0333, & 0.8 \le y < 2.9, \\ 0.0333 + \frac{2}{26} = 0.1103, & 2.9 \le y < 3.1, \\ 0.1103 + \frac{1}{26} = 0.1487, & 3.1 \le y < 4.0, \\ 0.1487 + \frac{2}{26} = 0.2256, & 4.0 \le y < 4.1, \\ 0.2256 + \frac{1}{23} = 0.2691, & 4.1 \le y < 4.8, \\ 0.2691 + \frac{1}{21} = 0.3167, & 4.8 \le y < 5.0. \end{cases}$$

$$\hat{S}(y) = \begin{cases} 1, & 0 \le y < 0.8, \\ e^{-0.0333} = 0.9672, & 0.8 \le y < 2.9, \\ e^{-0.1103} = 0.8956, & 2.9 \le y < 3.1, \\ e^{-0.1487} = 0.8618, & 3.1 \le y < 4.0, \\ e^{-0.2256} = 0.7980, & 4.0 \le y < 4.1, \\ e^{-0.2691} = 0.7641, & 4.1 \le y < 4.8, \\ e^{-0.3167} = 0.7285, & 4.8 \le y < 5.0. \end{cases}$$

□

In this section, the results were not formally developed, as was done for the case with only right censored data. However, all the results, including formulas for moment estimates and estimates of the variance of the estimators, hold when left truncation is added. However, it is important to note that when the data are truncated, the resulting distribution function is the distribution function of observations given that they are above the smallest truncation point (i.e. the smallest d value). Empirically, there is no information about observations below that value, and thus there can be no information for that range. Finally, if it turns out that there was no censoring or truncation, use of the formulas in this section will lead to the same results as when using the empirical formulas in Section 14.1.

14.5.1   Exercises

14.30 Repeat Example 14.14, treating "surrender" as "death." The easiest way to do this is to reverse the x and u labels. In this case, death produces censoring because those who die are lost to observation and thus their surrender time is never observed. Treat those who lasted the entire five years as surrenders at that time.

14.31 Determine the Kaplan–Meier estimate for the time to surrender for Data Set D. Treat those who lasted the entire five years as surrenders at that time.

14.32 Determine the Nelson–Åalen estimate of $H(t)$ and $S(t)$ for Data Set D, where the variable is time to surrender.

14.33 Determine the Kaplan–Meier and Nelson–Åalen estimates of the distribution function of the amount of a workers compensation loss. First use the raw data from Data Set B. Then repeat the exercise, modifying the data by left truncation at 100 and right censoring at 1,000.

14.34 (*) Three hundred mice were observed at birth. An additional 20 mice were first observed at age 2 (days) and 30 more were first observed at age 4. There were 6 deaths at age 1, 10 at age 3, 10 at age 4, a at age 5, b at age 9, and 6 at age 12. In addition, 45 mice were lost to observation at age 7, 35 at age 10, and 15 at age 13. The following product-limit estimates were obtained: $S_{350}(7) = 0.892$ and $S_{350}(13) = 0.856$. Determine the values of a and b.

14.35 Construct 95% confidence intervals for $H(3)$ by both the linear and log-transformed formulas using all 40 observations in Data Set D, with surrender being the variable of interest.

14.36 (*) For the interval from zero to one year, the exposure (r) is 15 and the number of deaths (s) is 3. For the interval from one to two years, the exposure is 80 and the number of deaths is 24. For two to three years, the values are 25 and 5; for three to four years, they are 60 and 6; and for four to five years, they are 10 and 3. Determine Greenwood's approximation to the variance of $\hat{S}(4)$.

14.37 (*) You are given the values in Table 14.18. Determine the standard deviation of the Nelson–Åalen estimator of the cumulative hazard function at time 20.

14.6   Kernel Density Models

One problem with empirical distributions is that they are always discrete. If it is known that the true distribution is continuous, the empirical distribution may be viewed as a poor approximation. In this section, a method of obtaining a smooth, empirical-like distribution, called a kernel density distribution, is introduced. We have the following definition.

Definition 14.11 *A **kernel smoothed distribution** is obtained by replacing each data point with a continuous random variable and then assigning probability $1/n$ to each such random variable. The random variables used must be identical except for a location or scale change that is related to its associated data point.*

Table 14.18   The data for Exercise 14.37.

y_j	r_j	s_j
1	100	15
8	65	20
17	40	13
25	31	31

Note that the empirical distribution is a special type of kernel smoothed distribution in which the random variable assigns probability 1 to the data point. With regard to kernel smoothing, there are several distributions that could be used, three of which are introduced here.

While not necessary, it is customary that the continuous variable have a mean equal to the value of the point it replaces, ensuring that, overall, the kernel estimate has the same mean as the empirical estimate. One way to think about such a model is that it produces the final observed value in two steps. The first step is to draw a value at random from the empirical distribution. The second step is to draw a value at random from a continuous distribution whose mean is equal to the value drawn at the first step. The selected continuous distribution is called the *kernel*.

For notation, let $p(y_j)$ be the probability assigned to the value y_j $(j = 1, \ldots, k)$ by the empirical distribution. Let $K_y(x)$ be a distribution function for a continuous distribution such that its mean is y. Let $k_y(x)$ be the corresponding density function.

Definition 14.12 *A **kernel density estimator** of a distribution function is*

$$\hat{F}(x) = \sum_{j=1}^{k} p(y_j) K_{y_j}(x),$$

and the estimator of the density function is

$$\hat{f}(x) = \sum_{j=1}^{k} p(y_j) k_{y_j}(x).$$

The function $k_y(x)$ is called the kernel. Three kernels are now introduced: uniform, triangular, and gamma.

Definition 14.13 *The **uniform kernel** is given by*

$$k_y(x) = \begin{cases} 0, & x < y - b, \\ \dfrac{1}{2b}, & y - b \le x \le y + b, \\ 0, & x > y + b, \end{cases}$$

$$K_y(x) = \begin{cases} 0, & x < y - b, \\ \dfrac{x - y + b}{2b}, & y - b \le x \le y + b, \\ 1, & x > y + b. \end{cases}$$

The **triangular kernel** is given by

$$
k_y(x) = \begin{cases} 0, & x < y - b, \\[2mm] \dfrac{x - y + b}{b^2}, & y - b \le x \le y, \\[3mm] \dfrac{y + b - x}{b^2}, & y \le x \le y + b, \\[3mm] 0, & x > y + b, \end{cases}
$$

$$
K_y(x) = \begin{cases} 0, & x < y - b, \\[2mm] \dfrac{(x - y + b)^2}{2b^2}, & y - b \le x \le y, \\[3mm] 1 - \dfrac{(y + b - x)^2}{2b^2}, & y \le x \le y + b, \\[3mm] 1, & x > y + b. \end{cases}
$$

The **gamma** kernel is given by letting the kernel have a gamma distribution with shape parameter α and scale parameter y/α. That is,

$$
k_y(x) = \frac{x^{\alpha-1} e^{-x\alpha/y}}{(y/\alpha)^\alpha \Gamma(\alpha)} \text{ and } K_y(x) = \Gamma(\alpha; \alpha x/y).
$$

Note that the gamma distribution has a mean of $\alpha(y/\alpha) = y$ and a variance of $\alpha(y/\alpha)^2 = y^2/\alpha$.

In each case, there is a parameter that relates to the spread of the kernel. In the first two cases, it is the value of $b > 0$, which is called the **bandwidth**. In the gamma case, the value of α controls the spread, with a larger value indicating a smaller spread. There are other kernels that cover the range from zero to infinity.

■ **EXAMPLE 14.17**

Determine the kernel density estimate for Example 14.2 using each of the three kernels.

The empirical distribution places probability $\frac{1}{8}$ at 1.0, $\frac{1}{8}$ at 1.3, $\frac{2}{8}$ at 1.5, $\frac{3}{8}$ at 2.1, and $\frac{1}{8}$ at 2.8. For a uniform kernel with a bandwidth of 0.1 we do not get much separation. The data point at 1.0 is replaced by a horizontal density function running from 0.9 to 1.1 with a height of $\frac{1}{8}\frac{1}{2(0.1)} = 0.625$. In comparison, with a bandwidth of 1.0, that same data point is replaced by a horizontal density function running from 0.0 to 2.0 with a height of $\frac{1}{8}\frac{1}{2(1)} = 0.0625$. Figures 14.3 and 14.4 provide plots of the density functions.

It should be clear that the larger bandwidth provides more smoothing. In the limit, as the bandwidth approaches zero, the kernel density estimate matches the empirical estimate. Note that, if the bandwidth is too large, probability will be assigned to negative values, which may be an undesirable result. Methods exist for dealing with that issue, but they are not presented here.
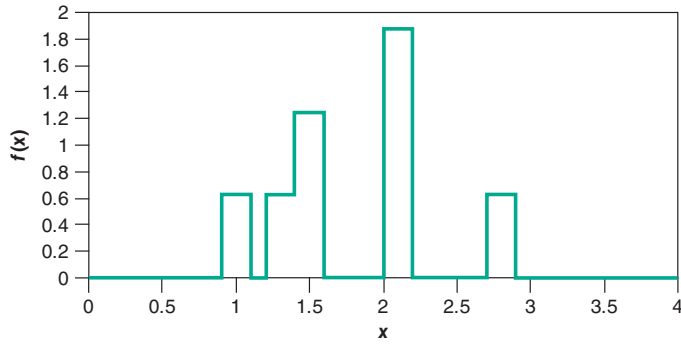
Figure 14.3   The uniform kernel density with bandwidth 0.1.
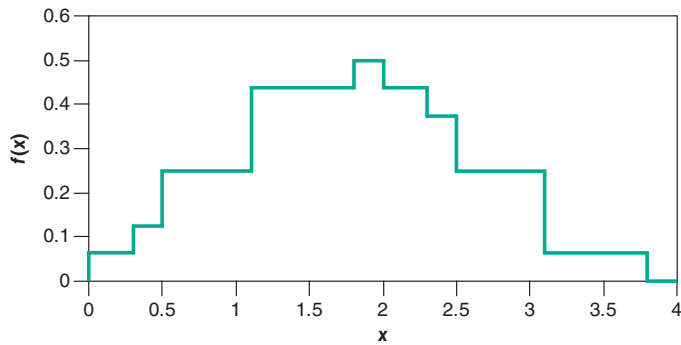
Figure 14.4   The uniform kernel density with bandwidth 1.0.

For the triangular kernel, each point is replaced by a triangle. Graphs for the same two bandwidths as used previously appear in Figures 14.5 and 14.6.

Once again, the larger bandwidth provides more smoothing. The gamma kernel simply provides a mixture of gamma distributions, where each data point provides the
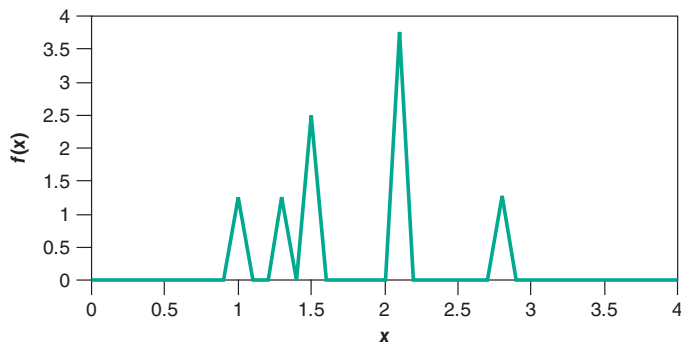
Figure 14.5   The triangular kernel density with bandwidth 0.1.
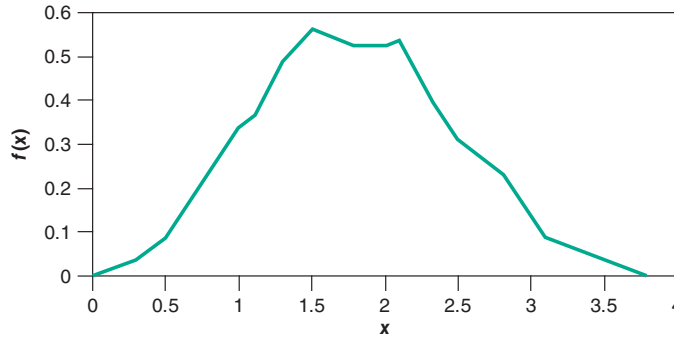
Figure 14.6   The triangular kernel density with bandwidth 1.0.
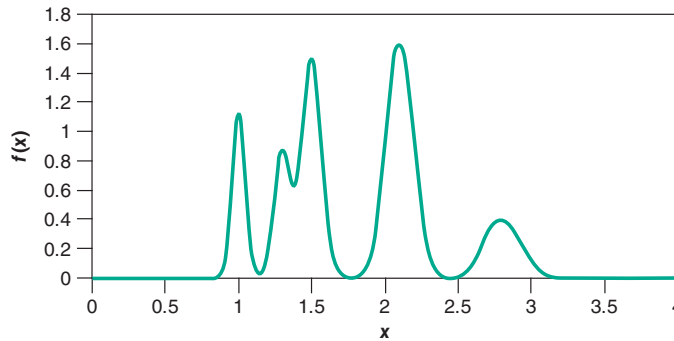
Figure 14.7   The gamma kernel density with $\alpha = 500$.

mean and the empirical probabilities provide the weights. The density function is

$$f_\alpha(x) = \sum_{j=1}^{5} p(y_j) \frac{x^{\alpha-1} e^{-x\alpha/y_j}}{(y_j/\alpha)^\alpha \Gamma(\alpha)}$$

and is graphed in Figures 14.7 and 14.8 for two α values.[7] For this kernel, decreasing the value of α increases the amount of smoothing. Further discussion of the gamma kernel can be found in Carriere [23], where the author recommends $\alpha = \sqrt{n}/(\hat{\mu}_4'/\hat{\mu}_2'-1)^{1/2}$. □

14.6.1 Exercises

14.38 Provide the formula for the Pareto kernel.

14.39 Construct a kernel density estimate for the time to surrender for Data Set D. Be aware of the fact that this is a mixed distribution (probability is continuous from 0 to 5 but is discrete at 5).

---

[7]When computing values of the density function, overflow and underflow problems can be reduced by computing the logarithm of the elements of the ratio, that is, $(\alpha-1)\ln x - x\alpha/y_j - \alpha \ln(y_j/\alpha) - \ln \Gamma(\alpha)$, and then exponentiating the result.
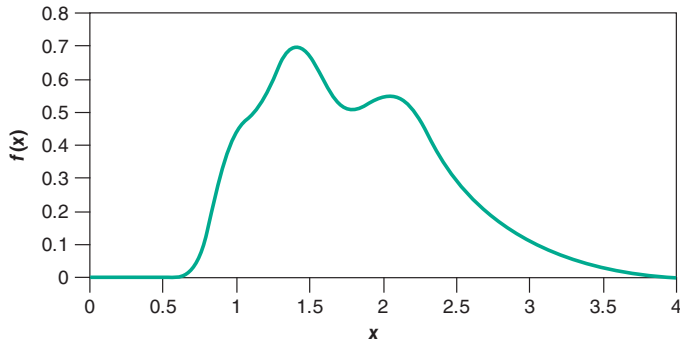
Figure 14.8   The gamma kernel density with $\alpha = 50$.

Table 14.19   The data for Exercise 14.40.

t_j	s_j	r_j
10	1	20
34	1	19
47	1	18
75	1	17
156	1	16
171	1	15

14.40 (*) You are given the data in Table 14.19 on time to death. Using the uniform kernel with a bandwidth of 60, determine $\hat{f}(100)$.

14.41 (*) You are given the following ages at time of death for 10 individuals: 25, 30, 35, 35, 37, 39, 45, 47, 49, and 55. Using a uniform kernel with a bandwidth of $b = 10$, determine the kernel density estimate of the probability of survival to age 40.

14.42 (*) Given the five observations 82, 126, 161, 294, and 384, determine each of the following estimates of $F(150)$:

   (a) The empirical estimate.

   (b) The kernel density estimate based on a uniform kernel with bandwidth $b = 50$.

   (c) The kernel density estimate based on a triangular kernel with bandwidth $b = 50$.

14.7   Approximations for Large Data Sets

14.7.1   Introduction

The discussion in this section is motivated by the circumstances that accompany the determination of a model for the time to death (or other decrement) for use in pricing, reserving, or funding insurance programs. The particular circumstances are as follows:

- Values of the survival function are required only at discrete values, normally integral ages measured in years.

- A large volume of data has been collected over a fixed time period, with most observations truncated, censored, or both.

- No parametric distribution is available that provides an adequate model given the volume of available data.

These circumstances typically apply when an insurance company (or a group of insurance companies) conducts a mortality study based on the historical experience of a very large portfolio of life insurance policies. (For the remainder of this section, we shall refer only to mortality. The results apply equally to the study of other decrements such as disablement or surrender.)

The typical mortality table is essentially a distribution function or a survival function with values presented only at integral ages. While there are parametric models that do well over parts of the age range (such as the Makeham model at ages over about 30), there are too many changes in the pattern from age 0 to ages well over 100 to allow for a simple functional description of the survival function.

The typical mortality study is conducted over a short period of time, such as three to five years. For example, all persons who are covered by an insurance company's policies at some time from January 1, 2014 through December 31, 2016 might be included. Some of these persons may have purchased their policies prior to 2014 and were still covered when the study period started. During the study period some persons will die, some will cancel (surrender) their policy, some will have their policy expire due to policy provisions (such as with term insurance policies that expire during the study period), and some will still be insured when the study ends. It is assumed that if a policy is cancelled or expires, the eventual age at death will not be known to the insurance company. Some persons will purchase their life insurance policy during the study period and be covered for some of the remaining part of the study period. These policies will be subject to the same decrements (death, surrender, expiration) as other policies. With regard to the age at death, almost every policy in the study will be left truncated.[8] If the policy was issued prior to 2014, the truncation point will be the age on January 1, 2014. For those who buy insurance during the study period, the truncation point is the age at which the contract begins. For any person who exits the study due to a cause other than death, their observation is right censored at the age of exit, because all that is known about them is that death will be at some unknown later age.

When no simple parametric distribution is appropriate and when large amounts of data are available, it is reasonable to use a nonparametric model because the large amount of data will ensure that key features of the survival function will be captured. Because there are both left truncation (due to the age at entry into the study) and right censoring (due to termination of the study at a fixed time), when there are large amounts of data, constructing the Kaplan–Meier estimate may require a very large amount of sorting and counting. Over the years, a variety of methods have been introduced and entire texts have been written about the problem of constructing mortality tables from this kind of data (e.g. [12, 81]). While the context for the examples presented here is the construction of mortality tables, the methods can apply any time the circumstances described previously apply.

---

[8]The only exception would be a policy issued during the study period to someone just born.

We begin by examining the two ways in which data are usually collected. Estimators will be presented for both situations. The formulas will be presented in this section and their derivation and properties will be provided in Section 14.8. In all cases, a set of values (ages), $c_0 < c_1 < \cdots < c_k$ has been established in advance and the goal is to estimate the survival function at these values and no others (with some sort of interpolation to be used to provide intermediate values as needed). All of the methods are designed to estimate the conditional one-period probability of death, $q_j = [S(c_j) - S(c_{j+1})]/S(c_j)$, where j may refer to the interval and not to a particular age. From those values, $S(c_j)/S(c_0)$ can be evaluated as follows:

$$\frac{S(c_j)}{S(c_0)} = \prod_{i=0}^{j-1}(1 - q_i).$$

14.7.2   Using Individual Data Points

In this setting, data are recorded for each person observed. This approach is sometimes referred to as a **seriatim** method, because the data points are analyzed as a series of individual observations. The estimator takes the form $\hat{q}_j = d_j/e_j$, where d_j is the number of observed deaths in the interval and e_j is a measure of exposure, representing the number of individuals who had a chance to be an observed death in that interval. Should a death occur at one of the boundary values between successive intervals, the death is counted in the preceding interval. When there are no entrants after age c_j into the interval and no exitants except for death during the interval (referred to as complete data), e_j represents the number of persons alive at age c_j and the number of deaths has a binomial distribution. With incomplete data, it is necessary to determine a suitable convenient approximation, preferably one that requires only a single pass through the data set. To illustrate this challenge, consider the following example.

■   **EXAMPLE 14.18**

A mortality study is based on observations during the period January 1, 2014 through December 31, 2016. Five policies were observed, with the information in the following list recorded. For simplicity, a date of 3-2000 is interpreted as March 1, 2000 and all events are treated as occurring on the first day of the month of occurrence. Furthermore, all months are treated as being one-twelfth of a year in length. Summarize the information in a manner that is sufficient for estimating mortality probabilities.

1. Born 4-1981, purchased insurance policy on 8-2013, was an active policyholder on 1-2017.

2. Born 6-1981, purchased insurance policy on 7-2013, died 9-2015.

3. Born 8-1981, purchased insurance policy on 2-2015, surrendered policy on 2-2016.

4. Born 5-1981, purchased insurance policy on 6-2014, died 3-2015.

5. Born 7-1981, purchased insurance policy on 3-2014, surrendered policy on 5-2016.

The key information is the age of the individual when first observed, the age when last observed, and the reason why observation ended. For the five policies, using "x" for death and "s" for any other reason, the values are (where an age of 33-6 means 33 years and 6 months) (32-9, 35-9, s), (32-7, 34-3, x), (33-6, 34-6, s), (33-1, 33-10, x),

and (32-8, 34-10, s). Note that policies 1 and 2 were purchased prior to the start of the study, so they are first observed when the study begins. No distinction needs to be made between those whose observation ends by surrender as opposed to the ending of the study. □

The next step is to tally information for each age interval, building up totals for d_j and e_j. Counting deaths is straightforward. For exposures, there are two approaches that are commonly used.

Exact exposure method

Following this method, we set the exposure equal to the exact total time under observation within the age interval. When a death occurs, that person's exposure ends at the exact age of death. It will be shown in Section 14.8 that d_j/e_j is the maximum likelihood estimator of the hazard rate, under the assumption that the hazard rate is constant over the interval $(c_j, c_{j+1}]$. Further properties of this estimator will also be discussed in that section. The estimated hazard rate can then be converted into a conditional probability of death using the formula $q_j = 1 - \exp(-d_j/e_j)$.

Actuarial exposure method

Under this method, the exposure period for deaths extends to the end of the age interval, rather than the exact age at death. This has the advantage of reproducing the empirical estimator for complete data, but has been shown to be an inconsistent estimator in other cases. In this case, the estimate of the conditional probability of death is obtained as $q_j = d_j/e_j$.

When the conditional probability of death is small, with a large number of observations, the choice of method is unlikely to materially affect the results.

■ EXAMPLE 14.19

Estimate all possible mortality rates at integral ages for Example 14.18 using both methods.

First observe that data are available for ages 32, 33, 34, and 35. The deaths are in the intervals 33–34 and 34–35. With a seriatim approach, each policy is analyzed and its contribution to the exposure for each interval added to the running total. For each age interval, the contributions for each interval are totaled, using the exact exposure method of recording time under observation. The numbers represent months.

32–33: $e_{32} = 3 + 5 + 0 + 0 + 4 = 12$.
33–34: $e_{33} = 12 + 12 + 6 + 9 + 12 = 51$.
34–35: $e_{34} = 12 + 3 + 6 + 0 + 10 = 31$.
35–36: $e_{35} = 9 + 0 + 0 + 0 + 0 = 9$.

As a check, the total contributions for the five policies are 36, 20, 12, 9, and 26, respectively, which matches the times on observation. The only two nonzero mortality probabilities are estimated as $\hat{q}_{33} = 1 - \exp[-1/(51/12)] = 0.20966$ and $\hat{q}_{34} = 1 - \exp[-1/(31/12)] = 0.32097$.

Under the actuarial exposure method, the exposure for the interval 33–34 for the fourth person increases from 9 to 11 months for a total of 53 (note that it is not a full year of exposure as observation did not begin until age 33-1). For the interval 34–35, the exposure for the second person increases from 3 to 12 months for a total of 40. The mortality probability estimates are $\hat{q}_{33} = 1/(53/12) = 0.2264$ and $\hat{q}_{34} = 1/(40/12) = 0.3$. ☐

■ EXAMPLE 14.20

Use the actuarial exposure method to estimate all mortality probabilities for Data Set D.

Noting that the deaths at time 4.0 are assigned to the interval 3–4, the estimated values are $\hat{q}_0 = 1/29.4 = 0.0340$, $\hat{q}_1 = 0/28.8 = 0$, $\hat{q}_2 = 2/27.5 = 0.0.0727$, $\hat{q}_3 = 3/27.3 = 0.1099$, and $\hat{q}_4 = 2/29.4 = 0.0930$. The corresponding survival probabilities are (with the Kaplan–Meier estimates in parentheses) $\hat{S}(1) = 0.9660$ (0.9667), $\hat{S}(2) = 0.9660$ (0.9667), $\hat{S}(3) = 0.8957$ (0.8923), $\hat{S}(4) = 0.7973$ (0.7920), and $\hat{S}(5) = 0.7231$ (0.7215). ☐

14.7.2.1 *Insuring Ages*   While the examples have been in a life insurance context, the methodology applies to any situation with left truncation and right censoring. However, there is a situation that is specific to life insurance studies. Consider a one-year term insurance policy. Suppose that an applicant was born on February 15, 1981 and applies for this insurance on October 15, 2016. Premiums are charged by whole-number ages. Some companies will use the age at the last birthday (35 in this case) and some will use the age at the nearest birthday (36 in this case). One company will base the premium on q_{35} and one on q_{36} when both should be using $q_{35.67}$, the applicant's true age. Suppose that a company uses age last birthday. When estimating q_{35}, it is not interested in the probability that a person exactly age 35 dies in the next year (the usual interpretation) but, rather, the probability that a random person who is assigned age 35 at issue (who can be anywhere between 35 and 36 years old) dies in the next year. One solution is to obtain a table based on exact ages, assume that the average applicant is 35.5, and use an interpolated value when determining premiums. A second solution is to perform the mortality study using the ages assigned by the company rather than the policyholder's true age. In the example, the applicant is considered to be exactly age 35 on October 15, 2016 and is thus assigned a new birthday of October 15, 1981. When this is done, the study is said to use **insuring ages** and the resulting values can be used directly for insurance calculations.

■ EXAMPLE 14.21

Suppose that the company in Example 14.18 assigned insuring ages by age last birthday. Use the actuarial exposure method to estimate all possible mortality values.

1. Born 4-1981, purchased insurance policy on 8-2013, was an active policyholder on 1-2017. New birthday is 8-1981, enters at 32-5, exits at 35-5.

2. Born 6-1981, purchased insurance policy on 7-2013, died 9-2015. New birthday is 7-1981, enters at 32-6, dies at 34-2.

3. Born 8-1981, purchased insurance policy on 2-2015, surrendered policy on 2-2016. New birthday is 2-1982, enters at 33-0, exits at 34-0.

4. Born 5-1981, purchased insurance policy on 6-2014, died 3-2015. New birthday is 6-1981, enters at 33-0, dies at 33-9.

5. Born 7-1981, purchased insurance policy on 3-2014, surrendered policy on 5-2016. New birthday is 3-1982, enters at 32-0, exits at 34-2.

The exposures are now:

32–33: $e_{32} = 7 + 6 + 0 + 0 + 12 = 25$.
33–34: $e_{33} = 12 + 12 + 12 + 12 + 12 = 60$.
34–35: $e_{34} = 12 + 12 + 0 + 0 + 2 = 26$.
35–36: $e_{35} = 5 + 0 + 0 + 0 + 0 = 5$.

As expected, the exposures are assigned to younger age intervals, reflecting the fact that each applicant is assigned an age that is less than their true age. The estimates are $\hat{q}_{33} = 1/(60/12) = 0.2$ and $\hat{q}_{34} = 1/(26/12) = 0.4615$. $\qquad\square$

Note that with insuring ages, those who enter observation after the study begins are first observed on their newly assigned birthday. Thus there are no approximation issues with regard to those numbers.

14.7.2.2 *Anniversary-Based Mortality Studies*

The mortality studies described so far in this section are often called calendar-based or **date-to-date** studies because the period of study runs from one calendar date to another calendar date. It is also common for mortality studies of insured persons to use a different setup.

Instead of having the observations run from one calendar date to another calendar date, observation for a particular policyholder begins on the first policy anniversary following the fixed start date of the study and ends on the last anniversary prior to the study's fixed end date. Such studies are often called **anniversary-to-anniversary** studies. We can illustrate this through a previous example.

Consider Example 14.18, with the study now running from anniversaries in 2014 to anniversaries in 2016. The first policy comes under observation on 8-2014 at insuring age 33-0 and exits the study on 8-2016 at insuring age 35-0. Policyholder 2 begins observation on 7-2014 at insuring age 33-0. Policyholder 5 surrendered after the 2016 anniversary, so observation ends on 3-2016 at age 34-0. All other ages remain the same. In this setting, all subjects begin observations at an integral age and all who are active policyholders at the end of the study do so at an integral age. Only the ages of death and surrender may be other than integers (and note that with the actuarial exposure method, in calculating the exposure, deaths are placed at the next integral age). There is a price to be paid for this convenience. In a three-year study such as the one in the example, no single policyholder can be observed for more than two years. In the date-to-date version, some policies will contribute three years of exposure.

All of the examples have used one-year time periods. If the length of an interval $(c_{j+1} - c_j)$ is not equal to 1, an adjustment is necessary. Exposures should be the fraction of the period under observation and not the length of time.

14.7.3 Interval-Based Methods

Instead of recording the exact age at which an event happens, all that is recorded is the age interval in which it took place and the nature of the event. As with the individual method,

Table 14.20 The exact age values for Example 14.22.

Age	Number at age	Number entering	Number dying	Number exiting	Number at next age
32	0	3	0	0	3
33	3	2	1	0	4
34	4	0	1	2	1
35	1	0	0	1	0

for a portfolio of insurance policies, only running totals need to be recorded, and the end result is just four to six[9] numbers for each age interval:

1. The number of persons at the beginning of the interval carried over from the previous interval.

2. The number of additional persons entering at the beginning of the interval.

3. The number of persons entering during the interval.

4. The number of persons exiting by death during the interval.

5. The number of persons exiting during the interval for reasons other than death.

6. The number of persons exiting at the end of the interval by other than death.

■ **EXAMPLE 14.22**

Consider two versions of Example 14.18. In the first one, use exact ages and a date-to-date study. Then use age last birthday insuring ages and an anniversary-to-anniversary study. For each, construct a table of the required values.

For the exact age study, the entry age, exit age, and cause of exit values for the five lives were $(32\text{-}9, 35\text{-}9, s)$, $(32\text{-}7, 34\text{-}3, x)$, $(33\text{-}6, 34\text{-}6, s)$, $(33\text{-}1, 33\text{-}10, x)$, and $(32\text{-}8, 34\text{-}10, s)$. There are no lives at age 32. Between ages 32 and 33, three lives enter and none exit. Between ages 33 and 34, two lives enter and one exits by death. Between ages 34 and 35, no lives enter, one dies, and two exit by other causes. Between ages 35 and 36, no lives enter and one life exits by a cause other than death. No lives enter or exit at age boundaries. The full set of values is in Table 14.20.

For the insuring age study, the values are $(33\text{-}0, 35\text{-}0, s)$, $(33\text{-}0, 34\text{-}2, x)$, $(33\text{-}0, 34\text{-}0, s)$, $(33\text{-}0, 33\text{-}9, x)$, and $(32\text{-}0, 34\text{-}0, s)$. In this case, there are several entries and exits at integral ages. The values are shown in Table 14.21. It is important to note that in the first table those who enter and exit do so during the age interval, while in the second table those who enter do so at the beginning of the interval and those who exit by reason other than death do so at the end of the age interval. □

[9]As will be seen in the examples, not every category need be present in a given situation.

Table 14.21 The insuring age values for Example 14.22.

Age	Number at age	Number entering	Number dying	Number exiting	Number at next age
32	0	1	0	0	1
33	1	4	1	2	2
34	2	0	1	1	0

The analysis of this situation is relatively simple. For the interval from age c_j to age c_{j+1}, let P_j be the number of lives under observation at age c_j. This number includes those carried over from the prior interval as well as those entering at age c_j. Let n_j, d_j, and w_j be the number entering, dying, and exiting during the interval. Note that, in general, $P_{j+1} \neq P_j + n_j - d_j - w_j$, as the right-hand side must be adjusted by those who exit or enter at exact age c_{j+1}. Estimating the mortality probability depends on the method selected and an assumption about when the events that occur during the age interval take place.

One approach is to assume a uniform distribution of the events during the interval. For the exact exposure method, the P_j who start the interval have the potential to contribute a full unit of exposure and the n_j entrants during the year add another half-year each (on average). Similarly, those who die or exit subtract one-half year on average. Thus the net exposure is $P_j + (n_j - d_j - w_j)/2$. For the actuarial exposure method, those who die do not reduce the exposure, and it becomes $P_j + (n_j - w_j)/2$.

Another approach is to adapt the Kaplan–Meier estimator to this situation. Suppose that the deaths all occur at midyear and all other decrements occur uniformly through the year. Then the risk set at midyear is $P_j + (n_j - w_j)/2$ and the estimator is the same as the actuarial estimator.

■ **EXAMPLE 14.23**

Apply both methods to the two data sets in Example 14.22.

For the exact age data, the exact exposures for ages 32–35 are $0 + (3 - 0 - 0)/2 = 1.5$, $3 + (2 - 1 - 0)/2 = 3.5$, $4 + (0 - 1 - 2)/2 = 2.5$, and $1 + (0 - 0 - 1)/2 = 0.5$, respectively. The use of actuarial exposures changes the values to $0 + (3 - 0)/2 = 1.5$, $3 + (2 - 0)/2 = 4.0$, $4 + (0 - 2)/2 = 3.0$, and $1 + (0 - 1)/2 = 0.5$.

For the insuring age data, the exact exposures for ages 32–34 are $1 + (0 - 0 - 0)/2 = 1.0$, $5 + (0 - 1 - 0)/2 = 4.5$, and $2 + (0 - 1 - 0)/2 = 1.5$. The actuarial exposures are $1 + (0 - 0)/2 = 1.0$, $5 + (0 - 0)/2 = 5.0$, and $2 + (0 - 0)/2 = 2.0$. Note that the entrants all occur at the beginning of the interval and so are included in P_j, while those who exit do so at the end of the interval and are not included in w_j. □

The goal of all the estimation procedures in this book is to deduce the probability distribution for the random variable in the absence of truncation and censoring. For loss data, that would be the probabilities if there were no deductible or limit, that is, ground-up losses. For lifetime data, it would be the probability distribution of the age at death if we could follow the person from birth to death. These are often referred to as **single-decrement probabilities** and are typically denoted q'_j in life insurance mathematics. In the life insurance context, the censoring rates are often as important as the mortality rates. For example, in the context of Data Set D, both time to death and time to withdrawal may be of interest. In the former case, withdrawals cause observations to be censored. In the latter

Table 14.22   The single-decrement mortality probabilities for Example 14.24.

j	P_j	n_j^b	n_j^m	d_j	w_j^m	w_j^e	$q_j'^{(d)}$
0	0	30	3	1	3	0	$1/30.0 = 0.0333$
1	29	0	1	0	2	0	$0/28.5 = 0.0000$
2	28	0	3	2	3	0	$2/28.0 = 0.0714$
3	26	0	3	3	3	0	$3/26.0 = 0.1154$
4	23	0	0	2	4	17	$2/21.0 = 0.0952$

case, censoring is caused by death. A superscript identifies the decrement of interest. For example, suppose that the decrements were death (d) and withdrawal (w). Then $q_j'^{(w)}$ is the actuarial notation for the probability that a person alive and insured at age c_j withdraws prior to age c_{j+1} in an environment where withdrawal is the only decrement, that is, that death is not possible. When the causes of censoring are other important decrements, an often-used assumption is that all the decrements are stochastically independent. That is, that they do not influence each other. For example, a person who withdraws at a particular age has the same probability of dying in the following month as a person who does not.

■ **EXAMPLE 14.24**

Estimate single-decrement probabilities using Data Set D and the actuarial method. Make reasonable assumptions.

First consider the decrement death. In the notation of this section, the relevant quantities are shown in Table 14.22. The notation n_j^b refers to entrants who do so at the beginning of the interval, while n_j^m refers to those entering during the interval. These categories are based not on the time of the event but, rather, on its nature. The 30 policies that were issued during the study are at time zero by definition. A policy that was at duration 1.0 when the study began could have entered at any duration. Such policies are included in n_j^m for the preceding interval. For those who exit other than by death, the notation is w_j^m for those exiting during the interval and w_j^e for those doing so at the end. Again, those who do so by chance are assigned to the previous interval. Deaths are also assigned to the previous interval.

For withdrawals, the values of $q_j'^{(w)}$ are given in Table 14.23. To keep the notation consistent, d_j now refers to withdrawals and w_j refers to other decrements.   □

Table 14.23   The single-decrement withdrawal probabilities for Example 14.24.

j	P_j	n_j^b	n_j^m	d_j	w_j^m	w_j^e	$q_j'^{(d)}$
0	0	30	3	3	1	0	$3/31.0 = 0.0968$
1	29	0	1	2	0	0	$2/29.5 = 0.0678$
2	28	0	3	3	2	0	$3/28.5 = 0.1053$
3	26	0	3	3	3	0	$3/26.0 = 0.1154$
4	23	0	0	4	2	17	$4/22.0 = 0.1818$

Table 14.24 The data for Example 14.25.

Range	Deductible 0	Deductible 250	Deductible 500	Total
0–100	15			15
100–250	16			16
250–500	34	96		130
500–1,000	73	175	251	499
1,000–2,500	131	339	478	948
2,500–5,000	83	213	311	607
5,000–7,500	12	48	88	148
7,500–10,000	1	4	11	16
At 5,000	7	17	18	42
At 7,500	5	10	15	30
At 10,000	2	1	4	7
Total	379	903	1,176	2,458

Table 14.25 The calculations for Example 14.25.

c_j	P_j	n_j^b	d_j	w_j^e	$q_j'^{(d)}$	$\hat{F}(c_j)$
0	0	379	15	0	$15/379 = 0.0396$	0.0000
100	364	0	16	0	$16/364 = 0.0440$	0.0396
250	348	903	130	0	$130/1{,}251 = 0.1039$	0.0818
500	1,121	1,176	499	0	$499/2{,}297 = 0.2172$	0.1772
1,000	1,798	0	948	0	$948/1{,}798 = 0.5273$	0.3560
2,500	850	0	607	42	$607/850 = 0.7141$	0.6955
5,000	201	0	148	30	$148/201 = 0.7363$	0.9130
7,500	23	0	16	7	$16/23 = 0.6957$	0.9770
10,000	0					0.9930

■ **EXAMPLE 14.25**

Loss data for policies with deductibles of 0, 250, and 500 and policy limits of 5,000, 7,500, and 10,000 were collected. The data appear in Table 14.24. Use the actuarial method to estimate the distribution function for losses.

The calculations are in Table 14.25. Because the deductibles and limits are at the endpoints of intervals, the only reasonable assumption is the first one presented. □

14.7.4 Exercises

14.43 Verify the calculations in Table 14.23.

14.44 For an anniversary-to-anniversary study, the values in Table 14.26 were obtained.

Table 14.26   The data for Exercise 14.44.

d	u	x	d	u	x
45	46.0		45	45.8	
45	46.0		46	47.0	
45		45.3	46	47.0	
45		46.7	46	46.3	
45		45.4	46		46.2
45	47.0		46		46.4
45	45.4		46	46.9	

Table 14.27   The data for Exercise 14.45.

Deductible	Payment[a]	Deductible	Payment
250	2,221	500	3,660
250	*2,500*	500	215
250	207	500	1,302
250	3,735	500	*10,000*
250	*5,000*	1,000	1,643
250	517	1,000	3,395
250	5,743	1,000	3,981
500	*2,500*	1,000	3,836
500	525	1,000	*5,000*
500	4,393	1,000	1,850
500	*5,000*	1,000	6,722

[a]Numbers in italics indicate that the amount paid was at the policy limit.

Estimate $q_{45}'^{(d)}$ and $q_{46}'^{(d)}$ using the exact Kaplan–Meier estimate, exact exposure, and actuarial exposure.

14.45 Twenty-two insurance payments are recorded in Table 14.27. Use the fewest reasonable number of intervals and an interval-based method with actuarial exposure to estimate the probability that a policy with a deductible of 500 will have a payment in excess of 5,000.

14.46 (*) Nineteen losses were observed. Six had a deductible of 250, six had a deductible of 500, and seven had a deductible of 1,000. Three losses were paid at a policy limit, those values being 1,000, 2,750, and 5,500. For the 16 losses not paid at the limit, one was in the interval (250, 500), two in (500, 1,000), four in (1,000, 2,750), seven in (2,750, 5,500), one in (5,500, 6,000), and one in (6,000, 10,000). Estimate the probability that a policy with a deductible of 500 will have a claim payment in excess of 5,500.

14.8   Maximum Likelihood Estimation of Decrement Probabilities

In Section 14.7, methods were introduced for estimating mortality probabilities with large data sets. One of the methods was a seriatim method using exact exposure. In this section,

that estimator will be shown to be maximum likelihood under a particular assumption. To do this, we need to develop some notation. Suppose that we are interested in estimating the probability that an individual alive at age a dies prior to age b, where $a < b$. This is denoted $q = [S(a) - S(b)]/S(a)$. Let X be the random variable with survival function $S(x)$, the probability of surviving from birth to age x. Now let Y be the random variable X conditioned on $X > a$. Its survival function is $S_Y(y) = \Pr(X > y | X > a) = S(y)/S(a)$.

We now introduce a critical assumption about the shape of the survival function within the interval under consideration. Assume that $S_Y(y) = \exp[-(y - a)\lambda]$ for $a < y \leq b$. This means that the survival function decreases exponentially within the interval. Equivalently, the hazard rate (called the force of mortality in life insurance mathematics) is assumed to be constant within the interval. Beyond b, a different hazard rate can be used. Our objective is to estimate the conditional probability q. Thus we can perform the estimation using only data from and a functional form for this interval. Values of the survival function beyond b will not be needed.

Now consider data collected on n individuals, all of whom were observed during the age interval $(a, b]$. For individual j, let g_j be the age at which the person was first observed within the interval and let h_j be the age at which the person was last observed within the interval (thus $a \leq g_j < h_j \leq b$). Let $\delta_j = 0$ if the individual was alive when last observed and $\delta_j = 1$ if the individual was last observed due to death. For this analysis, we assume that each individual's censoring age (everyone who does not die in the interval will be censored, either by reaching age b or through some event that removes them from observation) is known in advance. Thus the only random quantities are δ_j, and for individuals with $\delta_j = 1$, the age at death. The likelihood function is

$$L(\lambda) = \prod_{j=1}^{n} \left[\frac{S(h_j)}{S(g_j)} \right]^{1-\delta_j} \left[\frac{f(h_j)}{S(g_j)} \right]^{\delta_j} = \prod_{j=1}^{n} \left[e^{-(h_j-g_j)\lambda} \right]^{1-\delta_j} \left[\lambda e^{-(h_j-g_j)\lambda} \right]^{\delta_j}$$

$$= \lambda^d \prod_{j=1}^{n} e^{-(h_j-g_j)\lambda} = \lambda^d \exp\left[-\sum_{j=1}^{n}(h_j - g_j)\lambda \right] = \lambda^d \exp(-e\lambda),$$

where $d = \sum_{j=1}^{n} \delta_j$ is the number of observed deaths and $e = \sum_{j=1}^{n}(h_j - g_j)$ is the total time the individuals were observed in the interval (which was called exact exposure in Section 14.7). Taking logarithms, differentiating, and solving produces

$$l(\lambda) = d \ln \lambda - e\lambda,$$

$$l'(\lambda) = \frac{d}{\lambda} - e = 0,$$

$$\hat{\lambda} = \frac{d}{e}.$$

Finally, the maximum likelihood estimate of the probability of death is $\hat{q} = 1 - \exp[-(b - a)\hat{\lambda}] = 1 - \exp[-(b - a)d/e]$.

Studies often involve random censoring, where individuals may exit for reasons other than death at times that were not known in advance. If all decrements (e.g. death, disability, and retirement) are stochastically independent (that is, the timing of one event does not influence any of the others), then the maximum likelihood estimator turns out to be identical to the one derived in this section. Although we do not derive the result, note that it follows from the fact that the likelihood function can be decomposed into separate factors for each decrement.

The variance of this estimator can be approximated using the observed information approach. The second derivative of the loglikelihood function is

$$l''(\lambda) = -\frac{d}{\lambda^2}.$$

Substitution of the estimator produces

$$l''(\hat{\lambda}) = -\frac{d}{(d/e)^2} = -\frac{e^2}{d}$$

and so $\widehat{\mathrm{Var}}(\hat{\lambda}) = d/e^2$. Using the delta method,

$$\frac{dq}{d\lambda} = (b-a)\exp[-(b-a)\lambda],$$

$$\widehat{\mathrm{Var}}(\hat{q}) = (1-\hat{q})^2(b-a)^2\frac{d}{e^2}.$$

Recall from Section 14.7 that there is an alternative called actuarial exposure, with $\hat{q} = (b-a)d/e$ with e calculated in a different manner. When analyzing results from this approach, it is common to assume that d is the result of a binomial experiment with sample size $e/(b-a)$. Then,

$$\widehat{\mathrm{Var}}(\hat{q}) = \frac{\hat{q}(1-\hat{q})}{e/(b-a)}.$$

If the $1 - \hat{q}$ terms are dropped (and they are often close to 1), the two variance formulas are identical (noting that the values of e will be slightly different).

■ **EXAMPLE 14.26**

A pension plan assigns every employee an integral age on their date of hire (thus retirement age is always reached on an anniversary of being hired and not on a birthday). Because of the nature of the employee contracts, employees can only quit their job on annual anniversaries of being hired or six months after an anniversary. They can die at any age. Using assigned ages, a mortality study observed employees from age 35 to age 36. There were 10,000 who began at age 35. Of them, 100 died between ages 35 and 35.5 at an average age of 35.27, 400 quit at age 35.5, 110 died between ages 35.5 and 36 at an average age of 35.78, and the remaining 9,390 were alive and employed at age 36. Using both exact and actuarial exposure, estimate the single-decrement value of q_{35} and the standard deviation of the estimator.

The exact exposure is $100(0.27) + 400(0.5) + 110(0.78) + 9,390(1) = 9,702.8$. Then, $\hat{q}_{35} = 1 - \exp(-210/9,702.8) = 0.02141$. The estimated standard deviation is $0.97859(210)^{1/2}/9,702.8 = 0.00146$. Recall that actuarial exposure assumes deaths occur at the end of the interval. The exposure is now $100(1) + 400(0.5) + 110(1) + 9,390(1) = 9,800$. Then, $\hat{q}_{35} = 210/9,800 = 0.02143$. The estimated standard deviation is $\sqrt{0.02143(0.97857)/9,800} = 0.00146$. □

14.8.1 Exercise

14.47 In Exercise 14.44, mortality estimates for q_{45} and q_{46} were obtained by Kaplan–Meier, exact exposure, and actuarial exposure. Approximate the variances of these estimates (using Greenwood's formula in the Kaplan–Meier case).

14.9 Estimation of Transition Intensities

The discussion to this point has concerned estimating the probability of a decrement in the absence of other decrements. An unstated assumption was that the environment in which the observations are made is one where once any decrement occurs, the individual is no longer observed.

A common, and more complex, situation is one in which after a decrement occurs, the individual remains under observation, with the possibility of further decrements. A simple example is a disability income policy. A healthy individual can die, become disabled, or surrender their policy. Those who become disabled continue to be observed, with possible decrements being recovery or death. Scenarios such as this are referred to as **multistate models**. Such models are discussed in detail in Dickson et al. [28]. In this section, we cover estimation of the transition intensities associated with such models. The results presented are based on Waters [129].

For notation, let the possible states be $\{0, 1, \ldots, k\}$ and let μ_x^{ij} be the force of transition to state j for an individual who is currently between ages x and $x + 1$ and is in state i. This notation is based on an assumption that the force of transition is constant over an integral age. This is similar to the earlier assumption that the force of decrement is constant over a given age.

■ **EXAMPLE 14.27**

For the disability income policy previously described, identify the states and indicate which transition intensities likely have positive values. Of those with positive values, which can be estimated from available data?

The assigning of numbers to states is arbitrary. For this situation, consider $0 =$ healthy and an active policyholder, $1 =$ disabled and receiving benefits, $2 =$ surrendered, and $3 =$ dead. Possible transitions are $0 \to 1$, $0 \to 2$, $0 \to 3$, $1 \to 0$, $1 \to 3$, and $2 \to 3$. All but the last one can be observed (we are unlikely to know when a surrendered policyholder dies). States 0 and 1 have been carefully defined to exclude cases where a surrendered policyholder becomes disabled or healthy. □

While not shown here, maximum likelihood estimates turn out to be based on exact exposure for the time spent in each state. For those between ages x and $x + 1$ (which can be generalized for periods of other than one year), let T_i be the total time policyholders are observed in state i and d_{ij} be the number of observed transitions from state i to state j. Then, $\hat{\mu}_x^{ij} = d_{ij}/T_i$. Similarly, $\widehat{\text{Var}}(\hat{\mu}_x^{ij}) = d_{ij}/T_i^2$.

■ **EXAMPLE 14.28**

Consider five policyholders who, between ages 50 and 51, are observed to do the following (decimals are fractions of a year):

▪ Disabled at age 50, dies at age 50.27.

▪ Healthy at age 50, disabled at age 50.34, dies at age 50.78.

▪ Healthy at age 50, surrendered at age 50.80.

▪ Purchases policy (healthy) at age 50.31, healthy at age 51.

▪ Healthy at age 50, disabled at 50.12, healthy at 50.45, dies at age 50.91.

Calculate the maximum likelihood estimates of the transition intensities.

Time spent healthy is $50.34 - 50 + 50.80 - 50 + 51 - 50.31 + 50.12 - 50 + 50.91 - 50.45 = 2.41$. While healthy, two became disabled, one surrendered, and one died. Hence the estimated transition intensities are $\hat{\mu}_{50}^{01} = 2/2.41$, $\hat{\mu}_{50}^{02} = 1/2.41$, and $\hat{\mu}_{50}^{03} = 1/2.41$. Time spent disabled is $50.27 - 50 + 50.78 - 50.34 + 50.45 - 50.12 = 1.04$. While disabled, one became healthy and two died. The estimates are $\hat{\mu}_{50}^{10} = 2/1.04$ and $\hat{\mu}_{50}^{13} = 1/1.04$. □

The construction of interval-based methods is more difficult because it is unclear when to place the transitions. Those who make one transition in the year may be reasonably placed at mid-age. However, those who make two transitions would more reasonably be placed at the one-third and two-thirds points. This would require careful data-keeping and the counting of many different cases.

15

MODEL SELECTION

15.1 Introduction

When using data to build a model, the process must end with the announcement of a "winner." While qualifications, limitations, caveats, and other attempts to escape full responsibility are appropriate, and often necessary, a commitment to a solution is often required. In this chapter, we look at a variety of ways to evaluate a model and compare competing models. But we must also remember that whatever model we select, it is only an approximation of reality. This observation is reflected in the following modeler's motto:

> All models are wrong, but some models are useful.[1]

Thus, our goal is to determine a model that is good enough to use to answer the question. The challenge here is that the definition of "good enough" will depend on the particular application. Another important modeling point is that a solid understanding of the question will guide you to the answer. The following quote from John Tukey [122, pp. 13–14] sums up this point:

---

[1]It is usually attributed to George Box.

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

In this chapter, a specific modeling strategy is considered. Our preference is to have a single approach that can be used for any probabilistic modeling situation. A consequence is that for any particular modeling situation, there may be a better (more reliable or more accurate) approach. For example, while maximum likelihood is a good estimation method for most settings, it may not be the best[2] for certain distributions. A literature search will turn up methods that have been optimized for specific distributions, but they are not mentioned here. Similarly, many of the hypothesis tests used here give approximate results. For specific cases, better approximations, or maybe even exact results, are available. They are also bypassed. The goal here is to outline a method that will give reasonable answers most of the time and be adaptable to a variety of situations.

This chapter assumes that you have a basic understanding of mathematical statistics as reviewed in Chapters 10 and 11. The remaining sections cover a variety of evaluation and selection tools. Each tool has its own strengths and weaknesses, and it is possible for different tools to lead to different models, making modeling as much art as science. At times, in real-world applications, the model's purpose may lead the analyst to favor one tool over another.

15.2   Representations of the Data and Model

All the approaches to be presented compare the proposed model to the data or to another model. The proposed model is represented by either its density or distribution function, or perhaps some functional of these quantities such as the limited expected value function or the mean excess loss function. The data can be represented by the empirical distribution function or a histogram. The graphs are easy to construct when there is individual, complete data. When there is grouping or observations have been truncated or censored, difficulties arise. Here, the only cases covered are those where the data are all truncated at the same value (which could be zero) and are all censored at the same value (which could be infinity). Extensions to the case of multiple truncation or censoring points are detailed in Klugman and Rioux [75].[3] It should be noted that the need for such representations applies only to continuous models. For discrete data, issues of censoring, truncation, and grouping rarely apply. The data can easily be represented by the relative or cumulative frequencies at each possible observation.

With regard to representing the data, the empirical distribution function is used for individual data and the histogram will be used for grouped data.

To compare the model to truncated data, we begin by noting that the empirical distribution begins at the truncation point and represents conditional values (i.e. they are the distribution and density function given that the observation exceeds the truncation point). To make a comparison to the empirical values, the model must also be truncated.

---

[2]There are many definitions of "best." Combining the Cramér–Rao lower bound with Theorem 11.4 indicates that maximum likelihood estimators are asymptotically optimal using unbiasedness and minimum variance as the definition of "best."

[3]Because the Kaplan–Meier estimate can be used to represent data with multiple truncation or censoring points, constructing graphical comparisons of the model and data is not difficult. The major challenge is generalizing the hypothesis tests to this situation.

Let the truncation point in the data set be t. The modified functions are

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

and

$$f^*(x) = \begin{cases} 0, & x < t, \\ \dfrac{f(x)}{1 - F(t)}, & x \geq t. \end{cases}$$

In this chapter, when a distribution function or density function is indicated, a subscript equal to the sample size indicates that it is the empirical model (from Kaplan–Meier, Nelson–Åalen, the ogive, etc.), while no adornment or the use of an asterisk (*) indicates the estimated parametric model. There is no notation for the true, underlying distribution because it is unknown and unknowable.

15.3 Graphical Comparison of the Density and Distribution Functions

The most direct way to see how well the model and data match is to plot the respective density and distribution functions.

■ **EXAMPLE 15.1**

Consider Data Sets B and C as given in Tables 15.1 and 15.2. For this example and all that follow, in Data Set B replace the value at 15,743 with 3,476 (to allow the graphs to fit comfortably on a page). Truncate Data Set B at 50 and Data Set C at 7,500. Estimate the parameter of an exponential model for each data set. Plot the appropriate functions and comment on the quality of the fit of the model. Repeat this for Data Set B censored at 1,000 (without any truncation).

Table 15.1   Data Set B with the highest value changed.

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1,193	1,340	1,884	2,558	3,476

Table 15.2   Data Set C.

Payment range	Number of payments
0–7,500	99
7,500–17,500	42
17,500–32,500	29
32,500–67,500	28
67,500–125,000	17
125,000–300,000	9
Over 300,000	3

Exponential fit
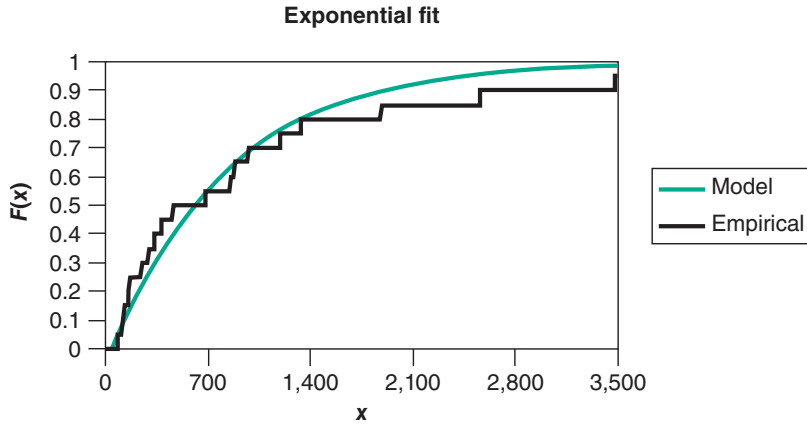
Figure 15.1   The model versus data cdf plot for Data Set B truncated at 50.

For Data Set B, there are 19 observations (the first observation is removed due to truncation). A typical contribution to the likelihood function is $f(82)/[1 - F(50)]$. The maximum likelihood estimate of the exponential parameter is $\hat{\theta} = 802.32$. The empirical distribution function starts at 50 and jumps 1/19 at each data point. The distribution function, using a truncation point of 50, is

$$F^*(x) = \frac{1 - e^{-x/802.32} - (1 - e^{-50/802.32})}{1 - (1 - e^{-50/802.32})} = 1 - e^{-(x-50)/802.32}.$$

Figure 15.1 presents a plot of these two functions.

The fit is not as good as we might like, because the model understates the distribution function at smaller values of x and overstates the distribution function at larger values of x. This result is not good because it means that tail probabilities are understated.

