WILEY SERIES IN PROBABILITY AND STATISTICS

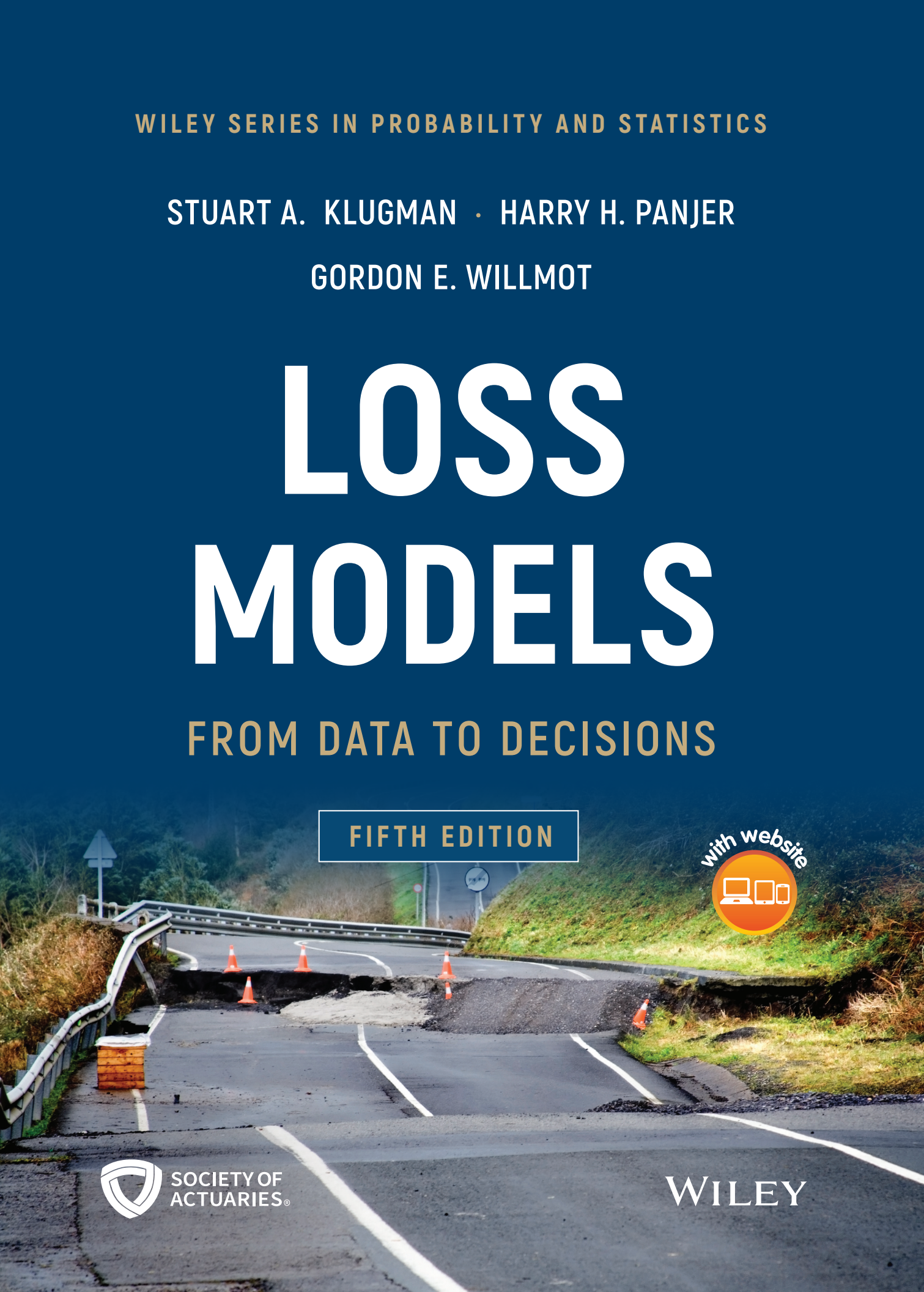
STUART A. KLUGMAN · HARRY H. PANJER

GORDON E. WILLMOT

LOSS MODELS

FROM DATA TO DECISIONS

FIFTH EDITION



SOCIETY OF ACTUARIES®

WILEY

LOSS MODELS

WILEY SERIES IN PROBABILITY AND STATISTICS
Established by *Walter A. Shewhart and Samuel S. Wilks*

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at
http://www.wiley.com/go/wsps

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

WILEY

This edition first published 2019

Edition History
Wiley (1e, 1998; 2e, 2004; 3e, 2008; and 4e, 2012)

