Loss Data Analytics

An open text authored by the Actuarial Community

Contents

Preface

Date: 24 January 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.

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.

IABA
INTERNATIONAL ASSOCIATION OF BLACK ACTUARIES

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
- **Jan Beirlant**, KU Leuven
- **Carolina Castro** - University of Buenos Aires

- **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 doctoral student at the University of Wisconsin-Madison.

- **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.

- **Shyamalkumar Nariankadu** - University of Iowa

- **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** - University of Wisconsin - Madison
- **Jianxi Su**, Purdue University
- **Tim Verdonck**, KU Leuven
- **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
- 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 the discussion features in the online text, 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 1.1 begins with a discussion of why the use of data is important in the insurance industry. Section 1.2 gives a general overview of the purposes of analyzing insurance data which is reinforced in the Section 1.3 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 Activites

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 a business. The seller, an insurance company, is also known as an insurer. Even insurance companies need insurance; this is known as reinsuranceInsurance purchased by an 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 insuranceCasualty 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-lifeNon-life insurance is any type of insurance where payments are not based on the death (or survivorship) of a named insured. Examples include automobile, homeowners, and so on. Also known as property and casualty or general insurance. 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 Insurance Information Institute (2016) 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, (Insurance Information Institute, 2016). 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 *analytics* Analytics 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 claims* An 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 *term*The 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 1.1 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 1.2; 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

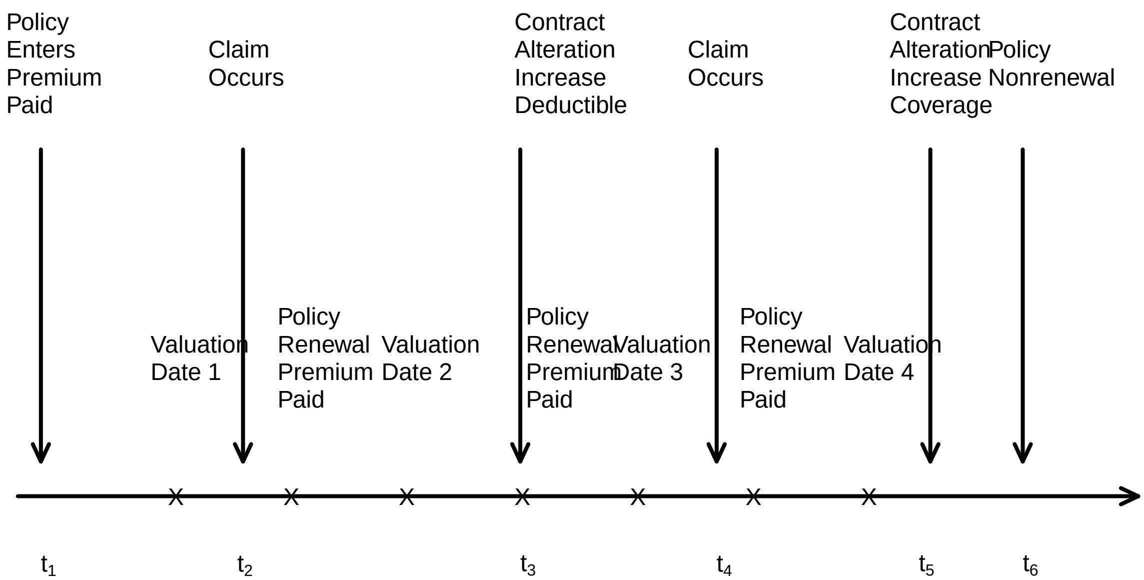


Figure 1.1: Timeline of a Typical Insurance Policy. Arrows mark the occurrences of random events. Each x marks the time of scheduled events that are typically non-random.

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 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 *underwriting*Underwriting 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., Earnix (2013)) 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* (Bailey and LeRoy, 1960), 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 dividendsA dividend is the refund of a portion of the premium paid by the insured from insurer surplus. 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 adjustment.*Claims adjustment is the process of determining coverage, legal liability, and settling claims.

Insurance managers sometimes use the phrase *claims leakage*Claims leakage represents 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., Gorman and Swenson (2013). 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 1.2.4.

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 3 and 10 describe common modifications including coinsurance,Coinsurance 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. deductibles,A 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 coinsurance). and policy upper limits.A policy limit is the maximum value covered by a policy.

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*;A loss reserve is an estimate of liability indicating the amount the insurer expects to pay for claims that have not yet been realized. This includes losses incurred but not yet reported (IBNR) and those claims that have been reported claims that haven't been paid (known by the acronym RBNS for reported but not settled). in some jurisdictions, reserves are also known as *technical provisions*. We saw in Figure 1.1 several times at which a company summarizes its financial position; these times are known as *valuation dates*.A valuation date is the date at which a company summarizes its financial position, typically quarterly or annually. 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, Bowers et al. (1986) or Dickson et al. (2013) 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. *Indemnification*Indemnification 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 1.1 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=FALSE)
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 premium*Pure 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 1.2 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 1.2 provides further information

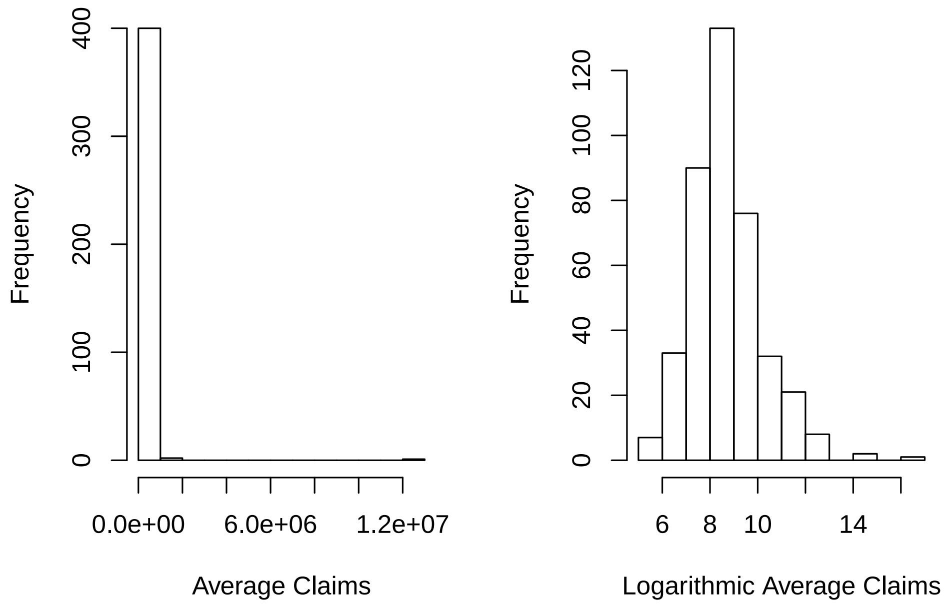


Figure 1.2: Distribution of Positive Average Severities

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("."), stringsAsFactors=FALSE
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 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* A variable that is useful to explain or predict a loss, typically used to classify risks for rating purposes. 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 1.3 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.

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("."), stringsAsFactors=FALSE
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 Policyholders")
Table1In
```

Table 1.3 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

[1] Note that the average severity in Table 1.3 differs from that reported in Table 1.2. 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.

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 1.4 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 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("."), stringsAsFactors=FALSE
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 Frees et al. (2016). 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 1.5 relates the claims outcomes to these categorical variables. Table 1.5 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 1.6 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 1.5 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$AC10==1)+ 4*(Insample$AC15==1)
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 1.2; 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 1.5 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 1.6 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 7 on *premium calculation fundamentals*. Selecting appropriate risks is introduced in Chapter 8 on *risk classification*.

Renewing Insurance

Although property insurance is typically a one-year contract, Table 1.3 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 1.5 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 9 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 10.

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 11 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 NAIC Glossary (2018). 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 claims are affected by both the frequency of insured events and the severity of the insured event. Decomposing aggregate claims into these two components, which each warrant significant attention, is essential for analysis and pricing. This chapter discusses frequency distributions, measures, and parameter estimation techniques.

2.1 Frequency Distributions

2.1.1 How Frequency Augments Severity Information

Basic Terminology

We use **claim** to denote the indemnification upon the occurrence of an insured event. Some authors use claim and loss interchangeably, while others think of loss as the amount suffered by the insured whereas claim is the amount paid by the insurer. **Frequency** 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 within a policy term. **Severity** 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 1 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.
- **Behavioral** - 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 characteristics, 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.
  - 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**. Many insurers keep 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 makes it natural for insurers to model 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.

2.2 Basic Frequency 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 integral values. Hence the distributions below are all discrete distributions supported on the set of non-negative integers ($\mathbb{Z}^+$).

2.2.1 Foundations

Since N is a discrete random variable taking values in $\mathbb{Z}^+$, the most natural full description of its distribution is through the specification of the probabilities with which it assumes each of the non-negative integral values. This leads us to the concept of the **probability mass function** (*pmf*) 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 function** 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 function** 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 **mean** 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 **moments*8 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 $\Pr(N \in \mathbb{Z}^+) = 1$; also, note that $\mu_N = \mu'_N(1)$.

Another basic aspect of a distribution is its dispersion, and of the various measures of dispersion studied in the literature, the **standard deviation** is the most popular. Towards defining it, we first define the **variance** of N, denoted by 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 Var N, by its definition as the average squared deviation from the mean, is non-negative; Var N is denoted by σ_N^2. Note that these two measures take values in $[0, \infty)$.

[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.

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 the **moment generating function** (mgf) 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 (Billingsley, 2008) (pages 285-6), encapsulates the reason for its name of moment generating functions.

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 moment of N are finite, i.e.

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

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

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

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*), and is a useful tool for random variables taking values in $\mathbb{Z}^+$. For a random variable N, by $P_N(\cdot)$ we denote its *pgf* and we define it as follows:

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

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 2.1. The following result for *pgf* is an analog of Theorem 2.1, 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 moment 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:

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

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 negative binomial and the Poisson 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 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 various 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 hulls 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 classes.
  - In insurance data, we may observe either a marginal or inordinate number of zeros, *i.e.* 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 outcomes. 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 distribution 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 [2]

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

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:

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

[2] 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 $p_N(\cdot)$.

Note that the above expression for the *pgf* confirms the fact that the binomial distribution is the m-convolution 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 $N_1, \dots, N_m$, where N_i are *iid* Bernoulli variates. Now using the moments of Bernoulli and linearity of the expectation, we see that

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

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 distribution (named after the French polymath Sim'eon 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. Relatedly, 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

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

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:

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

From the above, we derive its *mgf* as follows:

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

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 2.2, we see that

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

This, in particular, implies that

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

Note that interestingly for the Poisson distribution Var $N =$ E N.

Negative binomial Distribution

The third important count distribution is the negative binomial distribution. 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. In the following we will allow the parameter r to be any positive real, and unfortunately when r in not an integer the above random experiment would not be applicable. To then motivate the distribution more generally, and in the process explain its name, we recall the binomial series, *i.e.*

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

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

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

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 $\beta >= 0$, 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

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

The above implies that the *mgf* is given by

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

We derive its moments using Theorem 2.1 as follows:

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

We note that when $\beta > 0$, we have Var $N >$ E N. In other words, this distribution is **overdispersed** (relative to the Poisson); similarly, when $q > 0$ the binomial distribution is said to be **underdispersed** (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[3]

2.3 The (a, b, 0) Class

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.

To begin with let $a < 0$. In this case as $(a + b/k) \to a < 0$ as $k \to \infty$, and the ratio on the left is non-negative, 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

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

[3]For the theoretical basis underlying the above argument, see (Billingsley, 2008).

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.

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 a equivalent parametrization is $(\beta/(1 + \beta), (r - 1)\beta/(1 + \beta))$, for $r > 0$ and $0 < \beta$, we see from above that such distributions are negative binomial distributions.

From the above development we see that not only does the recurrence (2.6) 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, (2.6) 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 distribution 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 $\beta = 3$. Finally, since the mean of a negative binomial is $r\beta$ we have the mean of the given distribution equals 9.

2.4 Estimating Frequency Distributions

2.4.1 Parameter estimation

In Section 2.2 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 (*mle*s).

Let us suppose that we observe the *iid* random variables $X_1, X_2, \ldots, X_n$ from a distribution with *pmf* p_θ, where θ is an unknown value in $\Theta \subseteq \mathbb{R}^d$. 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 $x_1, \ldots, x_n$; 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*) for θ is any maximizer of the likelihood; in a sense the *mle* chooses the parameter value 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 2.1. As shown in the plot, the maximum value of the likelihood equals $4/27$ and is attained at $q = 1/3$, and hence the *mle* for q is $1/3$ for the given sample. In this case one can resort to algebra to show that

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

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 extrema. It is customary to equivalently maximize the logarithm of the likelihood[4] $L(\cdot)$, denoted by $l(\cdot)$, and look at the set of zeros of its first derivative[5] $l'(\cdot)$. 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 *mle*.

[4]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 $l(\cdot)$.

[5]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 $L'(\cdot)$.

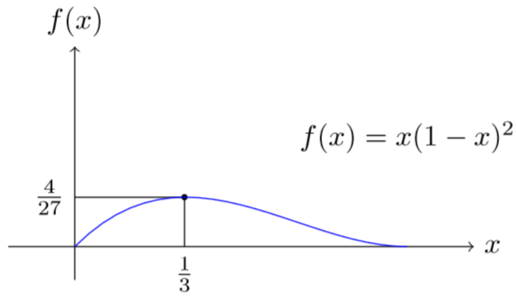


Figure 2.1: Likelihood of a $(0, 1, 0)$ 3-sample from Bernoulli

2.4.2 Frequency Distributions MLE

In the following, we derive the *mle* for the three members of the $(a, b, 0)$ class. We begin by summarizing the discussion above. In the setting of observing *iid* random variables $X_1, X_2, \ldots, X_n$ from a distribution with *pmf* p_θ, where θ is 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 maximum likelihood estimator (*mle*) of θ, denoted as $\hat{\theta}_{\text{MLE}}$ is a function which maps the observations to an element of the set of maximizers of $L(\cdot)$, namely

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

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 *mle* 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 is an information loss-less transformation of the data. 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

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

Since $l'' < 0$ if $\sum_{i=1}^{n} x_i > 0$, the maximum is attained at the sample mean. In the contrary, the maximum is attained at the least possible parameter value, that is the *mle* equals zero. Hence, we have

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

Note that the sample mean can be computed also as

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

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) 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 integral 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 (2.7). 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 (Olkin et al., 1981) 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 (2.8). In Figure 2.2 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 2.2, $\hat{m}_{\text{MLE}}$ 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 2.2 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 $\hat{\lambda} = \overline{x}$.
- *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 be available in closed form. A difference between them is that unlike the binomial parameter m which takes positive integral 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

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

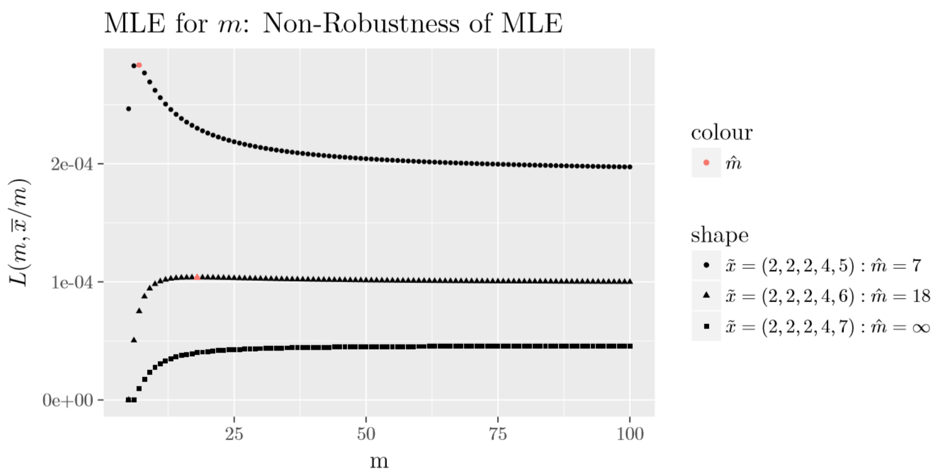


Figure 2.2: Plot of $L(m, \overline{x}/m)$ for binomial distribution

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 (Levin et al., 1977) it is show 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 any 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, the inner sum equals zero if $x_i = 0$. The *mle* for r is a zero of the last expression, and hence a root finding algorithm can be used 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. 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 MLE; else increment k by 1 and repeat *Step ii.*

For example, we simulated a 5 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

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

leads to the starting value 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 2.3 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 2.3, 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 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

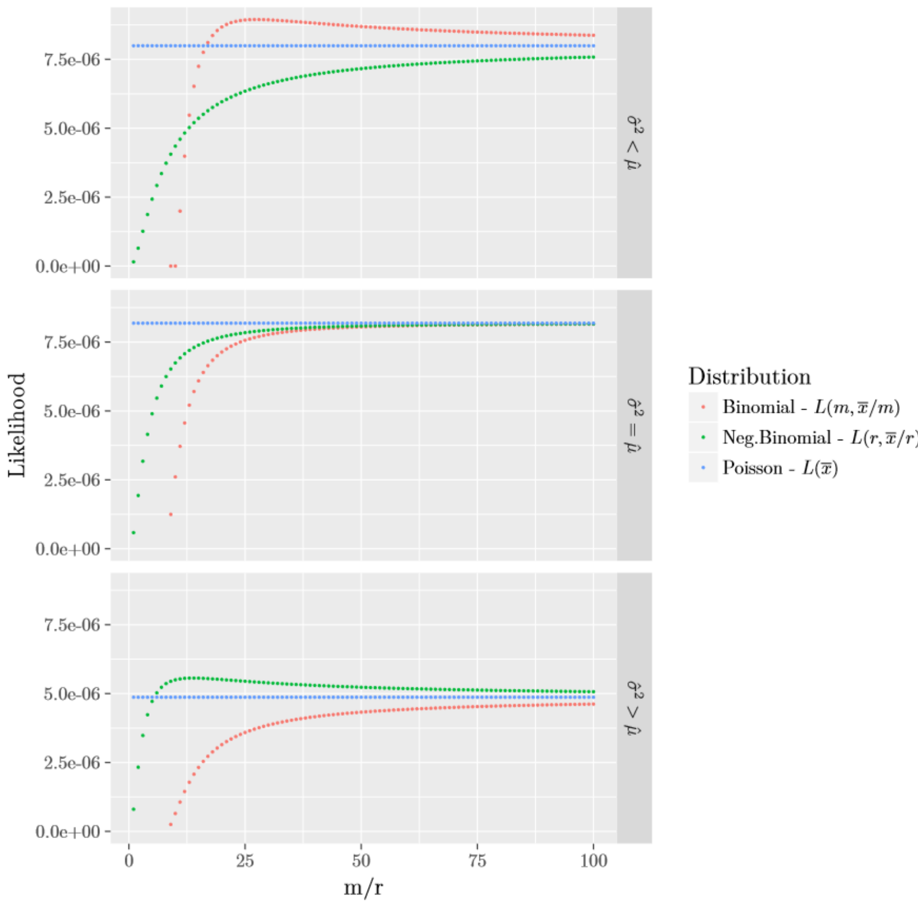


Figure 2.3: Plot of $(a, b, 0)$ Partially Maximized Likelihoods

Table 2.1 : Three Samples of Size 5

2.5 Other Frequency Distributions

In the above 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 variables 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 analysis** - it is an important topic of actuarial interest that we will not pursue in this book; see (Frees, 2009a).

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 homogeneous 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.

In the 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 distribution satisfies the recurrence (2.6), maintaining relative probabilities of non-zero counts implies that recurrence (2.6) 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 super-class 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 (2.10). 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 values for a and b than the $(a, b, 0)$ class. Recall from Section 2.3 that in the case of $a < 0$ we did not use the fact that the recurrence (2.6) 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 2.3, 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 truncation 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 $\{p_k\}_{k\geq 0}$, denoted by $\{p_k^M\}_{k\geq 0}$, can be written as a convex combination of the degenerate distribution 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	$p_2 = p_1\left(\frac{\lambda}{2}\right) = 0.27067$	$p_2^T = p_1^T\left(\frac{\lambda}{2}\right) = 0.313035$	$p_2^M = 0.125214$
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 $k \leq 3$

2.6 Mixture Distributions

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, then the distribution of a randomly chosen observation from the population, denoted by F, is given by

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

The above expression can be seen as a direct application of Bayes theorem. As an example, consider a population of drivers split broadly into two sub-groups, those with less than 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

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

An alternate definition of a mixture distribution 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 (2.11)[6].

In (2.11) 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 functionals of the distribution function. We note that this is not true for

[6] 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.

central moments like the variance, and conditional measures like the hazard rate function. In the case of variance it is easily seen as

$$\mathrm{E}N_I = \mathrm{E}\mathrm{E}N_I|I + \mathrm{E}\mathrm{E}N_I|I = \sum_{i=1}^{k} \alpha_i \mathrm{E}N_i + \mathrm{E}\mathrm{E}N_I|I,$$

and hence is not a convex function of the variances unless the group means are all equal.

Example 2.6.1. SOA 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 development above, 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 required probability can 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

$$F(x) = \mathbf{R}G_{\tilde{\vartheta}}(x), \quad \forall x \in \mathbb{R},$$

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:

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

It is worth mention that by considering mixtures of a parametric class of distributions we increase the richness of the class, resulting in the mixture class being able to cater well to more applications that the parametric class we started with. In the above case, this is seen as we have observed earlier that in a sense the Poisson distributions are on the boundary of negative binomial distributions and by mixing Poisson we get the interior distributions as well. 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 $N|\Lambda \sim \text{Poisson}(\Lambda)$ 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

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

2.7 Goodness of Fit

In the above we have discussed three basic frequency distributions, along with their ways to enhance the reach of these classes 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 *mle* estimate will converge to a value such that the fitted distribution will be a certain *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 fit* of the fitted distribution.

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 statistic test of the hypothesis that the underlying distribution is indeed Poisson. The *Pearson's chi-square statistic* is a goodness of fit statistical test that can be used for this purpose. To explain this statistic let use 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 distribution 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.

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. 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 w.r.t. 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 which uses calculus. 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*, its *mgf* and 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 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 MLEs 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 math operations needed to compute the n probabilities $p_0 \ldots p_{n-1}$ using the recurrence relationship. For the negative binomial distribution with non-integral r, count the number of such operations using the brute force approach. What do you observe?

Exercises with a Practical Focus:

Exercise 2.11. SOA Exam Question. You are given:

1. p_k denotes the probability that the number of claims equals k for $k = 0,1,2,\ldots$
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 2.12. SOA 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)

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 R Code for Plots in this Chapter

Code for Figure 2.3:

Show R Code

```
likbinm<-function(m){
  prod((dbinom(x,m,mean(x)/m)))
}
liknbinm<-function(r){
  prod(dnbinom(x,r,1-mean(x)/(mean(x)+r)))
}
x<-c(2,5,6,8,9)+2;
n<-(9:100);
r<-(1:100);
ll<-unlist(lapply(n,likbinm));
n[ll==max(ll[!is.na(ll)])]
y<-cbind(n,ll);
z<-cbind(rep("$\\hat{\\sigma}^2<\\hat{\\mu}$",length(n)),rep("Binomial - $L(m,\\overline{x}/m)$",length(
ll<-unlist(lapply(r,liknbinm));
ll[is.na(ll)]=0;
r[ll==max(ll[!is.na(ll)])];
y<-rbind(y,cbind(r,ll));
z<-rbind(z,cbind(rep("$\\hat{\\sigma}^2<\\hat{\\mu}$",length(r)),rep("Neg.binomial - $L(r,\\overline{x}/
y<-rbind(y,cbind(r,rep(prod(dpois(x,mean(x))),length(r))));
z<-rbind(z,cbind(rep("$\\hat{\\sigma}^2<\\hat{\\mu}$",length(r)),rep("Poisson - $L(\\overline{x})$",leng
x<-c(2,5,6,8,9);
ll<-unlist(lapply(n,likbinm));
n[ll==max(ll[!is.na(ll)])]
y<-rbind(y,cbind(n,ll));
z<-rbind(z,cbind(rep("$\\hat{\\sigma}^2=\\hat{\\mu}$",length(n)),rep("Binomial - $L(m,\\overline{x}/m)$"
ll<-unlist(lapply(r,liknbinm));
ll[is.na(ll)]=0;
r[ll==max(ll[!is.na(ll)])];
y<-rbind(y,cbind(r,ll));
z<-rbind(z,cbind(rep("$\\hat{\\sigma}^2=\\hat{\\mu}$",length(r)),rep("Neg.binomial - $L(r,\\overline{x}/
y<-rbind(y,cbind(r,rep(prod(dpois(x,mean(x))),length(r))));
z<-rbind(z,cbind(rep("$\\hat{\\sigma}^2=\\hat{\\mu}$",length(r)),rep("Poisson - $L(\\overline{x})$",leng
x<-c(2,3,6,8,9);
ll<-unlist(lapply(n,likbinm));
n[ll==max(ll[!is.na(ll)])]
y<-rbind(y,cbind(n,ll));
z<-rbind(z,cbind(rep("$\\hat{\\sigma}^2>\\hat{\\mu}$",length(n)),rep("Binomial - $L(m,\\overline{x}/m)$"
ll<-unlist(lapply(r,liknbinm));
ll[is.na(ll)]=0;
r[ll==max(ll[!is.na(ll)])];
y<-rbind(y,cbind(r,ll));
z<-rbind(z,cbind(rep("$\\hat{\\sigma}^2>\\hat{\\mu}$",length(r)),rep("Neg.binomial - $L(r,\\overline{x}/
y<-rbind(y,cbind(r,rep(prod(dpois(x,mean(x))),length(r))));
z<-rbind(z,cbind(rep("$\\hat{\\sigma}^2>\\hat{\\mu}$",length(r)),rep("Poisson - $L(\\overline{x})$",leng
colnames(y)<-c("x","lik");
colnames(z)<-c("dataset","Distribution");
dy<-cbind(data.frame(y),data.frame(z));

library(tikzDevice);
library(ggplot2);
options(tikzMetricPackages = c("\\usepackage[utf8]{inputenc}","\\usepackage[T1]{fontenc}", "\\usetikzlib
                      "\\usepackage{amssymb}","\\usepackage{amsmath}","\\usepackage[active]{preview}"))
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(dataset~.)+
```

```
  labs(x="m/r",y="Likelihood",title="");
dev.off();
```

Code for Figure 2.2:

Show R Code

```
likm<-function(m){
  prod((dbinom(x,m,mean(x)/m)))
}
x<-c(2,2,2,4,5);
n<-(5:100);
ll<-unlist(lapply(n,likm));
n[ll==max(ll)]
y<-cbind(n,ll);
x<-c(2,2,2,4,6);
ll<-unlist(lapply(n,likm));
n[ll==max(ll)]
y<-cbind(y,ll);
x<-c(2,2,2,4,7);
ll<-unlist(lapply(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}", "\\usetikzlibr
                     "\\usepackage{amssymb}","\\usepackage{amsmath}","\\usepackage[active]{preview}"))
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="$\\tilde{x}=(2,2,2,4,5):\\hat{m}=7$"),size=0.75
 geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.6..),shape="$\\tilde{x}=(2,2,2,4,6):\\hat{m}=18$"),size=0.7
 geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.7..),shape="$\\tilde{x}=(2,2,2,4,7):\\hat{m}=\\infty$"),siz
 geom_point(aes(x=c(7),y=dy$X..tilde.x...2.2.2.4.5..[3],colour="$\\hat{m}$",shape="$\\tilde{x}=(2,2,2,4
 geom_point(aes(x=c(18),y=dy$X..tilde.x...2.2.2.4.6..[14],colour="$\\hat{m}$",shape="$\\tilde{x}=(2,2,2
  labs(x="m",y="$L(m,\\overline{x}/m)$",title="MLE for $m$: Non-Robustness of MLE ");
dev.off();
```

2.10 Further Resources and Contributors

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.

Here are a few reference cited in the chapter.

Chapter 3

Modeling Loss Severity

Chapter Preview. The traditional loss distribution approach to modeling aggregate losses 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 2 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 distribution 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 insurance (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 2 will be combined with the ideas from this chapter to describe the aggregate losses over the whole portfolio in Chapter 5.

3.1 Basic Distributional Quantities

In this section we define the basic distributional quantities: moments, percentiles and generating functions.

3.1.1 Moments

Let X be a continuous random variable with probability density function $f_X(x)$. The k-th raw moment of X, denoted by μ_k', is the expected value 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.

The k-th central moment 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 variance of X, denoted by σ^2. The square root of the variance is the standard deviation σ. A 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 skewness 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 kurtosis. 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.

Example 3.1.1. SOA Exam Question. Assume that the *rv* X has a gamma distribution with mean 8 and skewness 1. Find the variance of X.

Show Example Solution

Solution. The probability density function of X is given by

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

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

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

Given $\Gamma\left(r+1\right) = r\Gamma\left(r\right)$ and $\Gamma\left(1\right) = 1$, then $\mu_1' = E\left(X\right) = \alpha\theta$, $\mu_2' = E\left(X^2\right) = \left(\alpha+1\right)\alpha\theta^2$, $\mu_3' = E\left(X^3\right) = \left(\alpha+2\right)\left(\alpha+1\right)\alpha\theta^3$, and $\mathrm{Var}\left(X\right) = (\alpha+1)\alpha\theta^2 - (\alpha\theta)^2 = \alpha\theta^2$.

$$\text{Skewness} = \frac{E\left[\left(X-\mu_1'\right)^3\right]}{\mathrm{Var}\left(X\right)^{3/2}} = \frac{\mu_3' - 3\mu_2'\mu_1' + 2{\mu_1'}^3}{\mathrm{Var}\left(X\right)^{3/2}} = \frac{\left(\alpha+2\right)\left(\alpha+1\right)\alpha\theta^3 - 3\left(\alpha+1\right)\alpha^2\theta^3 + 2\alpha^3\theta^3}{\left(\alpha\theta^2\right)^{3/2}} = \frac{2}{\alpha^{1/2}} = 1$$

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

3.1.2 Quantiles

Percentiles can also be used to describe the characteristics of the distribution of X. The 100p*th* percentile of the distribution of X, denoted by π_p, is the value of X which satisfies

$$F_X\left(\pi_p-\right) \leq p \leq F_X\left(\pi_p\right),$$

for $0 \leq p \leq 1$ where π_p- refers to the value of X as it increases approaching π_p from the left or from below.

The 50-th percentile or the middle point of the distribution, $\pi_{0.5}$, is the median. Unlike discrete random variables, percentiles of continuous variables are distinct.

Example 3.1.1. SOA Exam Question. Let X be a continuous random variable with density function $f_X\left(x\right) = \theta e^{-\theta x}$, 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 $F_X\left(x\right) = 1 - e^{-\theta x}$. So, $F_X\left(\pi_{0.5}\right) = 1 - e^{-\theta\pi_{0.5}} = 0.5$. As, $\pi_{0.5} = \frac{1}{3}$, we have $F_X\left(\frac{1}{3}\right) = 1 - e^{-\theta/3} = 0.5$ and $\theta = 3\ln 2$.

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 is a real function whose k-th derivative at zero is equal to the k-th raw moment of X. The moment generating function is given by

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

for all t for which the expected value exists.

Example 3.1.3. SOA 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.

Show Example Solution

Solution.

With $X \sim Exp\left(\frac{1}{b}\right)$,

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

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. SOA Exam Question. Let $X_1, \ldots, X_n$ be independent $\mathrm{Ga}(\alpha_i, \theta)$ random variables. Find the distribution of $S = \sum_{i=1}^n X_i$.

Show Example Solution

Solution.

The moment generating function of S is

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

using independence we get

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

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

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

indicating that $S \sim Ga\left(\sum_{i=1}^n \alpha_i, \theta\right)$. 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.

By finding the first and second derivatives of $M_S(t)$ at zero, we can show that $E(S) = \left.\frac{\partial M_S(t)}{\partial t}\right|_{t=0} = \alpha\theta$ where $\alpha = \sum_{i=1}^{n} \alpha_i$, and

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

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

3.1.4 Probability Generating Function

The probability generating function, denoted by $P_X(z)$, also uniquely characterizes the distribution of X. It is defined as

$$P_X(z) = \mathrm{E}\left(z^X\right) = \int_0^\infty z^x f_X(x)\,dx$$

for all z for which the expected value exists.

We can also use the probability generating function to generate moments of X. By taking the k-th derivative of $P_X(z)$ with respect to z and evaluating it at $z = 1$, we get $\mathrm{E}\left[X(X-1)\cdots(X-k+1)\right]$.

The probability generating function is more useful for discrete *rv*s and was introduced in Section 2.2.2.

3.2 Continuous Distributions for Modeling Loss Severity

In this section we explain the characteristics of distributions suitable for modeling severity of losses, including gamma, Pareto, Weibull and generalized beta distribution of the second kind. Applications for which each distribution may be used are identified.

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 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 4.

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 3.1 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
```

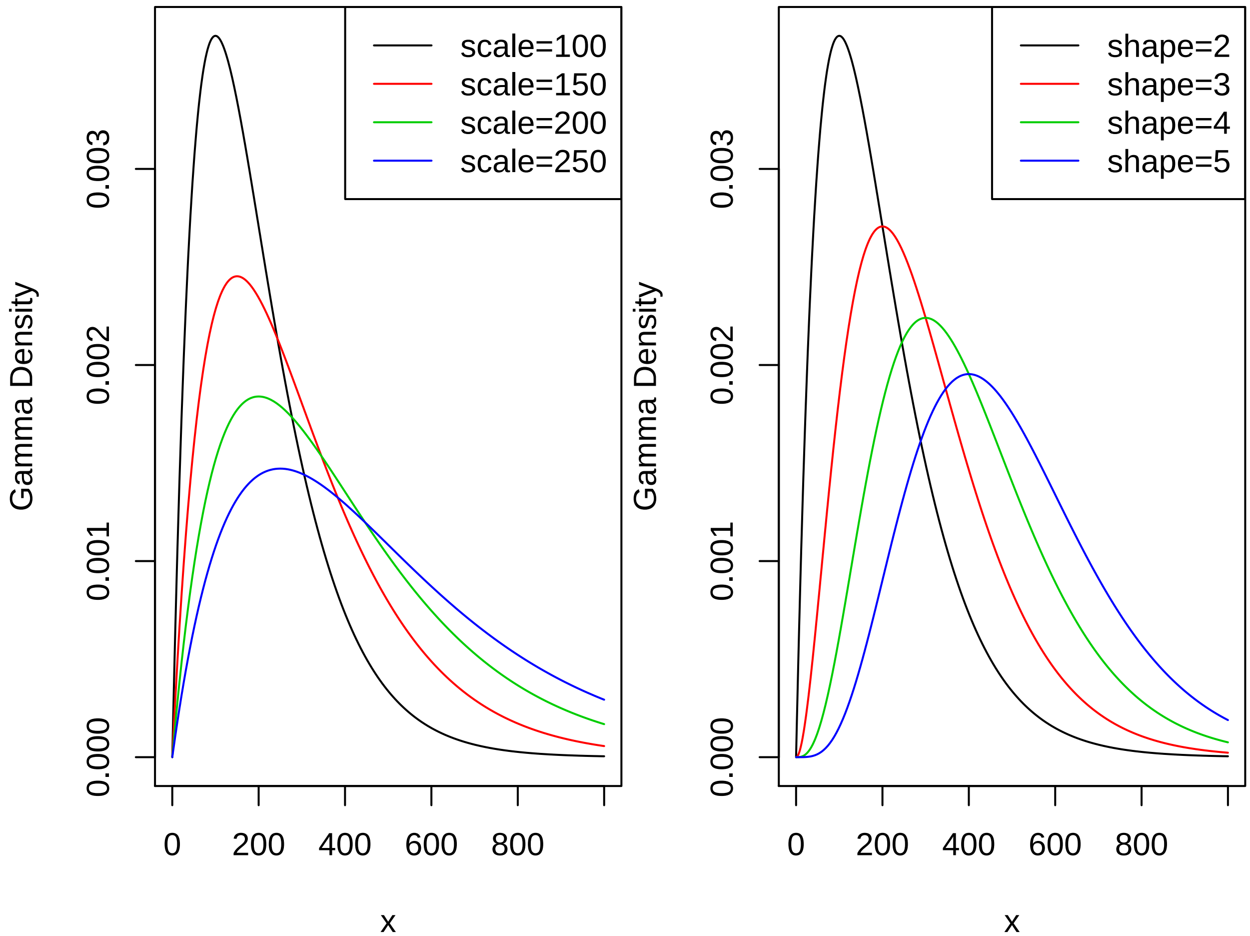


Figure 3.1: Gamma Densities. The left-hand panel is with shape=2 and Varying Scale. The right-hand panel is with scale=100 and Varying Shape.

```
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 = 1:4)

# 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 distribution and when $\alpha = \frac{n}{2}$ and $\theta = 2$ the gamma reduces to a chi-square distribution with n degrees of freedom. As we will see in Section 15.4, 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

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

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

$$\mathrm{E}\left(X^k\right) = \theta^k \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)} \quad \text{for } k > 0.$$

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 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 distribution, 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 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 3.2 demonstrate the effect of the scale and shape parameters on the Pareto density function.

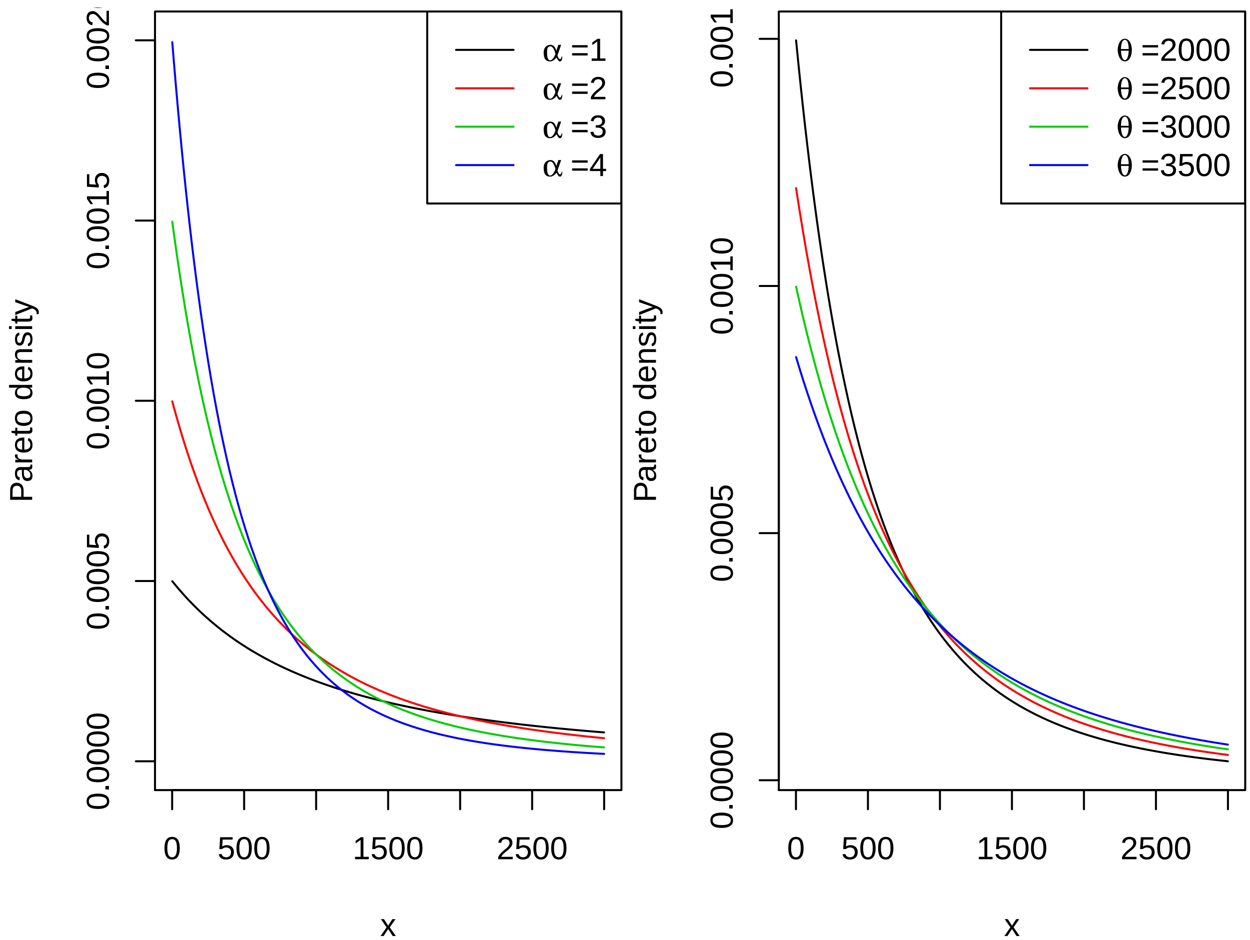


Figure 3.2: Pareto Densities. The left-hand panel is with scale=2000 and Varying Shape. The right-hand panel is with shape=3 and ,Varying Scale

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", ylab = "Pareto
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'), expression(a

# 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~'=3000'), exp
```

The distribution function of the Pareto distribution is given by

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

It can be easily seen that the hazard function of the Pareto distribution is a decreasing function in x, another indication that the distribution is heavy tailed. 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

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

The mean and variance are given by

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

and

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

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 distribution, 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 is particularly useful in modeling left-truncated claim severity distributions. Weibull was 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 3.3 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 density")
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 = 1:4)

# Varying Shape Weibull Densities
plot(z, dweibull(z, shape = shapeparam[1], scale = 100), ylim=c(0,0.012), type = "l", ylab = "Weibull densit
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 $\alpha < 1$, constant when $\alpha = 1$ and increasing when $\alpha > 1$. 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.

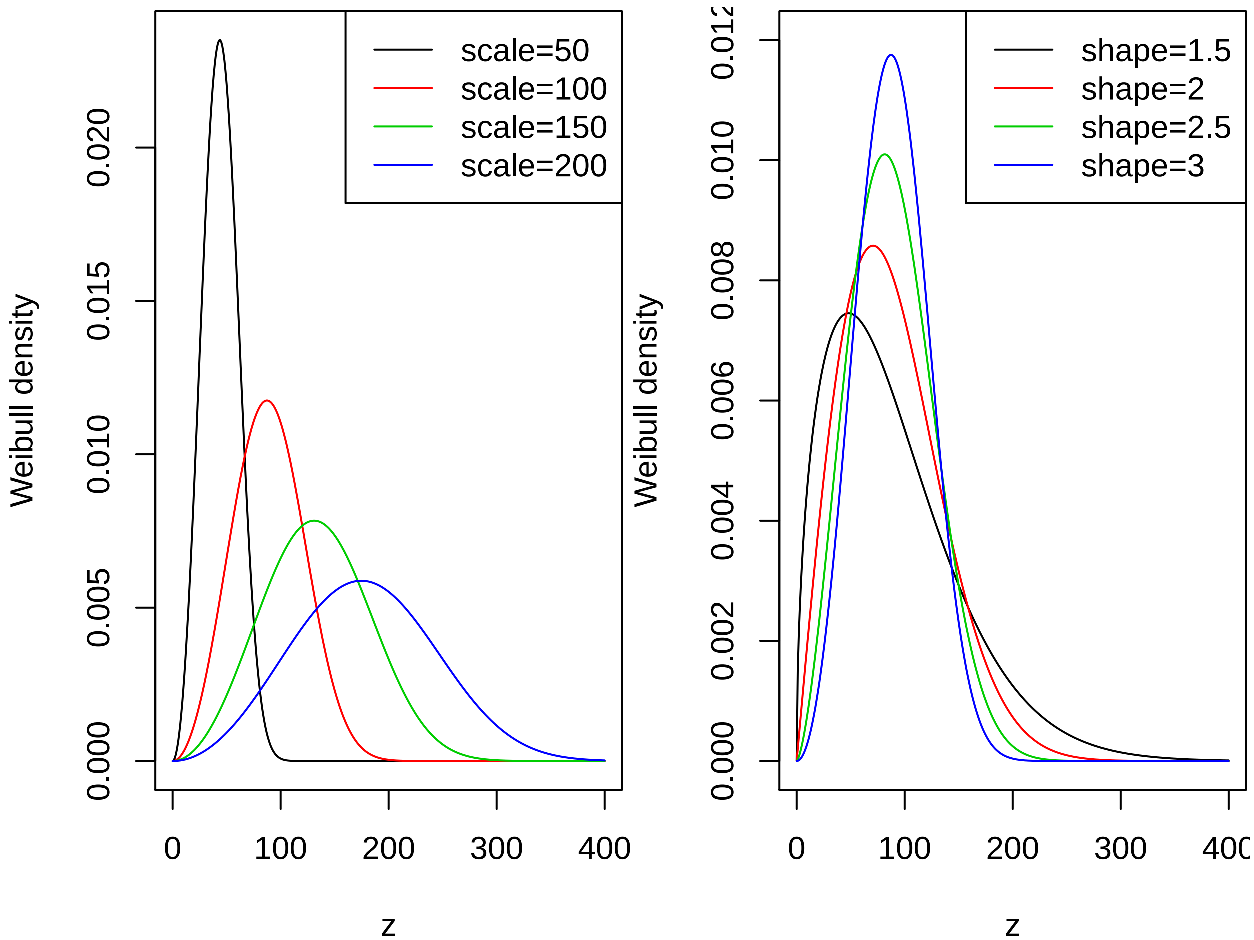


Figure 3.3: Weibull Densities. The left-hand panel is with shape=3 and Varying Scale. The right-hand panel is with scale=100 and Varying Shape.

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 $X \sim Wei\left(1.2, 33.33\right)$ be the lifetime of AIDS patients (in months). 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 Kind (GB2) was introduced by Venter (1983) in the context of insurance loss modeling and by McDonald (1984) 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 a, b, α and β if its probability density function is given by

$$f_X\left(x\right) = \frac{a x^{a\alpha - 1}}{b^{a\alpha} B\left(\alpha, \beta\right) \left[1 + \left(x/b\right)^a\right]^{\alpha + \beta}} \quad \text{for } x > 0,$$

$a, b, \alpha, \beta > 0$, and where the beta function $B\left(\alpha, \beta\right)$ is defined as

$$B\left(\alpha, \beta\right) = \int_0^1 t^{\alpha - 1}\left(1 - t\right)^{\beta - 1} \mathrm{dt}.$$

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 $a = \alpha = \beta = 1$, then the GB2 reduces to the log-logistic distribution. When $a = 1$ and $\beta \to \infty$, it reduces to the gamma distribution and when $\alpha = 1$ and $\beta \to \infty$, it reduces to the Weibull distribution.

The k-th moment of the GB2 distributed random variable is given by

$$\mathrm{E}\left(X^k\right) = \frac{b^k\left(\alpha + \frac{k}{a}, \beta - \frac{k}{a}\right)}{(\alpha, \beta)}, \quad k > 0.$$

Earlier applications of the GB2 were on income data and more recently have been used to model long-tailed claims data. GB2 was used to model different types of automobile insurance claims, severity of fire losses as well as medical insurance claim data.

3.3 Methods of Creating New Distributions

In this section we

- 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 3.2 we discussed some elementary known distributions. In this section we discuss means of creating new parametric probability distributions from existing ones. Let X be a continuous random variable with a known probability density function $f_X(x)$ and distribution function $F_X(x)$. Consider the transformation $Y = g(X)$, where $g(X)$ is a one-to-one transformation defining a new random variable Y. We can use the distribution function technique, the change-of-variable technique or the moment-generating function technique to find the probability density function of the variable of interest 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

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

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 distributions 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 distribution. 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. SOA 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^{μ} is the scale parameter which was multiplied by 1.3 while σ is the shape parameter that remained unchanged.

Example 3.3.2. SOA 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 3.2 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 the previous section we talked about the flexibility of the Weibull distribution in fitting reliability data. Looking to the origins of the Weibull distribution, we recognize that the Weibull is a power transformation 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

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

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

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

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 $X \sim Exp(\theta)$, we have

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

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,

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

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 distribution 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.

Consider the transformation $Y = e^X$, then the distribution function of Y is given by

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

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

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

Example 3.3.4. SOA Exam Question. X has a uniform distribution on the interval $(0,\ c)$. $Y = e^X$. 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,

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

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 population. 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

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

for $0 < a < 1$, where the mixing parameters 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

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

respectively.

Example 3.3.5. SOA 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 mixture* 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. SOA 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

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

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

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

Substituting the first two equations into the third, we get

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

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

.

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 mixture. 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

$$f_X(x) = \int_0^\infty f_X(x\,|\theta\ )\, g(\theta) d\theta,$$

where $f_X(x\,|\theta\ )$ is the conditional distribution of X at a particular value of θ and $g(\theta)$ is the probability statement made about the unknown parameter θ, known as the prior distribution 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

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

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

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

respectively.

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

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

Using the Double Expectation Theorem we can define the mean and variance of X as

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

and

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

Example 3.3.7. SOA 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

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

Example 3.3.8. SOA 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 distribution, mean and variance of X.

Show Example Solution

Solution.

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

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

The conditional mean and variance of X are given by

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

and

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

respectively.

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

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

and

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

The unconditional distribution of X is

$$f_X(x) = \int f_X(x|\theta)\ g(\theta)\, d\theta.$$

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

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 hazard. 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 loss variable, denoted by Y^L, includes losses for which a payment is made as well as losses less than the deductible and hence is defined as

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

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 payment variable, denoted by Y^P, is not defined when there is no payment and only includes losses for which a payment is made. The variable is defined as

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

Y^P is often referred to as left truncated 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 partly discrete and partly continuous. The discrete part of the distribution is concentrated at $Y = 0$ (when $X \le d$) 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,

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

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

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

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.

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*) 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 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. SOA 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

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

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,

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

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

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

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

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. SOA Exam Question. For an insurance:

Losses have a density function

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

The insurance has an ordinary deductible of 4 per loss.

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

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

Show Example Solution

Solution.

We define Y^P as follows

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

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. SOA 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

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

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

The LER for the coming year is

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

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 coverage 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. Then 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 a limit u, the per-loss variable Y^L is defined as

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

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 partly discrete and partly continuous. 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,

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

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

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

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

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

Example 3.4.4. SOA Exam Question. Under a group insurance 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

As we have seen in Section 3.4.1, 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 3.4.2 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 $Y^L = \alpha\left[(X \wedge u) - (X \wedge d)\right]$.

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

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

A growth factor $(1+r)$ 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

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

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

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

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. SOA 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,

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

$$= 356.026$$

.

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

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

$$= 361.659$$

.

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

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

Because $\frac{E\left(Y^P_{2007}\right)}{E\left(Y^P_{2006}\right)} - 1 = 0.011$, 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 3.4.1 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 reinsurance, 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 cedent 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 coverage. 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 retention.

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. SOA Exam Question. In 2005 a risk has a two-parameter Pareto distribution with $\alpha = 2$ and $\theta = 3000$. In 2006 losses inflate by 20%. Insurance on the risk has a deductible of 600 in each year. P_i, the premium in year i, equals 1.2 times expected claims. The risk is reinsured with a deductible that stays the same in each year. R_i, the reinsurance premium in year i, equals 1.1 times the expected reinsured claims. $\frac{R_{2005}}{P_{2005}} = 0.55$. Calculate $\frac{R_{2006}}{P_{2006}}$.

Show Example Solution

Solution.