For Data Set C, the likelihood function uses the truncated values. For example, the contribution to the likelihood function for the first interval is

$$\left[\frac{F(17,500) - F(7,500)}{1 - F(7,500)} \right]^{42}.$$

The maximum likelihood estimate is $\hat{\theta} = 44,253$. The height of the first histogram bar is

$$\frac{42}{128(17,500 - 7,500)} = 0.0000328,$$

and the last bar is for the interval from 125,000 to 300,000 (a bar cannot be constructed for the interval from 300,000 to infinity). The density function must be truncated at 7,500 and becomes

$$\begin{aligned} f^*(x) &= \frac{f(x)}{1 - F(7,500)} = \frac{44,253^{-1}e^{-x/44,253}}{1 - (1 - e^{-7,500/44,253})} \\ &= \frac{e^{-(x-7,500)/44,253}}{44,253}, \quad x > 7,500. \end{aligned}$$
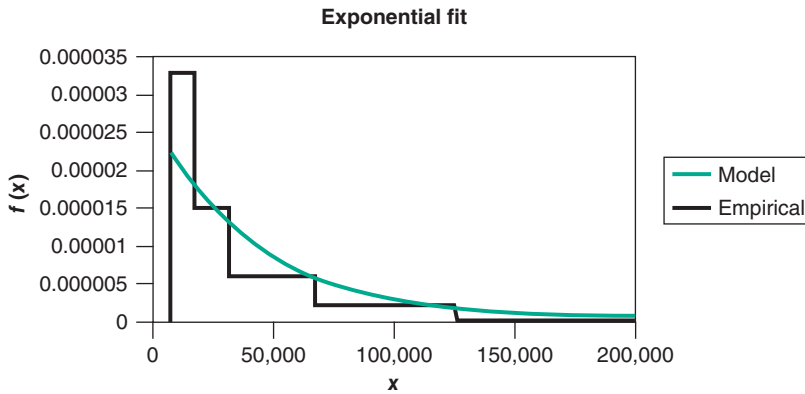
Figure 15.2   The model versus data density plot for Data Set C truncated at 7,500.
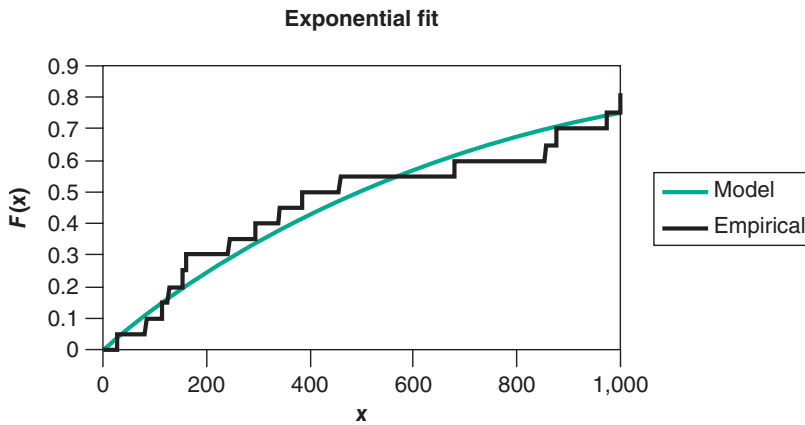
Figure 15.3   The model versus data cdf plot for Data Set B censored at 1,000.

The plot of the density function versus the histogram is given in Figure 15.2.

The exponential model understates the early probabilities. It is hard to tell from the plot how they compare above 125,000.

For Data Set B modified with a limit, the maximum likelihood estimate is $\hat{\theta} = 718.00$. When constructing the plot, the empirical distribution function must stop at 1,000. The plot appears in Figure 15.3.

Once again, the exponential model does not fit well. ☐

When the model's distribution function is close to the empirical distribution function, it is difficult to make small distinctions. Among the many ways to amplify those distinctions, two are presented here. The first is to simply plot the difference of the two functions. That is, if $F_n(x)$ is the empirical distribution function and $F^*(x)$ is the model distribution function, plot $D(x) = F_n(x) - F^*(x)$. There is no corresponding plot for grouped data.

■ **EXAMPLE 15.2**

Plot $D(x)$ for Data Set B for the previous example.
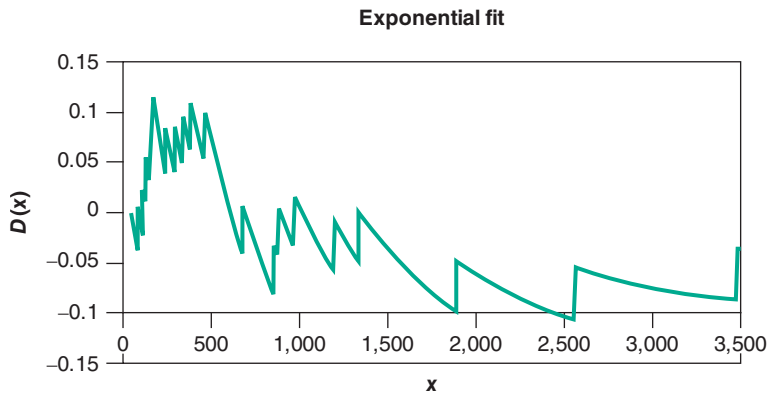
Figure 15.4   The model versus data $D(x)$ plot for Data Set B truncated at 50.
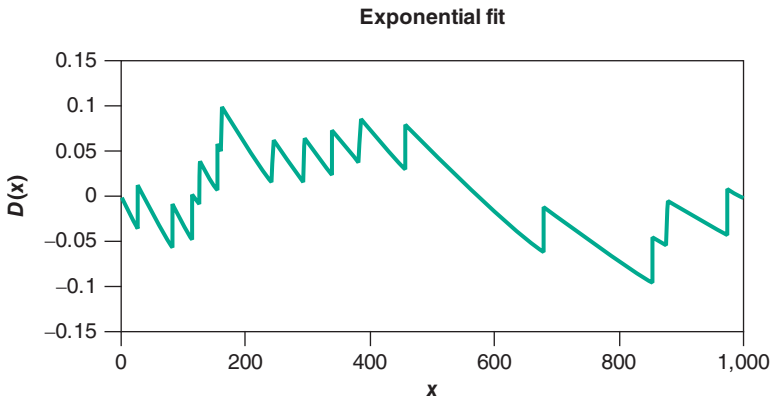
Figure 15.5   The model versus data $D(x)$ plot for Data Set B censored at 1,000.

For Data Set B truncated at 50, the plot appears in Figure 15.4. The lack of fit for this model is magnified in this plot.

For Data Set B censored at 1,000, the plot must again end at that value. It appears in Figure 15.5. The lack of fit continues to be apparent.   □

Another way to highlight any differences is the *p–p* plot, which is also called a **probability plot**. The plot is created by ordering the observations as $x_1 \leq \cdots \leq x_n$. A point is then plotted corresponding to each value. The coordinates to plot are $(F_n(x_j), F^*(x_j))$. If the model fits well, the plotted points will be near the 45° line running from $(0, 0)$ to $(1, 1)$. However, for this to be the case, a different definition of the empirical distribution function is needed. It can be shown that the expected value of $F(X_j)$ is $j/(n + 1)$ and, therefore, the empirical distribution should be that value and not the usual j/n. If two observations have the same value, either plot both points (they would have the same "*y*" value but different "*x*" values) or plot a single value by averaging the two "*x*" values.

■ **EXAMPLE 15.3**

Create a *p–p* plot for the continuing example.

Exponential fit
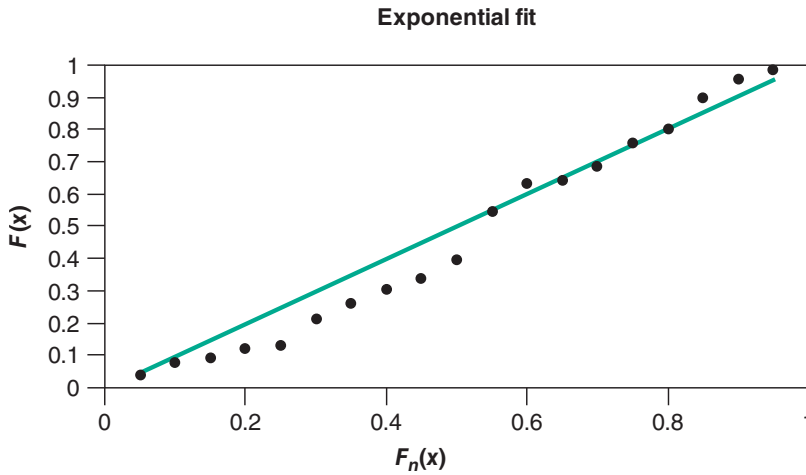
Figure 15.6   The p–p plot for Data Set B truncated at 50.

Exponential fit
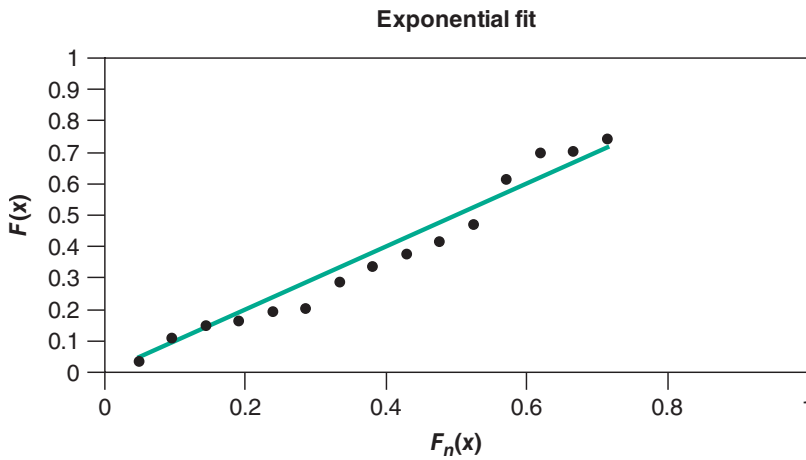
Figure 15.7   The p–p plot for Data Set B censored at 1,000.

For Data Set B truncated at 50, $n = 19$ and one of the observed values is $x = 82$. The empirical value is $F_{19}(82) = \frac{1}{20} = 0.05$. The other coordinate is

$$F^*(82) = 1 - e^{-(82-50)/802.32} = 0.0391.$$

One of the plotted points will be $(0.05, 0.0391)$. The complete graph appears in Figure 15.6.

From the lower left part of the plot, it is clear that the exponential model places less probability on small values than the data call for. A similar plot can be constructed for Data Set B censored at 1,000 and it appears in Figure 15.7.

This plot ends at about 0.75 because that is the highest probability observed prior to the censoring point at 1,000. There are no empirical values at higher probabilities. Again, the exponential model tends to underestimate the empirical values.   □

15.3.1   Exercises

15.1 Repeat Example 15.1 using a Weibull model in place of the exponential model.

15.2 Repeat Example 15.2 for a Weibull model.

15.3 Repeat Example 15.3 for a Weibull model.

15.4   Hypothesis Tests

A picture may be worth many words, but sometimes it is best to replace the impressions conveyed by pictures with mathematical demonstrations.[4]   One such demonstration is a test of the following hypotheses:

H_0 : The data came from a population with the stated model.

H_1 : The data did not come from such a population.

The test statistic is usually a measure of how close the model distribution function is to the empirical distribution function. When the null hypothesis completely specifies the model (e.g. an exponential distribution with mean 100), critical values are well known. However, it is more often the case that the null hypothesis states the name of the model but not its parameters. When the parameters are estimated from the data, the test statistic tends to be smaller than it would have been had the parameter values been prespecified. This relationship occurs because the estimation method itself tries to choose parameters that produce a distribution that is close to the data. When parameters are estimated from data, the tests become approximate. Because rejection of the null hypothesis occurs for large values of the test statistic, the approximation tends to increase the probability of a Type II error (declaring that the model is acceptable when it is not) while lowering the probability of a Type I error (rejecting an acceptable model).[5]   For actuarial modeling, this tendency is likely to be an acceptable trade-off.

   One method of avoiding the approximation is to randomly divide the sample into two sets. One is termed the training set. This set is used to estimate the parameters. The other set is called the test or validation set. This set is used to evaluate the quality of the model fit. This is more realistic because the model is being validated against new data. This approach is easier to do when there is a lot of data so that both sets are large enough to give useful results. These methods will not be discussed further in this text: for more details, see James et al. [61].

15.4.1   The Kolmogorov–Smirnov Test

Let t be the left truncation point ($t = 0$ if there is no truncation) and let u be the right censoring point ($u = \infty$ if there is no censoring). Then, the test statistic is

$$D = \max_{t \leq x \leq u} \left| F_n(x) - F^*(x) \right|.$$

---

[4]Thus this section is an application of the Society of Actuaries' motto (due to Ruskin): "The work of science is to substitute facts for appearances and demonstrations for impressions."

[5]Among the tests presented here, only the chi-square test has a built-in correction for this situation. Modifications for the other tests have been developed, but they are not presented here.

Table 15.3   The calculation of D for Example 15.4.

x	$F^*(x)$	$F_n(x-)$	$F_n(x)$	Maximum difference
82	0.0391	0.0000	0.0526	0.0391
115	0.0778	0.0526	0.1053	0.0275
126	0.0904	0.1053	0.1579	0.0675
155	0.1227	0.1579	0.2105	0.0878
161	0.1292	0.2105	0.2632	0.1340
243	0.2138	0.2632	0.3158	0.1020
294	0.2622	0.3158	0.3684	0.1062
340	0.3033	0.3684	0.4211	0.1178
384	0.3405	0.4211	0.4737	0.1332
457	0.3979	0.4737	0.5263	0.1284
680	0.5440	0.5263	0.5789	0.0349
855	0.6333	0.5789	0.6316	0.0544
877	0.6433	0.6316	0.6842	0.0409
974	0.6839	0.6842	0.7368	0.0529
1,193	0.7594	0.7368	0.7895	0.0301
1,340	0.7997	0.7895	0.8421	0.0424
1,884	0.8983	0.8421	0.8947	0.0562
2,558	0.9561	0.8947	0.9474	0.0614
3,476	0.9860	0.9474	1.0000	0.0386

This test as presented here should only be used on individual data to ensure that the step function $F_n(x)$ is well defined.[6] Also, the model distribution function $F^*(x)$ is assumed to be continuous over the relevant range.

■   **EXAMPLE 15.4**

Calculate D for Example 15.1.

Table 15.3 provides the needed values. Because the empirical distribution function jumps at each data point, the model distribution function must be compared both before and after the jump. The values just before the jump are denoted $F_n(x-)$ in the table. The maximum is $D = 0.1340$.

For Data Set B censored at 1,000, 15 of the 20 observations are uncensored. Table 15.4 illustrates the needed calculations. The maximum is $D = 0.0991$.   □

All that remains is to determine the critical value. Commonly used critical values for this test are $1.22/\sqrt{n}$ for $\alpha = 0.10$, $1.36/\sqrt{n}$ for $\alpha = 0.05$, and $1.63/\sqrt{n}$ for $\alpha = 0.01$. When $u < \infty$, the critical value should be smaller because there is less opportunity for the difference to become large. Modifications for this phenomenon exist in the literature

---

[6]It is possible to modify the Kolmogorov–Smirnov test for use with grouped data. For one approach to making the modifications, see Pettitt and Stephens [101].

Table 15.4   The calculation of D with censoring for Example 15.4.

x	$F^*(x)$	$F_n(x-)$	$F_n(x)$	Maximum difference
27	0.0369	0.00	0.05	0.0369
82	0.1079	0.05	0.10	0.0579
115	0.1480	0.10	0.15	0.0480
126	0.1610	0.15	0.20	0.0390
155	0.1942	0.20	0.25	0.0558
161	0.2009	0.25	0.30	0.0991
243	0.2871	0.30	0.35	0.0629
294	0.3360	0.35	0.40	0.0640
340	0.3772	0.40	0.45	0.0728
384	0.4142	0.45	0.50	0.0858
457	0.4709	0.50	0.55	0.0791
680	0.6121	0.55	0.60	0.0621
855	0.6960	0.60	0.65	0.0960
877	0.7052	0.65	0.70	0.0552
974	0.7425	0.70	0.75	0.0425
1,000	0.7516	0.75	0.75	0.0016

(see, e.g., Stephens [116], which also includes tables of critical values for specific null distribution models), and one such modification is given in Klugman and Rioux [75] but is not introduced here.

■ **EXAMPLE 15.5**

Complete the Kolmogorov–Smirnov test for the previous example.

For Data Set B truncated at 50, the sample size is 19. The critical value at a 5% significance level is $1.36/\sqrt{19} = 0.3120$. Because $0.1340 < 0.3120$, the null hypothesis is not rejected and the exponential distribution is a plausible model. While it is unlikely that the exponential model is appropriate for this population, the sample size is too small to lead to that conclusion. For Data Set B censored at 1,000, the sample size is 20, and so the critical value is $1.36/\sqrt{20} = 0.3041$ and the exponential model is again viewed as being plausible.   □

For both this test and the Anderson–Darling test that follows, the critical values are correct only when the null hypothesis completely specifies the model. When the data set is used to estimate parameters for the null hypothesized distribution (as in the example), the correct critical value is smaller. For both tests, the change depends on the particular distribution that is hypothesized and maybe even on the particular true values of the parameters. An indication of how simulation can be used for this situation is presented in Section 19.4.5.

15.4.2 The Anderson–Darling Test

This test is similar to the Kolmogorov–Smirnov test but uses a different measure of the difference between the two distribution functions. The test statistic is

$$A^2 = n \int_t^u \frac{[F_n(x) - F^*(x)]^2}{F^*(x)[1 - F^*(x)]} f^*(x) \, dx.$$

That is, it is a weighted average of the squared differences between the empirical and model distribution functions. Note that when x is close to t or to u, the weights might be very large due to the small value of one of the factors in the denominator. This test statistic tends to place more emphasis on good fit in the tails than in the middle of the distribution. Calculating with this formula appears to be challenging. However, for individual data (so this is another test that does not work for grouped data), the integral simplifies to

$$A^2 = -nF^*(u) + n \sum_{j=0}^{k} [1 - F_n(y_j)]^2 \{\ln[1 - F^*(y_j)] - \ln[1 - F^*(y_{j+1})]\}$$

$$+ n \sum_{j=1}^{k} F_n(y_j)^2 [\ln F^*(y_{j+1}) - \ln F^*(y_j)],$$

where the unique noncensored data points are $t = y_0 < y_1 < \cdots < y_k < y_{k+1} = u$. Note that when $u = \infty$, the last term of the first sum is zero (evaluating the formula as written will ask for $\ln(0)$). The critical values are 1.933, 2.492, and 3.857 for 10%, 5%, and 1% significance levels, respectively. As with the Kolmogorov–Smirnov test, the critical value should be smaller when $u < \infty$.

■ **EXAMPLE 15.6**

Perform the Anderson–Darling test for the continuing example.

For Data Set B truncated at 50, there are 19 data points. The calculation is shown in Table 15.5, where "summand" refers to the sum of the corresponding terms from the two sums. The total is 1.0226 and the test statistic is $-19(1) + 19(1.0226) = 0.4292$. Because the test statistic is less than the critical value of 2.492, the exponential model is viewed as plausible.

For Data Set B censored at 1,000, the results are shown in Table 15.6. The total is 0.7602 and the test statistic is $-20(0.7516) + 20(0.7602) = 0.1713$. Because the test statistic does not exceed the critical value of 2.492, the exponential model is viewed as plausible. □

15.4.3 The Chi-Square Goodness-of-Fit Test

Unlike the Kolmogorov–Smirnov and Anderson–Darling tests, this test allows for some discretion. It begins with the selection of $k-1$ arbitrary values, $t = c_0 < c_1 < \cdots < c_k = \infty$. Let $\hat{p}_j = F^*(c_j) - F^*(c_{j-1})$ be the probability a truncated observation falls in the interval from c_{j-1} to c_j. Similarly, let $p_{nj} = F_n(c_j) - F_n(c_{j-1})$ be the same probability according to the empirical distribution. The test statistic is then

$$\chi^2 = \sum_{j=1}^{k} \frac{n(\hat{p}_j - p_{nj})^2}{\hat{p}_j},$$

Table 15.5   The Anderson–Darling test for Example 15.6.

j	y_j	$F^*(x)$	$F_n(x)$	Summand
0	50	0.0000	0.0000	0.0399
1	82	0.0391	0.0526	0.0388
2	115	0.0778	0.1053	0.0126
3	126	0.0904	0.1579	0.0332
4	155	0.1227	0.2105	0.0070
5	161	0.1292	0.2632	0.0904
6	243	0.2138	0.3158	0.0501
7	294	0.2622	0.3684	0.0426
8	340	0.3033	0.4211	0.0389
9	384	0.3405	0.4737	0.0601
10	457	0.3979	0.5263	0.1490
11	680	0.5440	0.5789	0.0897
12	855	0.6333	0.6316	0.0099
13	877	0.6433	0.6842	0.0407
14	974	0.6839	0.7368	0.0758
15	1,193	0.7594	0.7895	0.0403
16	1,340	0.7997	0.8421	0.0994
17	1,884	0.8983	0.8947	0.0592
18	2,558	0.9561	0.9474	0.0308
19	3,476	0.9860	1.0000	0.0141
20	∞	1.0000	1.0000	

where n is the sample size. Another way to write the formula is to let $E_j = n\hat{p}_j$ be the number of expected observations in the interval (assuming that the hypothesized model is true) and let $O_j = np_{nj}$ be the number of observations in the interval. Then,

$$\chi^2 = \sum_{j=1}^{k} \frac{(E_j - O_j)^2}{E_j}.$$

The critical value for this test comes from the chi-square distribution with degrees of freedom equal to the number of terms in the sum (k) minus 1 minus the number of estimated parameters. There are a variety of rules that have been proposed for deciding when the test is reasonably accurate. They center around the values of $E_j = n\hat{p}_j$. The most conservative states that each must be at least 5. Some authors claim that values as low as 1 are acceptable. All agree that the test works best when the values are about equal from term to term. If the data are grouped, there is little choice but to use the groups as given, although adjacent groups could be combined to increase E_j. For individual data, the data can be grouped for the purpose of performing this test.[7]

[7]Moore [91] cites a number of rules. Among them are: (1) an expected frequency of at least 1 for all cells and an expected frequency of at least 5 for 80% of the cells; (2) an average count per cell of at least 4 when testing at the 1% significance level and an average count of at least 2 when testing at the 5% significance level; and (3) a sample size of at least 10, at least 3 cells, and the ratio of the square of the sample size to the number of cells of at least 10.

Table 15.6   The Anderson–Darling calculation for Example 15.6 with censored data.

j	y_j	$F^*(x)$	$F_n^*(x)$	Summand
0	0	0.0000	0.00	0.0376
1	27	0.0369	0.05	0.0718
2	82	0.1079	0.10	0.0404
3	115	0.1480	0.15	0.0130
4	126	0.1610	0.20	0.0334
5	155	0.1942	0.25	0.0068
6	161	0.2009	0.30	0.0881
7	243	0.2871	0.35	0.0493
8	294	0.3360	0.40	0.0416
9	340	0.3772	0.45	0.0375
10	384	0.4142	0.50	0.0575
11	457	0.4709	0.55	0.1423
12	680	0.6121	0.60	0.0852
13	855	0.6960	0.65	0.0093
14	877	0.7052	0.70	0.0374
15	974	0.7425	0.75	0.0092
16	1,000	0.7516	0.75	

Table 15.7   Data Set B truncated at 50 for Example 15.7.

Range	$\hat{p}$	Expected	Observed	χ^2
50–150	0.1172	2.227	3	0.2687
150–250	0.1035	1.966	3	0.5444
250–500	0.2087	3.964	4	0.0003
500–1,000	0.2647	5.029	4	0.2105
1,000–2,000	0.2180	4.143	3	0.3152
2,000–∞	0.0880	1.672	2	0.0644
Total	1	19	19	1.4034

■ **EXAMPLE 15.7**

Perform the chi-square goodness-of-fit test for the exponential distribution for the continuing example.

    All three data sets can be evaluated with this test. For Data Set B truncated at 50, establish boundaries at 50, 150, 250, 500, 1,000, 2,000, and infinity. The calculations appear in Table 15.7. The total is $\chi^2 = 1.4034$. With four degrees of freedom (6 rows minus 1 minus 1 estimated parameter), the critical value for a test at a 5% significance level is 9.4877 (this value can be obtained with the Excel$^{®}$ function CHISQ.INV(0.95,4)) and the p-value is 0.8436 (from $1-$CHISQ.DIST(1.4034,4,TRUE)). The exponential model is a good fit.

    For Data Set B censored at 1,000, the first interval is 0–150 and the last interval is 1,000–∞. Unlike the previous two tests, the censored observations can be used. The

Table 15.8   Data Set B censored at 1,000 for Example 15.7.

Range	$\hat{p}$	Expected	Observed	χ^2
0–150	0.1885	3.771	4	0.0139
150–250	0.1055	2.110	3	0.3754
250–500	0.2076	4.152	4	0.0055
500–1,000	0.2500	5.000	4	0.2000
1,000–∞	0.2484	4.968	5	0.0002
Total	1	20	20	0.5951

Table 15.9   Data Set C for Example 15.7.

Range	$\hat{p}$	Expected	Observed	χ^2
7,500–17,500	0.2023	25.889	42	10.026
17,500–32,500	0.2293	29.356	29	0.004
32,500–67,500	0.3107	39.765	28	3.481
67,500–125,000	0.1874	23.993	17	2.038
125,000–300,000	0.0689	8.824	9	0.003
300,000–∞	0.0013	0.172	3	46.360
Total	1	128	128	61.913

calculations are shown in Table 15.8. The total is $\chi^2 = 0.5951$. With three degrees of freedom (five rows minus one minus one estimated parameter), the critical value for a test at a 5% significance level is 7.8147 and the p-value is 0.8976. The exponential model is a good fit.

For Data Set C, the groups are already in place. The results are given in Table 15.9. The test statistic is $\chi^2 = 61.913$. There are four degrees of freedom, for a critical value of 9.488. The p-value is about 10^{-12}. There is clear evidence that the exponential model is not appropriate. A more accurate test would combine the last two groups (because the expected count in the last group is less than 1). The group from 125,000 to infinity has an expected count of 8.997 and an observed count of 12 for a contribution of 1.002. The test statistic is now 16.552, and with three degrees of freedom the p-value is 0.00087. The test continues to reject the exponential model.

□

Sometimes, the test can be modified to fit different situations. The following example illustrates this for aggregate frequency data.

■ **EXAMPLE 15.8**

Conduct an approximate goodness-of-fit test for the Poisson model determined in Example 12.9. The data are repeated in Table 15.10.

For each year, we are assuming that the number of claims is the result of the sum of a number (given by the exposure) of independent and identical random variables. In that case, the central limit theorem indicates that a normal approximation may be appropriate. The expected count (E_k) is the exposure times the estimated expected

Table 15.10   Automobile claims by year for Example 15.8.

Year	Exposure	Claims
1986	2,145	207
1987	2,452	227
1988	3,112	341
1989	3,458	335
1990	3,698	362
1991	3,872	359

value for one exposure unit, while the variance (V_k) is the exposure times the estimated variance for one exposure unit. The test statistic is then

$$Q = \sum_k \frac{(n_k - E_k)^2}{V_k}$$

and has an approximate chi-square distribution with degrees of freedom equal to the number of data points less the number of estimated parameters. The expected count is $E_k = \lambda e_k$ and the variance is $V_k = \lambda e_k$ also. The test statistic is

$$Q = \frac{(207 - 209.61)^2}{209.61} + \frac{(227 - 239.61)^2}{239.61} + \frac{(341 - 304.11)^2}{304.11}$$
$$+ \frac{(335 - 337.92)^2}{337.92} + \frac{(362 - 361.37)^2}{361.37} + \frac{(359 - 378.38)^2}{378.38}$$
$$= 6.19.$$

With five degrees of freedom, the 5% critical value is 11.07 and the Poisson hypothesis is accepted.   □

There is one important point to note about these tests. Suppose that the sample size were to double but that the sampled values were not much different (imagine each number showing up twice instead of once). For the Kolmogorov–Smirnov test, the test statistic would be unchanged, but the critical value would be smaller. For the Anderson–Darling and chi-square tests, the test statistic would double while the critical value would be unchanged. As a result, for larger sample sizes, it is more likely that the null hypothesis (and, thus, the proposed model) would be rejected. This outcome should not be surprising. We know that the null hypothesis is false (it is extremely unlikely that a simple distribution using a few parameters can explain the complex behavior that produced the observations), and with a large enough sample size we will have convincing evidence of that truth. When using these tests, we must remember that although all our models are wrong, some may be useful.

15.4.4   The Likelihood Ratio Test

An alternative question to "Could the population have distribution A?" is "Is distribution B a more appropriate representation of the population than distribution A?" More formally:

H_0 : The data came from a population with distribution A.
H_1 : The data came from a population with distribution B.

To perform a formal hypothesis test, distribution *A* must be a special case of distribution *B*, for example, exponential versus gamma. An easy way to complete this test is given as follows.

Definition 15.1 *The **likelihood ratio test** is conducted thus. First, let the likelihood function be written as $L(\theta)$. Let θ_0 be the value of the parameters that maximizes the likelihood function. However, only values of the parameters that are within the null hypothesis may be considered. Let $L_0 = L(\theta_0)$. Let θ_1 be the maximum likelihood estimator, where the parameters can vary over all possible values from the alternative hypothesis, and then let $L_1 = L(\theta_1)$. The test statistic is $T = 2\ln(L_1/L_0) = 2(\ln L_1 - \ln L_0)$. The null hypothesis is rejected if $T > c$, where c is calculated from $\alpha = \Pr(T > c)$, where T has a chi-square distribution with degrees of freedom equal to the number of free parameters in the model from the alternative hypothesis less the number of free parameters in the model from the null hypothesis.*

This test makes some sense. When the alternative hypothesis is true, forcing the parameter to be selected from the null hypothesis should produce a likelihood value that is significantly smaller.

■ **EXAMPLE 15.9**

You want to test the hypothesis that the population that produced Data Set B (using the original largest observation) has a mean that is other than 1,200. Assume that the population has a gamma distribution and conduct the likelihood ratio test at a 5% significance level. Also, determine the *p*-value.

The hypotheses are

$$H_0 : \text{gamma with } \mu = 1,200,$$
$$H_1 : \text{gamma with } \mu \neq 1,200.$$

From earlier work, the maximum likelihood estimates are $\hat{\alpha} = 0.55616$ and $\hat{\theta} = 2{,}561.1$. The loglikelihood at the maximum is $\ln L_1 = -162.293$. Next, the likelihood must be maximized, but only over those values α and θ for which $\alpha\theta = 1{,}200$. This restriction means that α can be free to range over all positive numbers, but that $\theta = 1{,}200/\alpha$. Thus, under the null hypothesis, there is only one free parameter. The likelihood function is maximized at $\hat{\alpha} = 0.54955$ and $\hat{\theta} = 2{,}183.6$. The loglikelihood at this maximum is $\ln L_0 = -162.466$. The test statistic is $T = 2(-162.293 + 162.466) = 0.346$. For a chi-square distribution with one degree of freedom, the critical value is 3.8415. Because $0.346 < 3.8415$, the null hypothesis is not rejected. The probability that a chi-square random variable with one degree of freedom exceeds 0.346 is 0.556, a *p*-value that indicates little support for the alternative hypothesis.   □

■ **EXAMPLE 15.10**

(*Example 6.3 continued*) Members of the $(a, b, 0)$ class were not sufficient to describe these data. Determine a suitable model.

Thirteen different distributions were fit to the data.[8] The results of that process

---

[8] For those reading this text to prepare for a professional examination, some of the distributions used in this

Table 15.11   Six useful models for Example 15.10.

Model	Number of parameters	Negative loglikelihood	χ^2	p-value
Negative binomial	2	5,348.04	8.77	0.0125
ZM logarithmic	2	5,343.79	4.92	0.1779
Poisson–inverse Gaussian	2	5,343.51	4.54	0.2091
ZM negative binomial	3	5,343.62	4.65	0.0979
Geometric–negative binomial	3	5,342.70	1.96	0.3754
Poisson–ETNB	3	5,342.51	2.75	0.2525

    revealed six models with p-values above 0.01 for the chi-square goodness-of-fit test. Information about those models is given in Table 15.11. The likelihood ratio test indicates that the three-parameter model with the smallest negative loglikelihood (Poisson–ETNB) is not significantly better than the two-parameter Poisson–inverse Gaussian model. The latter appears to be an excellent choice.   ☐

It is tempting to use this test when the alternative distribution simply has more parameters than the null distribution. In such cases, the test may not be appropriate. For example, it is possible for a two-parameter lognormal model to have a higher loglikelihood value than a three-parameter Burr model, resulting in a negative test statistic, indicating that a chi-square distribution is not appropriate. When the null distribution is a limiting (rather than special) case of the alternative distribution, the test may still be used, but the test statistic's distribution is now a mixture of chi-square distributions (see Self and Liang [112]). Regardless, it is still reasonable to use the "test" to make decisions in these cases, provided that it is clearly understood that a formal hypothesis test was not conducted.

15.4.5 Exercises

15.4 Use the Kolmogorov–Smirnov test to determine if a Weibull model is appropriate for the data used in Example 15.5.

15.5 (*) Five observations are made from a random variable. They are 1, 2, 3, 5, and 13. Determine the value of the Kolmogorov–Smirnov test statistic for the null hypothesis that $f(x) = 2x^{-2}e^{-2/x}$, $x > 0$.

15.6 (*) You are given the following five observations from a random sample: 0.1, 0.2, 0.5, 1.0, and 1.3. Calculate the Kolmogorov–Smirnov test statistic for the null hypothesis that the population density function is $f(x) = 2(1 + x)^{-3}$, $x > 0$.

15.7 Perform the Anderson–Darling test of the Weibull distribution for Example 15.6.

15.8 Repeat Example 15.7 for the Weibull model.

example (and some that follow) may not be covered in the required course of reading. While knowledge of these distributions is necessary to estimate the parameters and calculate the negative loglikelihood, once that is done, the methods discussed here can be applied without that knowledge.

Table 15.12   The data for Exercise 15.11.

Number of accidents	Days
0	209
1	111
2	33
3	7
4	3
5	2

15.9 (*) One hundred and fifty policyholders were observed from the time they arranged a viatical settlement until their death. No observations were censored. There were 21 deaths in the first year, 27 deaths in the second year, 39 deaths in the third year, and 63 deaths in the fourth year. The survival model

$$S(t) = 1 - \frac{t(t+1)}{20}, \quad 0 \le t \le 4,$$

is being considered. At a 5% significance level, conduct the chi-square goodness-of-fit test.

15.10 (*) Each day, for 365 days, the number of claims is recorded. The results were 50 days with no claims, 122 days with one claim, 101 days with two claims, 92 days with three claims, and no days with four or more claims. For a Poisson model, determine the maximum likelihood estimate of λ and then perform the chi-square goodness-of-fit test at a 2.5% significance level.

15.11 (*) During a one-year period, the number of accidents per day was distributed as given in Table 15.12. Test the hypothesis that the data are from a Poisson distribution with mean 0.6, using the maximum number of groups such that each group has at least five expected observations. Use a significance level of 5%.

15.12 (*) One thousand values were simulated from a uniform $(0, 1)$ distribution. The results were grouped into 20 ranges of equal width. The observed counts in each range were squared and added, resulting in a sum of 51,850. Determine the p-value for the chi-square goodness-of-fit test.

15.13 (*) Twenty claim amounts were sampled from a Pareto distribution with $\alpha = 2$ and θ unknown. The maximum likelihood estimate of θ is 7.0. Also, $\sum_{j=1}^{20} \ln(x_j + 7.0) = 49.01$ and $\sum_{j=1}^{20} \ln(x_j + 3.1) = 39.30$. The likelihood ratio test is used to test the null hypothesis that $\theta = 3.1$. Determine the p-value for this test.

15.14 Redo Example 15.8 assuming that each exposure unit has a geometric distribution. Conduct the approximate chi-square goodness-of-fit test. Is the geometric preferable to the Poisson model?

15.15 Using Data Set B (with the original largest value), determine if a gamma model is more appropriate than an exponential model. Recall that an exponential model is a gamma model with $\alpha = 1$. Useful values were obtained in Example 11.2.

Number of losses, k	Number of policies, n_k	Fitted model
0	20,592	20,596.76
1	2,651	2,631.03
2	297	318.37
3	41	37.81
4	7	4.45
5	0	0.52
6	1	0.06
≥ 7	0	0.00

Table 15.13 The data for Exercise 15.21.

15.16 Use Data Set C to choose a model for the population that produced those numbers. Choose from the exponential, gamma, and transformed gamma models. Information for the first two distributions was obtained in Example 11.3 and Exercise 11.17, respectively.

15.17 Conduct the chi-square goodness-of-fit test for each of the models obtained in Exercise 12.3.

15.18 Conduct the chi-square goodness-of-fit test for each of the models obtained in Exercise 12.5.

15.19 Conduct the chi-square goodness-of-fit test for each of the models obtained in Exercise 12.6.

15.20 For the data in Table 15.20, determine the method of moments estimates of the parameters of the Poisson–Poisson distribution where the secondary distribution is the ordinary (not zero-truncated) Poisson distribution. Perform the chi-square goodness-of-fit test using this model.

15.21 You are given the data in Table 15.13, which represent results from 23,589 automobile insurance policies. The third column, headed "Fitted model," represents the expected number of losses for a fitted (by maximum likelihood) negative binomial distribution.

(a) Perform the chi-square goodness-of-fit test at a significance level of 5%.

(b) Determine the maximum likelihood estimates of the negative binomial parameters r and β. This can be done from the given numbers without actually maximizing the likelihood function.

15.5 Selecting a Model

15.5.1 Introduction

Almost all of the tools are now in place for choosing a model. Before outlining a recommended approach, two important concepts must be introduced. The first is **parsimony**.

The principle of parsimony states that unless there is considerable evidence to do otherwise, a simpler model is preferred. The reason for this preference is that a complex model may do a great job of matching the data, but that is no guarantee that the model will match the population from which the observations were sampled. For example, given any set of 10 (x, y) pairs with unique x values, there will always be a polynomial of degree 9 or less that goes through all 10 points. But if these points were a random sample, it is highly unlikely that the population values would all lie on that polynomial. However, there may be a straight line that comes close to the sampled points as well as the other points in the population. This observation matches the spirit of most hypothesis tests. That is, do not reject the null hypothesis (and thus claim a more complex description of the population holds) unless there is strong evidence to do so.

The second concept does not have a name. It states that, if you try enough models, one will look good, even if it is not. Suppose that I have 900 models at my disposal. For most data sets, it is likely that one of them will fit extremely well, but it may not help us learn about the population.

Thus, in selecting models, there are two things to keep in mind:

1. Use a simple model if at all possible.
2. Restrict the universe of potential models.

The methods outlined in the remainder of this section help with the first point; the second one requires some experience. Certain models make more sense in certain situations, but only experience can enhance the modeler's senses so that only a short list of quality candidates is considered.

The section is split into two types of selection criteria. The first set is based on the modeler's judgment, while the second set is more formal in the sense that most of the time all analysts will reach the same conclusions because the decisions are made based on numerical measurements rather than charts or graphs.

15.5.2   Judgment-Based Approaches

Using judgment to select models involves one or more of the three concepts outlined herein. In all cases, the analyst's experience is critical.

First, the decision can be based on the various graphs (or tables based on the graphs) presented in this chapter, allowing the analyst to focus on aspects of the model that are important for the proposed application.[9] For example, it may be more important to fit the tail well or it may be more important to match the mode or modes. Even if a score-based approach is used, it may be appropriate to present a convincing picture to support the chosen model.

Second, the decision can be influenced by the success of particular models in similar situations or the value of a particular model for its intended use. For example, the 1941 CSO mortality table follows a Makeham distribution for much of its range of ages. In a time of limited computing power, such a distribution allowed for easier calculation of joint life values. As long as the fit of this model was reasonable, this advantage outweighed the use of a different, but better fitting, model. Similarly, if the Pareto distribution has been

---

[9]Besides the ones discussed here, there are other plots/tables that could be used, such as the q–q plot and a comparison of model and empirical limited expected values or mean excess loss functions.

used to model a particular line of liability insurance both by the analyst's company and by others, it may require more than the usual amount of evidence to change to an alternative distribution.

Third, the situation may completely determine the distribution. For example, suppose that a dental insurance contract provides for at most two checkups per year and suppose that individuals make two independent choices each year as to whether to have a checkup. If each time the probability is q, then the distribution must be binomial with $m = 2$.

Finally, it should be noted that the more algorithmic approaches outlined in this section do not always agree. In that case, judgment is most definitely required, if only to decide which algorithmic approach to use.

15.5.3 Score-Based Approaches

Some analysts might prefer an automated process for selecting a model. An easy way to do that would be to assign a score to each model and let the model with the best value win. The following scores are worth considering:

1. The Kolmogorov–Smirnov test statistic: Choose the model with the smallest value.

2. The Anderson–Darling test statistic: Choose the model with the smallest value.

3. The chi-square goodness-of-fit test statistic: Choose the model with the smallest value.

4. The chi-square goodness-of-fit test: Choose the model with the highest p-value.

5. The likelihood (or loglikelihood) function at its maximum: Choose the model with the largest value.

All but the chi-square p-value have a deficiency with respect to parsimony. First, consider the likelihood function. When comparing, say, an exponential to a Weibull model, the Weibull model must have a likelihood value that is at least as large as that of the exponential model. They would only be equal in the rare case that the maximum likelihood estimate of the Weibull parameter τ was equal to 1. Thus, the Weibull model would always win over the exponential model, a clear violation of the principle of parsimony. For the three test statistics, there is no assurance that the same relationship will hold, but it seems likely that, if a more complex model is selected, the fit measure will be better. The only reason the chi-square test p-value is immune from this problem is that with more complex models, the test has fewer degrees of freedom. It is then possible that the more complex model will have a smaller p-value. There is no comparable adjustment for the first two test statistics listed.

With regard to the likelihood value, there are two ways to proceed. One is to perform the likelihood ratio test and the other is to impose a penalty for employing additional parameters. The likelihood ratio test is technically only available when one model is a special case of another (e.g. Pareto versus generalized Pareto). The concept can be turned into an algorithm by using the test at a 5% significance level. Begin with the best one-parameter model (the one with the highest loglikelihood value). Add a second parameter only if the two-parameter model with the highest loglikelihood value shows an increase of at least 1.92 (so that twice the difference exceeds the critical value of 3.84). Then move to three-parameter models. If the comparison is to a two-parameter model, a 1.92 increase is again needed. If the early comparison led to keeping the one-parameter model, an increase

of 3.00 is needed (because the test has two degrees of freedom). To add three parameters requires a 3.91 increase; four parameters, a 4.74 increase; and so on. In the spirit of this chapter, this algorithm can be used even when one model is not a special case of the other model. However, it would not be appropriate to claim that a likelihood ratio test was being conducted.

Aside from the issue of special cases, the likelihood ratio test has the same problem as any hypothesis test. Were the sample size to double, the loglikelihoods would also double, making it more likely that a model with a higher number of parameters would be selected, tending to defeat the parsimony principle. Conversely, it could be argued that, if we possess a lot of data, we have the right to consider and fit more complex models. A method that effects a compromise between these positions is the Schwarz Bayesian Criterion (SBC) [110], which is also called the Bayesian Information Criterion (BIC). This method recommends that, when ranking models, a deduction of $(r/2) \ln n$ should be made from the loglikelihood value, where r is the number of estimated parameters and n is the sample size. Thus, adding a parameter requires an increase of $0.5 \ln n$ in the loglikelihood. For larger sample sizes, a greater increase is needed, but it is not proportional to the sample size itself.

An alternative penalty is the Akaike Information Criterion (AIC) [4]. This method deducts the number of parameters from the loglikelihood.[10] Section 3 of Brockett [17] promotes the AIC, while in a discussion of that paper Carlin provides support for the SBC. The difference in the two methods is that the SBC adjusts for the sample size while the AIC does not. To summarize, the scores are as follows:

$$\text{SBC/BIC: } \ln L - \frac{r}{2} \ln n; \quad \text{AIC: } \ln L - r. \tag{15.1}$$

■ **EXAMPLE 15.11**

For the continuing example in this chapter, choose between the exponential and Weibull models for the data.

Graphs were constructed in the various examples and exercises. Table 15.14 summarizes the numerical measures. For the truncated version of Data Set B, the SBC is calculated for a sample size of 19, while for the version censored at 1,000, there are 20 observations. For both versions of Data Set B, while the Weibull offers some improvement, it is not convincing. In particular, none of the likelihood ratio test, SBC, or AIC indicate value in the second parameter. For Data Set C, it is clear that the Weibull model is superior and provides an excellent fit. □

■ **EXAMPLE 15.12**

In Example 7.8, an *ad hoc* method was used to demonstrate that the Poisson–ETNB distribution provided a good fit. Use the methods of this chapter to determine a good model.

---

[10]When using software that computes the SBC and the AIC, it is important to note how they are being determined, as there are alternative formulas available. For example, when applying the AIC, the software may start with twice the loglikelihood and then subtract twice the number of parameters. It is also common to change the sign, in which case smaller values are preferred. For this text, The AIC and SBC/BIC values will be calculated as shown in (15.1).

Table 15.14 The results for Example 15.11.

Criterion	B truncated at 50		B censored at 1,000	
	Exponential	Weibull	Exponential	Weibull
K–S[a]	0.1340	0.0887	0.0991	0.0991
A–D[a]	0.4292	0.1631	0.1713	0.1712
χ^2	1.4034	0.3615	0.5951	0.5947
p-value	0.8436	0.9481	0.8976	0.7428
Loglikelihood	−146.063	−145.683	−113.647	−113.647
SBC	−147.535	−148.628	−115.145	−116.643
AIC	−147.063	−147.683	−114.647	−115.647
	C			
χ^2	61.913	0.3698		
p-value	10^{-12}	0.9464		
Loglikelihood	−214.924	−202.077		
SBC	−217.350	−206.929		
AIC	−215.924	−204.077		

[a]K–S and A–D refer to the Kolmogorov–Smirnov and Anderson–Darling test statistics, respectively.

The data set is very large and, as a result, requires a very close correspondence of the model to the data. The results are given in Table 15.15.

From Table 15.15, it is seen that the negative binomial distribution does not fit well, while the fit of the Poisson–inverse Gaussian is marginal at best ($p = 2.88\%$). The Poisson–inverse Gaussian is a special case ($r = -0.5$) of the Poisson–ETNB. Hence, a likelihood ratio test can be formally applied to determine if the additional parameter r is justified. Because the loglikelihood increases by 5, which is more than 1.92, the three-parameter model is a significantly better fit. The chi-square test shows that the Poisson–ETNB provides an adequate fit. In contrast, the SBC, but not the AIC, favors the Poisson–inverse Gaussian distribution. This illustrates that with large sample sizes, using the SBC makes it harder to add a parameter. Given the improved fit in the tail for the three-parameter model, it seems to be the best choice. ☐

■ **EXAMPLE 15.13**

This example is taken from Douglas [29, p. 254]. An insurance company's records for one year show the number of accidents per day that resulted in a claim to the insurance company for a particular insurance coverage. The results are shown in Table 15.16. Determine if a Poisson model is appropriate.

A Poisson model is fitted to these data. The method of moments and the maximum likelihood method both lead to the estimate of the mean:

$$\hat{\lambda} = \frac{742}{365} = 2.0329.$$

The results of a chi-square goodness-of-fit test are shown in Table 15.17. Any time such a table is made, the expected count for the last group is

$$E_{k+} = n\hat{p}_{k+} = n(1 - \hat{p}_0 - \cdots - \hat{p}_{k-1}).$$

Table 15.15 The results for Example 15.12.

Number of claims	Observed frequency	Fitted distributions		
		Negative binomial	Poisson–inverse Gaussian	Poisson–ETNB
0	565,664	565,708.1	565,712.4	565,661.2
1	68,714	68,570.0	68,575.6	68,721.2
2	5,177	5,317.2	5,295.9	5,171.7
3	365	334.9	344.0	362.9
4	24	18.7	20.8	29.6
5	6	1.0	1.2	3.0
6+	0	0.0	0.1	0.4
Parameters		$\beta = 0.0350662$	$\lambda = 0.123304$	$\lambda = 0.123395$
		$r = 3.57784$	$\beta = 0.0712027$	$\beta = 0.233862$
				$r = -0.846872$
Chi-square		12.13	7.09	0.29
Degrees of freedom		2	2	1
p-value		<1%	2.88%	58.9%
−Loglikelihood		251,117	251,114	251,109
SBC		−251,130	−251,127	−251,129
AIC		−251,119	−251,116	−251,112

Table 15.16 The data for Example 15.13.

Number of claims/day	Observed number of days
0	47
1	97
2	109
3	62
4	25
5	16
6	4
7	3
8	2
9+	0

The last three groups are combined to ensure an expected count of at least one for each row. The test statistic is 9.93 with six degrees of freedom. The critical value at a 5% significance level is 12.59 and the p-value is 0.1277. By this test, the Poisson distribution is an acceptable model; however, it should be noted that the fit is poorest at the large values, and with the model understating the observed values, this may be a risky choice. □

Table 15.17 The chi-square goodness-of-fit test for Example 15.13.

Claims/day	Observed	Expected	Chi-square
0	47	47.8	0.01
1	97	97.2	0.00
2	109	98.8	1.06
3	62	66.9	0.36
4	25	34.0	2.39
5	16	13.8	0.34
6	4	4.7	0.10
7+	5	1.8	5.66
Totals	365	365	9.93

Table 15.18 The test results for Example 15.14.

	Poisson	Geometric	ZM Poisson	ZM geometric
Chi-square	543.0	643.4	64.8	0.58
Degrees of freedom	2	4	2	2
p-value	< 1%	< 1%	< 1%	74.9%
Loglikelihood	−171,373	−171,479	−171,160	−171,133
SBC	−171,379.5	−171,485.5	−171,173	−171,146
AIC	−171,374	−171,480	−171,162	−171,135

■ **EXAMPLE 15.14**

The data in Table 12.7 come from Beard et al. [13] and are analyzed in Example 12.7. Determine a model that adequately describes the data.

Parameter estimates from fitting four models are shown in Table 12.7. Various fit measures are given in Table 15.18. Only the zero-modified geometric distribution passes the goodness-of-fit test. It is also clearly superior according to the SBC and AIC. A likelihood ratio test against the geometric has a test statistic of 2(171,479 − 171,133) = 692, which with one degree of freedom is clearly significant. This result confirms the qualitative conclusion in Example 12.7. □

■ **EXAMPLE 15.15**

The data in Table 15.19, from Simon [113], represent the observed number of claims per contract for 298 contracts. Determine an appropriate model.

The Poisson, negative binomial, and Polya–Aeppli distributions are fitted to the data. The Polya–Aeppli and the negative binomial are both plausible distributions. The p-value of the chi-square statistic and the loglikelihood both indicate that the Polya–Aeppli is slightly better than the negative binomial. The SBC and AIC verify that both models are superior to the Poisson distribution. The ultimate choice may depend on familiarity, prior use, and the computational convenience of the negative binomial versus the Polya–Aeppli model. □

Table 15.19   The fit of the Simon data for Example 15.15.

Number of claims/contract	Number of contracts	Fitted distributions		
		Poisson	Negative binomial	Polya–Aeppli
0	99	54.0	95.9	98.7
1	65	92.2	75.8	70.6
2	57	78.8	50.4	50.2
3	35	44.9	31.3	32.6
4	20	19.2	18.8	20.0
5	10	6.5	11.0	11.7
6	4	1.9	6.4	6.6
7	0	0.5	3.7	3.6
8	3	0.1	2.1	2.0
9	4	0.0	1.2	1.0
10	0	0.0	0.7	0.5
11	1	0.0	0.4	0.3
12+	0	0.0	0.5	0.3
Parameters		$\lambda = 1.70805$	$\beta = 1.15907$	$\lambda = 1.10551$
			$r = 1.47364$	$\beta = 0.545039$
Chi-square		72.64	4.06	2.84
Degrees of freedom		4	5	5
p-Value		<1%	54.05%	72.39%
Loglikelihood		−577.0	−528.8	−528.5
SBC		−579.8	−534.5	−534.2
AIC		−578.0	−530.8	−530.5

■ **EXAMPLE 15.16**

Consider the data in Table 15.20 on automobile liability policies in Switzerland, taken from Bühlmann [20]. Determine an appropriate model.

Three models are considered in Table 15.20. The Poisson distribution is a very bad fit. Its tail is far too light compared with the actual experience. The negative binomial distribution appears to be much better, but cannot be accepted because the p-value of the chi-square statistic is very small. The large sample size requires a better fit. The Poisson–inverse Gaussian distribution provides an almost perfect fit (the p-value is large). Note that the Poisson–inverse Gaussian has two parameters, like the negative binomial. The SBC and AIC also favor this choice. This example shows that the Poisson–inverse Gaussian can have a much heavier right-hand tail than the negative binomial.   □

■ **EXAMPLE 15.17**

Comprehensive medical claims were studied by Bevan [15] in 1963. Male (955 payments) and female (1,291 payments) claims were studied separately. The data

Table 15.20   The fit of the Bühlmann data for Example 15.16.

Number of accidents	Observed frequency	Fitted distributions		
		Poisson	Negative binomial	P.–i.G.[a]
0	103,704	102,629.6	103,723.6	103,710.0
1	14,075	15,922.0	13,989.9	14,054.7
2	1,766	1,235.1	1,857.1	1,784.9
3	255	63.9	245.2	254.5
4	45	2.5	32.3	40.4
5	6	0.1	4.2	6.9
6	2	0.0	0.6	1.3
7+	0	0.0	0.1	0.3
Parameters		$\lambda = 0.155140$	$\beta = 0.150232$	$\lambda = 0.144667$
			$r = 1.03267$	$\beta = 0.310536$
Chi-square		1,332.3	12.12	0.78
Degrees of freedom		2	2	3
p-Values		<1%	<1%	85.5%
Loglikelihood		$-55,108.5$	$-54,615.3$	$-54,609.8$
SBC		$-55,114.3$	$-54,627.0$	$-54,621.5$
AIC		$-55,109.5$	$-54,617.3$	$-54,611.8$

[a]P.–i.G. stands for Poisson–inverse Gaussian.

Table 15.21   The comprehensive medical losses for Example 15.17.

Loss	Male	Female
25–50	184	199
50–100	270	310
100–200	160	262
200–300	88	163
300–400	63	103
400–500	47	69
500–1,000	61	124
1,000–2,000	35	40
2,000–3,000	18	12
3,000–4,000	13	4
4,000–5,000	2	1
5,000–6,667	5	2
6,667–7,500	3	1
7,500–10,000	6	1

appear in Table 15.21, where there was a deductible of 25. Can a common model be used?

When using the combined data set, the lognormal distribution is the best two-parameter model. Its negative loglikelihood (NLL) is 4,580.20. This value is 19.09

Table 15.22   The number of actuaries per company for Example 15.18.

Number of actuaries	Number of companies – 1949	Number of companies – 1957
1	17	23
2	7	7
3–4	3	3
5–9	2	3
10+	0	1

better than the one-parameter inverse exponential model and 0.13 worse than the three-parameter Burr model. Because none of these models is a special case of the other, the likelihood ratio test (LRT) cannot be used, but it is clear that, using the 1.92 difference as a standard, the lognormal is preferred. The SBC requires an improvement of $0.5 \ln(2,246) = 3.86$, while the AIC requires 1.00, and again the lognormal is preferred. The parameters are $\mu = 4.5237$ and $\sigma = 1.4950$. When separate lognormal models are fitted to males ($\mu = 3.9686$ and $\sigma = 1.8432$) and females ($\mu = 4.7713$ and $\sigma = 1.2848$), the respective NLLs are 1,977.25 and 2,583.82 for a total of 4,561.07. This result is an improvement of 19.13 over a common lognormal model, which is significant by the LRT (3.00 needed), the SBC (7.72 needed), and the AIC (2.00 needed). Sometimes it is useful to be able to use the same nonscale parameter in both models. When a common value of σ is used, the NLL is 4,579.77, which is significantly worse than using separate models. □

■ **EXAMPLE 15.18**

In 1958, Longley-Cook [82] examined employment patterns of casualty actuaries. One of his tables listed the number of members of the Casualty Actuarial Society employed by casualty companies in 1949 (55 actuaries) and 1957 (78 actuaries). Using the data in Table 15.22, determine a model for the number of actuaries per company that employs at least one actuary and find out whether the distribution has changed over the eight-year period.

Because a value of zero is impossible, only zero-truncated distributions should be considered. In all three cases (1949 data only, 1957 data only, and combined data), the ZT logarithmic and ZT (extended) negative binomial distributions have acceptable goodness-of-fit test values. The improvements in NLL are 0.52, 0.02, and 0.94. The LRT can be applied (except that the ZT logarithmic distribution is a limiting case of the ZT negative binomial distribution with $r \to 0$), and the improvement is not significant in any of the cases. The same conclusions apply if the SBC or AIC are used. The parameter estimates (where β is the only parameter) are 2.0227, 2.8114, and 2.4479, respectively. The NLL for the combined data set is 74.35, while the total for the two separate models is 74.15. The improvement is only 0.20, which is not significant (there is one degree of freedom). Even though the estimated mean has increased from $2.0227/ \ln(3.0227) = 1.8286$ to $2.8114/ \ln(3.8114) = 2.1012$, there is not enough data to make a convincing case that the true mean has increased. □

Table 15.23   The data for exercise 15.22.

Number of accidents	Number of policies
0	100
1	267
2	311
3	208
4	87
5	23
6	4
Total	1,000

Table 15.24   The results for exercise 15.25.

