Loss Data Analytics

An open text authored by the Actuarial Community

Contents

Preface

Date: 12 August 2019

Book Description

Loss Data Analytics is an interactive, online, freely available text.

- The online version contains many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote deeper learning.
- A subset of the book is available for offline reading in pdf and EPUB formats.
- The online text will be available in multiple languages to promote access to a worldwide audience.

What will success look like?

The online text will be freely available to a worldwide audience. The online version will contain many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote deeper learning. Moreover, a subset of the book will be available in pdf format for low-cost printing. The online text will be available in multiple languages to promote access to a worldwide audience.

How will the text be used?

This book will be useful in actuarial curricula worldwide. It will cover the loss data learning objectives of the major actuarial organizations. Thus, it will be suitable for classroom use at universities as well as for use by independent learners seeking to pass professional actuarial examinations. Moreover, the text will also be useful for the continuing professional development of actuaries and other professionals in insurance and related financial risk management industries.

Why is this good for the profession?

An online text is a type of open educational resource (OER). One important benefit of an OER is that it equalizes access to knowledge, thus permitting a broader community to learn about the actuarial profession. Moreover, it has the capacity to engage viewers through active learning that deepens the learning process, producing analysts more capable of solid actuarial work.

Why is this good for students and teachers and others involved in the learning process? Cost is often cited as an important factor for students and teachers in textbook selection (see a recent post on the $400 textbook). Students will also appreciate the ability to "carry the book around" on their mobile devices.

Why loss data analytics?

The intent is that this type of resource will eventually permeate throughout the actuarial curriculum. Given the dramatic changes in the way that actuaries treat data, loss data seems like a natural place to start. The idea behind the name loss data analytics is to integrate classical loss data models from applied probability with modern analytic tools. In particular, we recognize that big data (including social media and usage based insurance) are here to stay and that high speed computation is readily available.

Project Goal

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. To get involved, please visit our Open Actuarial Textbooks Project Site.

Acknowledgements

Edward Frees acknowledges the John and Anne Oros Distinguished Chair for Inspired Learning in Business which provided seed money to support the project. Frees and his Wisconsin colleagues also acknowledge a Society of Actuaries Center of Excellence Grant that provided funding to support work in dependence modeling and health initiatives. Wisconsin also provided an education innovation grant that provided partial support for the many students who have worked on this project.

We acknowledge the Society of Actuaries for permission to use problems from their examinations.

We thank Rob Hyndman, Monash University, for allowing us to use his excellent style files to produce the online version of the book.

We thank Yihui Xie and his colleagues at Rstudio for the R bookdown package that allows us to produce this book.

We also wish to acknowledge the support and sponsorship of the International Association of Black Actuaries in our joint efforts to provide actuarial educational content to all.

Contributors

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. The following contributors have taken a leadership role in developing Loss Data Analytics.

- **Zeinab Amin** is the Director of the Actuarial Science Program and Associate Dean for Undergraduate Studies of the School of Sciences and Engineering at the American University in Cairo (AUC). Amin holds a PhD in Statistics and is an Associate of the Society of Actuaries. Amin is the recipient of the 2016 Excellence in Academic Service Award and the 2009 Excellence in Teaching Award from AUC. Amin has designed and taught a variety of statistics and actuarial science courses. Amin's current area of research includes quantitative risk assessment, reliability assessment, general statistical modelling, and Bayesian statistics.
- **Katrien Antonio**, KU Leuven
- **Arthur Charpentier** is a professor in the Department of Mathematics at the Université du Québec á Montréal. Prior to that, he worked at a large general insurance company in Hong Kong, China, and the French Federation of Insurers in Paris, France. He received a MS on mathematical economics at Université Paris Dauphine and a MS in actuarial science at ENSAE (National School of Statistics) in Paris, and a PhD degree from KU Leuven, Belgium. His research interests include econometrics, applied probability and actuarial science. He has published several books (the most recent one on Computational Actuarial Science with R, CRC) and papers on a variety of topics. He is a Fellow of the French Institute of Actuaries, and was in charge of the 'Data Science for Actuaries' program from 2015 to 2018.
- **Curtis Gary Dean** is the Lincoln Financial Distinguished Professor of Actuarial Science at Ball State University. He is a Fellow of the Casualty Actuarial Society and a CFA charterholder. He has extensive practical experience as an actuary at American States Insurance, SAFECO, and Travelers. He has served the CAS and actuarial profession as chair of the Examination Committee, first editor-in-chief for Variance: Advancing

the Science of Risk, and as a member of the Board of Directors and the Executive Council. He contributed a chapter to Predictive Modeling Applications in Actuarial Science published by Cambridge University Press.

- **Edward W. (Jed) Frees** is an emeritus professor, formerly the Hickman-Larson Chair of Actuarial Science at the University of Wisconsin-Madison. He is a Fellow of both the Society of Actuaries and the American Statistical Association. He has published extensively (a four-time winner of the Halmstad and Prize for best paper published in the actuarial literature) and has written three books. He also is a co-editor of the two-volume series Predictive Modeling Applications in Actuarial Science published by Cambridge University Press.

- **Guojun Gan** is an assistant professor in the Department of Mathematics at the University of Connecticut, where he has been since August 2014. Prior to that, he worked at a large life insurance company in Toronto, Canada for six years. He received a BS degree from Jilin University, Changchun, China, in 2001 and MS and PhD degrees from York University, Toronto, Canada, in 2003 and 2007, respectively. His research interests include data mining and actuarial science. He has published several books and papers on a variety of topics, including data clustering, variable annuity, mathematical finance, applied statistics, and VBA programming.

- **Lisa Gao** is a PhD candidate in the Risk and Insurance department at the University of Wisconsin-Madison. She holds a BMath in Actuarial Science and Statistics from the University of Waterloo and is an Associate of the Society of Actuaries.

- **José Garrido**, Concordia University

- **Lei (Larry) Hua** is an Associate Professor of Actuarial Science at Northern Illinois University. He earned a PhD degree in Statistics from the University of British Columbia. He is an Associate of the Society of Actuaries. His research work focuses on multivariate dependence modeling for non-Gaussian phenomena and innovative applications for financial and insurance industries.

- **Noriszura Ismail** is a Professor and Head of Actuarial Science Program, Universiti Kebangsaan Malaysia (UKM). She specializes in Risk Modelling and Applied Statistics. She obtained her BSc and MSc (Actuarial Science) in 1991 and 1993 from University of Iowa, and her PhD (Statistics) in 2007 from UKM. She also passed several papers from Society of Actuaries in 1994. She has received several research grants from Ministry of Higher Education Malaysia (MOHE) and UKM, totaling about MYR1.8 million. She has successfully supervised and co-supervised several PhD students (13 completed and 11 on-going). She currently has about 180 publications, consisting of 88 journals and 95 proceedings.

- **Joseph H.T. Kim**, Ph.D., FSA, CERA, is Associate Professor of Applied Statistics at Yonsei University, Seoul, Korea. He holds a Ph.D. degree in Actuarial Science from the University of Waterloo, at which he taught as Assistant Professor. He also worked in the life insurance industry. He has published papers in Insurance Mathematics and Economics, Journal of Risk and Insurance, Journal of Banking and Finance, ASTIN Bulletin, and North American Actuarial Journal, among others.
- **Nii-Armah Okine** is a dissertator at the business school of University of Wisconsin-Madison with a major in actuarial science. He obtained his master's degree in Actuarial science from Illinois State University. His research interests includes micro-level reserving, joint longitudinal-survival modeling, dependence modelling, micro insurance and machine learning.
- **Margie Rosenberg** - University of Wisconsin
- **Emine Selin Sarıdaş** is a doctoral candidate in the Statistics department of Mimar Sinan University. She holds a bachelor degree in Actuarial Science with a minor in Economics and a master degree in Actuarial Science from Hacettepe University. Her research interest includes dependence modeling, regression, loss models and life contingencies.
- **Peng Shi** is an associate professor in the Risk and Insurance Department at the Wisconsin School of Business. He is also the Charles & Laura Albright Professor in Business and Finance. Professor Shi is an Associate of the Casualty Actuarial Society (ACAS) and a Fellow of the Society of Actuaries (FSA). He received a Ph.D. in actuarial science from the University of Wisconsin-Madison. His research interests are problems at the intersection of insurance and statistics. He has won several research awards, including the Charles A. Hachemeister Prize, the Ronald Bornhuetter Loss Reserve Prize, and the American Risk and Insurance Association Prize.
- **Nariankadu D. Shyamalkumar (Shyamal)** is an associate professor in the Department of Statistics and Actuarial Science at The University of Iowa. He is an Associate of the Society of Actuaries, and has volunteered in various elected and non-elected roles within the SoA. Having a broad theoretical interest as well as interest in computing, he has published in prominent actuarial, computer science, probability theory, and statistical journals. Moreover, he has worked in the financial industry, and since then served as an independent consultant to the insurance industry. He has experience educating actuaries in both Mexico and the US, serving in the roles of directing an undergraduate program, and as a graduate adviser for both masters and doctoral students.
- **Jianxi Su** is an Assistant Professor at the Department of Statistics at Purdue University. He is the Associate Director of Purdue's Actuarial Science. Prior to joining Purdue in 2016, he completed the PhD at York University (2012-2015). He obtained the Fellow of the Society of Actuaries (FSA) in 2017. His research expertise are in dependence modelling, risk manage-

ment, and pricing. During the PhD candidature, Jianxi also worked as a research associate at the Model Validation and ORSA Implementation team of Sun Life Financial (Toronto office).

- **Tim Verdonck** is associate professor at the University of Antwerp. He has a degree in Mathematics and a PhD in Science: Mathematics, obtained at the University of Antwerp. During his PhD he successfully took the Master in Insurance and the Master in Financial and Actuarial Engineering, both at KU Leuven. His research focuses on the adaptation and application of robust statistical methods for insurance and finance data.
- **Krupa Viswanathan** is an Associate Professor in the Risk, Insurance and Healthcare Management Department in the Fox School of Business, Temple University. She is an Associate of the Society of Actuaries. She teaches courses in Actuarial Science and Risk Management at the undergraduate and graduate levels. Her research interests include corporate governance of insurance companies, capital management, and sentiment analysis. She received her Ph.D. from The Wharton School of the University of Pennsylvania.

Reviewers

Our goal is to have the actuarial community author our textbooks in a collaborative fashion. Part of the writing process involves many reviewers who generously donated their time to help make this book better. They are:

- Yair Babab
- Chunsheng Ban, Ohio State University
- Vytaras Brazauskas, University of Wisconsin - Milwaukee
- Chun Yong Chew, Universiti Tunku Abdul Rahman (UTAR)
- Eren Dodd, University of Southampton
- Gordon Enderle, University of Wisconsin - Madison
- Rob Erhardt, Wake Forest University
- Runhun Feng, University of Illinois
- Liang (Jason) Hong, Robert Morris University
- Fei Huang, Australian National University
- Hirokazu (Iwahiro) Iwasawa
- Himchan Jeong, University of Connecticut
- Min Ji, Towson University
- Paul Herbert Johnson, University of Wisconsin - Madison
- Samuel Kolins, Lebonan Valley College
- Andrew Kwon-Nakamura, Zurich North America
- Ambrose Lo, University of Iowa
- Mark Maxwell, University of Texas at Austin
- Tatjana Miljkovic, Miami University

- Bell Ouelega, American University in Cairo
- Zhiyu (Frank) Quan, University of Connecticut
- Jiandong Ren, Western University
- Rajesh V. Sahasrabuddhe, Oliver Wyman
- Ranee Thiagarajah, Illinois State University
- Ping Wang, Saint Johns University
- Chengguo Weng, University of Waterloo
- Toby White, Drake University
- Michelle Xia, Northern Illinois University
- Di (Cindy) Xu, University of Nebraska - Lincoln
- Lina Xu, Columbia University
- Lu Yang, University of Amsterdam
- Jorge Yslas, University of Copenhagen
- Jeffrey Zheng, Temple University
- Hongjuan Zhou, Arizona State University

For our Readers

We hope that you find this book worthwhile and even enjoyable. For your convenience, at our Github Landing site (https://openacttexts.github.io/), you will find links to the book that you can (freely) download for offline reading, including a pdf version (for Adobe Acrobat) and an EPUB version suitable for mobile devices. Data for running our examples are available at the same site.

In developing this book, we are emphasizing the online version that has lots of great features such as a glossary, code and solutions to examples that you can be revealed interactively. For example, you will find that the statistical code is hidden and can only be seen by clicking on terms such as

R Code for Frequency Table

```
Insample <- read.csv("Insample.csv", header=T,  na.strings=c("."),
                     stringsAsFactors=FALSE)
Insample2010 <- subset(Insample, Year==2010)
table(Insample2010$Freq)
```

We hide the code because we don't want to insist that you use the `R` statistical software (although we like it). Still, we encourage you to try some statistical code as you read the book – we have opted to make it easy to learn `R` as you go. We have even set up a separate R Code for Loss Data Analytics site to explain more of the details of the code.

Freely available, interactive textbooks represent a new venture in actuarial education and we need your input. Although a lot of effort has gone into the development, we expect hiccoughs. Please let your instructor know about opportunities for improvement, write us through our project site, or contact chapter

contributors directly with suggested improvements.

Chapter 1

Introduction to Loss Data Analytics

Chapter Preview. This book introduces readers to methods of analyzing insurance data. Section **??** begins with a discussion of why the use of data is important in the insurance industry. Section **??** gives a general overview of the purposes of analyzing insurance data which is reinforced in the Section **??** case study. Naturally, there is a huge gap between the broad goals summarized in the overview and a case study application; this gap is covered through the methods and techniques of data analysis covered in the rest of the text.

1.1 Relevance of Analytics to Insurance Activities

In this section, you learn how to:

- Summarize the importance of insurance to consumers and the economy
- Describe analytics
- Identify data generating events associated with the timeline of a typical insurance contract

---

1.1.1 Nature and Relevance of Insurance

This book introduces the process of using data to make decisions in an insurance context. It does not assume that readers are familiar with insurance but introduces insurance concepts as needed. If you are new to insurance, then it is probably easiest to think about an insurance policy that covers the contents of an apartment or house that you are renting (known as renters insuranceRenters insurance is an insurance policy that covers the contents of an apartment or house that you are renting.) or the contents and property of a building that is owned by you or a friend (known as homeowners insuranceHomeowners insurance is an insurance policy that covers the contents and property of a building that is owned by you or a friend.). Another common example is automobile insuranceAn insurance policy that covers damage to your vehicle, damage to other vehicles in the accident, as well as medical expenses of those injured in the accident.. In the event of an accident, this policy may cover damage to your vehicle, damage to other vehicles in the accident, as well as medical expenses of those injured in the accident.

One way to think about the nature of insurance is who buys it. Renters, homeowners, and auto insurance are examples of personal insuranceinsurance purchased by a person in that these are policies issued to people. Businesses also buy insurance, such as coverage on their properties, and this is known as commercial insuranceinsurance purchased by an insurer. The seller, an insurance company, is also known as an insurer. Even insurance companies need insurance; this is known as reinsurancea transaction where the primary insurer buys insurance from a re-insurer who will cover part of the losses and/or loss adjustment expenses of the primary insurer.

Another way to think about the nature of insurance is the type of risk being covered. In the U.S., policies such as renters and homeowners are known as property insuranceProperty insurance is a policy that protects the insured against loss or damage to real or personal property. The cause of loss might be fire, lightening, business interruption, loss of rents, glass breakage, tornado, windstorm, hail, water damage, explosion, riot, civil commotion, rain, or damage from aircraft or vehicles. whereas a policy such as auto that covers medical damages to people is known as casualty insuranceCausalty insurance is a form of liability insurance providing coverage for negligent acts and omissions. Examples include workers compensation, errors and omissions, fidelity, crime, glass, boiler, and various malpractice coverages.. In the rest of the world, these are both known as non-life or general insurance, to distinguish them from life insuranceLife insurance is a contract where the insurer promises to pay upon the death of an insured person. The person being paid is the beneficiary. .

Both life and non-life insurances are important components of the world economy. The **?** estimates that direct insurance premiums in the world for 2014 was 2,654,549 for life and 2,123,699 for non-life; these figures are in millions of U.S. dollars. As noted earlier, the total represents 6.2% of the world GDP. Put

another way, life accounts for 55.5% of insurance premiums and 3.4% of world GDP whereas non-life accounts for 44.5% of insurance premiums and 2.8% of world GDP. Both life and non-life represent important economic activities.

Insurance may not be as entertaining as the sports industry (another industry that depends heavily on data) but it does affect the financial livelihoods of many. By almost any measure, insurance is a major economic activity. On a global level, insurance premiums comprised about 6.2% of the world gross domestic product (GDP) in 2014, (**?**). As examples, premiums accounted for 18.9% of GDP in Taiwan (the highest in the study) and represented 7.3% of GDP in the United States. On a personal level, almost everyone owning a home has insurance to protect themselves in the event of a fire, hailstorm, or some other calamitous event. Almost every country requires insurance for those driving a car. In sum, although not particularly entertaining, insurance plays an important role in the economies of nations and the lives of individuals.

1.1.2 What is Analytics?

Insurance is a data-driven industry. Like all major corporations and organizations, insurers use data when trying to decide how much to pay employees, how many employees to retain, how to market their services and products, how to forecast financial trends, and so on. These represent general areas of activities that are not specific to the insurance industry. Although each industry has its own data nuances and needs, the collection, analysis and use of data is an activity shared by all, from the internet giants to the small business, by public and governmental organizations, and is not specific to the insurance industry. You will find that the data collection and analysis methods and tools introduced in this text are relevant for all.

In any data-driven industry, analytics is a key to deriving and extracting information from data. But what is analytics? Making data-driven business decisions has been described as business analytics, business intelligence, and data science. These terms, among others, are sometimes used interchangeably and sometimes refer to distinct applications. Business intelligence may focus on processes of collecting data, often through databases and data warehouses, whereas business analytics utilizes tools and methods for statistical analyses of data. In contrast to these two terms that emphasize business applications, the term data science can encompass broader data related applications in many scientific domains. For our purposes, we use the term analyticsAnalytics is the process of using data to make decisions. to refer to the process of using data to make decisions. This process involves gathering data, understanding concepts and models of uncertainty, making general inferences, and communicating results.

When introducing data methods in this text, we will focus on losses that arise from, or related to, obligations in insurance contracts. This could be the amount of damage to one's apartment under a renter's insurance agreement, the amount

needed to compensate someone that you hurt in a driving accident, and the like. We call these obligations insurance claimAn insurance claim is the compensation provided by the insurer for incurred hurt, loss, or damage that is covered by the policy.. With this focus, we will be able to introduce and directly use generally applicable statistical tools and techniques.

1.1.3 Insurance Processes

Yet another way to think about the nature of insurance is by the duration of an insurance contract, known as the termThe duration of an insurance contract. This text will focus on short-term insurance contracts. By short-term, we mean contracts where the insurance coverage is typically provided for a year or six months. Most commercial and personal contracts are for a year so that will be our default duration. An important exception is U.S. auto policies that are often six months in length.

In contrast, we typically think of life insurance as a long-term contract where the default is to have a multi-year contract. For example, if a person 25 years old purchases a whole life policy that pays upon death of the insured and that person does not die until age 100, then the contract is in force for 75 years.

There are other important differences between life and non-life products. In life insurance, the benefit amount is often stipulated in the contract provisions. In contrast, most non-life contracts provide for compensation of insured losses which are unknown before the accident. (There are usually limits placed on the compensation amounts.) In a life insurance contract that stretches over many years, the time value of money plays a prominent role. In a non-life contract, the random amount of compensation takes priority.

In both life and non-life insurances, the frequency of claims is very important. For many life insurance contracts, the insured event (such as death) happens only once. In contrast, for non-life insurances such as automobile, it is common for individuals (especially young male drivers) to get into more than one accident during a year. So, our models need to reflect this observation; we will introduce different frequency models that you may also see when studying life insurance.

For short-term insurance, the framework of the probabilistic model is straightforward. We think of a one-period model (the period length, e.g., one year, will be specified in the situation).

- At the beginning of the period, the insured pays the insurer a known premium that is agreed upon by both parties to the contract.
- At the end of the period, the insurer reimburses the insured for a (possibly multivariate) random loss.

This framework will be developed as we proceed; but we first focus on integrating this framework with concerns about how the data may arise. From an

insurer's viewpoint, contracts may be only for a year but they tend to be renewed. Moreover, payments arising from claims during the year may extend well beyond a single year. One way to describe the data arising from operations of an insurance company is to use a timeline granular approach. A **process** approach provides an overall view of the events occurring during the life of an insurance contract, and their nature – random or planned, loss events (claims) and contract changes events, and so forth. In this micro oriented view, we can think about what happens to a contract at various stages of its existence.

Figure **??** traces a timeline of a typical insurance contract. Throughout the life of the contract, the company regularly processes events such as premium collection and valuation, described in Section **??**; these are marked with an **x** on the timeline. Non-regular and unanticipated events also occur. To illustrate, t_2 and t_4 mark the event of an insurance claim (some contracts, such as life insurance, can have only a single claim). Times t_3 and t_5 mark events when a policyholder wishes to alter certain contract features, such as the choice of a deductible or the amount of coverage. From a company perspective, one can even think about the contract initiation (arrival, time t_1) and contract termination (departure, time t_6) as uncertain events. (Alternatively, for some purposes, you may condition on these events and treat them as certain.)

Show Quiz Solution

1.2 Insurance Company Operations

In this section, you learn how to:

- Describe five major operational areas of insurance companies.
- Identify the role of data and analytics opportunities within each operational area.

Armed with insurance data, the end goal is to use data to make decisions. We will learn more about methods of analyzing and extrapolating data in future chapters. To begin, let us think about why we want to do the analysis. We will take the insurance company's viewpoint (not the insured person) and introduce ways of bringing money in, paying it out, managing costs, and making sure that we have enough money to meet obligations. The emphasis is on insurance-specific operations rather than on general business activities such as advertising, marketing, and human resources management.

Specifically, in many insurance companies, it is customary to aggregate detailed insurance processes into larger operational units; many companies use these

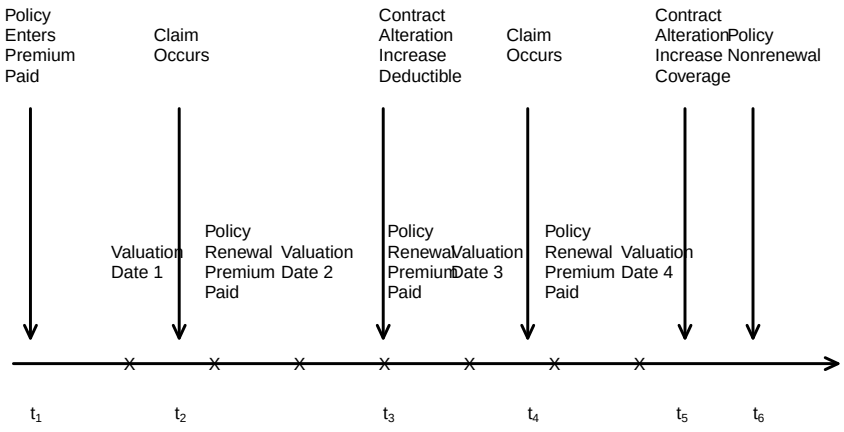


Figure 1.1: Timeline of a Typical Insurance Policy. Arrows mark the occurrences of random events. Each x marks the time of scheduled events that are typically non-random.

functional areas to segregate employee activities and areas of responsibilities. Actuaries, other financial analysts, and insurance regulators work within these units and use data for the following activities:

1. **Initiating Insurance**. At this stage, the company makes a decision as to whether or not to take on a risk (the underwritingUnderwriting is the process where the company makes a decision as to whether or not to take on a risk. stage) and assign an appropriate premium (or rate). Insurance analytics has its actuarial roots in ratemaking, where analysts seek to determine the right price for the right risk.

2. **Renewing Insurance**. Many contracts, particularly in general insurance, have relatively short durations such as 6 months or a year. Although there is an implicit expectation that such contracts will be renewed, the insurer has the opportunity to decline coverage and to adjust the premium. Analytics is also used at this policy renewal stage where the goal is to retain profitable customers.

3. **Claims Management**. Analytics has long been used in (1) detecting and preventing claims fraud, (2) managing claim costs, including identifying the appropriate support for claims handling expenses, as well as (3) understanding excess layers for reinsurance and retention.

4. **Loss Reserving**. Analytic tools are used to provide management with an appropriate estimate of future obligations and to quantify the uncertainty of those estimates.

5. **Solvency and Capital Allocation**. Deciding on the requisite amount of capital and on ways of allocating capital among alternative investments are also important analytics activities. Companies must understand how much capital is needed so that they will have sufficient flow of cash available to meet their obligations at the times they are expected to materialize (solvency). This is an important question that concerns not only company managers but also customers, company shareholders, regulatory authorities, as well as the public at large. Related to issues of how much capital is the question of how to allocate capital to differing financial projects, typically to maximize an investor's return. Although this question can arise at several levels, insurance companies are typically concerned with how to allocate capital to different lines of business within a firm and to different subsidiaries of a parent firm.

Although data represent a critical component of solvency and capital allocation, other components including the local and global economic framework, the financial investments environment, and quite specific requirements according to the regulatory environment of the day, are also important. Because of the background needed to address these components, we will not address solvency, capital allocation, and regulation issues in this text.

Nonetheless, for all operating functions, we emphasize that analytics in the

insurance industry is not an exercise that a small group of analysts can do by themselves. It requires an insurer to make significant investments in their information technology, marketing, underwriting, and actuarial functions. As these areas represent the primary end goals of the analysis of data, additional background on each operational unit is provided in the following subsections.

1.2.1 Initiating Insurance

Setting the price of an insurance product can be a perplexing problem. This is in contrast to other industries such as manufacturing where the cost of a product is (relatively) known and provides a benchmark for assessing a market demand price. Similarly, in other areas of financial services, market prices are available and provide the basis for a market-consistent pricing structure of products. However, for many lines of insurance, the cost of a product is uncertain and market prices are unavailable. Expectations of the random cost is a reasonable place to start for a price. (If you have studied finance, then you will recall that an expectation is the optimal price for a risk-neutral insurer.) It has been traditional in insurance pricing to begin with the expected cost. Insurers then add margins to this, to account for the product's riskiness, expenses incurred in servicing the product, and an allowance for profit/surplus of the company.

Use of expected costs as a foundation for pricing is prevalent in some lines of the insurance business. These include automobile and homeowners insurance. For these lines, analytics has served to sharpen the market by making the calculation of the product's expected cost more precise. The increasing availability of the internet to consumers has also promoted transparency in pricing; in today's marketplace, consumers have ready access to competing quotes from a host of insurers. Insurers seek to increase their market share by refining their risk classificationRisk classification is the process of grouping policyholders into categories, or classes, where each insured in the class has a risk profile that is similar to others in the class. systems, thus achieving a better approximation of the products' prices and enabling cream-skimming underwriting strategies ("cream-skimming" is a phrase used when the insurer underwrites only the best risks). Recent surveys (e.g., **?**) indicate that pricing is the most common use of analytics among insurers.

Underwriting, the process of classifying risks into homogeneous categories and assigning policyholders to these categories, lies at the core of ratemaking. Policyholders within a class (category) have similar risk profiles and so are charged the same insurance price. This is the concept of an actuarially fair premium; it is fair to charge different rates to policyholders only if they can be separated by identifiable risk factors. An early article, Two Studies in Automobile Insurance Ratemaking (**?**), provided a catalyst to the acceptance of analytic methods in the insurance industry. This paper addresses the problem of classification ratemaking. It describes an example of automobile insurance that has five use classes cross-classified with four merit rating classes. At that time, the

contribution to premiums for use and merit rating classes were determined independently of each other. Thinking about the interacting effects of different classification variables is a more difficult problem.

1.2.2 Renewing Insurance

Insurance is a type of financial service and, like many service contracts, insurance coverage is often agreed upon for a limited time period at which time coverage commitments are complete. Particularly for general insurance, the need for coverage continues and so efforts are made to issue a new contract providing similar coverage, when the existing contract comes to the end of its term. This is called policy renewal. Renewal issues can also arise in life insurance, e.g., term (temporary) life insurance. At the same time other contracts, such as life annuities, terminate upon the insured's death and so issues of renewability are irrelevant.

In the absence of legal restrictions, at renewal the insurer has the opportunity to:

- accept or decline to underwrite the risk; and
- determine a new premium, possibly in conjunction with a new classification of the risk.

Risk classification and rating at renewal is based on two types of information. First, at the initial stage, the insurer has available many rating variables upon which decisions can be made. Many variables will not change, e.g., sex, whereas others are likely to have changed, e.g., age, and still others may or may not change, e.g., credit score. Second, unlike the initial stage, at renewal the insurer has available a history of policyholder's loss experience, and this history can provide insights into the policyholder that are not available from rating variables. Modifying premiums with claims history is known as experience rating, also sometimes referred to as merit rating.

Experience rating methods are either applied retrospectively or prospectively. With retrospective methods, a refund of a portion of the premium is provided to the policyholder in the event of favorable (to the insurer) experience. Retrospective premiumsThe process of determining the cost of an insurance policy based on the actual loss experience determined as an adjustment to the initial premium payment. are common in life insurance arrangements (where policyholders earn dividends in the U.S., bonuses in the U.K., and profit sharing in Israeli term life coverage). In general insurance, prospective methods are more common, where favorable insured experience is rewarded through a lower renewal premium.

Claims history can provide information about a policyholder's risk appetite. For example, in personal lines it is common to use a variable to indicate whether or not a claim has occurred in the last three years. As another example, in

a commercial line such as worker's compensation, one may look to a policyholder's average claim frequency or severity over the last three years. Claims history can reveal information that is otherwise hidden (to the insurer) about the policyholder.

1.2.3 Claims and Product Management

In some of areas of insurance, the process of paying claims for insured events is relatively straightforward. For example, in life insurance, a simple death certificate is all that is needed to pay the benefit amount as provided in the contract. However, in non-life areas such as property and casualty insurance, the process can be much more complex. Think about even a relatively simple insured event such as automobile accident. Here, it is often required to determine which party is at fault, one needs to assess damage to all of the vehicles and people involved in the incident, both insured and non-insured, the expenses incurred in assessing the damages must be assessed, and so forth. The process of determining coverage, legal liability, and settling claims is known as claims adjustmentClaims adjustment is the process of determining coverage, legal liability, and settling claims..

Insurance managers sometimes use the phrase claims leakageClaims leakage respresents money lost through claims management inefficiencies. to mean dollars lost through claims management inefficiencies. There are many ways in which analytics can help manage the claims process, c.f., **?**. Historically, the most important has been fraud detection. The claim adjusting process involves reducing information asymmetry (the claimant knows what happened; the company knows some of what happened). Mitigating fraud is an important part of the claims management process.

Fraud detection is only one aspect of managing claims. More broadly, one can think about claims management as consisting of the following components:

- **Claims triaging**. Just as in the medical world, early identification and appropriate handling of high cost claims (patients, in the medical world), can lead to dramatic savings. For example, in workers compensation, insurers look to achieve early identification of those claims that run the risk of high medical costs and a long payout period. Early intervention into these cases could give insurers more control over the handling of the claim, the medical treatment, and the overall costs with an earlier return-to-work.
- **Claims processing**. The goal is to use analytics to identify routine situations that are anticipated to have small payouts. More complex situations may require more experienced adjusters and legal assistance to appropriately handle claims with high potential payouts.
- **Adjustment decisions**. Once a complex claim has been identified and

assigned to an adjusterAn adjuster is a person who investigates claims and recommends settlement options based on estimates of damage and insurance policies held., analytic driven routines can be established to aid subsequent decision-making
processes. Such processes can also be helpful for adjusters in developing case reserves, an estimate of the insurer's future liability. This is an important input to the insurer's loss reserves, described in Section **??**.

In addition to the insured's reimbursement for losses, the insurer also needs to be concerned with another source of revenue outflow, expenses. loss adjustment expensesLoss adjustment expenses are costs to the insurer that are directly attributable to settling a claims. For example, the cost of an adjuster is someone who assess the claim cost or a lawyer who becomes involve in settling an insurer's legal obligation on a claim are part of an insurer's cost of managing claims. Analytics can be used to reduce expenses directly related to claims handling (allocatedAllocated loss adjustment expenses, sometimes known by the acronym ALEA, are costs that can be directly attributed to settling a claim; for example, the cost of an adjuster) as well as general staff time for overseeing the claims processes (unallocatedUnallocated loss adjustment expenses are costs that can only be indirectly attributed to claim settlement; for example, the cost of an office to support claims staff). The insurance industry has high operating costs relative to other portions of the financial services sectors.

In addition to claims payments, there are many other ways in which insurers use data to manage their products. We have already discussed the need for analytics in underwriting, that is, risk classification at the initial acquisition and renewal stages. Insurers are also interested in which policyholders elect to renew their contract and, as with other products, monitor customer loyalty.

Analytics can also be used to manage the portfolio, or collection, of risks that an insurer has acquired. When the risk is initially obtained, the insurer's risk can be managed by imposing contract parameters that modify contract payouts. Chapters **??** and **??** describe common modifications including coinsuranceCoinsurance is an arrangement whereby the insured and insurer share the covered losses. Typically, a coinsurance parameter specified means that both parties receive a proportional share, e.g., 50%, of the loss.,deductiblesA deductible is a parameter specified in the contract. Typically, losses below the deductible are paid by the policyholder whereas losses in excess of the deductible are the insurer's responsibility (subject to policy limits and coninsurance). and policy upper limitsTemp.

After the contract has been agreed upon with an insured, the insurer may still modify its net obligation by entering into a reinsurance agreement. This type of agreement is with a reinsurer, an insurer of an insurer. It is common for insurance companies to purchase insurance on its portfolio of risks to gain protection from unusual events, just as people and other companies do.

1.2.4 Loss Reserving

An important feature that distinguishes insurance from other sectors of the economy is the timing of the exchange of considerations. In manufacturing, payments for goods are typically made at the time of a transaction. In contrast, for insurance, money received from a customer occurs in advance of benefits or services; these are rendered at a later date when the insured event occurs. This leads to the need to hold a reservoir of wealth to meet future obligations in respect to obligations made, and to gain the trust of the insureds that the company will be able to fulfill its commitments. The size of this reservoir of wealth, and the importance of ensuring its adequacy, is a major concern for the insurance industry.

Setting aside money for unpaid claims is known as loss reserving; in some jurisdictions, reserves are also known as technical provisions. We saw in Figure **??** several times at which a company summarizes its financial position; these times are known as valuation dates. Claims that arise prior to valuation dates have typically been paid, are in the process of being paid, or are about to be paid; claims in the future of these valuation dates are unknown. A company must estimate these outstanding liabilities when determining its financial strength. Accurately determining loss reserves is important to insurers for many reasons.

1. Loss reserves represent an anticipated claim that the insurer owes its customers. Under-reserving may result in a failure to meet claim liabilities. Conversely, an insurer with excessive reserves may present a weaker financial position than it truly has.

2. Reserves provide an estimate for the unpaid cost of insurance that can be used for pricing contracts.

3. Loss reserving is required by laws and regulations. The public has a strong interest in the financial strength and solvency of insurers.

4. In addition to insurance company management and regulators, other stakeholders such as investors and customers make decisions that depend on company loss reserves.

Loss reserving is a topic where there are substantive differences between life and general (also known as property and casualty, or non-life), insurance. In life insurance, the severity (amount of loss) is often not a source of uncertainty as payouts are specified in the contract. The frequency, driven by mortality of the insured, is a concern. However, because of the length of time for settlement of life insurance contracts, the time value of money uncertainty as measured from issue to date of payment can dominate frequency concerns. For example, for an insured who purchases a life contract at age 20, it would not be unusual for the contract to still be open in 60 years time, when the insured celebrates his or her 80th birthday. See, for example, **?** or **?** for introductions to reserving

for life insurance.

Show Quiz Solution

1.3 Case Study: Wisconsin Property Fund

In this section, we use the Wisconsin Property Fund as a case study. You learn how to:

- Describe how data generating events can produce data of interest to insurance analysts.
- Produce relevant summary statistics for each variable.
- Describe how these summary statistics can be used in each of the major operational areas of an insurance company.

Let us illustrate the kind of data under consideration and the goals that we wish to achieve by examining the Local Government Property Insurance Fund (LGPIF), an insurance pool administered by the Wisconsin Office of the Insurance Commissioner. The LGPIF was established to provide property insurance for local government entities that include counties, cities, towns, villages, school districts, and library boards. The fund insures local government property such as government buildings, schools, libraries, and motor vehicles. The fund covers all property losses except those resulting from flood, earthquake, wear and tear, extremes in temperature, mold, war, nuclear reactions, and embezzlement or theft by an employee.

The fund covers over a thousand local government entities who pay approximately \$25 million in premiums each year and receive insurance coverage of about \$75 billion. State government buildings are not covered; the LGPIF is for local government entities that have separate budgetary responsibilities and who need insurance to moderate the budget effects of uncertain insurable events. Coverage for local government property has been made available by the State of Wisconsin since 1911, thus providing a wealth of historical data.

In this illustration, we restrict consideration to claims from coverage of building and contents; we do not consider claims from motor vehicles and specialized equipment owned by local entities (such as snow plowing machines). We also consider only claims that are closed, with obligations fully met.

1.3.1 Fund Claims Variables: Frequency and Severity

At a fundamental level, insurance companies accept premiums in exchange for promises to compensate a policyholder upon the occurrence of an insured event. IndemnificationIndemnification is the compensation provided by the insurer. is the compensation provided by the insurer for incurred hurt, loss, or damage that is covered by the policy. This compensation is also known as a claim. The extent of the payout, known as the severity, is a key financial expenditure for an insurer.

In terms of money outgo, an insurer is indifferent to having ten claims of 100 when compared to one claim of 1,000. Nonetheless, it is common for insurers to study how often claims arise, known as the frequency of claims. The frequency is important for expenses, but it also influences contractual parameters (such as deductibles and policy limits that are described later) that are written on a per occurrence basis, is routinely monitored by insurance regulators, and can be a key driver in the overall indemnification obligation of the insurer. We shall consider the frequency and severity as the two main claim variables that we wish to understand, model, and manage.

To illustrate, in 2010 there were 1,110 policyholders in the property fund who experienced a total of 1,377 claims. Table **??** shows the distribution. Almost two-thirds (0.637) of the policyholders did not have any claims and an additional 18.8% had only one claim. The remaining 17.5% (=1 - 0.637 - 0.188) had more than one claim; the policyholder with the highest number recorded 239 claims. The average number of claims for this sample was 1.24 (=1377/1110).

Table 1.1: 2010 Claims Frequency Distribution

	Type									
Number	0	1	2	3	4	5	6	7	8	9 or more	Sum
Count	707	209	86	40	18	12	9	4	6	19	1,110
Claims	0	209	172	120	72	60	54	28	48	617	1,377
Proportion	0.637	0.188	0.077	0.036	0.016	0.011	0.008	0.004	0.005	0.017	1.000

R Code for Frequency Table

```
Insample <- read.csv("Insample.csv", header=T,  na.strings=c("."), stringsAsFactors=FA
Insample2010 <- subset(Insample, Year==2010)
table(Insample2010$Freq)
```

For the severity distribution, a common approach is to examine the distribution of the sample of 1,377 claims. However, another common approach is to examine the distribution of the average claims of those policyholders with claims. In our 2010 sample, there were 403 (=1110-707) such policyholders. For 209 of these policyholders with one claim, the average claim equals the only claim they

experienced. For the policyholder with highest frequency, the average claim is an average over 239 separately reported claim events. This average is also known as the pure premiumPure premium is the total severity divided by the number of claims. It does not include insurance company expenses, premium taxes, contingencies, nor an allowance for profits. Also called loss costs. Some definitions include allocated loss adjustment expenses (ALAE). or loss cost.

Table **??** summarizes the sample distribution of average severities from the 403 policyholders who made a claim; it shows that the average claim amount was 56,330 (all amounts are in U.S. Dollars). However, the average gives only a limited look at the distribution. More information can be gleaned from the summary statistics which show a very large claim in the amount of 12,920,000. Figure **??** provides further information about the distribution of sample claims, showing a distribution that is dominated by this single large claim so that the histogram is not very helpful. Even when removing the large claim, you will find a distribution that is skewed to the right. A generally accepted technique is to work with claims in logarithmic units especially for graphical purposes; the corresponding figure in the right-hand panel is much easier to interpret.

Table 1.2: 2010 Average Severity Distribution

Minimum	First Quartile	Median	Mean	Third Quartile	Maximum
167	2,226	4,951	56,330	11,900	12,920,000

R Code for Severity Distribution Table and Figures

```
Insample <- read.csv("Data/PropertyFundInsample.csv", header=T, na.strings=c("."), stringsAsFacto
Insample2010 <- subset(Insample, Year==2010)
InsamplePos2010 <- subset(Insample2010, yAvg>0)
# Table
summary(InsamplePos2010$yAvg)
length(InsamplePos2010$yAvg)
# Figures
par(mfrow=c(1, 2))
hist(InsamplePos2010$yAvg, main="", xlab="Average Claims")
hist(log(InsamplePos2010$yAvg), main="", xlab="Logarithmic Average Claims")
```

1.3.2 Fund Rating Variables

Developing models to represent and manage the two outcome variables, frequency and severity, is the focus of the early chapters of this text. However, when actuaries and other financial analysts use those models, they do so in the context of external variables. In general statistical terminology, one might call

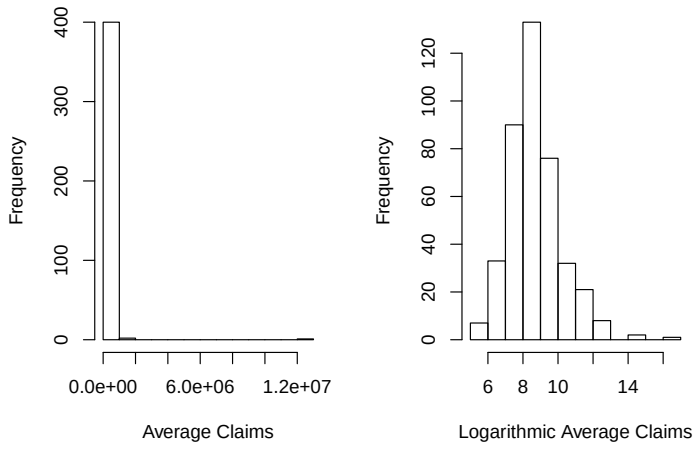


Figure 1.2: Distribution of Positive Average Severities

these explanatory or predictor variables; there are many other names in statistics, economics, psychology, and other disciplines. Because of our insurance focus, we call them rating variables as they will be useful in setting insurance rates and premiums.

We earlier considered observations from a sample of 1,110 policyholders which may seem like a lot. However, as we will see in our forthcoming applications, because of the preponderance of zeros and the skewed nature of claims, actuaries typically yearn for more data. One common approach that we adopt here is to examine outcomes from multiple years, thus increasing the sample size. We will discuss the strengths and limitations of this strategy later but, at this juncture, we just wish to show the reader how it works.

Specifically, Table **??** shows that we now consider policies over five years of data, 2006, ..., 2010, inclusive. The data begins in 2006 because there was a shift in claim coding in 2005 so that comparisons with earlier years are not helpful. To mitigate the effect of open claims, we consider policy years prior to 2011. An open claim means that not all of the obligations for the claim are known at the time of the analysis; for some claims, such an injury to a person in an auto accident or in the workplace, it can take years before costs are fully known.

Year	Average Frequency	Average Severity	Average Coverage	Number of Policyholders

Table 1.3: Claims Summary by Policyholder

Year	Average Frequency	Average Severity	Average Coverage	Number of Policyholders
2006	0.951	9,695	32,498,186	1,154
2007	1.167	6,544	35,275,949	1,138
2008	0.974	5,311	37,267,485	1,125
2009	1.219	4,572	40,355,382	1,112
2010	1.241	20,452	41,242,070	1,110

R Code for Claims Summary by Policyholder

```
Insample <- read.csv("Data/PropertyFundInsample.csv", header=T, na.strings=c("."), stringsAsFacto
library(doBy)
T1A <- summaryBy(Freq ~ Year, data = Insample,
   FUN = function(x) { c(m = mean(x), num=length(x)) } )
T1B <- summaryBy(yAvg    ~ Year, data = Insample,
   FUN = function(x) { c(m = mean(x), num=length(x)) } )
T1C <- summaryBy(BCcov    ~ Year, data = Insample,
   FUN = function(x) { c(m = mean(x), num=length(x)) } )
Table1In <- cbind(T1A[1],T1A[2],T1B[2],T1C[2],T1A[3])
names(Table1In) <- c("Year", "Average Frequency","Average Severity", "Average","Number of Policyh
Table1In
```

Table **??** shows that the average claim varies over time, especially with the high 2010 value (that we saw was due to a single large claim)[1]. The total number of policyholders is steadily declining and, conversely, the coverage is steadily increasing. The coverage variable is the amount of coverage of the property and contents. Roughly, you can think of it as the maximum possible payout of the insurer. For our immediate purposes, the coverage is our first rating variable. Other things being equal, we would expect that policyholders with larger coverage will have larger claims. We will make this vague idea much more precise as we proceed, and also justify this expectation with data.

For a different look at the 2006-2010 data, Table **??** summarizes the distribution of our two outcomes, frequency and claims amount. In each case, the average exceeds the median, suggesting that the two distributions are right-skewed. In addition, the table summarizes our continuous rating variables, coverage and

[1]Note that the average severity in Table **??** differs from that reported in Table **??**. This is because the former includes policyholders with zero claims where as the latter does not. This is an important distinction that we will address in later portions of the text.

deductible amount. The table also suggests that these variables also have right-skewed distributions.

Table 1.4: Summary of Claim Frequency and Severity, Deductibles, and Coverages

	Minimum	Median	Average	Maximum
Claim Frequency	0	0	1.109	263
Claim Severity	0	0	9,292	12,922,218
Deductible	500	1,000	3,365	100,000
Coverage (000's)	8.937	11,354	37,281	2,444,797

R Code for Summary of Claim Frequency and Severity, Deductibles, and Coverages

```
Insample <- read.csv("Data/PropertyFundInsample.csv", header=T, na.strings=c("."), stri
t1<- summaryBy(Insample$Freq ~ 1, data = Insample,
   FUN = function(x) { c(ma=min(x), m1=median(x),m=mean(x),mb=max(x)) } )
names(t1) <- c("Minimum", "Median","Average", "Maximum")
t2 <- summaryBy(Insample$yAvg ~ 1, data = Insample,
   FUN = function(x) { c(ma=min(x), m1=median(x), m=mean(x),mb=max(x)) } )
names(t2) <- c("Minimum", "Median","Average", "Maximum")
t3 <- summaryBy(Deduct ~ 1, data = Insample,
   FUN = function(x) { c(ma=min(x), m1=median(x), m=mean(x),mb=max(x)) } )
names(t3) <- c("Minimum", "Median","Average", "Maximum")
t4 <- summaryBy(BCcov/1000 ~ 1, data = Insample,
   FUN = function(x) { c(ma=min(x), m1=median(x), m=mean(x),mb=max(x)) } )
names(t4) <- c("Minimum", "Median","Average", "Maximum")
Table2 <- rbind(t1,t2,t3,t4)
Table2a <- round(Table2,3)
Rowlable <- rbind("Claim Frequency","Claim Severity","Deductible","Coverage (000's)")
Table2aa <- cbind(Rowlable,as.matrix(Table2a))
Table2aa
```

The following display describes the rating variables considered in this chapter. Hopefully, these are variables that you think might naturally be related to claims outcomes. You can learn more about them in **?**. To handle the skewness, we henceforth focus on logarithmic transformations of coverage and deductibles.

Description of Rating Variables

Variable	*Description*
EntityType	Categorical variable that is one of six types: (Village, City, County, Misc, School, or Town)
LnCoverage	Total building and content coverage, in logarithmic millions of dollars
LnDeduct	Deductible, in logarithmic dollars
AlarmCredit	Categorical variable that is one of four types: (0, 5, 10, or 15) for automatic smoke alarms in main rooms
NoClaimCredit	Binary variable to indicate no claims in the past two years
Fire5	Binary variable to indicate the fire class is below 5 (The range of fire class is 0 to 10)

To get a sense of the relationship between the non-continuous rating variables and claims, Table **??** relates the claims outcomes to these categorical variables. Table **??** suggests substantial variation in the claim frequency and average severity of the claims by entity type. It also demonstrates higher frequency and severity for the `Fire5` variable and the reverse for the `NoClaimCredit` variable. The relationship for the `Fire5` variable is counter-intuitive in that one would expect lower claim amounts for those policyholders in areas with better public protection (when the protection code is five or less). Naturally, there are other variables that influence this relationship. We will see that these background variables are accounted for in the subsequent multivariate regression analysis, which yields an intuitive, appealing (negative) sign for the `Fire5` variable.

Table 1.5: Claims Summary by Entity Type, Fire Class, and No Claim Credit

Variable	Number of Policies	Claim Frequency	Average Severity
EntityType			
Village	1,341	0.452	10,645
City	793	1.941	16,924
County	328	4.899	15,453
Misc	609	0.186	43,036
School	1,597	1.434	64,346
Town	971	0.103	19,831
Fire			
Fire5=0	2,508	0.502	13,935
Fire5=1	3,131	1.596	41,421
No Claims Credit			
NoClaimCredit=0	3,786	1.501	31,365
NoClaimCredit=1	1,853	0.310	30,499
Total	5,639	1.109	31,206

R Code for Claims Summary by Entity Type, Fire Class, and No Claim Credit

```
ByVarSumm<-function(datasub){
  tempA <- summaryBy(Freq    ~ 1 , data = datasub,
     FUN = function(x) { c(m = mean(x), num=length(x)) } )
  datasub1 <-  subset(datasub, yAvg>0)
  tempB <- summaryBy(yAvg   ~ 1, data = datasub1,FUN = function(x) { c(m = mean(x)) } )
  tempC <- merge(tempA,tempB,all.x=T)[c(2,1,3)]
  tempC1 <- as.matrix(tempC)
  return(tempC1)
  }
datasub <-  subset(Insample, TypeVillage == 1);
t1 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeCity == 1);
t2 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeCounty == 1);
t3 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeMisc == 1);
t4 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeSchool == 1);
t5 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeTown == 1);
t6 <- ByVarSumm(datasub)
datasub <-  subset(Insample, Fire5 == 0);
t7 <- ByVarSumm(datasub)
datasub <-  subset(Insample, Fire5 == 1);
t8 <- ByVarSumm(datasub)
datasub <-  subset(Insample, Insample$NoClaimCredit == 0);
t9 <- ByVarSumm(datasub)
datasub <-  subset(Insample, Insample$NoClaimCredit == 1);
t10 <- ByVarSumm(datasub)
t11 <- ByVarSumm(Insample)

Tablea <- rbind(t1,t2,t3,t4,t5,t6,t7,t8,t9,t10,t11)
Tableaa <- round(Tablea,3)
Rowlable <- rbind("Village","City","County","Misc","School",
          "Town","Fire5--No","Fire5--Yes","NoClaimCredit--No",
        "NoClaimCredit--Yes","Total")
Table4 <- cbind(Rowlable,as.matrix(Tableaa))
Table4
```

Table **??** shows the claims experience by alarm credit. It underscores the difficulty of examining variables individually. For example, when looking at the experience for all entities, we see that policyholders with no alarm credit have on average lower frequency and severity than policyholders with the highest (15%, with 24/7 monitoring by a fire station or security company) alarm credit.

In particular, when we look at the entity type School, the frequency is 0.422 and the severity 25,523 for no alarm credit, whereas for the highest alarm level it is 2.008 and 85,140. This may simply imply that entities with more claims are the ones that are likely to have an alarm system. Summary tables do not examine multivariate effects; for example, Table **??** ignores the effect of size (as we measure through coverage amounts) that affect claims.

Table 1.6: Claims Summary by Entity Type and Alarm Credit (AC) Category

Entity Type	AC0 Claim Frequency	AC0 Avg. Severity	AC0 Num. Policies	AC5 Claim Frequency	AC5 Avg. Severity	AC5 Num. Policies
Village	0.326	11,078	829	0.278	8,086	54
City	0.893	7,576	244	2.077	4,150	13
County	2.140	16,013	50	-	-	1
Misc	0.117	15,122	386	0.278	13,064	18
School	0.422	25,523	294	0.410	14,575	122
Town	0.083	25,257	808	0.194	3,937	31
Total	0.318	15,118	2,611	0.431	10,762	239

Table 1.7: Claims Summary by Entity Type and Alarm Credit (AC) Category

Entity Type	AC10 Claim Frequency	AC10 Avg. Severity	AC10 Num. Policies	AC15 Claim Frequency	AC15 Avg. Severity	AC15 Num. Policies
Village	0.500	8,792	50	0.725	10,544	408
City	1.258	8,625	31	2.485	20,470	505
County	2.125	11,688	8	5.513	15,476	269
Misc	0.077	3,923	26	0.341	87,021	179
School	0.488	11,597	168	2.008	85,140	1,013
Town	0.091	2,338	44	0.261	9,490	88
Total	0.517	10,194	327	2.093	41,458	2,462

R Code for Claims Summary by Entity Type and Alarm Credit Category

```
#Claims Summary by Entity Type and Alarm Credit
ByVarSumm<-function(datasub){
  tempA <- summaryBy(Freq    ~ AC00 , data = datasub,
                     FUN = function(x) { c(m = mean(x), num=length(x)) } )
  datasub1 <-  subset(datasub, yAvg>0)
  if(nrow(datasub1)==0) { n<-nrow(datasub)
```

```
    return(c(0,0,n))
  } else
  {
    tempB <- summaryBy(yAvg   ~ AC00, data = datasub1,
                       FUN = function(x) { c(m = mean(x)) } )
    tempC <- merge(tempA,tempB,all.x=T)[c(2,4,3)]
    tempC1 <- as.matrix(tempC)
    return(tempC1)
  }
}
AlarmC <- 1*(Insample$AC00==1) + 2*(Insample$AC05==1)+ 3*(Insample$AC10==1)+ 4*(Insampl
ByVarCredit<-function(ACnum){
datasub <-  subset(Insample, TypeVillage == 1 & AlarmC == ACnum);
  t1 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeCity == 1 & AlarmC == ACnum);
  t2 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeCounty == 1 & AlarmC == ACnum);
  t3 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeMisc == 1 & AlarmC == ACnum);
  t4 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeSchool == 1 & AlarmC == ACnum);
  t5 <- ByVarSumm(datasub)
datasub <-  subset(Insample, TypeTown == 1 & AlarmC ==ACnum);
  t6 <- ByVarSumm(datasub)
datasub <-  subset(Insample, AlarmC == ACnum);
  t7 <- ByVarSumm(datasub)
Tablea <- rbind(t1,t2,t3,t4,t5,t6,t7)
Tableaa <- round(Tablea,3)
Rowlable <- rbind("Village","City","County","Misc","School",
                  "Town","Total")
Table4 <- cbind(Rowlable,as.matrix(Tableaa))
}
Table4a <- ByVarCredit(1)    #Claims Summary by Entity Type and Alarm Credit==00
Table4b <- ByVarCredit(2)    #Claims Summary by Entity Type and Alarm Credit==05
Table4c <- ByVarCredit(3)    #Claims Summary by Entity Type and Alarm Credit==10
Table4d <- ByVarCredit(4)    #Claims Summary by Entity Type and Alarm Credit==15
```

1.3.3 Fund Operations

We have now seen the Fund's two outcome variables: a count variable for the number of claims, and a continuous variable for the claims amount. We have also introduced a continuous rating variable (coverage); a discrete quantitative variable (logarithmic deductibles); two binary rating variables (no claims credit and fire class); and two categorical rating variables (entity type and alarm credit).

Subsequent chapters will explain how to analyze and model the distribution of these variables and their relationships. Before getting into these technical details, let us first think about where we want to go. General insurance company functional areas are described in Section **??**; let us now think about how these areas might apply in the context of the property fund.

Initiating Insurance

Because this is a government sponsored fund, we do not have to worry about selecting good or avoiding poor risks; the fund is not allowed to deny a coverage application from a qualified local government entity. If we do not have to underwrite, what about how much to charge?

We might look at the most recent experience in 2010, where the total fund claims were approximately 28.16 million USD (= 1377 claims $\times$ 20452 average severity). Dividing that among 1,110 policyholders, that suggests a rate of 24,370 ($\approx$ 28,160,000/1110). However, 2010 was a bad year; using the same method, our premium would be much lower based on 2009 data. This swing in premiums would defeat the primary purpose of the fund, to allow for a steady charge that local property managers could utilize in their budgets.

Having a single price for all policyholders is nice but hardly seems fair. For example, Table **??** suggests that Schools have much higher claims than other entities and so should pay more. However, simply doing the calculation on an entity by entity basis is not right either. For example, we saw in Table **??** that had we used this strategy, entities with a 15% alarm credit (for good behavior, having top alarm systems) would actually wind up paying more.

So, we have the data for thinking about the appropriate rates to charge but will need to dig deeper into the analysis. We will explore this topic further in Chapter **??** on premium calculation fundamentals. Selecting appropriate risks is introduced in Chapter **??** on risk classification.

Renewing Insurance

Although property insurance is typically a one-year contract, Table **??** suggests that policyholders tend to renew; this is typical of general insurance. For renewing policyholders, in addition to their rating variables we have their claims history and this claims history can be a good predictor of future claims. For example, Table **??** shows that policyholders without a claim in the last two years had much lower claim frequencies than those with at least one accident (0.310 compared to 1.501); a lower predicted frequency typically results in a lower premium. This is why it is common for insurers to use variables such as `NoClaimCredit` in their rating. We will explore this topic further in Chapter **??** on experience rating.

Claims Management

Of course, the main story line of the 2010 experience was the large claim of over 12 million USD, nearly half the claims for that year. Are there ways that this could have been prevented or mitigated? Are their ways for the fund to purchase protection against such large unusual events? Another unusual feature of the 2010 experience noted earlier was the very large frequency of claims (239) for one policyholder. Given that there were only 1,377 claims that year, this means that a single policyholder had 17.4 % of the claims. This also suggests opportunities for managing claims, the subject of Chapter **??**.

Loss Reserving

In our case study, we look only at the one year outcomes of closed claims (the opposite of open). However, like many lines of insurance, obligations from insured events to buildings such as fire, hail, and the like, are not known immediately and may develop over time. Other lines of business, including those where there are injuries to people, take much longer to develop. Chapter **??** introduces this concern and loss reserving, the discipline of determining how much the insurance company should retain to meet its obligations.

Show Quiz Solution

1.4 Further Resources and Contributors

Contributor

- **Edward W. (Jed) Frees**, University of Wisconsin-Madison, is the principal author of the initial version of this chapter. Email: jfrees@bus.wisc.edu for chapter comments and suggested improvements.
- Chapter reviewers include: Yair Babad, Chunsheng Ban, Aaron Bruhn, Gordon Enderle, Hirokazu (Iwahiro) Iwasawa, Bell Ouelega.

This book introduces loss data analytic tools that are most relevant to actuaries and other financial risk analysts. We have also introduced you to many new insurance terms; more terms can be found at the **?**. Here are a few reference cited in the chapter.

Chapter 2

Frequency Modeling

Chapter Preview. A primary focus for insurers is estimating the magnitude of aggregate claims it must bear under its insurance contracts. Aggregate claimsThe sum of all claims observed in a period of time are affected by both the frequency of insured events and the severity of the insured event. Decomposing aggregate claims into these two components, each of which warrant significant attention, is essential for analysis and pricing. This chapter discusses frequency distributions, summary measures, and parameter estimation techniques.

In Section **??**, we present terminology and discuss reasons why we study frequency and severity separately. The foundations of frequency distributions and measures are presented in Section **??** along with three principal distributions: the binomial, the Poisson, and the negative binomial. These three distributions are members of what is known as the (a,b,0) class of distributions, a distinguishing, identifying feature which allows for efficient calculation of probabilities, further discussed in Section **??**. When fitting a dataset with a distribution, parameter values need to be estimated and in Section **??**, the procedure for maximum likelihood estimation is explained. For insurance datasets, the observation at 0, denoting the occurrence of zero of a particular event, is notable and may deserve additional attention. By nature of the dataset, and explained further in Section **??**, it may be impossible to have zero of the studied event, or the probability at zero could be of such a magnitude that direct fitting would lead to improper estimates. Zero truncation or modification techniques allow for more appropriate distribution fit. Noting that our insurance portfolio could consist of different sub-groups, each with its own set of individual characteristics, Section **??** introduces mixture distributions and methodology to allow for this heterogeneity within a portfolio. Section **??** describes Goodness of Fit which measures the reasonableness of the parameter estimates. Exercises are presented in Section **??** and Section **??** concludes the chapter with `R` Code for plots depicted in Section **??**.

2.1 Frequency Distributions

In this section, you learn how to summarize the importance of frequency modeling in terms of

- contractual,
- behavioral,
- database and
- regulatory/administrative motivations.

2.1.1 How Frequency Augments Severity Information

Basic Terminology

In this chapter, **loss**, also referred to as ground-up loss, denotes the amount suffered by the insured. We use **claim** to denote the indemnification upon the occurrence of an insured event, thus the amount paid by the insurer. While some texts use **loss** and **claim** interchangeably, we wish to make a distinction here to recognize how insurance contractual provisions, such as deductibles and limits, affect the size of the claim stemming from a loss. FrequencyCount random variables that represent the number of claims represents how often an insured event occurs, typically within a policy contract. Here, we focus on count random variables that represent the number of claims, that is, how frequently an event occurs. SeverityThe amount, or size, of each payment for an insured event denotes the amount, or size, of each payment for an insured event. In future chapters, the aggregate model, which combines frequency models with severity models, is examined.

The Importance of Frequency

Recall from Chapter **??** that setting the price of an insurance good can be a complex problem. In manufacturing, the cost of a good is (relatively) known. In other financial service areas, market prices are available. In insurance, we can generalize the price setting as follows: start with an expected cost. Add "margins" to account for the product's riskiness, expenses incurred in servicing the product, and a profit/surplus allowance for the insurer.

That expected cost for insurance can be defined as the expected number of claims times the expected amount per claim, that is, expected frequency times severity. The focus on claim count allows the insurer to consider those factors which directly affect the occurrence of a loss, thereby potentially generating a claim. The frequency process can then be modeled.

Why Examine Frequency Information

Insurers and other stakeholders, including governmental organizations, have various motivations for gathering and maintaining frequency datasets.

- **Contractual.** In insurance contracts, it is common for particular deductibles and policy limits to be listed and invoked for each occurrence of an insured event. Correspondingly, the claim count data generated would indicate the number of claims which meet these criteria, offering a unique claim frequency measure. Extending this, models of total insured losses would need to account for deductibles and policy limits for each insured event.
- **Behaviorial.** In considering factors that influence loss frequency, the risk-taking and risk-reducing behavior of individuals and companies should be considered. Explanatory (rating) variables can have different effects on models of how often an event occurs in contrast to the size of the event.
  - In healthcare, the decision to utilize healthcare by individuals, and minimize such healthcare utilization through preventive care and wellness measures, is related primarily to his or her personal characteristics. The cost per user is determined by those personal, the medical condition, potential treatment measures, and decisions made by the healthcare provider (such as the physician) and the patient. While there is overlap in those factors and how they affect total healthcare costs, attention can be focused on those separate drivers of healthcare visit frequency and healthcare cost severity.
  - In personal lines, prior claims history is an important underwriting factor. It is common to use an indicator of whether or not the insured had a claim within a certain time period prior to the contract. Also, the number of claims incurred by the insured in previous periods has predictive power.
  - In homeowners insurance, in modeling potential loss frequency, the insurer could consider loss prevention measures that the homeowner has adopted, such as visible security systems. Separately, when modeling loss severity, the insurer would examine those factors that affect repair and replacement costs.
- **Databases**. Insurers may hold separate data files that suggest developing separate frequency and severity models. For example, a policyholder file is established when a policy is written. This file records much underwriting information about the insured(s), such as age, gender, and prior claims experience, policy information such as coverage, deductibles and limitations, as well as the insurance claims event. A separate file, known as the "claims" file, records details of the claim against the insurer, including the amount. (There may also be a "payments" file that records the timing of

the payments although we shall not deal with that here.) This recording process could then extend to insurers modeling the frequency and severity as separate processes.

- **Regulatory and Administrative.** Insurance is a highly regulated and monitored industry, given its importance in providing financial security to individuals and companies facing risk. As part of its duties, regulators routinely require the reporting of claims numbers as well as amounts. This may be due to the fact that there can be alternative definitions of an "amount," e.g., paid versus incurred, and there is less potential error when reporting claim numbers. This continual monitoring helps ensure financial stability of these insurance companies.

Show Quiz Solution

2.2 Basic Frequency Distributions

In this section, you learn how to:

- Determine quantities that summarize a distribution such as the distribution and survival function, as well as moments such as the mean and variance
- Define and compute the moment and probability generating functions
- Describe and understand relationships among three important frequency distributions, the binomial, Poisson, and negative binomial distributions

---

In this section, we will introduce the distributions that are commonly used in actuarial practice to model count data. The claim count random variable is denoted by N; by its very nature it assumes only non-negative integer values. Hence the distributions below are all discrete distributions supported on the set of non-negative integers $\{0, 1, \ldots\}$.

2.2.1 Foundations

Since N is a discrete random variable taking values in $\{0, 1, \ldots\}$, the most natural full description of its distribution is through the specification of the probabilities with which it assumes each of the non-negative integer values. This leads us to the concept of the probability mass function (pmf)a function that gives the probability that a discrete random variable is exactly equal to some value of N, denoted as $p_N(\cdot)$ and defined as follows:

$$p_N(k) = \Pr(N = k), \quad \text{for } k = 0, 1, \ldots \tag{2.1}$$

We note that there are alternate complete descriptions, or characterizations, of the distribution of N; for example, the distribution functionthe chance that the random variable is less than or equal to x, as a function of x of N denoted by $F_N(\cdot)$ and defined below is another such:

$$F_N(x) := \begin{cases} \sum_{k=0}^{\lfloor x \rfloor} \Pr(N = k), & x \geq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

In the above, $\lfloor \cdot \rfloor$ denotes the floor function; $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. We note that the survival functionthe probability that the random variable takes on a value greater than a number x of N, denoted by $S_N(\cdot)$, is defined as the ones'-complement of $F_N(\cdot)$, i.e. $S_N(\cdot) := 1 - F_N(\cdot)$. Clearly, the latter is another characterization of the distribution of N.