Let us use the following notation:

X_i : The risk in year i

Y_i : The insured claim in year i

P_i : The insurance premium in year i

Y_i^R : The reinsured claim in year i

R_i : The reinsurance premium in year i

d : The insurance deductible in year i (the insurance deductible is fixed each year, equal to 600)

d^R : The reinsurance deductible or retention in year i (the reinsurance deductible is fixed each year, but unknown) where $i = 2005,\ 2006$

$$Y_i = \begin{cases} 0 & X_i \le 600 \\ X_i - 600 & X_i > 600 \end{cases}$$

where $i = 2005,\ 2006$

$$X_{2005} \sim Pa\,(2, 3000)$$

$$\mathrm{E}\,(Y_{2005}) = \mathrm{E}\,(X_{2005} - 600)_+ = \mathrm{E}\,(X_{2005}) - \mathrm{E}\,(X_{2005} \wedge 600)$$

$$= 3000 - 3000\left(1 - \frac{3000}{3600}\right) = 2500$$

$$P_{2005} = 1.2\mathrm{E}\,(Y_{2005}) = 3000$$

Since $X_{2006} = 1.2X_{2005}$ and Pareto is a scale distribution with scale parameter θ, then $X_{2006} \sim Pa\,(2, 3600)$

$$\mathrm{E}\,(Y_{2006}) = \mathrm{E}\,(X_{2006} - 600)_+ = \mathrm{E}\,(X_{2006}) - \mathrm{E}\,(X_{2006} \wedge 600)$$

$$= 3600 - 3600\left(1 - \frac{3600}{4200}\right) = 3085.714$$

$$P_{2006} = 1.2\mathrm{E}\,(Y_{2006}) = 3702.857$$

$$Y_i^R = \begin{cases} 0 & X_i - 600 \le d^R \\ X_i - 600 - d^R & X_i - 600 > d^R \end{cases}$$

Since $\frac{R_{2005}}{P_{2005}} = 0.55$, then $R_{2005} = 3000 \times 0.55 = 1650$

Since $R_{2005} = 1.1\mathrm{E}\left(Y_{2005}^R\right)$, then $\mathrm{E}\left(Y_{2005}^R\right) = \frac{1650}{1.1} = 1500$

$$\mathrm{E}\left(Y_{2005}^R\right) = \mathrm{E}\left(X_{2005} - 600 - d^R\right)_+ = \mathrm{E}\,(X_{2005}) - \mathrm{E}\left(X_{2005} \wedge \left(600 + d^R\right)\right)$$

$$= 3000 - 3000\left(1 - \frac{3000}{3600 + d^R}\right) = 1500 \Rightarrow d^R = 2400$$

$$\mathrm{E}\left(Y_{2006}^R\right) = \mathrm{E}\left(X_{2006} - 600 - d^R\right)_+ = \mathrm{E}\,(X_{2006} - 3000)_+ = \mathrm{E}\,(X_{2006}) - \mathrm{E}\,(X_{2006} \wedge 3000)$$

$$= 3600 - 3600\left(1 - \frac{3600}{6600}\right) = 1963.636$$

$$R_{2006} = 1.1\mathrm{E}\left(Y_{2006}^{R}\right) = 1.1 \times 1963.636 = 2160$$

Therefore $\frac{R_{2006}}{P_{2006}} = \frac{2160}{3702.857} = 0.583$

3.5 Maximum Likelihood Estimation

In this section we estimate statistical parameters using the method of maximum likelihood. Maximum likelihood estimates in the presence of grouping, truncation or censoring are calculated.

3.5.1 Maximum Likelihood Estimators for Complete Data

Pricing of insurance premiums and estimation of claim reserving are among many actuarial problems that involve modeling the severity of loss (claim size). Appendix Chapter 17 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 this section, we present a few examples to illustrate how actuaries fit a parametric distribution model to a set of claim data using maximum likelihood. In 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 3.5.1. SOA 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.

Approximate the variance of the maximum likelihood estimator.

Determine an approximate 95% confidence interval for θ.

Determine an approximate 95% confidence interval for $\Pr\left(X \leq 9,000\right)$.

Show Example Solution

Solution.

The probability density function is

$$f_X\left(x\right) = \frac{\theta e^{-\frac{\theta}{x}}}{x^2},$$

where $x > 0$.

a. The likelihood function, $L\left(\theta\right)$, can be viewed as the probability of the observed data, written as a function of the model's parameter θ

$$L\left(\theta\right) = \prod_{i=1}^{4} f_{X_i}\left(x_i\right) = \frac{\theta^4 e^{-\theta \sum_{i=1}^{4} \frac{1}{x_i}}}{\prod_{i=1}^{4} x_i^2}.$$

The log-likelihood function, $\ln L\left(\theta\right)$, is the sum of the individual logarithms.

$$\ln L\left(\theta\right) = 4ln\theta - \theta \sum_{i=1}^{4} \frac{1}{x_i} - 2\sum_{i=1}^{4} \ln x_i.$$

$$\frac{d \ln L\left(\theta\right)}{d\theta} = \frac{4}{\theta} - \sum_{i=1}^{4} \frac{1}{x_i}.$$

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.

b. Taking reciprocal of negative expectation of the second derivative of $\ln L(\theta)$, we obtain an estimate of the variance of $\hat{\theta}$, $\widehat{Var}\left(\hat{\theta}\right) = \left.\left[E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right)\right]^{-1}\right|_{\theta=\hat{\theta}} = \frac{\hat{\theta}^2}{4} = 28,446,222.$

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.

c. The 95% confidence interval for θ is given by

$$10,667 \pm 1.96\sqrt{28,446,222} = (213.34,\ 21,120.66).$$

d. The distribution function of X is $F(x) = 1 - e^{-\frac{x}{\theta}}$. Then, the maximum likelihood estimate of $g(\theta) = F(9,000)$ 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 $g\left(\hat{\theta}\right)$.

$$\frac{\mathrm{dg}(\theta)}{d\theta} = -\frac{9,000}{\theta^2} e^{-\frac{9,000}{\theta}}.$$

$$\widehat{Var}\left[g\left(\hat{\theta}\right)\right] = \left(-\frac{9,000}{\hat{\theta}^2} e^{-\frac{9,000}{\hat{\theta}}}\right)^2 \hat{V}\left(\hat{\theta}\right) = 0.0329.$$

The 95% confidence interval for $F(9,000)$ is given by

$$0.57 \pm 1.96\sqrt{0.0329} = (0.214,\ 0.926).$$

Example 3.5.2. SOA 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.

Estimate the covariance matrix 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

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 $x > 0$.

a. 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.

$$\ln(\mu, \sigma) = -6ln\sigma - 3ln(2\pi) - \sum_{i=1}^{6} \ln x_i - \frac{1}{2}\sum_{i=1}^{6}\left(\frac{\ln x_i - \mu}{\sigma}\right)^2.$$

The first partial derivatives are

$$\frac{\partial lnL(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{6}(\ln x_i - \mu).$$

$$\frac{\partial lnL(\mu, \sigma)}{\partial \sigma} = \frac{-6}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{6}(\ln x_i - \mu)^2.$$

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

$\hat{\mu} = \frac{\sum_{i=1}^{6} \ln x_i}{6} = 9.38$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{6} (\ln x_i - \hat{\mu})^2}{6} = 5.12.$

The second partial derivatives are

$\frac{\partial^2 \mathrm{lnL}(\mu,\sigma)}{\partial \mu^2} = \frac{-6}{\sigma^2}$, $\frac{\partial^2 \mathrm{lnL}(\mu,\sigma)}{\partial \mu \partial \sigma} = \frac{-2}{\sigma^3}\sum_{i=1}^{6}(\ln x_i - \mu)$ and $\frac{\partial^2 \mathrm{lnL}(\mu,\sigma)}{\partial \sigma^2} = \frac{6}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{6}(\ln x_i - \mu)^2$.

b. 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

$$\hat{\Sigma} = \begin{bmatrix} 0.8533 & 0 \\ 0 & 0.4267 \end{bmatrix}.$$

c. 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 $5.12 \pm 1.96\sqrt{0.4267} = (3.84,\ 6.40)$.

d. 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

$$\hat{V}(g(\hat{\mu}, \hat{\sigma})) = \left[\frac{\partial g(\mu,\sigma)}{\partial \mu} \quad \frac{\partial g(\mu,\sigma)}{\partial \sigma}\right] \Sigma \begin{bmatrix} \frac{\partial g(\mu,\sigma)}{\partial \mu} \\ \frac{\partial g(\mu,\sigma)}{\partial \sigma} \end{bmatrix} \Bigg|_{\mu=\hat{\mu}, \sigma=\hat{\sigma}}$$

$$= \begin{bmatrix} 153,277 & 346,826 \end{bmatrix} \begin{bmatrix} 0.8533 & 0 \\ 0 & 0.4267 \end{bmatrix} \begin{bmatrix} 153,277 \\ 346,826 \end{bmatrix} =$$

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.

3.5.2 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 data. 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

$$L(\theta) = \prod_{j=1}^{k} \left[F(\ c_j | \theta) - F(\ c_{j-1} | \theta)\right]^{n_j},$$

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 3.5.3. SOA Exam Question. For a group of policies, you are given that losses follow the distribution function $F(x) = 1 - \frac{\theta}{x}$, 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, $\Pr(X \leq 10) = F(10)$. Similarly, the contributions of each of 6 and 5 observations in the second and third intervals are $\Pr(10 < X \leq 25) = F(25) - F(10)$ and $P(X > 25) = 1 - F(25)$, respectively. The likelihood function is thus given by

$$\begin{aligned} L(\theta) &= [F(10)]^9 [F(25) - F(10)]^6 [1 - F(25)]^5 \\ &= \left(1 - \frac{\theta}{10}\right)^9 \left(\frac{\theta}{10} - \frac{\theta}{25}\right)^6 \left(\frac{\theta}{25}\right)^5 \\ &= \left(\frac{10-\theta}{10}\right)^9 \left(\frac{15\theta}{250}\right)^6 \left(\frac{\theta}{25}\right)^5. \end{aligned}$$

Then, $\ln L(\theta) = 9ln(10-\theta) + 6ln\theta + 5ln\theta - 9ln10 + 6ln15 - 6ln250 - 5ln25$.

$$\frac{d \ln L(\theta)}{d\theta} = \frac{-9}{(10-\theta)} + \frac{6}{\theta} + \frac{5}{\theta}.$$

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 $\hat{\theta} = 5.5$.

3.5.3 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 3.4.2 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.

Example 3.5.4. SOA 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

$$f_X(x) = \frac{4x\theta^4}{(\theta^2+x^2)^3}.$$

Thus,

$$L(\theta) = \frac{8\theta^4}{(\theta^2+4)^3}\frac{16\theta^4}{(\theta^2+16)^3}\frac{\theta^4}{(\theta^2+16)^2} = \frac{128\theta^{12}}{(\theta^2+4)^3(\theta^2+16)^5},$$

So,

$$\ln L(\theta) = ln128 + 12ln\theta - 3ln\left(\theta^2+4\right) - 5ln\left(\theta^2+16\right)$$

,

and

$\frac{\text{dlnL}(\theta)}{d\theta} = \frac{12}{\theta} - \frac{6\theta}{(\theta^2+4)} - \frac{10\theta}{(\theta^2+16)}$.

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 $\hat{\theta} = 5.7$.

3.5.4 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 3.4.1 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 $\frac{f_X(x)}{S_X(d)}$.

Example 3.5.5. SOA 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

$$f_X(x) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}} \text{ and } F_X(x) = 1 - \left(\frac{\theta}{x}\right)^\alpha,$$

for $x > \theta$, respectively. Then, the likelihood and loglikelihood functions are given by

$$L(\alpha) = \frac{\alpha^8}{\prod_{i=1}^{8} x_i^{\alpha+1}} \frac{5^{10\alpha}}{25^{2\alpha}},$$

$$\ln L(\alpha) = 8ln\alpha - (\alpha+1)\sum_{i=1}^{8} \ln x_i + 10\alpha ln5 - 2\alpha ln25.$$

$\frac{\mathrm{dlnL}(\alpha)}{d\theta} = \frac{8}{\alpha} - \sum_{i=1}^{8} \ln x_i + 10ln5 - 2ln25.$

The maximum likelihood estimator, $\hat{\alpha}$, is the solution to the equation

$$\frac{8}{\hat{\alpha}} - \sum_{i=1}^{8} \ln x_i + 10ln5 - 2ln25 = 0,$$

which yields

$$\hat{\alpha} = \frac{8}{\sum_{i=1}^{8} \ln x_i - 10ln5 + 2ln25} = \frac{8}{(ln7 + ln9 + \cdots + ln20) - 10ln5 + 2ln25} = 0.785.$$

The mean of the Pareto only exists for $\alpha > 1$. Since $\hat{\alpha} = 0.785 < 1$. Then, the mean does not exist.

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 .

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: Cummins and Derrig (2012), Frees and Valdez (2008), Klugman et al. (2012), Kreer et al. (2015), McDonald (1984), McDonald and Xu (1995), Tevet (2016), and Venter (1983).

Chapter 4

Model Selection and Estimation

Chapter Preview. Chapters 2 and 3 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 4.1 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 4.2. 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 4.3 describes estimation for grouped, censored and truncated data (following the Section 3.5 introduction). To see how they can be adapted to alternative models, the chapter closes with Section 4.4 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 2.2 for frequency and Section 3.1 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 **nonparametric** An 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 *iid* identically 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 2.2.2 for frequency and Section 3.1.1 for severity. In particular, the k-th moment, $\mathrm{E}~[X^k] = \mu'_k$, 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*, $\mathrm{E}~[(X-\mu)^k] = \mu_k$. (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, $\sigma^2 = \mathrm{E}~[(X-\mu)^2]$, 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 $\mathrm{E}~s^2 = \sigma^2$, 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 $\mathrm{E}~X^k$. In the same way, for any known function $\mathrm{g}(\cdot)$, we can estimate $\mathrm{E}~\mathrm{g}(X)$ using $n^{-1}\sum_{i=1}^{n} \mathrm{g}(X_i)$. This is sometimes known as the *analog principle*.

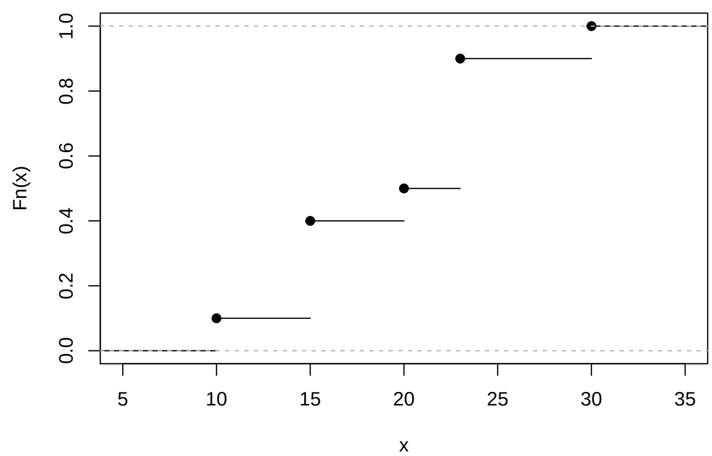


Figure 4.1: Empirical Distribution Function of a Toy Example

Now suppose that we fix a value of x and consider the function $\mathrm{g}(X) = I(X \leq x)$. 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 $\mathrm{g}(\cdot)$, the expected value is $\mathrm{E}\ I(X \leq x) = \Pr(X \leq x) = F(x)$, the distribution function evaluated at a fixed point x. Using the analog principle, we define the nonparametric estimator of the distribution function

$$\begin{aligned} F_n(x) &= \frac{1}{n} \sum_{i=1}^{n} I\left(X_i \leq x\right) \\ &= \frac{\text{number of observations less than or equal to } x}{n}. \end{aligned}$$

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 4.1.

Show R Code

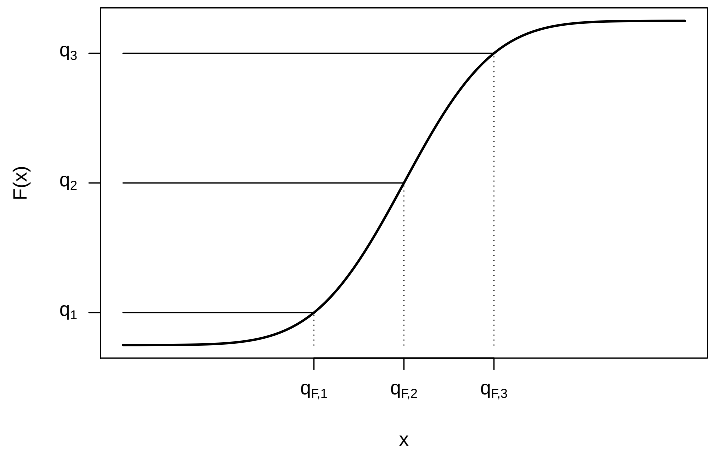


Figure 4.2: Continuous Quantile Case

```
(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 4.2. 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 4.3 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 4.1. For the value of q shown in this panel, we still have a unique value of the

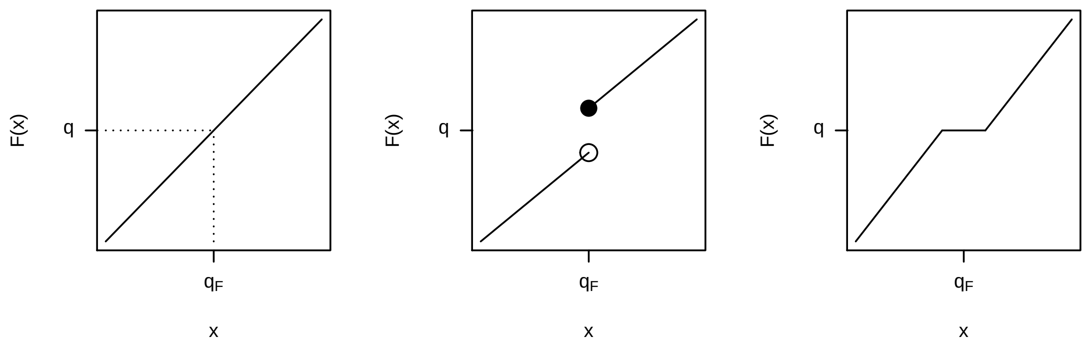


Figure 4.3: Three Quantile Cases

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 (4.1). 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 (4.1).

Example 4.1.2. Toy Data Set: Continued. Determine quantiles corresponding to the 20th, 50th, and 95th percentiles.

Show Example Solution

Solution. Consider Figure 4.1. 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 4.3. 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

$$f_n(x) = \frac{n_j}{n \times (c_j - c_{j-1})} \qquad c_{j-1} \leq x < c_j,$$

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

$$f_n(x) = \frac{1}{2nb} \sum_{i=1}^{n} I(x - b < X_i \leq x + b) \tag{4.2}$$

Show A Snippet of Theory

The idea is that the estimator $f_n(x)$ in equation (4.2) is the average over n *iid* identically and independently distributed realizations of a random variable with mean

$$\begin{aligned}
\mathrm{E}\ \frac{1}{2b} I(x - b < X \leq x + b) &= \frac{1}{2b} \left(F(x + b) - F(x - b) \right) \\
&= \frac{1}{2b} \left(\left\{ F(x) + bF'(x) + b^2 C_1 \right\} \left\{ F(x) - bF'(x) + b^2 C_2 \right\} \right) \\
&= F'(x) + b\frac{C_1 - C_2}{2} \to F'(x) = f(x),
\end{aligned}$$

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 *pdf* probability 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 (4.2) 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	$(1 - \lvert x\rvert) \times I(\lvert x\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - x^2) \times I(\lvert x\rvert \leq 1)$
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 4.4 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$Claim), main="", ylim=c(0,.35),xlab="Log Expenditures", freq=FALSE, col="lightgray")
lines(density(log(ClaimData$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)
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

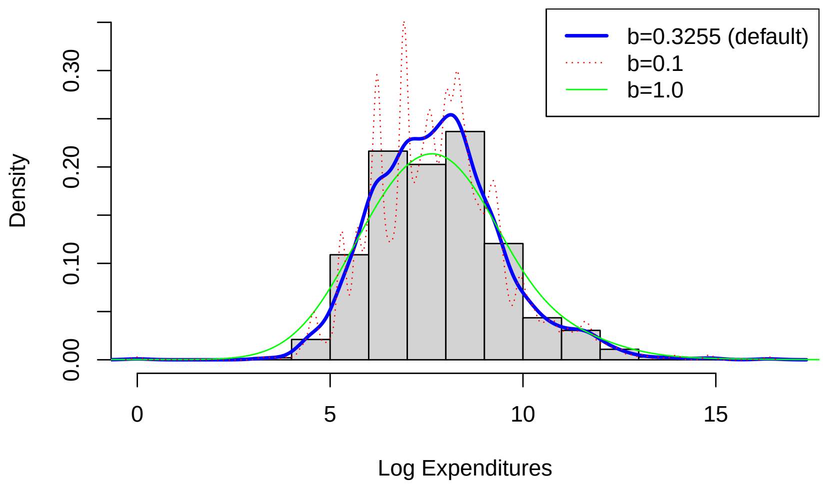


Figure 4.4: Histogram of Logarithmic Property Claims with Superimposed Kernel Density Estimators

$$W(y) = \begin{cases} 0 & y < -1 \\ \frac{y+1}{2} & -1 \le y < 1 \\ 1 & y \ge 1 \end{cases}.$$

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, $w(x) = (1 - |x|) \times I(|x| \le 1)$. 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

$$f_n(x) = \frac{1}{5(2)}\left(\frac{3}{4} + (3)\frac{3}{4} + 0\right) = \frac{3}{10}$$

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 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 4.5 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 4.6 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 4.5, 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 4.7.

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 4.7, for each point on the aforementioned grid, the horizontal axis displays the empirical quantile and the vertical axis displays the corresponding fitted parametric quantile (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

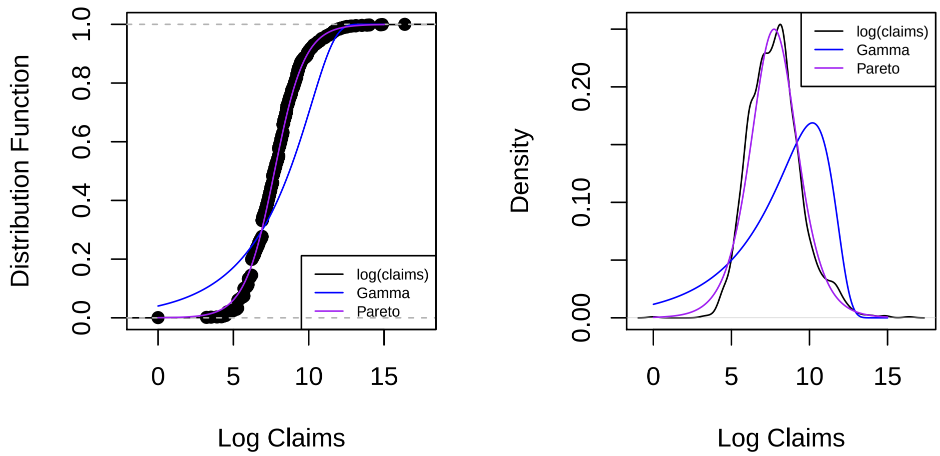


Figure 4.5: Nonparametric Versus Fitted Parametric Distribution and Density Functions. The left-hand panel compares distribution functions, with the dots corresponding to the empirical distribution, the thick blue curve corresponding to the fitted gamma and the light purple curve corresponding to the fitted Pareto. The right hand panel compares these three distributions summarized using probability density functions.

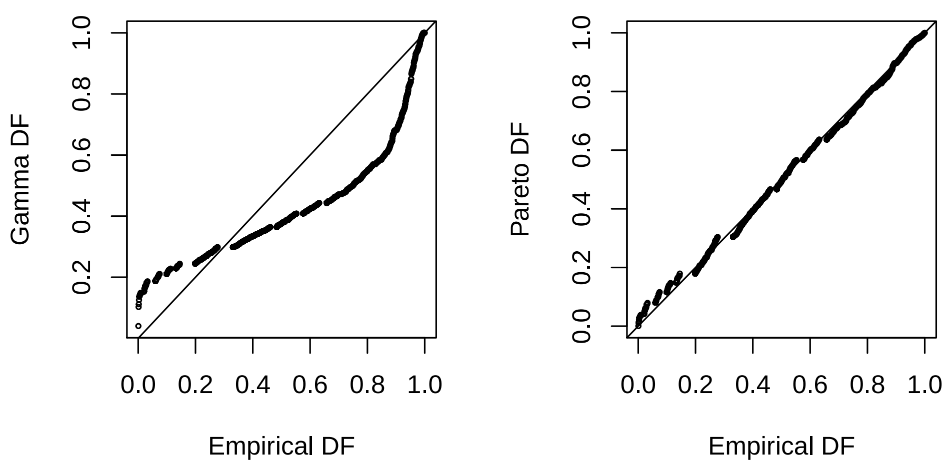


Figure 4.6: Probability-Probability (*pp*) Plots. The horizontal axes gives the empirical distribution function at each observation. In the left-hand panel, the corresponding distribution function for the gamma is shown in the vertical axis. The right-hand panel shows the fitted Pareto distribution. Lines of $y = x$ are superimposed.

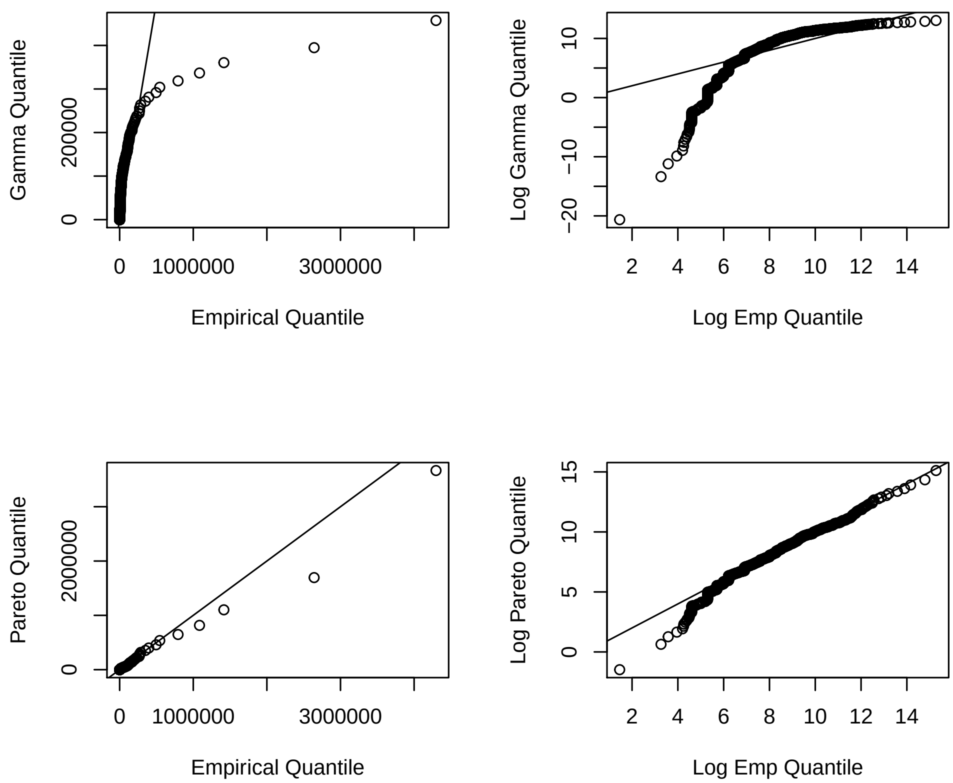


Figure 4.7: Quantile-Quantile (*qq*) Plots. The horizontal axes gives the empirical quantiles at each observation. The right-hand panels they are graphed on a logarithmic basis. The vertical axis gives the quantiles from the fitted distributions; gamma quantiles are in the upper panels, Pareto quantiles are in the lower panels.

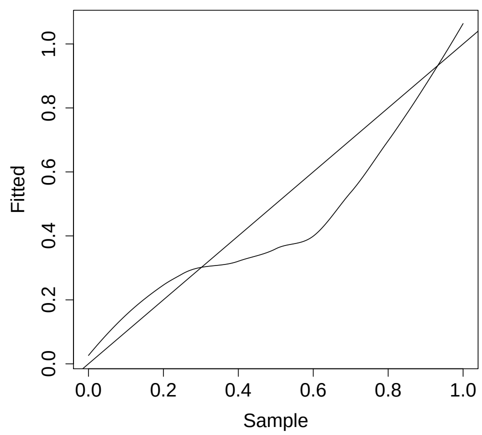
us to see where a fitted distribution is deficient. The straight line represents equality between the empirical distribution and fitted distribution. From these plots, we again see that the Pareto is an overall better fit than the gamma. Furthermore, the lower-right panel suggests that the Pareto distribution 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 17 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 4.1.1. 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

$$
\begin{aligned}
\hat{\alpha} &= \frac{26.62259^2}{136154.6 - 26.62259^2} = 0.005232809 \\
\hat{\theta} &= \frac{136154.6 - 26.62259^2}{26.62259} = 5,087.629.
\end{aligned}
$$

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

$$
\begin{aligned}
\hat{\alpha} &= 1 + \frac{136154.6}{136154.6 - 26,62259^2} = 2.005233 \\
\hat{\theta} &= (2.005233 - 1) \cdot 26.62259 = 26.7619
\end{aligned}
$$

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 4.1.1.

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:

$$10 \quad 35 \quad 80 \quad 86 \quad 90 \quad 120 \quad 158 \quad 180 \quad 200 \quad 210 \quad 1500$$

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.

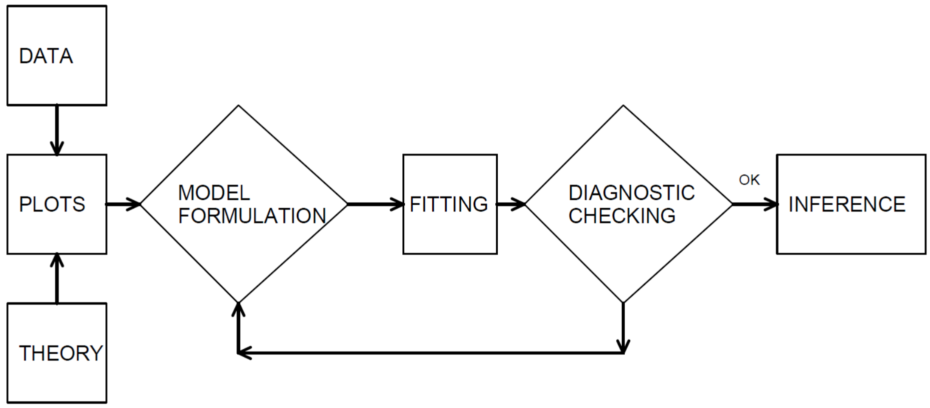


Figure 4.8: The iterative model specification process.

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**.

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. Box (1980) describes this as an *iterative process* which is shown in Figure 4.8.

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 4.1. 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 4.1.2 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 15.4.3 and 17.3.2.

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 15.4.4 provides a summary of these statistics.

For selecting the appropriate distribution, statistics that compare a parametric fit to a nonparametric alternative, summarized in Section 4.1.2, are useful for model comparison. For discrete data, a *goodness of fit* statistic (as described in Section 2.7) is generally preferred as it is more intuitive and simpler to explain.

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.

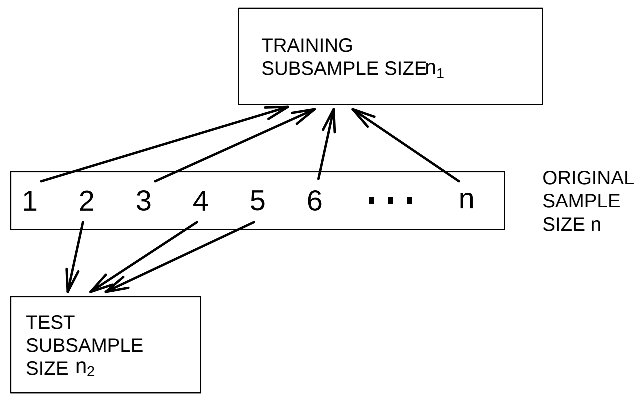


Figure 4.9: Model Validation. A data set is randomly split into two subsamples.

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 4.9 illustrates this splitting of the data.

Various researchers recommend different proportions for the allocation. Snee (1977) suggests that data-splitting not be done unless the sample size is moderately large. The guidelines of Picard and Berk (1990) 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 (James et al. (2013)). 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). The Chapter 4 *Technical Supplement A* 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 *AIC* Akaike's information criterion and *BIC* Bayesian 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 (4.4) 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 (4.4) 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.

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 3.5 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 10, 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

$$\begin{aligned} f(x) &= \begin{cases} F(c_1) - F(c_0) & \text{if } \; x \in (c_0, c_1] \\ \vdots & \vdots \\ F(c_k) - F(c_{k-1}) & \text{if } \; x \in (c_{k-1}, c_k] \end{cases} \\ &= \prod_{j=1}^{k} \{F(c_j) - F(c_{j-1})\}^{I(x \in (c_{j-1}, c_j])} \end{aligned}$$

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

$$\mathcal{L}(\theta) = \prod_{j=1}^{n} f(x_i) = \prod_{j=1}^{k} \{F(c_j) - F(c_{j-1})\}^{n_j}$$

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 3.4.2, 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 3.4.4 and will see more in Chapter 10. 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

$$Y = \left\{ \begin{array}{cc} \text{we do not observe } X & X < C_L \\ X & X \geq C_L \end{array} \right. .$$

In other words, if X is less than the threshold C_L, then it is not observed.

For an example we saw in Section 3.4.1, 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

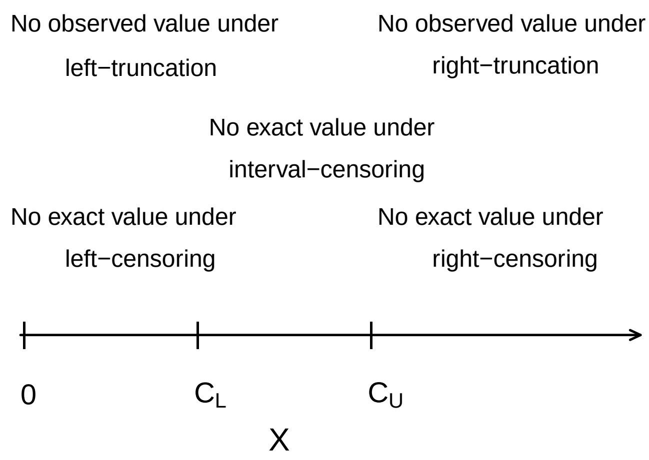


Figure 4.10: Censoring and Truncation

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 3.4.1, we defined the per payment loss to be

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

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, 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

$$Y = \begin{cases} X & X \le C_U \\ \text{we do not observe } X & X > C_U. \end{cases}$$

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 4.10 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 4.11 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 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.

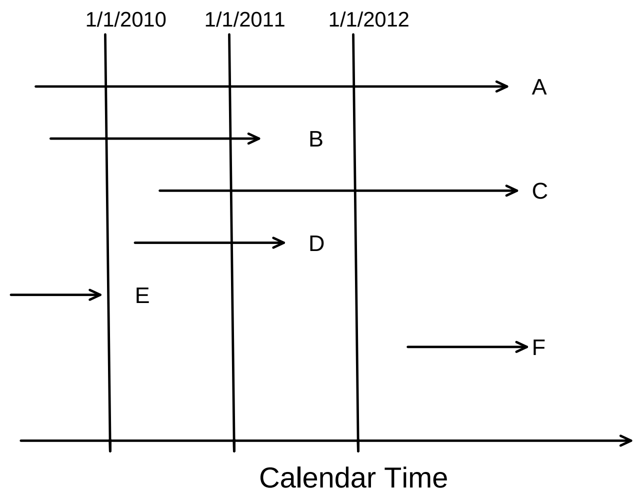


Figure 4.11: Timeline for Several Subjects on Test in a Mortality Study

- **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	$X_U^* = \min(X, C_U)$	$\delta_U = I(X \le C_U)$
left censoring	$X_L^* = \max(X, C_L)$	$\delta_L = I(X \ge C_L)$
interval censoring		
right truncation	X	observe X if $X \le C_U$
left truncation	X	observe X if $X \ge C_L$

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 $\delta = I(X \le C_U)$. If censoring occurs so that $\delta = 0$, then $X \ge C_U$ and the likelihood is $\Pr(X \ge C_U) = 1 - F(C_U)$. 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

$$\left\{ \begin{array}{ll} 1 - F(C_U) & \text{if } \delta = 0 \\ f(x) & \text{if } \delta = 1 \end{array} \right. = \{f(x)\}^{\delta} \{1 - F(C_U)\}^{1-\delta}.$$

The right-hand expression allows us to present the likelihood more compactly. Now, for an *iid* sample of size n, the likelihood is

$$\mathcal{L} = \prod_{i=1}^{n} \{f(x_i)\}^{\delta_i} \{1 - F(C_{Ui})\}^{1-\delta_i} = \prod_{\delta_i=1} f(x_i) \prod_{\delta_i=0} \{1 - F(C_{Ui})\},$$

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 *E*xact values, and similarly for *R*ight-, *L*eft- and *I*nterval-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

$$\begin{aligned} L(\theta) &= \sum_E \{\ln f(x_i) - \ln(1 - F(C_{Li}))\} + \sum_R \{\ln(1 - F(C_{Ui})) - \ln(1 - \mathrm{F}(C_{Li}))\} \\ &= \sum_E (-\ln \theta - (x_i - C_{Li})/\theta) - \sum_R (C_{Ui} - C_{Li})/\theta. \end{aligned}$$

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 4.3.1, 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 (Kaplan and Meier, 1958) 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 (Kaplan and Meier, 1958) 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

$$\hat{F}(x) = \left\{ \begin{array}{ll} 0 & x < t_1 \\ 1 - \prod_{j: t_j \leq x} \left(1 - \frac{s_j}{R_j}\right) & x \geq t_1 \end{array} \right. . \tag{4.6}$$

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 4 5+ 5+ 5+ 8 10+ 10+ 12 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

$$\begin{aligned}\hat{S}(11) &= \prod_{j:t_j\leq 11}\left(1-\frac{s_j}{R_j}\right) = \prod_{j=1}^{2}\left(1-\frac{s_j}{R_j}\right)\\ &= \left(1-\frac{2}{10}\right)\left(1-\frac{1}{5}\right) = (0.8)(0.8) = 0.64.\end{aligned}$$

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 (4.6).

Greenwood's Formula. (Greenwood, 1926) 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 (Aalen, 1978)) 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

$$\hat{F}_{NA}(x) = \begin{cases} 0 & x < t_1 \\ 1 - \exp\left(-\sum_{j:t_j \le x} \frac{s_j}{R_j}\right) & x \ge t_1 \end{cases}.$$

Note that the above expression is a result of the Nelson-Äalen estimator of the cumulative hazard function

$$\hat{H}(x) = \sum_{j:t_j \le x} \frac{s_j}{R_j}$$

and the relationship between the survival function and cumulative hazard function, $\hat{S}_{NA}(x) = e^{-\hat{H}(x)}$.

Example 4.3.6. 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 $R_j = \sum_{i=1}^{n} \{I(x_i \ge t_j) + I(u_i \ge t_j) - I(d_i \ge t_j)\}$. 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 4.3.7. 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

$$\begin{aligned}\hat{H}(11) &= \sum_{j:t_j \le 11} \frac{s_j}{R_j} = \sum_{j=1}^{2} \frac{s_j}{R_j} \\ &= \frac{2}{10} + \frac{1}{5} = 0.2 + 0.2 = 0.4.\end{aligned}$$

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

$$\widehat{Var}(\hat{S}(11)) = (\hat{S}(11))^2 \sum_{j:t_j \le 11} \frac{s_j}{R_j(R_j - s_j)} = (0.64)^2 \left(\frac{2}{10(8)} + \frac{1}{5(4)} \right) = 0.0307.$$

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 2. 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 16.3, 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 *pdf* probability 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 16 on iterated expectations, we have that the *pmf* probability 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 2.2.3, 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 6 but more intensive techniques such as *MCMC* requires yet more background. See Hartman (2016) for an introduction to computational Bayesian methods from an actuarial perspective.

4.4.2 Bayesian Model

With the intuition developed in the preceding Section 4.4.1, 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 *pdf* probability density function by a *pmf* probability 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

$$\Pr(a \le \theta \le b|\mathbf{x}) \ge 1-\alpha.$$

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 $\mathrm{E}\ [l(\hat{\theta}, \theta)]$. Some important special cases include:

Loss function $l(\hat{\theta}, \theta)$	Descriptor	Bayes Estimate
$(\hat{\theta}-\theta)^2$	squared error loss	$\mathrm{E}(\theta|X)$
$|\hat{\theta}-\theta|$	absolute deviation loss	median of $\pi(\theta|x)$
$I(\hat{\theta}=\theta)$	zero-one loss (for discrete probabilities)	mode of $\pi(\theta|x)$

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

$$\pi(\theta|3) = \frac{f(3|\theta)\pi(\theta)}{\int_1^\infty f(3|\theta)\pi(\theta)d\theta} = \frac{\frac{2\theta^2}{(3+\theta)^3}\frac{1}{\theta^2}}{\int_1^\infty 2(3+\theta)^{-3}d\theta} = \frac{2(3+\theta)^{-3}}{-(3+\theta)^{-2}|_1^\infty} = 32(3+\theta)^{-3}, \quad \theta > 1$$

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

$$
\begin{aligned}
\Pr(X_2>8|\theta) &= \int_8^\infty f(x|\theta)dx\\
&= \int_8^\infty \frac{\theta}{(x+\theta)^2}dx = -\frac{\theta}{x+\theta}\bigg|_8^\infty = \frac{\theta}{8+\theta}
\end{aligned}
$$

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

$$
\begin{aligned}
\pi(p|1,1,1,1,1,1,1,1) &\propto \Pr(1,1,1,1,1,1,1,1|p)\ \pi(p) = p^8(2) \propto p^8\\
\Rightarrow \pi(p|1,1,1,1,1,1,1,1) &= \frac{p^8}{\int_0^5 p^8dp} = \frac{p^8}{(0.5^9)/9} = 9(0.5^{-9})p^8
\end{aligned}
$$

Thus, the posterior probability that the insured will have at least one loss during Year 9 is

$$
\begin{aligned}
\Pr(X_9=1|1,1,1,1,1,1,1,1) &= \int_0^5 \Pr(X_9=1|p)\pi(p|1,1,1,1,1,1,1,1)dp\\
&= \int_0^5 p(9)(0.5^{-9})p^8dp = 9(0.5^{-9})(0.5^{10})/10 = 0.45
\end{aligned}
$$

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.

$$\begin{aligned}
\Pr(3|\mathrm{I}) &= \binom{6}{3}(0.1^3)(0.9^3) = 0.01458 \\
\Pr(3|\mathrm{II}) &= \binom{6}{3}(0.2^3)(0.8^3) = 0.08192 \\
\Pr(3|\mathrm{III}) &= \binom{6}{3}(0.4^3)(0.6^3) = 0.27648
\end{aligned}$$

The probability of observing three successes is

$$\begin{aligned}
\Pr(3) &= \Pr(3|\mathrm{I})\Pr(\mathrm{I}) + \Pr(3|\mathrm{II})\Pr(\mathrm{II}) + \Pr(3|\mathrm{III})\Pr(\mathrm{III}) \\
&= 0.7(0.01458) + 0.2(0.08192) + 0.1(0.27648) = 0.054238
\end{aligned}$$

The three posterior probabilities are

$$\begin{aligned}
\Pr(\mathrm{I}|3) &= \frac{\Pr(3|\mathrm{I})\Pr(\mathrm{I})}{\Pr(3)} = \frac{0.7(0.01458)}{0.054238} = 0.18817 \\
\Pr(\mathrm{II}|3) &= \frac{\Pr(3|\mathrm{II})\Pr(\mathrm{II})}{\Pr(3)} = \frac{0.2(0.08192)}{0.054238} = 0.30208 \\
\Pr(\mathrm{III}|3) &= \frac{\Pr(3|\mathrm{III})\Pr(\mathrm{III})}{\Pr(3)} = \frac{0.1(0.27648)}{0.054238} = 0.50975
\end{aligned}$$

The posterior probability of a claim is then

$$\begin{aligned}
\Pr(\mathrm{claim}|3) &= \Pr(\mathrm{claim}|\mathrm{I})\Pr(\mathrm{I}|3) + \Pr(\mathrm{claim}|\mathrm{II})\Pr(\mathrm{II}|3) + \Pr(\mathrm{claim}|\mathrm{III})\Pr(\mathrm{III}|3) \\
&= 0.1(0.18817) + 0.2(0.30208) + 0.4(0.50975) = 0.28313
\end{aligned}$$

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 4.4.1, modern data analysis using Bayesian methods utilize computationally intensive techniques such as *MCMC* Markov 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 Chapter 16.

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.

Technical Supplement A. Gini Statistic

TS A.1. The Classic Lorenz Curve

In welfare economics, it is common to compare distributions via the **Lorenz curve**, developed by Max Otto Lorenz (Lorenz, 1905). A Lorenz curve is a graph of the proportion of a population on the horizontal axis and a distribution function of interest on the vertical axis. It is typically used to represent income distributions. When the income distribution is perfectly aligned with the population distribution, the Lorenz curve results in a 45 degree line that is known as the *line of equality*. Because the graph compares two distribution functions, one can also think of a Lorenz curve as a type of *pp* plot that was introduced in Section 4.1.2. The area between the Lorenz curve and the line of equality is a measure of the discrepancy between the income and population distributions. Two times this area is known as the **Gini index**, introduced by Corrado Gini in 1912.

Example – Classic Lorenz Curve. For an insurance example, Figure 4.12 shows a distribution of insurance losses. This figure is based on a random sample of 2000 losses. The left-hand panel shows a right-skewed histogram of losses. The right-hand panel provides the corresponding Lorenz curve, showing again a skewed distribution. For example, the arrow marks the point where 60 percent of the policyholders have 30 percent of losses. The 45 degree line is the line of equality; if each policyholder has the same loss,

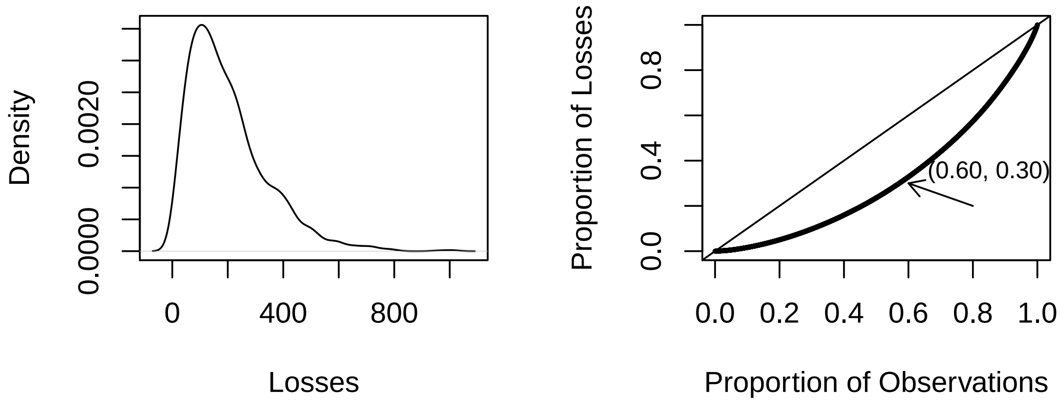


Figure 4.12: Distribution of insurance losses.

then the loss distribution would be at this line. The Gini index, twice the area between the Lorenz curve and the 45 degree line, is 37.6 percent for this data set.

TS A.2. Ordered Lorenz Curve and the Gini Index

We now introduce a modification of the classic Lorenz curve and Gini statistic that is useful in insurance applications. Specifically, we introduce an *ordered* Lorenz curve which is a graph of the distribution of losses versus premiums, where both losses and premiums are ordered by relativities. Intuitively, the relativities point towards aspects of the comparison where there is a mismatch between losses and premiums. To make the ideas concrete, we first provide some notation. We will consider $i = 1, \ldots, n$ policies. For the ith policy, let

- y_i denote the insurance loss,
- $\mathbf{x}_i$ be the set of policyholder characteristics known to the analyst,
- $P_i = P(\mathbf{x}_i)$ be the associated premium that is a function of $\mathbf{x}_i$,
- $S_i = S(\mathbf{x}_i)$ be an insurance score under consideration for rate changes, and
- $R_i = R(\mathbf{x}_i) = S(\mathbf{x}_i)/P(\mathbf{x}_i)$ is the relativity, or relative premium.

Thus, the set of information used to calculate the ordered Lorenz curve for the ith policy is (y_i, P_i, S_i, R_i).

Ordered Lorenz Curve

We now sort the set of policies based on relativities (from smallest to largest) and compute the premium and loss distributions. Using notation, the premium distribution is

$$\hat{F}_P(s) = \frac{\sum_{i=1}^n P(\mathbf{x}_i)\mathrm{I}(R_i \leq s)}{\sum_{i=1}^n P(\mathbf{x}_i)}, \tag{4.7}$$

and the loss distribution is

$$\hat{F}_L(s) = \frac{\sum_{i=1}^n y_i\mathrm{I}(R_i \leq s)}{\sum_{i=1}^n y_i}, \tag{4.8}$$

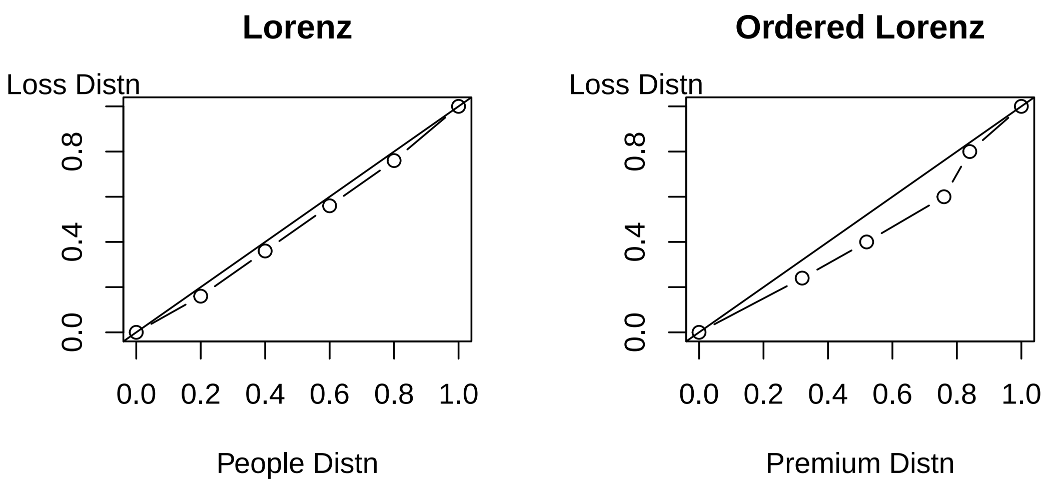


Figure 4.13: Lorenz versus Ordered Lorenz Curve

where I(·) is the indicator function, returning a 1 if the event is true and zero otherwise. The graph $\left(\hat{F}_P(s), \hat{F}_L(s)\right)$ is an **ordered Lorenz curve**.

The classic Lorenz curve shows the proportion of policyholders on the horizontal axis and the loss distribution function on the vertical axis. The ordered Lorenz curve extends the classical Lorenz curve in two ways, (1) through the ordering of risks and prices by relativities and (2) by allowing prices to vary by observation. We summarize the ordered Lorenz curve in the same way as the classic Lorenz curve using a Gini index, defined as twice the area between the curve and a 45 degree line. The analyst seeks ordered Lorenz curves that approach passing through the southeast corner (1,0); these have greater separation between the loss and premium distributions and therefore larger Gini indices.

Example – Loss Distribution.

Suppose we have $n = 5$ policyholders with experience as:

Variable	i	1	2	3	4	5	Sum
Loss	y_i	5	5	5	4	6	25
Premium	$P(\mathbf{x}_i)$	4	2	6	5	8	25
Relativity	$R(\mathbf{x}_i)$	5	4	3	2	1	

Determine the Lorenz curve and the ordered Lorenz curve.

Show Example Solution

Figure 4.13 compares the Lorenz curve to the ordered version based on this data. The left-hand panel shows the Lorenz curve. The horizontal axis is the cumulative proportion of policyholders (0, 0.2, 0.4, and so forth) and the vertical axis is the cumulative proportion of losses (0, 4/25, 9/25, and so forth). This figure shows little separation between the distributions of losses and policyholders.

The right-hand panel shows the ordered Lorenz curve. Because observations are sorted by relativities, the first point after the origin (reading from left to right) is (8/25, 6/25). The second point is (13/25, 10/25), with the pattern continuing. For the ordered Lorenz curve, the horizontal axis uses premium weights, the vertical axis uses loss weights, and both axes are ordered by relativities. From the figure, we see that there is greater separation between losses and premiums when viewed through this relativity.

Gini Index

Specifically, the Gini index can be calculated as follows. Suppose that the empirical ordered Lorenz curve is given by $\{(a_0 = 0, b_0 = 0), (a_1, b_1), \ldots, (a_n = 1, b_n = 1)\}$ for a sample of n observations. Here, we use $a_j = \hat{F}_P(R_j)$ and $b_j = \hat{F}_L(R_j)$. Then, the empirical Gini index is

$$\begin{aligned} \widehat{Gini} &= 2\sum_{j=0}^{n-1}(a_{j+1} - a_j)\left\{\frac{a_{j+1} + a_j}{2} - \frac{b_{j+1} + b_j}{2}\right\} \\ &= 1 - \sum_{j=0}^{n-1}(a_{j+1} - a_j)(b_{j+1} + b_j). \end{aligned} \tag{4.9}$$

Example – Loss Distribution: Continued. In the figure, the Gini index for the left-hand panel is 5.6%. In contrast, the Gini index for the right-hand panel is 14.9%. ☐

TS A.3. Out-of-Sample Validation

The Gini statistics based on an ordered Lorenz curve can be used for out-of-sample validation. The procedure follows:

1. Use an in-sample data set to estimate several competing models.
2. Designate an out-of-sample, or validation, data set of the form $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$.
3. Establish one of the models as the base model. Use this estimated model and explanatory variables from the validation sample to form premiums of the form $P(\mathbf{x}_i)$.
4. Pick one of the competing models. Use this estimated model and explanatory variables from the validation sample to form scores of the form $S(\mathbf{x}_i)$.
5. From the premiums and scores, develop relativities $R_i = S(\mathbf{x}_i)/P(\mathbf{x}_i)$.
6. Use the validation sample outcomes y_i to compute the Gini statistic.

Example – Out-of-Sample Validation

Suppose that we have experience from 25 states and that, for each state, we have available 500 observations that can be used to predict future losses. For simplicity, assume that the analyst knows that these losses were generated by a gamma distribution with a common shape parameter equal to 5. Unknown to the analyst, the scale parameters vary by state, from a low of 20 to 66. To compute base premiums, the analyst assumes a scale parameter that is common to all states that is to be estimated from the data. As an alternative, the analyst allows the scale parameters to vary by state and will again use the data to estimate these parameters.

An out of sample validation set of 200 from each state is available. Determine the ordered Lorenz curve and the corresponding Gini statistic to compare the two rate procedures.

Show Example Solution

Recall for the gamma distribution that the mean equals the shape times the scale or, 5 times the scale parameter, for our example. So, you can check that the maximum likelihood estimates are simple the average experience.

For our base premium, we assume a common distribution among all states. For these simulated data, the average is **P**=219.96. You can think of this common premium as based on a *community rating* principle.

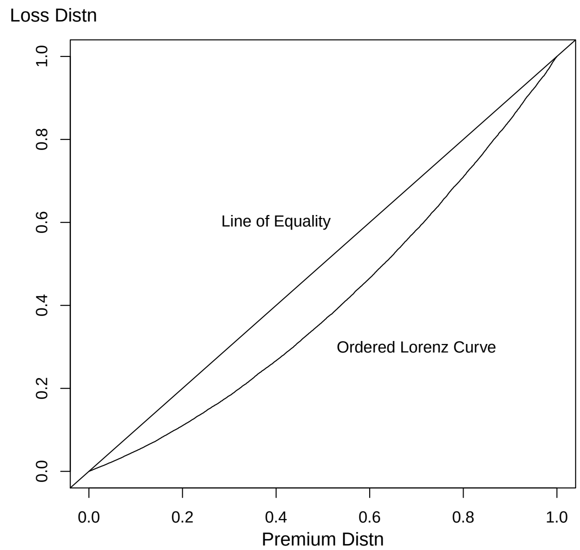
As an alternative, we use averages that are state-specific; these averages form our scores **S**. Because this illustration uses means that vary by states, we anticipate this alternative rating procedure to be preferred to the community rating procedure.

Out of sample claims were generated from the same gamma distribution, with 200 observations for each state. The following `R` code shows how to calculate the ordered Lorenz curve.

Show R Code