Model	Number of parameters	Negative loglikelihood
Generalized Pareto	3	219.1
Burr	3	219.2
Pareto	2	221.2
Lognormal	2	221.4
Inverse exponential	1	224.3

15.5.4 Exercises

15.22 (*) One thousand policies were sampled and the number of accidents for each recorded. The results are shown in Table 15.23. Without doing any formal tests, determine which of the following five models is most appropriate: binomial, Poisson, negative binomial, normal, or gamma.

15.23 For Example 15.1, determine if a transformed gamma model is more appropriate than either the exponential model or the Weibull model for each of the three data sets.

15.24 (*) From the data in Exercise 15.11, the maximum likelihood estimates are $\hat{\lambda} = 0.60$ for the Poisson distribution and $\hat{r} = 2.9$ and $\hat{\beta} = 0.21$ for the negative binomial distribution. Conduct the likelihood ratio test for choosing between these two models.

15.25 (*) From a sample of size 100, five models are fitted with the results given in Table 15.24. Use the SBC and then the AIC to select the best model.

15.26 Refer to Exercise 11.27. Use the likelihood ratio test (at a 5% significance level), the SBC, and the AIC to decide if Sylvia's claim is true.

15.27 (*) Five models were fitted to a sample of 260 observations. The following are the number of parameters in the model followed by the loglikelihood value: 1, −414, 2, −412, 3, −411, 4, −409, 6, −409. According to the SBC, which model (identified by the number of parameters) should be selected? Does the decision change if the AIC is used?

15.28 Using results from Exercises 12.3 and 15.17, use the chi-square goodness-of-fit test, the likelihood ratio test, the SBC, and the AIC to determine the best model from the members of the $(a, b, 0)$ class.

Table 15.25   The data for exercise 15.31.

Number of medical claims	Number of accidents
0	529
1	146
2	169
3	137
4	99
5	87
6	41
7	25
8+	0

15.29 Using results from Exercises 12.5 and 15.18, use the chi-square goodness-of-fit test, the likelihood ratio test, the SBC, and the AIC to determine the best model from the members of the $(a, b, 0)$ class.

15.30 Using results from Exercises 12.6 and 15.19, use the chi-square goodness-of-fit test, the likelihood ratio test, the SBC, and the AIC to determine the best model from the members of the $(a, b, 0)$ class.

15.31 Table 15.25 gives the number of medical claims per reported automobile accident.

(a) Construct a plot similar to Figure 6.1. Does it appear that a member of the $(a, b, 0)$ class will provide a good model? If so, which one?

(b) Determine the maximum likelihood estimates of the parameters for each member of the $(a, b, 0)$ class.

(c) Based on the chi-square goodness-of-fit test, the likelihood ratio test, the SBC, and the AIC, which member of the $(a, b, 0)$ class provides the best fit? Is this model acceptable?

15.32 For the four data sets introduced in Exercises 12.3, 12.5, 12.6, and 15.31, you have determined the best model from among members of the $(a, b, 0)$ class. For each data set, determine the maximum likelihood estimates of the zero-modified Poisson, geometric, logarithmic, and negative binomial distributions. Use the chi-square goodness-of-fit test and likelihood ratio tests to determine the best of the eight models considered and state whether the selected model is acceptable.

15.33 A frequency model that has not been mentioned to this point is the **zeta distribution**. It is a zero-truncated distribution with $p_k^T = k^{-(\rho+1)}/\zeta(\rho+1), \ k = 1, 2, \ldots, \rho > 0$. The denominator is the zeta function, which must be evaluated numerically as $\zeta(\rho+1) = \sum_{k=1}^{\infty} k^{-(\rho+1)}$. The zero-modified zeta distribution can be formed in the usual way. More information can be found in Luong and Doray [84].

(a) Determine the maximum likelihood estimates of the parameters of the zero-modified zeta distribution for the data in Example 12.7.

Table 15.26 The data for exercise 15.35(a).

Number of claims	Number of policies
0	96,978
1	9,240
2	704
3	43
4	9
5+	0

(b) Is the zero-modified zeta distribution acceptable?

15.34 In Exercise 15.32, the best model from among the members of the $(a, b, 0)$ and $(a, b, 1)$ classes was selected for the data sets in Exercises 12.3, 12.5, 12.6, and 15.31. Fit the Poisson–Poisson, Polya–Aeppli, Poisson–inverse Gaussian, and Poisson–ETNB distributions to these data and determine if any of these distributions should replace the one selected in Exercise 15.32. Is the current best model acceptable?

15.35 The five data sets presented in this problem are all taken from Lemaire [79]. For each data set, compute the first three moments and then use the ideas in Section 7.2 to make a guess at an appropriate model from among the compound Poisson collection [(Poisson, geometric, negative binomial, Poisson–binomial (with $m = 2$ and $m = 3$), Polya–Aeppli, Neyman Type A, Poisson–inverse Gaussian, and Poisson–ETNB)]. From the selected model (if any) and members of the $(a, b, 0)$ and $(a, b, 1)$ classes, determine the best model.

(a) The data in Table 15.26 represent counts from third-party automobile liability coverage in Belgium.

(b) The data in Table 15.27 represent the number of deaths due to horse kicks in the Prussian army between 1875 and 1894. The counts are the number of deaths in a corps (there were 10 of them) in a given year, and thus there are 200 observations. This data set is often cited as the inspiration for the Poisson distribution. For using any of our models, what additional assumption about the data must be made?

(c) The data in Table 15.28 represent the number of major international wars per year from 1500 through 1931.

(d) The data in Table 15.29 represent the number of runs scored in each half-inning of World Series baseball games played from 1947 through 1960.

(e) The data in Table 15.30 represent the number of goals per game per team in the 1966–1967 season of the National Hockey League.

Table 15.27 The data for exercise 15.35(b).

Number of deaths	Number of corps
0	109
1	65
2	22
3	3
4	1
5+	0

Table 15.28 The data for exercise 15.35(c).

Number of wars	Number of years
0	223
1	142
2	48
3	15
4	4
5+	0

Table 15.29 The data for exercise 15.35(d).

Number of runs	Number of half innings
0	1,023
1	222
2	87
3	32
4	18
5	11
6	6
7+	3

15.36 Verify that the estimates presented in Example 7.14 are the maximum likelihood estimates. (Because only two decimals are presented, it is probably sufficient to observe that the likelihood function takes on smaller values at each of the nearby points.) The negative binomial distribution was fitted to these data in Example 12.5. Which of these two models is preferable?

Table 15.30 The data for exercise 15.35(e).

Number of goals	Number of games
0	29
1	71
2	82
3	89
4	65
5	45
6	24
7	7
8	4
9	1
10+	3

CREDIBILITY

16

INTRODUCTION TO LIMITED FLUCTUATION CREDIBILITY

16.1   Introduction

Credibility theory is a set of quantitative tools that allows an insurer to perform prospective experience rating (adjust future premiums based on past experience) on a risk or group of risks. If the experience of a policyholder is consistently better than that assumed in the underlying manual rate (sometimes called the **pure premium**), then the policyholder may demand a rate reduction.

The policyholder's argument is as follows. The manual rate is designed to reflect the expected experience (past and future) of the entire rating class and implicitly assumes that the risks are homogeneous. However, no rating system is perfect, and there always remains some heterogeneity in the risk levels after all the underwriting criteria are accounted for. Consequently, some policyholders will be better risks than that assumed in the underlying manual rate. Of course, the same logic dictates that a rate increase should be applied to a poor risk, but in this situation the policyholder is certainly not going to ask for a rate increase! Nevertheless, an increase may be necessary, due to considerations of equity and the economics of the situation.

The insurer is then forced to answer the following question: How much of the difference in experience of a given policyholder is due to random variation in the underlying claims experience and how much is due to the fact that the policyholder really is a better or worse risk than average? In other words, how credible is the policyholder's own experience? Two facts must be considered in this regard:

1. The more past information the insurer has on a given policyholder, the more **credible** is the policyholder's own experience, all else being equal. In the same manner, in group insurance the experience of larger groups is more credible than that of smaller groups.

2. Competitive considerations may force the insurer to give full (using the past experience of the policyholder only and not the manual rate) or nearly full credibility to a given policyholder in order to retain the business.

Another use for credibility is in the setting of rates for classification systems. For example, in workers compensation insurance, there may be hundreds of occupational classes, some of which may provide very little data. To accurately estimate the expected cost for insuring these classes, it may be appropriate to combine the limited actual experience with some other information, such as past rates, or the experience of occupations that are closely related.

From a statistical perspective, credibility theory leads to a result that would appear to be counterintuitive. If experience from an insured or group of insureds is available, our statistical training may convince us to use the sample mean or some other unbiased estimator. But credibility theory tells us that it is optimal to give only partial weight to this experience and give the remaining weight to an estimator produced from other information. We will discover that what we sacrifice in terms of bias, we gain in terms of reducing the average (squared) error.

Credibility theory allows an insurer to quantitatively formulate the problem of combining data with other information, and this part provides an introduction to this theory. This chapter deals with **limited fluctuation credibility theory**, a subject developed in the early part of the twentieth century. This theory provides a mechanism for assigning full (Section 16.3) or partial (Section 16.4) credibility to a policyholder's experience. The difficulty with this approach is the lack of a sound underlying mathematical theory to justify the use of these methods. Nevertheless, this approach provided the original treatment of the subject and is still in use today.

A classic paper by Bühlmann in 1967 [19] provides a statistical framework within which credibility theory has developed and flourished. While this approach, termed **greatest accuracy credibility theory**,[1] was formalized by Bühlmann, the basic ideas had been around for some time. This approach is introduced in Chapter 17. The simplest model, that of Bühlmann [19], is discussed in Section 17.5. Practical improvements were made by Bühlmann and Straub in 1970 [21]. Their model is discussed in Section 17.6. The concept of exact credibility is presented in Section 17.7.

Practical use of the theory requires that unknown model parameters be estimated from data. Chapter 18 covers two estimation approaches. Nonparametric estimation (where the problem is somewhat model free and the parameters are generic, such as the mean and

---

[1]The terms *limited fluctuation* and *greatest accuracy* go back at least as far as a 1943 paper by Arthur Bailey [8].

variance) is considered in Section 18.2. Semiparametric estimation (where some of the parameters are based on assuming particular distributions) is covered in Section 18.3.

We close this introduction with a quote from Arthur Bailey in 1950 [9, p. 8] that aptly summarizes much of the history of credibility. We, too, must tip our hats to the early actuaries, who, with unsophisticated mathematical tools at their disposal, were able to come up with formulas that not only worked but also were very similar to those that we carefully develop in this part:

> It is at this point in the discussion that the ordinary individual has to admit that, while there seems to be some hazy logic behind the actuaries' contentions, it is too obscure for him to understand. The trained statistician cries "Absurd! Directly contrary to any of the accepted theories of statistical estimation." The actuaries themselves have to admit that they have gone beyond anything that has been proven mathematically, that all of the values involved are still selected on the basis of judgment, and that the only demonstration they can make is that, in actual practice, it works. Let us not forget, however, that they have made this demonstration many times. It does work!

16.2 Limited Fluctuation Credibility Theory

This branch of credibility theory represents the first attempt to quantify the credibility problem. This approach was suggested in the early 1900s in connection with workers compensation insurance. The original paper on the subject was by Mowbray in 1914 [92]. The problem may be formulated as follows. Suppose that a policyholder has experienced X_j claims or losses[2] in past experience period j, where $j \in \{1, 2, 3, \ldots, n\}$. Another view is that X_j is the experience from the jth policy in a group or from the jth member of a particular class in a rating scheme. Suppose that $E(X_j) = \xi$, that is, the mean is stable over time or across the members of a group or class.[3] This quantity would be the premium to charge (net of expenses, profits, and a provision for adverse experience) if only we knew its value. Also suppose that $\text{Var}(X_j) = \sigma^2$, again, the same for all j. The past experience may be summarized by the average $\bar{X} = n^{-1}(X_1 + \cdots + X_n)$. We know that $E(\bar{X}) = \xi$, and if the X_j are independent, $\text{Var}(\bar{X}) = \sigma^2/n$. The insurer's goal is to decide on the value of ξ. One possibility is to ignore the past data (no credibility) and simply charge M, a value obtained from experience on other similar, but not identical, policyholders. This quantity is often called the **manual premium** because it would come from a book (manual) of premiums. Another possibility is to ignore M and charge $\bar{X}$ (full credibility). A third possibility is to choose some combination of M and $\bar{X}$ (partial credibility).

From the insurer's standpoint, it seems sensible to "lean toward" the choice $\bar{X}$ if the experience is more "stable" (less variable, σ^2 small). Stable values imply that $\bar{X}$ is of more use as a predictor of next year's results. Conversely, if the experience is more volatile (variable), then $\bar{X}$ is of less use as a predictor of next year's results and the choice of M makes more sense.

Also, if we have an a priori reason to believe that the chances are great that this policyholder is unlike those who produced the manual premium M, then more weight

---

[2]"Claims" refers to the number of claims and "losses" refers to payment amounts. In many cases, such as in this introductory paragraph, the ideas apply equally whether we are counting claims or losses.

[3]The customary symbol for the mean, μ, is not used here because that symbol is used for a different but related mean in Chapter 17. We have chosen this particular symbol ("xi") because it is the most difficult Greek letter to write and pronounce. It is an unwritten rule of textbook writing that it appear at least once.

should be given to $\bar{X}$ because, as an unbiased estimator, $\bar{X}$ tells us something useful about ξ while M is likely to be of little value. In contrast, if all of our other policyholders have similar values of ξ, there is no point in relying on the (perhaps limited) experience of any one of them when M is likely to provide an excellent description of the propensity for claims or losses.

While reference is made to policyholders, the entity contributing to each X_j could arise from a single policyholder, a class of policyholders possessing similar underwriting characteristics, or a group of insureds assembled for some other reason. For example, for a given year j, X_j could be the number of claims filed in respect of a single automobile policy in one year, the average number of claims filed by all policyholders in a certain ratings class (e.g. single, male, under age 25, living in an urban area, driving over 7,500 miles per year), or the average amount of losses per vehicle for a fleet of delivery trucks owned by a food wholesaler.

We first present one approach to decide whether to assign full credibility (charge $\bar{X}$), and then we present an approach to assign partial credibility if there is evidence that full credibility is inappropriate.

16.3 Full Credibility

One method of quantifying the stability of $\bar{X}$ is to infer that $\bar{X}$ is stable if the difference between $\bar{X}$ and ξ is small relative to ξ with high probability. In statistical terms, stability can be defined by selecting two numbers $r > 0$ and $0 < p < 1$ (with r close to 0 and p close to 1, common choices being $r = 0.05$ and $p = 0.9$) and assigning full credibility if

$$\Pr(-r\xi \le \bar{X} - \xi \le r\xi) \ge p. \tag{16.1}$$

It is convenient to restate (16.1) as

$$\Pr\left(\left| \frac{\bar{X} - \xi}{\sigma/\sqrt{n}} \right| \le \frac{r\xi\sqrt{n}}{\sigma} \right) \ge p.$$

Now let y_p be defined by

$$y_p = \inf_y \left\{ \Pr\left(\left| \frac{\bar{X} - \xi}{\sigma/\sqrt{n}} \right| \le y \right) \ge p \right\}. \tag{16.2}$$

That is, y_p is the smallest value of y that satisfies the probability statement in braces in (16.2). If $\bar{X}$ has a continuous distribution, the "$\ge$" sign in (16.2) may be replaced by an "$=$" sign, and y_p satisfies

$$\Pr\left(\left| \frac{\bar{X} - \xi}{\sigma/\sqrt{n}} \right| \le y_p \right) = p. \tag{16.3}$$

Then, the condition for full credibility is $r\xi\sqrt{n}/\sigma \ge y_p$,

$$\frac{\sigma}{\xi} \le \frac{r}{y_p}\sqrt{n} = \sqrt{\frac{n}{\lambda_0}}, \tag{16.4}$$

where $\lambda_0 = (y_p/r)^2$. Condition (16.4) states that full credibility is assigned if the coefficient of variation σ/ξ is no larger than $\sqrt{n/\lambda_0}$, an intuitively reasonable result.

Also of interest is that (16.4) can be rewritten to show that full credibility occurs when

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \leq \frac{\xi^2}{\lambda_0}. \tag{16.5}$$

Alternatively, solving (16.4) for n gives the number of exposure units required for full credibility, namely

$$n \geq \lambda_0 \left(\frac{\sigma}{\xi} \right)^2. \tag{16.6}$$

In many situations, it is reasonable to approximate the distribution of $\bar{X}$ by a normal distribution with mean ξ and variance σ^2/n. For example, central limit theorem arguments may be applicable if n is large. In that case, $(\bar{X} - \xi)/(\sigma/\sqrt{n})$ has a standard normal distribution. Then, (16.3) becomes (where Z has a standard normal distribution and $\Phi(y)$ is its cdf)

$$\begin{aligned} p &= \Pr(|Z| \leq y_p) \\ &= \Pr(-y_p \leq Z \leq y_p) \\ &= \Phi(y_p) - \Phi(-y_p) \\ &= \Phi(y_p) - 1 + \Phi(y_p) \\ &= 2\Phi(y_p) - 1. \end{aligned}$$

Therefore, $\Phi(y_p) = (1 + p)/2$, and y_p is the $(1 + p)/2$ percentile of the standard normal distribution.

For example, if $p = 0.9$, then standard normal tables give $y_{0.9} = 1.645$. If, in addition, $r = 0.05$, then $\lambda_0 = (1.645/0.05)^2 = 1{,}082.41$ and (16.6) yields $n \geq 1{,}082.41\sigma^2/\xi^2$. Note that this answer assumes we know the coefficient of variation of X_j. It is possible that we have some idea of its value, even though we do not know the value of ξ (remember, that is the quantity we want to estimate).

The important thing to note when using (16.6) is that the coefficient of variation is for the estimator of the quantity to be estimated. The right-hand side gives the standard for full credibility when measuring it in terms of exposure units. If some other unit is desired, it is usually sufficient to multiply both sides by an appropriate quantity. Finally, any unknown quantities will have be to estimated from the data, which implies that the credibility question can be posed in a variety of ways. The following examples cover the most common cases.

■ **EXAMPLE 16.1**

Suppose that past losses $X_1, \dots, X_n$ are available for a particular policyholder. The sample mean is to be used to estimate $\xi = \mathrm{E}(X_j)$. Determine the standard for full credibility. Then suppose that there were 10 observations, with six being zero and the others being 253, 398, 439, and 756. Determine the full-credibility standard for this case with $r = 0.05$ and $p = 0.9$.

The solution is available directly from (16.6) as

$$n \geq \lambda_0 \left(\frac{\sigma}{\xi} \right)^2.$$

For this specific case, the mean and standard deviation can be estimated from the data as 184.6 and 267.89 (where the variance estimate is the unbiased version using $n - 1$). With $\lambda_0 = 1082.41$, the standard is

$$n \geq 1082.41 \left(\frac{267.89}{184.6} \right)^2 = 2279.51,$$

and the 10 observations do not deserve full credibility. ☐

In the next example, it is further assumed that the observations are from a particular type of distribution.

■ **EXAMPLE 16.2**

Suppose that past losses $X_1, \ldots, X_n$ are available for a particular policyholder and it is reasonable to assume that the X_js are independent and compound Poisson distributed, that is, $X_j = Y_{j1} + \cdots + Y_{jN_j}$, where each N_j is Poisson with parameter λ and the claim size distribution Y has mean θ_Y and variance σ_Y^2. Determine the standard for full credibility when estimating the expected number of claims per policy and then when estimating the expected dollars of claims per policy. Then determine if these standards are met for the data in Example 16.1, where it is now known that the first three nonzero payments came from a single claim but the final one was from two claims, one for 129 and the other for 627.

Case 1: Accuracy is to be measured with regard to the average number of claims. Then, using the N_js rather than the X_js, we have $\xi = E(N_j) = \lambda$ and $\sigma^2 = \text{Var}(N_j) = \lambda$, implying from (16.6) that

$$n \geq \lambda_0 \left(\frac{\lambda^{1/2}}{\lambda} \right)^2 = \frac{\lambda_0}{\lambda}.$$

Thus, if the standard is in terms of the number of policies, it will have to exceed λ_0 / λ for full credibility and λ will have to be estimated from the data. If the standard is in terms of the number of expected claims, that is, $n\lambda$, we must multiply both sides by λ. Doing so sets the standard as

$$n\lambda \geq \lambda_0.$$

While it appears that no estimation is needed for this standard, it is in terms of the expected number of claims needed. In practice, the standard is set in terms of the actual number of claims experienced, effectively replacing $n\lambda$ on the left by its estimate $N_1 + \cdots + N_n$.

For the given data, there were five claims, for an estimate of λ of 0.5 per policy. The standard is then

$$n \geq \frac{1,082.41}{0.5} = 2,164.82,$$

and the 10 policies are far short of this standard. Or the five actual claims could be compared to $\lambda_0 = 1,082.41$, which leads to the same result.

Case 2: When accuracy is with regard to the average total payment, we have $\xi = E(X_j) = \lambda\theta_Y$ and $\text{Var}(X_j) = \lambda(\theta_Y^2 + \sigma_Y^2)$, formulas developed in Chapter 9. In terms

of the sample size, the standard is

$$n \geq \lambda_0 \frac{\lambda(\theta_Y^2 + \sigma_Y^2)}{\lambda^2 \theta_Y^2} = \frac{\lambda_0}{\lambda}\left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right].$$

If the standard is in terms of the expected number of claims, multiply both sides by λ to obtain

$$n\lambda \geq \lambda_0\left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right].$$

Finally, if the standard is in terms of the expected total dollars of claims, multiply both sides by θ_Y to obtain

$$n\lambda\theta_Y \geq \lambda_0\left(\theta_Y + \frac{\sigma_Y^2}{\theta_Y}\right).$$

For the given data, the five claims have mean 369.2 and standard deviation 189.315, and thus

$$n \geq \frac{\lambda_0}{\lambda}\left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right] = \frac{1,082.41}{0.5}\left[1 + \left(\frac{189.315}{369.2}\right)^2\right] = 2,734.02.$$

Again, the 10 observations are far short of what is needed. If the standard is to be set in terms of claims (of which there are five), multiply both sides by 0.5 to obtain a standard of 1,367.01. Finally, the standard could be set in terms of total dollars of claims. To do so, multiply both sides by 369.2 to obtain 504,701. Note that in all three cases, the ratio of the observed quantity to the corresponding standard is unchanged:

$$\frac{10}{2,734.02} = \frac{5}{1,367.01} = \frac{1,846}{504,701} = 0.003658.$$

□

In these examples, the standard for full credibility is not met, and so the sample means are not sufficiently accurate to be used as estimates of the expected value. We need a method for dealing with this situation.

16.4 Partial Credibility

If it is decided that full credibility is inappropriate, then for competitive reasons (or otherwise), it may be desirable to reflect the past experience $\bar{X}$ in the net premium as well as the externally obtained mean, M. An intuitively appealing method for combining the two quantities is through a weighted average, that is, through the credibility premium

$$P_c = Z\bar{X} + (1 - Z)M, \tag{16.7}$$

where the credibility factor $Z \in [0, 1]$ needs to be chosen. There are many formulas for Z that have been suggested in the actuarial literature, usually justified on intuitive rather than theoretical grounds. (We remark that Mowbray [92] considered full but not partial credibility.) One important choice is

$$Z = \frac{n}{n + k}, \tag{16.8}$$

where k needs to be determined. This particular choice will be shown to be theoretically justified on the basis of a statistical model to be presented in Chapter 17. Another choice, based on the same idea as full credibility (and including the full-credibility case $Z = 1$), is now discussed.

A variety of arguments have been used for developing the value of Z, many of which lead to the same answer. All of them are flawed in one way or another. The development we have chosen to present is also flawed but is at least simple. Recall that the goal of the full-credibility standard is to ensure that the difference between the net premium we are considering ($\bar{X}$) and what we should be using (ξ) is small with high probability. Because $\bar{X}$ is unbiased, achieving this standard is essentially (and exactly if $\bar{X}$ has the normal distribution) equivalent to controlling the variance of the proposed net premium, $\bar{X}$, in this case. We see from (16.5) that there is no assurance that the variance of $\bar{X}$ will be small enough. However, it is possible to control the variance of the credibility premium, P_c, as follows:

$$
\begin{aligned}
\frac{\xi^2}{\lambda_0} &= \mathrm{Var}(P_c) \\
&= \mathrm{Var}[Z\bar{X} + (1 - Z)M] \\
&= Z^2 \, \mathrm{Var}(\bar{X}) \\
&= Z^2 \frac{\sigma^2}{n}.
\end{aligned}
$$

Thus $Z = (\xi/\sigma)\sqrt{n/\lambda_0}$, provided that it is less than 1, which can be expressed using the single formula

$$
Z = \min\left\{ \frac{\xi}{\sigma} \sqrt{\frac{n}{\lambda_0}}, 1 \right\}. \tag{16.9}
$$

One interpretation of (16.9) is that the credibility factor Z is the ratio of the coefficient of variation required for full credibility ($\sqrt{n/\lambda_0}$) to the actual coefficient of variation. For obvious reasons, this formula is often called the square-root rule for partial credibility, regardless of what is being counted.

While we could do the algebra with regard to (16.9), it is sufficient to note that it always turns out that Z is the square root of the ratio of the actual count to the count required for full credibility.

■ **EXAMPLE 16.3**

Suppose, in Example 16.1, that the manual premium M is 225. Determine the credibility estimate.

The average of the payments is 184.6. With the square-root rule the credibility factor is

$$
Z = \sqrt{\frac{10}{2{,}279.51}} = 0.06623.
$$

Then the credibility premium is

$$
P_c = 0.06623(184.6) + 0.93377(225) = 222.32.
$$

□

■ **EXAMPLE 16.4**

Suppose, in Example 16.2, that the manual premium M is 225. Determine the credibility estimate using both cases.

For the first case, the credibility factor is

$$Z = \sqrt{\frac{5}{1,082.41}} = 0.06797$$

and applying it yields

$$P_c = 0.06797(184.6) + 0.93203(225) = 222.25.$$

At first glance, this approach may appear inappropriate. The standard was set in terms of estimating the frequency but was applied to the aggregate claims. Often, individuals are distinguished more by differences in the frequency with which they have claims rather than by differences in the cost per claim. So this factor captures the most important feature.

For the second case, we can use any of the three calculations:

$$Z = \sqrt{\frac{10}{2,734.02}} = \sqrt{\frac{5}{1,367.01}} = \sqrt{\frac{1,846}{504,701}} = 0.06048.$$

Then,

$$P_c = 0.06048(184.6) + 0.93952(225) = 222.56.$$

□

Earlier, we mentioned a flaw in the approach. Other than assuming that the variance captures the variability of $\bar{X}$ in the right way, all of the mathematics is correct. The flaw is in the goal. Unlike $\bar{X}$, P_c is not an unbiased estimator of ξ. In fact, one of the qualities that allows credibility to work is its use of biased estimators. But for biased estimators the appropriate measure of its quality is not its variance, but its MSE. However, the MSE requires knowledge of the bias and, in turn, that requires knowledge of the relationship of ξ and M. However, we know nothing about that relationship, and the data we have collected are of little help. As noted in Section 16.5, this is not only a problem with our determination of Z; it is a problem that is characteristic of the limited fluctuation approach. A model for this relationship is introduced in Chapter 17.

This section closes with a few additional examples. In the first two examples, $\lambda_0 = 1,082.41$ is used.

■ **EXAMPLE 16.5**

For group dental insurance, historical experience on many groups has revealed that annual losses per person insured have a mean of 175 and a standard deviation of 140. A particular group has been covered for two years, with 100 lives insured in year 1 and 110 in year 2, and has experienced average claims of 150 over that period. Determine if full or partial credibility is appropriate, and determine the credibility premium for next year's losses if there will be 125 lives insured.

We apply the credibility on a per-person-insured basis. We have observed $100 + 110 = 210$ exposure units (assume that experience is independent for different lives and years), and $\bar{X} = 150$. Now $M = 175$, and we assume that σ will be 140 for this group. Although this is an aggregate loss situation (each person has a random number of claims of random amounts), there is no information about the frequency and severity components. Thus (16.6) applies, where we estimate ξ with the sample mean of 150 to obtain the standard for full credibility as

$$n \geq 1,082.41 \left(\frac{140}{150}\right)^2 = 942.90$$

and then calculate

$$Z = \sqrt{\frac{210}{942.90}} = 0.472$$

(note that $\bar{X}$ is the average of 210 claims, so approximate normality is assumed by the central limit theorem). Thus, the net premium per person insured is

$$P_c = 0.472(150) + 0.528(175) = 163.2.$$

The net premium for the whole group is $125(163.2) = 20,400$. □

■ **EXAMPLE 16.6**

An insurance coverage involves credibility based on the number of claims only. For a particular group, 715 claims have been observed. Determine an appropriate credibility factor, assuming that the number of claims is Poisson distributed.

This is Case 1 from Example 16.4, and the standard for full credibility with regard to the number of claims is $n\lambda \geq \lambda_0 = 1,082.41$. Then,

$$Z = \sqrt{\frac{715}{1,082.41}} = 0.813.$$

□

■ **EXAMPLE 16.7**

Past data on a particular group are $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$, where the X_j are i.i.d. compound Poisson random variables with exponentially distributed claim sizes. If the credibility factor based on claim numbers is 0.8, determine the appropriate credibility factor based on total claims.

When based on Poisson claim numbers, from Example 16.2, $Z = 0.8$ implies that $\lambda n / \lambda_0 = (0.8)^2 = 0.64$, where λn is the observed number of claims. For exponentially distributed claim sizes, $\sigma_Y^2 = \theta_Y^2$. From Case 2 of Example 16.2, the standard for full credibility in terms of the number of claims is

$$n\lambda \geq \lambda_0 \left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right] = 2\lambda_0.$$

Then,

$$Z = \sqrt{\frac{\lambda n}{2\lambda_0}} = \sqrt{0.32} = 0.566.$$

□

16.5 Problems with the Approach

While the limited fluctuation approach yields simple solutions to the problem, there are theoretical difficulties. First, there is no underlying theoretical model for the distribution of the X_js and, thus, no reason why a premium of the form (16.7) is appropriate and preferable to M. Why not just estimate ξ from a collection of homogeneous policyholders and charge all policyholders the same rate? While there is a practical reason for using (16.7), no model has been presented to suggest that this formula may be appropriate. Consequently, the choice of Z (and hence P_c) is completely arbitrary.

Second, even if (16.7) were appropriate for a particular model, there is no guidance for the selection of r and p.

Finally, the limited fluctuation approach does not examine the difference between ξ and M. When (16.7) is employed, we are essentially stating that the value of M is accurate as a representation of the expected value, given no information about this particular policyholder. However, it is usually the case that M is also an estimate and, therefore, unreliable in itself. The correct credibility question should be "how much more reliable is $\bar{X}$ compared to M?" and not "how reliable is $\bar{X}$?"

In the next chapter, a systematic modeling approach is presented for the claims experience of a particular policyholder that suggests that the past experience of the policyholder is relevant for prospective rate making. Furthermore, the intuitively appealing formula (16.7) is a consequence of this approach, and Z is often obtained from relations of the form (16.8).

16.6 Notes and References

The limited fluctuation approach is discussed by Herzog [52] and Longley-Cook [83]. See also Norberg [94].

16.7 Exercises

16.1 An insurance company has decided to establish its full-credibility requirements for an individual state rate filing. The full-credibility standard is to be set so that the observed total amount of claims underlying the rate filing would be within 5% of the true value with probability 0.95. The claim frequency follows a Poisson distribution and the severity distribution has pdf

$$f(x) = \frac{100 - x}{5,000}, \quad 0 \le x \le 100.$$

Determine the expected number of claims necessary to obtain full credibility using the normal approximation.

16.2 For a particular policyholder, the past total claims experience is given by $X_1, \ldots, X_n$, where the X_js are i.i.d. compound random variables with Poisson parameter λ and gamma

<div align="center">**Table 16.1**   The data for Exercise 16.3.</div>

Year	1	2	3
Claims	475	550	400

claim size distribution with pdf

$$f_Y(y) = \frac{y^{\alpha-1}e^{-y/\beta}}{\Gamma(\alpha)\beta^\alpha}, \quad y > 0.$$

You also know the following:

1. The credibility factor based on number of claims is 0.9.
2. The expected claim size $\alpha\beta = 100$.
3. The credibility factor based on total claims is 0.8.

Determine α and β.

16.3 For a particular policyholder, the manual premium is 600 per year. The past claims experience is given in Table 16.1. Assess whether full or partial credibility is appropriate and determine the net premium for next year's claims assuming the normal approximation. Use $r = 0.05$ and $p = 0.9$.

16.4 Redo Example 16.2 assuming that X_j is a compound negative binomial distribution rather than compound Poisson.

16.5 (*) The total number of claims for a group of insureds is Poisson with mean λ. Determine the value of λ such that the observed number of claims will be within 3% of λ with a probability of 0.975 using the normal approximation.

16.6 (*) An insurance company is revising rates based on old data. The expected number of claims for full credibility is selected so that observed total claims will be within 5% of the true value 90% of the time. Individual claim amounts have pdf $f(x) = 1/200,000$, $0 < x < 200,000$, and the number of claims has the Poisson distribution. The recent experience consists of 1,082 claims. Determine the credibility, Z, to be assigned to the recent experience. Use the normal approximation.

16.7 (*) The average claim size for a group of insureds is 1,500, with a standard deviation of 7,500. Assume that claim counts have the Poisson distribution. Determine the expected number of claims so that the total loss will be within 6% of the expected total loss with probability 0.90.

16.8 (*) A group of insureds had 6,000 claims and a total loss of 15,600,000. The prior estimate of the total loss was 16,500,000. Determine the limited fluctuation credibility estimate of the total loss for the group. Use the standard for full credibility determined in Exercise 16.7.

16.9 (*) The full-credibility standard is set so that the total number of claims is within 5% of the true value with probability p. This standard is 800 claims. The standard is then

altered so that the total cost of claims is to be within 10% of the true value with probability p. The claim frequency has a Poisson distribution and the claim severity distribution has pdf $f(x) = 0.0002(100 - x)$, $0 < x < 100$. Determine the expected number of claims necessary to obtain full credibility under the new standard.

16.10 (*) A standard for full credibility of 1,000 claims has been selected so that the actual pure premium will be within 10% of the expected pure premium 95% of the time. The number of claims has the Poisson distribution. Determine the coefficient of variation of the severity distribution.

16.11 (*) For a group of insureds, you are given the following information:

1. The prior estimate of expected total losses is 20,000,000.

2. The observed total losses are 25,000,000.

3. The observed number of claims is 10,000.

4. The number of claims required for full credibility is 17,500.

Determine the credibility estimate of the group's expected total losses based on all the given information. Use the credibility factor that is appropriate if the goal is to estimate the expected number of losses.

16.12 (*) A full-credibility standard is determined so that the total number of claims is within 5% of the expected number with probability 98%. If the same expected number of claims for full credibility is applied to the total cost of claims, the actual total cost would be within $100K\%$ of the expected cost with 95% probability. Individual claims have severity pdf $f(x) = 2.5x^{-3.5}$, $x > 1$, and the number of claims has a Poisson distribution. Determine K.

16.13 (*) The number of claims has a Poisson distribution. The number of claims and the claim severity are independent. Individual claim amounts can be for 1, 2, or 10, with probabilities 0.5, 0.3, and 0.2, respectively. Determine the expected number of claims needed so that the total cost of claims is within 10% of the expected cost with 90% probability.

16.14 (*) The number of claims has a Poisson distribution. The coefficient of variation of the severity distribution is 2. The standard for full credibility in estimating total claims is 3,415. With this standard, the observed pure premium will be within $100k\%$ of the expected pure premium 95% of the time. Determine k.

16.15 (*) You are given the following:

1. $P = $ prior estimate of pure premium for a particular class of business.

2. $O = $ observed pure premium during the latest experience period for the same class of business.

3. $R = $ revised estimate of pure premium for the same class following the observations.

4. F = number of claims required for full credibility of the pure premium.

Express the observed number of claims as a function of these four items.

16.16 (*) A company's standard for full credibility is 2,000 claims when it is assumed that the number of claims follows a Poisson distribution and the total number of claims is to be within 3% of the true value with probability p. The standard is changed so that the total cost of claims is to be within 5% of the true value with the same probability p and a severity distribution that is uniform from 0 to 10,000, and the frequency distribution is Poisson. Determine the expected number of claims necessary to obtain full credibility under the new standard.

17

GREATEST ACCURACY CREDIBILITY

17.1 Introduction

In this and Chapter 18, we consider a model-based approach to the solution of the credibility problem. This approach, referred to as greatest accuracy credibility theory, is the outgrowth of a classic 1967 paper by Bühlmann [19]. Many of the ideas are also found in Whitney [131] and Bailey [9].

    We return to the basic problem. For a particular policyholder, we have observed n exposure units of past claims $\mathbf{X} = (X_1, \ldots, X_n)^T$. We have a manual rate μ (we no longer use M for the manual rate) applicable to this policyholder, but the past experience indicates that it may not be appropriate ($\bar{X} = n^{-1}\left(X_1 + \cdots + X_n\right)$, as well as $\mathrm{E}(X)$, could be quite different from μ). This difference raises the question of whether next year's net premium (per exposure unit) should be based on μ, on $\bar{X}$, or on a combination of the two.

    The insurer needs to consider the following question: Is the policyholder really different from what has been assumed in the calculation of μ, or is it just random chance that is responsible for the difference between μ and $\bar{X}$?

    While it is difficult to definitively answer that question, it is clear that no underwriting system is perfect. The manual rate μ has presumably been obtained by (a) evaluation of

the underwriting characteristics of the policyholder and (b) assignment of the rate on the basis of inclusion of the policyholder in a rating class. Such a class should include risks with similar underwriting characteristics. In other words, the rating class is viewed as homogeneous with respect to the underwriting characteristics used. Surely, not all risks in the class are truly homogeneous, however. No matter how detailed the underwriting procedure, there still remains some heterogeneity with respect to risk characteristics within the rating class (good and bad risks, relatively speaking).

Thus, it is possible that the given policyholder may be different from what has been assumed. If this is the case, how should an appropriate rate for the policyholder be determined?

To proceed, let us assume that the risk level of each policyholder in the rating class may be characterized by a risk parameter θ (possibly vector valued), but that the value of θ varies by policyholder. This assumption allows us to quantify the differences between policyholders with respect to the risk characteristics. Because all observable underwriting characteristics have already been used, θ may be viewed as representative of the residual, unobserved factors that affect the risk level. Consequently, we shall assume the existence of θ, but we shall further assume that it is not observable and that we can never know its true value.

Because θ varies by policyholder, there is a probability distribution with pf $\pi(\theta)$ of these values across the rating class. Thus, if θ is a scalar parameter, the cumulative distribution function $\pi(\theta)$ may be interpreted as the proportion of policyholders in the rating class with risk parameter Θ less than or equal to θ. (In statistical terms, Θ is a random variable with distribution function $\pi(\theta) = \Pr(\Theta \leq \theta)$.) Stated another way, $\pi(\theta)$ represents the probability that a policyholder picked at random from the rating class has a risk parameter less than or equal to θ (to accommodate the possibility of new insureds, we slightly generalize the "rating class" interpretation to include the population of all potential risks, whether insured or not).

While the θ value associated with an individual policyholder is not (and cannot be) known, we assume (for this chapter) that $\pi(\theta)$ is known. That is, the structure of the risk characteristics within the population is known. This assumption can be relaxed, and we shall decide later how to estimate the relevant characteristics of $\pi(\theta)$.

Because risk levels vary within the population, it is clear that the experience of the policyholder varies in a systematic way with θ. Imagine that the experience of a policyholder picked (at random) from the population arises from a two-stage process. First, the risk parameter θ is selected from the distribution $\pi(\theta)$. Then the claims or losses X arise from the conditional distribution of X given θ, $f_{X|\Theta}(x|\theta)$. Thus the experience varies with θ via the distribution given the risk parameter θ. The distribution of claims thus differs from policyholder to policyholder to reflect the differences in the risk parameters.

■ **EXAMPLE 17.1**

Consider a rating class for automobile insurance, where θ represents the expected number of claims for a policyholder with risk parameter θ. To accommodate the variability in claims incidence, we assume that the values of θ vary across the rating class. Relatively speaking, the good drivers are those with small values of θ, whereas the poor drivers are those with larger values of θ. A convenient and often reasonable assumption is that the number of claims for a policyholder with risk parameter θ is Poisson distributed with mean θ. Further assume that the random variable Θ is gamma distributed with parameters α and β. Suppose it is known that the average

Table 17.1 The probabilities for Example 17.2.

x	$\Pr(X = x|\Theta = G)$	$\Pr(X = x|\Theta = B)$	θ	$\Pr(\Theta = \theta)$
0	0.7	0.5	G	0.75
1	0.2	0.3	B	0.25
2	0.1	0.2		

number of expected claims for this rating class is 0.15 [$E(\Theta) = 0.15$], and 95% of the policyholders have expected claims between 0.10 and 0.20. Determine α and β.

Assuming the normal approximation to the gamma, where it is known that 95% of the probability lies within about two standard deviations of the mean, it follows that Θ has standard deviation 0.025. Thus $E(\Theta) = \alpha\beta = 0.15$ and $Var(\Theta) = \alpha\beta^2 = (0.025)^2$. Solving for α and β yields $\beta = 1/240 = 0.004167$ and $\alpha = 36$.[1] □

■ **EXAMPLE 17.2**

There are two types of driver. Good drivers make up 75% of the population and in one year have zero claims with probability 0.7, one claim with probability 0.2, and two claims with probability 0.1. Bad drivers make up the other 25% of the population and have zero, one, or two claims with probabilities 0.5, 0.3, and 0.2, respectively. Describe this process and how it relates to an unknown risk parameter.

When a driver buys our insurance, we do not know if the individual is a good or bad driver. So the risk parameter Θ can be one of two values. We can set $\Theta = G$ for good drivers and $\Theta = B$ for bad drivers. The probability model for the number of claims, X, and risk parameter Θ is given in Table 17.1. □

■ **EXAMPLE 17.3**

The amount of a claim has an exponential distribution with mean $1/\Theta$. Among the class of insureds and potential insureds, the parameter Θ varies according to the gamma distribution with $\alpha = 4$ and scale parameter $\beta = 0.001$. Provide a mathematical description of this model.

For claims,

$$f_{X|\Theta}(x|\theta) = \theta e^{-\theta x}, \quad x, \theta > 0,$$

and for the risk parameter,

$$\pi_\Theta(\theta) = \frac{\theta^3 e^{-1,000\theta} 1,000^4}{6}, \quad \theta > 0.$$

□

---

[1] The exact solution can be determined numerically as $\beta = 0.004408$ and $\alpha = 34.03$.

17.2 Conditional Distributions and Expectation

The formulation of the problem just presented involves the use of conditional distributions, given the risk parameter θ of the insured. Subsequent analyses of mathematical models of this nature will be seen to require a good working knowledge of conditional distributions and conditional expectation. A discussion of these topics is now presented.

Much of the material is of a review nature and, hence, may be quickly glossed over if you have a good background in probability. Nevertheless, there may be some material not seen before, and so this section should not be completely ignored.

Suppose that X and Y are two random variables with joint probability function (pf) or probability density function (pdf)[2] $f_{X,Y}(x, y)$ and marginal pfs $f_X(x)$ and $f_Y(y)$, respectively. The conditional pf of X given that $Y = y$ is

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

If X and Y are discrete random variables, then $f_{X|Y}(x|y)$ is the conditional probability of the event $X = x$ under the hypothesis that $Y = y$. If X and Y are continuous, then $f_{X|Y}(x|y)$ may be interpreted as a definition. When X and Y are independent random variables,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y),$$

and, in this case,

$$f_{X|Y}(x|y) = f_X(x).$$

We observe that the conditional and marginal distributions of X are identical.

Note that

$$f_{X,Y}(x, y) = f_{X|Y}(x|y)f_Y(y),$$

demonstrating that joint distributions may be constructed from products of conditional and marginal distributions. Because the marginal distribution of X may be obtained by integrating (or summing) y out of the joint distribution,

$$f_X(x) = \int f_{X,Y}(x, y)\, dy,$$

we find that

$$f_X(x) = \int f_{X|Y}(x|y)f_Y(y)\, dy. \tag{17.1}$$

Formula (17.1) has an interesting interpretation as a mixed distribution (see Section 5.2.4). Assume that the conditional distribution $f_{X|Y}(x|y)$ is one of the usual parametric distributions, where y is the realization of a random parameter Y with distribution $f_Y(y)$. Section 6.3 shows that if, given $\Theta = \theta$, X has a Poisson distribution with mean θ and Θ has a gamma distribution, then the marginal distribution of X will be negative binomial. Also, Example 5.5 shows that if $X|\Theta$ has a normal distribution with mean Θ and variance v and Θ has a normal distribution with mean μ and variance a, then the marginal distribution of X is normal with mean μ and variance $a + v$.

---

[2]When it is unclear or when the random variable may be continuous, discrete, or a mixture of the two, the term *probability function* and its abbreviation pf are used. The term *probability density function* and its abbreviation pdf are used only when the random variable is known to be continuous.

Note that the roles of X and Y can be interchanged, yielding

$$f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x),$$

because both sides of this equation equal the joint distribution of X and Y. Division by $f_Y(y)$ yields Bayes' theorem, namely

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

We now turn our attention to conditional expectation. Consider the conditional pf of X given that $Y = y$, $f_{X|Y}(x|y)$. Clearly, this is a valid probability distribution, and its mean is denoted by

$$E(X|Y = y) = \int x\, f_{X|Y}(x|y)\, dx, \tag{17.2}$$

with the integral replaced by a sum in the discrete case.[3] Clearly, (17.2) is a function of y, and it is often of interest to view this conditional expectation as a random variable obtained by replacing y by Y in the right-hand side of (17.2). Thus we can write $E(X|Y)$ instead of the left-hand side of (17.2), and so $E(X|Y)$ is itself a random variable because it is a function of the random variable Y. The expectation of $E(X|Y)$ is given by

$$E[E(X|Y)] = E(X). \tag{17.3}$$

Equation (17.3) can be proved using (17.2) as follows:

$$\begin{aligned}
E[E(X|Y)] &= \int E(X|Y = y)f_Y(y)\, dy \\
&= \int \int xf_{X|Y}(x|y)\, dx f_Y(y)\, dy \\
&= \int x \int f_{X|Y}(x|y)f_Y(y)\, dy\, dx \\
&= \int xf_X(x)\, dx \\
&= E(X),
\end{aligned}$$

with a similar derivation in the discrete case.

■ **EXAMPLE 17.4**

Derive the mean of the negative binomial distribution by conditional expectation, recalling that, if $X|\Theta \sim \text{Poisson}(\Theta)$ and $\Theta \sim \text{gamma}(\alpha, \beta)$, then $X \sim$ negative binomial with $r = \alpha$ and $\beta = \beta$.

We have
$$E(X|\Theta) = \Theta,$$

and so
$$E(X) = E[E(X|\Theta)] = E(\Theta).$$

---

[3] All formulas assume that the integral (or sum) exists.

From Appendix A, the mean of the gamma distribution of Θ is $\alpha\beta$, and so $E(X) = \alpha\beta$.

$\square$

It is often convenient to replace X by an arbitrary function $h(X, Y)$ in (17.2), yielding the more general definition

$$E[h(X, Y)|Y = y] = \int h(x, y) f_{X|Y}(x|y) \, dx.$$

Similarly, $E[h(X, Y)|Y]$ is the conditional expectation viewed as a random variable that is a function of Y. Then, (17.3) generalizes to

$$E\{E[h(X, Y)|Y]\} = E[h(X, Y)]. \tag{17.4}$$

To see (17.4), note that

$$
\begin{aligned}
E\{E[h(X, Y)|Y]\} &= \int E[h(X, Y)|Y = y] f_Y(y) \, dy \\
&= \iint h(x, y) f_{X|Y}(x|y) \, dx \, f_Y(y) \, dy \\
&= \iint h(x, y)[f_{X|Y}(x|y) f_Y(y)] \, dx \, dy \\
&= \iint h(x, y) f_{X,Y}(x, y) \, dx \, dy \\
&= E[h(X, Y)].
\end{aligned}
$$

If we choose $h(X, Y) = [X - E(X|Y)]^2$, then its expected value, based on the conditional distribution of X given Y, is the variance of this conditional distribution,

$$\text{Var}(X|Y) = E\{[X - E(X|Y)]^2|Y\}. \tag{17.5}$$

Clearly, (17.5) is a function of the random variable Y.

It is instructive now to analyze the variance of X where X and Y are two random variables. To begin, note that (17.5) may be written as

$$\text{Var}(X|Y) = E(X^2|Y) - [E(X|Y)]^2.$$

Thus,

$$
\begin{aligned}
E[\text{Var}(X|Y)] &= E\{E(X^2|Y) - [E(X|Y)]^2\} \\
&= E[E(X^2|Y)] - E\{[E(X|Y)]^2\} \\
&= E(X^2) - E\{[E(X|Y)]^2\}.
\end{aligned}
$$

Also, because $\text{Var}[h(Y)] = E\{[h(Y)]^2\} - \{E[h(Y)]\}^2$, we may use $h(Y) = E(X|Y)$ to obtain

$$
\begin{aligned}
\text{Var}[E(X|Y)] &= E\{[E(X|Y)]^2\} - \{E[E(X|Y)]\}^2 \\
&= E\{[E(X|Y)]^2\} - [E(X)]^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathrm{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathrm{E}(X|Y)] &= \mathrm{E}(X^2) - \mathrm{E}\{[\mathrm{E}(X|Y)]^2\} \\
&\quad + \mathrm{E}\{[\mathrm{E}(X|Y)]^2\} - [\mathrm{E}(X)]^2 \\
&= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \\
&= \mathrm{Var}(X).
\end{aligned}
$$

Finally, we have established the important formula

$$
\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathrm{E}(X|Y)]. \tag{17.6}
$$

Formula (17.6) states that the variance of X is composed of the sum of two parts: the mean of the conditional variance plus the variance of the conditional mean.

■ **EXAMPLE 17.5**

Derive the variance of the negative binomial distribution.

The Poisson distribution has equal mean and variance, that is,

$$
\mathrm{E}(X|\Theta) = \mathrm{Var}(X|\Theta) = \Theta,
$$

and so, from (17.6),

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}[\mathrm{Var}(X|\Theta)] + \mathrm{Var}[\mathrm{E}(X|\Theta)] \\
&= \mathrm{E}(\Theta) + \mathrm{Var}(\Theta).
\end{aligned}
$$

Because Θ itself has a gamma distribution with parameters α and β, $\mathrm{E}(\Theta) = \alpha\beta$ and $\mathrm{Var}(\Theta) = \alpha\beta^2$. Thus the variance of the negative binomial distribution is

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(\Theta) + \mathrm{Var}(\Theta) \\
&= \alpha\beta + \alpha\beta^2 \\
&= \alpha\beta(1 + \beta).
\end{aligned}
$$

■ **EXAMPLE 17.6**

It is shown in Example 5.5 that, if $X|\Theta$ is normally distributed with mean Θ and variance v, where Θ is itself normally distributed with mean μ and variance a, then X is (unconditionally) normally distributed with mean μ and variance $a + v$. Use (17.3) and (17.6) to obtain the mean and variance of X directly.

For the mean, we have

$$
\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X|\Theta)] = \mathrm{E}(\Theta) = \mu,
$$

and for the variance, we obtain

$$
\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}[\mathrm{Var}(X|\Theta)] + \mathrm{Var}[\mathrm{E}(X|\Theta)] \\
&= \mathrm{E}(v) + \mathrm{Var}(\Theta) \\
&= v + a,
\end{aligned}
$$

because v is a constant.

17.3 The Bayesian Methodology

Continue to assume that the distribution of the risk characteristics in the population may be represented by $\pi(\theta)$, and the experience of a particular policyholder with risk parameter θ arises from the conditional distribution $f_{X|\Theta}(x|\theta)$ of claims or losses, given θ.

We now return to the problem introduced in Section 16.2. That is, for a particular policyholder, we have observed $\mathbf{X} = \mathbf{x}$, where $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{x} = (x_1, \ldots, x_n)^T$, and are interested in setting a rate to cover X_{n+1}. We assume that the risk parameter associated with the policyholder is θ (which is unknown). Furthermore, the experience of the policyholder corresponding to different exposure periods is assumed to be independent. In statistical terms, conditional on θ, the claims or losses $X_1, \ldots, X_n, X_{n+1}$ are independent (although not necessarily identically distributed).

Let X_j have conditional pf

$$f_{X_j|\Theta}(x_j|\theta), \quad j = 1, \ldots, n, n+1.$$

Note that, if the X_js are identically distributed (conditional on $\Theta = \theta$), then $f_{X_j|\Theta}(x_j|\theta)$ does not depend on j. Ideally, we are interested in the conditional distribution of X_{n+1}, given $\Theta = \theta$, in order to predict the claims experience X_{n+1} of the same policyholder (whose value of θ has been assumed not to have changed). If we knew θ, we could use $f_{X_{n+1}|\Theta}(x_{n+1}|\theta)$. Unfortunately, we do not know θ, but we do know $\mathbf{x}$ for the same policyholder. The obvious next step is to condition on $\mathbf{x}$ rather than θ. Consequently, we calculate the conditional distribution of X_{n+1} given $\mathbf{X} = \mathbf{x}$, termed the **predictive distribution**, as defined in Chapter 13.

The predictive distribution of X_{n+1} given $\mathbf{X} = \mathbf{x}$ is the relevant distribution for risk analysis, management, and decision making. It combines the uncertainty about the claims losses with that of the parameters associated with the risk process.

Here, we repeat the development in Chapter 13, noting that if Θ has a discrete distribution, the integrals are replaced by sums. Because the X_js are independent conditional on $\Theta = \theta$, we have

$$f_{\mathbf{X},\Theta}(\mathbf{x}, \theta) = f(x_1, \ldots, x_n|\theta)\pi(\theta) = \left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta).$$

The joint distribution of $\mathbf{X}$ is thus the marginal distribution obtained by integrating θ out, that is,

$$f_{\mathbf{X}}(\mathbf{x}) = \int \left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta)\,d\theta. \tag{17.7}$$

Similarly, the joint distribution of $X_1, \ldots, X_{n+1}$ is the right-hand side of (17.7) with n replaced by $n+1$ in the product. Finally, the conditional density of X_{n+1} given $\mathbf{X} = \mathbf{x}$ is the joint density of $(X_1, \ldots, X_{n+1})$ divided by that of $\mathbf{X}$, namely

$$f_{X_{n+1}|\mathbf{X}}(x_{n+1}|\mathbf{x}) = \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \int \left[\prod_{j=1}^{n+1} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta)\,d\theta. \tag{17.8}$$

There is a hidden mathematical structure underlying (17.8) that may be exploited. The posterior density of Θ given $\mathbf{X}$ is

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X},\Theta}(\mathbf{x}, \theta)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{1}{f_{\mathbf{X}}(\mathbf{x})} \left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta). \tag{17.9}$$

In other words, $\left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta) = \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$, and substitution in the numerator of (17.8) yields

$$f_{X_{n+1}|\mathbf{X}}(x_{n+1}|\mathbf{x}) = \int f_{X_{n+1}|\Theta}(x_{n+1}|\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\,d\theta. \tag{17.10}$$

Equation (17.10) provides the additional insight that the conditional distribution of X_{n+1} given $\mathbf{X}$ may be viewed as a mixture distribution, with the mixing distribution being the posterior distribution $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$.

The posterior distribution combines and summarizes the information about θ contained in the prior distribution and the likelihood, and consequently (17.10) reflects this information. As noted in Theorem 13.18, the posterior distribution admits a convenient form when the likelihood is derived from the linear exponential family and $\pi(\theta)$ is the natural conjugate prior. When both are in place, there is an easy method to evaluate the conditional distribution of X_{n+1} given $\mathbf{X}$.

■ **EXAMPLE 17.7**

(*Example 17.2 continued*) For a particular policyholder, suppose that we have observed $x_1 = 0$ and $x_2 = 1$. Determine the predictive distribution of $X_3|X_1 = 0, X_2 = 1$ and the posterior distribution of $\Theta|X_1 = 0, X_2 = 1$.