Often one is interested in quantifying a certain aspect of the distribution and not in its complete description. This is particularly useful when comparing distributions. A center of location of the distribution is one such aspect, and there are many different measures that are commonly used to quantify it. Of these, the meanaverage is the most popular; the mean of N, denoted by μ_N,[1] is defined as

$$\mu_N = \sum_{k=0}^{\infty} k \; p_N(k). \tag{2.3}$$

We note that μ_N is the expected value of the random variable N, i.e. $\mu_N = \mathrm{E}[N]$. This leads to a general class of measures, the momentsThe rth moment of a list is the average value of the random variable raised to the rth power of the distribution; the r-th moment of N, for $r > 0$, is defined as $\mathrm{E}[N^r]$ and denoted by $\mu'_N(r)$. Hence, for $r > 0$, we have

$$\mu'_N(r) = \mathrm{E}[N^r] = \sum_{k=0}^{\infty} k^r \; p_N(k). \tag{2.4}$$

We note that $\mu'_N(\cdot)$ is a well-defined non-decreasing function taking values in $[0, \infty]$, as $\Pr(N \in \{0, 1, \ldots\}) = 1$; also, note that $\mu_N = \mu'_N(1)$. In the following, when we refer to a moment it will be implicit that it is finite unless mentioned otherwise.

[1] For convenience, we have indexed μ_N with the random variable N instead of F_N or p_N, even though it is solely a function of the distribution of the random variable.

Another basic aspect of a distribution is its dispersion, and of the various measures of dispersion studied in the literature, the standard deviationthe square-root of variance is the most popular. Towards defining it, we first define the variancesecond central moment of a random variable X, measuring the expected squared deviation of between the variable and its mean of N, denoted by $\mathrm{Var}[N]$, as $\mathrm{Var}[N] := \mathrm{E}[(N - \mu_N)^2]$ when μ_N is finite. By basic properties of the expected value of a random variable, we see that $\mathrm{Var}[N] := \mathrm{E}[N^2] - [\mathrm{E}(N)]^2$. The standard deviation of N, denoted by σ_N, is defined as the square-root of Var N. Note that the latter is well-defined as $\mathrm{Var}[N]$, by its definition as the average squared deviation from the mean, is non-negative; $\mathrm{Var}[N]$ is denoted by σ_N^2. Note that these two measures take values in $[0, \infty]$.

2.2.2 Moment and Probability Generating Functions

Now we will introduce two generating functions that are found to be useful when working with count variables. Recall that for a discrete random variable, the moment generating function (mgf)the mgf of random variable N is defined the expectation of exp(tN), as a function of t of N, denoted as $M_N(\cdot)$, is defined as

$$M_N(t) = \mathrm{E}\,[e^{tN}] = \sum_{k=0}^{\infty} e^{tk}\, p_N(k), \quad t \in \mathbb{R}.$$

We note that while $M_N(\cdot)$ is well defined as it is the expectation of a non-negative random variable (e^{tN}), though it can assume the value ∞. Note that for a count random variable, $M_N(\cdot)$ is finite valued on $(-\infty, 0]$ with $M_N(0) = 1$. The following theorem, whose proof can be found in (**?**) (pages 285-6), encapsulates the reason for its name.

Theorem 2.1. *Let N be a count random variable such that* $\mathrm{E}\,[e^{t^* N}]$ *is finite for some $t^* > 0$. We have the following:*

All moments of N are finite, i.e.

$$\mathrm{E}[N^r] < \infty, \quad r > 0.$$

The mgf *can be used to* generate *its moments as follows:*

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m} M_N(t)\bigg|_{t=0} = \mathrm{E}N^m, \quad m \geq 1.$$

The mgf $M_N(\cdot)$ *characterizes the distribution; in other words it uniquely specifies the distribution.*

Another reason that the mgf is very useful as a tool is that for two independent random variables X and Y, with their mgfs existing in a neighborhood of 0, the mgf of $X + Y$ is the product of their respective mgfs.

A related generating function to the mgf is called the probability generating function (pgf)For a random variable N, its pgf is defined as the expectation of ŝN, as a function of s, and is a useful tool for random variables taking values in the non-negative integers. For a random variable N, by $P_N(\cdot)$ we denote its pgf and we define it as follows[2]:

$$P_N(s) := \mathrm{E}\,[s^N], \quad s \geq 0. \tag{2.5}$$

It is straightforward to see that if the mgf $M_N(\cdot)$ exists on $(-\infty, t^*)$ then

$$P_N(s) = M_N(\log(s)), \quad s < e^{t^*}.$$

Moreover, if the pgf exists on an interval $[0, s^*)$ with $s^* > 1$, then the mgf $M_N(\cdot)$ exists on $(-\infty, \log(s^*))$, and hence uniquely specifies the distribution of N by Theorem **??**. The following result for pgf is an analog of Theorem **??**, and in particular justifies its name.

Theorem 2.2. *Let N be a count random variable such that* $\mathrm{E}\,(s^*)^N$ *is finite for some $s^* > 1$. We have the following:*

All moments of N are finite, i.e.

$$\mathrm{E}\,N^r < \infty, \quad r \geq 0.$$

The pmf *of N can be derived from the* pgf *as follows:*

$$p_N(m) = \begin{cases} P_N(0), & m = 0; \\ \left(\frac{1}{m!}\right) \frac{\mathrm{d}^m}{\mathrm{d}s^m} P_N(s)\big|_{s=0}\ , & m \geq 1. \end{cases}$$

The factorial moments of N can be derived as follows:

$$\frac{\mathrm{d}^m}{\mathrm{d}s^m} P_N(s)\bigg|_{s=1} = \mathrm{E}\,\prod_{i=0}^{m-1}(N-i), \quad m \geq 1.$$

The pgf $P_N(\cdot)$ *characterizes the distribution; in other words it uniquely specifies the distribution.*

2.2.3 Important Frequency Distributions

In this sub-section we will study three important frequency distributions used in statistics, namely the binomial, the Poisson, and the negative binomial distributions. In the following, a risk denotes a unit covered by insurance. A risk could be an individual, a building, a company, or some other identifier for which insurance coverage is provided. For context, imagine an insurance data

[2] $0^0 = 1$

set containing the number of claims by risk or stratified in some other manner. The above mentioned distributions also happen to be the most commonly used in insurance practice for reasons, some of which we mention below.

- These distributions can be motivated by natural random experiments which are good approximations to real life processes from which many insurance data arise. Hence, not surprisingly, they together offer a reasonable fit to many insurance data sets of interest. The appropriateness of a particular distribution for the set of data can be determined using standard statistical methodologies, as we discuss later in this chapter.
- They provide a rich enough basis for generating other distributions that even better approximate or well cater to more real situations of interest to us.
  - The three distributions are either one-parameter or two-parameter distributions. In fitting to data, a parameter is assigned a particular value. The set of these distributions can be enlarged to their convex hullsthe convex hull of a set of points X is the smallest convex set that contains X by treating the parameter(s) as a random variable (or vector) with its own probability distribution, with this larger set of distributions offering greater flexibility. A simple example that is better addressed by such an enlargement is a portfolio of claims generated by insureds belonging to many different risk classesThe formation of different premiums for the same coverage based on each homogeneous group's characteristics..
  - In insurance data, we may observe either a marginal or inordinate number of zeros, that is, zero claims by risk. When fitting to the data, a frequency distribution in its standard specification often fails to reasonably account for this occurrence. The natural modification of the above three distributions, however, accommodate this phenomenon well towards offering a better fit.
  - In insurance we are interested in total claims paid, whose distribution results from compounding the fitted frequency distribution with a severity distribution. These three distributions have properties that make it easy to work with the resulting aggregate severity distribution.

Binomial Distribution

We begin with the binomial distribution which arises from any finite sequence of identical and independent experiments with binary outcomesOutcomes whose unit can take on only two possible states, traditionally labeled as 0 and 1. The most canonical of such experiments is the (biased or unbiased) coin tossing experiment with the outcome being heads or tails. So if N denotes the number of

heads in a sequence of m independent coin tossing experiments with an identical coin which turns heads up with probability q, then the distribution of N is called the binomial distributionA random variable has a binomial distribution (with parameters m and q) if it is the number of "successes" in a fixed number m of independent random trials, all of which have the same probability q of resulting in "success." with parameters (m, q), with m a positive integer and $q \in [0, 1]$. Note that when $q = 0$ (resp., $q = 1$) then the distribution is degenerate with $N = 0$ (resp., $N = m$) with probability 1. Clearly, its support when $q \in (0, 1)$ equals $\{0, 1, \ldots, m\}$ with pmf given by [3]

$$p_k := \binom{m}{k} q^k (1-q)^{m-k}, \quad k = 0, \ldots, m.$$

where

$$\binom{m}{k} = \frac{m!}{k!(m-k)!}$$

The reason for its name is that the pmf takes values among the terms that arise from the binomial expansion of $(q + (1 - q))^m$. This realization then leads to the the following expression for the pgf of the binomial distribution:

$$P_N(z) := \sum_{k=0}^{m} z^k \binom{m}{k} q^k (1-q)^{m-k} = \sum_{k=0}^{m} \binom{m}{k} (zq)^k (1-q)^{m-k} = (qz+(1-q))^m = (1+q(z-1))^m.$$

Note that the above expression for the pgf confirms the fact that the binomial distribution is the m-convolutionthe addition of m independent random variables of the Bernoulli distribution, which is the binomial distribution with $m = 1$ and pgf $(1+q(z-1))$. Also, note that the mgf of the binomial distribution is given by $(1 + q(e^t - 1))^m$.

The central moments of the binomial distribution can be found in a few different ways. To emphasize the key property that it is a m-convolution of the Bernoulli distribution, we derive below the moments using this property. We begin by observing that the Bernoulli distribution with parameter q assigns probability of q and $1-q$ to 1 and 0, respectively. So its mean equals q $(= 0 \times (1-q) + 1 \times q)$; note that its raw second moment equals its mean as $N^2 = N$ with probability 1. Using these two facts we see that the variance equals $q(1 - q)$. Moving on to the binomial distribution with parameters m and q, using the fact that it is the m-convolution of the Bernoulli distribution, we write N as the sum of $N_1, \ldots, N_m$, where N_i are iid Bernoulli variates. Now using the moments of Bernoulli and linearity of the expectation, we see that

$$\mathrm{E}[N] = \mathrm{E}[\sum_{i=1}^{m} N_i] = \sum_{i=1}^{m} \mathrm{E}[N_i] = mq.$$

[3] In the following we will suppress the reference to N and denote the pmf by the sequence $\{p_k\}_{k \geq 0}$, instead of the function $p_N(\cdot)$.

Also, using the fact that the variance of the sum of independent random variables is the sum of their variances, we see that

$$\mathrm{Var}[N] = \mathrm{Var}\left(\sum_{i=1}^{m} N_i\right) = \sum_{i=1}^{m} \mathrm{Var}[N_i] = mq(1-q).$$

Alternate derivations of the above moments are suggested in the exercises. One important observation, especially from the point of view of applications, is that the mean is greater than the variance unless $q = 0$.

Poisson Distribution

After the binomial distribution, the Poisson distributiona discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event (named after the French polymath Simeon Denis Poisson) is probably the most well known of discrete distributions. This is partly due to the fact that it arises naturally as the distribution of the count of the random occurrences of a type of event in a certain time period, if the rate of occurrences of such events is a constant. It also arises as the asymptotic limit of the binomial distribution with $m \to \infty$ and $mq \to \lambda$.

The Poisson distribution is parametrized by a single parameter usually denoted by λ which takes values in $(0, \infty)$. Its pmf is given by

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

It is easy to check that the above specifies a pmf as the terms are clearly non-negative, and that they sum to one follows from the infinite Taylor series expansion of e^{λ}. More generally, we can derive its pgf, $P(\cdot)$, as follows:

$$P_N(z) = \sum_{k=0}^{\infty} p_k z^k = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k z^k}{k!} = e^{-\lambda}e^{\lambda z} = e^{\lambda(z-1)}, \forall z \in \mathbb{R}.$$

From the above, we derive its mgf as follows:

$$M_N(t) = P_N(e^t) = e^{\lambda(e^t-1)}, t \in \mathbb{R}.$$

Towards deriving its mean, we note that for the Poisson distribution

$$kp_k = \begin{cases} 0, & k = 0; \\ \lambda\ p_{k-1}, & k \geq 1. \end{cases}$$

this can be checked easily. In particular, this implies that

$$\mathrm{E}[N] = \sum_{k \geq 0} k\ p_k = \lambda \sum_{k \geq 1} p_{k-1} = \lambda \sum_{j \geq 0} p_j = \lambda.$$

In fact, more generally, using either a generalization of the above or using Theorem **??**, we see that

$$\mathrm{E}\prod_{i=0}^{m-1}(N-i) = \left.\frac{\mathrm{d}^m}{\mathrm{d}s^m}P_N(s)\right|_{s=1} = \lambda^m, \quad m \geq 1.$$

This, in particular, implies that

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

Note that interestingly for the Poisson distribution $\mathrm{Var}[N] = \mathrm{E}[N]$.

Negative Binomial Distribution

The third important count distribution is the negative binomial distributionthe number of successes until we observe the rth failure in independent repetitions of an experiment with binary outcomes. Recall that the binomial distribution arose as the distribution of the number of successes in m independent repetition of an experiment with binary outcomes. If we instead consider the number of successes until we observe the r-th failure in independent repetitions of an experiment with binary outcomes, then its distribution is a negative binomial distribution. A particular case, when $r = 1$, is the geometric distribution. However when r in not an integer, the above random experiment would not be applicable. In the following, we will allow the parameter r to be any positive real number to then motivate the distribution more generally. To explain its name, we recall the binomial series, i.e.

$$(1+x)^s = 1 + sx + \frac{s(s-1)}{2!}x^2 + \ldots ..., \quad s \in \mathbb{R}; |x| < 1.$$

If we define $\binom{s}{k}$, the generalized binomial coefficient, by

$$\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!},$$

then we have

$$(1+x)^s = \sum_{k=0}^{\infty}\binom{s}{k}x^k, \quad s \in \mathbb{R}; |x| < 1.$$

If we let $s = -r$, then we see that the above yields

$$(1-x)^{-r} = 1 + rx + \frac{(r+1)r}{2!}x^2 + \ldots ... = \sum_{k=0}^{\infty}\binom{r+k-1}{k}x^k, \quad r \in \mathbb{R}; |x| < 1.$$

This implies that if we define p_k as

$$p_k = \binom{k+r-1}{k}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^k, \quad k = 0, 1, \ldots$$

for $r > 0$ and $\beta \geq 0$, then it defines a valid pmf. Such defined distribution is called the negative binomial distribution with parameters (r, β) with $r > 0$ and $\beta \geq 0$. Moreover, the binomial series also implies that the pgf of this distribution is given by

$$P_N(z) = (1 - \beta(z-1))^{-r}, \quad |z| < 1 + \frac{1}{\beta}, \beta \geq 0.$$

The above implies that the mgf is given by

$$M_N(t) = (1 - \beta(e^t - 1))^{-r}, \quad t < \log\left(1 + \frac{1}{\beta}\right), \beta \geq 0.$$

We derive its moments using Theorem **??** as follows:

$$\begin{aligned}
\mathrm{E}[N] &= M'(0) = r\beta e^t (1 - \beta(e^t - 1))^{-r-1}\big|_{t=0} = r\beta; \\
\mathrm{E}[N^2] &= M''(0) = \left[r\beta e^t (1 - \beta(e^t - 1))^{-r-1} + r(r+1)\beta^2 e^{2t}(1 - \beta(e^t - 1))^{-r-2}\right]\big|_{t=0} \\
&= r\beta(1 + \beta) + r^2\beta^2; \\
\text{and } \mathrm{Var}[N] &= \mathrm{E}[N^2] - (\mathrm{E}[N])^2 = r\beta(1+\beta) + r^2\beta^2 - r^2\beta^2 = r\beta(1+\beta)
\end{aligned}$$

We note that when $\beta > 0$, we have $\mathrm{Var}[N] > \mathrm{E}[N]$. In other words, this distribution is overdispersedthe presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model (relative to the Poisson); similarly, when $q > 0$ the binomial distribution is said to be underdispersedthere was less variation in the data than predicted (relative to the Poisson).

Finally, we observe that the Poisson distribution also emerges as a limit of negative binomial distributions. Towards establishing this, let β_r be such that as r approaches infinity $r\beta_r$ approaches $\lambda > 0$. Then we see that the mgfs of negative binomial distributions with parameters (r, β_r) satisfies

$$\lim_{r \to 0} (1 - \beta_r(e^t - 1))^{-r} = \exp\{\lambda(e^t - 1)\},$$

with the right hand side of the above equation being the mgf of the Poisson distribution with parameter λ.[4]

Show Quiz Solution

2.3 The (a, b, 0) Class

---

[4] For the theoretical basis underlying the above argument, see (**?**).

In this section, you learn how to:

- Define the (a,b,0) class of frequency distributions
- Discuss the importance of the recursive relationship underpinning this class of distributions
- Identify conditions under which this general class reduces to each of the binomial, Poisson, and negative binomial distributions

---

In the previous section we studied three distributions, namely the binomial, the Poisson and the negative binomial distributions. In the case of the Poisson, to derive its mean we used the the fact that

$$kp_k = \lambda p_{k-1}, \quad k \geq 1,$$

which can be expressed equivalently as

$$\frac{p_k}{p_{k-1}} = \frac{\lambda}{k}, \quad k \geq 1.$$

Interestingly, we can similarly show that for the binomial distribution

$$\frac{p_k}{p_{k-1}} = \frac{-q}{1-q} + \left(\frac{(m+1)q}{1-q}\right)\frac{1}{k}, \quad k = 1, \ldots, m,$$

and that for the negative binomial distribution

$$\frac{p_k}{p_{k-1}} = \frac{\beta}{1+\beta} + \left(\frac{(r-1)\beta}{1+\beta}\right)\frac{1}{k}, \quad k \geq 1.$$

The above relationships are all of the form

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k \geq 1; \tag{2.6}$$

this raises the question if there are any other distributions which satisfy this seemingly general recurrence relation. Note that the ratio on the left, the ratio of two probabilities, is non-negative.

To begin with, let $a < 0$. In this case as $k \to \infty$, $(a + b/k) \to a < 0$. It follows that if $a < 0$ then b should satisfy $b = -ka$, for some $k \geq 1$. Any such pair (a, b) can be written as

$$\left(\frac{-q}{1-q}, \frac{(m+1)q}{1-q}\right), \quad q \in (0,1), m \geq 1;$$

note that the case $a < 0$ with $a + b = 0$ yields the degenerate at 0 distribution which is the binomial distribution with $q = 0$ and arbitrary $m \geq 1$.

In the case of $a = 0$, again by non-negativity of the ratio p_k/p_{k-1}, we have $b \geq 0$. If $b = 0$ the distribution is degenerate at 0, which is a binomial with $q = 0$ or a

Poisson distribution with $\lambda = 0$ or a negative binomial distribution with $\beta = 0$. If $b > 0$, then clearly such a distribution is a Poisson distribution with mean (i.e. λ) equal to b, as presented at the beginning of this section.

In the case of $a > 0$, again by non-negativity of the ratio p_k/p_{k-1}, we have $a + b/k \geq 0$ for all $k \geq 1$. The most stringent of these is the inequality $a + b \geq 0$. Note that $a + b = 0$ again results in degeneracy at 0; excluding this case we have $a + b > 0$ or equivalently $b = (r - 1)a$ with $r > 0$. Some algebra easily yields the following expression for p_k:

$$p_k = \binom{k + r - 1}{k} p_0 a^k, \quad k = 1, 2, \ldots.$$

The above series converges for $a < 1$ when $r > 0$, with the sum given by $p_0 * ((1 - a)^{(-r)} - 1)$. Hence, equating the latter to $1 - p_0$ we get $p_0 = (1 - a)^{(r)}$. So in this case the pair (a, b) is of the form $(a, (r-1)a)$, for $r > 0$ and $0 < a < 1$; since an equivalent parametrization is $(\beta/(1 + \beta), (r - 1)\beta/(1 + \beta))$, for $r > 0$ and $\beta > 0$, we see from above that such distributions are negative binomial distributions.

From the above development we see that not only does the recurrence (**??**) tie these three distributions together, but also it characterizes them. For this reason these three distributions are collectively referred to in the actuarial literature as (a,b,0) class of distributions, with 0 referring to the starting point of the recurrence. Note that the value of p_0 is implied by (a, b) since the probabilities have to sum to one. Of course, (**??**) as a recurrence relation for p_k makes the computation of the pmf efficient by removing redundancies. Later, we will see that it does so even in the case of compound distributions with the frequency distribution belonging to the $(a, b, 0)$ class - this fact is the more important motivating reason to study these three distributions from this viewpoint.

Example 2.3.1. A discrete probability distribution has the following properties

$$p_k = c\left(1 + \frac{2}{k}\right) p_{k-1} \quad k = 1, 2, 3, \ldots$$
$$p_1 = \frac{9}{256}$$

Determine the expected value of this discrete random variable.

Show Example Solution

Solution: Since the pmf satisfies the $(a, b, 0)$ recurrence relation we know that the underlying distribution is one among the binomial, Poisson, and negative binomial distributions. Since the ratio of the parameters (i.e. b/a) equals 2, we know that it is negative binomial and that $r = 3$. Moreover, since for a negative

binomial $p_1 = r(1+\beta)^{-(r+1)}\beta$, we have

$$\begin{aligned} \frac{9}{256} &= 3\frac{\beta}{(1+\beta)^4} \\ \Longrightarrow \quad \frac{3}{(1+3)^4} &= \frac{\beta}{(1+\beta)^4} \\ \Longrightarrow \quad \beta &= 3. \end{aligned}$$

Finally, since the mean of a negative binomial is $r\beta$ we have the mean of the given distribution equals 9.

Show Quiz Solution

2.4 Estimating Frequency Distributions

In this section, you learn how to:

- Define a likelihood for a sample of observations from a discrete distribution
- Define the maximum likelihood estimator for a random sample of observations from a discrete distribution
- Calculate the maximum likelihood estimator for the binomial, Poisson, and negative binomial distributions

---

2.4.1 Parameter Estimation

In Section **??** we introduced three distributions of importance in modeling various types of count data arising from insurance. Let us now suppose that we have a set of count data to which we wish to fit a distribution, and that we have determined that one of these $(a, b, 0)$ distributions is more appropriate than the others. Since each one of these forms a class of distributions if we allow its parameter(s) to take any permissible value, there remains the task of determining the **best** value of the parameter(s) for the data at hand. This is a statistical point estimation problem, and in parametric inference problems the statistical inference paradigm of maximum likelihood usually yields efficient estimators. In this section we will describe this paradigm and derive the maximum likelihood estimators.

Let us suppose that we observe the independent and identically distributed, iid, random variables $X_1, X_2, \ldots, X_n$ from a distribution with pmf p_θ, where θ is a

parameter and an unknown value in the parameter space $\Theta \subseteq \mathbb{R}^d$. For example, in the case of the Poisson distribution

$$p_\theta(x) = e^{-\theta}\frac{\theta^x}{x!}, \quad x = 0, 1, \ldots,$$

with $\theta = (0, \infty)$. In the case of the binomial distribution we have

$$p_\theta(x) = \binom{m}{x} q^x (1-q)^{m-x}, \quad x = 0, 1, \ldots, m,$$

with $\theta := (m, q) \in \{0, 1, 2, \ldots\} \times [0, 1]$. Let us suppose that the observations are $x_1, \ldots, x_n$, observed values of the random sample $X_1, X_2, \ldots, X_n$ presented earlier. In this case, the probability of observing this sample from p_θ equals

$$\prod_{i=1}^{n} p_\theta(x_i).$$

The above, denoted by $L(\theta)$, viewed as a function of θ is called the likelihood. Note that we suppressed its dependence on the data, to emphasize that we are viewing it as a function of the parameter. For example, in the case of the Poisson distribution we have

$$L(\lambda) = e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}\left(\prod_{i=1}^{n} x_i!\right)^{-1};$$

in the case of the binomial distribution we have

$$L(m, q) = \left(\prod_{i=1}^{n}\binom{m}{x_i}\right) q^{\sum_{i=1}^n x_i}(1-q)^{nm-\sum_{i=1}^n x_i}.$$

The maximum likelihood estimator (MLE)the possible value of the parameter for which the chance of observing the data largest for θ is any maximizer of the likelihood; in a sense the MLE chooses the set of parameter values that best explains the observed observations. Consider a sample of size 3 from a Bernoulli distribution (binomial with $m = 1$) with values $0, 1, 0$. The likelihood in this case is easily checked to equal

$$L(q) = q(1-q)^2,$$

and the plot of the likelihood is given in Figure **??**. As shown in the plot, the maximum value of the likelihood equals $4/27$ and is attained at $q = 1/3$, and hence the maximum likelihood estimate for q is $1/3$ for the given sample. In this case one can resort to algebra to show that

$$q(1-q)^2 = \left(q - \frac{1}{3}\right)^2\left(q - \frac{4}{3}\right) + \frac{4}{27},$$

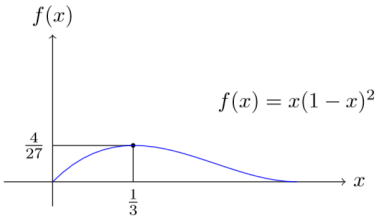


Figure 2.1: Likelihood of a $(0, 1, 0)$ 3-sample from Bernoulli

and conclude that the maximum equals $4/27$, and is attained at $q = 1/3$ (using the fact that the first term is non-positive in the interval $[0, 1]$). But as is apparent, this way of deriving the mle using algebra does not generalize. In general, one resorts to calculus to derive the mle - note that for some likelihoods one may have to resort to other optimization methods, especially when the likelihood has many local extremathe largest and smallest value of the function within a given range. It is customary to equivalently maximize the logarithm of the likelihood[5] $L(\cdot)$, denoted by $l(\cdot)$, and look at the set of zeros of its first derivative[6] $l'(\cdot)$. In the case of the above likelihood, $l(q) = \log(q) + 2\log(1-q)$, and

$$l'(q) := \frac{\mathrm{d}}{\mathrm{d}q} l(q) = \frac{1}{q} - \frac{2}{1-q}.$$

The unique zero of $l'(\cdot)$ equals $1/3$, and since $l''(\cdot)$ is negative, we have $1/3$ is the unique maximizer of the likelihood and hence its maximum likelihood estimate.

2.4.2 Frequency Distributions MLE

In the following, we derive the maximum likelihood estimator, MLE, for the three members of the $(a, b, 0)$ class. We begin by summarizing the discussion above. In the setting of observing iid, independent and identically distributed, random variables $X_1, X_2, \ldots, X_n$ from a distribution with pmf p_θ, where θ takes an unknown value in $\Theta \subseteq \mathbb{R}^d$, the likelihood $L(\cdot)$, a function on Θ is defined as

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

where $x_1, \ldots, x_n$ are the observed values. The MLE of θ, denoted as $\hat{\theta}_{\mathrm{MLE}}$, is a function which maps the observations to an element of the set of maximizers of $L(\cdot)$, namely

$$\{\theta | L(\theta) = \max_{\eta \in \Theta} L(\eta)\}.$$

[5]The set of maximizers of $L(\cdot)$ are the same as the set of maximizers of any strictly increasing function of $L(\cdot)$, and hence the same as those for $l(\cdot)$.

[6]A slight benefit of working with $l(\cdot)$ is that constant terms in $L(\cdot)$ do not appear in $l'(\cdot)$ whereas they do in $L'(\cdot)$.

Note the above set is a function of the observations, even though this dependence is not made explicit. In the case of the three distributions that we will study, and quite generally, the above set is a singleton with probability tending to one (with increasing sample size). In other words, for many commonly used distributions and when the sample size is large, the likelihood estimate is uniquely defined with high probability.

In the following, we will assume that we have observed n iid random variables $X_1, X_2, \ldots, X_n$ from the distribution under consideration, even though the parametric value is unknown. Also, $x_1, x_2, \ldots, x_n$ will denote the observed values. We note that in the case of count data, and data from discrete distributions in general, the likelihood can alternately be represented as

$$L(\theta) := \prod_{k \geq 0} \left(p_\theta(k)\right)^{m_k},$$

where

$$m_k := |\{i | x_i = k, 1 \leq i \leq n\}| = \sum_{i=1}^{n} I(x_i = k), \quad k \geq 0.$$

Note that this transformation retains all of the data, compiling it in a streamlined manner. For large n it leads to compression of the data in the sense of sufficiency. Below, we present expressions for the MLE in terms of $\{m_k\}_{k \geq 1}$ as well.

MLE - Poisson Distribution: In this case, as noted above, the likelihood is given by

$$L(\lambda) = \left(\prod_{i=1}^{n} x_i!\right)^{-1} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i},$$

which implies that

$$l(\lambda) = -\sum_{i=1}^{n} \log(x_i!) - n\lambda + \log(\lambda) \cdot \sum_{i=1}^{n} x_i,$$

and

$$l'(\lambda) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i.$$

In evaluating $l''(\lambda)$, when $\sum_{i=1}^{n} x_i > 0$, $l'' < 0$. Consequently, the maximum is attained at the sample mean, $\overline{x}$, presented below. When $\sum_{i=1}^{n} x_i = 0$, the likelihood is a decreasing function and hence the maximum is attained at the least possible parameter value; this results in the maximum likelihood estimate being zero. Hence, we have

$$\overline{x} = \hat{\lambda}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Note that the sample mean can be computed also as

$$\frac{1}{n}\sum_{k\geq 1} km_k.$$

It is noteworthy that in the case of the Poisson, the exact distribution of $\hat{\lambda}_{\text{MLE}}$ is available in closed form - it is a scaled Poisson - when the underlying distribution is a Poisson. This is so as the sum of independent Poisson random variables is a Poisson as well. Of course, for large sample size one can use the ordinary Central Limit Theorem (CLT)in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed to derive a normal approximation. Note that the latter approximation holds even if the underlying distribution is any distribution with a finite second moment.

MLE - Binomial Distribution: Unlike the case of the Poisson distribution, the parameter space in the case of the binomial is 2-dimensional. Hence the optimization problem is a bit more challenging. We begin by observing that the likelihood is given by

$$L(m,q) = \left(\prod_{i=1}^{n}\binom{m}{x_i}\right) q^{\sum_{i=1}^{n} x_i}(1-q)^{nm-\sum_{i=1}^{n} x_i},$$

and the log-likelihood by

$$l(m,q) = \sum_{i=1}^{n}\log\left(\binom{m}{x_i}\right) + \left(\sum_{i=1}^{n} x_i\right)\log(q) + \left(nm - \sum_{i=1}^{n} x_i\right)\log(1-q).$$

Note that since m takes only non-negative integer values, we cannot use multivariate calculus to find the optimal values. Nevertheless, we can use single variable calculus to show that

$$\hat{q}_{\text{MLE}} \times \hat{m}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.7}$$

Towards this we note that for a fixed value of m,

$$\frac{\delta}{\delta q} l(m,q) = \left(\sum_{i=1}^{n} x_i\right)\frac{1}{q} - \left(nm - \sum_{i=1}^{n} x_i\right)\frac{1}{1-q},$$

and that

$$\frac{\delta^2}{\delta q^2} l(m,q) = -\left[\left(\sum_{i=1}^{n} x_i\right)\frac{1}{q^2} + \left(nm - \sum_{i=1}^{n} x_i\right)\frac{1}{(1-q)^2}\right] \leq 0.$$

The above implies that for any fixed value of m, the maximizing value of q satisfies

$$mq = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and hence we establish equation (**??**). The above reduces the task to the search for $\hat{m}_{\text{MLE}}$, which is member of the set of the maximizers of

$$L\left(m, \frac{1}{nm}\sum_{i=1}^{n} x_i\right). \tag{2.8}$$

Note the likelihood would be zero for values of m smaller than $\max\limits_{1\le i\le n} x_i$, and hence

$$\hat{m}_{\text{MLE}} \ge \max_{1\le i\le n} x_i.$$

Towards specifying an algorithm to compute $\hat{m}_{\text{MLE}}$, we first point out that for some data sets $\hat{m}_{\text{MLE}}$ could equal ∞, indicating that a Poisson distribution would render a better fit than any binomial distribution. This is so as the binomial distribution with parameters $(m, \overline{x}/m)$ approaches the Poisson distribution with parameter $\overline{x}$ with m approaching infinity. The fact that some data sets will **prefer** a Poisson distribution should not be surprising since in the above sense the set of Poisson distribution is on the boundary of the set of binomial distributions.

Interestingly, in (**?**) they show that if the sample mean is less than or equal to the sample variance then $\hat{m}_{\text{MLE}} = \infty$; otherwise, there exists a finite m that maximizes equation (**??**). In Figure **??** below we display the plot of $L\left(m, \frac{1}{nm}\sum_{i=1}^{n} x_i\right)$ for three different samples of size 5; they differ only in the value of the sample maximum. The first sample of $(2, 2, 2, 4, 5)$ has the ratio of sample mean to sample variance greater than 1 (1.875), the second sample of $(2, 2, 2, 4, 6)$ has the ratio equal to 1.25 which is closer to 1, and the third sample of $(2, 2, 2, 4, 7)$ has the ratio less than 1 (0.885). For these three samples, as shown in Figure **??**, $\hat{m}_{\text{MLE}}$ equals 7, 18 and ∞, respectively. Note that the limiting value of $L\left(m, \frac{1}{nm}\sum_{i=1}^{n} x_i\right)$ as m approaches infinity equals

$$\left(\prod_{i=1}^{n} x_i!\right)^{-1} \exp\left\{-\sum_{i=1}^{n} x_i\right\} \overline{x}^{n\overline{x}}. \tag{2.9}$$

Also, note that Figure **??** shows that the MLE of m is non-robust, i.e. changes in a small proportion of the data set can cause large changes in the estimator.

The above discussion suggests the following simple algorithm:

- Step 1. If the sample mean is less than or equal to the sample variance, $\hat{m}_{MLE} = \infty$. The MLE suggested distribution is a Poisson distribution with $\hat{\lambda} = \overline{x}$.

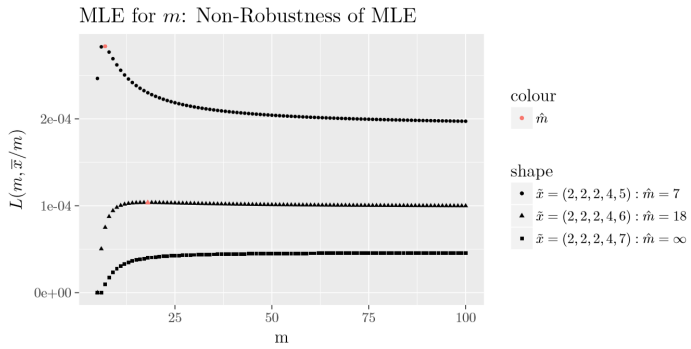


Figure 2.2: Plot of $L(m, \overline{x}/m)$ for binomial distribution

- Step 2. If the sample mean is greater than the sample variance, then compute $L(m, \overline{x}/m)$ for m values greater than or equal to the sample maximum until $L(m, \overline{x}/m)$ is close to the value of the Poisson likelihood given in (**??**). The value of m that corresponds to the maximum value of $L(m, \overline{x}/m)$ among those computed equals $\hat{m}_{MLE}$.

We note that if the underlying distribution is the binomial distribution with parameters (m, q) (with $q > 0$) then $\hat{m}_{MLE}$ will equal m for large sample sizes. Also, $\hat{q}_{MLE}$ will have an asymptotically normal distribution and converge with probability one to q.

MLE - Negative Binomial Distribution: The case of the negative binomial distribution is similar to that of the binomial distribution in the sense that we have two parameters and the MLEs are not available in closed form. A difference between them is that unlike the binomial parameter m which takes positive integer values, the parameter r of the negative binomial can assume any positive real value. This makes the optimization problem a tad more complex. We begin by observing that the likelihood can be expressed in the following form:

$$L(r, \beta) = \left(\prod_{i=1}^{n} \binom{r + x_i - 1}{x_i} \right) (1 + \beta)^{-n(r+\overline{x})} \beta^{n\overline{x}}.$$

The above implies that log-likelihood is given by

$$l(r, \beta) = \sum_{i=1}^{n} \log \binom{r + x_i - 1}{x_i} - n(r + \overline{x}) \log(1 + \beta) + n\overline{x} \log \beta,$$

and hence

$$\frac{\delta}{\delta\beta} l(r, \beta) = -\frac{n(r + \overline{x})}{1 + \beta} + \frac{n\overline{x}}{\beta}.$$

Equating the above to zero, we get

$$\hat{r}_{MLE} \times \hat{\beta}_{MLE} = \overline{x}.$$

The above reduces the two dimensional optimization problem to a one-dimensional problem - we need to maximize

$$l(r, \overline{x}/r) = \sum_{i=1}^{n} \log \binom{r + x_i - 1}{x_i} - n(r + \overline{x}) \log(1 + \overline{x}/r) + n\overline{x} \log(\overline{x}/r),$$

with respect to r, with the maximizing r being its MLE and $\hat{\beta}_{MLE} = \overline{x}/\hat{r}_{MLE}$. In (**?**) it is shown that if the sample variance is greater than the sample mean then there exists a unique $r > 0$ that maximizes $l(r, \overline{x}/r)$ and hence a unique MLE for r and β. Also, they show that if $\hat{\sigma}^2 \leq \overline{x}$, then the negative binomial likelihood will be dominated by the Poisson likelihood with $\hat{\lambda} = \overline{x}$. In other words, a Poisson distribution offers a better fit to the data. The guarantee in the case of $\hat{\sigma}^2 > \hat{\mu}$ permits us to use some algorithm to maximize $l(r, \overline{x}/r)$. Towards an alternate method of computing the likelihood, we note that

$$l(r, \overline{x}/r) = \sum_{i=1}^{n} \sum_{j=1}^{x_i} \log(r-1+j) - \sum_{i=1}^{n} \log(x_i!) - n(r+\overline{x}) \log(r+\overline{x}) + nr \log(r) + n\overline{x} \log(\overline{x}),$$

which yields

$$\left(\frac{1}{n}\right) \frac{\delta}{\delta r} l(r, \overline{x}/r) = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \sum_{j=1}^{x_i} \frac{1}{r - 1 + j} - \log(r + \overline{x}) + \log(r).$$

We note that, in the above expressions for the terms involving a double summation, the inner sum equals zero if $x_i = 0$. The maximum likelihood estimate for r is a root of the last expression and we can use a root finding algorithm to compute it. Also, we have

$$\left(\frac{1}{n}\right) \frac{\delta^2}{\delta r^2} l(r, \overline{x}/r) = \frac{\overline{x}}{r(r + \overline{x})} - \left(\frac{1}{n}\right) \sum_{i=1}^{n} \sum_{j=1}^{x_i} \frac{1}{(r - 1 + j)^2}.$$

A simple but quickly converging iterative root finding algorithm is the Newton's method, which incidentally the Babylonians are believed to have used for computing square roots. Under this method, an initial approximation is selected for the root and new approximations for the root are successively generated until convergence. Applying the Newton's method to our problem results in the following algorithm:
Step i. Choose an approximate solution, say r_0. Set k to 0.
Step ii. Define r_{k+1} as

$$r_{k+1} := r_k - \frac{\left(\frac{1}{n}\right) \sum_{i=1}^{n} \sum_{j=1}^{x_i} \frac{1}{r_k - 1 + j} - \log(r_k + \overline{x}) + \log(r_k)}{\frac{\overline{x}}{r_k(r_k + \overline{x})} - \left(\frac{1}{n}\right) \sum_{i=1}^{n} \sum_{j=1}^{x_i} \frac{1}{(r_k - 1 + j)^2}}$$

Step iii. If $r_{k+1} \sim r_k$, then report r_{k+1} as maximum likelihood estimate; else increment k by 1 and repeat Step ii.

For example, we simulated a 5 observation sample of $41, 49, 40, 27, 23$ from the negative binomial with parameters $r = 10$ and $\beta = 5$. Choosing the starting value of r such that

$$r\beta = \hat{\mu} \quad \text{and} \quad r\beta(1+\beta) = \hat{\sigma}^2$$

where $\hat{\mu}$ represents the estimated mean and $hat\sigma^2$ is the estimated variance. This leads to the starting value for r of 23.14286. The iterates of r from the Newton's method are

$$21.39627, 21.60287, 21.60647, 21.60647;$$

the rapid convergence seen above is typical of the Newton's method. Hence in this example, $\hat{r}_{MLE} \sim 21.60647$ and $\hat{\beta}_{MLE} = 8.3308$.

R Implementation of Newton's Method - Negative Binomial MLE for r

Show R Code

```
Newton<-function(x,abserr){
mu<-mean(x);
sigma2<-mean(x^2)-mu^2;
r<-mu^2/(sigma2-mu);
b<-TRUE;
iter<-0;
while (b) {
tr<-r;
m1<-mean(c(x[x==0],sapply(x[x>0],function(z){sum(1/(tr:(tr-1+z)))})));
m2<-mean(c(x[x==0],sapply(x[x>0],function(z){sum(1/(tr:(tr-1+z))^2)})));
r<-tr-(m1-log(1+mu/tr))/(mu/(tr*(tr+mu))-m2);
b<-!(abs(tr-r)<abserr);
iter<-iter+1;
}
c(r,iter)
}
```

To summarize our discussion of MLE for the $(a, b, 0)$ class of distributions, in Figure **??** below we plot the maximum value of the Poisson likelihood, $L(m, \overline{x}/m)$ for the binomial, and $L(r, \overline{x}/r)$ for the negative binomial, for the three samples of size 5 given in Table 2.1. The data was constructed to cover the three orderings of the sample mean and variance. As show in the Figure **??**, and supported by theory, if $\hat{\mu} < \hat{\sigma}^2$ then the negative binomial will result in a higher maximum likelihood value; if $\hat{\mu} = \hat{\sigma}^2$ the Poisson will have the highest likelihood value; and finally in the case that $\hat{\mu} > \hat{\sigma}^2$ the binomial will give a better fit than the others. So before fitting a frequency data with an $(a, b, 0,)$ distribution, it is best to start with examining the ordering of $\hat{\mu}$ and $\hat{\sigma}^2$. We again emphasize that the Poisson is on the **boundary** of the negative binomial and binomial distributions. So in the case that $\hat{\mu} \geq \hat{\sigma}^2$ ($\hat{\mu} \leq \hat{\sigma}^2$, resp.) the Poisson will

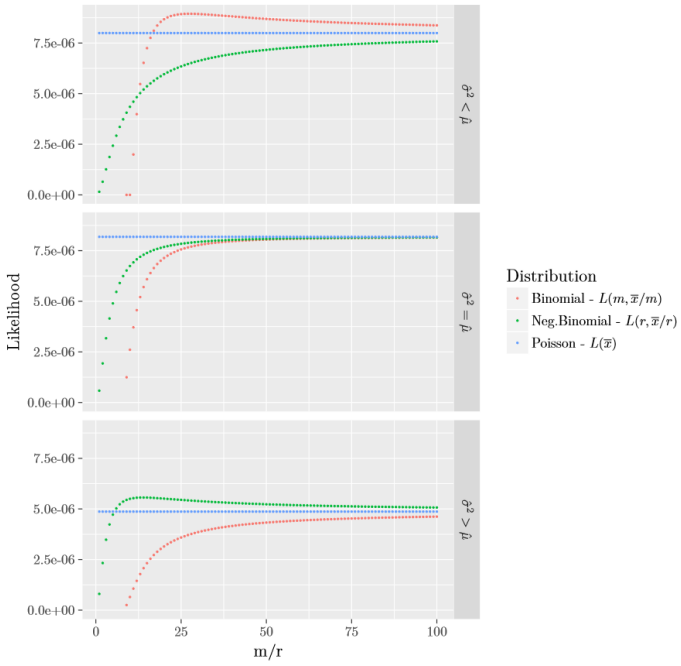


Figure 2.3: Plot of $(a, b, 0)$ Partially Maximized Likelihoods

yield a better fit than the negative binomial (binomial, resp.), which will also be indicated by $\hat{r} = \infty$ ($\hat{m} = \infty$, resp.).

Data	Mean ($\hat{\mu}$)	Variance ($\hat{\sigma}^2$)
$(2, 3, 6, 8, 9)$	5.60	7.44
$(2, 5, 6, 8, 9)$	6	6
$(4, 7, 8, 10, 11)$	8	6

Table 2.1 : Three Samples of Size 5

Show Quiz Solution

2.5 Other Frequency Distributions

In this section, you learn how to:

- Define the (a,b,1) class of frequency distributions and discuss the importance of the recursive relationship underpinning this class of distributions
- Interpret zero truncated and modifed versions of the binomial, Poisson, and negative binomial distributions
- Compute probabilities using the recursive relationship

In the previous sections, we discussed three distributions with supports contained in the set of non-negative integers, which well cater to many insurance applications. Moreover, typically by allowing the parameters to be a function of known (to the insurer) explanatory variablesIn regression, the explanatory variable is the one that is supposed to "explain" the other. such as age, sex, geographic location (territory), and so forth, these distributions allow us to explain claim probabilities in terms of these variables. The field of statistical study that studies such models is known as regression analysisa set of statistical processes for estimating the relationships among variables - it is an important topic of actuarial interest that we will not pursue in this book; see (**?**).

There are clearly infinitely many other count distributions, and more importantly the above distributions by themselves do not cater to all practical needs. In particular, one feature of some insurance data is that the proportion of zero counts can be out of place with the proportion of other counts to be explainable by the above distributions. In the following we modify the above distributions to allow for arbitrary probability for zero count irrespective of the assignment of relative probabilities for the other counts. Another feature of a data set which is naturally comprised of homogeneousUnits of exposure that face approximately the same expected frequency and severity of loss. subsets is that while the above distributions may provide good fits to each subset, they may fail to do so to the whole data set. Later we naturally extend the $(a, b, 0)$ distributions to be able to cater to, in particular, such data sets.

2.5.1 Zero Truncation or Modification

Let us suppose that we are looking at auto insurance policies which appear in a database of auto claims made in a certain period. If one is to study the number of claims that these policies have made during this period, then clearly the distribution has to assign a probability of zero to the count variable assuming the value zero. In other words, by restricting attention to count data from policies in the database of claims, we have in a sense zero-truncated the count data of all policies. In personal lines (like auto), policyholders may not want to report that first claim because of fear that it may increase future insurance rates - this behavior will inflate the proportion of zero counts. Examples such as the latter modify the proportion of zero counts. Interestingly, natural modifications of the three distributions considered above are able to provide good fits to zero-modified/truncated data sets arising in insurance.

As presented below, we modify the probability assigned to zero count by the $(a, b, 0)$ class while maintaining the relative probabilities assigned to non-zero counts - zero modification. Note that since the $(a, b, 0)$ class of distributions satisfies the recurrence (**??**), maintaining relative probabilities of non-zero counts implies that recurrence (**??**) is satisfied for $k \geq 2$. This leads to the definition of the following class of distributions.

Definition. A count distribution is a member of the $(a, b, 1)$ class if for some constants a and b the probabilities p_k satisfy

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k \geq 2. \tag{2.10}$$

Note that since the recursion starts with p_1, and not p_0, we refer to this superclass of $(a, b, 0)$ distributions by (a,b,1). To understand this class, recall that each valid pair of values for a and b of the $(a, b, 0)$ class corresponds to a unique vector of probabilities $\{p_k\}_{k\geq 0}$. If we now look at the probability vector $\{\tilde{p}_k\}_{k\geq 0}$ given by

$$\tilde{p}_k = \frac{1 - \tilde{p}_0}{1 - p_0} \cdot p_k, \quad k \geq 1,$$

where $\tilde{p}_0 \in [0, 1)$ is arbitrarily chosen, then since the relative probabilities for positive values according to $\{p_k\}_{k\geq 0}$ and $\{\tilde{p}_k\}_{k\geq 0}$ are the same, we have $\{\tilde{p}_k\}_{k\geq 0}$ satisfies recurrence (**??**). This, in particular, shows that the class of $(a, b, 1)$ distributions is strictly wider than that of $(a, b, 0)$.

In the above, we started with a pair of values for a and b that led to a valid $(a, b, 0)$ distribution, and then looked at the $(a, b, 1)$ distributions that corresponded to this $(a, b, 0)$ distribution. We will now argue that the $(a, b, 1)$ class allows for a larger set of permissible distributions for a and b than the $(a, b, 0)$ class. Recall from Section **??** that in the case of $a < 0$ we did not use the fact that the recurrence (**??**) started at $k = 1$, and hence the set of pairs (a, b) with $a < 0$ that are permissible for the $(a, b, 0)$ class is identical to those that are permissible for the $(a, b, 1)$ class. The same conclusion is easily drawn for pairs with $a = 0$. In the case that $a > 0$, instead of the constraint $a + b > 0$ for the $(a, b, 0)$ class we now have the weaker constraint of $a + b/2 > 0$ for the $(a, b, 1)$ class. With the parametrization $b = (r - 1)a$ as used in Section **??**, instead of $r > 0$ we now have the weaker constraint of $r > -1$. In particular, we see that while zero modifying a $(a, b, 0)$ distribution leads to a distribution in the $(a, b, 1)$ class, the conclusion does not hold in the other direction.

Zero modification of a count distribution F such that it assigns zero probability to zero count is called a zero truncationZero modification of a count distribution such that it assigns zero probability to zero count of F. Hence, the zero truncated version of probabilities $\{p_k\}_{k\geq 0}$ is given by

$$\tilde{p}_k = \begin{cases} 0, & k = 0; \\ \frac{p_k}{1-p_0}, & k \geq 1. \end{cases}$$

In particular, we have that a zero modification of a count distribution $\{p_k^T\}_{k\geq 0}$, denoted by $\{p_k^M\}_{k\geq 0}$, can be written as a convex combinationa linear combination of points where all coefficients are non-negative and sum to 1 of the degenerate distributiona deterministic distribution and takes only a single value at 0 and the zero truncation of $\{p_k\}_{k\geq 0}$, denoted by $\{p_k^T\}_{k\geq 0}$. That is we have

$$p_k^M = p_0^M \cdot \delta_0(k) + (1 - p_0^M) \cdot p_k^T, \quad k \geq 0.$$

Example 2.5.1. Zero Truncated/Modified Poisson. Consider a Poisson distribution with parameter $\lambda = 2$. Calculate $p_k, k = 0, 1, 2, 3$, for the usual (unmodified), truncated and a modified version with ($p_0^M = 0.6$).

Show Example Solution

Solution. For the Poisson distribution as a member of the $(a, b,0)$ class, we have $a = 0$ and $b = \lambda = 2$. Thus, we may use the recursion $p_k = \lambda p_{k-1}/k = 2p_{k-1}/k$ for each type, after determining starting probabilities. The calculation of probabilities for $k \leq 3$ is shown in Table 2.2.

k	p_k	p_k^T	p_k^M
0	$p_0 = e^{-\lambda} = 0.135335$	0	0.6
1	$p_1 = p_0(0 + \frac{\lambda}{1}) = 0.27067$	$\frac{p_1}{1-p_0} = 0.313035$	$\frac{1-p_0^M}{1-p_0}\ p_1 = 0.125214$
2	$p_2 = p_1\left(\frac{\lambda}{2}\right) = 0.27067$	$p_2^T = p_1^T\left(\frac{\lambda}{2}\right) = 0.313035$	$p_2^M = p_1^M\left(\frac{\lambda}{2}\right) = 0.125214$
3	$p_3 = p_2\left(\frac{\lambda}{3}\right) = 0.180447$	$p_3^T = p_2^T\left(\frac{\lambda}{3}\right) = 0.208690$	$p_3^M = p_2^M\left(\frac{\lambda}{3}\right) = 0.083476$

Table 2.2 : Calculation of probabilities for $k \leq 3$

Show Quiz Solution

2.6 Mixture Distributions

In this section, you learn how to:

- Define a mixture distribution when the mixing component is based on a finite number of sub-groups
- Compute mixture distribution probabilities from mixing proportions and knowledge of the distribution of each subgroup
- Define a mixture distribution when the mixing component is continuous

---

In many applications, the underlying population consists of naturally defined sub-groups with some homogeneity within each sub-group. In such cases it is convenient to model the individual sub-groups, and in a ground-up manner model the whole population. As we shall see below, beyond the aesthetic appeal of the approach, it also extends the range of applications that can be catered to by standard parametric distributions.

Let k denote the number of defined sub-groups in a population, and let F_i denote the distribution of an observation drawn from the i-th subgroup. If we let α_i denote the proportion of the population in the i-th subgroup, with $\sum_{i=1}^{k} \alpha_i = 1$, then the distribution of a randomly chosen observation from the population, denoted by F, is given by

$$F(x) = \sum_{i=1}^{k} \alpha_i \cdot F_i(x). \tag{2.11}$$

The above expression can be seen as a direct application of the Law of Total Probability. As an example, consider a population of drivers split broadly into two sub-groups, those with at most 5-years of driving experience and those with more than 5-years experience. Let α denote the proportion of drivers with less than 5 years experience, and $F_{\leq 5}$ and $F_{>5}$ denote the distribution of the count of claims in a year for a driver in each group, respectively. Then the distribution of claim count of a randomly selected driver is given by

$$\alpha \cdot F_{\leq 5}(x) + (1 - \alpha) F_{>5}(x).$$

An alternate definition of a mixture distributionThe probability distribution of a random variable that is derived from a collection of other random variables as follows: first, a random variable is selected by chance from the collection according to given probabilities of selection, and then the value of the selected random variable is realized is as follows. Let N_i be a random variable with distribution F_i, $i = 1, \ldots, k$. Let I be a random variable taking values $1, 2, \ldots, k$ with probabilities $\alpha_1, \ldots, \alpha_k$, respectively. Then the random variable N_I has a distribution given by equation (**??**)[7].

In (**??**) we see that the distribution function is a convex combination of the component distribution functions. This result easily extends to the density function, the survival function, the raw moments, and the expectation as these are all linear mappings of the distribution function. We note that this is not true for central moments like the variance, and conditional measures like the hazard rate function. In the case of variance it is easily seen as

[7]This in particular lays out a way to simulate from a mixture distribution that makes use of efficient simulation schemes that may exist for the component distributions.

$$\text{Var}[N_I] = \text{E}[\text{Var}[N_I|I]] + \text{Var}[\text{E}[N_I|I]] = \sum_{i=1}^{k} \alpha_i \text{Var}[N_i] + \text{Var}[\text{E}[N_I|I]]. \quad (2.12)$$

Appendix **??** provides additonal background about this important expression.

Example 2.6.1. Actuarial Exam Question. In a certain town the number of common colds an individual will get in a year follows a Poisson distribution that depends on the individual's age and smoking status. The distribution of the population and the mean number of colds are as follows:

	Proportion of population	Mean number of colds
Children	0.3	3
Adult Non-Smokers	0.6	1
Adult Smokers	0.1	4

Table 2.3 : The distribution of the population and the mean number of colds

1. Calculate the probability that a randomly drawn person has 3 common colds in a year.
2. Calculate the conditional probability that a person with exactly 3 common colds in a year is an adult smoker.

Show Example Solution

Solution.

1. Using Law of Total Probability, we can write the required probability as $\Pr(N_I = 3)$, with I denoting the group of the randomly selected individual with $1, 2$ and 3 signifying the groups Children, Adult Non-Smoker, and Adult Smoker, respectively. Now by conditioning we get

$$\Pr(N_I = 3) = 0.3 \cdot \Pr(N_1 = 3) + 0.6 \cdot \Pr(N_2 = 3) + 0.1 \cdot \Pr(N_3 = 3),$$

   with N_1, N_2 and N_3 following Poisson distributions with means $3, 1$, and 4, respectively. Using the above, we get $\Pr(N_I = 3) \sim 0.1235$
2. The conditional probability of event A given event B, $\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}$. The required conditional probability in this problem can then be written as $\Pr(I = 3|N_I = 3)$, which equals

$$\Pr(I = 3|N_I = 3) = \frac{\Pr(I = 3, N_3 = 3)}{\Pr(N_I = 3)} \sim \frac{0.1 \times 0.1954}{0.1235} \sim 0.1581.$$

In the above example, the number of subgroups k was equal to three. In general, k can be any natural number, but when k is large it is parsimonious from a

modeling point of view to take the following infinitely many subgroup approach. To motivate this approach, let the i-th subgroup be such that its component distribution F_i is given by $G_{\tilde{\theta}_i}$, where $G_\cdot$ is a parametric family of distributions with parameter space $\Theta \subseteq \mathbb{R}^d$. With this assumption, the distribution function F of a randomly drawn observation from the population is given by

$$F(x) = \sum_{i=1}^{k} \alpha_i G_{\tilde{\theta}_i}(x), \quad \forall x \in \mathbb{R}.$$

which can be alternately written as

$$F(x) = \mathrm{E}[G_{\tilde{\vartheta}}(x)], \quad \forall x \in \mathbb{R},$$

where $\tilde{\vartheta}$ takes values $\tilde{\theta}_i$ with probability α_i, for $i = 1, \ldots, k$. The above makes it clear that when k is large, one could model the above by treating $\tilde{\vartheta}$ as continuous random variable.

To illustrate this approach, suppose we have a population of drivers with the distribution of claims for an individual driver being distributed as a Poisson. Each person has their own (personal) expected number of claims λ - smaller values for good drivers, and larger values for others. There is a distribution of λ in the population; a popular and convenient choice for modeling this distribution is a gamma distribution with parameters (α, θ). With these specifications it turns out that the resulting distribution of N, the claims of a randomly chosen driver, is a negative binomial with parameters $(r = \alpha, \beta = \theta)$. This can be shown in many ways, but a straightforward argument is as follows:

$$\begin{aligned}\Pr(N = k) &= \int_0^\infty \frac{e^{-\lambda}\lambda^k}{k!} \frac{\lambda^{\alpha-1}e^{-\lambda/\theta}}{\Gamma(\alpha)\theta^\alpha} \mathrm{d}\lambda = \frac{1}{k!\Gamma(\alpha)\theta^\alpha} \int_0^\infty \lambda^{\alpha+k-1} e^{-\lambda(1+1/\theta)} \mathrm{d}\lambda = \frac{\Gamma(\alpha+k)}{k!\Gamma(\alpha)\theta^\alpha(1+1/\theta)^{\alpha+k}} \\ &= \binom{\alpha+k-1}{k} \left(\frac{1}{1+\theta}\right)^\alpha \left(\frac{\theta}{1+\theta}\right)^k, \quad k = 0, 1, \ldots\end{aligned}$$

Note that the above derivation implicitly uses the following:

$$f_{N|\Lambda=\lambda}(N = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \geq 0; \quad \text{and} \quad f_\Lambda(\lambda) = \frac{\lambda^{\alpha-1}e^{-\lambda/\theta}}{\Gamma(\alpha)\theta^\alpha}, \quad \lambda > 0.$$

It is worth mentioning that by considering mixtures of a parametric class of distributions we increase the richness of the class. This expansion of distributions results in the mixture class being able to cater well to more applications that the parametric class we started with. Mixture modeling is a very important modeling technique in insurance applications, and later chapters will cover more aspects of this modeling technique.

Example 2.6.2. Suppose that $N|\Lambda \sim$ Poisson(Λ) and that $\Lambda \sim$ gamma with mean of 1 and variance of 2. Determine the probability that $N = 1$.

Show Example Solution

Solution. For a gamma distribution with parameters (α, θ), we have that the mean is $\alpha\theta$ and the variance is $\alpha\theta^2$. Using these expressions we have

$$\alpha = \frac{1}{2} \text{ and } \theta = 2.$$

Now, one can directly use the above result to conclude that N is distributed as a negative binomial with $r = \alpha = \frac{1}{2}$ and $\beta = \theta = 2$. Thus

$$\begin{aligned}
\Pr(N = 1) &= \binom{1 + r - 1}{1} \left(\frac{1}{(1+\beta)^r}\right) \left(\frac{\beta}{1+\beta}\right)^1 \\
&= \binom{1 + \frac{1}{2} - 1}{1} \frac{1}{(1+2)^{1/2}} \left(\frac{2}{1+2}\right)^1 \\
&= \frac{1}{3^{3/2}} = 0.19245.
\end{aligned}$$

Show Quiz Solution

2.7 Goodness of Fit

In this section, you learn how to:

- Calculate a goodness of fit statistic to compare a hypothesized discrete distribution to a sample of discrete observations
- Compare the statistic to a reference distribution to assess the adequacy of the fit

---

In the above we have discussed three basic frequency distributions, along with their extensions through zero modification/truncation and by looking at mixtures of these distributions. Nevertheless, these classes still remain parametric and hence by their very nature a small subset of the class of all possible frequency distributions (i.e. the set of distributions on non-negative integers.) Hence, even though we have talked about methods for estimating the unknown parameters, the fitted distribution will not be a good representation of the underlying distribution if the latter is **far** from the class of distribution used for modeling. In fact, it can be shown that the maximum likelihood estimator will converge to a value such that the corresponding distribution will be a Kullback-Leibler projection of the underlying distribution on the class of distributions used for

modeling. Below we present one testing method - Pearson's chi-square statistic - to check for the goodness of fitThe goodness of fit of a statistical model describes how well it fits a set of observations. of the fitted distribution. For more details on the Pearson's chi-square test, at an introductory mathematical statistics level, we refer the reader to Section 9.1 of (**?**).

In 1993, a portfolio of $n = 7,483$ automobile insurance policies from a major Singaporean insurance company had the distribution of auto accidents per policyholder as given in Table 2.4.

Count (k)	0	1	2	3	4	Total
No. of Policies with k accidents (m_k)	6,996	455	28	4	0	7483

Table 2.4 : Singaporean Automobile Accident Data

If we a fit a Poisson distribution, then the MLE for λ, the Poisson mean, is the sample mean which is given by

$$\overline{N} = \frac{0 \cdot 6996 + 1 \cdot 455 + 2 \cdot 28 + 3 \cdot 4 + 4 \cdot 0}{7483} = 0.06989.$$

Now if we use Poisson ($\hat{\lambda}_{MLE}$) as the fitted distribution, then a tabular comparison of the fitted counts and observed counts is given by Table 2.5 below, where $\hat{p}_k$ represents the estimated probabilities under the fitted Poisson distribution.

Count (k)	Observed (m_k)	Fitted Counts Using Poisson ($n\hat{p}_k$)
0	6,996	6,977.86
1	455	487.70
2	28	17.04
3	4	0.40
≥ 4	0	0.01
Total	7,483	7,483.00

Table 2.5 : Comparison of Observed to Fitted Counts: Singaporean Auto Data

While the fit seems reasonable, a tabular comparison falls short of a statistical test of the hypothesis that the underlying distribution is indeed Poisson. The Pearson's chi-square statistic is a goodness of fit statistical measure that can be used for this purpose. To explain this statistic let us suppose that a dataset of size n is grouped into k cells with m_k/n and $\hat{p}_k$, for $k = 1 \ldots, K$ being the observed and estimated probabilities of an observation belonging to the k-th cell, respectively. The Pearson's chi-square test statistic is then given by

$$\sum_{k=1}^{K} \frac{(m_k - n\widehat{p}_k)^2}{n\widehat{p}_k}.$$

The motivation for the above statistic derives from the fact that

$$\sum_{k=1}^{K} \frac{(m_k - np_k)^2}{np_k}$$

has a limiting chi-square distributionthe chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables with $K-1$ degrees of freedom if $p_k,\ k = 1, \ldots, K$ are the true cell probabilities. Now suppose that only the summarized data represented by $m_k,\ k = 1, \ldots, K$ is available. Further, if p_k's are functions of s parameters, replacing p_k's by any efficiently estimated probabilities $\widehat{p}_k$'s results in the statistic continuing to have a limiting chi-square distribution but with degrees of freedom given by $K - 1 - s$. Such efficient estimates can be derived for example by using the MLE method (with a multinomial likelihood) or by estimating the s parameters which minimizes the Pearson's chi-square statistic above. For example, the `R` code below does calculate an estimate for λ doing the latter and results in the estimate 0.06623153, close but different from the MLE of λ using the full data:

```
m<-c(6996,455,28,4,0);
op<-m/sum(m);
g<-function(lam){sum((op-c(dpois(0:3,lam),1-ppois(3,lam)))^2)};
optim(sum(op*(0:4)),g,method="Brent",lower=0,upper=10)$par
```

When one uses the full data to estimate the probabilities, the asymptotic distribution is in between chi-square distributions with parameters $K-1$ and $K-1-s$. In practice it is common to ignore this subtlety and assume the limiting chi-square has $K-1-s$ degrees of freedom. Interestingly, this practical shortcut works quite well in the case of the Poisson distribution.

For the Singaporean auto data the Pearson's chi-square statistic equals 41.98 using the full data MLE for λ. Using the limiting distribution of chi-square with $5-1-1=3$ degrees of freedom, we see that the value of 41.98 is way out in the tail (99-th percentile is below 12). Hence we can conclude that the Poisson distribution provides an inadequate fit for the data.

In the above, we started with the cells as given in the above tabular summary. In practice, a relevant question is how to define the cells so that the chi-square distribution is a good approximation to the finite sample distribution of the statistic. A rule of thumb is to define the cells in such a way to have at least 80%, if not all, of the cells having expected counts greater than 5. Also, it is clear that a larger number of cells results in a higher power of the test, and hence a simple rule of thumb is to maximize the number of cells such that each cell has at least 5 observations.

Show Quiz Solution

2.8 Exercises

Theoretical Exercises

Exercise 2.1. Derive an expression for $p_N(\cdot)$ in terms of $F_N(\cdot)$ and $S_N(\cdot)$.

Exercise 2.2. A measure of center of location must be **equi-variant** with respect to shifts, or location transformations. In other words, if N_1 and N_2 are two random variables such that $N_1 + c$ has the same distribution as N_2, for some constant c, then the difference between the measures of the center of location of N_2 and N_1 must equal c. Show that the mean satisfies this property.

Exercise 2.3. Measures of dispersion should be invariant with respect to shifts and scale equi-variant. Show that standard deviation satisfies these properties by doing the following:

- Show that for a random variable N, its standard deviation equals that of $N + c$, for any constant c.
- Show that for a random variable N, its standard deviation equals $1/c$ times that of cN, for any positive constant c.

Exercise 2.4. Let N be a random variable with probability mass function given by

$$p_N(k) := \begin{cases} \left(\frac{6}{\pi^2}\right)\left(\frac{1}{k^2}\right), & k \geq 1; \\ 0, & \text{otherwise.} \end{cases}$$

Show that the mean of N is ∞.

Exercise 2.5. Let N be a random variable with a finite second moment. Show that the function $\psi(\cdot)$ defined by

$$\psi(x) := \mathrm{E}(N - x)^2. \quad x \in \mathbb{R}$$

is minimized at μ_N without using calculus. Also, give a proof of this fact using derivatives. Conclude that the minimum value equals the variance of N.

Exercise 2.6. Derive the first two central moments of the $(a, b, 0)$ distributions using the methods mentioned below:

- For the binomial distribution, derive the moments using only its pmf, then its mgf, and then its pgf.
- For the Poisson distribution, derive the moments using only its mgf.
- For the negative binomial distribution, derive the moments using only its pmf, and then its pgf.

Exercise 2.7. Let N_1 and N_2 be two independent Poisson random variables with means λ_1 and λ_2, respectively. Identify the conditional distribution of N_1 given $N_1 + N_2$.