```
GiniCalc <- function(Claims,PIx,Sx){
   y   <- Claims/mean(Claims)
   PIx <- PIx/mean(PIx)
   Sx  <- Sx/mean(Sx)
   Rx  <- Sx/PIx       #Relativity
   n   <- length(PIx)
   origorder <- (1:n)
   PSRmat <- data.frame(cbind(PIx,Sx,Rx,y,origorder))
   PSRmatOrder <- PSRmat[order(Rx),]  #  Sort by relativity
#  PREMIUM, LOSS DFs
   DFPrem <- cumsum(PSRmatOrder$PIx)/n
   DFLoss <- cumsum(PSRmatOrder$y)/n
#  GINI CALC
   DFPremdiff <- DFPrem[2:n]-DFPrem[1:(n-1)]
   DFPremavg  <- (DFPrem[2:n]+DFPrem[1:(n-1)])/2
   DFLossavg  <- (DFLoss[2:n]+DFLoss[1:(n-1)])/2
   (Gini <- 2*crossprod(DFPremdiff,DFPremavg-DFLossavg))
#  STANDARD ERROR CALC
   h1 <- 0.5* (PSRmatOrder$PIx*DFLoss + PSRmatOrder$y*(1-DFPrem) ) #  PROJECTION CALC
   h1bar   <- mean(h1)
   sigmah  <- var(h1)
   sigmahy <- cov(h1,PSRmatOrder$y)
   sigmahpi <- cov(h1,PSRmatOrder$PIx)
   sigmay  <- var(y)
   sigmapi <- var(PIx)
   sigmaypi <- cov(PSRmatOrder$y,PSRmatOrder$PIx)
   temp1= 4*sigmah + h1bar^2*sigmay + h1bar^2*sigmapi -
          4*h1bar*sigmahy - 4*h1bar*sigmahpi +
          2*h1bar^2*sigmaypi
   sigmaGini  <- 4*temp1
   stderrGini <- sqrt(sigmaGini/n)
   check <- var(PIx-Sx)
   Gini  <- Gini*(check != 0)
   stderrGini <- stderrGini*(check != 0)
   Retmat <- data.frame(cbind(DFPrem,DFLoss))
   RetmatGini<-list(Retmat,Gini,stderrGini)
     return(RetmatGini)
}
temp=GiniCalc(Claims=Outy,PIx=Flatpred,Sx=Predvec)
Results=temp[[1]]
Gini <- temp[[2]];#Gini
stderrorGini <- temp[[3]];#Standard Error
```

For these data, the Gini index is 0.187 with a standard error equal to 0.00381. This suggests that the



state-specific alternative procedure is strongly preferred to the base community rating procedure.

Discussion

In insurance claims modeling, standard out-of-sample validation measures are not the most informative due to the high proportions of zeros (corresponding to no claim) and the skewed fat-tailed distribution of the positive values. In contrast, the Gini index works well with many zeros (see the demonstration in (Frees et al., 2014)). Moreover, the Gini index can be motivated by the economics of insurance. Intuitively, the Gini index measures the negative covariance between a policy's "profit" ($P-y$, premium minus loss) and the rank of the relativity ($\mathbf{R}$, score divided by premium). That is, the close approximation

$$\widehat{Gini} \approx -\frac{2}{n}\widehat{Cov}\left\{(P-y), rank(R)\right\}.$$

This observation leads an insurer to seek a score and resulting relativity that produces to a large Gini index. In this way, the Gini index and associated ordered Lorenz curve are useful for identifying profitable blocks of insurance business.

Unlike classical measures of association, the Gini index assumes that a premium base $\mathbf{P}$ is currently in place and seeks to assess vulnerabilities of this structure. This approach is more akin to hypothesis testing (when compared to goodness of fit) where one identifies a "null hypothesis" as the current state of the world and uses decision-making criteria/statistics to compare this with an "alternative hypothesis."

Properties of the insurance version of the Gini statistic were developed by (Frees et al., 2011) and (Frees et al., 2014). In these articles you can find formulas for the standard errors and other additional background information.

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.

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 $\mathrm{Cov}\left(X_i, X_j\right) = 0$ for all $i \neq j$), it can further be shown that

$$\begin{aligned} \mathrm{Var}(S_n) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \\ P_{S_n}(z) &= \prod_{i=1}^{n} P_{X_i}(z) \\ M_{S_n}(t) &= \prod_{i=1}^{n} M_{X_i}(t), \end{aligned}$$

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:

$$X_i = I_i \times B_i = \left\{ \begin{array}{ll} 0 \ , & \text{if} \ \ I_i = 0 \\ B_i \ , & \text{if} \ \ I_i = 1 \end{array} \right.$$

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 $\Pr(I_i = 1) = q_i$, $\mu_i = \mathrm{E}(B_i)$, and $\sigma_i^2 = \mathrm{Var}(B_i)$. It can be shown (see *Technical Supplement B.1* for details) that

$$\begin{aligned} \mathrm{E}(S_n) &= \sum_{i=1}^{n} \ q_i \ \mu_i \\ \mathrm{Var}(S_n) &= \sum_{i=1}^{n} \left(q_i \sigma_i^2 + q_i(1-q_i)\mu_i^2\right) \\ P_{S_n}(z) &= \prod_{i=1}^{n} \left(1 - q_i + q_i P_{B_i}(z)\right) \\ M_{S_n}(t) &= \prod_{i=1}^{n} \left(1 - q_i + q_i M_{B_i}(t)\right) \end{aligned}$$

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

$$\mathrm{E}\ (S_{300}) = \sum_{i=1}^{300} q_i \mu_i = 100\left\{0.05(200)\right\} + 200\left\{0.06(150)\right\} = 2,800$$

The variance of the aggregate claims is

$$\begin{aligned}
\mathrm{Var}\ (S_{300}) &= \sum_{i=1}^{300} \left(q_i \sigma_i^2 + q_i(1-q_i)\mu_i^2\right) \qquad \text{since } X_i\text{'s are independent} \\
&= 100\left\{0.05\left(\frac{400^2}{12}\right) + 0.05(1-0.05)200^2\right\} + 200\left\{0.06\left(\frac{300^2}{12}\right) + 0.06(1-0.06)150^2\right\} \\
&= 600,467.
\end{aligned}$$

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 2.3). 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 2.5.1). Using the relationship between their probability generating functions (see *Technical Supplement B.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 2.5.1 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

$$\begin{aligned}\mu_i &= \text{E}(B_i) = \text{E}(N_i^T) = \frac{\lambda}{1 - e^{-\lambda}}\\ \sigma_i^2 &= \text{Var}(B_i) = \text{Var}(N_i^T) = \frac{\lambda[1 - (\lambda + 1)e^{-\lambda}]}{(1 - e^{-\lambda})^2}\end{aligned}$$

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 2.3).

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 $\sigma_{S_n}^2 = \mathrm{Var}(S_n)$ 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

$$\begin{aligned}
\Pr(S_{300} > 3,500) &= 1 - \Pr(S_{300} \le 3,500) \\
&\approx 1 - \Phi\left(\frac{3,500-2,800}{\sqrt{600,467}}\right) = 1 - \Phi\,(0.90334) \\
&= 1 - 0.8168 = 0.1832
\end{aligned}$$

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

$$F_X^{*n}(x) = \Pr(X_1 + \cdots + X_n \le x)$$

the n-fold convolution of F_X. More generally, we can compute F_X^{*n} recursively. Begin the recursion at $k = 1$ using $F_X^{*1}(x) = F_X(x)$. 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

$$\begin{aligned}
P_{S_n}(z) &= \mathrm{E}\ \left[z^{S_n}\right] = \mathrm{E}\ \left[z^{X_1+\cdots+X_n}\right] = \mathrm{E}\ \left[z^{X_1} z^{X_2} \cdots z^{X_n}\right] \\
&= \mathrm{E}\ \left[z^{X_1}\right] \cdots \mathrm{E}\ \left[z^{X_n}\right] \qquad \text{under the independence of } X_i\text{'s} \\
&= \prod_{i=1}^{n} P_{X_i}(z) = \prod_{i=1}^{n} [1-\beta(z-1)]^{-r_i} = [1-\beta(z-1)]^{-\sum_{i=1}^{n} r_i} \ ,
\end{aligned}$$

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

$$\begin{aligned}
\Pr(S_4 < 4) &= \Pr(S_4 = 0) + \Pr(S_4 = 1) + \Pr(S_4 = 2) + \Pr(S_4 = 3) \\
&= (1+1.5)^{-4} + \frac{4(1.5)}{(1+1.5)^5} + \frac{4(5)(1.5^2)}{2(1+1.5)^6} + \frac{4(5)(6)(1.5^3)}{3!(1+1.5)^7} \\
&= 0.0256 + 0.0614 + 0.0922 + 0.1106 = 0.2898
\end{aligned}$$

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 $\mu = \mathrm{E}\left(X_i\right)$ and $\sigma^2 = \mathrm{Var}\left(X_i\right)$ 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

$$\begin{aligned}
\mathrm{Var}(S_N) &= \mathrm{E}_N[\mathrm{Var}(S_N|N)] + \mathrm{Var}_N[\mathrm{E}(S_N|N)] \\
&= \mathrm{E}_N\left[\mathrm{Var}(X_1 + \cdots + X_N)\right] + \mathrm{Var}_N\left[\mathrm{E}(X_1 + \cdots + X_N)\right] \\
&= \mathrm{E}_N\left[\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_N) + 2\mathrm{Cov}(X_1, X_2) + \cdots + \mathrm{Cov}(X_{N-1}, X_N)\right] + \mathrm{Var}_N\left[\mathrm{E}(X_1) + \cdots + \mathrm{E}(X_N)\right] \\
&= \mathrm{E}_N[N\sigma^2] + \mathrm{Var}_N[N\mu] \qquad \text{since } \mathrm{Cov}(X_i, X_j) = 0 \text{ for all } i \neq j \text{ by independence} \\
&= \sigma^2 \mathrm{E}[N] + \mu^2 \mathrm{Var}[N]
\end{aligned}$$

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

$$\begin{aligned}
M'_{S_N}(t) &= e^{\lambda(M_X(t)-1)}\ \lambda\ M'_X(t) = M_{S_N}(t)\ \lambda\ M'_X(t) \\
M''_{S_N}(t) &= M_S(t)\ \lambda\ M''_X(t) + [\ M_S(t)\ \lambda\ M'_X(t)\ ]\ \lambda\ M'_X(t)
\end{aligned}$$

Evaluating these at $t = 0$ yields

$$\mathrm{E}(S_N) = M'_{S_N}(0) = \lambda \mathrm{E}(X) = \lambda\mu$$

and

$$\begin{aligned}
M''_{S_N}(0) &= \lambda \mathrm{E}(X^2) + \lambda^2\mu^2 \\
\Rightarrow \mathrm{Var}(S_N) &= \lambda \mathrm{E}(X^2) + \lambda^2\mu^2 - (\lambda\mu)^2 = \lambda\ \mathrm{E}(X^2).
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{E}(N) &= 1(0.8) + 2(0.2) = 1.2 \\
\mathrm{E}(N^2) &= 1^2(0.8) + 2^2(0.2) = 1.6 \\
\mathrm{Var}(N) &= \mathrm{E}(N^2) - [\mathrm{E}(N)]^2 = 0.16
\end{aligned}$$

The moments of the severity distribution X are

$$\begin{aligned}
\mathrm{E}(X) &= 0(0.2) + 100(0.7) + 1000(0.1) = 170 = \mu \\
\mathrm{E}(X^2) &= 0^2(0.2) + 100^2(0.7) + 1000^2(0.1) = 107,000 \\
\mathrm{Var}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = 78,100 = \sigma^2
\end{aligned}$$

Thus, the mean and variance of the aggregate cash prize are

$$\begin{aligned}
\mathrm{E}(S_N) &= \mu \mathrm{E}(N) = 170(1.2) = 204 \\
\mathrm{Var}(S_N) &= \sigma^2 \mathrm{E}(N) + \mu^2 \mathrm{Var}(N) \\
&= 78,100(1.2) + 170^2(0.16) = 98,344
\end{aligned}$$

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:

$$\begin{aligned} F_{S_N}(s) &= \Pr\left(X_1 + \cdots + X_N \le s\right) \\ &= \mathrm{E}\left[\Pr\left(X_1 + \cdots + X_N \le s | N = n\right)\right] \\ &= \mathrm{E}\left[F_X^{*N}(s)\right] \\ &= p_0 + \sum_{n=1}^{\infty} p_n F_X^{*n}(s) \end{aligned}$$

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

$$\begin{aligned} F_{S_N}(3) &= \Pr\left(\sum_{i=1}^{N} X_i \le 3\right) = \sum_{n=0}^{\infty} \Pr\left(\sum_{i=1}^{n} X_i \le 3 | N = n\right) \Pr(N = n) \\ &= \sum_{n} F^{*n}(3)\, p_n = \sum_{n=0}^{3} F^{*n}(3) p_n \\ &= p_0 + F^{*1}(3)\ p_1 + F^{*2}(3)\ p_2 + F^{*3}(3)\ p_3 \end{aligned}$$

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

$$\begin{aligned} F^{*1}(3) &= \Pr(X \le 3) = \frac{3}{4} \\ F^{*2}(3) &= \sum_{y \le 3} F^{*1}(3-y) f(y) = F^{*1}(2) f(1) + F^{*1}(1) f(2) \\ &= \frac{1}{4}\left[F^{*1}(2) + F^{*1}(1)\right] = \frac{1}{4}\left[\Pr(X \le 2) + \Pr(X \le 1)\right] \\ &= \frac{1}{4}\left(\frac{2}{4} + \frac{1}{4}\right) = \frac{3}{16} \\ F^{*3}(3) &= \Pr(X_1 + X_2 + X_3 \le 3) = \Pr(X_1 = X_2 = X_3 = 1) = \left(\frac{1}{4}\right)^3 \end{aligned}$$

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 $X_1 + X_2 + X_3 \le 3$ is to have $X_1 = X_2 = X_3 = 1$. Additionally, for $n \ge 4$, $F^{*n}(3) = 0$ 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

$$\begin{array}{ll} \mathrm{E}(N) = 25 & \mathrm{Var}(N) = 25 \\ \mathrm{E}(X) = \frac{5+95}{2} = 50 = \mu & \mathrm{Var}(X) = \frac{(95-5)^2}{12} = 675 = \sigma^2 \end{array}$$

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 3.4. 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

$$\begin{aligned}
\mathrm{E}\left[(S_N-(j+1))_+\right]-\mathrm{E}\ \left[(S_N-j)_+\right] &= \mathrm{E}\left(S_N\wedge j\right)-\mathrm{E}\left[S\wedge(j+1)\right] \\
&= \left\{\sum_{x=0}^{j-1} x f_{S_N}(x)+j\ \Pr(S_N\geq j)\right\}-\left\{\sum_{x=0}^{j-1} x f_{S_N}(x)+j\ \Pr(S_N=j)+(j+1)\ \Pr(S_N \geq\right. \\
&= j\ \left[\Pr(S_N\geq j)-\Pr(S_N=j)\right]-(j+1)\ \Pr(S_N\geq j+1) \\
&= j\ \Pr(S_N>j)-(j+1)\ \Pr(S_N\geq j+1) \qquad (\text{note } \Pr(S_N>j)=\Pr(S_N\geq j+1)) \\
&= -\Pr(S_N\geq j+1)=-\left[1-F_{S_N}(j)\right],
\end{aligned}$$

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:

$$\begin{aligned}
\mathrm{E}\ (S_N-15)_+ &= \mathrm{E}\ (S_N-10)_+-5[1-\Pr(S_N\leq 10)] \\
&= \mathrm{E}\ (S_N-10)_+-5\Pr(S_N\geq 15) \\
&= 21.555-5(0.5496)=18.807
\end{aligned}$$

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 $\mathrm{E}(N) = \beta$, and that claim amount X is exponential with $\mathrm{E}(X) = \theta$. Recall that the *pgf* of N and the *mgf* of X are:

$$\begin{aligned} P_N(z) &= \frac{1}{1-\beta(z-1)} \\ M_X(t) &= \frac{1}{1-\theta t} \end{aligned}$$

Thus, the *mgf* of aggregate loss S_N can be expressed two ways (for details, see *Technical Supplement B.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

$$\begin{aligned} P_{N^*}(z) &= 1+\frac{\beta}{1+\beta}(z-1), \\ M_{X^*}(t) &= \frac{1}{1-\theta(1+\beta)t}. \end{aligned}$$

From (5.2), we note that S_N is also equivalent to a 2-point mixture of 0 and X^*. Specifically,

$$S_N = \left\{ \begin{array}{ll} 0 & \text{with probability } \Pr(N^*=0) = 1/(1+\beta) \\ Y^* & \text{with probability } \Pr(N^*=1) = \beta/(1+\beta) \end{array} \right. .$$

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:

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \qquad \alpha = \frac{2-p}{p-1}, \qquad \frac{1}{\gamma} = \phi(p-1)\mu^{p-1}$$

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.

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 2.3 and 2.5.1) 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:

$$\begin{aligned} f_{S_N}(s) &= \Pr(S = s) \\ &= \frac{1}{1 - af_X(0)} \left\{ [p_1 - (a+b)p_0]\, f_X(s) + \sum_{x=1}^{s \wedge m} \left(a + \frac{bx}{s}\right) f_X(x) f_{S_N}(s-x) \right\}. \end{aligned}$$

If N is in the $(a, b, 0)$ class, then $p_1 = (a + b)p_0$ and so

$$f_{S_N}(s) = \frac{1}{1 - af_X(0)} \left\{ \sum_{x=1}^{s \wedge m} \left(a + \frac{bx}{s}\right) f_X(x) f_{S_N}(s-x) \right\}.$$

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

$$
\begin{aligned}
f_{S_N}(0) &= \Pr(N=0) = p_0 = \frac{1}{5} \\
f_{S_N}(1) &= \frac{4}{5}\sum_{y=1}^{1} f_X(y) f_{S_N}(1-y) = \frac{4}{5} f_X(1) f_{S_N}(0) \\
&= \frac{4}{5}\left(\frac{1}{4}\right)\left(\frac{1}{5}\right) = \frac{1}{25} \\
f_{S_N}(2) &= \frac{4}{5}\sum_{y=1}^{2} f_X(y) f_{S_N}(2-y) = \frac{4}{5}\left[f_X(1) f_{S_N}(1) + f_X(2) f_{S_N}(0)\right] \\
&= \frac{4}{5}\left[\frac{1}{4}\left(\frac{1}{25}+\frac{1}{5}\right)\right] = \frac{4}{5}\left(\frac{6}{100}\right) = \frac{6}{125} \\
f_{S_N}(3) &= \frac{4}{5}\left[f_X(1) f_{S_N}(2) + f_X(2) f_{S_N}(1) + f_X(3) f_{S_N}(0)\right] \\
&= \frac{4}{5}\left[\frac{1}{4}\left(\frac{1}{25}+\frac{1}{5}+\frac{6}{125}\right)\right] = \frac{1}{5}\left(\frac{5+25+6}{125}\right) = 0.0576 \\
\Rightarrow \quad F_{S_N}(3) &= f_{S_N}(0) + f_{S_N}(1) + f_{S_N}(2) + f_{S_N}(3) = 0.3456
\end{aligned}
$$

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 4). 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 uniform
  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
```

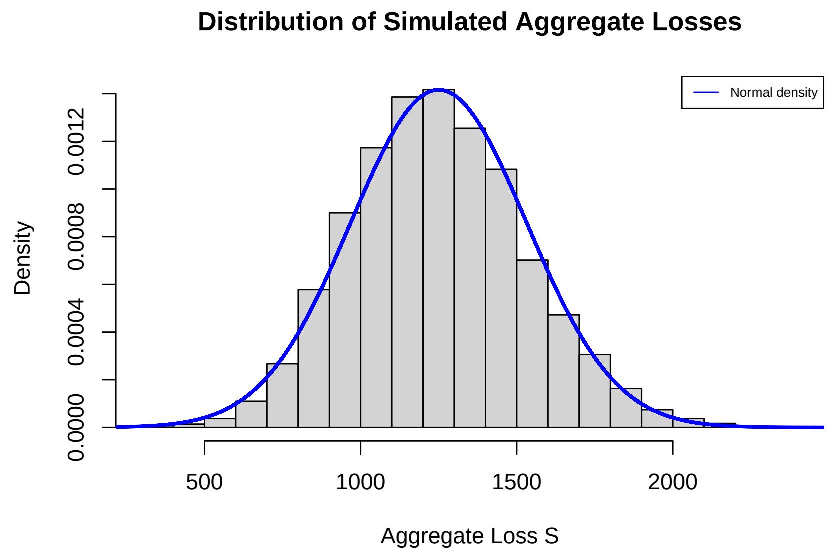
```
mean(S>2000) # Proportion of simulated observations s_j that are > 2000
```

```
## [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.

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

$$v_2 = \Pr(X > 100) = \left(\frac{150}{100+150}\right)^4 = \left(\frac{3}{5}\right)^4$$
$$\Rightarrow \lambda_2 = \lambda v_2 = 0.4\left(\frac{6}{5}\right)^4\left(\frac{3}{5}\right)^4 = 0.1075$$

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

$$0.4 = r\beta_1 = r\left(\beta v_1\right)$$
$$\Rightarrow r\beta = \frac{0.4}{v_1} = 0.4\left(\frac{6}{5}\right)^4$$

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 2.5.1). 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, $P_{N^P}(z) = P_{N^L}\left(1 + v(z-1)\right)$.

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 5.5.2). Also, define the amount of payment on a per-loss basis as

$$X^L = \begin{cases} 0 \ , & \text{if } \ X < \dfrac{d}{1+r} \\ \alpha[(1+r)X - d] \ , & \text{if } \ \dfrac{d}{1+r} \leq X < \dfrac{u}{1+r} \\ \alpha(u-d) \ , & \text{if } \ X \geq \dfrac{u}{1+r} \end{cases} ,$$

and the the amount of payment on a per-payment basis as

$$X^P = \begin{cases} \text{undefined} \ , & \text{if } \ X < \dfrac{d}{1+r} \\ \alpha[(1+r)X - d] \ , & \text{if } \ \dfrac{d}{1+r} \leq X < \dfrac{u}{1+r} \\ \alpha(u-d) \ , & \text{if } \ X \geq \dfrac{u}{1+r} \end{cases} .$$

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

$$X^L = \begin{cases} 0 & 1.5x < 100 \\ 1.5x - 100 & 1.5x \geq 100 \end{cases}$$

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

$$\begin{aligned}
\mathrm{E}\left[(X^L)^2\right] &= \frac{1}{4}\left[0^2+20^2+80^2+200^2\right] = 11,700 \\
\Rightarrow \mathrm{Var}(X^L) &= \mathrm{E}\left[(X^L)^2\right] - \left[\mathrm{E}(X^L)\right]^2 = 11,700 - 75^2 = 6,075
\end{aligned}$$

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 5.5.2) 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

$$X^P = \begin{cases} \text{undefined}, & \text{if } 1.5x < 100 \\ 1.5x - 100, & \text{if } 1.5x \geq 100 \end{cases}$$

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

$$\mathrm{E}(S) = \mathrm{E}(X^P)\ \mathrm{E}(N^P) = 200(9.7) = 1,941$$

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.

Technical Supplement B. Aggregate Loss Models

TS B.1. Individual Risk Model Properties

For the expected value of the aggregate loss under the individual risk model,

$$\begin{aligned}
\mathrm{E}(S_n) &= \sum_{i=1}^{n}~\mathrm{E}(X_i) = \sum_{i=1}^{n}~\mathrm{E}(I_i \times B_i) = \sum_{i=1}^{n}~\mathrm{E}(I_i)~~\mathrm{E}(B_i) \quad \text{from the independence of } I_i\text{'s and } B_i\text{'s} \\
&= \sum_{i=1}^{n} \mathrm{Pr}(I_i = 1)~\mu_i \quad \text{since the expectation of an indicator variable is the probability it equals 1} \\
&= \sum_{i=1}^{n}~q_i~\mu_i
\end{aligned}$$

For the variance of the aggregate loss under the individual risk model,

$$\begin{aligned}
\mathrm{Var}(S_n) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \quad \text{from the independence of } X_i\text{'s} \\
&= \sum_{i=1}^{n}~~(~\mathrm{E}\left[\mathrm{Var}(X_i|I_i)\right] + \mathrm{Var}\left[\mathrm{E}(X_i|I_i)\right]~~) \quad \text{from the conditional variance formulas} \\
&= \sum_{i=1}^{n} \left(q_i~\sigma_i^2~~+~~q_i~(1-q_i)~\mu_i^2\right)
\end{aligned}$$

To see this, note that

$$\begin{aligned}
\mathrm{E}\left[\mathrm{Var}(X_i|I_i)\right] &= \mathrm{Var}(X_i|I_i=0)~~\mathrm{Pr}(I_i=0) + \mathrm{Var}(X_i|I_i=1)~~\mathrm{Pr}(I_i=1) \\
&= q_i~\sigma_i^2 + (1-q_i)~(0) = q_i~\sigma_i^2,
\end{aligned}$$

and

$$\mathrm{Var}\left[\mathrm{E}(X_i|I_i)\right] = q_i~(1-q_i)~\mu_i^2~,$$

using the Bernoulli variance shortcut since $\mathrm{E}(X_i|I_i) = 0$ when $I_i = 0$ (probability $\mathrm{Pr}(I_i = 0) = 1 - q_i$) and $\mathrm{E}(X_i|I_i) = \mu_i$ when $I_i = 1$ (probability $\mathrm{Pr}(I_i = 1) = q_i$).

For the probability generating function of the aggregate loss under the individual risk model,

$$\begin{aligned}
P_{S_n}(z) &= \prod_{i=1}^{n} \ P_{X_i}(z) \quad \text{ from the independence of } X_i\text{'s} \\
&= \prod_{i=1}^{n} \ \mathrm{E}(z^{\ X_i}) = \prod_{i=1}^{n} \ \mathrm{E}(z^{\ I_i \times B_i}) = \mathrm{E}\left[\mathrm{E}(z^{\ I_i \times B_i}|I_i)\right] \quad \text{ from the law of iterated expectations} \\
&= \prod_{i=1}^{n} \left[\ E\left(z^{\ I_i \times B_i}|I_i = 0\right) \ \Pr(I_i = 0) + E\left(z^{\ I_i \times B_i}|I_i = 1\right) \ \Pr(I_i = 1)\ \right] \\
&= \prod_{i=1}^{n} \ [\ (1)\ (1 - q_i) + P_{B_i}(z)\ q_i\ ] = \prod_{i=1}^{n} (\ 1 - q_i + q_i\ P_{B_i}(z)\ )
\end{aligned}$$

Lastly, for the moment generating function of the aggregate loss under the individual risk model,

$$\begin{aligned}
M_{S_n}(t) &= \prod_{i=1}^{n} \ M_{X_i}(t) \quad \text{ from the independence of } X_i\text{'s} \\
&= \prod_{i=1}^{n} \ \mathrm{E}(e^{t\ X_i}) = \prod_{i=1}^{n} \ \mathrm{E}\left(e^{\ t\ (I_i \times B_i)}\right) = \prod_{i=1}^{n} \ \mathrm{E}\left[\mathrm{E}\left(e^{\ t\ (I_i \times B_i)}|I_i\right)\right] \quad \text{ from the law of iterated expectations} \\
&= \prod_{i=1}^{n} \ \left[\ \mathrm{E}\left(e^{\ t\ (I_i \times B_i)}|I_i = 0\right) \ \Pr(I_i = 0) + \mathrm{E}\left(e^{\ t\ (I_i \times B_i)}|I_i = 1\right) \ \Pr(I_i = 1)\ \right] \\
&= \prod_{i=1}^{n} \ [\ (1)\ (1 - q_i) + M_{B_i}(t)\ q_i\ ] = \prod_{i=1}^{n} (\ 1 - q_i + q_i\ M_{B_i}(t)\ )
\end{aligned}$$

TS B.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 B.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

Simulation is a computer-based, computationally intensive, method of solving difficult problems, such as analyzing business processes. 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. Performing the same analysis with a physical system is not only expensive and time-consuming but, in many cases, impossible. A drawback of simulation is that computer models are not perfect representations of business processes.

There are three basic steps for producing a simulation study:

- Generating approximately independent realizations that are uniformly distributed
- Transforming the uniformly distributed realizations to observations from a probability distribution of interest
- With the generated observations as inputs, designing a structure to produce interesting and reliable results.

Designing the structure can be a difficult step, where the degree of difficulty depends on the problem being studied. There are many resources, including this tutorial, to help the actuary with the first two steps.

6.1 Generating Independent Uniform Observations

We begin with 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." Update it 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 http://support.microsoft.com/kb/231847). 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. 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). The ability to replicate results can be a tremendous tool as you use simulation while trying to uncover patterns in a business process.

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.

6.2 Inverse Transform

With the sequence of uniform random numbers, we next transform them to a distribution of interest. Let F represent a distribution function of interest. Then, use the *inverse transform*

$$X_i = F^{-1}\left(U_i\right).$$

The result is that the sequence $\{X_i\}$ is approximately i.i.d. with distribution function F.

To interpret the result, recall that a distribution function, F, is monotonically increasing and so the inverse function, F^{-1}, is well-defined. The inverse distribution function (also known as the *quantile function*), is defined as

$$F^{-1}(y) = \inf_x \{F(x) \geq y\},$$

where "inf" stands for "infimum", or the greatest lower bound.

Inverse Transform Visualization. Here is a graph to help you visualize the inverse transform. When the random variable is continuous, the distribution function is strictly increasing and we can readily identify a unique inverse at each point of the distribution.

The inverse transform result is available when the underlying random variable is continuous, discrete or a mixture. Here is a series of examples to illustrate its scope of applications.

Exponential Distribution Example. 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.

Some Numbers. Take $\theta = 10$ and generate three random numbers to get

U	0.26321364	0.196884752	0.897884218
$X = -10\ln(1-U)$	1.32658423	0.952221285	9.909071325

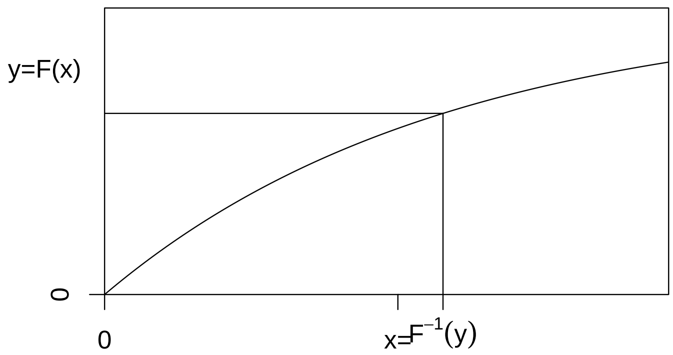


Figure 6.1: Inverse of a Distribution Function

Pareto Distribution Example. 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"?

This is easy to establish in the continuous case. Because U is a Uniform random variable on (0,1), we know that $\Pr(U \le y) = y$, for $0 \le y \le 1$. Thus,

$$\begin{aligned} \Pr(X \le x) &= \Pr(F^{-1}(U) \le x) \\ &= \Pr(F(F^{-1}(U)) \le F(x)) \\ &= \Pr(U \le F(x)) = F(x) \end{aligned}$$

as required. The key step is that \$ F(F^{-1}(u)) = u\$ for each u, which is clearly true when F is strictly increasing.

Bernoulli Distribution Example. Suppose that we wish to simulate random variables from a Bernoulli distribution with parameter $p = 0.85$. A graph of the cumulative distribution function shows that the quantile function can be written as

$$F^{-1}(y) = \left\{ \begin{array}{cc} 0 & 0 < y \le 0.85 \\ 1 & 0.85 < y \le 1.0. \end{array} \right.$$

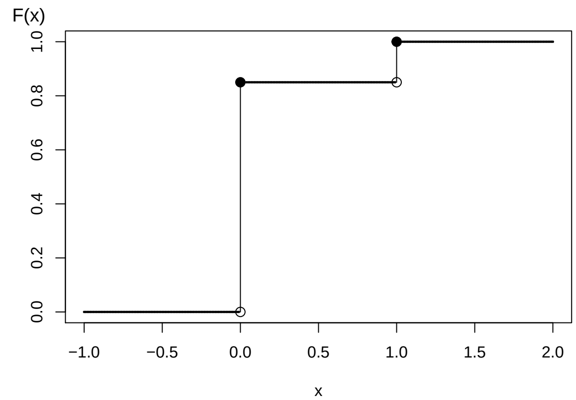


Figure 6.2: Distribution Function of a Binary Random Variable

Thus, with the inverse transform we may define

$$X = \begin{cases} 0 & 0 < U \leq 0.85 \\ 1 & 0.85 < U \leq 1.0 \end{cases}$$

Some Numbers. Generate three random numbers to get

U	0.26321364	0.196884752	0.897884218
$X = F^{-1}(U)$	0	0	1

Discrete Distribution Example. Consider the time of a machine failure in the first five years. The distribution of failure times is given as:

Time (x)	1	2	3	4	5
probability	0.1	0.2	0.1	0.4	0.2
$F(x)$	0.1	0.3	0.4	0.8	1.0

Using the graph of the distribution function, with the inverse transform we may define

$$X = \begin{cases} 1 & 0 < U \leq 0.1 \\ 2 & 0.1 < U \leq 0.3 \\ 3 & 0.3 < U \leq 0.4 \\ 4 & 0.4 < U \leq 0.8 \\ 5 & 0.8 < U \leq 1.0. \end{cases}$$

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:

$$X = \begin{cases} \text{whole life} & 0.9 < U < 1.0 \\ \text{endowment} & 0.7 \leq U < 0.9 \\ \text{term life} & 0.6 \leq U < 0.7 \\ \text{universal life} & 0.2 \leq U < 0.6 \\ \text{variable life} & 0 \leq U < 0.2 \end{cases}$$

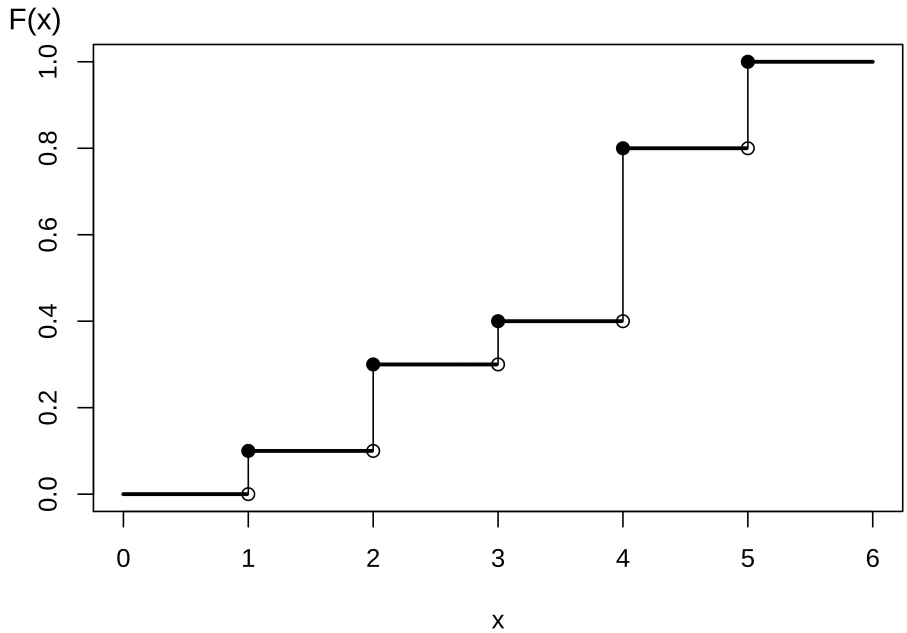


Figure 6.3: Distribution Function of a Discrete Random Variable

$$F(y) = \begin{cases} 0 & x < 0 \\ 1 - 0.3\exp(-x/10000) & x \geq 0. \end{cases}$$

From the graph, we can see that the inverse transform for generating random variables with this distribution function is

$$X = F^{-1}(U) = \begin{cases} 0 & 0 < U \leq 0.7 \\ -1000\ln(\frac{1-U}{0.3}) & 0.7 < U < 1. \end{cases}$$

As you have seen, for the discrete and mixed random variables, the key is to draw a graph of the distribution function that allows you to visualize potential values of the inverse function.

6.3 How Many Simulated Values?

There are many topics to be described in the study of simulation (and fortunately many good sources to help you). The best way to appreciate simulation is to experience it. One topic that inevitably comes up is the number of simulated trials needed to rid yourself of sampling variability so that you may focus on patterns of interest.

How many simulated values are recommended? 100? 1,000,000? We can use the central limit theorem to respond to this question. Suppose that we wish to use simulation to calculate E $h(X)$, where $h(\cdot)$ is some known function. Then, based on R simulations (replications), we get \$ X_1,…,X_R\$. From this simulated sample, we calculate a sample average

$$\overline{h}_R = \frac{1}{R}\sum_{i=1}^{R} h(X_i)$$

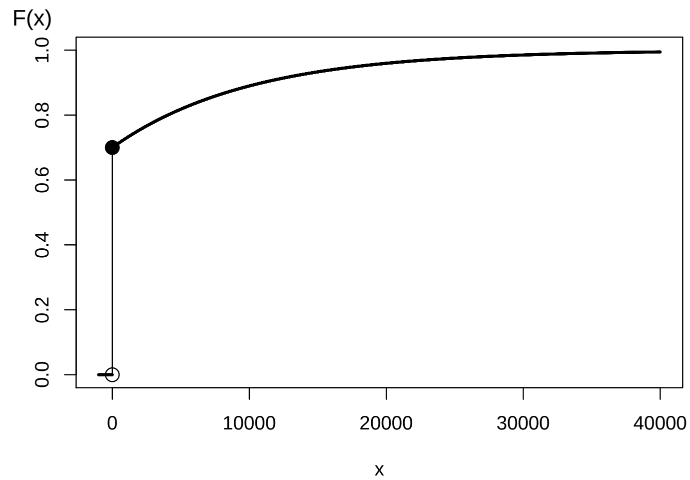


Figure 6.4: Distribution Function of a Hybrid Random Variable

and a sample standard deviation

$$s_{h,R}^2 = \frac{1}{R}\sum_{i=1}^{R}\left(h(X_i) - \overline{h}_R\right)^2.$$

So, $\overline{h}_R$ is your best estimate of E $h(X)$ and $s_{h,R}^2$ provides an indication of the uncertainty of your estimate. As one criterion for your confidence in the result, suppose that you wish to be within 1% of the mean with 95% certainty. According to the central limit theorem, your estimate should be approximately normally distributed. Thus, you should continue your simulation until

$$\frac{.01\overline{h}_R}{s_{h,R}/\sqrt{R}} \geq 1.96$$

or equivalently

$$R \geq 38,416\frac{s_{h,R}^2}{\overline{h}_R^2}.$$

This criterion is a direct application of the approximate normality (recall that 1.96 is the 97.5th percentile of the standard normal curve). Note that $\overline{h}_R$ and $s_{h,R}$ are not known in advance, so you will have to come up with estimates as you go (sequentially), either by doing a little pilot study in advance or by interrupting your procedure intermittently to see if the criterion is satisfied.

Chapter 7

Premium Calculation Fundamentals

This is a placeholder file

Chapter 8

Risk Classification

Chapter Preview. This chapter motivates the use of risk classification in insurance pricing and introduces readers to the Poisson regression as a prominent example of risk classification. In Section 8.1 we explain why insurers need to incorporate various risk characteristics, or rating factors, of individual policyholders in pricing insurance contracts. We then introduce Section 8.2 the Poisson regression as a pricing tool to achieve such premium differentials. The concept of exposure is also introduced in this section. As most rating factors are categorical, we show in Section 8.3 how the multiplicative tariff model can be incorporated in the Poisson regression model in practice, along with numerical examples for illustration.

8.1 Introduction

In this section, you learn:

- Why premiums should vary across policyholders with different risk characteristics.
- The meaning of the adverse selection spiral.
- The need for risk classification.

Through insurance contracts, the policyholders effectively transfer their risks to the insurer in exchange for premiums. For the insurer to stay in business, the premium income collected from a pool of policyholders must at least equal the benefit outgo. In general insurance products where a premium is charged for a single period, say annual, the gross insurance premium based on the equivalence principle is stated as

$$\text{Gross Premium} = \text{Expected Losses} + \text{Expected Expenses} + \text{Profit}.$$

Thus, ignoring the frictional expenses associated with the administrative expenses and the profit, the net or pure premium charged by the insurer should be equal to the expected losses occurring from the risk that is transferred from the policyholder.

If all policyholders in the insurance pool have identical risk profiles, the insurer simply charges the same premium for all policyholders because they have the same expected loss. In reality, however, the policyholders are hardly homogeneous. For example, mortality risk in life insurance depends on the characteristics of the policyholder, such as, age, sex and life style. In auto insurance, those characteristics may include age, occupation, the type or use of the car, and the area where the driver resides. The knowledge of these characteristics or variables can enhance the ability of calculating fair premiums for individual policyholders, as they can be used to estimate or predict the expected losses more accurately.

Adverse Selection. Indeed, if the insurer does not differentiate the risk characteristics of individual policyholders and simply charges the same premium to all insureds based on the average loss in the portfolio, the insurer would face *adverse selection*, a situation where individuals with a higher chance of loss are attracted in the portfolio and low-risk individuals are repelled. For example, consider a health insurance industry where smoking status is an important risk factor for mortality and morbidity. Most health insurers in the market require different premiums depending on smoking status, so smokers pay higher premiums than non-smokers, with other characteristics being identical. Now suppose that there is an insurer, we will call EquitabAll, that offers the same premium to all insureds regardless of smoking status, unlike other competitors. The net premium of EquitabAll is naturally an average mortality loss accounting for both smokers and non-smokers. That is, the net premium is a weighted average of the losses with the weights being the proportion of smokers and non-smokers, respectively. Thus it is easy to see that that a smoker would have a good incentive to purchase insurance from EquitabAll than from other insurers as the offered premium by EquitabAll is relatively lower. At the same time non-smokers would prefer buying insurance from somewhere else where lower premiums, computed from the non-smoker group only, are offered. As a result, there will be more smokers and less non-smokers in the EquitabAll's portfolio, which leads to larger-than-expected losses and hence a higher premium for insureds in the next period to cover the higher costs. With the raised new premium in the next period, non-smokers in EquitabAll will have even greater incentives to switch the insurer. As this cycle continues over time, EquitabAll would gradually retain more smokers and less non-smokers in its portfolio with the premium continually raised, eventually leading to a collapsing of business. In the literature, this phenomenon is known as the *adverse selection spiral* or death spiral. Therefore, incorporating and differentiating important risk characteristics of individuals in the insurance pricing process are a pertinent component for both the determination of fair premium for individual policyholders and the long term sustainability of insurers.

Rating Factors. In order to incorporate relevant risk characteristics of policyholders in the pricing process, insurers maintain some classification system that assigns each policyholder to one of the risk classes based on a relatively small number of risk characteristics that are deemed most relevant. These characteristics used in the classification system are called the *rating factors*, which are *a priori* variables in the sense that they are known before the contract begins (e.g., sex, health status, vehicle type, etc, are known during the underwriting). All policyholders sharing identical risk factors thus are assigned to the same risk class, and are considered homogeneous from the pricing viewpoint; the insurer consequently charge them the same premium or rate.

Regarding the risk factors and premiums, the *Actuarial Standard of Practice* (ASOP No. 12) of the Actuarial Standards Board (2018) states that the actuary should select risk characteristics that are related to expected outcomes, and that rates within a risk classification system would be considered equitable if differences in rates reflect material differences in expected cost for risk characteristics. In the process of choosing risk factors, ASOP also requires the actuary to consider the following: relationship of risk characteristics and expected outcomes, causality, objectivity, practicality, applicable law, industry practices, and business practices.

On the quantitative side, an important task for the actuary in building any risk classification is to construct a statistical model that can determine the expected loss given various rating factors of a policyholder. The standard approach is to adopt a regression model which produces the expected loss as the output when the relevant risk factors are given as the inputs. In this chapter we learn the Poisson regression, which can be used when the loss is a count variable, as a prominent example of an insurance pricing tool.

8.2 Poisson Regression Model

The Poisson regression model has been successfully used in a wide range of applications and has an advantage of allowing closed-form expressions for important quantities, which provides a informative intuition and interpretation. In this section we introduce the Poisson regression as a natural extension of the Poisson distribution.

In this section you will:

- Understand Poisson regressions as convenient tool to combine individual Poisson distributions in a unified fashion.
- Learn the concept of exposure and its importance.
- Formally learn how to formulate the Poisson regression model using indicator variables when the explanatory variables are categorical.

8.2.1 Need for Poisson Regression

Poisson Distribution

To introduce the Poisson regression, let us consider a hypothetical health insurance portfolio where all policyholders are of the same age and only one risk factor, smoking status, is relevant. Smoking status thus is a categorical variable containing two different types: smoker and non-smoker. In the statistical literature different types in a given categorical variable are commonly called *levels*. As there are two levels for the smoking status, we may denote smoker and non-smoker by level 1 and 2, respectively. Here the numbering is arbitrary and nominal. Suppose now that we are interested in pricing a health insurance where the premium for each policyholder is determined by the number of outpatient visits to doctor's office during a year. The amount of medical cost for each visit is assumed to be the same regardless of the smoking status for simplicity. Thus if we believe that smoking status is a valid risk factor in this health insurance, it is natural to consider the data separately for each smoking status. In Table 8.1 we present the data for this portfolio.

Smoker	(level 1)	Non-smoker	(level 2)		Both
Count	Observed	Count	Observed	Count	Observed
0	2213	0	6671	0	8884
1	178	1	430	1	608
2	11	2	25	2	36
3	6	3	9	3	15
4	0	4	4	4	4
5	1	5	2	5	3
Total	2409	Total	7141	Total	9550
Mean	0.0926	Mean	0.0746	Mean	0.0792

Table 8.1 : Number of visits to doctor's office in last year

As this dataset contains random counts, we try to fit a Poisson distribution for each level.

As introduced in Section 2.2.3, the probability mass function of the Poisson with mean μ is given by

$$\Pr(Y=y)=\frac{\mu^y e^{-\mu}}{y!}, \qquad y=0,1,2,\ldots \tag{8.1}$$

and $\mathrm{E}\,(Y)=\mathrm{Var}\,(Y)=\mu$. In regression contexts, it is common to use μ for mean parameters instead of the Poisson parameter λ although certainly both symbols are suitable. As we saw in Section 2.4, the *mle* of the Poisson distribution is given by the sample mean. Thus if we denote the Poisson mean parameter for each level by $\mu_{(1)}$ (smoker) and $\mu_{(2)}$ (non-smoker), we see from Table 8.1 that $\hat{\mu}_{(1)}=0.0926$ and $\hat{\mu}_{(2)}=0.0746$. This simple example shows the basic idea of risk classification. Depending on the smoking status a policyholder will have a different risk characteristic and it can be incorporated through varying Poisson parameter in computing the fair premium. In this example the ratio of expected loss frequencies is $\hat{\mu}_{(1)}/\hat{\mu}_{(2)} = 1.2402$, implying that smokers tend to visit doctor's office 24.02% times more frequently compared to non-smokers.

It is also informative to note that if the insurer charges the same premium to all policyholders regardless of the smoking status, based on the average characteristic of the portfolio, as was the case for EquitabAll described in Introduction, the expected frequency (or the premium) $\hat{\mu}$ is 0.0792, obtained from the last column of Table 8.1. It is easily verified that

$$\hat{\mu} = \left(\frac{n_1}{n_1+n_2}\right)\hat{\mu}_{(1)} + \left(\frac{n_2}{n_1+n_2}\right)\hat{\mu}_{(2)} = 0.0792, \tag{8.2}$$

where n_i is the number of observations in each level. Clearly, this premium is a weighted average of the premiums for each level with the weight equal to the proportion of the insureds in that level.

A simple Poisson regression

In the example above, we have fitted a Poisson distribution for each level separately, but we can actually combine them together in a unified fashion so that a single Poisson model can encompass both smoking and non-smoking statuses. This can be done by relating the Poisson mean parameter with the risk factor. In other words, we make the Poisson mean, which is the expected loss frequency, respond to the change in the smoking status. The conventional approach to deal with a categorical variable is to adopt indicator or dummy variables that take either 1 or 0, so that we turn the switch on for one level and off for others. Therefore we may propose to use

$$\mu = \beta_0 + \beta_1 x_1 \tag{8.3}$$

or, more commonly, a log linear form

$$\log \mu = \beta_0 + \beta_1 x_1, \tag{8.4}$$

where x_1 is an indicator variable with

$$x_1 = \begin{cases} 1 & \text{if smoker,} \\ 0 & \text{otherwise.} \end{cases} \tag{8.5}$$

We generally prefer the log linear relation (8.4) to the linear one in (8.3) to prevent undesirable events of producing negative μ values, which may happen when there are many different risk factors and levels. The setup (8.4) and (8.5) then results in different Poisson frequency parameters depending on the level in the risk factor:

$$\log \mu = \begin{cases} \beta_0 + \beta_1 \\ \beta_0 \end{cases} \quad \text{or equivalently,} \quad \mu = \begin{cases} e^{\beta_0+\beta_1} & \text{if smoker (level 1),} \\ e^{\beta_0} & \text{if non-smoker (level 2),} \end{cases} \tag{8.6}$$

achieving what we aim for. This is the simplest form of the Poisson regression. Note that we require a single indicator variable to model two levels in this case. Alternatively, it is also possible to use two indicator variables through a different coding scheme. This scheme requires dropping the intercept term so that (8.4) is modified to

$$\log \mu = \beta_1 x_1 + \beta_2 x_2, \tag{8.7}$$

where x_2 is the second indicator variable with

$$x_2 = \begin{cases} 1 & \text{if non-smoker,} \\ 0 & \text{otherwise.} \end{cases} \tag{8.8}$$

Then we have, from (8.7),

$$\log \mu = \begin{cases} \beta_1 \\ \beta_2 \end{cases} \qquad \text{or} \qquad \mu = \begin{cases} e^{\beta_1} & \text{if smoker (level 1),} \\ e^{\beta_2} & \text{if non-smoker (level 2).} \end{cases} \tag{8.9}$$

The numerical result of (8.6) is the same as (8.9) as all the coefficients are given as numbers in actual estimation, with the former setup more common in most texts; we also stick to the former.

With this Poisson regression model we can easily understand how the coefficients β_0 and β_1 are linked to the expected loss frequency in each level. According to (8.6), the Poisson mean of the smokers, $\mu_{(1)}$, is given by

$$\mu_{(1)} = e^{\beta_0+\beta_1} = \mu_{(2)}\, e^{\beta_1} \quad \text{or} \quad \mu_{(1)}/\mu_{(2)} = e^{\beta_1} \tag{8.10}$$

where $\mu_{(2)}$ is the Poisson mean for the non-smokers. This relation between the smokers and non-smokers suggests a useful way to compare the risks embedded in different levels of a given risk factor. That is, the proportional increase in the expected loss frequency of the smokers compared to that of the non-smokers is simply given by a multiplicative factor e^{β_1}. Putting another way, if we set the expected loss frequency of the non-smokers as the base value, the expected loss frequency of the smokers is obtained by applying e^{β_1} to the base value.

Dealing with multi-level case

We can readily extend the two-level case to a multi-level one where l different levels are involved for a single rating factor. For this we generally need $l-1$ indicator variables to formulate

$$\log \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_{l-1} x_{l-1}, \tag{8.11}$$

where x_k is an indicator variable that takes 1 if the policy belongs to level k and 0 otherwise, for $k = 1, 2, \ldots, l-1$. By omitting the indicator variable associated with the last level in (8.11) we effectively chose level l as the base case, but this choice is arbitrary and does not matter numerically. The resulting Poisson parameter for policies in level k then becomes, from (8.11),

$$\mu = \begin{cases} e^{\beta_0+\beta_k} & \text{if the policy belongs to level } k \text{ (k=1,2, ..., l-1),} \\ e^{\beta_0} & \text{if the policy belongs to level } l. \end{cases}$$

Thus if we denote the Poisson parameter for policies in level k by $\mu_{(k)}$, we can relate the Poisson parameter for different levels through $\mu_{(k)} = \mu_{(l)}\, e^{\beta_k}$, $k = 1, 2, \ldots, l-1$. This indicates that, just like the two-level case, the expected loss frequency of the kth level is obtained from the base value multiplied by the relative factor e^{β_k}. This relative interpretation becomes more powerful when there are many risk factors with multi-levels, and leads us to a better understanding of the underlying risk and more accurate prediction of future losses. Finally, we note that the varying Poisson mean is completely driven by the coefficient parameters β_k's, which are to be estimated from the dataset; the procedure of the parameter estimation will be discussed later in this chapter.

8.2.2 Poisson Regression

We now describe the Poisson regression in a formal and more general setting. Let us assume that there are n independent policyholders with a set of rating factors characterized by a k-variate vector[1]. The ith policyholder's rating factor is thus denoted by vector $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$, and the policyholder has recorded the loss count $y_i \in \{0, 1, 2, \ldots\}$ from the last period of loss observation, for $i = 1, \ldots, n$. In the regression literature, the values $x_{i1}, \ldots, x_{ik}$ are generally known as the *explanatory variables*, as these are measurements providing information about the variable of interest y_i. In essence, regression analysis is a method to quantify the relationship between a variable of interest and explanatory variables.

We also assume, for now, that all policyholders have the same one unit period for loss observation, or equal exposure of 1, to keep things simple; we will discuss more details on the exposure in the following subsection.

As done before, we describe the Poisson regression through its mean function. For this we first denote μ_i to be the expected loss count of the ith policyholder under the Poisson specification (8.1):

$$\mu_i = \mathrm{E}\ (y_i|\mathbf{x}_i), \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n. \tag{8.12}$$

The condition inside the expectation operation in (8.12) indicates that the loss frequency μ_i is the model output responding to the given set of risk factors or explanatory variables. In principle the conditional mean $\mathrm{E}\ (y_i|\mathbf{x}_i)$ in (8.12) can take different forms depending on how we specify the relationship between $\mathbf{x}$ and y. The standard choice for the Poisson regression is to adopt the exponential function, as we mentioned previously, so that

$$\mu_i = \mathrm{E}\ (y_i|\mathbf{x}_i) = e^{\mathbf{x}_i'\beta}, \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n. \tag{8.13}$$

Here $\beta = (\beta_0, \ldots, \beta_k)'$ is the vector of coefficients so that $\mathbf{x}_i'\beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$. The exponential function in (8.13) ensures that $\mu_i > 0$ for any set of rating factors $\mathbf{x}_i$. Often (8.13) is rewritten as a log linear form

$$\log \mu_i = \log \mathrm{E}\ (y_i|\mathbf{x}_i) = \mathbf{x}_i'\beta, \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n \tag{8.14}$$

to reveal the relationship when the right side is set as the linear form, $\mathbf{x}_i'\beta$. Again, we see that the mapping works well as both sides of (8.14), $\log \mu_i$ and $\mathbf{x}_i\beta$, can now cover the entire real values. This is the formulation of the Poisson regression, assuming that all policyholders have the same unit period of exposure. When the exposures differ among the policyholders, however, as is the case in most practical cases, we need to revise this formulation by adding exposure component as an additional term in (8.14).

8.2.3 Incorporating Exposure

Concept of Exposure

In order to determine the size of potential losses in any type of insurance, one must always know the corresponding exposure. The concept of exposure is an extremely important ingredient in insurance pricing, though we usually take it for granted. For example, when we say the expected claim frequency of a health insurance policy is 0.2, it does not mean much without the specification of the exposure such as, in this case, per month or per year. In fact, all premiums and losses need the exposure precisely specified and must be quoted accordingly; otherwise all subsequent statistical analyses and predictions will be distorted.

In the previous section we assumed the same unit of exposure across all policyholders, but this is hardly realistic in practice. In health insurance, for example, two different policyholders with different lengths of

[1] For example, if there are 3 risk factors each of which the number of levels are 2, 3 and 4, respectively, we have $k = (2-1) \times (3-1) \times (4-1) = 6$.

insurance coverage (e.g., 3 months and 12 months, respectively) could have recorded the same number of claim counts. As the expected number of claim counts would be proportional to the length of coverage, we should not treat these two policyholders' loss experiences identically in the modeling process. This motivates the need of the concept of *exposure* in the Poisson regression.

The Poisson distribution in (8.1) is parametrized via its mean. To understand the exposure, we alternatively parametrize the Poisson *pmf* in terms of the *rate* parameter λ, based on the definition of the Poisson process:

$$\Pr(Y = y) = \frac{(\lambda t)^y e^{-\lambda t}}{y!}, \qquad y = 0, 1, 2, \ldots \tag{8.15}$$

with $\mathrm{E}\ (Y) = \mathrm{Var}\ (Y) = \lambda t$. Here λ is known as the rate or intensity per unit period of the Poisson process and t represents the length of time or *exposure*, a known constant value. For given λ the Poisson distribution (8.15) produces a larger expected loss count as the exposure t gets larger. Clearly, (8.15) reduces to (8.1) when $t = 1$, which means that the mean and the rate become the same for the unit exposure, the case we considered in the previous subsection.

In principle, the exposure does not need to be measured in units of time and may represent different things depending the problem at hand. For example:

1. In health insurance, the rate may be the occurrence of a specific disease per 1,000 people and the exposure is the number of people considered in the unit of 1,000.
2. In auto insurance, the rate may be the number of accidents per year of a driver and the exposure is the length of the observed period for the driver in the unit of year.

3. For workers compensation that covers lost wages resulting from an employee's work-related injury or illness, the rate may be the probability of injury in the course of employment per dollar and the exposure is the payroll amount in dollars.
4. In marketing, the rate may be the number of customers who enter a store per hour and the exposure is the number of hours observed.

5. In civil engineering, the rate may be the number of major cracks on the paved road per 10 kms and the exposure is the length of road considered in the unit of 10 kms.

6. In credit risk modelling, the rate may be the number of default events per 1000 firms and the exposure is the number of firms under consideration in the unit of 1,000.

Actuaries may be able to use different exposure bases for a given insurable loss. For example, in auto insurance, both the number of kilometers driven and the number of months covered by insurance can be used as exposure bases. Here the former is more accurate and useful in modelling the losses from car accidents, but more difficult to measure and manage for insurers. Thus, a good exposure base may not be the theoretically best one due to various practical constraints. As a rule, an exposure base must be easy to determine, accurately measurable, legally and socially acceptable, and free from potential manipulation by policyholders.

Incorporating exposure in Poisson regression

As exposures affect the Poisson mean, constructing Poisson regressions requires us to carefully separate the rate and exposure in the modelling process. Focusing on the insurance context, let us denote the rate of the loss event of the ith policyholder by λ_i, the known exposure (the length of coverage) by m_i and the expected loss count under the given exposure by μ_i. Then the Poisson regression formulation in (8.13) and (8.14) should be revised in light of (8.15) as

$$\mu_i = \mathrm{E}\ (y_i|\mathbf{x}_i) = m_i\, \lambda_i = m_i\, e^{\mathbf{x}_i'\beta}, \qquad y_i \sim Pois(\mu_i),\, i = 1, \ldots, n, \tag{8.16}$$

which gives

$$\log \mu_i = \log m_i + \mathbf{x}_i'\beta, \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, \tag{8.17}$$

Adding $\log m_i$ in (8.17) does not pose a problem in fitting as we can always specify this as an extra explanatory variable, as it is a known constant, and fix its coefficient to 1. In the literature the log of exposure, $\log m_i$, is commonly called the **offset**.

8.2.4 Exercises

1. Regarding Table 8.1 answer the following.
   (a) Verify the mean values in the table.

   (b) Verify the number in equation (8.2).

   (c) Produce the fitted Poisson counts for each smoking status in the table.
2. In the Poisson regression formulation (8.12), consider using $\mu_i = \mathrm{E}\,(y_i|\mathbf{x}_i) = (\mathbf{x}_i'\beta)^2$, for $i = 1, \ldots, n$, instead of the exponential function. What potential issue would you have?

8.3 Categorical Variables and Multiplicative Tariff

In this section you will learn:

- The multiplicative tariff model when the rating factors are categorical.
- How to construct the Poisson regression model based on the multiplicative tariff structure.

8.3.1 Rating Factors and Tariff

In practice most rating factors in insurance are *categorical variables*, meaning that they take one of the pre-determined number of possible values. Examples of categorical variables include sex, type of cars, the driver's region of residence and occupation. Continuous variables, such as age or auto mileage, can also be grouped by bands and treated as categorical variables. Thus we can imagine that, with a small number of rating factors, there will be many policyholders falling into the same risk class, charged with the same premium. For the remaining of this chapter we assume that all rating factors are categorical variables.

To illustrate how categorical variables are used in the pricing process, we consider a hypothetical auto insurance with only two rating factors:

- Type of vehicle: Type A (personally owned) and B (owned by corporations). We use index $j = 1$ and 2 to respectively represent each level of this rating factor.
- Age band of the driver: Young (age < 25), middle ($25 \leq$ age < 60) and old age (age ≥ 60). We use index $k = 1, 2$ and 3, respectively, for this rating factor.

From this classification rule, we may create an organized table or list, such as the one shown in Table 8.2, collected from all policyholders. Clearly there are $2 \times 3 = 6$ different risk classes in total. Each row of the table shows a combination of different risk characteristics of individual policyholders. Our goal is to compute six different premiums for each of these combinations. Once the premium for each row has been determined using the given exposure and claim counts, the insurer can replace the last two columns in Table 8.2 with a single column containing the computed premiums. This new table then can serve as a manual to determine the premium for a new policyholder given the rating factors during the underwriting process. In non-life insurance, a table (or a set of tables) or list that contains each set of rating factors and the associated premium is referred to as a *tariff*. Each unique combination of the rating factors in a tariff is called a *tariff cell*; thus, in Table 8.2 the number of tariff cells is six, same as the number of risk classes.

Rating Type (j)	factors Age (k)	Exposure in year	Claim count observed
$j = 1$	$k = 1$	89.1	9
1	2	208.5	8
1	3	155.2	6
2	1	19.3	1
2	2	360.4	13
2	3	276.7	6

Table 8.2 : Loss record of the illustrative auto insurer

Let us now look at the loss information in Table 8.2 more closely. The exposure in each row represents the sum of the length of insurance coverages, or in-force times, in the unit of year, of all the policyholders in that tariff cell. Similarly the claim counts in each row is the number of claims at each cell. Naturally the exposures and claim counts vary due to the different number of drivers across the cells, as well as different in-force time periods among the drivers within each cell.

In light of the Poisson regression framework, we denote the exposure and claim count of cell (j, k) as m_{jk} and y_{jk}, respectively, and define the claim count per unit exposure as

$$z_{jk} = \frac{y_{jk}}{m_{jk}}, \qquad j = 1, 2;\ k = 1, 2, 3.$$

For example, $z_{12} = 8/208.5 = 0.03837$, meaning that a policyholder in tariff cell (1,2) would have 0.03837 accidents if insured for a full year on average. The set of z_{ij} values then corresponds to the rate parameter in the Poisson distribution (8.15) as they are the event occurrence rates per unit exposure. That is, we have $z_{jk} = \hat{\lambda}_{jk}$ where λ_{jk} is the Poisson rate parameter. Producing z_{ij} values however does not do much beyond comparing the average loss frequencies across risk classes. To fully exploit the dataset, we will construct a pricing model from Table 8.2 using the Poisson regression, for the remaining part of the chapter.

We comment that actual loss records used by insurers typically include much more risk factors, in which case the number of cells grows exponentially. The tariff would then consist of a set of tables, instead of one, separated by some of the basic rating factors, such as sex or territory.

8.3.2 Multiplicative Tariff Model

In this subsection, we introduce the multiplicative tariff model, a popular pricing structure that can be naturally used within the Poisson regression framework. The developments here are based on Table 8.2. Recall that the loss count of a policyholder is described by the Poisson regression model with rate λ and the exposure m, so that the expected loss count becomes $m\lambda$. As m is a known constant, we are essentially concerned with modelling λ, so that it responds to the change in the rating factors. Among other possible functional forms, we commonly choose the multiplicative[2] relation to model the Poisson rate λ_{jk} for rating factor (j, k):

$$\lambda_{jk} = f_0 \times f_{1j} \times f_{2k}, \qquad j = 1, 2;\ k = 1, 2, 3. \tag{8.18}$$

Here $\{f_{1j}, j = 1, 2\}$ are the parameters associated with the two levels in the first rating factor, car type, and $\{f_{2k}, k = 1, 2, 3\}$ associated with the three levels in the age band, the second rating factor. For instance, the

[2] Preferring the multiplicative form to others (e.g., additive one) was already hinted in (8.4).

Poisson rate for a mid-aged policyholder with a Type B vehicle is given by $\lambda_{22} = f_0 \times f_{12} \times f_{22}$. The first term f_0 is some base value to be discussed shortly. Thus these six parameters are understood as numerical representations of the levels within each rating factor, and are to be estimated from the dataset.

The multiplicative form (8.18) is easy to understand and use, because it clearly shows how the expected loss count (per unit exposure) changes as each rating factor varies. For example, if $f_{11} = 1$ and $f_{12} = 1.2$, then the expected loss count of a policyholder with a vehicle of type B would be 20% larger than type A, when the other factors are the same. In non-life insurance, the parameters f_{1j} and f_{2k} are known as *relativities* as they determine how much expected loss should change relative to the base value f_0. The idea of relativity is quite convenient in practice, as we can decide the premium for a policyholder by simply multiplying a series of corresponding relativities to the base value.

Dropping an existing rating factor or adding a new one is also transparent with this multiplicative structure. In addition, the insurer may easily adjust the overall premium for all policyholders by controlling the base value f_0 without changing individual relativities. However, by adopting the multiplicative form, we implicitly assume that there is no serious interaction among the risk factors.

When the multiplicative form is used we need to address an identification issue. That is, for any $c > 0$, we can write

$$\lambda_{jk} = f_0 \times \frac{f_{1j}}{c} \times c\, f_{2k}. \tag{8.19}$$

By comparing with (8.18), we see that the identical rate parameter λ_{jk} can be obtained for very different individual relativities. This over-parametrization, meaning that many different sets of parameters arrive at the identical model, obviously calls for some restriction on f_{1j} and f_{2k}. The standard practice is to make one relativity in each rating factor equal to one. This can be made arbitrarily in theory, but the standard practice is to make the relativity of most common class (base class) equals to one. We will assume that *type A vehicles* and *young drivers* to be the most common classes, that is, $f_{11} = 1$ and $f_{21} = 1$. This way all other relativities are uniquely determined. The tariff cell $(j, k) = (1, 1)$ is then called the *base tariff cell*, where the rate simply becomes $\lambda_{11} = f_0$, corresponding to the base value according to (8.18). Thus the base value f_0 is generally interpreted as the Poisson rate of the base tariff cell.

Again, (8.18) is log-transformed and rewritten as

$$\log \lambda_{jk} = \log f_0 + \log f_{1j} + \log f_{2k}, \tag{8.20}$$

as it is easier to work with in estimating process, similar to (8.14). This log linear form makes the log relativities of the base level in each rating factor equal to zero, i.e., $\log f_{11} = \log f_{21} = 0$, and leads to the following alternative, more explicit expression for (8.20):

$$\log \lambda = \begin{cases} \log f_0 + \quad 0 \quad + \quad 0 & \text{for a policy in cell } (1,1), \\ \log f_0 + \quad 0 \quad + \log f_{22} & \text{for a policy in cell } (1,2), \\ \log f_0 + \quad 0 \quad + \log f_{23} & \text{for a policy in cell } (1,3), \\ \log f_0 + \log f_{12} + \quad 0 & \text{for a policy in cell } (2,1), \\ \log f_0 + \log f_{12} + \log f_{22} & \text{for a policy in cell } (2,2), \\ \log f_0 + \log f_{12} + \log f_{23} & \text{for a policy in cell } (2,3). \end{cases} \tag{8.21}$$

This clearly shows that the Poisson rate parameter λ varies across different tariff cells, with the same log linear form used in the Poisson regression framework. In fact the reader may see that (8.21) is an extended version of the early expression (8.6) with multiple risk factors and that the log relativities now play the role of β_i parameters. Therefore all the relativities can be readily estimated via fitting a Poisson regression with a suitably chosen set of indicator variables.

8.3.3 Poisson Regression for Multiplicative Tariff

Indicator Variables for Tariff Cells

We now explain how the relativities can be incorporated in the Poisson regression. As seen early in this chapter we use indicator variables to deal with categorical variables. For our illustrative auto insurer, therefore, we define an indicator variable for the first rating factor as

$$x_1 = \begin{cases} 1 & \text{for vehicle type B,} \\ 0 & \text{otherwise.} \end{cases} \tag{8.22}$$

For the second rating factor, we employ two indicator variables for the age band, that is,

$$x_2 = \begin{cases} 1 & \text{for age band 2,} \\ 0 & \text{otherwise.} \end{cases} \tag{8.23}$$

and

$$x_3 = \begin{cases} 1 & \text{for age band 3,} \\ 0 & \text{otherwise.} \end{cases} \tag{8.24}$$

The triple (x_1, x_2, x_3) then can effectively and uniquely determine each risk class. By observing that the indicator variables associated with Type A and Age band 1 are omitted, we see that tariff cell $(j, k) = (1, 1)$ plays the role of the base cell. We emphasize that our choice of the three indicator variables above has been carefully made so that it is consistent with the choice of the base levels in the multiplicative tariff model in the previous subsection (i.e., $f_{11} = 1$ and $f_{21} = 1$).

With the proposed indicator variables we can rewrite the log rate (8.20) as

$$\log \lambda = \log f_0 + \log f_{12} \times x_1 + \log f_{22} \times x_2 + \log f_{23} \times x_3, \tag{8.25}$$

which is identical to (8.21) when each triple value is actually applied. For example, we can verify that the base tariff cell $(j, k) = (1, 1)$ corresponds to $(x_1, x_2, x_3) = (0, 0, 0)$, and in turn produces $\log \lambda = \log f_0$ or $\lambda = f_0$ in (8.25) as required.

Poisson regression for the tariff model

Under this specification, let us consider n policyholders in the portfolio with the ith policyholder's risk characteristic given by a vector of explanatory variables $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})'$, for $i = 1, \ldots, n$. We then recognize (8.25) as

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} = \mathbf{x}_i' \beta, \qquad i = 1, \ldots, n, \tag{8.26}$$

where $\beta_0, \ldots, \beta_3$ can be mapped to the corresponding log relativities in (8.25). This is exactly the same setup as in (8.17) except for the exposure component. Therefore, by incorporating the exposure in each risk class, the Poisson regression model for this multiplicative tariff model finally becomes

$$\log \mu_i = \log \lambda_i + \log m_i = \log m_i + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} = \log m_i + \mathbf{x}_i' \beta, \tag{8.27}$$

for $i = 1, \ldots, n$. As a result, the relativities are given by

$$f_0 = e^{\beta_0}, \quad f_{12} = e^{\beta_1}, \quad f_{22} = e^{\beta_2} \quad \text{and} \quad f_{23} = e^{\beta_3}, \tag{8.28}$$

with $f_{11} = 1$ and $f_{21} = 1$ from the original construction. For the actual dataset, β_i, $i = 0, 1, 2, 3$, is replaced with the *mle* b_i using the method in the technical supplement at the end of this chapter (Section 8.5).

8.3.4 Numerical Examples

We present two numerical examples of the Poisson regression. In the first example we construct a Poisson regression model from Table 8.2, which is a dataset of a hypothetical auto insurer. The second example uses an actual industry dataset with more risk factors. As our purpose is to show how the Poisson regression model can be used under a given classification rule, we are not concerned with the quality of the Poisson model fit in this chapter.

Example 8.1: Poisson regression for the illustrative auto insurer

In the last few subsections we considered a dataset of a hypothetical auto insurer with two risk factors, as given in Table 8.2. We now apply the Poisson regression model to this dataset. As done before, we have set $(j, k) = (1, 1)$ as the base tariff cell, so that $f_{11} = f_{21} = 1$. The result of the regression gives the coefficient estimates $(b_0, b_1, b_2, b_3) = (-2.3359, -0.3004, -0.7837, -1.0655)$, which in turn produces the corresponding relativities

$$f_0 = 0.0967, \quad f_{12} = 0.7405, \quad f_{22} = 0.4567 \quad \text{and} \quad f_{23} = 0.3445.$$

from the relation given in (8.28). The `R` script and the output are as follows.

Show R Code