From (17.7), the marginal probability is

$$f_{\mathbf{X}}(0, 1) = \sum_{\theta} f_{X_1|\Theta}(0|\theta)f_{X_2|\Theta}(1|\theta)\pi(\theta)$$
$$= 0.7(0.2)(0.75) + 0.5(0.3)(0.25)$$
$$= 0.1425.$$

Similarly, the joint probability of all three variables is

$$f_{\mathbf{X},X_3}(0, 1, x_3) = \sum_{\theta} f_{X_1|\Theta}(0|\theta)f_{X_2|\Theta}(1|\theta)f_{X_3|\Theta}(x_3|\theta)\pi(\theta).$$

Thus,

$$f_{\mathbf{X},X_3}(0, 1, 0) = 0.7(0.2)(0.7)(0.75) + 0.5(0.3)(0.5)(0.25) = 0.09225,$$
$$f_{\mathbf{X},X_3}(0, 1, 1) = 0.7(0.2)(0.2)(0.75) + 0.5(0.3)(0.3)(0.25) = 0.03225,$$
$$f_{\mathbf{X},X_3}(0, 1, 2) = 0.7(0.2)(0.1)(0.75) + 0.5(0.3)(0.2)(0.25) = 0.01800.$$

The predictive distribution is then

$$f_{X_3|\mathbf{X}}(0|0, 1) = \frac{0.09225}{0.1425} = 0.647368,$$
$$f_{X_3|\mathbf{X}}(1|0, 1) = \frac{0.03225}{0.1425} = 0.226316,$$
$$f_{X_3|\mathbf{X}}(2|0, 1) = \frac{0.01800}{0.1425} = 0.126316.$$

The posterior probabilities are, from (17.9),

$$\pi(G|0,1) = \frac{f(0|G)f(1|G)\pi(G)}{f(0,1)} = \frac{0.7(0.2)(0.75)}{0.1425} = 0.736842,$$

$$\pi(B|0,1) = \frac{f(0|B)f(1|B)\pi(B)}{f(0,1)} = \frac{0.5(0.3)(0.25)}{0.1425} = 0.263158.$$

(From this point forward, the subscripts on f and π are dropped unless needed for clarity.) The predictive probabilities could also have been obtained using (17.10). This method is usually simpler from a computational viewpoint:

$$f(0|0,1) = \sum_\theta f(0|\theta)\pi(\theta|0,1)$$

$$= 0.7(0.736842) + 0.5(0.263158) = 0.647368,$$

$$f(1|0,1) = 0.2(0.736842) + 0.3(0.263158) = 0.226316,$$

$$f(2|0,1) = 0.1(0.736842) + 0.2(0.263158) = 0.126316,$$

which matches the previous calculations. $\square$

■ **EXAMPLE 17.8**

(*Example 17.3 continued*) Suppose that a person had claims of 100, 950, and 450. Determine the predictive distribution of the fourth claim and the posterior distribution of Θ.

The marginal density at the observed values is

$$f(100,950,450) = \int_0^\infty \theta e^{-100\theta}\theta e^{-950\theta}\theta e^{-450\theta}\frac{1,000^4}{6}\theta^3 e^{-1,000\theta}\,d\theta$$

$$= \frac{1,000^4}{6}\int_0^\infty \theta^6 e^{-2,500\theta}\,d\theta = \frac{1,000^4}{6}\frac{720}{2,500^7}.$$

Similarly,

$$f(100,950,450,x_4) = \int_0^\infty \theta e^{-100\theta}\theta e^{-950\theta}\theta e^{-450\theta}\theta e^{-\theta x_4}\frac{1,000^4}{6}\theta^3 e^{-1,000\theta}\,d\theta$$

$$= \frac{1,000^4}{6}\int_0^\infty \theta^7 e^{-(2,500+x_4)\theta}\,d\theta$$

$$= \frac{1,000^4}{6}\frac{5,040}{(2,500+x_4)^8}.$$

Then, the predictive density is

$$f(x_4|100,950,450) = \frac{\dfrac{1,000^4}{6}\dfrac{5,040}{(2,500+x_4)^8}}{\dfrac{1,000^4}{6}\dfrac{720}{2,500^7}} = \frac{7(2,500)^7}{(2,500+x_4)^8},$$

which is a Pareto density with parameters 7 and 2,500.

For the posterior distribution, we take a shortcut. The denominator is an integral that produces a number and can be ignored for now. The numerator satisfies

$$\pi(\theta|100, 950, 450) \propto \theta e^{-100\theta} \theta e^{-950\theta} \theta e^{-450\theta} \frac{1{,}000^4}{6} \theta^3 e^{-1{,}000\theta},$$

which was the term to be integrated in the calculation of the marginal density. Because there are constants in the denominator that have been ignored, we might as well ignore constants in the numerator. Only multiplicative terms involving the variable (θ in this case) need to be retained. Then,

$$\pi(\theta|100, 950, 450) \propto \theta^6 e^{-2{,}500\theta}.$$

We could integrate this expression to determine the constant needed to make this a density function (i.e. make the integral equal 1). But we recognize this function as that of a gamma distribution with parameters 7 and $1/2{,}500$. Therefore,

$$\pi(\theta|100, 950, 450) = \frac{\theta^6 e^{-2{,}500\theta} 2{,}500^7}{\Gamma(7)}.$$

Then, the predictive density can be alternatively calculated from

$$
\begin{aligned}
f(x_4|100, 950, 450) &= \int_0^\infty \theta e^{-\theta x_4} \frac{\theta^6 e^{-2{,}500\theta} 2{,}500^7}{\Gamma(7)} d\theta \\
&= \frac{2{,}500^7}{6!} \int_0^\infty \theta^7 e^{-(2{,}500+x_4)\theta} d\theta \\
&= \frac{2{,}500^7}{6!} \frac{7!}{(2{,}500 + x_4)^8},
\end{aligned}
$$

matching the answer previously obtained.   □

Note that the posterior distribution is of the same type (gamma) as the prior distribution. The concept of a conjugate prior distribution is introduced in Section 13.3. This result also implies that $X_{n+1}|\mathbf{x}$ is a mixture distribution with a simple mixing distribution, facilitating evaluation of the density of $X_{n+1}|\mathbf{x}$. Further examples of this idea are found in the exercises at the end of this section.

To return to the original problem, we have observed $\mathbf{X} = \mathbf{x}$ for a particular policyholder and we wish to predict X_{n+1} (or its mean). An obvious choice would be the hypothetical mean (or individual premium)

$$\mu_{n+1}(\theta) = \mathrm{E}(X_{n+1}|\Theta = \theta) = \int x_{n+1} f_{X_{n+1}|\Theta}(x_{n+1}|\theta) \, dx_{n+1} \qquad (17.11)$$

if we knew θ. Note that replacement of θ by Θ in (17.11) yields, on taking the expectation,

$$\mu_{n+1} = \mathrm{E}(X_{n+1}) = \mathrm{E}[\mathrm{E}(X_{n+1}|\Theta)] = \mathrm{E}[\mu_{n+1}(\Theta)],$$

so that the pure, or collective, premium is the mean of the hypothetical means. This is the premium we would use if we knew nothing about the individual. It does not depend on the individual's risk parameter, θ; nor does it use $\mathbf{x}$, the data collected from the individual.

Because θ is unknown, the best we can do is try to use the data, which suggest the use of the Bayesian premium (the mean of the predictive distribution):

$$E(X_{n+1}|\mathbf{X} = \mathbf{x}) = \int x_{n+1} f_{X_{n+1}|\mathbf{X}}(x_{n+1}|\mathbf{x}) \, dx_{n+1}. \tag{17.12}$$

A computationally more convenient form is

$$E\left(X_{n+1}|\mathbf{X} = \mathbf{x}\right) = \int \mu_{n+1}(\theta) \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \, d\theta. \tag{17.13}$$

In other words, the Bayesian premium is the expected value of the hypothetical means, with expectation taken over the posterior distribution $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$. Recall that in the discrete case, the integrals are replaced by sums. To prove (17.13), we see from (17.10) that

$$\begin{aligned}
E(X_{n+1}|\mathbf{X} = \mathbf{x}) &= \int x_{n+1} f_{X_{n+1}|\mathbf{X}}(x_{n+1}|\mathbf{x}) \, dx_{n+1} \\
&= \int x_{n+1} \left[\int f_{X_{n+1}|\Theta}(x_{n+1}|\theta) \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \, d\theta \right] dx_{n+1} \\
&= \int \left[\int x_{n+1} f_{X_{n+1}|\Theta}(x_{n+1}|\theta) \, dx_{n+1} \right] \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \, d\theta \\
&= \int \mu_{n+1}(\theta) \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \, d\theta.
\end{aligned}$$

■ **EXAMPLE 17.9**

(*Example 17.7 continued*) Determine the Bayesian premium using both (17.12) and (17.13).

The (unobservable) hypothetical means are

$$\mu_3(G) = (0)(0.7) + 1(0.2) + 2(0.1) = 0.4,$$
$$\mu_3(B) = (0)(0.5) + 1(0.3) + 2(0.2) = 0.7.$$

If, as in Example 17.7, we have observed $x_1 = 0$ and $x_2 = 1$, we have the Bayesian premium obtained directly from (17.12):

$$E(X_3|0, 1) = 0(0.647368) + 1(0.226316) + 2(0.126316) = 0.478948.$$

The (unconditional) pure premium is

$$\mu_3 = E\left(X_3\right) = \sum_\theta \mu_3(\theta)\pi(\theta) = (0.4)(0.75) + (0.7)(0.25) = 0.475.$$

To verify (17.13) with $x_1 = 0$ and $x_2 = 1$, we have the posterior distribution $\pi(\theta|0, 1)$ from Example 17.7. Thus, (17.13) yields

$$E(X_3|0, 1) = 0.4(0.736842) + 0.7(0.263158) = 0.478947,$$

with the difference being due to rounding. In general, the latter approach utilizing (17.13) is simpler than the direct approach using the conditional distribution of $X_{n+1}|\mathbf{X} = \mathbf{x}$. □

As expected, the revised value based on two observations is between the prior value (0.475) based on no data and the value based only on the data (0.5).

■ **EXAMPLE 17.10**

(*Example 17.8 continued*) Determine the Bayesian premium.

From Example 17.8, we have $\mu_4(\theta) = \theta^{-1}$. Then, (17.13) yields

$$E(X_4|100, 950, 450) = \int_0^\infty \theta^{-1} \frac{\theta^6 e^{-2,500\theta} 2,500^7}{720} d\theta$$

$$= \frac{2,500^7}{720} \frac{120}{2,500^6} = 416.67.$$

This result could also have been obtained from the formula for the moments of the gamma distribution in Appendix A. From the prior distribution,

$$\mu = E(\Theta^{-1}) = \frac{1,000}{3} = 333.33$$

and once again the Bayesian estimate is between the prior estimate and one based solely on the data (the sample mean of 500).
From (17.12),

$$E(X_4|100, 950, 450) = \frac{2,500}{6} = 416.67,$$

the mean of the predictive Pareto distribution.   □

■ **EXAMPLE 17.11**

Generalize the result of Example 17.10 for an arbitrary sample size of n and an arbitrary prior gamma distribution with parameters α and β, where β is the reciprocal of the usual scale parameter.

The posterior distribution can be determined from

$$\pi(\theta|\mathbf{x}) \propto \left(\prod_{j=1}^n \theta e^{-\theta x_j} \right) \frac{\theta^{\alpha-1} e^{-\beta\theta} \beta^\alpha}{\Gamma(\alpha)}$$

$$\propto \theta^{n+\alpha-1} e^{-(\Sigma x_j + \beta)\theta}.$$

The second line follows because the posterior density is a function of θ and thus all multiplicative terms not involving θ may be dropped. Rather than perform the integral to determine the constant, we recognize that the posterior distribution is gamma with first parameter $n + \alpha$ and scale parameter $(\Sigma x_j + \beta)^{-1}$. The Bayes estimate of X_{n+1} is the expected value of Θ^{-1} using the posterior distribution. It is

$$\frac{\Sigma x_j + \beta}{n + \alpha - 1} = \frac{n}{n + \alpha - 1} \bar{x} + \frac{\alpha - 1}{n + \alpha - 1} \frac{\beta}{\alpha - 1}.$$

Note that the estimate is a weighted average of the observed values and the unconditional mean. This formula is of the credibility-weighted type (16.7).   □

Example 17.12 is one where the random variables do not have identical distributions.

■ **EXAMPLE 17.12**

Suppose that the number of claims N_j in year j for a group policyholder with (unknown) risk parameter θ and m_j individuals in the group is Poisson distributed with mean $m_j\theta$, that is, for $j = 1, \dots, n$,

$$\Pr(N_j = x|\Theta = \theta) = \frac{(m_j\theta)^x e^{-m_j\theta}}{x!}, \quad x = 0, 1, 2, \dots.$$

This result would be the case if, per individual, the number of claims were independently Poisson distributed with mean θ. Determine the Bayesian expected number of claims for the m_{n+1} individuals to be insured in year $n + 1$.

With these assumptions, the average number of claims per individual in year j is

$$X_j = \frac{N_j}{m_j}, \quad j = 1, \dots, n.$$

Therefore,

$$f_{X_j|\Theta}(x_j|\theta) = \Pr[N_j = m_j x_j|\Theta = \theta].$$

Assume that Θ is gamma distributed with parameters α and β,

$$\pi(\theta) = \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}, \quad \theta > 0.$$

Then, the posterior distribution $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ is proportional (as a function of θ) to

$$\left[\prod_{j=1}^n f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta),$$

which is itself proportional to

$$\left[\prod_{j=1}^n \theta^{m_j x_j} e^{-m_j\theta}\right]\theta^{\alpha-1} e^{-\theta/\beta} = \theta^{\alpha+\sum_{j=1}^n m_j x_j - 1} e^{-\theta\left(\beta^{-1}+\sum_{j=1}^n m_j\right)}.$$

This function is proportional to a gamma density with parameters $\alpha_* = \alpha + \sum_{j=1}^n m_j x_j$ and $\beta_* = (1/\beta + \sum_{j=1}^n m_j)^{-1}$, and so $\Theta|\mathbf{X}$ is also gamma, but with α and β replaced by α_* and β_*, respectively.

Now,

$$E(X_j|\Theta = \theta) = E\left(\frac{1}{m_j}N_j|\Theta = \theta\right) = \frac{1}{m_j}E(N_j|\Theta = \theta) = \theta.$$

Thus $\mu_{n+1}(\theta) = E(X_{n+1}|\Theta = \theta) = \theta$ and $\mu_{n+1} = E(X_{n+1}) = E[\mu_{n+1}(\Theta)] = \alpha\beta$ because Θ is gamma distributed with parameters α and β. From (17.13) and because $\Theta|\mathbf{X}$ is also gamma distributed with parameters α_* and β_*,

$$
\begin{aligned}
E(X_{n+1}|\mathbf{X} = \mathbf{x}) &= \int_0^\infty \mu_{n+1}(\theta)\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \\
&= E[\mu_{n+1}(\Theta)|\mathbf{X} = \mathbf{x}] \\
&= E(\Theta|\mathbf{X} = \mathbf{x}) \\
&= \alpha_*\beta_*.
\end{aligned}
$$

Define the total number of lives observed to be $m = \sum_{j=1}^{n} m_j$.

Then,

$$E(X_{n+1}|\mathbf{X} = \mathbf{x}) = Z\bar{x} + (1 - Z)\mu_{n+1},$$

where $Z = m/(m + \beta^{-1})$, $\bar{x} = m^{-1} \sum_{j=1}^{n} m_j x_j$, and $\mu_{n+1} = \alpha\beta$, again an expression of the form (16.7).

The total Bayesian expected number of claims for m_{n+1} individuals in the group for the next year would be $m_{n+1}E(X_{n+1}|\mathbf{X} = \mathbf{x})$.

The analysis based on i.i.d. Poisson claim counts is obtained with $m_j = 1$. Then, $X_j \equiv N_j$ for $j = 1, 2, \ldots, n$ are independent (given θ) Poisson random variables with mean θ. In this case,

$$E(X_{n+1}|\mathbf{X} = \mathbf{x}) = Z\bar{x} + (1 - Z)\mu,$$

where $Z = n/(n + \beta^{-1})$, $\bar{x} = n^{-1} \sum_{j=1}^{n} x_j$, and $\mu = \alpha\beta$. $\qquad\square$

In each of Examples 17.11 and 17.12, the Bayesian estimate was a weighted average of the sample mean $\bar{x}$ and the pure premium μ_{n+1}. This result is appealing from a credibility standpoint. Furthermore, the credibility factor Z in each case is an increasing function of the number of exposure units. The greater the amount of past data observed, the closer Z is to 1, consistent with our intuition.

17.4 The Credibility Premium

In Section 17.3, a systematic approach is suggested for treatment of the past data of a particular policyholder. Ideally, rather than the pure premium $\mu_{n+1} = E(X_{n+1})$, we would like to charge the individual premium (or hypothetical mean) $\mu_{n+1}(\theta)$, where θ is the (hypothetical) parameter associated with the policyholder. Because θ is unknown, the hypothetical mean is impossible to determine, but we could instead condition on $\mathbf{x}$, the past data from the policyholder. This leads to the Bayesian premium $E(X_{n+1}|\mathbf{x})$.

The major challenge with this approach is that it may be difficult to evaluate the Bayesian premium. Of course, in simple examples such as those in Section 17.3, the Bayesian premium is not difficult to evaluate numerically. But these examples can hardly be expected to capture the essential features of a realistic insurance scenario. More realistic models may well introduce analytic difficulties with respect to evaluation of $E(X_{n+1}|\mathbf{x})$, whether we use (17.12) or (17.13). Often, numerical integration may be required. There are exceptions, such as Examples 17.11 and 17.12.

We now present an alternative suggested by Bühlmann [19] in 1967. Recall the basic problem. We wish to use the conditional distribution $f_{X_{n+1}|\Theta}(x_{n+1}|\theta)$ or the hypothetical mean $\mu_{n+1}(\theta)$ for estimation of next year's claims. Because we have observed $\mathbf{x}$, one suggestion is to approximate $\mu_{n+1}(\theta)$ by a linear function of the past data. (After all, the formula $Z\bar{X} + (1 - Z)\mu$ is of this form.) Thus, let us restrict ourselves to estimators of the form $\alpha_0 + \sum_{j=1}^{n} \alpha_j X_j$, where $\alpha_0, \alpha_1, \ldots, \alpha_n$ need to be chosen. To this end, we choose the αs to minimize expected squared-error loss, that is,

$$Q = E\left\{\left[\mu_{n+1}(\Theta) - \alpha_0 - \sum_{j=1}^{n} \alpha_j X_j\right]^2\right\}, \qquad (17.14)$$

and the expectation is over the joint distribution of $X_1, \ldots, X_n$ and Θ. That is, the squared error is averaged over all possible values of Θ and all possible observations. To minimize Q, we take derivatives. Thus,

$$\frac{\partial Q}{\partial \alpha_0} = E \left\{ 2 \left[\mu_{n+1}(\Theta) - \alpha_0 - \sum_{j=1}^{n} \alpha_j X_j \right] (-1) \right\}.$$

We shall denote by $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ the values of $\alpha_0, \alpha_1, \ldots, \alpha_n$ that minimize (17.14). Then, equating $\partial Q / \partial \alpha_0$ to 0 yields

$$E[\mu_{n+1}(\Theta)] = \tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j E(X_j).$$

But $E(X_{n+1}) = E[E(X_{n+1}|\Theta)] = E[\mu_{n+1}(\Theta)]$, and so $\partial Q / \partial \alpha_0 = 0$ implies that

$$E(X_{n+1}) = \tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j E(X_j). \tag{17.15}$$

Equation (17.15) may be termed the **unbiasedness equation** because it requires that the estimate $\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j$ be unbiased for $E(X_{n+1})$. However, the credibility estimate may be biased as an estimator of $\mu_{n+1}(\theta) = E(X_{n+1}|\theta)$, the quantity we are trying to estimate. This bias will average out over the members of Θ. By accepting this bias, we are able to reduce the overall MSE. For $i = 1, \ldots, n$, we have

$$\frac{\partial Q}{\partial \alpha_i} = E \left\{ 2 \left[\mu_{n+1}(\Theta) - \alpha_0 - \sum_{j=1}^{n} \alpha_j X_j \right] (-X_i) \right\}$$

and setting this expression equal to zero yields

$$E[\mu_{n+1}(\Theta) X_i] = \tilde{\alpha}_0 E(X_i) + \sum_{j=1}^{n} \tilde{\alpha}_j E(X_i X_j).$$

The left-hand side of this equation may be reexpressed as

$$\begin{aligned} E[\mu_{n+1}(\Theta) X_i] &= E\{E[X_i \mu_{n+1}(\Theta)|\Theta]\} \\ &= E\{\mu_{n+1}(\Theta) E[X_i|\Theta]\} \\ &= E[E(X_{n+1}|\Theta) E(X_i|\Theta)] \\ &= E[E(X_{n+1} X_i|\Theta)] \\ &= E(X_i X_{n+1}), \end{aligned}$$

where the second from last step follows by independence of X_i and X_{n+1} conditional on Θ. Thus $\partial Q / \partial \alpha_i = 0$ implies

$$E(X_i X_{n+1}) = \tilde{\alpha}_0 E(X_i) + \sum_{j=1}^{n} \tilde{\alpha}_j E(X_i X_j). \tag{17.16}$$

Next, multiply (17.15) by $E(X_i)$ and subtract from (17.16) to obtain

$$Cov(X_i, X_{n+1}) = \sum_{j=1}^{n} \tilde{\alpha}_j Cov(X_i, X_j), \quad i = 1, \ldots, n. \tag{17.17}$$

Equation (17.15) and the n equations (17.17) together are called the **normal equations**. These equations may be solved for $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ to yield the credibility premium

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j. \tag{17.18}$$

While it is straightforward to express the solution $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ to the normal equations in matrix notation (if the covariance matrix of the X_js is nonsingular), we shall be content with solutions for some special cases.

Note that exactly one of the terms on the right-hand side of (17.17) is a variance term, that is, $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$. The other $n - 1$ terms are true covariance terms.

As an added bonus, the values $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ also minimize

$$Q_1 = \mathrm{E}\left\{ \left[\mathrm{E}(X_{n+1}|\mathbf{X}) - \alpha_0 - \sum_{j=1}^{n} \alpha_j X_j \right]^2 \right\} \tag{17.19}$$

and

$$Q_2 = \mathrm{E}\left[\left(X_{n+1} - \alpha_0 - \sum_{j=1}^{n} \alpha_j X_j \right)^2 \right]. \tag{17.20}$$

To see this, differentiate (17.19) or (17.20) with respect to $\alpha_0, \alpha_1, \ldots, \alpha_n$ and observe that the solutions still satisfy the normal equations (17.15) and (17.17). Thus the credibility premium (17.18) is the best linear estimator of each of the hypothetical mean $\mathrm{E}(X_{n+1}|\Theta)$, the Bayesian premium $\mathrm{E}(X_{n+1}|\mathbf{X})$, and X_{n+1}.

■ **EXAMPLE 17.13**

If $\mathrm{E}(X_j) = \mu$, $\mathrm{Var}(X_j) = \sigma^2$, and, for $i \neq j$, $\mathrm{Cov}(X_i, X_j) = \rho\sigma^2$, where the correlation coefficient ρ satisfies $-1 < \rho < 1$, determine the credibility premium $\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j$.

The unbiasedness equation (17.15) yields

$$\mu = \tilde{\alpha}_0 + \mu \sum_{j=1}^{n} \tilde{\alpha}_j,$$

or

$$\sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}.$$

The n equations (17.17) become, for $i = 1, \ldots, n$,

$$\rho = \sum_{\substack{j=1 \\ j \neq i}}^{n} \tilde{\alpha}_j \rho + \tilde{\alpha}_i,$$

or, stated another way,

$$\rho = \sum_{j=1}^{n} \tilde{\alpha}_j \rho + \tilde{\alpha}_i(1 - \rho), \quad i = 1, \ldots, n.$$

Thus

$$\tilde{\alpha}_i = \frac{\rho \left(1 - \sum_{j=1}^{n} \tilde{\alpha}_j\right)}{1 - \rho} = \frac{\rho \tilde{\alpha}_0}{\mu(1 - \rho)}$$

using the unbiasedness equation. Summation over i from 1 to n yields

$$\sum_{i=1}^{n} \tilde{\alpha}_i = \sum_{j=1}^{n} \tilde{\alpha}_j = \frac{n\rho\tilde{\alpha}_0}{\mu(1 - \rho)},$$

which, combined with the unbiasedness equation, gives an equation for $\tilde{\alpha}_0$, namely

$$1 - \frac{\tilde{\alpha}_0}{\mu} = \frac{n\rho\tilde{\alpha}_0}{\mu(1 - \rho)}.$$

Solving for $\tilde{\alpha}_0$ yields

$$\tilde{\alpha}_0 = \frac{(1 - \rho)\mu}{1 - \rho + n\rho}.$$

Thus,

$$\tilde{\alpha}_j = \frac{\rho\tilde{\alpha}_0}{\mu(1 - \rho)} = \frac{\rho}{1 - \rho + n\rho}.$$

The credibility premium is then

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = \frac{(1 - \rho)\mu}{1 - \rho + n\rho} + \sum_{j=1}^{n} \frac{\rho X_j}{1 - \rho + n\rho}$$

$$= (1 - Z)\mu + Z\bar{X},$$

where $Z = n\rho/(1 - \rho + n\rho)$ and $\bar{X} = n^{-1} \sum_{j=1}^{n} X_j$. Thus, if $0 < \rho < 1$, then $0 < Z < 1$, and the credibility premium is a weighted average of $\mu = E(X_{n+1})$ and $\bar{X}$, that is, the premium is of the form (16.7). □

We now turn to some models that specify the conditional means and variances of $X_j | \Theta$ and, hence, the means $E(X_j)$, variances $\text{Var}(X_j)$, and covariances $\text{Cov}(X_i, X_j)$.

17.5   The Bühlmann Model

The simplest credibility model, the Bühlmann model, specifies that, for each policyholder (conditional on Θ), past losses $X_1, \dots, X_n$ have the same mean and variance and are i.i.d. conditional on Θ.

Thus, define

$$\mu(\theta) = E(X_j | \Theta = \theta)$$

and

$$v(\theta) = \text{Var}(X_j | \Theta = \theta).$$

As discussed previously, $\mu(\theta)$ is referred to as the **hypothetical mean**, whereas $v(\theta)$ is called the **process variance**. Define

$$\mu = E[\mu(\Theta)], \tag{17.21}$$

$$v = E[v(\Theta)], \tag{17.22}$$

and

$$a = \text{Var}[\mu(\Theta)]. \tag{17.23}$$

The quantity μ in (17.21) is the **expected value of the hypothetical means**, v in (17.22) is the **expected value of the process variance**, and a in (17.23) is the **variance of the hypothetical means**. Note that μ is the estimate to use if we have no information about θ (and thus no information about $\mu(\theta)$). It will also be referred to as the **collective premium**.

The mean, variance, and covariance of the X_js may now be obtained. First,

$$\text{E}(X_j) = \text{E}[\text{E}(X_j|\Theta)] = \text{E}[\mu(\Theta)] = \mu. \tag{17.24}$$

Second,

$$\begin{aligned} \text{Var}(X_j) &= \text{E}[\text{Var}(X_j|\Theta)] + \text{Var}[\text{E}(X_j|\Theta)] \\ &= \text{E}[v(\Theta)] + \text{Var}[\mu(\Theta)] \\ &= v + a. \end{aligned} \tag{17.25}$$

Finally, for $i \neq j$,

$$\begin{aligned} \text{Cov}(X_i, X_j) &= \text{E}(X_i X_j) - \text{E}(X_i)\text{E}(X_j) \\ &= \text{E}[\text{E}(X_i X_j|\Theta)] - \mu^2 \\ &= \text{E}[\text{E}(X_i|\Theta)\text{E}(X_j|\Theta)] - \{\text{E}[\mu(\Theta)]\}^2 \\ &= \text{E}\left\{[\mu(\Theta)]^2\right\} - \{\text{E}[\mu(\Theta)]\}^2 \\ &= \text{Var}[\mu(\Theta)] \\ &= a. \end{aligned} \tag{17.26}$$

This result is exactly of the form of Example 17.13 with parameters $\mu, \sigma^2 = v + a$, and $\rho = a/(v + a)$. Thus the credibility premium is

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = Z\bar{X} + (1 - Z)\mu, \tag{17.27}$$

where

$$Z = \frac{n}{n + k} \tag{17.28}$$

and

$$k = \frac{v}{a} = \frac{\text{E}[\text{Var}(X_j|\Theta)]}{\text{Var}[\text{E}(X_j|\Theta)]}. \tag{17.29}$$

The credibility factor Z in (17.28) with k given by (17.29) is referred to as the **Bühlmann credibility factor**. Note that (17.27) is of the form (16.7), and (17.28) is exactly (16.8). Now, however, we know how to obtain k, namely, from (17.29).

Formula (17.27) has many appealing features. First, the credibility premium (17.27) is a weighted average of the sample mean $\bar{X}$ and the collective premium μ, a formula we find desirable. Furthermore, Z approaches 1 as n increases, giving more credit to $\bar{X}$ rather than μ as more past data accumulate, a feature that agrees with intuition. Also, if the population is fairly homogeneous with respect to the risk parameter Θ, then (relatively speaking) the hypothetical means $\mu(\Theta) = \text{E}(X_j|\Theta)$ do not vary greatly with Θ (i.e. they are close in value) and hence have small variability. Thus, a is small relative to v, that is,

k is large and Z is closer to zero. This observation agrees with intuition because, for a homogeneous population, the overall mean μ is of more value in helping to predict next year's claims for a particular policyholder. Conversely, for a heterogeneous population, the hypothetical means $E(X_j|\Theta)$ are more variable, that is, a is large and k is small, and so Z is closer to 1. Again this observation makes sense because, in a heterogeneous population, the experience of other policyholders is of less value in predicting the future experience of a particular policyholder than is the past experience of that policyholder.

We now present some examples.

■ **EXAMPLE 17.14**

(*Example 17.9 continued*) Determine the Bühlmann estimate of $E(X_3|0, 1)$.

From earlier work,

$$\mu(G) = E(X_j|G) = 0.4, \quad \mu(B) = E(X_j|B) = 0.7,$$
$$\pi(G) = 0.75, \qquad\qquad \pi(B) = 0.25,$$

and, therefore,

$$\mu = \sum_\theta \mu(\theta)\pi(\theta) = 0.4(0.75) + 0.7(0.25) = 0.475,$$

$$a = \sum_\theta \mu(\theta)^2 \pi(\theta) - \mu^2 = 0.16(0.75) + 0.49(0.25) - 0.475^2 = 0.016875.$$

For the process variance,

$$v(G) = \text{Var}(X_j|G) = 0^2(0.7) + 1^2(0.2) + 2^2(0.1) - 0.4^2 = 0.44,$$
$$v(B) = \text{Var}(X_j|B) = 0^2(0.5) + 1^2(0.3) + 2^2(0.2) - 0.7^2 = 0.61,$$
$$v = \sum_\theta v(\theta)\pi(\theta) = 0.44(0.75) + 0.61(0.25) = 0.4825.$$

Then, (17.29) gives
$$k = \frac{v}{a} = \frac{0.4825}{0.016875} = 28.5926$$

and (17.28) gives
$$Z = \frac{2}{2 + 28.5926} = 0.0654.$$

The expected next value is then $0.0654(0.5) + 0.9346(0.475) = 0.4766$. This is the best linear approximation to the Bayesian premium (given in Example 17.9).   □

■ **EXAMPLE 17.15**

Suppose, as in Example 17.12 (with $m_j = 1$), that $X_j|\Theta, j = 1, \ldots, n$, are independently and identically Poisson distributed with (given) mean Θ and Θ is gamma distributed with parameters α and β. Determine the Bühlmann premium.

We have

$$\mu(\theta) = E(X_j|\Theta = \theta) = \theta, \quad v(\theta) = \text{Var}(X_j|\Theta = \theta) = \theta,$$

and so

$$\mu = \mathrm{E}[\mu(\Theta)] = \mathrm{E}(\Theta) = \alpha\beta, \quad v = \mathrm{E}[v(\Theta)] = \mathrm{E}(\Theta) = \alpha\beta,$$

and

$$a = \mathrm{Var}[\mu(\Theta)] = \mathrm{Var}(\Theta) = \alpha\beta^2.$$

Then,

$$k = \frac{v}{a} = \frac{\alpha\beta}{\alpha\beta^2} = \frac{1}{\beta}, \quad Z = \frac{n}{n+k} = \frac{n}{n+1/\beta} = \frac{n\beta}{n\beta+1},$$

and the credibility premium is

$$Z\bar{X} + (1-Z)\mu = \frac{n\beta}{n\beta+1}\bar{X} + \frac{1}{n\beta+1}\alpha\beta.$$

But, as shown at the end of Example 17.12, this result is also the Bayesian estimate $\mathrm{E}(X_{n+1}|\mathbf{X})$. Thus, the credibility premium equals the Bayesian estimate in this case. $\square$

■ EXAMPLE 17.16

Determine the Bühlmann estimate for the setting in Example 17.11.

For this model,

$$\mu(\Theta) = \Theta^{-1}, \mu = \mathrm{E}(\Theta^{-1}) = \frac{\beta}{\alpha-1},$$

$$v(\Theta) = \Theta^{-2}, v = \mathrm{E}(\Theta^{-2}) = \frac{\beta^2}{(\alpha-1)(\alpha-2)},$$

$$a = \mathrm{Var}(\Theta^{-1}) = \frac{\beta^2}{(\alpha-1)(\alpha-2)} - \left(\frac{\beta}{\alpha-1}\right)^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)},$$

$$k = \frac{v}{a} = \alpha - 1,$$

$$Z = \frac{n}{n+k} = \frac{n}{n+\alpha-1},$$

$$P_c = \frac{n}{n+\alpha-1}\bar{X} + \frac{\alpha-1}{n+\alpha-1}\frac{\beta}{\alpha-1},$$

which again matches the Bayesian estimate. $\square$

An alternative analysis for this problem could have started with a single observation of $S = X_1 + \cdots + X_n$. From the assumptions of the problem, S has a mean of $n\Theta^{-1}$ and a variance of $n\Theta^{-2}$. While it is true that S has a gamma distribution, that information is not needed because the Bühlmann approximation requires only moments. Following the preceding calculations,

$$\mu = \frac{n\beta}{\alpha-1}, v = \frac{n\beta^2}{(\alpha-1)(\alpha-2)}, a = \frac{n^2\beta^2}{(\alpha-1)^2(\alpha-2)},$$

$$k = \frac{\alpha-1}{n}, Z = \frac{1}{1+k} = \frac{n}{n+\alpha-1}.$$

The key is to note that in calculating Z the sample size is now 1, reflecting the single observation of S. Because $S = n\bar{X}$, the Bühlmann estimate is

$$P_c = \frac{n}{n + \alpha - 1} n\bar{X} + \frac{\alpha - 1}{n + \alpha - 1} \frac{n\beta}{\alpha - 1},$$

which is n times the previous answer. That is because we are now estimating the next value of S rather than the next value of X. However, the credibility factor itself (i.e. Z) is the same whether we are predicting X_{n+1} or the next value of S.

17.6   The Bühlmann–Straub Model

The Bühlmann model is the simplest of the credibility models because it effectively requires that the past claims experience of a policyholder comprise i.i.d. components with respect to each past year. An important practical difficulty with this assumption is that it does not allow for variations in exposure or size.

For example, what if the first year's claims experience of a policyholder reflected only a portion of a year due to an unusual policyholder anniversary? What if a benefit change occurred part way through a policy year? For group insurance, what if the size of the group changed over time?

To handle these variations, we consider the following generalization of the Bühlmann model. Assume that $X_1, \ldots, X_n$ are independent, conditional on Θ, with common mean (as before)

$$\mu(\theta) = \mathrm{E}(X_j|\Theta = \theta)$$

but with conditional variances

$$\mathrm{Var}(X_j|\Theta = \theta) = \frac{v(\theta)}{m_j},$$

where m_j is a known constant measuring exposure. Note that m_j need only be proportional to the size of the risk. This model would be appropriate if each X_j were the average of m_j independent (conditional on Θ) random variables each with mean $\mu(\theta)$ and variance $v(\theta)$. In the preceding situations, m_j could be the number of months the policy was in force in past year j, or the number of individuals in the group in past year j, or the amount of premium income for the policy in past year j.

As in the Bühlmann model, let

$$\mu = \mathrm{E}[\mu(\Theta)], \quad v = \mathrm{E}[v(\Theta)],$$

and

$$a = \mathrm{Var}[\mu(\Theta)].$$

Then, for the unconditional moments, from (17.24) $\mathrm{E}(X_j) = \mu$, and from (17.26) $\mathrm{Cov}(X_i, X_j) = a$, but

$$\mathrm{Var}(X_j) = \mathrm{E}[\mathrm{Var}(X_j|\Theta)] + \mathrm{Var}[\mathrm{E}(X_j|\Theta)]$$

$$= \mathrm{E}\left[\frac{v(\Theta)}{m_j}\right] + \mathrm{Var}[\mu(\Theta)]$$

$$= \frac{v}{m_j} + a.$$

To obtain the credibility premium (17.18), we solve the normal equations (17.15) and (17.17) to obtain $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$. For notational convenience, define

$$m = m_1 + m_2 + \cdots + m_n$$

to be the total exposure. Then, using (17.24), the unbiasedness equation (17.15) becomes

$$\mu = \tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j \mu,$$

which implies

$$\sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}. \tag{17.30}$$

For $i = 1, \ldots, n$, (17.17) becomes

$$a = \sum_{\substack{j=1 \\ j \neq i}}^{n} \tilde{\alpha}_j a + \tilde{\alpha}_i \left(a + \frac{v}{m_i} \right) = \sum_{j=1}^{n} \tilde{\alpha}_j a + \frac{v \tilde{\alpha}_i}{m_i},$$

which may be rewritten as

$$\tilde{\alpha}_i = \frac{a}{v} m_i \left(1 - \sum_{j=1}^{n} \tilde{\alpha}_j \right) = \frac{a}{v} \frac{\tilde{\alpha}_0}{\mu} m_i, \quad i = 1, \ldots, n. \tag{17.31}$$

Then, using (17.30) and (17.31),

$$1 - \frac{\tilde{\alpha}_0}{\mu} = \sum_{j=1}^{n} \tilde{\alpha}_j = \sum_{i=1}^{n} \tilde{\alpha}_i = \frac{a}{v} \frac{\tilde{\alpha}_0}{\mu} \sum_{i=1}^{n} m_i = \frac{a \tilde{\alpha}_0 m}{\mu v},$$

and so

$$\tilde{\alpha}_0 = \frac{\mu}{1 + am/v} = \frac{v/a}{m + v/a} \mu.$$

As a result,

$$\tilde{\alpha}_j = \frac{a \tilde{\alpha}_0}{\mu v} m_j = \frac{m_j}{m + v/a}.$$

The credibility premium (17.18) becomes

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = Z \bar{X} + (1 - Z) \mu, \tag{17.32}$$

where, with $k = v/a$ from (17.29),

$$Z = \frac{m}{m + k}$$

and

$$\bar{X} = \sum_{j=1}^{n} \frac{m_j}{m} X_j. \tag{17.33}$$

Clearly, the credibility premium (17.32) is still of the form (16.7). In this case, m is the total exposure associated with the policyholder, and the Bühlmann–Straub credibility factor Z depends on m. Furthermore, $\bar{X}$ is a weighted average of the X_j, with weights proportional to m_j. Following the group interpretation, X_j is the average loss of the m_j group members in year j, and so $m_j X_j$ is the total loss of the group in year j. Then, $\bar{X}$ is the overall average loss per group member over the n years. The credibility premium to be charged to the group in year $n+1$ would thus be $m_{n+1}[Z\bar{X} + (1-Z)\mu]$ for m_{n+1} members in the next year.

Had we known that (17.33) would be the correct weighting of the X_j to receive the credibility weight Z, the rest would have been easy. For the single observation $\bar{X}$, the process variance is

$$\text{Var}(\bar{X}|\theta) = \sum_{j=1}^{n} \frac{m_j^2}{m^2} \frac{v(\theta)}{m_j} = \frac{v(\theta)}{m},$$

and so the expected process variance is v/m. The variance of the hypothetical means is still a, and therefore $k = v/(am)$. There is only one observation of $\bar{X}$, and so the credibility factor is

$$Z = \frac{1}{1 + v/(am)} = \frac{m}{m + v/a} \tag{17.34}$$

as before. Equation (17.33) should not have been surprising because the weights are simply inversely proportional to the (conditional) variance of each X_j.

■ **EXAMPLE 17.17**

As in Example 17.12, assume that in year j there are N_j claims from m_j policies, $j = 1, \ldots, n$. An individual policy has a Poisson distribution with parameter Θ, and the parameter itself has a gamma distribution with parameters α and β. Determine the Bühlmann–Straub estimate of the number of claims in year $n+1$ if there will be m_{n+1} policies.

To meet the conditions of this model, let $X_j = N_j/m_j$. Because N_j has a Poisson distribution with mean $m_j\Theta$, $\text{E}(X_j|\Theta) = \Theta = \mu(\Theta)$ and $\text{Var}(X_j|\Theta) = \Theta/m_j = v(\Theta)/m_j$. Then,

$$\mu = \text{E}(\Theta) = \alpha\beta, \quad a = \text{Var}(\Theta) = \alpha\beta^2, \quad v = \text{E}(\Theta) = \alpha\beta,$$

$$k = \frac{1}{\beta}, \quad Z = \frac{m}{m + 1/\beta} = \frac{m\beta}{m\beta + 1},$$

and the estimate for one policyholder is

$$P_c = \frac{m\beta}{m\beta + 1}\bar{X} + \frac{1}{m\beta + 1}\alpha\beta,$$

where $\bar{X} = m^{-1}\sum_{j=1}^{n} m_j X_j$. For year $n+1$, the estimate is $m_{n+1}P_c$, matching the answer to Example 17.12. □

The assumptions underlying the Bühlmann–Straub model may be too restrictive to represent reality. In a 1967 paper, Hewitt [54] observed that large risks do not behave the same as an independent aggregation of small risks and, in fact, are more variable than would be indicated by independence. A model that reflects this observation is created in the following example.

■ **EXAMPLE 17.18**

Let the conditional mean be $E(X_j|\Theta) = \mu(\Theta)$ and the conditional variance be $Var(X_j|\Theta) = w(\Theta) + v(\Theta)/m_j$. Further assume that $X_1, \ldots, X_n$ are conditionally independent given Θ. Show that this model supports Hewitt's observation and determine the credibility premium.

Consider independent risks i and j with exposures m_i and m_j and with a common value of Θ. When aggregated, the variance of the average loss is

$$
\begin{aligned}
Var\left(\frac{m_i X_i + m_j X_j}{m_i + m_j} \,\middle|\, \Theta \right) &= \left(\frac{m_i}{m_i + m_j} \right)^2 Var(X_i|\Theta) \\
&\quad + \left(\frac{m_j}{m_i + m_j} \right)^2 Var(X_j|\Theta) \\
&= \frac{m_i^2 + m_j^2}{(m_i + m_j)^2} w(\Theta) + \frac{1}{m_i + m_j} v(\Theta),
\end{aligned}
$$

while a single risk with exposure $m_i + m_j$ has variance $w(\Theta) + v(\Theta)/(m_i + m_j)$, which is larger.

With regard to the credibility premium, we have

$$
\begin{aligned}
E(X_j) &= E[E(X_j|\Theta)] = E[\mu(\Theta)] = \mu, \\
Var(X_j) &= E[Var(X_j|\Theta)] + Var[E(X_j|\Theta)] \\
&= E\left[w(\Theta) + \frac{v(\Theta)}{m_j} \right] + Var[\mu(\Theta)] \\
&= w + \frac{v}{m_j} + a,
\end{aligned}
$$

and, for $i \neq j$, $Cov(X_i, X_j) = a$ as in (17.26). The unbiasedness equation is still

$$
\mu = \tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j \mu,
$$

and so

$$
\sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}.
$$

Equation (17.17) becomes

$$
\begin{aligned}
a &= \sum_{j=1}^{n} \tilde{\alpha}_j a + \tilde{\alpha}_i \left(w + \frac{v}{m_i} \right) \\
&= a \left(1 - \frac{\tilde{\alpha}_0}{\mu} \right) + \tilde{\alpha}_i \left(w + \frac{v}{m_i} \right), \quad i = 1, \ldots, n.
\end{aligned}
$$

Therefore,

$$
\tilde{\alpha}_i = \frac{a \tilde{\alpha}_0 / \mu}{w + v/m_i}.
$$

Summing both sides yields

$$\frac{a\tilde{\alpha}_0}{\mu} \sum_{j=1}^{n} \frac{m_j}{v + wm_j} = \sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu},$$

and so

$$\tilde{\alpha}_0 = \frac{1}{(a/\mu) \sum_{j=1}^{n} \dfrac{m_j}{v + wm_j} + \dfrac{1}{\mu}} = \frac{\mu}{1 + am^*},$$

where

$$m^* = \sum_{j=1}^{n} \frac{m_j}{v + wm_j}.$$

Then,

$$\tilde{\alpha}_j = \frac{am_j}{v + wm_j} \frac{1}{1 + am^*}.$$

The credibility premium is

$$\frac{\mu}{1 + am^*} + \frac{a}{1 + am^*} \sum_{j=1}^{n} \frac{m_j X_j}{v + wm_j}.$$

The sum can be made to define a weighted average of the observations by letting

$$\bar{X} = \frac{\sum_{j=1}^{n} \dfrac{m_j}{v + wm_j} X_j}{\sum_{j=1}^{n} \dfrac{m_j}{v + wm_j}} = \frac{1}{m^*} \sum_{j=1}^{n} \frac{m_j}{v + wm_j} X_j.$$

If we now set

$$Z = \frac{am^*}{1 + am^*},$$

the credibility premium is

$$Z\bar{X} + (1 - Z)\mu.$$

Observe what happens as the exposures m_j go to infinity. The credibility factor becomes

$$Z \to \frac{an/w}{1 + an/w} < 1.$$

Contrast this limit to the Bühlmann–Straub model, where the limit is 1. Thus, no matter how large the risk, there is a limit to its credibility. A further generalization of this result is provided in Exercise 17.12.   □

Another generalization is provided by letting the variance of $\mu(\Theta)$ depend on the exposure, which may be reasonable if we believe that the extent to which a given risk's propensity to produce claims that differ from the mean is related to its size. For example, larger risks may be underwritten more carefully. In this case, extreme variations from the mean are less likely because we ensure that the risk not only meets the underwriting requirements but also appears to be exactly what it claims to be.

■ **EXAMPLE 17.19**

(*Example 17.18 continued*) In addition to the specification presented in Example 17.18, let $\text{Var}[\mu(\Theta)] = a + b/m$, where $m = \sum_{j=1}^{n} m_j$ is the total exposure for the group. Develop the credibility formula.

We now have

$$E(X_j) = E[E(X_j|\Theta)] = E[\mu(\Theta)] = \mu$$

$$\text{Var}(X_j) = E[\text{Var}(X_j|\Theta)] + \text{Var}[E(X_j|\Theta)]$$

$$= E\left[w(\Theta) + \frac{v(\Theta)}{m_j}\right] + \text{Var}[\mu(\Theta)]$$

$$= w + \frac{v}{m_j} + a + \frac{b}{m}$$

and, for $i \neq j$,

$$\text{Cov}(X_i, X_j) = E[E(X_i X_j|\Theta)] - \mu^2$$

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

$$= a + \frac{b}{m}.$$

It can be seen that all the calculations used in Example 17.18 apply here with a replaced by $a + b/m$. The credibility factor is

$$Z = \frac{(a + b/m)m^*}{1 + (a + b/m)m^*},$$

and the credibility premium is

$$Z\bar{X} + (1 - Z)\mu,$$

with $\bar{X}$ and m^* defined as in Example 17.18. This particular credibility formula has been used in workers compensation experience rating. One example of this application is presented in detail in Gillam [45]. □

17.7 Exact Credibility

In Examples 17.15–17.17, we found that the credibility premium and the Bayesian premium are equal. From (17.19), we may view the credibility premium as the best linear approximation to the Bayesian premium in the sense of squared-error loss. In these examples, the approximation is exact because the two premiums are equal. The term **exact credibility** is used to describe the situation in which the credibility premium equals the Bayesian premium.

At first glance, it appears to be unnecessary to discuss the existence and finiteness of the credibility premium in this context, because exact credibility as defined is clearly not possible otherwise. However, in what follows, there are some technical issues to be considered, and their treatment is clearer if it is tacitly remembered that the credibility premium must be well defined, which requires that $E(X_j) < \infty$, $\text{Var}(X_j) < \infty$, and

$\text{Cov}(X_i, X_j) < \infty$, as is obvious from the normal equations (17.15) and (17.17). Exact credibility typically occurs in Bühlmann (and Bühlmann–Straub) situations involving linear exponential family members and their conjugate priors. It is clear that the existence of the credibility premium requires that the structural parameters $E[\mu(\Theta)]$, $E[\text{Var}(\Theta)]$, and $\text{Var}[\mu(\Theta)]$ be finite.

Consider $E[\mu(\Theta)]$ in this situation. Recall from (5.8) that, for the linear exponential family, the mean is

$$\mu(\theta) = E(X_j | \Theta = \theta) = \frac{q'(\theta)}{r'(\theta)q(\theta)}, \tag{17.35}$$

and the conjugate prior pdf is, from Theorem 13.18, given by

$$\pi(\theta) = \frac{[q(\theta)]^{-k} e^{\mu k r(\theta)} r'(\theta)}{c(\mu, k)}, \quad \theta_0 < \theta < \theta_1, \tag{17.36}$$

where the interval of support (θ_0, θ_1) is explicitly identified. Also, for now, μ and k should be viewed as known parameters associated with the prior pdf $\pi(\theta)$. To determine $E[\mu(\Theta)]$, note that from (17.36) it follows that

$$\ln[\pi(\theta)/r'(\theta)] = -k \ln q(\theta) + \mu k r(\theta) - \ln c(\mu, k),$$

and differentiating with respect to θ yields

$$\frac{\frac{d}{d\theta}[\pi(\theta)/r'(\theta)]}{\pi(\theta)/r'(\theta)} = -k\frac{q'(\theta)}{q(\theta)} + \mu k r'(\theta).$$

Multiplication by $\pi(\theta)/r'(\theta)$ results in, using (17.35),

$$\frac{d}{d\theta}\left[\frac{\pi(\theta)}{r'(\theta)}\right] = -k[\mu(\theta) - \mu]\pi(\theta). \tag{17.37}$$

Next, integrate both sides of (17.37) with respect to θ over the interval (θ_0, θ_1), to obtain

$$\frac{\pi(\theta_1)}{r'(\theta_1)} - \frac{\pi(\theta_0)}{r'(\theta_0)} = -k \int_{\theta_0}^{\theta_1} [\mu(\theta) - \mu]\pi(\theta)\, d\theta$$

$$= -k E[\mu(\Theta) - \mu].$$

Therefore, it follows that

$$E[\mu(\Theta)] = \mu + \frac{\pi(\theta_0)}{kr'(\theta_0)} - \frac{\pi(\theta_1)}{kr'(\theta_1)}. \tag{17.38}$$

Note that, if

$$\frac{\pi(\theta_1)}{r'(\theta_1)} = \frac{\pi(\theta_0)}{r'(\theta_0)}, \tag{17.39}$$

then

$$E[\mu(\Theta)] = \mu, \tag{17.40}$$

demonstrating that the choice of the symbol μ in (17.36) is not coincidental. If (17.40) holds, as is often the case, it is normally because both sides of (17.39) are equal to zero.

Regardless, it is possible to have $E[\mu(\Theta)] < \infty$ but $E[\mu(\theta)] \neq \mu$. Also, $E[\mu(\Theta)] = \infty$ may result if either $\pi(\theta_0)/r'(\theta_0)$ or $\pi(\theta_1)/r'(\theta_1)$ fails to be finite.

Next, consider the posterior distribution in the Bühlmann situation with

$$f_{X_j|\Theta}(x_j|\theta) = \frac{p(x_j)e^{r(\theta)x_j}}{q(\theta)}$$

and $\pi(\theta)$ given by (17.36). From Theorem 13.18, the posterior pdf is

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{[q(\theta)]^{-k_*}e^{\mu_* k_* r(\theta)}r'(\theta)}{c(\mu_*, k_*)}, \quad \theta_0 < \theta < \theta_1, \tag{17.41}$$

with

$$k_* = k + n \tag{17.42}$$

and

$$\mu_* = \frac{\mu k + n\bar{x}}{k + n}. \tag{17.43}$$

Because (17.41) is of the same form as (17.36), the Bayesian premium (17.13) is

$$E(X_{n+1}|\mathbf{X} = \mathbf{x}) = \mu_* + \frac{\pi_{\Theta|\mathbf{X}}(\theta_0|\mathbf{x})}{k_* r'(\theta_0)} - \frac{\pi_{\Theta|\mathbf{X}}(\theta_1|\mathbf{x})}{k_* r'(\theta_1)}, \tag{17.44}$$

with μ_* given by (17.43). Because μ_* is a linear function of the x_js, the same is true of the Bayesian premium if

$$\frac{\pi_{\Theta|\mathbf{X}}(\theta_0|\mathbf{x})}{r'(\theta_0)} = \frac{\pi_{\Theta|\mathbf{X}}(\theta_1|\mathbf{x})}{r'(\theta_1)}, \tag{17.45}$$

that is, (17.45) implies that (17.44) becomes

$$E(X_{n+1}|\mathbf{X} = \mathbf{x}) = \mu_* = \frac{\mu k + n\bar{x}}{k + n}. \tag{17.46}$$

Clearly, for (17.45) to hold for all vectors $\mathbf{x}$, both sides should be equal to zero. Also, note that (17.46) is of the form (16.7).

To summarize, posterior linearity of the Bayesian premium results (i.e. (17.46) holds) if (17.45) is true (usually with both sides equal to zero). It is instructive to note that posterior linearity of the Bayesian premium may occur even if $E[\mu(\Theta)] = \infty$. However, as long as the credibility premium is well defined (all three of $E[\mu(\Theta)]$, $E[v(\Theta)]$, and $\text{Var}[\mu(\Theta)]$ are finite), the posterior linearity of the Bayesian premium implies equality with the credibility premium, that is, exact credibility. To see this equivalence, note that, if the Bayesian premium is a linear function of $X_1, \ldots, X_n$, that is,

$$E(X_{n+1}|\mathbf{X}) = a_0 + \sum_{j=1}^{n} a_j X_j,$$

then it is clear that in (17.19) the quantity Q_1 attains its minimum value of zero with $\tilde{\alpha}_j = a_j$ for $j = 0, 1, \ldots, n$. Thus the credibility premium is $\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = a_0 + \sum_{j=1}^{n} a_j X_j = E(X_{n+1}|\mathbf{X})$, and credibility is exact.

The following example clarifies these concepts.

■ **EXAMPLE 17.20**

Suppose that the $f_{X_j|\Theta}(x_j|\theta) = \theta e^{-\theta x_j}$, $x_j > 0$. Then, $r(\theta) = -\theta$ and $q(\theta) = 1/\theta$, which implies, using (17.35), that $\mu(\theta) = 1/\theta$. Similarly, $v(\theta) = \text{Var}(X_j|\Theta = \theta) = \mu'(\theta)/r'(\theta) = 1/\theta^2$ from (5.9). Next, (17.36) becomes

$$\pi(\theta) = \frac{\theta^k e^{-\mu k \theta}}{\int_{\theta_0}^{\theta_1} t^k e^{-\mu k t}\,dt}, \quad \theta_0 < \theta < \theta_1. \tag{17.47}$$

In the truncated situation, with $0 < \theta_0 < \theta_1 < \infty$, there are no restrictions on the prior model parameters k and μ for $\pi(\theta)$ to be a valid pdf. Furthermore, in this case, $E[\mu(\Theta)]$, $E[v(\Theta)]$, and $\text{Var}[\mu(\Theta)]$ are all finite, and therefore the credibility premium is well defined. In fact, (17.38) becomes

$$E[\mu(\Theta)] = \mu + \frac{\theta_1^k e^{-\mu k \theta_1} - \theta_0^k e^{-\mu k \theta_0}}{k \int_{\theta_0}^{\theta_1} t^k e^{-\mu k t}\,dt}. \tag{17.48}$$

The posterior pdf from (17.41) is of the same form as (17.47), with μ and k replaced by μ_* and k_* in (17.43) and (17.42), respectively. Therefore, the Bayesian premium (17.44) in this truncated situation is, by analogy with (17.48),

$$E(X_{n+1}|\mathbf{X} = \mathbf{x}) = \mu_* + \frac{\theta_1^{k_*} e^{-\mu_* k_* \theta_1} - \theta_0^{k_*} e^{-\mu_* k_* \theta_0}}{k_* \int_{\theta_0}^{\theta_1} t^{k_*} e^{-\mu_* k_* t}\,dt}. \tag{17.49}$$

Because μ_* is a linear function of the x_js, (17.49) is nonlinear in the x_js, and therefore credibility cannot be exact. Furthermore, this truncated example demonstrates that the endpoint conditions (17.39) and (17.45) needed for exact credibility are model assumptions and, so, cannot be omitted just to obtain a nicer result.