Exercise 2.8. (**Non-Uniqueness of the MLE**) Consider the following parametric family of densities indexed by the parameter p taking values in $[0, 1]$:

$$f_p(x) = p \cdot \phi(x + 2) + (1 - p) \cdot \phi(x - 2), \quad x \in \mathbb{R},$$

where $\phi(\cdot)$ represents the standard normal density.

- Show that for all $p \in [0, 1]$, $f_p(\cdot)$ above is a valid density function.
- Find an expression in p for the mean and the variance of $f_p(\cdot)$.
- Let us consider a sample of size one consisting of x. Show that when x equals 0, the set of maximum likelihood estimates for p equals $[0, 1]$; also show that the MLE is unique otherwise.

Exercise 2.9. Graph the region of the plane corresponding to values of (a, b) that give rise to valid $(a, b, 0)$ distributions. Do the same for $(a, b, 1)$ distributions.

Exercise 2.10. (**Computational Complexity**) For the $(a, b, 0)$ class of distributions, count the number of basic mathematical operations (addition, subtraction, multiplication, division) needed to compute the n probabilities $p_0 \dots p_{n-1}$ using the recurrence relationship. For the negative binomial distribution with non-integer r, count the number of such operations. What do you observe?

Exercise 2.11. (** **) Using the development of Section 2.3 rigorously show that not only does the recurrence (**??**) tie the binomial, the Poisson and the negative binomial distributions together, but that it also characterizes them.

Exercises with a Practical Focus

Exercise 2.12. Actuarial Exam Question. You are given:

1. p_k denotes the probability that the number of claims equals k for $k = 0, 1, 2, \dots$
2. $\frac{p_n}{p_m} = \frac{m!}{n!}, m \geq 0, n \geq 0$

Using the corresponding zero-modified claim count distribution with $p_0^M = 0.1$, calculate p_1^M.

Exercise 2.13. Actuarial Exam Question. During a one-year period, the number of accidents per day was distributed as follows:

No. of Accidents	0	1	2	3	4	5
No. of Days	209	111	33	7	5	2

You use a chi-square test to measure the fit of a Poisson distribution with mean 0.60. The minimum expected number of observations in any group should be 5. The maximum number of groups should be used. Determine the value of the chi-square statistic.

A discrete probability distribution has the following properties

$$\Pr(N=k) = \left(\frac{3k+9}{8k}\right)\Pr(N=k-1), \quad k=1,2,3,\ldots$$

Determine the value of $\Pr(N=3)$. (Ans: 0.1609)

Additional Exercises

Here are a set of exercises that guide the viewer through some of the theoretical foundations of **Loss Data Analytics**. Each tutorial is based on one or more questions from the professional actuarial examinations – typically the Society of Actuaries Exam C.

Frequency Distribution Guided Tutorials

2.9 Further Resources and Contributors

Appendix Chapter **??** gives a general introduction to maximum likelihood theory regarding estimation of parameters from a parametric family. Appendix Chapter **??** gives more specific examples and expands some of the concepts.

Contributors

- **N.D. Shyamalkumar**, The University of Iowa, and **Krupa Viswanathan**, Temple University, are the principal authors of the initial version of this chapter. Email: shyamal-kumar@uiowa.edu for chapter comments and suggested improvements.
- Chapter reviewers include: Paul Johnson, Hirokazu (Iwahiro) Iwasawa, Rajesh Sahasrabuddhe, Michelle Xia.

2.9.1 TS 2.A. R Code for Plots

Code for Figure ??:

Show R Code

```
likbinm<-function(m){
  # binomial likelihood maximized w.r.t. p
  prod((dbinom(x,m,mean(x)/m)))
}

liknbinm<-function(r){
  # negative binomial likelihood maximized w.r.t. beta
```

```
  prod(dnbinom(x,r,1-mean(x)/(mean(x)+r)))
}

# Data Matrix; Three Samples, one in each Column;
# First Sample has Var<Mean
# Second Sample has Var=Mean
# Third Sample has Var>Mean

  X<-cbind(c(2,5,6,8,9)+2,c(2,5,6,8,9),c(2,3,6,8,9));

# Used for creating the labels in the z matrix
  ord_char<-c("<","=",">");

# Empty matrices;
  Y<-matrix(1,ncol=2,nrow=0);
  Z<-matrix(1,ncol=2,nrow=0);

for (i in (1:3)) {
  # Work with data in the i-th sample
    x<-X[,i];

  # Binomial Likelihood
      # Interval of n values covering the MLE
        n<-(9:100);
      # Evaluating the Likelihood at various values of n
        ll<-sapply(n,likbinm);
      # Finding the MLE of n
        n[ll==max(ll[!is.na(ll)])]
      # Storing the data and the labels
        Y<-rbind(Y,cbind(n,ll));
        Z<-rbind(Z,cbind(rep(paste("$\\hat{\\sigma}^2",ord_char[i],"\\hat{\\mu}$"),length(n)),rep

  # Negative Binomial Likelihood
    # Interval of r values
      r<-(1:100);
    # Evaluating the Likelihood at various values of r
      ll<-sapply(r,liknbinm);
    # Finding the MLE of r
      ll[is.na(ll)]=0;
      r[ll==max(ll[!is.na(ll)])];
    # Storing the data and the labels
      Y<-rbind(Y,cbind(r,ll));
      Z<-rbind(Z,cbind(rep(paste("$\\hat{\\sigma}^2",ord_char[i],"\\hat{\\mu}$"),length(r)),rep('

  # Poisson Likelihood
```

```
    # Storing the data and the labels
    # In the Poisson case MLE is the sample mean
      Y<-rbind(Y,cbind(r,rep(prod(dpois(x,mean(x))),length(r))));
      Z<-rbind(Z,cbind(rep(paste("$\\hat{\\sigma}^2",ord_char[i],"\\hat{\\mu}$"),lengt
}

  # Assigning Column Names
    colnames(Y)<-c("x","lik");
    colnames(Z)<-c("dataset","Distribution");
  # Creating a Dataframe for using ggplot
    dy<-cbind(data.frame(Y),data.frame(Z));


library(tikzDevice);
library(ggplot2);
options(tikzMetricPackages = c("\\usepackage[utf8]{inputenc}","\\usepackage[T1]{fontenc
                               "\\usepackage{amssymb}","\\usepackage{amsmath}","\\usepa
tikz(file = "plot_test_2.tex", width = 6.25, height = 6.25);
ggplot(data=dy,aes(x=x,y=lik,col=Distribution)) + geom_point(size=0.25) + facet_grid(da
  labs(x="m/r",y="Likelihood",title="");
dev.off();
```

Code for Figure ??:

Show R Code

```
likm<-function(m){
  prod((dbinom(x,m,mean(x)/m)))
}
x<-c(2,2,2,4,5);
n<-(5:100);
# Computing the Likelihood
ll<-sapply(n,likm);
# Computing the MLE
n[ll==max(ll)]
# Storing the Likelihood Curve
y<-cbind(n,ll);

# Second Dataset
x<-c(2,2,2,4,6);
ll<-sapply(n,likm);
n[ll==max(ll)]
y<-cbind(y,ll);

# Third Dataset
```

```
x<-c(2,2,2,4,7);
ll<-sapply(n,likm);
n[ll==max(ll)]
y<-cbind(y,ll);

colnames(y)<-c("m","$\\tilde{x}=(2,2,2,4,5)$","$\\tilde{x}=(2,2,2,4,6)$","$\\tilde{x}=(2,2,2,4,7)
dy<-data.frame(y);
library(tikzDevice);
library(ggplot2);
options(tikzMetricPackages = c("\\usepackage[utf8]{inputenc}","\\usepackage[T1]{fontenc}", "\\use
                               "\\usepackage{amssymb}","\\usepackage{amsmath}","\\usepackage[acti
tikz(file = "plot_test.tex", width = 6.25, height = 3.125);
ggplot(dy) +
  geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.5..),shape="$\\tilde{x}=(2,2,2,4,5):\\hat{m}=7$"),s
  geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.6..),shape="$\\tilde{x}=(2,2,2,4,6):\\hat{m}=18$"),
  geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.7..),shape="$\\tilde{x}=(2,2,2,4,7):\\hat{m}=\\inft
  geom_point(aes(x=c(7),y=dy$X..tilde.x...2.2.2.4.5..[3],colour="$\\hat{m}$",shape="$\\tilde{x}=(
  geom_point(aes(x=c(18),y=dy$X..tilde.x...2.2.2.4.6..[14],colour="$\\hat{m}$",shape="$\\tilde{x}
  labs(x="m",y="$L(m,\\overline{x}/m)$",title="MLE for $m$: Non-Robustness of MLE ");
dev.off();
```

Chapter 3

Modeling Loss Severity

Chapter Preview. The traditional loss distribution approach to modeling aggregate lossesaggregate claims, or total claims observed in the time period starts by separately fitting a frequency distribution to the number of losses and a severity distribution to the size of losses. The estimated aggregate loss distribution combines the loss frequency distribution and the loss severity distribution by convolution. Discrete distributions often referred to as counting or frequency distributions were used in Chapter **??** to describe the number of events such as number of accidents to the driver or number of claims to the insurer. Lifetimes, asset values, losses and claim sizes are usually modeled as continuous random variables and as such are modeled using continuous distributions, often referred to as loss or severity distributions. A mixture distributionThe probability distribution of a random variable that is derived from a collection of other random variables as follows: first, a random variable is selected by chance from the collection according to given probabilities of selection, and then the value of the selected random variable is realized is a weighted combination of simpler distributions that is used to model phenomenon investigated in a heterogeneous population, such as modelling more than one type of claims in liability insuranceinsurance that compensates an insured for loss due to legal liability towards others (small frequent claims and large relatively rare claims). In this chapter we explore the use of continuous as well as mixture distributions to model the random size of loss. We present key attributes that characterize continuous models and means of creating new distributions from existing ones. We also explore the effect of coverage modifications, which change the conditions that trigger a payment, such as applying deductibles, limits, or adjusting for inflation, on the distribution of individual loss amounts. The frequency distributions from Chapter **??** will be combined with the ideas from this chapter to describe the aggregate losses over the whole portfolio in Chapter **??**.

3.1 Basic Distributional Quantities

In this section, you learn how to define some basic distributional quantities:

- moments,
- percentiles, and
- generating functions.

3.1.1 Moments

Let X be a continuous random variablerandom variable which can take infinitely many values in its specified domain with probability density function $f_X(x)$. The k-th raw momentthe kth moment of a random variable X is the average (expected) value of Xk of X, denoted by μ'_k, is the expected valueaverage of the k-th power of X, provided it exists. The first raw moment μ'_1 is the mean of X usually denoted by μ. The formula for μ'_k is given as

$$\mu'_k = \mathrm{E}\left(X^k\right) = \int_0^\infty x^k f_X(x)\, dx.$$

The support of the random variable X is assumed to be nonnegative since actuarial phenomena are rarely negative. An easy integration by parts shows that the raw moments for nonnegative variables can also be computed using

$$\mu'_k = \int_0^\infty k\ x^{k-1}\left[1 - F_X(x)\right] dx,$$

that is based on the survival function, denoted as $S_X(x) = 1 - F_X(x)$. This formula is particularly useful when $k = 1$.

The k-th central momentthe kth central moment of a random variable X is the expected value of (X-its mean)k of X, denoted by μ_k, is the expected value of the k-th power of the deviation of X from its mean μ. The formula for μ_k is given as

$$\mu_k = \mathrm{E}\left[(X-\mu)^k\right] = \int_0^\infty (x-\mu)^k f_X(x)\, dx.$$

The second central moment μ_2 defines the variancesecond central moment of a random variable X, measuring the expected squared deviation of between the variable and its mean of X, denoted by σ^2. The square root of the variance is the standard deviationthe square-root of variance σ.

From a classical perspective, further characterization of the shape of the distribution includes its degree of symmetry as well as its flatness compared to the normal distribution. The ratio of the third central moment to the cube of the

standard deviation $\left(\mu_3/\sigma^3\right)$ defines the coefficient of skewnessmeasure of the symmetry of a distribution, 3rd central moment/standard deviatiôn3 which is a measure of symmetry. A positive coefficient of skewness indicates that the distribution is skewed to the right (positively skewed). The ratio of the fourth central moment to the fourth power of the standard deviation $\left(\mu_4/\sigma^4\right)$ defines the coefficient of kurtosismeasure of the peaked-ness of a distribution, 4th central moment/standard deviatiôn4. The normal distribution has a coefficient of kurtosis of 3. Distributions with a coefficient of kurtosis greater than 3 have heavier tails and higher peak than the normal, whereas distributions with a coefficient of kurtosis less than 3 have lighter tails and are flatter. Section **??** describes the tails of distributions from an insurance and actuarial perspective.

Example 3.1.1. Actuarial Exam Question. Assume that the rv X has a gamma distribution with mean 8 and skewness 1. Find the variance of X. (Hint: The gamma distribution is reviewed in Section **??**.)

Show Example Solution

Solution. The probability density function of X is given by

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

for $x > 0$. For $\alpha > 0$, the k-th raw moment is

$$\mu'_k = \mathrm{E}\left(X^k\right) = \int_0^\infty \frac{1}{\Gamma(\alpha)\,\theta^\alpha} x^{k+\alpha-1} e^{-x/\theta} dx = \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)} \theta^k$$

Given $\Gamma(r+1) = r\Gamma(r)$ and $\Gamma(1) = 1$, then $\mu'_1 = \mathrm{E}(X) = \alpha\theta$, $\mu'_2 = \mathrm{E}\left(X^2\right) = (\alpha+1)\,\alpha\theta^2$, $\mu'_3 = \mathrm{E}\left(X^3\right) = (\alpha+2)\,(\alpha+1)\,\alpha\theta^3$, and $\mathrm{Var}(X) = (\alpha+1)\alpha\theta^2 - (\alpha\theta)^2 = \alpha\theta^2$.

$$\text{Skewness} = \frac{\mathrm{E}\left[(X-\mu'_1)^3\right]}{\mathrm{Var}(X)^{3/2}} = \frac{\mu'_3 - 3\mu'_2\mu'_1 + 2{\mu'_1}^3}{\mathrm{Var}(X)^{3/2}} = \frac{(\alpha+2)\,(\alpha+1)\,\alpha\theta^3 - 3\,(\alpha+1)\,\alpha^2\theta^3 + 2\alpha^3\theta^3}{(\alpha\theta^2)^{3/2}} = \frac{2}{\alpha^{1/2}} = 1.$$

Hence, $\alpha = 4$. Since, $\mathrm{E}(X) = \alpha\theta = 8$, then $\theta = 2$ and finally, $\mathrm{Var}(X) = \alpha\theta^2 = 16$.

3.1.2 Quantiles

Quantiles can also be used to describe the characteristics of the distribution of X. When the distribution of X is continuous, for a given fraction $0 \le p \le 1$ the corresponding quantile is the solution of the equation

$$F_X(\pi_p) = p.$$

For example, the middle point of the distribution, $\pi_{0.5}$, is the median50th percentile of a definition, or middle value where half of the distribution lies below. A percentilethe pth percentile of a random variable X is the smallest value x_p such that the probability of not exceeding it is p% is a type of quantile; a $100p$ percentile is the number such that $100 \times p$ percent of the data is below it.

Example 3.1.1. Actuarial Exam Question. Let X be a continuous random variable with density function $f_X(x) = \theta e^{-\theta x}$, for $x > 0$ and 0 elsewhere. If the median of this distribution is $\frac{1}{3}$, find θ.

Show Example Solution

Solution.

The distribution function is $F_X(x) = 1 - e^{-\theta x}$. So, $F_X(\pi_{0.5}) = 1 - e^{-\theta \pi_{0.5}} = 0.5$. As, $\pi_{0.5} = \frac{1}{3}$, we have $F_X\left(\frac{1}{3}\right) = 1 - e^{-\theta/3} = 0.5$ and $\theta = 3 \ln 2$.

Section **??** will extend the definition of quantiles to include distributions that are discrete, continuous, or a hybrid combination.

3.1.3 Moment Generating Function

The moment generating function, denoted by $M_X(t)$ uniquely characterizes the distribution of X. While it is possible for two different distributions to have the same moments and yet still differ, this is not the case with the moment generating function. That is, if two random variables have the same moment generating function, then they have the same distribution. The moment generating function is given by

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \int_0^\infty e^{\mathrm{tx}} f_X(x)\, dx$$

for all t for which the expected value exists. The moment generating is a real function whose k-th derivative at zero is equal to the k-th raw moment of X. In symbols, this is

$$\frac{d^k}{dt^k} M_X(t)\bigg|_{t=0} = \mathrm{E}\left(X^k\right).$$

Example 3.1.3. Actuarial Exam Question. The random variable X has an exponential distribution with mean $\frac{1}{b}$. It is found that $M_X\left(-b^2\right) = 0.2$. Find b. (Hint: The exponential is a special case of the gamma distribution which is reviewed in Section **??**.)

Show Example Solution

Solution.

With X having an exponential distribution with mean $\frac{1}{b}$, we have that

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \int_0^\infty e^{\mathrm{tx}} b e^{-bx} dx = \int_0^\infty b e^{-x(b-t)} dx = \frac{b}{(b-t)}.$$

Then,

$$M_X\left(-b^2\right) = \frac{b}{(b+b^2)} = \frac{1}{(1+b)} = 0.2.$$

Thus, $b = 4$.

Example 3.1.4. Actuarial Exam Question. Let $X_1, \ldots, X_n$ be independenttwo variables are independent if conditional information given about one variable provides no information regarding the other variable random variables, where X_i has a gamma distribution with parameters α_i and θ. Find the distribution of $S = \sum_{i=1}^n X_i$, the mean $\mathrm{E}(S)$, and the variance $\mathrm{Var}(S)$.

Show Example Solution

Solution.

The moment generating function of S is

$$M_S(t) = \mathrm{E}\left(e^{\mathrm{tS}}\right) = \mathrm{E}\left(e^{t\sum_{i=1}^n X_i}\right) = \mathrm{E}\left(\prod_{i=1}^n e^{tX_i}\right).$$

Using independence, we get

$$M_S(t) = \prod_{i=1}^n \mathrm{E}\left(e^{tX_i}\right) = \prod_{i=1}^n M_{X_i}(t).$$

The moment generating function of the gamma distribution X_i is $M_{X_i}(t) = (1-\theta)^{\alpha_i}$. . Then,

$$M_S(t) = \prod_{i=1}^n (1-\theta t)^{-\alpha_i} = (1-\theta t)^{-\sum_{i=1}^n \alpha_i}.$$

This indicates that the distribution of S is gamma with parameters $\sum_{i=1}^n \alpha_i$ and θ.

This is a demonstration of how we can use the uniqueness property of the moment generating function to determine the probability distribution of a random variable.

We can find the mean and variance from the properties of the gamma distribution. Alternatively, by finding the first and second derivatives of $M_S(t)$ at zero,

we can show that $\mathrm{E}(S) = \left.\frac{\partial M_S(t)}{\partial t}\right|_{t=0} = \alpha\theta$ where $\alpha = \sum_{i=1}^{n} \alpha_i$, and

$$\mathrm{E}\left(S^2\right) = \left.\frac{\partial^2 M_S(t)}{\partial t^2}\right|_{t=0} = (\alpha+1)\,\alpha\theta^2.$$

Hence, $\mathrm{Var}(S) = \alpha\theta^2$.

One can also use the moment generating function to compute the probability generating function

$$P_X(z) = \mathrm{E}\left(z^X\right) = M_X(\ln z).$$

As introduced in Section **??**, the probability generating function is more useful for discrete rvs.

Show Quiz Solution

3.2 Continuous Distributions for Modeling Loss Severity

In this section, you learn how to define and apply four fundamental severity distributions:

- gamma,
- Pareto,
- Weibull, and
- generalized beta distribution of the second kind.

3.2.1 Gamma Distribution

Recall that the traditional approach in modelling losses is to fit separate models for claim frequency and claim severity. When frequency and severity are modeled separately it is common for actuaries to use the Poisson distribution (introduced in Section **??**) for claim count and the gamma distribution to model severity. An alternative approach for modelling losses that has recently gained popularity is to create a single model for pure premium (average claim cost) that will be described in Chapter **??**.

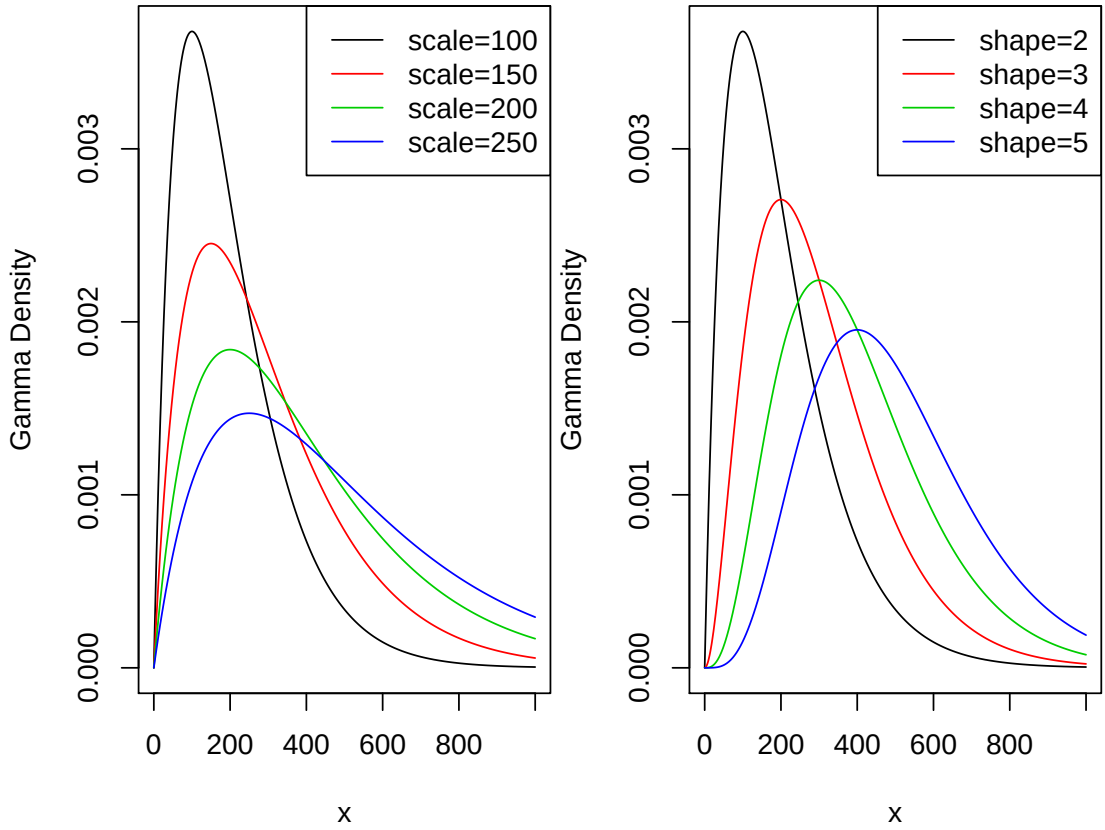


Figure 3.1: Gamma Densities. The left-hand panel is with shape=2 and Varying Scale. The right-hand panel is with scale=100 and Varying Shape.

The continuous variable X is said to have the gamma distribution with shape parameter α and scale parameter θ if its probability density function is given by

$$f_X(x) = \frac{(x/\theta)^{\alpha}}{x\,\Gamma(\alpha)} \exp(-x/\theta) \quad \text{for } x > 0.$$

Note that $\alpha > 0,\ \theta > 0$.

The two panels in Figure **??** demonstrate the effect of the scale and shape parameters on the gamma density function.

R Code for Gamma Density Plots

```
par(mfrow=c(1, 2), mar = c(4, 4, .1, .1))

# Varying Scale Gamma Densities
scaleparam <- seq(100, 250, by = 50)
```

```
shapeparam <- 2:5
x <- seq(0, 1000, by = 1)
fgamma <- dgamma(x, shape = 2, scale = scaleparam[1])
plot(x, fgamma, type = "l", ylab = "Gamma Density")
for(k in 2:length(scaleparam)){
  fgamma <- dgamma(x,shape = 2, scale = scaleparam[k])
  lines(x,fgamma, col = k)
}
legend("topright", c("scale=100", "scale=150", "scale=200", "scale=250"), lty=1, col =

# Varying Shape Gamma Densities
fgamma <- dgamma(x, shape = shapeparam[1], scale = 100)
plot(x, fgamma, type = "l", ylab = "Gamma Density")
for(k in 2:length(shapeparam)){
  fgamma <- dgamma(x,shape = shapeparam[k], scale = 100)
  lines(x,fgamma, col = k)
}
legend("topright", c("shape=2", "shape=3", "shape=4", "shape=5"), lty=1, col = 1:4)
```

When $\alpha = 1$ the gamma reduces to an exponential distributiona single parameter continous probability distribution that is defined by its rate parameter and when $\alpha = \frac{n}{2}$ and $\theta = 2$ the gamma reduces to a chi-square distributionthe chi-squared distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables with n degrees of freedom. As we will see in Section **??**, the chi-square distribution is used extensively in statistical hypothesis testing.

The distribution function of the gamma model is the incomplete gamma function, denoted by $\Gamma\left(\alpha; \frac{x}{\theta}\right)$, and defined as

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

$\alpha > 0$, $\theta > 0$. For an integer α, it can be written as $\Gamma\left(\alpha; \frac{x}{\theta}\right) = 1 - e^{-x/\theta} \sum_{k=0}^{\alpha-1} \frac{(x/\theta)^k}{k!}$.

The k-th moment of the gamma distributed random variable for any positive k is given by

$$\mathrm{E}\left(X^k\right) = \theta^k \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)}.$$

The mean and variance are given by $\mathrm{E}(X) = \alpha\theta$ and $\mathrm{Var}(X) = \alpha\theta^2$, respectively.

Since all moments exist for any positive k, the gamma distribution is considered a light tailed distributiona distribution with thinner tails than the benchmark exponential distribution, which may not be suitable for modeling risky assets as it will not provide a realistic assessment of the likelihood of severe losses.

3.2.2 Pareto Distribution

The Pareto distributiona heavy-tailed and positively skewed distribution with 2 parameters, named after the Italian economist Vilfredo Pareto (1843-1923), has many economic and financial applications. It is a positively skewed and heavy-tailed distribution which makes it suitable for modeling income, high-risk insurance claims and severity of large casualty losses. The survival function of the Pareto distribution which decays slowly to zero was first used to describe the distribution of income where a small percentage of the population holds a large proportion of the total wealth. For extreme insurance claims, the tail of the severity distribution (losses in excess of a threshold) can be modeled using a Generalized Pareto distribution.

The continuous variable X is said to have the Pareto distribution with shape parameter α and scale parameter θ if its pdf is given by

$$f_X(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} \quad x > 0,\ \alpha > 0,\ \theta > 0.$$

The two panels in Figure **??** demonstrate the effect of the scale and shape parameters on the Pareto density function.

R Code for Pareto Density Plots

```
library(VGAM)

par(mfrow=c(1, 2), mar = c(4, 4, .1, .1))

# Varying Shape Pareto Densities
x <- seq(1, 3000, by = 1)
scaleparam <- seq(2000, 3500, 500)
shapeparam <- 1:4

# varying the shape parameter
plot(x, dparetoII(x, loc=0, shape = shapeparam[1], scale = 2000), ylim=c(0,0.002),type = "l", yla
for(k in 2:length(shapeparam)){
  lines(x, dparetoII(x, loc=0, shape = shapeparam[k], scale = 2000), col = k)
}
legend("topright", c(expression(alpha~'=1'), expression(alpha~'=2'), expression(alpha~'=3'), expr

# Varying Scale Pareto Densities
plot(x, dparetoII(x, loc=0, shape = 3, scale = scaleparam[1]), type = "l", ylab = "Pareto density
for(k in 2:length(scaleparam)){
  lines(x, dparetoII(x, loc=0, shape = 3, scale = scaleparam[k]), col = k)
}
legend("topright", c(expression(theta~'=2000'), expression(theta~'=2500'), expression(theta~'=300
```

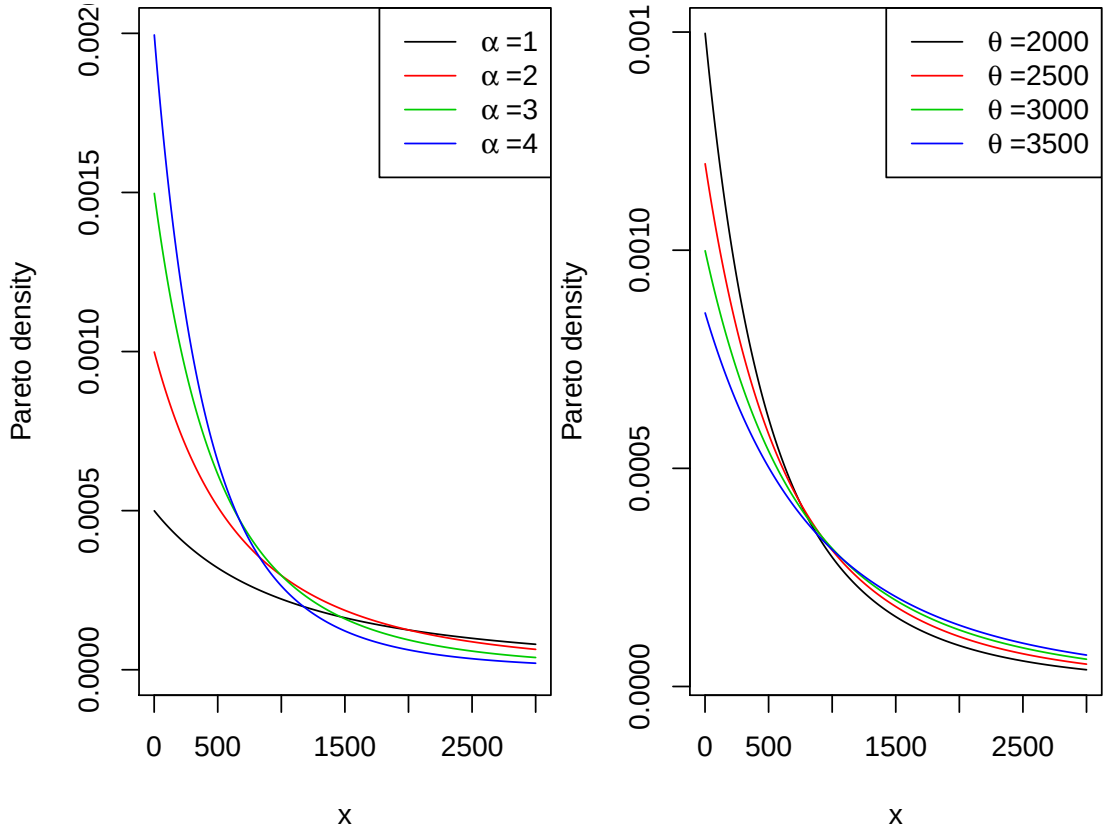


Figure 3.2: Pareto Densities. The left-hand panel is with scale=2000 and Varying Shape. The right-hand panel is with shape=3 and ,Varying Scale

The distribution function of the Pareto distribution is given by

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

It can be easily seen that the hazard functionratio of the probability density function and the survival function: f(x)/S(x), and represents an instantaneous probability within a small time frame of the Pareto distribution is a decreasing function in x, another indication that the distribution is heavy tailed. When the hazard function decreases over time the population dies off at a decreasing rate resulting in a heavier tail for the distribution. The hazard function reveals information about the tail distribution and is often used to model data distributions in survival analysis. The hazard function is defined as the instantaneous potential that the event of interest occurs within a very narrow time frame.

The k-th moment of the Pareto distributed random variable exists, if and only if, $\alpha > k$. If k is a positive integer then

$$\mathrm{E}\left(X^k\right) = \frac{\theta^k\ k!}{(\alpha-1)\cdots(\alpha-k)} \quad \alpha > k.$$

The mean and variance are given by

$$\mathrm{E}(X) = \frac{\theta}{\alpha-1} \quad \text{for } \alpha > 1$$

and

$$\mathrm{Var}(X) = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} \quad \text{for } \alpha > 2,$$

respectively.

Example 3.2.1. The claim size of an insurance portfolio follows the Pareto distribution with mean and variance of 40 and 1800 respectively. Find

The shape and scale parameters.

The 95-th percentile of this distribution.

Show Example Solution

Solution.

a. As, $X \sim Pa(\alpha, \theta)$, we have $\mathrm{E}(X) = \frac{\theta}{\alpha-1} = 40$ and $\mathrm{Var}(X) = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} = 1800$. By dividing the square of the first equation by the second we get $\frac{\alpha-2}{\alpha} = \frac{40^2}{1800}$. Thus, $\alpha = 18.02$ and $\theta = 680.72$.
b. The 95-th percentile, $\pi_{0.95}$, satisfies the equation

$$F_X(\pi_{0.95}) = 1 - \left(\frac{680.72}{\pi_{0.95}+680.72}\right)^{18.02} = 0.95.$$

Thus, $\pi_{0.95} = 122.96$.

3.2.3 Weibull Distribution

The Weibull distributiona positively skewed continuous distribution with 2 parameters that can have an increasing or decreasing hazard function depending on the shape parameter, named after the Swedish physicist Waloddi Weibull (1887-1979) is widely used in reliability, life data analysis, weather forecasts and general insurance claims. Truncated data arise frequently in insurance studies. The Weibull distribution has been used to model excess of loss treaty over automobile insurance as well as earthquake inter-arrival times.

The continuous variable X is said to have the Weibull distribution with shape parameter α and scale parameter θ if its probability density function is given by

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

The two panels Figure **??** demonstrate the effects of the scale and shape parameters on the Weibull density function.

R Code for Weibull Density Plots

```
par(mfrow=c(1, 2), mar = c(4, 4, .1, .1))

# Varying Scale Weibull Densities
z<- seq(0,400,by=1)
scaleparam <- seq(50,200,50)
shapeparam <- seq(1.5,3,0.5)
plot(z, dweibull(z, shape = 3, scale = scaleparam[1]), type = "l", ylab = "Weibull dens
for(k in 2:length(scaleparam)){
  lines(z,dweibull(z,shape = 3, scale = scaleparam[k]), col = k)}
legend("topright", c("scale=50", "scale=100", "scale=150", "scale=200"), lty=1, col = 

# Varying Shape Weibull Densities
plot(z, dweibull(z, shape = shapeparam[1], scale = 100), ylim=c(0,0.012), type = "l", 
for(k in 2:length(shapeparam)){
  lines(z,dweibull(z,shape = shapeparam[k], scale = 100), col = k)}
legend("topright", c("shape=1.5", "shape=2", "shape=2.5", "shape=3"), lty=1, col = 1:4)
```

The distribution function of the Weibull distribution is given by

$$F_X(x) = 1 - e^{-(x/\theta)^{\alpha}} \quad x>0,\ \alpha>0,\ \theta>0.$$

It can be easily seen that the shape parameter α describes the shape of the hazard function of the Weibull distribution. The hazard function is a decreasing function when $\alpha < 1$ (heavy tailed distribution), constant when $\alpha = 1$ and increasing when $\alpha > 1$ (light tailed distribution). This behavior of the hazard function makes the Weibull distribution a suitable model for a wide variety of phenomena such as weather forecasting, electrical and industrial engineering, insurance modeling, and financial risk analysis.

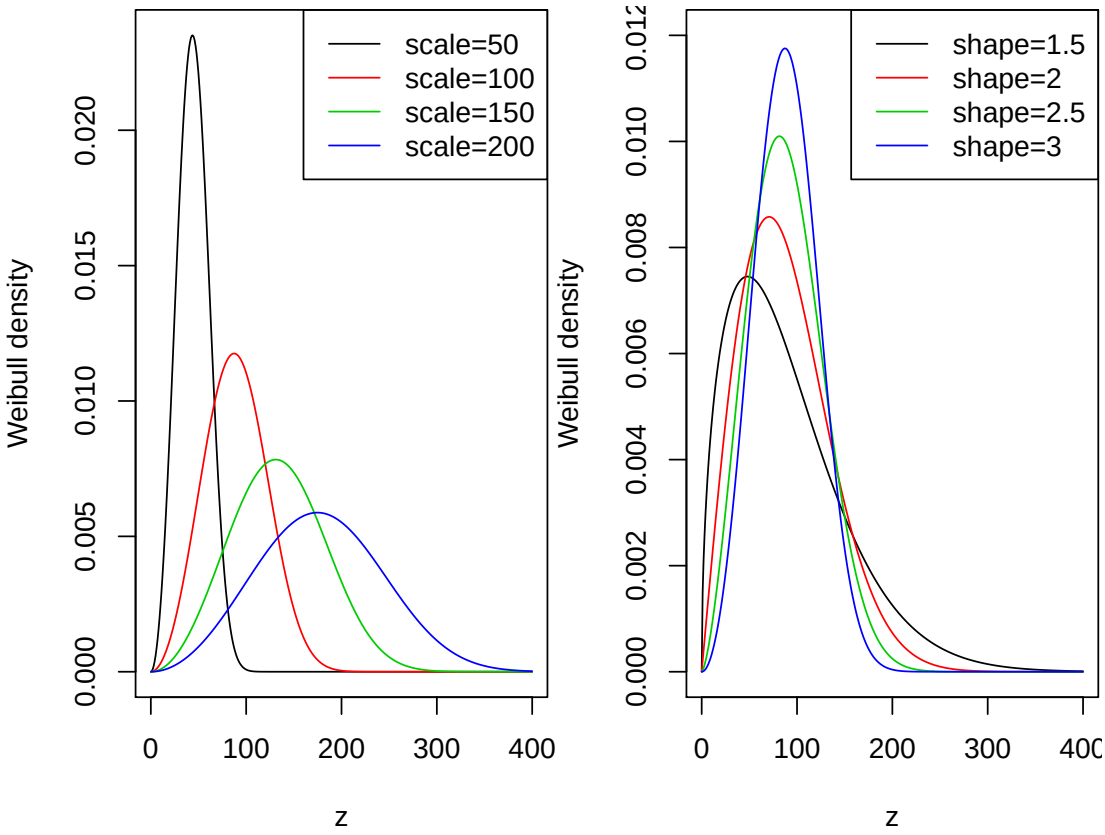


Figure 3.3: Weibull Densities. The left-hand panel is with shape=3 and Varying Scale. The right-hand panel is with scale=100 and Varying Shape.

The k-th moment of the Weibull distributed random variable is given by

$$\mathrm{E}\left(X^k\right) = \theta^k \, \Gamma\left(1 + \frac{k}{\alpha}\right).$$

The mean and variance are given by

$$\mathrm{E}\left(X\right) = \theta \, \Gamma\left(1 + \frac{1}{\alpha}\right)$$

and

$$\mathrm{Var}(X) = \theta^2 \left(\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2\right),$$

respectively.

Example 3.2.2. Suppose that the probability distribution of the lifetime of AIDS patients (in months) from the time of diagnosis is described by the Weibull distribution with shape parameter 1.2 and scale parameter 33.33.

Find the probability that a randomly selected person from this population survives at least 12 months,

A random sample of 10 patients will be selected from this population. What is the probability that at most two will die within one year of diagnosis.

Find the 99-th percentile of the distribution of lifetimes.

Show Example Solution

Solution.

a. Let X be the lifetime of AIDS patients (in months) having a Weibull distribution with parameters $(1.2, 33.33)$. We have,

$$\Pr\left(X \geq 12\right) = S_X\left(12\right) = e^{-\left(\frac{12}{33.33}\right)^{1.2}} = 0.746.$$

b. Let Y be the number of patients who die within one year of diagnosis. Then, $Y \sim Bin\left(10,\ 0.254\right)$ and $\Pr\left(Y \leq 2\right) = 0.514$.

c. Let $\pi_{0.99}$ denote the 99-th percentile of this distribution. Then,

$$S_X\left(\pi_{0.99}\right) = \exp\left\{-\left(\frac{\pi_{0.99}}{33.33}\right)^{1.2}\right\} = 0.01.$$

Solving for $\pi_{0.99}$, we get $\pi_{0.99} = 118.99$.

3.2.4 The Generalized Beta Distribution of the Second Kind

The Generalized Beta Distribution of the Second Kinda 4-parameter flexible distribution that encompasses many common distributions (GB2) was introduced by **?** in the context of insurance loss modeling and by **?** as an income and wealth distribution. It is a four-parameter very flexible distribution that can model positively as well as negatively skewed distributions.

The continuous variable X is said to have the GB2 distribution with parameters σ, θ, α_1 and α_2 if its probability density function is given by

$$f_X(x) = \frac{(x/\theta)^{\alpha_2/\sigma}}{x\sigma\ \mathrm{B}(\alpha_1,\alpha_2)\left[1+(x/\theta)^{1/\sigma}\right]^{\alpha_1+\alpha_2}} \quad \text{for } x>0, \tag{3.1}$$

$\sigma, \theta, \alpha_1, \alpha_2 > 0$, and where the beta function $\mathrm{B}(\alpha_1, \alpha_2)$ is defined as

$$\mathrm{B}(\alpha_1,\alpha_2) = \int_0^1 t^{\alpha_1-1}(1-t)^{\alpha_2-1}\,dt.$$

The GB2 provides a model for heavy as well as light tailed data. It includes the exponential, gamma, Weibull, Burr, Lomax, F, chi-square, Rayleigh, lognormal and log-logistic as special or limiting cases. For example, by setting the parameters $\sigma = \alpha_1 = \alpha_2 = 1$, then the GB2 reduces to the log-logistic distribution. When $\sigma = 1$ and $\alpha_2 \to \infty$, it reduces to the gamma distribution and when $\alpha = 1$ and $\alpha_2 \to \infty$, it reduces to the Weibull distribution.

A GB2 random variable can be defined as follows. Suppose that G_1 and G_2 are independent random variables where G_i has a gamma distribution with parameters α_i and scale parameter 1. Then, one can show that the random variable $X = \theta\left(\frac{G_1}{G_2}\right)^{\sigma}$ has a GB2 distribution with pdf summarized in equation (**??**). This theoretical result has several implications. For example, when the moments exist, one can show that the k-th moment of the GB2 distributed random variable is given by

$$\mathrm{E}\left(X^k\right) = \frac{\theta^k\ \mathrm{B}(\alpha_1+k\sigma, \alpha_2-k\sigma)}{\mathrm{B}(\alpha_1,\alpha_2)}, \quad k>0.$$

Earlier applications of the GB2 were on income data and more recently have been used to model long-tailed claims data (Section **??** describes different interpretations of the descriptor "long-tail"). GB2 was used to model different types of automobile insurance claims, severity of fire losses as well as medical insurance claim data.

Show Quiz Solution

3.3 Methods of Creating New Distributions

In this section, you learn how to:

- Understand connections among the distributions
- Give insights into when a distribution is preferred when compared to alternatives
- Provide foundations for creating new distributions

3.3.1 Functions of Random Variables and their Distributions

In Section **??** we discussed some elementary known distributions. In this section we discuss means of creating new parametric probability distributions from existing ones. Specifically, let X be a continuous random variable with a known probability density function $f_X(x)$ and distribution function $F_X(x)$. We are interested in the distribution of $Y = g(X)$, where $g(X)$ is a one-to-one transformationa function or method that turns one distribution into another defining a new random variable Y. In this section we apply the following techniques for creating new families of distributions: (a) multiplication by a constant (b) raising to a power, (c) exponentiation and (d) mixing.

3.3.2 Multiplication by a Constant

If claim data show change over time then such transformation can be useful to adjust for inflation. If the level of inflation is positive then claim costs are rising, and if it is negative then costs are falling. To adjust for inflation we multiply the cost X by 1+ inflation rate (negative inflation is deflation). To account for currency impact on claim costs we also use a transformation to apply currency conversion from a base to a counter currency.

Consider the transformation $Y = cX$, where $c > 0$, then the distribution function of Y is given by

$$F_Y(y) = \Pr(Y \leq y) = \Pr(cX \leq y) = \Pr\left(X \leq \frac{y}{c}\right) = F_X\left(\frac{y}{c}\right).$$

Hence, the probability density function of interest $f_Y(y)$ can be written as

$$f_Y(y) = \frac{1}{c} f_X\left(\frac{y}{c}\right).$$

Suppose that X belongs to a certain set of parametric distributionsprobability distribution defined by a fixed set of parameters and define a rescaled version

$Y = cX$, $c > 0$. If Y is in the same set of distributions then the distribution is said to be a scale distributiona distribution with the property that multiplying all values by a constant leads to the same distribution family with only the scale parameter changed. When a member of a scale distribution is multiplied by a constant c ($c > 0$), the scale parameter for this scale distribution meets two conditions:

The parameter is changed by multiplying by c;

All other parameters remain unchanged.

Example 3.3.1. Actuarial Exam Question. The aggregate losses of Eiffel Auto Insurance are denoted in Euro currency and follow a lognormal distribution with $\mu = 8$ and $\sigma = 2$. Given that 1 euro = 1.3 dollars, find the set of lognormal parameters which describe the distribution of Eiffel's losses in dollars.

Show Example Solution

Solution.

Let X and Y denote the aggregate losses of Eiffel Auto Insurance in euro currency and dollars respectively. As $Y = 1.3X$, we have,

$$F_Y(y) = \Pr(Y \leq y) = \Pr(1.3X \leq y) = \Pr\left(X \leq \frac{y}{1.3}\right) = F_X\left(\frac{y}{1.3}\right).$$

X follows a lognormal distribution with parameters $\mu = 8$ and $\sigma = 2$. The probability density function of X is given by

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right\} \quad \text{for } x > 0.$$

As $\left|\frac{dx}{dy}\right| = \frac{1}{1.3}$, the probability density function of interest $f_Y(y)$ is

$$f_Y(y) = \frac{1}{1.3} f_X\left(\frac{y}{1.3}\right) = \frac{1}{1.3}\frac{1.3}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln(y/1.3) - \mu}{\sigma}\right)^2\right\} = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln y - (\ln 1.3 + \mu)}{\sigma}\right)^2\right\}.$$

Then Y follows a lognormal distribution with parameters $\ln 1.3 + \mu = 8.26$ and $\sigma = 2.00$. If we let $\mu = ln(m)$ then it can be easily seen that $m{=}e^{\mu}$ is the scale parameter which was multiplied by 1.3 while σ is the shape parameter that remained unchanged.

Example 3.3.2. Actuarial Exam Question. Demonstrate that the gamma distribution is a scale distribution.

Show Example Solution

Solution.

Let $X \sim Ga(\alpha, \theta)$ and $Y = cX$. As $\left|\frac{dx}{dy}\right| = \frac{1}{c}$, then

$$f_Y(y) = \frac{1}{c} f_X\left(\frac{y}{c}\right) = \frac{\left(\frac{y}{c\theta}\right)^{\alpha}}{y\,\Gamma(\alpha)} \exp\left(-\frac{y}{c\theta}\right).$$

We can see that $Y \sim Ga(\alpha, c\theta)$ indicating that gamma is a scale distribution and θ is a scale parameter.

Using the same approach you can demonstrate that other distributions introduced in Section **??** are also scale distributions. In actuarial modeling, working with a scale distribution is very convenient because it allows to incorporate the effect of inflation and to accommodate changes in the currency unit.

3.3.3 Raising to a Power

In Section **??** we talked about the flexibility of the Weibull distribution in fitting reliability dataa dataset consisting of failure times for failed units and run times for units still functioning. Looking to the origins of the Weibull distribution, we recognize that the Weibull is a power transformationa transformation type that involves raising a random variable to a power of the exponential distribution. This is an application of another type of transformation which involves raising the random variable to a power.

Consider the transformation $Y = X^{\tau}$, where $\tau > 0$, then the distribution function of Y is given by

$$F_Y(y) = \Pr(Y \le y) = \Pr(X^{\tau} \le y) = \Pr\left(X \le y^{1/\tau}\right) = F_X\left(y^{1/\tau}\right).$$

Hence, the probability density function of interest $f_Y(y)$ can be written as

$$f_Y(y) = \frac{1}{\tau} y^{1/\tau - 1} f_X\left(y^{1/\tau}\right).$$

On the other hand, if $\tau < 0$, then the distribution function of Y is given by

$$F_Y(y) = \Pr(Y \le y) = \Pr(X^{\tau} \le y) = \Pr\left(X \ge y^{1/\tau}\right) = 1 - F_X\left(y^{1/\tau}\right),$$

and

$$f_Y(y) = \left|\frac{1}{\tau}\right| y^{1/\tau - 1} f_X\left(y^{1/\tau}\right).$$

Example 3.3.3. We assume that X follows the exponential distribution with mean θ and consider the transformed variable $Y = X^{\tau}$. Show that Y follows the Weibull distribution when τ is positive and determine the parameters of the Weibull distribution.

Show Example Solution

Solution.

As X follows the exponential distribution with mean θ, we have

$$f_X(x) = \frac{1}{\theta} e^{-x/\theta} \quad x > 0.$$

Solving for x yields $x = y^{1/\tau}$. Taking the derivative, we have

$$\left| \frac{dx}{dy} \right| = \frac{1}{\tau} y^{\frac{1}{\tau}-1}.$$

Thus,

$$f_Y(y) = \frac{1}{\tau} y^{\frac{1}{\tau}-1} f_X\left(y^{\frac{1}{\tau}}\right) = \frac{1}{\tau\theta} y^{\frac{1}{\tau}-1} e^{-\frac{y^{\frac{1}{\tau}}}{\theta}} = \frac{\alpha}{\beta}\left(\frac{y}{\beta}\right)^{\alpha-1} e^{-(y/\beta)^{\alpha}}.$$

where $\alpha = \frac{1}{\tau}$ and $\beta = \theta^{\tau}$. Then, Y follows the Weibull distribution with shape parameter α and scale parameter β.

3.3.4 Exponentiation

The normal distribution is a very popular model for a wide number of applications and when the sample size is large, it can serve as an approximate distribution for other models. If the random variable X has a normal distribution with mean μ and variance σ^2, then $Y = e^X$ has lognormal distributiona heavy-tailed, positively skewed 2-parameter continuous distribution such that the natural log of the random variable is normally distributed with the same parameter values with parameters μ and σ^2. The lognormal random variable has a lower bound of zero, is positively skewed and has a long right tail. A lognormal distribution is commonly used to describe distributions of financial assets such as stock prices. It is also used in fitting claim amounts for automobile as well as health insurance. This is an example of another type of transformation which involves exponentiation.

In general, consider the transformation $Y = e^X$. Then, the distribution function of Y is given by

$$F_Y(y) = \Pr(Y \le y) = \Pr\left(e^X \le y\right) = \Pr(X \le \ln y = F_X(\ln y).$$

Taking derivatives, we see that the probability density function of interest $f_Y(y)$ can be written as

$$f_Y(y) = \frac{1}{y} f_X(\ln y).$$

As an important special case, suppose that X is normally distributed with mean μ and variance σ^2. Then, the distribution of $Y = e^X$ is

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

This is known as a lognormal distribution.

Example 3.3.4. Actuarial Exam Question. Assume that X has a uniform distribution on the interval $(0,\ c)$ and define $Y = e^X$. Find the distribution of Y.

Show Example Solution

Solution.

We begin with the cdf of Y,

$$F_Y(y) = \Pr(Y \leq y) = \Pr\left(e^X \leq y\right) = \Pr(X \leq \ln y) = F_X(\ln y).$$

Taking the derivative, we have,

$$f_Y(y) = \frac{1}{y} f_X(\ln y) = \frac{1}{cy}.$$

Since $0 < x < c$, then $1 < y < e^c$.

3.3.5 Finite Mixtures

Mixture distributions represent a useful way of modelling data that are drawn from a heterogeneous populationa dataset where the subpopulations are represented by separate distinct distributions. This parent population can be thought to be divided into multiple subpopulations with distinct distributions.

Two-point Mixture

If the underlying phenomenon is diverse and can actually be described as two phenomena representing two subpopulations with different modes, we can construct the two point mixture random variable X. Given random variables X_1 and X_2, with probability density functions $f_{X_1}(x)$ and $f_{X_2}(x)$ respectively, the probability density function of X is the weighted average of the component probability density function $f_{X_1}(x)$ and $f_{X_2}(x)$. The probability density function and distribution function of X are given by

$$f_X(x) = a f_{X_1}(x) + (1-a) f_{X_2}(x),$$

and

$$F_X(x) = aF_{X_1}(x) + (1-a)F_{X_2}(x),$$

for $0 < a < 1$, where the mixing parametersproportion weight given to each subpopulation in a mixture a and $(1-a)$ represent the proportions of data points that fall under each of the two subpopulations respectively. This weighted average can be applied to a number of other distribution related quantities. The k-th moment and moment generating function of X are given by $\mathrm{E}\left(X^k\right) = a\mathrm{E}\left(X_1^K\right) + (1-a)\,\mathrm{E}\left(X_2^k\right)$, and

$$M_X(t) = aM_{X_1}(t) + (1-a)M_{X_2}(t),$$

respectively.

Example 3.3.5. Actuarial Exam Question. A collection of insurance policies consists of two types. 25% of policies are Type 1 and 75% of policies are Type 2. For a policy of Type 1, the loss amount per year follows an exponential distribution with mean 200, and for a policy of Type 2, the loss amount per year follows a Pareto distribution with parameters $\alpha = 3$ and $\theta = 200$. For a policy chosen at random from the entire collection of both types of policies, find the probability that the annual loss will be less than 100, and find the average loss.

Show Example Solution

Solution.

The two types of losses are the random variables X_1 and X_2. X_1 has an exponential distribution with mean 100, so $F_{X_1}(100) = 1 - e^{-\frac{100}{200}} = 0.393$. X_2 has a Pareto distribution with parameters $\alpha = 3$ and $\theta = 200$, so $F_{X_1}(100) = 1 - \left(\frac{200}{100+200}\right)^3 = 0.704$. Hence, $F_X(100) = (0.25 \times 0.393) + (0.75 \times 0.704) = 0.626$.

The average loss is given by

$$\mathrm{E}(X) = 0.25\mathrm{E}(X_1) + 0.75\mathrm{E}(X_2) = (0.25 \times 200) + (0.75 \times 100) = 125$$

.

---

k-point Mixture

In case of finite mixture distributions, the random variable of interest X has a probability p_i of being drawn from homogeneous subpopulation i, where $i = 1, 2, \ldots, k$ and k is the initially specified number of subpopulations in our mixture. The mixing parameter p_i represents the proportion of observations from subpopulation i. Consider the random variable X generated from

k distinct subpopulations, where subpopulation i is modeled by the continuous distribution $f_{X_i}(x)$. The probability distribution of X is given by

$$f_X(x) = \sum_{i=1}^{k} p_i f_{X_i}(x),$$

where $0 < p_i < 1$ and $\sum_{i=1}^{k} p_i = 1$.

This model is often referred to as a finite mixturea mixture distribution with a finite k number of subpopulations or a k-point mixture. The distribution function, r-th moment and moment generating functions of the k-th point mixture are given as

$$F_X(x) = \sum_{i=1}^{k} p_i F_{X_i}(x),$$

$$\mathrm{E}(X^r) = \sum_{i=1}^{k} p_i \mathrm{E}(X_i^r), \text{and}$$

$$M_X(t) = \sum_{i=1}^{k} p_i M_{X_i}(t),$$

respectively.

Example 3.3.6. Actuarial Exam Question. Y_1 is a mixture of X_1 and X_2 with mixing weights a and $(1-a)$. Y_2 is a mixture of X_3 and X_4 with mixing weights b and $(1-b)$. Z is a mixture of Y_1 and Y_2 with mixing weights c and $(1-c)$.

Show that Z is a mixture of X_1, X_2, X_3 and X_4, and find the mixing weights.

Show Example Solution

Solution. Applying the formula for a mixed distribution, we get

$$f_{Y_1}(x) = a f_{X_1}(x) + (1-a) f_{X_2}(x)$$

$$f_{Y_2}(x) = b f_{X_3}(x) + (1-b) f_{X_4}(x)$$

$$f_Z(x) = c f_{Y_1}(x) + (1-c) f_{Y_2}(x)$$

Substituting the first two equations into the third, we get

$$f_Z(x) = c\left[a f_{X_1}(x) + (1-a) f_{X_2}(x)\right] + (1-c)\left[b f_{X_3}(x) + (1-b) f_{X_4}(x)\right]$$

$$= caf_{X_1}(x) + c(1-a) f_{X_2}(x) + (1-c) bf_{X_3}(x) + (1-c)(1-b) f_{X_4}(x)$$

.

Then, Z is a mixture of X_1, X_2, X_3 and X_4, with mixing weights ca, $c(1-a)$, $(1-c)b$ and $(1-c)(1-b)$, respectively. It can be easily seen that the mixing weights sum to one.

3.3.6 Continuous Mixtures

A mixture with a very large number of subpopulations (k goes to infinity) is often referred to as a continuous mixturea mixture distribution with an infinite number of subpopulations, where the mixing parameter is itself a continuous distribution. In a continuous mixture, subpopulations are not distinguished by a discrete mixing parameter but by a continuous variable Θ, where Θ plays the role of p_i in the finite mixture. Consider the random variable X with a distribution depending on a parameter Θ, where Θ itself is a continuous random variable. This description yields the following model for X

$$f_X(x) = \int_{-\infty}^{\infty} f_X(x|\theta)\, g_\Theta(\theta) d\theta,$$

where $f_X(x|\theta)$ is the conditional distributiona probability distribution that applies to a subpopulation satisfying the condition of X at a particular value of $\Theta = \theta$ and $g_\Theta(\theta)$ is the probability statement made about the unknown parameter θ. In a Bayesian context (described in Section **??**), this is known as the prior distributiona probability distribution assigned prior to observing additional data of Θ (the prior information or expert opinion to be used in the analysis).

The distribution function, k-th moment and moment generating functions of the continuous mixture are given as

$$F_X(x) = \int_{-\infty}^{\infty} F_X(x|\theta)\, g_\Theta(\theta) d\theta,$$

$$\mathrm{E}\left(X^k\right) = \int_{-\infty}^{\infty} \mathrm{E}\left(X^k|\theta\right) g_\Theta(\theta) d\theta,$$

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \int_{-\infty}^{\infty} \mathrm{E}\left(e^{tx}|\theta\right) g_\Theta(\theta) d\theta,$$

respectively.

The k-th moment of the mixture distribution can be rewritten as

$$\mathrm{E}\left(X^k\right) = \int_{-\infty}^{\infty} \mathrm{E}\left(X^k\,|\theta\right) g_\Theta(\theta) d\theta \;=\; \mathrm{E}\left[\mathrm{E}\left(X^k\,|\Theta\right)\right].$$

Using the law of iterated expectations (see Appendix Chapter **??**), we can define the mean and variance of X as

$$\mathrm{E}\left(X\right) = \mathrm{E}\left[\mathrm{E}\left(X\,|\Theta\right)\right]$$

and

$$\mathrm{Var}\left(X\right) = \mathrm{E}\left[\mathrm{Var}\left(X\,|\Theta\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(X\,|\Theta\right)\right].$$

Example 3.3.7. Actuarial Exam Question. X has a normal distribution with a mean of Λ and variance of 1. Λ has a normal distribution with a mean of 1 and variance of 1. Find the mean and variance of X.

Show Example Solution

Solution.

X is a continuous mixture with mean

$$\mathrm{E}\left(X\right) = \mathrm{E}\left[\mathrm{E}\left(X|\Lambda\right)\right] = \mathrm{E}\left(\Lambda\right) = 1 \text{ and } \mathrm{V}\left(X\right) = \mathrm{V}\left[\mathrm{E}\left(X|\Lambda\right)\right] + \mathrm{E}\left[\mathrm{V}\left(X|\Lambda\right)\right] = \mathrm{V}\left(\Lambda\right) + \mathrm{E}\left(1\right) = 1+1 = 2.$$

Example 3.3.8. Actuarial Exam Question. Claim sizes, X, are uniform on the interval $(\Theta, \Theta + 10)$ for each policyholder. Θ varies by policyholder according to an exponential distribution with mean 5. Find the unconditional distributiona probability distribution independent of any another imposed conditions, mean and variance of X.

Show Example Solution

Solution.

The conditional distribution of X is $f_X\left(\ x|\,\theta\right) = \frac{1}{10}$ for $\theta < x < \theta + 10$.

The prior distribution of θ is $g_\Theta(\theta) = \frac{1}{5}e^{-\frac{\theta}{5}}$ for $0 < \theta < \infty$.

The conditional mean and variance of X are given by

$$\mathrm{E}\left(\ X|\,\theta\right) = \frac{\theta + \theta + 10}{2} = \theta + 5$$

and

$$\mathrm{Var}\left(\ X|\,\theta\right) = \frac{\left[(\theta + 10) - \theta\right]^2}{12} = \frac{100}{12},$$

respectively.

Hence, the unconditional mean and variance of X are given by

$$\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X|\Theta)] = \mathrm{E}(\Theta + 5) = \mathrm{E}(\Theta) + 5 = 5 + 5 = 10,$$

and

$$\mathrm{Var}(X) = \mathrm{E}[V(X|\Theta)] + \mathrm{Var}[\mathrm{E}(X|\Theta)] = \mathrm{E}\left(\frac{100}{12}\right) + \mathrm{Var}(\Theta + 5) = 8.33 + \mathrm{Var}(\Theta) = 33.33.$$

The unconditional distribution of X is

$$f_X(x) = \int f_X(x|\theta)\ g_\Theta(\theta) d\theta.$$

$$f_X(x) = \begin{cases} \int_0^x \frac{1}{50} e^{-\frac{\theta}{5}} d\theta = \frac{1}{10}\left(1 - e^{-\frac{x}{5}}\right) & 0 \le x \le 10, \\ \int_{x-10}^x \frac{1}{50} e^{-\frac{\theta}{5}} d\theta = \frac{1}{10}\left(e^{-\frac{(x-10)}{5}} - e^{-\frac{x}{5}}\right) & 10 < x < \infty. \end{cases}$$

Show Quiz Solution

3.4 Coverage Modifications

In this section we evaluate the impacts of coverage modifications: a) deductibles, b) policy limit, c) coinsurance and inflation on insurer's costs.

3.4.1 Policy Deductibles

Under an ordinary deductible policy, the insured (policyholder) agrees to cover a fixed amount of an insurance claim before the insurer starts to pay. This fixed expense paid out of pocket is called the deductible and often denoted by d. If the loss exceeds d then the insurer is responsible for covering the loss X less the deductible d. Depending on the agreement, the deductible may apply to each covered loss or to the total losses during a defined benefit period (month, year, etc.)

Deductibles eliminate a large number of small claims, reduce costs of handling and processing these claims, reduce premiums for the policyholders and reduce moral hazardsituation where an insured is more likely to be risk seeking if they do not bear sufficient consequences for a loss. Moral hazard occurs when the insured takes more risks, increasing the chances of loss due to perils insured

against, knowing that the insurer will incur the cost (e.g. a policyholder with collision insurance may be encouraged to drive recklessly). The larger the deductible, the less the insured pays in premiums for an insurance policy.

Let X denote the loss incurred to the insured and Y denote the amount of paid claim by the insurer. Speaking of the benefit paid to the policyholder, we differentiate between two variables: The payment per loss and the payment per payment. The payment per lossamount insurer pays when a loss occurs and can be 0 variable, denoted by Y^L or $(X-d)_+$ is left censoredvalues below a threshold d are not ignored but converted to 0 because values of X that are less than d are not ignored but are set equal to zero. This variable includes losses for which a payment is made as well as losses less than the deductible and hence is defined as

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

Y^L is often referred to as left censored and shifted variable because the values below d are not ignored and all losses are shifted by a value d.

On the other hand, the payment per paymentamount insurer pays given a payment is needed and is greater than 0 variable, denoted by Y^P, is defined only when there is a payment. Specifically, Y^P equals $X-d$ on the event $\{X > d\}$, denoted as $Y^P = X-d||X > d$. Another way of expressing this that is commonly used is

$$Y^P = \begin{cases} \text{Undefined} & X \leq d \\ X-d & X > d \end{cases}.$$

Here, Y^P is often referred to as left truncatedvalues below a threshold d are not reported and unknown and shifted variable or excess loss variable because the claims smaller than d are not reported and values above d are shifted by d.

Even when the distribution of X is continuous, the distribution of Y^L is a hybrid combination of discrete and continuous components. The discrete part of the distribution is concentrated at $Y = 0$ (when $X \leq d$) and the continuous part is spread over the interval $Y > 0$ (when $X > d$). For the discrete part, the probability that no payment is made is the probability that losses fall below the deductible; that is,

$$\Pr\left(Y^L = 0\right) = \Pr\left(X \leq d\right) = F_X\left(d\right).$$

Using the transformation $Y^L = X - d$ for the continuous part of the distribution, we can find the probability density function of Y^L given by

$$f_{Y^L}\left(y\right) = \begin{cases} F_X\left(d\right) & y = 0, \\ f_X\left(y+d\right) & y > 0 \end{cases}$$

We can see that the payment per payment variable is the payment per loss variable conditioned on the loss exceeding the deductible; that is, $Y^P = \left. Y^L \right| X > d$.

Hence, the probability density function of Y^P is given by

$$f_{Y^P}(y) = \frac{f_X(y+d)}{1-F_X(d)},$$

for $y > 0$. Accordingly, the distribution functions of $Y^L$and Y^P are given by

$$F_{Y^L}(y) = \begin{cases} F_X(d) & y = 0, \\ F_X(y+d) & y > 0. \end{cases}$$

and

$$F_{Y^P}(y) = \frac{F_X(y+d) - F_X(d)}{1-F_X(d)},$$

for $y > 0$, respectively.

The raw moments of Y^L and Y^P can be found directly using the probability density function of X as follows

$$\mathrm{E}\left[\left(Y^L\right)^k\right] = \int_d^\infty (x-d)^k f_X(x)\,dx,$$

and

$$\mathrm{E}\left[\left(Y^P\right)^k\right] = \frac{\int_d^\infty (x-d)^k f_X(x)\,dx}{1-F_X(d)} = \frac{\mathrm{E}\left[\left(Y^L\right)^k\right]}{1-F_X(d)},$$

respectively. For $k = 1$, we can use the survival function to calculate $\mathrm{E}(Y^L)$ as

$$\mathrm{E}(Y^L) = \int_d^\infty [1 - F_X(x)]\,dx.$$

This could be easily proved if we start with the initial definition of $\mathrm{E}(Y^L)$ and use integration by parts.

We have seen that the deductible d imposed on an insurance policy is the amount of loss that has to be paid out of pocket before the insurer makes any payment. The deductible d imposed on an insurance policy reduces the insurer's payment. The loss elimination ratio (LER)% decrease of the expected payment by the insurer as a result of the deductible is the percentage decrease in the expected payment of the insurer as a result of imposing the deductible. LER is defined as

$$LER = \frac{\mathrm{E}(X) - \mathrm{E}\left(Y^L\right)}{\mathrm{E}(X)}.$$

A little less common type of policy deductible is the franchise deductible. The franchise deductibleinsurer pays nothing for losses below the deductible, but pays the full amount for any loss above the deductible will apply to the policy in the same way as ordinary deductible except that when the loss exceeds the

deductible d, the full loss is covered by the insurer. The payment per loss and payment per payment variables are defined as

$$Y^L = \begin{cases} 0 & X \leq d, \\ X & X > d, \end{cases}$$

and

$$Y^P = \begin{cases} \text{Undefined} & X \leq d, \\ X & X > d, \end{cases}$$

respectively.