```
> mydat1<- read.csv("eg1_v1a.csv")
> mydat1
  Vtype Agebnd Expsr Claims
1     1      1  89.1      9
2     1      2 208.5      8
3     1      3 155.2      6
4     2      1  19.3      1
5     2      2 360.4     13
6     2      3 276.7      6
> VtypeF <- relevel(factor(Vtype), ref="1") # treat Vtype as factors with 1 as base.
> AgebndF <- relevel(factor(Agebnd), ref="1") # treat Age band as factors.
> Pois_reg1 = glm(Claims ~ VtypeF + AgebndF,
                    data = mydat1, family = poisson(link = log), offset = log(Expsr) )
> Pois_reg1

Coefficients:
(Intercept)      VtypeF2     AgebndF2     AgebndF3
    -2.3359      -0.3004      -0.7837      -1.0655

Degrees of Freedom: 5 Total (i.e. Null);  2 Residual
Null Deviance:      8.774
Residual Deviance: 0.6514   AIC: 30.37
```

Example 8.2. Poisson regression for Singapore insurance claims data

This actual dataset is a subset of the data used by (Frees and Valdez, 2008). The data are from the General Insurance Association of Singapore, an organisation consisting of non-life insurers in Singapore. The data contains the number of car accidents for $n = 7,483$ auto insurance policies with several categorical explanatory variables and the exposure for each policy. The explanatory variables include four risk factors:

the type of the vehicle insured (either automobile (A) or other (O), denoted by `Vtype`), the age of the vehicle in years (`Vage`), gender of the policyholder (`Sex`) and the age of the policyholder (in years, grouped into seven categories, denoted `Age`).

Based on the data description, there are several things to remember before constructing a model. First, there are 3,842 policies with vehicle type A (automobile) and 3,641 policies with other vehicle types. However, age and sex information is available for the policies of vehicle type A only; the drivers of all other types of vehicles are recorded to be aged 21 or less with sex unspecified, except for one policy, indicating that no driver information has been collected for non-automobile vehicles. Second, type A vehicles are all classified as private vehicles and all the other types are not.

When we include these risk factors, we assume all unspecified sex to be male. As the age information is only applicable to type A vehicles, we set the model accordingly. That is, we apply the age variable only to vehicles of type A. Also we used five vehicle age bands, simplifying the original seven bands, by combining vehicle ages 0,1 and 2; the combined band is marked as level 2[3] in the data file. Thus our Poisson model has the following explicit form:

$$
\begin{aligned}
\log \mu_i = \mathbf{x}_i'\beta + \log m_i &= \beta_0 + \beta_1 I(Sex_i = M) + \sum_{t=2}^{6} \beta_t \, I(Vage_i = t) \\
&+ \sum_{t=7}^{13} \beta_t \, I(Vtype_i = A) \times I(Age_i = t-7) + \log m_i.
\end{aligned}
$$

The fitting result is given in Table 8.3, for which we have several comments.

- The claim frequency is higher for male by 17.3%, when other rating factors are held fixed. However, this may have been affected by the fact that all unspecified sex has been assigned to male.
- Regarding the vehicle age, the claim frequency gradually decreases as the vehicle gets old, when other rating factors are held fixed. The level starts from 2 for this variable but, again, the numbering is nominal and does not affect the numerical result.
- The policyholder age variable only applies to type A (automobile) vehicle, and there is no policy in the first age band. We may speculate that younger drivers less than age 21 drive their parents' cars rather than having their own because of high insurance premiums or related regulations. The missing relativity may be estimated by some interpolation or the professional judgement of the actuary. The claim frequency is the lowest for age band 3 and 4, but gets substantially higher for older age bands, a reasonable pattern seen in many auto insurance loss datasets.

We also note that there is no base level in the policyholder age variable, in the sense that no relativity is equal to 1. This is because the variable is only applicable to vehicle type A. This does not cause a problem numerically, but one may set the base relativity as follows if necessary for other purposes. Since there is no policy in age band 0, we consider band 1 as the base case. Specifically, we treat its relativity as a product of 0.918 and 1, where the former is the common relativity (that is, the common premium reduction) applied to all policies with vehicle type A and the latter is the base value for age band 1. Then the relativity of age band 2 can be seen as $0.917 = 0.918 \times 0.999$, where 0.999 is understood as the relativity for age band 2. The remaining age bands can be treated similarly.

[3] corresponding to `VAgecat1`

Rating factor	Level	Relativity in the tariff	Note
Base value		0.167	f_0
Sex	$1(F)$	1.000	Base level
	$2(M)$	1.173	
Vehicle age	$2(0-2$ yrs$)$	1.000	Base level
	$3(3-5$ yrs$)$	0.843	
	$4(6-10$ yrs$)$	0.553	
	$5(11-15$ yrs$)$	0.269	
	$6(16+$ yrs$)$	0.189	
Policyholder age	$0(0-21)$	N/A	No policy
(Only applicable to	$1(22-25)$	0.918	
vehicle type A)	$2(26-35)$	0.917	
	$3(36-45)$	0.758	
	$4(46-55)$	0.632	
	$5(56-65)$	1.102	
	$6(65+)$	1.179	

Table 8.3 : Singapore insurance claims data

Let us try several examples based on Table 8.3. Suppose a male policyholder aged 40 who owns a 7-year-old vehicle of type A. The expected claim frequency for this policyholder is then given by

$$\lambda = 0.167 \times 1.173 \times 0.553 \times 0.758 = 0.082. \tag{8.29}$$

As another example consider a female policyholder aged 60 who owns a 3-year-old vehicle of type O. The expected claim frequency for this policyholder is

$$\lambda = 0.167 \times 1 \times 0.843 = 0.141. \tag{8.30}$$

Note that for this policy the age band variable is not used as the vehicle type is not A. The `R` script is given as follows.

Show R Code

```
mydat <- read.csv("SingaporeAuto.csv",  quote = "", header = TRUE)
attach(mydat)

# create vehicle type as factor
TypeA = 1 * (VehicleType == "A")
table(VehicleType)
VtypeF <- as.character(VehicleType)
VtypeF[VtypeF != "A"] <- "O"
VtypeF = relevel(factor(VtypeF), ref="A")

# create gender as factor
Female = 1 * (SexInsured == "F" )
Sex = as.character(SexInsured)
Sex[Sex != "F"] <- "M"
SexF = relevel(factor(Sex), ref = "F")

# create driver age as factor
AgeCat = pmax(AgeCat - 1, 0)
AgeCatF = relevel(factor(AgeCat), ref = "0")
table(AgeCatF) # No policy in the first age band
```

```
# create vehicle age as factor
VAgeCatF = relevel( factor(VAgeCat), ref = "0" )
VAgecat1 = factor(VAgecat1, labels =
                    c("Vage0-2", "Vage3-5", "Vage6-10", "Vage11-15", "Vage15+") )
VAgecat1F = relevel( factor(VAgecat1), ref = "Vage0-2" )

# Poisson reg model
Pois_reg2 = glm(Clm_Count ~ SexF + TypeA:AgeCatF + VAgecat1F,
                   offset = LNWEIGHT, poisson(link = log) )
summary(Pois_reg2)