Next, consider the more usual (untruncated) situation with $\theta_0 = 0$ and $\theta_1 = \infty$. Then, (17.47) becomes the gamma pdf with

$$\pi(\theta) = \frac{\mu k(\mu k \theta)^k e^{-\mu k \theta}}{\Gamma(k+1)}, \quad \theta > 0, \tag{17.50}$$

which is a valid pdf as long as $k > -1$ and $\mu k > 0$. There are three cases:

Case	Result
$-1 < k \leq 0$	$E[\mu(\Theta)] = E[v(\Theta)] = \text{Var}[\mu(\Theta)] = \infty$
$0 < k \leq 1$	$E[\mu(\Theta)] = \mu < \infty$, $E[v(\Theta)] = \text{Var}[\mu(\Theta)] = \infty$
$k > 1$	$E[\mu(\Theta)] = \mu < \infty$, $E[v(\Theta)] < \infty$, $\text{Var}[\mu(\Theta)] < \infty$

Hence, there is no credibility premium unless $k > 1$. However, because $k_* = k+n > 0$ regardless of the value of k, the Bayesian premium is

$$E(X_{n+1}|\mathbf{X} = \mathbf{x}) = \mu_* = \frac{\mu k + n\bar{x}}{k + n},$$

a linear function of the x_js. To summarize, in the exponential-gamma model with prior pdf (17.50), the Bayesian premium is a linear function of the x_js regardless of the

value of k, whereas if $k \leq 1$ there is no credibility premium. If $k > 1$, then credibility is exact. $\square$

There is one last technical point worth noting. It was mentioned previously that the choice of the symbol μ as a parameter associated with the prior pdf $\mu(\theta)$ is not a coincidence because it is often the case that $E[\mu(\Theta)] = \mu$. A similar comment applies to the parameter k. Because $v(\theta) = \mu'(\theta)/r'(\theta)$ from (5.9), it follows from (17.37) and the product rule for differentiation that

$$\frac{d}{d\theta}\left\{[\mu(\theta) - \mu]\frac{\pi(\theta)}{r'(\theta)}\right\} = \mu'(\theta)\left[\frac{\pi(\theta)}{r'(\theta)}\right] + [\mu(\theta) - \mu]\frac{d}{d\theta}\left[\frac{\pi(\theta)}{r'(\theta)}\right]$$
$$= v(\theta)\pi(\theta) - k[\mu(\theta) - \mu]^2\pi(\theta).$$

Integrating with respect to θ over (θ_0, θ_1) yields

$$[\mu(\theta) - \mu]\frac{\pi(\theta)}{r'(\theta)}\bigg|_{\theta_0}^{\theta_1} = E[v(\Theta)] - kE\left\{[\mu(\Theta) - \mu]^2\right\},$$

and solving for k yields

$$k = \frac{E[v(\Theta)] + [\mu(\theta_0) - \mu]\dfrac{\pi(\theta_0)}{r'(\theta_0)} - [\mu(\theta_1) - \mu]\dfrac{\pi(\theta_1)}{r'(\theta_0)}}{E\left\{[\mu(\Theta) - \mu]^2\right\}}. \tag{17.51}$$

If, in addition, (17.39) holds, then (17.40) holds, and (17.51) simplifies to

$$k = \frac{E[v(\Theta)] + \mu(\theta_0)\dfrac{\pi(\theta_0)}{r'(\theta_0)} - \mu(\theta_1)\dfrac{\pi(\theta_1)}{r'(\theta_1)}}{\text{Var}[\mu(\Theta)]}, \tag{17.52}$$

in turn simplifying to the well-known result $k = E[v(\Theta)]/\text{Var}[\mu(\Theta)]$ if

$$\frac{\mu(\theta_0)\pi(\theta_0)}{r'(\theta_0)} = \frac{\mu(\theta_1)\pi(\theta_1)}{r'(\theta_1)},$$

which typically holds with both sides equal to zero.

17.8   Notes and References

In this section, one of the two major criticisms of limited fluctuation credibility has been addressed. Through the use of the variance of the hypothetical means, we now have a means of relating the mean of the group of interest, $\mu(\theta)$, to the manual, or collective, premium, μ. The development is also mathematically sound in that the results follow directly from a specific model and objective. We have also seen that the additional restriction of a linear solution is not as bad as it might be in that we still often obtain the exact Bayesian solution. There has subsequently been a great deal of effort expended to generalize the model. With a sound basis for obtaining a credibility premium, we have but one remaining obstacle: how to numerically estimate the quantities a and v in the Bühlmann formulation, or how to specify the prior distribution in the Bayesian formulation. Those matters are addressed in Chapter 18.

Table 17.2   The data for Exercise 17.4.

x	y 0	1	2
0	0.20	0	0.10
1	0	0.15	0.25
2	0.05	0.15	0.10

A historical review of credibility theory including a description of the limited fluctuation and greatest accuracy approaches is provided by Norberg [94]. Since the classic paper of Bühlmann [19], there has developed a vast literature on credibility theory in the actuarial literature. Other elementary introductions are given by Herzog [52] and Waters [130]. Other more advanced treatments are Goovaerts and Hoogstad [46] and Sundt [118]. An important generalization of the Bühlmann–Straub model is the Hachemeister [48] regression model, which is not discussed here. See also Klugman [71]. The material on exact credibility is motivated by Jewell [62]. See also Ericson [36]. A special issue of *Insurance: Abstracts and Reviews* (Sundt [117]) contains an extensive list of papers on credibility.

17.9 Exercises

17.1 Suppose that X and Z are independent Poisson random variables with means λ_1 and λ_2, respectively. Let $Y = X + Z$. Demonstrate that $X|Y = y$ is binomial.

17.2 Suppose X is binomially distributed with parameters n_1 and p, that is,

$$f_X(x) = \binom{n_1}{x} p^x (1-p)^{n_1-x}, \quad x = 0, 1, 2, \dots, n_1.$$

Suppose also that Z is binomially distributed with parameters n_2 and p independently of X. Then, $Y = X + Z$ is binomially distributed with parameters $n_1 + n_2$ and p. Demonstrate that $X|Y = y$ has the hypergeometric distribution.

17.3 Consider a compound Poisson distribution with Poisson mean λ, where $X = Y_1 + \cdots + Y_N$ with $E(Y_i) = \mu_Y$ and $Var(Y_i) = \sigma_Y^2$. Determine the mean and variance of X.

17.4 Let X and Y have joint probability distribution as given in Table 17.2.

(a) Compute the marginal distributions of X and Y.

(b) Compute the conditional distribution of X given $Y = y$ for $y = 0, 1, 2$.

(c) Compute $E(X|y)$, $E(X^2|y)$, and $Var(X|y)$ for $y = 0, 1, 2$.

(d) Compute $E(X)$ and $Var(X)$ using (17.3), (17.6), and (c).

17.5 Suppose that X and Y are two random variables with bivariate normal joint density function

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)\right.\right.$$
$$\left.\left.+ \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

Show the following:

(a) The conditional density function is

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}\left[\frac{x-\mu_1-\rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)}{\sigma_1\sqrt{1-\rho^2}}\right]^2\right\}.$$

Hence,
$$E(X|Y = y) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2).$$

(b) The marginal pdf is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right].$$

(c) The variables X and Y are independent if and only if $\rho = 0$.

17.6 Suppose that, given $\Theta = (\Theta_1, \Theta_2)$, the random variable X is normally distributed with mean Θ_1 and variance Θ_2.

(a) Show that $E(X) = E(\Theta_1)$ and $Var(X) = E(\Theta_2) + Var(\Theta_1)$.
(b) If Θ_1 and Θ_2 are independent, show that X has the same distribution as $\Theta_1 + Y$, where Θ_1 and Y are independent and Y conditional on Θ_2 is normally distributed with mean zero and variance Θ_2.

17.7 Suppose that Θ has pdf $\pi(\theta)$, $\theta > 0$, and Θ_1 has pdf $\pi_1(\theta) = \pi(\theta - \alpha)$, $\theta > \alpha > 0$. If, given Θ_1, X is Poisson distributed with mean Θ_1, show that X has the same distribution as $Y + Z$, where Y and Z are independent, Y is Poisson distributed with mean α, and $Z|\Theta$ is Poisson distributed with mean Θ.

17.8 Consider a die–spinner model. The first die has one "marked" face and five "unmarked" faces, whereas the second die has four "marked" faces and two "unmarked" faces. There are three spinners, each with five equally spaced sectors marked 3 or 8. The first spinner has one sector marked 3 and four marked 8, the second has two marked 3 and three marked 8, and the third has four marked 3 and one marked 8. One die and one spinner are selected at random. If rolling the die produces an unmarked face, no claim occurs. If a marked face

Table 17.3   The data for Exercise 17.9.

Urn	0s	1s	2s
1	0.40	0.35	0.25
2	0.25	0.10	0.65
3	0.50	0.15	0.35

occurs, there is a claim and then the spinner is spun once to determine the amount of the claim.

(a) Determine $\pi(\theta)$ for each of the six die–spinner combinations.

(b) Determine the conditional distributions $f_{X|\Theta}(x|\theta)$ for the claim sizes for each die–spinner combination.

(c) Determine the hypothetical means $\mu(\theta)$ and the process variances $v(\theta)$ for each θ.

(d) Determine the marginal probability that the claim X_1 on the first iteration equals 3.

(e) Determine the posterior distribution $\pi_{\Theta|X_1}(\theta|3)$ of Θ using Bayes' theorem.

(f) Use (17.10) to determine the conditional distribution $f_{X_2|X_1}(x_2|3)$ of the claims X_2 on the second iteration given that $X_1 = 3$ was observed on the first iteration.

(g) Use (17.13) to determine the Bayesian premium $E(X_2|X_1 = 3)$.

(h) Determine the joint probability that $X_2 = x_2$ and $X_1 = 3$ for $x_2 = 0, 3, 8$.

(i) Determine the conditional distribution $f_{X_2|X_1}(x_2|3)$ directly using (17.8) and compare your answer to that of (f).

(j) Determine the Bayesian premium directly using (17.12) and compare your answer to that of (g).

(k) Determine the structural parameters μ, v, and a.

(l) Compute the Bühlmann credibility factor and the Bühlmann credibility premium to approximate the Bayesian premium $E(X_2|X_1 = 3)$.

17.9 Three urns have balls marked 0, 1, and 2 in the proportions given in Table 17.3. An urn is selected at random, and two balls are drawn from that urn with replacement. A total of 2 on the two balls is observed. Two more balls are then drawn with replacement from the same urn, and it is of interest to predict the total on these next two balls.

(a) Determine $\pi(\theta)$.

(b) Determine the conditional distributions $f_{X|\Theta}(x|\theta)$ for the totals on the two balls for each urn.

(c) Determine the hypothetical means $\mu(\theta)$ and the process variances $v(\theta)$ for each θ.

(d) Determine the marginal probability that the total X_1 on the first two balls equals 2.

(e) Determine the posterior distribution $\pi_{\Theta|X_1}(\theta|2)$ using Bayes' theorem.

(f) Use (17.10) to determine the conditional distribution $f_{X_2|X_1}(x_2|2)$ of the total X_2 on the next two balls drawn, given that $X_1 = 2$ was observed on the first two draws.

Table 17.4 The data for Exercise 17.10.

Type	Number of claims		Severity	
	Mean	Variance	Mean	Variance
A	0.2	0.2	200	4,000
B	0.7	0.3	100	1,500

(g) Use (17.13) to determine the Bayesian premium $E(X_2|X_1 = 2)$.

(h) Determine the joint probability that the total X_2 on the next two balls equals x_2 and the total X_1 on the first two balls equals 2 for $x_2 = 0, 1, 2, 3, 4$.

(i) Determine the conditional distribution $f_{X_2|X_1}(x_2|2)$ directly using (17.8) and compare your answer to that of (f).

(j) Determine the Bayesian premium directly using (17.12) and compare your answer to that of (g).

(k) Determine the structural parameters μ, v, and a.

(l) Determine the Bühlmann credibility factor and the Bühlmann credibility premium.

(m) Show that the Bühlmann credibility factor is the same if each "exposure unit" consists of one draw from the urn rather than two draws.

17.10 Suppose that there are two types of policyholder: type A and type B. Two-thirds of the total number of the policyholders are of type A and one-third are of type B. For each type, the information on annual claim numbers and severity are given in Table 17.4. A policyholder has a total claim amount of 500 in the past four years. Determine the credibility factor Z and the credibility premium for next year for this policyholder.

17.11 Let Θ_1 represent the risk factor for claim numbers and let Θ_2 represent the risk factor for the claim severity for a line of insurance. Suppose that Θ_1 and Θ_2 are independent. Suppose also that, given $\Theta_1 = \theta_1$, the claim number N is Poisson distributed and, given $\Theta_2 = \theta_2$, the severity Y is exponentially distributed. The expectations of the hypothetical means and process variances for the claim number and severity as well as the variance of the hypothetical means for frequency are, respectively,

$$\mu_N = 0.1, \qquad v_N = 0.1, \qquad a_N = 0.05,$$
$$\mu_Y = 100, \qquad v_Y = 25{,}000.$$

Three observations are made on a particular policyholder and we observe total claims of 200. Determine the Bühlmann credibility factor and the Bühlmann premium for this policyholder.

17.12 Suppose that $X_1, \ldots, X_n$ are independent (conditional on Θ) and that

$$E(X_j|\Theta) = \beta_j \mu(\Theta) \text{ and } \text{Var}(X_j|\Theta) = \tau_j(\Theta) + \psi_j v(\Theta), \quad j = 1, \ldots, n.$$

Let

$$\mu = E[\mu(\Theta)], \quad v = E[v(\Theta)], \quad \tau_j = E[\tau_j(\Theta)], \quad a = \text{Var}[\mu(\Theta)].$$

(a) Show that

$$E(X_j) = \beta_j \mu, \quad \text{Var}(X_j) = \tau_j + \psi_j v + \beta_j^2 a,$$

and

$$\text{Cov}(X_i, X_j) = \beta_i \beta_j a, \quad i \neq j.$$

(b) Solve the normal equations for $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ to show that the credibility premium satisfies

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = (1 - Z) \, \text{E}(X_{n+1}) + Z \beta_{n+1} \bar{X},$$

where

$$m_j = \beta_j^2 (\tau_j + \psi_j v)^{-1}, \quad j = 1, \ldots, n,$$

$$m = m_1 + \cdots + m_n,$$

$$Z = am(1 + am)^{-1},$$

$$\bar{X} = \sum_{j=1}^{n} \frac{m_j}{m} \frac{X_j}{\beta_j}.$$

17.13 For the situation described in Exercise 13.5, determine $\mu(\theta)$ and the Bayesian premium $\text{E}(X_{n+1}|\mathbf{x})$. Why is the Bayesian premium equal to the credibility premium?

17.14 Suppose that, for $j = 1, 2, \ldots, n$,

$$f_{X_j|\Theta}(x_j|\theta) = \frac{\Gamma(\alpha + x_j)}{\Gamma(\alpha)x_j!}(1 - \theta)^\alpha \theta^{x_j}, \quad x_j = 0, 1, \ldots ,$$

a negative binomial pf with $\alpha > 0$ a known quantity.

(a) Demonstrate that the conjugate prior from Theorem 13.18 is the beta pf

$$\pi(\theta) = \frac{\Gamma(\mu k + \alpha k + 1)}{\Gamma(\mu k)\Gamma(\alpha k + 1)} \theta^{\mu k - 1}(1 - \theta)^{\alpha k}, \quad 0 < \theta < 1,$$

where $k > -1/\alpha$ and $\mu k > 0$ are the acceptable parameter values for $\pi(\theta)$ to be a valid pdf.

(b) Show that $\text{E}[\mu(\Theta)] = \infty$ if $-1/\alpha < k < 0$ and $\text{E}[\mu(\theta)] = \mu < \infty$ if $k > 0$.

(c) Show that there is no credibility premium if $k \leq 1/\alpha$. Next, show that if $k > 1/\alpha$, then $\text{Var}[\mu(\Theta)] = \mu(\mu + \alpha)/(\alpha k - 1)$ and $\text{E}[v(\Theta)]/\text{Var}[\mu(\Theta)] = k$.

(d) Prove that there is no Bayesian premium if the number of observations n satisfies $n < 1/\alpha$ and $-1/\alpha < k < -n$, and that if $k > -n$, then the Bayesian premium is linear in the x_js. What happens if $k = -n$?

(e) Show that credibility is exact if $k > 1/\alpha$.

17.15 Consider the generalization of the linear exponential family given by

$$f(x; \theta, m) = \frac{p(m, x)e^{mr(\theta)x}}{[q(\theta)]^m}.$$

If m is a parameter, this is called the **exponential dispersion family**. In Exercise 5.25, it is shown that the mean of this random variable is $q'(\theta)/[r'(\theta)q(\theta)]$. For this exercise, assume that m is known.

(a) Consider the prior distribution

$$\pi(\theta) = \frac{[q(\theta)]^{-k} \exp[\mu k r(\theta)] r'(\theta)}{c(\mu, k)}, \quad \theta_0 < \theta < \theta_1.$$

Determine the Bayesian premium.

(b) Using the same prior, determine the Bühlmann premium.

(c) Show that the inverse Gaussian distribution is a member of the exponential dispersion family.

17.16 Suppose that $X_1, \ldots, X_n$ are independent (conditional on Θ) and

$$E(X_j|\Theta) = \tau^j \mu(\Theta) \quad \text{and} \quad \text{Var}(X_j|\Theta) = \frac{\tau^{2j} v(\Theta)}{m_j}, \quad j = 1, \ldots, n.$$

Let $\mu = E[\mu(\Theta)]$, $v = E[v(\Theta)]$, $a = \text{Var}[\mu(\Theta)]$, $k = v/a$, and $m = m_1 + \cdots + m_n$.

(a) Discuss when these assumptions may be appropriate.

(b) Show that

$$E(X_j) = \tau^j \mu, \quad \text{Var}(X_j) = \tau^{2j}(a + v/m_j),$$

and

$$\text{Cov}(X_i, X_j) = \tau^{i+j} a, \quad i \neq j.$$

(c) Solve the normal equations for $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ to show that the credibility premium satisfies

$$\tilde{\alpha}_0 + \sum_{j=1}^n \tilde{\alpha}_j X_j = \frac{k}{k+m} \tau^{n+1} \mu + \frac{m}{k+m} \sum_{j=1}^n \frac{m_j}{m} \tau^{n+1-j} X_j.$$

(d) Give a verbal interpretation of the formula in (c).

(e) Suppose that

$$f_{X_j|\Theta}(x_j|\theta) = \frac{p(x_j, m_j, \tau) e^{m_j \tau^{-j} x_j r(\theta)}}{[q(\theta)]^{m_j}}.$$

Show that $E(X_j|\Theta) = \tau^j \mu(\Theta)$ and that $\text{Var}(X_j|\Theta) = \tau^{2j} v(\Theta)/m_j$, where $\mu(\theta) = q'(\theta)/[r'(\theta)q(\theta)]$ and $v(\theta) = \mu'(\theta)/r'(\theta)$.

(f) Determine the Bayesian premium if

$$\pi(\theta) = \frac{[q(\theta)]^{-k} e^{\mu k r(\theta)}}{c(\mu, k)}, \quad \theta_0 < \theta < \theta_1.$$

17.17 Suppose that, given $\Theta = \theta$, the random variables $X_1, \ldots, X_n$ are independent with Poisson pf

$$f_{X_j|\Theta}(x_j|\theta) = \frac{\theta^{x_j} e^{-\theta}}{x_j!}, \quad x_j = 0, 1, 2, \ldots.$$

(a) Let $S = X_1 + \cdots + X_n$. Show that S has pf

$$f_S(s) = \int_0^\infty \frac{(n\theta)^s e^{-n\theta}}{s!} \pi(\theta) \, d\theta, \quad s = 0, 1, 2, \dots,$$

where Θ has pdf $\pi(\theta)$.

(b) Show that the Bayesian premium is

$$E(X_{n+1}|\mathbf{X} = \mathbf{x}) = \frac{1 + n\bar{x}}{n} \frac{f_S(1 + n\bar{x})}{f_S(n\bar{x})}.$$

(c) Evaluate the distribution of S in (a) when $\pi(\theta)$ is a gamma distribution. What type of distribution is this?

17.18 Suppose that, given $\Theta = \theta$, the random variables $X_1, X_2, \dots, X_n$ are independent with Poisson pf

$$f_{X_j|\Theta}(x_j|\theta) = \frac{\theta^{x_j} e^{-\theta}}{x_j!}, \quad x_j = 0, 1, \dots,$$

and Θ has the inverse Gaussian pdf from Appendix A (with θ replaced by γ),

$$\pi(\theta) = \sqrt{\frac{\gamma}{2\pi\theta^3}} \exp\left[-\frac{\gamma}{2\theta}\left(\frac{\theta - \mu}{\mu}\right)^2\right], \quad \theta > 0.$$

Define $\alpha(m) = m + \gamma/(2\mu^2)$.

(a) Use Exercise 5.20(g) to show that the posterior distribution of Θ given that $\mathbf{X} = \mathbf{x}$ is the **generalized inverse Gaussian distribution** with pdf

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{[2\alpha(n)/\gamma]^{0.5n\bar{x}-0.25} \theta^{n\bar{x}-1.5} \exp\left[-\alpha(n)\theta - \frac{\gamma}{2\theta}\right]}{2K_{n\bar{x}-0.5}\left(\sqrt{2\alpha(n)\gamma}\right)}, \quad \theta > 0.$$

(b) Use part (a) and Exercise 5.20(g) to prove that the predictive distribution of X_{n+1} given $\mathbf{X} = \mathbf{x}$ is the **Sichel distribution** with pdf

$$f_{X_{n+1}|\mathbf{X}}(x_{n+1}|\mathbf{x}) = \left[\frac{\gamma}{2\alpha(n+1)}\right]^{0.5x_{n+1}} \left[\frac{\alpha(n)}{\alpha(n+1)}\right]^{0.5n\bar{x}-0.25}$$

$$\times \frac{K_{x_{n+1}+n\bar{x}-0.5}\left(\sqrt{2\alpha(n+1)\gamma}\right)}{(x_{n+1}!)K_{n\bar{x}-0.5}\left(\sqrt{2\alpha(n)\gamma}\right)}$$

for $x_{n+1} = 0, 1, \dots$.

(c) Use Example 7.16 to evaluate the pf

$$f_S(s) = \int_0^\infty \frac{(n\theta)^s e^{-n\theta}}{s!} \pi(\theta) \, d\theta, \quad s = 0, 1, \dots,$$

and, hence, use Exercise 17.17(b) to describe how to calculate the Bayesian premium.

17.19 Suppose that $X_j | \Theta$ is normally distributed with mean Θ and variance v for $j = 1, 2, \ldots, n + 1$. Further suppose that Θ is normally distributed with mean μ and variance a. Thus,

$$f_{X_j | \Theta}(x_j | \theta) = (2 \pi v)^{-1/2} \exp \left[-\frac{1}{2v}(x_j - \theta)^2 \right], \quad -\infty < x_j < \infty,$$

and

$$\pi(\theta) = (2 \pi a)^{-1/2} \exp \left[-\frac{1}{2a}(\theta - \mu)^2 \right], \quad -\infty < \theta < \infty.$$

Determine the posterior distribution of $\Theta | \mathbf{X}$ and the predictive distribution of $X_{n+1} | \mathbf{X}$. Then determine the Bayesian estimate of $E(X_{n+1} | \mathbf{X})$. Finally, show that the Bayesian and Bühlmann estimates are equal.

17.20 (*) Your friend selected at random one of two urns and then she pulled a ball with number 4 on it from the urn. Then she replaced the ball in the urn. One of the urns contains four balls, numbered 1–4. The other urn contains six balls, numbered 1–6. Your friend will make another random selection from the same urn.

(a) Estimate the expected value of the number on the next ball using the Bayesian method.

(b) Estimate the expected value of the number on the next ball using Bühlmann credibility.

17.21 The number of claims for a randomly selected insured has a Poisson distribution with parameter θ. The parameter θ is distributed across the population with pdf $\pi(\theta) = 3\theta^{-4}$, $\theta > 1$. For an individual, the parameter does not change over time. A particular insured experienced a total of 20 claims in the previous two years.

(a) (*) Determine the Bühlmann credibility estimate for the future expected claim frequency for this particular insured.

(b) Determine the Bayesian credibility estimate for the future expected claim frequency for this particular insured.

17.22 (*) The distribution of payments to an insured is constant over time. If the Bühlmann credibility assigned for one-half year of observation is 0.5, determine the Bühlmann credibility to be assigned for three years.

17.23 (*) Three urns contain balls marked either 0 or 1. In urn A, 10% are marked 0; in urn B, 60% are marked 0; and in urn C, 80% are marked 0. An urn is selected at random and three balls selected with replacement. The total of the values is 1. Three more balls are selected with replacement from the same urn.

(a) Determine the expected total of the three balls using Bayes' theorem.

(b) Determine the expected total of the three balls using Bühlmann credibility.

17.24 (*) The number of claims follows a Poisson distribution with parameter λ. A particular insured had three claims in the past three years.

    (a) The value of λ has pdf $f(\lambda) = 4\lambda^{-5}$, $\lambda > 1$. Determine the value of K used in Bühlmann's credibility formula. Then use Bühlmann credibility to estimate the claim frequency for this insured.

    (b) The value of λ has pdf $f(\lambda) = 1$, $0 < \lambda < 1$. Determine the value of K used in Bühlmann's credibility formula. Then use Bühlmann credibility to estimate the claim frequency for this insured.

17.25 (*) The number of claims follows a Poisson distribution with parameter h. The value of h has the gamma distribution with pdf $f(h) = he^{-h}$, $h > 0$. Determine the Bühlmann credibility to be assigned to a single observation. (The Bayes solution is obtained in Exercise 13.22.)

17.26 Consider the situation of Exercise 13.24.

    (a) Determine the expected number of claims in the second year using Bayesian credibility.

    (b) (*) Determine the expected number of claims in the second year using Bühlmann credibility.

17.27 (*) One spinner is selected at random from a group of three spinners. Each spinner is divided into six equally likely sectors. The number of sectors marked 0, 12, and 48, respectively, on each spinner is as follows: spinner A, 2,2,2; spinner B, 3,2,1; and spinner C, 4,1,1. A spinner is selected at random and a zero is obtained on the first spin.

    (a) Determine the Bühlmann credibility estimate of the expected value of the second spin using the same spinner.

    (b) Determine the Bayesian credibility estimate of the expected value of the second spin using the same spinner.

17.28 The number of claims in a year has a Poisson distribution with mean λ. The parameter λ has the uniform distribution over the interval $(1, 3)$.

    (a) (*) Determine the probability that a randomly selected individual will have no claims.

    (b) (*) If an insured had one claim during the first year, estimate the expected number of claims for the second year using Bühlmann credibility.

    (c) If an insured had one claim during the first year, estimate the expected number of claims for the second year using Bayesian credibility.

17.29 (*) Each of two classes, A and B, has the same number of risks. In class A, the number of claims per risk per year has mean $\frac{1}{6}$ and variance $\frac{5}{36}$, while the amount of a single claim has mean 4 and variance 20. In class B, the number of claims per risk per year has mean $\frac{5}{6}$ and variance $\frac{5}{36}$, while the amount of a single claim has mean 2 and variance 5. A risk is selected at random from one of the two classes and is observed for four years.

Table 17.5   The data for Exercise 17.30.

Outcome, T	$\Pr(X_1 = T)$	Bühlmann estimate of $E(X_2\|X_1 = T)$	Bayesian estimate of $E(X_2\|X_1 = T)$
1	1/3	2.72	2.6
8	1/3	7.71	7.8
12	1/3	10.57	–

(a) Determine the value of Z for Bühlmann credibility for the observed pure premium.

(b) Suppose that the pure premium calculated from the four observations is 0.25. Determine the Bühlmann credibility estimate for the risk's pure premium.

17.30 (*) Let X_1 be the outcome of a single trial and let $E(X_2|X_1)$ be the expected value of the outcome of a second trial. You are given the information in Table 17.5. Determine the Bayesian estimate for $E(X_2|X_1 = 12)$.

17.31 Consider the situation of Exercise 13.25.

(a) Determine the expected number of claims in the second year using Bayesian credibility.

(b) (*) Determine the expected number of claims in the second year using Bühlmann credibility.

17.32 Consider the situation of Exercise 13.26.

(a) Use Bayesian credibility to determine the expected number of claims in the second year.

(b) Use Bühlmann credibility to determine the expected number of claims in the second year.

17.33 Two spinners, A_1 and A_2, are used to determine the number of claims. For spinner A_1, there is a 0.15 probability of one claim and 0.85 of no claim. For spinner A_2, there is a 0.05 probability of one claim and 0.95 of no claim. If there is a claim, one of two spinners, B_1 and B_2, is used to determine the amount. Spinner B_1 produces a claim of 20 with probability 0.8 and 40 with probability 0.2. Spinner B_2 produces a claim of 20 with probability 0.3 and 40 with probability 0.7. A spinner is selected at random from each of A_1, A_2 and from B_1, B_2. Three observations from the selected pair yield claim amounts of 0, 20, and 0.

(a) (*) Use Bühlmann credibility to separately estimate the expected number of claims and the expected severity. Use these estimates to estimate the expected value of the next observation from the same pair of spinners.

(b) Use Bühlmann credibility once on the three observations to estimate the expected value of the next observation from the same pair of spinners.

(c) (*) Repeat parts (a) and (b) using Bayesian estimation.

(d) (*) For the same selected pair of spinners, determine

$$\lim_{n\to\infty} \mathrm{E}(X_n | X_1 = X_2 = \cdots = X_{n-1} = 0).$$

17.34 (*) A portfolio of risks is such that all risks are normally distributed. Those of type A have a mean of 0.1 and a standard deviation of 0.03. Those of type B have a mean of 0.5 and a standard deviation of 0.05. Those of type C have a mean of 0.9 and a standard deviation of 0.01. There are an equal number of each type of risk. The observed value for a single risk is 0.12. Determine the Bayesian estimate of the same risk's expected value.

17.35 (*) You are given the following:

1. The conditional distribution $f_{X|\Theta}(x|\theta)$ is a member of the linear exponential family.

2. The prior distribution $\pi(\theta)$ is a conjugate prior for $f_{X|\Theta}(x|\theta)$.

3. $\mathrm{E}(X) = 1$.

4. $\mathrm{E}(X|X_1 = 4) = 2$, where X_1 is the value of a single observation.

5. The expected value of the process variance $\mathrm{E}[\mathrm{Var}(X|\Theta)] = 3$.

   Determine the variance of the hypothetical means $\mathrm{Var}[\mathrm{E}(X|\Theta)]$.

17.36 (*) You are given the following:

1. X is a random variable with mean μ and variance v.

2. μ is a random variable with mean 2 and variance 4.

3. v is a random variable with mean 8 and variance 32.

   Determine the value of the Bühlmann credibility factor Z after three observations of X.

17.37 The amount of an individual claim has an exponential distribution with pdf $f_{Y|\Lambda}(y|\lambda) = \lambda^{-1} e^{-y/\lambda}$, $y, \lambda > 0$. The parameter λ has an inverse gamma distribution with pdf $\pi(\lambda) = 400 \lambda^{-3} e^{-20/\lambda}$.

(a) (*) Determine the unconditional expected value, $\mathrm{E}(X)$.
(b) Suppose that two claims were observed with values 15 and 25. Determine the Bühlmann credibility estimate of the expected value of the next claim from the same insured.
(c) Repeat part (b), but determine the Bayesian credibility estimate.

17.38 The distribution of the number of claims is binomial with $n = 1$ and θ unknown. The parameter θ is distributed with mean 0.25 and variance 0.07. Determine the value of Z for a single observation using Bühlmann's credibility formula.

17.39 (*) Consider four marksmen. Each is firing at a target that is 100 feet away. The four targets are 2 feet apart (i.e. they lie on a straight line at positions 0, 2, 4, and 6 in feet). The marksmen miss to the left or right, never high or low. Each marksman's shot follows a normal distribution with mean at his target and a standard deviation that is a constant times the distance to the target. At 100 feet, the standard deviation is 3 feet. By observing where an unknown marksman's shot hits the straight line, you are to estimate the location of the next shot by the same marksman.

(a) Determine the Bühlmann credibility assigned to a single shot of a randomly selected marksman.

(b) Which of the following will increase Bühlmann credibility the most?

    i. Revise the targets to 0, 4, 8, and 12.

    ii. Move the marksmen to 60 feet from the targets.

    iii. Revise the targets to 2, 2, 10, and 10.

    iv. Increase the number of observations from the same marksman to three.

    v. Move two of the marksmen to 50 feet from the targets and increase the number of observations from the same marksman to two.

17.40 (*) Risk 1 produces claims of amounts 100, 1,000, and 20,000 with probabilities 0.5, 0.3, and 0.2, respectively. For risk 2, the probabilities are 0.7, 0.2, and 0.1. Risk 1 is twice as likely as risk 2 of being observed. A claim of 100 is observed, but the observed risk is unknown.

(a) Determine the Bayesian credibility estimate of the expected value of the second claim amount from the same risk.

(b) Determine the Bühlmann credibility estimate of the expected value of the second claim amount from the same risk.

17.41 (*) You are given the following:

1. The number of claims for a single insured follows a Poisson distribution with mean M.

2. The amount of a single claim has an exponential distribution with pdf

$$f_{X|\Lambda}(x|\lambda) = \lambda^{-1}e^{-x/\lambda}, \ x, \lambda > 0.$$

3. M and Λ are independent.

4. $E(M) = 0.10$ and $Var(M) = 0.0025$.

5. $E(\Lambda) = 1,000$ and $Var(\Lambda) = 640,000$.

6. The number of claims and the claim amounts are independent.

(a) Determine the expected value of the pure premium's process variance for a single risk.

(b) Determine the variance of the hypothetical means for the pure premium.

17.42 (*) The number of claims has a Poisson distribution. For 75% of risks, $\lambda = 1$, and for 25% of risks, $\lambda = 3$. A randomly selected risk had r claims in year 1. The Bayesian estimate of the expected number of claims in year 2 is 2.98. Determine the Bühlmann estimate of the expected number of claims in year 2.

17.43 (*) Claim sizes have an exponential distribution with mean θ. For 80% of risks, $\theta = 8$, and for 20% of risks, $\theta = 2$. A randomly selected policy had a claim of size 5 in year 1. Determine both the Bayesian and Bühlmann estimates of the expected claim size in year 2.

17.44 (*) A portfolio has 100 risks with identical and independent numbers of claims. The number of claims for one risk has a Poisson distribution with mean λ. The prior distribution is $\pi(\lambda) = (50\lambda)^4 e^{-50\lambda}/(6\lambda)$, $\lambda > 0$. During year 1, 90 risks had 0 claims, 7 had 1 claim, 2 had 2 claims, and 1 had 3 claims. Determine both the Bayesian and Bühlmann estimates of the expected number of claims for the portfolio in year 2.

17.45 (*) For a portfolio of risks, all members' aggregate losses per year per exposure have a normal distribution with a standard deviation of 1,000. For these risks, 60% have a mean of 2,000, 30% have a mean of 3,000, and 10% have a mean of 4,000. A randomly selected risk had the following experience over three years. In year 1, there were 24 exposures with total losses of 24,000. In year 2, there were 30 exposures with total losses of 36,000. In year 3, there were 26 exposures with total losses of 28,000. Determine the Bühlmann–Straub estimate of the mean aggregate loss per year per exposure for year 4.

17.46 (*) The number of claims for each policyholder has a binomial distribution with $m = 8$ and q unknown. The prior distribution of q is beta with parameters a unknown, $b = 9$, and $\theta = 1$. A randomly selected policyholder had two claims in year 1 and k claims in year 2. Based on the year 1 experience, the Bayesian estimate of the expected number of claims in year 2 is 2.54545. Based on years 1 and 2, the Bayesian estimate of the expected number of claims in year 3 is 3.73333. Determine k.

17.47 In Example 17.13, if $\rho = 0$, then $Z = 0$, and the estimator is μ. That is, the data should be ignored. However, as ρ increases toward 1, Z increases to 1, and the sample mean becomes the preferred predictor of X_{n+1}. Explain why this is a reasonable result.

17.48 In the following, let the random vector $\mathbf{X}$ represent all the past data and let X_{n+1} represent the next observation. Let $g(\mathbf{X})$ be any function of the past data.

(a) Prove that the following is true:

$$\mathrm{E}\left\{[X_{n+1} - g(\mathbf{X})]^2\right\} = \mathrm{E}\{[X_{n+1} - \mathrm{E}(X_{n+1}|\mathbf{X})]^2\}$$
$$+ \mathrm{E}\{[\mathrm{E}(X_{n+1}|\mathbf{X}) - g(\mathbf{X})]^2\},$$

where the expectation is taken over $(X_{n+1}, \mathbf{X})$.

(b) Show that setting $g(\mathbf{X})$ equal to the Bayesian premium (the mean of the predictive distribution) minimizes the expected squared error,

$$\mathrm{E}\left\{[X_{n+1} - g(\mathbf{X})]^2\right\}.$$

(c) Show that, if $g(\mathbf{X})$ is restricted to be a linear function of the past data, then the expected squared error is minimized by the credibility premium.

18

EMPIRICAL BAYES PARAMETER ESTIMATION

18.1 Introduction

In Chapter 17, a modeling methodology was proposed that suggests the use of either the Bayesian or the credibility premium as a way to incorporate past data into the prospective rate. There is a practical problem associated with the use of these models that has not yet been addressed.

In the examples seen so far, we have been able to obtain numerical values for the quantities of interest because the input distributions $f_{X_j|\Theta}(x_j|\theta)$ and $\pi(\theta)$ have been assumed to be known. These examples, while useful for illustration of the methodology, can hardly be expected to accurately represent the business of an insurance portfolio. More practical models of necessity involve the use of parameters that must be chosen to ensure a close agreement between the model and reality. Examples of this include: the Poisson–gamma model (Example 17.1), where the gamma parameters α and β need to be selected; or the Bühlmann or Bühlmann–Straub parameters μ, v, and a. The assignment of numerical values to the Bayesian or credibility premium requires that these parameters be replaced by numerical values.

In general, the unknown parameters are those associated with the structure density $\pi(\theta)$ and, hence, we refer to these as **structural parameters**. The terminology we use follows the Bayesian framework of the previous chapter. Strictly speaking, in the Bayesian context all structural parameters are assumed known and there is no need for estimation. An example is the Poisson–gamma, where our prior information about the structural density is quantified by the choice of $\alpha = 36$ and $\beta = \frac{1}{240}$. For our purposes, this fully Bayesian approach is often unsatisfactory (e.g. when there is little or no prior information available, such as with a new line of insurance) and we may need to use the data at hand to estimate the structural (prior) parameters. This approach is called **empirical Bayes estimation**.

We refer to the situation in which $\pi(\theta)$ and $f_{X_j|\Theta}(x_j|\theta)$ are left largely unspecified (e.g. in the Bühlmann or Bühlmann–Straub models, where only the first two moments need be known) as the **nonparametric** case. This situation is dealt with in Section 18.2. If $f_{X_j|\Theta}(x_j|\theta)$ is assumed to be of parametric form (e.g. Poisson, normal, etc.) but not $\pi(\theta)$, then we refer to the problem as being of a **semiparametric** nature and it is considered in Section 18.3. A third, and technically more difficult, case is called **fully parametric**, where both $f_{X_j|\Theta}(x_j|\theta)$ and $\pi(\theta)$ are assumed to be of parametric form. That case is not covered.

This decision as to whether or not to select a parametric model depends partially on the situation at hand and partially on the judgment and knowledge of the person doing the analysis. For example, an analysis based on claim counts might involve the assumption that $f_{X_j|\Theta}(x_j|\theta)$ is of Poisson form, whereas the choice of a parametric model for $\pi(\theta)$ may not be reasonable.

Any parametric assumptions should be reflected (as far as possible) in parametric estimation. For example, in the Poisson case, because the mean and variance are equal, the same estimate would normally be used for both. Nonparametric estimators would normally be no more efficient than estimators appropriate for the parametric model selected, assuming that the model selected is appropriate. This notion is relevant for the decision as to whether to select a parametric model.

Finally, nonparametric models have the advantage of being appropriate for a wide variety of situations, a fact that may well eliminate the extra burden of a parametric assumption (often a stronger assumption than is reasonable).

In this section, the data are assumed to be of the following form. For each of $r \geq 1$ policyholders, we have the observed losses per unit of exposure $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})^T$ for $i = 1, \ldots, r$. The random vectors $\{\mathbf{X}_i, i = 1, \ldots, r\}$ are assumed to be statistically independent (experience of different policyholders is assumed to be independent). The (unknown) risk parameter for the ith policyholder is θ_i, $i = 1, \ldots, r$, and it is assumed further that $\theta_1, \ldots, \theta_r$ are realizations of the i.i.d. random variables Θ_i with structural density $\pi(\theta_i)$. For fixed i, the (conditional) random variables $X_{ij}|\Theta_i$ are assumed to be independent with pf $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$, $j = 1, \ldots, n_i$.

Two particularly common cases produce this data format. The first is classification ratemaking or experience rating. In either, i indexes the classes or groups and j indexes the individual members. The second case is like the first, where i continues to index the class or group, but now j is the year and the observation is the average loss for that year. An example of the second setting is Meyers [86], where $i = 1, \ldots, 319$ employment classifications are studied over $j = 1, 2, 3$ years. Regardless of the potential settings, we refer to the r entities as policyholders.

There may also be a known exposure vector $\mathbf{m}_i = (m_{i1}, m_{i2}, \ldots, m_{in_i})^T$ for policyholder i, where $i = 1, \ldots, r$. If not (and if it is appropriate), we may set $m_{ij} = 1$ in what follows

for all i and j. For notational convenience, let

$$m_i = \sum_{j=1}^{n_i} m_{ij}, \quad i = 1, \dots, r$$

be the total past exposure for policyholder i, and let

$$\bar{X}_i = \frac{1}{m_i} \sum_{j=1}^{n_i} m_{ij} X_{ij}, \quad i = 1, \dots, r$$

be the past weighted average loss experience. Furthermore, the total exposure is

$$m = \sum_{i=1}^{r} m_i = \sum_{i=1}^{r} \sum_{j=1}^{n_i} m_{ij}$$

and the overall weighted average losses are

$$\bar{X} = \frac{1}{m} \sum_{i=1}^{r} m_i \bar{X}_i = \frac{1}{m} \sum_{i=1}^{r} \sum_{j=1}^{n_i} m_{ij} X_{ij}. \tag{18.1}$$

The parameters that need to be estimated depend on what is assumed about the distributions $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$ and $\pi(\theta)$.

For the Bühlmann–Straub formulation, there are additional quantities of interest. The hypothetical mean (assumed not to depend on j) is

$$E(X_{ij}|\Theta_i = \theta_i) = \mu(\theta_i)$$

and the process variance is

$$\mathrm{Var}(X_{ij}|\Theta_i = \theta_i) = \frac{v(\theta_i)}{m_{ij}}.$$

The structural parameters are

$$\mu = E[\mu(\Theta_i)], \quad v = E[v(\Theta_i)],$$

and

$$a = \mathrm{Var}[\mu(\Theta_i)].$$

The approach is to estimate μ, v, and a (when unknown) from the data. The credibility premium for next year's losses (per exposure unit) for policyholder i is

$$Z_i \bar{X}_i + \left(1 - Z_i\right) \mu, \quad i = 1, \dots, r, \tag{18.2}$$

where

$$Z_i = \frac{m_i}{m_i + k}, \quad k = \frac{v}{a}.$$

If estimators of μ, v, and a are denoted by $\hat{\mu}$, $\hat{v}$, and $\hat{a}$, respectively, then we would replace the credibility premium (18.2) by its estimator

$$\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i)\hat{\mu}, \tag{18.3}$$

where

$$\hat{Z}_i = \frac{m_i}{m_i + \hat{k}}, \quad \hat{k} = \frac{\hat{v}}{\hat{a}}.$$

Note that, even if $\hat{v}$ and $\hat{a}$ are unbiased estimators of v and a, the same cannot be said of $\hat{k}$ and $\hat{Z}_i$. Finally, the credibility premium to cover all m_{i,n_i+1} exposure units for policyholder i in the next year would be (18.3) multiplied by m_{i,n_i+1}.

18.2   Nonparametric Estimation

In this section, we consider unbiased estimation of μ, v, and a. To illustrate the ideas, let us begin with the following simple Bühlmann-type example.

■   **EXAMPLE 18.1**

Suppose that $n_i = n > 1$ for all i and $m_{ij} = 1$ for all i and j. That is, for policyholder i, we have the loss vector

$$\mathbf{X}_i = \left(X_{i1}, \dots, X_{in} \right)^T, \quad i = 1, \dots, r.$$

Furthermore, conditional on $\Theta_i = \theta_i$, X_{ij} has mean

$$\mu(\theta_i) = \mathrm{E}(X_{ij}|\Theta_i = \theta_i)$$

and variance

$$v(\theta_i) = \mathrm{Var}(X_{ij}|\Theta_i = \theta_i),$$

and $X_{i1}, \dots, X_{in}$ are independent (conditionally). Also, different policyholders' past data are independent, so that if $i \neq s$, then X_{ij} and X_{st} are independent. In this case,

$$\bar{X}_i = n^{-1} \sum_{j=1}^{n} X_{ij} \text{ and } \bar{X} = r^{-1} \sum_{i=1}^{r} \bar{X}_i = (rn)^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n} X_{ij}.$$

Determine unbiased estimators of the Bühlmann quantities.

An unbiased estimator of μ is

$$\hat{\mu} = \bar{X}$$

because

$$\mathrm{E}(\hat{\mu}) = (rn)^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n} \mathrm{E}(X_{ij}) = (rn)^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n} \mathrm{E}[\mathrm{E}(X_{ij}|\Theta_i)]$$

$$= (rn)^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n} \mathrm{E}[\mu(\Theta_i)] = (rn)^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n} \mu = \mu.$$

For estimation of v and a, we use the following result. Suppose that $Y_1, Y_2, \dots, Y_k$ are independent (but not necessarily identically distributed) random variables with

identical means and variances $\mu = E(Y_j)$ and $\sigma^2 = \text{Var}(Y_j)$. Let $\bar{Y} = k^{-1} \sum_{j=1}^{k} Y_j$. Then,

$$E(\bar{Y}) = k^{-1} \sum_{j=1}^{k} E(Y_j) = \mu,$$

$$\text{Var}(\bar{Y}) = k^{-2} \sum_{j=1}^{k} \text{Var}(Y_j) = \frac{\sigma^2}{k}.$$

Next, consider the statistic $\sum_{j=1}^{k}(Y_j - \bar{Y})^2$. It can be rewritten

$$\sum_{j=1}^{k}(Y_j - \bar{Y})^2 = \sum_{j=1}^{k}[(Y_j - \mu) + (\mu - \bar{Y})]^2$$

$$= \sum_{j=1}^{k}[(Y_j - \mu)^2 + 2(Y_j - \mu)(\mu - \bar{Y}) + (\mu - \bar{Y})^2]$$

$$= \sum_{j=1}^{k}(Y_j - \mu)^2 + 2(\mu - \bar{Y})\sum_{j=1}^{k}(Y_j - \mu) + \sum_{j=1}^{k}(\bar{Y} - \mu)^2$$

$$= \sum_{j=1}^{k}(Y_j - \mu)^2 + 2(\mu - \bar{Y})(k\bar{Y} - k\mu) + k(\bar{Y} - \mu)^2,$$

which simplifies to

$$\sum_{j=1}^{k}(Y_j - \bar{Y})^2 = \sum_{j=1}^{k}(Y_j - \mu)^2 - k(\bar{Y} - \mu)^2. \tag{18.4}$$

Taking expectations of both sides yields

$$E\left[\sum_{j=1}^{k}(Y_j - \bar{Y})^2\right] = \sum_{j=1}^{k} E[(Y_j - \mu)^2] - kE[(\bar{Y} - \mu)^2]$$

$$= \sum_{j=1}^{k} \text{Var}(Y_j) - k\,\text{Var}(\bar{Y})$$

$$= k\sigma^2 - k\left(\frac{\sigma^2}{k}\right) = (k-1)\sigma^2.$$

Therefore,

$$E\left[\frac{1}{k-1}\sum_{j=1}^{k}(Y_j - \bar{Y})^2\right] = \sigma^2, \tag{18.5}$$

and thus $\sum_{j=1}^{k}(Y_j - \bar{Y})^2/(k-1)$ is an unbiased estimator of the variance of Y_j.

To estimate v, consider

$$\hat{v}_i = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2. \tag{18.6}$$

Recall that for fixed i, the random variables $X_{i1}, \dots, X_{in}$ are independent, conditional on $\Theta_i = \theta_i$. Thus, $\hat{v}_i$ is an unbiased estimate of $\text{Var}(X_{ij}|\Theta_i = \theta_i) = v(\theta_i)$. Unconditionally,

$$E(\hat{v}_i) = E[E(\hat{v}_i|\Theta_i)] = E[v(\Theta_i)] = v,$$

and $\hat{v}_i$ is unbiased for v. Hence, an unbiased estimator of v is

$$\hat{v} = \frac{1}{r} \sum_{i=1}^{r} \hat{v}_i = \frac{1}{r(n-1)} \sum_{i=1}^{r} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2. \tag{18.7}$$

We now turn to estimation of the parameter a. Begin with

$$E(\bar{X}_i|\Theta_i = \theta_i) = n^{-1} \sum_{j=1}^{n} E(X_{ij}|\Theta_i = \theta_i) = n^{-1} \sum_{j=1}^{n} \mu(\theta_i) = \mu(\theta_i).$$

Thus,

$$E(\bar{X}_i) = E[E(\bar{X}_i|\Theta_i)] = E[\mu(\Theta_i)] = \mu$$

and

$$\text{Var}(\bar{X}_i) = \text{Var}[E(\bar{X}_i|\Theta_i)] + E[\text{Var}(\bar{X}_i|\Theta_i)]$$

$$= \text{Var}[\mu(\Theta_i)] + E\left[\frac{v(\Theta_i)}{n}\right] = a + \frac{v}{n}.$$

Therefore, $\bar{X}_1, \dots, \bar{X}_r$ are independent with common mean μ and common variance $a + v/n$. Their sample average is $\bar{X} = r^{-1} \sum_{i=1}^{r} \bar{X}_i$. Consequently, an unbiased estimator of $a + v/n$ is $(r-1)^{-1} \sum_{i=1}^{r} (\bar{X}_i - \bar{X})^2$. Because we already have an unbiased estimator of v as just given, an unbiased estimator of a is

$$\hat{a} = \frac{1}{r-1} \sum_{i=1}^{r} (\bar{X}_i - \bar{X})^2 - \frac{\hat{v}}{n}$$

$$= \frac{1}{r-1} \sum_{i=1}^{r} (\bar{X}_i - \bar{X})^2 - \frac{1}{rn(n-1)} \sum_{i=1}^{r} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2. \tag{18.8}$$

☐

These estimators might look familiar. Consider a one-factor analysis of variance in which each policyholder represents a treatment. The estimator for v (18.7) is *the within* (also called the error) *mean square*. The first term in the estimator for a (18.8) is *the between* (also called the treatment) *mean square* divided by n. The hypothesis that all treatments have the same mean is accepted when the between mean square is small relative to the within mean square – that is, when $\hat{a}$ is small relative to $\hat{v}$. But that relationship implies $\hat{Z}$ will be near zero and little credibility will be given to each $\bar{X}_i$. This is as it should be when the policyholders are essentially identical.

Due to the subtraction in (18.8), it is possible that $\hat{a}$ could be negative. When that happens, it is customary to set $\hat{a} = \hat{Z} = 0$. This case is equivalent to the F-test statistic in the analysis of variance being less than 1, a case that always leads to an acceptance of the hypothesis of equal means.

■ **EXAMPLE 18.2**

(*Example 18.1 continued*) As a numerical illustration, suppose that we have $r = 2$ policyholders with $n = 3$ years' experience for each. Let the losses be $\mathbf{x}_1 = (3, 5, 7)^T$ and $\mathbf{x}_2 = (6, 12, 9)^T$. Estimate the Bühlmann credibility premiums for each policyholder.

We have

$$\bar{X}_1 = \tfrac{1}{3}(3 + 5 + 7) = 5, \ \bar{X}_2 = \tfrac{1}{3}(6 + 12 + 9) = 9,$$

and so $\bar{X} = \tfrac{1}{2}(5 + 9) = 7$. Then, $\hat{\mu} = 7$. We next have

$$\hat{v}_1 = \tfrac{1}{2}[(3 - 5)^2 + (5 - 5)^2 + (7 - 5)^2] = 4,$$
$$\hat{v}_2 = \tfrac{1}{2}[(6 - 9)^2 + (12 - 9)^2 + (9 - 9)^2] = 9,$$

and so $\hat{v} = \tfrac{1}{2}(4 + 9) = \tfrac{13}{2}$. Then,

$$\hat{a} = [(5 - 7)^2 + (9 - 7)^2] - \tfrac{1}{3}\hat{v} = \tfrac{35}{6}.$$

Next, $\hat{k} = \hat{v}/\hat{a} = \tfrac{39}{35}$ and the estimated credibility factor is $\hat{Z} = 3/(3 + \hat{k}) = \tfrac{35}{48}$. The estimated credibility premiums are

$$\hat{Z}\bar{X}_1 + (1 - \hat{Z})\hat{\mu} = \left(\tfrac{35}{48}\right)(5) + \left(\tfrac{13}{48}\right)(7) = \tfrac{133}{24},$$
$$\hat{Z}\bar{X}_2 + (1 - \hat{Z})\hat{\mu} = \left(\tfrac{35}{48}\right)(9) + \left(\tfrac{13}{48}\right)(7) = \tfrac{203}{24}$$

for policyholders 1 and 2, respectively. □

We now turn to the more general Bühlmann–Straub setup described earlier in this section. We have $E(X_{ij}) = E[E(X_{ij}|\Theta_i)] = E[\mu(\Theta_i)] = \mu$. Thus,

$$E(\bar{X}_i|\Theta_i) = \sum_{j=1}^{n_i} \frac{m_{ij}}{m_i} E(X_{ij}|\Theta_i) = \sum_{j=1}^{n_i} \frac{m_{ij}}{m_i} \mu(\Theta_i) = \mu(\Theta_i),$$

implying that

$$E(\bar{X}_i) = E[E(\bar{X}_i|\Theta_i)] = E[\mu(\Theta_i)] = \mu.$$

Finally,

$$E(\bar{X}) = \frac{1}{m} \sum_{i=1}^{r} m_i E(\bar{X}_i) = \frac{1}{m} \sum_{i=1}^{r} m_i \mu = \mu$$

and so an obvious unbiased estimator of μ is

$$\hat{\mu} = \bar{X}. \tag{18.9}$$

To estimate v and a in the Bühlmann–Straub framework, a more general statistic than that in (18.5) is needed. The following example provides the needed results.

■ **EXAMPLE 18.3**

Suppose that $X_1, \ldots, X_n$ are independent with common mean $\mu = \mathrm{E}(X_j)$ and variance $\mathrm{Var}(X_j) = \beta + \alpha/m_j$, $\alpha, \beta > 0$ and all $m_j \geq 1$. The values of m_j are assumed to be known. Let $m = \sum_{j=1}^{n} m_j$ and consider the estimators

$$\bar{X} = \frac{1}{m} \sum_{j=1}^{n} m_j X_j \text{ and } \hat{\mu}_1 = \frac{1}{n} \sum_{j=1}^{n} X_j.$$

Show that both estimators are unbiased for μ and then compare their MSEs. Also obtain the expected value of a sum of squares that may be useful for estimating α and β.

First, consider $\bar{X}$:

$$\mathrm{E}(\bar{X}) = m^{-1} \sum_{j=1}^{n} m_j \mathrm{E}(X_j) = m^{-1} \sum_{j=1}^{n} m_j \mu = \mu,$$

$$\mathrm{Var}(\bar{X}) = m^{-2} \sum_{j=1}^{n} m_j^2 \mathrm{Var}(X_j)$$

$$= m^{-2} \sum_{j=1}^{n} m_j^2 \left(\beta + \frac{\alpha}{m_j} \right)$$

$$= \alpha m^{-1} + \beta m^{-2} \sum_{j=1}^{n} m_j^2.$$

The estimator $\hat{\mu}_1$ is easily shown to be unbiased. We also have

$$\mathrm{Var}(\hat{\mu}_1) = n^{-2} \sum_{j=1}^{n} \mathrm{Var}(X_j)$$

$$= n^{-2} \sum_{j=1}^{n} \left(\beta + \frac{\alpha}{m_j} \right)$$

$$= \beta n^{-1} + n^{-2} \alpha \sum_{j=1}^{n} m_j^{-1}.$$

We now consider the relative ranking of these variances (because both estimators are unbiased, their MSEs equal their variances, so it is sufficient to rank the variances). It turns out that it is not possible to order $\mathrm{Var}(\hat{\mu}_1)$ and $\mathrm{Var}(\bar{X})$. The difference is

$$\mathrm{Var}(\bar{X}) - \mathrm{Var}(\hat{\mu}_1) = \alpha \left(m^{-1} - n^{-2} \sum_{j=1}^{n} m_j^{-1} \right) + \beta \left(m^{-2} \sum_{j=1}^{n} m_j^2 - n^{-1} \right).$$

The coefficient of β must be nonnegative. To see this, note that

$$\frac{1}{n} \sum_{j=1}^{n} m_j^2 \geq \left(\frac{1}{n} \sum_{j=1}^{n} m_j \right)^2 = \frac{m^2}{n^2}$$

(the left-hand side is like a sample second moment and the right-hand side is like the square of the sample mean) and multiply both sides by nm^{-2}. To show that the coefficient of α must be nonpositive, note that

$$\frac{n}{\sum_{j=1}^{n} m_j^{-1}} \leq \frac{1}{n} \sum_{j=1}^{n} m_j = \frac{m}{n}$$

(the harmonic mean is always less than or equal to the arithmetic mean), then multiply both sides by n, and then invert both sides. Therefore, by suitable choice of α and β, the difference in the variances can be made positive or negative.