Example 3.4.1. Actuarial Exam Question. A claim severity distribution is exponential with mean 1000. An insurance company will pay the amount of each claim in excess of a deductible of 100. Calculate the variance of the amount paid by the insurance company for one claim, including the possibility that the amount paid is 0.

Show Example Solution

Solution.

Let Y^L denote the amount paid by the insurance company for one claim.

$$Y^L = (X - 100)_+ = \begin{cases} 0 & X \leq 100, \\ X - 100 & X > 100. \end{cases}$$

The first and second moments of Y^L are

$$E\left(Y^L\right) = \int_{100}^{\infty} (x - 100) f_X(x)\, dx = \int_{100}^{\infty} S_X(x) dx = 1000 e^{-\frac{100}{1000}},$$

and

$$E\left[\left(Y^L\right)^2\right] = \int_{100}^{\infty} (x - 100)^2 f_X(x)\, dx = 2 \times 1000^2 e^{-\frac{100}{1000}}.$$

So,

$$\text{Var}\left(Y^L\right) = \left(2 \times 1000^2 e^{-\frac{100}{1000}}\right) - \left(1000 e^{-\frac{100}{1000}}\right)^2 = 990,944.$$

An arguably simpler path to the solution is to make use of the relationship between X and Y^P. If X is exponentially distributed with mean 1000, then Y^P is also exponentially distributed with the same mean, because of the memoryless property of the exponential distribution. Hence, $E\left(Y^P\right)$=1000 and

$$E\left[\left(Y^P\right)^2\right] = 2 \times 1000^2.$$

Using the relationship between Y^L and Y^P we find

$$E\left(Y^L\right) = E\left(Y^P\right) S_X(100) = 1000 e^{-\frac{100}{1000}}$$

$$E\left[\left(Y^L\right)^2\right] = E\left[\left(Y^P\right)^2\right] S_X\left(100\right) = 2 \times 1000^2 e^{-\frac{100}{1000}}.$$

The relationship between X and Y^P can also be used when dealing with the uniform or the Pareto distributions. You can easily show that if X is uniform over the interval $(0, \theta)$ then Y^P is uniform over the interval $(0, \theta - d)$ and if X is Pareto with parameters α and θ then Y^P is Pareto with parameters α and $\theta + d$.

Example 3.4.2. Actuarial Exam Question. For an insurance:

Losses have a density function

$$f_X\left(x\right) = \begin{cases} 0.02x & 0 < x < 10, \\ 0 & \text{elsewhere.} \end{cases}$$

The insurance has an ordinary deductible of 4 per loss.

Y^P is the claim payment per payment random variable.

Calculate $\mathrm{E}\left(Y^P\right)$.

Show Example Solution

Solution.

We define Y^P as follows

$$Y^P = \begin{cases} \text{Undefined} & X \leq 4, \\ X - 4 & X > 4. \end{cases}$$

So, $E\left(Y^P\right) = \frac{\int_4^{10}(x-4)0.02x dx}{1-F_X(4)} = \frac{2.88}{0.84} = 3.43$.

Note that we divide by $S_X(4) = 1 - F_X(4)$, as this is the range where the variable Y^P is defined.

Example 3.4.3. Actuarial Exam Question. You are given:

Losses follow an exponential distribution with the same mean in all years.

The loss elimination ratio this year is 70%.

The ordinary deductible for the coming year is 4/3 of the current deductible.

Compute the loss elimination ratio for the coming year.

Show Example Solution

Solution.

Let the losses $X \sim Exp(\theta)$ and the deductible for the coming year $d' = \frac{4}{3}d$, the deductible of the current year. The LER for the current year is

$$\frac{E(X) - E\left(Y^L\right)}{E(X)} = \frac{\theta - \theta e^{-d/\theta}}{\theta} = 1 - e^{-d/\theta} = 0.7.$$

Then, $e^{-d/\theta} = 0.3$.

The LER for the coming year is

$$\frac{\theta - \theta \exp(-\frac{d'}{\theta})}{\theta} = \frac{\theta - \theta \exp(-\frac{\left(\frac{4}{3}d\right)}{\theta})}{\theta}$$
$$= 1 - \exp\left(-\frac{\frac{4}{3}d}{\theta}\right) = 1 - \left(e^{-d/\theta}\right)^{4/3} = 1 - 0.3^{4/3} = 0.8.$$

3.4.2 Policy Limits

Under a limited policy, the insurer is responsible for covering the actual loss X up to the limit of its coverage. This fixed limit of coveragepolicy limit, or maximum contractual financial obligation of the insurer for a loss is called the policy limit and often denoted by u. If the loss exceeds the policy limit, the difference $X - u$ has to be paid by the policyholder. While a higher policy limit means a higher payout to the insured, it is associated with a higher premium.

Let X denote the loss incurred to the insured and Y denote the amount of paid claim by the insurer. The variable Y known as the limited loss variable and denoted by $X \wedge u$. It is a right censored variablevalues above a threshold u are not ignored but converted to u because values above u are set equal to u. The limited loss random variable Y is defined as

$$Y = X \wedge u = \begin{cases} X & X \leq u, \\ u & X > u. \end{cases}$$

It can be seen that the distinction between Y^L and Y^P is not needed under limited policy as the insurer will always make a payment.

Using the definitions of $(X - d)_+$ and $(X \wedge d)$, it can be easily seen that the expected payment without any coverage modification, X, is equal to the sum of the expected payments with deductible d and limit d. That is, $X = (X - d)_+ + (X \wedge d)$.

When a loss is subject to a deductible d and a limit u, the per-loss variable Y^L is defined as

$$Y^L = \begin{cases} 0 & X \leq d, \\ X - d & d < X \leq u, \\ u - d & X > u. \end{cases}$$

Hence, Y^L can be expressed as $Y^L = (X \wedge u) - (X \wedge d)$.

Even when the distribution of X is continuous, the distribution of Y is a hybrid combination of discrete and continuous components. The discrete part of the distribution is concentrated at $Y = u$ (when $X > u$), while the continuous part is spread over the interval $Y < u$ (when $X \leq u$). For the discrete part, the probability that the benefit paid is u, is the probability that the loss exceeds the policy limit u; that is,

$$\Pr(Y = u) = \Pr(X > u) = 1 - F_X(u).$$

For the continuous part of the distribution $Y = X$, hence the probability density function of Y is given by

$$f_Y(y) = \begin{cases} f_X(y) & 0 < y < u, \\ 1 - F_X(u) & y = u. \end{cases}$$

Accordingly, the distribution function of Y is given by

$$F_Y(y) = \begin{cases} F_X(x) & 0 < y < u, \\ 1 & y \geq u. \end{cases}$$

The raw moments of Y can be found directly using the probability density function of X as follows

$$\mathrm{E}\left(Y^k\right) = \mathrm{E}\left[(X \wedge u)^k\right] = \int_0^u x^k f_X(x)\, dx + \int_u^\infty u^k f_X(x) dx = \int_0^u x^k f_X(x)\, dx + u^k \left[1 - F_X(u)\right].$$

For $k = 1$, we can use the survival function to calculate $\mathrm{E}(Y)$ as follows

$$\mathrm{E}(Y) = \mathrm{E}(X \wedge u) = \int_0^u [1 - F_X(x)] dx.$$

This could be easily proved if we start with the initial definition of $\mathrm{E}(Y)$ and use integration by parts.

Example 3.4.4. Actuarial Exam Question. Under a group insuranceinsurance provided to groups of people to take advantage of lower administrative costs vs. individual policies policy, an insurer agrees to pay 100% of the medical bills incurred during the year by employees of a small company, up to a maximum total of one million dollars. The total amount of bills incurred, X, has probability density function

$$f_X(x) = \begin{cases} \frac{x(4-x)}{9} & 0 < x < 3, \\ 0 & \text{elsewhere.} \end{cases}$$

where x is measured in millions. Calculate the total amount, in millions of dollars, the insurer would expect to pay under this policy.

Show Example Solution

Solution.

Define the total amount of bills paid by the insurer as

$$Y = X \wedge 1 = \begin{cases} X & X \leq 1, \\ 1 & X > 1. \end{cases}$$

So $\mathrm{E}(Y) = \mathrm{E}(X \wedge 1) = \int_0^1 \frac{x^2(4-x)}{9} dx + 1 * \int_1^3 \frac{x(4-x)}{9} dx = 0.935$.

3.4.3 Coinsurance and Inflation

As we have seen in Section **??**, the amount of loss retained by the policyholder can be losses up to the deductible d. The retained loss can also be a percentage of the claim. The percentage α, often referred to as the coinsurance factor, is the percentage of claim the insurance company is required to cover. If the policy is subject to an ordinary deductible and policy limit, coinsurance refers to the percentage of claim the insurer is required to cover, after imposing the ordinary deductible and policy limit. The payment per loss variable, Y^L, is defined as

$$Y^L = \begin{cases} 0 & X \leq d, \\ \alpha(X-d) & d < X \leq u, \\ \alpha(u-d) & X > u. \end{cases}$$

The policy limit (the maximum amount paid by the insurer) in this case is $\alpha(u-d)$, while u is the maximum covered loss.

We have seen in Section **??** that when a loss is subject to both a deductible d and a limit u the per-loss variable Y^L can be expressed as $Y^L = (X \wedge u) - (X \wedge d)$. With coinsurance, this becomes Y^L can be expressed as $Y^L = \alpha[(X \wedge u) - (X \wedge d)]$.

The k-th moment of Y^L is given by

$$\mathrm{E}\left[\left(Y^L\right)^k\right] = \int_d^u [\alpha(x-d)]^k f_X(x)\, dx + [\alpha(u-d)]^k [1 - F_X(u)].$$

A growth factormultiplicative factor applied to a distribution to account for the impact of inflation, typically (1+rate) $(1+r)$ may be applied to X resulting in an inflated loss random variable $(1+r)X$ (the prespecified d and u remain unchanged). The resulting per loss variable can be written as

$$Y^L = \begin{cases} 0 & X \leq \frac{d}{1+r}, \\ \alpha[(1+r)X - d] & \frac{d}{1+r} < X \leq \frac{u}{1+r}, \\ \alpha(u-d) & X > \frac{u}{1+r}. \end{cases}$$

The first and second moments of Y^L can be expressed as

$$\mathrm{E}\left(Y^L\right) = \alpha\left(1+r\right)\left[\mathrm{E}\left(X \wedge \frac{u}{1+r}\right) - \mathrm{E}\left(X \wedge \frac{d}{1+r}\right)\right],$$

and

$$\mathrm{E}\left[\left(Y^L\right)^2\right] = \alpha^2\left(1+r\right)^2\left\{\mathrm{E}\left[\left(X \wedge \frac{u}{1+r}\right)^2\right] - \mathrm{E}\left[\left(X \wedge \frac{d}{1+r}\right)^2\right] \qquad -2\left(\frac{d}{1+r}\right)\left[\mathrm{E}\left(X \wedge \frac{u}{1+r}\right) - \mathrm{E}\left(X \wedge \frac{d}{1+r}\right)\right]\right.$$

respectively.

The formulas given for the first and second moments of Y^L are general. Under full coverage, $\alpha = 1$, $r = 0$, $u = \infty$, $d = 0$ and $\mathrm{E}\left(Y^L\right)$ reduces to $\mathrm{E}\left(X\right)$. If only an ordinary deductible is imposed, $\alpha = 1$, $r = 0$, $u = \infty$ and $\mathrm{E}\left(Y^L\right)$ reduces to $\mathrm{E}\left(X\right) - \mathrm{E}\left(X \wedge d\right)$. If only a policy limit is imposed $\alpha = 1$, $r = 0$, $d = 0$ and $\mathrm{E}\left(Y^L\right)$ reduces to $\mathrm{E}\left(X \wedge u\right)$.

Example 3.4.5. Actuarial Exam Question. The ground up loss random variable for a health insurance policy in 2006 is modeled with X, an exponential distribution with mean 1000. An insurance policy pays the loss above an ordinary deductible of 100, with a maximum annual payment of 500. The ground up loss random variable is expected to be 5% larger in 2007, but the insurance in 2007 has the same deductible and maximum payment as in 2006. Find the percentage increase in the expected cost per payment from 2006 to 2007.

Show Example Solution

Solution.

We define the amount per loss Y^L in both years as

$$Y^L_{2006} = \begin{cases} 0 & X \leq 100, \\ X - 100 & 100 < X \leq 600, \\ 500 & X > 600. \end{cases}$$

$$Y^L_{2007} = \begin{cases} 0 & X \leq 95.24, \\ 1.05X - 100 & 95.24 < X \leq 571.43, \\ 500 & X > 571.43. \end{cases}$$

So,

$$E\left(Y^L_{2006}\right) = E\left(X \wedge 600\right) - E\left(X \wedge 100\right) = 1000\left(1 - e^{-\frac{600}{1000}}\right) - 1000\left(1 - e^{-\frac{100}{1000}}\right)$$

$$= 356.026$$

.

$$E\left(Y_{2007}^{L}\right) = 1.05\left[E\left(X \wedge 571.43\right) - E\left(X \wedge 95.24\right)\right]$$

$$= 1.05\left[1000\left(1 - e^{-\frac{571.43}{1000}}\right) - 1000\left(1 - e^{-\frac{95.24}{1000}}\right)\right]$$

$$= 361.659$$

.

$E\left(Y_{2006}^{P}\right) = \frac{356.026}{e^{-\frac{100}{1000}}} = 393.469.$

$E\left(Y_{2007}^{P}\right) = \frac{361.659}{e^{-\frac{95.24}{1000}}} = 397.797.$

Because $\frac{E\left(Y_{2007}^{P}\right)}{E\left(Y_{2006}^{P}\right)} - 1 = 0.011$, there is an increase of 1.1% from 2006 to 2007. Due to the policy limit, the cost per payment event grew by only 1.1% between 2006 and 2007 even though the ground up losses increased by 5% between the two years.

3.4.4 Reinsurance

In Section **??** we introduced the policy deductible, which is a contractual arrangement under which an insured transfers part of the risk by securing coverage from an insurer in return for an insurance premium. Under that policy, the insured must pay all losses up to the deductible, and the insurer only pays the amount (if any) above the deductible. We now introduce reinsurancea transaction where the primary insurer buys insurance from a re-insurer who will cover part of the losses and/or loss adjustment expenses of the primary insurer, a mechanism of insurance for insurance companies. Reinsurance is a contractual arrangement under which an insurer transfers part of the underlying insured risk by securing coverage from another insurer (referred to as a reinsurer) in return for a reinsurance premium. Although reinsurance involves a relationship between three parties: the original insured, the insurer (often referred to as cedentparty that is transferring the risk to a reinsurer or cedant) and the reinsurer, the parties of the reinsurance agreement are only the primary insurer and the reinsurer. There is no contractual agreement between the original insured and the reinsurer. Though many different types of reinsurance contracts exist, a common form is excess of loss coveragecontract where an insurer pays all claims up to a specified amount and then the reinsurer pays claims in excess of stated reinsurance deductible. In such contracts, the primary insurer must make all required payments to the insured until the primary insurer's total payments reach a fixed reinsurance deducible. The reinsurer is then only responsible for paying losses above the reinsurance deductible. The maximum

amount retained by the primary insurer in the reinsurance agreement (the reinsurance deductible) is called retentionmaximum amount payable by the primary insurer in a reinsurance arrangement .

Reinsurance arrangements allow insurers with limited financial resources to increase the capacity to write insurance and meet client requests for larger insurance coverage while reducing the impact of potential losses and protecting the insurance company against catastrophic losses. Reinsurance also allows the primary insurer to benefit from underwriting skills, expertise and proficient complex claim file handling of the larger reinsurance companies.

Example 3.4.6. Actuarial Exam Question. Losses arising in a certain portfolio have a two-parameter Pareto distribution with $\alpha = 5$ and $\theta = 3,600$. A reinsurance arrangement has been made, under which (a) the reinsurer accepts 15% of losses up to $u = 5,000$ and all amounts in excess of 5,000 and (b) the insurer pays for the remaining losses.

a) Express the random variables for the reinsurer's and the insurer's payments as a function of X, the portfolio losses.
b) Calculate the mean amount paid on a single claim by the insurer.
c) By assuming that the upper limit is $u = \infty$, calculate an upper bound on the standard deviation of the amount paid on a single claim by the insurer (retaining the 15% copayment).

Show Example Solution

Solution.

a). The reinsurer's portion is

$$Y_{reinsurer} = \begin{cases} 0.15X & X < 5000, \\ 0.15(5000) + X - 5000 & X \geq 5000 \end{cases}.$$

and the insurer's portion is

$$Y_{insurer} = \begin{cases} 0.85X & X < 5000, \\ 0.85(5000) & X \geq 5000 \end{cases} = 0.85(X \wedge 5000).$$

b) Using the limited expected value tables for the Pareto distribution, we have

$$\mathrm{E}\, Y_{insurer} = 0.85\, \mathrm{E}\, (X \wedge 5000) = 0.85\, \frac{\theta}{\alpha - 1} \left[1 - \left(\frac{\theta}{5000 + \theta} \right)^{\alpha - 1} \right] = 0.85\, \frac{3600}{5 - 1} \left[1 - \left(\frac{3600}{5000 + 3600} \right)^{5-1} \right] = 741.510$$

c) For the first moment of the unlimited variable, we have

$$\mathrm{E}\, Y_{insurer}(u = \infty) = 0.85\, \mathrm{E}\, X = 0.85\, \frac{\theta}{\alpha - 1} = 0.85\, \frac{3600}{5 - 1} = 765.$$

For the second moment of the unlimited variable, we use the table of distributions to get

$$\mathrm{E}\, Y_{insurer}(u=\infty)^2 = 0.85^2\, \mathrm{E}\, X^2 = 0.85^2\, \frac{\theta^2\Gamma(2+1)\Gamma(\alpha-2)}{\Gamma(\alpha)} = 0.85^2\, \frac{3600^2 * 2 * 2}{24} = 1560600.$$

Thus, the variance is $1560600 - 765^2 = 975375$. Alternatively, you can use the formula

$$0.85^2\ \mathrm{Var}\ X = 0.85^2\, \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} = 0.85^2\, \frac{5(3600^2)}{(5-1)^2(5-2)} = 975375.$$

Taking square roots, the standard deviation is $\sqrt{975375} \approx 987.6108$.

Further discussions of reinsurance will be provided in Section **??**.

Show Quiz Solution

3.5 Maximum Likelihood Estimation

In this section, you learn how to:

- Define a likelihood for a sample of observations from a continuous distribution
- Define the maximum likelihood estimator for a random sample of observations from a continuous distribution
- Estimate parametric distributions based on grouped, censored, and truncated data

---

3.5.1 Maximum Likelihood Estimators for Complete Data

Up to this point, the chapter has focused on parametric distributions that are commonly used in insurance applications. However, to be useful in applied work, these distributions must use "realistic" values for the parameters and for this we turn to data. At a foundational level, we assume that the analyst has available a random sample $X_1, \ldots, X_n$ from a distribution with distribution

function F_X (for brevity, we sometimes drop the subscript X). As is common, we use the vector $\boldsymbol{\theta}$ to denote the set of parameters for F.This basic sample scheme is reviewed in Appendix Section **??**. Although basic, this sampling scheme provides the foundations for understanding more complex schemes that are regularly used in practice, and so it is important to master the basics.

Before a draw from a distribution, we consider potential outcomes summarized by the random variable X_i (here, i is 1, 2, ..., n). After the draw, we observe x_i. Notationally, we use upper case roman letters for random variables and lower case ones for realizations. We have seen this set-up already in Section **??**, where we used $\Pr(X_1 = x_1, \ldots, X_n = x_n)$ to quantify the "likelihood" of drawing a sample $\{x_1, \ldots, x_n\}$. With continuous data, we use the joint probability density function (pdf) instead of joint probabilities. With the independence assumption, the joint pdf may be written as the product of pdfs. Thus, we define the **likelihood** to be

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i). \tag{3.2}$$

From the notation, note that we consider this to be a function of the parameters in $\boldsymbol{\theta}$, with the data $\{x_1, \ldots, x_n\}$ held fixed. The maximum likelihood estimator is that value of the parameters in $\boldsymbol{\theta}$ that maximize $L(\boldsymbol{\theta})$.

From calculus, we know that maximizing a function produces the same results as maximizing the logarithm of a function (this is because the logarithm is a monotone, convex function). Because we get the same results, to ease computation considerations, it is common to consider the **logarithmic likelihood**, denoted as

$$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i). \tag{3.3}$$

Example 3.5.1. Actuarial Exam Question. You are given the following five observations: 521, 658, 702, 819, 1217. You use the single-parameter Pareto with cumulative distribution function:

$$F(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500.$$

With $n = 5$, the log-likelihood function is

$$l(\alpha|\mathbf{x}) = \sum_{i=1}^{5} \ln f(x_i; \alpha) = 5\alpha \ln 500 + 5 \ln \alpha - (\alpha + 1) \sum_{i=1}^{5} \ln x_i.$$

Figure **??** shows the logarithmic likelihood as a function of the parameter α.

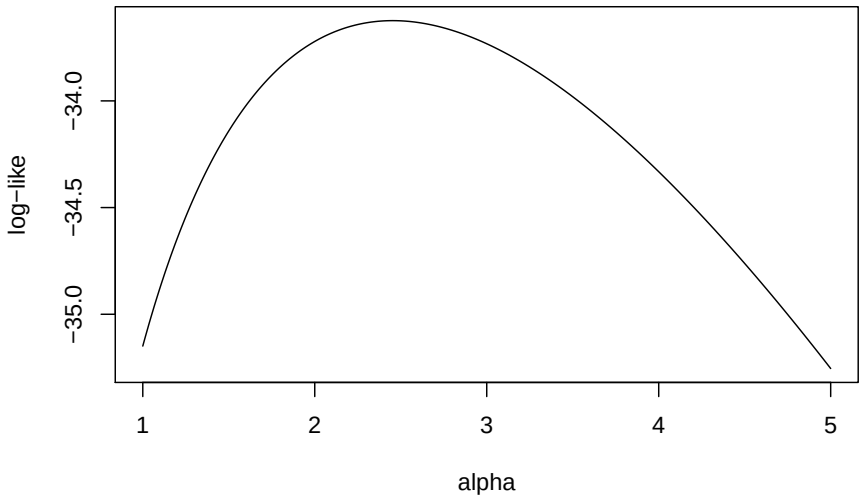


Figure 3.4: Logarithmic Likelihood for a One Parameter Pareto

We can determine the maximum value of the logarithmic likelihood by taking derivatives and setting it equal to zero. This yields

$$\frac{\partial}{\partial\alpha}l(\alpha|\mathbf{x}) = 5\ln 500 + 5/\alpha - \sum_{i=1}^{5}\ln x_i =_{set} 0 \Rightarrow \hat{\alpha}_{MLE} = \frac{5}{\sum_{i=1}^{5}\ln x_i - 5\ln 500} = 2.453.$$

Naturally, there are many problems where it is not practical to use hand calculations for optimization. Fortunately there are many statistical routines available such as the `R` function `optim`.

R Code for Optimization

```
c1 <- log(521)+log(658)+log(702)+log(819)+log(1217)
nloglike <- function(alpha){-(5*alpha*log(500)+5*log(alpha)-(alpha+1)*c1)}
MLE <- optim(par=1, fn=nloglike)$par
```

This code confirms our hand calculation result where the maximum likelihood estimator is $\alpha_{MLE} = 2.453125$.

We present a few additional examples to illustrate how actuaries fit a parametric distribution model to a set of claim data using maximum likelihood.

Example 3.5.2. Actuarial Exam Question. Consider a random sample of claim amounts: 8,000 10,000 12,000 15,000. You assume that claim amounts follow an inverse exponential distribution, with parameter θ. Calculate the maximum likelihood estimator for θ.

Show Example Solution

Solution.

The probability density function is

$$f_X(x) = \frac{\theta e^{-\frac{\theta}{x}}}{x^2},$$

where $x > 0$.

The likelihood function, $L(\theta)$, can be viewed as the probability of the observed data, written as a function of the model's parameter θ

$$L(\theta) = \prod_{i=1}^{4} f_{X_i}(x_i) = \frac{\theta^4 e^{-\theta \sum_{i=1}^{4} \frac{1}{x_i}}}{\prod_{i=1}^{4} x_i^2}.$$

The log-likelihood function, $\ln L(\theta)$, is the sum of the individual logarithms.

$$\ln L(\theta) = 4 \ln \theta - \theta \sum_{i=1}^{4} \frac{1}{x_i} - 2 \sum_{i=1}^{4} \ln x_i.$$

$$\frac{d \ln L(\theta)}{d\theta} = \frac{4}{\theta} - \sum_{i=1}^{4} \frac{1}{x_i}.$$

The maximum likelihood estimator of θ, denoted by $\hat{\theta}$, is the solution to the equation

$$\frac{4}{\hat{\theta}} - \sum_{i=1}^{4} \frac{1}{x_i} = 0.$$

Thus, $\hat{\theta} = \frac{4}{\sum_{i=1}^{4} \frac{1}{x_i}} = 10,667$

The second derivative of $\ln L(\theta)$ is given by

$$\frac{d^2 \ln L(\theta)}{d\theta^2} = \frac{-4}{\theta^2}.$$

Evaluating the second derivative of the loglikelihood function at $\hat{\theta} = 10,667$ gives a negative value, indicating $\hat{\theta}$ as the value that maximizes the loglikelihood function.

Example 3.5.3. Actuarial Exam Question. A random sample of size 6 is from a lognormal distribution with parameters μ and σ. The sample values are 200, 3,000, 8,000, 60,000, 60,000, 160,000. Calculate the maximum likelihood estimator for μ and σ.

Show Example Solution

Solution.

The probability density function is

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2,$$

where $x > 0$.

The likelihood function, $L(\mu, \sigma)$, is the product of the pdf for each data point.

$$L(\mu, \sigma) = \prod_{i=1}^{6} f_{X_i}(x_i) = \frac{1}{\sigma^6 (2\pi)^3 \prod_{i=1}^{6} x_i} exp - \frac{1}{2}\sum_{i=1}^{6}\left(\frac{\ln x_i - \mu}{\sigma}\right)^2.$$

The loglikelihood function, $\ln L(\mu, \sigma)$, is the sum of the individual logarithms.

$$\ln(\mu, \sigma) = -6\ln\sigma - 3\ln(2\pi) - \sum_{i=1}^{6}\ln x_i - \frac{1}{2}\sum_{i=1}^{6}\left(\frac{\ln x_i - \mu}{\sigma}\right)^2.$$

The first partial derivatives are

$$\frac{\partial \ln L(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{6}(\ln x_i - \mu).$$

$$\frac{\partial \ln L(\mu, \sigma)}{\partial \sigma} = \frac{-6}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{6}(\ln x_i - \mu)^2.$$

The maximum likelihood estimators of μ and σ, denoted by $\hat{\mu}$ and $\hat{\sigma}$, are the solutions to the equations

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{6}(\ln x_i - \hat{\mu}) = 0.$$

$$\frac{-6}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{6}(\ln x_i - \hat{\mu})^2 = 0.$$

These yield the estimates

$$\hat{\mu} = \frac{\sum_{i=1}^{6}\ln x_i}{6} = 9.38 \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{6}(\ln x_i - \hat{\mu})^2}{6} = 5.12.$$

.

The second partial derivatives are

$$\frac{\partial^2 \ln L(\mu,\sigma)}{\partial \mu^2} = \frac{-6}{\sigma^2}, \qquad \frac{\partial^2 \ln L(\mu,\sigma)}{\partial \mu \partial \sigma} = \frac{-2}{\sigma^3}\sum_{i=1}^{6}(\ln x_i - \mu)$$

and

$$\frac{\partial^2 \ln L(\mu,\sigma)}{\partial \sigma^2} = \frac{6}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{6}(\ln x_i - \mu)^2$$

.

Two follow-up questions rely on large sample properties that you may have seen in an earlier course. Appendix Chapter **??** reviews the definition of the likelihood function, introduces its properties, reviews the maximum likelihood estimators, extends their large-sample properties to the case where there are multiple parameters in the model, and reviews statistical inference based on maximum likelihood estimators. In the solutions of these examples we derive the asymptotic variance of maximum-likelihood estimators of the model parameters. We use the delta method to derive the asymptotic variances of functions of these parameters.

Example 3.5.2 - Follow - Up. Refer to **Example 3.5.2.**

Approximate the variance of the maximum likelihood estimator.

Determine an approximate 95% confidence interval for θ.

Determine an approximate 95% confidence interval for $\Pr(X \leq 9,000)$.

Show Example Solution

Solution.

a. Taking reciprocal of negative expectation of the second derivative of $\ln L(\theta)$, we obtain an estimate of the variance of $\hat{\theta}$, $\widehat{Var}\left(\hat{\theta}\right) = \left[E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right)\right]^{-1}\Big|_{\theta=\hat{\theta}} = \frac{\hat{\theta}^2}{4} = 28,446,222.$

It should be noted that as the sample size $n \to \infty$, the distribution of the maximum likelihood estimator $\hat{\theta}$ converges to a normal distribution with mean θ and variance $\hat{V}\left(\hat{\theta}\right)$. The approximate confidence interval in this example is based on the assumption of normality, despite the small sample size, only for the purpose of illustration.

b. The 95% confidence interval for θ is given by

$$10,667 \pm 1.96\sqrt{28,446,222} = (213.34,\ 21,120.66)\,.$$

c. The distribution function of X is $F(x) = 1 - e^{-\frac{x}{\theta}}$. Then, the maximum likelihood estimate of $g_\Theta(\theta) = F(9,000)$ is

$$g\left(\hat{\theta}\right) = 1 - e^{-\frac{9,000}{10,667}} = 0.57.$$

We use the delta method to approximate the variance of $g\left(\hat{\theta}\right)$.

$$\frac{\mathrm{dg}(\theta)}{d\theta} = -\frac{9,000}{\theta^2} e^{-\frac{9,000}{\theta}}.$$

$$\widehat{Var}\left[g\left(\hat{\theta}\right)\right] = \left(-\tfrac{9,000}{\hat{\theta}^2} e^{-\frac{9,000}{\hat{\theta}}}\right)^2 \hat{V}\left(\hat{\theta}\right) = 0.0329.$$

The 95% confidence interval for $F(9,000)$ is given by

$$0.57 \pm 1.96\sqrt{0.0329} = (0.214,\ 0.926)\,.$$

Example 3.5.3 - Follow - Up. Refer to **Example 3.5.3.**

Estimate the covariance matrixmatrix where the (i,j)th element represents the covariance between the ith and jth random variables of the maximum likelihood estimator.

Determine approximate 95% confidence intervals for μ and σ.

Determine an approximate 95% confidence interval for the mean of the lognormal distribution.

Show Example Solution

a. To derive the covariance matrix of the mle we need to find the expectations of the second derivatives. Since the random variable X is from a lognormal distribution with parameters μ and σ, then lnX is normally distributed with mean μ and variance σ^2.

$\mathrm{E}\left(\frac{\partial^2 \mathrm{lnL}(\mu,\sigma)}{\partial\mu^2}\right) = \mathrm{E}\left(\frac{-6}{\sigma^2}\right) = \frac{-6}{\sigma^2}$,

$\mathrm{E}\left(\frac{\partial^2 \mathrm{lnL}(\mu,\sigma)}{\partial\mu\partial\sigma}\right) = \frac{-2}{\sigma^3}\sum_{i=1}^{6}\mathrm{E}(\ln x_i - \mu) = \frac{-2}{\sigma^3}\sum_{i=1}^{6}[\mathrm{E}(\ln x_i) - \mu] = \frac{-2}{\sigma^3}\sum_{i=1}^{6}(\mu - \mu) = 0$,

and

$\mathrm{E}\left(\frac{\partial^2 \mathrm{lnL}(\mu,\sigma)}{\partial\sigma^2}\right) = \frac{6}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{6}\mathrm{E}(\ln x_i - \mu)^2 = \frac{6}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{6}\mathrm{V}(\ln x_i) = \frac{6}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{6}\sigma^2 = \frac{-12}{\sigma^2}$.

Using the negatives of these expectations we obtain the Fisher information matrix

$$\begin{bmatrix} \frac{6}{\sigma^2} & 0 \\ 0 & \frac{12}{\sigma^2} \end{bmatrix}.$$

The covariance matrix, Σ, is the inverse of the Fisher information matrix

$$\Sigma = \begin{bmatrix} \frac{\sigma^2}{6} & 0 \\ 0 & \frac{\sigma^2}{12} \end{bmatrix}.$$

The estimated matrix is given by

$$\hat{\Sigma} = \begin{bmatrix} 0.8533 & 0 \\ 0 & 0.4267 \end{bmatrix}.$$

b. The 95% confidence interval for μ is given by $9.38 \pm 1.96\sqrt{0.8533} = (7.57,\ 11.19)$.

The 95% confidence interval for σ^2 is given by $5.12 \pm 1.96\sqrt{0.4267} = (3.84,\ 6.40)$.

c. The mean of X is $\exp\left(\mu + \frac{\sigma^2}{2}\right)$. Then, the maximum likelihood estimate of

$$g(\mu, \sigma) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

is

$$g(\hat{\mu}, \hat{\sigma}) = \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) = 153,277.$$

We use the delta method to approximate the variance of the mle $g(\hat{\mu}, \hat{\sigma})$.

$\frac{\partial g(\mu,\sigma)}{\partial \mu} = exp\left(\mu + \frac{\sigma^2}{2}\right)$ and $\frac{\partial g(\mu,\sigma)}{\partial \sigma} = \sigma exp\left(\mu + \frac{\sigma^2}{2}\right)$.

Using the delta method, the approximate variance of $g(\hat{\mu}, \hat{\sigma})$ is given by

$$\hat{V}(g(\hat{\mu}, \hat{\sigma})) = \begin{bmatrix} \frac{\partial g(\mu,\sigma)}{\partial \mu} & \frac{\partial g(\mu,\sigma)}{\partial \sigma} \end{bmatrix} \Sigma \begin{bmatrix} \frac{\partial g(\mu,\sigma)}{\partial \mu} \\ \frac{\partial g(\mu,\sigma)}{\partial \sigma} \end{bmatrix} \Bigg|_{\mu=\hat{\mu},\sigma=\hat{\sigma}}$$

$$= \begin{bmatrix} 153,277 & 346,826 \end{bmatrix} \begin{bmatrix} 0.8533 & 0 \\ 0 & 0.4267 \end{bmatrix} \begin{bmatrix} 153,277 \\ 346,826 \end{bmatrix} =$$

71,374,380,000

The 95% confidence interval for $\exp\left(\mu + \frac{\sigma^2}{2}\right)$ is given by

$153,277 \pm 1.96\sqrt{71,374,380,000} = (-370,356,\ 676,910)$.

Since the mean of the lognormal distribution cannot be negative, we should replace the negative lower limit in the previous interval by a zero.

Example 3.5.4. Wisconsin Property Fund. To see how maximum likelihood estimators work with real data, we return to the 2010 claims data introduced in Section **??**.

The following snippet of code shows how to fit the exponential, gamma, Pareto, lognormal, and GB2 models. For consistency, the code employs the `R` package `VGAM`. The acronym stands for Vector Generalized Linear and Additive Models; as suggested by the name, this package can do far more than fit these models although it suffices for our purposes. The one exception is the GB2 density which is not widely used outside of insurance applications; however, we can code this density and compute maximum likelihood estimators using the `optim` general purpose optimizer.

Show Example Solution

```
library(VGAM)
claim_lev <- read.csv("Data/CLAIMLEVEL.csv", header = TRUE)
claim_data <- subset(claim_lev, Year == 2010);

# Inference assuming a GB2 Distribution - this is more complicated
# The likelihood functon of GB2 distribution (negative for optimization)
lik_gb2 <- function (param) {
  a_1 <- param[1]
  a_2 <- param[2]
  mu <- param[3]
  sigma <- param[4]
  yt <- (log(claim_data$Claim) - mu) / sigma
  logexpyt <- ifelse(yt > 23, yt, log(1 + exp(yt)))
  logdens <- a_1 * yt - log(sigma) - log(beta(a_1,a_2)) -
    (a_1+a_2) * logexpyt - log(claim_data$Claim)
  return(-sum(logdens))
}
# "optim" is a general purpose minimization function
gb2_bop <- optim(c(1, 1, 0, 1), lik_gb2, method = c("L-BFGS-B"),
                 lower = c(0.01, 0.01, -500, 0.01),
                 upper = c(500, 500, 500, 500), hessian = TRUE)
# Nonparametric Plot
plot(density(log(claim_data$Claim)), main = "", xlab = "Log Expenditures",
     ylim = c(0 ,0.37))
x <- seq(0, 15, by = 0.01)
#Exponential
fit.exp <- vglm(Claim ~ 1, exponential, data = claim_data)
theta = 1 / exp(coef(fit.exp))
fexp_ex <- dgamma(exp(x), scale = exp(-coef(fit.exp)), shape = 1) * exp(x)
lines(x, fexp_ex, col = "red", lty =2)
```

```
# Inference assuming a gamma distribution
fit.gamma <- vglm(Claim ~ 1, family = gamma2, data = claim_data)
theta <- exp(coef(fit.gamma)[1]) / exp(coef(fit.gamma)[2])  # theta = mu / alpha
alpha <- exp(coef(fit.gamma)[2])
fgamma_ex <- dgamma(exp(x), shape = alpha, scale = theta) * exp(x)
lines(x, fgamma_ex, col = "blue", lty =3)
#Pareto
fit.pareto <- vglm(Claim ~ 1, paretoII, loc = 0, data = claim_data)
fpareto_ex <- dparetoII(exp(x), loc = 0, shape = exp(coef(fit.pareto)[2]),
                        scale = exp(coef(fit.pareto)[1])) * exp(x)
lines(x, fpareto_ex, col = "purple")
# Lognormal
fit.LN <- vglm(Claim ~ 1, family = lognormal, data = claim_data)
flnorm_ex <- dlnorm(exp(x), mean = coef(fit.LN)[1],
                    sd = exp(coef(fit.LN)[2])) * exp(x)
lines(x, flnorm_ex, col = "lightblue")
# Density for GB II
gb2_density <- function (x) {
  a_1 <- gb2_bop$par[1]
  a_2 <- gb2_bop$par[2]
  mu <- gb2_bop$par[3]
  sigma <- gb2_bop$par[4]
  xt <- (log(x) - mu) / sigma
  logexpxt <- ifelse (xt > 23, yt, log(1 + exp(xt)))
  logdens <- a_1 * xt - log(sigma) - log(beta(a_1, a_2)) -
    (a_1+a_2) * logexpxt -log(x)
  exp(logdens)
  }
fGB2_ex = gb2_density(exp(x)) * exp(x)
lines(x, fGB2_ex, col="green")
legend("topleft", c("log(Expend)", "Exponential", "Gamma", "Pareto",
                    "Lognormal", "GB2"), cex=0.8,
       lty = c(4,2,3,1,1,1), #4 is "longdash"
       col = c("black","red","blue","purple","lightblue","green"))
```

Results from the fitting exercise are summarized in Figure **??**. Here, the black "longdash" curve is a smoothed histogram of the actual data (that we will introduce in Section **??**); the other curves are parametric curves where the parameters are computed via maximum likelihood. We see poor fits in the red dashed line from the exponential distribution fit and the blue dotted line from the gamma distribution fit. Fits of the other curves, Pareto, lognormal, and GB2, all seem to provide reasonably good fits to the actual data. Chapter **??** describes in more detail the principles of model selection.

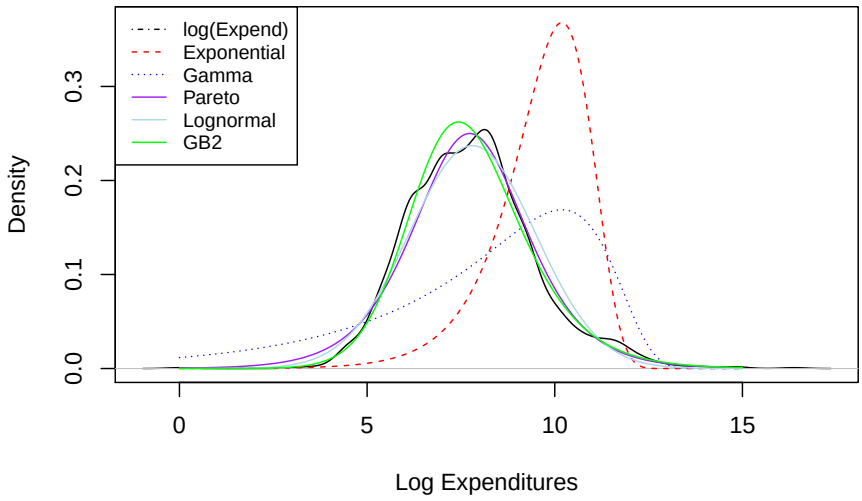


Figure 3.5: Density Comparisons for the Wisconsin Property Fund

3.5.2 Maximum Likelihood Estimators using Modified Data

In many applications, actuaries and other analysts wish to estimate model parameters based on individual data that are not limited. However, there are also important applications when only limited, or modified, data are available. This section introduces maximum likelihood estimation for grouped, censored, and truncated data. Later, we will follow up with additional details in Section **??**.

Maximum Likelihood Estimators for Grouped Data

In the previous section we considered the maximum likelihood estimation of continuous models from complete (individual) data. Each individual observation is recorded, and its contribution to the likelihood function is the density at that value. In this section we consider the problem of obtaining maximum likelihood estimates of parameters from grouped datadata bucketed into categories with ranges, such as for use in histograms or frequency tables. The observations are only available in grouped form, and the contribution of each observation to the likelihood function is the probability of falling in a specific group (interval). Let n_j represent the number of observations in the interval $(\ c_{j-1}, c_j]$ The grouped

data likelihood function is thus given by

$$L(\theta) = \prod_{j=1}^{k} \left[F_X(\ c_j|\,\theta) - F_X(\ c_{j-1}|\,\theta)\right]^{n_j},$$

where c_0 is the smallest possible observation (often set to zero) and c_k is the largest possible observation (often set to infinity).

Example 3.5.5. Actuarial Exam Question. For a group of policies, you are given that losses follow the distribution function $F_X(x) = 1 - \frac{\theta}{x}$, for $\theta < x < \infty$. Further, a sample of 20 losses resulted in the following:

Interval	Number of Losses
$(\theta, 10]$	9
$(10, 25]$	6
$(25, \infty)$	5

Calculate the maximum likelihood estimate of θ.

Show Example Solution

Solution.

The contribution of each of the 9 observations in the first interval to the likelihood function is the probability of $X \leq 10$; that is, $\Pr(X \leq 10) = F_X(10)$. Similarly, the contributions of each of 6 and 5 observations in the second and third intervals are $\Pr(10 < X \leq 25) = F_X(25) - F_X(10)$ and $P(X > 25) = 1 - F_X(25)$, respectively. The likelihood function is thus given by

$$\begin{aligned} L(\theta) &= \left[F_X(10)\right]^9 \left[F_X(25) - F_X(10)\right]^6 \left[1 - F_X(25)\right]^5 \\ &= \left(1 - \frac{\theta}{10}\right)^9 \left(\frac{\theta}{10} - \frac{\theta}{25}\right)^6 \left(\frac{\theta}{25}\right)^5 \\ &= \left(\frac{10-\theta}{10}\right)^9 \left(\frac{15\theta}{250}\right)^6 \left(\frac{\theta}{25}\right)^5. \end{aligned}$$

Then, $\ln L(\theta) = 9\ln(10-\theta) + 6\ln\theta + 5\ln\theta - 9\ln 10 + 6\ln 15 - 6\ln 250 - 5\ln 25$.

$$\frac{d\ln L(\theta)}{d\theta} = \frac{-9}{(10-\theta)} + \frac{6}{\theta} + \frac{5}{\theta}.$$

The maximum likelihood estimator, $\hat{\theta}$, is the solution to the equation

$$\frac{-9}{\left(10-\hat{\theta}\right)} + \frac{11}{\hat{\theta}} = 0$$

and $\hat{\theta} = 5.5$.

Maximum Likelihood Estimators for Censored Data

Another possible distinguishing feature of a data gathering mechanism is censoring. While for some events of interest (losses, claims, lifetimes, etc.) the complete data maybe available, for others only partial information is available; all that may be known is that the observation exceeds a specific value. The limited policy introduced in Section **??** is an example of right censoring. Any loss greater than or equal to the policy limit is recorded at the limit. The contribution of the censored observation to the likelihood function is the probability of the random variable exceeding this specific limit. Note that contributions of both complete and censored data share the survivor function, for a complete point this survivor function is multiplied by the hazard function, but for a censored observation it is not. The likelihood function for censored data is then given by

$$L(\theta) = \left[\prod_{i=1}^{r} f_X(x_i)\right] [S_X(u)]^m,$$

where r is the number of known loss amounts below the limit u and m is the number of loss amounts larger than the limit u.

Example 3.5.6. Actuarial Exam Question. The random variable X has survival function:

$$S_X(x) = \frac{\theta^4}{\left(\theta^2 + x^2\right)^2}.$$

Two values of X are observed to be 2 and 4. One other value exceeds 4. Calculate the maximum likelihood estimate of θ.

Show Example Solution

Solution.

The contributions of the two observations 2 and 4 are $f_X(2)$ and $f_X(4)$ respectively. The contribution of the third observation, which is only known to exceed 4 is $S_X(4)$. The likelihood function is thus given by

$$L(\theta) = f_X(2) f_X(4) S_X(4).$$

The probability density function of X is given by

$$f_X(x) = \frac{4x\theta^4}{\left(\theta^2 + x^2\right)^3}.$$

Thus,

$$L(\theta) = \frac{8\theta^4}{\left(\theta^2+4\right)^3} \frac{16\theta^4}{\left(\theta^2+16\right)^3} \frac{\theta^4}{\left(\theta^2+16\right)^2} = \frac{128\theta^{12}}{\left(\theta^2+4\right)^3 \left(\theta^2+16\right)^5},$$

So,

$$\ln L(\theta) = \ln 128 + 12\ln\theta - 3\ln\left(\theta^2+4\right) - 5\ln\left(\theta^2+16\right),$$

and

$\frac{d\ln L(\theta)}{d\theta} = \frac{12}{\theta} - \frac{6\theta}{(\theta^2+4)} - \frac{10\theta}{(\theta^2+16)}$.

The maximum likelihood estimator, $\hat{\theta}$, is the solution to the equation

$$\frac{12}{\hat{\theta}} - \frac{6\hat{\theta}}{\left(\hat{\theta}^2+4\right)} - \frac{10\hat{\theta}}{\left(\hat{\theta}^2+16\right)} = 0$$

or

$$12\left(\hat{\theta}^2+4\right)\left(\hat{\theta}^2+16\right) - 6\hat{\theta}^2\left(\hat{\theta}^2+16\right) - 10\hat{\theta}^2\left(\hat{\theta}^2+4\right) = -4\hat{\theta}^4 + 104\hat{\theta}^2 + 768 = 0,$$

which yields $\hat{\theta}^2 = 32$ and $\hat{\theta} = 5.7$.

Maximum Likelihood Estimators for Truncated Data

This section is concerned with the maximum likelihood estimation of the continuous distribution of the random variable X when the data is incomplete due to truncation. If the values of X are truncated at d, then it should be noted that we would not have been aware of the existence of these values had they not exceeded d. The policy deductible introduced in Section **??** is an example of left truncation. Any loss less than or equal to the deductible is not recorded. The contribution to the likelihood function of an observation x truncated at d will be a conditional probability and the $f_X(x)$ will be replaced by $\frac{f_X(x)}{S_X(d)}$. The likelihood function for truncated data is then given by

$$L(\theta) = \prod_{i=1}^{k} \frac{f_X(x_i)}{S_X(d)},$$

where k is the number of loss amounts larger than the deductible d.

Example 3.5.7. Actuarial Exam Question. For the single parameter Pareto distribution with $\theta = 2$, maximum likelihood estimation is applied to estimate the parameter α. Find the estimated mean of the ground up loss distribution based on the maximum likelihood estimate of α for the following data set:

Ordinary policy deductible of 5, maximum covered loss of 25 (policy limit 20)

8 insurance payment amounts: 2, 4, 5, 5, 8, 10, 12, 15

2 limit payments: 20, 20.

Show Example Solution

Solution.

The contributions of the different observations can be summarized as follows:

For the exact loss: $f_X(x)$

For censored observations: $S_X(25)$.

For truncated observations: $\frac{f_X(x)}{S_X(5)}$.

Given that ground up losses smaller than 5 are omitted from the data set, the contribution of all observations should be conditional on exceeding 5. The likelihood function becomes

$$L(\alpha) = \frac{\prod_{i=1}^{8} f_X(x_i)}{[S_X(5)]^8} \left[\frac{S_X(25)}{S_X(5)}\right]^2.$$

For the single parameter Pareto the probability density and distribution functions are given by

$$f_X(x) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}} \quad \text{and} \quad F_X(x) = 1 - \left(\frac{\theta}{x}\right)^\alpha,$$

for $x > \theta$, respectively. Then, the likelihood and loglikelihood functions are given by

$$L(\alpha) = \frac{\alpha^8}{\prod_{i=1}^{8} x_i^{\alpha+1}} \frac{5^{10\alpha}}{25^{2\alpha}},$$

$$\ln L(\alpha) = 8\ln\alpha - (\alpha+1)\sum_{i=1}^{8} \ln x_i + 10\alpha\ln 5 - 2\alpha\ln 25.$$

$\frac{d\ln L(\alpha)}{d\theta} = \frac{8}{\alpha} - \sum_{i=1}^{8} \ln x_i + 10\ln 5 - 2\ln 25.$

The maximum likelihood estimator, $\hat{\alpha}$, is the solution to the equation

$$\frac{8}{\hat{\alpha}} - \sum_{i=1}^{8} \ln x_i + 10\ln 5 - 2\ln 25 = 0,$$

which yields

$$\hat{\alpha} = \frac{8}{\sum_{i=1}^{8} \ln x_i - 10\ln 5 + 2\ln 25} = \frac{8}{(\ln 7 + \ln 9 + \cdots + \ln 20) - 10\ln 5 + 2\ln 25} = 0.785.$$

The mean of the Pareto only exists for $\alpha > 1$. Since $\hat{\alpha} = 0.785 < 1$. Then, the mean does not exist.

Show Quiz Solution

3.6 Further Resources and Contributors

Contributors

- **Zeinab Amin**, The American University in Cairo, is the principal author of this chapter. Email: zeinabha@aucegypt.edu for chapter comments and suggested improvements.
- Many helpful comments have been provided by Hirokazu (Iwahiro) Iwasawa, iwahiro@bb.mbn.or.jp .
- Other chapter reviewers include: Rob Erhardt, Jorge Yslas, Tatjana Miljkovic, and Samuel Kolins.

Exercises

Here are a set of exercises that guide the viewer through some of the theoretical foundations of **Loss Data Analytics**. Each tutorial is based on one or more questions from the professional actuarial examinations – typically the Society of Actuaries Exam C.

Severity Distribution Guided Tutorials

Further Readings and References

Notable contributions include: **?**, **?**, **?**, **?**, **?**, **?**, **?**, and **?**.

Chapter 4

Model Selection and Estimation

Chapter Preview. Chapters **??** and **??** have described how to fit parametric models to frequency and severity data, respectively. This chapter begins with the selection of models. To compare alternative parametric models, it is helpful to summarize data without reference to a specific parametric distribution. Section **??** describes nonparametric estimation, how we can use it for model comparisons and how it can be used to provide starting values for parametric procedures. The process of model selection is then summarized in Section **??**. Although our focus is on continuous data, the same process can be used for discrete data or data that come from a hybrid combination of discrete and continuous data.

Model selection and estimation are fundamental aspects of statistical modeling. To provide a flavor as to how they can be adapted to alternative sampling schemes, Section **??** describes estimation for grouped, censored and truncated data (following the Section **??** introduction). To see how they can be adapted to alternative models, the chapter closes with Section **??** on Bayesian inference, an alternative procedure where the (typically unknown) parameters are treated as random variables.

4.1 Nonparametric Inference

In this section, you learn how to:

- Estimate moments, quantiles, and distributions without reference to a parametric distribution

- Summarize the data graphically without reference to a parametric distribution
- Determine measures that summarize deviations of a parametric from a nonparametric fit
- Use nonparametric estimators to approximate parameters that can be used to start a parametric estimation procedure

---

4.1.1 Nonparametric Estimation

In Section **??** for frequency and Section **??** for severity, we learned how to summarize a distribution by computing means, variances, quantiles/percentiles, and so on. To approximate these summary measures using a dataset, one strategy is to:

i. assume a parametric form for a distribution, such as a negative binomial for frequency or a gamma distribution for severity,
ii. estimate the parameters of that distribution, and then
iii. use the distribution with the estimated parameters to calculate the desired summary measure.

This is the **parametric** approach. Another strategy is to estimate the desired summary measure directly from the observations without reference to a parametric model. Not surprisingly, this is known as the nonparametricAn approach to inference that does not rely on references to a parametric model. approach.

Let us start by considering the most basic type of sampling scheme and assume that observations are realizations from a set of random variables $X_1, \ldots, X_n$ that are iidIdentically and independently distributed. draws from an unknown population distribution $F(\cdot)$. An equivalent way of saying this is that $X_1, \ldots, X_n$, is a random sample (with replacement) from $F(\cdot)$. To see how this works, we now describe nonparametric estimators of many important measures that summarize a distribution.

Moment Estimators

We learned how to define moments in Section **??** for frequency and Section **??** for severity. In particular, the k-th moment, $\mathrm{E}\left[X^k\right] = \mu'_k$, summarizes many aspects of the distribution for different choices of k. Here, μ'_k is sometimes called the kth population moment to distinguish it from the kth sample moment,

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

which is the corresponding nonparametric estimator. In typical applications, k is a positive integer, although it need not be.

An important special case is the first moment where k=1. In this case, the prime symbol ($\prime$) and the 1 subscript are usually dropped and one uses $\mu = \mu_1'$ to denote the population mean, or simply the mean. The corresponding sample estimator for μ is called the sample mean, denoted with a bar on top of the random variable:

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

Another type of summary measure of interest is the the k-th central moment, E $[(X-\mu)^k] = \mu_k$. (Sometimes, μ_k' is called the k-th raw moment to distinguish it from the central moment μ_k.). A nonparametric, or sample, estimator of μ_k is

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^k.$$

The second central moment ($k = 2$) is an important case for which we typically assign a new symbol, $\sigma^2 =$ E $[(X-\mu)^2]$, known as the variance. Properties of sample moment estimator of the variance such as $n^{-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ have been studied extensively and so it is natural that many variations have been proposed. The most widely used variation is one where the effective sample size is reduced by one, and so we define

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

Here, the statistic s^2 known as the sample variance. Dividing by n-1 instead of n matters little when you have a sample size n in the thousands as is common in insurance applications. Still, the resulting estimator is unbiased in the sense that E $s^2 = \sigma^2$, a desirable property particularly when interpreting results of an analysis.

Empirical Distribution Function

We have seen how to compute nonparametric estimators of the kth moment E X^k. In the same way, for any known function g$(\cdot)$, we can estimate E g(X) using $n^{-1}\sum_{i=1}^{n}$ g(X_i).

Now suppose that we fix a value of x and consider the function g$(X) = I(X \le x)$. Here, the notation $I(\cdot)$ is the indicator function; it returns 1 if the event $(\cdot)$ is true and 0 otherwise. For this choice of g$(\cdot)$, the expected value is E $I(X \le x) = \Pr(X \le x) = F(x)$, the distribution function evaluated at a fixed point x. Using the analog principle, we define the nonparametric estimator of the distribution function

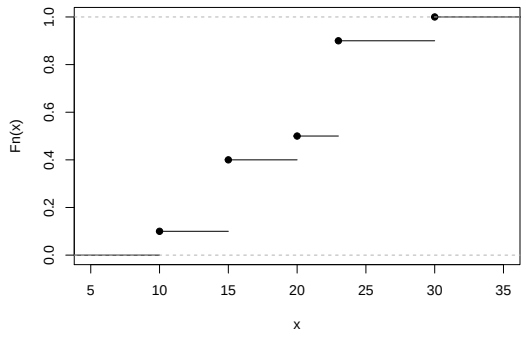


Figure 4.1: Empirical Distribution Function of a Toy Example

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right)$$
$$= \frac{\text{number of observations less than or equal to } x}{n}.$$

As a nonparametric estimator, $F_n(\cdot)$ is based on only observations and does not assume a parametric family for the distribution, it is also known as the **empirical distribution function**.

Example 4.1.1. Toy Data Set. To illustrate, consider a fictitious, or "toy," data set of $n = 10$ observations. Determine the empirical distribution function.

i	1	2	3	4	5	6	7	8	9	10
X_i	10	15	15	15	20	23	23	23	23	30

Show Example Solution

You should check that the sample mean is $\bar{X} = 19.7$ and that the sample variance is $s^2 = 34.45556$. The corresponding empirical distribution function is

$$F_n(x) = \begin{cases} 0 & \text{for } x < 10 \\ 0.1 & \text{for } 10 \leq x < 15 \\ 0.4 & \text{for } 15 \leq x < 20 \\ 0.5 & \text{for } 20 \leq x < 23 \\ 0.9 & \text{for } 23 \leq x < 30 \\ 1 & \text{for } x \geq 30, \end{cases}$$

which is shown in the following graph in Figure **??**.

Show R Code

```
(xExample <- c(10,rep(15,3),20,rep(23,4),30))
PercentilesxExample <- ecdf(xExample)
plot(PercentilesxExample, main="",xlab="x")
```

Quartiles, Percentiles and Quantiles

We have already seen the median, which is the number such that approximately half of a data set is below (or above) it. The **first quartile** is the number such that approximately 25% of the data is below it and the third quartile is the number such that approximately 75% of the data is below it. A $100p$ **percentile** is the number such that $100 \times p$ percent of the data is below it.

To generalize this concept, consider a distribution function $F(\cdot)$, which may or may not be continuous, and let q be a fraction so that $0 < q < 1$. We want to define a quantile, say q_F, to be a number such that $F(q_F) \approx q$. Notice that when $q = 0.5$, q_F is the median; when $q = 0.25$, q_F is the first quartile, and so on. So, a quantile generalizes the concepts of median, quartiles, and percentiles.

To be precise, for a given $0 < q < 1$, define the q**th quantile** q_F to be any number that satisfies

$$F(q_F-) \leq q \leq F(q_F) \tag{4.1}$$

Here, the notation $F(x-)$ means to evaluate the function $F(\cdot)$ as a left-hand limit.

To get a better understanding of this definition, let us look at a few special cases. First, consider the case where X is a continuous random variable so that the distribution function $F(\cdot)$ has no jump points, as illustrated in Figure **??**. In this figure, a few fractions, q_1, q_2, and q_3 are shown with their corresponding quantiles $q_{F,1}$, $q_{F,2}$, and $q_{F,3}$. In each case, it can be seen that $F(q_F-) = F(q_F)$ so that there is a unique quantile. Because we can find a unique inverse of the distribution function at any $0 < q < 1$, we can write $q_F = F^{-1}(q)$.

Figure **??** shows three cases for distribution functions. The left panel corresponds to the continuous case just discussed. The middle panel displays a jump point similar to those we already saw in the empirical distribution function of Figure **??**. For the value of q shown in this panel, we still have a unique value of the quantile q_F. Even though there are many values of q such that $F(q_F-) \leq q \leq F(q_F)$, for a particular value of q, there is only one solution to equation (**??**). The right panel depicts a situation in which the quantile cannot be uniquely determined for the q shown as there is a range of q_F's satisfying equation (**??**).

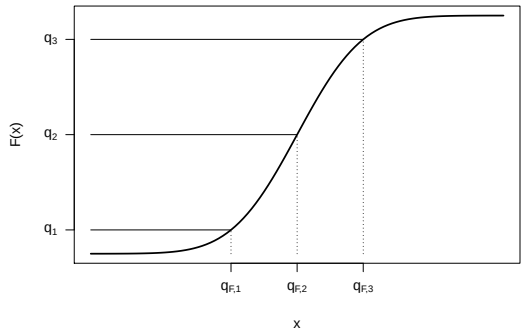


Figure 4.2: Continuous Quantile Case

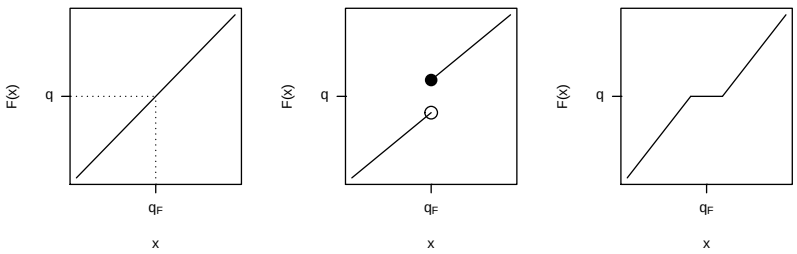


Figure 4.3: Three Quantile Cases

Example 4.1.2. Toy Data Set: Continued. Determine quantiles corresponding to the 20th, 50th, and 95th percentiles.

Show Example Solution

Solution. Consider Figure **??**. The case of $q = 0.20$ corresponds to the middle panel, so the 20th percentile is 15. The case of $q = 0.50$ corresponds to the right panel, so the median is any number between 20 and 23 inclusive. Many software packages use the average 21.5 (e.g. `R`, as seen below). For the 95th percentile, the solution is 30. We can see from the graph that 30 also corresponds to the 99th and the 99.99th percentiles.

```
quantile(xExample, probs=c(0.2, 0.5, 0.95), type=6)
```

```
##  20%  50%  95%
## 15.0 21.5 30.0
```

By taking a weighted average between data observations, smoothed empirical quantiles can handle cases such as the right panel in Figure **??**. The qth smoothed empirical quantile is defined as

$$\hat{\pi}_q = (1-h)X_{(j)} + hX_{(j+1)}$$

where $j = \lfloor (n+1)q \rfloor$, $h = (n+1)q - j$, and $X_{(1)}, \ldots, X_{(n)}$ are the ordered values (known as the order statistics) corresponding to $X_1, \ldots, X_n$. Note that $\hat{\pi}_q$ is simply a linear interpolation between $X_{(j)}$ and $X_{(j+1)}$.

Example 4.1.3. Toy Data Set: Continued. Determine the 50th and 20th smoothed percentiles.

Show Example Solution

Solution Take $n = 10$ and $q = 0.5$. Then, $j = \lfloor (11)0.5 \rfloor = \lfloor 5.5 \rfloor = 5$ and $h = (11)(0.5) - 5 = 0.5$. Then the 0.5-th smoothed empirical quantile is

$$\hat{\pi}_{0.5} = (1-0.5)X_{(5)} + (0.5)X_{(6)} = 0.5(20) + (0.5)(23) = 21.5.$$

Now take $n = 10$ and $q = 0.2$. In this case, $j = \lfloor (11)0.2 \rfloor = \lfloor 2.2 \rfloor = 2$ and $h = (11)(0.2) - 2 = 0.2$. Then the 0.2-th smoothed empirical quantile is

$$\hat{\pi}_{0.2} = (1-0.2)X_{(2)} + (0.2)X_{(3)} = 0.2(15) + (0.8)(15) = 15.$$

Density Estimators

Discrete Variable. When the random variable is discrete, estimating the probability mass function $f(x) = \Pr(X = x)$ is straightforward. We simply use the sample average, defined to be

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = x).$$

Continuous Variable within a Group. For a continuous random variable, consider a discretized formulation in which the domain of $F(\cdot)$ is partitioned by constants $\{c_0 < c_1 < \cdots < c_k\}$ into intervals of the form $[c_{j-1}, c_j)$, for $j = 1, \ldots, k$. The data observations are thus "grouped" by the intervals into which they fall. Then, we might use the basic definition of the empirical mass function, or a variation such as

$$f_n(x) = \frac{n_j}{n \times (c_j - c_{j-1})} \qquad c_{j-1} \le x < c_j,$$

where n_j is the number of observations (X_i) that fall into the interval $[c_{j-1}, c_j)$.

Continuous Variable (not grouped). Extending this notion to instances where we observe individual data, note that we can always create arbitrary groupings and use this formula. More formally, let $b > 0$ be a small positive constant, known as a **bandwidth**, and define a density estimator to be

$$f_n(x) = \frac{1}{2nb}\sum_{i=1}^{n} I(x - b < X_i \le x + b) \tag{4.2}$$

Show A Snippet of Theory

The idea is that the estimator $f_n(x)$ in equation (**??**) is the average over n iidIdentically and independently distributed. realizations of a random variable with mean

$$\begin{aligned}
\mathrm{E}\ \frac{1}{2b} I(x - b < X \le x + b) &= \frac{1}{2b}\left(F(x + b) - F(x - b)\right) \\
&= \frac{1}{2b}\left(\left\{F(x) + bF'(x) + b^2 C_1\right\}\left\{F(x) - bF'(x) + b^2 C_2\right\}\right) \\
&= F'(x) + b\frac{C_1 - C_2}{2} \to F'(x) = f(x),
\end{aligned}$$

as $b \to 0$. That is, $f_n(x)$ is an asymptotically unbiased estimator of $f(x)$ (its expectation approaches the true value as sample size increases to infinity). This development assumes some smoothness of $F(\cdot)$, in particular, twice differentiability at x, but makes no assumptions on the form of the distribution function F. Because of this, the density estimator f_n is said to be nonparametric.

More generally, define the **kernel density estimator** of the pdfprobability density function at x as

$$f_n(x) = \frac{1}{nb} \sum_{i=1}^{n} w\left(\frac{x - X_i}{b}\right), \tag{4.3}$$

where w is a probability density function centered about 0. Note that equation (**??**) simply becomes the kernel density estimator where $w(x) = \frac{1}{2} I(-1 < x \leq 1)$, also known as the uniform kernel. Other popular choices are shown in Table 4.1.

Table 4.1: Popular Choices for the Kernel Density Estimator

Kernel	$w(x)$
Uniform	$\frac{1}{2} I(-1 < x \leq 1)$
Triangle	$(1 - \|x\|) \times I(\|x\| \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - x^2) \times I(\|x\| \leq 1)$
Gaussian	$\phi(x)$

Here, $\phi(\cdot)$ is the standard normal density function. As we will see in the following example, the choice of bandwidth b comes with a bias-variance tradeoff between matching local distributional features and reducing the volatility.