# compute relativities
exp(Pois_reg2$coefficients)

detach(mydat)
```

As a concluding remark, we comment that the Poisson regression is not the only possible count regression model. Actually, the Poisson distribution can be restrictive in the sense that it has a single parameter and its mean and the variance are always equal. There are other count regression models that allow more flexible distributional structure, such as negative binomial regressions and zero-inflated (ZI) regressions; details of these alternative regressions can be found in other texts listed in the next section.

8.4 Contributors and Further Resources

Further Reading and References

The Poisson regression is a special member of a more general regression model class known as the generalized linear model (glm). The glm develops a unified regression framework for datasets when the response variables are continuous, binary or discrete. The classical linear regression model with normal error is also a member of the glm. There are many standard statistical texts dealing with the glm, including (McCullagh and Nelder, 1989). More accessible texts are (Dobson and Barnett, 2008), (Agresti, 1996) and (Faraway, 2016). For actuarial and insurance applications of the glm see (Frees, 2009b), (De Jong and Heller, 2008). Also, (Ohlsson and Johansson, 2010) discusses the glm in non-life insurance pricing context with tariff analyses.

Contributor

- **Joseph H. T. Kim**, Yonsei University, is the principal author of the initial version of this chapter. Email: jhtkim@yonsei.ac.kr for chapter comments and suggested improvements.

8.5 Technical Supplement – Estimating Poisson Regression Models

The principles of maximum likelihood estimation (*mle*) are introduced in Sections 2.4.1 and 3.5, defined in Section 15.2.2, and theoretically developed in Chapter 17. Here we present the *mle* procedure of the Poisson regression so that the reader can see how the explanatory variables are treated in maximizing the likelihood function in the regression setting.

Maximum Likelihood Estimation for Individual Data

In the Poisson regression the varying Poisson mean is determined by parameters β_i's, as shown in (8.17). In this subsection we use the maximum likelihood method to estimate these parameters. Again, we assume that there are n policyholders and the ith policyholder is characterized by $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$ with the observed loss count y_i. Then, from (8.16) and (8.17), the log-likelihood function of vector $\beta = (\beta_0, \dots, \beta_k)$ is given by

$$\begin{aligned}\log L(\beta) = l(\beta) &= \sum_{i=1}^{n} (-\mu_i + y_i \log \mu_i - \log y_i!) \\ &= \sum_{i=1}^{n} (-m_i \exp(\mathbf{x}_i'\beta) + y_i (\log m_i + \mathbf{x}_i'\beta) - \log y_i!) \end{aligned} \tag{8.31}$$

To obtain the *mle* of $\beta = (\beta_0, \dots, \beta_k)'$, we differentiate[4] $l(\beta)$ with respect to vector β and set it to zero:

$$\left.\frac{\partial}{\partial \beta} l(\beta)\right|_{\beta = \mathbf{b}} = \sum_{i=1}^{n} (y_i - m_i \exp(\mathbf{x}_i'\mathbf{b}))\, \mathbf{x}_i = \mathbf{0}. \tag{8.32}$$

Numerically solving this equation system gives the *mle* of β, denoted by $\mathbf{b} = (b_0, b_1, \dots, b_k)'$. Note that, as $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$ is a column vector, equation (8.32) is a system of $k+1$ equations with both sides written as column vectors of size $k+1$. If we denote $\hat{\mu}_i = m_i \exp(\mathbf{x}_i'\mathbf{b})$, we can rewrite (8.32) as

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i)\, \mathbf{x}_i = \mathbf{0}. \tag{8.33}$$

Since the solution $\mathbf{b}$ satisfies this equation, it follows that the first among the array of $k+1$ equations, corresponding to the first constant element of $\mathbf{x}_i$, yields

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i) \times 1 = 0, \tag{8.34}$$

which implies that we must have

$$n^{-1} \sum_{i=1}^{n} y_i = \bar{y} = n^{-1} \sum_{i=1}^{n} \hat{\mu}_i. \tag{8.35}$$

This is an interesting property saying that the average of the individual losses, $\bar{y}$, is same as the average of the estimated values. That is, the sample mean is preserved under the fitted Poisson regression model.

Maximum Likelihood Estimation for Grouped Data

Sometimes the data are not available at the individual policy level. For example, Table 8.2 provides collective loss information for each risk class after grouping individual policies. When this is the case, y_i and m_i, the quantities needed for the *mle* calculation in (8.32), are unavailable for each i. However this does not pose a problem as long as we have the total loss counts and total exposure for each risk class.

To elaborate, let us assume that there are K different risk classes, and further that, in the kth risk class, we have n_k policies with the total exposure $m_{(k)}$ and the average loss count $\bar{y}_{(k)}$, for $k = 1, \dots, K$; the total loss count for the kth risk class is then $n_k\, \bar{y}_{(k)}$. We denote the set of indices of the policies belonging to the kth class by C_k. As all policies in a given risk class share the same risk characteristics, we may denote $\mathbf{x}_i = \mathbf{x}_{(k)}$ for all $i \in C_k$. With this notation, we can rewrite (8.32) as

[4] We use matrix derivative here.

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - m_i \exp(\mathbf{x}_i'\mathbf{b}))\, \mathbf{x}_i &= \sum_{k=1}^{K} \Big\{ \sum_{i \in C_k} (y_i - m_i \exp(\mathbf{x}_i'\mathbf{b}))\, \mathbf{x}_i \Big\} \\
&= \sum_{k=1}^{K} \Big\{ \sum_{i \in C_k} \Big(y_i - m_i \exp(\mathbf{x}_{(k)}'\mathbf{b})\Big)\, \mathbf{x}_{(k)} \Big\} \\
&= \sum_{k=1}^{K} \Big\{ \Big(\sum_{i \in C_k} y_i - \sum_{i \in C_k} m_i \exp(\mathbf{x}_{(k)}'\mathbf{b})\Big) \mathbf{x}_{(k)} \Big\} \\
&= \sum_{k=1}^{K} \Big(n_k\, \bar{y}_{(k)} - m_{(k)} \exp(\mathbf{x}_{(k)}'\mathbf{b})\Big) \mathbf{x}_{(k)} = 0. \qquad (8.36)
\end{aligned}$$

Since $n_k\, \bar{y}_{(k)}$ in (8.36) represents the total loss count for the kth risk class and $m_{(k)}$ is its total exposure, we see that for the Poisson regression the *mle* $\mathbf{b}$ is the same whether if we use the individual data or the grouped data.

Information matrix

Section 17.1 defines information matrices. Taking second derivatives to (8.31) gives the information matrix of the *mle* estimators,

$$\mathbf{I}(\beta) = -\mathrm{E}\left(\frac{\partial^2}{\partial \beta \partial \beta'} l(\beta) \right) = \sum_{i=1}^{n} m_i \exp(\mathbf{x}_i'\beta) \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i'. \qquad (8.37)$$

For actual datasets, μ_i in (8.37) is replaced with $\hat{\mu}_i = m_i \exp(\mathbf{x}_i'\mathbf{b})$ to estimate the relevant variances and covariances of the *mle* $\mathbf{b}$ or its functions.

For grouped datasets, we have

$$\mathbf{I}(\beta) = \sum_{k=1}^{K} \Big\{ \sum_{i \in C_k} m_i \exp(\mathbf{x}_i'\beta) \mathbf{x}_i \mathbf{x}_i' \Big\} = \sum_{k=1}^{K} m_{(k)} \exp(\mathbf{x}_{(k)}'\beta) \mathbf{x}_{(k)} \mathbf{x}_{(k)}'. \qquad (8.38)$$

Chapter 9

Experience Rating Using Credibility Theory

Chapter Preview. This chapter introduces credibility theory which is an important actuarial tool for estimating pure premiums, frequencies, and severities for individual risks or classes of risks. Credibility theory provides a convenient framework for combining the experience for an individual risk or class with other data to produce more stable and accurate estimates. Several models for calculating credibility estimates will be discussed including limited fluctuation, Bühlmann, Bühlmann-Straub, and nonparametric and semiparametric credibility methods. The chapter will also show a connection between credibility theory and Bayesian estimation which was introduced in Chapter 4.

9.1 Introduction to Applications of Credibility Theory

What premium should be charged to provide insurance? The answer depends upon the exposure to the risk of loss. A common method to compute an insurance premium is to rate an insured using a classification rating plan. A classification plan is used to select an insurance rate based on an insured's rating characteristics such as geographic territory, age, etc. All classification rating plans use a limited set of criteria to group insureds into a "class" and there will be variation in the risk of loss among insureds within the class.

An experience rating plan attempts to capture some of the variation in the risk of loss among insureds within a rating class by using the insured's own loss experience to complement the rate from the classification rating plan. One way to do this is to use a credibility weight Z with $0 \leq Z \leq 1$ to compute

$$\hat{R} = Z\bar{X} + (1 - Z)M,$$

$$\begin{aligned}
\hat{R} &= \text{credibility weighted rate for risk,} \\
\bar{X} &= \text{average loss for the risk over a specified time period,} \\
M &= \text{the rate for the classification group, often called the manual rate.}
\end{aligned}$$

For a large risk whose loss experience is stable from year to year, Z might be close to 1. For a smaller risk whose losses vary widely from year to year, Z may be close to 0.

Credibility theory is also used for computing rates for individual classes within a classification rating plan. When classification plan rates are being determined, some or many of the groups may not have sufficient data to produce stable and reliable rates. The actual loss experience for a group will be assigned a credibility weight Z and the complement of credibility $1 - Z$ may be given to the average experience for risk across all classes. Or, if a class rating plan is being updated, the complement of credibility may be assigned to the current class rate. Credibility theory can also be applied to the calculation of expected frequencies and severities.

Computing numeric values for Z requires analysis and understanding of the data. What are the variances in the number of losses and sizes of losses for risks? What is the variance between expected values across risks?

9.2 Limited Fluctuation Credibility

In this section, you learn how to:

- Calculate full credibility standards for number of claims, average size of claims, and aggregate losses.
- Learn how the relationship between means and variances of underlying distributions affects full credibility standards.
- Determine credibility-weight Z using the square-root partial credibility formula.

Limited fluctuation credibility, also called "classical credibility", was given this name because the method explicitly attempts to limit fluctuations in estimates for claim frequencies, severities, or losses. For example, suppose that you want to estimate the expected number of claims for a group of risks in an insurance rating class. How many risks are needed in the class to ensure that a specified level of accuracy is attained in the estimate? First the question will be considered from the perspective of how many claims are needed.

9.2.1 Full Credibility for Claim Frequency

Let N be a random variable representing the number of claims for a group of risks. The observed number of claims will be used to estimate $\mu_N = \mathrm{E}[N]$, the expected number of claims. How big does μ_N need to be to get a good estimate? One way to quantify the accuracy of the estimate would be a statement like: "The observed value of N should be within 5% of μ_N at least 90% of the time." Writing this as a mathematical expression would give $\Pr[0.95\mu_N \leq N \leq 1.05\mu_N] \geq 0.90$. Generalizing this statement by letting k replace 5% and probability p replace 0.90 produces a confidence interval

$$\Pr[(1-k)\mu_N \leq N \leq (1+k)\mu_N] \geq p. \tag{9.1}$$

The expected number of claims required for the probability on the left-hand side of (9.1) to equal p is called the **full credibility** standard.

If the expected number of claims is greater than or equal to the full credibility standard then full credibility can be assigned to the data so $Z = 1$. Usually the expected value μ_N is not known so full credibility will be assigned to the data if the actual observed value of N is greater than or equal to the full credibility standard. The k and p values must be selected and the actuary may rely on experience, judgment, and other factors in making the choices.

Subtracting μ_N from each term in (9.1) and dividing by the standard deviation σ_N of N gives

$$\Pr\left[\frac{-k\mu_N}{\sigma_N} \leq \frac{N-\mu_N}{\sigma_N} \leq \frac{k\mu_N}{\sigma_N}\right] \geq p. \tag{9.2}$$

For large values of $\mu_N = \mathrm{E}[N]$ it may be reasonable to approximate the distribution for $Z = (N - \mu_N)/\sigma_N$ with the standard normal distribution.

Let y_p be the value such that $\Pr[-y_p \leq Z \leq y_p] = \Phi(y_p) - \Phi(-y_p) = p$ where $\Phi()$ is the cumulative standard normal distribution. Because $\Phi(-y_p) = 1 - \Phi(y_p)$, the equality can be rewritten as $2\Phi(y_p) - 1 = p$. Solving for y_p gives $y_p = \Phi^{-1}((p+1)/2)$ where $\Phi^{-1}()$ is the inverse of the cumulative normal.

Equation (9.2) will be satisfied if $k\mu_N/\sigma_N \geq y_p$ assuming the normal approximation. First we will consider this inequality for the case when N has a Poisson distribution: $\Pr[N = n] = \lambda^n \mathrm{e}^{\lambda}/n!$. Because $\lambda = \mu_N = \sigma_N^2$ for the Poisson, taking square roots yields $\mu_N^{1/2} = \sigma_N$. So, $k\mu_N/\mu_N^{1/2} \geq y_p$ which is equivalent to $\mu_N \geq (y_p/k)^2$. Let's define λ_{kp} to be the value of μ_N for which equality holds. Then the full credibility standard for the Poisson distribution is

$$\lambda_{kp} = \left(\frac{y_p}{k}\right)^2 \text{with } y_p = \Phi^{-1}((p+1)/2). \tag{9.3}$$

If the expected number of claims μ_N is greater than or equal to λ_{kp} then equation (9.1) is assumed to hold and full credibility can be assigned to the data. As noted previously, because μ_N is usually unknown, full credibility is given if the observed value of N satisfies $N \geq \lambda_{kp}$.

Example 9.2.1. The full credibility standard is set so that the observed number of claims is to be within 5% of the expected value with probability $p = 0.95$. If the number of claims has a Poisson distribution find the number of claims needed for full credibility.

Show Example Solution

Solution Referring to a normal table, $y_p = \Phi^{-1}((p+1)/2) = \Phi^{-1}((0.95+1)/2)$=$\Phi^{-1}(0.975) = 1.960$. Using this value and $k = .05$ then $\lambda_{kp} = (y_p/k)^2 = (1.960/0.05)^2 = 1,536.64$. After rounding up the full credibility standard is 1,537.

If claims are not Poisson distributed then equation (9.2) does not imply (9.3). Setting the upper bound of Z in (9.2) equal to y_p gives $k\mu_N/\sigma_N = y_p$. Squaring both sides and moving everything to the right side except for one of the μ_N's gives $\mu_N = (y_p/k)^2(\sigma_N^2/\mu_N)$. This is the full credibility standard for frequency and will be denoted by n_f,

$$n_f = \left(\frac{y_p}{k}\right)^2 \left(\frac{\sigma_N^2}{\mu_N}\right) = \lambda_{kp}\left(\frac{\sigma_N^2}{\mu_N}\right). \tag{9.4}$$

This is the same equation as the Poisson full credibility standard except for the (σ_N^2/μ_N) multiplier. When the claims distribution is Poisson this extra term is one because the variance equals the mean.

Example 9.2.2. The full credibility standard is set so that the total number of claims is to be within 5% of the observed value with probability $p = 0.95$. The number of claims has a negative binomial distribution

$$\Pr(N = x) = \binom{x+r-1}{x}\left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^x$$

with $\beta = 1$. Calculate the full credibility standard.

Show Example Solution

Solution From the prior example, $\lambda_{kp} = 1,536.64$. The mean and variance for the negative binomial are $\mathrm{E}(N) = r\beta$ and $\mathrm{Var}(N) = r\beta(1+\beta)$ so $(\sigma_N^2/\mu_N) = (r\beta(1+\beta)/(r\beta)) = 1+\beta$ which equals 2 when $\beta = 1$. So, $n_f = \lambda_{kp}(\sigma_N^2/\mu_N) = 1,536.64(2) = 3,073.28$ and rounding up gives a full credibility standard of 3,074.

We see that the negative binomial distribution with $(\sigma_N^2/\mu_N) > 1$ requires more claims for full credibility than a Poission distribution for the same k and p values. The next example shows that a binomial distribution which has $(\sigma_N^2/\mu_N) < 1$ will need fewer claims for full credibility.

Example 9.2.3. The full credibility standard is set so that the total number of claims is to be within 5% of the observed value with probability $p = 0.95$. The number of claims has a binomial distribution

$$\Pr(N = x) = \binom{m}{x} q^x (1-q)^{m-x}.$$

Calculate the full credibility standard for $q = 1/4$.

Show Example Solution

Solution From the first example in this section $\lambda_{kp} = 1,536.64$. The mean and variance for a binomial are $\mathrm{E}(N) = mq$ and $\mathrm{Var}(N) = mq(1-q)$ so $(\sigma_N^2/\mu_N) = (mq(1-q)/(mq)) = 1-q$ which equals 3/4 when $q = 1/4$. So, $n_f = \lambda_{kp}(\sigma_N^2/\mu_N) = 1,536.64(3/4) = 1,152.48$ and rounding up gives a full credibility standard of 1,153.

Rather than use expected number of claims to define the full credibility standard, the number of exposures can be used for the full credibility standard. An exposure is a measure of risk. For example, one car insured for a full year would be one car-year. Two cars each insured for exactly one-half year would also result in one car-year. Car-years attempt to quantify exposure to loss. Two car-years would be expected to generate twice as many claims as one car-year if the vehicles have the same risk of loss. To translate a full credibility standard denominated in terms of number of claims to a full credibility standard denominated in exposures one needs a reasonable estimate of the expected number of claims per exposure.

Example 9.2.4. The full credibility standard should be selected so that the observed number of claims will be within 5% of the expected value with probability $p = 0.95$. The number of claims has a Poisson distribution. If one exposure is expected to have about 0.20 claims per year, find the number of exposures needed for full credibility.

Show Example Solution

Solution With $p = 0.95$ and $k = .05$, $\lambda_{kp} = (y_p/k)^2 = (1.960/0.05)^2 = 1,536.64$ claims are required for full credibility. The claims frequency rate is 0.20 claims/exposures. To convert the full credibility standard to a standard denominated in exposures the calculation is: (1,536.64 claims)/(0.20 claims/exposures) = 7,683.20 exposures. This can be rounded up to 7,684.

Frequency can be defined as the number of claims per exposure. Letting m represent number of exposures then the observed claim frequency is N/m which is used to estimate $\mathrm{E}(N/m)$:

$$\Pr[(1-k)\mathrm{E}(N/m) \leq N/m \leq (1+k)\mathrm{E}(N/m)] \geq p.$$

.

Because the number of exposures is not a random variable, $\mathrm{E}(N/m) = \mathrm{E}(N)/m = \mu_N/m$ and the prior equation becomes

$$\Pr\left[(1-k)\frac{\mu_N}{m} \leq \frac{N}{m} \leq (1+k)\frac{\mu_N}{m}\right] \geq p.$$

Mulitplying through by m results in equation (9.1) at the beginning of the section. The full credibility standards that were developed for estimating expected number of claims also apply to frequency.

9.2.2 Full Credibility for Aggregate Losses and Pure Premium

Aggregate losses are the total of all loss amounts for a risk or group of risks. Letting S represent aggregate losses then

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

The random variable N represents the number of losses and random variables $X_1, X_2, \ldots, X_N$ are the individual loss amounts. In this section it is assumed that N is independent of the loss amounts and that $X_1, X_2, \ldots, X_N$ are *iid*.

The mean and variance of S are

$$\mu_S = \mathrm{E}(S) = \mathrm{E}(N)\mathrm{E}(X) = \mu_N \mu_X \text{ and}$$

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

where X is the amount of a single loss.

Observed losses S will be used to estimate expected losses $\mu_S = \mathrm{E}(S)$. As with the frequency model in the previous section, the observed losses must be close to the expected losses as quantified in the equation

$$\Pr[(1-k)\mu_S \leq S \leq (1+k)\mu_S] \geq p.$$

After subtracting the mean and dividing by the standard deviation,

$$\Pr\left[\frac{-k\mu_S}{\sigma_S} \leq Z \leq \frac{k\mu_S}{\sigma_S}\right] \geq p$$

with $Z = (S - \mu_S)/\sigma_S$. As done in the previous section the distribution for Z is assumed to be normal and $k\mu_S/\sigma_S = y_p = \Phi^{-1}((p+1)/2)$. This equation can be rewritten as $\mu_S^2 = (y_p/k)^2\sigma_S^2$. Using the prior formulas for μ_S and σ_S^2 gives $(\mu_N \mu_X)^2 = (y_p/k)^2(\mu_N\sigma_X^2 + \mu_X^2\sigma_N^2)$. Dividing both sides by $\mu_N \mu_X^2$ and reordering terms on the right side results in a full credibility standard n_S for aggregate losses

$$n_S = \left(\frac{y_p}{k}\right)^2 \left[\left(\frac{\sigma_N^2}{\mu_N}\right) + \left(\frac{\sigma_X}{\mu_X}\right)^2\right] = \lambda_{kp}\left[\left(\frac{\sigma_N^2}{\mu_N}\right) + \left(\frac{\sigma_X}{\mu_X}\right)^2\right]. \tag{9.5}$$

Example 9.2.5. The number of claims has a Poisson distribution. Individual loss amounts are independently and identically distributed with a Pareto distribution $F(x) = 1 - [\theta/(x+\theta)]^\alpha$. The number of claims and loss amounts are independent. If observed aggregate losses should be within 5% of the expected value with probability $p = 0.95$, how many losses are required for full credibility?

Show Example Solution

Solution Because the number of claims is Poission, $(\sigma_N^2/\mu_N) = 1$. The mean of the Pareto is $\mu_X = \theta/(\alpha-1)$ and the variance is $\sigma_X^2 = \theta^2\alpha/[(\alpha-1)^2(\alpha-2)]$ so $(\sigma_X/\mu_X)^2 = \alpha/(\alpha-2)$. Combining the frequency and severity terms gives $[(\sigma_N^2/\mu_N)+(\sigma_X/\mu_X)^2] = 2(\alpha-1)/(\alpha-2)$. From a normal table $y_p = \Phi^{-1}((0.95+1)/2) = 1.960$. The full credibility standard is $n_S = (1.96/0.05)^2[2(\alpha-1)/(\alpha-2)] = 3,073.28(\alpha-1)/(\alpha-2)$. Suppose $\alpha = 3$ then $n_S = 6,146.56$ for a full credibility standard of 6,147. Note that considerably more claims are needed for full credibility for aggregate losses that frequency alone.

When the number of claims are Poisson distributed then equation (9.5) can be simplified using $(\sigma_N^2/\mu_N) = 1$. It follows that $[(\sigma_N^2/\mu_N) + (\sigma_X/\mu_X)^2] = [1 + (\sigma_X/\mu_X)^2] = [(\mu_x^2 + \sigma_X^2)/\mu_X^2] = \mathrm{E}(X^2)/\mathrm{E}(X)^2$ using the relationship $\mu_X^2 + \sigma_X^2 = \mathrm{E}(X^2)$. The full credibility standard is $n_S = \lambda_{kp}\mathrm{E}(X^2)/\mathrm{E}(X)^2$.

The pure premium PP is equal to aggregate losses S divided by exposures m: $PP = S/m$. The full credibility standard for pure premium will require

$$\Pr\left[(1-k)\mu_{PP} \leq PP \leq (1+k)\mu_{PP}\right] \geq p.$$

The number of exposures m is assumed fixed and not a random variable so $\mu_{PP} = \mathrm{E}(S/m) = \mathrm{E}(S)/m = \mu_S/m$.

$$\Pr\left[(1-k)\left(\frac{\mu_S}{m}\right) \leq \left(\frac{S}{m}\right) \leq (1+k)\left(\frac{\mu_S}{m}\right)\right] \geq p.$$

Multiplying through by exposures m returns the confidence interval for losses

$$\Pr[(1-k)\mu_S \leq S \leq (1+k)\mu_S] \geq p.$$

This means that the full credibility standard n_{PP} for the pure premium is the same as that for aggregate losses

$$n_{PP} = n_S = \lambda_{kp}\left[\left(\frac{\sigma_N^2}{\mu_n}\right) + \left(\frac{\sigma_X}{\mu_X}\right)^2\right].$$

9.2.3 Full Credibility for Severity

Let X be a random variable representing the size of one claim. Claim severity is $\mu_X = \mathrm{E}(X)$. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of n claims that will be used to estimate claim severity μ_X. The claims are assumed to be *iid*. The average value of the sample is

$$\bar{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right).$$

How big does n need to be to get a good estimate? Note that n is not a random variable whereas it is in the aggregate loss model.

In Section 9.2.1 the accuracy of an estimator was defined in terms of a confidence interval. For severity this confidence interval is

$$\Pr[(1-k)\mu_X \leq \bar{X} \leq (1+k)\mu_X] \geq p$$

where k and p need to be specified. Following the steps in Section 9.2.1, mean claim severity μ_X is subtracted from each term and the standard deviation of the claim severity estimator $\sigma_{\bar{X}}$ is divided into each term yielding

$$\Pr\left[\frac{-k\mu_X}{\sigma_{\bar{X}}} \leq Z \leq \frac{k\mu_X}{\sigma_{\bar{X}}}\right] \geq p$$

with $Z = (\bar{X} - \mu_X)/\sigma_X$. As in prior sections, it is assumed that Z is approximately normally distributed and the prior equation is satistifed if $k\mu_X/\sigma_{\bar{X}} \geq y_p$ with $y_p = \Phi^{-1}((p+1)/2)$. Because $\bar{X}$ is the average of individual claims $X_1, X_2, \ldots, X_n$, its standard deviation is equal to the standard deviation of an individual claim divided by $\sqrt{n}$: $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$. So, $k\mu_X/(\sigma_X/\sqrt{n}) \geq y_p$ and with a little algebra this can be rewritten as $n \geq (y_p/k)^2(\sigma_X/\mu_X)^2$. The full credibility standard for severity is

$$n_X = \left(\frac{y_p}{k}\right)^2 \left(\frac{\sigma_X}{\mu_X}\right)^2 = \lambda_{kp} \left(\frac{\sigma_X}{\mu_X}\right)^2. \tag{9.6}$$

Note that the term σ_X/μ_X is the coefficient of variation for an individual claim. Even though λ_{kp} is the full credibility standard for frequency given a Poisson distribution, there is no assumption about the distribution for the number of claims.

Example 9.2.6. Individual loss amounts are independently and identically distributed with a Pareto distribution $F(x) = 1 - [\theta/(x+\theta)]^\alpha$. How many claims are required for the average severity of observed claims to be within 5% of the expected severity with probability $p = 0.95$?

Show Example Solution

Solution The mean of the Pareto is $\mu_X = \theta/(\alpha-1)$ and the variance is $\sigma_X^2 = \theta^2\alpha/[(\alpha-1)^2(\alpha-2)]$ so $(\sigma_X/\mu_X)^2 = \alpha/(\alpha-2)$. From a normal table $y_p = \Phi^{-1}((0.95+1)/2) = 1.960$. The full credibility standard is $n_X = (1.96/0.05)^2[\alpha/(\alpha-2)] = 1{,}536.64\alpha/(\alpha-2)$. Suppose $\alpha = 3$ then $n_X = 4{,}609.92$ for a full credibility standard of 4,610.

9.2.4 Partial Credibility

In prior sections full credibility standards were calculated for estimating frequency (n_f), pure premium (n_{PP}), and severity (n_X) - in this section these full credibility standards will be denoted by n_0. In each case the full credibility standard was the expected number of claims required to achieve a defined level of accuracy when using empirical data to estimate an expected value. If the observed number of claims is greater than or equal to the full credibility standard then a full credibility weight $Z = 1$ is given to the data.

In limited fluctuation credibility, credibility weights Z assigned to data are

$$Z = \sqrt{\frac{n}{n_0}} \quad \text{if} \quad n < n_0 \quad \text{and} \quad Z = 1 \quad \text{for} \quad n \geq n_0$$

where n_0 is the full credibility standard. The quantity n is the number of claims for the data that is used to estimate the expected frequency, severity, or pure premium.

Example 9.2.7. The number of claims has a Poisson distribution. Individual loss amounts are independently and identically distributed with a Pareto distribution $F(x) = 1 - [\theta/(x+\theta)]^\alpha$. Assume that $\alpha = 3$. The number of claims and loss amounts are independent. The full credibility standard is that the observed pure premium should be within 5% of the expected value with probability $p = 0.95$. What credibility Z is assigned to a pure premium computed from 1,000 claims?

Show Example Solution

Solution Because the number of claims is Poisson,

$$\frac{\mathrm{E}(X^2)}{[\mathrm{E}\ X]^2} = \frac{\sigma_N^2}{\mu_N} + \left(\frac{\sigma_X}{\mu_X}\right)^2.$$

The mean of the Pareto is $\mu_X = \theta/(\alpha-1)$ and the second moment is $\mathrm{E}(X^2) = 2\theta^2/[(\alpha-1)(\alpha-2)]$ so $\mathrm{E}(X^2)/[\mathrm{E}\ X]^2 = 2(\alpha-1)/(\alpha-2)$. From a normal table $y_p = \Phi^{-1}((0.95+1)/2) = 1.960$. The full credibility standard is

$$n_{PP} = (1.96/0.05)^2[2(\alpha-1)/(\alpha-2)] = 3,073.28(\alpha-1)/(\alpha-2)$$

and if $\alpha = 3$ then $n_0 = n_{PP} = 6,146.56$ or 6,147 if rounded up. The credibility assigned to 1,000 claims is $Z = (1,000/6,147)^{1/2} = 0.40$.

Limited fluctuation credibility uses the formula $Z = \sqrt{n/n_0}$ to limit the fluctuation in the credibility-weighted estimate to match the fluctuation allowed for data with expected claims at the full credibility standard. Variance or standard deviation is used as the measure of fluctuation. Rather than derive the square-root formula an example is shown

Suppose that average claim severity is being estimated from a sample of size n that is less that the full credibility standard $n_0 = n_X$. Applying credibility theory the estimate $\hat{\mu}_X$ would be

$$\hat{\mu}_X = Z\bar{X} + (1-Z)M_X$$

with $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ and independent random variables X_i representing the sizes of individual claims. The complement of credibility is applied to M_X which could be last year's estimated average severity adjusted for inflation, the average severity for a much larger pool of risks, or some other relevant quantity selected by the actuary. It is assumed that the variance of M_X is zero or negligible. With this assumption

$$\mathrm{Var}(\hat{\mu}_X) = \mathrm{Var}(Z\bar{X}) = Z^2\mathrm{Var}(\bar{X}) = \frac{n}{n_0}\mathrm{Var}(\bar{X}).$$

Because $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ it follows that $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$ where random variable X is one claim. So,

$$\mathrm{Var}(\hat{\mu}_X) = \frac{n}{n_0}\mathrm{Var}(\bar{X}) = \frac{n}{n_0}\frac{\mathrm{Var}(X)}{n} = \frac{\mathrm{Var}(X)}{n_0}.$$

The last term is exactly the variance of a sample mean $\bar{X}$ when the sample size is equal to the full credibility standard $n_0 = n_X$.

9.3 Bühlmann Credibility

In this section, you learn how to:

- Compute a credibility-weighted estimate for the expected loss for a risk or group of risks.
- Determine the credibility Z assigned to observations.
- Calculate the values required in Bühlmann credibility including the Expected Value of the Process Variance (EPV), Variance of the Hypothetical Means (VHM) and collective mean μ.
- Recognize situations when the Bühlmann model is appropriate.

A classification rating plan groups policyholders together into classes based on risk characteristics. Although policyholders within a class have similarities, they are not identical and their expected losses will not be exactly the same. An experience rating plan can supplement a class rating plan by credibility weighting an individual policyholder's loss experience with the class rate to produce a more accurate rate for the policyholder.

In the presentation of Bühlmann credibility it is convenient to assign a risk parameter θ to each policyholder. Losses X for the policyholder will have a common distribution function $F_\theta(x)$ with mean $\mu(\theta) = \mathrm{E}(X|\theta)$ and variance $\sigma^2(\theta) = \mathrm{Var}(X|\theta)$. In the prior sentence *losses* can represent pure premiums, aggegrate losses, number of claims, claim severities, or some other measure of loss. Parameter θ can be continuous, discrete, or multivariate depending on the model.

If the policyholder had losses $x_1, \ldots, x_n$ during n observation periods then we want to find $\mathrm{E}(\mu(\theta)|x_1, \ldots, \ldots, x_n)$, the conditional expectation of $\mu(\theta)$ given $x_1, \ldots, x_n$. Another way to view this is to consider random variable X_{n+1} which is the observation during period $n+1$. Finding $\mathrm{E}(X_{n+1}|x_1, x_2, \ldots, x_n)$ is equivalent to finding $\mathrm{E}(\mu(\theta)|x_1, x_2, \ldots, x_n)$ assuming that $X_1, \ldots, X_n, X_{n+1}$ are *iid*.

The Bühlmann credibility-weighted estimate for $\mathrm{E}(\mu(\theta)|X_1, \ldots, X_n)$ for the policyholder is

$$\hat{\mu}(\theta) = Z\bar{X} + (1 - Z)\mu \tag{9.7}$$

with

$$\begin{aligned}
\theta &= \text{a risk parameter that identifies a policyholder's risk level} \\
\hat{\mu}(\theta) &= \text{estimated expected loss for a policyholder with parameter } \theta \\
& \quad\ \text{and loss experience } \bar{X} \\
\bar{X} &= (X_1 + \cdots + X_n)/n \text{ is the average of } n \text{ observations of the policyholder} \\
Z &= \text{credibility assigned to } n \text{ observations} \\
\mu &= \text{the expected loss for a randomly chosen policyholder in the class.}
\end{aligned}$$

Random variables X_j are assumed to be *iid* for $j = 1, \ldots, n$. The quantity $\bar{X}$ is the average of n observations and $\mathrm{E}(\bar{X}|\theta) = \mathrm{E}(X_j|\theta) = \mu(\theta)$.

If a policyholder is randomly chosen from the class and there is no loss information about the risk then it's expected loss is $\mu = \mathrm{E}(\mu(\theta))$ where the expectation is taken over all θ's in the class. In this situation $Z = 0$ and the expected loss is $\hat{\mu}(\theta) = \mu$ for the risk. The quantity μ can also be written as $\mu = \mathrm{E}(X_j)$ or $\mu = \mathrm{E}(\bar{X})$ and is often called the overall mean or collective mean. Note that $\mathrm{E}(X_j)$ is evaluated with the "law of total expectation": $\mathrm{E}(X_j)$=$\mathrm{E}(\mathrm{E}(X_j|\theta))$.

Example 9.3.1. The number of claims X for an insured in a class has a Poisson distribution with mean $\theta > 0$. The risk parameter θ is exponentially distributed within the class with *pdf* $f(\theta) = e^{-\theta}$. What is the expected number of claims for an insured chosen at random from the class?

Show Example Solution

Solution Random variable X is Poisson with parameter θ and $\mathrm{E}(X|\theta) = \theta$. The expected number of claims for a randomly chosen insured is $\mu = \mathrm{E}(\mu(\theta)) = \mathrm{E}(\mathrm{E}(X|\theta))$ =$\mathrm{E}(\theta) = \int_0^\infty \theta e^{-\theta} d\theta$. Integration by parts gives $\mu = 1$.

The prior example has risk parameter θ as a positive real number but the risk parameter can be a categorical variable as shown in the next example.

Example 9.3.2. For any risk (policyholder) in a population the number of losses N in a year has a Poisson distribution with parameter λ. Individual loss amounts X_i for a risk are independent of N and are *iid* with Pareto distribution $F(x) = 1 - [\theta/(x + \theta)]^\alpha$. There are three types of risks in the population as follows:

Risk Type	Percentage of Population	Poisson Parameter	Pareto Parameters
A	50%	$\lambda = 0.5$	$\theta = 1000, \alpha = 2.0$
B	30%	$\lambda = 1.0$	$\theta = 1500, \alpha = 2.0$
C	20%	$\lambda = 2.0$	$\theta = 2000, \alpha = 2.0$

If a risk is selected at random from the population, what is the expected aggregate loss in a year?

Show Example Solution

Solution The expected number of claims for a risk is E(N)=λ. The expected value for a Pareto distributed random variable is E(X)=$\theta/(\alpha-1)$. The expected value of the aggregate loss random variable $S = X_1 + \cdots + X_N$ for a risk is E(S)=E(N)E(X)=$\lambda\theta/(\alpha-1)$. The expected aggregate loss for a risk of type A is E(S_{A})=(0.5)(1000)/(2-1)=500. The expected aggregate loss for a risk selected at random from the population is E(S)=0.5[(0.5)(1000)]+0.3[(1.0)(1500)]+0.2[(2.0)(2000)]=1500.

Although formula (9.7) was introduced using experience rating as an example, the Bühlmann credibility model has wider application. Suppose that a rating plan has multiple classes. Credibility formula (9.7) can be used to determine individual class rates. The overall mean μ would be the average loss for all classes combined, $\bar{X}$ would be the experience for the individual class, and $\hat{\mu}(\theta)$ would be the estimated loss for the class.

9.3.1 Credibility Z, *EPV*, and *VHM*

When computing the credibility estimate $\hat{\mu}(\theta) = Z\bar{X} + (1-Z)\mu$, how much weight Z should go to experience $\bar{X}$ and how much weight $(1-Z)$ to the overall mean μ? In Bühlmann credibility there are three factors that need to be considered:

- How much variation is there in a single observation X_j for a selected risk? With $\bar{X} = (X_1 + \cdots + X_n)/n$ and assuming that the observations are *iid*, it follows that Var($\bar{X}|\theta$)=Var($X_j|\theta$)$/n$. For larger Var($\bar{X}|\theta$) less credibility weight Z should be given to experience $\bar{X}$. The Expected Value of the Process Variance, abbreviated *EPV*, is the expected value of Var($X_j|\theta$) across all risks:

$$EPV = \mathrm{E}(\mathrm{Var}(X_j|\theta)).$$

Because Var($\bar{X}|\theta$)=Var($X_j|\theta$)$/n$ it follows that E(Var($\bar{X}|\theta$))=EPV$/n$.

- How homogeneous is the population of risks whose experience was combined to compute the overall mean μ? If all the risks are similar in loss potential then more weight $(1-Z)$ would be given to the overall mean μ because μ is the average for a group of similar risks whose means $\mu(\theta)$ are not far apart. The homogeneity or heterogeneity of the population is measured by the Variance of the Hypothetical Means with abbreviation *VHM*:

$$VHM = \mathrm{Var}(\mathrm{E}(X_j|\theta)) = \mathrm{Var}(\mathrm{E}(\bar{X}|\theta)).$$

Note that we used $\mathrm{E}(\bar{X}|\theta) = \mathrm{E}(X_j|\theta)$ for the second equality. *How many observations n were used to compute $\bar{X}$? More observations would infer a larger Z.

Example 9.3.3. The number of claims N in a year for a risk in a population has a Poisson distribution with mean $\lambda > 0$. The risk parameters λ for the population are uniformly distributed over the interval (0,2). Calculate the EPV and *VHM* for the population.

Show Example Solution

Solution Random variable N is Poisson with parameter λ so $\mathrm{Var}(N|\lambda) = \lambda$. The Expected Value of the Process variance is EPV=E(Var($N|\lambda$))=E(λ) $= \int_0^2 \lambda \frac{1}{2} d\lambda = 1$. The Variance of the Hypothetical Means is VHM=Var(E(N|λ))= Var(λ)=E(λ^2) − (E(λ))$^2 = \int_0^2 \lambda^2 \frac{1}{2} d\lambda - (1)^2 = \frac{1}{3}$.

The Bühlmann credibility formula includes values for n, EPV, and VHM:

$$Z = \frac{n}{n+K} \quad , \quad K = \frac{EPV}{VHM}. \tag{9.8}$$

If n increases then so does Z. If the VHM increases then Z increases. If the EPV increases then Z gets smaller. Unlike limited fluctuation credibility where $Z = 1$ when the expected number of claims is greater than the full credibility standard, Z can approach but not equal 1 as the number of observations n goes to infinity.

If you multiply the numerator and denominator of the Z formula by (VHM/n) then Z can be rewritten as

$$Z = \frac{VHM}{VHM + (EPV/n)}.$$

The number of observations n is captured in the term (EPV/n). As shown in bullet (1) at the beginning of the section, E(Var($\bar{X}|\theta$))=EPV/n. As the number of observations get larger, the expected variance of $\bar{X}$ gets smaller and credibility Z increases so that more weight gets assigned to $\bar{X}$ in the credibility-weighted estimate $\hat{\mu}(\theta)$.

Example 9.3.4. Use the "law of total variance" to show that Var($\bar{X}$) = VHM + (EPV/n) and derive a formula for Z in terms of $\bar{X}$.

Show Example Solution

Solution The quantity Var($\bar{X}$) is called the unconditional variance or the total variance of $\bar{X}$. The law of total variance says

$$\mathrm{Var}(\bar{X}) = \mathrm{E}(\mathrm{Var}(\bar{X}|\theta)) + \mathrm{Var}(\mathrm{E}(\bar{X}|\theta)).$$

In bullet (1) at the beginning of this section we showed E(Var($\bar{X}|\theta$))=EPV/n. In the second bullet (2), Var(E($\bar{X}|\theta$))=VHM. Reordering the right hand side gives Var($\bar{X}$)= VHM +(EPV/n). Another way to write the formula for credibility Z is Z=Var(E($\bar{X}|\theta$))/Var($\bar{X}$). This implies (1 − Z)=E(Var($\bar{X}|\theta$))/Var($\bar{X}$).

The following long example and solution demonstrates how to compute the credibility-weighted estimate with frequency and severity data.

Example 9.3.5. For any risk in a population the number of losses N in a year has a Poisson distribution with parameter λ. Individual loss amounts X for a selected risk are independent of N and are *iid* with exponential distribution $F(x) = 1 - e^{x/\beta}$. There are three types of risks in the population as shown below. A risk was selected at random from the population and all losses were recorded over a five-year period. The total amount of losses over the five-year period was 5,000. Use Bühlmann credibility to estimate the annual expected aggregate loss for the risk.

Risk Type	Percentage of Population	Poisson Parameter	Exponential Parameter
A	50%	$\lambda = 0.5$	$\beta = 1000$
B	30%	$\lambda = 1.0$	$\beta = 1500$
C	20%	$\lambda = 2.0$	$\beta = 2000$

Show Example Solution

Solution Because individual loss amounts X are exponentially distributed, E(X)=β and Var(X)=β^2. For aggregate loss $S = X_1 + \cdots + X_N$, the mean is E(S)=E(N)E(X) and process variance is Var(S)=E(N)Var(X)+[E(X)]^{2}Var(N). With Poisson frequency and exponentially distributed loss amounts, E(S)=$\lambda\beta$ and Var(S)=$\lambda\beta^2 + \beta^2\lambda = 2\lambda\beta^2$.
Population mean μ: Risk means are μ(A)=0.5(1000)=500; μ(B)=1.0(1500)=1500; μ(C)=2.0(2000)=4000; and μ=0.50(500)+0.30(1500)+0.20(4000)=1,500.
VHM: VHM=$0.50(500-1500)^2 + 0.30(1500-1500)^2 + 0.20(4000-1500)^2$=1,750,000.
EPV: Process variances are $\sigma^2(A) = 2(0.5)(1000)^2 = 1,000,000$; $\sigma^2(B) = 2(1.0)(1500)^2 = 4,500,000$; $\sigma^2(C) = 2(2.0)(2000)^2 = 16,000,000$; and EPV=0.50(1,000,000)+0.30(4,500,000)+0.20(16,000,000)=5,050,000.
$\bar{\mathbf{X}}$: $\bar{X}_5 = 5,000/5$=1,000.
K: $K = 5,050,000/1,750,000$=2.89.
Z: There are five years of observations so $n = 5$. $Z = 5/(5 + 2.89)$=0.63.
$\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$: $\hat{\mu}(\theta) = 0.63(1,000) + (1 - 0.63)1,500 = \boxed{\mathbf{1,185.00}}$.

In real world applications of Bühlmann credibility the value of $K = EPV/VHM$ must be estimated. Sometimes a value for K is selected using judgment. A smaller K makes estimator $\hat{\mu}(\theta)$ more responsive to actual experience $\bar{X}$ whereas a larger K produces a more stable estimate by giving more weight to μ. Judgment may be used to balance responsiveness and stability. A later section in this chapter will discuss methods for determining K from data.

For a policyholder with risk parameter θ, Bühlmann credibility uses a linear approximation $\hat{\mu}(\theta) = Z\bar{X} + (1-Z)\mu$ to estimate E($\mu(\theta)|X_1, \ldots, X_n$), the expected loss for the policyholder given prior losses $X_1, \ldots, X_n$. We can rewrite this as $\hat{\mu}(\theta) = a + b\bar{X}$ which makes it obvious that the credibility estimate is a linear function of $\bar{X}$.

If E($\mu(\theta)|X_1, \ldots, X_n$) is approximated by the linear function $a + b\bar{X}$ and constants a and b are chosen so that $\mathrm{E}[(\mathrm{E}(\mu(\theta)|X_1, \ldots, X_n) - (a + b\bar{X}))^2]$ is minimized, what are a and b? The answer is $b = n/(n + K)$ and $a = (1 - b)\mu$ with $K = EPV/VHM$ and $\mu = E(\mu(\theta))$. More detail can be found in references (Buhlmann, 1967), (Buhlmann and Gisler, 2005), (Klugman et al., 2012), and (Tse, 2009).

Bühlmann credibility is also called least-squares credibility, greatest accuracy credibility, or Bayesian credibility.

9.4 Bühlmann-Straub Credibility

In this section, you learn how to:

- Compute a credibility-weighted estimate for the expected loss for a risk or group of risks using the Bühlmann-Straub model.
- Determine the credibility Z assigned to observations.
- Calculate required values including the Expected Value of the Process Variance (EPV), Variance of the Hypothetical Means (VHM) and collective mean μ.
- Recognize situations when the Bühlmann-Straub model is appropriate.

With standard Bühlmann or least-squares credibility as described in the prior section, losses $X_1, \ldots, X_n$ for a policyholder are assumed to be *iid*. If the subscripts indicate year 1, year 2 and so on up to year n, then the *iid* assumption means that the policyholder has the same exposure to loss every year. For commercial insurance this assumption is frequently violated.

Consider a commercial policyholder that uses a fleet of vehicles in its business. In year 1 there are m_1 vehicles in the fleet, m_2 vehicles in year 2, .., and m_n vehicles in year n. The exposure to loss from ownership and use of this fleet is not constant from year to year. The annual losses for the fleet are not *iid*.

Define Y_{jk} to be the loss for the k^{th} vehicle in the fleet for year j. Then, the total losses for the fleet in year j are $Y_{j1}+\cdots+Y_{jm_j}$ where we are adding up the losses for each of the m_j vehicles. In the Bühlmann-Straub model it is assumed that random variables Y_{jk} are *iid* across all vehicles and years for the policyholder. With this assumption the means $\mathrm{E}(Y_{jk}|\theta)=\mu(\theta)$ and variances $\mathrm{Var}(Y_{jk}|\theta)=\sigma^2(\theta)$ are the same for all vehicles and years. The quantity $\mu(\theta)$ is the expected loss and $\sigma^2(\theta)$ is the variance in the loss for one year for one vehicle for a policyholder with risk parameter θ.

If X_j is the average loss per unit of exposure in year j, $X_j=(Y_1+\cdots+Y_{m_j})/m_j$, then $\mathrm{E}(X_j)=\mu(\theta)$ and $\mathrm{Var}(X_j)=\sigma^2(\theta)/m_j$ for policyholder with risk parameter θ. The average loss per vehicle for the entire n-year period is

$$\bar{X}=\frac{1}{m}\sum_{j=1}^{n} m_j X_j \quad , \quad m=\sum_{j=1}^{n} m_j .$$

It follows that $\mathrm{E}(\bar{X}|\theta)=\mu(\theta)$ and $\mathrm{Var}(\bar{X}|\theta)=\sigma^2(\theta)/m$ where $\mu(\theta)$ and $\sigma^2(\theta)$ are the mean and variance for a single vehicle for one year for the policyholder.

Example 9.4.1. Prove that $\mathrm{Var}(\bar{X}|\theta)=\sigma^2(\theta)/m$ for a risk with risk parameter θ.

Show Example Solution

Solution

$$\begin{aligned}
\mathrm{Var}(\bar{X}|\theta) &= \mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^{n} m_j X_j|\theta\right) \\
&= \frac{1}{m^2}\sum_{j=1}^{n}\mathrm{Var}(m_j X_j|\theta)=\frac{1}{m^2}\sum_{j=1}^{n} m_j^2\mathrm{Var}(X_j|\theta) \\
&= \frac{1}{m^2}\sum_{j=1}^{n} m_j^2(\sigma^2(\theta)/m_j)=\frac{\sigma^2(\theta)}{m^2}\sum_{j=1}^{n} m_j=\sigma^2(\theta)/m.
\end{aligned}$$

The Bühlmann-Straub credibility estimate is:

$$\hat{\mu}(\theta)=Z\bar{X}+(1-Z)\mu \tag{9.9}$$

with

$$
\begin{aligned}
\theta &= \text{a risk parameter that identifies a policyholder's risk level} \\
\hat{\mu}(\theta) &= \text{estimated expected loss for one exposure for the policyholder} \\
&\quad\ \text{with loss experience } \bar{X} \\
\bar{X} &= \frac{1}{m}\sum_{j=1}^{n} m_j X_j \text{ is the average loss per exposure for m exposures} \\
Z &= \text{credibility assigned to } m \text{ exposures} \\
\mu &= \text{expected loss for one exposure for randomly chosen} \\
&\quad\ \text{policyholder from population.}
\end{aligned}
$$

Note that $\hat{\mu}(\theta)$ is the estimator for the expected loss for one exposure. If the policyholder has m_j exposures then the expected loss is $m_j\hat{\mu}(\theta)$.

In an example in the prior section it was shown that Z=Var(E($\bar{X}|\theta$))/Var($\bar{X}$) where $\bar{X}$ is the average loss for n observations. In equation (9.9) the $\bar{X}$ is the average loss for m exposures and the same Z formula can be used:

$$Z = \frac{\mathrm{Var}(\mathrm{E}(\bar{X}))}{\mathrm{Var}(\bar{X})} = \frac{\mathrm{Var}(\mathrm{E}(\bar{X}))}{\mathrm{E}(\mathrm{Var}(\bar{X}|\theta)) + \mathrm{Var}(\mathrm{E}(\bar{X}|\theta))}.$$

The denominator was expanded using "the law of total variance." As noted above $\mathrm{E}(\bar{X}|\theta) = \mu(\theta)$ so $\mathrm{Var}(\mathrm{E}(\bar{X}|\theta)) = \mathrm{Var}(\mu(\theta)) = VHM$. Because $\mathrm{Var}(\bar{X}|\theta) = \sigma^2(\theta)/m$ it follows that E(Var($\bar{X}|\theta$))=E($\sigma^2(\theta)$)$/m$=EPV/m. Making these substitutions and a little algebra gives

$$Z = \frac{m}{m+K} \quad , \quad K = \frac{EPV}{VHM}. \tag{9.10}$$

This is the same Z as for Bühlmann credibility except number of exposures m replaces number of years or observations n.

Example 9.4.2.
A commercial automobile policyholder had the following exposures and claims over a three-year period:

Year	Number of Vehicles	Number of Claims
1	9	5
2	12	4
3	15	4

- The number of claims in a year for each vehicle in the policyholder's fleet is Poisson distributed with the same mean (parameter) λ.
- Parameter λ is distributed among the policyholders in the population with *pdf* $f(\lambda) = 6\lambda(1-\lambda)$ with $0 < \lambda < 1$.

The policyholder has 18 vehicles in its fleet in year 4. Use Bühlmann-Straub credibility to estimate the expected number of policyholder claims in year 4.

Show Example Solution

Solution The expected number of claims for one vehicle for a randomly chosen policyholder is $\mu = \mathrm{E}(\lambda) = \int_0^1 \lambda[6\lambda(1-\lambda)]d\lambda = 1/2$. The average number of claims per vehicle for the policyholder is $\bar{X}$=13/36. The Expected Value of the Process Variance for a single vehicle is EPV=E(λ) = 1/2. The Variance of the Hypothetical Means across policyholders is VHM=Var(λ)=E(λ^2)-(E(λ))^2 $= \int_0^1 \lambda^2[6\lambda(1-\lambda)]d\lambda - (1/2)^2 = (3/10)-$

$(1/4) = (6/20) - (5/20) = 1/20$. So, $K{=}EPV/VHM{=}(1/2)/(1/20){=}10$. The number of exposures in the experience period is $m = 9{+}12{+}15 = 36$. The credibility is $Z = 36/(36{+}10) = 18/23$. The credibility-weighted estimate for the number of claims for one vehicle is $\hat{\mu}(\theta) = Z\bar{X}{+}(1{-}Z)\mu{=}(18/23)(13/36){+}(5/23)(1/2){=}9/23$. With 18 vehicles in the fleet in year 4 the expected number of claims is 18(9/23)=162/23=7.04 .

9.5 Bayesian Inference and Bühlmann

In this section, you learn how to:

- Use Bayes Theorem to determine a formula for the expected loss of a risk when given a likelihood and prior distribution.
- Determine the posterior distributions for the Gamma-Poisson and Beta-Binomial Bayesian models and compute expected values.
- Understand the connection between the Bühlmann and Bayesian estimates for the Gamma-Poisson and Beta-Binomial models.

Section 4.4 reviews Bayesian inference and it is assumed that the reader is familiar with that material. This section will compare Bayesian inference and Bühlmann credibility and show connections between the two models.

A risk with risk parameter θ has expected loss $\mu(\theta) = E(X|\theta)$ with random variable X representing pure premium, aggegrate loss, number of claims, claim severity, or some other measure of loss. If the risk had n losses $x_1, \ldots, x_n$ then $\mathrm{E}(\mu(\theta)|x_1, \ldots, x_n)$ is the conditional expectation of $\mu(\theta)$. The Bühlmann credibility formula $\hat{\mu}(\theta) = Z\bar{X}{+}(1{-}Z)\mu$ is a linear function of $\bar{X} = (x_1{+}\cdots{+}x_n)/n$ used to estimate $E(\mu(\theta)|x_1, \ldots, x_n)$.

Expectation $E(\mu(\theta)|x_1, \ldots, x_n)$ can be calculated from the conditional density function $f(x|\theta)$ and the posterior distribution $\pi(\theta|x_1, \ldots, x_n)$:

$$\begin{aligned}
\mathrm{E}(\mu(\theta)|x_1, \ldots, x_n) &= \int \mu(\theta)\pi(\theta|x_1, ..x_n)d\theta \\
\mu(\theta) &= \mathrm{E}(X|\theta) = \int x f(x|\theta)dx.
\end{aligned}$$

The posterior distribution comes from Bayes theorem

$$\pi(\theta|x_1, \ldots, x_n) = \frac{\prod_{j=1}^n f(x_j|\theta)}{f(x_1, ..x_n)}\pi(\theta).$$

The conditional density function $f(x|\theta)$ and the prior distribution $\pi(\theta)$ must be specified. The numerator on the right-hand side is called the likelihood.

9.5.1 Gamma-Poisson Model

In the Gamma-Poisson model the number of claims X has a Poisson distribution $\Pr(X = x|\lambda) = \lambda^x e^{-\lambda}/x!$ for a risk with risk parameter λ. The prior distribution for λ is gamma with $\pi(\lambda) = \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}/\Gamma(\alpha)$. (Note that a rate parameter β is being used in the gamma distribution rather than a scale parameter.) The mean of the gamma is $\mathrm{E}(\lambda) = \alpha/\beta$ and the variance is $\mathrm{Var}(\lambda) = \alpha/\beta^2$. In this section we will assume that λ is the expected number of claims per year though we could have chosen another time interval.

If a risk is selected at random from the population then the expected number of claims in a year is $\mathrm{E}(N)$=$\mathrm{E}(\mathrm{E}(N|\lambda))$=$\mathrm{E}(\lambda)$=$\alpha/\beta$. If we had no observations for the selected risk then the expected number of claims for the risk is α/β.

During n years the following number of claims by year was observed for the randomly selected risk: $x_1, \ldots, x_n$. From Bayes theorem the posterior distribution is

$$\pi(\lambda|x_1, \ldots, x_n) = \frac{\prod_{j=1}^{n} (\lambda^{x_j} e^{-\lambda}/x_j!)}{\Pr(x_1, \ldots, x_n)} \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}/\Gamma(\alpha).$$

Combining terms that have a λ and putting all other terms into constant C gives

$$\pi(\lambda|x_1, \ldots, x_n) = C\lambda^{(\alpha+\sum_{j=1}^{n} x_j)-1} e^{-(\beta+n)\lambda}.$$

This is a gamma distribution with parameters $\alpha' = \alpha + \sum_{j=1}^{n} x_i$ and $\beta' = \beta + n$. The constant must be $C = \beta'^{\alpha'}/\Gamma(\alpha')$ so that $\int_0^\infty \pi(\lambda|x_1, \ldots, x_n)d\lambda = 1$ though we do not need to know C. As explained in chapter four the gamma distribution is a conjugate prior for the Poisson distribution so the posterior distribution is also gamma.

Because the posterior distribtution is gamma the expected number of claims for the selected risk is

$$\mathrm{E}(\lambda|x_1, \ldots, x_n) = \frac{\alpha + \sum_{j=1}^{n} x_j}{\beta + n} = \frac{\alpha + \text{number of claims}}{\beta + \text{number of years}}.$$

This formula is slightly different from chapter four because β is multiplied times λ in the exponential of the gamma *pdf* whereas in chapter four λ is divided by parameter θ.

Now we will compute the Bühlmann credibility estimate for the Gamma-Poisson model. The variance for a Poisson distribution with parameter λ is λ so EPV=$\mathrm{E}(\mathrm{Var}(X|\lambda))$=$\mathrm{E}(\lambda)$=$\alpha/\beta$. The mean number claims for the risk is λ so VHM=$\mathrm{Var}(\mathrm{E}(X|\lambda))$=$\mathrm{Var}(\lambda)$=$\alpha/\beta^2$. The credibility parameter is K=EPV/VHM=$(\alpha/\beta)/(\alpha/\beta^2) = \beta$. The overall mean is $\mathrm{E}(\mathrm{E}(X|\lambda))$=$\mathrm{E}(\lambda)$=$\alpha/\beta$. The sample mean is $\bar{X} = (\sum_{j=1}^{n} x_j)/n$. The credibility-weighted estimate for the expected number of claims for the risk is

$$\hat{\mu} = \frac{n}{n+\beta}\frac{\sum_{j=1}^{n} x_j}{n} + (1 - \frac{n}{n+\beta})\frac{\alpha}{\beta} = \frac{\alpha + \sum_{j=1}^{n} x_j}{\beta + n}$$

For the Gamma-Poisson model the Bühlmann credibility estimate equals the Bayesian analysis answer.

9.5.2 Exact Credibility

For the Gamma-Poisson claims model the Bühlmann credibility estimate for the expected number of claims exactly matches the Bayesian answer. The term *exact credibility* is applied in this situation. Exact credibility may occur if the probability distribution for X_j is in the linear exponential family and the prior distribution is a conjugate prior. Besides the Gamma-Poisson model other examples include Gamma-Exponential, Normal-Normal, and Beta-Binomial. More information about exact credibility can be found in (Buhlmann and Gisler, 2005), (Klugman et al., 2012), and (Tse, 2009).

The beta-binomial model is useful for modeling the probability of an event. Assume that random variable X is the number of successes in n trials and that X has a binomial distribution $\Pr(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}$. In the beta-binomial model the prior distribution for probability p is a beta distribution with *pdf*

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 < p < 1, \alpha > 0, \beta > 0.$$

The posterior distribution for p given outcome x is

$$\pi(p|x) = \frac{\binom{n}{x} p^x (1-p)^{n-x}}{\Pr(x)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

Combining terms that have a p and putting everything else into the constant C yields

$$\pi(p|x) = C p^{\alpha+x-1}(1-p)^{\beta+(n-x)-1}.$$

This is a beta distribtuion with new parameters $\alpha' = \alpha + x$ and $\beta' = \beta + (n - x)$. The constant must be $C = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}$.

The mean for the beta distribution with parameters α and β is $\mathrm{E}(p) = \frac{\alpha}{\alpha+\beta}$. Given x successes in n trials in the beta-binomial model the mean of the posterior distribution is $\mathrm{E}(p|x) = \frac{\alpha+x}{\alpha+\beta+n}$. As the number of trials n and successes x increase, the expected value of p approaches x/n. The Bühlmann credibility estimate for $\mathrm{E}(p|x)$ is exactly the same as shown in the following example.

Example 9.5.1 The probability that a coin toss will yield heads is p. The prior distribution for probability p is beta with parameters α and β. On n tosses of the coin there were exactly x heads. Use Bühlmann credibility to estimate the expected value of p.

Show Example Solution

Solution Define random variables Y_j such that $Y_j = 1$ if the j^{th} coin toss is heads and $Y_j = 0$ if tails for $j = 1, \ldots, n$. Random variables Y_j are *iid* with $\Pr[Y = 1|p] = p$ and $\Pr[Y = 0|p] = 1-p$ The number of heads in n tosses can be represented by the random variable $X = Y_1 + \cdots + Y_n$. We want to estimate $p = E[Y_j]$ using Bühlmann credibility: $\hat{p} = Z\bar{Y} + (1-Z)\mu$. The overall mean is $\mu = E(E(Y_j|p)) = E(p) = \alpha/(\alpha+\beta)$. The sample mean is $\bar{y} = x/n$. The credibility is $Z = n/(n+K)$ and K=EPV/VHM. With $\mathrm{Var}(Y_j|p) = p(1-p)$ it follows that EPV=E(Var($Y_j|p$))=E($p(1-p)$). Because $\mathrm{E}(Y_j) = p$ then VHM=Var($E(Y_j|p)$)=Var(p). For the beta distribution

$$\mathrm{E}(p) = \frac{\alpha}{\alpha+\beta}, \mathrm{E}(p^2) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}, \text{ and } \mathrm{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Parameter K=EPV/VHM=[E(p)-E(p^2)]/Var(p). With some algebra this reduces to $K = \alpha + \beta$. The Bühlmann credibility-weighted estimate is

$$\hat{p} = \frac{n}{n+\alpha+\beta}\left(\frac{x}{n}\right) + \left(1 - \frac{n}{n+\alpha+\beta}\right)\frac{\alpha}{\alpha+\beta}$$
$$\hat{p} = \frac{\alpha+x}{\alpha+\beta+n}$$

which is the same as the Bayesian posterior mean.

9.6 Estimating Credibility Parameters

In this section, you learn how to:

- Perform nonparametric estimation with the Bühlmann and Bühlmann-Straub credibility models.
- Identify situations when semiparametric estimation is appropriate.
- Use data to approximate the *EPV* and *VHM*.
- Balance credibility-weighted estimates.

The examples in this chapter have provided assumptions for calculating credibility parameters. In actual practice the actuary must use real world data and judgment to determine credibility parameters.

9.6.1 Full Credibility Standard for Limited Fluctuation Credibility

Limited-fluctuation credibility requires a full credibility standard. The general formula for aggregate losses or pure premium is

$$n_S = \left(\frac{y_p}{k}\right)^2 \left[\left(\frac{\sigma_N^2}{\mu_N}\right) + \left(\frac{\sigma_X}{\mu_X}\right)^2\right]$$

with N representing number of claims and X the size of claims. If one assumes $\sigma_X = 0$ then the full credibility standard for frequency results. If $\sigma_N = 0$ then the full credibility formula for severity follows. Probability p and k value are often selected using judgment and experience.

In practice it is often assumed that the number of claims is Poisson distributed so that $\sigma_N^2/\mu_N = 1$. In this case the formula can be simplified to

$$n_S = \left(\frac{y_p}{k}\right)^2 \left[\frac{\mathrm{E}(X^2)}{(\mathrm{E}(X))^2}\right].$$

An empirical mean and second moment for the sizes of individual claim losses can be computed from past data, if available.

9.6.2 Nonparametric Estimation for Bühlmann and Bühlmann-Straub Models

Bayesian analysis as described previously requires assumptions about a prior distribution and likelihood. It is possible to produce estimates without these assumptions and these methods are often referred to as empirical Bayes methods. Bühlmann and Bühlmann-Straub credibility with parameters estimated from the data are included in category of empirical Bayes methods.