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Klugman, Stuart A., 1949- author. | Panjer, Harry H., 1946- author. | Willmot, Gordon E., 1957- author.
Title: Loss models : from data to decisions / Stuart A. Klugman, Society of Actuaries, Harry H. Panjer, University of Waterloo, Gordon E. Willmot, University of Waterloo.
Description: 5th edition. | Hoboken, NJ : John Wiley and Sons, Inc., [2018] | Series: Wiley series in probability and statistics | Includes bibliographical references and index. |
Identifiers: LCCN 2018031122 (print) | LCCN 2018033635 (ebook) | ISBN 9781119523734 (Adobe PDF) | ISBN 9781119523758 (ePub) | ISBN 9781119523789 (hardcover)
Subjects: LCSH: Insurance–Statistical methods. | Insurance–Mathematical models.
Classification: LCC HG8781 (ebook) | LCC HG8781 .K583 2018 (print) | DDC 368/.01–dc23
LC record available at https://lccn.loc.gov/2018031122

Cover image: © iStock.com/hepatus
Cover design by Wiley

Set in 10/12 pt TimesLTStd-Roman by Thomson Digital, Noida, India
"Printed in the United States of America"

10 9 8 7 6 5 4 3 2 1

CONTENTS

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:

Loss Models: From Data to Decisions, Fifth Edition.
Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot.

Companion website: www.wiley.com/go/klugman/lossmodels5e

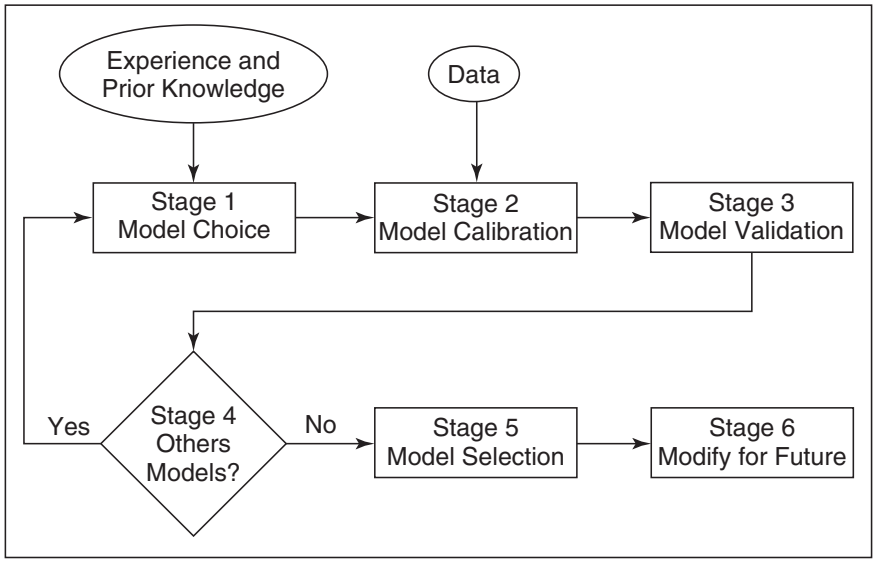


Figure 1.1 The modeling process.

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.

Loss Models: From Data to Decisions, Fifth Edition.
Stuart A. Klugman, Harry H. Panjer, and Gordon E. Willmot.

Companion website: www.wiley.com/go/klugman/lossmodels5e

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, $F_X(x) = \Pr(X \le x)$. The abbreviation **cdf** is often used.*

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

- $0 \le F(x) \le 1$ 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:

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

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.

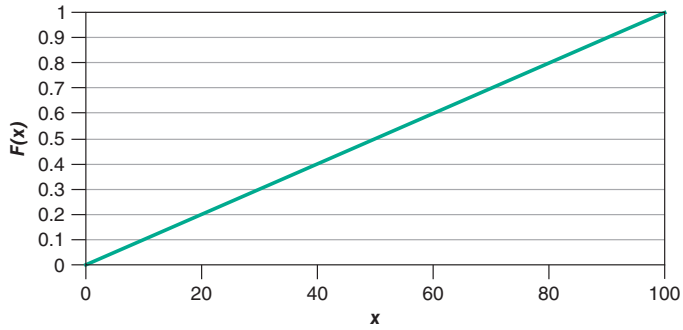


Figure 2.1 The distribution function for Model 1.

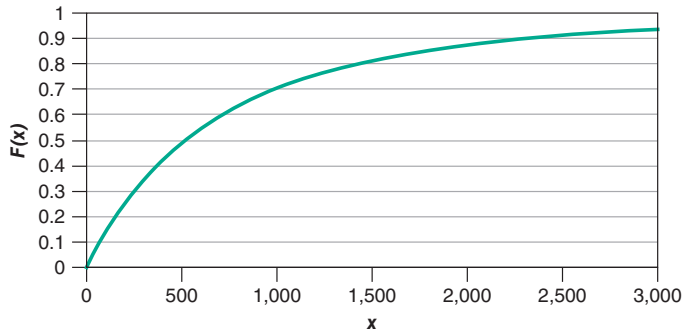


Figure 2.2 The distribution function for Model 2.

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.