Example 4.1.4. Property Fund. Figure **??** shows a histogram (with shaded gray rectangles) of logarithmic property claims from 2010. The (blue) thick curve represents a Gaussian kernel density where the bandwidth was selected automatically using an ad hoc rule based on the sample size and volatility of the data. For this dataset, the bandwidth turned out to be $b = 0.3255$. For comparison, the (red) dashed curve represents the density estimator with a bandwidth equal to 0.1 and the green smooth curve uses a bandwidth of 1. As anticipated, the smaller bandwidth (0.1) indicates taking local averages over less data so that we get a better idea of the local average, but at the price of higher volatility. In contrast, the larger bandwidth (1) smooths out local fluctuations, yielding a smoother curve that may miss perturbations in the local average. For actuarial applications, we mainly use the kernel density estimator to get a quick visual impression of the data. From this perspective, you can simply use the default ad hoc rule for bandwidth selection, knowing that you have the ability to change it depending on the situation at hand.

Show R Code

```
#Density Comparison
hist(log(ClaimData$Claim), main="", ylim=c(0,.35),xlab="Log Expenditures", freq=FALSE, col="light
lines(density(log(ClaimData$Claim)), col="blue",lwd=2.5)
lines(density(log(ClaimData$Claim), bw=1), col="green")
lines(density(log(ClaimData$Claim), bw=.1), col="red", lty=3)
```

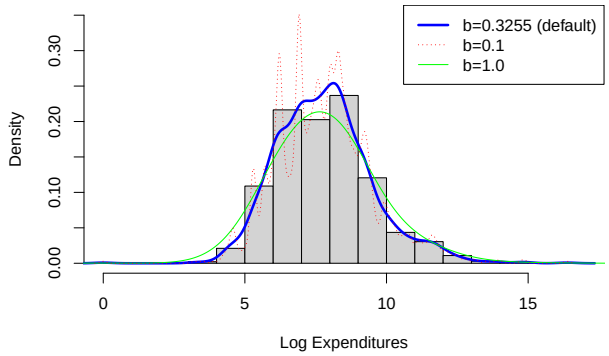


Figure 4.4: Histogram of Logarithmic Property Claims with Superimposed Kernel Density Estimators

```
legend("topright", c("b=0.3255 (default)", "b=0.1", "b=1.0"), lty=c(1,3,1),
           lwd=c(2.5,1,1), col=c("blue", "red", "green"), cex=1)
```

Nonparametric density estimators, such as the kernel estimator, are regularly used in practice. The concept can also be extended to give smooth versions of an empirical distribution function. Given the definition of the kernel density estimator, the kernel estimator of the distribution function can be found as

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{b}\right).$$

where W is the distribution function associated with the kernel density w. To illustrate, for the uniform kernel, we have $w(y) = \frac{1}{2}I(-1 < y \leq 1)$, so

$$W(y) = \begin{cases} 0 & y < -1 \\ \frac{y+1}{2} & -1 \leq y < 1 \\ 1 & y \geq 1 \end{cases}.$$

Example 4.1.5. Actuarial Exam Question.

You study five lives to estimate the time from the onset of a disease to death. The times to death are:

2 3 3 3 7

Using a triangular kernel with bandwith 2, calculate the density function estimate at 2.5.

Show Example Solution

Solution. For the kernel density estimate, we have

$$f_n(x) = \frac{1}{nb} \sum_{i=1}^{n} w\left(\frac{x - X_i}{b}\right),$$

where $n = 5$, $b = 2$, and $x = 2.5$. For the triangular kernel, $w(x) = (1 - |x|) \times I(|x| \leq 1)$. Thus,

X_i	$\frac{x-X_i}{b}$	$w\left(\frac{x-X_i}{b}\right)$
2	$\frac{2.5-2}{2} = \frac{1}{4}$	$(1 - \frac{1}{4})(1) = \frac{3}{4}$
3		
3	$\frac{2.5-3}{2} = \frac{-1}{4}$	$\left(1 - \left\lvert\frac{-1}{4}\right\rvert\right)(1) = \frac{3}{4}$
3		
7	$\frac{2.5-7}{2} = -2.25$	$(1 - \lvert -2.25\rvert)(0) = 0$

Then the kernel density estimate is

$$f_n(x) = \frac{1}{5(2)}\left(\frac{3}{4} + (3)\frac{3}{4} + 0\right) = \frac{3}{10}$$

Plug-in Principle

One way to create a nonparametric estimator is to use the analog or plug-in principle where one replaces the unknown cdf F with a known estimate such as the empirical cdf F_n. So, if we are trying to estimate $\mathrm{E}\ \mathrm{g}(X) = \mathrm{E}_F\ \mathrm{g}(X)$ for a generic function g, then we define a nonparametric estimator to be $\mathrm{E}_{F_n}\ \mathrm{g}(X) = n^{-1} \sum_{i=1}^{n} \mathrm{g}(X_i)$.

To see how this works, as a special case of g we consider the loss elimination ratio introduced in Section 3.4.1,

$$LER(d) = \frac{\mathrm{E}\ (\min(X, d))}{\mathrm{E}\ (X)}$$

for a fixed deductible d.

Example. 4.1.11. Bodily Injury Claims and Loss Elimination Ratios

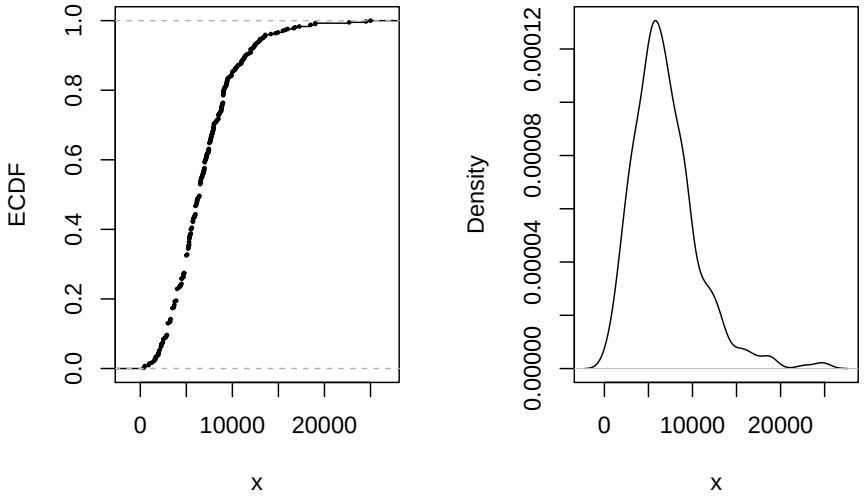


Figure 4.5: Bodily Injury Claims. The left-hand panel gives the empirical distribution function. The right-hand panel presents a nonparametric density plot.

We use a sample of 432 closed auto claims from Boston from **?**. Losses are recorded for payments due to bodily injuries in auto accidents. Losses are not subject to deductibles but are subject to various policy limits also available in the data. It turns out that only 17 out of 432 ($\approx 4\%$) were subject to policy limit and so we ignore these data for this illustration.

The average loss paid is 6906. Figure **??** shows other aspects of the distribution. Specifically, the left-hand panel shows the empirical distribution function, the right-hand panel gives a nonparametric density plot.

The impact of bodily injury losses can be mitigated by the imposition of limits or purchasing reinsurance policies (see Section 10.3). To quantify the impact of these risk mitigation tools, it is common to compute the loss elimination ratio (LER) as introduced in Section 3.4.1. The distribution function is not available and so much be estimated in some way. Using the plug-in principle, a nonparametric estimator can be defined as

$$LER_n(d) = \frac{n^{-1}\sum_{i=1}^n \min(X_i, d)}{n^{-1}\sum_{i=1}^n X_i} = \frac{\sum_{i=1}^n \min(X_i, d)}{\sum_{i=1}^n X_i}.$$

Figure **??** shows the estimator $LER_n(d)$ for various choices of d. For example, if $d = 14,000$, then it turns out that $LER_n(14000) \approx 0.9768$. Imposing a limit

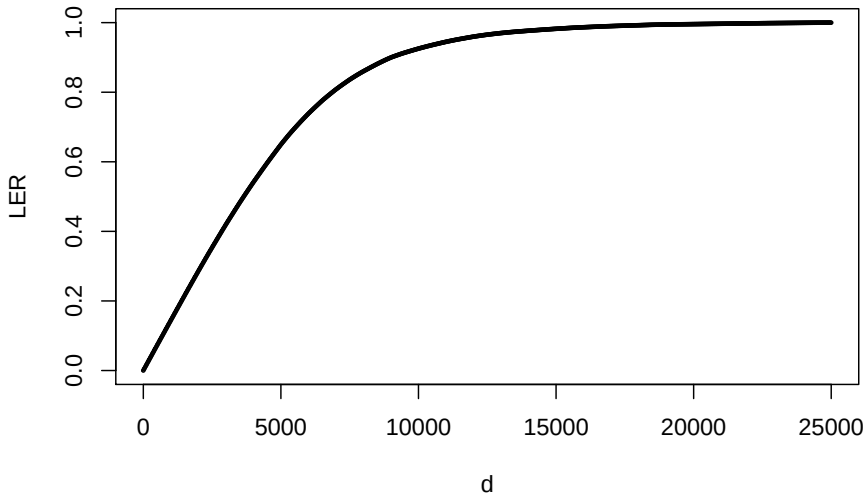


Figure 4.6: LER for Bodily Injury Claims. The figure presents the loss elimination ratio (LER) as a function of deductible d.

of 14,000 means that we expect to retain 97.68 percent of claims.

4.1.2 Tools for Model Selection and Diagnostics

The previous section introduced nonparametric estimators in which there was no parametric form assumed about the underlying distributions. However, in many actuarial applications, analysts seek to employ a parametric fit of a distribution for ease of explanation and the ability to readily extend it to more complex situations such as including explanatory variables in a regression setting. When fitting a parametric distribution, one analyst might try to use a gamma distribution to represent a set of loss data. However, another analyst may prefer to use a Pareto distribution. How does one know which model to **select**?

Nonparametric tools can be used to corroborate the selection of parametric models. Essentially, the approach is to compute selected summary measures under a fitted parametric model and to compare it to the corresponding quantity under the nonparametric model. As the nonparametric does not assume a specific distribution and is merely a function of the data, it is used as a benchmark to assess how well the parametric distribution/model represents the data. This comparison may alert the analyst to deficiencies in the parametric model and

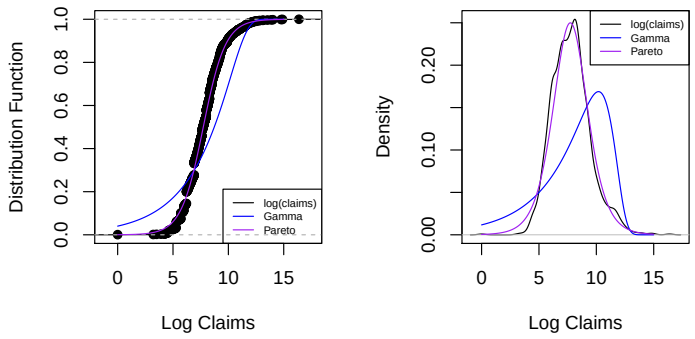


Figure 4.7: Nonparametric Versus Fitted Parametric Distribution and Density Functions. The left-hand panel compares distribution functions, with the dots corresponding to the empirical distribution, the thick blue curve corresponding to the fitted gamma and the light purple curve corresponding to the fitted Pareto. The right hand panel compares these three distributions summarized using probability density functions.

sometimes point ways to improving the parametric specification. Procedures geared towards assessing the validity of a model are known as **model diagnostics**.

Graphical Comparison of Distributions

We have already seen the technique of overlaying graphs for comparison purposes. To reinforce the application of this technique, Figure **??** compares the empirical distribution to two parametric fitted distributions. The left panel shows the distribution functions of claims distributions. The dots forming an "S-shaped" curve represent the empirical distribution function at each observation. The thick blue curve gives corresponding values for the fitted gamma distribution and the light purple is for the fitted Pareto distribution. Because the Pareto is much closer to the empirical distribution function than the gamma, this provides evidence that the Pareto is the better model for this data set. The right panel gives similar information for the density function and provides a consistent message. Based (only) on these figures, the Pareto distribution is the clear choice for the analyst.

For another way to compare the appropriateness of two fitted models, consider the **probability-probability (*pp*) plot**. A *pp* plot compares cumulative probabilities under two models. For our purposes, these two models are the nonparametric empirical distribution function and the parametric fitted model. Figure

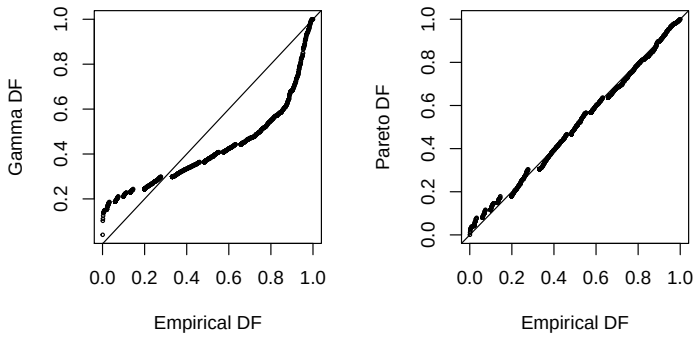


Figure 4.8: Probability-Probability (*pp*) Plots. The horizontal axes gives the empirical distribution function at each observation. In the left-hand panel, the corresponding distribution function for the gamma is shown in the vertical axis. The right-hand panel shows the fitted Pareto distribution. Lines of $y = x$ are superimposed.

?? shows *pp* plots for the Property Fund data. The fitted gamma is on the left and the fitted Pareto is on the right, compared to the same empirical distribution function of the data. The straight line represents equality between the two distributions being compared, so points close to the line are desirable. As seen in earlier demonstrations, the Pareto is much closer to the empirical distribution than the gamma, providing additional evidence that the Pareto is the better model.

A *pp* plot is useful in part because no artificial scaling is required, such as with the overlaying of densities in Figure **??**, in which we switched to the log scale to better visualize the data. The Chapter 4 Technical Supplement A.1 introduces a variation of the *pp* plot known as a Lorenz curve; this is an important tool for assessing income inequality. Furthermore, *pp* plots are available in multivariate settings where more than one outcome variable is available. However, a limitation of the *pp* plot is that, because it is a plot cumulative distribution functions, it can sometimes be difficult to detect where a fitted parametric distribution is deficient. As an alternative, it is common to use a **quantile-quantile (*qq*) plot**, as demonstrated in Figure **??**.

The *qq* plot compares two fitted models through their quantiles. As with *pp* plots, we compare the nonparametric to a parametric fitted model. Quantiles may be evaluated at each point of the data set, or on a grid (e.g., at $0, 0.001, 0.002, \ldots, 0.999, 1.000$), depending on the application. In Figure **??**, for each point on the aforementioned grid, the horizontal axis displays the empirical quantile and the vertical axis displays the corresponding fitted parametric quan-

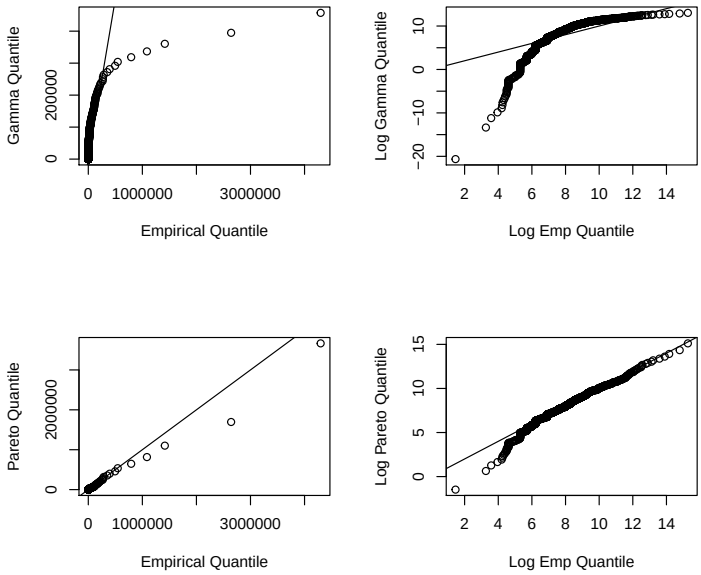


Figure 4.9: Quantile-Quantile (qq) Plots. The horizontal axes gives the empirical quantiles at each observation. The right-hand panels they are graphed on a logarithmic basis. The vertical axis gives the quantiles from the fitted distributions; gamma quantiles are in the upper panels, Pareto quantiles are in the lower panels.

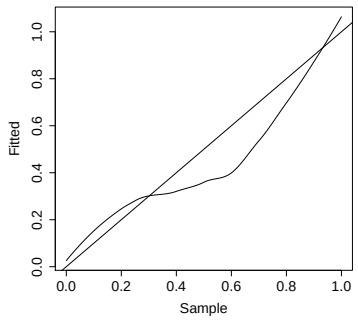
tile (gamma for the upper two panels, Pareto for the lower two). Quantiles are plotted on the original scale in the left panels and on the log scale in the right panels to allow us to see where a fitted distribution is deficient. The straight line represents equality between the empirical distribution and fitted distribution. From these plots, we again see that the Pareto is an overall better fit than the gamma. Furthermore, the lower-right panel suggests that the Pareto distribution does a good job with large observations, but provides a poorer fit for small observations.

Example 4.1.6. Actuarial Exam Question. The graph below shows a *pp* plot of a fitted distribution compared to a sample.

Comment on the two distributions with respect to left tail, right tail, and median probabilities.

Show Example Solution

Solution. The tail of the fitted distribution is too thick on the left, too thin



on the right, and the fitted distribution has less probability around the median than the sample. To see this, recall that the *pp* plot graphs the cumulative distribution of two distributions on its axes (empirical on the x-axis and fitted on the y-axis in this case). For small values of x, the fitted model assigns greater probability to being below that value than occurred in the sample (i.e. $F(x) > F_n(x)$). This indicates that the model has a heavier left tail than the data. For large values of x, the model again assigns greater probability to being below that value and thus less probability to being above that value (i.e. $S(x) < S_n(x)$. This indicates that the model has a lighter right tail than the data. In addition, as we go from 0.4 to 0.6 on the horizontal axis (thus looking at the middle 20% of the data), the *pp* plot increases from about 0.3 to 0.4. This indicates that the model puts only about 10% of the probability in this range.

Statistical Comparison of Distributions

When selecting a model, it is helpful to make the graphical displays presented. However, for reporting results, it can be effective to supplement the graphical displays with selected statistics that summarize model goodness of fit. Table 4.2 provides three commonly used goodness of fit statistics. In this table, F_n is the empirical distribution, F is the fitted or hypothesized distribution, and $F_i = F(x_i)$.

Table 4.2: Three Goodness of Fit Statistics

Statistic	Definition	Computational Expression
Kolmogorov-Smirnov	$\max_x \lvert F_n(x) - F(x)\rvert$	$\max(D^+, D^-)$ where $D^+ = \max_{i=1,\ldots,n} \left\lvert \frac{i}{n} - F_i \right\rvert$ $D^- = \max_{i=1,\ldots,n} \left\lvert F_i - \frac{i-1}{n} \right\rvert$
Cramer-von Mises	$n\int (F_n(x) - F(x))^2 f(x)dx$	$\frac{1}{12n} + \sum_{i=1}^n \left(F_i - (2i-1)/n\right)^2$
Anderson-Darling	$n\int \frac{(F_n(x)-F(x))^2}{F(x)(1-F(x))} f(x)dx$	$-n - \frac{1}{n}\sum_{i=1}^n (2i-1)\log\left(F_i(1-F_{n+1-i})\right)^2$

The Kolmogorov-Smirnov statistic is the maximum absolute difference between the fitted distribution function and the empirical distribution function. Instead of comparing differences between single points, the Cramer-von Mises statistic integrates the difference between the empirical and fitted distribution functions over the entire range of values. The Anderson-Darling statistic also integrates this difference over the range of values, although weighted by the inverse of the variance. It therefore places greater emphasis on the tails of the distribution (i.e when $F(x)$ or $1 - F(x) = S(x)$ is small).

Exaxmple 4.1.7. Actuarial Exam Question (modified). A sample of claim payments is:

$$29 \quad 64 \quad 90 \quad 135 \quad 182$$

Compare the empirical claims distribution to an exponential distribution with mean 100 by calculating the value of the Kolmogorov-Smirnov test statistic.

Show Example Solution

Solution. For an exponential distribution with mean 100, the cumulative distribution function is $F(x) = 1 - e^{-x/100}$. Thus,

x	$F(x)$	$F_n(x)$	$F_n(x-)$	$\max(\lvert F(x) - F_n(x)\rvert, \lvert F(x) - F_n(x-)\rvert)$
29	0.2517	0.2	0	$\max(0.0517, 0.2517) = 0.2517$
64	0.4727	0.4	0.2	$\max(0.0727, 0.2727) = 0.2727$
90	0.5934	0.6	0.4	$\max(0.0066, 0.1934) = 0.1934$
135	0.7408	0.8	0.6	$\max(0.0592, 0.1408) = 0.1408$
182	0.8380	1	0.8	$\max(0.1620, 0.0380) = 0.1620$

The Kolmogorov-Smirnov test statistic is therefore $KS = \max(0.2517, 0.2727, 0.1934, 0.1408, 0.1620) = 0.2727$.

4.1.3 Starting Values

The method of moments and percentile matching are nonparametric estimation methods that provide alternatives to maximum likelihood. Generally, maximum likelihood is the preferred technique because it employs data more efficiently. (See Appendix Chapter **??** for precise definitions of efficiency.) However, methods of moments and percentile matching are useful because they are easier to interpret and therefore allow the actuary or analyst to explain procedures to others. Additionally, the numerical estimation procedure (e.g. if performed in

R) for the maximum likelihood is iterative and requires starting values to begin the recursive process. Although many problems are robust to the choice of the starting values, for some complex situations, it can be important to have a starting value that is close to the (unknown) optimal value. Method of moments and percentile matching are techniques that can produce desirable estimates without a serious computational investment and can thus be used as a starting value for computing maximum likelihood.

Method of Moments

Under the **method of moments**, we approximate the moments of the parametric distribution using the empirical (nonparametric) moments described in Section **??**. We can then algebraically solve for the parameter estimates.

Example 4.1.8. Property Fund. For the 2010 property fund, there are $n = 1,377$ individual claims (in thousands of dollars) with

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = 26.62259 \quad \text{and} \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = 136154.6.$$

Fit the parameters of the gamma and Pareto distributions using the method of moments.

Show Example Solution

Solution. To fit a gamma distribution, we have $\mu_1 = \alpha\theta$ and $\mu_2' = \alpha(\alpha+1)\theta^2$. Equating the two yields the method of moments estimators, easy algebra shows that

$$\alpha = \frac{\mu_1^2}{\mu_2' - \mu_1^2} \quad \text{and} \quad \theta = \frac{\mu_2' - \mu_1^2}{\mu_1}.$$

Thus, the method of moment estimators are

$$\hat{\alpha} = \frac{26.62259^2}{136154.6 - 26.62259^2} = 0.005232809$$
$$\hat{\theta} = \frac{136154.6 - 26.62259^2}{26.62259} = 5,087.629.$$

For comparison, the maximum likelihood values turn out to be $\hat{\alpha}_{MLE} = 0.2905959$ and $\hat{\theta}_{MLE} = 91.61378$, so there are big discrepancies between the two estimation procedures. This is one indication, as we have seen before, that the gamma model fits poorly.

In contrast, now assume a Pareto distribution so that $\mu_1 = \theta/(\alpha - 1)$ and $\mu_2' = 2\theta^2/((\alpha - 1)(\alpha - 2))$. Easy algebra shows

$$\alpha = 1 + \frac{\mu_2'}{\mu_2' - \mu_1^2} \quad \text{and} \quad \theta = (\alpha - 1)\mu_1.$$

Thus, the method of moment estimators are

$$\hat{\alpha} = 1 + \frac{136154.6}{136154.6 - 26,62259^2} = 2.005233$$
$$\hat{\theta} = (2.005233 - 1) \cdot 26.62259 = 26.7619$$

The maximum likelihood values turn out to be $\hat{\alpha}_{MLE} = 0.9990936$ and $\hat{\theta}_{MLE} = 2.2821147$. It is interesting that $\hat{\alpha}_{MLE} < 1$; for the Pareto distribution, recall that $\alpha < 1$ means that the mean is infinite. This is another indication that the property claims data set is a long tail distribution.

As the above example suggests, there is flexibility with the method of moments. For example, we could have matched the second and third moments instead of the first and second, yielding different estimators. Furthermore, there is no guarantee that a solution will exist for each problem. You will also find that matching moments is possible for a few problems where the data are censored or truncated, but in general, this is a more difficult scenario. Finally, for distributions where the moments do not exist or are infinite, method of moments is not available. As an alternative, one can use the percentile matching technique.

Percentile Matching

Under **percentile matching**, we approximate the quantiles or percentiles of the parametric distribution using the empirical (nonparametric) quantiles or percentiles described in Section **??**.

Example 4.1.9. Property Fund. For the 2010 property fund, we illustrate matching on quantiles. In particular, the Pareto distribution is intuitively pleasing because of the closed-form solution for the quantiles. Recall that the distribution function for the Pareto distribution is

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

Easy algebra shows that we can express the quantile as

$$F^{-1}(q) = \theta \left((1 - q)^{-1/\alpha} - 1\right).$$

for a fraction q, $0 < q < 1$.

Determine estimates of the Pareto distribution parameters using the 25th and 95th empirical quantiles.

Show Example Solution

Solution.

The 25th percentile (the first quartile) turns out to be 0.78853 and the 95th percentile is 50.98293 (both in thousands of dollars). With two equations

$$0.78853 = \theta\left(1-(1-.25)^{-1/\alpha}\right) \quad \text{and} \quad 50.98293 = \theta\left(1-(1-.75)^{-1/\alpha}\right)$$

and two unknowns, the solution is

$$\hat{\alpha} = 0.9412076 \quad \text{and} \quad \hat{\theta} = 2.205617.$$

We remark here that a numerical routine is required for these solutions as no analytic solution is available. Furthermore, recall that the maximum likelihood estimates are $\hat{\alpha}_{MLE} = 0.9990936$ and $\hat{\theta}_{MLE} = 2.2821147$, so the percentile matching provides a better approximation for the Pareto distribution than the method of moments.

Example 4.1.10. Actuarial Exam Question. You are given:

(i) Losses follow a loglogistic distribution with cumulative distribution function:

$$F(x) = \frac{(x/\theta)^{\gamma}}{1+(x/\theta)^{\gamma}}$$

(ii) The sample of losses is:

10 35 80 86 90 120 158 180 200 210 1500

Calculate the estimate of θ by percentile matching, using the 40th and 80th empirically smoothed percentile estimates.

Show Example Solution

Solution. With 11 observations, we have $j = \lfloor (n+1)q \rfloor = \lfloor 12(0.4) \rfloor = \lfloor 4.8 \rfloor = 4$ and $h = (n+1)q - j = 12(0.4) - 4 = 0.8$. By interpolation, the 40th empirically smoothed percentile estimate is $\hat{\pi}_{0.4} = (1-h)X_{(j)} + hX_{(j+1)} = 0.2(86) + 0.8(90) = 89.2$.

Similarly, for the 80th empirically smoothed percentile estimate, we have $12(0.8) = 9.6$ so the estimate is $\hat{\pi}_{0.8} = 0.4(200) + 0.6(210) = 206$.

Using the loglogistic cumulative distribution, we need to solve the following two equations for parameters θ and γ:

$$0.4 = \frac{(89.2/\theta)^\gamma}{1+(89.2/\theta)^\gamma} \quad \text{and} \quad 0.8 = \frac{(206/\theta)^\gamma}{1+(206+\theta)^\gamma}$$

Solving for each parenthetical expression gives $\frac{2}{3} = (89.2/\theta)^\gamma$ and $4 = (206/\theta)^\gamma$. Taking the ratio of the second equation to the first gives $6 = (206/89.2)^\gamma \Rightarrow \gamma = \frac{\ln(6)}{\ln(206/89.2)} = 2.1407$. Then $4^{1/2.1407} = 206/\theta \Rightarrow \theta = 107.8$

Like the method of moments, percentile matching is almost too flexible in the sense that many estimators can be based on percentile matches; for example, one actuary can base estimation on the 25th and 95th percentiles whereas another actuary uses the 20th and 80th percentiles. In general these estimators will differ and there is no compelling reason to prefer one over the other. Also as with the method of moments, percentile matching is appealing because it provides a technique that can be readily applied in selected situations and has an intuitive basis. Although most actuarial applications use maximum likelihood estimators, it can be convenient to have alternative approaches such as method of moments and percentile matching available.

Show Quiz Solution

4.2 Model Selection

In this section, you learn how to:

- Describe the iterative model selection specification process
- Outline steps needed to select a parametric model
- Describe pitfalls of model selection based purely on insample data when compared to the advantages of out-of-sample model validation

---

This section underscores the idea that model selection is an iterative process in which models are cyclically (re)formulated and tested for appropriateness before using them for inference. After an overview, we describe the model selection process based on:

- an in-sample or training dataset,
- an out-of-sample or test dataset, and
- a method that combines these approaches known as **cross-validation**.

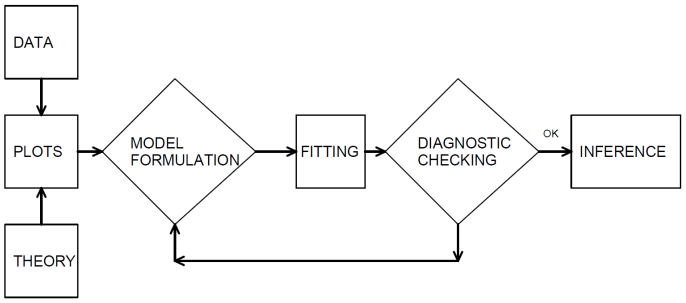


Figure 4.10: The iterative model specification process.

4.2.1 Iterative Model Selection

In our development, we examine the data graphically, hypothesize a model structure, and compare the data to a candidate model in order to formulate an improved model. **?** describes this as an iterative process which is shown in Figure **??**.

This iterative process provides a useful recipe for structuring the task of specifying a model to represent a set of data.

1. The first step, the model formulation stage, is accomplished by examining the data graphically and using prior knowledge of relationships, such as from economic theory or industry practice.
2. The second step in the iteration is fitting based on the assumptions of the specified model. These assumptions must be consistent with the data to make valid use of the model.
3. The third step is diagnostic checking; the data and model must be consistent with one another before additional inferences can be made. Diagnostic checking is an important part of the model formulation; it can reveal mistakes made in previous steps and provide ways to correct these mistakes.

The iterative process also emphasizes the skills you need to make analytics work. First, you need a willingness to summarize information numerically and portray this information graphically. Second, it is important to develop an understanding of model properties. You should understand how a probabilistic model behaves in order to match a set of data to it. Third, theoretical properties of the model are also important for inferring general relationships based on the behavior of the data.

4.2.2 Model Selection Based on a Training Dataset

It is common to refer to a dataset used for analysis as an in-sample or training dataset. Techniques available for selecting a model depend upon whether the outcomes X are discrete, continuous, or a hybrid of the two, although the principles are the same.

Graphical and other Basic Summary Measures. Begin by summarizing the data graphically and with statistics that do not rely on a specific parametric form, as summarized in Section **??**. Specifically, you will want to graph both the empirical distribution and density functions. Particularly for loss data that contain many zeros and that can be skewed, deciding on the appropriate scale (e.g., logarithmic) may present some difficulties. For discrete data, tables are often preferred. Determine sample moments, such as the mean and variance, as well as selected quantiles, including the minimum, maximum, and the median. For discrete data, the mode (or most frequently occurring value) is usually helpful.

These summaries, as well as your familiarity of industry practice, will suggest one or more candidate parametric models. Generally, start with the simpler parametric models (for example, one parameter exponential before a two parameter gamma), gradually introducing more complexity into the modeling process.

Critique the candidate parametric model numerically and graphically. For the graphs, utilize the tools introduced in Section **??** such as *pp* and *qq* plots. For the numerical assessments, examine the statistical significance of parameters and try to eliminate parameters that do not provide additional information.

Likelihood Ratio Tests. For comparing model fits, if one model is a subset of another, then a likelihood ratio test may be employed; the general approach to likelihood ratio testing is described in Sections **??** and **??**.

Goodness of Fit Statistics. Generally, models are not proper subsets of one another so overall goodness of fit statistics are helpful for comparing models. Information criteria are one type of goodness of statistic. The most widely used examples are Akaike's Information Criterion (AIC) and the (Schwarz) Bayesian Information Criterion (BIC); they are widely cited because they can be readily generalized to multivariate settings. Section **??** provides a summary of these statistics.

For selecting the appropriate distribution, statistics that compare a parametric fit to a nonparametric alternative, summarized in Section **??**, are useful for model comparison. For discrete data, a goodness of fit statistic (as described in Section **??**) is generally preferred as it is more intuitive and simpler to explain.

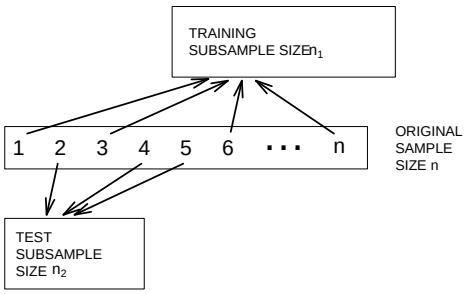


Figure 4.11: Model Validation. A data set is randomly split into two subsamples.

4.2.3 Model Selection Based on a Test Dataset

Model validation is the process of confirming that the proposed model is appropriate, especially in light of the purposes of the investigation. An important limitation of the model selection process based only on insample data is that it can be susceptible to data-snooping, that is, fitting a great number of models to a single set of data. By looking at a large number of models, we may overfit the data and understate the natural variation in our representation.

Selecting a model based only on insample data also does not support the goal of **predictive inference**. Particularly in actuarial applications, our goal is to make statements about new experience rather than a dataset at hand. For example, we use claims experience from one year to develop a model that can be used to price insurance contracts for the following year. As an analogy, we can think about the training data set as experience from one year that is used to predict the behavior of the next year's test data set.

We can respond to these criticisms by using a technique sometimes known as **out-of-sample validation**. The ideal situation is to have available two sets of data, one for training, or model development, and one for testing, or model validation. We initially develop one or several models on the first data set that we call our candidate models. Then, the relative performance of the candidate models can be measured on the second set of data. In this way, the data used to validate the model is unaffected by the procedures used to formulate the model.

Random Split of the Data. Unfortunately, rarely will two sets of data be available to the investigator. However, we can implement the validation process by splitting the data set into **training** and **test** subsamples, respectively. Figure **??** illustrates this splitting of the data.

Various researchers recommend different proportions for the allocation. **?** suggests that data-splitting not be done unless the sample size is moderately large.

The guidelines of **?** show that the greater the number of parameters to be estimated, the greater the proportion of observations needed for the model development subsample.

Model Validation Statistics. Much of the literature supporting the establishment of a model validation process is based on regression and classification models that you can think of as an input-output problem (**?**). That is, we have several inputs $x_1, \ldots, x_k$ that are related to an output y through a function such as

$$y = \mathrm{g}\left(x_1, \ldots, x_k\right).$$

One uses the training sample to develop an estimate of g, say, $\hat{\mathrm{g}}$, and then calibrate the distance from the observed outcomes to the predictions using a criterion of the form

$$\sum_i \mathrm{d}(y_i, \hat{\mathrm{g}}\left(x_{i1}, \ldots, x_{ik}\right)). \tag{4.4}$$

Here, the sum i is over the test data. In many regression applications, it is common to use squared Euclidean distance of the form $\mathrm{d}(y_i, \mathrm{g}) = (y_i - \mathrm{g})^2$. In actuarial applications, Euclidean distance $\mathrm{d}(y_i, \mathrm{g}) = |y_i - \mathrm{g}|$ is often preferred because of the skewed nature of the data (large outlying values of y can have a large effect on the measure). Chapter **??** describes another measure, the Gini index, that is useful in actuarial applications particularly when there is a large proportion of zeros in claims data (corresponding to no claims).

Selecting a Distribution. Still, our focus so far has been to select a distribution for a data set that can be used for actuarial modeling without additional inputs $x_1, \ldots, x_k$. Even in this more fundamental problem, the model validation approach is valuable. If we base all inference on only in-sample data, then there is a tendency to select more complicated models then needed. For example, we might select a four parameter GB2, generalized beta of the second kind, distribution when only a two parameter Pareto is needed. Information criteria such as AICAkaike's information criterion and BICBayesian information criterion included penalties for model complexity and so provide some protection but using a test sample is the best guarantee to achieve parsimonious models. From a quote often attributed to Albert Einstein, we want to "use the simplest model as possible but no simpler."

4.2.4 Model Selection Based on Cross-Validation

Although out-of-sample validation is the gold standard in predictive modeling, it is not always practical to do so. The main reason is that we have limited sample sizes and the out-of-sample model selection criterion in equation (**??**) depends on a random split of the data. This means that different analysts, even when working the same data set and same approach to modeling, may select different

models. This is likely in actuarial applications because we work with skewed data sets where there is a large chance of getting some very large outcomes and large outcomes may have a great influence on the parameter estimates.

Cross-Validation Procedure. Alternatively, one may use **cross-validation**, as follows.

- The procedure begins by using a random mechanism to split the data into K subsets known as folds, where analysts typically use 5 to 10.
- Next, one uses the first K-1 subsamples to estimate model parameters. Then, "predict" the outcomes for the Kth subsample and use a measure such as in equation (**??**) to summarize the fit.
- Now, repeat this by holding out each of the K sub-samples, summarizing with a cumulative out-of-sample statistic.

Repeat these steps for several candidate models and choose the model with the lowest cumulative out-of-sample statistic.

Cross-validation is widely used because it retains the predictive flavor of the out-of-sample model validation process but, due to the re-use of the data, is more stable over random samples.

Show Quiz Solution

4.3 Estimation using Modified Data

In this section, you learn how to:

- Describe grouped, censored, and truncated data
- Estimate parametric distributions based on grouped, censored, and truncated data
- Estimate distributions nonparametrically based on grouped, censored, and truncated data

---

4.3.1 Parametric Estimation using Modified Data

Basic theory and many applications are based on individual observations that are "complete" and "unmodified," as we have seen in the previous section. Section **??** introduced the concept of observations that are "modified" due to two common types of limitations: **censoring** and **truncation**. For example, it is common to think about an insurance deductible as producing data that are truncated (from the left) or policy limits as yielding data that are censored (from the

right). This viewpoint is from the primary insurer (the seller of the insurance). However, as we will see in Chapter **??**, a reinsurer (an insurer of an insurance company) may not observe claims smaller than an amount, only that a claim exists, an example of censoring from the left. So, in this section, we cover the full gamut of alternatives. Specifically, this section will address parametric estimation methods for three alternatives to individual, complete, and unmodified data: **interval-censored** data available only in groups, data that are limited or **censored**, and data that may not be observed due to **truncation**.

Parametric Estimation using Grouped Data

Consider a sample of size n observed from the distribution $F(\cdot)$, but in groups so that we only know the group into which each observation fell, not the exact value. This is referred to as **grouped** or **interval-censored** data. For example, we may be looking at two successive years of annual employee records. People employed in the first year but not the second have left sometime during the year. With an exact departure date (individual data), we could compute the amount of time that they were with the firm. Without the departure date (grouped data), we only know that they departed sometime during a year-long interval.

Formalizing this idea, suppose there are k groups or intervals delimited by boundaries $c_0 < c_1 < \cdots < c_k$. For each observation, we only observe the interval into which it fell (e.g. (c_{j-1}, c_j)), not the exact value. Thus, we only know the number of observations in each interval. The constants $\{c_0 < c_1 < \cdots < c_k\}$ form some partition of the domain of $F(\cdot)$. Then the probability of an observation X_i falling in the jth interval is

$$\Pr\left(X_i \in (c_{j-1}, c_j]\right) = F(c_j) - F(c_{j-1}).$$

The corresponding probability mass function for an observation is

$$\begin{aligned} f(x) &= \begin{cases} F(c_1) - F(c_0) & \text{if } \ x \in (c_0, c_1] \\ \vdots & \vdots \\ F(c_k) - F(c_{k-1}) & \text{if } \ x \in (c_{k-1}, c_k] \end{cases} \\ &= \prod_{j=1}^{k} \left\{F(c_j) - F(c_{j-1})\right\}^{I(x \in (c_{j-1}, c_j])} \end{aligned}$$

Now, define n_j to be the number of observations that fall in the jth interval, $(c_{j-1}, c_j]$. Thus, the likelihood function (with respect to the parameter(s) θ) is

$$\mathcal{L}(\theta) = \prod_{j=1}^{n} f(x_i) = \prod_{j=1}^{k} \left\{F(c_j) - F(c_{j-1})\right\}^{n_j}$$

And the log-likelihood function is

$$L(\theta) = \ln \mathcal{L}(\theta) = \ln \prod_{j=1}^{n} f(x_i) = \sum_{j=1}^{k} n_j \ln \{F(c_j) - F(c_{j-1})\}$$

Maximizing the likelihood function (or equivalently, maximizing the log-likelihood function) would then produce the maximum likelihood estimates for grouped data.

Example 4.3.1. Actuarial Exam Question.

You are given:

(i) Losses follow an exponential distribution with mean θ.
(ii) A random sample of 20 losses is distributed as follows:

Loss Range	Frequency
$[0, 1000]$	7
$(1000, 2000]$	6
$(2000, \infty)$	7

Calculate the maximum likelihood estimate of θ.

Show Example Solution

Solution.

$$\begin{aligned}
\mathcal{L}(\theta) &= F(1000)^7[F(2000) - F(1000)]^6[1 - F(2000)]^7 \\
&= (1 - e^{-1000/\theta})^7(e^{-1000/\theta} - e^{-2000/\theta})^6(e^{-2000/\theta})^7 \\
&= (1-p)^7(p-p^2)^6(p^2)^7 \\
&= p^{20}(1-p)^{13}
\end{aligned}$$

where $p = e^{-1000/\theta}$. Maximizing this expression with respect to p is equivalent to maximizing the likelihood with respect to θ. The maximum occurs at $p = \frac{20}{33}$ and so $\hat{\theta} = \frac{-1000}{\ln(20/33)} = 1996.90$.

Censored Data

Censoring occurs when we record only a limited value of an observation. The most common form is **right-censoring**, in which we record the smaller of the "true" dependent variable and a censoring variable. Using notation, let X represent an outcome of interest, such as the loss due to an insured event or time until an event. Let C_U denote the censoring amount. With right-censored observations, we record $X_U^* = \min(X, C_U) = X \wedge C_U$. We also record whether or

not censoring has occurred. Let $\delta_U = I(X \leq C_U)$ be a binary variable that is 0 if censoring occurs and 1 if it does not.

For an example that we saw in Section **??**, C_U may represent the upper limit of coverage of an insurance policy (we used u for the upper limit in that section). The loss may exceed the amount C_U, but the insurer only has C_U in its records as the amount paid out and does not have the amount of the actual loss X in its records.

Similarly, with **left-censoring**, we record the larger of a variable of interest and a censoring variable. If C_L is used to represent the censoring amount, we record $X_L^* = \max(X, C_L)$ along with the censoring indicator $\delta_L = I(X \geq C_L)$.

As an example, you got a brief introduction to reinsurance, insurance for insurers, in Section **??** and will see more in Chapter **??**. Suppose a reinsurer will cover insurer losses greater than C_L; this means that the reinsurer is responsible for the excess of X_L^* over C_L. Using notation, this is $Y = X_L^* - C_L$. To see this, first consider the case where the policyholder loss $X < C_L$. Then, the insurer will pay the entire claim and $Y = C_L - C_L = 0$, no loss for the reinsurer. For the second case, if the loss $X \geq C_L$, then $Y = X - C_L$ represents the reinsurer's retained claims. Put another way, if a loss occurs, the reinsurer records the actual amount if it exceeds the limit C_L and otherwise it only records that it had a loss of 0.

Truncated Data

Censored observations are recorded for study, although in a limited form. In contrast, **truncated** outcomes are a type of missing data. An outcome is potentially truncated when the availability of an observation depends on the outcome.

In insurance, it is common for observations to be **left-truncated** at C_L when the amount is

$$Y = \begin{cases} \text{we do not observe } X & X < C_L \\ X & X \geq C_L \end{cases}.$$

In other words, if X is less than the threshold C_L, then it is not observed.

For an example we saw in Section **??**, C_L may represent the deductible of an insurance policy (we used d for the deductible in that section). If the insured loss is less than the deductible, then the insurer may not observe or record the loss at all. If the loss exceeds the deductible, then the excess $X - C_L$ is the claim that the insurer covers. In Section **??**, we defined the per payment loss to be

$$Y^P = \begin{cases} \text{Undefined} & X \leq d \\ X - d & X > d \end{cases},$$

so that if a loss exceeds a deductible, we record the excess amount $X - d$. This is very important when considering amounts that the insurer will pay. However,

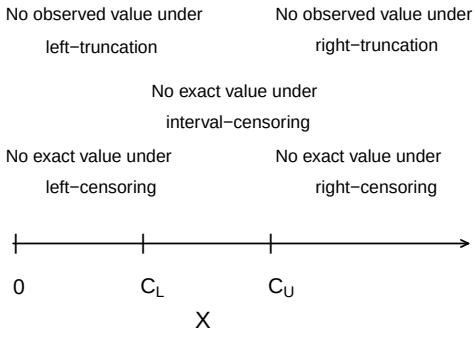


Figure 4.12: Censoring and Truncation

for estimation purposes of this section, it matters little if we substract a known constant such as $C_L = d$. So, for our truncated variable Y, we use the simpler convention and do not subract d.

Similarly for **right-truncated** data, if X exceeds a threshold C_U, then it is not observed. In this case, the amount is

$$Y = \begin{cases} X & X \leq C_U \\ \text{we do not observe } X & X > C_U. \end{cases}$$

Classic examples of truncation from the right include X as a measure of distance to a star. When the distance exceeds a certain level C_U, the star is no longer observable.

Figure **??** compares truncated and censored observations. Values of X that are greater than the "upper" censoring limit C_U are not observed at all (right-censored), while values of X that are smaller than the "lower" truncation limit C_L are observed, but observed as C_L rather than the actual value of X (left-truncated).

Show Mortality Study Example

Example – Mortality Study. Suppose that you are conducting a two-year study of mortality of high-risk subjects, beginning January 1, 2010 and finishing January 1, 2012. Figure **??** graphically portrays the six types of subjects recruited. For each subject, the beginning of the arrow represents that the the subject was recruited and the arrow end represents the event time. Thus, the arrow represents exposure time.

- **Type A - Right-censored.** This subject is alive at the beginning and

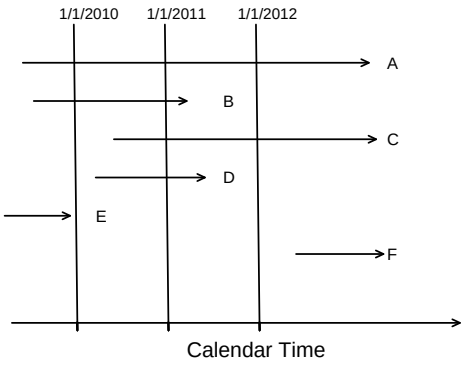


Figure 4.13: Timeline for Several Subjects on Test in a Mortality Study

the end of the study. Because the time of death is not known by the end of the study, it is right-censored. Most subjects are Type A.

- **Type B - Complete** information is available for a type B subject. The subject is alive at the beginning of the study and the death occurs within the observation period.
- **Type C - Right-censored and left-truncated.** A type C subject is right-censored, in that death occurs after the observation period. However, the subject entered after the start of the study and is said to have a delayed entry time. Because the subject would not have been observed had death occurred before entry, it is left-truncated.
- **Type D - Left-truncated.** A type D subject also has delayed entry. Because death occurs within the observation period, this subject is not right censored.
- **Type E - Left-truncated.** A type E subject is not included in the study because death occurs prior to the observation period.
- **Type F - Right-truncated.** Similarly, a type F subject is not included because the entry time occurs after the observation period.

To summarize, for outcome X and constants C_L and C_U,

Limitation Type	Limited Variable	Recording Information
right censoring	$X_U^* = \min(X, C_U)$	$\delta_U = I(X \leq C_U)$
left censoring	$X_L^* = \max(X, C_L)$	$\delta_L = I(X \geq C_L)$
interval censoring		
right truncation	X	observe X if $X \leq C_U$
left truncation	X	observe X if $X \geq C_L$

Parametric Estimation using Censored and Truncated Data

For simplicity, we assume non-random censoring amounts and a continuous outcome X. To begin, consider the case of right-censored data where we record $X_U^* = \min(X, C_U)$ and censoring indicator $\delta = I(X \leq C_U)$. If censoring occurs so that $\delta = 0$, then $X \geq C_U$ and the likelihood is $\Pr(X \geq C_U) = 1 - F(C_U)$. If censoring does not occur so that $\delta = 1$, then $X < C_U$ and the likelihood is $f(x)$. Summarizing, we have the likelihood of a single observation as

$$\begin{cases} 1 - F(C_U) & \text{if } \delta = 0 \\ f(x) & \text{if } \delta = 1 \end{cases}$$

$$\mathcal{L} = \prod_{i=1}^{n} \{f(x_i)\}^{\delta_i} \{1 - F(C_{Ui})\}^{1-\delta_i} = \prod_{\delta_i=1} f(x_i) \prod_{\delta_i=0} \{1 - F(C_{Ui})\},$$

with potential censoring times $\{C_{U1}, \ldots, C_{Un}\}$. Here, the notation "$\prod_{\delta_i=1}$" means to take the product over uncensored observations, and similarly for "$\prod_{\delta_i=0}$."

On the other hand, truncated data are handled in likelihood inference via conditional probabilities. Specifically, we adjust the likelihood contribution by dividing by the probability that the variable was observed. To summarize, we have the following contributions to the likelihood function for six types of outcomes:

Outcome	Likelihood Contribution
exact value	$f(x)$
right-censoring	$1 - F(C_U)$
left-censoring	$F(C_L)$
right-truncation	$f(x)/F(C_U)$
left-truncation	$f(x)/(1 - F(C_L))$
interval-censoring	$F(C_U) - F(C_L)$

For known outcomes and censored data, the likelihood is

$$\mathcal{L}(\theta) = \prod_{E} f(x_i) \prod_{R} \{1 - F(C_{Ui})\} \prod_{L} F(C_{Li}) \prod_{I} (F(C_{Ui}) - F(C_{Li})),$$

where "$\prod_E$" is the product over observations with Exact values, and similarly for Right-, Left- and Interval-censoring.

For right-censored and left-truncated data, the likelihood is

$$\mathcal{L} = \prod_{E} \frac{f(x_i)}{1 - F(C_{Li})} \prod_{R} \frac{1 - F(C_{Ui})}{1 - F(C_{Li})},$$

and similarly for other combinations. To get further insights, consider the following.

Show Special Case - Exponential Distribution

Special Case: Exponential Distribution. Consider data that are right-censored and left-truncated, with random variables X_i that are exponentially distributed with mean θ. With these specifications, recall that $f(x) = \theta^{-1} \exp(-x/\theta)$ and $F(x) = 1 - \exp(-x/\theta)$.

For this special case, the log-likelihood is

$$L(\theta) = \sum_E \{\ln f(x_i) - \ln(1 - F(C_{Li}))\} + \sum_R \{\ln(1 - F(C_{Ui})) - \ln(1 - \mathrm{F}(C_{Li}))\}$$
$$= \sum_E (-\ln\theta - (x_i - C_{Li})/\theta) - \sum_R (C_{Ui} - C_{Li})/\theta.$$

To simplify the notation, define $\delta_i = I(X_i \geq C_{Ui})$ to be a binary variable that indicates right-censoring. Let $X_i^{**} = \min(X_i, C_{Ui}) - C_{Li}$ be the amount that the observed variable exceeds the lower truncation limit. With this, the log-likelihood is

$$L(\theta) = -\sum_{i=1}^{n} ((1 - \delta_i)\ln\theta + \frac{x_i^{**}}{\theta}) \tag{4.5}$$

Taking derivatives with respect to the parameter θ and setting it equal to zero yields the maximum likelihood estimator

$$\widehat{\theta} = \frac{1}{n_u}\sum_{i=1}^{n} x_i^{**},$$

where $n_u = \sum_i (1 - \delta_i)$ is the number of uncensored observations.

Example 4.3.2. Actuarial Exam Question. You are given:

(i) A sample of losses is: 600 700 900
(ii) No information is available about losses of 500 or less.
(iii) Losses are assumed to follow an exponential distribution with mean θ.

Calculate the maximum likelihood estimate of θ.

Show Example Solution

Solution. These observations are truncated at 500. The contribution of each observation to the likelihood function is

$$\frac{f(x)}{1 - F(500)} = \frac{\theta^{-1}e^{-x/\theta}}{e^{-500/\theta}}$$

Then the likelihood function is

$$\mathcal{L}(\theta) = \frac{\theta^{-1}e^{-600/\theta}\theta^{-1}e^{-700/\theta}\theta^{-1}e^{-900/\theta}}{(e^{-500/\theta})^3} = \theta^{-3}e^{-700/\theta}$$

The log-likelihood is

$$L(\theta) = \ln \mathcal{L}(\theta) = -3 \ln \theta - 700\theta^{-1}$$

Maximizing this expression by setting the derivative with respect to θ equal to 0, we have

$$L'(\theta) = -3\theta^{-1} + 700\theta^{-2} = 0 \;\Rightarrow\; \hat{\theta} = \frac{700}{3} = 233.33$$

Example 4.3.3. Actuarial Exam Question. You are given the following information about a random sample:

(i) The sample size equals five.
(ii) The sample is from a Weibull distribution with $\tau = 2$.
(iii) Two of the sample observations are known to exceed 50, and the remaining three observations are 20, 30, and 45.

Calculate the maximum likelihood estimate of θ.

Show Example Solution

Solution. The likelihood function is

$$\begin{aligned}
\mathcal{L}(\theta) &= f(20)f(30)f(45)[1 - F(50)]^2 \\
&= \frac{2(20/\theta)^2 e^{-(20/\theta)^2}}{20} \frac{2(30/\theta)^2 e^{-(30/\theta)^2}}{30} \frac{2(45/\theta)^2 e^{-(45/\theta)^2}}{45} (e^{-(50/\theta)^2})^2 \\
&\propto \frac{1}{\theta^6} e^{-8325/\theta^2}
\end{aligned}$$

The natural logarithm of the above expression is $-6 \ln \theta - \frac{8325}{\theta^2}$. Maximizing this expression by setting its derivative to 0, we get

$$\frac{-6}{\theta} + \frac{16650}{\theta^3} = 0 \;\Rightarrow\; \hat{\theta} = \left(\frac{16650}{6}\right)^{\frac{1}{2}} = 52.6783$$

4.3.2 Nonparametric Estimation using Modified Data

Nonparametric estimators provide useful benchmarks, so it is helpful to understand the estimation procedures for grouped, censored, and truncated data.

Grouped Data

As we have seen in Section **??**, observations may be grouped (also referred to as interval censored) in the sense that we only observe them as belonging in one of k intervals of the form $(c_{j-1}, c_j]$, for $j = 1, \ldots, k$. At the boundaries, the empirical distribution function is defined in the usual way:

$$F_n(c_j) = \frac{\text{number of observations } \leq c_j}{n}.$$

For other values of $x \in (c_{j-1}, c_j)$, we can estimate the distribution function with the ogive estimator, which linearly interpolates between $F_n(c_{j-1})$ and $F_n(c_j)$, i.e. the values of the boundaries $F_n(c_{j-1})$ and $F_n(c_j)$ are connected with a straight line. This can formally be expressed as

$$F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j) \quad \text{for } c_{j-1} \leq x < c_j$$

The corresponding density is

$$f_n(x) = F_n'(x) = \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} \quad \text{for } c_{j-1} \leq x < c_j.$$

Example 4.3.4. Actuarial Exam Question.

You are given the following information regarding claim sizes for 100 claims:

Claim Size	Number of Claims
$0 - 1,000$	16
$1,000 - 3,000$	22
$3,000 - 5,000$	25
$5,000 - 10,000$	18
$10,000 - 25,000$	10
$25,000 - 50,000$	5
$50,000 - 100,000$	3
over $100,000$	1

Using the ogive, calculate the estimate of the probability that a randomly chosen claim is between 2000 and 6000.

Show Example Solution

Solution. At the boundaries, the empirical distribution function is defined in the usual way, so we have

$$F_{100}(1000) = 0.16,\ F_{100}(3000) = 0.38,\ F_{100}(5000) = 0.63,\ F_{100}(10000) = 0.81$$

For other claim sizes, the ogive estimator linearly interpolates between these values:

$$F_{100}(2000) = 0.5F_{100}(1000) + 0.5F_{100}(3000) = 0.5(0.16) + 0.5(0.38) = 0.27$$

$$F_{100}(6000) = 0.8F_{100}(5000) + 0.2F_{100}(10000) = 0.8(0.63) + 0.2(0.81) = 0.666$$

Thus, the probability that a claim is between 2000 and 6000 is $F_{100}(6000) - F_{100}(2000) = 0.666 - 0.27 = 0.396$.

Right-Censored Empirical Distribution Function

It can be useful to calibrate parametric estimators with nonparametric methods that do not rely on a parametric form of the distribution. The product-limit estimator due to (**?**) is a well-known estimator of the distribution function in the presence of censoring.

Motivation for the Kaplan-Meier Product Limit Estimator. To explain why the product-limit works so well with censored observations, let us first return to the "usual" case without censoring. Here, the empirical distribution function $F_n(x)$ is an unbiased estimator of the distribution function $F(x)$. This is because $F_n(x)$ is the average of indicator variables each of which are unbiased, that is, $\mathrm{E}\ I(X_i \leq x) = \Pr(X_i \leq x) = F(x)$.

Now suppose the the random outcome is censored on the right by a limiting amount, say, C_U, so that we record the smaller of the two, $X^* = \min(X, C_U)$. For values of x that are smaller than C_U, the indicator variable still provides an unbiased estimator of the distribution function before we reach the censoring limit. That is, $\mathrm{E}\ I(X^* \leq x) = F(x)$ because $I(X^* \leq x) = I(X \leq x)$ for $x < C_U$. In the same way, $\mathrm{E}\ I(X^* > x) = 1 - F(x) = S(x)$. But, for $x > C_U$, $I(X^* \leq x)$ is in general not an unbiased estimator of $F(x)$.

As an alternative, consider two random variables that have different censoring limits. For illustration, suppose that we observe $X_1^* = \min(X_1, 5)$ and $X_2^* = \min(X_2, 10)$ where X_1 and X_2 are independent draws from the same distribution. For $x \leq 5$, the empirical distribution function $F_2(x)$ is an unbiased estimator of $F(x)$. However, for $5 < x \leq 10$, the first observation cannot be used for the distribution function because of the censoring limitation. Instead, the strategy developed by (**?**) is to use $S_2(5)$ as an estimator of $S(5)$ and then to use the second observation to estimate the survival function conditional on survival to time 5, $\Pr(X > x | X > 5) = \frac{S(x)}{S(5)}$. Specifically, for $5 < x \leq 10$, the estimator of the survival function is

$$\hat{S}(x) = S_2(5) \times I(X_2^* > x).$$

Kaplan-Meier Product Limit Estimator. Extending this idea, for each observation i, let u_i be the upper censoring limit ($= \infty$ if no censoring). Thus, the recorded value is x_i in the case of no censoring and u_i if there is censoring. Let $t_1 < \cdots < t_k$ be k distinct points at which an uncensored loss occurs, and let s_j be the number of uncensored losses x_i's at t_j. The corresponding **risk set** is the number of observations that are active (not censored) at a value less than t_j, denoted as $R_j = \sum_{i=1}^{n} I(x_i \geq t_j) + \sum_{i=1}^{n} I(u_i \geq t_j)$.

With this notation, the **product-limit estimator** of the distribution function is

$$\hat{F}(x) = \begin{cases} 0 & x < t_1 \\ 1 - \prod_{j:t_j \leq x} \left(1 - \frac{s_j}{R_j}\right) & x \geq t_1 \end{cases} . \tag{4.6}$$

As usual, the corresponding estimate of the survival function is $\hat{S}(x) = 1 - \hat{F}(x)$.

Example 4.3.5. Actuarial Exam Question. The following is a sample of 10 payments:

$$4 \quad 4 \quad 5+ \quad 5+ \quad 5+ \quad 8 \quad 10+ \quad 10+ \quad 12 \quad 15$$

where + indicates that a loss has exceeded the policy limit.

Using the Kaplan-Meier product-limit estimator, calculate the probability that the loss on a policy exceeds 11, $\hat{S}(11)$.

Show Example Solution

Solution. There are four event times (non-censored observations). For each time t_j, we can calculate the number of events s_j and the risk set R_j as the following:

j	t_j	s_j	R_j
1	4	2	10
2	8	1	5
3	12	1	2
4	15	1	1

Thus, the Kaplan-Meier estimate of $S(11)$ is

$$\begin{aligned} \hat{S}(11) &= \prod_{j:t_j \leq 11} \left(1 - \frac{s_j}{R_j}\right) = \prod_{j=1}^{2} \left(1 - \frac{s_j}{R_j}\right) \\ &= \left(1 - \frac{2}{10}\right)\left(1 - \frac{1}{5}\right) = (0.8)(0.8) = 0.64. \end{aligned}$$

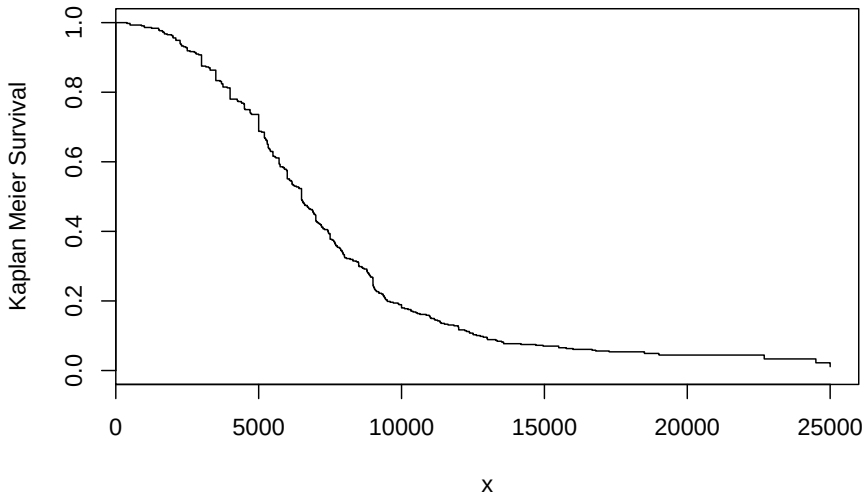


Figure 4.14: Kaplan-Meier Estimate of the Survival Function for Bodily Injury Claims

---

Example. 4.3.6. Bodily Injury Claims. We consider again the Boston auto bodily injury claims data from **?** that was introduced in Example 4.1.11. In that example, we omitted the 17 claims that were censored by policy limits. Now, we include the full dataset and use the Kaplan-Meier product limit to estimate the survival function. This is given in Figure **??**.

Show R Code

```
library(survival)                # for Surv(), survfit()
## KM estimate
KM0 <- survfit(Surv(AmountPaid, UnCensored) ~ 1,  type="kaplan-meier", data=BIData)
plot(KM0, conf.int=FALSE, xlab="x",ylab="Kaplan Meier Survival")
```

Right-Censored, Left-Truncated Empirical Distribution Function

In addition to right-censoring, we now extend the framework to allow for left-truncated data. As before, for each observation i, let u_i be the upper censoring limit ($= \infty$ if no censoring). Further, let d_i be the lower truncation limit (0 if

no truncation). Thus, the recorded value (if it is greater than d_i) is x_i in the case of no censoring and u_i if there is censoring. Let $t_1 < \cdots < t_k$ be k distinct points at which an event of interest occurs, and let s_j be the number of recorded events x_i's at time point t_j. The corresponding risk set is

$$R_j = \sum_{i=1}^{n} I(x_i \geq t_j) + \sum_{i=1}^{n} I(u_i \geq t_j) - \sum_{i=1}^{n} I(d_i \geq t_j).$$

With this new definition of the risk set, the product-limit estimator of the distribution function is as in equation (**??**).

Greenwood's Formula. (**?**) derived the formula for the estimated variance of the product-limit estimator to be

$$\widehat{Var}(\hat{F}(x)) = (1 - \hat{F}(x))^2 \sum_{j:t_j \leq x} \frac{s_j}{R_j(R_j - s_j)}.$$

R's `survfit` method takes a survival data object and creates a new object containing the Kaplan-Meier estimate of the survival function along with confidence intervals. The Kaplan-Meier method (`type='kaplan-meier'`) is used by default to construct an estimate of the survival curve. The resulting discrete survival function has point masses at the observed event times (discharge dates) t_j, where the probability of an event given survival to that duration is estimated as the number of observed events at the duration s_j divided by the number of subjects exposed or 'at-risk' just prior to the event duration R_j.

Two alternate types of estimation are also available for the `survfit` method. The alternative (`type='fh2'`) handles ties, in essence, by assuming that multiple events at the same duration occur in some arbitrary order. Another alternative (`type='fleming-harrington'`) uses the Nelson-Äalen (see (**?**)) estimate of the **cumulative hazard function** to obtain an estimate of the survival function. The estimated cumulative hazard $\hat{H}(x)$ starts at zero and is incremented at each observed event duration t_j by the number of events s_j divided by the number at risk R_j. With the same notation as above, the **Nelson-Äalen** estimator of the distribution function is

$$\hat{F}_{NA}(x) = \begin{cases} 0 & x < t_1 \\ 1 - \exp\left(-\sum_{j:t_j \leq x} \frac{s_j}{R_j}\right) & x \geq t_1 \end{cases}.$$

Note that the above expression is a result of the Nelson-Äalen estimator of the cumulative hazard function

$$\hat{H}(x) = \sum_{j:t_j \leq x} \frac{s_j}{R_j}$$

and the relationship between the survival function and cumulative hazard function, $\hat{S}_{NA}(x) = e^{-\hat{H}(x)}$.

Example 4.3.7. Actuarial Exam Question.

For observation i of a survival study:

- d_i is the left truncation point
- x_i is the observed value if not right censored
- u_i is the observed value if right censored

You are given:

Observation (i)	1	2	3	4	5	6	7	8	9	10
d_i	0	0	0	0	0	0	0	1.3	1.5	1.6
x_i	0.9	–	1.5	–	–	1.7	–	2.1	2.1	–
u_i	–	1.2	–	1.5	1.6	–	1.7	–	–	2.3

Calculate the Kaplan-Meier product-limit estimate, $\hat{S}(1.6)$

Show Example Solution

Solution. Recall the risk set $R_j = \sum_{i=1}^{n} \{I(x_i \geq t_j) + I(u_i \geq t_j) - I(d_i \geq t_j)\}$. Then

j	t_j	s_j	R_j	$\hat{S}(t_j)$
1	0.9	1	$10-3=7$	$1-\frac{1}{7}=\frac{6}{7}$
2	1.5	1	$8-2=6$	$\frac{6}{7}\left(1-\frac{1}{6}\right)=\frac{5}{7}$
3	1.7	1	$5-0=5$	$\frac{5}{7}\left(1-\frac{1}{5}\right)=\frac{4}{7}$
4	2.1	2	3	$\frac{4}{7}\left(1-\frac{2}{3}\right)=\frac{4}{21}$

The Kaplan-Meier estimate is therefore $\hat{S}(1.6) = \frac{5}{7}$.