Bühlmann Model First we will address the simpler Bühlmann model. Assume that there are r risks in a population. For risk i with risk parameter θ_i the losses for n periods are $X_{i1}, \ldots, X_{in}$. The losses for a risk are *iid* across periods as assumed in the Bühlmann model. For risk i the sample mean is $\bar{X}_i = \sum_{j=1}^n X_{ij}/n$ and the unbiased sample process variance is $s_i^2 = \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2/(n-1)$. An unbiased estimator for the *EPV* can be calculated by taking the average of s_i^2 for the r risks in the population:

$$\widehat{EPV} = \frac{1}{r}\sum_{i=1}^r s_i^2 = \frac{1}{r(n-1)}\sum_{i=1}^r\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2. \tag{9.11}$$

The individual risk means $\bar{X}_i$ for $i = 1, \ldots, r$ can be used to estimate the *VHM*. An unbiased estimator of $\mathrm{Var}(\bar{X}_i)$ is

$$\widehat{\mathrm{Var}}(\bar{X}_i) = \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2 \text{ and } \bar{X} = \frac{1}{r}\sum_{i=1}^{r}\bar{X}_i,$$

but $\mathrm{Var}(\bar{X}_i)$ is not the *VHM*. The total variance formula is

$$\mathrm{Var}(\bar{X}_i) = \mathrm{E}(\mathrm{Var}(\bar{X}_i|\Theta = \theta_i)) + \mathrm{Var}(\mathrm{E}(\bar{X}_i|\Theta = \theta_i)).$$

The *VHM* is the second term on the right because $\mu(\theta_i) = \mathrm{E}(\bar{X}_i|\Theta = \theta_i)$ is the hypothetical mean for risk i. So,

$$VHM = \mathrm{Var}(\mathrm{E}(\mu(\theta_i)) = \mathrm{Var}(\bar{X}_i) - \mathrm{E}(\mathrm{Var}(\bar{X}_i|\Theta = \theta_i)).$$

As discussed previously in Section 9.3.1, $EPV/\mathrm{n} = \mathrm{E}(\mathrm{Var}(\bar{X}_i|\Theta = \theta_i))$ and using the above estimators gives an unbiased estimator for the *VHM*:

$$\widehat{VHM} = \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2 - \frac{\widehat{EPV}}{n}. \tag{9.12}$$

Although the expected loss for a risk with parameter θ_i is $\mu(\theta_i)$=$\mathrm{E}(\bar{X}_i|\Theta = \theta_i)$, the variance of the sample mean $\bar{X}_i$ is greater that the variance of the hypothetical means: $\mathrm{Var}(\bar{X}_i) \geq \mathrm{Var}(\mu(\theta_i))$. The variance in the sample means $\mathrm{Var}(\bar{X}_i)$ includes both the variance in the hypothetical means plus a process variance term because for individual observations X_{ij}, $Var(X_{ij}|\Theta = \theta_i) > 0$.

In some cases formula (9.12) can produce a negative value for $\widehat{VHM}$ because of the subtraction of $\widehat{EPV}/n$, but a variance cannot be negative. The process variance within risks is so large that it overwhelms the measurement of the variance in means between risks. We cannot use this method to determine the values needed for Bühlmann credibility.

Example 9.6.1. Two policyholders had claims over a three-year period as shown in the table below. Estimate the expected number of claims for each policyholder using Bühlmann credibility and calculating necessary parameters from the data.

Year	Risk A	Risk B
1	0	2
2	1	1
3	0	2

Show Example Solution

Solution $\bar{x}_A = \frac{1}{3}(0+1+0) = \frac{1}{3}$, $\bar{x}_B = \frac{1}{3}(2+1+2) = \frac{5}{3}$, $\bar{x} = \frac{1}{2}(\frac{1}{3}+\frac{5}{3}) = 1$
$s_A^2 = \frac{1}{3-1}\left[(0-\frac{1}{3})^2 + (1-\frac{1}{3})^2 + (0-\frac{1}{3})^2\right] = \frac{1}{3}$
$s_B^2 = \frac{1}{3-1}\left[(2-\frac{5}{3})^2 + (1-\frac{5}{3})^2 + (2-\frac{5}{3})^2\right] = \frac{1}{3}$
$\widehat{EPV} = \frac{1}{2}\left(\frac{1}{3}+\frac{1}{3}\right) = \frac{1}{3}$, $\widehat{VHM} = \frac{1}{2-1}\left[(\frac{1}{3}-1)^2 + (\frac{5}{3}-1)^2\right] - \frac{1/3}{3} = \frac{7}{9}$
$K = \frac{1/3}{7/9} = \frac{3}{7}$, $Z = \frac{3}{3+(3/7))} = \frac{7}{8}$
$\hat{\mu}_A = \frac{7}{8}\left(\frac{1}{3}\right) + (1-\frac{7}{8})1 = \frac{5}{12}$, $\hat{\mu}_B = \frac{7}{8}\left(\frac{5}{3}\right) + (1-\frac{7}{8})1 = \frac{19}{12}$

Example9.6.2. Two policyholders had claims over a three-year period as shown in the table below. Calculate the nonparametric estimate for the *VHM*.

Year	Risk A	Risk B
1	3	3
2	0	0
3	0	3

Show Example Solution

Solution $\bar{x}_A = \frac{1}{3}(3+0+0) = 1$, $\bar{x}_B = \frac{1}{3}(3+0+3) = 2$, $\bar{x} = \frac{1}{2}(1+2) = \frac{3}{2}$
$s_A^2 = \frac{1}{3-1}\left[(3-1)^2 + (0-1)^2 + (0-1)^2\right] = 3$
$s_B^2 = \frac{1}{3-1}\left[(3-2)^2 + (0-2)^2 + (3-2)^2\right] = 3$
$\widehat{EPV} = \frac{1}{2}(3+3) = 3$
$\widehat{VHM} = \frac{1}{2-1}\left[(1-\frac{3}{2})^2 + (2-\frac{3}{2})^2\right] - \frac{3}{3} = -\frac{1}{2}$.
The process variance is so large that it is not possible to estimate the *VHM*.

Bühlmann-Straub Model Empirical formulas for *EPV* and *VHM* in the Bühlmann-Straub model are more complicated because a risk's number of exposures can change from one period to another. Also, the number of experience periods does not have to be constant across the population because exposure rather than time measures loss potential. First some definitions:

- X_{ij} is the losses per exposure for risk i in period j. Losses can refer to number of claims or amount of loss. There are r risks so $i = 1, \ldots, r$.
- n_i is the number of observation periods for risk i
- m_{ij} is the number of exposures for risk i in period j for $j = 1, \ldots, n_i$

Risk i with risk parameter θ_i has m_{ij} exposures in period j which means that the losses per exposure random variable can be written as $X_{ij} = (Y_{i1} + \cdots + Y_{im_{ij}})/m_{ij}$. Random variable Y_{ik} is the loss for one exposure. For risk i losses Y_{ik} are *iid* with mean $\mathrm{E}(Y_{ik})=\mu(\theta_i)$ and process variance $\mathrm{Var}(Y_{ik})=\sigma^2(\theta_i)$. It follows that $\mathrm{Var}(X_{ij})=\sigma^2(\theta_i)/m_{i,j}$.

Two more important definitions are:

- $\bar{X}_i = \frac{1}{m_i}\sum_{j=1}^{n_i} m_{ij}X_{ij}$ with $m_i = \sum_{j=1}^{n_i} m_{ij}$. $\bar{X}_i$ is the average loss per exposure for risk i for all observation periods combined.
- $\bar{X} = \frac{1}{m}\sum_{i=1}^{r} m_i\bar{X}_i$ with $m = \sum_{i=1}^{r} m_i$. $\bar{X}$ is the average loss per exposure for all risks for all observation periods combined.

Random variable $\bar{X}_i$ is the average loss for all m_i exposures for risk i for all years combined. Random variable $\bar{X}$ is the average loss for all exposures for all risks for all years combined.

An unbiased estimator for the process variance $\sigma^2(\theta_i)$ of one exposure for risk i is

$$s_i{}^2 = \frac{\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{n_i - 1}.$$

The m_{ij} weights are applied to the squared differences because the X_{ij} are the averages of m_{ij} exposures. The weighted average of the sample variances $s_i{}^2$ for each risk i in the population with weights proportional to the number of $(n_i - 1)$ observation periods will produce the expected value of the process variance (EPV) estimate

$$\widehat{EPV} = \frac{\sum_{i=1}^{r}(n_i-1)s_i{}^2}{\sum_{i=1}^{r}(n_i-1)} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^{r}(n_i-1)}.$$

The quantity $*\widehat{EPV}*$ is an unbiased estimator for the process variance of one exposure for a risk chosen at random from the population.

To calculate an estimator for the variance in the hypothetical means (VHM) the squared differences of the individual risk sample means $\bar{X}_i$ and population mean $\bar{X}$ are used. An unbiased estimator for the VHM is

$$\widehat{VHM} = \frac{\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2 - (r-1)*\widehat{EPV}*}{m - \frac{1}{m}\sum_{i=1}^{r} m_i^2}.$$

This complicated formula is necessary because of the varying number of exposures. Proofs that the EPV and VHM estimators shown above are unbiased can be found in several references mentioned at the end of this chapter including (Buhlmann and Gisler, 2005), (Klugman et al., 2012), and (Tse, 2009).

Example 9.6.3. Two policyholders had claims shown in the table below. Estimate the expected number of claims for each policyholder using Bü hlmann-Straub credibility and calculating parameters from the data.

Policyholder		Year 1	Year 2	Year 3	Year 4
A	Number of claims	0	2	2	3
A	Insured vehicles	1	2	2	2
B	Number of claims	0	0	1	2
B	Insured vehicles	0	2	3	4

Show Example Solution

Solution $\bar{x}_A = \frac{0+2+2+3}{1+2+2+2} = 1$, $\bar{x}_B = \frac{0+1+2}{2+3+4} = \frac{1}{3}$, $\bar{x} = \frac{7(1)+9(1/3)}{7+9} = \frac{5}{8}$
$s_A^2 = \frac{1}{4-1}\left[1(0-1)^2 + 2(1-1)^2 + 2(1-1)^2 + 2(\frac{3}{2}-1)^2\right] = \frac{1}{2}$
$s_B^2 = \frac{1}{3-1}\left[2(0-\frac{1}{3})^2 + 3(\frac{1}{3}-\frac{1}{3})^2 + 4(\frac{1}{2}-\frac{1}{3})^2\right] = \frac{1}{6}$
$\widehat{EPV} = \left[3\left(\frac{1}{2}\right) + 2\left(\frac{1}{6}\right)\right]/(3+2) = \frac{11}{30} = 0.3667$
$\widehat{VHM} = \left[(7(1-\frac{5}{8})^2 + 9(\frac{1}{3}-\frac{5}{8})^2 - (2-1)\frac{11}{30}\right] / \left[16 - \left(\frac{1}{16}\right)(7^2+9^2)\right] = 0.1757$
$K = \frac{0.3667}{0.1757} = 2.0871$
$Z_A = \frac{7}{7+2.0871)} = 0.7703$, $Z_B = \frac{9}{9+2.0871)} = 0.8118$,
$\hat{\mu}_A = 0.7703(1) + (1-0.7703)(5/8) = 0.9139$
$\hat{\mu}_B = 0.8118(1/3) + (1-0.8118)(5/8) = 0.3882$

9.6.3 Semiparametric Estimation for Bühlmann and Bühlmann-Straub Models

In the prior section on nonparametric estimation, there were no assumptions about the distribution of the losses per exposure random variables X_{ij}. Assuming that the X_{ij} have a particular distribution and using properties of the distribution along with the data to determine credibility parameters is referred to as semiparametric estimation.

An example of semiparametric estimation would be the assumption of a Poisson distribution when estimating claim frequencies. The Poisson distribution has the property that the mean and variance are identical and this property can simplify calculations. The following simple example comes from the prior section but now includes a Poisson assumption about claim frequencies.

Example 9.6.4. Two policyholders had claims over a three-year period as shown in the table below. Assume that the number of claims for each risk has a Poisson distribution. Estimate the expected number of claims for each policyholder using Bühlmann credibility and calculating necessary parameters from the data.

Year	Risk A	Risk B
1	0	2
2	1	1
3	0	2

Show Example Solution

Solution $\bar{x}_A = \frac{1}{3}(0+1+0) = \frac{1}{3}$, $\bar{x}_B = \frac{1}{3}(2+1+2) = \frac{5}{3}$, $\bar{x} = \frac{1}{2}(\frac{1}{3}+\frac{5}{3}) = 1$
With Poisson assumption the estimated variance for risk A is $\hat{\sigma}_A^2 = \bar{x}_A = \frac{1}{3}$
Similarly, $\hat{\sigma}_B^2 = \bar{x}_B = \frac{5}{3}$
$\widehat{EPV} = \frac{1}{2}(\frac{1}{3}) + \frac{1}{2}(\frac{5}{3}) = 1$. This is also $\bar{x}$ because of Poisson assumption.
$\widehat{VHM} = \frac{1}{2-1}\left[(\frac{1}{3}-1)^2 + (\frac{5}{3}-1)^2\right] - \frac{1}{3} = \frac{5}{9}$
$K = \frac{1}{5/9} = \frac{9}{5}$, $Z_A = Z_B = \frac{3}{3+(9/5)} = \frac{5}{8}$
$\hat{\mu}_A = \frac{5}{8}\left(\frac{1}{3}\right) + (1-\frac{5}{8})1 = \frac{7}{12}$, $\hat{\mu}_B = \frac{5}{8}\left(\frac{5}{3}\right) + (1-\frac{5}{8})1 = \frac{17}{12}$.

We did not have to make the Poisson assumption in the prior example because there was enough data to use nonparametric estimation but the following example is commonly used to demonstrate a situation where semiparametric estimation is needed. There is insufficient data for nonparametric estimation but with the Poisson assumption estimates can be calculated.

Example 9.6.5. A portfolio of 2,000 policyholders generated the following claims profile during a five-year period:

Number of Claims In 5 Years	Number of policies
0	923
1	682
2	249
3	70
4	51
5	25

In your model you assume that the number of claims for each policyholder has a Poisson distribution and that a policyholder's expected number of claims is constant through time. Use Bühlmann credibility to estimate the annual expected number of claims for policyholders with 3 claims during the five-year period.

Show Example Solution

Solution Let θ_i be the risk parameter for the i^{th} risk in the portfolio with mean $\mu(\theta_i)$ and variance $\sigma^2(\theta_i)$. With the Poisson assumption $\mu(\theta_i) = \sigma^2(\theta_i)$. The expected value of the process variance is EPV=E($\sigma^2(\theta_i)$) where the expectation is taken across all risks in the population. Because of the Poisson assumption for all risks it follows that EPV=E($\sigma^2(\theta_i)$)=E($\mu(\theta_i)$). An estimate for the annual expected number of claims is $\hat{\mu}(\theta_i)$= (observed number of claims)/5. This can also serve as the estimate for the process variance for a risk. Weighting the process variance estimates (or means) by the number of policies in each group gives the estimators

$$\widehat{EPV} = \bar{x} = \frac{923(0) + 682(1) + 249(2) + 70(3) + 51(4) + 25(5)}{(5)(2000)} = 0.1719.$$

The *VHM* estimator is

$$
\begin{aligned}
V\hat{H}M &= \frac{1}{2000-1}[923(0-0.1719)^2 + 682(0.20-0.1719)^2 + 249(0.40-0.1719)^2 \\
&\quad +70(0.60-0.1719)^2 + 51(0.80-0.1719)^2 + 25(1-0.1719)^2] - \frac{0.1719}{5} \\
&= 0.0111 \\
\hat{K} &= *E\hat{P}V*/V\hat{H}M = 0.1719/0.0111 = 15.49 \\
\hat{Z} &= \frac{5}{5+15.49} = 0.2440 \\
\hat{\mu}_{3\text{ claims}} &= 0.2440(3/5) + (1-0.2440)0.1719 = 0.2764.
\end{aligned}
$$

9.6.4 Balancing Credibility Estimators

The estimated loss for risk i in a credibility weighted model is $\hat{\mu}(\theta_i) = Z_i\bar{X}_i + (1-Z_i)\bar{X}$ where $\bar{X}_i$ is the loss per exposure for risk i and $\bar{X}$ is loss per exposure for the population. The overall mean in the Bühlmann-Straub model is $\bar{X} = \sum_{i=1}^{r}(m_i/m)\bar{X}_i$ where m_i and m are number of exposures for risk i and population, respectively. The same formula works for the simpler Bühlmann model by setting $m_i = 1$ and $m = r$ where r is the number of risks.

For the credilility weighted estimators to be in balance we want

$$\bar{X} = \sum_{i=1}^{r}(m_i/m)\bar{X}_i = \sum_{i=1}^{r}(m_i/m)\hat{\mu}(\theta_i).$$

If this equation is satisfied then the estimated losses for each risk will add up to the population total, an important goal in ratemating, but this may not happen if $\bar{X}$ is used for the complement of credibility.

In order to find a complement of credibility that will bring the credibility-weighted estimators into balance we will set $\hat{\mu}$ as the complement of crediblity:

$$\sum_{i=1}^{r}(m_i/m)\bar{X}_i = \sum_{i=1}^{r}(m_i/m)(Z_i\bar{X}_i + (1-Z_i)\hat{\mu}).$$

A little algebra gives

$$\sum_{i=1}^{r} m_i\bar{X}_i = \sum_{i=1}^{r} m_i Z_i\bar{X}_i + \hat{\mu}\sum_{i=1}^{r} m_i(1-Z_i),$$

and

$$\hat{\mu} = \frac{\sum_{i=1}^{r} m_i(1-Z_i)\bar{X}_i}{\sum_{i=1}^{r} m_i(1-Z_i)}.$$

This can be simplified using the following relationship

$$m_i(1-Z_i) = m_i\left(1 - \frac{m_i}{m_i+K}\right) = m_i\left(\frac{(m_i+K)-m_i}{m_i+K}\right) = KZ_i.$$

A complement of credibility that will bring the credibility-weighed estimators into balance with the overall mean loss per exposure is

$$\hat{\mu} = \frac{\sum_{i=1}^{r} Z_i \bar{X}_i}{\sum_{i=1}^{r} Z_i}.$$

Example 9.6.6. An example from the nonparametric Bühlmann-Straub section had the following data for two risks. Find the complement of credibility $\hat{\mu}$ that will produce crediblity-weighted estimates that are in balance.

Policyholder		Year 1	Year 2	Year 3	Year 4
A	Number of claims	0	2	2	3
A	Insured vehicles	1	2	2	2
B	Number of claims	0	0	1	2
B	Insured vehicles	0	2	3	4

Show Example Solution

Solution The credibilities from the prior example are $Z_A = \frac{7}{7+2.0871)} = 0.7703$ and $Z_B = \frac{9}{9+2.0871)} = 0.8118$. The sample means are $\bar{X}_A = 1$ and $\bar{X}_B = 1/3$. The balanced complement of credibility is

$$\hat{\mu} = \frac{0.7703(1) + 0.8118(1/3)}{0.7703 + 0.8118} = 0.6579.$$

The updated credibility estimates are $\hat{\mu}_A = 0.7703(1) + (1 - 0.7703)(.6579) = 0.9214$ versus the previous 0.9139 and $\hat{\mu}_B = 0.8118(1/3) + (1 - 0.8118)(.6579) = 0.3944$ versus previous 0.3882. Checking the balance on the new estimators: (7/16)(0.9214)+(9/16)0.3944)=0.6250. This exactly matches $\bar{X} = 10/16 = 0.6250$.

9.7 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.

Credibility Guided Tutorials

Contributors

- **Gary Dean**, Ball State University is the author of the initial version of this chapter. Email: cgdean@bsu.edu for chapter comments and suggested improvements.

Chapter 10

Insurance Portfolio Management including Reinsurance

Chapter Preview. An insurance portfolio is simply a collection of insurance contracts. To help manage the uncertainty of the portfolio, this chapter

- quantifies unusually large obligations by examining the tail of the distribution,
- quantifies the overall riskiness by introducing summaries known as risk measures, and
- discusses options of spreading portfolio risk through reinsurance, the purchase of insurance protection by an insurer.

Overview

Most of our analyses in prior chapters has been at the contract level which is an agreement between a policyholder and an insurer. Insurers hold, and manage, *portfolios*a collection of contracts that are simply collections of contracts. As in other areas of finance, there are management decision-making choices that occur only at the portfolio level. For example, strategic decision-making does not occur at the contract level. It happens in the conference room, where management reviews available data and possibly steers a new course. From the portfolio perspective, insurers want to do capacity planning, set management policies, and balance the mix of products being booked to grow revenue while controlling volatility.

Conceptually, one can think about an insurance company as nothing more than a collection, or portfolio, of insurance contracts. In Chapter 5 we learned about modeling insurance portfolios as the sum of individual contracts based on assumptions of independence among contracts. Because of their importance, this chapter focuses directly on portfolio distributions.

- Insurance portfolios represent obligations of insurers and so we are particularly interested in probabilities of large outcomes as these represent unusually large obligations. To formalize this concept, we introduce the notion of a heavy-tail distribution in Section 10.1.

- Insurance portfolios represent company obligations and so insurers keep an equivalent amount of assets to meet these obligations. *Risk measures*, introduced in Section 10.2, summarize the distribution of the insurance portfolio and these summary measures are used to quantify the amount of assets that an insurer needs to retain to meet obligations.

- In Section 3.4, we learned about mechanisms that policyholders use to spread risks such as deductibles and policy limits. In the same way, insurers use similar mechanisms in order to spread portfolio risks. They purchase risk protection from reinsurers, an insurance company for insurers. This sharing of insurance portfolio risk is described in Section 10.3.

10.1 Tails of Distributions

In this section, you learn how to:

- Describe a heavy tail distribution intuitively.
- Classify the heaviness of a distribution's tails based on moments.
- Compare the tails of two distributions.

In 1998 freezing rain fell on eastern Ontario, southwestern Quebec and lasted for six days. The event was double the amount of precipitation in the area experienced in any prior ice storm and resulted in a catastrophe that produced in excess of 840,000 insurance claims. This number is 20% more than that of claims caused by the Hurricane Andrew - one of the largest natural disasters in the history of North America. The catastrophe caused approximately 1.44 billion Canadian dollars in insurance settlements which is the highest loss burden in the history of Canada. This is not an isolated example - similar catastrophic events that caused extreme insurance losses are Hurricanes Harvey and Sandy, the 2011 Japanese earthquake and tsunami, and so forth.

In the context of insurance, a few large losses hitting a portfolio and then converting into claims usually represent the greatest part of the indemnities paid by insurance companies. The aforementioned losses, also called 'extremes', are quantitatively modelled by the tails of the associated probability distributions. From the quantitative modelling standpoint, relying on probabilistic models with improper tails is rather daunting. For instance, periods of financial stress may appear with a higher frequency than expected, and insurance losses may occur with worse severity. Therefore, the study of probabilistic behavior in the tail portion of actuarial models is of utmost importance in the modern framework of quantitative risk management. For this reason, this section is devoted to the introduction of a few mathematical notions that characterize the tail weight of random variables (*rv*'s). The applications of these notions will benefit us in the construction and selection of appropriate models with desired mathematical properties in the tail portion, that are suitable for a given task.

Formally, define X to be the (random) obligations that arise from a collection (portfolio) of insurance contracts. We are particularly interested in studying the right tail of the distribution of X, which represents the occurrence of large losses. Informally, *a rv is said to be heavy-tailed if high probabilities are assigned to large values.* Note that this by no mean implies the probability density/mass functions are increasing as the value of *rv* goes to infinity. Ineed for a real-valued *rv*, the *pdf*probability density function/*pmf*probability mass function must diminish at infinity in order to guarantee the total probability to be equal to one. Instead, what we concern about is the rate of decaying of the probability function. Unwelcome outcomes are more likely to occur for an insurance portfolio that is described by a loss *rv* possessing heavier (right) tail. Tail weight can be an absolute or a relative concept. Specifically, for the former, we may consider a *rv* to be heavy-tailed if certain mathematical properties of the probability distribution are met. For the latter, we can say the tail of one distribution is heavier than the other if some tail measures are larger/smaller.

Several quantitative approaches have been proposed to classify and compare tail weight. Among most of these approaches, the *survival function*one minus the distribution function. It gives the probability that a rv exceeds a specific value serves as the building block. In what follows, we introduce two simple yet useful tail classification methods both of which are based on the behavior of the survival function of X.

10.1.1 Classification Based on Moments

One way of classifying the tail weight of distribution is by assessing the existence of raw moments. Since our major interest lies in the right tails of distributions, we henceforth assume the obligation or loss *rv* X to be positive. At the outset, the $k-$th raw moment of a continuous *rv* X, introduced in Section 3.1, can be computed as

$$\mu'_k = k\int_0^\infty x^{k-1}S(x)dx,$$

where $S(\cdot)$ denotes the survival function of X. This expression emphasizes that the existence of the raw moments depends on the asymptotic behavior of the survival function at infinity. Namely, the faster the survival function decays to zero, the higher the order of finite moment (k) the associated *rv* possesses.You may interpret k^* to be the largest value of k so that the moment is finite. Formally, define $k^* := \sup\{k > 0 : \mu'_k < \infty\}$, where *sup* represents the supremum operator. This observation leads us to the moment-based tail weight classification method, which is defined formally next.

Definition 10.1. Consider a positive loss random variable X.

- If all the positive raw moments exist, namely the maximal order of finite moment $k^* = \infty$, then X is said to be **light tailed** based on the moment method.
- If $k^* < \infty$, then X is said to be **heavy tailed** based on the moment method.
- Moreover, for two positive loss random variables X_1 and X_2 with maximal orders of moment k_1^* and k_2^* respectively, we say X_1 has a **heavier (right) tail** than X_2 if $k_1^* \leq k_2^*$.

The first part of Definition 10.1 is an absolute concept of tail weight, while the second part is a relative concept of tail weight which compares the (right) tails between two distributions. Next, we present a few examples that illustrate the applications of the moment-based method for comparing tail weight.

Example 10.1.1. Light tail nature of the gamma distribution. Let $X \sim gamma(\alpha, \theta)$, with $\alpha > 0$ and $\theta > 0$, then for all $k > 0$, show that $\mu'_k < \infty$.

Show Example Solution

Solution.

$$\begin{aligned}
\mu'_k &= \int_0^\infty x^k \frac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}dx \\
&= \int_0^\infty (y\theta)^k \frac{(y\theta)^{\alpha-1}e^{-y}}{\Gamma(\alpha)\theta^\alpha}\theta dy \\
&= \frac{\theta^k}{\Gamma(\alpha)}\Gamma(\alpha+k) < \infty.
\end{aligned}$$

Since all the positive moments exist, i.e., $k^* = \infty$, in accordance with the moment-based classification method in Definition 10.1, the gamma distribution is light-tailed.

Example 10.1.2. Light tail nature of the Weibull distribution. Let $X \sim Weibull(\theta, \tau)$, with $\theta > 0$ and $\tau > 0$, then for all $k > 0$, show that $\mu'_k < \infty$.

Show Example Solution

Solution.

$$\begin{aligned}
\mu'_k &= \int_0^\infty x^k \frac{\tau x^{\tau-1}}{\theta^\tau}e^{-(x/\theta)^\tau}dx \\
&= \int_0^\infty \frac{y^{k/\tau}}{\theta^\tau}e^{-y/\theta^\tau}dy \\
&= \theta^k\Gamma(1+k/\tau) < \infty.
\end{aligned}$$

Again, due to the existence of all the positive moments, the Weibull distribution is light-tailed.

The gamma and Weibull distributions are used quite extensively in the actuarial practice. Applications of these two distributions are vast which include, but are not limited to, insurance claim severity modelling, solvency assessment, loss reserving, aggregate risk approximation, reliability engineering and failure analysis. We have thus far seen two examples of using the moment-based method to analyze light-tailed distributions. We document a heavy-tailed example in what follows.

Example 10.1.3. Heavy tail nature of the Pareto distribution. Let $X \sim Pareto(\alpha, \theta)$, with $\alpha > 0$ and $\theta > 0$, then for $k > 0$

$$\begin{aligned} \mu_k^{'} &= \int_0^\infty x^k \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} dx \\ &= \alpha\theta^\alpha \int_\theta^\infty (y-\theta)^k y^{-(\alpha+1)} dy. \end{aligned}$$

Consider a similar integration:

$$g_k := \int_\theta^\infty y^{k-\alpha-1} dy = \begin{cases} < \infty, & \text{for } k < \alpha; \\ = \infty, & \text{for } k \geq \alpha. \end{cases}$$

Meanwhile,

$$\lim_{y\to\infty} \frac{(y-\theta)^k y^{-(\alpha+1)}}{y^{k-\alpha-1}} = \lim_{y\to\infty} (1-\theta/y)^k = 1.$$

Application of the limit comparison theorem for improper integrals yields μ_k' is finite if and only if g_k is finite. Hence we can conclude that the raw moments of Pareto *rv*'s exist only up to $k < \alpha$, i.e., $k^* = \alpha$, and thus the distribution is heavy-tailed. What is more, the maximal order of finite moment depends only on the shape parameter α and it is an increasing function of α. In other words, based on the moment method, the tail weight of Pareto *rv*'s is solely manipulated by α – the smaller the value of α, the heavier the tail weight becomes. Since $k^* < \infty$, the tail of Pareto distribution is heavier than those of the gamma and Weibull distributions.

We conclude this section with an open discussion on the limitations of the moment-based method. Despite its simple implementation and intuitive interpretation, there are certain circumstances in which the application of the moment-based method is not suitable. First, for more complicated probabilistic models, the k-th raw moment may not be simple to derive, and thus the identification of the maximal order of finite moment can be challenging. Second, the moment-based method does not well comply with main body of the well established heavy tail theory in the literature. Specifically, the existence of moment generating functions is arguably the most popular method for classifying heavy tail versus light tail within the community of academic actuaries. However, for some *rv*'s such as the lognormal *rv*'s, their moment generating functions do not exist even that all the positive moments are finite. In these cases, applications of the moment-based methods can lead to different tail weight assessment. Third, when we need to compare the tail weight between two light-tailed distributions both having all positive moments exist, the moment-based method is no longer informative (see, e.g., Examples 10.1.1 and 10.1.2).

10.1.2 Comparison Based on Limiting Tail Behavior

In order to resolve the aforementioned issues of the moment-based classification method, an alternative approach for comparing tail weight is to directly study the limiting behavior of the survival functions.

Definition 10.2. For two *rv*'s X and Y, let

$$\gamma := \lim_{t\to\infty} \frac{S_X(t)}{S_Y(t)}.$$

We say that

- X has a **heavier right tail** than Y if $\gamma = \infty$;
- X and Y are **proportionally equivalent in the right tail** if $\gamma = c \in \mathbf{R}_+$;
- X has a **lighter right tail** than Y if $\gamma = 0$.

Example 10.1.4. Comparison of Pareto to Weibull distributions. Let $X \sim Pareto(\alpha, \theta)$ and $Y \sim Weibull(\tau, \theta)$, for $\alpha > 0$, $\tau > 0$, and $\theta > 0$. Show that the Pareto has a heavier right tail than the Weibull.

Show Example Solution

Solution.

$$\begin{aligned}
\lim_{t\to\infty} \frac{S_X(t)}{S_Y(t)} &= \lim_{t\to\infty} \frac{(1+t/\theta)^{-\alpha}}{\exp\{-(t/\theta)^{\tau}\}} \\
&= \lim_{t\to\infty} \frac{\exp\{t/\theta^{\tau}\}}{(1+t^{1/\tau}/\theta)^{\alpha}} \\
&= \lim_{t\to\infty} \frac{\sum_{i=0}^{\infty}\left(\frac{t}{\theta^{\tau}}\right)^{i}/i!}{(1+t^{1/\tau}/\theta)^{\alpha}} \\
&= \lim_{t\to\infty} \sum_{i=0}^{\infty}\left(t^{-i/\alpha}+\frac{t^{(1/\tau-i/\alpha)}}{\theta}\right)^{-\alpha} /\theta^{\tau i} i! \\
&= \infty.
\end{aligned}$$

Therefore, the Pareto distribution has a heavier tail than the Weibull distribution. One may also realize that exponentials go to infinity faster than polynomials, thus the aforementioned limit must be infinite.

For some distributions of which the survival functions do not admit explicit expressions, we may find the following alternative formula useful:

$$\begin{aligned}
\lim_{t\to\infty} \frac{S_X(t)}{S_Y(t)} &= \lim_{t\to\infty} \frac{S_X^{'}(t)}{S_Y^{'}(t)} \\
&= \lim_{t\to\infty} \frac{-f_X(t)}{-f_Y(t)} \\
&= \lim_{t\to\infty} \frac{f_X(t)}{f_Y(t)}.
\end{aligned}$$

given that the density functions exist.

Example 10.1.5. Comparison of Pareto to gamma distributions. Let $X \sim Pareto(\alpha, \theta)$ and $Y \sim gamma(\alpha, \theta)$, for $\alpha > 0$ and $\theta > 0$. Show that the Pareto has a heavier right tail than the gamma.

Show Example Solution

Solution.

$$\begin{aligned}
\lim_{t\to\infty}\frac{f_X(t)}{f_Y(t)} &= \lim_{t\to\infty}\frac{\alpha\theta^{\alpha}(t+\theta)^{-\alpha-1}}{t^{\tau-1}e^{-t/\lambda}\lambda^{-\tau}\Gamma(\tau)^{-1}} \\
&\propto \lim_{t\to\infty}\frac{e^{t/\lambda}}{(t+\theta)^{\alpha+1}t^{\tau-1}} \\
&= \infty,
\end{aligned}$$

as exponentials go to infinity faster than polynomials.

10.2 Risk Measures

In this section, you learn how to:

- Define the idea of *coherence* and determine whether or not a risk measure is coherent.
- Define the value-at-risk and calculate this quantity for a given distribution.
- Define the tail value-at-risk and calculate this quantity for a given distribution.

In the previous section, we studied two methods for classifying the weight of distribution tails. We may claim that the risk associated with one distribution is more dangerous (asymptotically) than the others if the tail is heavier. However, knowing one risk is more dangerous (asymptotically) than the others may not provide sufficient information for a sophisticated risk management purpose, and in addition, one is also interested in quantifying how much more. In fact, the magnitude of risk associated with a given loss distribution is an essential input for many insurance applications, such as actuarial pricing, reserving, hedging, insurance regulatory oversight, and so forth.

10.2.1 Coherent Risk Measures

To compare the magnitude of risk in a practically convenient manner, we seek a function that maps the loss *rv* of interest to a numerical value indicating the level of riskiness, which is termed the *risk measure*a measure that summarizes the riskiness, or uncertainty, of a distribution. Put mathematically, the risk measure simply summarizes the distribution function of a *rv* as a single number. Two simple risk measures are the mean $\mathrm{E}[X]$ and the standard deviation $\mathrm{SD}(X)=\sqrt{\mathrm{Var}(X)}$. Other classical examples of risk measures include the standard deviation *principle*

$$H_{\mathrm{SD}}(X):=\mathrm{E}[X]+\alpha\mathrm{SD}(X),\ \text{for } \alpha\geq 0, \tag{10.1}$$

and the variance principle

$$H_{\mathrm{Var}}(X):=\mathrm{E}[X]+\alpha\mathrm{Var}(X),\ \text{for } \alpha\geq 0.$$

It is a simple matter to check that all the aforementioned functions are risk measures in which we input the loss *rv* and the functions output a numerical value. On a different note, the function $H^*(X):=\alpha X^{\beta}$ for any real-valued $\alpha,\beta\neq 0$, is not a risk measure because H^* produces another *rv* rather than a single numerical value.

Since risk measures are scalar measures which aim to use a single numerical value to describe the stochastic nature of loss *rv*'s, it should not be surprising to us that there is no risk measure which can capture all the risk information of the associated *rv*'s. Therefore, when seeking useful risk measures, it is important for us to keep in mind that the measures should be at least

- interpretable practically;
- computable conveniently; and
- able to reflect the most critical information of risk underpinning the loss distribution.

Several risk measures have been developed in the literature. Unfortunately, there is no best risk measure that can outperform the others, and the selection of appropriate risk measure depends mainly on the application questions at hand. In this respect, it is imperative to emphasize that *risk* is a subjective concept, and thus even given the same problem, there are multifarious approaches to assess risk. However, for many risk management applications, there is a wide agreement that economically sounded risk measures should satisfy four major axioms which we are going to describe in detail next. Risk measures that satisfy these axioms are termed *coherent risk measures*a risk measure that is is subadditive, monontonic, has positive homogeneity, and is translation invariant.

Consider in what follows a risk measure $H(\cdot)$, then H is a **coherent risk measure** if the following axioms are satisfied.

- **Axiom 1.** *Subadditivity:* $H(X+Y) \leq H(X) + H(Y)$. The economic implication of this axiom is that diversification benefits exist if different risks are combined.
- **Axiom 2.** *Monotonicity:* if $\Pr[X \leq Y] = 1$, then $H(X) \leq H(Y)$. Recall that X and Y are *rv*'s representing losses, the underlying economic implication is that higher losses essentially leads to a higher level of risk.
- **Axiom 3.** *Positive homogeneity:* $H(cX) = cH(X)$ for any positive constant c. A potential economic implication about this axiom is that risk measure should be independent of the monetary units in which the risk is measured. For example, let c be the currency exchange rate between the US and Canadian dollars, then the risk of random losses measured in terms of US dollars (i.e., X) and Canadian dollars (i.e., cX) should be different only up to the exchange rate c (i.e., $cH(x) = H(cX)$).
- **Axiom 4.** *Translation invariance:* $H(X+c) = H(X) + c$ for any positive constant c. If the constant c is interpreted as risk-free cash, this axiom tells that no additional risk is created for adding cash to an insurance portfolio, and injecting risk-free capital of c can only reduce the risk by the same amount.

Verifying the coherent properties for some risk measures can be quite straightforward, but it can be very challenging sometimes. For example, it is a simple matter to check that the mean is a coherent risk measure.

Example. The Mean is a Coherent Risk Measure.

For any pair of *rv*'s X and Y having finite means and constant $c > 0$,

- validation of *subadditivity*: $\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$;
- validation of *monotonicity*: if $\Pr[X \leq Y] = 1$, then $\mathrm{E}[X] \leq \mathrm{E}[Y]$;
- validation of *positive homogeneity*: $\mathrm{E}[cX] = c\mathrm{E}[X]$;
- validation of *translation invariance*: $\mathrm{E}[X+c] = \mathrm{E}[X] + c$

With a little more effort, we can determine the following.

Example. The Standard Deviation is not a Coherent Risk Measure.

Show Example Verification

On a different note, the standard deviation is not a coherent risk measure. Specifically, one can check that the standard deviation satisfies

- validation of *subadditivity*:

$$\begin{aligned}\mathrm{SD}[X+Y] &= \sqrt{\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)} \\ &\leq \sqrt{\mathrm{SD}(X)^2+\mathrm{SD}(Y)^2+2\mathrm{SD}(X)\mathrm{SD}(Y)} \\ &= \mathrm{SD}(X)+\mathrm{SD}(Y);\end{aligned}$$

- validation of *positive homogeneity*: $\mathrm{SD}[cX] = c\ \mathrm{SD}[X]$.

However, the standard deviation does not comply with translation invariance property as for any positive constant c,

$$\mathrm{SD}(X+c) = \mathrm{SD}(X) < \mathrm{SD}(X) + c.$$

Moreover, the standard deviation also does not satisfy the monotonicity property. To see this, consider the following two *rv*'s:

$$X = \begin{cases} 0, & \text{with probability } 0.25; \\ 4, & \text{with probability } 0.75, \end{cases} \tag{10.2}$$

and Y is a degenerate *rv* such that

$$\Pr[Y=4]=1. \tag{10.3}$$

It is easy to check that $\Pr[X \leq Y] = 1$, but $\mathrm{SD}(X) = \sqrt{4^2 \cdot 0.25 \cdot 0.75} = \sqrt{3} > \mathrm{SD}(Y) = 0$.

We have so far checked that $\mathrm{E}[\cdot]$ is a coherent risk measure, but not $\mathrm{SD}(\cdot)$. Let us now proceed to study the coherent property for the standard deviation principle (10.1) which is a linear combination of coherent and incoherent risk measures.

Example. The Standard Deviation Principle (10.1) **is a Coherent Risk Measure.**

Show Example Verification

To this end, for a given $\alpha > 0$, we check the four axioms for $H_{\mathrm{SD}}(X+Y)$ one by one:

- validation of *subadditivity:*

$$\begin{aligned}H_{\mathrm{SD}}(X+Y) &= \mathrm{E}[X+Y] + \alpha\mathrm{SD}(X+Y) \\ &\leq \mathrm{E}[X]+\mathrm{E}[Y]+\alpha[\mathrm{SD}(X)+\mathrm{SD}(Y)] \\ &= H_{\mathrm{SD}}(X)+H_{\mathrm{SD}}(Y);\end{aligned}$$

- validation of *positive homogeneity:*

$$H_{\mathrm{SD}}(cX) = c\mathrm{E}[X] + c\alpha\mathrm{SD}(X) = cH_{\mathrm{SD}}(X);$$

- validation of *translation invariance:*

$$H_{\mathrm{SD}}(X+c) = \mathrm{E}[X] + c + \alpha \mathrm{SD}(X) = H_{\mathrm{SD}}(X) + c.$$

It only remains to verify the monotonicity property, which may or may not be satisfied depending on the value of α. To see this, consider again the setup of (10.2) and (10.3) in which $\Pr[X \le Y] = 1$. Let $\alpha = 0.1 \cdot \sqrt{3}$, then $H_{\mathrm{SD}}(X) = 3 + 0.3 = 3.3 < H_{\mathrm{SD}}(Y) = 4$ and the monotonicity condition is met. On the other hand, let $\alpha = \sqrt{3}$, then $H_{\mathrm{SD}}(X) = 3 + 3 = 6 > H_{\mathrm{SD}}(Y) = 4$ and the monotonicity condition is not satisfied. More precisely, by setting

$$H_{\mathrm{SD}}(X) = 3 + \alpha\sqrt{3} \le 4 = H_{\mathrm{SD}}(Y),$$

we find that the monotonicity condition is only satisfied for $0 \le \alpha \le 1/\sqrt{3}$, and thus the standard deviation principle H_{SD} is coherent. This result appears to be very intuitive to us since the standard deviation principle H_{SD} is a linear combination of two risk measures of which one is coherent and the other is incoherent. If $\alpha \le 1/\sqrt{3}$, then the coherent measure dominates the incoherent one, thus the resulting measure H_{SD} is coherent and vice versa. Note that the aforementioned conclusion may not be generalized to any pair of *rv*'s X and Y.

The literature on risk measures has been growing rapidly in popularity and importance. In the succeeding two subsections, we introduce two indices which have recently earned an unprecedented amount of interest among theoreticians, practitioners, and regulators. They are namely the *Value-at-Risk* (*VaR*) and the *Tail Value-at-Risk* (*TVaR*) measures. The economic rationale behind these two popular risk measures is similar to that for the tail classification methods introduced in the previous section, with which we hope to capture the risk of extremal losses represented by the distribution tails.

10.2.2 Value-at-Risk

In Section 4.1.1, we defined the quantile of a distribution. We now look to a special case of this and offer the formal definition of the *value-at-risk*a risk measure based on a quantile function, or *VaR*.

Definition 10.3. Consider an insurance loss random variable X. The value-at-risk measure of X with confidence level $q \in (0,1)$ is formulated as

$$VaR_q[X] := \inf\{x : F_X(x) \ge q\}. \tag{10.4}$$

Here, inf is the infimum operator so that the *VaR* measure outputs the smallest value of X such that the associated *cdf*cumulative distribution function first exceeds or equates to q.

Here is how we should interpret *VaR* in the context of actuarial applications. The *VaR* is a measure of the 'maximal' probable loss for an insurance product/portfolio or a risky investment occurring $q \times 100\%$ of times, over a specific time horizon (typically, one year). For instance, let X be the annual loss *rv* of an insurance product, $VaR_{0.95}[X] = 100$ million means that there is a 5% chance that the loss will exceed 100 million over a given year. Owing to this meaningful interpretation, *VaR* has become the industrial standard to measuring financial and insurance risks since 1990's. Financial conglomerates, regulators, and academics often utilize *VaR* to measure risk capital, ensure the compliance with regulatory rules, and disclose the financial positions.

Next, we present a few examples about the computation of *VaR*.

Example 10.2.1. *VaR* for the exponential distribution. Consider an insurance loss *rv* $X \sim Exp(\theta)$ for $\theta > 0$, then the *cdf* of X is given by

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

Give a closed-form expression for the *VaR*.

Show Example Solution

Solution.

Because exponential distribution is a continuous distribution, the smallest value such that the *cdf* first exceeds or equates to $q \in (0,1)$ must be at the point x_q satisfying

$$q = F_X(x_q) = 1 - \exp\{-x_q/\theta\}.$$

Thus

$$VaR_q[X] = F_X^{-1}(q) = -\theta[\log(1-q)].$$

The result reported in Example 10.2.1 can be generalized to any continuous *rv*'s having strictly increasing *cdf*. Specifically, the *VaR* of any continuous *rv*'s is simply the inverse of the corresponding *cdf*. Let us consider another example of continuous *rv* which has the support from negative infinity to positive infinity.

Example 10.2.2. ***VaR* for the normal distribution.** Consider an insurance loss *rv* $X \sim Normal(\mu, \sigma^2)$ with $\sigma > 0$. In this case, one may interpret the negative values of X as profit or revenue. Give a closed-form expression for the *VaR*.

Show Example Solution

Solution.

Because normal distribution is a continuous distribution, the *VaR* of X must satisfy

$$\begin{aligned} q &= F_X(VaR_q[X]) \\ &= \Pr\left[(X-\mu)/\sigma \leq (VaR_q[X]-\mu)/\sigma\right] \\ &= \Phi((VaR_q[X]-\mu)/\sigma). \end{aligned}$$

Therefore, we have

$$VaR_q[X] = \Phi^{-1}(q)\ \sigma + \mu.$$

In many insurance applications, we have to deal with transformations of *rv*'s. For instance, in Example 10.2.2, the loss *rv* $X \sim Normal(\mu, \sigma^2)$ can be viewed as a linear transformation of a standard normal *rv* $Z \sim Normal(0,1)$, namely $X = Z\sigma + \mu$. By setting $\mu = 0$ and $\sigma = 1$, it is straightforward for us to check $VaR_q[Z] = \Phi^{-1}(q)$. A useful finding revealed from Example 10.2.2 is that the *VaR* of a linear transformation of the normal *rv*'s is equivalent to the linear transformation of the *VaR* of the original *rv*'s. This finding can be further generalized to any *rv*'s as long as the transformations are strictly increasing.

Example 10.2.3. ***VaR* for transformed variables.** Consider an insurance loss *rv* $Y \sim lognormal(\mu, \sigma^2)$, for $\mu \in \mathbf{R}$ and $\sigma > 0$. Give an expression of the VaR of Y in terms of the standard normal inverse *cdf*.

Show Example Solution

Solution.

Note that $\log Y \sim Normal(\mu, \sigma^2)$, or equivalently let $X \sim Normal(\mu, \sigma^2)$, then $Y \stackrel{d}{=} e^X$ which is strictly increasing transformation. Here, the notation '$\stackrel{d}{=}$' means equality in distribution. The *VaR* of Y is thus given by the exponential transformation of the *VaR* of X. Precisely, for $q \in (0,1)$,

$$VaR_q[Y] = e^{VaR_q[X]} = \exp\{\Phi^{-1}(q)\ \sigma + \mu\}.$$

We have thus far seen a number of examples about the *VaR* for continuous *rv*'s, let us consider an example concerning the *VaR* for a discrete *rv*.

Example 10.2.4. ***VaR* for a discrete random variable.** Consider an insurance loss *rv* with the following probability distribution:

$$\Pr[X = x] = \begin{cases} 1, & \text{with probability } 0.75 \\ 3, & \text{with probability } 0.20 \\ 4, & \text{with probability } 0.05. \end{cases}$$

Determine the *VaR* at $q = 0.6, 0.9, 0.95, 0.95001$.

Show Example Solution

Solution.

The corresponding *cdf* of X is

$$F_X(x) = \begin{cases} 0, & x < 1; \\ 0.75, & 1 \leq x < 3; \\ 0.95, & 3 \leq x < 4; \\ 1, & 4 \leq x. \end{cases}$$

By the definition of *VaR*, we thus have then

- $VaR_{0.6}[X] = 1$;
- $VaR_{0.9}[X] = 3$;
- $VaR_{0.95}[X] = 3$;
- $VaR_{0.950001}[X] = 4$.

Let us now conclude the current subsection by an open discussion of the *VaR* measure. Some advantages of utilizing *VaR* include

- possessing a practically meaningful interpretation;
- relatively simple to compute for many distributions with closed-form distribution functions;
- no additional assumption is required for the computation of *VaR*.

On the other hand, the limitations of *VaR* can be particularly pronounced for some risk management practices. We report some of them herein:

- the selection of the confidence level $q \in (0, 1)$ is highly subjective, while the *VaR* can be very sensitive to the choice of q (e.g., in Example 10.2.4, $VaR_{0.95}[X] = 3$ and $VaR_{0.950001}[X] = 4$);
- the scenarios/loss information that are above the $(1-p) \times 100\%$ worst event, are completely neglected;
- *VaR* is not a coherent risk measure (specifically, the *VaR* measure does not satisfy the subadditivity axiom, meaning that diversification benefits may not be fully reflected).

10.2.3 Tail Value-at-Risk

Recall that the *VaR* represents the $(1-p) \times 100\%$ chance maximal loss. As we mentioned in the previous section, one major drawback of the *VaR* measure is that it does not reflect the extremal losses occurring beyond the $(1-p) \times 100\%$ chance worst scenario. For illustrative purposes, let us consider the following slightly unrealistic yet inspiring example.

Example 10.2.5. Consider two loss *rv*'s $X \sim Uniform[0, 100]$, and $Y \sim Exp(31.71)$. We use *VaR* at 95% confidence level to measure the riskiness of X and Y. Simple calculation yields (see, also, Example 10.2.1),

$$VaR_{0.95}[X] = VaR_{0.95}[Y] = 95,$$

and thus these two loss distributions have the same level of risk according to $VaR_{0.95}$. However, it is clear that Y is riskier than X if extremal losses are of major concern since X is bounded above while Y is unbounded. Simply quantifying risk by using *VaR* at a specific confidence level could be misleading and may not reflect the true nature of risk.

As a remedy, the *Tail Value-at-Risk* (*TVaR*) was proposed to measure the extremal losses that are above a given level of *VaR* as an average. We document the definition of *TVaR* in what follows. For the sake of simplicity, we are going to confine ourselves to continuous positive *rv*'s only, which are more frequently used in the context of insurance risk management. We refer the interested reader to Hardy (2006) for a more comprehensive discussion of *TVaR* for both discrete and continuous *rv*'s.

Definition 10.4. Fix $q \in (0,1)$, the *tail value-at-risk*the expected value of a risk given that the risk exceeds a value-at-risk of a (continuous) *rv* X is formulated as

$$TVaR_q[X] \quad := \quad \mathrm{E}[X|X > VaR_q[X]],$$

given that the expectation exists.

In light of Definition 10.4, the computation of *TVaR* typically consists of two major components - the *VaR* and the average of losses that are above the *VaR*. The *TVaR* can be computed via a number of formulas. Consider a continuous positive *rv* X, for notional convenience, henceforth let us write $\pi_q := VaR_q[X]$. By definition, the *TVaR* can be computed via

$$TVaR_q[X] = \frac{1}{(1-q)} \int_{\pi_q}^{\infty} x f_X(x) dx. \tag{10.5}$$

Example 10.2.6. *TVaR* for a normal distribution. Consider an insurance loss *rv* $X \sim Normal(\mu, \sigma^2)$ with $\mu \in \mathbf{R}$ and $\sigma > 0$. Give an expression for *TVaR*.

Show Example Solution

Solution.

Let Z be the standard normal *rv*. For $q \in (0,1)$, the *TVaR* of X can be computed via

$$\begin{aligned} TVaR_q[X] &= \mathrm{E}[X|X > VaR_q[X]] \\ &= \mathrm{E}[\sigma Z + \mu | \sigma Z + \mu > VaR_q[X]] \\ &= \sigma \mathrm{E}[Z|Z > (VaR_q[X] - \mu)/\sigma] + \mu \\ &\overset{(1)}{=} \sigma \mathrm{E}[Z|Z > VaR_q[Z]] + \mu, \end{aligned}$$

where '$\overset{(1)}{=}$' holds because of the results reported in Example 10.2.2. Next, we turn to study $TVaR_q[Z] = \mathrm{E}[Z|Z > VaR_q[Z]]$. Let $\omega(q) = (\Phi^{-1}(q))^2/2$, we have

$$\begin{aligned} (1-q)\ TVaR_q[Z] &= \int_{\Phi^{-1}(q)}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \int_{\omega(q)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x} dx \\ &= \frac{1}{\sqrt{2\pi}} e^{-\omega(q)} \\ &= \phi(\Phi^{-1}(q)). \end{aligned}$$

Thus,

$$TVaR_q[X] = \sigma \frac{\phi(\Phi^{-1}(q))}{1-q} + \mu.$$

We mentioned earlier in the previous subsection that the *VaR* of a strictly increasing function of *rv* is equal to the function of *VaR* of the original *rv*. Motivated by the results in Example 10.2.6, one can show that the *TVaR* of a strictly increasing linear transformation of *rv* is equal to the function of *VaR* of the original *rv* This is due to the linearity property of expectations. However, the aforementioned finding cannot be extended to non-linear functions. The following example of lognormal *rv* serves as a counter example.

Example 10.2.7. *TVaR* of a lognormal distribution. Consider an insurance loss *rv* $X \sim lognormal(\mu, \sigma^2)$, with $\sigma > 0$. Show that

$$TVaR_q[X] = \frac{e^{\mu+\sigma^2/2}}{(1-q)} \Phi(\Phi^{-1}(q) - \sigma).$$

Show Example Solution

Solution.

Recall that the *pdf* of lognormal distribution is formulated as

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

Fix $q \in (0,1)$, then the *TVaR* of X can be computed via

$$\begin{aligned}
TVaR_q[X] &= \frac{1}{(1-q)} \int_{\pi_q}^{\infty} x f_X(x) dx \\
&= \frac{1}{(1-q)} \int_{\pi_q}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\} dx \\
&\overset{(1)}{=} \frac{1}{(1-q)} \int_{\omega(q)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2 + \sigma w + \mu} dw \\
&= \frac{e^{\mu+\sigma^2/2}}{(1-q)} \int_{\omega(q)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(w-\sigma)^2} dw \\
&= \frac{e^{\mu+\sigma^2/2}}{(1-q)} \Phi(\omega(q) - \sigma), \qquad (10.6)
\end{aligned}$$

where $\overset{(1)}{=}$ holds by applying change of variable $w = (\log x - \mu)/\sigma$, and $\omega(q) = (\log \pi_q - \mu)/\sigma$. Evoking the formula of *VaR* for lognormal *rv* reported in Example 10.2.2, we can simplify the expression (10.6) into

$$TVaR_q[X] = \frac{e^{\mu+\sigma^2/2}}{(1-q)} \Phi(\Phi^{-1}(q) - \sigma).$$

Clearly, the *TVaR* of lognormal *rv* is not the exponential of the *TVaR* of normal *rv*.

For distributions of which the distribution functions are more tractable to work with, we may apply the integration by parts technique to rewrite equation (10.5) as

$$
\begin{aligned}
TVaR_q[X] &= \left[-xS_X(x)\big|_{\pi_q}^{\infty} + \int_{\pi_q}^{\infty} S_X(x)dx \right] \frac{1}{(1-q)} \\
&= \pi_q + \frac{1}{(1-q)} \int_{\pi_q}^{\infty} S_X(x)dx.
\end{aligned}
$$

Example 10.2.8. *TVaR* of an exponential distribution. Consider an insurance loss *rv* $X \sim Exp(\theta)$ for $\theta > 0$. Give an expression for the *TVaR*.

Show Example Solution

Solution.

We have seen from the previous subsection that

$$\pi_q = -\theta[\log(1-q)].$$

Let us now consider the *TVaR*:

$$
\begin{aligned}
TVaR_q[X] &= \pi_q + \int_{\pi_q}^{\infty} e^{-x/\theta} dx/(1-q) \\
&= \pi_q + \theta e^{-\pi_q/\theta}/(1-q) \\
&= \pi_q + \theta.
\end{aligned}
$$

It can also be helpful to express the *TVaR* in terms of limited expected values. Specifically, we have

$$
\begin{aligned}
TVaR_q[X] &= \int_{\pi_q}^{\infty} (x - \pi_q + \pi_q) f_X(x)dx/(1-q) \\
&= \pi_q + \frac{1}{(1-q)} \int_{\pi_q}^{\infty} (x-\pi_q) f_X(x)dx \\
&= \pi_q + e_X(\pi_q) \\
&= \pi_q + \frac{(\mathrm{E}[X] - \mathrm{E}[X \wedge \pi_q])}{(1-q)},
\end{aligned}
\tag{10.7}
$$

where $e_X(d) := \mathrm{E}[X - d|X > d]$ for $d > 0$ denotes the *mean excess loss function*. For many commonly used parametric distributions, the formulas for calculating $\mathrm{E}[X]$ and $\mathrm{E}[X \wedge \pi_q]$ can be found in a table of distributions.

Example 10.2.9. *TVaR* of the Pareto distribution. Consider a loss *rv* $X \sim Pareto(\theta, \alpha)$ with $\theta > 0$ and $\alpha > 0$. The *cdf* of X is given by

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

Fix $q \in (0,1)$ and set $F_X(\pi_q) = q$, we readily obtain

$$\pi_q = \theta\left[(1-q)^{-1/\alpha} - 1\right]. \tag{10.8}$$

According to the distribution table provided in the Society of Actuaries, we know

$$\mathrm{E}[X] = \frac{\theta}{\alpha - 1},$$

and

$$\mathrm{E}[X \wedge \pi_q] = \frac{\theta}{\alpha - 1}\left[1 - \left(\frac{\theta}{\theta + \pi_q}\right)^{\alpha-1}\right].$$

Evoking equation (10.7) yields

$$\begin{aligned} TVaR_q[X] &= \pi_q + \frac{\theta}{\alpha - 1}\frac{(\theta/(\theta+\pi_q))^{\alpha-1}}{(\theta/(\theta+\pi_q))^{\alpha}} \\ &= \pi_q + \frac{\theta}{\alpha - 1}\left(\frac{\pi_q + \theta}{\theta}\right) \\ &= \pi_q + \frac{\pi_q + \theta}{\alpha - 1}, \end{aligned}$$

where π_q is given by (10.8).

Via a change of variables, we can also rewrite equation (10.5) as

$$TVaR_q[X] = \frac{1}{(1-q)}\int_q^1 VaR_\alpha[X]\, d\alpha. \tag{10.9}$$

What this alternative formula (10.9) tells is that *TVaR* in fact is the average of $VaR_\alpha[X]$ with varying degree of confidence level over $\alpha \in [q, 1]$. Therefore, the *TVaR* effectively resolves most of the limitations of *VaR* outlined in the previous subsection. First, due to the averaging effect, the *TVaR* may be less sensitive to the change of confidence level compared with *VaR*. Second, all the extremal losses that are above the $(1-q) \times 100\%$ worst probable event are taken in account.

In this respect, it is a simple matter for us to see that for any given $q \in (0, 1)$

$$TVaR_q[X] \geq VaR_q[X].$$

Third and perhaps foremost, *TVaR* is a coherent risk measure and thus is able to more accurately capture the diversification effects of insurance portfolio. Herein, we do not intend to provide the proof of the coherent feature for *TVaR*, which is considered to be challenging technically.

10.3 Reinsurance

In this section, you learn how to:

- Define basic reinsurance treaties including proportional, quota share, non-proportional, stop-loss, excess of loss, and surplus share.
- Interpret the optimality of quota share for reinsurers and compute optimal quota share agreements.
- Interpret the optimality of stop-loss for insurers.
- Interpret and calculate optimal excess of loss retention limits.

Recall that *reinsurance*insurance purchased by an insurer is simply insurance purchased by an insurer. Insurance purchased by non-insurers is sometimes known as *primary insurance*insurance purchased by an non-insurer to distinguish it from reinsurance. Reinsurance differs from personal insurance purchased by individuals, such as auto and homeowners insurance, in contract flexibility. Like insurance purchased by major corporations, reinsurance programs are generally tailored more closely to the buyer. For contrast, in personal insurance buyers typically cannot negotiate on the contract terms although they may have a variety of different options (contracts) from which to choose.

The two broad types are proportional and non-proportional reinsurance. A *proportional reinsurance*an agreement between a reinsurer and a ceding company (also known as the reinsured) in which the reinsurer assumes a given percent of losses and premium contract is an agreement between a reinsurer and a *ceding company*a company that purchases reinsurance (also known as the reinsured) (also known as the *reinsured*) in which the reinsurer assumes a given percent of losses and premium. A reinsurance contract is also known as a *treaty*a reinsurance contract. Non-proportional agreements are simply everything else. As examples of non-proportional agreements, this chapter focuses on *stop-loss*Under a stop-loss arrangement, the insurer sets a retention level and pays in full total claims less than the level with the reinsurer paying the excess and *excess of loss* Under an excess of loss arrangement, the insurer sets a retention level for each claim and pays claim amounts less than the level with the reinsurer paying the excesscontracts. For all types of agreements, we split the total risk X into the portion taken on by the reinsurer, $Y_{reinsurer}$, and that retained by the insurer, $Y_{insurer}$, that is, $X = Y_{insurer} + Y_{reinsurer}$.

The mathematical structure of a basic reinsurance treaty is the same as the coverage modifications of personal insurance introduced in Chapter 3. For a proportional reinsurance, the transformation $Y_{insurer} = cX$ is identical to a coinsurance adjustment in personal insurance. For stop-loss reinsurance, the transformation $Y_{reinsurer} = \max(0, X - M)$ is the same as an insurer's payment with a deductible M and $Y_{insurer} = \min(X, M) = X \wedge M$ is equivalent to what a policyholder pays with deductible M. For practical applications of the mathematics, in personal insurance the focus is generally upon the expectation as this is a key ingredient used in pricing. In contrast, for reinsurance the focus is on the entire distribution of the risk, as the extreme events are a primary concern of the financial stability of the insurer and reinsurer.

This section describes the foundational and most basic of reinsurance treaties: Section 10.3.1 for proportional and Section 10.3.2 for non-proportional reinsurance. Section 10.3.3 gives a flavor of more complex contracts.

10.3.1 Proportional Reinsurance

The simplest example of a proportional treaty is called *quota share*A proportional treaty where the reinsurer receives a flat percent of the premium for the book of business reinsured and pays a percentage of losses, including allocated loss adjustment expenses. The reinsurer may also pays the ceding company a ceding commission which is designed to reflect the differences in underwriting expenses incurred..

- In a quota share treaty, the reinsurer receives a flat percent, say 50%, of the premium for the book of business reinsured.
- In exchange, the reinsurer pays 50% of losses, including allocated loss adjustment expenses
- The reinsurer also pays the ceding company a ceding commission which is designed to reflect the differences in underwriting expenses incurred.

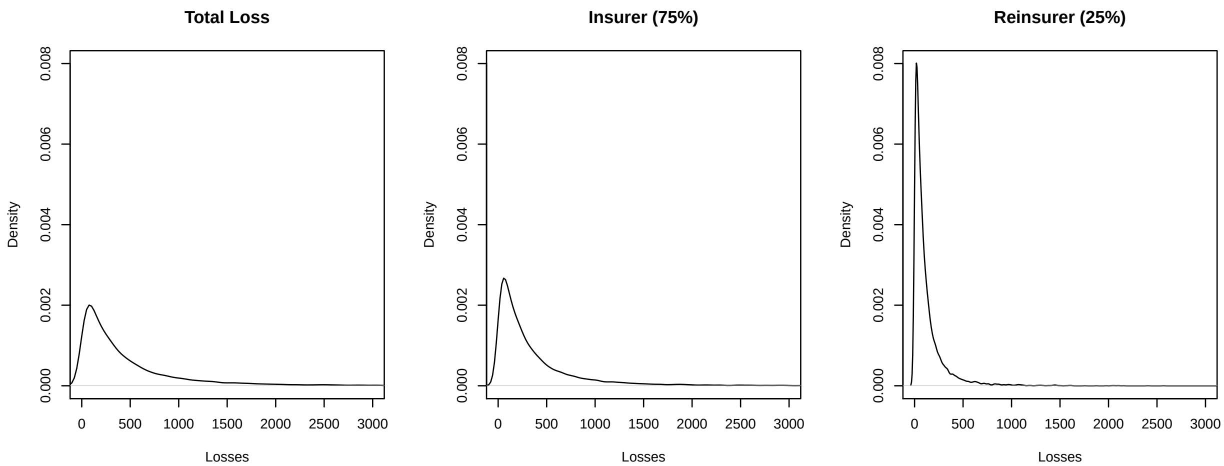
The amounts paid by the direct insurer and the reinsurer are summarized as

$$Y_{insurer} = cX \quad \text{and} \quad Y_{reinsurer} = (1-c)X,$$

where $c \in (0,1)$ denotes the proportion retained by the insurer. Note that $Y_{insurer} + Y_{reinsurer} = X$.

Example 10.3.1. Distribution of losses under quota share. To develop an intuition for the effect of quota-share agreement on the distribution of losses, the following is a short `R` demonstration using simulation. Note the relative shapes of the distributions of total losses, the retained portion (of the insurer), and the reinsurer's portion.