With regard to a sum of squares, consider

$$
\begin{aligned}
\sum_{j=1}^{n} m_j (X_j - \bar{X})^2 &= \sum_{j=1}^{n} m_j (X_j - \mu + \mu - \bar{X})^2 \\
&= \sum_{j=1}^{n} m_j (X_j - \mu)^2 + 2 \sum_{j=1}^{n} m_j (X_j - \mu)(\mu - \bar{X}) \\
&\quad + \sum_{j=1}^{n} m_j (\mu - \bar{X})^2 \\
&= \sum_{j=1}^{n} m_j (X_j - \mu)^2 + 2(\mu - \bar{X}) \sum_{j=1}^{n} m_j (X_j - \mu) \\
&\quad + m(\mu - \bar{X})^2 \\
&= \sum_{j=1}^{n} m_j (X_j - \mu)^2 + 2(\mu - \bar{X})m(\bar{X} - \mu) + m(\mu - \bar{X})^2 \\
&= \sum_{j=1}^{n} m_j (X_j - \mu)^2 - m(\bar{X} - \mu)^2. \tag{18.10}
\end{aligned}
$$

Taking expectations yields

$$
\begin{aligned}
E\left[\sum_{j=1}^{n} m_j (X_j - \bar{X})^2\right] &= \sum_{j=1}^{n} m_j E[(X_j - \mu)^2] - mE[(\bar{X} - \mu)^2] \\
&= \sum_{j=1}^{n} m_j \ \text{Var}(X_j) - m \, \text{Var}(\bar{X}) \\
&= \sum_{j=1}^{n} m_j \left(\beta + \frac{\alpha}{m_j}\right) - \beta \left(m^{-1} \sum_{j=1}^{n} m_j^2\right) - \alpha,
\end{aligned}
$$

and thus

$$
E\left[\sum_{j=1}^{n} m_j (X_j - \bar{X})^2\right] = \beta\left(m - m^{-1} \sum_{j=1}^{n} m_j^2\right) + \alpha(n - 1). \tag{18.11}
$$

In addition to being of interest in its own right, (18.11) provides an unbiased estimator in situations more general than in (18.5). The latter is recovered with the choice $\alpha = 0$

and $m_j = 1$ for $j = 1, 2, \ldots, n$, implying that $m = n$. Also, if $\beta = 0$, (18.11) allows us to derive an estimator of α when each X_j is the average of m_j independent observations, each with mean μ and variance α. In any event, the m_js (and hence m) are known. $\square$

We now return to the problem of estimation of v in the Bühlmann–Straub framework. Clearly, $E(X_{ij}|\Theta_i) = \mu(\Theta_i)$ and $\text{Var}(X_{ij}|\Theta_i) = v(\Theta_i)/m_{ij}$ for $j = 1, \ldots, n_i$. Consider

$$\hat{v}_i = \frac{\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{n_i - 1}, \quad i = 1, \ldots, r. \tag{18.12}$$

Condition on Θ_i and use (18.11) with $\beta = 0$ and $\alpha = v(\Theta_i)$. Then, $E(\hat{v}_i|\Theta_i) = v(\Theta_i)$, which implies that, unconditionally,

$$E(\hat{v}_i) = E[E(\hat{v}_i|\Theta_i)] = E[v(\Theta_i)] = v,$$

and so $\hat{v}_i$ is unbiased for v for $i = 1, \ldots, r$. Another unbiased estimator for v is then the weighted average $\hat{v} = \sum_{i=1}^r w_i \hat{v}_i$, where $\sum_{i=1}^r w_i = 1$. If we choose weights proportional to $n_i - 1$, we weight the original X_{ij}s by m_{ij}. That is, with $w_i = (n_i - 1) / \sum_{i=1}^r (n_i - 1)$, we obtain an unbiased estimator of v, namely

$$\hat{v} = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^r (n_i - 1)}. \tag{18.13}$$

We now turn to estimation of a. Recall that, for fixed i, the random variables $X_{i1}, \ldots, X_{in_i}$ are independent, conditional on Θ_i. Thus,

$$\text{Var}(\bar{X}_i|\Theta_i) = \sum_{j=1}^{n_i} \left(\frac{m_{ij}}{m_i}\right)^2 \text{Var}(X_{ij}|\Theta_i) = \sum_{j=1}^{n_i} \left(\frac{m_{ij}}{m_i}\right)^2 \frac{v(\Theta_i)}{m_{ij}}$$

$$= \frac{v(\Theta_i)}{m_i^2} \sum_{j=1}^{n_i} m_{ij} = \frac{v(\Theta_i)}{m_i}.$$

Then, unconditionally,

$$\text{Var}(\bar{X}_i) = \text{Var}[E(\bar{X}_i|\Theta_i)] + E[\text{Var}(\bar{X}_i|\Theta_i)]$$

$$= \text{Var}[\mu(\Theta_i)] + E\left[\frac{v(\Theta_i)}{m_i}\right] = a + \frac{v}{m_i}. \tag{18.14}$$

To summarize, $\bar{X}_1, \ldots, \bar{X}_r$ are independent with common mean μ and variances $\text{Var}(\bar{X}_i) = a + v/m_i$. Furthermore, $\bar{X} = m^{-1} \sum_{i=1}^r m_i \bar{X}_i$. Now, (18.11) may again be used with $\beta = a$ and $\alpha = v$ to yield

$$E\left[\sum_{i=1}^r m_i (\bar{X}_i - \bar{X})^2\right] = a\left(m - m^{-1} \sum_{i=1}^r m_i^2\right) + v(r - 1).$$

An unbiased estimator for a may be obtained by replacing v by an unbiased estimator $\hat{v}$ and "solving" for a. That is, an unbiased estimator of a is

$$\hat{a} = \left(m - m^{-1} \sum_{i=1}^r m_i^2\right)^{-1} \left[\sum_{i=1}^r m_i(\bar{X}_i - \bar{X})^2 - \hat{v}(r - 1)\right] \tag{18.15}$$

with $\hat{v}$ given by (18.13). An alternative form of (18.15) is given in Exercise 18.9.

Some remarks are in order at this point. Equations (18.9), (18.13), and (18.15) provide unbiased estimators for μ, v, and a, respectively. They are nonparametric, requiring no distributional assumptions. They are certainly not the only (unbiased) estimators that could be used, and it is possible that $\hat{a} < 0$. In this case, a is likely to be close to zero, and it makes sense to set $\hat{Z} = 0$. Furthermore, the ordinary Bühlmann estimators of Example 18.1 are recovered with $m_{ij} = 1$ and $n_i = n$. Finally, these estimators are essentially maximum likelihood estimators in the case where $X_{ij}|\Theta_i$ and Θ_i are both normally distributed, and thus the estimators have good statistical properties.

There is one problem with the use of the formulas just developed. In the past, the data from the ith policyholder were collected on an exposure of m_i. Total losses on all policyholders was $TL = \sum_{i=1}^{r} m_i \bar{X}_i$. If we had charged the credibility premium as previously given, the total premium would have been

$$TP = \sum_{i=1}^{r} m_i [\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i)\hat{\mu}]$$

$$= \sum_{i=1}^{r} m_i (1 - \hat{Z}_i)(\hat{\mu} - \bar{X}_i) + \sum_{i=1}^{r} m_i \bar{X}_i$$

$$= \sum_{i=1}^{r} m_i \frac{\hat{k}}{m_i + \hat{k}}(\hat{\mu} - \bar{X}_i) + \sum_{i=1}^{r} m_i \bar{X}_i.$$

It is often desirable for TL to equal TP, because any premium increases that will meet the approval of regulators will be based on the total claim level from past experience. While credibility adjustments make both practical and theoretical sense, it is usually a good idea to keep the total unchanged. Thus, we need

$$0 = \sum_{i=1}^{r} m_i \frac{\hat{k}}{m_i + \hat{k}}(\hat{\mu} - \bar{X}_i)$$

or

$$\hat{\mu} \sum_{i=1}^{r} \hat{Z}_i = \sum_{i=1}^{r} \hat{Z}_i \bar{X}_i$$

or

$$\hat{\mu} = \frac{\sum_{i=1}^{r} \hat{Z}_i \bar{X}_i}{\sum_{i=1}^{r} \hat{Z}_i}. \tag{18.16}$$

That is, rather than using (18.9) to compute $\hat{\mu}$, use a credibility-weighted average of the individual sample means. Either method provides an unbiased estimator (given the $\hat{Z}_i$s), but this latter one has the advantage of preserving total claims. It should be noted that when using (18.15), the value of $\bar{X}$ from (18.1) should still be used. It can also be derived by least squares arguments. Finally, from Example 18.3 and noting the form of $\mathrm{Var}(\bar{X}_j)$ in (18.14), the weights in (18.16) provide the smallest unconditional variance for $\hat{\mu}$.

■ **EXAMPLE 18.4**

Past data on two group policyholders are available and are given in Table 18.1. Determine the estimated credibility premium to be charged to each group in year 4.

Table 18.1   The data for Example 18.4.

	Policyholder	Year 1	Year 2	Year 3	Year 4
Total claims	1	—	10,000	13,000	—
Number in group		—	50	60	75
Total claims	2	18,000	21,000	17,000	—
Number in group		100	110	105	90

We first need to determine the average claims per person for each group in each past year. We have $n_1 = 2$ years experience for group 1 and $n_2 = 3$ for group 2. It is immaterial which past years' data we have for policyholder 1, so for notational purposes we choose

$$m_{11} = 50 \text{ and } X_{11} = \frac{10,000}{50} = 200.$$

Similarly,

$$m_{12} = 60 \text{ and } X_{12} = \frac{13,000}{60} = 216.67.$$

Then,

$$m_1 = m_{11} + m_{12} = 50 + 60 = 110,$$
$$\bar{X}_1 = \frac{10,000 + 13,000}{110} = 209.09.$$

For policyholder 2,

$$m_{21} = 100, \ X_{21} = \frac{18,000}{100} = 180,$$
$$m_{22} = 110, \ X_{22} = \frac{21,000}{110} = 190.91,$$
$$m_{23} = 105, \ X_{23} = \frac{17,000}{105} = 161.90.$$

Then,

$$m_2 = m_{21} + m_{22} + m_{23} = 100 + 110 + 105 = 315,$$
$$\bar{X}_2 = \frac{18,000 + 21,000 + 17,000}{315} = 177.78.$$

Now, $m = m_1 + m_2 = 110 + 315 = 425$. The overall mean is

$$\hat{\mu} = \bar{X} = \frac{10,000 + 13,000 + 18,000 + 21,000 + 17,000}{425} = 185.88.$$

The alternative estimate (18.16) of μ cannot be computed until later. Now,

$$\hat{v} = \frac{\begin{array}{c} 50(200 - 209.09)^2 + 60(216.67 - 209.09)^2 + 100(180 - 177.78)^2 \\ +110(190.91 - 177.78)^2 + 105(161.90 - 177.78)^2 \end{array}}{(2-1) + (3-1)}$$

$$= 17,837.87,$$

and so

$$\hat{a} = \frac{110(209.09 - 185.88)^2 + 315(177.78 - 185.88)^2 - (17{,}837.87)(1)}{425 - \left(110^2 + 315^2\right)/425}$$

$$= 380.76.$$

Then, $\hat{k} = \hat{v}/\hat{a} = 46.85$. The estimated credibility factors for the two policyholders are

$$\hat{Z}_1 = \frac{110}{110 + 46.85} = 0.70, \quad \hat{Z}_2 = \frac{315}{315 + 46.85} = 0.87.$$

Per individual, the estimated credibility premium for policyholder 1 is

$$\hat{Z}_1 \bar{X}_1 + (1 - \hat{Z}_1)\hat{\mu} = (0.70)(209.09) + (0.30)(185.88) = 202.13,$$

and so the total estimated credibility premium for the whole group is

$$75(202.13) = 15{,}159.75.$$

For policyholder 2,

$$\hat{Z}_2 \bar{X}_2 + (1 - \hat{Z}_2)\hat{\mu} = (0.87)(177.78) + (0.13)(185.88) = 178.83,$$

and the total estimated credibility premium is

$$90(178.83) = 16{,}094.70.$$

For the alternative estimator we would use

$$\hat{\mu} = \frac{0.70(209.09) + 0.87(177.78)}{0.70 + 0.87} = 191.74.$$

The credibility premiums are

$$0.70(209.09) + 0.30(191.74) = 203.89, \quad 0.87(177.78) + 0.13(191.74) = 179.59.$$

The total past credibility premium is $110(203.89) + 315(179.59) = 78{,}998.75$. Except for rounding error, this total matches the actual total losses of 79,000. ☐

The preceding analysis assumes that the parameters μ, v, and a are all unknown and need to be estimated, which may not always be the case. Also, it is assumed that $n_i > 1$ and $r > 1$. If $n_i = 1$, so that there is only one exposure unit's experience for policyholder i, it is difficult to obtain information on the process variance $v(\Theta_i)$ and, thus, on v. Similarly, if $r = 1$, there is only one policyholder, and it is difficult to obtain information on the variance of the hypothetical means a. In these situations, stronger assumptions are needed, such as knowledge of one or more of the parameters (e.g. the pure premium or manual rate μ, discussed in the following) or parametric assumptions that imply functional relationships between the parameters (discussed in Section 18.3).

To illustrate these ideas, suppose, for example, that the manual rate μ may be already known, but estimates of a and v may be needed. In that case, (18.13) can still be used to estimate v as it is unbiased whether μ is known or not. (Why is $\left[\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \mu)^2\right]/n_i$

not unbiased for v in this case?) Similarly, (18.15) is still an unbiased estimator for a. However, if μ is known, an alternative unbiased estimator for a is

$$\tilde{a} = \sum_{i=1}^{r} \frac{m_i}{m}(\bar{X}_i - \mu)^2 - \frac{r}{m}\hat{v},$$

where $\hat{v}$ is given by (18.13). To verify unbiasedness, note that

$$\begin{aligned}
\mathrm{E}(\tilde{a}) &= \sum_{i=1}^{r} \frac{m_i}{m}\mathrm{E}[(\bar{X}_i - \mu)^2] - \frac{r}{m}\mathrm{E}(\hat{v}) \\
&= \sum_{i=1}^{r} \frac{m_i}{m}\mathrm{Var}(\bar{X}_i) - \frac{r}{m}v \\
&= \sum_{i=1}^{r} \frac{m_i}{m}\left(a + \frac{v}{m_i}\right) - \frac{r}{m}v = a.
\end{aligned}$$

   If there are data on only one policyholder, an approach like this is necessary. Clearly, (18.12) provides an estimator for v based on data from policyholder i alone, and an unbiased estimator for a based on data from policyholder i alone is

$$\tilde{a}_i = (\bar{X}_i - \mu)^2 - \frac{\hat{v}_i}{m_i} = (\bar{X}_i - \mu)^2 - \frac{\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{m_i(n_i - 1)},$$

which is unbiased because $\mathrm{E}[(\bar{X}_i - \mu)^2] = \mathrm{Var}(\bar{X}_i) = a + v/m_i$ and $\mathrm{E}(\hat{v}_i) = v$.

■ **EXAMPLE 18.5**

For a group policyholder, we have the data as given in Table 18.2. If the manual rate per person is 500 per year, estimate the total credibility premium for year 3.

   In the preceding notation, we have (assuming for notational purposes that this group is policyholder i) $m_{i1} = 125$, $X_{i1} = 60{,}000/125 = 480$, $m_{i2} = 150$, $X_{i2} = 70{,}000/150 = 466.67$, $m_i = m_{i1} + m_{i2} = 275$, and $\bar{X}_i = (60{,}000 + 70{,}000)/275 = 472.73$. Then,

$$\hat{v}_i = \frac{125(480 - 472.73)^2 + 150(466.67 - 472.73)^2}{2 - 1} = 12{,}115.15,$$

and, with $\mu = 500$, $\tilde{a}_i = (472.73 - 500)^2 - (12{,}115.15/275) = 699.60$. We then estimate k by $\hat{v}_i/\tilde{a}_i = 17.32$. The estimated credibility factor is $m_i/(m_i + \hat{v}_i/\tilde{a}_i) = 275/(275 + 17.32) = 0.94$. The estimated credibility premium per person is then $0.94(472.73) + 0.06(500) = 474.37$, and the estimated total credibility premium for year 3 is $200(474.37) = 94{,}874$. □

Table 18.2   The data for Example 18.5.

	Year 1	Year 2	Year 3
Total claims	60,000	70,000	—
Number in group	125	150	200

It is instructive to note that estimation of the parameters a and v based on data from a single policyholder (as in Example 18.5) is not advised unless there is no alternative because the estimators $\hat{v}_i$ and $\tilde{a}_i$ have high variability. In particular, we are effectively estimating a from one observation ($\bar{X}_i$). It is strongly suggested that an attempt be made to obtain more data.

18.3 Semiparametric Estimation

In some situations it may be reasonable to assume a parametric form for the conditional distribution $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$. The situation at hand may suggest that such an assumption is reasonable or prior information may imply its appropriateness.

For example, in dealing with numbers of claims, it may be reasonable to assume that the number of claims $m_{ij}X_{ij}$ for policyholder i in year j is Poisson distributed with mean $m_{ij}\theta_i$ given $\Theta_i = \theta_i$. Thus $\mathrm{E}(m_{ij}X_{ij}|\Theta_i) = \mathrm{Var}(m_{ij}X_{ij}|\Theta_i) = m_{ij}\Theta_i$, implying that $\mu(\Theta_i) = v(\Theta_i) = \Theta_i$, and so $\mu = v$ in this case. Rather than use (18.13) to estimate v, we could use $\hat{\mu} = \bar{X}$ to estimate v.

■ **EXAMPLE 18.6**

In the past year, the distribution of automobile insurance policyholders by number of claims is given in Table 18.3. For each policyholder, obtain a credibility estimate for the number of claims next year based on the past year's experience, assuming a (conditional) Poisson distribution of number of claims for each policyholder.

Assume that we have $r = 1,875$ policyholders, $n_i = 1$ year experience on each, and exposures $m_{ij} = 1$. For policyholder i (where $i = 1, \dots, 1,875$), assume that $X_{i1}|\Theta_i = \theta_i$ is Poisson distributed with mean θ_i so that $\mu(\theta_i) = v(\theta_i) = \theta_i$ and $\mu = v$. As in Example 18.1,

$$\bar{X} = \frac{1}{1,875}\left(\sum_{i=1}^{1,875} X_{i1}\right)$$

$$= \frac{0(1,563) + 1(271) + 2(32) + 3(7) + 4(2)}{1,875} = 0.194.$$

Now,

$$\mathrm{Var}(X_{i1}) = \mathrm{Var}[\mathrm{E}(X_{i1}|\Theta_i)] + \mathrm{E}[\mathrm{Var}(X_{i1}|\Theta_i)]$$

$$= \mathrm{Var}[\mu(\Theta_i)] + \mathrm{E}[v(\Theta_i)] = a + v = a + \mu.$$

Table 18.3   The data for Example 18.6.

Number of claims	Number of insureds
0	1,563
1	271
2	32
3	7
4	2
Total	1,875

Thus, an unbiased estimator of $a + v$ is the sample variance

$$\frac{\sum_{i=1}^{1,875} \left(X_{i1} - \bar{X} \right)^2}{1,874} = \frac{\begin{aligned}&1{,}563(0 - 0.194)^2 + 271(1 - 0.194)^2 \\ &+ 32(2 - 0.194)^2 + 7(3 - 0.194)^2 + 2(4 - 0.194)^2\end{aligned}}{1,874}$$

$$= 0.226.$$

Consequently, $\hat{a} = 0.226 - 0.194 = 0.032$ and $\hat{k} = 0.194/0.032 = 6.06$, and the credibility factor Z is $1/(1 + 6.06) = 0.14$. The estimated credibility premium for the number of claims for each policyholder is $(0.14)X_{i1} + (0.86)(0.194)$, where X_{i1} is 0, 1, 2, 3, or 4, depending on the policyholder.   □

Note in this case that $v = \mu$ identically, so that only one year's experience per policyholder is needed.

■ **EXAMPLE 18.7**

Suppose we are interested in the probability that an individual in a group makes a claim (e.g. group life insurance), and the probability is believed to vary by policyholder. Then, $m_{ij} X_{ij}$ could represent the number of the m_{ij} individuals in year j for policyholder i who made a claim. Develop a credibility model for this situation.

If the claim probability is θ_i for policyholder i, then a reasonable model to describe this effect is that $m_{ij} X_{ij}$ is binomially distributed with parameters m_{ij} and θ_i, given $\Theta_i = \theta_i$. Then,

$$E(m_{ij} X_{ij} | \Theta_i) = m_{ij} \Theta_i \quad \text{and} \quad \text{Var}(m_{ij} X_{ij} | \Theta_i) = m_{ij} \Theta_i (1 - \Theta_i),$$

and so $\mu(\Theta_i) = \Theta_i$ with $v(\Theta_i) = \Theta_i(1 - \Theta_i)$. Thus

$$\mu = E(\Theta_i), \quad v = \mu - E[(\Theta_i)^2],$$
$$a = \text{Var}(\Theta_i) = E[(\Theta_i)^2] - \mu^2 = \mu - v - \mu^2.$$

□

In these examples, there is a functional relationship between the parameters μ, v, and a that follows from the parametric assumptions made, and this often facilitates estimation of parameters.

18.4 Notes and References

In this section, a simple approach is employed to find parameter estimates. No attempt is made to find optimum estimators in the sense of minimum variance. A good deal of research has been done on this problem. For more details and further references, see Goovaerts and Hoogstad [46].

18.5 Exercises

18.1 Past claims data on a portfolio of policyholders are given in Table 18.4. Estimate the Bühlmann credibility premium for each of the three policyholders for year 4.

Table 18.4   The data for Exercise 18.1.

Policyholder	Year 1	Year 2	Year 3
1	750	800	650
2	625	600	675
3	900	950	850

Table 18.5   The data for Exercise 18.2.

	Policyholder	Year 1	Year 2	Year 3	Year 4
Claims	1	—	20,000	25,000	—
Number in group		—	100	120	110
Claims	2	19,000	18,000	17,000	—
Number in group		90	75	70	60
Claims	3	26,000	30,000	35,000	—
Number in group		150	175	180	200

18.2 Past data on a portfolio of group policyholders are given in Table 18.5. Estimate the Bühlmann–Straub credibility premiums to be charged to each group in year 4.

18.3 For the situation in Exercise 16.3, estimate the Bühlmann credibility premium for the next year for the policyholder.

18.4 Consider the Bühlmann model in Example 18.1.

(a) Prove that $\mathrm{Var}(X_{ij}) = a + v$.

(b) If $\{X_{ij} : i = 1, \ldots, r \text{ and } j = 1, \ldots, n\}$ are unconditionally independent for all i and j, argue that an unbiased estimator of $a + v$ is

$$\frac{1}{nr - 1} \sum_{i=1}^{r} \sum_{j=1}^{n} (X_{ij} - \bar{X})^2.$$

(c) Prove the algebraic identity

$$\sum_{i=1}^{r} \sum_{j=1}^{n} (X_{ij} - \bar{X})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2 + n \sum_{i=1}^{r} (\bar{X}_i - \bar{X})^2.$$

(d) Show that, conditionally,

$$\mathrm{E}\left[\frac{1}{nr - 1} \sum_{i=1}^{r} \sum_{j=1}^{n} (X_{ij} - \bar{X})^2 \right] = (v + a) - \frac{n - 1}{nr - 1} a.$$

(e) Comment on the implications of (b) and (d).

Table 18.6 The data for Exercise 18.6.

Number of claims	Number of insureds
0	2,500
1	250
2	30
3	5
4	2
Total	2,787

18.5 Suppose that the random variables $Y_1, \dots, Y_n$ are independent, with

$$E(Y_j) = \gamma \quad \text{and} \quad \text{Var}(Y_j) = a_j + \sigma^2/b_j, \quad j = 1, 2, \dots, n.$$

Define $b = b_1 + b_2 + \cdots + b_n$ and $\bar{Y} = \sum_{j=1}^{n} \dfrac{b_j}{b} Y_j$. Prove that

$$E\left[\sum_{j=1}^{n} b_j(Y_j - \bar{Y})^2\right] = (n-1)\sigma^2 + \sum_{j=1}^{n} a_j\left(b_j - \frac{b_j^2}{b}\right).$$

18.6 The distribution of automobile insurance policyholders by number of claims is given in Table 18.6. Assuming a (conditional) Poisson distribution for the number of claims per policyholder, estimate the Bühlmann credibility premiums for the number of claims next year.

18.7 Suppose that, given Θ, $X_1, \dots, X_n$ are independently geometrically distributed with pf

$$f_{X_j|\Theta}(x_j|\theta) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{x_j}, \quad x_j = 0, 1, \dots.$$

(a) Show that $\mu(\theta) = \theta$ and $v(\theta) = \theta(1+\theta)$.

(b) Prove that $a = v - \mu - \mu^2$.

(c) Rework Exercise 18.6 assuming a (conditional) geometric distribution.

18.8 Suppose that

$$\Pr(m_{ij}X_{ij} = t_{ij}|\Theta_i = \theta_i) = \frac{(m_{ij}\theta_i)^{t_{ij}} e^{-m_{ij}\theta_i}}{t_{ij}!}$$

and

$$\pi(\theta_i) = \frac{1}{\mu}e^{-\theta_i/\mu}, \quad \theta_i > 0.$$

Write down the equation satisfied by the mle $\hat{\mu}$ of μ for Bühlmann–Straub-type data.

Table 18.7   The data for Exercise 18.10.

Number of claims	Number of insureds
0	200
1	80
2	50
3	10

18.9   (a) Prove the algebraic identity

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2.$$

(b) Use part (a) and (18.13) to show that (18.15) may be expressed as

$$\hat{a} = m_*^{-1} \left[\frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X})^2}{\sum_{i=1}^{r} n_i - 1} - \hat{v} \right],$$

where

$$m_* = \frac{\sum_{i=1}^{r} m_i \left(1 - \dfrac{m_i}{m} \right)}{\sum_{i=1}^{r} n_i - 1}.$$

18.10 (*) In a one-year period, a group of 340 insureds in a high-crime area submit 210 theft claims as given in Table 18.7. Each insured is assumed to have a Poisson distribution for the number of thefts, but the mean of such a distribution may vary from one insured to another. If a particular insured experienced two claims in the observation period, determine the Bühlmann credibility estimate for the number of claims for this insured in the next period.

18.11 (*) Three individual policyholders were observed for four years. Policyholder X had claims of 2, 3, 3, and 4. Policyholder Y had claims of 5, 5, 4, and 6. Policyholder Z had claims of 5, 5, 3, and 3. Use nonparametric empirical Bayes estimation to obtain estimated claim amounts for each policyholder in year 5.

18.12 (*) Two insureds own delivery vans. Insured A had two vans in year 1 and one claim, two vans in year 2 and one claim, and one van in year 3 with no claims. Insured B had no vans in year 1, three vans in year 2 and two claims, and two vans in year 3 and three claims. The number of claims for insured each year has a Poisson distribution. Use semiparametric empirical Bayes estimation to obtain the estimated number of claims for each insured in year 4.

18.13 (*) One hundred policies were in force for a five-year period. Each policyholder has a Poisson distribution for the number of claims, but the parameters may vary. During the five years, 46 policies had no claims, 34 had one claim, 13 had two claims, 5 had three claims, and 2 had four claims. For a policy with three claims in this period, use semiparametric empirical Bayes estimation to estimate the number of claims in year 6 for that policy.

SIMULATION

19

SIMULATION

19.1 Basics of Simulation

Simulation has had an on-again, off-again history in actuarial practice. For example, in the 1970s, aggregate loss calculations were commonly done by simulation, because the analytic methods available at the time were not adequate. However, the typical simulation often took a full day on the company's mainframe computer, a serious drag on resources. In the 1980s, analytic methods such as the recursive formula discussed in Chapter 9 and others were developed and found to be significantly faster and more accurate. Today, desktop computers have sufficient power to run complex simulations that allow for the analysis of models not amenable to analytic approaches.

In a similar vein, as investment vehicles become more complex, insurance contracts may include interest-sensitive components and financial guarantees related to stock market performance. These products must be analyzed on a stochastic basis. To accommodate these and other complexities, simulation has become the technique of choice.

In this chapter, we provide some illustrations of how simulation can be used to address some complex modeling problems in insurance, or as an alternative to other methods. It is not our intention to cover the subject in great detail but, rather, to give you an idea of

how simulation can help. Study of simulation texts such as Ripley [105] and Ross [108] provides many important additional insights. In addition, simulation can also be an aid in evaluating some of the statistical techniques covered in earlier chapters. This use of simulation will be covered here, with an emphasis on the bootstrap method.

19.1.1   The Simulation Approach

The beauty of simulation is that once a model is created, little additional creative thought is required.[1] When the goal is to determine values relating to the distribution of a random variable S, the entire process can be summarized in the following four steps:

1. Build a model for S that depends on random variables $X, Y, Z, \ldots$, where their distributions and any dependencies are known.

2. For $j = 1, \ldots, n$ generate pseudorandom values $x_j, y_j, z_j, \ldots$ and then compute s_j using the model from step 1.

3. The cdf of S may be approximated by $F_n(s)$, the empirical cdf based on the pseudo-random sample $s_1, \ldots, s_n$.

4. Compute quantities of interest, such as the mean, variance, percentiles, or probabilities, using the empirical cdf.

   Two questions remain. One is postponed until Section 19.3. The other one is: What does it mean to generate a pseudorandom variable? Consider a random variable X with cdf $F_X(x)$. This is the real random variable produced by some phenomenon of interest. For example, it may be the result of the experiment "collect one automobile bodily injury medical payment at random and record its value." We assume that the cdf is known. For example, it may be the Pareto cdf,

$$F_X(x) = 1 - \left(\frac{1,000}{1,000 + x} \right)^3 .$$

Now consider a second random variable, X^*, resulting from some other process but with the same Pareto distribution. A random sample from X^*, say $x_1^*, \ldots, x_n^*$, is impossible to distinguish from one taken from X. That is, given the n numbers, we could not tell if they arose from automobile claims or something else. Thus, instead of learning about X by observing automobile claims, we could learn about it by observing X^*. Obtaining a random sample from a Pareto distribution is still probably difficult, so we have not yet accomplished much.
   We can make some progress by making a concession. Let us accept as a replacement for a random sample from X^* a sequence of numbers $x_1^{**}, \ldots, x_n^{**}$, which is not a random sample at all, but simply a sequence of numbers that may not be independent, or even random, but was generated by some known process that is related to the random variable X^*. Such a sequence is called a **pseudorandom** sequence because anyone who did not know how the sequence was created could not distinguish it from a random sample from X^* (and, therefore, from X). Such a sequence is satisfactory for our purposes.

---

[1]This statement is not entirely true. A great deal of creativity may be employed in designing an efficient simulation. The brute force approach used here will work; it just may take your computer longer to produce the answer.

Table 19.1   The chi-square test of simulated Pareto observations.

Interval	Observed	Expected	Chi-square
0–100	2,519	2,486.85	0.42
100–250	2,348	2,393.15	0.85
250–500	2,196	2,157.04	0.70
500–750	1,071	1,097.07	0.62
750–1,000	635	615.89	0.59
1,000–1,500	589	610.00	0.72
1,500–2,500	409	406.76	0.01
2,500–5,000	192	186.94	0.14
5,000–10,000	36	38.78	0.20
10,000–	5	7.51	0.84
Total	10,000	10,000	5.10

The field of developing processes for generating pseudorandom sequences of numbers has been well developed. One fact that makes it easier to provide such sequences is that it is sufficient to be able to generate them for the uniform distribution on the interval $(0, 1)$. That is because, if U has the uniform$(0, 1)$ distribution, then $X = F_X^{-1}(U)$ when the inverse exists will have $F_X(x)$ as its cdf. Therefore, we simply obtain uniform pseudorandom numbers $u_1^{**}, \ldots, u_n^{**}$ and then let $x_j^{**} = F_X^{-1}(u_j^{**})$. This is called the **inversion method** of generating random variates. Specific methods for particular distributions have been developed and some will be discussed here. There is a considerable literature on the best ways to generate pseudorandom uniform numbers and a variety of tests have been proposed to evaluate them. Make sure the method you use is a good one.

■   **EXAMPLE 19.1**

Generate 10,000 pseudo-Pareto (with $\alpha = 3$ and $\theta = 1,000$) variates and verify that they are indistinguishable from real Pareto observations.

The pseudouniform values were obtained using the built-in generator supplied with a commercial programming language. The pseudo-Pareto values are calculated from

$$u^{**} = 1 - \left(\frac{1,000}{1,000 + x^{**}} \right)^3.$$

That is,

$$x^{**} = 1,000[(1 - u^{**})^{-1/3} - 1].$$

So, for example, if the first value generated is $u_1^{**} = 0.54246$, we have $x_1^{**} = 297.75$. This calculation was repeated 10,000 times. The results are displayed in Table 19.1, where a chi-square goodness-of-fit test is conducted. The expected counts are calculated using the Pareto distribution with $\alpha = 3$ and $\theta = 1,000$. Because the parameters are known, there are nine degrees of freedom. At a significance level of 5%, the critical value is 16.92, and we conclude that the pseudorandom sample mimics a random sample from this Pareto distribution.   □
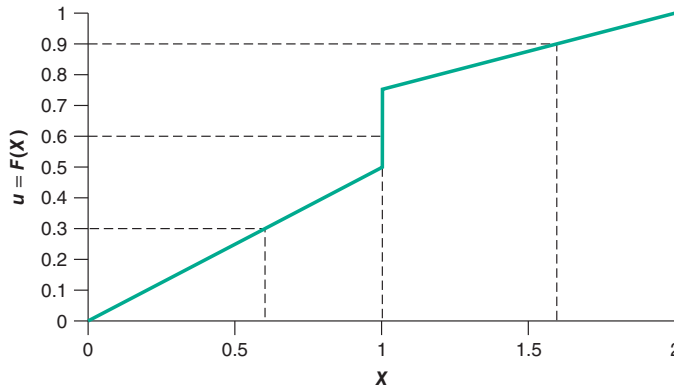
Figure 19.1   The inversion of the distribution function for Example 19.2.

When the distribution function of X is continuous and strictly increasing, the equation $u = F_X(x)$ will have a unique value of x for any given value of u and a unique value of u for any given x. In that case, the inversion method reduces to solving the equation for x. In other cases, some care must be taken. Suppose that $F_X(x)$ has a jump at $x = c$ so that $F_X(c-) = a$ and $F_X(c) = b > a$. If the uniform number is such that $a \leq u < b$, the equation has no solution. In that situation, choose c as the simulated value.

■ **EXAMPLE 19.2**

Suppose that

$$F_X(x) = \begin{cases} 0.5x, & 0 \leq x < 1, \\ 0.5 + 0.25x & 1 \leq x \leq 2. \end{cases}$$

Determine the simulated values of x resulting from the uniform numbers 0.3, 0.6, and 0.9.

In the first interval, the distribution function ranges from 0 to 0.5 and in the second interval from 0.75 to 1. With $u = 0.3$, we are in the first interval, and the equation to solve is $0.3 = 0.5x$, producing $x = 0.6$. With the distribution function jumping from 0.5 to 0.75 at $x = 1$, any u in that interval will lead to a simulated value of 1, so for $u = 0.6$, the simulated value is $x = 1$. Note that $\Pr(0.5 \leq U < 0.75) = 0.25$, so the value of $x = 1$ will be simulated 25% of the time, matching its true probability. Finally, with 0.9 in the second interval, solve $0.9 = 0.5 + 0.25x$ for $x = 1.6$. Figure 19.1 illustrates this process, showing how drawing vertical bars on the function makes the inversion obvious.   □

It is also possible for the distribution function to be constant over some interval. In that case, the equation $u = F_X(x)$ will have multiple solutions for x if u corresponds to the constant value of $F_X(x)$ over that interval. Our convention (to be justified shortly) is to choose the largest possible value in the interval.

■ **EXAMPLE 19.3**

Suppose that

$$F_X(x) = \begin{cases} 0.5x, & 0 \le x < 1, \\ 0.5, & 1 \le x < 2, \\ 0.5x - 0.5, & 2 \le x < 3. \end{cases}$$

Determine the simulated values of x resulting from the uniform numbers 0.3, 0.5, and 0.9.

The first interval covers values of the distribution function from 0 to 0.5 and the final interval covers the range 0.5 to 1. For $u = 0.3$, use the first interval and solve $0.3 = 0.5x$ for $x = 0.6$. The function is constant at 0.5 from 1 to 2, and so for $u = 0.5$, choose the largest value, $x = 2$. For $u = 0.9$, use the final interval and solve $0.9 = 0.5x - 0.5$ for $x = 2.8$. □

Discrete distributions have both features. The distribution function has jumps at the possible values of the variable and is constant in between.

■ **EXAMPLE 19.4**

Simulate values from a binomial distribution with $m = 4$ and $q = 0.5$ using the uniform numbers 0.3, 0.6875, and 0.95.

The distribution function is

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 0.0625, & 0 \le x < 1, \\ 0.3125, & 1 \le x < 2, \\ 0.6875, & 2 \le x < 3, \\ 0.9375, & 3 \le x < 4, \\ 1, & x \ge 4. \end{cases}$$

For $u = 0.3$, the function has a jump at $x = 1$. For $u = 0.6875$, the function is constant from 2 to 3 (as the limiting value of the interval), and so $x = 3$. For $u = 0.95$, the function has a jump at $x = 4$. It is usually easier to present the simulation algorithm using a table based on the distribution function. Then, a simple lookup function (such as the VLOOKUP function in Excel®) can be used to obtain simulated values. For this example, Table 19.2 has the values. □

Table 19.2   The simulation lookup for Example 19.4.

For u in this range,	the simulated value is
$0 \le u < 0.0625$,	0
$0.0625 \le u < 0.3125$,	1
$0.3125 \le u < 0.6875$,	2
$0.6875 \le u < 0.9375$,	3
$0.9375 \le u < 1$,	4

Many random number generators can produce a value of 0 but not a value of 1 (though some produce neither one). This is the motivation for choosing the largest value in an interval where the cdf is constant.

The second question is: What value of n should be used? This will be answered after some special simulation cases are discussed.

19.1.2 Exercises

19.1 Use the inversion method to simulate three values from the Poisson(3) distribution. Use 0.1247, 0.9321, and 0.6873 for the uniform random numbers.

19.2 Use the uniform random numbers 0.2, 0.5, and 0.7 to simulate values from

$$f_X(x) = \begin{cases} 0.25, & 0 \le x \le 2, \\ 0.1, & 4 \le x \le 9, \\ 0, & \text{otherwise.} \end{cases}$$

19.2 Simulation for Specific Distributions

In this section, we will look at a few special cases where either the inversion method may not be the best (or easiest) choice or the situation warrants some additional thoughts.

19.2.1 Discrete Mixtures

Recall from Section 4.2.3 that the distribution function for a discrete mixture can be written as

$$F_Y(y) = a_1 F_{X_1}(y) + a_2 F_{X_2}(y) + \cdots + a_k F_{X_k}(y).$$

It may be difficult to invert this function, but it may be easy to invert the individual *cdf* s. This suggests a two-step process for simulating from a mixture distribution.

1. Simulate a value from the discrete random variable J, where $\Pr(J = j) = a_j$.

2. Use an appropriate method (usually inversion) to simulate an observation from a random variable with distribution function $F_{X_J}(y)$.

■ **EXAMPLE 19.5**

Consider the following mixture-of-exponentials distribution: $F(x) = 0.3(1 - e^{-0.02x}) + 0.5(1 - e^{-0.04x}) + 0.2(1 - e^{-0.05x})$. Use the outlined approach to simulate an observation from this distribution, using $u_1 = 0.45$ and $u_2 = 0.76$ as pseudouniform random numbers.

For simulating values of J, $0 \le u < 0.3$ simulates a 1, $0.3 \le u < 0.8$ simulates a 2, and $0.8 \le u < 1$ simulates a 3. In this case, the simulated value of J is 2. The second step requires simulation of an exponential variable with mean 25. The equation to solve is $0.76 = 1 - e^{-0.04x}$ resulting in $x = -\ln(1 - 0.76)/0.04 = 35.68$. □

19.2.2 Time or Age of Death from a Life Table

When following the progress of a portfolio of life insurance or annuity policies, it is necessary to simulate the time or age of death or other decrements. Two approaches will be discussed here.

First, suppose that the portfolio of policies is to be followed from period to period. This may be necessary because other random factors, such as investment earning, may need to be updated as well. If we are looking at a single policy, then there is a probability for each decrement that applies. For example, a policy that is a certain age and duration may have, in the next period, a death probability of 0.01, a lapse probability of 0.09, a disability probability of 0.03, and a probability of continuing as a healthy policyholder of 0.87. Simulating the outcome is simply obtaining a value from a discrete random variable with four possible outcomes. Now suppose that the portfolio contains 250 policyholders with these probabilities and the individual outcomes are independent. This is a multinomial distribution, but it can be broken down into three steps using a result from this distribution.

■ **EXAMPLE 19.6**

Simulate the outcome from these 250 policies using the following pseudouniform random numbers: 0.34, 0.73, and 0.49.

First, simulate the number of deaths. The marginal distribution is binomial with $m = 250$ and $q = 0.01$. From the exact binomial probabilities and a lookup table, the simulated number of deaths is 2. A normal approximation could also be used with the result rounded. In this case, with a mean of 2.5 and a variance of 2.475, use of the inversion method produces 1.85, which rounds to 2. A property of the multinomial distribution is that the conditional distribution of one value given some of the other values is binomial. In this case, the distribution of the number of lapses will have $m = 250 - 2 = 248$ and $q = 0.09/(1 - 0.01)$. The simulated value from a lookup table is 25. Next, the simulated number of disability cases is binomial with $m = 248 - 25 = 223$ and $q = 0.03/(1 - 0.01 - 0.09)$. The simulated value is 7. For these last two cases, the normal approximation again produces the same answer. The number who continue as healthy policyholders is $250 - 2 - 25 - 7 = 216$. □

The second case is where all that is needed is the age or time at which an event occurs. In a single-decrement setting, this is a discrete distribution and a single lookup will provide the answer. The multiple-decrement setting is a bit more complicated. However, simulation can be accomplished with a single lookup provided that some care is taken.

■ **EXAMPLE 19.7**

Suppose that the policy is a two-year term insurance with annual premiums, where the second premium is waived for those who are disabled at the end of the first year. For year 1, the probabilities from Example 19.6 hold. For those who are healthy entering the second year, the probability of death is 0.02 (all other outcomes are not relevant) and for those who are disabled the probability of death is 0.05 (again, all other outcomes are not relevant). Use the single pseudouniform random number 0.12 to simulate the results of one policy.

Table 19.3 The life insurance simulation for Example 19.7.

j	Description	Probability	Cumulative probability
1	Dies in year 1	0.0100	0.0100
2	Lapses in year 1	0.0900	0.1000
3	Disabled in year 1, dies in year 2	0.0015	0.1015
4	Disabled in year 1, survives year 2	0.0285	0.1300
4	Healthy and insured at time 1, dies in year 2	0.0174	0.1474
5	Healthy and insured at time 1, survives year 2	0.8526	1.0000

There are several outcomes during the two-year period. They are identified by providing a number and a name. The information is provided in Table 19.3. This is a discrete distribution. Use of the inversion method with $u = 0.12$ provides a simulation of outcome 4, which becomes disabled in year 1 (so only pays the first premium) but survives to the end of the two years. The ordering of the outcomes is not relevant to the process, though changing the ordering may produce a different outcome for this simulation. The same technique as in Example 19.6 can be used for a portfolio of similar policies. □

19.2.3 Simulating from the $(a, b, 0)$ Class

Consider Example 19.6, where it was necessary to simulate a value from a binomial distribution with $m = 250$ and $q = 0.01$. The creation of a table requires 251 rows, one for each possible outcome. While some computer applications make the table lookup process easy, writing code for this situation will require a looping process where each value in the table is checked until the desired answer is reached. In some cases, there may be a more efficient approach based on a stochastic process. Such processes generate a series of events and the time of each event. Counting of the number of outcomes in a fixed time period, such as one year, produces the simulated result. The simulation of the event times may be of use in some situations. However, it should be noted that it is the timing that ensures the modeled distribution, not the reverse. That is, for example, it is possible to have a binomially distributed number of events in a given time period with a different process (from the one used here) generating the random times. While the random variable is being interpreted as the number of events in a fixed time period, no such time period is required for the actual situation being simulated. The theory that supports this method is available in Section 6.6.3 of the third edition of this book [73].

The process that creates the timing of the events starts by having an exponential distribution for the time of the first event. The time for the second event will also have an exponential distribution, with the mean depending on the number of previous events. To simplify matters, the time period in question will always be of length 1. If the period is other than 1, the times that are simulated can be viewed as the proportion of the relevant period. The process is, in general, noting that the first event carries an index of 0, the second event an index of 1, and so on:

1. Simulate the time of the first event as an exponential variable with mean $1/\lambda_0$. Determine this time as $t_0 = -\ln(1 - u_0)/\lambda_0$, where u_0 is a pseudouniform random number.

2. Let t_{k-1} be the time of the most recently simulated event. Simulate the time to the next event using an exponential variable with mean $1/\lambda_k$. Determine this time as $s_k = -\ln(1 - u_k)/\lambda_k$.

3. The time of the next event is then $t_k = t_{k-1} + s_k$.

4. Repeat steps 2 and 3 until $t_k > 1$.

5. The simulated value is k.

All that remains is to determine the formulas for the exponential means.

19.2.3.1 Poisson   This is the simplest case. To simulate values from a Poisson distribution with mean λ, set $\lambda_k = \lambda$ for all k.

■ **EXAMPLE 19.8**

Simulate an observation from a Poisson distribution with mean 2.5, using as many of the following pseudouniform random numbers as necessary: 0.12, 0.79, 0.48, 0.62, and 0.29.

The time of the first event is $t_0 = -\ln(0.88)/2.5 = 0.0511$. The time to the second event is $s_1 = -\ln(0.21)/2.5 = 0.6243$ and so the time of the second event is $t_1 = 0.0511 + 0.6243 = 0.6754$. The time to the third event is $s_2 = -\ln(0.52)/2.5 = 0.2616$ and $t_2 = 0.6754 + 0.2616 = 0.9370$. The time to the fourth event is $s_3 = -\ln(0.38)/2.5 = 0.3870$ and $t_3 = 0.9370 + 0.3870 = 1.3240$. Thus the simulated value is 3. □

19.2.3.2 Binomial   As usual, let the binomial parameters be m and q. From them, calculate $d = \ln(1 - q)$ and $c = -md$. Then, $\lambda_k = c + dk$.

■ **EXAMPLE 19.9**

Repeat Example 19.8 for a binomial distribution with parameters $m = 250$ and $q = 0.01$.

For these parameters, $d = \ln(0.99) = -0.01005$ and $c = -250(-0.01005) = 2.51258$. Then, $\lambda_0 = 2.51258$ and $t_0 = -\ln(0.88)/2.51258 = 0.0509$. For the next value $\lambda_1 = 2.51258 - 0.01005(1) = 2.50253$ and $t_1 = 0.0509 - \ln(0.21)/2.50253 = 0.6745$. Continuing, $\lambda_2 = 2.51258 - 0.01005(2) = 2.49248$ and $t_2 = 0.6745 - \ln(0.52)/2.49248 = 0.9369$. Finally, $\lambda_3 = 2.48243$ and $t_3 = 0.9369 - \ln(0.38)/2.48243 = 1.3267$. Once again the simulated value is 3. This is not surprising, as for small values of mq the binomial and Poisson distributions are similar. □

It should be noted that because the binomial distribution cannot produce a value greater than m, if $t_{m-1} < 1$, then the simulation stops and the simulated value is set equal to m. Note that $\lambda_m = 0$ for all binomial distributions, so if the algorithm were to continue, the next simulated time would be at infinity, regardless of the value of t_{m-1}.

19.2.3.3 Negative Binomial The process is the same as for the binomial distribution, but with different formulas for c and d. With parameters r and β, the formulas are $d = \ln(1+\beta)$ and $c = rd$.

■ **EXAMPLE 19.10**

Repeat Example 19.8 for a negative binomial distribution with parameters $r = 250$ and $\beta = 0.01$.

For these parameters, $d = \ln(1.01) = 0.00995$ and $c = 250(0.00995) = 2.48758$. Then, $\lambda_0 = 2.48758$ and $t_0 = -\ln(0.88)/2.48758 = 0.0514$. For the next value, $\lambda_1 = 2.48758 + 0.00995(1) = 2.49753$ and $t_1 = 0.0514 - \ln(0.21)/2.49753 = 0.6763$. Continuing, $\lambda_2 = 2.48758 + 0.00995(2) = 2.50748$ and $t_2 = 0.6763 - \ln(0.52)/2.50748 = 0.9371$. Finally, $\lambda_3 = 2.51743$ and $t_3 = 0.9371 - \ln(0.38)/2.51743 = 1.3215$. □

This procedure provides additional insight for the negative binomial distribution. In Section 6.3, the distribution was derived as a gamma mixture of Poisson distributions. A motivation is that the distribution is the result for a portfolio of drivers, each of which has a Poisson-distributed number of claims, but their Poisson means vary by a gamma distribution. However, there is also some evidence that individual drivers have a negative binomial distribution. In the above simulation, the value of d is always positive. Thus, with each claim the parameter λ increases in value. This reduces the expected time to the next claim. Thus, if we believe that drivers who have claims are more likely to have further claims, then the negative binomial distribution may be a reasonable model.[2]

19.2.4 Normal and Lognormal Distributions

It is always sufficient to be able to simulate Z, a standard normal random variable. Then, if $X \sim N(\mu, \sigma^2)$, let $X = \mu + \sigma Z$. If X is lognormal with parameters μ and σ, let $X = \exp(\mu + \sigma Z)$. The inversion method is usually available (for example, the NORM.INV function in Excel®). However, this method is not as good in the tails as it is in the central part of the distribution, being likely to underrepresent more extreme values. A simple alternative is the Box–Muller transformation [16]. The method begins with the generation of two independent pseudouniform random numbers u_1 and u_2. Then, two independent standard normal values are obtained from $z_1 = \sqrt{-2\ln(u_1)}\cos(2\pi u_2)$ and $z_1 = \sqrt{-2\ln(u_1)}\sin(2\pi u_2)$. An improvement is the polar method, which also begins with two pseudouniform values. The steps are as follows:

1. Calculate $x_1 = 2u_1 - 1$ and $x_2 = 2u_2 - 1$.
2. Calculate $w = x_1^2 + x_2^2$.
3. If $w \geq 1$, repeat steps 1 and 2; else proceed to step 4.

---

[2]There is a subtle distinction needed here. The fact that a driver who has had claims provides evidence of being a bad driver is not what motivates this model. Such a driver could have a constant value of λ and the claims are revealing that it is higher than we thought. The motivation here is that having had a claim, the driver's behavior changes in a way that produces more claims. This phenomenon is called **positive contagion** and has been used to model the increasing frequency of windstorms over time, possibly as a result of global warming.

4. Calculate $y = \sqrt{-2\ln(w)/w}$.

5. Calculate $z_1 = x_1 y$ and $z_2 = x_2 y$.

The polar method requires more programming work due to the rejection possibility at step 3, but is superior to other methods.

■ **EXAMPLE 19.11**

Simulate values of two standard normal random variables using all three methods. Use $u_1 = 0.35$ and $u_2 = 0.57$.

For the inversion method, looking up the values in a standard normal distribution table or using software produces $z_1 = -0.385$ and $z_2 = 0.176$. The Box–Muller transform produces $z_1 = \sqrt{-2\ln(0.35)}\cos[2\pi(0.57)] = -1.311$ and $z_2 = \sqrt{-2\ln(0.35)}\sin[2\pi(0.57)] = -0.617$. For the polar method, $x_1 = -0.3$ and $x_2 = 0.14$. Then, $w = (-0.3)^2 + (0.14)^2 = 0.1096$. Because the value is less than 1, we can proceed. Then, $y = \sqrt{-2\ln(0.1096)/0.1096} = 6.3518$ and $z_1 = -0.3(6.3518) = -1.906$ and $z_2 = 0.14(6.3518) = 0.889$. □

19.2.5 Exercises

19.3 Simulate two observations from a distribution that is a mixture of a Pareto distribution with $\alpha = 3$ and $\theta = 100$ and an inverse Weibull distribution with $\tau = 2$ and $\theta = 200$. The weight on the Pareto distribution is 0.4. Use the pairs of uniform random numbers $(0.372, 0.693)$ and $(0.701, 0.284)$, where the first number is used to determine the distribution to use and the second number is used to simulate an outcome from that distribution.

19.4 At any time, a member of a pension plan is in one of three states: employed (e), alive but no longer employed (n), or dead (d). Let q^{ab} denote the probability that a current or former member of the plan is in state b at the end of a year given that the member was in state a at the beginning of that year. The probabilities are constant over time and independent of age. They are: $q^{ee} = 0.90$, $q^{en} = 0.08$, $q^{ed} = 0.02$, $q^{nn} = 0.95$, $q^{nd} = 0.05$, and $q^{dd} = 1$. Any probabilities not listed are zero. At the beginning of a particular year there are 200 members, all of whom are employed. Using the approach from Example 19.6, simulate the number of members in each of the three states two years from now. Use the uniform random numbers 0.123, 0.876, 0.295, 0.623, and 0.426.

19.5 Use the method of this section to simulate one observation from a binomial distribution with $m = 3$ and $q = 0.07$. Use the uniform random numbers 0.143, 0.296, 0.003, and 0.192.

19.6 Simulate two values from a lognormal distribution with $\mu = 5$ and $\sigma = 1.5$. Use the polar method and the uniform random numbers 0.942, 0.108, 0.217, and 0.841.

19.3 Determining the Sample Size

A question asked at the beginning of this chapter remains unanswered: How many simulations are needed to achieve a desired level of accuracy? We know that any consistent

estimator will be arbitrarily close to the true value with high probability as the sample size is increased. In particular, empirical estimators have this attribute. With a little effort, we should be able to determine the number of simulated values needed to get us as close as we want with a specified probability. Often, the central limit theorem will help, as in the following example.

■ **EXAMPLE 19.12**

(*Example 19.1 continued*) Use simulation to estimate the mean, $F_X(1,000)$, and $\pi_{0.9}$, the 90th percentile of the Pareto distribution with $\alpha = 3$ and $\theta = 1,000$. In each case, stop the simulations when you are 95% confident that the answer is within $\pm 1\%$ of the true value.

In this example, we know the values. Here, $\mu = 500$, $F_X(1,000) = 0.875$, and $\pi_{0.9} = 1,154.43$. For instructional purposes, we behave as if we do not know these values.

The empirical estimate of μ is $\bar{x}$. The central limit theorem tells us that for a sample of size n

$$0.95 = \Pr(0.99\mu \le \bar{X}_n \le 1.01\mu)$$

$$= \Pr\left(-\frac{0.01\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le \frac{0.01\mu}{\sigma/\sqrt{n}}\right)$$

$$\doteq \Pr\left(-\frac{0.01\mu}{\sigma/\sqrt{n}} \le Z \le \frac{0.01\mu}{\sigma/\sqrt{n}}\right),$$

where Z has the standard normal distribution. Our goal is achieved when

$$\frac{0.01\mu}{\sigma/\sqrt{n}} = 1.96, \tag{19.1}$$

which means $n = 38{,}416(\sigma/\mu)^2$. Because we do not know the values of σ and μ, we estimate them using the sample standard deviation and mean. The estimates improve with n, so our stopping rule is to cease simulating when

$$n \ge \frac{38{,}416s^2}{\bar{x}^2}.$$

For a particular simulation that we conducted, the criterion was met when $n = 106{,}934$, at which point $\bar{x} = 501.15$, a relative error of 0.23%, well within our goal.

We now turn to the estimation of $F_X(1,000)$. The empirical estimator is the sample proportion at or below 1,000, say, P_n/n, where P_n is the number at or below 1,000 after n simulations. The central limit theorem tells us that P_n/n is approximately normal with mean $F_X(1,000)$ and variance $F_X(1,000)[1 - F_X(1,000)]/n$. Using P_n/n as an estimate of $F_X(1,000)$ and arguing as previously yields

$$n \ge 38{,}416\frac{(P_n/n)(1 - P_n/n)}{(P_n/n)^2}$$

$$= 38{,}416\frac{n - P_n}{P_n}.$$

For our simulation, the criterion was met at $n = 5,548$, at which point the estimate was $4,848/5,548 = 0.87383$, which has a relative error of 0.13%.

Finally, for $\pi_{0.9}$, begin with

$$0.95 = \Pr(Y_a \leq \pi_{0.9} \leq Y_b),$$

where $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ are the order statistics from the simulated sample; a is the greatest integer less than or equal to $0.9n + 0.5 - 1.96\sqrt{0.9(0.1)n}$; b is the smallest integer greater than or equal to $0.9n + 0.5 + 1.96\sqrt{0.9(0.1)n}$; and the process terminates when both

$$\hat{\pi}_{0.9} - Y_a \leq 0.01\hat{\pi}_{0.9}$$

and

$$Y_b - \hat{\pi}_{0.9} \leq 0.01\hat{\pi}_{0.9}.$$

For the example, this occurred when $n = 126,364$, and the estimated 90th percentile is 1,153.97, with a relative error of 0.04%. □

The method for working with percentiles is not as satisfying as the other two examples. When the goal is to estimate the mean or a probability, we were able to work directly with a normal approximation and an estimate of the standard deviation of the estimator. A similar approach can be taken with the estimated percentiles. However, the formula for the asymptotic variance is

$$\mathrm{Var}(\hat{\pi}_p) \doteq \frac{p(1-p)}{n[f(\pi_p)]^2}.$$

The problem is that while p is known and π_p can be replaced by its estimate, the density function of the simulated variable is not known (recall that we are performing simulations because basic quantities such as the pdf and cdf are not available). Thus, it is likely to be difficult to obtain an estimated value of the variance that can, in turn, be used to estimate the required sample size.

19.3.1 Exercises

19.7 Demonstrate that $0.95 = \Pr(Y_a \leq \pi_{0.9} \leq Y_b)$ for Y_a and Y_b as defined in Example 19.12.

19.8 You are simulating observations from an exponential distribution with $\theta = 100$. How many simulations are needed to be 90% certain of being within 2% of each of the mean and the probability of being below 200? Conduct the required number of simulations and note if the 2% goal has been reached.

19.9 Simulate 1,000 observations from a gamma distribution with $\alpha = 2$ and $\theta = 500$. Perform the chi-square goodness-of-fit and Kolmogorov–Smirnov tests to see if the simulated values were actually from that distribution.