Example 4.3.8. Actuarial Exam Question. - Continued.

a) Using the Nelson-Äalen estimator, calculate the probability that the loss on a policy exceeds 11, $\hat{S}_{NA}(11)$.
b) Calculate Greenwood's approximation to the variance of the product-limit estimate $\hat{S}(11)$.

Show Example Solution

Solution. As before, there are four event times (non-censored observations). For each time t_j, we can calculate the number of events s_j and the risk set R_j as the following:

j	t_j	s_j	R_j
1	4	2	10
2	8	1	5
3	12	1	2
4	15	1	1

The Nelson-Äalen estimate of $S(11)$ is $\hat{S}_{NA}(11) = e^{-\hat{H}(11)} = e^{-0.4} = 0.67$, since

$$\begin{aligned}\hat{H}(11) &= \sum_{j:t_j \leq 11} \frac{s_j}{R_j} = \sum_{j=1}^{2} \frac{s_j}{R_j} \\ &= \frac{2}{10} + \frac{1}{5} = 0.2 + 0.2 = 0.4.\end{aligned}$$

From earlier work, the Kaplan-Meier estimate of $S(11)$ is $\hat{S}(11) = 0.64$. Then Greenwood's estimate of the variance of the product-limit estimate of $S(11)$ is

$$\widehat{Var}(\hat{S}(11)) = (\hat{S}(11))^2 \sum_{j:t_j \leq 11} \frac{s_j}{R_j(R_j - s_j)} \quad = (0.64)^2 \left(\frac{2}{10(8)} + \frac{1}{5(4)} \right) = 0.0307.$$

Show Quiz Solution

4.4 Bayesian Inference

In this section, you learn how to:

- Describe the Bayesian model as an alternative to the frequentist approach and summarize the five components of this modeling approach.
- Summarize posterior distributions of parameters and use these posterior distributions to predict new outcomes.
- Use conjugate distributions to determine posterior distributions of parameters.

4.4.1 Introduction to Bayesian Inference

Up to this point, our inferential methods have focused on the **frequentist** setting, in which samples are repeatedly drawn from a population. The vector of

parameters $\boldsymbol{\theta}$ is fixed yet unknown, whereas the outcomes X are realizations of random variables.

In contrast, under the **Bayesian** framework, we view both the model parameters and the data as random variables. We are uncertain about the parameters $\boldsymbol{\theta}$ and use probability tools to reflect this uncertainty.

To get a sense of the Bayesian framework, begin by recalling Bayes' rule

$$\Pr(parameters|data) = \frac{\Pr(data|parameters) \times \Pr(parameters)}{\Pr(data)}$$

where

- $\Pr(parameters)$ is the distribution of the parameters, known as the prior distribution.
- $\Pr(data|parameters)$ is the sampling distribution. In a frequentist context, it is used for making inferences about the parameters and is known as the likelihood.
- $\Pr(parameters|data)$ is the distribution of the parameters having observed the data, known as the posterior distribution.
- $\Pr(data)$ is the marginal distribution of the data. It is generally obtained by integrating (or summing) the joint distribution of data and parameters over parameter values.

Why Bayes? There are several advantages of the Bayesian approach. First, we can describe the entire distribution of parameters conditional on the data. This allows us, for example, to provide probability statements regarding the likelihood of parameters. Second, the Bayesian approach provides a unified approach for estimating parameters. Some non-Bayesian methods, such as least squares, require a separate approach to estimate variance components. In contrast, in Bayesian methods, all parameters can be treated in a similar fashion. This is convenient for explaining results to consumers of the data analysis. Third, this approach allows analysts to blend prior information known from other sources with the data in a coherent manner. This topic is developed in detail in the credibility chapter. Fourth, Bayesian analysis is particularly useful for forecasting future responses.

Poisson - Gamma Special Case. To develop intuition, we consider the Poisson-Gamma case that holds a prominent position in actuarial applications. The idea is to consider a set of random variables $X_1, \ldots, X_n$ where each X_i could represent the number of claims for the ith policyholder. Assume that X_i has a Poisson distribution with parameter λ, analogous to the likelihood that we first saw in Chapter **??**. In a non-Bayesian (or frequentist) context, the parameter λ is viewed as an unknown quantity that is not random (it is said to be "fixed"). In the Bayesian context, the unknown parameter λ is viewed as uncertain and is modeled as a random variable. In this special case, we use the gamma distribution to reflect this uncertainty, the prior distribution.

Think of the following two-stage sampling scheme to motivate our probabilistic set-up.

1. In the first stage, the parameter λ is drawn from a gamma distribution.
2. In the second stage, for that value of λ, there are n draws from the same (identical) Poisson distribution that are independent, conditional on λ.

From this simple set-up, some important conclusions emerge.

- The distribution of X_i is no longer Poisson. For a special case, it turns out to be a negative binomial distribution (see the following "Snippet of Theory").
- The random variables $X_1, \ldots, X_n$ are not independent. This is because they share the common random variable λ.
- As in the frequentist context, the goal is to make statments about likely values of parameters such as λ given the observed data $X_1, \ldots, X_n$. However, because now both the parameter and the data are random variables, we can use the language of conditional probability to make such statements. As we will see in Section **??**, it turns out that the distribution of λ given the data $X_1, \ldots, X_n$ is also gamma (with updated parameters), a result that simplifies the task of inferring likely values of the parameter λ.

Show A Snippet of Theory

Let us demonstrate that the distribution of X is negative binomial. We assume that the distribution of X given λ is Poisson, so that

$$\Pr(X = x|\lambda) = \frac{\lambda^x}{\Gamma(x+1)} e^{-\lambda},$$

using notation $\Gamma(x+1) = x!$ for integer x. Assume that λ is a draw from a gamma distribution with fixed parameters, say, α and θ, so this has pdfprobability density function

$$f(\lambda) = \frac{\lambda^{\alpha-1}}{\theta^\alpha \Gamma(\alpha)} \exp(-\lambda/\theta).$$

We know that a pdf integrates to one. To make the development easier, define the reciprocal parameter $\theta_r = 1/\theta$ and so we have

$$\int_0^\infty f(\lambda)\ d\lambda = 1 \quad ==> \quad \theta_r^{-\alpha}\Gamma(\alpha) = \int_0^\infty \lambda^{\alpha-1} \exp\left(-\lambda\theta_r\right)\ d\lambda.$$

From Appendix Chapter **??** on iterated expectations, we have that the pmfprobability mass function of X can be computed in an iterated fashion as

$$\begin{aligned}
\Pr(X=x) &= \mathrm{E}\left\{\Pr(X=x|\lambda)\right\} \\
&= \int_0^\infty \Pr(X=x|\lambda) f(\lambda)\ d\lambda \\
&= \int_0^\infty \frac{\lambda^x}{\Gamma(x+1)} e^{-\lambda} \frac{\lambda^{\alpha-1}}{\theta^\alpha \Gamma(\alpha)} \exp(-\lambda/\theta)\ d\lambda \\
&= \frac{1}{\theta^\alpha \Gamma(x+1)\Gamma(\alpha)} \int_0^\infty \lambda^{x+\alpha-1} \exp\left(-\lambda(1+\frac{1}{\theta})\right)\ d\lambda \\
&= \frac{1}{\theta^\alpha \Gamma(x+1)\Gamma(\alpha)} \Gamma(x+\alpha) \left(1+\frac{1}{\theta}\right)^{-(x+\alpha)} \\
&= \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)} \left(\frac{1}{1+\theta}\right)^\alpha \left(\frac{\theta}{1+\theta}\right)^x .
\end{aligned}$$

Here, we used the gamma distribution equality with the substitution $\theta_r = 1+1/\theta$. As can be seen from Section **??**, this is a negative binomial distribution with parameter $r = \alpha$ and $\beta = \theta$.

In this chapter, we use small examples that can be done by hand in order to focus on the foundations. For practical implementation, analysts rely heavily on simulation methods using modern computational methods such as Markov Chain Monte Carlo (MCMC) simulation. We will get an exposure to simulation techniques in Chapter **??** but more intensive techniques such as MCMC requires yet more background. See **?** for an introduction to computational Bayesian methods from an actuarial perspective.

4.4.2 Bayesian Model

With the intuition developed in the preceding Section **??**, we now restate the Bayesian model with a bit more precision using mathematical notation. For simplicity, this summary assumes both the outcomes and parameters are continuous random variables. In the examples, we sometimes ask the viewer to apply these same principles to discrete versions. Conceptually both the continuous and discrete cases are the same; mechanically, one replaces a pdfprobability density function by a pmfprobability mass function and an integral by a sum.

As stated earlier, under the Bayesian perspective, the model parameters and data are both viewed as random. Our uncertainty about the parameters of the underlying data generating process is reflected in the use of probability tools.

Prior Distribution. Specifically, think about parameters $\boldsymbol{\theta}$ as a random vector and let $\pi(\boldsymbol{\theta})$ denote the distribution of possible outcomes. This is knowledge that we have before outcomes are observed and is called the prior distribution. Typically, the prior distribution is a regular distribution and so integrates or sums to one, depending on whether $\boldsymbol{\theta}$ is continuous or discrete. However, we

may be very uncertain (or have no clue) about the distribution of $\boldsymbol{\theta}$; the Bayesian machinery allows the following situation

$$\int \pi(\theta) d\theta = \infty,$$

in which case $\pi(\cdot)$ is called an **improper prior**.

Model Distribution. The distribution of outcomes given an assumed value of $\boldsymbol{\theta}$ is known as the model distribution and denoted as $f(x|\boldsymbol{\theta}) = f_{X|\boldsymbol{\theta}}(x|\boldsymbol{\theta})$. This is the usual frequentist mass or density function. This is simply the likelihood in the frequentist context and so it is also convenient to use this as a descriptor for the model distribution.

Joint Distribution. The distribution of outcomes and model parameters is, unsurprisingly, known as the joint distribution and denoted as $f(x, \boldsymbol{\theta}) = f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Marginal Outcome Distribution. The distribution of outcomes can be expressed as

$$f(x) = \int f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \; d\boldsymbol{\theta}.$$

This is analogous to a frequentist mixture distribution. In the mixture distribution, we combined (or "mixed") different subpopulations. In the Bayesian context, the marginal distribution is a combination of different realizations of parameters (in some literatures, you can think about this as combining different "states of nature").

Posterior Distribution of Parameters. After outcomes have been observed (hence the terminology "posterior"), one can use Bayes theorem to write the distribution as

$$\pi(\boldsymbol{\theta}|x) = \frac{f(x, \boldsymbol{\theta})}{f(x)} = \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(x)}$$

The idea is to update your knowledge of the distribution of $\boldsymbol{\theta}$ ($\pi(\boldsymbol{\theta})$) with the data x. Making statements about potential values of parameters is an important aspect of statistical inference.

4.4.3 Bayesian Inference

Summarizing the Posterior Distribution of Parameters

One way to summarize a distribution is to use a confidence interval type statement. To summarize the posterior distribution of parameters, the interval $[a, b]$ is said to be a $100(1-\alpha)\%$ credibility interval for $\boldsymbol{\theta}$ if

$$\Pr(a \leq \theta \leq b|\mathbf{x}) \geq 1 - \alpha.$$

For another approach to summarization, we can look to classical decision analysis. In this set-up, the loss function $l(\hat{\theta}, \theta)$ determines the penalty paid for using the estimate $\hat{\theta}$ instead of the true θ. The **Bayes estimate** is the value that minimizes the expected loss E $[l(\hat{\theta}, \theta)]$. Some important special cases include:

Loss function $l(\hat{\theta}, \theta)$	Descriptor	Bayes Estimate
$(\hat{\theta} - \theta)^2$	squared error loss	E$(\theta \vert X)$
$\vert\hat{\theta} - \theta\vert$	absolute deviation loss	median of $\pi(\theta \vert x)$
$I(\hat{\theta} = \theta)$	zero-one loss (for discrete probabilities)	mode of $\pi(\theta \vert x)$

Minimizing expected loss is a rigorous method for providing a single "best guess" about a likely value of a parameter, comparable to a frequentist estimator of the unknown (fixed) parameter.

Example 4.4.1. Actuarial Exam Question. You are given:

(i) In a portfolio of risks, each policyholder can have at most one claim per year.
(ii) The probability of a claim for a policyholder during a year is q.
(iii) The prior density is

$$\pi(q) = q^3/0.07, \quad 0.6 < q < 0.8$$

A randomly selected policyholder has one claim in Year 1 and zero claims in Year 2. For this policyholder, calculate the posterior probability that $0.7 < q < 0.8$.

Show Example Solution

Solution. The posterior density is proportional to the product of the likelihood function and prior density. Thus,

$$\pi(q|1,0) \propto f(1|q)\ f(0|q)\ \pi(q) \propto q(1-q)q^3 = q^4 - q^5$$

To get the exact posterior density, we integrate the above function over its range $(0.6, 0.8)$

$$\int_{0.6}^{0.8} q^4 - q^5 dq = \frac{q^5}{5} - \frac{q^6}{6}\Big|_{0.6}^{0.8} = 0.014069 \ \Rightarrow\ \pi(q|1,0) = \frac{q^4 - q^5}{0.014069}$$

Then

$$\Pr(0.7 < q < 0.8|1,0) = \int_{0.7}^{0.8} \frac{q^4 - q^5}{0.014069} dq = 0.5572$$

Example 4.4.2. Actuarial Exam Question. You are given:

(i) The prior distribution of the parameter Θ has probability density function:

$$\pi(\theta) = \frac{1}{\theta^2}, \quad 1 < \theta < \infty$$

(ii) Given $\Theta = \theta$, claim sizes follow a Pareto distribution with parameters $\alpha = 2$ and θ.

A claim of 3 is observed. Calculate the posterior probability that Θ exceeds 2.

Show Example Solution

Solution: The posterior density, given an observation of 3 is

$$\pi(\theta|3) = \frac{f(3|\theta)\pi(\theta)}{\int_1^\infty f(3|\theta)\pi(\theta)d\theta} = \frac{\frac{2\theta^2}{(3+\theta)^3}\frac{1}{\theta^2}}{\int_1^\infty 2(3+\theta)^{-3}d\theta} = \frac{2(3+\theta)^{-3}}{-(3+\theta)^{-2}|_1^\infty} = 32(3+\theta)^{-3}, \ \theta > 1$$

Then

$$\Pr(\Theta > 2|3) = \int_2^\infty 32(3+\theta)^{-3}d\theta = -16(3+\theta)^{-2}\big|_2^\infty = \frac{16}{25} = 0.64$$

Bayesian Predictive Distribution

For another type of statistical inference, it is often of interest to "predict" the value of a random outcome that is yet to be observed. Specifically, for new data y, the **predictive distribution** is

$$f(y|x) = \int f(y|\theta)\pi(\theta|x)d\theta.$$

It is also sometimes called a "posterior" distribution as the distribution of the new data is conditional on a base set of data.

Using squared error loss for the loss function, the **Bayesian prediction** of Y is

$$\begin{aligned} \mathrm{E}(Y|X) &= \int yf(y|X)dy = \int y\left(\int f(y|\theta)\pi(\theta|X)d\theta\right)dy \\ &= \int \mathrm{E}(Y|\theta)\pi(\theta|X)\ d\theta. \end{aligned}$$

As noted earlier, for some situations the distribution of parameters is discrete, not continuous. Having a discrete set of possible parameters allow us to think of them as alternative "states of nature," a helpful interpretation.

Example 4.4.3. Actuarial Exam Question. For a particular policy, the conditional probability of the annual number of claims given $\Theta = \theta$, and the probability distribution of Θ are as follows:

Number of Claims	0	1	2
Probability	2θ	θ	$1-3\theta$

θ	0.05	0.30
Probability	0.80	0.20

Two claims are observed in Year 1. Calculate the Bayesian prediction of the number of claims in Year 2.

Show Example Solution

Solution. Start with the posterior distribution of the parameter

$$\Pr(\theta|X) = \frac{\Pr(X|\theta)\Pr(\theta)}{\sum_\theta \Pr(X|\theta)\Pr(\theta)}$$

so

$$\begin{aligned}\Pr(\theta = 0.05|X = 2) &= \frac{\Pr(X = 2|\theta = 0.05)\Pr(\theta = 0.05)}{\Pr(X = 2|\theta = 0.05)\Pr(\theta = 0.05) + \Pr(X = 2|\theta = 0.3)\Pr(\theta = 0.3)} \\ &= \frac{(1 - 3 \times 0.05)(0.8)}{(1 - 3 \times 0.05)(0.8) + (1 - 3 \times 0.3)(0.2)} = \frac{68}{70}.\end{aligned}$$

Thus, $\Pr(\theta = 0.3|X = 1) = 1 - \Pr(\theta = 0.05|X = 1) = \frac{2}{70}$.

From the model distribution, we have

$$E(X|\theta) = 0 \times 2\theta + 1 \times \theta + 2 \times (1 - 3\theta) = 2 - 5\theta.$$

Thus,

$$\begin{aligned}E(Y|X) &= \sum_\theta \mathrm{E}(Y|\theta)\pi(\theta|X) \\ &= \mathrm{E}(Y|\theta = 0.05)\pi(\theta = 0.05|X) + \mathrm{E}(Y|\theta = 0.3)\pi(\theta = 0.3|X) \\ &= (2 - 5(0.05))\frac{68}{70} + (2 - 5(0.3))\frac{2}{70} = 1.714.\end{aligned}$$

Example 4.4.4. Actuarial Exam Question.

You are given:

(i) Losses on a company's insurance policies follow a Pareto distribution with probability density function:

$$f(x|\theta) = \frac{\theta}{(x+\theta)^2}, \quad 0 < x < \infty$$

(ii) For half of the company's policies $\theta = 1$, while for the other half $\theta = 3$.

For a randomly selected policy, losses in Year 1 were 5. Calculate the posterior probability that losses for this policy in Year 2 will exceed 8.

Show Example Solution

Solution. We are given the prior distribution of θ as $\Pr(\theta = 1) = \Pr(\theta = 3) = \frac{1}{2}$, the conditional distribution $f(x|\theta)$, and the fact that we observed $X_1 = 5$. The goal is to find the predictive probability $\Pr(X_2 > 8|X_1 = 5)$.

The posterior probabilities are

$$\begin{aligned}\Pr(\theta = 1|X_1 = 5) &= \frac{f(5|\theta = 1)\Pr(\theta = 1)}{f(5|\theta = 1)\Pr(\theta = 1) + f(5|\theta = 3)\Pr(\theta = 3)} \\ &= \frac{\frac{1}{36}(\frac{1}{2})}{\frac{1}{36}(\frac{1}{2}) + \frac{3}{64}(\frac{1}{2})} = \frac{\frac{1}{72}}{\frac{1}{72} + \frac{3}{128}} = \frac{16}{43}\end{aligned}$$

$$\begin{aligned}\Pr(\theta = 3|X_1 = 5) &= \frac{f(5|\theta = 3)\Pr(\theta = 3)}{f(5|\theta = 1)\Pr(\theta = 1) + f(5|\theta = 3)\Pr(\theta = 3)} \\ &= 1 - \Pr(\theta = 1|X_1 = 5) = \frac{27}{43}\end{aligned}$$

Note that the conditional probability that losses exceed 8 is

$$\begin{aligned}\Pr(X_2 > 8|\theta) &= \int_8^\infty f(x|\theta)dx \\ &= \int_8^\infty \frac{\theta}{(x+\theta)^2}dx = -\frac{\theta}{x+\theta}\Big|_8^\infty = \frac{\theta}{8+\theta}\end{aligned}$$

The predictive probability is therefore

$$\begin{aligned}\Pr(X_2 > 8|X_1 = 5) &= \Pr(X_2 > 8|\theta = 1)\Pr(\theta = 1|X_1 = 5) + \Pr(X_2 > 8|\theta = 3)\Pr(\theta = 3|X_1 = 5) \\ &= \frac{1}{8+1}\left(\frac{16}{43}\right) + \frac{3}{8+3}\left(\frac{27}{43}\right) = 0.2126\end{aligned}$$

Example 4.4.5. Actuarial Exam Question.

You are given:

(i) The probability that an insured will have at least one loss during any year is p.
(ii) The prior distribution for p is uniform on $[0, 0.5]$.
(iii) An insured is observed for 8 years and has at least one loss every year.

Calculate the posterior probability that the insured will have at least one loss during Year 9.

Show Example Solution

Solution. The posterior probability density is

$$\pi(p|1,1,1,1,1,1,1,1) \propto \Pr(1,1,1,1,1,1,1,1|p)\ \pi(p) = p^8(2) \propto p^8$$
$$\Rightarrow \pi(p|1,1,1,1,1,1,1,1) = \frac{p^8}{\int_0^5 p^8 dp} = \frac{p^8}{(0.5^9)/9} = 9(0.5^{-9})p^8$$

Thus, the posterior probability that the insured will have at least one loss during Year 9 is

$$\Pr(X_9 = 1|1,1,1,1,1,1,1,1) = \int_0^5 \Pr(X_9 = 1|p)\pi(p|1,1,1,1,1,1,1,1)dp$$
$$= \int_0^5 p(9)(0.5^{-9})p^8 dp = 9(0.5^{-9})(0.5^{10})/10 = 0.45$$

Example 4.4.6. Actuarial Exam Question. You are given:

(i) Each risk has at most one claim each year.

Type of Risk	Prior Probability	Annual Claim Probability
I	0.7	0.1
II	0.2	0.2
III	0.1	0.4

One randomly chosen risk has three claims during Years 1-6. Calculate the posterior probability of a claim for this risk in Year 7.

Show Example Solution

Solution. The probabilities are from a binomial distribution with 6 trials in which 3 successes were observed.

$$\Pr(3|\text{I}) = \binom{6}{3}(0.1^3)(0.9^3) = 0.01458$$
$$\Pr(3|\text{II}) = \binom{6}{3}(0.2^3)(0.8^3) = 0.08192$$
$$\Pr(3|\text{III}) = \binom{6}{3}(0.4^3)(0.6^3) = 0.27648$$

The probability of observing three successes is

$$\begin{aligned}\Pr(3) &= \Pr(3|\text{I})\Pr(\text{I}) + \Pr(3|\text{II})\Pr(\text{II}) + \Pr(3|\text{III})\Pr(\text{III}) \\ &= 0.7(0.01458) + 0.2(0.08192) + 0.1(0.27648) = 0.054238\end{aligned}$$

The three posterior probabilities are

$$\begin{aligned}\Pr(\text{I}|3) &= \frac{\Pr(3|\text{I})\Pr(\text{I})}{\Pr(3)} = \frac{0.7(0.01458)}{0.054238} = 0.18817 \\ \Pr(\text{II}|3) &= \frac{\Pr(3|\text{II})\Pr(\text{II})}{\Pr(3)} = \frac{0.2(0.08192)}{0.054238} = 0.30208 \\ \Pr(\text{III}|3) &= \frac{\Pr(3|\text{III})\Pr(\text{III})}{\Pr(3)} = \frac{0.1(0.27648)}{0.054238} = 0.50975\end{aligned}$$

The posterior probability of a claim is then

$$\begin{aligned}\Pr(\text{claim}|3) &= \Pr(\text{claim}|\text{I})\Pr(\text{I}|3) + \Pr(\text{claim}|\text{II})\Pr(\text{II}|3) + \Pr(\text{claim}|\text{III})\Pr(\text{III}|3) \\ &= 0.1(0.18817) + 0.2(0.30208) + 0.4(0.50975) = 0.28313\end{aligned}$$

4.4.4 Conjugate Distributions

In the Bayesian framework, the key to statistical inference is understanding the posterior distribution of the parameters. As described in Section **??**, modern data analysis using Bayesian methods utilize computationally intensive techniques such as MCMCMarkov Chain Monte Carlo simulation. Another approach for computing posterior distributions are based on **conjugate distributions.** Although this approach is available only for a limited number of distributions, it has the appeal that it provides closed-form expressions for the distributions, allowing for easy interpretations of results.

To relate the prior and posterior distributions of the parameters, we have the relationship

$$\begin{array}{ccc}\pi(\boldsymbol{\theta}|x) & = & \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(x)} \\ & \propto & f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ \text{Posterior} & \text{is proportional to} & \text{likelihood} \times \text{prior.}\end{array}$$

For conjugate distributions, the posterior and the prior come from the same family of distributions. The following illustration looks at the Poisson-gamma special case, the most well-known in actuarial applications.

Special Case – Poisson-Gamma - Continued. Assume a Poisson(λ) model distribution and that λ follows a gamma(α, θ) prior distribution. Then, the

posterior distribution of λ given the data follows a gamma distribution with new parameters $\alpha_{post} = \sum_i x_i + \alpha$ and $\theta_{post} = 1/(n + 1/\theta)$.

Show Special Case Details

Special Case – Poisson-Gamma - Continued. The model distribution is

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}.$$

The prior distribution is

$$\pi(\lambda) = \frac{(\lambda/\theta)^{\alpha} \exp(-\lambda/\theta)}{\lambda \Gamma(\alpha)}.$$

Thus, the posterior distribution is proportional to

$$\begin{aligned} \pi(\lambda|\mathbf{x}) &\propto f(\mathbf{x}|\theta)\pi(\lambda) \\ &= C\lambda^{\sum_i x_i + \alpha - 1} \exp(-\lambda(n + 1/\theta)) \end{aligned}$$

where C is a constant. We recognize this to be a gamma distribution with new parameters $\alpha_{post} = \sum_i x_i + \alpha$ and $\theta_{post} = 1/(n + 1/\theta)$. Thus, the gamma distribution is a conjugate prior for the Poisson model distribution.

Example 4.4.7. Actuarial Exam Question.

You are given:

(i) The conditional distribution of the number of claims per policyholder is Poisson with mean λ.
(ii) The variable λ has a gamma distribution with parameters α and θ.
(iii) For policyholders with 1 claim in Year 1, the Bayes prediction for the number of claims in Year 2 is 0.15.
(iv) For policyholders with an average of 2 claims per year in Year 1 and Year 2, the Bayes prediction for the number of claims in Year 3 is 0.20.

Calculate θ.

Show Example Solution

Solution.

Since the conditional distribution of the number of claims per policyholder, $E(X|\lambda) = Var(X|\lambda) = \lambda$, the Bayes prediction is

$$\mathrm{E}(X_2|X_1) = \int \mathrm{E}(X_2|\lambda)\pi(\lambda|X_1)d\lambda = \alpha_{new}\theta_{new}$$

because the posterior distribution is gamma with parameters α_{new} and θ_{new}.

For year 1, we have

$$0.15 = (X_1 + \alpha) \times \frac{1}{n + 1/\theta} = (1 + \alpha) \times \frac{1}{1 + 1/\theta},$$

so $0.15(1 + 1/\theta) = 1 + \alpha$. For year 2, we have

$$0.2 = (X_1 + X_2 + \alpha) \times \frac{1}{n + 1/\theta} = (4 + \alpha) \times \frac{1}{2 + 1/\theta},$$

so $0.2(2 + 1/\theta) = 4 + \alpha$. Equating these yields

$$0.2(2 + 1/\theta) = 3 + 0.15(1 + 1/\theta)$$

resulting in $\theta = 1/55 = 0.018182$.

Closed-form expressions means that results can be readily interpreted and easily computed; hence, conjugate distributions are useful in actuarial practice. Two other special cases used extensively are:

- The uncertainty of parameters is summarized using a beta distribution and the outcomes have a (conditional on the parameter) binomial distribution.
- The uncertainty of parameters is summarized using a normal distribution and the outcomes are conditionally normally distributed.

Additional results on conjugate distributions are summarized in the Appendix Section **??**.

Show Quiz Solution

4.5 Further Resources and Contributors

Exercises

Here are a set of exercises that guide the viewer through some of the theoretical foundations of **Loss Data Analytics**. Each tutorial is based on one or more questions from the professional actuarial examinations, typically the Society of Actuaries Exam C.

Model Selection Guided Tutorials

Contributors

- **Edward W. (Jed) Frees** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter.

Email: jfrees@bus.wisc.edu for chapter comments and suggested improvements.

- Chapter reviewers include: Andrew Kwon-Nakamura, Hirokazu (Iwahiro) Iwasawa, Eren Dodd.

Chapter 5

Aggregate Loss Models

Chapter Preview. This chapter introduces probability models for describing the aggregate (total) claims that arise from a portfolio of insurance contracts. We present two standard modeling approaches, the individual risk model and the collective risk model. Further, we discuss strategies for computing the distribution of the aggregate claims, including exact methods for special cases, recursion, and simulation. Finally, we examine the effects of individual policy modifications such as deductibles, coinsurance, and inflation, on the frequency and severity distributions, and thus the aggregate loss distribution.

5.1 Introduction

The objective of this chapter is to build a probability model to describe the aggregate claims by an insurance system occurring in a fixed time period. The insurance system could be a single policy, a group insurance contract, a business line, or an entire book of an insurer's business. In this chapter, aggregate claims refer to either the number or the amount of claims from a portfolio of insurance contracts. However, the modeling framework can be readily applied in the more general setup.

Consider an insurance portfolio of n individual contracts, and let S denote the aggregate losses of the portfolio in a given time period. There are two approaches to modeling the aggregate losses S, the individual risk model and the collective risk model. The individual risk model emphasizes the loss from each individual contract and represents the aggregate losses as:

$$S_n = X_1 + X_2 + \cdots + X_n,$$

where X_i $(i = 1, \ldots, n)$ is interpreted as the loss amount from the ith contract. It is worth stressing that n denotes the number of contracts in the portfolio and

thus is a fixed number rather than a random variable. For the individual risk model, one usually assumes X_i's are independent. Because of different contract features such as coverage and exposure, X_i's are not necessarily identically distributed. A notable feature of the distribution of each X_i is the probability mass at zero corresponding to the event of no claims.

The collective risk model represents the aggregate losses in terms of a frequency distribution and a severity distribution:

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

Here, one thinks of a random number of claims N that may represent either the number of losses or the number of payments. In contrast, in the individual risk model, we use a fixed number of contracts n. We think of $X_1, X_2, \ldots, X_N$ as representing the amount of each loss. Each loss may or may not correspond to a unique contract. For instance, there may be multiple claims arising from a single contract. It is natural to think about $X_i > 0$ because if $X_i = 0$ then no claim has occurred. Typically we assume that conditional on $N = n$, $X_1, X_2, \ldots, X_n$ are iid random variables. The distribution of N is known as the frequency distribution, and the common distribution of X is known as the severity distribution. We further assume N and X are independent. With the collective risk model, we may decompose the aggregate losses into the frequency (N) process and the severity (X) model. This flexibility allows the analyst to comment on these two separate components. For example, sales growth due to lower underwriting standards could lead to higher frequency of losses but might not affect severity. Similarly, inflation or other economic forces could have an impact on severity but not on frequency.

Show Quiz Solution

5.2 Individual Risk Model

As noted earlier, for the individual risk model, we think of X_i as the loss from ith contract and interpret

$$S_n = X_1 + X_2 + \cdots + X_n$$

to be the aggregate loss from all contracts in a portfolio or group of contracts. Here, the X_i's are not necessarily identically distributed and we have

$$\mathrm{E}(S_n) = \sum_{i=1}^{n} \mathrm{E}(X_i) \ .$$

Under the independence assumption on X_i's (which implies $\text{Cov}\,(X_i, X_j) = 0$ for all $i \neq j$), it can further be shown that

$$\begin{aligned}
\text{Var}(S_n) &= \sum_{i=1}^{n} \text{Var}(X_i) \\
P_{S_n}(z) &= \prod_{i=1}^{n} P_{X_i}(z) \\
M_{S_n}(t) &= \prod_{i=1}^{n} M_{X_i}(t),
\end{aligned}$$

where $P_{S_n}(\cdot)$ and $M_{S_n}(\cdot)$ are the probability generating function (pgf) and the moment generating function (mgf) of S_n, respectively. The distribution of each X_i contains a probability mass at zero, corresponding to the event of no claims from the ith contract. One strategy to incorporate the zero mass in the distribution is to use the two-part framework:

$$X_i = I_i \times B_i = \begin{cases} 0\,, & \text{if } \ I_i = 0 \\ B_i\,, & \text{if } \ I_i = 1 \end{cases}$$

Here, I_i is a Bernoulli variable indicating whether or not a loss occurs for the ith contract, and B_i is a random variable with nonnegative support representing the amount of losses of the contract given loss occurrence. Assume that $I_1, \ldots, I_n, B_1, \ldots, B_n$ are mutually independent. Denote $\Pr(I_i = 1) = q_i$, $\mu_i = \text{E}(B_i)$, and $\sigma_i^2 = \text{Var}(B_i)$. It can be shown (see Technical Supplement 5.A.1 for details) that

$$\begin{aligned}
\text{E}(S_n) &= \sum_{i=1}^{n} \ q_i \ \mu_i \\
\text{Var}(S_n) &= \sum_{i=1}^{n} \left(q_i \sigma_i^2 + q_i (1 - q_i) \mu_i^2 \right) \\
P_{S_n}(z) &= \prod_{i=1}^{n} \left(1 - q_i + q_i P_{B_i}(z) \right) \\
M_{S_n}(t) &= \prod_{i=1}^{n} \left(1 - q_i + q_i M_{B_i}(t) \right)
\end{aligned}$$

A special case of the above model is when B_i follows a degenerate distribution with $\mu_i = b_i$ and $\sigma_i^2 = 0$. One example is term life insurance or a pure endowment insurance where b_i represents the insurance benefit amount of the ith contract.

Another strategy to accommodate the zero mass in the loss from each contract is to consider them in aggregate on the portfolio level, as in the collective risk model. Here, the aggregate loss is $S_N = X_1 + \cdots X_N$, where N is a random

variable representing the number of non-zero claims that occurred out of the entire group of contracts. Thus, not every contract in the portfolio may be represented in this sum, and $S_N = 0$ when $N = 0$. The collective risk model will be discussed in detail in the next section.

Example 5.2.1. Actuarial Exam Question. An insurance company sold 300 fire insurance policies as follows:

Number of Policies	Policy Maximum (M_i)	Probability of Claim Per Policy (q_i)
100	400	0.05
200	300	0.06

You are given:
(i) The claim amount for each policy, X_i, is uniformly distributed between 0 and the policy maximum M_i.
(ii) The probability of more than one claim per policy is 0.
(iii) Claim occurrences are independent.

Calculate the mean, E (S_{300}), and variance, Var (S_{300}), of the aggregate claims. How would these results change if every claim is equal to the policy maximum?

Show Example Solution

Solution. The aggregate claims are $S_{300} = X_1 + \cdots + X_{300}$, where $X_1, \ldots, X_{300}$ are independent but not identically distributed. Policy claims amounts are uniformly distributed on $(0, M_i)$, so the mean claim amount is $M_i/2$ and the variance is$M_i^2/12$. Thus, for policy $i = 1, \ldots, 300$, we have

Number of Policies	Policy Maximum (M_i)	Probability of Claim Per Policy (q_i)	Mean Amount (μ_i)	Variance Amount (σ_i^2)
100	400	0.05	200	$400^2/12$
200	300	0.06	150	$300^2/12$

The mean of the aggregate claims is

$$\text{E}\ (S_{300}) = \sum_{i=1}^{300} q_i \mu_i = 100\left\{0.05(200)\right\} + 200\left\{0.06(150)\right\} = 2,800$$

The variance of the aggregate claims is

$$\begin{aligned}\text{Var }(S_{300}) &= \sum_{i=1}^{300}\left(q_i\sigma_i^2+q_i(1-q_i)\mu_i^2\right) \quad \text{since } X_i\text{'s are independent}\\ &= 100\left\{0.05\left(\frac{400^2}{12}\right)+0.05(1-0.05)200^2\right\}+200\left\{0.06\left(\frac{300^2}{12}\right)+0.06(1-0.06)150^2\right\}\\ &= 600,467.\end{aligned}$$

Follow-Up. Now suppose everybody receives the policy maximum M_i if a claim occurs. What is the expected aggregate loss E $(\tilde{S})$ and variance of the aggregate loss Var $(\tilde{S})$?

Each policy claim amount X_i is now deterministic and fixed at M_i instead of a randomly distributed amount, so $\sigma_i^2 = \text{Var }(X_i) = 0$ and $\mu_i = M_i$. Again, the probability of a claim occurring for each policy is q_i. Under these circumstances, the expected aggregate loss is

$$\text{E }(\tilde{S}) = \sum_{i=1}^{300} q_i\mu_i = 100\left\{0.05(400)\right\}+200\left\{0.06(300)\right\} = 5,600$$

The variance of the aggregate loss is

$$\begin{aligned}\text{Var }(\tilde{S}) &= \sum_{i=1}^{300}\left(q_i\sigma_i^2+q_i(1-q_i)\mu_i^2\right) = \sum_{i=1}^{300}\left(q_i(1-q_i)\mu_i^2\right)\\ &= 100\left\{(0.05)(1-0.05)400^2\right\}+200\left\{(0.06)(1-0.06)300^2\right\}\\ &= 1,775,200\end{aligned}$$

The individual risk model can also be used for claim frequency. If X_i denotes the number of claims from the ith contract, then S_n is interpreted as the total number of claims from the portfolio. In this case, the above two-part framework still applies since there is a probability mass at zero for contracts that do not experience any claims. Assume X_i belongs to the $(a, b, 0)$ class with pmf denoted by $p_{ik} = \Pr(X_i = k)$ for $k = 0, 1, \ldots$ (see Section **??**). Let X_i^T denote the associated zero-truncated distribution in the $(a, b, 1)$ class with pmf $p_{ik}^T = p_{ik}/(1 - p_{i0})$ for $k = 1, 2, \ldots$ (see Section **??**). Using the relationship between their probability generating functions (see Technical Supplement 5.A.2 for details):

$$P_{X_i}(z) = p_{i0} + (1 - p_{i0})P_{X_i^T}(z),$$

we can write $X_i = I_i \times B_i$ with $q_i = \Pr(I_i = 1) = \Pr(X_i > 0) = 1 - p_{i0}$ and $B_i = X_i^T$. Notice that in this case, we have a zero-modified distribution since the I_i variable covers the modified probability mass at zero with $q_i = \Pr(I_i = 1)$, while the $B_i = X_i^T$ covers the discrete non-zero frequency portion. See Section **??** for the relationship between zero-truncated and zero-modified distributions.

Example 5.2.2. An insurance company sold a portfolio of 100 independent homeowners insurance policies, each of which has claim frequency following a zero-modified Poisson distribution, as follows:

Type of Policy	Number of Policies	Probability of At Least 1 Claim	λ
Low-risk	40	0.03	1
High-risk	60	0.05	2

Find the expected value and variance of the claim frequency for the entire portfolio.

Show Example Solution

Solution. For each policy, we can write the zero-modified Poisson claim frequency N_i as $N_i = I_i \times B_i$, where

$$q_i = \Pr(I_i = 1) = \Pr(N_i > 0) = 1 - p_{i0}$$

For the low-risk policies, we have $q_i = 0.03$ and for the high-risk policies, we have $q_i = 0.05$. Further, $B_i = N_i^T$, the zero-truncated version of N_i. Thus, we have

$$\mu_i = \mathrm{E}(B_i) = \mathrm{E}(N_i^T) = \frac{\lambda}{1 - e^{-\lambda}}$$

$$\sigma_i^2 = \mathrm{Var}(B_i) = \mathrm{Var}(N_i^T) = \frac{\lambda[1 - (\lambda + 1)e^{-\lambda}]}{(1 - e^{-\lambda})^2}$$

Let the portfolio claim frequency be $S_n = \sum_{i=1}^{n} N_i$. Using the formulas above, the expected claim frequency of the portfolio is

$$\begin{aligned}
\mathrm{E}\ (S_n) &= \sum_{i=1}^{100} q_i \mu_i \\
&= 40 \left[0.03 \left(\frac{1}{1 - e^{-1}} \right) \right] + 60 \left[0.05 \left(\frac{2}{1 - e^{-2}} \right) \right] \\
&= 40(0.03)(1.5820) + 60(0.05)(2.3130) = 8.8375
\end{aligned}$$

The variance of the claim frequency of the portfolio is

$$\begin{aligned}
\mathrm{Var}\ (S_n) &= \sum_{i=1}^{100} \left(q_i \sigma_i^2 + q_i (1 - q_i) \mu_i^2 \right) \\
&= 40 \left[0.03 \left(\frac{1 - 2e^{-1}}{(1 - e^{-1})^2} \right) + 0.03(0.97)(1.5820^2) \right] + 60 \left[0.05 \left(\frac{2[1 - 3e^{-2}]}{(1 - e^{-2})^2} \right) + 0.05(0.95)(2.3130^2) \right] \\
&= 23.7214
\end{aligned}$$

Note that equivalently, we could have calculated the mean and variance of an individual policy directly using the relationship between the zero-modified and zero-truncated Poisson distributions (see Section **??**).

To understand the distribution of the aggregate loss, one could use the central limit theorem to approximate the distribution of S_n for large n. Denote $\mu_{S_n} = \mathrm{E}(S_n)$ and $\sigma^2_{S_n} = \mathrm{Var}(S_n)$ and let $Z \sim N(0,1)$, a standard normal random variable with cdf Φ. Then the cdf of S_n can be approximated as follows:

$$\begin{aligned} F_{S_n}(s) = \Pr(S_n \le s) &= \Pr\left(\frac{S_n - \mu_{S_n}}{\sigma_{S_n}} \le \frac{s - \mu_{S_n}}{\sigma_{S_n}}\right) \\ &\approx \Pr\left(Z \le \frac{s - \mu_{S_n}}{\sigma_{S_n}}\right) = \Phi\left(\frac{s - \mu_S}{\sigma_S}\right). \end{aligned}$$

Example 5.2.3. Actuarial Exam Question - Follow-Up. As in the Example 5.2.1 earlier, an insurance company sold 300 fire insurance policies, with claim amounts X_i uniformly distributed between 0 and the policy maximum M_i. Using the normal approximation, calculate the probability that the aggregate claim amount S_{300} exceeds $\$3,500$.

Show Example Solution

Solution. We have seen earlier that $\mathrm{E}(S_{300}) = 2,800$ and $\mathrm{Var}(S_{300}) = 600,467$. Then

$$\begin{aligned} \Pr(S_{300} > 3,500) &= 1 - \Pr(S_{300} \le 3,500) \\ &\approx 1 - \Phi\left(\frac{3,500 - 2,800}{\sqrt{600,467}}\right) = 1 - \Phi\left(0.90334\right) \\ &= 1 - 0.8168 = 0.1832 \end{aligned}$$

For small n, the distribution of S_n is likely skewed, and the normal approximation would be a poor choice. To examine the aggregate loss distribution, we go back to the basics and first principles. Specifically, the distribution can be derived recursively. Define $S_k = X_1 + \cdots + X_k, k = 1, \ldots, n$.

For $k = 1$:

$$F_{S_1}(s) = \Pr(S_1 \le s) = \Pr(X_1 \le s) = F_{X_1}(s).$$

For $k = 2, \ldots, n$:

$$\begin{aligned} F_{S_k}(s) &= \Pr(X_1 + \cdots + X_k \le s) = \Pr(S_{k-1} + X_k \le s) \\ &= \mathrm{E}_{X_k}\left[\Pr(S_{k-1} \le s - X_k | X_k)\right] = \mathrm{E}_{X_k}\left[F_{S_{k-1}}(s - X_k)\right]. \end{aligned}$$

A special case is when X_i's are identically distributed. Let $F_X(x) = \Pr(X \le x)$ be the common distribution of $X_i,\ i = 1, \ldots, n$. We define

$$F^{*n}_X(x) = \Pr(X_1 + \cdots + X_n \le x)$$

the n-fold convolution of F_X. More generally, we can compute F^{*n}_X recursively. Begin the recursion at $k = 1$ using $F^{*1}_X(x) = F_X(x)$. Next, for $k = 2$, we have

$$\begin{aligned} F_X^{*2}(x) &= \Pr(X_1 + X_2 \le x) = \mathrm{E}_{X_2}\left[\Pr(X_1 \le x - X_2 | X_2)\right] \\ &= \mathrm{E}_{X_2}\left[F(x - X_2)\right] \\ &= \begin{cases} \int_0^x F(x-y)f(y)dy & \text{for continuous } X_i\text{'s} \\ \sum_{y \le x} F(x-y)f(y) & \text{for discrete } X_i\text{'s} \end{cases} \end{aligned}$$

Recall $F(0) = 0$.

Similarly for $k = n$, we have $S_n = X_1 + X_2 + \cdots + X_n$ and

$$\begin{aligned} F^{*n}(x) &= \Pr(S_n \le x) = \Pr(S_{n-1} + X_n \le x) \\ &= \mathrm{E}_{X_n}\left[\Pr(S_{n-1} \le x - X_n | X_n)\right] \\ &= \mathrm{E}_X\left[F^{*(n-1)}(x - X)\right] \\ &= \begin{cases} \int_0^x F^{*(n-1)}(x-y)f(y)dy & \text{for continuous } X_i\text{'s} \\ \sum_{y \le x} F^{*(n-1)}(x-y)f(y) & \text{for discrete } X_i\text{'s} \end{cases} \end{aligned}$$

When X_i's are independent and belong to the same family of distributions, there are some simple cases where S_n has a closed form. This makes it easy to compute $\Pr(S_n \le x)$. This property is known as closed under convolution, meaning the distribution of the sum of independent random variables belongs to the same family of distributions as that of the component variables, just with different parameters. Examples include:

Table of Closed Form Partial Sum Distributions

Distribution of X_i	Abbreviation	Distribution of S_n
Normal with mean μ_i and variance σ_i^2	$N(\mu_i, \sigma_i^2)$	$N\left(\sum_{i=1}^n \mu_i,\ \sum_{i=1}^n \sigma_i^2\right)$
Exponential with mean θ	$Exp(\theta)$	$Gam(n, \theta)$
Gamma with shape α_i and scale θ	$Gam(\alpha_i, \theta)$	$Gam\left(\sum_{i=1}^n \alpha_i, \theta\right)$
Poisson with mean (and variance) λ_i	$Poi(\lambda_i)$	$Poi\left(\sum_{i=1}^n \lambda_i\right)$
Binomial with m_i trials and q success probability	$Bin(m_i, q)$	$Bin\left(\sum_{i=1}^n m_i, q\right)$
Geometric with mean β	$Geo(\beta)$	$NB(\beta, n)$
Negative binomial with mean $r_i\beta$ and variance $r_i\beta(1+\beta)$	$NB(\beta, r_i)$	$NB\left(\beta, \sum_{i=1}^n r_i\right)$

Example 5.2.4. Gamma Distribution. Assume that $X_1, \ldots, X_n$ are independent random variables with $X_i \sim Gam(\alpha_i, \theta)$. The mgf of X_i is $M_{X_i}(t) = (1 - \theta t)^{-\alpha_i}$. Thus, the mgf of the sum $S_n = X_1 + \cdots + X_n$ is

$$\begin{aligned} M_{S_n}(t) &= \prod_{i=1}^n M_{X_i}(t) \quad \text{from the independence of } X_i\text{'s} \\ &= \prod_{i=1}^n (1 - \theta t)^{-\alpha_i} = (1 - \theta t)^{-\sum_{i=1}^n \alpha_i}, \end{aligned}$$

which is the mgf of a gamma random variable with parameters $(\sum_{i=1}^{n}\alpha_i, \theta)$. Thus, $S_n \sim Gam(\sum_{i=1}^{n}\alpha_i, \theta)$.

Example 5.2.5. Negative Binomial Distribution. Assume that $X_1, \ldots, X_n$ are independent random variables with $X_i \sim NB(\beta, r_i)$. The pgf of X_i is $P_{X_i}(z) = [1-\beta(z-1)]^{-r_i}$. Thus, the pgf of the sum $S_n = X_1 + \cdots + X_n$ is

$$\begin{aligned} P_{S_n}(z) &= \mathrm{E}\left[z^{S_n}\right] = \mathrm{E}\left[z^{X_1+\cdots+X_n}\right] = \mathrm{E}\left[z^{X_1}z^{X_2}\cdots z^{X_n}\right] \\ &= \mathrm{E}\left[z^{X_1}\right]\cdots\mathrm{E}\left[z^{X_n}\right] \quad \text{under the independence of } X_i\text{'s} \\ &= \prod_{i=1}^{n} P_{X_i}(z) = \prod_{i=1}^{n}[1-\beta(z-1)]^{-r_i} = [1-\beta(z-1)]^{-\sum_{i=1}^{n} r_i}\ , \end{aligned}$$

which is the pgf of a negative binomial random variable with parameters $(\beta, \sum_{i=1}^{n} r_i)$. Thus, $S_n \sim NB(\beta, \sum_{i=1}^{n} r_i)$.

Example 5.2.6. Actuarial Exam Question (modified). The annual number of doctor visits for each individual in a family of 4 has geometric distribution with mean 1.5. The annual numbers of visits for the family members are mutually independent. An insurance pays 100 per doctor visit beginning with the 4th visit per family. Calculate the probability that the family will receive an insurance payment this year.

Show Example Solution

Solution. Let $X_i \sim Geo(\beta = 1.5)$ be the number of doctor visits for one individual in the family and $S_4 = X_1 + X_2 + X_3 + X_4$ be the number of doctor visits for the family. The sum of 4 independent geometric random variables each with mean $\beta = 1.5$ follows a negative binomial distribution, i.e. $S_4 \sim NB(\beta = 1.5, r = 4)$.

If the insurance pays 100 per visit beginning with the 4th visit for the family, then the family will not receive an insurance payment if they have less than 4 claims. This probability is

$$\begin{aligned} \Pr(S_4 < 4) &= \Pr(S_4 = 0) + \Pr(S_4 = 1) + \Pr(S_4 = 2) + \Pr(S_4 = 3) \\ &= (1+1.5)^{-4} + \frac{4(1.5)}{(1+1.5)^5} + \frac{4(5)(1.5^2)}{2(1+1.5)^6} + \frac{4(5)(6)(1.5^3)}{3!(1+1.5)^7} \\ &= 0.0256 + 0.0614 + 0.0922 + 0.1106 = 0.2898 \end{aligned}$$

Show Quiz Solution

5.3 Collective Risk Model

5.3.1 Moments and Distribution

Under the collective risk model $S_N = X_1 + \cdots + X_N$, $\{X_i\}$ are iid, and independent of N. Let $\mu = \mathrm{E}(X_i)$ and $\sigma^2 = \mathrm{Var}(X_i)$ for all i. Using the law of iterated expectations, the mean of the aggregate loss is

$$\mathrm{E}(S_N) = \mathrm{E}_N[\mathrm{E}_S(S|N)] = \mathrm{E}_N(N\mu) = \mu\mathrm{E}(N).$$

Using the law of total variance, the variance of the aggregate loss is

$$\begin{aligned}
\mathrm{Var}(S_N) &= \mathrm{E}_N[\mathrm{Var}(S_N|N)] + \mathrm{Var}_N[\mathrm{E}(S_N|N)] \\
&= \mathrm{E}_N\left[\mathrm{Var}(X_1 + \cdots + X_N)\right] + \mathrm{Var}_N\left[\mathrm{E}(X_1 + \cdots + X_N)\right] \\
&= \mathrm{E}_N\left[\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_N) + 2\mathrm{Cov}(X_1, X_2) + \cdots + \mathrm{Cov}(X_{N-1}, X_N)\right] + \mathrm{Var}_N\left[\mathrm{E}(X_1) + \cdots\right. \\
&= \mathrm{E}_N[N\sigma^2] + \mathrm{Var}_N[N\mu] \quad \text{since } \mathrm{Cov}(X_i, X_j) = 0 \text{ for all } i \neq j \text{ by independence} \\
&= \sigma^2\mathrm{E}[N] + \mu^2\mathrm{Var}[N]
\end{aligned}$$

Special Case: Poisson Distributed Frequency. If $N \sim Poi(\lambda)$, then

$$\begin{aligned}
\mathrm{E}(N) = \mathrm{Var}(N) &= \lambda \\
\mathrm{E}(S) &= \lambda\mathrm{E}(X) \\
\mathrm{Var}(S) &= \lambda(\sigma^2 + \mu^2) = \lambda\ \mathrm{E}(X^2).
\end{aligned}$$

Example 5.3.1. Actuarial Exam Question. The number of accidents follows a Poisson distribution with mean 12. Each accident generates 1, 2, or 3 claimants with probabilities 1/2, 1/3, and 1/6 respectively.

Calculate the variance in the total number of claimants.

Show Example Solution

Solution.

$$\begin{aligned}
\mathrm{E}(X^2) &= 1^2\left(\frac{1}{2}\right) + 2^2\left(\frac{1}{3}\right) + 3^2\left(\frac{1}{6}\right) = \frac{10}{3} \\
\Rightarrow \mathrm{Var}(S_N) &= \lambda\ \mathrm{E}(X^2) = 12\left(\frac{10}{3}\right) = 40
\end{aligned}$$

Alternatively, using the general approach, $\mathrm{Var}(S_N) = \sigma^2\mathrm{E}(N) + \mu^2\mathrm{Var}(N)$, where

$$\begin{aligned}
\mathrm{E}(N) &= \mathrm{Var}(N) = 12 \\
\mu &= \mathrm{E}(X) = 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{3}\right) + 3\left(\frac{1}{6}\right) = \frac{5}{3} \\
\sigma^2 &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \frac{10}{3} - \frac{25}{9} = \frac{5}{9} \\
\Rightarrow\ \mathrm{Var}(S) &= \left(\frac{5}{9}\right)(12) + \left(\frac{5}{3}\right)^2 (12) = 40.
\end{aligned}$$

In general, the moments of S_N can be derived from its moment generating function (mgf). Because X_i's are iid, we denote the mgf of X as $M_X(t) = \mathrm{E}\ (e^{tX})$. Using the law of iterated expectations, the mgf of S_N is

$$\begin{aligned}
M_{S_N}(t) &= \mathrm{E}(e^{tS_N}) = \mathrm{E}_N[\ \mathrm{E}(e^{tS_N}|N)\ ] \\
&= \mathrm{E}_N\left[\ \mathrm{E}\left(e^{t(X_1+\cdots+X_N)}\right)\ \right] = \mathrm{E}_N\left[\mathrm{E}(e^{tX_1})\cdots\mathrm{E}(e^{tX_N})\right] \quad \text{since } X_i\text{'s are independent} \\
&= \mathrm{E}_N[\ (M_X(t))^N\ ]
\end{aligned}$$

Now, recall that the probability generating function (pgf) of N is $P_N(z) = \mathrm{E}(z^N)$. Denote $M_X(t) = z$. Substituting into the expression for the mgf of S_N above, it is shown

$$M_{S_N}(t) = \mathrm{E}\ (z^N) = P_N(z) = P_N[M_X(t)].$$

Similarly, if S_N is discrete, one can show the pgf of S_N is:

$$P_{S_N}(z) = P_N[P_X(z)].$$

To get $\mathrm{E}(S_N) = M'_{S_N}(0)$, we use the chain rule

$$M'_{S_N}(t) = \frac{\partial}{\partial t} P_N(M_X(t)) = P'_N(M_X(t))M'_X(t)$$

and recall $M_X(0) = 1, M'_X(0) = \mathrm{E}(X) = \mu, P'_N(1) = \mathrm{E}(N)$. So,

$$\mathrm{E}(S_N) = M'_{S_N}(0) = P'_N(M_X(0))M'_X(0) = \mu\mathrm{E}(N)$$

Similarly, one could use relation $\mathrm{E}(S_N^2) = M''_{S_N}(0)$ to get

$$\mathrm{Var}(S_N) = \sigma^2\mathrm{E}(N) + \mu^2\mathrm{Var}(N).$$

Special Case. Poisson Frequency. Let $N \sim Poi(\lambda)$. Thus, the pgf of N is $P_N(z) = e^{\lambda(z-1)}$ and the mgf of S_N is

$$M_{S_N}(t) = P_N[M_X(t)] = e^{\lambda(M_X(t)-1)}.$$

Taking derivatives yields

$$M'_{S_N}(t) = e^{\lambda(M_X(t)-1)}\ \lambda\ M'_X(t) = M_{S_N}(t)\ \lambda\ M'_X(t)$$
$$M''_{S_N}(t) = M_S(t)\ \lambda\ M''_X(t) + [\ M_S(t)\ \lambda\ M'_X(t)\ ]\ \lambda\ M'_X(t)$$

Evaluating these at $t = 0$ yields

$$\mathrm{E}(S_N) = M'_{S_N}(0) = \lambda \mathrm{E}(X) = \lambda\mu$$

and

$$M''_{S_N}(0) = \lambda \mathrm{E}(X^2) + \lambda^2\mu^2$$
$$\Rightarrow \mathrm{Var}(S_N) = \lambda \mathrm{E}(X^2) + \lambda^2\mu^2 - (\lambda\mu)^2 = \lambda\ \mathrm{E}(X^2).$$

Example 5.3.2. Actuarial Exam Question. You are the producer of a television quiz show that gives cash prizes. The number of prizes, N, and prize amount, X, have the following distributions:

n	$\Pr(N = n)$	x	$\Pr(X = x)$
1	0.8	0	0.2
2	0.2	100	0.7
		1000	0.1

Your budget for prizes equals the expected aggregate cash prizes plus the standard deviation of aggregate cash prizes. Calculate your budget.

Show Example Solution

Solution. We need to calculate the mean and standard deviation of the aggregate (sum) of cash prizes. The moments of the frequency distribution N are

$$\mathrm{E}(N) = 1(0.8) + 2(0.2) = 1.2$$
$$\mathrm{E}(N^2) = 1^2(0.8) + 2^2(0.2) = 1.6$$
$$\mathrm{Var}(N) = \mathrm{E}(N^2) - [\mathrm{E}(N)]^2 = 0.16$$

The moments of the severity distribution X are

$$\mathrm{E}(X) = 0(0.2) + 100(0.7) + 1000(0.1) = 170 = \mu$$
$$\mathrm{E}(X^2) = 0^2(0.2) + 100^2(0.7) + 1000^2(0.1) = 107,000$$
$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = 78,100 = \sigma^2$$

Thus, the mean and variance of the aggregate cash prize are

$$\mathrm{E}(S_N) = \mu\mathrm{E}(N) = 170(1.2) = 204$$
$$\mathrm{Var}(S_N) = \sigma^2\mathrm{E}(N) + \mu^2\mathrm{Var}(N)$$
$$= 78,100(1.2) + 170^2(0.16) = 98,344$$

This gives the following required budget

$$\begin{aligned} Budget &= \mathrm{E}(S_N) + \sqrt{\mathrm{Var}(S_N)} \\ &= 204 + \sqrt{98,344} = 517.60. \end{aligned}$$

The distribution of S_N is called a compound distribution, and it can be derived based on the convolution of F_X as follows:

$$\begin{aligned} F_{S_N}(s) &= \Pr\left(X_1 + \cdots + X_N \leq s\right) \\ &= \mathrm{E}\left[\Pr\left(X_1 + \cdots + X_N \leq s | N = n\right)\right] \\ &= \mathrm{E}\left[F_X^{*N}(s)\right] \\ &= p_0 + \sum_{n=1}^{\infty} p_n F_X^{*n}(s) \end{aligned}$$

Example 5.3.3. Actuarial Exam Question. The number of claims in a period has a geometric distribution with mean 4. The amount of each claim X follows $\Pr(X = x) = 0.25,\ x = 1, 2, 3, 4$, i.e. a discrete uniform distribution on $\{1, 2, 3, 4\}$. The number of claims and the claim amounts are independent. Let S_N denote the aggregate claim amount in the period. Calculate $F_{S_N}(3)$.

Show Example Solution

Solution. By definition, we have

$$\begin{aligned} F_{S_N}(3) &= \Pr\left(\sum_{i=1}^{N} X_i \leq 3\right) = \sum_{n=0}^{\infty} \Pr\left(\sum_{i=1}^{n} X_i \leq 3 | N = n\right) \Pr(N = n) \\ &= \sum_{n} F^{*n}(3)\, p_n = \sum_{n=0}^{3} F^{*n}(3) p_n \\ &= p_0 + F^{*1}(3)\ p_1 + F^{*2}(3)\ p_2 + F^{*3}(3)\ p_3 \end{aligned}$$

Because $N \sim Geo(\beta = 4)$, we know that

$$p_n = \frac{1}{1+\beta}\left(\frac{\beta}{1+\beta}\right)^n = \frac{1}{5}\left(\frac{4}{5}\right)^n$$

For the claim severity distribution, recursively, we have

$$\begin{aligned}
F^{*1}(3) &= \Pr(X \leq 3) = \frac{3}{4} \\
F^{*2}(3) &= \sum_{y \leq 3} F^{*1}(3-y) f(y) = F^{*1}(2) f(1) + F^{*1}(1) f(2) \\
&= \frac{1}{4}\left[F^{*1}(2) + F^{*1}(1)\right] = \frac{1}{4}\left[\Pr(X \leq 2) + \Pr(X \leq 1)\right] \\
&= \frac{1}{4}\left(\frac{2}{4} + \frac{1}{4}\right) = \frac{3}{16} \\
F^{*3}(3) &= \Pr(X_1 + X_2 + X_3 \leq 3) = \Pr(X_1 = X_2 = X_3 = 1) = \left(\frac{1}{4}\right)^3
\end{aligned}$$

Notice that we did not need to recursively calculate $F^{*3}(3)$ by recognizing that each $X \in \{1, 2, 3, 4\}$, so the only way of obtaining $X_1 + X_2 + X_3 \leq 3$ is to have $X_1 = X_2 = X_3 = 1$. Additionally, for $n \geq 4$, $F^{*n}(3) = 0$ since it is impossible for the sum of 4 or more X's to be less than 3. For $n = 0$, $F^{*0}(3) = 1$ since the sum of 0 X's is 0, which is always less than 3. Laying out the probabilities systematically,

x	$F^{*1}(x)$	$F^{*2}(x)$	$F^{*3}(x)$
0			
1	$\frac{1}{4}$	0	
2	$\frac{2}{4}$	$\left(\frac{1}{4}\right)^2$	
3	$\frac{3}{4}$	$\frac{3}{16}$	$\left(\frac{1}{4}\right)^3$

Finally,

$$\begin{aligned}
F_{S_N}(3) &= p_0 + F^{*1}(3)\, p_1 + F^{*2}(3)\, p_2 + F^{*3}(3)\, p_3 \\
&= \frac{1}{5} + \frac{3}{4}\left(\frac{4}{25}\right) + \frac{3}{16}\left(\frac{16}{125}\right) + \frac{1}{64}\left(\frac{64}{625}\right) = 0.3456
\end{aligned}$$

When $\mathrm{E}(N)$ and $\mathrm{Var}(N)$ are known, one may also use the central limit theorem to approximate the distribution of S_N as in the individual risk model. That is, $\frac{S_N - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}}$ approximately follows the standard normal distribution $N(0, 1)$.

Example 5.3.4. Actuarial Exam Question.. You are given:

	Mean	Standard Deviation
Number of Claims	8	3
Individual Losses	10,000	3,937

Using the normal approximation, determine the probability that the aggregate loss will exceed 150% of the expected loss.

Show Example Solution

Solution. To use the normal approximation, we must first find the mean and variance of the aggregate loss S

$$\begin{aligned}
\mathrm{E}(S_N) &= \mu\ \mathrm{E}(N) = 10,000(8) = 80,000 \\
\mathrm{Var}(S_N) &= \sigma^2\ \mathrm{E}(N) + \mu^2\ \mathrm{Var}(N) \\
&= 3937^2(8) + 10000^2(3^2) = 1,023,999,752 \\
\sqrt{\mathrm{Var}(S_N)} &= 31,999.996 \approx 32,000
\end{aligned}$$

Then under the normal approximation, aggregate loss S_N is approximately normal with mean 80,000 and standard deviation 32,000. The probability that S_N will exceed 150% of the expected aggregate loss is therefore

$$\begin{aligned}
\Pr(S_N > 1.5\mathrm{E}(S_N)\ ) &= \Pr\left(\frac{S_N - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}} > \frac{1.5\ \mathrm{E}(S_N) - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}}\right) \\
&\approx \Pr\left(Z > \frac{0.5\ \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}}\right), \quad \text{where } Z \sim N(0,1) \\
&= \Pr\left(Z > \frac{0.5(80,000)}{32,000}\right) = \Pr(Z > 1.25) \\
&= 1 - \Phi(1.25) = 0.1056
\end{aligned}$$

Example 5.3.5. Actuarial Exam Question. For an individual over 65:
(i) The number of pharmacy claims is a Poisson random variable with mean 25.
(ii) The amount of each pharmacy claim is uniformly distributed between 5 and 95.
(iii) The amounts of the claims and the number of claims are mutually independent.
Estimate the probability that aggregate claims for this individual will exceed 2000 using the normal approximation.

Show Example Solution

Solution. We have claim frequency $N \sim Poi(\lambda = 25)$ and claim severity $X \sim U(5, 95)$. To use the normal approximation, we need to find the mean and variance of the aggregate claims S_N. Note

$$\begin{aligned}
&\mathrm{E}(N) = 25 && \mathrm{Var}(N) = 25 \\
&\mathrm{E}(X) = \tfrac{5+95}{2} = 50 = \mu && \mathrm{Var}(X) = \tfrac{(95-5)^2}{12} = 675 = \sigma^2
\end{aligned}$$

Then for S_N,

$$\begin{aligned}
\mathrm{E}(S_N) &= \mu\ \mathrm{E}(N) = 50(25) = 1,250 \\
\mathrm{Var}(S_N) &= \sigma^2\ \mathrm{E}(N) + \mu^2\ \mathrm{Var}(N) \\
&= 675(25) + 50^2(25) = 79,375
\end{aligned}$$

Using the normal approximation, S_N is approximately normal with mean 1,250 and variance 79,375. The probability that S_N exceeds 2,000 is

$$\begin{aligned}\Pr(S_N > 2,000) &= \Pr\left(\frac{S_N - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}} > \frac{2,000 - \mathrm{E}(S_N)}{\sqrt{\mathrm{Var}(S_N)}}\right) \\ &= \Pr\left(Z > \frac{2,000 - 1,250}{\sqrt{79,375}}\right), \quad \text{where } Z \sim N(0,1) \\ &= \Pr(Z > 2.662) = 1 - \Phi(2.662) = 0.003884\end{aligned}$$

5.3.2 Stop-loss Insurance

Recall the coverage modifications on the individual policy level in Section **??**. Insurance on the aggregate loss S_N, subjected to a deductible d, is called *net stop-loss insurance*. The expected value of the amount of the aggregate loss in excess of the deductible,

$$\mathrm{E}[(S-d)_+]$$

is known as the net stop-loss premium.