```
Warning: package 'actuar' was built under R version 3.3.3
```



Show the R Code

```
set.seed(2018)
theta = 1000
alpha = 3
nSim = 10000
library(actuar)
X <-  rpareto(nSim, shape = alpha, scale = theta)

par(mfrow=c(1,3))
plot(density(X), xlim=c(0,3*theta), ylim=c(0,0.008), main="Total Loss", xlab="Losses")
plot(density(0.75*X), xlim=c(0,3*theta), ylim=c(0,0.008), main="Insurer (75%)", xlab="Losses")
plot(density(0.25*X), xlim=c(0,3*theta), ylim=c(0,0.008), main="Reinsurer (25%)", xlab="Losses")
```

Quota Share is Desirable for Reinsurers

The quota share contract is particularly desirable for the reinsurer. To see this, suppose that an insurer and reinsurer wish to enter a contract to share total losses X such that

$$Y_{insurer} = g(X) \quad \text{and} \quad Y_{reinsurer} = X - g(X),$$

for some generic function $g(\cdot)$ (known as the *retention* function). Suppose further that the insurer only cares about the variability of retained claims and is indifferent to the choice of g as long as $Var(Y_{insurer})$ stays the same and equals, say, Q. Then, the following result shows that the quota share reinsurance treaty minimizes the reinsurer's uncertainty as measured by $Var(Y_{reinsurer})$.

Proposition. Suppose that $Var(Y_{insurer}) = Q$. Then, $Var((1-c)X) \le Var(g(X))$ for all $g(.)$, where $c = Q/Var(X)$.

Show the Justification of the Proposition

Proof of the Proposition. With $Y_{reinsurer} = X - Y_{insurer}$ and the law of total variation

$$\begin{aligned} Var(Y_{reinsurer}) &= Var(X - Y_{insurer}) \\ &= Var(X) + Var(Y_{insurer}) - 2Cov(X, Y_{insurer}) \\ &= Var(X) + Q - 2Corr(X, Y_{insurer}) \times \sqrt{Q}\sqrt{Var(X)} \end{aligned}$$

In this expression, we see that Q and $Var(X)$ do not change with the choice of g. Thus, we can minimize $Var(Y_{reinsurer})$ by maximizing the correlation $Corr(X, Y_{insurer})$. If we use a quota share reinsurance agreement, then $Corr(X, Y_{insurer}) = Corr(X, (1-c)X) = 1$, the maximum possible correlation. This establishes the proposition.

□‘

The proposition is intuitively appealing - with quota share insurance, the reinsurer shares the responsibility for very large claims in the tail of the distribution. This is in contrast to non-proportional agreements where reinsurers take responsibility for the very large claims.

Optimizing Quota Share Agreements for Insurers

Now assume n risks in the porfolio, $X_1, \ldots, X_n$, so that the portfolio sum is $X = X_1 + \cdots + X_n$. For simplicity, we focus on the case of independent risks. Let us consider a variation of the basic quota share agreement where the amount retained by the insurer may vary with each risk, say c_i. Thus, the insurer's portion of the portfolio risk is $Y_{insurer} = \sum_{i=1}^n c_i X_i$. What is the best choice of the proportions c_i?

To formalize this question, we seek to find those values of c_i that minimize $Var(Y_{insurer})$ subject to the constraint that $E(Y_{insurer}) = K$. The requirement that $E(Y_{insurer}) = K$ suggests that the insurers wishes to retain a revenue in at least the amount of the constant K. Subject to this revenue constraint, the insurer wishes to minimize the uncertainty of the retained risks as measured by the variance.

Show the Optimal Retention Proportions

The Optimal Retention Proportions

Minimizing $Var(Y_{insurer})$ subject to $E(Y_{insurer}) = K$ is a constrained optimization problem - we can use the method of Lagrange multipliers, a calculus technique, to solve this. To this end, define the Lagrangian

$$\begin{aligned} L &= Var(Y_{insurer}) - \lambda(E(Y_{insurer}) - K) \\ &= \sum_{i=1}^n c_i^2 ~Var(X_i) - \lambda(\sum_{i=1}^n c_i ~E(X_i) - K) \end{aligned}$$

Taking a partial derivative with respect to λ and setting this equal to zero simply means that the constraint, $E(Y_{insurer}) = K$, is enforced and we have to choose the proportions c_i to satisfy this constraint. Moreover, taking the partial derivative with respect to each proportion c_i yields

$$\frac{\partial}{\partial c_i} L = 2 c_i ~Var(X_i) - \lambda ~E(X_i) = 0$$

so that

$$c_i = \frac{\lambda}{2} \frac{E(X_i)}{Var(X_i)} .$$

With our constraint, we may determine λ as the solution of

$$\begin{aligned} K &= \sum_{i=1}^3 c_i \mathrm{E}(X_i) \\ &= \frac{\lambda}{2} \sum_{i=1}^3 \frac{\mathrm{E}(X_i)^2}{Var(X_i)} \end{aligned}$$

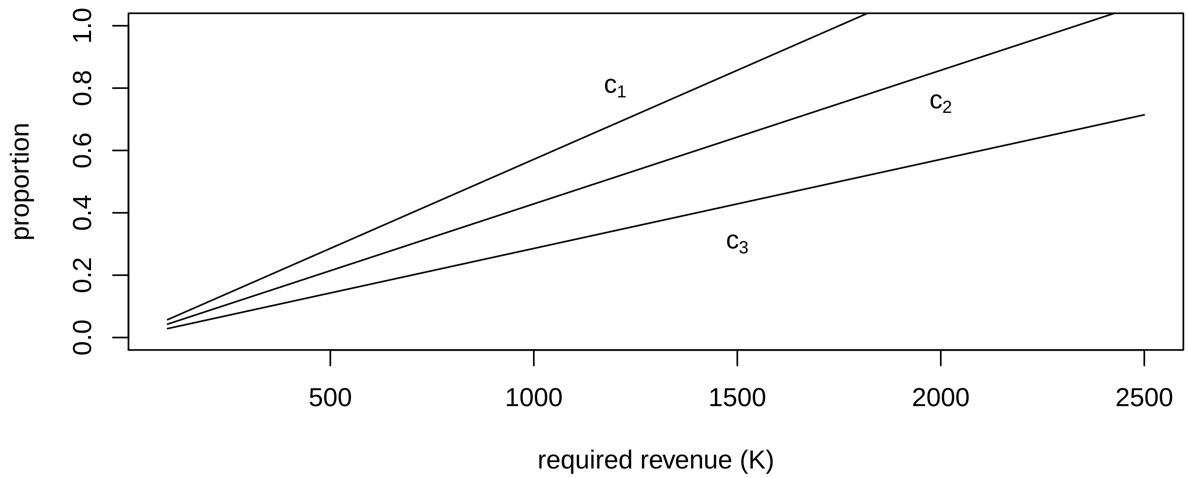
and use this value of λ to determine the proportions.

From the math, it turns out that the constant for the ith risk, c_i is proportional to $\frac{E(X_i)}{Var(X_i)}$. This is intuitively appealing. Other things being equal, a higher revenue as measured by $E(X_i)$ means a higher value of c_i. In the same way, a higher value of uncertainty as measured by $Var(X_i)$ means a lower value of c_i. The proportional scaling factor is determined by the revenue requirement $E(Y_{insurer}) = K$. The following example helps to develop a feel for this relationship.

Example 10.3.2. Three Pareto risks. Consider three risks that have a Pareto distribution. Provide a graph, and supporting code, that give values of c_1, c_2, and c_3 for a required revenue K. Note that these values increase linearly with K.

Show an Example with Three Pareto Risks

```
theta1 = 1000; theta2 = 2000; theta3 = 3000;
alpha1 = 3; alpha2 = 3; alpha3 = 4;
library(actuar)
propnfct <- function(alpha,theta){
  mu    <- mpareto(shape=alpha, scale=theta, order=1)
  var   <- mpareto(shape=alpha, scale=theta, order=2) - mu^2
  mu/var
}
temp <- propnfct(alpha1, theta1)*mpareto(shape=alpha1, scale=theta1, order=1)+
        propnfct(alpha2, theta2)*mpareto(shape=alpha2, scale=theta2, order=1)+
        propnfct(alpha3, theta3)*mpareto(shape=alpha3, scale=theta3, order=1)
KVec <- seq(100, 2500, length.out=20)
Lambdavec <- 2*KVec/temp
c1 <- propnfct(alpha1, theta1)
c2 <- propnfct(alpha2, theta2)
c3 <- propnfct(alpha3, theta3)
c1Vec <- c2Vec <- c3Vec <- 0*KVec
for (j in 1:20) {
  c1Vec[j] <- (Lambdavec[j]/2) * propnfct(alpha1, theta1)
  c2Vec[j] <- (Lambdavec[j]/2) * propnfct(alpha2, theta2)
  c3Vec[j] <- (Lambdavec[j]/2) * propnfct(alpha3, theta3)
  }
plot(KVec, c1Vec, type="l", ylab="proportion", xlab="required revenue (K)", ylim=c(0,1))
lines(KVec, c2Vec)
lines(KVec, c3Vec)
text(1200,0.80, expression(c[1]))
text(2000,0.75, expression(c[2]))
text(1500,0.30, expression(c[3]))
```



10.3.2 Non-Proportional Reinsurance

The Optimality of Stop-Loss Insurance

Under a **stop-loss** arrangement, the insurer sets a retention level $M(>0)$ and pays in full total claims for which $X \leq M$. Further, for claims for which $X > M$, the direct insurer pays M and the reinsurer pays the remaining amount $X - M$. Thus, the insurer retains an amount M of the risk. Summarizing, the amounts paid by the direct insurer and the reinsurer are

$$Y_{insurer} = \begin{cases} X & \text{for } X \leq M \\ M & \text{for } X > M \end{cases} \quad = \min(X, M) = X \wedge M$$

and

$$Y_{reinsurer} = \begin{cases} 0 & \text{for } X \leq M \\ X - M & \text{for } X > M \end{cases} \quad = \max(0, X - M).$$

As before, note that $Y_{insurer} + Y_{reinsurer} = X$.

The stop-loss type of contract is particularly desirable for the insurer. Similar to earlier, suppose that an insurer and reinsurer wish to enter a contract so that $Y_{insurer} = g(X)$ and $Y_{reinsurer} = X - g(X)$ for some generic retention function $g(\cdot)$. Suppose further that the insurer only cares about the variability of retained claims and is indifferent to the choice of g as long as $Var(Y_{insurer})$ can be minimized. Again, we impose the constraint that $E(Y_{insurer}) = K$; the insurer needs to retain a revenue K. Subject to this revenue constraint, the insurer wishes to minimize uncertainty of the retained risks (as measured by the variance). Then, the following result shows that the stop-loss reinsurance treaty minimizes the reinsurer's uncertainty as measured by $Var(Y_{reinsurer})$.

Proposition. Suppose that $E(Y_{insurer}) = K$. Then, $Var(X \wedge M) \leq Var(g(X))$ for all $g(.)$, where M is such that $E(X \wedge M) = K$.

Show the Justification of the Proposition

Proof of the Proposition. Add and subtract a constant M and expand the square to get

$$\begin{aligned}Var(g(X)) &= E(g(X)-K)^2 = E(g(X)-M+M-K)^2 \\ &= E(g(X)-M)^2 + (M-K)^2 + 2E(g(X)-M)(M-K) \\ &= E(g(X)-M)^2 - (M-K)^2,\end{aligned}$$

because $E(g(X)) = K$.

Now, for any retention function, we have $g(X) \leq X$, that is, the insurer's retained claims are less than or equal to total claims. Using the notation $g_{SL}(X) = X \wedge M$ for stop-loss insurance, we have

$$\begin{aligned}M - g_{SL}(X) &= M - (X \wedge M) \\ &= (M - X) \wedge 0 \\ &\leq (M - g(X)) \wedge 0.\end{aligned}$$

Squaring each side yields

$$(M - g_{SL}(X))^2 \leq (M - g(X))^2 \wedge 0 \leq (M - g(X))^2.$$

Returning to our expression for the variance, we have

$$\begin{aligned}Var(g_{SL}(X)) &= E(g_{SL}(X) - M)^2 - (M-K)^2 \\ &\leq E(g(X)-M)^2 - (M-K)^2 = Var(g(X)),\end{aligned}$$

for any retention function g. This establishes the proposition.

□‘

The proposition is intuitively appealing - with stop-loss insurance, the reinsurer takes the responsibility for very large claims in the tail of the distribution, not the insurer.

Excess of Loss

A closely related form of non-proportional reinsurance is the *excess of loss* Under an excess of loss arrangement, the insurer sets a retention level for each claim and pays claim amounts less than the level with the reinsurer paying the excess coverage. Under this contract, we assume that the total risk X can be thought of as composed as n separate risks $X_1, \ldots, X_n$ and that each of these risks are subject to an upper limit, say, M_i. So the insurer retains

$$Y_{i,insurer} = X_i \wedge M_i \qquad Y_{insurer} = \sum_{i=1}^{n} Y_{i,insurer}$$

and the reinsurer is responsible for the excess, $Y_{reinsurer} = X - Y_{insurer}$. The retention limits may vary by risk or may be the same for all risks, $M_i = M$, for all i.

Optimal Choice for Excess of Loss Retention Limits

What is the best choice of the excess of loss retention limits M_i? To formalize this question, we seek to find those values of M_i that minimize $Var(Y_{insurer})$ subject to the constraint that $E(Y_{insurer}) = K$. Subject to this revenue constraint, the insurer wishes to minimize the uncertainty of the retained risks (as measured by the variance).

Show the Optimal Retention Proportions

The Optimal Retention Limits

Minimizing $Var(Y_{insurer})$ subject to $E(Y_{insurer}) = K$ is a constrained optimization problem - we can use the method of Lagrange multipliers, a calculus technique, to solve this. As before, define the Lagrangian

$$\begin{aligned} L &= Var(Y_{insurer}) - \lambda(E(Y_{insurer}) - K) \\ &= \sum_{i=1}^{n} \ Var(X_i \wedge M_i) - \lambda(\sum_{i=1}^{n} \ E(X_i \wedge M_i) - K). \end{aligned}$$

We first recall the relationships

$$E(X \wedge M) = \int_0^M \ (1 - F(x))dx$$

and

$$E(X \wedge M)^2 = 2\int_0^M \ x(1 - F(x))dx.$$

Taking a partial derivative with respect to λ and setting this equal to zero simply means that the constraint, $E(Y_{insurer}) = K$, is enforced and we have to choose the limits M_i to satisfy this constraint. Moreover, taking the partial derivative with respect to each limit M_i yields

$$\begin{aligned} \frac{\partial}{\partial M_i} L &= \frac{\partial}{\partial M_i} \ Var(X_i \wedge M_i) - \lambda \frac{\partial}{\partial M_i} \ E(X_i \wedge M_i) \\ &= \frac{\partial}{\partial M_i} \left(E(X_i \wedge M_i)^2 - (E(X_i \wedge M_i))^2\right) - \lambda(1 - F_i(M_i)) \\ &= 2M_i(1 - F_i(M_i)) - 2E(X_i \wedge M_i)(1 - F_i(M_i)) - \lambda(1 - F_i(M_i)). \end{aligned}$$

Setting $\frac{\partial}{\partial M_i} L = 0$ and solving for λ, we get

$$\lambda = 2(M_i - E(X_i \wedge M_i)).$$

From the math, it turns out that the retention limit less the expected insurer's claims, $M_i - E(X_i \wedge M_i)$, is the same for *all* risks. This is intuitively appealing.

Example 10.3.3. Excess of loss for three Pareto risks. Consider three risks that have a Pareto distribution, each having a different set of parameters (so they are independent but non-identical). Show numerically that the optimal retention limits M_1, M_2, and M_3 resulting retention limit minus expected insurer's claims, $M_i - E(X_i \wedge M_i)$, is the same for all risks, as we derived theoretically. Further, graphically compare the distribution of total risks to that retained by the insurer and by the reinsurer.

Show an Example with Three Pareto Risks

We first optimize the Lagrangian using the `R` package `alabama` for *Augmented Lagrangian Adaptive Barrier Minimization Algorithm*.

```
theta1 = 1000;theta2 = 2000;theta3 = 3000;
alpha1 = 3;   alpha2 = 3;   alpha3 = 4;
Pmin <- 2000
library(actuar)
VarFct <- function(M){
  M1=M[1];M2=M[2];M3=M[3]
  mu1    <- levpareto(limit=M1,shape=alpha1, scale=theta1, order=1)
  var1   <- levpareto(limit=M1,shape=alpha1, scale=theta1, order=2)-mu1^2
  mu2    <- levpareto(limit=M2,shape=alpha2, scale=theta2, order=1)
  var2   <- levpareto(limit=M2,shape=alpha2, scale=theta2, order=2)-mu2^2
  mu3    <- levpareto(limit=M3,shape=alpha3, scale=theta3, order=1)
  var3   <- levpareto(limit=M3,shape=alpha3, scale=theta3, order=2)-mu3^2
  varFct <- var1 +var2+var3
```

```
  meanFct <- mu1+mu2+mu3
  c(meanFct,varFct)
  }
f <- function(M){VarFct(M)[2]}
h <- function(M){VarFct(M)[1] - Pmin}
library(alabama)
par0=rep(1000,3)
op <- auglag(par=par0,fn=f,hin=h,control.outer=list(trace=FALSE))
```

The optimal retention limits M_1, M_2, and M_3 resulting retention limit minus expected insurer's claims, $M_i - E(X_i \wedge M_i)$, is the same for all risks, as we derived theoretically.

```
M1star = op$par[1];M2star = op$par[2];M3star = op$par[3]
M1star -levpareto(M1star,shape=alpha1, scale=theta1,order=1)
```

```
[1] 1344.135
```

```
M2star -levpareto(M2star,shape=alpha2, scale=theta2,order=1)
```

```
[1] 1344.133
```

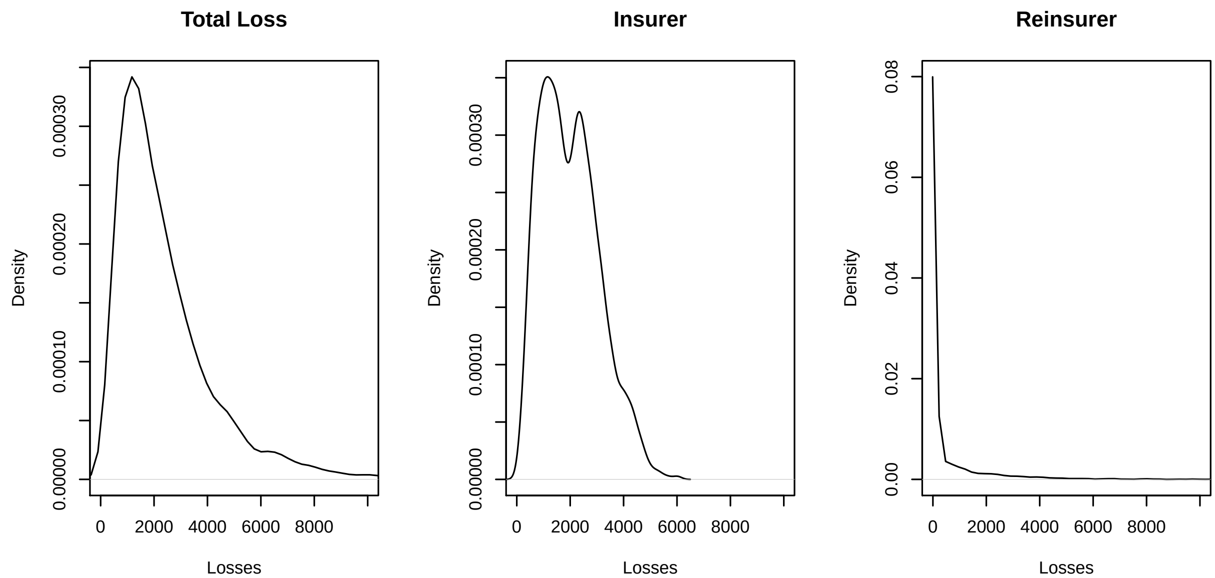
```
M3star -levpareto(M3star,shape=alpha3, scale=theta3,order=1)
```

```
[1] 1344.133
```

We graphically compare the distribution of total risks to that retained by the insurer and by the reinsurer.

```
set.seed(2018)
nSim = 10000
library(actuar)
Y1 <- rpareto(nSim, shape = alpha1, scale = theta1)
Y2 <- rpareto(nSim, shape = alpha2, scale = theta2)
Y3 <- rpareto(nSim, shape = alpha3, scale = theta3)
YTotal <- Y1 + Y2 + Y3
Yinsur <-  pmin(Y1,M1star)+pmin(Y2,M2star)+pmin(Y3,M3star)
Yreinsur <- YTotal - Yinsur

par(mfrow=c(1,3))
plot(density(YTotal),   xlim=c(0,10000), main="Total Loss", xlab="Losses")
plot(density(Yinsur),   xlim=c(0,10000), main="Insurer",    xlab="Losses")
plot(density(Yreinsur), xlim=c(0,10000), main="Reinsurer",  xlab="Losses")
```



10.3.3 Additional Reinsurance Treaties

Surplus Share Proportional Treaty

Another proportional treaty is known as *surplus share*A proportional reinsurance treaty that is common in commercial property insurance. A surplus share treaty allows the reinsured to limit its exposure on any one risk to a given amount (the retained line). The reinsurer assumes a part of the risk in proportion to the amount that the insured value exceeds the retained line, up to a given limit (expressed as a multiple of the retained line, or number of lines).; this type of contract is common in commercial property insurance.

- A surplus share treaty allows the reinsured to limit its exposure on any one risk to a given amount (the *retained line*).
- The reinsurer assumes a part of the risk in proportion to the amount that the insured value exceeds the retained line, up to a given limit (expressed as a multiple of the retained line, or number of lines).

For example, let the retained line be $100,000 and the given limit be 4 lines ($400,000). Then, if X is the loss, the reinsurer's portion is $\min(400000, (X - 100000)_+)$.

Layers of Coverage

One can also extend non-proportional stop-loss treaties by introducing additional parties to the contract. For example, instead of simply an insurer and reinsurer or an insurer and a policyholder, think about the situation with all three parties, a policyholder, insurer, and reinsurer, who agree on how to share a risk. More generally, we consider k parties. If $k = 3$, it could be an insurer and two different reinsurers.

Example 10.3.4. Layers of coverage for three parties.

- Suppose that there are $k = 3$ parties. The first party is responsible for the first 100 of claims, the second responsible for claims from 100 to 3000, and the third responsible for claims above 3000.

- If there are four claims in the amounts 50, 600, 1800 and 4000, then they would be allocated to the parties as follows:

Layer	Claim 1	Claim 2	Claim 3	Claim 4	Total
(0, 100]	50	100	100	100	350
(100, 3000]	0	500	1700	2900	5100
(3000, ∞)	0	0	0	1000	1000
Total	50	600	1800	4000	6450

To handle the general situation with k groups, partition the positive real line into k intervals using the cut-points

$$0 = M_0 < M_1 < \cdots < M_{k-1} < M_k = \infty.$$

Note that the jth interval is $(M_{j-1}, M_j]$. Now let Y_j be the amount of risk shared by the jth party. To illustrate, if a loss x is such that $M_{j-1} < x \leq M_j$, then

$$\left(\begin{array}{c} Y_1 \\ Y_2 \\ \vdots \\ Y_j \\ Y_{j+1} \\ \vdots \\ Y_k \end{array} \right) = \left(\begin{array}{c} M_1 - M_0 \\ M_2 - M_1 \\ \vdots \\ x - M_{j-1} \\ 0 \\ \vdots \\ 0 \end{array} \right)$$

More succinctly, we can write

$$Y_j = \min(X, M_j) - \min(X, M_{j-1}).$$

With the expression $Y_j = \min(X, M_j) - \min(X, M_{j-1})$, we see that the jth party is responsible for claims in the interval $(M_{j-1}, M_j]$. With this, it is easy to check that $X = Y_1 + Y_2 + \cdots + Y_k$. As emphasized in the following example, we also remark that the parties need not be different.

Example 10.3.5. - Suppose that a policyholder is responsible for the first 500 of claims and all claims in excess of 100,000. The insurer takes claims between 100 and 100,000. - Then, we would use $M_1 = 100$, $M_2 = 100000$. - The policyholder is responsible for $Y_1 = \min(X, 100)$ and $Y_3 = X - \min(X, 100000) = \max(0, X - 100000)$.

For additional reading, see the Wisconsin Property Fund site for an example on layers of reinsurance.

Portfolio Management Example

Many other variations of the foundational contracts are possible. For one more illustration, consider the following.

Example. 10.3.6. Portfolio management. You are the Chief Risk Officer of a telecommunications firm. Your firm has several property and liability risks. We will consider:

- 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.
- X_3 - directors and executive officers risk, modeled using a Pareto distribution with mean 1000 and scale parameter 1000.
- X_4 - cyber risks, modeled using a Pareto distribution with mean 1000 and scale parameter 2000.

Denote the total risk as

$$X = X_1 + X_2 + X_3 + X_4.$$

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_1 and M_2, respectively. You seek insurance to cover all other risks. Specifically, the insurer's portion is

$$Y_{insurer} = (X_1 - M_1)_+ + (X_2 - M_2)_+ + X_3 + X_4,$$

so that your retained risk is $Y_{retained} = X - Y_{insurer} = \min(X_1, M_1) + \min(X_2, M_2)$. Using deductibles $M_1 = 100$ and $M_2 = 200$:

a. Determine the expected claim amount of (i) that retained, (ii) that accepted by the insurer, and (iii) the total overall amount.
b. Determine the 80th, 90th, 95th, and 99th percentiles for (i) that retained, (ii) that accepted by the insurer, and (iii) the total overall amount.
c. Compare the distributions by plotting the densities for (i) that retained, (ii) that accepted by the insurer, and (iii) the total overall amount.

Show Example Solution with R Code

In preparation, here is the code needed to set the parameters.

```
# For the gamma distributions, use
alpha1 <- 2;      theta1 <- 100
alpha2 <- 2;      theta2 <- 200
# For the Pareto distributions, use
alpha3 <- 2;      theta3 <- 1000
alpha4 <- 3;      theta4 <- 2000
# Limits
M1     <- 100
M2     <- 200
```

With these parameters, we can now simulate realizations of the portfolio risks.

```
# Simulate the risks
nSim <- 10000  #number of simulations
set.seed(2017) #set seed to reproduce work
X1 <- rgamma(nSim,alpha1,scale = theta1)
X2 <- rgamma(nSim,alpha2,scale = theta2)
# For the Pareto Distribution, use
library(actuar)
X3 <- rpareto(nSim,scale=theta3,shape=alpha3)
X4 <- rpareto(nSim,scale=theta4,shape=alpha4)
# Portfolio Risks
X         <- X1 + X2 + X3 + X4
Yretained <- pmin(X1,M1) + pmin(X2,M2)
Yinsurer  <- X - Yretained
```

(a) Here is the code for the expected claim amounts.

```
# Expected Claim Amounts
ExpVec <- t(as.matrix(c(mean(Yretained),mean(Yinsurer),mean(X))))
colnames(ExpVec) <- c("Retained", "Insurer","Total")
round(ExpVec,digits=2)
```

```
     Retained Insurer   Total
[1,]   269.05 2274.41 2543.46
```

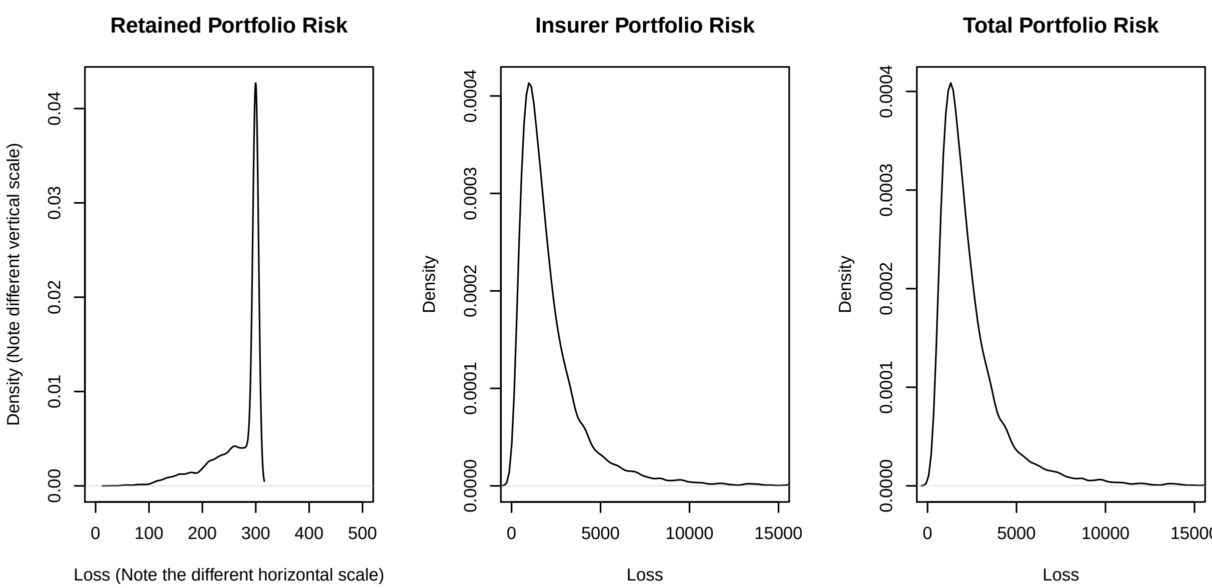
(b) Here is the code for the quantiles.

```
# Quantiles
quantMat <- rbind(
  quantile(Yretained, probs=c(0.80, 0.90, 0.95, 0.99)),
  quantile(Yinsurer,  probs=c(0.80, 0.90, 0.95, 0.99)),
  quantile(X        ,  probs=c(0.80, 0.90, 0.95, 0.99)))
rownames(quantMat) <- c("Retained", "Insurer","Total")
round(quantMat,digits=2)
```

```
             80%     90%     95%      99%
Retained  300.00  300.00  300.00   300.00
Insurer  3075.67 4399.80 6172.69 11859.02
Total    3351.35 4675.04 6464.20 12159.02
```

(c) Here is the code for the density plots of the retained, insurer, and total portfolio risk.