19.10 (*) To estimate $E(X)$, you have simulated five observations from the random variable X. The values are 1, 2, 3, 4, and 5. Your goal is to have the standard deviation of the estimate of $E(X)$ be less than 0.05. Estimate the total number of simulations needed.

19.4   Examples of Simulation in Actuarial Modeling

19.4.1   Aggregate Loss Calculations

The recursive method for calculating aggregate loss distribution, presented in Chapter 9, has three features. First, the recursive method is exact up to the level of the approximation introduced. The only approximation involves replacing the true severity distribution with an arithmetized approximation. The approximation error can be reduced by increasing the number of points (that is, reducing the span). Second, it assumes that aggregate claims can be written as $S = X_1 + \cdots + X_N$ with $N, X_1, X_2, \ldots$ independent and the X_js identically distributed. Third, the recursive method assumes that the frequency distribution is in the $(a, b, 0)$ or $(a, b, 1)$ classes.

There is no need to be concerned about the first feature because the approximation error can be made as small as desired, though at the expense of increased computing time. However, the second restriction may prevent the model from reflecting reality. The third restriction means that if the frequency distribution is not in one of these classes (or can be constructed from them, such as with compound distributions), we will need to find an alternative to the recursive method. Simulation is one such alternative.

In this section, we indicate some common ways in which the independence or identical distribution assumptions may fail to hold and then demonstrate how simulation can be used to obtain numerical values of the distribution of aggregate losses.

When the X_js are i.i.d., it does not matter how we go about labeling the losses, that is, which loss is called X_1, which one X_2, and so on. With the assumption removed, the labels become important. Because S is the aggregate loss for one year, time is a factor. One way of identifying the losses is to let X_1 be the first loss, X_2 be the second loss, and so on. Then let T_j be the random variable that records the time of the jth loss. Without going into much detail about the claims-paying process, we do want to note that T_j may be the time at which the loss occurred, the time it was reported, or the time payment was made. In the latter two cases, it may be that $T_j > 1$, which occurs when the report of the loss or the payment of the claim takes place at a time subsequent to the end of the time period of the coverage, usually one year. If the timing of the losses is important, we will need to know the joint distribution of $(T_1, T_2, \ldots, X_1, X_2, \ldots)$.

19.4.2   Examples of Lack of Independence

There are two common situations in which the assumption of independence does not hold. One is through accounting for time (and, in particular, the time value of money) and the other is through coverage modifications. The latter may have a time factor as well. The following examples provide some illustrations.

■   **EXAMPLE 19.13**

(*Time value of loss payments*) Suppose that the quantity of interest, S, is the present value of all payments made in respect of a policy issued today and covering loss events that occur in the next year. Develop a model for S.

Let T_j be the time of the payment of the jth loss. While T_j records the time of the payment, the subscripts are selected in order of the loss events. Let $T_j = C_j + L_j$, where C_j is the time of the event and L_j is the time from occurrence to payment.

Assume that they are independent and the L_js are independent of each other. Let the time between events, $C_j - C_{j-1}$ (where $C_0 = 0$), be i.i.d. with an exponential distribution with mean 0.2 years.

Let X_j be the amount paid at time T_j on the loss that occurred at time C_j. Assume that X_j and C_j are independent (the amount of the claim does not depend on when in the year it occurred) but X_j and L_j are positively correlated (a specific distributional model is specified when the example is continued). This assumption is reasonable because the more expensive losses may take longer to settle.

Finally, let V_t be a random variable that represents the value, which, if invested today, will accumulate to 1 in t years. It is independent of all X_j, C_j, and L_j. But clearly, for $s \neq t$, V_s and V_t are dependent.

We then have

$$S = \sum_{j=1}^{N} X_j V_{T_j},$$

where $N = \max_{C_j < 1}\{j\}$. □

■ EXAMPLE 19.14

(*Out-of-pocket maximum*) Suppose that there is a deductible, d, on individual losses. However, in the course of a year, the policyholder will pay no more than u. Develop a model for the insurer's aggregate payments.

Let X_j be the amount of the jth loss. Here, the assignment of j does not matter. Let $W_j = X_j \wedge d$ be the amount paid by the policyholder due to the deductible and let $Y_j = X_j - W_j$ be the amount paid by the insurer. Then, $R = W_1 + \cdots + W_N$ is the total amount paid by the policyholder prior to imposing the out-of-pocket maximum. It follows that the amount actually paid by the policyholder is $R_u = R \wedge u$. Let $S = X_1 + \cdots + X_N$ be the total losses; then, the aggregate amount paid by the insurer is $T = S - R_u$. Note that the distributions of S and R_u are based on i.i.d. severity distributions. The analytic methods described earlier can be used to obtain their distributions. But because they are dependent, their individual distributions cannot be combined to produce the distribution of T. There is also no way to write T as a random sum of i.i.d. variables. At the beginning of the year, it appears that T will be the sum of i.i.d. Y_js, but at some point the Y_js may be replaced by X_js as the out-of-pocket maximum is reached. □

19.4.3 Simulation Analysis of the Two Examples

We now complete the two examples using the simulation approach. The models have been selected arbitrarily, but we should assume they were determined by a careful estimation process using the techniques presented earlier in this text.

■ EXAMPLE 19.15

(*Example 19.13 continued*) The model is completed with the following specifications. The amount of a payment (X_j) has a Pareto distribution with parameters $\alpha = 3$ and $\theta = 1,000$. The time from the occurrence of a claim to its payment (L_j) has a Weibull distribution with $\tau = 1.5$ and $\theta = \ln(X_j)/6$. This models the dependence by having

the scale parameter depend on the size of the loss. The discount factor is modeled by assuming that, for $t > s$, $[\ln(V_s/V_t)]/(t - s)$ has a normal distribution with mean 0.06 and variance $0.0004(t - s)$. We do not need to specify a model for the number of losses. Instead, we use the model given earlier for the time between losses. Use simulation to determine the expected present value of aggregate payments.

The mechanics of a single simulation are presented in detail to indicate how the process is to be done. Begin by generating i.i.d. exponential interloss times until their sum exceeds 1 (in order to obtain one year's worth of claims). The individual variates are generated from pseudouniform numbers using

$$u = 1 - e^{-5x},$$

which yields

$$x = -0.2 \ln(1 - u).$$

For the first simulation, the uniform pseudorandom numbers and the corresponding x values are (0.25373, 0.0585), (0.46750, 0.1260), (0.23709, 0.0541), (0.75780, 0.2836), and (0.96642, 0.6788). At this point, the simulated xs total 1.2010, and therefore there are four loss events, occurring at times $c_1 = 0.0585$, $c_2 = 0.1845$, $c_3 = 0.2386$, and $c_4 = 0.5222$.

The four loss amounts are found from inverting the Pareto cdf. That is,

$$x = 1,000[(1 - u)^{-1/3} - 1].$$

The four pseudouniform numbers are 0.71786, 0.47779, 0.61084, and 0.68579, producing the four losses $x_1 = 524.68$, $x_2 = 241.80$, $x_3 = 369.70$, and $x_4 = 470.93$.

The times from occurrence to payment have a Weibull distribution. The equation to solve is

$$u = 1 - e^{-[6l/\ln(x)]^{1.5}},$$

where x is the loss. Solving for the lag time l yields

$$l = \frac{1}{6}\ln(x)[-\ln(1 - u)]^{2/3}.$$

For the first lag, we have $u = 0.23376$ and so

$$l_1 = \frac{1}{6}\ln(524.68)[-\ln 0.76624]^{2/3} = 0.4320.$$

Similarly, with the next three values of u being 0.85799, 0.12951, and 0.72085, we have $l_2 = 1.4286$, $l_3 = 0.2640$, and $l_4 = 1.2068$. The payment times of the four losses are the sum of c_j and l_j, namely $t_1 = 0.4905$, $t_2 = 1.6131$, $t_3 = 0.5026$, and $t_4 = 1.7290$.

Finally, we generate the discount factors. They must be generated in order of increasing t_j, so we first obtain $v_{0.4905}$. We begin with a normal variate with mean 0.06 and variance $0.0004(0.4905) = 0.0001962$. Using inversion, the simulated value is $0.0592 = [\ln(1/v_{0.4905})]/0.4905$, and so $v_{0.4905} = 0.9714$. Note that for the first value we have $s = 0$ and $v_0 = 1$. For the second value, we require a normal variate with mean 0.06 and variance $(0.5026 - 0.4905)(0.0004) = 0.00000484$. The simulated value is

$$0.0604 = \frac{\ln(0.9714/v_{0.5026})}{0.0121} \text{ for } v_{0.5026} = 0.9707.$$

Table 19.4   Negative binomial cumulative probabilities.

n	$F_N(n)$	n	$F_N(n)$
0	0.03704	8	0.76589
1	0.11111	9	0.81888
2	0.20988	10	0.86127
3	0.31962	11	0.89467
4	0.42936	12	0.92064
5	0.53178	13	0.94062
6	0.62282	14	0.95585
7	0.70086	15	0.96735

For the next two payments, we have

$$0.0768 = \frac{\ln(0.9707/v_{1.6131})}{1.1105} \text{ for } v_{1.6131} = 0.8913,$$

$$0.0628 = \frac{\ln(0.8913/v_{1.7290})}{0.1159} \text{ for } v_{1.7290} = 0.8848.$$

We are now ready to determine the first simulated value of the aggregate present value. It is

$$s_1 = 524.68(0.9714) + 241.80(0.8913) + 369.70(0.9707) + 470.93(0.8848)$$
$$= 1,500.74.$$

The process was then repeated until there was 95% confidence that the estimated mean was within 1% of the true mean. This took 26,944 simulations, producing a sample mean of 2,299.16.   □

■ **EXAMPLE 19.16**

(*Example 19.14 continued*) For this example, set the deductible d at 250 and the out-of-pocket maximum at $u = 1,000$. Assume that the number of losses has a negative binomial distribution with $r = 3$ and $\beta = 2$. Further assume that individual losses have a Weibull distribution with $\tau = 2$ and $\theta = 600$. Determine the 95th percentile of the insurer's losses.

To simulate the negative binomial claim counts, we require the cdf of the negative binomial distribution. There is no closed form that does not involve a summation operator. However, a table can be constructed, and one appears here as Table 19.4. The number of losses for the year is generated by obtaining one pseudouniform value – for example, $u = 0.47515$ – and then determining the smallest entry in the table that is larger than 0.47515. The simulated value appears to its left. In this case, our first simulation produced $n = 5$ losses.

The amounts of the five losses are obtained from the Weibull distribution. Inversion of the cdf produces

$$x = 600[-\ln(1-u)]^{1/2}.$$

The five simulated values are 544.04, 453.67, 217.87, 681.98, and 449.83. The total loss is 2,347.39. The policyholder pays $250.00 + 250.00 + 217.87 + 250.00 + 250.00 = 1,217.87$, but the out-of-pocket maximum limits the payment to 1,000. Thus our first simulated value has the insurer paying 1,347.39.

The goal was set to be 95% confident that the estimated 95th percentile would be within 2% of the true value. Achieving this goal requires 11,476 simulations, producing an estimated 95th percentile of 6,668.18. □

19.4.4   The Use of Simulation to Determine Risk Measures

If the distribution of interest is too complex to admit an analytic form, simulation may be used to estimate risk measures such as VaR and TVaR. Because VaR is simply a specific percentile of the distribution, this case has already been discussed. The estimation of TVaR is also fairly straightforward. Suppose that $y_1 \leq y_2 \leq \cdots \leq y_n$ is an ordered simulated sample from the random variable of interest. If the percentile being used is p, let $k = [pn] + 1$, where $[\cdot]$ indicates the greatest integer function. Then, the two estimators are

$$\widehat{\mathrm{VaR}}_p(X) = y_k \quad \text{and} \quad \widehat{\mathrm{TVaR}}_p(X) = \frac{1}{n-k+1} \sum_{j=k}^{n} y_j.$$

We know that the variance of a sample mean can be estimated by the sample variance divided by the sample size. While $\mathrm{T\hat{V}aR}_p(X)$ is a sample mean, this estimator will underestimate the true value. This is because the observations being averaged are dependent and, as a result, there is more variability than is reflected by the sample variance. Let the sample variance be

$$s_p^2 = \frac{1}{n-k} \sum_{j=k}^{n} [y_j - \mathrm{T\hat{V}aR}_p(X)]^2.$$

Manistre and Hancock [85] show that an asymptotically unbiased estimator of the variance of the estimator of TVaR is

$$\widehat{\mathrm{Var}}\left[\widehat{\mathrm{TVaR}}_p(X)\right] = \frac{s_p^2 + p\left[\widehat{\mathrm{TVaR}}_p(X) - \widehat{\mathrm{VaR}}_p(X)\right]^2}{n-k+1}.$$

■ **EXAMPLE 19.17**

Consider a Pareto distribution with $\alpha = 2$ and $\theta = 100$. Use 10,000 simulations to estimate the risk measures with $p = 0.95$.

The true values are $\mathrm{VaR}_{0.95}(X) = 347.21$ and $\mathrm{TVaR}_{0.95}(X) = 794.42$. For the simulation, $k = [9,500] + 1 = 9,501$. The simulation produced $\widehat{\mathrm{VaR}}_{0.95}(X) = y_{9,501} = 363.09$ and $\widehat{\mathrm{TVaR}}_{0.95}(X) = \frac{1}{500} \sum_{j=9,501}^{10,000} y_j = 816.16$. The variance of the estimator of TVaR is estimated to be 2,935.36. A 95% confidence interval for the true value is 816.16 ± 106.19 and the true value is well within this confidence interval. □

19.4.5   Statistical Analyses

Simulation can help in a variety of ways when analyzing data. Two are discussed here, both of which have to do with evaluating a statistical procedure. The first is the determination

of the p-value (or critical value) for a hypothesis test. The second is to evaluate the MSE of an estimator. We begin with the hypothesis testing situation.

■ **EXAMPLE 19.18**

It is conjectured that losses have a lognormal distribution. One hundred observations have been collected and the Kolmogorov–Smirnov test statistic is 0.06272. Determine the p-value for this test, first with the null hypothesis being that the distribution is lognormal with $\mu = 7$ and $\sigma = 1$ and then with the parameters unspecified.

For the null hypothesis with each parameter specified, one simulation involves first simulating 100 lognormal observations from the specified lognormal distribution. Then the Kolmogorov–Smirnov test statistic is calculated. The estimated p-value is the proportion of simulations for which the test statistic exceeds 0.06272. After 1,000 simulations, the estimate of the p-value is 0.836.

With the parameters unspecified, it is not clear which lognormal distribution should be used. It turns out that, for the observations actually collected, $\hat{\mu} = 7.2201$ and $\hat{\sigma} = 0.80893$. These values were used as the basis for each simulation. The only change is that after the simulated observations have been obtained, the results are compared to a lognormal distribution with parameters estimated (by maximum likelihood) from the simulated data set. For 1,000 simulations, the test of 0.06272 was exceeded 491 times, for an estimated p-value of 0.491.

As indicated in Section 15.4.1, not specifying the parameters makes a considerable difference in the interpretation of the test statistic.   □

When testing hypotheses, p-values and significance levels are calculated assuming the null hypothesis to be true. In other situations, there is no known population distribution from which to simulate. For such situations, a technique called the **bootstrap** may help (for thorough coverage of this subject, see Efron and Tibshirani [34]). The key is to use the empirical distribution from the data as the population from which to simulate values. Theoretical arguments show that the bootstrap estimate will converge asymptotically to the true value. This result is reasonable because as the sample size increases, the empirical distribution becomes more and more like the true distribution. The following example shows how the bootstrap works and also indicates that, at least in the case illustrated, it gives a reasonable answer.

■ **EXAMPLE 19.19**

A sample (with replacement) of size 3 from a population produced the values 2, 3, and 7. Determine the bootstrap estimate of the MSE of the sample mean as an estimator of the population mean.

The bootstrap approach assumes that the population places probability 1/3 on each of the three values 2, 3, and 7. The mean of this distribution is 4. From this population, there are 27 samples of size 3 that might be drawn. Sample means can be 2 (sample values 2, 2, 2, with probability 1/27), 7/3 (sample values 2, 2, 3, 2, 3, 2, and 3, 2, 2, with probability 3/27), and so on, up to 7 with probability 1/27. The MSE is

$$(2-4)^2 \left(\frac{1}{27}\right) + \left(\frac{7}{3} - 4\right)^2 \left(\frac{3}{27}\right) + \cdots + (7-4)^2 \left(\frac{1}{27}\right) = \frac{14}{9}.$$

The usual approach is to note that the sample mean is unbiased and therefore

$$\text{MSE}(\bar{X}) = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

With the variance unknown, a reasonable choice is to use the sample variance. With a denominator of n, for this example, the estimated MSE is

$$\frac{\frac{1}{3}[(2-4)^2 + (3-4)^2 + (7-4)^2]}{3} = \frac{14}{9},$$

the same as the bootstrap estimate. □

In many situations, determination of the MSE is not so easy, and then the bootstrap becomes an extremely useful tool. While simulation was not needed for the example, note that an original sample size of 3 led to 27 possible bootstrap values. Once the sample size gets beyond 6, it becomes impractical to enumerate all the cases. In that case, simulating observations from the empirical distribution becomes the only feasible choice.

■ **EXAMPLE 19.20**

In Example 14.15, an empirical model for time to death was obtained. The empirical probabilities are 0.0333, 0.0744, 0.0343, 0.0660, 0.0344, and 0.0361 that death is at times 0.8, 2.9, 3.1, 4.0, 4.1, and 4.8, respectively. The remaining 0.7215 probability is that the person will be alive five years from now. The expected present value for a five-year term insurance policy that pays 1,000 at the moment of death is estimated as

$$1{,}000(0.0333v^{0.8} + \cdots + 0.0361v^{4.8}) = 223.01,$$

where $v = 1.07^{-1}$. Simulate 10,000 bootstrap samples to estimate the MSE of this estimator.

A method for conducting a bootstrap simulation with the Kaplan–Meier estimate is given by Efron [32]. Rather than simulating from the empirical distribution (as given by the Kaplan–Meier estimate), simulate from the original sample. In this example, that means assigning probability $\frac{1}{40}$ to each of the original observations. Then each bootstrap observation is a left-truncation point along with the accompanying censored or uncensored value. After 40 such observations are recorded, the Kaplan–Meier estimate is constructed from the bootstrap sample and then the quantity of interest computed. This is relatively easy because the bootstrap estimate places probability only at the six original points. Ten thousand simulations were quickly done. The mean was 222.05 and the MSE was 4,119. Efron also notes that the bootstrap estimate of the variance of $\hat{S}(t)$ is asymptotically equal to Greenwood's estimate, thus giving credence to both methods. □

19.4.6 Exercises

19.11 (*) Insurance for a city's snow removal costs covers four winter months. There is a deductible of 10,000 per month. Monthly costs are independent and normally distributed with $\mu = 15{,}000$ and $\sigma = 2{,}000$. Monthly costs are simulated using the inversion method.

For one simulation of a year's payments, the four uniform pseudorandom numbers are 0.5398, 0.1151, 0.0013, and 0.7881. Calculate the insurer's cost for this simulated year.

19.12 (*) After one period, the price of a stock is X times its price at the beginning of the period, where X has a lognormal distribution with $\mu = 0.01$ and $\sigma = 0.02$. The price at time zero is 100. The inversion method is used to simulate price movements. The pseudouniform random numbers are 0.1587 and 0.9332 for periods 1 and 2. Determine the simulated prices at the end of each of the first two periods.

19.13 (*) You have insured 100 people, each age 70. Each person has probability 0.03318 of dying in the next year and the deaths are independent. Therefore, the number of deaths has a binomial distribution with $m = 100$ and $q = 0.03318$. Use the inversion method to determine the simulated number of deaths in the next year based on $u = 0.18$.

19.14 (*) For a surplus process, claims occur according to a Poisson process at the rate of two per year. Thus the time between claims has the exponential distribution with $\theta = 1/2$. Claims have a Pareto distribution with $\alpha = 2$ and $\theta = 1,000$. The initial surplus is 2,000 and premiums are collected at a rate of 2,200. Ruin occurs any time the surplus is negative, at which time no further premiums are collected or claims paid. All simulations are done with the inversion method. For the time between claims, use 0.83, 0.54, 0.48, and 0.14 as the pseudorandom numbers. For claim amounts, use 0.89, 0.36, 0.70, and 0.61. Determine the surplus at time 1.

19.15 (*) You are given a random sample of size 2 from some distribution. The values are 1 and 3. You plan to estimate the population variance with the estimator $[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2]/2$. Determine the bootstrap estimate of the MSE of this estimator.

19.16 A sample of three items from the uniform(0,10) distribution produced the following values: 2, 4, and 7.

(a) Calculate the Kolmogorov–Smirnov test statistic for the null hypothesis that the data came from the uniform(0,10) distribution.

(b) Simulate 10,000 samples of size 3 from the uniform(0,10) distribution and compute the Kolmogorov–Smirnov test statistic for each. The proportion of times the value equals or exceeds your answer to part (a) is an estimate of the p-value.

19.17 A sample of three items from the uniform(0, θ) distribution produced the following values: 2, 4, and 7. Consider the estimator of θ,

$$\hat{\theta} = \tfrac{4}{3} \max(x_1, x_2, x_3).$$

From Example 10.12 the MSE of this unbiased estimator was shown to be $\theta^2/15$.

(a) Estimate the MSE by replacing θ with its estimate.

(b) Obtain the bootstrap estimate of the variance of the estimator. (It is not possible to use the bootstrap to estimate the MSE because you cannot obtain the true value of θ from the empirical distribution, but you can obtain the expected value of the estimator.)

19.18 Losses on an insurance contract have a Pareto distribution with parameters $\alpha = 3$ and $\theta = 10{,}000$. Expenses to process claims have an exponential distribution with mean 400. The dependence structure is modeled with a Gaussian copula with correlation $\rho_{12} = 0.6$. Losses have a deductible of 500. When the deductible is not met, there are no processing expenses. Also, when there is a payment in excess of 10,000, a reinsurer pays the excess. In addition, the primary insurer and reinsurer split the processing expenses in proportion to their share of the payments to the insured. Use the uniform random pairs $(0.983, 0.453)$ and $(0.234, 0.529)$, where the first number simulates the loss and the second the expense, to simulate the results of two loss events. Calculate the total amounts of these losses paid by the insured, the primary insurer, and the reinsurer.

19.19 Repeat Exercise 19.18 using a t copula with $\nu = 6$. Use the same uniform numbers from that exercise to generate the multivariate normal values. Use 0.319 and 0.812 to simulate the scaling factors required for this simulation.

19.20 (*) A dental benefit has a deductible of 100 applied to annual charges. The insured is then reimbursed for 80% of excess charges to a maximum reimbursement of 1,000. Annual charges have an exponential distribution with mean 1,000. Four years' charges are simulated by the inversion method using the uniform random numbers 0.30, 0.92, 0.70, and 0.08. Determine the average annual reimbursement for this simulation.

19.21 (*) Paid losses have a lognormal distribution with parameters $\mu = 13.294$ and $\sigma = 0.494$. The ratio, y, of unpaid losses to paid losses is $y = 0.801x^{0.851}e^{-0.747x}$, where $x = 2006$ minus the contract purchase year. The inversion method is used, with the uniform random numbers 0.2877, 0.1210, 0.8238, and 0.6179 to simulate paid losses. Estimate the average unpaid losses for purchase year 2005.

19.22 (*) You plan to use simulation to estimate the mean of a nonnegative random variable. The population standard deviation is known to be 20% larger than the population mean. Use the central limit theorem to estimate the smallest number of trials needed so that you will be at least 95% confident that your simulated mean is within 5% of the population mean.

19.23 (*) Simulation is used to estimate the value of the cumulative distribution function at 300 of the exponential distribution with mean 100. Determine the minimum number of simulations so that there is at least a 99% probability that the estimate is within 1% of the correct value.

19.24 (*) For a policy that covers both fire and wind losses, you are given that a sample of fire losses was 3 and 4 and a sample of wind losses for the same period was 0 and 3. Fire and wind losses are independent and do not have identical distributions. Based on the sample, you estimate that adding a deductible of 2 per wind claim will eliminate 20% of total losses. Determine the bootstrap approximation to the MSE of the estimate.

APPENDIX A

AN INVENTORY OF CONTINUOUS DISTRIBUTIONS

A.1 Introduction

Descriptions of the models are given starting in Section A.2. First, a few mathematical preliminaries are presented that indicate how the various quantities can be computed.

The incomplete gamma function[1] is given by

$$\Gamma(\alpha; x) = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t}\, dt, \quad \alpha > 0, \ x > 0,$$

with

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt, \quad \alpha > 0.$$

A useful fact is $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$. Also, define

$$G(\alpha; x) = \int_x^\infty t^{\alpha-1} e^{-t}\, dt, \quad x > 0.$$

---

[1]Some references, such as [2], denote this integral $P(\alpha, x)$ and define $\Gamma(\alpha, x) = \int_x^\infty t^{\alpha-1} e^{-t}\, dt$. Note that this definition does not normalize by dividing by $\Gamma(\alpha)$. When using software to evaluate the incomplete gamma function, be sure to note how it is defined.

At times, we will need this integral for nonpositive values of α. Integration by parts produces the relationship

$$G(\alpha; x) = -\frac{x^\alpha e^{-x}}{\alpha} + \frac{1}{\alpha} G(\alpha + 1; x).$$

This process can be repeated until the first argument of G is $\alpha + k$, a positive number. Then, it can be evaluated from

$$G(\alpha + k; x) = \Gamma(\alpha + k)[1 - \Gamma(\alpha + k; x)].$$

However, if α is a negative integer or zero, the value of $G(0; x)$ is needed. It is

$$G(0; x) = \int_x^\infty t^{-1} e^{-t}\, dt = E_1(x),$$

which is called the **exponential integral**. A series expansion for this integral is

$$E_1(x) = -0.57721566490153 - \ln x - \sum_{n=1}^\infty \frac{(-1)^n x^n}{n(n!)}.$$

When α is a positive integer, the incomplete gamma function can be evaluated exactly as given in the following theorem.

Theorem A.1 *For integer α,*

$$\Gamma(\alpha; x) = 1 - \sum_{j=0}^{\alpha-1} \frac{x^j e^{-x}}{j!}.$$

Proof: For $\alpha = 1$, $\Gamma(1; x) = \int_0^x e^{-t}\, dt = 1 - e^{-x}$, and so the theorem is true for this case. The proof is completed by induction. Assume that it is true for $\alpha = 1, \dots, n$. Then,

$$\begin{aligned}
\Gamma(n + 1; x) &= \frac{1}{n!} \int_0^x t^n e^{-t}\, dt \\
&= \frac{1}{n!} \left(-t^n e^{-t} \big|_0^x + \int_0^x n t^{n-1} e^{-t}\, dt \right) \\
&= \frac{1}{n!} (-x^n e^{-x}) + \Gamma(n; x) \\
&= -\frac{x^n e^{-x}}{n!} + 1 - \sum_{j=0}^{n-1} \frac{x^j e^{-x}}{j!} \\
&= 1 - \sum_{j=0}^{n} \frac{x^j e^{-x}}{j!}.
\end{aligned}$$

$\square$

The incomplete beta function is given by

$$\beta(a, b; x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1} (1 - t)^{b-1}\, dt, \quad a > 0,\ b > 0,\ 0 < x < 1,$$

where

$$\beta(a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}$$

is the beta function, and when $b < 0$ (but $a > 1 + \lfloor -b \rfloor$), repeated integration by parts produces

$$
\begin{aligned}
\Gamma(a)\Gamma(b)\beta(a, b; x) = -\Gamma(a + b) &\left[\frac{x^{a-1}(1 - x)^b}{b} \right. \\
&+ \frac{(a - 1)x^{a-2}(1 - x)^{b+1}}{b(b + 1)} + \cdots \\
&+ \left. \frac{(a - 1)\cdots(a - r)x^{a-r-1}(1 - x)^{b+r}}{b(b + 1)\cdots(b + r)} \right] \\
&+ \frac{(a - 1)\cdots(a - r - 1)}{b(b + 1)\cdots(b + r)}\Gamma(a - r - 1) \\
&\times \Gamma(b + r + 1)\beta(a - r - 1, b + r + 1; x),
\end{aligned}
$$

where r is the smallest integer such that $b + r + 1 > 0$. The first argument must be positive (that is, $a - r - 1 > 0$).

Numerical approximations for both the incomplete gamma and the incomplete beta function are available in many statistical computing packages as well as in many spreadsheets, because they are just the distribution functions of the gamma and beta distributions. The following approximations are taken from [2]. The suggestion regarding using different formulas for small and large x when evaluating the incomplete gamma function is from [103]. That reference also contains computer subroutines for evaluating these expressions. In particular, it provides an effective way of evaluating continued fractions.

For $x \leq \alpha + 1$, use the series expansion

$$\Gamma(\alpha; x) = \frac{x^\alpha e^{-x}}{\Gamma(\alpha)} \sum_{n=0}^{\infty} \frac{x^n}{\alpha(\alpha + 1)\cdots(\alpha + n)}$$

while for $x > \alpha + 1$, use the continued-fraction expansion

$$1 - \Gamma(\alpha; x) = \frac{x^\alpha e^{-x}}{\Gamma(\alpha)} \cfrac{1}{x + \cfrac{1 - \alpha}{1 + \cfrac{1}{x + \cfrac{2 - \alpha}{1 + \cfrac{2}{x + \cdots}}}}}.$$

The incomplete gamma function can also be used to produce cumulative probabilities from the standard normal distribution. Let $\Phi(z) = \Pr(Z \leq z)$, where Z has the standard normal distribution. Then, for $z \geq 0$, $\Phi(z) = 0.5 + \Gamma(0.5; z^2/2)/2$, while for $z < 0$, $\Phi(z) = 1 - \Phi(-z)$.

The incomplete beta function can be evaluated by the series expansion

$$
\begin{aligned}
\beta(a, b; x) = &\frac{\Gamma(a + b)x^a(1 - x)^b}{a\Gamma(a)\Gamma(b)} \\
&\times \left[1 + \sum_{n=0}^{\infty} \frac{(a + b)(a + b + 1)\cdots(a + b + n)}{(a + 1)(a + 2)\cdots(a + n + 1)}x^{n+1} \right].
\end{aligned}
$$

The gamma function itself can be found from

$$\ln \Gamma(\alpha) \doteq (\alpha - \tfrac{1}{2}) \ln \alpha - \alpha + \frac{\ln(2\pi)}{2}$$

$$+ \frac{1}{12\alpha} - \frac{1}{360\alpha^3} + \frac{1}{1{,}260\alpha^5} - \frac{1}{1{,}680\alpha^7} + \frac{1}{1{,}188\alpha^9} - \frac{691}{360{,}360\alpha^{11}}$$

$$+ \frac{1}{156\alpha^{13}} - \frac{3{,}617}{122{,}400\alpha^{15}} + \frac{43{,}867}{244{,}188\alpha^{17}} - \frac{174{,}611}{125{,}400\alpha^{19}}.$$

For values of α above 10, the error is less than 10^{-19}. For values below 10, use the relationship

$$\ln \Gamma(\alpha) = \ln \Gamma(\alpha + 1) - \ln \alpha.$$

The distributions are presented in the following way. First, the name is given along with the parameters. Many of the distributions have other names, which are noted in parentheses. Next, the density function $f(x)$ and distribution function $F(x)$ are given. For some distributions, formulas for starting values are given. Within each family, the distributions are presented in decreasing order with regard to the number of parameters. The Greek letters used are selected to be consistent. Any Greek letter that is not used in the distribution means that that distribution is a special case of one with more parameters but with the missing parameters set equal to 1. *Unless specifically indicated, all parameters must be positive.*

Except for two distributions, inflation can be recognized by simply inflating the scale parameter θ. That is, if X has a particular distribution, then cX has the same distribution type, with all parameters unchanged except that θ is changed to $c\theta$. For the lognormal distribution, μ changes to $\mu + \ln(c)$ with σ unchanged, while for the inverse Gaussian, both μ and θ are multiplied by c.

For several of the distributions, starting values are suggested. They are not necessarily good estimators, but just places from which to start an iterative procedure to maximize the likelihood or other objective function. These are found by either the methods of moments or percentile matching. The quantities used are

$$\text{Moments:} \quad m = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad t = \frac{1}{n} \sum_{i=1}^{n} x_i^2,$$

$$\text{Percentile matching:} \quad p = 25\text{th percentile}, \ q = 75\text{th percentile}.$$

For grouped data or data that have been truncated or censored, these quantities may have to be approximated. Because the purpose is to obtain starting values and not a useful estimate, it is often sufficient to just ignore modifications. For three- and four-parameter distributions, starting values can be obtained by using estimates from a special case, then making the new parameters equal to 1. An all-purpose starting value rule (for when all else fails) is to set the scale parameter (θ) equal to the mean and set all other parameters equal to 2.

Risk measures may be calculated as follows. For $\text{VaR}_p(X)$, the Value at Risk, solve the equation $p = F[\text{Var}_p(X)]$. Where there are convenient explicit solutions, they are provided. For $\text{TVaR}_p(X)$, the Tail Value at Risk, use the formula

$$\text{TVaR}_p(X) = \text{Var}_p(X) + \frac{\text{E}(X) - \text{E}[X \wedge \text{Var}_p(X)]}{1 - p}.$$

Explicit formulas are provided in a few cases.

All the distributions listed here (and many more) are discussed in great detail in Kleiber and Kotz [69]. In many cases, alternatives to maximum likelihood estimators are presented.

A.2 The Transformed Beta Family

A.2.1 The Four-Parameter Distribution

A.2.1.1 Transformed Beta – $\alpha, \theta, \gamma, \tau$ (generalized beta of the second kind, Pearson Type VI)[2]

$$f(x) = \frac{\Gamma(\alpha + \tau)}{\Gamma(\alpha)\Gamma(\tau)} \frac{\gamma(x/\theta)^{\gamma\tau}}{x[1 + (x/\theta)^{\gamma}]^{\alpha+\tau}},$$

$$F(x) = \beta(\tau, \alpha; u), \quad u = \frac{(x/\theta)^{\gamma}}{1 + (x/\theta)^{\gamma}},$$

$$E[X^k] = \frac{\theta^k \Gamma(\tau + k/\gamma)\Gamma(\alpha - k/\gamma)}{\Gamma(\alpha)\Gamma(\tau)}, \quad -\tau\gamma < k < \alpha\gamma,$$

$$E[(X \wedge x)^k] = \frac{\theta^k \Gamma(\tau + k/\gamma)\Gamma(\alpha - k/\gamma)}{\Gamma(\alpha)\Gamma(\tau)} \beta(\tau + k/\gamma, \alpha - k/\gamma; u)$$
$$+ x^k[1 - F(x)], \quad k > -\tau\gamma,$$

$$\text{Mode} = \theta \left(\frac{\tau\gamma - 1}{\alpha\gamma + 1} \right)^{1/\gamma}, \quad \tau\gamma > 1, \text{ else } 0.$$

A.2.2 Three-Parameter Distributions

A.2.2.1 Generalized Pareto – α, θ, τ   (beta of the second kind)

$$f(x) = \frac{\Gamma(\alpha + \tau)}{\Gamma(\alpha)\Gamma(\tau)} \frac{\theta^{\alpha} x^{\tau-1}}{(x + \theta)^{\alpha+\tau}},$$

$$F(x) = \beta(\tau, \alpha; u), \quad u = \frac{x}{x + \theta},$$

$$E[X^k] = \frac{\theta^k \Gamma(\tau + k)\Gamma(\alpha - k)}{\Gamma(\alpha)\Gamma(\tau)}, \quad -\tau < k < \alpha,$$

$$E[X^k] = \frac{\theta^k \tau(\tau + 1) \cdots (\tau + k - 1)}{(\alpha - 1) \cdots (\alpha - k)} \quad \text{if } k \text{ is a positive integer,}$$

$$E[(X \wedge x)^k] = \frac{\theta^k \Gamma(\tau + k)\Gamma(\alpha - k)}{\Gamma(\alpha)\Gamma(\tau)} \beta(\tau + k, \alpha - k; u),$$
$$+ x^k[1 - F(x)], \quad k > -\tau,$$

$$\text{Mode} = \theta \frac{\tau - 1}{\alpha + 1}, \quad \tau > 1, \text{ else } 0.$$

---

[2]There is no inverse transformed beta distribution because the reciprocal has the same distribution, with α and τ interchanged and θ replaced with $1/\theta$.

A.2.2.2   *Burr* – α, θ, γ (Burr Type XII, Singh–Maddala)

$$f(x) = \frac{\alpha\gamma(x/\theta)^\gamma}{x[1 + (x/\theta)^\gamma]^{\alpha+1}},$$

$$F(x) = 1 - u^\alpha, \quad u = \frac{1}{1 + (x/\theta)^\gamma},$$

$$\text{VaR}_p(X) = \theta[(1 - p)^{-1/\alpha} - 1]^{1/\gamma},$$

$$\text{E}[X^k] = \frac{\theta^k \Gamma(1 + k/\gamma)\Gamma(\alpha - k/\gamma)}{\Gamma(\alpha)}, \quad -\gamma < k < \alpha\gamma,$$

$$\text{E}[(X \wedge x)^k] = \frac{\theta^k \Gamma(1 + k/\gamma)\Gamma(\alpha - k/\gamma)}{\Gamma(\alpha)} \beta(1 + k/\gamma, \alpha - k/\gamma; 1 - u)$$

$$+ x^k u^\alpha, \quad k > -\gamma,$$

$$\text{Mode} = \theta\left(\frac{\gamma - 1}{\alpha\gamma + 1}\right)^{1/\gamma}, \quad \gamma > 1, \text{ else } 0.$$

A.2.2.3   *Inverse Burr* – τ, θ, γ (Dagum)

$$f(x) = \frac{\tau\gamma(x/\theta)^{\gamma\tau}}{x[1 + (x/\theta)^\gamma]^{\tau+1}},$$

$$F(x) = u^\tau, \quad u = \frac{(x/\theta)^\gamma}{1 + (x/\theta)^\gamma},$$

$$\text{VaR}_p(X) = \theta(p^{-1/\tau} - 1)^{-1/\gamma},$$

$$\text{E}[X^k] = \frac{\theta^k \Gamma(\tau + k/\gamma)\Gamma(1 - k/\gamma)}{\Gamma(\tau)}, \quad -\tau\gamma < k < \gamma,$$

$$\text{E}[(X \wedge x)^k] = \frac{\theta^k \Gamma(\tau + k/\gamma)\Gamma(1 - k/\gamma)}{\Gamma(\tau)} \beta(\tau + k/\gamma, 1 - k/\gamma; u)$$

$$+ x^k[1 - u^\tau], \quad k > -\tau\gamma,$$

$$\text{Mode} = \theta\left(\frac{\tau\gamma - 1}{\gamma + 1}\right)^{1/\gamma}, \quad \tau\gamma > 1, \text{ else } 0.$$

A.2.3   Two-Parameter Distributions

A.2.3.1   *Pareto* – α, θ (Pareto Type II, Lomax)

$$f(x) = \frac{\alpha\theta^\alpha}{(x + \theta)^{\alpha+1}},$$

$$F(x) = 1 - \left(\frac{\theta}{x + \theta}\right)^\alpha,$$

$$\text{VaR}_p(X) = \theta[(1 - p)^{-1/\alpha} - 1],$$

$$\text{E}[X^k] = \frac{\theta^k \Gamma(k + 1)\Gamma(\alpha - k)}{\Gamma(\alpha)}, \quad -1 < k < \alpha,$$

$$\text{E}[X^k] = \frac{\theta^k k!}{(\alpha - 1) \cdots (\alpha - k)} \quad \text{if } k \text{ is a positive integer,}$$

$$\text{E}[X \wedge x] = \frac{\theta}{\alpha - 1}\left[1 - \left(\frac{\theta}{x + \theta}\right)^{\alpha-1}\right], \quad \alpha \neq 1,$$

$$\mathrm{E}[X \wedge x] = -\theta \ln\left(\frac{\theta}{x+\theta}\right), \quad \alpha = 1,$$

$$\mathrm{TVaR}_p(X) = \mathrm{VaR}_p(X) + \frac{\theta(1-p)^{-1/\alpha}}{\alpha-1}, \quad \alpha > 1,$$

$$\mathrm{E}[(X \wedge x)^k] = \frac{\theta^k \Gamma(k+1)\Gamma(\alpha-k)}{\Gamma(\alpha)} \beta[k+1, \alpha-k; x/(x+\theta)]$$

$$+ x^k \left(\frac{\theta}{x+\theta}\right)^\alpha, \quad k > -1,\ k \neq \alpha,$$

$$\mathrm{E}[(X \wedge x)^\alpha] = \theta^\alpha \left(\frac{x}{x+\theta}\right)^\alpha \left[1 + \alpha \sum_{n=0}^{\infty} \frac{[x/(x+\theta)]^{n+1}}{\alpha+n+1}\right],$$

$$\mathrm{Mode} = 0,$$

$$\hat{\alpha} = 2\frac{t - m^2}{t - 2m^2}, \quad \hat{\theta} = \frac{mt}{t - 2m^2}.$$

A.2.3.2 *Inverse Pareto* – τ, θ

$$f(x) = \frac{\tau\theta x^{\tau-1}}{(x+\theta)^{\tau+1}},$$

$$F(x) = \left(\frac{x}{x+\theta}\right)^\tau,$$

$$\mathrm{VaR}_p(X) = \theta[p^{-1/\tau} - 1]^{-1},$$

$$\mathrm{E}[X^k] = \frac{\theta^k \Gamma(\tau+k)\Gamma(1-k)}{\Gamma(\tau)}, \quad -\tau < k < 1,$$

$$\mathrm{E}[X^k] = \frac{\theta^k(-k)!}{(\tau-1)\cdots(\tau+k)} \quad \text{if } k \text{ is a negative integer,}$$

$$\mathrm{E}[(X \wedge x)^k] = \theta^k \tau \int_0^{x/(x+\theta)} y^{\tau+k-1}(1-y)^{-k} dy$$

$$+ x^k \left[1 - \left(\frac{x}{x+\theta}\right)^\tau\right], \quad k > -\tau,$$

$$\mathrm{Mode} = \theta\frac{\tau-1}{2}, \quad \tau > 1,\ \text{else } 0.$$

A.2.3.3 *Loglogistic* – γ, θ (Fisk)

$$f(x) = \frac{\gamma(x/\theta)^\gamma}{x[1 + (x/\theta)^\gamma]^2},$$

$$F(x) = u, \quad u = \frac{(x/\theta)^\gamma}{1 + (x/\theta)^\gamma},$$

$$\mathrm{VaR}_p(X) = \theta(p^{-1} - 1)^{-1/\gamma},$$

$$\mathrm{E}[X^k] = \theta^k \Gamma(1 + k/\gamma)\Gamma(1 - k/\gamma), \quad -\gamma < k < \gamma,$$

$$\mathrm{E}[(X \wedge x)^k] = \theta^k \Gamma(1 + k/\gamma)\Gamma(1 - k/\gamma)\beta(1 + k/\gamma, 1 - k/\gamma; u)$$

$$+ x^k(1 - u), \quad k > -\gamma,$$

$$\mathrm{Mode} = \theta\left(\frac{\gamma-1}{\gamma+1}\right)^{1/\gamma}, \quad \gamma > 1,\ \text{else } 0,$$

$$\hat{\gamma} = \frac{2\ln(3)}{\ln(q) - \ln(p)}, \quad \hat{\theta} = \exp\left(\frac{\ln(q) + \ln(p)}{2}\right).$$

A.2.3.4   *Paralogistic* – α, θ   This is a Burr distribution with $\gamma = \alpha$.

$$f(x) = \frac{\alpha^2 (x/\theta)^\alpha}{x[1 + (x/\theta)^\alpha]^{\alpha+1}},$$

$$F(x) = 1 - u^\alpha, \quad u = \frac{1}{1 + (x/\theta)^\alpha},$$

$$\text{VaR}_p(X) = \theta[(1 - p)^{-1/\alpha} - 1]^{1/\alpha},$$

$$\text{E}[X^k] = \frac{\theta^k \Gamma(1 + k/\alpha)\Gamma(\alpha - k/\alpha)}{\Gamma(\alpha)}, \quad -\alpha < k < \alpha^2,$$

$$\text{E}[(X \wedge x)^k] = \frac{\theta^k \Gamma(1 + k/\alpha)\Gamma(\alpha - k/\alpha)}{\Gamma(\alpha)} \beta(1 + k/\alpha, \alpha - k/\alpha; 1 - u)$$

$$+ x^k u^\alpha, \quad k > -\alpha,$$

$$\text{Mode} = \theta \left(\frac{\alpha - 1}{\alpha^2 + 1} \right)^{1/\alpha}, \quad \alpha > 1, \text{ else } 0.$$

Starting values can use estimates from the loglogistic (use γ for α) or Pareto (use α) distributions.

A.2.3.5   *Inverse Paralogistic* – τ, θ   This is an inverse Burr distribution with $\gamma = \tau$.

$$f(x) = \frac{\tau^2 (x/\theta)^{\tau^2}}{x[1 + (x/\theta)^\tau]^{\tau+1}},$$

$$F(x) = u^\tau, \quad u = \frac{(x/\theta)^\tau}{1 + (x/\theta)^\tau},$$

$$\text{VaR}_p(X) = \theta(p^{-1/\tau} - 1)^{-1/\tau},$$

$$\text{E}[X^k] = \frac{\theta^k \Gamma(\tau + k/\tau)\Gamma(1 - k/\tau)}{\Gamma(\tau)}, \quad -\tau^2 < k < \tau,$$

$$\text{E}[(X \wedge x)^k] = \frac{\theta^k \Gamma(\tau + k/\tau)\Gamma(1 - k/\tau)}{\Gamma(\tau)} \beta(\tau + k/\tau, 1 - k/\tau; u)$$

$$+ x^k[1 - u^\tau], \quad k > -\tau^2,$$

$$\text{Mode} = \theta (\tau - 1)^{1/\tau}, \quad \tau > 1, \text{ else } 0.$$

Starting values can use estimates from the loglogistic (use γ for τ) or inverse Pareto (use τ) distributions.

A.3   The Transformed Gamma Family

A.3.1   Three-Parameter Distributions

A.3.1.1   *Transformed Gamma* – α, θ, τ   (generalized gamma)

$$f(x) = \frac{\tau u^\alpha e^{-u}}{x \Gamma(\alpha)}, \quad u = (x/\theta)^\tau,$$

$$F(x) = \Gamma(\alpha; u),$$

$$\text{E}[X^k] = \frac{\theta^k \Gamma(\alpha + k/\tau)}{\Gamma(\alpha)}, \quad k > -\alpha\tau,$$

$$\mathrm{E}[(X \wedge x)^k] = \frac{\theta^k \Gamma(\alpha + k/\tau)}{\Gamma(\alpha)} \Gamma(\alpha + k/\tau; u)$$
$$+ x^k [1 - \Gamma(\alpha; u)], \quad k > -\alpha\tau,$$
$$\mathrm{Mode} = \theta \left(\frac{\alpha\tau - 1}{\tau} \right)^{1/\tau}, \quad \alpha\tau > 1, \text{ else } 0.$$

A.3.1.2 *Inverse Transformed Gamma – α, θ, τ* (inverse generalized gamma)

$$f(x) = \frac{\tau u^\alpha e^{-u}}{x\Gamma(\alpha)}, \quad u = (\theta/x)^\tau,$$
$$F(x) = 1 - \Gamma(\alpha; u),$$
$$\mathrm{E}[X^k] = \frac{\theta^k \Gamma(\alpha - k/\tau)}{\Gamma(\alpha)}, \quad k < \alpha\tau,$$
$$\mathrm{E}[(X \wedge x)^k] = \frac{\theta^k \Gamma(\alpha - k/\tau)}{\Gamma(\alpha)} [1 - \Gamma(\alpha - k/\tau; u)] + x^k \Gamma(\alpha; u)$$
$$= \frac{\theta^k G(\alpha - k/\tau; u)}{\Gamma(\alpha)} + x^k \Gamma(\alpha; u), \quad \text{all } k,$$
$$\mathrm{Mode} = \theta \left(\frac{\tau}{\alpha\tau + 1} \right)^{1/\tau}.$$

A.3.2   Two-Parameter Distributions

A.3.2.1 *Gamma – α, θ* (When $\alpha = n/2$ and $\theta = 2$, it is a chi-square distribution with n degrees of freedom.)

$$f(x) = \frac{(x/\theta)^\alpha e^{-x/\theta}}{x\Gamma(\alpha)},$$
$$F(x) = \Gamma(\alpha; x/\theta),$$
$$\mathrm{E}[X^k] = \frac{\theta^k \Gamma(\alpha + k)}{\Gamma(\alpha)}, \quad k > -\alpha,$$
$$\mathrm{E}[X^k] = \theta^k (\alpha + k - 1) \cdots \alpha \quad \text{if } k \text{ is a positive integer,}$$
$$\mathrm{E}[(X \wedge x)^k] = \frac{\theta^k \Gamma(\alpha + k)}{\Gamma(\alpha)} \Gamma(\alpha + k; x/\theta) + x^k [1 - \Gamma(\alpha; x/\theta)], \quad k > -\alpha,$$
$$\mathrm{E}[(X \wedge x)^k] = \alpha(\alpha + 1) \cdots (\alpha + k - 1)\theta^k \Gamma(\alpha + k; x/\theta)$$
$$+ x^k [1 - \Gamma(\alpha; x/\theta)] \quad \text{if } k \text{ is a positive integer,}$$
$$M(t) = (1 - \theta t)^{-\alpha}, \quad t < 1/\theta,$$
$$\mathrm{Mode} = \theta(\alpha - 1), \quad \alpha > 1, \text{ else } 0,$$
$$\hat{\alpha} = \frac{m^2}{t - m^2}, \quad \hat{\theta} = \frac{t - m^2}{m}.$$

A.3.2.2 *Inverse Gamma – α, θ* (Vinci)

$$f(x) = \frac{(\theta/x)^\alpha e^{-\theta/x}}{x\Gamma(\alpha)},$$
$$F(x) = 1 - \Gamma(\alpha; \theta/x),$$
$$\mathrm{E}[X^k] = \frac{\theta^k \Gamma(\alpha - k)}{\Gamma(\alpha)}, \quad k < \alpha,$$

$$E[X^k] = \frac{\theta^k}{(\alpha - 1) \cdots (\alpha - k)} \quad \text{if } k \text{ is a positive integer,}$$

$$E[(X \wedge x)^k] = \frac{\theta^k \Gamma(\alpha - k)}{\Gamma(\alpha)}[1 - \Gamma(\alpha - k; \theta/x)] + x^k \Gamma(\alpha; \theta/x)$$

$$= \frac{\theta^k G(\alpha - k; \theta/x)}{\Gamma(\alpha)} + x^k \Gamma(\alpha; \theta/x), \text{ all } k,$$

$$\text{Mode} = \theta/(\alpha + 1),$$

$$\hat{\alpha} = \frac{2t - m^2}{t - m^2}, \quad \hat{\theta} = \frac{mt}{t - m^2}.$$

A.3.2.3   Weibull – θ, τ

$$f(x) = \frac{\tau(x/\theta)^\tau e^{-(x/\theta)^\tau}}{x},$$

$$F(x) = 1 - e^{-(x/\theta)^\tau},$$

$$\text{VaR}_p(X) = \theta[-\ln(1 - p)]^{1/\tau},$$

$$E[X^k] = \theta^k \Gamma(1 + k/\tau), \quad k > -\tau,$$

$$E[(X \wedge x)^k] = \theta^k \Gamma(1 + k/\tau)\Gamma[1 + k/\tau; (x/\theta)^\tau] + x^k e^{-(x/\theta)^\tau}, \quad k > -\tau,$$

$$\text{Mode} = \theta \left(\frac{\tau - 1}{\tau} \right)^{1/\tau}, \quad \tau > 1, \text{ else } 0,$$

$$\hat{\theta} = \exp \left(\frac{g \ln(p) - \ln(q)}{g - 1} \right), \quad g = \frac{\ln(\ln(4))}{\ln(\ln(4/3))},$$

$$\hat{\tau} = \frac{\ln(\ln(4))}{\ln(q) - \ln(\hat{\theta})}.$$

A.3.2.4   Inverse Weibull – θ, τ (log-Gompertz)

$$f(x) = \frac{\tau(\theta/x)^\tau e^{-(\theta/x)^\tau}}{x},$$

$$F(x) = e^{-(\theta/x)^\tau},$$

$$\text{VaR}_p(X) = \theta(-\ln p)^{-1/\tau},$$

$$E[X^k] = \theta^k \Gamma(1 - k/\tau), \quad k < \tau,$$

$$E[(X \wedge x)^k] = \theta^k \Gamma(1 - k/\tau)\{1 - \Gamma[1 - k/\tau; (\theta/x)^\tau]\}$$

$$+ x^k \left[1 - e^{-(\theta/x)^\tau} \right],$$

$$= \theta^k G[1 - k/\tau; (\theta/x)^\tau] + x^k \left[1 - e^{-(\theta/x)^\tau} \right], \quad \text{all } k,$$

$$\text{Mode} = \theta \left(\frac{\tau}{\tau + 1} \right)^{1/\tau},$$

$$\hat{\theta} = \exp \left(\frac{g \ln(q) - \ln(p)}{g - 1} \right), \quad g = \frac{\ln[\ln(4)]}{\ln[\ln(4/3)]},$$

$$\hat{\tau} = \frac{\ln(\ln(4))}{\ln(\hat{\theta}) - \ln(p)}.$$

A.3.3   One-Parameter Distributions

A.3.3.1   Exponential – θ

$$f(x) = \frac{e^{-x/\theta}}{\theta},$$
$$F(x) = 1 - e^{-x/\theta},$$
$$\text{VaR}_p(X) = -\theta \ln(1 - p),$$
$$\text{E}[X^k] = \theta^k \Gamma(k + 1), \quad k > -1,$$
$$\text{E}[X^k] = \theta^k k! \quad \text{if } k \text{ is a positive integer,}$$
$$\text{E}[X \wedge x] = \theta(1 - e^{-x/\theta}),$$
$$\text{TVaR}_p(X) = -\theta \ln(1 - p) + \theta,$$
$$\text{E}[(X \wedge x)^k] = \theta^k \Gamma(k + 1)\Gamma(k + 1; x/\theta) + x^k e^{-x/\theta}, \quad k > -1,$$
$$\text{E}[(X \wedge x)^k] = \theta^k k! \Gamma(k + 1; x/\theta) + x^k e^{-x/\theta} \quad \text{if } k > -1 \text{ is an integer,}$$
$$M(z) = (1 - \theta z)^{-1}, \quad z < 1/\theta,$$
$$\text{Mode} = 0,$$
$$\hat{\theta} = m.$$

A.3.3.2   Inverse Exponential – θ

$$f(x) = \frac{\theta e^{-\theta/x}}{x^2},$$
$$F(x) = e^{-\theta/x},$$
$$\text{VaR}_p(X) = \theta(-\ln p)^{-1},$$
$$\text{E}[X^k] = \theta^k \Gamma(1 - k), \quad k < 1,$$
$$\text{E}[(X \wedge x)^k] = \theta^k G(1 - k; \theta/x) + x^k(1 - e^{-\theta/x}), \quad \text{all } k,$$
$$\text{Mode} = \theta/2,$$
$$\hat{\theta} = -q \ln(3/4).$$

A.4   Distributions for Large Losses

The general form of most of these distributions has probability starting or ending at an arbitrary location. The versions presented here all use zero for that point. The distribution can always be shifted to start or end elsewhere.

A.4.1   Extreme Value Distributions

A.4.1.1   Gumbel – θ, μ   (μ can be negative)

$$f(x) = \frac{1}{\theta} \exp(-y) \exp\left[-\exp(-y)\right], \ y = \frac{x - \mu}{\theta}, \ -\infty < x < \infty,$$
$$F(x) = \exp\left[-\exp(-y)\right],$$
$$\text{VaR}_p(X) = \mu + \theta[-\ln(-\ln p)],$$

$$M(z) = e^{\mu t}\Gamma(1 - \theta z), \ z < 1/\theta,$$
$$E[X] = \mu + 0.577215664901530\theta,$$
$$\text{Var}(X) = \frac{\pi^2\theta^2}{6}.$$

A.4.1.2 *Frechet* – α, θ This is the inverse Weibull distribution of Section A.3.2.4.

$$f(x) = \frac{\alpha(x/\theta)^{-\alpha}e^{-(x/\theta)^{-\alpha}}}{x},$$
$$F(x) = e^{-(x/\theta)^{-\alpha}},$$
$$\text{VaR}_p(X) = \theta(-\ln p)^{1/\alpha},$$
$$E[X^k] = \theta^k\Gamma(1 - k/\alpha), \quad k < \alpha,$$
$$E[(X \wedge x)^k] = \theta^k\Gamma(1 - k/\alpha)\{1 - \Gamma[1 - k/\alpha; (x/\theta)^{-\alpha}]\}$$
$$+ x^k\left[1 - e^{-(x/\theta)^{-\alpha}}\right],$$
$$= \theta^k G[1 - k/\alpha; (x/\theta)^{-\alpha}] + x^k\left[1 - e^{-(x/\theta)^{-\alpha}}\right], \quad \text{all } k.$$

A.4.1.3 *Weibull*[3] – α, θ

$$f(x) = -\frac{\alpha(-x/\theta)^{\alpha}e^{-(-x/\theta)^{\alpha}}}{x}, \quad x \leq 0,$$
$$F(x) = e^{-(-x/\theta)^{\alpha}}, \quad x \leq 0,$$
$$E[X^k] = (-1)^k\theta^k\Gamma(1 + k/\alpha), \quad k > -\alpha, \ k \text{ an integer},$$
$$\text{Mode} = -\theta\left(\frac{\alpha - 1}{\alpha}\right)^{1/\alpha}, \quad \alpha > 1, \text{ else } 0.$$

A.4.2 Generalized Pareto Distributions

A.4.2.1 *Generalized Pareto* – γ, θ This is the Pareto distribution of Section A.2.3.1 with α replaced by $1/\gamma$ and θ replaced by $\alpha\theta$.

$$F(x) = 1 - \left(1 + \gamma\frac{x}{\theta}\right)^{-1/\gamma}, \ x \geq 0.$$

A.4.2.2 *Exponential* – θ This is the same as the exponential distribution of Section A.3.3.1 and is the limiting case of the above distribution as $\gamma \to 0$.