To calculate the net stop-loss premium, we have

$$\begin{aligned}\mathrm{E}(S_N - d)_+ &= \begin{cases} \int_d^\infty (s-d) f_{S_N}(s) ds & \text{for continuous } S_N \\ \sum_{s>d} (s-d) f_{S_N}(s) & \text{for discrete } S_N \end{cases} \\ &= \mathrm{E}(S_N) - \mathrm{E}(S_N \wedge d)\end{aligned}$$

Example 5.3.6. Actuarial Exam Question. In a given week, the number of projects that require you to work overtime has a geometric distribution with $\beta = 2$. For each project, the distribution of the number of overtime hours in the week, X, is as follows:

x	$f(x)$
5	0.2
10	0.3
20	0.5

The number of projects and the number of overtime hours are independent. You will get paid for overtime hours in excess of 15 hours in the week. Calculate the expected number of overtime hours for which you will get paid in the week.

Show Example Solution

Solution. The number of projects in a week requiring overtime work has distribution $N \sim Geo(\beta = 2)$, while the number of overtime hours worked per project has distribution X as described above. The aggregate number of overtime hours in a week is S_N and we are therefore looking for

$$\mathrm{E}(S_N - 15)_+ = \mathrm{E}(S_N) - \mathrm{E}(S_N \wedge 15).$$

To find $\mathrm{E}(S_N) = \mathrm{E}(X)\ \mathrm{E}(N)$, we have

$$\begin{aligned} \mathrm{E}(X) &= 5(0.2) + 10(0.3) + 20(0.5) = 14 \\ \mathrm{E}(N) &= 2 \\ \Rightarrow \mathrm{E}(S) &= \mathrm{E}(X)\ \mathrm{E}(N) = 14(2) = 28 \end{aligned}$$

To find $\mathrm{E}(S_N \wedge 15) = 0\Pr(S_N = 0) + 5\Pr(S_N = 5) + 10\Pr(S_N = 10) + 15\Pr(S_N \geq 15)$, we have

$$\begin{aligned} \Pr(S_N = 0) &= \Pr(N = 0) = \frac{1}{1+\beta} = \frac{1}{3} \\ \Pr(S_N = 5) &= \Pr(X = 5,\ N = 1) = 0.2\left(\frac{2}{9}\right) = \frac{0.4}{9} \\ \Pr(S_N = 10) &= \Pr(X = 10,\ N = 1) + \Pr(X_1 = X_2 = 5, N = 2) \\ &= 0.3\left(\frac{2}{9}\right) + (0.2)(0.2)\left(\frac{4}{27}\right) = 0.0726 \\ \Pr(S_N \geq 15) &= 1 - \left(\frac{1}{3} + \frac{0.4}{9} + 0.0726\right) = 0.5496 \\ \Rightarrow \mathrm{E}(S_N \wedge 15) &= 0\Pr(S_N = 0) + 5\Pr(S_N = 5) + 10\Pr(S_N = 10) + 15\Pr(S_N \geq 15) \\ &= 0\left(\frac{1}{3}\right) + 5\left(\frac{0.4}{9}\right) + 10(0.0726) + 15(0.5496) = 9.193 \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{E}(S_N - 15)_+ &= \mathrm{E}(S_N) - \mathrm{E}(S_N \wedge 15) \\ &= 28 - 9.193 = 18.807 \end{aligned}$$

Recursive Net Stop-Loss Premium Calculation. For the discrete case, this can be computed recursively as

$$\mathrm{E}\left[(S_N - (j+1)h)_+\right] = \mathrm{E}\left[(S_N - jh)_+\right] - h\left(1 - F_{S_N}(jh)\right).$$

This assumes that the support of S_N is equally spaced over units of h.

To establish this, we assume that $h = 1$. We have

$$\begin{aligned} \mathrm{E}\left[(S_N - (j+1))_+\right] &= \mathrm{E}(S_N) - \mathrm{E}[S_N \wedge (j+1)]\ , \quad \text{and} \\ \mathrm{E}\left[(S_N - j)_+\right] &= \mathrm{E}(S_N) - \mathrm{E}[S_N \wedge j] \end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{E}\left[(S_N-(j+1))_+\right]-\mathrm{E}\left[(S_N-j)_+\right] &= \{\mathrm{E}(S_N)-\mathrm{E}(S_N\wedge(j+1))\}-\{\mathrm{E}(S_N)-\mathrm{E}(S_N\wedge j)\}\\ &= \mathrm{E}\left(S_N\wedge j\right)-\mathrm{E}\left[S\wedge(j+1)\right]\end{aligned}$$

We can write

$$\begin{aligned}\mathrm{E}\left[S_N\wedge(j+1)\right] &= \sum_{x=0}^{j} x f_{S_N}(x)+(j+1)\ \Pr(S_N\geq j+1)\\ &= \sum_{x=0}^{j-1} x f_{S_N}(x)+j\ \Pr(S_N=j)+(j+1)\ \Pr(S_N\geq j+1)\end{aligned}$$

Similarly,

$$\mathrm{E}(S_N\wedge j)=\sum_{x=0}^{j-1} x f_{S_N}(x)+j\ \Pr(S_N\geq j)$$

With these, expressions, we have

$$\begin{aligned}\mathrm{E}\left[(S_N-(j+1))_+\right]-\mathrm{E}\ \left[(S_N-j)_+\right] &= \mathrm{E}\left(S_N\wedge j\right)-\mathrm{E}\left[S\wedge(j+1)\right]\\ &= \left\{\sum_{x=0}^{j-1} x f_{S_N}(x)+j\ \Pr(S_N\geq j)\right\}-\left\{\sum_{x=0}^{j-1} x f_{S_N}(x)+j\ \Pr(S_N=\right.\\ &= j\ [\Pr(S_N\geq j)-\Pr(S_N=j)]-(j+1)\ \Pr(S_N\geq j+1)\\ &= j\ \Pr(S_N>j)-(j+1)\ \Pr(S_N\geq j+1)\qquad(\text{note }\Pr(S_N>j)=\\ &= -\Pr(S_N\geq j+1)=-\left[1-F_{S_N}(j)\right],\end{aligned}$$

as required.

Example 5.3.7. Actuarial Exam Question - Continued. Recall that the goal of this question was to calculate $\mathrm{E}\ (S_N-15)_+$. Note that the support of S_N is equally spaced over units of 5, so this question can also be done recursively, using the expression above with steps of $h=5$:

- Step 1:

$$\begin{aligned}\mathrm{E}\ (S_N-5)_+ &= \mathrm{E}(S_N)-5[1-\Pr(S_N\leq 0)]\\ &= 28-5\left(1-\frac{1}{3}\right)=\frac{74}{3}=24.6667\end{aligned}$$

- Step 2:

$$\begin{aligned}\mathrm{E}\ (S_N-10)_+ &= \mathrm{E}\ (S_N-5)_+-5[1-\Pr(S_N\leq 5)]\\ &= \frac{74}{3}-5\left(1-\frac{1}{3}-\frac{0.4}{9}\right)=21.555\end{aligned}$$

- Step 3:

$$\begin{aligned}\text{E}\ (S_N - 15)_+ &= \text{E}\ (S_N - 10)_+ - 5[1 - \Pr(S_N \leq 10)] \\ &= \text{E}\ (S_N - 10)_+ - 5\Pr(S_N \geq 15) \\ &= 21.555 - 5(0.5496) = 18.807\end{aligned}$$

5.3.3 Analytic Results

There are a few combinations of claim frequency and severity distributions that result in an easy-to-compute distribution for aggregate losses. This section provides some simple examples. Although these examples are computationally convenient, they are generally too simple to be used in practice.

Example 5.3.8. One has a closed-form expression for the aggregate loss distribution by assuming a geometric frequency distribution and an exponential severity distribution.

Assume that claim count N is geometric with mean $\text{E}(N) = \beta$, and that claim amount X is exponential with $\text{E}(X) = \theta$. Recall that the pgf of N and the mgf of X are:

$$P_N(z) = \frac{1}{1 - \beta(z-1)}$$
$$M_X(t) = \frac{1}{1 - \theta t}$$

Thus, the mgf of aggregate loss S_N can be expressed two ways (for details, see Technical Supplement 5.A.3)

$$\begin{aligned}M_{S_N}(t) &= P_N[M_X(t)] = \frac{1}{1 - \beta\left(\frac{1}{1-\theta t} - 1\right)} \\ &= 1 + \frac{\beta}{1+\beta}\left([1 - \theta(1+\beta)t]^{-1} - 1\right) \qquad (5.1) \\ &= \frac{1}{1+\beta}(1) + \frac{\beta}{1+\beta}\left(\frac{1}{1 - \theta(1+\beta)t}\right) \qquad (5.2)\end{aligned}$$

From (5.1), we note that S_N is equivalent to the compound distribution of $S_N = X_1^* + \cdots + X_{N^*}^*$, where N^* is a Bernoulli with mean $\beta/(1+\beta)$ and X^* is an exponential with mean $\theta(1+\beta)$. To see this, we examine the mgf of S:

$$M_{S_N}(t) = P_N[M_X(t)] = P_{N^*}[M_{X^*}(t)],$$

where

$$P_{N^*}(z) = 1 + \frac{\beta}{1+\beta}(z-1),$$
$$M_{X^*}(t) = \frac{1}{1-\theta(1+\beta)t}.$$

From (5.2), we note that S_N is also equivalent to a 2-point mixture of 0 and X^*. Specifically,

$$S_N = \begin{cases} 0 & \text{with probability } \Pr(N^*=0) = 1/(1+\beta) \\ Y^* & \text{with probability } \Pr(N^*=1) = \beta/(1+\beta) \end{cases}.$$

The distribution function of S_N is:

$$\begin{aligned} \Pr(S_N = 0) &= \frac{1}{1+\beta} \\ \Pr(S_N > s) &= \Pr(X^* > s) = \frac{\beta}{1+\beta}\exp\left(-\frac{s}{\theta(1+\beta)}\right) \end{aligned}$$

with pdf

$$f_{S_N}(s) = \frac{\beta}{\theta(1+\beta)^2}\exp\left(-\frac{s}{\theta(1+\beta)}\right).$$

Example 5.3.9. Consider a collective risk model with an exponential severity and an arbitrary frequency distribution. Recall that if $X_i \sim Exp(\theta)$, then the sum of iid exponential, $S_n = X_1 + \cdots + X_n$, has a gamma distribution, i.e. $S_n \sim Gam(n,\theta)$. This has cdf:

$$\begin{aligned} F_X^{*n}(s) &= \Pr(S_n \le s) = \int_0^s \frac{1}{\Gamma(n)\theta^n} s^{n-1}\exp\left(-\frac{s}{\theta}\right) ds \\ &= 1 - \sum_{j=0}^{n-1}\frac{1}{j!}\left(\frac{s}{\theta}\right)^j e^{-s/\theta}. \end{aligned}$$

The last equality is derived by applying integration by parts $n-1$ times.

For the aggregate loss distribution, we can interchange the order of summations in the second line below to get

$$\begin{aligned} F_S(s) &= p_0 + \sum_{n=1}^{\infty} p_n F_X^{*n}(s) \\ &= 1 - \sum_{n=1}^{\infty} p_n \sum_{j=0}^{n-1} \frac{1}{j!} \left(\frac{s}{\theta}\right)^j e^{-s/\theta} \\ &= 1 - e^{-s/\theta} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{s}{\theta}\right)^j \overline{P}_j \end{aligned}$$

where $\overline{P}_j = p_{j+1} + p_{j+2} + \cdots = \Pr(N > j)$ is the "survival function" of the claims count distribution.

5.3.4 Tweedie Distribution

In this section, we examine a particular compound distribution where the number of claims has a Poisson distribution and the amount of claims has a gamma distribution. This specification leads to what is known as a Tweedie distribution. The Tweedie distribution has a mass probability at zero and a continuous component for positive values. Because of this feature, it is widely used in insurance claims modeling, where the zero mass is interpreted as no claims and the positive component as the amount of claims.

Specifically, consider the collective risk model $S_N = X_1 + \cdots + X_N$. Suppose that N has a Poisson distribution with mean λ, and each X_i has a gamma distribution shape parameter α and scale parameter γ. The Tweedie distribution is derived as the Poisson sum of gamma variables. To understand the distribution of S_N, we first examine the mass probability at zero. The aggregate loss is zero when no claims occurred, i.e.

$$\Pr(S_N = 0) = \Pr(N = 0) = e^{-\lambda}.$$

In addition, note that S_N conditional on $N = n$, denoted by $S_n = X_1 + \cdots + X_n$, follows a gamma distribution with shape $n\alpha$ and scale γ. Thus, for $s > 0$, the density of a Tweedie distribution can be calculated as

$$\begin{aligned} f_{S_n}(s) &= \sum_{n=1}^{\infty} p_n f_{S_n}(s) \\ &= \sum_{n=1}^{\infty} e^{-\lambda} \frac{(\lambda)^n}{n!} \frac{\gamma^{na}}{\Gamma(n\alpha)} s^{n\alpha-1} e^{-s\gamma} \end{aligned}$$

Thus, the Tweedie distribution can be thought of a mixture of zero and a positive valued distribution, which makes it a convenient tool for modeling insurance claims and for calculating pure premiums. The mean and variance of the

Tweedie compound Poisson model are:

$$\mathrm{E}(S_N) = \lambda\frac{\alpha}{\gamma} \quad \text{and} \quad \mathrm{Var}(S) = \lambda\frac{\alpha(1+\alpha)}{\gamma^2}.$$

As another important feature, the Tweedie distribution is a special case of exponential dispersion models, a class of models used to describe the random component in generalized linear models. To see this, we consider the following reparameterization:

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \frac{1}{\gamma} = \phi(p-1)\mu^{p-1}$$

With the above relationships, one can show that the distribution of S_N is

$$f_{S_N}(s) = \exp\left[\frac{1}{\phi}\left(\frac{-s}{(p-1)\mu^{p-1}} - \frac{\mu^{2-p}}{2-p}\right) + C(s;\phi)\right]$$

where

$$C(s;\phi/\omega_i) = \begin{cases} 0 & \text{if } y = 0 \\ \ln\sum_{n\geq 1}\left\{\frac{(1/\phi)^{1/(p-1)}y^{(2-p)/(p-1)}}{(2-p)(p-1)^{(2-p)/(p-1)}}\right\}^n \frac{1}{n!\Gamma(n(2-p)/(p-1))s} & \text{if } y > 0 \end{cases}$$

Hence, the distribution of S_N belongs to the exponential family with parameters μ, ϕ, and $1 < p < 2$, and we have

$$\mathrm{E}(S_N) = \mu \quad \text{and} \quad \mathrm{Var}(S_N) = \phi\mu^p.$$

This allows us to use the Tweedie distribution with generalized linear models to model claims. It is also worth mentioning the two limiting cases of the Tweedie model: $p \to 1$ results in the Poisson distribution and $p \to 2$ results in the gamma distribution. Thus, the Tweedie model accommodates the situations in between the gamma and Poisson distributions, which makes intuitive sense as it is the Poisson sum of gamma random variables.

Show Quiz Solution

5.4 Computing the Aggregate Claims Distribution

Computing the distribution of aggregate losses is a difficult, yet important, problem. As we have seen, for both individual risk model and collective risk model,

computing the distribution frequently involves the evaluation of a n-fold convolution. To make the problem tractable, one strategy is to use a distribution that is easy to evaluate to approximate the aggregate loss distribution. For instance, normal distribution is a natural choice based on central limit theorem where parameters of the normal distribution can be estimated by matching the moments. This approach has its strength and limitations. The main advantage is the ease of computation. The disadvantage are: first, the size and direction of approximation error are unknown; second, the approximation may fail to capture some special features of the aggregate loss such as mass point at zero.

This section discusses two practical approaches to computing the distribution of aggregate loss, the recursive method and the simulation.

5.4.1 Recursive Method

The recursive method applies to compound models where the frequency component N belongs to either $(a, b, 0)$ or $(a, b, 1)$ class (see Sections **??** and **??**) and the severity component X has a discrete distribution. For continuous X, a common practice is to first discretize the severity distribution, after which the recursive method is ready to apply.

Assume that N is in the $(a, b, 1)$ class so that $p_k = \left(a + \frac{b}{k}\right) p_{k-1}, k = 2, 3, \ldots$. Further assume that the support of X is $\{0, 1, \ldots, m\}$, discrete and finite. Then, the probability function of S_N is:

$$\begin{aligned} f_{S_N}(s) &= \Pr(S = s) \\ &= \frac{1}{1 - a f_X(0)} \left\{ [p_1 - (a+b)p_0] f_X(s) + \sum_{x=1}^{s \wedge m} \left(a + \frac{bx}{s}\right) f_X(x) f_{S_N}(s-x) \right\}. \end{aligned}$$

If N is in the $(a, b, 0)$ class, then $p_1 = (a + b)p_0$ and so

$$f_{S_N}(s) = \frac{1}{1 - a f_X(0)} \left\{ \sum_{x=1}^{s \wedge m} \left(a + \frac{bx}{s}\right) f_X(x) f_{S_N}(s-x) \right\}.$$

Special Case: Poisson Frequency. If $N \sim Poi(\lambda)$, then $a = 0$ and $b = \lambda$, and thus

$$f_{S_N}(s) = \frac{\lambda}{s} \left\{ \sum_{x=1}^{s \wedge m} x f_X(x) f_{S_N}(s-x) \right\}.$$

Example 5.4.1. Actuarial Exam Question. The number of claims in a period N has a geometric distribution with mean 4. The amount of each claim X follows $\Pr(X = x) = 0.25$, for $x = 1, 2, 3, 4$. The number of claims and the claim amount are independent. S_N is the aggregate claim amount in the period. Calculate $F_{S_N}(3)$.

Show Example Solution

Solution. The severity distribution X follows

$$f_X(x) = \frac{1}{4}, \quad x = 1, 2, 3, 4.$$

The frequency distribution N is geometric with mean 4, which is a member of the $(a, b, 0)$ class with $b = 0$, $a = \frac{\beta}{1+\beta} = \frac{4}{5}$, and $p_0 = \frac{1}{1+\beta} = \frac{1}{5}$. The support of severity component X is $\{1, \ldots, m = 4\}$, discrete and finite. Thus, we can use the recursive method

$$\begin{aligned} f_{S_N}(x) &= 1 \sum_{y=1}^{x \wedge m} (a + 0) f_X(y) f_{S_N}(x - y) \\ &= \frac{4}{5} \sum_{y=1}^{x \wedge m} f_X(y) f_{S_N}(x - y) \end{aligned}$$

Specifically, we have

$$\begin{aligned} f_{S_N}(0) &= \Pr(N = 0) = p_0 = \frac{1}{5} \\ f_{S_N}(1) &= \frac{4}{5} \sum_{y=1}^{1} f_X(y) f_{S_N}(1 - y) = \frac{4}{5} f_X(1) f_{S_N}(0) \\ &= \frac{4}{5} \left(\frac{1}{4}\right) \left(\frac{1}{5}\right) = \frac{1}{25} \\ f_{S_N}(2) &= \frac{4}{5} \sum_{y=1}^{2} f_X(y) f_{S_N}(2 - y) = \frac{4}{5} \left[f_X(1) f_{S_N}(1) + f_X(2) f_{S_N}(0)\right] \\ &= \frac{4}{5} \left[\frac{1}{4} \left(\frac{1}{25} + \frac{1}{5}\right)\right] = \frac{4}{5} \left(\frac{6}{100}\right) = \frac{6}{125} \\ f_{S_N}(3) &= \frac{4}{5} \left[f_X(1) f_{S_N}(2) + f_X(2) f_{S_N}(1) + f_X(3) f_{S_N}(0)\right] \\ &= \frac{4}{5} \left[\frac{1}{4} \left(\frac{1}{25} + \frac{1}{5} + \frac{6}{125}\right)\right] = \frac{1}{5} \left(\frac{5 + 25 + 6}{125}\right) = 0.0576 \\ \Rightarrow \; F_{S_N}(3) &= f_{S_N}(0) + f_{S_N}(1) + f_{S_N}(2) + f_{S_N}(3) = 0.3456 \end{aligned}$$

5.4.2 Simulation

The distribution of aggregate loss can be evaluated using Monte Carlo simulation. The idea is that one can calculate the empirical distribution of S_N using a random sample. The expected value and variance of the aggregate loss can also be estimated using the sample mean and sample variance of the simulated values. Below we summarize the simulation procedures for the aggregate loss models. Let m be the size of the generated random sample of aggregate losses.

1. Individual Risk Model $S_n = X_1 + \cdots + X_n$
   - Let $j = 1, \ldots, m$ be a counter. Start by setting $j = 1$.
   - Generate each individual loss realization x_{ij} for $i = 1, \ldots, n$. For example, this can be done using the inverse transformation method (Section 6.2).
   - Calculate the aggregate loss $s_j = x_{1j} + \cdots + x_{nj}$.
   - Repeat the above two steps for $j = 2, \ldots, m$ to obtain a size-m sample of S_n, i.e. $\{s_1, \ldots, s_m\}$.
2. Collective Risk Model $S_N = X_1 + \cdots + X_N$
   - Let $j = 1, \ldots, m$ be a counter. Start by setting $j = 1$.
   - Generate the number of claims n_j from the frequency distribution N.
   - Given n_j, generate the amount of each claim independently from severity distribution X, denoted by $x_{1j}, \ldots, x_{n_j j}$.
   - Calculate the aggregate loss $s_j = x_{1j} + \cdots + x_{n_j j}$.
   - Repeat the above three steps for $j = 2, \ldots, m$ to obtain a size-m sample of S_N, i.e. $\{s_1, \ldots, s_m\}$.

Given the random sample of S, the empirical distribution can be calculated as

$$\hat{F}_S(s) = \frac{1}{m} \sum_{i=1}^{m} I(s_i \leq s),$$

where $I(\cdot)$ is an indicator function. The empirical distribution $\hat{F}_S(s)$ will converge to $F_S(s)$ almost surely as the sample size $m \to \infty$.

The above procedure assumes that the probability distributions, including the parameter values, of the frequency and severity distributions are known. In practice, one would need to first assume these distributions, estimate their parameters from the data, and then assess the quality of model fit using various model validation tools (see Chapter **??**). For instance, the assumptions in the collective risk model suggest a two-stage estimation where one model is developed for the number of claims N from the data on claim counts, and another model is developed for the severity of claims X from the data on the amount of claims.

Example 5.4.2. Recall Example 5.3.5 with an individual's claim frequency $N \sim Poi(\lambda = 25)$ and claim severity $X \sim U(5, 95)$. Using a simulated sample of 10000 observations, estimate the mean and variance of the aggregate loss S_N. In addition, use the simulated sample to estimate the probability that aggregate claims for this individual will exceed 2000 and compare with the normal approximation estimates from Example 5.3.5.

Show Example Solution

Solution. We follow the algorithm for the collective risk model, where we first

simulate frequencies $n_1, \ldots, n_{10000}$, and conditional on n_j, $j = 1, \ldots, 10000$, simulate each individual loss x_{ij}, $i = 1, \ldots n_j$.

```
set.seed(4321)  # For reproducibility of results
m <- 10000      # Number of observations to simulate
lambda <- 25    # Parameter for frequency distribution N
a <- 5; b <- 95 # Parameters for severity distribution X
S <- rep(NA, m) # Initalize an empty vector to store S observations

n <- rpois(m, lambda) # Generate m=10000 observations of N from Poisson
for(j in 1:m){
  n_j <- n[j] # Given each n_j (j=1,...,m), generate n_j observations of X from unifor
  x_j <- runif(n_j, min=a, max=b)
  s_j <- sum(x_j) # Calculate the aggregate loss s_j
  S[j] <- s_j # Store s_j in the vector of observations
}
mean(S) # Compare to theoretical value of 1,250
```

```
## [1] 1248.09
```

```
var(S)  # Compare to theoretical value of 79,375
```

```
## [1] 77441.22
```

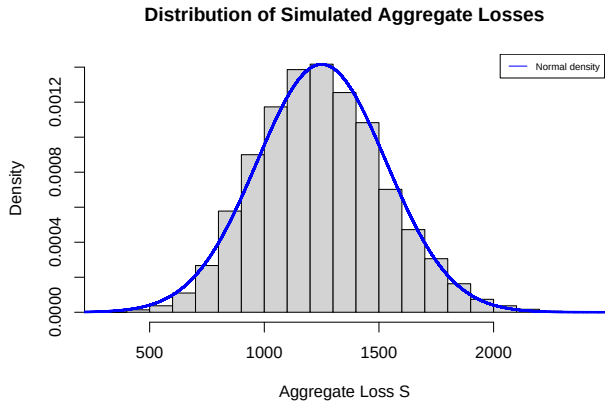
```
mean(S>2000) # Proportion of simulated observations s_j that are > 2000
```

```
## [1] 0.0062
```

```
# Compare to normal approximation method of 0.003884
```

Using simulation, we estimate the mean and variance of the aggregate claims to be approximately 1248 and 77441 respectively, compared to the theoretical values of 1,250 and 79,375. In addition, we estimate the probability that aggregate losses exceed 2000 to be 0.0062, compared to the normal approximation estimate of 0.003884.

We can assess the appropriateness of the normal approximation by comparing the empirical distribution of the simulated aggregate losses to the density of the normal distribution used for the normal approximation, $N(\mu = 1,250\ ,\ \sigma^2 = 79,375)$:



The simulated losses are slightly more right-skewed than the normal distribution, with a longer right tail. This explains why the normal approximation estimate of $\Pr(S_N > 2000)$ is lower than the simulated estimate.

Show Quiz Solution

5.5 Effects of Coverage Modifications

5.5.1 Impact of Exposure on Frequency

This section focuses on an individual risk model for claim counts. Recall the individual risk model involves a fixed n number of contracts and independent loss random variables X_i. Consider the number of claims from a group of n policies:

$$S = X_1 + \cdots + X_n$$

where we assume X_i are iid representing the number of claims from policy i. In this case, the exposure for the portfolio is n, using policy as exposure base. The pgf of S is

$$\begin{aligned} P_S(z) &= \mathrm{E}(z^S) = \mathrm{E}\left(z^{\sum_{i=1}^n X_i}\right) \\ &= \prod_{i=1}^n \mathrm{E}(z^{X_i}) = [P_X(z)]^n \end{aligned}$$

Special Case: Poisson. If $X_i \sim Poi(\lambda)$, its pgf is $P_X(z) = e^{\lambda(z-1)}$. Then the pgf of S is

$$P_S(z) = [e^{\lambda(z-1)}]^n = e^{n\lambda(z-1)}.$$

So $S \sim Poi(n\lambda)$. That is, the sum of n independent Poisson random variables each with mean λ has a Poisson distribution with mean $n\lambda$.

Special Case: Negative Binomial. If $X_i \sim NB(\beta, r)$, its pgf is $P_X(z) = [1-\beta(z-1)]^{-r}$. Then the pgf of S is

$$P_S(z) = [[1-\beta(z-1)]^{-r}]^n = [1-\beta(z-1)]^{-nr}.$$

So $S \sim NB(\beta, nr)$.

Example 5.5.1. Assume that the number of claims for each vehicle is Poisson with mean λ. Given the following data on the observed number of claims for each household, calculate the MLE of λ.

Household ID	Number of vehicles	Number of claims
1	2	0
2	1	2
3	3	2
4	1	0
5	1	1

Show Example Solution

Solution. Each of the 5 households has number of exposures n_j (number of vehicles) and number of claims S_j, $j = 1, ..., 5$. Note for each household, the number of claims $S_j \sim Poi(n_j\lambda)$. The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{j=1}^{5} \Pr(S_j = s_j) = \prod_{j=1}^{5} \frac{e^{-n_j\lambda}(n_j\lambda)^{s_j}}{s_j!} \\ &= \left(\frac{e^{-2\lambda}(2\lambda)^0}{0!}\right)\left(\frac{e^{-1\lambda}(1\lambda)^2}{2!}\right)\left(\frac{e^{-3\lambda}(3\lambda)^2}{2!}\right)\left(\frac{e^{-1\lambda}(1\lambda)^0}{0!}\right)\left(\frac{e^{-1\lambda}(1\lambda)^1}{1!}\right) \\ &\propto e^{-8\lambda}\lambda^5 \end{aligned}$$

Taking the log-likehood, we have

$$l(\lambda) = \log L(\lambda) = -8\lambda + 5\log(\lambda)$$

Setting the first derivative of the log-likehood to 0, we get $\hat{\lambda} = \frac{5}{8}$

If the exposure of the portfolio changes from n_1 to n_2, we can establish the following relation between the aggregate claim counts:

$$P_{S_{n_2}}(z) = [P_X(z)]^{n_2} = [P_X(z)^{n_1}]^{n_2/n_1} = P_{S_{n_1}}(z)^{n_2/n_1}.$$

5.5.2 Impact of Deductibles on Claim Frequency

This section examines the effect of deductibles on claim frequency. Intuitively, there will be fewer claims filed when a policy deductible is imposed because a loss below the deductible level may not result in a claim. Even if an insured does file a claim, this may not result in a payment by the policy, since the claim may be denied or the loss amount may ultimately be determined to be below deductible. Let N^L denote the number of losses (i.e. the number of claims with no deductible), and N^P denote the number of payments when a deductible d is imposed. Our goal is to identify the distribution of N^P given the distribution of N^L. We show below that the relationship between N^L and N^P can be established within an aggregate risk model framework.

Note that sometimes changes in deductibles will affect policyholder claim behavior. We assume that this is not the case, i.e. the underlying distributions of losses for both frequency and severity remain unchanged when the deductible changes.

Given there are N^L losses, let $X_1, X_2 \ldots, X_{N^L}$ be the associated amount of losses. For $j = 1, \ldots, N^L$, define

$$I_j = \begin{cases} 1 & \text{if } X_j > d \\ 0 & \text{otherwise} \end{cases}.$$

Then we establish

$$N^P = I_1 + I_2 + \cdots + I_{N_L},$$

that is, the total number of payments is equal to the number of losses above the deductible level. Given that I_j's are independent Bernoulli random variables with probability of success $v = \Pr(X > d)$, the sum of a fixed number of such variables is then a binomial random variable. Thus, conditioning on N^L, N^P has a binomial distribution, i.e. $N^P|N^L \sim Bin(N^L, v)$, where $v = \Pr(X > d)$. This implies that

$$\mathrm{E}\left(z^{N^P}|N^L\right) = [1 + v(z-1)]^{N^L}$$

So the pgf of N^P is

$$\begin{aligned} P_{N^P}(z) &= \mathrm{E}_{N^P}\left(z^{N^P}\right) = \mathrm{E}_{N^L}\left[\mathrm{E}_{N^P}\left(z^{N^P}|N^L\right)\right] \\ &= \mathrm{E}_{N^L}\left[(1 + v(z-1))^{N^L}\right] \\ &= P_{N^L}\left(1 + v(z-1)\right) \end{aligned}$$

Thus, we can write the pgf of N^P as the pgf of N^L, evaluated at a new argument $z^* = 1 + v(z-1)$. That is, $P_{N^P}(z) = P_{N^L}(z^*)$.

Special Cases:

- $N^L \sim Poi(\lambda)$. The pgf of N^L is $P_{N^L} = e^{\lambda(z-1)}$. Thus the pgf of N^P is

$$\begin{aligned} P_{N^P}(z) &= e^{\lambda(1+v(z-1)-1)} \\ &= e^{\lambda v(z-1)}, \end{aligned}$$

  So $N^P \sim Poi(\lambda v)$. This means the number of payments has the same distribution as the number of losses, but with the expected number of payments equal to $\lambda v = \lambda \Pr(X > d)$.

- $N^L \sim NB(\beta, r)$. The pgf of N^L is $P_{N^L}(z) = [1 - \beta(z-1)]^{-r}$. Thus the pgf of N^P is

$$\begin{aligned} P_{N^P}(z) &= (1 - \beta(1 + v(z-1) - 1))^{-r} \\ &= (1 - \beta v(z-1))^{-r}, \end{aligned}$$

  So $N^P \sim NB(\beta v, r)$. This means the number of payments has the same distribution as the number of losses, but with parameters βv and r.

Example 5.5.2. Suppose that loss amounts $X_i \sim Pareto(\alpha = 4,\ \theta = 150)$. You are given that the loss frequency is $N^L \sim Poi(\lambda)$ and the payment frequency distribution is $N_1^P \sim Poi(0.4)$ at deductible level $d_1 = 30$. Find the distribution of the payment frequency N_2^P when the deductible level is $d_2 = 100$.

Show Example Solution

Solution. Because the loss frequency N^L is Poisson, we can relate the means of the loss distribution N^L and the first payment distribution N_1^P (under deductible $d_1 = 30$) through $0.4 = \lambda v_1$, where

$$\begin{aligned} v_1 = \Pr(X > 30) &= \left(\frac{150}{30+150}\right)^4 = \left(\frac{5}{6}\right)^4 \\ \Rightarrow \lambda &= 0.4\left(\frac{6}{5}\right)^4 \end{aligned}$$

With this, we can assess the second payment distribution N_2^P (under deductible $d_2 = 100$) as being Poisson with mean $\lambda_2 = \lambda v_2$, where

$$\begin{aligned} v_2 = \Pr(X > 100) &= \left(\frac{150}{100+150}\right)^4 = \left(\frac{3}{5}\right)^4 \\ \Rightarrow \lambda_2 = \lambda v_2 &= 0.4\left(\frac{6}{5}\right)^4\left(\frac{3}{5}\right)^4 = 0.1075 \end{aligned}$$

Example 5.5.3. Follow-Up. Now suppose instead that the loss frequency is $N^L \sim NB(\beta,\ r)$ and for deductible $d_1 = 30$, the payment frequency N_1^P is negative binomial with mean 0.4. Find the mean of the payment frequency N_2^P for deductible $d_2 = 100$.

Show Example Solution

Solution. Because the loss frequency N^L is negative binomial, we can relate the parameter β of the N^L distribution and the parameter β_1 of the first payment distribution N_1^P using $\beta_1 = \beta v_1$, where

$$v_1 = \Pr(X > 30) = \left(\frac{5}{6}\right)^4$$

Thus, the mean of N_1^P and the mean of N^L are related via

$$\begin{aligned} 0.4 = r\beta_1 &= r\,(\beta v_1) \\ \Rightarrow r\beta &= \frac{0.4}{v_1} = 0.4\left(\frac{6}{5}\right)^4 \end{aligned}$$

Note that $v_2 = \Pr(X > 100) = \left(\frac{3}{5}\right)^4$ as in the original example. Then the second payment frequency distribution under deductible $d_2 = 100$ is $N_2^P \sim NegBin(\beta v_2,\ r)$ with mean

$$r(\beta v_2) = (r\beta)v_2 = 0.4\left(\frac{6}{5}\right)^4\left(\frac{3}{5}\right)^4 = 0.1075$$

Next, we examine the more general case where N^L is a zero-modified distribution. Recall that a zero-modified distribution can be defined in terms of an unmodified one (as was shown in Section **??**). That is,

$$p_k^M = c\ p_k^0,\ \text{for } k = 1, 2, 3, \ldots,\ \text{with } c = \frac{1 - p_0^M}{1 - p_0^0},$$

where p_k^0 is the pmf of the unmodified distribution. In the case that $p_0^M = 0$, we call this a zero-truncated distribution, or ZT. For other arbitrary values of p_0^M, this is a zero-modified, or ZM, distribution. The pgf for the modified distribution is shown as

$$P^M(z) = 1 - c + c\ P^0(z),$$

expressed in terms of the pgf of the unmodified distribution, $P^0(z)$. When N^L follows a zero-modified distribution, the distribution of N^P is established using the same relation from earlier, $P_{N^P}(z) = P_{N^L}\,(1 + v(z-1))$.

Special Cases:

- N^L is a ZM-Poisson random variable with parameters λ and p_0^M. The pgf of N^L is
$$P_{N^L}(z) = 1 - \frac{1 - p_0^M}{1 - e^{-\lambda}} + \frac{1 - p_0^M}{1 - e^{-\lambda}}\left(e^{\lambda(z-1)}\right).$$

Thus the pgf of N^P is

$$P_{N^P}(z) = 1 - \frac{1-p_0^M}{1-e^{-\lambda}} + \frac{1-p_0^M}{1-e^{-\lambda}}\left(e^{\lambda v(z-1)}\right).$$

So the number of payments is also a ZM-Poisson distribution with parameters λv and p_0^M. The probability at zero can be evaluated using $\Pr(N^P = 0) = P_{N^P}(0)$.

- N^L is a ZM-negative binomial random variable with parameters β, r, and p_0^M. The pgf of N^L is

$$P_{N^L}(z) = 1 - \frac{1-p_0^M}{1-(1+\beta)^{-r}} + \frac{1-p_0^M}{1-(1+\beta)^{-r}}\left[1-\beta\left(z-1\right)\right]^{-r}.$$

Thus the pgf of N^P is

$$P_{N^P}(z) = 1 - \frac{1-p_0^M}{1-(1+\beta)^{-r}} + \frac{1-p_0^M}{1-(1+\beta)^{-r}}\left[1-\beta v\left(z-1\right)\right]^{-r}.$$

So the number of payments is also a ZM-negative binomial distribution with parameters βv, r, and p_0^M. Similarly, the probability at zero can be evaluated using $\Pr(N^P = 0) = P_{N^P}(0)$.

Example 5.5.4. Aggregate losses are modeled as follows:
(i) The number of losses follows a zero-modified Poisson distribution with $\lambda = 3$ and $p_0^M = 0.5$.
(ii) The amount of each loss has a Burr distribution with $\alpha = 3, \theta = 50, \gamma = 1$.
(iii) There is a deductible of $d = 30$ on each loss.
(iv) The number of losses and the amounts of the losses are mutually independent.

Calculate $\mathrm{E}(N^P)$ and $\mathrm{Var}(N^P)$.

Show Example Solution

Solution. Since N^L follows a ZM-Poisson distribution with parameters λ and p_0^M, we know that N^P also follows a ZM-Poisson distribution, but with parameters λv and p_0^M, where

$$v = \Pr(X > 30) = \left(\frac{1}{1+(30/50)}\right)^3 = 0.2441$$

Thus, N^P follows a ZM-Poisson distribution with parameters $\lambda^* = \lambda v = 0.7324$

and $p_0^M = 0.5$. Finally,

$$\begin{aligned}
\mathrm{E}(N^P) &= (1-p_0^M)\frac{\lambda^*}{1-e^{-\lambda^*}} = 0.5\left(\frac{0.7324}{1-e^{-0.7324}}\right) \\
&= 0.7053 \\
\mathrm{Var}(N^P) &= (1-p_0^M)\left(\frac{\lambda^*[1-(\lambda^*+1)e^{-\lambda^*}]}{(1-e^{-\lambda^*})^2}\right) + p_0^M(1-p_0^M)\left(\frac{\lambda^*}{1-e^{-\lambda^*}}\right)^2 \\
&= 0.5\left(\frac{0.7324(1-1.7324e^{-0.7324})}{(1-e^{-0.7324})^2}\right) + 0.5^2\left(\frac{0.7324}{1-e^{-0.7324}}\right)^2 \\
&= 0.7244
\end{aligned}$$

5.5.3 Impact of Policy Modifications on Aggregate Claims

In this section, we examine how a change in the deductible affects the aggregate payments from an insurance portfolio. We assume that the presence of policy limits (u), coinsurance (α), and inflation (r) have no effect on the underlying distribution of frequency of payments made by an insurer. As in the previous section, we further assume that deductible changes do not impact the underlying distributions of losses for both frequency and severity.

Recall the notation N^L for the number of losses. With ground-up loss amount X and policy deductible d, we use N^P for the number of payments (as defined in the previous section **??**). Also, define the amount of payment on a per-loss basis as

$$X^L = \begin{cases} 0\ , & \text{if } X < \dfrac{d}{1+r} \\ \alpha[(1+r)X-d]\ , & \text{if } \dfrac{d}{1+r} \le X < \dfrac{u}{1+r} \\ \alpha(u-d)\ , & \text{if } X \ge \dfrac{u}{1+r} \end{cases},$$

and the the amount of payment on a per-payment basis as

$$X^P = \begin{cases} \text{undefined}\ , & \text{if } X < \dfrac{d}{1+r} \\ \alpha[(1+r)X-d]\ , & \text{if } \dfrac{d}{1+r} \le X < \dfrac{u}{1+r} \\ \alpha(u-d)\ , & \text{if } X \ge \dfrac{u}{1+r} \end{cases}.$$

In the above, r, u, and α represent the inflation rate, policy limit, and coinsurance, respectively. Hence, aggregate costs (payment amounts) can be expressed either on a per loss or per payment basis:

$$\begin{aligned} S &= X_1^L + \cdots + X_{N^L}^L \\ &= X_1^P + \cdots + X_{N^P}^P \ . \end{aligned}$$

The fundamentals regarding collective risk models are ready to apply. For instance, we have:

$$\begin{aligned} \mathrm{E}(S) &= \mathrm{E}\left(N^L\right)\mathrm{E}\left(X^L\right) = \mathrm{E}\left(N^P\right)\mathrm{E}\left(X^P\right) \\ \mathrm{Var}(S) &= \mathrm{E}\left(N^L\right)\mathrm{Var}\left(X^L\right) + \left[\mathrm{E}\left(X^L\right)\right]^2\mathrm{Var}(N^L) \\ &= \mathrm{E}\left(N^P\right)\mathrm{Var}\left(X^P\right) + \left[\mathrm{E}\left(X^P\right)\right]^2\mathrm{Var}(N^P) \\ M_S(z) &= P_{N^L}\left[M_{X^L}(z)\right] = P_{N^P}\left[M_{X^P}(z)\right] \end{aligned}$$

Example 5.5.5. Actuarial Exam Question. A group dental policy has a negative binomial claim count distribution with mean 300 and variance 800. Ground-up severity is given by the following table:

Severity	Probability
40	0.25
80	0.25
120	0.25
200	0.25

You expect severity to increase 50% with no change in frequency. You decide to impose a per claim deductible of 100. Calculate the expected total claim payment S after these changes.

Show Example Solution

Solution. The cost per loss with a 50% increase in severity and a 100 deductible per claim is

$$X^L \quad = \quad \begin{cases} 0 & 1.5x < 100 \\ 1.5x - 100 & 1.5x \geq 100 \end{cases}$$

This has expectation

$$\begin{aligned} \mathrm{E}(X^L) &= \frac{1}{4}\left[(1.5(40)-100)_+ + (1.5(80)-100)_+ + (1.5(120)-100)_+ + (1.5(200)-100)_+\right] \\ &= \frac{1}{4}\left[(60-100)_+ + (120-100)_+ + (180-100)_+ + (300-100)_+\right] \\ &= \frac{1}{4}\left[0+20+80+200\right] = 75 \end{aligned}$$

Thus, the expected aggregate loss is

$$\mathrm{E}(S) = \mathrm{E}(N)\ \mathrm{E}\left(X^L\right) = 300(75) = 22,500.$$

Example 5.5.6. Follow-Up. What is the variance of the total claim payment, Var S?

Show Example Solution

Solution. On a per loss basis, we have

$$\mathrm{Var}(S) = \mathrm{E}(N)\ \mathrm{Var}\left(X^L\right) + \left[\mathrm{E}\left(X^L\right)\right]^2\ \mathrm{Var}(N)$$

where $\mathrm{E}(N) = 300$ and $\mathrm{Var}(N) = 800$. We find

$$\mathrm{E}\left[(X^L)^2\right] = \frac{1}{4}\left[0^2 + 20^2 + 80^2 + 200^2\right] = 11,700$$

$$\Rightarrow \mathrm{Var}(X^L) = \mathrm{E}\left[(X^L)^2\right] - \left[\mathrm{E}(X^L)\right]^2 = 11,700 - 75^2 = 6,075$$

Thus, the variance of the aggregate claim payment is

$$\mathrm{Var}(S) = 300(6,075) + 75^2(800) = 6,322,500$$

Alternative Method: Using the Per Payment Basis. Previously, we calculated the expected total claim payment by multiplying the expected number of losses by the expected payment per loss. Recall that we can also multiply the expected number of payments by the expected payment per payment. In this case, we have

$$S = X_1^P + \cdots + X_{N_P}^P$$

The probability of a payment is

$$\Pr(1.5X \geq 100) = \Pr(X \geq 66.\bar{6}) = \frac{3}{4}.$$

Thus, the number of payments, N^P has a negative binomial distribution (see negative binomial special case in Section **??**) with mean

$$\mathrm{E}(N^P) = \mathrm{E}(N^L)\ \Pr(1.5X \geq 100) = 300\left(\frac{3}{4}\right) = 225$$

The cost per payment is

$$X^P \quad = \quad \begin{cases} \text{undefined}\ , & \text{if}\ \ 1.5x < 100 \\ 1.5x - 100\ , & \text{if}\ \ 1.5x \geq 100 \end{cases}$$

This has expectation

$$\mathrm{E}(X^P) = \frac{\mathrm{E}(X^L)}{\Pr(1.5X > 100)} = \frac{75}{(3/4)} = 100$$

Thus, as before, the expected aggregate loss is

$$\mathrm{E}(S) = \mathrm{E}(X^P)\ \mathrm{E}(N^P) = 100(225) = 22,500$$

Example 5.5.7. Actuarial Exam Question. A company insures a fleet of vehicles. Aggregate losses have a compound Poisson distribution. The expected number of losses is 20. Loss amounts, regardless of vehicle type, have exponential distribution with $\theta = 200$. To reduce the cost of the insurance, two modifications are to be made:
(i) A certain type of vehicle will not be insured. It is estimated that this will reduce loss frequency by 20%.
(ii) A deductible of 100 per loss will be imposed.

Calculate the expected aggregate amount paid by the insurer after the modifications.

Show Example Solution

Solution. On a per loss basis, we have a 100 deductible. Thus, the expectation per loss is

$$\begin{aligned}\mathrm{E}(X^L) &= E[(X-100)_+] = E(X) - E(X \wedge 100) \\ &= 200 - 200(1 - e^{-100/200}) = 121.31\end{aligned}$$

Loss frequency has been reduced by 20%, resulting in an expected number of losses

$$\mathrm{E}(N^L) = 0.8(20) = 16$$

Thus, the expected aggregate amount paid after the modifications is

$$\mathrm{E}(S) = \mathrm{E}(X^L)\ \mathrm{E}(N^L) = 121.31(16) = 1,941$$

Alternative Method: Using the Per Payment Basis. We can also use the per payment basis to find the expected aggregate amount paid after the modifications. With the deductible of 100, the probability that a payment occurs is $\Pr(X > 100) = e^{-100/200}$. For the per payment severity, plugging in the expression for $\mathrm{E}(X^L)$ from the original example, we have

$$\mathrm{E}(X^P) = \frac{\mathrm{E}(X^L)}{\Pr(X > 100)} = \frac{200 - 200(1 - e^{-100/200})}{e^{-100/200}} = 200$$

This is not surprising – recall that the exponential distribution is memoryless, so the expected claim amount paid in excess of 100 is still exponential with mean 200.

Now we look at the payment frequency

$$\mathrm{E}(N^P) = \mathrm{E}(N^L)\ \Pr(X > 100) = 16\ e^{-100/200} = 9.7$$

Putting this together, we produce the same answer using the per payment basis as the per loss basis from earlier

$$\mathrm{E}(S) = \mathrm{E}(X^P)\ \mathrm{E}(N^P) = 200(9.7) = 1,941$$

Show Quiz Solution

5.6 Further Resources and Contributors

Exercises

Here are a set of exercises that guide the viewer through some of the theoretical foundations of **Loss Data Analytics**. Each tutorial is based on one or more questions from the professional actuarial examinations, typically the Society of Actuaries Exam C.

Aggregate Loss Guided Tutorials

Contributors

- **Peng Shi** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter. Email: pshi@bus.wisc.edu for chapter comments and suggested improvements.
- Chapter reviewers include: Vytaras Brazauska, Mark Maxwell, Jiadong Ren, Di (Cindy) Xu.

TS 5.A.1. Individual Risk Model Properties

For the expected value of the aggregate loss under the individual risk model,

$$\begin{aligned}
\mathrm{E}(S_n) &= \sum_{i=1}^{n} \mathrm{E}(X_i) = \sum_{i=1}^{n} \mathrm{E}(I_i \times B_i) = \sum_{i=1}^{n} \mathrm{E}(I_i)\ \ \mathrm{E}(B_i) \quad \text{from the independence of } I_i\text{'s and } B_i\text{'s} \\
&= \sum_{i=1}^{n} \Pr(I_i = 1)\ \mu_i \quad \text{since the expectation of an indicator variable is the probability it equals 1} \\
&= \sum_{i=1}^{n} q_i\ \mu_i
\end{aligned}$$

For the variance of the aggregate loss under the individual risk model,

$$\begin{aligned}
\text{Var}(S_n) &= \sum_{i=1}^{n} \text{Var}(X_i) \quad \text{from the independence of } X_i\text{'s} \\
&= \sum_{i=1}^{n} \left(\text{E}\left[\text{Var}(X_i|I_i)\right] + \text{Var}\left[\text{E}(X_i|I_i)\right] \right) \quad \text{from the conditional variance formulas} \\
&= \sum_{i=1}^{n} \left(q_i \ \sigma_i^2 \ + \ q_i \ (1-q_i) \ \mu_i^2 \right)
\end{aligned}$$

To see this, note that

$$\begin{aligned}
\text{E}\left[\text{Var}(X_i|I_i)\right] &= \text{Var}(X_i|I_i=0) \ \Pr(I_i=0) + \text{Var}(X_i|I_i=1) \ \Pr(I_i=1) \\
&= q_i \ \sigma_i^2 + (1-q_i) \ (0) = q_i \ \sigma_i^2,
\end{aligned}$$

and

$$\text{Var}\left[\text{E}(X_i|I_i)\right] = q_i \ (1-q_i) \ \mu_i^2 \ ,$$

using the Bernoulli variance shortcut since $\text{E}(X_i|I_i) = 0$ when $I_i = 0$ (probability $\Pr(I_i = 0) = 1 - q_i$) and $\text{E}(X_i|I_i) = \mu_i$ when $I_i = 1$ (probability $\Pr(I_i = 1) = q_i$).

For the probability generating function of the aggregate loss under the individual risk model,

$$\begin{aligned}
P_{S_n}(z) &= \prod_{i=1}^{n} \ P_{X_i}(z) \quad \text{from the independence of } X_i\text{'s} \\
&= \prod_{i=1}^{n} \ \text{E}(z^{\ X_i}) = \prod_{i=1}^{n} \ \text{E}(z^{\ I_i \times B_i}) = \text{E}\left[\text{E}(z^{\ I_i \times B_i}|I_i)\right] \quad \text{from the law of iterated expectations} \\
&= \prod_{i=1}^{n} \left[\ E\left(z^{\ I_i \times B_i}|I_i=0\right) \ \Pr(I_i=0) + E\left(z^{\ I_i \times B_i}|I_i=1\right) \ \Pr(I_i=1) \ \right] \\
&= \prod_{i=1}^{n} \ \left[\ (1) \ (1-q_i) + P_{B_i}(z) \ q_i \ \right] = \prod_{i=1}^{n} \left(\ 1 - q_i + q_i \ P_{B_i}(z) \ \right)
\end{aligned}$$

Lastly, for the moment generating function of the aggregate loss under the individual risk model,

$$
\begin{aligned}
M_{S_n}(t) &= \prod_{i=1}^{n} M_{X_i}(t) \quad \text{from the independence of } X_i\text{'s} \\
&= \prod_{i=1}^{n} \mathrm{E}(e^{t\ X_i}) = \prod_{i=1}^{n} \mathrm{E}\left(e^{\ t\ (I_i \times B_i)}\right) = \prod_{i=1}^{n} \mathrm{E}\left[\mathrm{E}\left(e^{\ t\ (I_i \times B_i)} | I_i\right)\right] \quad \text{from the law of iterated expectations} \\
&= \prod_{i=1}^{n} \left[\ \mathrm{E}\left(e^{\ t\ (I_i \times B_i)} | I_i = 0\right)\ \Pr(I_i = 0) + \mathrm{E}\left(e^{\ t\ (I_i \times B_i)} | I_i = 1\right)\ \Pr(I_i = 1)\ \right] \\
&= \prod_{i=1}^{n} \left[\ (1)\ (1 - q_i) + M_{B_i}(t)\ q_i\ \right] = \prod_{i=1}^{n} (\ 1 - q_i + q_i\ M_{B_i}(t)\ )
\end{aligned}
$$

TS 5.A.2. Relationship Between Probability Generating Functions of X_i and X_i^T

Let X_i belong to the $(a, b, 0)$ class with pmf $p_{ik} = \Pr(X_i = k)$ for $k = 0, 1, \ldots$ and X_i^T be the associated zero-truncated distribution in the $(a, b, 1)$ class with pmf $p_{ik}^T = p_{ik}/(1 - p_{i0})$ for $k = 1, 2, \ldots$. Then the relationship between the pgf of X_i and the pgf of X_i^T is shown by

$$
\begin{aligned}
P_{X_i}(z) &= \mathrm{E}\ (z^{X_i}) = \mathrm{E}\left[\mathrm{E}\left(z^{X_i} | X_i\right)\right] \quad \text{from the law of iterated expectations} \\
&= \mathrm{E}\left(z^{X_i} | X_i = 0\right)\ \Pr(X_i = 0) + \mathrm{E}\left(z^{X_i} | X_i > 0\right)\ \Pr(X_i > 0) \\
&= (1)\ p_{i0} + \mathrm{E}(z^{X_i^T})\ (1 - p_{i0}) \quad \text{since } (X_i | X_i > 0) \text{ is the zero-truncated random variable } X_i^T \\
&= p_{i0} + (1 - p_{i0}) P_{X_i^T}(z)
\end{aligned}
$$

TS 5.A.3. Example 5.3.8 Moment Generating Function of Aggregate Loss S_N

For $N \sim Geo(\beta)$ and $X \sim Exp(\theta)$, we have

$$
\begin{aligned}
P_N(z) &= \frac{1}{1 - \beta(z - 1)} \\
M_X(t) &= \frac{1}{1 - \theta t}
\end{aligned}
$$

Thus, the mgf of aggregate loss S_N is

$$\begin{aligned}
M_{S_N}(t) = P_N[M_X(t)] &= \frac{1}{1-\beta\left(\frac{1}{1-\theta t}-1\right)} \\
&= \frac{1}{1-\beta\left(\frac{\theta t}{1-\theta t}\right)} + 1 - 1 = 1 + \frac{\beta\left(\frac{\theta t}{1-\theta t}\right)}{1-\beta\left(\frac{\theta t}{1-\theta t}\right)} \\
&= 1 + \frac{\beta\theta t}{(1-\theta t)-\beta\theta t} = 1 + \frac{\beta\theta t}{1-\theta t(1+\beta)} \cdot \frac{1+\beta}{1+\beta} \\
&= 1 + \frac{\beta}{1+\beta}\left[\frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\
&= 1 + \frac{\beta}{1+\beta}\left[\frac{1}{1-\theta(1+\beta)t} - 1\right],
\end{aligned}$$

which gives the expression (5.1). For the alternate expression of the mgf (5.2), we continue from where we just left off:

$$\begin{aligned}
M_{S_N}(t) &= 1 + \frac{\beta}{1+\beta}\left[\frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\
&= \frac{1+\beta}{1+\beta} + \frac{\beta}{1+\beta}\left[\frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\
&= \frac{1}{1+\beta} + \frac{\beta}{1+\beta} + \frac{\beta}{1+\beta}\left[\frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\
&= \frac{1}{1+\beta} + \frac{\beta}{1+\beta}\left[1 + \frac{\theta(1+\beta)t}{1-\theta(1+\beta)t}\right] \\
&= \frac{1}{1+\beta} + \frac{\beta}{1+\beta}\left[\frac{1}{1-\theta(1+\beta)t}\right]
\end{aligned}$$

Chapter 6

Simulation and Resampling

Chapter Preview. Simulation is a computationally intensive method used to solve difficult problems. Instead of creating physical processes and experimenting with them in order to understand their operational characteristics, a simulation study is based on a computer representation - it considers various hypothetical conditions as inputs and summarizes the results. Through simulation, a vast number of hypothetical conditions can be quickly and inexpensively examined. Section **??** introduces simulation, a wonderful computational tool that is especially useful in complex, multivariate settings.

We can also use simulation to draw from an empirical distribution - this process is known as resampling. Resampling allows us to assess the uncertainty of estimates in complex models. Section **??** introduces resampling in the context of bootstrapping to determine the precision of estimators.

Subsequent sections introduce other topics in resampling. Section **??** on cross-validation shows how to use it for model selection and validation. Section **??** on importance sampling describes resampling in specific regions of interest, such as long-tailed actuarial applications. Section **??** on Monte Carlo Markov Chain (MCMC) introduces the simulation and resampling engine underpinning much of modern Bayesian analysis.

This chapter is being written and is not yet complete nor edited. It is here to give you a flavor of what will be in the final version.

6.1 Simulation Fundamentals

In this section, you learn how to:

- Generate approximately independent realizations that are uniformly distributed
- Transform the uniformly distributed realizations to observations from a probability distribution of interest
- Calculate quantities of interest and determining the precision of the calculated quantities

---

6.1.1 Generating Independent Uniform Observations

The simulations that we consider are generated by computers. A major strength of this approach is that they can be replicated, allowing us to check and improve our work. Naturally, this also means that they are not really random. Nonetheless, algorithms have been produced so that results appear to be random for all practical purposes. Specifically, they pass sophisticated tests of independence and can be designed so that they come from a single distribution - our iid assumption, identically and independently distributed.

To get a sense as to what these algorithms do, we consider a historically prominent method.

Linear Congruential Generator. To generate a sequence of random numbers, start with B_0, a starting value that is known as a seed. This value is updated using the recursive relationship

$$B_{n+1} = aB_n + c \text{ modulo } m, \quad n = 0, 1, 2, \ldots.$$

This algorithm is called a linear congruential generator. The case of $c = 0$ is called a multiplicative congruential generator; it is particularly useful for really fast computations.

For illustrative values of a and m, Microsoft's Visual Basic uses $m = 2^{24}$, $a = 1,140,671,485$, and $c = 12,820,163$ (see https://en.wikipedia.org/wiki/Linear_congruential_generator). This is the engine underlying the random number generation in Microsoft's Excel program.

The sequence used by the analyst is defined as $U_n = B_n/m$. The analyst may interpret the sequence $\{U_i\}$ to be (approximately) identically and independently uniformly distributed on the interval (0,1). To illustrate the algorithm, consider the following.

Example 6.1.1. Illustrative Sequence. Take $m = 15$, $a = 3$, $c = 2$ and $B_0 = 1$. Then we have:

step n	B_n	U_n
0	$B_0 = 1$	
1	$B_1 = \mod (3 \times 1 + 2) = 5$	$U_1 = \frac{5}{15}$
2	$B_2 = \mod (3 \times 5 + 2) = 2$	$U_2 = \frac{2}{15}$
3	$B_3 = \mod (3 \times 2 + 2) = 8$	$U_3 = \frac{8}{15}$
4	$B_4 = \mod (3 \times 8 + 2) = 11$	$U_4 = \frac{11}{15}$

Sometimes computer generated random results are known as pseudo-random numbers to reflect the fact that they are machine generated and can be replicated. That is, despite the fact that $\{U_i\}$ appears to be i.i.d, it can be reproduced by using the same seed number (and the same algorithm).

Example 6.1.2. Generating uniform random numbers in `R`. The following code shows how to generate three uniform (0,1) numbers in `R` using the `runif` command. The `set.seed()` function sets the initial seed. In many computer packages, the initial seed is set using the system clock unless specified otherwise.

Three Uniform Random Variates

```
set.seed(2017)
U <- runif(3)
knitr::kable(U, digits=5, align = "c", col.names = "Uniform")
```

Uniform
0.92424
0.53718
0.46920

The linear congruential generator is just one method of producing pseudo-random outcomes. It is easy to understand and is (still) widely used. The linear congruential generator does have limitations, including the fact that it is possible to detect long-run patterns over time in the sequences generated (recall that we can interpret independence to mean a total lack of functional patterns). Not surprisingly, advanced techniques have been developed that address some of this method's drawbacks.

6.1.2 Inverse Transform Method

With the sequence of uniform random numbers, we next transform them to a distribution of interest, say F. A prominent technique is the inverse transform method, defined as

$$X_i = F^{-1}(U_i).$$

Here, recall from Section 4.1.1 that we introduced the inverse of the distribution function, F^{-1}, and referred to it also as the quantile function. Specifically, it is defined to be

$$F^{-1}(y) = \inf_x \{F(x) \geq y\}.$$

Recall that inf stands for infimum or the greatest lower bound. It is essentially the smallest value of x that satisfies the inequality $\{F(x) \geq y\}$. The result is that the sequence $\{X_i\}$ is approximately iid with distribution function F.

The inverse transform result is available when the underlying random variable is continuous, discrete or a hybrid combination of the two. We now present a series of examples to illustrate its scope of applications.

Example 6.1.3. Generating exponential random numbers. Suppose that we would like to generate observations from an exponential distribution with scale parameter θ so that $F(x) = 1 - e^{-x/\theta}$. To compute the inverse transform, we can use the following steps:

$$\begin{aligned} y = F(x) &\Leftrightarrow y = 1 - e^{-x/\theta} \\ &\Leftrightarrow -\theta \ln(1-y) = x = F^{-1}(y). \end{aligned}$$

Thus, if U has a uniform (0,1) distribution, then $X = -\theta \ln(1-U)$ has an exponential distribution with parameter θ.

The following `R` code shows how we can start with the same three uniform random numbers as in Example 6.1.2 and transform them to independent exponentially distributed random variables with a mean of 10. Alternatively, you can directly use the `rexp` function in `R` to generate random numbers from the exponential distribution. The algorithm built into this routine is different so even with the same starting seed number, individual realizations will differ.

```
set.seed(2017)
U <- runif(3)
X1 <- -10*log(1-U)
set.seed(2017)
X2 <- rexp(3, rate = 1/10)
```

Three Uniform Random Variates

Uniform	Exponential 1	Exponential 2
0.92424	25.80219	3.25222
0.53718	7.70409	8.47652
0.46920	6.33362	5.40176

Example 6.1.4. Generating Pareto random numbers. Suppose that we would like to generate observations from a Pareto distribution with parameters α and θ so that $F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha}$. To compute the inverse transform, we can use the following steps:

$$\begin{aligned} y = F(x) &\Leftrightarrow 1 - y = \left(\frac{\theta}{x+\theta}\right)^{\alpha} \\ &\Leftrightarrow (1-y)^{-1/\alpha} = \frac{x+\theta}{\theta} = \frac{x}{\theta} + 1 \\ &\Leftrightarrow \theta\left((1-y)^{-1/\alpha} - 1\right) = x = F^{-1}(y). \end{aligned}$$

Thus, $X = \theta\left((1-U)^{-1/\alpha} - 1\right)$ has a Pareto distribution with parameters α and θ.

Inverse Transform Justification. Why does the random variable $X = F^{-1}(U)$ have a distribution function F?

Show A Snippet of Theory

This is easy to establish in the continuous case. Because U is a Uniform random variable on (0,1), we know that $\Pr(U \leq y) = y$, for $0 \leq y \leq 1$. Thus,

$$\begin{aligned} \Pr(X \leq x) &= \Pr(F^{-1}(U) \leq x) \\ &= \Pr(F(F^{-1}(U)) \leq F(x)) \\ &= \Pr(U \leq F(x)) = F(x) \end{aligned}$$

as required. The key step is that $F(F^{-1}(u)) = u$ for each u, which is clearly true when F is strictly increasing.

We now consider some discrete examples.

Example 6.1.5. Generating Bernoulli random numbers. Suppose that we wish to simulate random variables from a Bernoulli distribution with parameter $p = 0.85$.

A graph of the cumulative distribution function in Figure **??** shows that the quantile function can be written as

$$F^{-1}(y) = \begin{cases} 0 & 0 < y \leq 0.85 \\ 1 & 0.85 < y \leq 1.0. \end{cases}$$

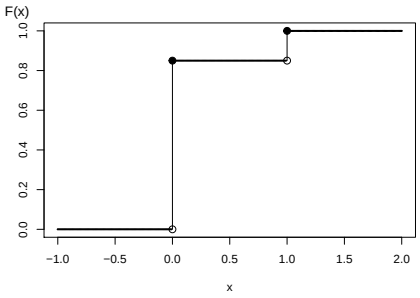


Figure 6.1: Distribution Function of a Binary Random Variable

Thus, with the inverse transform we may define

$$X = \begin{cases} 0 & 0 < U \leq 0.85 \\ 1 & 0.85 < U \leq 1.0 \end{cases}$$

For illustration, we generate three random numbers to get

```
set.seed(2017)
U <- runif(3)
X <- 1*(U > 0.85)
```

Three Random Variates

Uniform	Binary X
0.92424	1
0.53718	0
0.46920	0

Example 6.1.6. Generating random numbers from a discrete distribution. Consider the time of a machine failure in the first five years. The distribution of failure times is given as:

Discrete Distribution

	$~~~~~~~~~~$	$~~~~~~~~~~$	$~~~~~~~~~~$	$~~~~~~~~~~$	$~~~~~~~~~~$
Time	1.0	2.0	3.0	4.0	5.0
Probability	0.1	0.2	0.1	0.4	0.2
Distribution Function $F(x)$	0.1	0.3	0.4	0.8	1.0

Using the graph of the distribution function in Figure **??**, with the inverse transform we may define

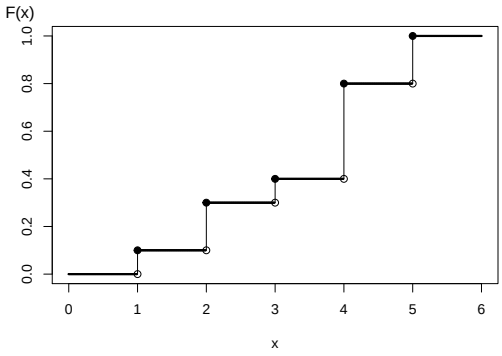


Figure 6.2: Distribution Function of a Discrete Random Variable

$$X = \begin{cases} 1 & 0 < U \leq 0.1 \\ 2 & 0.1 < U \leq 0.3 \\ 3 & 0.3 < U \leq 0.4 \\ 4 & 0.4 < U \leq 0.8 \\ 5 & 0.8 < U \leq 1.0. \end{cases}$$

For general discrete random variables there may not be an ordering of outcomes. For example, a person could own one of five types of life insurance products and we might use the following algorithm to generate random outcomes:

$$X = \begin{cases} \text{whole life} & 0 < U \leq 0.1 \\ \text{endowment} & 0.1 < U \leq 0.3 \\ \text{term life} & 0.3 < U \leq 0.4 \\ \text{universal life} & 0.4 < U \leq 0.8 \\ \text{variable life} & 0.8 < U \leq 1.0. \end{cases}$$

Another analyst may use an alternative procedure such as:

$$X = \begin{cases} \text{whole life} & 0.9 < U < 1.0 \\ \text{endowment} & 0.7 \leq U < 0.9 \\ \text{term life} & 0.6 \leq U < 0.7 \\ \text{universal life} & 0.2 \leq U < 0.6 \\ \text{variable life} & 0 \leq U < 0.2. \end{cases}$$

Both algorithms produce (in the long-run) the same probabilities, e.g., Pr(whole life) = 0.1, and so forth. So, neither is incorrect. You should be aware that "there is more than one way to skin a cat." (What an old expression!) Similarly, you could use an alternative algorithm for ordered outcomes (such as failure times 1, 2, 3, 4, or 5, above).

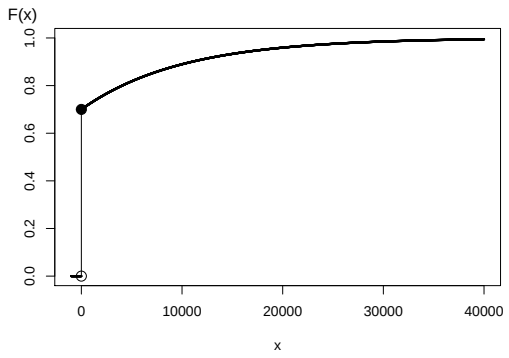


Figure 6.3: Distribution Function of a Hybrid Random Variable

Example 6.1.7. Generating random numbers from a hybrid distribution. Consider a random variable that is 0 with probability 70% and is exponentially distributed with parameter $\theta = 10,000$ with probability 30%. In an insurance application, this might correspond to a 70% chance of having no insurance claims and a 30% chance of a claim - if a claim occurs, then it is exponentially distributed. The distribution function, depicted in Figure **??**, is given as

$$F(y) = \begin{cases} 0 & x < 0 \\ 1 - 0.3\exp(-x/10000) & x \geq 0. \end{cases}$$

From Figure **??**, we can see that the inverse transform for generating random variables with this distribution function is

$$X = F^{-1}(U) = \begin{cases} 0 & 0 < U \leq 0.7 \\ -1000\ln(\frac{1-U}{0.3}) & 0.7 < U < 1. \end{cases}$$

For discrete and hybrid random variables, the key is to draw a graph of the distribution function that allows you to visualize potential values of the inverse function.

6.1.3 Simulation Precision

From the prior subsections, we now know how to generate independent simulated realizations from a distribution of interest. With these realizations, we can construct an empirical distribution and approximate the underlying distribution as precisely as needed. As we introduce more actuarial applications in this book, you will see that simulation can be applied in a wide variety of contexts.

Many of these applications can be reduced to the problem of approximating E $h(X)$, where $h(\cdot)$ is some known function. Based on R simulations (replications), we get $X_1, \ldots, X_R$. From this simulated sample, we calculate an average

$$\overline{h}_R = \frac{1}{R}\sum_{i=1}^{R} h(X_i)$$

that we use as our simulated approximate (estimate) of E $h(X)$. To estimate the precision of this approximation, we use the simulation variance

$$s^2_{h,R} = \frac{1}{R-1}\sum_{i=1}^{R} \left(h(X_i) - \overline{h}_R\right)^2.$$

From the independence, the standard error of the estimate is $s_{h,R}/\sqrt{R}$. This can be made as small as we like by increasing the number of replications R.

Example. 6.1.8. Portfolio management. In Section 3.4, we learned how to calculate the expected value of policies with deductibles. For an example of something that cannot be done with closed form expressions, we now consider two risks. This is a variation of a more complex example that will be covered as Example 10.3.6.

We consider two property risks of a telecommunications firm:

- X_1 - buildings, modeled using a gamma distribution with mean 200 and scale parameter 100.
- X_2 - motor vehicles, modeled using a gamma distribution with mean 400 and scale parameter 200.

Denote the total risk as $X = X_1 + X_2$. For simplicity, you assume that these risks are independent.

To manage the risk, you seek some insurance protection. You wish to manage internally small building and motor vehicles amounts, up to M, say. Your retained risk is $Y_{retained} = \min(X_1+X_2, M)$. The insurer's portion is $Y_{insurer} = X - Y_{retained}$.

To be specific, we use $M = 400$ as well as $R = 1000000$ simulations.

a. With the settings, we wish to determine the expected claim amount and the associated standard deviation of (i) that retained, (ii) that accepted by the insurer, and (iii) the total overall amount.

```
# Simulate the risks
nSim <- 1e6  #number of simulations
set.seed(2017) #set seed to reproduce work
X1 <- rgamma(nSim,alpha1,scale = theta1)
X2 <- rgamma(nSim,alpha2,scale = theta2)
```

```
# Portfolio Risks
X         <- X1 + X2
Yretained <- pmin(X, M)
Yinsurer  <- X - Yretained
```

Here is the code for the expected claim amounts.

```
# Expected Claim Amounts
ExpVec <- t(as.matrix(c(mean(Yretained),mean(Yinsurer),mean(X))))
sdVec <- t(as.matrix(c(sd(Yretained),sd(Yinsurer),sd(X))))
outMat <- rbind(ExpVec, sdVec)
colnames(outMat) <- c("Retained", "Insurer","Total")
row.names(outMat) <- c("Mean","Standard Deviation")
round(outMat,digits=2)
```

```
                  Retained Insurer  Total
Mean                365.17  235.01 600.18
Standard Deviation   69.51  280.86 316.36
```

b. For insured claims, the standard error of the simulation approximation is $s_{h,R}/\sqrt{1000000} = 280.86\ /\sqrt{1000000} = 0.281$. For this example, simulation is quick and so a large value such as 1000000 is an easy choice. However, for more complex problems, the simulation size may be an issue. Figure **??** allows us to visualize the development of the approximation as the number of simulations increases.

```
Yinsurefct <- function(numSim){
X1 <- rgamma(numSim,alpha1,scale = theta1)
X2 <- rgamma(numSim,alpha2,scale = theta2)
# Portfolio Risks
X         <- X1 + X2
Yinsurer <- X - pmin(X, M)
return(Yinsurer)
}

R <- 1e3
nPath <- 20
set.seed(2017)
simU <- matrix(Yinsurefct(R*nPath),R,nPath)
sumP2 <- apply(simU, 2, cumsum)/(1:R)
matplot(1:R,sumP2[,1:20],type="l",col=rgb(1,0,0,.2), ylim=c(100, 400),
        xlab=expression(paste("Number of Simulations (", italic('R'), ")")), ylab="Expected Insur
abline(h=mean(Yinsurer),lty=2)
bonds <- cbind(1.96*sd(Yinsurer)*sqrt(1/(1:R)),-1.96*sd(Yinsurer)*sqrt(1/(1:R)))
matlines(1:R,bonds+mean(Yinsurer),col="red",lty=1)
```

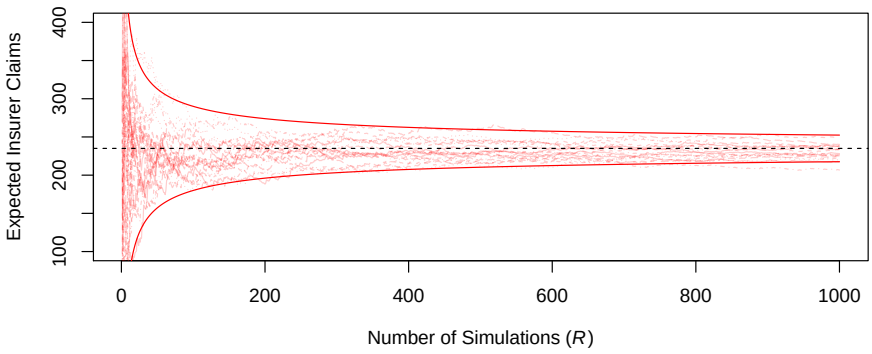


Figure 6.4: Estimated Expected Insurer Claims versus Number of Simulations.

Determination of Number of Simulations

How many simulated values are recommended? 100? 1,000,000? We can use the central limit theorem to respond to this question.

As one criterion for your confidence in the result, suppose that you wish to be within 1% of the mean with 95% certainty. That is, you want $\Pr\left(|\overline{h}_R - \mathrm{E}\ h(X)| \leq 0.01\mathrm{E}\ h(X)\right) \leq 0.95$. According to the central limit theorem, your estimate should be approximately normally distributed and so we want to have R large enough to satisfy $0.01\mathrm{E}\ h(X)/\sqrt{\mathrm{Var}\ h(X)/R} \geq 1.96$. (Recall that 1.96 is the 97.5th percentile from the standard normal distribution.) Replacing E $h(X)$ and Var $h(X)$ with estimates, you continue your simulation until

$$\frac{.01\overline{h}_R}{s_{h,R}/\sqrt{R}} \geq 1.96$$

or equivalently

$$R \geq 38,416 \frac{s_{h,R}^2}{\overline{h}_R^2}. \tag{6.1}$$

This criterion is a direct application of the approximate normality. Note that $\overline{h}_R$ and $s_{h,R}$ are not known in advance, so you will have to come up with estimates, either by doing a small pilot study in advance or by interrupting your procedure intermittently to see if the criterion is satisfied.

Example. 6.1.8. Portfolio management - continued

For our example, the average insurance claim is 235.011 and the corresponding standard deviation is 280.862. Using equation (**??**), to be within 10% of the mean, we would only require at least 54.87 thousand simulations. However, to be within 1% we would want at least 5.49 million simulations.

Example. 6.1.9. Approximation choices. An important application of simulation is the approximation of E $h(X)$. In this example, we show that the choice of the $h(\cdot)$ function and the distribution of X can play a role.

Consider the following question : what is $\Pr[X > 2]$ when X has a Cauchy distribution, with density $f(x) = \left(\pi(1+x^2)\right)^{-1}$, on the real line? The true value is

$$\Pr[X > 2] = \int_2^\infty \frac{dx}{\pi(1+x^2)}.$$

One can use an R numerical integration function (which works usually well on improper integrals)

```
true_value <- integrate(function(x) 1/(pi*(1+x^2)),lower=2,upper=Inf)$value
```

which is equal to 0.14758.

Alternatively, one can use simulation techniques to approximate that quantity. From calculus, you can check that the quantile function of the Cauchy distribution is $F^{-1}(y) = \tan(\pi(y-0.5))$. Then, with simulated uniform (0,1) variates, $U_1, \ldots, U_R$, we can construct the estimator

$$p_1 = \frac{1}{R}\sum_{i=1}^{R} \mathrm{I}(F^{-1}(U_i) > 2) = \frac{1}{R}\sum_{i=1}^{R} \mathrm{I}(\tan(\pi(U_i - 0.5)) > 2).$$

```
Q <- function(u) tan(pi*(u-.5))
R <- 1e6
set.seed(1)
X <- Q(runif(R))
p1 <- mean(X>2)
se.p1 <- sd(X>2)/sqrt(R)
p1
```

```
[1] 0.147439
```

```
se.p1
```

```
[1] 0.0003545432
```

With one million simulations, we obtain an estimate of 0.14744 with standard error 0.355 (divided by 1000). One can prove that the variance of p_1 is of order $0.127/R$.

With other choices of $h(\cdot)$ and $F(\cdot)$, it is actually possible to reduce uncertainty even using the same number of simulations R. To begin, one can use the symmetry of the Cauchy distribution to write $\Pr[X > 2] = 0.5 \cdot \Pr[|X| > 2]$. With this, can construct a new estimator

$$p_2 = \frac{1}{2R} \sum_{i=1}^{R} \mathrm{I}(|F^{-1}(U_i)| > 2).$$

With one million simulations, we obtain an estimate of 0.14748 with standard error 0.228 (divided by 1000). One can prove that the variance of p_2 is of order $0.052/R$.

But one can go one step further. The improper integral can be written as a proper one by a simple symmetry property (since the function is symmetry and the integral on the real line is equal to 1)

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \frac{1}{2} - \int_0^2 \frac{dx}{\pi(1+x^2)}.$$

From this expression, a natural approximation would be

$$p_3 = \frac{1}{2} - \frac{1}{R} \sum_{i=1}^{R} h_3(2U_i), \qquad \text{where } h_3(x) = \frac{2}{\pi(1+x^2)}.$$

With one million simulations, we obtain an estimate of 0.14756 with standard error 0.169 (divided by 1000). One can prove that the variance of p_3 is of order $0.0285/R$.

Finally, one can also consider some change of variable in the integral

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \int_0^{1/2} \frac{y^{-2}dy}{\pi(1-y^{-2})}.$$

From this expression, a natural approximation would be

$$p_4 = \frac{1}{R} \sum_{i=1}^{R} h_4(U_i/2), \qquad \text{where } h_4(x) = \frac{1}{2\pi(1+x^2)}.$$

The expression seems rather similar to the previous one,

With one million simulations, we obtain an estimate of 0.14759 with standard error 0.01 (divided by 1000). One can prove that the variance of p_4 is of order $0.00009/R$, which is much smaller than what we had so far !

Table 6.1 summarizes the four choices of $h(\cdot)$ and $F(\cdot)$ to approximate $\Pr[X > 2] = 0.14758$. The standard error varies dramatically. Thus, if we have a desired degree of accuracy, then the number of simulations depends strongly on how we write the integrals we try to approximate.

Table 6.1. Summary of Four Choices to Approximate $\Pr[X > 2]$.

Estimator	Definition	Support Function	Estimate	Standard Error
p_1	$\frac{1}{R}\sum_{i=1}^{R} \mathrm{I}(F^{-1}(U_i) > 2)$	$F^{-1}(u) = \tan(\pi(u - 0.5))$	0.147439	0.000355
p_2	$\frac{1}{2R}\sum_{i=1}^{R} \mathrm{I}(\lvert F^{-1}(U_i)\rvert > 2)$	$F^{-1}(u) = \tan(\pi(u - 0.5))$	0.147477	0.000228
p_3	$\frac{1}{2} - \frac{1}{R}\sum_{i=1}^{R} h_3(2U_i)$	$h_3(x) = \frac{2}{\pi(1+x^2)}$	0.147558	0.000169
p_4	$\frac{1}{R}\sum_{i=1}^{R} h_4(U_i/2)$	$h_4(x) = \frac{1}{2\pi(1+x^2)}$	0.147587	0.000010

6.1.4 Simulation and Statistical Inference

Simulations not only help us approximate expected values but are also useful in calculating other aspects of distribution functions. In particular, they are very useful when distributions of test statistics are too complicated to derive; in this

case, one can use simulations to approximate the reference distribution. We now illustrate this with the Kolmogorov-Smirnov test that we learned about in Section 4.1.2.2.

Example. 6.1.10. Kolmogorov-Smirnov Test of Distribution. Suppose that we have available $n = 100$ observations $\{x_1, \cdots, x_n\}$ that, unknown to the analyst, was generated from a gamma distribution with parameters $\alpha = 6$ and $\theta = 2$. The analyst believes that the data come from a lognormal distribution with parameters 1 and 0.4 and would like to test this assumption.

The first step is to visualize the data. Figure **??** provides a graph of a histogram and empirical distribution. For reference, superimposed are red dashed lines from the lognormal distribution.

```
set.seed(1)
n <- 100
x <- rgamma(n, 6, 2)
par(mfrow=c(1,2))
hist(x,probability = TRUE,main="Histogram", col="light blue",border="white",xlim=c(0,7)
u=seq(0,7,by=.01)
lines(u,dlnorm(u,1,.4),col="red",lty=2)
vx = c(0,sort(x))
vy = (0:n)/n
plot(vx,vy,type="l",xlab="x",ylab="Cumulative Distribution",main="Empirical cdf")
lines(u,plnorm(u,1,.4),col="red",lty=2)
```

Recall that the Kolmogorov-Smirnov statistic equals the largest discrepancy between the empirical and the hypothesized distribution. This is $\max_x |F_n(x) - F_0(x)|$, where F_0 is the hypothesized lognormal distribution. We can calculate this directly as:

```
# test statistic
D <- function(data, F0){
   F <- Vectorize(function(x) mean((data<=x)))
   n <- length(data)
   x <- sort(data)
   d1=abs(F(x+1e-6)-F0(x+1e-6))
   d2=abs(F(x-1e-6)-F0(x-1e-6))
   return(max(c(d1,d2)))
}
D(x,function(x) plnorm(x,1,.4))
```

```
[1] 0.09703627
```

Fortunately, for the lognormal distribution, R has built-in tests that allow us to determine this without complex programming:

```
ks.test(x, plnorm, mean=1, sd=0.4)
```

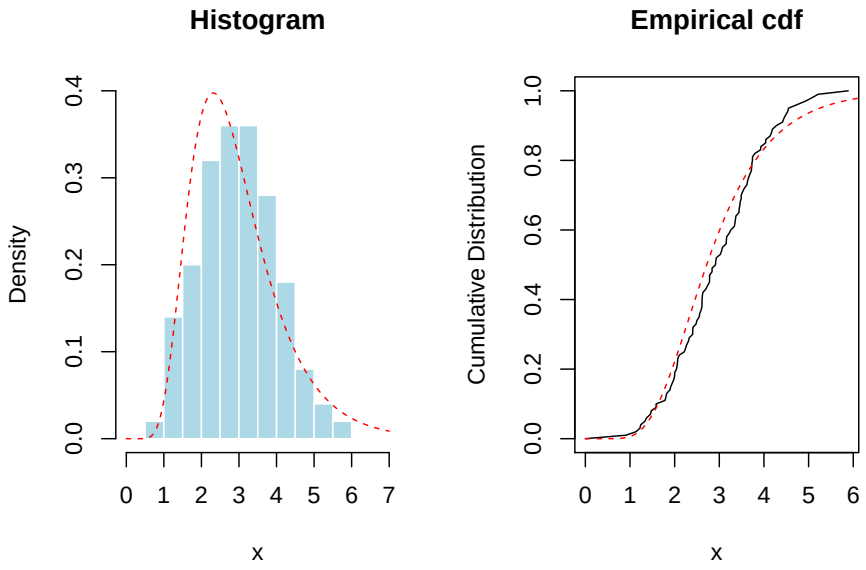


Figure 6.5: Histogram and Empirical Distribution Function of Data used in Kolmogorov-Smirnov Test. The red dashed lines are fits based on (incorrectly) hypothesized lognormal distribution.

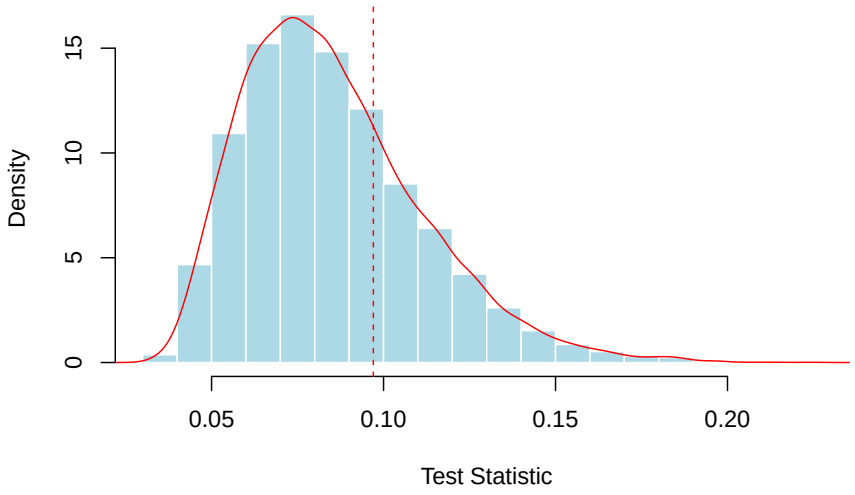


Figure 6.6: Simulated Distribution of the Kolmogorov-Smirnov Test Statistic. The vertical red dashed line marks the test statistic for the sample of 100.

```
    One-sample Kolmogorov-Smirnov test

data:  x
D = 0.097037, p-value = 0.3031
alternative hypothesis: two-sided
```

However, for many distributions of actuarial interest, pre-built programs are not available. We can use simulation to test the relevance of the test statistic. Specifically, to compute the p-value, let us generate thousands of random samples from a $LN(1, 0.4)$ distribution (with the same size), and compute empirically the distribution of the statistic,

```
ns <- 1e4
d_KS <- rep(NA,ns)
# compute the test statistics for a large (ns) number of simulated samples
for(s in 1:ns) d_KS[s] <- D(rlnorm(n,1,.4),function(x) plnorm(x,1,.4))
hist(d_KS,probability = TRUE,col="light blue",border="white",xlab="Test Statistic",mai
lines(density(d_KS),col="red")
abline(v=D(x,function(x) plnorm(x,1,.4)),lty=2,col="red")
```

```
mean(d_KS>D(x,function(x) plnorm(x,1,.4)))
```

```
[1] 0.2843
```

The simulated distribution based on 10,000 random samples is summarized in Figure **??**. Here, the statistic exceeded the empirical value (0.09704) in 28.43% of the scenarios, while the theoretical p-value is 0.3031. For both the simulation and the theoretical p-values, the conclusions are the same; the data do not provide sufficient evidence to reject the hypothesis of a lognormal distribution.

Although only an approximation, the simulation approach works in a variety of distributions and test statistics without needing to develop the nuances of the underpinning theory for each situation. We summarize the procedure for developing simulated distributions and p-values as follows:

1. Draw a sample of size n, say, $X_1, \ldots, X_n$, from a known distribution function F. Compute a statistic of interest, denoted as $\hat{\theta}(X_1, \ldots, X_n)$. Call this $\hat{\theta}^r$ for the rth replication.
2. Repeat this $r = 1, \ldots, R$ times to get a sample of statistics, $\hat{\theta}^1, \ldots, \hat{\theta}^R$.
3. From the sample of statistics in Step 2, $\{\hat{\theta}^1, \ldots, \hat{\theta}^R\}$, compute a summary measure of interest, such as a p-value.

6.2 Bootstrapping and Resampling

In this section, you learn how to:

- Generate a nonparametric bootstrap distribution for a statistic of interest
- Use the bootstrap distribution to generate estimates of precision for the statistic of interest, including bias, standard deviations, and confidence intervals
- Perform bootstrap analyses for parametric distributions

6.2.1 Bootstrap Foundations

Simulation presented up to now is based on sampling from a known distribution. Section **??** showed how to use simulation techniques to sample and compute quantities from known distributions. However, statistical science is dedicated to providing inferences about distributions that are unknown. We gather summary statistics based on this unknown population distribution. But how do we sample from an unknown distribution?

Naturally, we cannot simulate draws from an unknown distribution but we can draw from a sample of observations. If the sample is a good representation from the population, then our simulated draws from the sample should well approximate the simulated draws from a population. The process of sampling from a sample is called resampling or bootstrapping. The term bootstrap comes from the phrase "pulling oneself up by one's bootstraps" (Efron, 1979). With resampling, the original sample plays the role of the population and estimates from the sample play the role of true population parameters.

The resampling algorithm is that same as introduced in Section **??** except that now we use simulated draws from a sample. It is common to use $\{X_1, \ldots, X_n\}$ to denote the original sample and let $\{X_1^*, \ldots, X_n^*\}$ denote the simulated draws from this sample. We draw them with replacement so that the simulated draws will be independent from one another, the same assumption as with the original sample. We also use n simulated draws, the same number as the original sample size. To distinguish this procedure from the simulation, it is common to use B (for bootstrap) to be the number of simulated samples. We could also write $\{X_1^{(b)}, \ldots, X_n^{(b)}\}$, $b = 1, \ldots, B$ to clarify this.

There are two basic resampling methods, model-free and model-based, which are also known, respectively, as nonparametric and parametric. In the nonparametric approach, no assumption is made about the distribution of the parent population. The simulated draws come from the empirical distribution function $F_n(\cdot)$, so each draw comes from $\{X_1, \ldots, X_n\}$ with probability 1/n.

In contrast, for the parametric approach, we assume that we have knowledge of the distribution family F. The original sample $X_1, \ldots, X_n$ is used to estimate parameters of that family, say, $\hat{\theta}$. Then, simulated draws are taken from the $F(\hat{\theta})$. Section **??** will discuss this approach in further detail.

Nonparametric Bootstrap

The idea of the nonparametric bootstrap is to use the inverse method on F_n, the empirical cumulative distribution function, depicted in Figure **??**.

Because F_n is a step-function, F_n^{-1} will take values in $\{x_1, \cdots, x_n\}$. More precisely, as illustrated in Figure **??**.

- if $y \in (0, 1/n)$ (with probability $1/n$) we will draw the smallest value ($\min\{x_i\}$)
- if $y \in (1/n, 2/n)$ (with probability $1/n$) we will draw the second smallest value,
- ...
- if $y \in ((n-1)/n, 1)$ (with probability $1/n$) we will draw the largest value ($\max\{x_i\}$)

So finally, using the inverse method with F_n mean sampling from $\{x_1, \cdots, x_n\}$, with probability $1/n$. And generating a bootstrap sample of size B means

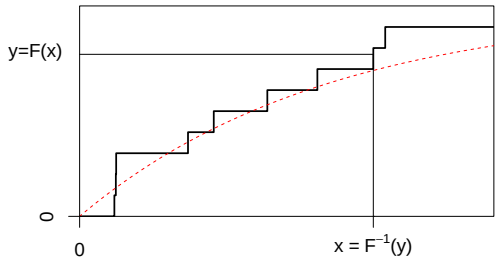


Figure 6.7: Inverse of an Empirical Distribution Function

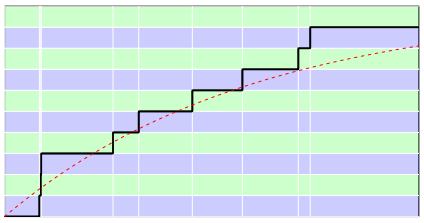


Figure 6.8: Inverse of an Empirical Distribution Function

sampling from $\{x_1, \cdots, x_n\}$, with probability $1/n$, with replacement,

```
set.seed(1)
n <- 10
x <- rexp(n, 1/6)
m <- 8
round(sample(x, size=m, replace=TRUE),digits=4)
```

```
[1] 7.0899 0.8742 0.8388 4.5311 0.8388 5.7394 0.8388 2.6164
```

Observe that value 0.8388 was obtained three times here.

6.2.2 Bootstrap Precision: Bias, Standard Deviation, and MSE

We summarize the nonparametric bootstrap procedure as follows:

1. From the sample $\{X_1, \ldots, X_n\}$, draw a sample of size n (with replacement), say, $X_1^*, \ldots, X_n^*$. From the simulated draws compute a statistic of interest, denoted as $\hat{\theta}(X_1^*, \ldots, X_n^*)$. Call this $\hat{\theta}_b^*$ for the bth replicate.
2. Repeat this $b = 1, \ldots, B$ times to get a sample of statistics, $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
3. From the sample of statistics in Step 2, $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$, compute a summary measure of interest.

In this section, we focus on three summary measures, the bias, the standard deviation, and the mean square error (MSE). Table 6.2 summarizes these three measures. Here, $\overline{\hat{\theta}^*}$ is the average of $\{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$.

Table 6.2. Bootstrap Summary Measures

Population Measure	Population Definition	Bootstrap Definition	Bootstrap Symbol
Bias	$\mathrm{E}(\hat{\theta}) - \theta$	$\overline{\hat{\theta}^*} - \hat{\theta}$	$Bias_{boot}(\hat{\theta})$
Standard Deviation	$\sqrt{\mathrm{Var}(\hat{\theta})}$	$\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2}$	$s_{boot}(\hat{\theta})$
Mean Square Error	$\mathrm{E}(\hat{\theta} - \theta)^2$	$\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}\right)^2$	$MSE_{boot}(\hat{\theta})$

Example 6.2.1. Bodily Injury Claims and Loss Elimination Ratios. To show how the bootstrap can be used to quantify the precision of estimators, we return to the Example 4.1.11 bodily injury claims data where we introduced a nonparametric estimator of the loss elimination ratio.

Table 6.3 summarizes the results of the bootstrap estimation. For example, at $d = 14000$, we saw in Example 4.1.11 that the nonparametric estimate of LER is 0.97678. This has an estimated bias of 0.00018 with a standard deviation of 0.00701. For some applications, you may wish to apply the estimated bias to the original estimate to give a bias-corrected estimator. This is the focus of the

next example. For this illustration, the bias is small and so such a correction is not relevant.

The bootstrap standard deviation gives a measure of precision. For one application of standard deviations, we can use the normal approximation to create a confidence interval. For example, the R function boot.ci produces the normal confidence intervals at 95%. These are produced by creating an interval of twice the length of 1.95994 bootstrap standard deviations, centered about the bias-corrected estimator (1.95994 is the 97.5th quantile of the standard normal distribution). For example, the lower normal 95% CI at $d = 14000$ is $(0.97678 - 0.00018) - 1.95994 * 0.00701 = 0.96286$. We give additional discussion on how to created bootstrap confidence intervals in the next section.

Table 6.3. Bootstrap Estimates of LER at Selected Deductibles

```
# Example from Derrig et al
BIData <- read.csv("../../Data/DerrigResampling.csv", header =T)
BIData$Censored <- 1*(BIData$AmountPaid >= BIData$PolicyLimit)
BIDataUncensored <- subset(BIData, Censored == 0)
LER.boot <- function(ded, data, indices){
  resample.data <- data[indices,]
  sumClaims <- sum(resample.data$AmountPaid)
  sumClaims_d <- sum(pmin(resample.data$AmountPaid,ded))
  LER <-   sumClaims_d/sumClaims
  return(LER)
}
#indices <- 1:nrow(BIDataUncensored)
#LER.boot(data=BIDataUncensored, indices)
#results <- boot(data=BIDataUncensored, statistic=LER.boot, R=1000, ded=4000)
#plot(results)
#results$t0
#boot.ci(results, type="bca")

##Derrig et al
set.seed(2019)
dVec2 <- c(4000, 5000, 10500, 11500, 14000, 18500)
OutBoot <- matrix(0,length(dVec2),6)
colnames(OutBoot) <- c("d","NP Estimate","Bootstrap Bias", "Bootstrap SD", "Lower Normal 95% CI",
  for (i in 1:length(dVec2)) {
OutBoot[i,1] <- dVec2[i]
results <- boot(data=BIDataUncensored, statistic=LER.boot, R=1000, ded=dVec2[i])
OutBoot[i,2] <- results$t0
biasboot <- mean(results$t)-results$t0 -> OutBoot[i,3]
sdboot <- sd(results$t) -> OutBoot[i,4]
temp <- boot.ci(results)
OutBoot[i,5] <- temp$normal[2]
```

```
OutBoot[i,6] <- temp$normal[3]
}
knitr::kable(OutBoot, digits=5)
```

d	NP Estimate	Bootstrap Bias	Bootstrap SD	Lower Normal 95% CI	Upper Normal 95%
4000	0.54113	-0.00012	0.01228	0.51719	0.56
5000	0.64960	-0.00038	0.01400	0.62254	0.67
10500	0.93563	0.00018	0.01065	0.91458	0.95
11500	0.95281	-0.00016	0.00918	0.93497	0.97
14000	0.97678	0.00018	0.00701	0.96286	0.99
18500	0.99382	0.00006	0.00342	0.98706	1.00

Example. 6.2.2. Estimating $\exp(\mu)$**.** The bootstrap can be used to quantify the bias of an estimator, for instance. Consider here a sample $\mathbf{x} = \{x_1, \cdots, x_n\}$ iid with mean μ.

```
sample_x <- c(2.46,2.80,3.28,3.86,2.85,3.67,3.37,3.40,5.22,2.55,2.79,4.50,3.37,2.88,1.4
```

Suppose that the quantity of interest is $\theta = \exp[\mu]$. A natural estimator would be $\widehat{\theta}_1 = \exp(\overline{x})$. This estimator is biased (Jensen inequality) but asymptotically unbiased.

```
(theta_1 <- exp(mean(sample_x)))
```

```
[1] 19.13463
```

One can use the Central Limit Theorem to get a correction since

$$\overline{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ where } \sigma^2 = \text{Var}[X_i],$$

so that, with the normal moment generating function, we have

$$\text{E } \exp[\overline{X}] \approx \exp\left(\mu + \frac{\sigma^2}{2n}\right).$$

Hence, one can consider naturally

$$\widehat{\theta}_2 = \exp\left(\overline{x} - \frac{\widehat{\sigma}^2}{2n}\right).$$

```
n <- length(sample_x)
(theta_2 <- exp(mean(sample_x)-var(sample_x)/(2*n)))
```

```
[1] 18.73334
```

Note that one can also use Taylor's approximation to get a more accurate estimator (as in the delta method),

$$g(\overline{x}) = g(\mu) + [\overline{x} - \mu]g'(\mu) + [\overline{x} - \mu]^2 \frac{g''(\mu)}{2} + \cdots$$

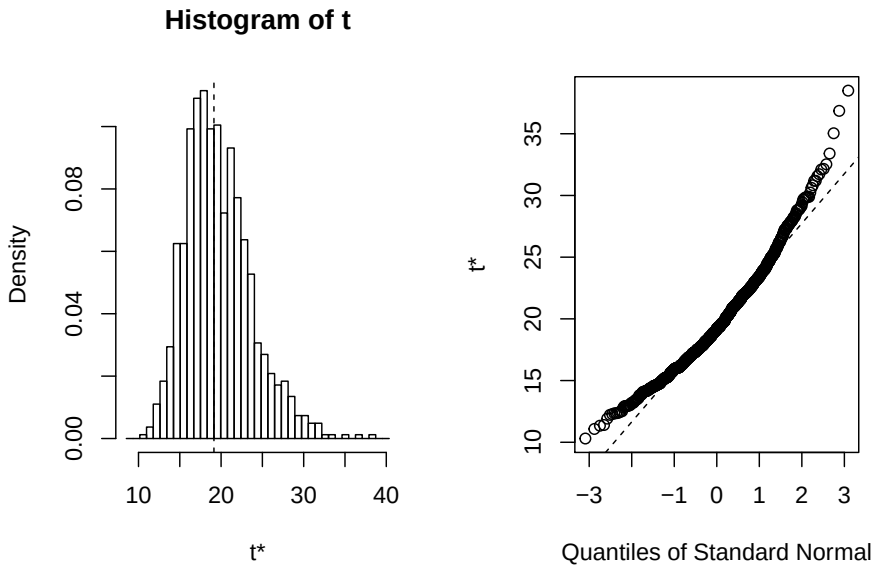


Figure 6.9: Distribution of Bootstrap Replicates. The left-hand panel is a histogram of replicates. The right-hand panel is a quantile-quantile plot, comparing the bootstrap distribution to the standard normal distribution.

Finally, an alternative is to use a bootstrap strategy: given a bootstrap sample, $\mathbf{x}_b^*$, let $\overline{x}_b^*$ denotes its mean, and set

$$\widehat{\theta}_3 = \frac{1}{B}\sum_{b=1}^{B} \exp[\overline{x}_b^*].$$

```
library(boot)
results <- boot(data=sample_x,
                statistic=function(y,indices) exp(mean(y[indices])),
                R=1000)
plot(results)
```

```
print(results)
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = sample_x, statistic = function(y, indices) exp(mean(y[indices])),
```

```
    R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 19.13463 0.5589631    4.021793
```

```
param <- function(x){
  n <- length(x)
  theta_1 <- exp(mean(x))
  theta_2 <- exp(mean(x)-var(x)/(2*n))
  results <- boot(data=x,
                  statistic=function(y,indices) exp(mean(y[indices])),
                  R=999)
  theta_3 <- mean(results$t)
  return(c(theta_1,theta_2,theta_3))
}
param(sample_x)
```

```
[1] 19.13463 18.73334 19.63824
```

How does this work with differing sample sizes? We now suppose that the x_i's are generated from a log normal distribution $LN(0,1)$, so that $\mu = \exp(0 + 1/2) = 1.648721$ and $\theta = \exp(1.648721) = 5.200326$. We use simulation to draw the sample sizes but then act as if they were a realized set of observations.

The results of the comparison are summarized in Figure **??**. This figure shows that the bootstrap estimator is closer to the true parameter value for almost all sample sizes. The bias of all three estimators decreases as the sample size increases.

```
set.seed(2074)
ns<- 200
est <- function(n){
call_param <- function(i) param(rlnorm(n,0,1))
V <- Vectorize(call_param)(1:ns)
apply(V,1,median)
}
VN=seq(15,100,by=5)
Est <- Vectorize(est)(VN)
matplot(VN,t(Est),type="l", col=2:4, lty=2:4, ylim=exp(exp(1/2))+c(-1,1),
        xlab="sample size (n)", ylab="estimator")
abline(h=exp(exp(1/2)),lty=1, col=1)
legend("topleft", c("raw estimator", "second order correction", "bootstrap"),
       col=2:4,lty=2:4, bty="n")
```

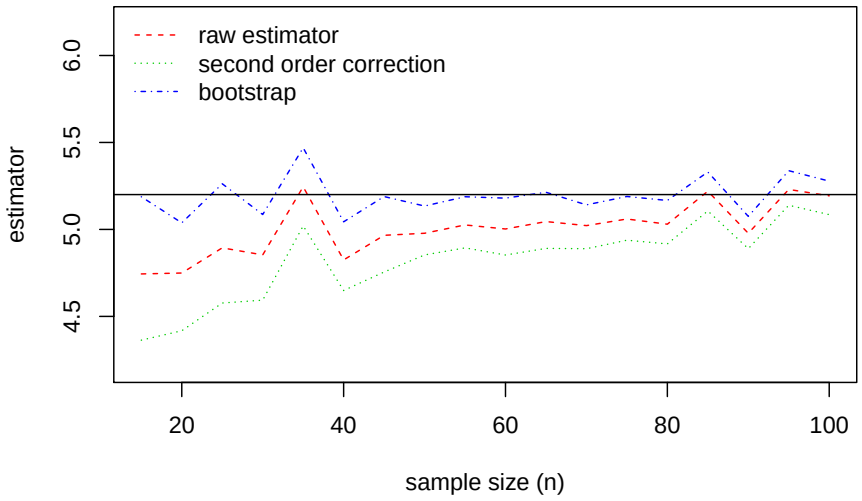


Figure 6.10: Comparision of Estimates. True value of the parameter is given by the solid horizontal line at 5.20.

6.2.3 Confidence Intervals

The bootstrap procedure generates B replicates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ of the estimator $\hat{\theta}$. In Example 6.2.1, we saw how to use standard normal approximations to create a confidence interval for parameters of interest. However, given that a major point is to use bootstrapping to avoid relying on assumptions of approximate normality, it is not surprising that there are alternative confidence intervals available.

For an estimator $\hat{\theta}$, the basic bootstrap confidence interval is

$$\left(2\hat{\theta} - q_U, 2\hat{\theta} - q_L\right), \tag{6.2}$$

where q_L and q_U are lower and upper 2.5% quantiles from the bootstrap sample $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

To see where this comes from, start with the idea that q_L, q_U is a 95% interval for $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. So, for a random $\hat{\theta}_b^*$, there is a 95% chance that $q_L \leq \hat{\theta}_b^* \leq q_U$. Reversing the inequalities and adding $\hat{\theta}$ to each side gives a 95% interval

$$\hat{\theta} - q_U \leq \hat{\theta} - \hat{\theta}_b^* \leq \hat{\theta} - q_L.$$

So, $\left(\hat{\theta} - q_U, \hat{\theta} - q_L\right)$ is an 95% interval for $\hat{\theta} - \hat{\theta}_b^*$. The bootstrap approximation idea says that this is also a 95% interval for $\theta - \hat{\theta}$. Adding $\hat{\theta}$ to each side gives the 95% interval in equation (**??**).

Many alternative bootstrap intervals are available. The easiest to explain is the percentile bootstrap interval which is defined as (q_L, q_U). However, this has the drawback of potentially poor behavior in the tails which can be of concern in some actuarial problems of interest.

Example 6.2.3. Bodily Injury Claims and Risk Measures. To see how the bootstrap confidence intervals work, we return to the bodily injury auto claims considered in Example 6.2.1. Instead of the loss elimination ratio, we consider the 95th percentile $F^{-1}(0.95)$ and a measure defined as

$$TVaR_{0.95)}[X] = \mathrm{E}[X|X > F^{-1}(0.95)].$$

This measure is called the tail value at risk; it is the expected value of X conditional on X exceeding the 95th percentile. In Section 10.2 we will explain how quantiles and the tail value at risk are the two most important examples of so-called risk measures. For now, we will simply think of these as measures that we wish to estimate. For the percentile, we use the nonparametric estimator $F_n^{-1}(0.95)$ defined in Section 4.1.1.3. For the tail value at risk, we use the plug-in principle to define the nonparametric estimator

$$TVaR_{n,0.95}[X] = \frac{\sum_{i=1}^{n} X_i I(X_i > F_n^{-1}(0.95))}{\sum_{i=1}^{n} I(X_i > F_n^{-1}(0.95))} .$$

In this expression, the denominator counts the number of observations that exceed the 95th percentile $F_n^{-1}(0.95)$. The numerator adds up losses for those observations that exceed $F_n^{-1}(0.95)$. Table 6.4 summarizes the estimator for selected fractions.

Table 6.4. Bootstrap Estimates of Quantiles at Selected Fractions

```
# Example from Derrig et al
#BIData <- read.csv("./Data/DerrigResampling.csv", header =T)
BIData$Censored <- 1*(BIData$AmountPaid >= BIData$PolicyLimit)
BIDataUncensored <- subset(BIData, Censored == 0)
#quantile(BIDataUncensored$AmountPaid, 0.8)
#set.seed(2017)
#results <- boot(data=BIDataUncensored$AmountPaid,
#           statistic=function(X,indices) quantile(X[indices],0.80),
#                R=100)
#results

set.seed(2017)
PercentVec <- c(0.50, 0.80, 0.90, 0.95, 0.98)
```

```
OutBoot1 <- matrix(0,5,10)
colnames(OutBoot1) <- c("Fraction","NP Estimate", "Bootstrap Bias", "Bootstrap SD", "Lower Normal
       "Lower Basic 95% CI", "Upper Basic 95% CI",
       "Lower Percentile 95% CI", "Upper  Percentile 95% CI")
  for (i in 1:length(PercentVec)) {
OutBoot1[i,1] <- PercentVec[i]
results <- boot(data=BIDataUncensored$AmountPaid,
                statistic=function(X,indices)
                    quantile(X[indices],PercentVec[i]),
                 R=1000)
if (i==1){bootreal <- results$t}
OutBoot1[i,2] <- results$t0
OutBoot1[i,3] <- mean(results$t)-results$t0
OutBoot1[i,4] <- sd(results$t)
temp <- boot.ci(results, type = c("norm", "basic", "perc"))
OutBoot1[i,5] <- temp$normal[2]
OutBoot1[i,6] <- temp$normal[3]
OutBoot1[i,7] <- temp$basic[4]
OutBoot1[i,8] <- temp$basic[5]
OutBoot1[i,9] <- temp$percent[4]
OutBoot1[i,10] <- temp$percent[5]
}
knitr::kable(OutBoot1,digits=2)
```

Fraction	NP Estimate	Bootstrap Bias	Bootstrap SD	Lower Normal 95% CI	Upper Normal 95% CI	Lo
0.50	6500.00	-126.66	205.29	6224.30	7029.01	
0.80	9078.40	92.14	205.90	8582.70	9389.82	
0.90	11454.00	38.17	462.87	10508.62	12323.04	
0.95	13313.40	47.27	721.87	11851.29	14680.96	
0.98	16758.72	58.44	1272.69	14205.86	19194.70	

For example, when the fraction is 0.50, we see that lower and upper 2.5th quantiles of the bootstrap simulations are $q_L = 6000$ and $q_u = 6703$, respectively. These form the percentile bootstrap confidence interval. With the nonparametric estimator 6500, these yield the lower and upper bounds of the basic confidence interval 6297 and 7000, respectively. Table 6.4 also shows bootstrap estimates of the bias, standard deviation, and a normal confidence interval, concepts introduced in the prior section.

Table 6.5 shows similar calculations for the tail value at risk. In each case, we see that the bootstrap standard deviation increases as the fraction increases. This is a reflection of the fewer observations available to estimate quantiles as the fraction increases, hence greater imprecision. Width of confidence intervals also increase. Interestingly, there does not seem to be the same pattern in the estimates of the bias.

Table 6.5. Bootstrap Estimates of TVaR at Selected Risk Levels

```
CTE.boot <- function(data, indices, RiskLevel){
  resample.data <- data[indices,]
  X <- resample.data$AmountPaid
  cutoff <- quantile(X, RiskLevel)
  CTE <- sum(X*(X > cutoff))/sum(X > cutoff)
  return(CTE)
}
#indices <- 1:nrow(BIDataUncensored)
#CTE.boot(BIDataUncensored, indices, RiskLevel = 0.8)
set.seed(2017)
PercentVec <- c(0.50, 0.80, 0.90, 0.95, 0.98)
OutBoot1 <- matrix(0,5,10)
colnames(OutBoot1) <- c("Fraction","NP Estimate", "Bootstrap Bias", "Bootstrap SD", "Lo
       "Lower Basic 95% CI", "Upper Basic 95% CI",
       "Lower Percentile 95% CI", "Upper  Percentile 95% CI")
  for (i in 1:length(PercentVec)) {
OutBoot1[i,1] <- PercentVec[i]
results <- boot(data=BIDataUncensored, statistic=CTE.boot, R=1000, RiskLevel=PercentVec
OutBoot1[i,2] <- results$t0
OutBoot1[i,3] <- mean(results$t)-results$t0
OutBoot1[i,4] <- sd(results$t)
temp <- boot.ci(results, type = c("norm", "basic", "perc"))
OutBoot1[i,5] <- temp$normal[2]
OutBoot1[i,6] <- temp$normal[3]
OutBoot1[i,7] <- temp$basic[4]
OutBoot1[i,8] <- temp$basic[5]
OutBoot1[i,9] <- temp$percent[4]
OutBoot1[i,10] <- temp$percent[5]
  }
#quantile(results$t,.975, type=6)
#sd(results$t)
#temp
#2*results$t0-quantile(results$t,.025, type=6)
#2*results$t0-quantile(results$t,.975, type=6)
knitr::kable(OutBoot1,digits=2)
```

Fraction	NP Estimate	Bootstrap Bias	Bootstrap SD	Lower Normal 95% CI	Upper Normal 9
0.50	9794.69	-126.05	293.39	9345.71	10
0.80	12454.18	58.79	475.60	11463.23	13
0.90	14720.05	-40.96	704.89	13379.45	16
0.95	17072.43	-7.29	1141.01	14843.37	19
0.98	20140.56	81.22	1623.89	16876.56	23

6.2.4 Parametric Bootstrap

As described in the section on bootstrap, the idea of (nonparametric) bootstrap is to resample, or equivalently, to use the inverse method on the empirical cumulative distribution function F_n. With parametric bootstrap, we draw independent variables from $F_{\widehat{\theta}}$ where the underlying distribution is assume to be in a parametric family $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$. Typically, parameters from this distribution are estimated based on a sample and denoted as $\hat{\theta}$.

Example 6.2.4. Lognormal distribution. Consider again the dataset

```
sample_x <- c(2.46,2.80,3.28,3.86,2.85,3.67,3.37,3.40,5.22,2.55,2.79,4.50,3.37,2.88,1.44,2.56,2.0
```

The classical (nonparametric) bootstrap was based on samples

```
x <- sample(sample_x,replace=TRUE)
```

while for the parametric bootstrap, we have to assume that the distribution of x_i's is from a specific family, for instance a lognormal distribution

```
library(MASS)
fit <- fitdistr(sample_x, dlnorm, list(meanlog = 1, sdlog = 1))
fit
```

```
    meanlog       sdlog
  1.03630697   0.30593440
 (0.06840901) (0.04837027)
```

then we draw from that distribution

```
x <- rlnorm(length(sample_x), meanlog=fit$estimate[1], sdlog=fit$estimate[2])
```

Figure **??** compares the bootstrap distributions of the variation coefficient, one based on the parametric approach and the other based on a parametric approach, assuming a lognormal distribution.

```
set.seed(2074)
CV <- matrix(NA,1e5,2)
for(s in 1:nrow(CV)){
x1 <- sample(sample_x,replace=TRUE)
x2 <- rlnorm(length(sample_x), meanlog=fit$estimate[1], sdlog=fit$estimate[2])
CV[s,] <- c(sd(x1)/mean(x1),sd(x2)/mean(x2))
}
plot(density(CV[,1]),col="red",main="",xlab="Coefficient of Variation", ,lty=1)
lines(density(CV[,2]),col="blue",lty=2)
abline(v=sd(sample_x)/mean(sample_x),lty=3)
legend("topright",c("nonparametric","parametric (LN)"),col=c("red","blue"),lty=1:2,bty="n")
```

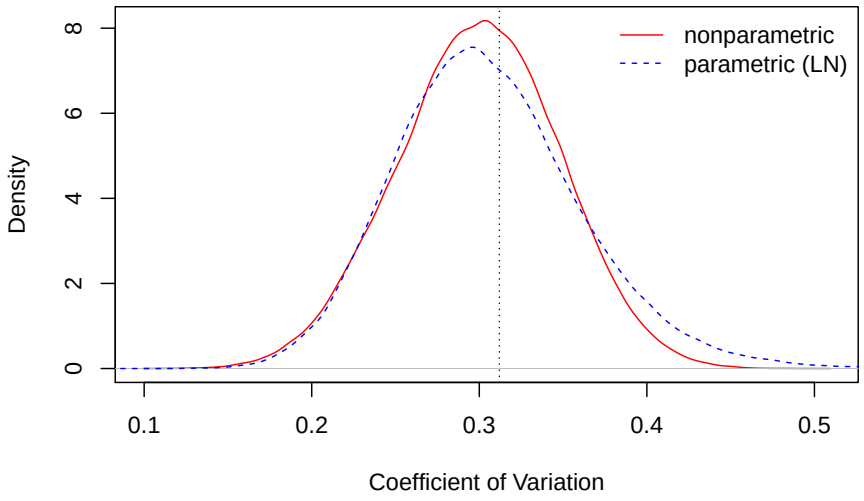


Figure 6.11: Bootstrap Distributions of the Coeffiecient of Variation.

Example 6.2.5. Bootstrapping Censored Observations. The parametric bootstrap draws simulated realizations from a parametric estimate of the distribution function. In the same way, we can draw simulated realizations from estimates of a distribution function. As one example, we might draw from smoothed estimates of a distribution function introduced in Section 4.1.1.4. Another special case, considered here, is to draw an estimate from the Kaplan-Meier estimator introduced in Section 4.3.2.2. In this way, we can handle observations that are censored.

Specifically, return to the bodily injury data in Examples 6.2.1 and 6.2.3 but now we include the 17 claims that were censored by policy limits. In Example 4.3.6, we used this full dataset to estimate the Kaplan-Meier estimator of the survival function introduced in Section 4.3.2.2. Table 6.6 present bootstrap estimates of the quantiles from the Kaplan-Meier survival function estimator. These include the bootstrap precision estimates, bias and standard deviation, as well as the basic 95% confidence interval. xx

Table 6.6. Bootstrap Kaplan-Meier Estimates of Quantiles at Selected Fractions

```
# Example from Derrig et al
library(survival)                    # for Surv(), survfit()
```

```
#BIData <- read.csv("./Data/DerrigResampling.csv", header =T)
#BIData$Censored <- 1*(BIData$AmountPaid >= BIData$PolicyLimit)
BIData$UnCensored <- 1*(BIData$AmountPaid < BIData$PolicyLimit)
## KM estimate
KM0 <- survfit(Surv(AmountPaid, UnCensored) ~ 1,  type="kaplan-meier", data=BIData)


# #summary(KM0)
# plot(KM0, conf.int=FALSE, xlab="x",ylab="Kaplan Meier Survival")
# quantile(KM0, 0.80)$quantile

set.seed(2019)
PercentVec <- c(0.50, 0.80, 0.90, 0.95, 0.98)
OutBoot1 <- matrix(NA,5,6)
colnames(OutBoot1) <- c("Fraction","KM NP Estimate", "Bootstrap Bias", "Bootstrap SD",
       "Lower Basic 95% CI", "Upper Basic 95% CI")
KM.survobj <- Surv(BIData$AmountPaid, BIData$UnCensored)
for (i in 1:length(PercentVec)) {
OutBoot1[i,1] <- PercentVec[i]
results <- bootkm(KM.survobj, q=1-PercentVec[i], B=1000, pr = FALSE)
if (i==1){bootreal <- results}
OutBoot1[i,2] <- quantile(KM0, PercentVec[i])$quantile
OutBoot1[i,3] <- mean(results)-OutBoot1[i,2]
OutBoot1[i,4] <- sd(results)
# temp <- boot.ci(results, type = c("norm",  "basic","perc"))
OutBoot1[i,5] <- 2*OutBoot1[i,2]-quantile(results,.975, type=6)
OutBoot1[i,6] <- 2*OutBoot1[i,2]-quantile(results,.025, type=6)
}
knitr::kable(OutBoot1,digits=2)
```

Fraction	KM NP Estimate	Bootstrap Bias	Bootstrap SD	Lower Basic 95% CI	Upper Basic 95% CI
0.50	6500	13.58	181.23	6093	6923
0.80	9500	173.61	423.41	8445	9949
0.90	12756	20.17	675.86	10812	14012
0.95	18500	Inf	NaN	12500	22300
0.98	25000	Inf	NaN	-Inf	27308

Results in Table 6.6 are consistent with the results for the uncensored subsample in Table 6.4. In Table 6.6, we note the difficulty in estimating quantiles at large fractions due to the censoring. However, for moderate size fractions (0.50, 0.80, and 0.90), the Kaplan-Meier nonparametric (KM NP) estimates of the quantile are consistent with those Table 6.4. The bootstrap standard deviation is smaller at the 0.50 (corresponding to the median) but larger at the 0.80 and 0.90 levels. The censored data analysis summarized in Table 6.6 uses more data than the uncensored subsample analysis in Table 6.4 but also has difficulty extracting information for large quantiles.

6.3 Cross-Validation

In this section, you learn how to:

- Compare and contrast cross-validation to simulation techniques and bootstrap methods.
- Use cross-validation techniques for model selection
- Explain the jackknife method as a special case of cross-validation and calculate jackknife estimates of bias and standard errors

---

Cross-validation, briefly introduced in Section 4.2.4, is a technique based on simulated outcomes and so let's think about its purposes in contrast to other simulation techniques already introduced in this chapter.

- Simulation, or Monte-Carlo, introduced in Section **??**, allow us to compute expected values and other summaries of statistical distributions, such as p-values, readily.
- Bootstrap, and other resampling methods introduced in Section **??**, provide estimators of the precision, or variability, of statistics.
- Cross-validation is important when assessing how accurately a predictive model will perform in practice.

Overlap exists but nonetheless it is helpful to think about these broad goals as associated with each statistical method.

To discuss cross-validation, let us recall from Section 4.2 some of the key ideas of model validation. When assessing, or validating, a model, we look to performance measured on new data, or at least not those that were used to fit the model. A classical approach, described in Section 4.2.3, is to split the sample in two: a subpart (the training dataset) is used to fit the model and another one (the testing dataset) is used to validate. However, a limitation of this approach is that results depend on the split; even though the overall sample is fixed, the split between training and test sub samples varies randomly. A different training sample means that model estimated parameters will differ. Different model parameters and a different test sample means that validation statistics will differ. Two analysts may use the same data and same models yet reach different conclusions about the viability of a model (based on different random splits), a frustrating situation.

6.3.1 k-Fold Cross-Validation

To mitigate this difficulty, it is common to use a cross-validation approach as introduced in Section 4.2.4. The key idea is to replicate the basic test/training approach to model validation but to repeat it many times by averaging over different splits of the data. A key advantage is that the validation statistic

is not tied to a specific parametric (or nonparametric) model - one can use a nonparametric statistic or a statistic that has economic interpretations - and so this can be used to compare models that are not nested (unlike likelihood ratio procedures).

Example 6.3.1. Wisconsin Property Fund. For the 2010 property fund data introduced in Section 1.3, we fit gamma and Pareto distributions to the 1,377 claims data. For details of the related goodness of fit, see Appendix Section 15.4.4. We now consider the Kolmogorov-Smirnov statistic introduced in Section 4.1.2.2. When the entire dataset was fit, the Kolmogorov-Smirnov goodness of fit statistic for the gamma distribution turns out to be 0.0478 and for the Pareto distribution is 0.2639. The lower value for the Pareto distribution indicates that this distribution is a better fit than the gamma.

To see how k-fold cross-validation works, we randomly split the data into $k = 8$ groups, or folds, each having about $1377/8 \approx 172$ observations. Then, we fit gamma and Pareto models to a data set with the first seven folds (about $172 * 7 = 1204$ observations), determined estimated parameters, and then used these fitted models with the held-out data to determine the Kolmogorov-Smirnov statistic. The results appear in Figure **??** where horizontal axis is Fold=1. This process was repeated for the other seven folds. The results summarized in Figure **??** show that the Pareto is a consistently provides a more reliable predictive distribution than the gamma.

```
## Cross - Validation
# Randomly re-order the data - "shuffle it"
n <- nrow(claim_data)
set.seed(12347)
cvdata <- claim_data[sample(n), ]
# Number of folds
k <- 8
cvalvec <- matrix(0,2,k)
for (i in 1:k) {
  indices <- (((i-1) * round((1/k)*nrow(cvdata))) + 1):((i*round((1/k) * nrow(cvdata))))
# Pareto
  fit.pareto <- vglm(Claim ~ 1, paretoII, loc = 0, data = cvdata[-indices,])
  ksResultPareto <- ks.test(cvdata[indices,]$Claim, "pparetoII", loc = 0, shape = exp(coef(fit.pa
        scale = exp(coef(fit.pareto)[1]))
  cvalvec[1,i] <- ksResultPareto$statistic
# Gamma
  fit.gamma <- glm(Claim ~ 1, data = cvdata[-indices,], family = Gamma(link = log))
  gamma_theta <- exp(coef(fit.gamma)) * gamma.dispersion(fit.gamma)
  alpha <- 1 / gamma.dispersion(fit.gamma)
  ksResultGamma <- ks.test(cvdata[indices,]$Claim, "pgamma", shape = alpha, scale = gamma_theta)
  cvalvec[2,i] <- ksResultGamma$statistic
}
# Plot the statistics
```

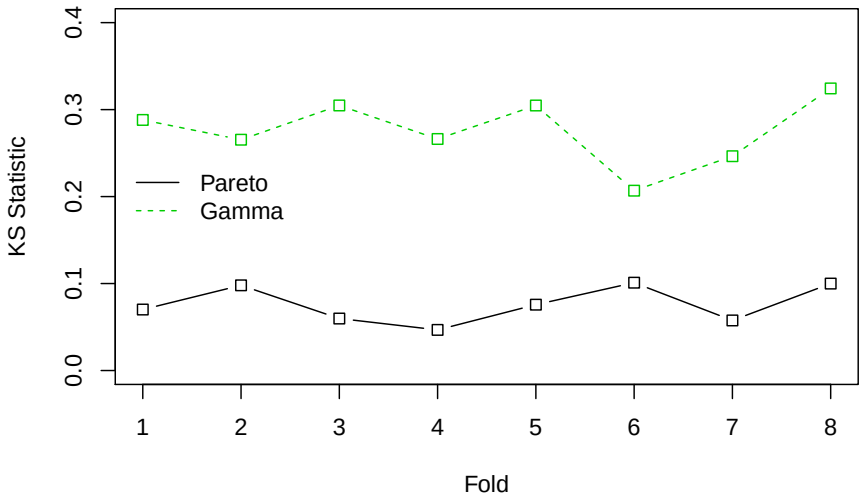


Figure 6.12: Cross Validated Kolmogorov-Smirnov (KS) Statistics for the Property Fund Claims Data. The solid black line is for the Pareto distribution, the green dashed line is for the gamma distribution. The KS statistic measures the largest deviation between the fitted distribution and the empirical distribution for each of 8 groups, or folds, of randomly selected data.

```
matplot(1:k,t(cvalvec),type="b", col=c(1,3), lty=1:2, ylim=c(0,0.4), pch = 0, xlab="Fo
legend("left", c("Pareto", "Gamma"), col=c(1,3),lty=1:2, bty="n")
```

```
KScv <- rowSums(cvalvec)/k
```

6.3.2 Leave-One-Out Cross-Validation

A special case where $k = n$ is known as leave-one-out cross validation. This case is historically prominent and is closely related to **jackknife statistics**, a precursor of the bootstrap technique.

Even though we present it as a special case of cross-validation, it is helpful to given an explicit definition. Consider a generic statistic $\widehat{\theta} = t(\boldsymbol{x})$ that is an estimator for a parameter of interest θ. The idea of jackknife is to compute n values $\widehat{\theta}_{-i} = t(\boldsymbol{x}_{-i})$, where $\boldsymbol{x}_{-i}$ is the subsample of $\boldsymbol{x}$ with the i-th value

removed. The average of these values is denoted a

$$\bar{\widehat{\theta}}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{-i}.$$

These values can be used to create estimates of the bias of the statistic $\widehat{\theta}$

$$Bias_{jack} = (n-1)\left(\bar{\widehat{\theta}}_{(\cdot)} - \widehat{\theta}\right) \tag{6.3}$$

as well as a standard deviation

$$s_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{-i} - \bar{\widehat{\theta}}_{(\cdot)}\right)^2}. \tag{6.4}$$

Example 6.3.2. Coefficient of Variation. To illustrate, consider a small fictitious sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ with realizations

```
sample_x <- c(2.46,2.80,3.28,3.86,2.85,3.67,3.37,3.40,5.22,2.55,2.79,4.50,3.37,2.88,1.44,2.56,2.0
```

Suppose that we are interested in the coefficient of variation $\theta = CV = \sqrt{\text{Var } X}/\text{E } X$.

With this dataset, the estimator of the coefficient of variation turns out to be 0.31196. But how reliable is it? To answer this question, we can compute the jackknife estimates of bias and its standard deviation. The following code shows that the jackknife estimator of the bias is $Bias_{jack}$ = -0.00627 and the jackknife standard deviation is s_{jack} = 0.01293.

```
CVar <- function(x) sqrt(var(x))/mean(x)
JackCVar <- function(i) sqrt(var(sample_x[-i]))/mean(sample_x[-i])
JackTheta <- Vectorize(JackCVar)(1:length(sample_x))
BiasJack <- (length(sample_x)-1)*(mean(JackTheta) - CVar(sample_x))
sd(JackTheta)
```

Example 6.3.3. Bodily Injury Claims and Loss Elimination Ratios. In Example 6.2.1, we showed how to compute bootstrap estimates of the bias and standard deviation for the loss elimination ratio using the Example 4.1.11 bodily injury claims data. We follow up now by providing comparable quantities using jackknife statistics.

Table 6.7 summarizes the results of the jackknife estimation. It shows that jackknife estimates of the bias and standard deviation of the loss elimination ratio E $\min(X, d)/\text{E } X$ are largely consistent with the bootstrap methodology. Moreover, one can use the standard deviations to construct normal based confidence intervals, centered around a bias-corrected estimator. For example, at

$d = 14000$, we saw in Example 4.1.11 that the nonparametric estimate of LER is 0.97678. This has an estimated bias of 0.00010, resulting in the (jackknife) bias-corrected estimator 0.97688. The 95% confidence intervals are produced by creating an interval of twice the length of 1.96 jackknife standard deviations, centered about the bias-corrected estimator (1.96 is the approximate 97.5th quantile of the standard normal distribution).

Table 6.7. Jackkife Estimates of LER at Selected Deductibles

d	NP Estimate	Bootstrap Bias	Bootstrap SD	Jackknife Bias	Jackknife SD	Lower Jack
4000	0.54113	-0.00012	0.01228	0.00031	0.00061	
5000	0.64960	-0.00038	0.01400	0.00033	0.00068	
10500	0.93563	0.00018	0.01065	0.00019	0.00053	
11500	0.95281	-0.00016	0.00918	0.00016	0.00047	
14000	0.97678	0.00018	0.00701	0.00010	0.00034	
18500	0.99382	0.00006	0.00342	0.00003	0.00017	

Discussion. One of the many interesting things about the leave-one-out special case is the ability to replicate estimates exactly. That is, when the size of the fold is only one, then there is no additional uncertainty induced by the cross-validation. This means that analysts can exactly replicate work of one another, an important consideration.

Jackknife statistics were developed to understand precision of estimators, producing estimators of bias and standard deviation in equations (**??**) and (**??**). This crosses into goals that we have associated with bootstrap techniques, not cross-validation methods. This demonstrates how statistical techniques can be used to achieve different goals.

6.3.3 Cross-Validation and Bootstrap

Arthur. I'm not sure about the point of this section. Is it supposed to point out connections, as per Efron (1982) Chapter 7?

One can mix the idea of cross-validation and bootstrap :

- create a bootstrap sample with re-sampling (with replacement) n indices in $\{1, \cdots, n\}$: let $\mathcal{I}_b$ denote that set of indices. That will be our training sample
- the validation sample will observations that were not

```
indices <- sample(nrow(cars))
List_indices = List_indices_C = list()
for(i in 1:100){
  List_indices[[i]] <- sample(indices, size=length(indices), replace= TRUE)
```

```
  List_indices_C[[i]] <- (1:nrow(cars))[-sort(unique(List_indices[[i]]))]}
List_indices[[1]]
```

```
FALSE  [1] 47 47 21 43 38  3 36 29 50 10  1 32 49 10 45 48 16 32 21 18 21 40  9
FALSE [24] 36 32 33 18 29 12 40 22  9 10 43  6 40 34 19  2 39  4 23 43  7 27 24
FALSE [47] 25 28 33 32
```

```
List_indices_C[[1]]
```

```
FALSE  [1]  5  8 11 13 14 15 17 20 26 30 31 35 37 41 42 44 46
```

```
epsilon <- list()
mean_RSS <- rep(NA,100)
for(i in 1:length(List_indices)){
  model <- lm(dist~speed,data=cars[List_indices[[i]],])
  epsilon[[i]] <- cars[List_indices_C[[i]],"dist"]-predict(model, newdata=cars[List_indices_C[[i]
  mean_RSS[i] <- mean( (epsilon[[i]])^2 )
}
nrow(cars)*mean(mean_RSS)
```

```
FALSE [1] 12083.28
```

6.4 Importance Sampling

We have introduced Monte Carlo techniques using the inversion technique : to generate a random variable X with distribution F, apply F^{-1} to calls of a random generator (uniform on the unit interval). What if we what to draw according to X, conditional on $X \in [a, b]$?

One can use an **accept-reject** mechanism : draw x from distribution F

- if $x \in [a, b]$: keep it ("accept")
- if $x \notin [a, b]$: draw another one ("reject")

Observe that from n values initially generated, we keep here only $[F(b)-F(a)]\cdot n$ draws, on average.

```
#install.packages('gifski')
#if (packageVersion('knitr') < '1.20.14') {
#  remotes::install_github('yihui/knitr')
#}
pic_ani = function(){
  u=seq(0,5,by=.01)
  plot(u,pnorm(u,2.5,1),col="white",ylab="",xlab="")
  rect(-1,-1,6,2,col=rgb(1,0,0,.2),border=NA)
  rect(2,pnorm(2,2.5,1),4,pnorm(4,2.5,1),col="white",border=NA)
  lines(u,pnorm(u,2.5,1),lwd=2)
```