A.4.2.3 *Pareto* – γ, θ This is the single-parameter Pareto distribution of Section A.5.1.4. From the above distribution, shift the probability to start at θ.

A.4.2.4 *Beta* – α, θ This is the beta distribution of Section A.6.1.2 with $a = 1$.

---

[3]This is not the same Weibull distribution as in Section A.3.2.3. It is the negative of a Weibull distribution.

A.5   Other Distributions

A.5.1.1 *Lognormal – μ, σ*   (μ can be negative)

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp(-z^2/2) = \phi(z)/(\sigma x), \quad z = \frac{\ln x - \mu}{\sigma},$$

$$F(x) = \Phi(z),$$

$$E[X^k] = \exp\left(k\mu + \tfrac{1}{2}k^2\sigma^2\right),$$

$$E[(X \wedge x)^k] = \exp\left(k\mu + \tfrac{1}{2}k^2\sigma^2\right)\Phi\left(\frac{\ln x - \mu - k\sigma^2}{\sigma}\right) + x^k[1 - F(x)],$$

$$\text{Mode} = \exp(\mu - \sigma^2),$$

$$\hat{\sigma} = \sqrt{\ln(t) - 2\ln(m)}, \quad \hat{\mu} = \ln(m) - \tfrac{1}{2}\hat{\sigma}^2.$$

A.5.1.2 *Inverse Gaussian – μ, θ*

$$f(x) = \left(\frac{\theta}{2\pi x^3}\right)^{1/2}\exp\left(-\frac{\theta z^2}{2x}\right), \quad z = \frac{x - \mu}{\mu},$$

$$F(x) = \Phi\left[z\left(\frac{\theta}{x}\right)^{1/2}\right] + \exp\left(\frac{2\theta}{\mu}\right)\Phi\left[-y\left(\frac{\theta}{x}\right)^{1/2}\right], \quad y = \frac{x + \mu}{\mu},$$

$$E[X] = \mu, \quad \text{Var}[X] = \mu^3/\theta,$$

$$E[X^k] = \sum_{n=0}^{k-1}\frac{(k+n-1)!}{(k-n-1)!n!}\frac{\mu^{n+k}}{(2\theta)^n}, \quad k = 1, 2, \ldots,$$

$$E[X \wedge x] = x - \mu z\Phi\left[z\left(\frac{\theta}{x}\right)^{1/2}\right] - \mu y\exp(2\theta/\mu)\Phi\left[-y\left(\frac{\theta}{x}\right)^{1/2}\right],$$

$$M(z) = \exp\left[\frac{\theta}{\mu}\left(1 - \sqrt{1 - \frac{2\mu^2}{\theta}z}\right)\right], \quad z < \frac{\theta}{2\mu^2},$$

$$\hat{\mu} = m, \quad \hat{\theta} = \frac{m^3}{t - m^2}.$$

A.5.1.3 *Log-t – r, μ, σ*   (μ can be negative) Let Y have a t distribution with r degrees of freedom. Then, $X = \exp(\sigma Y + \mu)$ has the log-t distribution. Positive moments do not exist for this distribution. Just as the t distribution has a heavier tail than the normal distribution, this distribution has a heavier tail than the lognormal distribution.

$$f(x) = \frac{\Gamma\left(\dfrac{r+1}{2}\right)}{x\sigma\sqrt{\pi r}\,\Gamma\left(\dfrac{r}{2}\right)\left[1 + \dfrac{1}{r}\left(\dfrac{\ln x - \mu}{\sigma}\right)^2\right]^{(r+1)/2}},$$

$$F(x) = F_r\left(\frac{\ln x - \mu}{\sigma}\right) \text{ with } F_r(t) \text{ the cdf of a } t \text{ distribution with } r \text{ degrees of freedom,}$$

$$F(x) = \begin{cases} \dfrac{1}{2}\beta\left[\dfrac{r}{2}, \dfrac{1}{2}; \dfrac{r}{r + \left(\dfrac{\ln x - \mu}{\sigma}\right)^2}\right], & 0 < x \le e^{\mu}, \\[3em] 1 - \dfrac{1}{2}\beta\left[\dfrac{r}{2}, \dfrac{1}{2}; \dfrac{r}{r + \left(\dfrac{\ln x - \mu}{\sigma}\right)^2}\right], & x \ge e^{\mu}. \end{cases}$$

A.5.1.4   Single-Parameter Pareto – α, θ

$$f(x) = \frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}, \quad x > \theta,$$

$$F(x) = 1 - \left(\frac{\theta}{x}\right)^{\alpha}, \quad x > \theta,$$

$$\mathrm{VaR}_p(X) = \theta(1 - p)^{-1/\alpha},$$

$$\mathrm{E}[X^k] = \frac{\alpha\theta^k}{\alpha - k}, \quad k < \alpha,$$

$$\mathrm{E}[(X \wedge x)^k] = \frac{\alpha\theta^k}{\alpha - k} - \frac{k\theta^{\alpha}}{(\alpha - k)x^{\alpha-k}}, \quad x \ge \theta, \, k \ne \alpha,$$

$$\mathrm{E}[(X \wedge x)^{\alpha}] = \theta^{\alpha}[1 + \alpha\ln(x/\theta)],$$

$$\mathrm{TVaR}_p(X) = \frac{\alpha\theta(1 - p)^{-1/\alpha}}{\alpha - 1}, \quad \alpha > 1,$$

$$\mathrm{Mode} = \theta,$$

$$\hat{\alpha} = \frac{m}{m - \theta}.$$

Note: Although there appear to be two parameters, only α is a true parameter. The value of θ must be set in advance.

A.6   Distributions with Finite Support

For these two distributions, the scale parameter θ is assumed to be known.

A.6.1.1   Generalized Beta – a, b, θ, τ

$$f(x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}u^a(1 - u)^{b-1}\frac{\tau}{x}, \quad 0 < x < \theta, \quad u = (x/\theta)^{\tau},$$

$$F(x) = \beta(a, b; u),$$

$$\mathrm{E}[X^k] = \frac{\theta^k\Gamma(a + b)\Gamma(a + k/\tau)}{\Gamma(a)\Gamma(a + b + k/\tau)}, \quad k > -a\tau,$$

$$\mathrm{E}[(X \wedge x)^k] = \frac{\theta^k\Gamma(a + b)\Gamma(a + k/\tau)}{\Gamma(a)\Gamma(a + b + k/\tau)}\beta(a + k/\tau, b; u) + x^k[1 - \beta(a, b; u)].$$

A.6.1.2 Beta – a, b, θ   The case $\theta = 1$ has no special name but is the commonly used version of this distribution.

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^a (1-u)^{b-1} \frac{1}{x}, \quad 0 < x < \theta, \quad u = x/\theta,$$

$$F(x) = \beta(a, b; u),$$

$$E[X^k] = \frac{\theta^k \Gamma(a+b)\Gamma(a+k)}{\Gamma(a)\Gamma(a+b+k)}, \quad k > -a,$$

$$E[X^k] = \frac{\theta^k a(a+1) \cdots (a+k-1)}{(a+b)(a+b+1) \cdots (a+b+k-1)} \quad \text{if } k \text{ is a positive integer,}$$

$$E[(X \wedge x)^k] = \frac{\theta^k a(a+1) \cdots (a+k-1)}{(a+b)(a+b+1) \cdots (a+b+k-1)} \beta(a+k, b; u)$$
$$+ x^k [1 - \beta(a, b; u)],$$

$$\hat{a} = \frac{\theta m^2 - mt}{\theta t - \theta m^2}, \quad \hat{b} = \frac{(\theta m - t)(\theta - m)}{\theta t - \theta m^2}.$$

APPENDIX B

AN INVENTORY OF DISCRETE DISTRIBUTIONS

B.1 Introduction

The 16 models presented in this appendix fall into three classes. The divisions are based on the algorithm used to compute the probabilities. For some of the more familiar distributions, these formulas will look different from the ones you may have learned, but they produce the same probabilities. After each name, the parameters are given. All parameters are positive unless otherwise indicated. In all cases, p_k is the probability of observing k losses.

For finding moments, the most convenient form is to give the factorial moments. The jth factorial moment is $\mu_{(j)} = \mathrm{E}[N(N-1)\cdots(N-j+1)]$. We have $\mathrm{E}[N] = \mu_{(1)}$ and $\mathrm{Var}(N) = \mu_{(2)} + \mu_{(1)} - \mu_{(1)}^2$.

The estimators presented are not intended to be useful estimators but, rather, provide starting values for maximizing the likelihood (or other) function. For determining starting values, the following quantities are used (where n_k is the observed frequency at k [if, for the last entry, n_k represents the number of observations at k or more, assume it was at exactly k] and n is the sample size):

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{\infty} k n_k, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{\infty} k^2 n_k - \hat{\mu}^2.$$

When the method of moments is used to determine the starting value, a circumflex (e.g. $\hat{\lambda}$) is used. For any other method, a tilde (e.g. $\tilde{\lambda}$) is used. When the starting value formulas do not provide admissible parameter values, a truly crude guess is to set the product of all λ and β parameters equal to the sample mean and set all other parameters equal to 1. If there are two λ or β parameters, an easy choice is to set each to the square root of the sample mean.

The last item presented is the probability generating function,

$$P(z) = E[z^N].$$

B.2   The $(a, b, 0)$ Class

The distributions in this class have support on $0, 1, \ldots$. For this class, a particular distribution is specified by setting p_0 and then using $p_k = (a + b/k)p_{k-1}$. Specific members are created by setting p_0, a, and b. For any member, $\mu_{(1)} = (a + b)/(1 - a)$, and for higher j, $\mu_{(j)} = (aj + b)\mu_{(j-1)}/(1 - a)$. The variance is $(a + b)/(1 - a)^2$.

B.2.1.1   Poisson – λ

$$p_0 = e^{-\lambda}, \quad a = 0, \quad b = \lambda,$$
$$p_k = \frac{e^{-\lambda}\lambda^k}{k!},$$
$$E[N] = \lambda, \quad Var[N] = \lambda,$$
$$\hat{\lambda} = \hat{\mu},$$
$$P(z) = e^{\lambda(z-1)}.$$

B.2.1.2   Geometric – β

$$p_0 = \frac{1}{1 + \beta}, \quad a = \frac{\beta}{1 + \beta}, \quad b = 0,$$
$$p_k = \frac{\beta^k}{(1 + \beta)^{k+1}},$$
$$E[N] = \beta, \quad Var[N] = \beta(1 + \beta),$$
$$\hat{\beta} = \hat{\mu},$$
$$P(z) = [1 - \beta(z - 1)]^{-1}, \quad -(1 + 1/\beta) < z < 1 + 1/\beta.$$

This is a special case of the negative binomial with $r = 1$.

B.2.1.3   Binomial – q, m $(0 < q < 1, m$ an integer$)$

$$p_0 = (1 - q)^m, \quad a = -\frac{q}{1 - q}, \quad b = \frac{(m + 1)q}{1 - q},$$
$$p_k = \binom{m}{k}q^k(1 - q)^{m-k}, \quad k = 0, 1, \ldots, m,$$
$$E[N] = mq, \quad Var[N] = mq(1 - q),$$
$$\hat{q} = \hat{\mu}/m,$$
$$P(z) = [1 + q(z - 1)]^m.$$

B.2.1.4   Negative Binomial – β, r

$$p_0 = (1 + \beta)^{-r}, \quad a = \frac{\beta}{1 + \beta}, \quad b = \frac{(r - 1)\beta}{1 + \beta},$$

$$p_k = \frac{r(r + 1) \cdots (r + k - 1)\beta^k}{k!(1 + \beta)^{r+k}},$$

$$\mathrm{E}[N] = r\beta, \quad \mathrm{Var}[N] = r\beta(1 + \beta),$$

$$\hat{\beta} = \frac{\hat{\sigma}^2}{\hat{\mu}} - 1, \quad \hat{r} = \frac{\hat{\mu}^2}{\hat{\sigma}^2 - \hat{\mu}},$$

$$P(z) = [1 - \beta(z - 1)]^{-r}, \quad -(1 + 1/\beta) < z < 1 + 1/\beta.$$

B.3   The $(a, b, 1)$ Class

To distinguish this class from the $(a, b, 0)$ class, the probabilities are denoted $\Pr(N = k) = p_k^M$ or $\Pr(N = k) = p_k^T$. depending on which subclass is being represented. For this class, p_0^M is arbitrary (i.e. it is a parameter), and then p_1^M or p_1^T is a specified function of the parameters a and b. Subsequent probabilities are obtained recursively as in the $(a, b, 0)$ class: $p_k^M = (a + b/k)p_{k-1}^M$, $k = 2, 3, \ldots$, with the same recursion for p_k^T. There are two subclasses of this class. When discussing their members, we often refer to the "corresponding" member of the $(a, b, 0)$ class. This refers to the member of that class with the same values for a and b. The notation p_k will continue to be used for probabilities for the corresponding $(a, b, 0)$ distribution.

B.3.1   The Zero-Truncated Subclass

The members of this class have $p_0^T = 0$, and therefore it need not be estimated. These distributions should only be used when a value of zero is impossible. The first factorial moment is $\mu_{(1)} = (a + b)/[(1 - a)(1 - p_0)]$, where p_0 is the value for the corresponding member of the $(a, b, 0)$ class. For the logarithmic distribution (which has no corresponding member), $\mu_{(1)} = \beta/\ln(1 + \beta)$. Higher factorial moments are obtained recursively with the same formula as with the $(a, b, 0)$ class. The variance is $(a+b)[1-(a+b+1)p_0]/[(1-a)(1-p_0)]^2$. For those members of the subclass that have corresponding $(a, b, 0)$ distributions, $p_k^T = p_k/(1 - p_0)$.

B.3.1.1   Zero-Truncated Poisson – λ

$$p_1^T = \frac{\lambda}{e^\lambda - 1}, \quad a = 0, \quad b = \lambda,$$

$$p_k^T = \frac{\lambda^k}{k!(e^\lambda - 1)},$$

$$\mathrm{E}[N] = \lambda/(1 - e^{-\lambda}), \quad \mathrm{Var}[N] = \lambda[1 - (\lambda + 1)e^{-\lambda}]/(1 - e^{-\lambda})^2,$$

$$\tilde{\lambda} = \ln(n\hat{\mu}/n_1),$$

$$P(z) = \frac{e^{\lambda z} - 1}{e^\lambda - 1}.$$

B.3.1.2   Zero-Truncated Geometric – β

$$p_1^T = \frac{1}{1+\beta}, \quad a = \frac{\beta}{1+\beta}, \quad b = 0,$$

$$p_k^T = \frac{\beta^{k-1}}{(1+\beta)^k},$$

$$E[N] = 1 + \beta, \quad \text{Var}[N] = \beta(1+\beta),$$

$$\hat{\beta} = \hat{\mu} - 1,$$

$$P(z) = \frac{[1 - \beta(z-1)]^{-1} - (1+\beta)^{-1}}{1 - (1+\beta)^{-1}}, \quad -(1+1/\beta) < z < 1 + 1/\beta.$$

This is a special case of the zero-truncated negative binomial with $r = 1$.

B.3.1.3   Logarithmic – β

$$p_1^T = \frac{\beta}{(1+\beta)\ln(1+\beta)}, \quad a = \frac{\beta}{1+\beta}, \quad b = -\frac{\beta}{1+\beta},$$

$$p_k^T = \frac{\beta^k}{k(1+\beta)^k \ln(1+\beta)},$$

$$E[N] = \beta/\ln(1+\beta), \quad \text{Var}[N] = \frac{\beta[1 + \beta - \beta/\ln(1+\beta)]}{\ln(1+\beta)},$$

$$\tilde{\beta} = \frac{n\hat{\mu}}{n_1} - 1 \quad \text{or} \quad \frac{2(\hat{\mu}-1)}{\hat{\mu}},$$

$$P(z) = 1 - \frac{\ln[1 - \beta(z-1)]}{\ln(1+\beta)}, \quad -(1+1/\beta) < z < 1 + 1/\beta.$$

This is a limiting case of the zero-truncated negative binomial as $r \to 0$.

B.3.1.4   Zero-Truncated Binomial – q, m, $(0 < q < 1,\ m$ an integer)

$$p_1^T = \frac{m(1-q)^{m-1}q}{1 - (1-q)^m}, \quad a = -\frac{q}{1-q}, \quad b = \frac{(m+1)q}{1-q},$$

$$p_k^T = \frac{\binom{m}{k} q^k (1-q)^{m-k}}{1 - (1-q)^m}, \quad k = 1, 2, \ldots, m,$$

$$E[N] = \frac{mq}{1 - (1-q)^m},$$

$$\text{Var}[N] = \frac{mq[(1-q) - (1-q+mq)(1-q)^m]}{[1 - (1-q)^m]^2},$$

$$\tilde{q} = \frac{\hat{\mu}}{m},$$

$$P(z) = \frac{[1 + q(z-1)]^m - (1-q)^m}{1 - (1-q)^m}.$$

B.3.1.5 Zero-Truncated Negative Binomial – β, r $(r > -1, r \neq 0)$

$$p_1^T = \frac{r\beta}{(1+\beta)^{r+1} - (1+\beta)}, \quad a = \frac{\beta}{1+\beta}, \quad b = \frac{(r-1)\beta}{1+\beta},$$

$$p_k^T = \frac{r(r+1)\cdots(r+k-1)}{k![(1+\beta)^r - 1]} \left(\frac{\beta}{1+\beta}\right)^k,$$

$$E[N] = \frac{r\beta}{1 - (1+\beta)^{-r}},$$

$$Var[N] = \frac{r\beta[(1+\beta) - (1+\beta+r\beta)(1+\beta)^{-r}]}{[1 - (1+\beta)^{-r}]^2},$$

$$\tilde{\beta} = \frac{\hat{\sigma}^2}{\hat{\mu}} - 1, \quad \tilde{r} = \frac{\hat{\mu}^2}{\hat{\sigma}^2 - \hat{\mu}},$$

$$P(z) = \frac{[1 - \beta(z-1)]^{-r} - (1+\beta)^{-r}}{1 - (1+\beta)^{-r}}, \quad -(1+1/\beta) < z < 1 + 1/\beta.$$

This distribution is sometimes called the extended truncated negative binomial distribution because the parameter r can extend below zero.

B.3.2 The Zero-Modified Subclass

A zero-modified distribution is created by starting with a truncated distribution and then placing an arbitrary amount of probability at zero. This probability, p_0^M, is a parameter. The remaining probabilities are adjusted accordingly. Values of p_k^M can be determined from the corresponding zero-truncated distribution as $p_k^M = (1 - p_0^M)p_k^T$ or from the corresponding $(a, b, 0)$ distribution as $p_k^M = (1 - p_0^M)p_k/(1 - p_0)$. The same recursion used for the zero-truncated subclass applies.

The mean is $1 - p_0^M$ times the mean for the corresponding zero-truncated distribution. The variance is $1 - p_0^M$ times the zero-truncated variance plus $p_0^M(1 - p_0^M)$ times the square of the zero-truncated mean. The probability generating function is $P^M(z) = p_0^M + (1 - p_0^M)P(z)$, where $P(z)$ is the probability generating function for the corresponding zero-truncated distribution.

The maximum likelihood estimator of p_0^M is always the sample relative frequency at zero.

B.4 The Compound Class

Members of this class are obtained by compounding one distribution with another. That is, let N be a discrete distribution, called the **primary distribution,** and let $M_1, M_2, \ldots$ be i.i.d. with another discrete distribution, called the **secondary distribution**. The compound distribution is $S = M_1 + \cdots + M_N$. The probabilities for the compound distributions are found from

$$p_k = \frac{1}{1 - af_0} \sum_{y=1}^{k} (a + by/k)f_y p_{k-y}$$

for $n = 1, 2, \ldots$, where a and b are the usual values for the primary distribution (which must be a member of the $(a, b, 0)$ class) and f_y is p_y for the secondary distribution. The

only two primary distributions used here are Poisson (for which $p_0 = \exp[-\lambda(1 - f_0)]$) and geometric [for which $p_0 = 1/(1 + \beta - \beta f_0)$]. Because this information completely describes these distributions, only the names and starting values are given in the following sections.

The moments can be found from the moments of the individual distributions:

$$\mathrm{E}[S] = \mathrm{E}[N]\mathrm{E}[M] \quad \text{and} \quad \mathrm{Var}[S] = \mathrm{E}[N]\,\mathrm{Var}[M] + \mathrm{Var}[N]\mathrm{E}[M]^2.$$

The pgf is $P(z) = P_{\text{primary}}[P_{\text{secondary}}(z)]$.

In the following list, the primary distribution is always named first. For the first, second, and fourth distributions, the secondary distribution is the $(a, b, 0)$ class member with that name. For the third and the last three distributions (the Poisson–ETNB and its two special cases), the secondary distribution is the zero-truncated version.

B.4.1 Some Compound Distributions

B.4.1.1 Poisson–Binomial – λ, q, m $(0 < q < 1, m$ an integer)

$$\hat{q} = \frac{\hat{\sigma}^2/\hat{\mu} - 1}{m - 1}, \quad \hat{\lambda} = \frac{\hat{\mu}}{m\hat{q}} \quad \text{or} \quad \tilde{q} = 0.5, \ \tilde{\lambda} = \frac{2\hat{\mu}}{m}.$$

B.4.1.2 Poisson–Poisson – λ_1, λ_2 The parameter λ_1 is for the primary Poisson distribution, and λ_2 is for the secondary Poisson distribution. This distribution is also called the *Neyman Type A*:

$$\tilde{\lambda}_1 = \tilde{\lambda}_2 = \sqrt{\hat{\mu}}.$$

B.4.1.3 Geometric–Extended Truncated Negative Binomial – β_1, β_2, r $(r > -1)$ The parameter β_1 is for the primary geometric distribution. The last two parameters are for the secondary distribution, noting that for $r = 0$ the secondary distribution is logarithmic. The truncated version is used so that the extension of r is available.

$$\tilde{\beta}_1 = \tilde{\beta}_2 = \sqrt{\hat{\mu}}.$$

B.4.1.4 Geometric–Poisson – β, λ

$$\tilde{\beta} = \tilde{\lambda} = \sqrt{\hat{\mu}}.$$

B.4.1.5 Poisson–Extended Truncated Negative Binomial – λ, β $(r > -1, r \neq 0)$
When $r = 0$ the secondary distribution is logarithmic, resulting in the negative binomial distribution.

$$\tilde{r} = \frac{\hat{\mu}(K - 3\hat{\sigma}^2 + 2\hat{\mu}) - 2(\hat{\sigma}^2 - \hat{\mu})^2}{\hat{\mu}(K - 3\hat{\sigma}^2 + 2\hat{\mu}) - (\hat{\sigma}^2 - \hat{\mu})^2}, \quad \tilde{\beta} = \frac{\hat{\sigma}^2 - \hat{\mu}}{\hat{\mu}(1 + \hat{r})}, \quad \tilde{\lambda} = \frac{\hat{\mu}}{\hat{r}\hat{\beta}},$$

or,

$$\tilde{r} = \frac{\hat{\sigma}^2 n_1/n - \hat{\mu}^2 n_0/n}{(\hat{\sigma}^2 - \hat{\mu}^2)(n_0/n)\ln(n_0/n) - \hat{\mu}(\hat{\mu}n_0/n - n_1/n)},$$

$$\tilde{\beta} = \frac{\hat{\sigma}^2 - \hat{\mu}}{\hat{\mu}(1 + \hat{r})}, \quad \tilde{\lambda} = \frac{\hat{\mu}}{\hat{r}\hat{\beta}},$$

where

$$K = \frac{1}{n} \sum_{k=0}^{\infty} k^3 n_k - 3\hat{\mu} \frac{1}{n} \sum_{k=0}^{\infty} k^2 n_k + 2\hat{\mu}^3.$$

This distribution is also called the *generalized Poisson–Pascal*.

$$\hat{\beta} = \frac{\hat{\sigma}^2 - \hat{\mu}}{2\hat{\mu}}, \quad \hat{\lambda} = \frac{\hat{\mu}}{1 + \hat{\beta}}.$$

This is a special case of the Poisson–extended truncated negative binomial with $r = 1$. It is actually a Poisson–truncated geometric.

$$\tilde{\lambda} = -\ln(n_0/n), \quad \tilde{\beta} = \frac{4(\hat{\mu} - \hat{\lambda})}{\hat{\mu}}.$$

This is a special case of the Poisson–extended truncated negative binomial with $r = -0.5$.

B.5   A Hierarchy of Discrete Distributions

Table B.1 indicates which distributions are special or limiting cases of others. For the special cases, one parameter is set equal to a constant to create the special case. For the limiting cases, two parameters go to infinity or zero in some special way.

Table B.1   The hierarchy of discrete distributions.

Distribution	Is a special case of	Is a limiting case of
Poisson	ZM Poisson	Negative binomial, Poisson–binomial, Poisson–inverse Gaussian, Polya–Aeppli, Neyman–Type A
ZT Poisson	ZM Poisson	ZT negative binomial
ZM Poisson		ZM negative binomial
Geometric	Negative binomial ZM geometric	Geometric–Poisson
ZT geometric	ZT negative binomial	
ZM geometric	ZM negative binomial	
Logarithmic		ZT negative binomial
ZM logarithmic		ZM negative binomial
Binomial	ZM binomial	
Negative binomial	ZM negative binomial	Poisson–ETNB
Poisson–inverse Gaussian	Poisson–ETNB	
Polya–Aeppli	Poisson–ETNB	
Neyman–Type A		Poisson–ETNB

APPENDIX C

FREQUENCY AND SEVERITY RELATIONSHIPS

Let N^L be the number of losses random variable and let X be the severity random variable. If there is a deductible of d imposed, there are two ways to modify X. One is to create Y^L, the amount paid per loss:

$$Y^L = \begin{cases} 0, & X \le d, \\ X - d, & X > d. \end{cases}$$

In this case, the appropriate frequency distribution continues to be N^L.

An alternative approach is to create Y^P, the amount paid per payment:

$$Y^P = \begin{cases} \text{undefined}, & X \le d, \\ X - d, & X > d. \end{cases}$$

In this case, the frequency random variable must be altered to reflect the number of payments. Let this variable be N^P. Assume that, for each loss, the probability is $v = 1 - F_X(d)$ that a payment will result. Further assume that the incidence of making a payment is independent of the number of losses. Then, $N^P = L_1 + L_2 + \cdots + L_N$, where

Table C.1 Parameter adjustments.

N^L	Parameters for N^P
Poisson	$\lambda^* = v\lambda$
ZM Poisson	$p_0^{M*} = \dfrac{p_0^M - e^{-\lambda} + e^{-v\lambda} - p_0^M e^{-v\lambda}}{1 - e^{-\lambda}}, \ \lambda^* = v\lambda$
Binomial	$q^* = vq$
ZM binomial	$p_0^{M*} = \dfrac{p_0^M - (1-q)^m + (1-vq)^m - p_0^M(1-vq)^m}{1-(1-q)^m}$ $q^* = vq$
Negative binomial	$\beta^* = v\beta, \ r^* = r$
ZM negative binomial	$p_0^{M*} = \dfrac{p_0^M - (1+\beta)^{-r} + (1+v\beta)^{-r} - p_0^M(1+v\beta)^{-r}}{1-(1+\beta)^{-r}}$ $\beta^* = v\beta, \ r^* = r$
ZM logarithmic	$p_0^{M*} = 1 - (1 - p_0^M)\ln(1+v\beta)/\ln(1+\beta)$ $\beta^* = v\beta$

L_j is 0 with probability $1 - v$ and is 1 with probability v. Probability generating functions yield the relationships in Table C.1.

The geometric distribution is not presented as it is a special case of the negative binomial with $r = 1$. For zero-truncated distributions, the same formulas are still used as the distribution for N^P will now be zero modified. For compound distributions, modify only the secondary distribution. For ETNB secondary distributions, the parameter for the primary distribution is multiplied by $1 - p_0^{M*}$ as obtained in Table C.1, while the secondary distribution remains zero truncated (however, $\beta^* = v\beta$).

There are occasions on which frequency data are collected that provide a model for N^P. There would have to have been a deductible d in place and therefore v is available. It is possible to recover the distribution for N^L, although there is no guarantee that reversing the process will produce a legitimate probability distribution. The solutions are the same as in Table C.1, only now $v = 1/[1 - F_X(d)]$.

Now suppose that the current frequency model is N^d, which is appropriate for a deductible of d. Also suppose that the deductible is to be changed to d^*. The new frequency for payments is N^{d^*} and is of the same type. Then use Table C.1 with $v = [1 - F_X(d^*)]/[1 - F_X(d)]$.

APPENDIX D

THE RECURSIVE FORMULA

The recursive formula is (where the frequency distribution is a member of the $(a, b, 1)$ class),

$$f_S(x) = \frac{[p_1 - (a+b)p_0]f_X(x) + \sum_{y=1}^{x \wedge m} \left(a + \frac{by}{x}\right) f_X(y) f_S(x-y)}{1 - a f_X(0)},$$

where $f_S(x) = \Pr(S = x)$, $x = 0, 1, 2, \ldots$, $f_X(x) = \Pr(X = x)$, $x = 0, 1, 2, \ldots$, $p_0 = \Pr(N = 0)$, and $p_1 = \Pr(N = 1)$. Note that the severity distribution (X) must place probability on nonnegative integers. The formula must be initialized with the value of $f_S(0)$. These values are given in Table D.1. It should be noted that, if N is a member of the $(a, b, 0)$ class, $p_1 - (a+b)p_0 = 0$, and so the first term will vanish. If N is a member of the compound class, the recursion must be run twice. The first pass uses the secondary distribution for p_0, p_1, a, and b. The second pass uses the output from the first pass as $f_X(x)$ and uses the primary distribution for p_0, p_1, a, and b.

Table D.1   Starting values $[f_S(0)]$ for recursions.

Distribution	$f_S(0)$
Poisson	$\exp[\lambda(f_0 - 1)]$
Geometric	$[1 + \beta(1 - f_0)]^{-1}$
Binomial	$[1 + q(f_0 - 1)]^m$
Negative binomial	$[1 + \beta(1 - f_0)]^{-r}$
ZM Poisson	$p_0^M + (1 - p_0^M)\dfrac{\exp(\lambda f_0) - 1}{\exp(\lambda) - 1}$
ZM geometric	$p_0^M + (1 - p_0^M)\dfrac{f_0}{1 + \beta(1 - f_0)}$
ZM binomial	$p_0^M + (1 - p_0^M)\dfrac{[1 + q(f_0 - 1)]^m - (1 - q)^m}{1 - (1 - q)^m}$
ZM negative binomial	$p_0^M + (1 - p_0^M)\dfrac{[1 + \beta(1 - f_0)]^{-r} - (1 + \beta)^{-r}}{1 - (1 + \beta)^{-r}}$
ZM logarithmic	$p_0^M + (1 - p_0^M)\left\{1 - \dfrac{\ln[1 + \beta(1 - f_0)]}{\ln(1 + \beta)}\right\}$

APPENDIX E

DISCRETIZATION OF THE SEVERITY DISTRIBUTION

There are two relatively simple ways to discretize the severity distribution. One is the method of rounding and the other is a mean-preserving method.

E.1 The Method of Rounding

This method has two features: All probabilities are positive and the probabilities add to 1. Let h be the span and let Y be the discretized version of X. If there are no modifications, then

$$
\begin{aligned}
f_j = \Pr(Y = jh) &= \Pr\left[\left(j - \tfrac{1}{2}\right)h \le X < \left(j + \tfrac{1}{2}\right)h\right] \\
&= F_X\left[\left(j + \tfrac{1}{2}\right)h\right] - F_X\left[\left(j - \tfrac{1}{2}\right)h\right].
\end{aligned}
$$

The recursive formula is then used with $f_X(j) = f_j$. Suppose that a deductible of d, limit of u, and coinsurance of α are to be applied. If the modifications are to be applied before

the discretization, then

$$g_0 = \frac{F_X(d + h/2) - F_X(d)}{1 - F_X(d)},$$

$$g_j = \frac{F_X[d + (j + 1/2)h] - F_X[d + (j - 1/2)h]}{1 - F_X(d)},$$

$$j = 1, \dots, \frac{u - d}{h} - 1,$$

$$g_{(u-d)/h} = \frac{1 - F_X(u - h/2)}{1 - F_X(d)},$$

where $g_j = \Pr(Z = j\alpha h)$ and Z is the modified distribution. This method does not require that the limits be multiples of h, but does require that $u - d$ be a multiple of h. This method gives the probabilities of payments per payment.

Finally, if there is truncation from above at u, change all denominators to $F_X(u) - F_X(d)$ and also change the numerator of $g_{(u-d)/h}$ to $F_X(u) - F_X(u - h/2)$.

E.2   Mean Preserving

This method ensures that the discretized distribution has the same mean as the original severity distribution. With no modifications, the discretization is

$$f_0 = 1 - \frac{E[X \wedge h]}{h},$$

$$f_j = \frac{2E[X \wedge jh] - E[X \wedge (j - 1)h] - E[X \wedge (j + 1)h]}{h}, \quad j = 1, 2, \dots.$$

For the modified distribution,

$$g_0 = 1 - \frac{E[X \wedge d + h] - E[X \wedge d]}{h[1 - F_X(d)]},$$

$$g_j = \frac{2E[X \wedge d + jh] - E[X \wedge d + (j - 1)h] - E[X \wedge d + (j + 1)h]}{h[1 - F_X(d)]},$$

$$j = 1, \dots, \frac{u - d}{h} - 1,$$

$$g_{(u-d)/h} = \frac{E[X \wedge u] - E[X \wedge u - h]}{h[1 - F_X(d)]}.$$

To incorporate truncation from above, change the denominators to

$$h[F_X(u) - F_X(d)]$$

and subtract $h[1 - F_X(u)]$ from the numerators of each of g_0 and $g_{(u-d)/h}$.

E.3   Undiscretization of a Discretized Distribution

Assume that we have $g_0 = \Pr(S = 0)$, the true probability that the random variable is zero. Let $p_j = \Pr(S^* = jh)$, where S^* is a discretized distribution and h is the span.

The following are approximations for the cdf and limited expected value of S, the true distribution that was discretized as S^*. They are all based on the assumption that S has a uniform distribution over the interval from $(j - \frac{1}{2})h$ to $(j + \frac{1}{2})h$ for integral j. The first interval is from 0 to $h/2$, and the probability $p_0 - g_0$ is assumed to be uniformly distributed over it. Let S^{**} be the random variable with this approximate mixed distribution. (It is continuous, except for discrete probability g_0 at zero.) The approximate distribution function can be found by interpolation as follows. First, let

$$F_j = F_{S^{**}}\left[\left(j + \tfrac{1}{2}\right)h\right] = \sum_{i=0}^{j} p_i, \quad j = 0, 1, \dots.$$

Then, for x in the interval $(j - \frac{1}{2})h$ to $(j + \frac{1}{2})h$,

$$F_{S^{**}}(x) = F_{j-1} + \int_{(j-1/2)h}^{x} h^{-1} p_j \, dt = F_{j-1} + \left[x - \left(j - \tfrac{1}{2}\right)h\right] h^{-1} p_j$$

$$= F_{j-1} + \left[x - \left(j - \tfrac{1}{2}\right)h\right] h^{-1}(F_j - F_{j-1})$$

$$= (1 - w)F_{j-1} + wF_j, \quad w = \frac{x}{h} - j + \frac{1}{2}.$$

Because the first interval is only half as wide, the formula for $0 \le x \le h/2$ is

$$F_{S^{**}}(x) = (1 - w)g_0 + wp_0, \quad w = \frac{2x}{h}.$$

It is also possible to express these formulas in terms of the discrete probabilities:

$$F_{S^{**}}(x) = \begin{cases} g_0 + \dfrac{2x}{h}[p_0 - g_0], & 0 < x \le \dfrac{h}{2}, \\[2ex] \displaystyle\sum_{i=0}^{j-1} p_i + \dfrac{x - (j - 1/2)h}{h} p_j, & (j - \tfrac{1}{2})h < x \le (j + \tfrac{1}{2})h. \end{cases}$$

With regard to the limited expected value, expressions for the first and kth LEVs are

$$E(S^{**} \wedge x) = \begin{cases} x(1 - g_0) - \dfrac{x^2}{h}(p_0 - g_0), & 0 < x \le \dfrac{h}{2}, \\[2ex] \dfrac{h}{4}(p_0 - g_0) + \displaystyle\sum_{i=1}^{j-1} ihp_i + \dfrac{x^2 - [(j - 1/2)h]^2}{2h} p_j \\[2ex] \quad + x[1 - F_{S^{**}}(x)], & (j - \tfrac{1}{2})h < x \le (j + \tfrac{1}{2})h, \end{cases}$$

and, for $0 < x \le h/2$,

$$E[(S^{**} \wedge x)^k] = \frac{2x^{k+1}}{h(k + 1)}(p_0 - g_0) + x^k[1 - F_{S^{**}}(x)],$$

while for $(j - \frac{1}{2})h < x \le (j + \frac{1}{2})h$,

$$E[(S^{**} \wedge x)^k] = \frac{(h/2)^k(p_0 - g_0)}{k + 1} + \sum_{i=1}^{j-1} \frac{h^k[(i + \tfrac{1}{2})^{k+1} - (i - \tfrac{1}{2})^{k+1}]}{k + 1} p_i$$

$$+ \frac{x^{k+1} - [(j - \tfrac{1}{2})h]^{k+1}}{h(k + 1)} p_j + x^k[1 - F_{S^{**}}(x)].$$

REFERENCES

1. Åalen, O. (1978), "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics*, **6**, 701–726.

2. Abramowitz, M. and Stegun, I. (1964), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Wiley.

3. Acerbi, C. and Tasche, D. (2002), "On the Coherence of Expected Shortfall," *Journal of Banking and Finance*, **26**, 1487–1503.

4. Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, **19(6)**, 716–723.

5. Allaben, M., Diamantoukos, C., Dicke, A., Gutterman, S., Klugman, S., Lord, R., Luckner, W., Miccolis, R., and Tan, J. (2008), "Principles Underlying Actuarial Science," *Actuarial Practice Forum*, **August 2008**.

6. Arnold, B. (1983), *Pareto Distributions (Statistical Distributions in Scientific Work)*, Vol. 5, Fairland, MD: International

7. Artzner, P., Delbaen, F., Eber, J., and Heath, D. (1997), "Thinking Coherently," *RISK*, **10**, 11, 68–71.

8. Bailey, A. (1943), "Sampling Theory in Casualty Insurance, Parts III through VII," *Proceedings of the Casualty Actuarial Society*, **XXX**, 31–65.

9. Bailey, A. (1950), "Credibility Procedures," *Proceedings of the Casualty Actuarial Society*, **XXXVII**, 7–23, 94–115.

10. Baker, C. (1977), *The Numerical Treatment of Integral Equations*, Oxford: Clarendon Press.

11. Balkema, A. and de Haan, L. (1974), "Residual Life at Great Ages," *Annals of Probability*, **2**, 792–804.

12. Batten, R. (1978), *Mortality Table Construction*, Englewood Cliffs, NJ: Prentice-Hall.

13. Beard, R., Pentikainen, T., and Pesonen, E. (1984), *Risk Theory*, 3rd ed., London: Chapman & Hall.

14. Berger, J. (1985), *Bayesian Inference in Statistical Analysis*, 2nd ed., New York: Springer-Verlag.

15. Bevan, J. (1963), "Comprehensive Medical Insurance – Statistical Analysis for Ratemaking," *Proceedings of the Casualty Actuarial Society*, **L**, 111–128.

16. Box, G. and Muller, M. (1958), "A Note on the Generation of Random Normal Deviates," *Annals of Mathematical Statistics*, **29**, 610–611.

17. Brockett, P. (1991), "Information Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications," with discussion, *Transactions of the Society of Actuaries*, **XLIII**, 73–135.

18. Brown, J., Hollander, M., and Korwar, R. (1974), "Nonparametric Tests of Independence for Censored Data, with Applications to Heart Transplant Studies," in Proschen, F. and Serfling, R., eds., *Reliability and Biometry: Statistical Analysis of Lifelength*, SIAM, 327–354.

19. Bühlmann, H. (1967), "Experience Rating and Credibility," *ASTIN Bulletin,* **4**, 199–207.

20. Bühlmann, H. (1970), *Mathematical Methods in Risk Theory*, Berlin: Springer-Verlag.

21. Bühlmann, H. and Straub, E. (1970), "Glaubwürdigkeit für Schadensätze (credibility for loss ratios)," *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker*, **70**, 111–133.

22. Carlin, B. and Louis, T. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Boca Raton, FL: CRC Press.

23. Carriere, J. (1993), "Nonparametric Estimators of a Distribution Function Based on Mixtures of Gamma Distributions," *Actuarial Research Clearing House*, **1993.3**, 1–11.

24. Clark, D. and Thayer, C. (2004), "A Primer on the Exponential Family of Distributions," *Casualty Actuarial Society Discussion Paper Program*, Arlington, VA: Casualty Actuarial Society, 117–148.

25. Cunningham, R., Herzog, T., and London, R. (2011), *Models for Quantifying Risk*, 4th ed., Winsted, CT: ACTEX.

26. deAlba, E. (2002), "Bayesian Estimation of Outstanding Claim Reserves," *North American Actuarial Journal*, **6**, 1–20.

27. DePril, N. (1986), "On the Exact Computation of the Aggregate Claims Distribution in the Individual Life Model," *ASTIN Bulletin*, **16**, 109–112.

28. Dickson, D., Hardy, M., and Waters, H. (2013), *Actuarial Mathematics for Life Contingent Risks*, 2nd ed., Cambridge: Cambridge University Press.

29. Douglas, J. (1980), *Analysis with Standard Contagious Distributions*, Fairland, MD: International Co-operative Publishing House.

30. Dropkin, L. (1959), "Some Considerations on Automobile Rating Systems Utilizing Individual Driving Records," *Proceedings of the Casualty Actuarial Society*, **XLVI**, 165–176.

31. Efron, B. (1967), "The Two Sample Problem with Censored Data," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **4**, 831–853.

32. Efron, B. (1981), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, **76**, 312–319.

33. Efron, B. (1986), "Why Isn't Everyone a Bayesian?" *The American Statistician*, **40**, 1–11 (including comments and reply).

34. Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

35. Embrechts, P. and Wang, R. (2015), "Seven Proofs for the Subadditivity of Expected Shortfall," *Dependence Modeling*, **3**, 126–140.

36. Ericson, W. (1969), "A Note on the Posterior Mean of a Population Mean," *Journal of the Royal Statistical Society, Series B*, **31**, 332–334.

37. Feller, W. (1968), *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed. rev., New York: Wiley.

38. Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd ed., New York: Wiley.

39. Fellingham, G., Kottas, A., and Hartman, B. (2015), "Bayesian Nonparametric Predictive Modeling of Group Health Claims, *Insurance: Mathematics and Economics*, **60**, 1–10.

40. Fisher, R. and Tippett, L. (1928), "Limiting Forms of the Largest or Smallest Member of a Sample," *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

41. Frees, E. W. (2010), *Regression Modeling with Actuarial and Financial Applications*, New York: Cambridge.

42. Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013), *Bayesian Data Analysis*, 3rd ed., Boca Raton, FL: CRC Press.

43. Gerber, H. (1982), "On the Numerical Evaluation of the Distribution of Aggregate Claims and Its Stop-Loss Premiums," *Insurance: Mathematics and Economics*, **1**, 13–18.

44. Gerber, H. and Jones, D. (1976), "Some Practical Considerations in Connection with the Calculation of Stop-Loss Premiums," *Transactions of the Society of Actuaries*, **XXVIII**, 215–231.

45. Gillam, W. (1992), "Parametrizing the Workers Compensation Experience Rating Plan," *Proceedings of the Casualty Actuarial Society*, **LXXIX**, 21–56.

46. Goovaerts, M. J. and Hoogstad, W. J. (1987), *Credibility Theory, Surveys of Actuarial Studies No. 4*, Rotterdam: Nationale-Nederlanden.

47. Grandell, J. (1997), *Mixed Poisson Processes*, London: Chapman & Hall.

48. Hachemeister, C. A. (1975), "Credibility for Regression Models with Application to Trend," in P. Kahn, ed., *Credibility: Theory and Applications*, New York: Academic Press, 129–163.

49. Hartman, B. (2014), "Bayesian Computational Methods," in Frees, J., Meyers, G., and Derrig, R., eds., *Predictive Modeling Applications in Actuarial Science*, New York: Cambridge University Press.

50. Hartman, B., Richardson, R., and Bateman, R. (2017), "Parameter Uncertainty," Research Report published by the Canadian Institute of Actuaries, the Casualty Actuarial Society, and the Society of Actuaries. Available at https://www.soa.org/research-reports/2017/parameter-uncertainty.

51. Hayne, R. (1994), "Extended Service Contracts," *Proceedings of the Casualty Actuarial Society*, **LXXXI**, 243–302.

52. Herzog, T. (1999), *Introduction to Credibility Theory*, 3rd ed., Winsted, CT: ACTEX.

53. Herzog, T. and Laverty, J. (1995), "Experience of Refinanced FHA Section 203(b) Single Family Mortgages," *Actuarial Research Clearing House*, **1995.1**, 97–129.

54. Hewitt, C., Jr. (1967), "Loss Ratio Distributions – A Model," *Proceedings of the Casualty Actuarial Society*, **LIV**, 70–88.

55. Hogg, R. and Klugman, S. (1984), *Loss Distributions*, New York: Wiley.

56. Hogg, R., McKean, J., and Craig, A. (2005), *Introduction to Mathematical Statistics*, 6th ed., Upper Saddle River, NJ: Prentice-Hall.

57. Holgate, P. (1970), "The Modality of Some Compound Poisson Distributions," *Biometrika*, **57**, 666–667.

58. Hossack, I., Pollard, J., and Zehnwirth, B. (1983), *Introductory Statistics with Applications in General Insurance*, Cambridge: Cambridge University Press.

59. Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, New York: Springer.

60. Hyndman, R. and Fan, Y. (1996), "Sample Quantiles in Statistical Packages," *The American Statistician*, **50**, 361–365.

61. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning, with Applications in R*, New York: Springer.

62. Jewell, W. (1974), "Credibility Is Exact Bayesian for Exponential Families," *ASTIN Bulletin*, **8**, 77–90.

63. Johnson, N., Kotz, S., and Balakrishnan, N. (1994), *Continuous Univariate Distributions*, Vol. 1, 2nd ed., New York: Wiley.

64. Johnson, N., Kotz, S., and Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2, 2nd ed., New York: Wiley.

65. Johnson, N., Kotz, S., and Kemp, A. (1993), *Univariate Discrete Distributions*, 2nd ed., New York: Wiley.

66. Kaplan, E. and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, **53**, 457–481.

67. Karlin, S. and Taylor, H. (1981), *A Second Course in Stochastic Processes*, New York: Academic Press.

68. Keatinge, C. (1999), "Modeling Losses with the Mixed Exponential Distribution," *Proceedings of the Casualty Actuarial Society*, **LXXXVI**, 654–698.

69. Kleiber, C. and Kotz, S. (2003), *Statistical Size Distributions in Economics and Actuarial Sciences*, New York: Wiley.

70. Klein, J. and Moeschberger, M. (2003), *Survival Analysis, Techniques for Censored and Truncated Data*, 2nd ed., New York: Springer.

71. Klugman, S. (1987), "Credibility for Classification Ratemaking via the Hierarchical Linear Model," *Proceedings of the Casualty Actuarial Society*, **LXXIV**, 272–321.

72. Klugman, S. (1992), *Bayesian Statistics in Actuarial Science with Emphasis on Credibility*, Boston: Kluwer.

73. Klugman, S., Panjer, H., and Willmot, G. (2008), *Loss Models: From Data to Decisions*, 3rd ed., New York: Wiley.

74. Klugman, S., Panjer, H., and Willmot, G. (2013), *Loss Models: Further Topics*, New York: Wiley.

75. Klugman, S. and Rioux, J. (2006), "Toward a Unified Approach to Fitting Loss Models," *North American Actuarial Journal*, **10**, 1, 63–83.

76. Kornya, P. (1983), "Distribution of Aggregate Claims in the Individual Risk Model," *Transactions of the Society of Actuaries*, **XXXV**, 837–858.

77. Lawless, J. (2003), *Statistical Models and Methods for Lifetime Data*, 2nd ed., New York: Wiley.

78. Leemis, L. and McQueston, J. (2008), "Univariate Distribution Relationships," *The American Statistician*, **62**, 1, 45-53.

79. Lemaire, J. (1995), *Automobile Insurance: Actuarial Models*, 2nd ed., Boston: Kluwer.

80. Lindley, D. (1987), "The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence and Expert Systems," *Statistical Science*, **2**, 17–24 (also related articles in that issue).

81. London, D. (1988), *Survival Models and Their Estimation*, 3rd ed., Winsted, CT: ACTEX.

82. Longley-Cook, L. (1958), "The Employment of Property and Casualty Actuaries," *Proceedings of the Casualty Actuarial Society*, **XLV**, 9–10.

83. Longley-Cook, L. (1962), "An Introduction to Credibility Theory," *Proceeding of the Casualty Actuarial Society*, **XLIX**, 194–221.

84. Luong, A. and Doray, L. (1996), "Goodness of Fit Test Statistics for the Zeta Family," *Insurance: Mathematics and Economics*, **10**, 45–53.

85. Manistre, B. and Hancock, G. (2005), "Variance of the CTE Estimator," *North American Actuarial Journal*, **9**, 129–156.

86. Meyers, G. (1984), "Empirical Bayesian Credibility for Workers' Compensation Classification Ratemaking," *Proceedings of the Casualty Actuarial Society*, **LXXI**, 96–121.

87. Meyers, G. (1994), "Quantifying the Uncertainty in Claim Severity Estimates for an Excess Layer When Using the Single Parameter Pareto," *Proceedings of the Casualty Actuarial Society*, **LXXXI**, 91–122 (including discussion).

88. Meyers, G. (2007), "Estimating Predictive Distributions for Loss Reserve Models," *Variance*, **1:2**, 248–272.

89. Meyers, G. (2016), *Stochastic Loss Reserving Using Bayesian MCMC Models*, CAS Monograph Series #1, Arlington, VA: Casualty Actuarial Society.

90. Mildenhall, S. (2006), "A Multivariate Bayesian Claim Count Development Model with Closed Form Posterior and Predictive Distributions," *Casualty Actuarial Society Forum*, **2006:Winter**, 451–493.

91. Moore, D. (1986), "Tests of Chi-Squared Type," in D'Agostino, R. and Stephens, M., eds., *Goodness-of-Fit Techniques*, New York: Marcel Dekker, 63–95.

92. Mowbray, A. H. (1914), "How Extensive a Payroll Exposure Is Necessary to Give a Dependable Pure Premium?" *Proceedings of the Casualty Actuarial Society*, **I**, 24–30.

93. Nelson, W. (1972), "Theory and Applications of Hazard Plotting for Censored Failure Data," *Technometrics*, **14**, 945–965.

94. Norberg, R. (1979), "The Credibility Approach to Experience Rating," *Scandinavian Actuarial Journal*, 181–221.

95. Ntzoufras, I. and Dellaportas, P. (2002), "Bayesian Modeling of Outstanding Liabilities Incorporating Claim Count Uncertainty," *North American Actuarial Journal*, **6**, 113–128.

96. Overbeck, L. (2000), "Allocation of Economic Capital in Loan Portfolios," in Franke, J., Haerdle, W., and Stahl, G., eds., *Measuring Risk in Complex Systems*, New York: Springer.

97. Panjer, H. and Lutek, B. (1983), "Practical Aspects of Stop-Loss Calculations," *Insurance: Mathematics and Economics*, **2**, 159–177.

98. Panjer, H. and Wang, S. (1993), "On the Stability of Recursive Formulas," *ASTIN Bulletin*, **23**, 227–258.

99. Panjer, H. and Willmot, G. (1986), "Computational Aspects of Recursive Evaluation of Compound Distributions," *Insurance: Mathematics and Economics*, **5**, 113–116.

100. Panjer, H. and Willmot, G. (1992), *Insurance Risk Models*, Chicago: Society of Actuaries.

101. Pettitt, A. and Stephens, M. (1977), "The Kolmogorov–Smirnov Goodness-of-Fit Statistic with Discrete and Grouped Data," *Technometrics*, **19**, 205–210.

102. Pickands, J. (1975), "Statistical Inference Using Extreme Order Statistics," *Annals of Statistics*, **3**, 119–131.

103. Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1988), *Numerical Recipes in C*, Cambridge: Cambridge University Press.

104. Rao, C. (1965), *Linear Statistical Inference and Its Applications*, New York: Wiley.

105. Ripley, B. (1987), *Stochastic Simulation*, New York: Wiley.

106. Rohatgi, V. (1976), *An Introduction to Probability Theory and Mathematical Statistics*, New York: Wiley.

107. Ross, S. (1996), *Stochastic Processes*, 2nd ed., New York: Wiley.

108. Ross, S. (2006), *Simulation*, 4th ed., San Diego: Academic Press.

109. Ross, S. (2007), *Introduction to Probability Models*, 9th ed., San Diego: Academic Press.

110. Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, **6**, 461–464.

111. Scollnik, D. (2002), "Modeling Size-of-Loss Distributions for Exact Data in WinBUGS," *Journal of Actuarial Practice*, **10**, 193–218.

112. Self, S. and Liang, K. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, **82**, 605–610.

113. Simon, L. (1961), "Fitting Negative Binomial Distributions by the Method of Maximum Likelihood," *Proceedings of the Casualty Actuarial Society*, **XLVIII**, 45–53.

114. Society of Actuaries Committee on Actuarial Principles (1992), "Principles of Actuarial Science," *Transactions of the Society of Actuaries*, **XLIV**, 565–628.

115. Society of Actuaries Committee on Actuarial Principles (1995), "Principles Regarding Provisions for Life Risks," *Transactions of the Society of Actuaries*, **XLVII**, 775–793.

116. Stephens, M. (1986), "Tests Based on EDF Statistics," in D'Agostino, R. and Stephens, M., eds., *Goodness-of-Fit Techniques*, New York: Marcel Dekker, 97–193.

117. Sundt, B. (1986), Special issue on credibility theory, *Insurance: Abstracts and Reviews*, **2**.

118. Sundt, B. (1999), *An Introduction to Non-Life Insurance Mathematics*, 4th ed., Karlsruhe: Verlag Versicherungswirtschaft (VVW).

119. Tasche, D. (2002), "Expected Shortfall and Beyond," *Journal of Banking and Finance*, **26**, 1519–1533.

120. Thyrion, P. (1961), "Contribution a l'Etude du Bonus pour non Sinistre en Assurance Automobile," *ASTIN Bulletin*, **1**, 142–162.

121. Tröbliger, A. (1961), "Mathematische Untersuchungen zur Beitragsruckgewahr in der Kraftfahrversicherung," *Blätter der Deutsche Gesellschaft fur Versicherungsmathematik*, **5**, 327–348.

122. Tukey, J. (1962), "The Future of Data Analysis," *Annals of Mathematical Statistics*, **33**, 1–67.

123. Tweedie, M. C. K. (1984), "An Index which Distinguishes Between some Important Exponential Families," in Ghosh, J. K. and Roy, J., eds., *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, Calcutta: Indian Statistical Institute, 579–604.

124. Venter, G. (1983), "Transformed Beta and Gamma Distributions and Aggregate Losses," *Proceedings of the Casualty Actuarial Society*, **LXX**, 156–193.

125. Verrall, R. (1990), "Bayes and Empirical Bayes Estimation for the Chain Ladder Method," *ASTIN Bulletin*, **20**, 217–243.

126. Wang, S. (1996), "Premium Calculation by Transforming the Layer Premium Density," *ASTIN Bulletin*, **26**, 71–92.

127. Wang, S. (1998), "Implementation of PH Transforms in Ratemaking," *Proceedings of the Casualty Actuarial Society*, **85**, 940–979 .

128. Wang, S., Young, V., and Panjer, H. (1997), "Axiomatic Characterization of Insurance Prices," *Insurance: Mathematics and Economics*, **21**, 173–183.

129. Waters, H. (1984). "An Approach to the Study of Multiple State Models," *Journal of the Institute of Actuaries*, **111(2)**, 363–374.

130. Waters, H. R. (1993), *Credibility Theory*, Edinburgh: Department of Actuarial Mathematics & Statistics, Heriot-Watt University.

131. Whitney, A. W. (1918), "The Theory of Experience Rating," *Proceedings of the Casualty Actuarial Society*, **IV**, 274–292.

132. Wirch, J. (1999), "Raising Value at Risk," *North American Actuarial Journal*, **3**, 106–115.

133. Wuthrich, M. (2007), "Using a Bayesian Approach for Claim Reserving," *Variance*, **1:2**, 292–301.

INDEX

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.