```
par(mfrow=c(1,3))
plot(density(Yretained), xlim=c(0,500), main="Retained Portfolio Risk", xlab="Loss (Note the different h
plot(density(Yinsurer), xlim=c(0,15000), main="Insurer Portfolio Risk", xlab="Loss")
plot(density(X), xlim=c(0,15000), main="Total Portfolio Risk", xlab="Loss")
```



Further Resources and Contributors

- **Edward W. (Jed) Frees**, University of Wisconsin-Madison, and **Jianxi Su**, Purdue University are the principal authors of the initial version of this chapter. Email: jfrees@bus.wisc.edu and/or jianxi@purdue.edu for chapter comments and suggested improvements.
- Chapter reviewers include: Fei Huang, Hirokazu (Iwahiro) Iwasawa, Peng Shi, Ping Wang, Chengguo Weng.

Some of the examples from this chapter were borrowed from Clark (1996), Klugman et al. (2012), and Bahnemann (2015). These resources provide excellent sources for additional discussions and examples.

Chapter 11

Loss Reserving

This is a placeholder file

Chapter 12

Experience Rating using Bonus-Malus

This is a placeholder file

Bonus-Malus

Bonus-malus system, which is used interchangeably as “no-fault discount”, “merit rating”, “experience rating” or “no-claim discount” in different countries, is based on penalizing insureds who are responsible for one or more claims by a premium surcharge, and awarding insureds with a premium discount if they do not have any claims (Frangos and Vrontos, 2001). Insurers use bonus-malus systems for two main purposes; firstly, to encourage drivers to drive more carefully in a year without any claims, and secondly, to ensure insureds to pay premiums proportional to their risks which are based on their claims experience.

NCD and Experience Rating

No Claim Discount (NCD) system is an experience rating system commonly used in motor insurance. NCD system represents an attempt to categorize insureds into homogeneous groups who pay premiums based on their claims experience. Depending on the rules in the scheme, new policyholders may be required to pay full premium initially, and obtain discounts in the future years as a results of claim-free years.

Hunger for Bonus

An NCD system rewards policyholders for not making any claims during a year, or in other words, it grants a bonus to a careful driver. This bonus principle may affect policy holders’ decisions whether to claim or not to claim, especially when involving accidents with slight damages, which is known as ‘hunger for bonus’ phenomenon (Philipson, 1960). The option of ‘hunger for bonus’ implemented on insureds under an NCD system may reduce insurers’ claim costs, and may be able to offset the expected decrease in premium income.

Chapter 13

Data Systems

Chapter Preview. This chapter covers the learning areas on data and systems outlined in the IAA (International Actuarial Association) Education Syllabus published in September 2015.

13.1 Data

13.1.1 Data Types and Sources

In terms of how data are collected, data can be divided into two types (Hox and Boeije, 2005): primary data and secondary data. Primary data are original data that are collected for a specific research problem. Secondary data are data originally collected for a different purpose and reused for another research problem. A major advantage of using primary data is that the theoretical constructs, the research design, and the data collection strategy can be tailored to the underlying research question to ensure that the data collected indeed help to solve the problem. A disadvantage of using primary data is that data collection can be costly and time-consuming. Using secondary data has the advantage of lower cost and faster access to relevant information. However, using secondary data may not be optimal for the research question under consideration.

In terms of the degree of organization of the data, data can be also divided into two types (Inmon and Linstedt, 2014; O'Leary, 2013; Hashem et al., 2015; Abdullah and Ahmad, 2013; Pries and Dunnigan, 2015): structured data and unstructured data. Structured data have a predictable and regularly occurring format. In contrast, unstructured data are unpredictable and have no structure that is recognizable to a computer. Structured data consists of records, attributes, keys, and indices and are typically managed by a database management system (DBMS) such as IBM DB2, Oracle, MySQL, and Microsoft SQL Server. As a result, most units of structured data can be located quickly and easily. Unstructured data have many different forms and variations. One common form of unstructured data is text. Accessing unstructured data is clumsy. To find a given unit of data in a long text, for example, sequentially search is usually performed.

In terms of how the data are measured, data can be classified as qualitative or quantitative. Qualitative data is data about qualities, which cannot be actually measured. As a result, qualitative data is extremely varied in nature and includes interviews, documents, and artifacts (Miles et al., 2014). Quantitative data is data about quantities, which can be measured numerically with numbers. In terms of the level of measurement, quantitative data can be further classified as nominal, ordinal, interval, or ratio (Gan, 2011). Nominal data, also called categorical data, are discrete data without a natural ordering. Ordinal data are discrete data with a natural order. Interval data are continuous data with a specific order and equal intervals. Ratio data are interval data with a natural zero.

There exist a number of data sources. First, data can be obtained from university-based researchers who collect primary data. Second, data can be obtained from organizations that are set up for the purpose of releasing secondary data for general research community. Third, data can be obtained from national and regional statistical institutes that collect data. Finally, companies have corporate data that can be obtained for research purpose.

While it might be difficult to obtain data to address a specific research problem or answer a business question, it is relatively easy to obtain data to test a model or an algorithm for data analysis. In nowadays, readers can obtain datasets from the Internet easily. The following is a list of some websites to obtain real-world data:

- **UCI Machine Learning Repository** This website (url: http://archive.ics.uci.edu/ml/index.php) maintains more than 400 datasets that can be used to test machine learning algorithms.
- **Kaggle** The Kaggle website (url: https://www.kaggle.com/) include real-world datasets used for data science competition. Readers can download data from Kaggle by registering an account.
- **DrivenData** DrivenData aims at bringing cutting-edge practices in data science to solve some of the world's biggest social challenges. In its website (url: https://www.drivendata.org/), readers can participate data science competitions and download datasets.
- **Analytics Vidhya** This website (url: https://datahack.analyticsvidhya.com/contest/all/) allows you to participate and download datasets from practice problems and hackathon problems.
- **KDD Cup** KDD Cup is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge Discovery and Data Mining. This website (url: http://www.kdd.org/kdd-cup) contains the datasets used in past KDD Cup competitions since 1997.
- **U.S. Government's open data** This website (url: https://www.data.gov/) contains about 200,000 datasets covering a wide range of areas including climate, education, energy, and finance.
- **AWS Public Datasets** In this website (url: https://aws.amazon.com/datasets/), Amazon provides a centralized repository of public datasets, including some huge datasets.

13.1.2 Data Structures and Storage

As mentioned in the previous subsection, there are structured data as well as unstructured data. Structured data are highly organized data and usually have the following tabular format:

	V_1	V_2	$\cdots$	V_d
$\mathbf{x}_1$	x_{11}	x_{12}	$\cdots$	x_{1d}
$\mathbf{x}_2$	x_{21}	x_{22}	$\cdots$	x_{2d}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
$\mathbf{x}_n$	x_{n1}	x_{n2}	$\cdots$	x_{nd}

In other words, structured data can be organized into a table consists of rows and columns. Typically, each row represents a record and each column represents an attribute. A table can be decomposed into several tables that can be stored in a relational database such as the Microsoft SQL Server. The SQL (Structured Query Language) can be used to access and modify the data easily and efficiently.

Unstructured data do not follow a regular format (Abdullah and Ahmad, 2013). Examples of unstructured data include documents, videos, and audio files. Most of the data we encounter are unstructured data. In fact, the term "big data" was coined to reflect this fact. Traditional relational databases cannot meet the challenges on the varieties and scales brought by massive unstructured data nowadays. NoSQL databases have been used to store massive unstructured data.

There are three main NoSQL databases (Chen et al., 2014): key-value databases, column-oriented databases, and document-oriented databases. Key-value databases use a simple data model and store data according to key-values. Modern key-value databases have higher expandability and smaller query response time than relational databases. Examples of key-value databases include Dynamo used by Amazon and Voldemort used by LinkedIn. Column-oriented databases store and process data according to columns rather than rows. The columns and rows are segmented in multiple nodes to achieve expandability. Examples of column-oriented databases include BigTable developed by Google and Cassandra developed by FaceBook. Document databases are designed to support more complex data forms than those stored in key-value databases. Examples of document databases include MongoDB, SimpleDB, and CouchDB. MongoDB is an open-source document-oriented database that stores documents as binary objects. SimpleDB is a distributed NoSQL database used by Amazon. CouchDB is an another open-source document-oriented database.

13.1.3 Data Quality

Accurate data are essential to useful data analysis. The lack of accurate data may lead to significant costs to organizations in areas such as correction activities, lost customers, missed opportunities, and incorrect decisions (Olson, 2003).

Data has quality if it satisfies its intended use, that is, the data is accurate, timely, relevant, complete, understood, and trusted (Olson, 2003). As a result, we first need to know the specification of the intended uses and then judge the suitability for those uses in order to assess the quality of the data. Unintended uses of data can arise from a variety of reasons and lead to serious problems.

Accuracy is the single most important component of high-quality data. Accurate data have the following properties (Olson, 2003):

- The data elements are not missing and have valid values.
- The values of the data elements are in the right ranges and have the right representations.

Inaccurate data arise from different sources. In particular, the following areas are common areas where inaccurate data occur:

- Initial data entry. Mistakes (including deliberate errors) and system errors can occur during the initial data entry. Flawed data entry processes can result in inaccurate data.
- Data decay. Data decay, also known as data degradation, refers to the gradual corruption of computer data due to an accumulation of non-critical failures in a storage device.
- Data moving and restructuring. Inaccurate data can also arise from data extracting, cleaning, transforming, loading, or integrating.
- Data using. Faulty reporting and lack of understanding can lead to inaccurate data.

Reverification and analysis are two approaches to find inaccurate data elements. To ensure that the data elements are 100% accurate, we must use reverification. However, reverification can be time-consuming and may not be possible for some data. Analytical techniques can also be used to identify inaccurate data elements. There are five types of analysis that can be used to identify inaccurate data (Olson, 2003): data element analysis, structural analysis, value correlation, aggregation correlation, and value inspection

Companies can create a data quality assurance program to create high-quality databases. For more information about data quality issues management and data profiling techniques, readers are referred to (Olson, 2003).

13.1.4 Data Cleaning

Raw data usually need to be cleaned before useful analysis can be conducted. In particular, the following areas need attention when preparing data for analysis (Janert, 2010):

- **Missing values** It is common to have missing values in raw data. Depending on the situations, we can discard the record, discard the variable, or impute the missing values.

- **Outliers** Raw data may contain unusual data points such as outliers. We need to handle outliers carefully. We cannot just remove outliers without knowing the reason for their existence. Sometimes the outliers are caused by clerical errors. Sometimes outliers are the effect we are looking for.
- **Junk** Raw data may contain junks such as nonprintable characters. Junks are typically rare and not easy to get noticed. However, junks can cause serious problems in downstream applications.
- **Format** Raw data may be formated in a way that is inconvenient for subsequent analysis. For example, components of a record may be split into multiple lines in a text file. In such cases, lines corresponding to a single record should be merged before loading to a data analysis software such as R.
- **Duplicate records** Raw data may contain duplicate records. Duplicate records should be recognized and removed. This task may not be trivial depending on what you consider "duplicate."
- **Merging datasets** Raw data may come from different sources. In such cases, we need to merge the data from different sources to ensure compatibility.

For more information about how to handle data in R, readers are referred to (Forte, 2015) and (Buttrey and Whitaker, 2017).

13.2 Data Analysis Preliminary

Data analysis involves inspecting, cleansing, transforming, and modeling data to discover useful information to suggest conclusions and make decisions. Data analysis has a long history. In 1962, statistician John Tukey defined data analysis as:

> procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.
>
> — (Tukey, 1962)

Recently, Judd and coauthors defined data analysis as the following equation(Judd et al., 2017):

$$\text{Data} = \text{Model} + \text{Error},$$

where Data represents a set of basic scores or observations to be analyzed, Model is a compact representation of the data, and Error is simply the amount the model fails to represent accurately. Using the above equation for data analysis, an analyst must resolve the following two conflicting goals:

- to add more parameters to the model so that the model represents the data better.
- to remove parameters from the model so that the model is simple and parsimonious.

In this section, we give a high-level introduction to data analysis, including different types of methods.

13.2.1 Data Analysis Process

Data analysis is part of an overall study. For example, Figure 13.1 shows the process of a typical study in behavioral and social sciences as described in (Albers, 2017). The data analysis part consists of the following steps:

- **Exploratory analysis** The purpose of this step is to get a feel of the relationships with the data and figure out what type of analysis for the data makes sense.
- **Statistical analysis** This step performs statistical analysis such as determining statistical significance and effect size.

Figure 13.1: The process of a typical study in behavioral and social sciences.

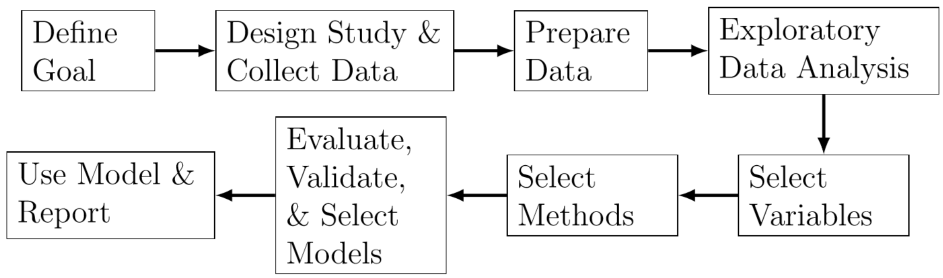


Figure 13.2: The process of statistical modeling.

- **Make sense of the results** This step interprets the statistical results in the context of the overall study.
- **Determine implications** This step interprets the data by connecting it to the study goals and the larger field of this study.

The goal of the data analysis as described above focuses on explaining some phenomenon (See Section 13.2.5).

Shmueli (2010) described a general process for statistical modeling, which is shown in Figure 13.2. Depending on the goal of the analysis, the steps differ in terms of the choice of methods, criteria, data, and information.

13.2.2 Exploratory versus Confirmatory

There are two phases of data analysis (Good, 1983): exploratory data analysis (EDA) and confirmatory data analysis (CDA). Table 13.1 summarizes some differences between EDA and CDA. EDA is usually applied to observational data with the goal of looking for patterns and formulating hypotheses. In contrast, CDA is often applied to experimental data (i.e., data obtained by means of a formal design of experiments) with the goal of quantifying the extent to which discrepancies between the model and the data could be expected to occur by chance (Gelman, 2004).

	EDA	**CDA**
Data	Observational data	Experimental data
Goal	Pattern recognition, formulate hypotheses	Hypothesis testing, estimation, prediction
Techniques	Descriptive statistics, visualization, clustering	Traditional statistical tools of inference, significance, and confidence

Table 13.1: Comparison of exploratory data analysis and confirmatory data analysis.

Techniques for EDA include descriptive statistics (e.g., mean, median, standard deviation, quantiles), distributions, histograms, correlation analysis, dimension reduction, and cluster analysis. Techniques for CDA include the traditional statistical tools of inference, significance, and confidence.

13.2.3 Supervised versus Unsupervised

Methods for data analysis can be divided into two types (Abbott, 2014; Igual and Segu, 2017): supervised learning methods and unsupervised learning methods. Supervised learning methods work with labeled data, which include a target variable. Mathematically, supervised learning methods try to approximate the following function:

$$Y = f(X_1, X_2, \ldots, X_p),$$

where Y is a target variable and X_1, X_2, ..., X_p are explanatory variables. Other terms are also used to mean a target variable. Table 13.2 gives a list of common names for different types of variables (Frees, 2009c). When the target variable is a categorical variable, supervised learning methods are called classification methods. When the target variable is continuous, supervised learning methods are called regression methods.

Target Variable	**Explanatory Variable**
Dependent variable	Independent variable
Response	Treatment
Output	Input
Endogenous variable	Exogenous variable
Predicted variable	Predictor variable
Regressand	Regressor

Table 13.2: Common names of different variables.

Unsupervised learning methods work with unlabeled data, which include explanatory variables only. In other words, unsupervised learning methods do not use target variables. As a result, unsupervised learning methods are also called descriptive modeling methods.

13.2.4 Parametric versus Nonparametric

Methods for data analysis can be parametric or nonparametric (Abbott, 2014). Parametric methods assume that the data follow a certain distribution. Nonparametric methods do not assume distributions for the data and therefore are called distribution-free methods.

Parametric methods have the advantage that if the distribution of the data is known, properties of the data and properties of the method (e.g., errors, convergence, coefficients) can be derived. A disadvantage of parametric methods is that analysts need to spend considerable time on figuring out the distribution. For example, analysts may try different transformation methods to transform the data so that it follows a certain distribution.

Since nonparametric methods make fewer assumptions, nonparametric methods have the advantage that they are more flexible, more robust, and applicable to non-quantitative data. However, a drawback of nonparametric methods is that the conclusions drawn from nonparametric methods are not as powerful as those drawn from parametric methods.

13.2.5 Explanation versus Prediction

There are two goals in data analysis (Breiman, 2001; Shmueli, 2010): explanation and prediction. In some scientific areas such as economics, psychology, and environmental science, the focus of data analysis is to explain the causal relationships between the input variables and the response variable. In other scientific areas such as natural language processing and bioinformatics, the focus of data analysis is to predict what the responses are going to be given the input variables.

Shmueli (2010) discussed in detail the distinction between explanatory modeling and predictive modeling, which reflect the process of using data and methods for explaining or predicting, respectively. Explanatory modeling is commonly used for theory building and testing. However, predictive modeling is rarely used in many scientific fields as a tool for developing theory.

Explanatory modeling is typically done as follows:

- State the prevailing theory.
- State causal hypotheses, which are given in terms of theoretical constructs rather than measurable variables. A causal diagram is usually included to illustrate the hypothesized causal relationship between the theoretical constructs.
- Operationalize constructs. In this step, previous literature and theoretical justification are used to build a bridge between theoretical constructs and observable measurements.
- Collect data and build models alongside the statistical hypotheses, which are operationalized from the research hypotheses.
- Reach research conclusions and recommend policy. The statistical conclusions are converted into research conclusions. Policy recommendations are often accompanied.

Shmueli (2010) defined predictive modeling as the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations. Predictions include point predictions, interval predictions, regions, distributions, and rankings of new observations. Predictive model can be any method that produces predictions.

13.2.6 Data Modeling versus Algorithmic Modeling

Breiman (2001) discussed two cultures in the use of statistical modeling to reach conclusions from data: the data modeling culture and the algorithmic modeling culture. In the data modeling culture, the data are assumed to be generated by a given stochastic data model. In the algorithmic modeling culture, the data mechanism is treated as unknown and algorithmic models are used.

Data modeling gives the statistics field many successes in analyzing data and getting information about the data mechanisms. However, Breiman (2001) argued that the focus on data models in the statistical community has led to some side effects such as

- Produced irrelevant theory and questionable scientific conclusions.
- Kept statisticians from using algorithmic models that might be more suitable.
- Restricted the ability of statisticians to deal with a wide range of problems.

Algorithmic modeling was used by industrial statisticians long time ago. However, the development of algorithmic methods was taken up by a community outside statistics (Breiman, 2001). The goal of algorithmic modeling is predictive accuracy. For some complex prediction problems, data models are not suitable. These prediction problems include speech recognition, image recognition, handwriting recognition, nonlinear time series prediction, and financial market prediction. The theory in algorithmic modeling focuses on the properties of algorithms, such as convergence and predictive accuracy.

13.2.7 Big Data Analysis

Unlike traditional data analysis, big data analysis employs additional methods and tools that can extract information rapidly from massive data. In particular, big data analysis uses the following processing methods

(Chen et al., 2014):

- **Bloom filter** A bloom filter is a space-efficient probabilistic data structure that is used to determine whether an element belongs to a set. It has the advantages of high space efficiency and high query speed. A drawback of using bloom filter is that there is a certain misrecognition rate.
- **Hashing** Hashing is a method that transforms data into fixed-length numerical values through a hash function. It has the advantages of rapid reading and writing. However, sound hash functions are difficult to find.
- **Indexing** Indexing refers to a process of partitioning data in order to speed up reading. Hashing is a special case of indexing.
- **Tries** A trie, also called digital tree, is a method to improve query efficiency by using common prefixes of character strings to reduce comparison on character strings to the greatest extent.
- **Parallel computing** Parallel computing uses multiple computing resources to complete a computation task. Parallel computing tools include MPI (Message Passing Interface), MapReduce, and Dryad.

Big data analysis can be conducted in the following levels (Chen et al., 2014): memory-level, business intelligence (BI) level, and massive level. Memory-level analysis is conducted when the data can be loaded to the memory of a cluster of computers. Current hardware can handle hundreds of gigabytes (GB) of data in memory. BI level analysis can be conducted when the data surpass the memory level. It is common for BI level analysis products to support data over terabytes (TB). Massive level analysis is conducted when the data surpass the capabilities of products for BI level analysis. Usually Hadoop and MapReduce are used in massive level analysis.

13.2.8 Reproducible Analysis

As mentioned in Section 13.2.1, a typical data analysis workflow includes collecting data, analyzing data, and reporting results. The data collected are saved in a database or files. The data are then analyzed by one or more scripts, which may save some intermediate results or always work on the raw data. Finally a report is produced to describe the results, which include relevant plots, tables, and summaries of the data. The workflow may subject to the following potential issues (Mailund, 2017, Chapter 2):

- The data are separated from the analysis scripts.
- The documentation of the analysis is separated from the analysis itself.

If the analysis is done on the raw data with a single script, then the first issue is not a major problem. If the analysis consists of multiple scripts and a script saves intermediate results that are read by the next script, then the scripts describe a workflow of data analysis. To reproduce an analysis, the scripts have to be executed in the right order. The workflow may cause major problems if the order of the scripts is not documented or the documentation is not updated or lost. One way to address the first issue is to write the scripts so that any part of the workflow can be run completely automatically at any time.

If the documentation of the analysis is synchronized with the analysis, then the second issue is not a major problem. However, the documentation may become completely useless if the scripts are changed but the documentation is not updated.

Literate programming is an approach to address the two issues mentioned above. In literate programming, the documentation of a program and the code of the program are written together. To do literate programming in R, one way is to use the R Markdown and the `knitr` package.

13.2.9 Ethical Issues

Analysts may face ethical issues and dilemmas during the data analysis process. In some fields, for example, ethical issues and dilemmas include participant consent, benefits, risk, confidentiality, and data ownership

(Miles et al., 2014). For data analysis in actuarial science and insurance in particular, we face the following ethical matters and issues (Miles et al., 2014):

- **Worthness of the project** Is the project worth doing? Will the project contribute in some significant way to a domain broader than my career? If a project is only opportunistic and does not have a larger significance, then it might be pursued with less care. The result may be looked good but not right.
- **Competence** Do I or the whole team have the expertise to carry out the project? Incompetence may lead to weakness in the analytics such as collecting large amounts of data poorly and drawing superficial conclusions.
- **Benefits, costs, and reciprocity** Will each stakeholder gain from the project? Are the benefit and the cost equitable? A project will likely to fail if the benefit and the cost for a stakeholder do not match.
- **Privacy and confidentiality** How do we make sure that the information is kept confidentially? Where raw data and analysis results are stored and how will have access to them should be documented in explicit confidentiality agreements.

13.3 Data Analysis Techniques

Techniques for data analysis are drawn from different but overlapping fields such as statistics, machine learning, pattern recognition, and data mining. Statistics is a field that addresses reliable ways of gathering data and making inferences based on them (Bandyopadhyay and Forster, 2011; Bluman, 2012). The term machine learning was coined by Samuel in 1959 (Samuel, 1959). Originally, machine learning refers to the field of study where computers have the ability to learn without being explicitly programmed. Nowadays, machine learning has evolved to the broad field of study where computational methods use experience (i.e., the past information available for analysis) to improve performance or to make accurate predictions (Bishop, 2007; Clarke et al., 2009; Mohri et al., 2012; Kubat, 2017). There are four types of machine learning algorithms (See Table 13.3 depending on the type of the data and the type of the learning tasks.

	Supervised	**Unsupervised**
Discrete Label	Classification	Clustering
Continuous Label	Regression	Dimension reduction

Table 13.3: Types of machine learning algorithms.

Originating in engineering, pattern recognition is a field that is closely related to machine learning, which grew out of computer science. In fact, pattern recognition and machine learning can be considered to be two facets of the same field (Bishop, 2007). Data mining is a field that concerns collecting, cleaning, processing, analyzing, and gaining useful insights from data (Aggarwal, 2015).

13.3.1 Exploratory Techniques

Exploratory data analysis techniques include descriptive statistics as well as many unsupervised learning techniques such as data clustering and principal component analysis.

13.3.2 Descriptive Statistics

In the mass noun sense, descriptive statistics is an area of statistics that concerns the collection, organization, summarization, and presentation of data (Bluman, 2012). In the count noun sense, descriptive statistics are summary statistics that quantitatively describe or summarize data.

	Descriptive Statistics
Measures of central tendency	Mean, median, mode, midrange
Measures of variation	Range, variance, standard deviation
Measures of position	Quantile

Table 13.4: Some commonly used descriptive statistics.

Table 13.4 lists some commonly used descriptive statistics. In R, we can use the function `summary` to calculate some of the descriptive statistics. For numeric data, we can visualize the descriptive statistics using a boxplot.

In addition to these quantitative descriptive statistics, we can also qualitatively describe shapes of the distributions (Bluman, 2012). For example, we can say that a distribution is positively skewed, symmetric, or negatively skewed. To visualize the distribution of a variable, we can draw a histogram.

Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that transforms a dataset described by possibly correlated variables into a dataset described by linearly uncorrelated variables, which are called principal components and are ordered according to their variances. PCA is a technique for dimension reduction. If the original variables are highly correlated, then the first few principal components can account for most of the variation of the original data.

To describe PCA, let $X_1, X_2, \ldots, X_d$ be a set of variables. The first principal component is defined to be the normalized linear combination of the variables that has the largest variance, that is, the first principal component is defined as

$$Z_1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1d}X_d,$$

where $\mathbf{w}_1 = (w_{11}, w_{12}, \ldots, w_{1d})'$ is a vector of loadings such that Var (Z_1) is maximized subject to the following constraint:

$$\mathbf{w}_1'\mathbf{w}_1 = \sum_{j=1}^{d} w_{1j}^2 = 1.$$

For $i = 2, 3, \ldots, d$, the ith principal component is defined as

$$Z_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{id}X_d,$$

where $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{id})'$ is a vector of loadings such that Var (Z_i) is maximized subject to the following constraints:

$$\mathbf{w}_i'\mathbf{w}_i = \sum_{j=1}^{d} w_{ij}^2 = 1,$$

$$\text{cov } (Z_i, Z_j) = 0, \quad j = 1, 2, \ldots, i-1.$$

The principal components of the variables are related to the eigenvectors and eigenvectors of the covariance matrix of the variables. For $i = 1, 2, \ldots, d$, let $(\lambda_i, \mathbf{e}_i)$ be the ith eigenvalue-eigenvector pair of the covariance matrix Σ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ and the eigenvectors are normalized. Then the ith principal component is given by

$$Z_i = \mathbf{e}_i'\mathbf{X} = \sum_{j=1}^{d} e_{ij}X_j,$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_d)'$. It can be shown that Var $(Z_i) = \lambda_i$. As a result, the proportion of variance explained by the ith principal component is calculated as

$$\frac{\text{Var }(Z_i)}{\sum_{j=1}^{d} \text{Var }(Z_j)} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_d}.$$

For more information about PCA, readers are referred to (Mirkin, 2011).

13.3.3 Cluster Analysis

Cluster analysis (aka data clustering) refers to the process of dividing a dataset into homogeneous groups or clusters such that points in the same cluster are similar and points from different clusters are quite distinct (Gan et al., 2007; Gan, 2011). Data clustering is one of the most popular tools for exploratory data analysis and has found applications in many scientific areas.

During the past several decades, many clustering algorithms have been proposed. Among these clustering algorithms, the k-means algorithm is perhaps the most well-known algorithm due to its simplicity. To describe the k-means algorithm, let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a dataset containing n points, each of which is described by d numerical features. Given a desired number of clusters k, the k-means algorithm aims at minimizing the following objective function:

$$P(U, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} \|\mathbf{x}_i - \mathbf{z}_l\|^2,$$

where $U = (u_{il})_{n\times k}$ is an $n \times k$ partition matrix, $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$ is a set of cluster centers, and $\|\cdot\|$ is the L^2 norm or Euclidean distance. The partition matrix U satisfies the following conditions:

$$u_{il} \in \{0, 1\}, \quad i = 1, 2, \ldots, n,\ l = 1, 2, \ldots, k,$$

$$\sum_{l=1}^{k} u_{il} = 1, \quad i = 1, 2, \ldots, n.$$

The k-means algorithm employs an iterative procedure to minimize the objective function. It repeatedly updates the partition matrix U and the cluster centers Z alternately until some stop criterion is met. When the cluster centers Z are fixed, the partition matrix U is updated as follows:

$$u_{il} = \begin{cases} 1, & \text{if } \|\mathbf{x}_i - \mathbf{z}_l\| = \min_{1\le j\le k} \|\mathbf{x}_i - \mathbf{z}_j\|; \\ 0, & \text{if otherwise,} \end{cases}$$

When the partition matrix U is fixed, the cluster centers are updated as follows:

$$z_{lj} = \frac{\sum_{i=1}^{n} u_{il} x_{ij}}{\sum_{i=1}^{n} u_{il}}, \quad l = 1, 2, \ldots, k,\ j = 1, 2, \ldots, d,$$

where z_{lj} is the jth component of $\mathbf{z}_l$ and x_{ij} is the jth component of $\mathbf{x}_i$.

For more information about k-means, readers are referred to (Gan et al., 2007) and (Mirkin, 2011).

13.3.4 Confirmatory Techniques

Confirmatory data analysis techniques include the traditional statistical tools of inference, significance, and confidence.

Linear Models

Linear models, also called linear regression models, aim at using a linear function to approximate the relationship between the dependent variable and independent variables. A linear regression model is called a simple linear regression model if there is only one independent variable. When more than one independent variables are involved, a linear regression model is called a multiple linear regression model.

Let X and Y denote the independent and the dependent variables, respectively. For $i = 1, 2, \ldots, n$, let (x_i, y_i) be the observed values of (X, Y) in the ith case. Then the simple linear regression model is specified as follows (Frees, 2009c):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where β_0 and β_1 are parameters and ϵ_i is a random variable representing the error for the ith case.

When there are multiple independent variables, the following multiple linear regression model is used:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where β_0, β_1, ..., β_k are unknown parameters to be estimated.

Linear regression models usually make the following assumptions:

(a) $x_{i1}, x_{i2}, \ldots, x_{ik}$ are nonstochastic variables.

(b) Var $(y_i) = \sigma^2$, where Var (y_i) denotes the variance of y_i.

(c) $y_1, y_2, \ldots, y_n$ are independent random variables.

For the purpose of obtaining tests and confidence statements with small samples, the following strong normality assumption is also made:

(d) $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are normally distributed.

Generalized Linear Models

The generalized linear model (GLM) is a wide family of regression models that include linear regression models as special cases. In a GLM, the mean of the response (i.e., the dependent variable) is assumed to be a function of linear combinations of the explanatory variables, i.e.,

$$\mu_i = E[y_i],$$
$$\eta_i = \mathbf{x}_i' \boldsymbol{\beta} = g(\mu_i),$$

where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})'$ is a vector of regressor values, μ_i is the mean response for the ith case, and η_i is a systematic component of the GLM. The function $g(\cdot)$ is known and is called the link function. The mean response can vary by observations by allowing some parameters to change. However, the regression parameters $\boldsymbol{\beta}$ are assumed to be the same among different observations.

GLMs make the following assumptions:

(a) $x_{i1}, x_{i2}, \ldots, x_{in}$ are nonstochastic variables.

(b) $y_1, y_2, \ldots, y_n$ are independent.

(c) The dependent variable is assumed to follow a distribution from the linear exponential family.

(d) The variance of the dependent variable is not assumed to be constant but is a function of the mean, i.e.,

$$\text{Var}\ (y_i) = \phi\nu(\mu_i),$$

where ϕ denotes the dispersion parameter and $\nu(\cdot)$ is a function.

As we can see from the above specification, the GLM provides a unifying framework to handle different types of dependent variables, including discrete and continuous variables. For more information about GLMs, readers are referred to (de Jong and Heller, 2008) and (Frees, 2009c).

Tree-based Models

Decision trees, also known as tree-based models, involve dividing the predictor space (i.e., the space formed by independent variables) into a number of simple regions and using the mean or the mode of the region for prediction (Breiman et al., 1984). There are two types of tree-based models: classification trees and regression trees. When the dependent variable is categorical, the resulting tree models are called classification trees. When the dependent variable is continuous, the resulting tree models are called regression trees.

The process of building classification trees is similar to that of building regression trees. Here we only briefly describe how to build a regression tree. To do that, the predictor space is divided into non-overlapping regions such that the following objective function

$$f(R_1, R_2, \ldots, R_J) = \sum_{j=1}^{J} \sum_{i=1}^{n} I_{R_j}(\mathbf{x}_i)(y_i - \mu_j)^2$$

is minimized, where I is an indicator function, R_j denotes the set of indices of the observations that belong to the jth box, μ_j is the mean response of the observations in the jth box, $\mathbf{x}_i$ is the vector of predictor values for the ith observation, and y_i is the response value for the ith observation.

In terms of predictive accuracy, decision trees generally do not perform to the level of other regression and classification models. However, tree-based models may outperform linear models when the relationship between the response and the predictors is nonlinear. For more information about decision trees, readers are referred to (Breiman et al., 1984) and (Mitchell, 1997).

13.4 Some R Functions

R is an open-source software for statistical computing and graphics. The R software can be downloaded from the R project website at https://www.r-project.org/. In this section, we give some R function for data analysis, especially the data analysis tasks mentioned in previous sections.

Data Analysis Task	R package	R Function
Descriptive Statistics	`base`	`summary`
Principal Component Analysis	`stats`	`prcomp`
Data Clustering	`stats`	`kmeans`, `hclust`
Fitting Distributions	`MASS`	`fitdistr`
Linear Regression Models	`stats`	`lm`
Generalized Linear Models	`stats`	`glm`
Regression Trees	`rpart`	`rpart`
Survival Analysis	`survival`	`survfit`

Table 13.5: Some R functions for data analysis.

Table 13.5 lists a few R functions for different data analysis tasks. Readers can read the R documentation for examples of using these functions. There are also other R functions from other packages to do similar things. However, the functions listed in this table provide good start points for readers to conduct data analysis in R. For analyzing large datasets in R in an efficient way, readers are referred to (Daroczi, 2015).

13.5 Summary

In this chapter, we gave a high-level overview of data analysis. The overview is divided into three major parts: data, data analysis, and data analysis techniques. In the first part, we introduced data types, data structures, data storages, and data sources. In particular, we provided several websites where readers can obtain real-world datasets to horn their data analysis skills. In the second part, we introduced the process of data analysis and various aspects of data analysis. In the third part, we introduced some commonly used techniques for data analysis. In addition, we listed some R packages and functions that can be used to perform various data analysis tasks.

13.6 Further Resources and Contributors

Contributor

- **Guojun Gan**, University of Connecticut, is the principal author of the initial version of this chapter. Email: guojun.gan@uconn.edu for chapter comments and suggested improvements.

Chapter 14

Dependence Modeling

Chapter Preview. In practice, there are many types of variables that one encounter and the first step in dependence modeling is identifying the type of variable you are dealing with to help direct you to the appropriate technique.This chapter introduces readers to variable types and techniques for modeling dependence or association of multivariate distributions. Section 14.1 provides an overview of the types of variables. Section 14.2 then elaborates basic measures for modeling the dependence between variables.

Section 14.3 introduces a novel approach to modeling dependence using Copulas which is reinforced with practical illustrations in Section 14.4. The types of Copula families and basic properties of Copula functions is explained Section 14.5. The chapter concludes by explaining why the study of dependence modeling is important in Section 14.6.

14.1 Variable Types

In this section, you learn how to:

- Classify variables as qualitative or quantitative.
- Describe multivariate variables.

People, firms, and other entities that we want to understand are described in a dataset by numerical characteristics. As these characteristics vary by entity, they are commonly known as *variables*. To manage insurance systems, it will be critical to understand the distribution of each variable and how they are associated with one another. It is common for data sets to have many variables (high dimensional) and so it useful to begin by classifying them into different types. As will be seen, these classifications are not strict; there is overlap among the groups. Nonetheless, the grouping summarized in Table 14.1 and explained in the remainder of this section provide a solid first step in framing a data set.

Variable Type	**Example**
Qualitative	
Binary	Sex
Categorical (Unordered, Nominal)	Territory (e.g., state/province) in which an insured resides
Ordered Category (Ordinal)	Claimant satisfaction (five point scale ranging from 1=dissatisfied to 5 =satisfied)
Quantitative	
Continuous	Policyholder's age, weight, income
Discrete	Amount of deductible
Count	Number of insurance claims
Combinations of Discrete and Continuous	Policy losses, mixture of 0's (for no loss) and positive claim amount
Interval Variable	Driver Age: 16-24 (young), 25-54 (intermediate), 55 and over (senior)
Circular Data	Time of day measures of customer arrival
Multivariate Variable	
High Dimensional Data	Characteristics of a firm purchasing worker's compensation insurance (location of plants, industry, number of employees, and so on)
Spatial Data	Longitude/latitude of the location an insurance hailstorm claim
Missing Data	Policyholder's age (continuous/interval) and "-99" for "not reported," that is, missing
Censored and Truncated Data	Amount of insurance claims in excess of a deductible
Aggregate Claims	Losses recorded for each claim in a motor vehicle policy.
Stochastic Process Realizations	The time and amount of each occurrence of an insured loss

Table 14.1 : Variable types

In data analysis, it is important to understand what type of variable you are working with. For example, Consider a pair of random variables *(Coverage,Claim)* from the LGPIF data introduced in chapter 1 as displayed in Figure 14.1 below. We would like to know whether the distribution of *Coverage* depends on the distribution of *Claim* or whether they are statistically independent. We would also want to know how the *Claim* distribution depends on the *EntityType* variable. Because the *EntityType* variable belongs to a different class of variables, modeling the dependence between *Claim* and *Coverage* may require a different technique from that of *Claim* and *EntityType*.

14.1.1 Qualitative Variables

In this sub-section, you learn how to:

- Classify qualitative variables as nominal or ordinal
- Describe binary variable

A *qualitative*, or *categorical*, variable is one for which the measurement denotes membership in a set of groups, or categories. For example, if you were coding which area of the country an insured resides, you might use a 1 for the northern part, 2 for southern, and 3 for everything else. This location variable is an example of a *nominal* variable, one for which the levels have no natural ordering. Any analysis of nominal variables should not depend on the labeling of the categories. For example, instead of using a 1,2,3 for north, south, other, I should arrive at the same set of summary statistics if I used a 2,1,3 coding instead, interchanging north and south.

In contrast, an *ordinal* variable is a type of categorical variable for which an ordering does exist. For example, with a survey to see how satisfied customers are with our claims servicing department, we might use a five

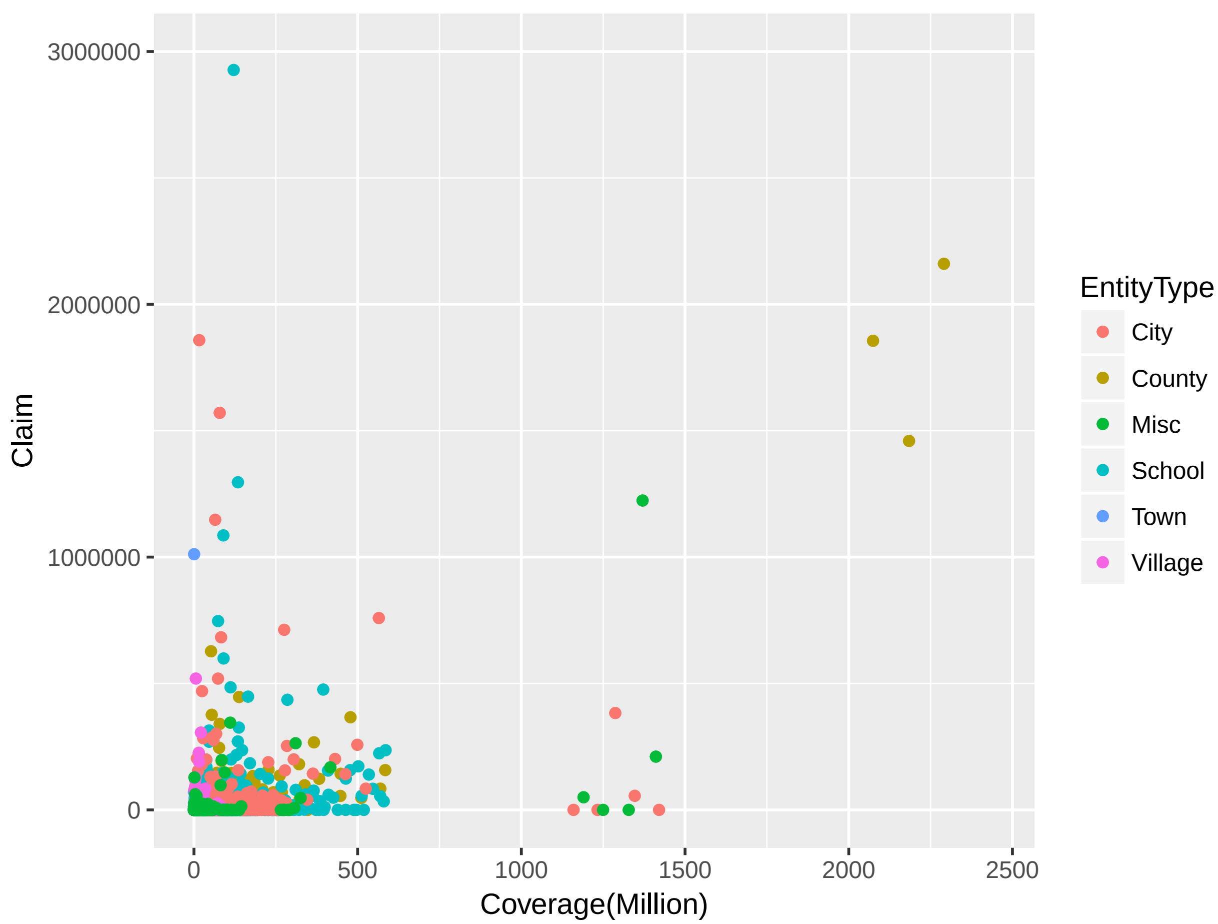


Figure 14.1: Scatter plot of *(Coverage,Claim)* from LGPIF data

point scale that ranges from 1 meaning dissatisfied to a 5 meaning satisfied. Ordinal variables provide a clear ordering of levels of a variable but the amount of separation between levels is unknown.

A *binary* variable is a special type of categorical variable where there are only two categories commonly taken to be a 0 and a 1. For example, we might code a variable in a dataset to be a 1 if an insured is female and a 0 if male.

14.1.2 Quantitative Variables

In this sub-section, you learn how to:

- Differentiate between continuous and discrete variable
- Use a combination of continuous and discrete variable
- Describe circular data

Unlike a qualitative variable, a quantitative variable is one in which numerical level is a realization from some scale so that the distance between any two levels of the scale takes on meaning. A *continuous variable* is one that can take on any value within a finite interval. For example, it is common to represent a policyholder's age, weight, or income, as a continuous variable. In contrast, a *discrete variable* is one that takes on only a finite number of values in any finite interval. Like an ordinal variable, these represent distinct categories that are ordered. Unlike an ordinal variable, the numerical difference between levels takes on economic meaning. A special type of discrete variable is a *count variable*, one with values on the nonnegative integers. For example, we will be particularly interested in the number of claims arising from a policy during a given period.

Some variables are inherently a *combination of discrete and continuous* components. For example, when we analyze the insured loss of a policyholder, we will encounter a discrete outcome at zero, representing no insured loss, and a continuous amount for positive outcomes, representing the amount of the insured loss. Another interesting variation is an *interval variable*, one that gives a range of possible outcomes.

Circular data represent an interesting category typically not analyzed by insurers. As an example of circular data, suppose that you monitor calls to your customer service center and would like to know when is the peak time of the day for calls to arrive. In this context, one can think about the time of the day as a variable with realizations on a circle, e.g., imagine an analog picture of a clock. For circular data, the distance between observations at 00:15 and 00:45 are just as close as observations 23:45 and 00:15 (here, we use the convention *HH:MM* means hours and minutes).

14.1.3 Multivariate Variables

In this sub-section, you learn how to:

- Differentiate between univariate and multivariate data
- Handle missing variables

Insurance data typically are *multivariate* in the sense that we can take many measurements on a single entity. For example, when studying losses associated with a firm's worker's compensation plan, we might want to know the location of its manufacturing plants, the industry in which it operates, the number of employees, and so forth. The usual strategy for analyzing multivariate data is to begin by examining each variable in isolation of the others. This is known as a *univariate* approach.

In contrast, for some variables, it makes little sense to only look at one dimensional aspects. For example, insurers typically organize *spatial* data by longitude and latitude to analyze the location of weather related insurance claims due hailstorms. Having only a single number, either longitude or latitude, provides little information in understanding geographical location.

Another special case of a multivariate variable, less obvious, involves coding for *missing data*. When data are missing, it is better to think about the variable as two dimensions, one to indicate whether or not the variable is reported and the second providing the age (if reported). In the same way, insurance data are commonly *censored* and *truncated*. We refer you to Chapter 4 for more on censored and truncated data. *Aggregate claims* can also be coded as another special type of multivariate variable. We refer you to Chapter 5 for more Aggregate claims.

Perhaps the most complicated type of multivariate variable is a *realization of a stochastic process*. You will recall that a stochastic process is little more than a collection of random variables. For example, in insurance, we might think about the times that claims arrive to an insurance company in a one year time horizon. This is a high dimensional variable that theoretically is infinite dimensional. Special techniques are required to understand realizations of stochastic processes that will not be addressed here.

14.2 Classic Measures of Scalar Associations

In this section, you learn how to:

- Estimate correlation using Pearson method
- Use rank based measures like Spearman, Kendall to estimate correlation
- Measure dependence using odds ratio,Pearson chi-square and likelihood ratio test statistic
- Use normal-based correlations to quantify associations involving ordinal variables

---

14.2.1 Association Measures for Quantitative Variables

For this section, consider a pair of random variables (X, Y) having joint distribution function $F(\cdot)$ and a random sample $(X_i, Y_i), i = 1, \ldots, n$. For the continuous case, suppose that $F(\cdot)$ is absolutely continuous with absolutely continuous marginals.

Pearson Correlation

Define the sample covariance function $Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$, where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y, respectively. Then, the product-moment (Pearson) correlation can be written as

$$r = \frac{Cov(X, Y)}{\sqrt{Cov(X, X)Cov(Y, Y)}}.$$

The correlation statistic r is widely used to capture association between random variables. It is a (nonparametric) estimator of the correlation parameter ρ, defined to be the covariance divided by the product of standard deviations. In this sense, it captures association for any pair of random variables.

This statistic has several important features. Unlike regression estimators, it is symmetric between random variables, so the correlation between X and Y equals the correlation between Y and X. It is unchanged by linear transformations of random variables (up to sign changes) so that we can multiply random variables or add constants as is helpful for interpretation. The range of the statistic is $[-1, 1]$ which does not depend on the distribution of either X or Y.

Further, in the case of independence, the correlation coefficient r is 0. However, it is well known that zero correlation does not imply independence, except for normally distributed random variables. The correlation statistic r is also a (maximum likelihood) estimator of the association parameter for bivariate normal distribution. So, for normally distributed data, the correlation statistic r can be used to assess independence. For additional interpretations of this well-known statistic, readers will enjoy (Lee Rodgers and Nicewander, 1998).

You can obtain the correlation statistic r using the `cor()` function in `R` and selecting the `pearson` method. This is demonstrated below by using the *Coverage* rating variable in millions of dollars and *Claim* amount variable in dollars from the LGPIF data introduced in chapter 1.

R Code for Pearson Correlation Statistic

```
### Pearson correlation between Claim and Coverage ###
r<-cor(Claim,Coverage, method = c("pearson"))
round(r,2)

Output:
[1] 0.31

### Pearson correlation between Claim and log(Coverage) ###
r<-cor(Claim,log(Coverage), method = c("pearson"))
round(r,2)

Output:
[1] 0.1
```

From `R` output above, $r = 0.31$, which indicates a positive association between *Claim* and *Coverage*. This means that as the coverage amount of a policy increases we expect claim to increase.

14.2.2 Rank Based Measures

Spearman's Rho

The Pearson correlation coefficient does have the drawback that it is not invariant to nonlinear transforms of the data. For example, the correlation between X and $\ln Y$ can be quite different from the correlation between X and Y. As we see from the `R` code for Pearson correlation statistic above, the correlation statistic r between *Coverage* rating variable in logarithmic millions of dollars and *Claim* amounts variable in dollars is 0.1 as compared to 0.31 when we calculate the correlation between *Coverage* rating variable in millions of dollars and *Claim* amounts variable in dollars. This limitation is one reason for considering alternative statistics.

Alternative measures of correlation are based on ranks of the data. Let $R(X_j)$ denote the rank of X_j from the sample $X_1, \ldots, X_n$ and similarly for $R(Y_j)$. Let $R(X) = (R(X_1), \ldots, R(X_n))'$ denote the vector of ranks, and similarly for $R(Y)$. For example, if $n = 3$ and $X = (24, 13, 109)$, then $R(X) = (2, 1, 3)$. A comprehensive introduction of rank statistics can be found in, for example, (Hettmansperger, 1984). Also, ranks can be used to obtain the empirical distribution function, refer to section 4.1.1 for more on the empirical distribution function.

With this, the correlation measure of (Spearman, 1904) is simply the product-moment correlation computed on the ranks:

$$r_S = \frac{Cov(R(X), R(Y))}{\sqrt{Cov(R(X), R(X))Cov(R(Y), R(Y))}} = \frac{Cov(R(X), R(Y))}{(n^2-1)/12}.$$

You can obtain the Spearman correlation statistic r_S using the `cor()` function in `R` and selecting the `spearman` method. From below, the Spearman correlation between the *Coverage* rating variable in millions of dollars and *Claim* amount variable in dollars is 0.41.

R Code for Spearman Correlation Statistic

```
### Spearman correlation between Claim and Coverage ###
rs<-cor(Claim,Coverage, method = c("spearman"))
round(rs,2)

Output:
[1] 0.41

### Spearman correlation between Claim and log(Coverage) ###
rs<-cor(Claim,log(Coverage), method = c("spearman"))
round(rs,2)

Output:
[1] 0.41
```

To show that the Spearman correlation statistic is invariate under strictly increasing transformations , from the `R` Code for Spearman correlation statistic above, $r_S = 0.41$ between the *Coverage* rating variable in logarithmic millions of dollars and *Claim* amount variable in dollars.

Kendall's Tau

An alternative measure that uses ranks is based on the concept of *concordance*. An observation pair (X, Y) is said to be concordant (discordant) if the observation with a larger value of X has also the larger (smaller) value of Y. Then $\Pr(concordance) = \Pr[(X_1-X_2)(Y_1-Y_2) > 0]$, $\Pr(discordance) = \Pr[(X_1-X_2)(Y_1-Y_2) < 0]$ and

$$\tau(X, Y) = \Pr(concordance) - \Pr(discordance) = 2\Pr(concordance) - 1 + \Pr(tie).$$

To estimate this, the pairs (X_i, Y_i) and (X_j, Y_j) are said to be concordant if the product $sgn(X_j-X_i)sgn(Y_j-Y_i)$ equals 1 and discordant if the product equals -1. Here, $sgn(x) = 1, 0, -1$ as $x > 0$, $x = 0$, $x < 0$, respectively. With this, we can express the association measure of (Kendall, 1938), known as *Kendall's tau*, as

$$\begin{array}{rl} \tau & = \frac{2}{n(n-1)} \sum_{i<j} sgn(X_j - X_i)sgn(Y_j - Y_i) \\ & = \frac{2}{n(n-1)} \sum_{i<j} sgn(R(X_j) - R(X_i))sgn(R(Y_j) - R(Y_i)) \end{array}.$$

Interestingly, (Hougaard, 2000), page 137, attributes the original discovery of this statistic to (Fechner, 1897), noting that Kendall's discovery was independent and more complete than the original work.

You can obtain the Kendall's tau, using the `cor()` function in `R` and selecting the `kendall` method. From below, $\tau = 0.32$ between the *Coverage* rating variable in millions of dollars and *Claim* amount variable in dollars.

R Code for Kendall's Tau

```
### Kendall's tau correlation between Claim and Coverage ###
tau<-cor(Claim,Coverage, method = c("kendall"))
round(tau,2)

Output:
[1]  0.32

### Kendall's tau correlation between Claim and log(Coverage) ###
tau<-cor(Claim,log(Coverage), method = c("kendall"))
round(tau,2)

Output:
[1] 0.32
```

Also,to show that the Kendall's tau is invariate under strictly increasing transformations , $\tau = 0.32$ between the *Coverage* rating variable in logarithmic millions of dollars and *Claim* amount variable in dollars.

14.2.3 Nominal Variables

Bernoulli Variables

To see why dependence measures for continuous variables may not be the best for discrete variables, let us focus on the case of Bernoulli variables that take on simple binary outcomes, 0 and 1. For notation, let $\pi_{jk} = \Pr(X = j, Y = k)$ for $j, k = 0, 1$ and let $\pi_X = \Pr(X = 1)$ and similarly for π_Y. Then, the population version of the product-moment (Pearson) correlation can be easily seen to be

$$\rho = \frac{\pi_{11} - \pi_X \pi_Y}{\sqrt{\pi_X(1-\pi_X)\pi_Y(1-\pi_Y)}}.$$

Unlike the case for continuous data, it is not possible for this measure to achieve the limiting boundaries of the interval $[-1, 1]$. To see this, students of probability may recall the Fréchet-Höeffding bounds for a joint distribution that turn out to be $\max\{0, \pi_X + \pi_Y - 1\} \leq \pi_{11} \leq \min\{\pi_X, \pi_Y\}$ for this joint probability. This limit on the joint probability imposes an additional restriction on the Pearson correlation. As an illustration, assume equal probabilities $\pi_X = \pi_Y = \pi > 1/2$. Then, the lower bound is

$$\frac{2\pi - 1 - \pi^2}{\pi(1-\pi)} = -\frac{1-\pi}{\pi}.$$

For example, if $\pi = 0.8$, then the smallest that the Pearson correlation could be is -0.25. More generally, there are bounds on ρ that depend on π_X and π_Y that make it difficult to interpret this measure.

As noted by (Bishop et al., 1975) (page 382), squaring this correlation coefficient yields the Pearson chi-square statistic. Despite the boundary problems described above, this feature makes the Pearson correlation coefficient a good choice for describing dependence with binary data. The other is the odds ratio, described as follows.

As an alternative measure for Bernoulli variables, the *odds ratio* is given by

$$OR(\pi_{11}) = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}} = \frac{\pi_{11}\left(1 + \pi_{11} - \pi_1 - \pi_2\right)}{(\pi_1 - \pi_{11})(\pi_2 - \pi_{11})}.$$

Pleasant calculations show that $OR(z)$ is 0 at the lower Fréchet-Höeffding bound $z = \max\{0, \pi_1 + \pi_2 - 1\}$ and is ∞ at the upper bound $z = \min\{\pi_1, \pi_2\}$. Thus, the bounds on this measure do not depend on the marginal probabilities π_X and π_Y, making it easier to interpret this measure.

As noted by (Yule, 1900), odds ratios are invariant to the labeling of 0 and 1. Further, they are invariant to the marginals in the sense that one can rescale π_1 and π_2 by positive constants and the odds ratio remains unchanged. Specifically, suppose that a_i, b_j are sets of positive constants and that

$$\pi_{ij}^{new} = a_i b_j \pi_{ij}$$

and $\sum_{ij} \pi_{ij}^{new} = 1$. Then,

$$OR^{new} = \frac{(a_1 b_1 \pi_{11})(a_0 b_0 \pi_{00})}{(a_0 b_1 \pi_{01})(a_1 b_0 \pi_{10})} = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}} = OR^{old}.$$

For additional help with interpretation, Yule proposed two transforms for the odds ratio, the first in (Yule, 1900),

$$\frac{OR - 1}{OR + 1},$$

and the second in (Yule, 1912),

$$\frac{\sqrt{OR} - 1}{\sqrt{OR} + 1}.$$

Although these statistics provide the same information as is the original odds ration OR, they have the advantage of taking values in the interval $[-1, 1]$, making them easier to interpret.

In a later section, we will also see that the marginal distributions have no effect on the Fréchet-Höeffding of the tetrachoric correlation, another measure of association, see also, (Joe, 2014), page 48.

	Fire5		
NoClaimCredit	0	1	Total
0	1611	2175	3786
1	897	956	1853
Total	2508	3131	5639

Table 14.2 : 2 × 2 table of counts for *Fire5* and *NoClaimCredit*

From Table 14.2, $OR(\pi_{11}) = \frac{1611(956)}{897(2175)} = 0.79$. You can obtain the $OR(\pi_{11})$, using the `oddsratio()` function from the `epitools` library in `R`. From the output below, $OR(\pi_{11}) = 0.79$ for the binary variables *NoClaim-Credit* and *Fier5* from the LGPIF data.

R Code for Odds Ratios

```
library(epitools)
oddsratio(NoClaimCredit, Fire5,method = c("wald"))$measure

Output:
[1]  0.79
```

Categorical Variables

More generally, let (X, Y) be a bivariate pair having $ncat_X$ and $ncat_Y$ numbers of categories, respectively. For a two-way table of counts, let n_{jk} be the number in the jth row, k column. Let $n_{j\cdot}$ be the row margin total and $n_{\cdot k}$ be the column margin total. Define Pearson chi-square statistic as

$$chi^2 = \sum_{jk} \frac{(n_{jk} - n_{j\cdot} n_{\cdot k}/n)^2}{n_{j\cdot} n_{\cdot k}/n}.$$

The likelihood ratio test statistic is

$$G^2 = 2 \sum_{jk} n_{jk} \ln \frac{n_{jk}}{n_{j\cdot} n_{\cdot k}/n}.$$

Under the assumption of independence, both chi^2 and G^2 have an asymptotic chi-square distribution with $(ncat_X - 1)(ncat_Y - 1)$ degrees of freedom.

To help see what these statistics are estimating, let $\pi_{jk} = \Pr(X = j, Y = k)$ and let $\pi_{X,j} = \Pr(X = j)$ and similarly for $\pi_{Y,k}$. Assuming that $n_{jk}/n \approx \pi_{jk}$ for large n and similarly for the marginal probabilities, we have

$$\frac{chi^2}{n} \approx \sum_{jk} \frac{(\pi_{jk} - \pi_{X,j}\pi_{Y,k})^2}{\pi_{X,j}\pi_{Y,k}}$$

and

$$\frac{G^2}{n} \approx 2 \sum_{jk} \pi_{jk} \ln \frac{\pi_{jk}}{\pi_{X,j}\pi_{Y,k}}.$$

Under the null hypothesis of independence, we have $\pi_{jk} = \pi_{X,j}\pi_{Y,k}$ and it is clear from these approximations that we anticipate that these statistics will be small under this hypothesis.

Classical approaches, as described in (Bishop et al., 1975) (page 374), distinguish between tests of independence and measures of associations. The former are designed to detect whether a relationship exists whereas the latter are meant to assess the type and extent of a relationship. We acknowledge these differing purposes but also less concerned with this distinction for actuarial applications.

	NoClaimCredit	
EntityType	0	1
City	644	149
County	310	18
Misc	336	273
School	1103	494
Town	492	479
Village	901	440

Table 14.3 : Two-way table of counts for *EntityType* and *NoClaimCredit*

You can obtain the Pearson chi-square statistic, using the `chisq.test()` function from the `MASS` library in `R`. Here, we test whether the *EntityType* variable is independent of *NoClaimCredit* variable using Table 14.3.

R Code for Pearson Chi-square Statistic

```
library(MASS)
table = table(EntityType, NoClaimCredit)
chisq.test(table)
```

```
Output:
-------------------------------------
 Test statistic   df     P value
---------------- ---- --------------
     344.2        5   3.15e-72 * * *
-------------------------------------

Table: Pearson's Chi-squared test
```

As the p-value is less than the .05 significance level, we reject the null hypothesis that the *EntityType* is independent of *NoClaimCredit*.

Furthermore, you can obtain the likelihood ratio test statistic , using the `likelihood.test()` function from the `Deducer` library in `R`. From below, we test whether the *EntityType* variable is independent of *NoClaimCredit* variable from the LGPIF data. Same conclusion is drawn as the Pearson chi-square test.

R Code for Likelihood Ratio Test Statistic

```
library(Deducer)
likelihood.test(EntityType, NoClaimCredit)
```

```
Output:
-----------------------------------------
 Test statistic   X-squared df   P value
---------------- -------------- ---------
     378.7             5         0 * * *
-----------------------------------------

Table: Log likelihood ratio (G-test) test of independence without correction
```

Ordinal Variables

As the analyst moves from the continuous to the nominal scale, there are two main sources of loss of information (Bishop et al., 1975) (page 343). The first is breaking the precise continuous measurements into groups. The second is losing the ordering of the groups. So, it is sensible to describe what we can do with variables that in discrete groups but where the ordering is known.

As described in Section 14.1.1, ordinal variables provide a clear ordering of levels of a variable but distances between levels are unknown. Associations have traditionally been quantified parametrically using normal-based correlations and nonparametrically using Spearman correlations with tied ranks.

Parametric Approach Using Normal Based Correlations

Refer to page 60, Section 2.12.7 of (Joe, 2014). Let (y_1, y_2) be a bivariate pair with discrete values on $m_1, \ldots, m_2$. For a two-way table of ordinal counts, let n_{st} be the number in the sth row, t column. Let $(n_{m_1 *}, \ldots, n_{m_2 *})$ be the row margin total and $(n_{*m_1}, \ldots, n_{*m_2})$ be the column margin total.

Let $\hat{\xi}_{1s} = \Phi^{-1}((n_{m_1} + \cdots + n_{s*})/n)$ for $s = m_1, \ldots, m_2$ be a cutpoint and similarly for $\hat{\xi}_{2t}$. The *polychoric* correlation, based on a two-step estimation procedure, is

$$\begin{aligned}\hat{\rho_N} &= \text{argmax}_\rho \textstyle\sum_{s=m_1}^{m_2} \sum_{t=m_1}^{m_2} n_{st} \log\Big\{ \Phi_2(\hat{\xi}_{1s}, \hat{\xi}_{2t}; \rho) - \Phi_2(\hat{\xi}_{1,s-1}, \hat{\xi}_{2t}; \rho) \\ &\qquad -\Phi_2(\hat{\xi}_{1s}, \hat{\xi}_{2,t-1}; \rho) + \Phi_2(\hat{\xi}_{1,s-1}, \hat{\xi}_{2,t-1}; \rho) \Big\}\end{aligned}$$

It is called a tetrachoric correlation for binary variables.

	NoClaimCredit	
AlarmCredit	0	1
1	1669	942
2	121	118
3	195	132
4	1801	661

Table 14.4 : Two-way table of counts for *AlarmCredit* and *NoClaimCredit*

You can obtain the polychoric or tetrachoric correlation using the `polychoric()` or `tetrachoric()` function from the `psych` library in `R`. The polychoric correlation is illustrated using Table 14.4. $\hat{\rho_N} = -0.14$, which means that there is a negative relationship between *AlarmCredit* and *NoClaimCredit*.

R Code for Polychoric Correlation

```
library(psych)
AlarmCredit<-as.numeric(ifelse(Insample$AC00==1,"1",
                    ifelse(Insample$AC05==1,"2",
                           ifelse(Insample$AC10==1,"3",
                                  ifelse(Insample$AC15==1,"4",0)))))
x <- table(AlarmCredit,NoClaimCredit)
rhoN<-polychoric(x,correct=FALSE)$rho
round(rhoN,2)

Output:
[1] -0.14
```

Interval Variables

As described in Section 14.1.2, interval variables provide a clear ordering of levels of a variable and the numerical distance between any two levels of the scale can be readily interpretable. For example, a claims count variable is an interval variable.

For measuring association, both the continuous variable and ordinal variable approaches make sense. The former takes advantage of knowledge of the ordering although assumes continuity. The latter does not rely on the continuity but also does not make use of the information given by the distance between scales.

For applications, one type is a count variable, a random variable on the discrete integers. Another is a mixture variable, on that has discrete and continuous components.

Discrete and Continuous Variables

The polyserial correlation is defined similarly, when one variable (y_1) is continuous and the other (y_2) ordinal. Define z to be the normal score of y_1. The polyserial correlation is

$$\hat{\rho_N} = \text{argmax}_\rho \sum_{i=1}^{n} \log\left\{ \phi(z_{i1}) \left[\Phi(\frac{\hat{\xi}_{2,y_{i2}} - \rho z_{i1}}{(1-\rho^2)^{1/2}}) - \Phi(\frac{\hat{\xi}_{2,y_{i2-1}} - \rho z_{i1}}{(1-\rho^2)^{1/2}}) \right] \right\}$$

The biserial correlation is defined similarly, when one variable is continuous and the other binary.

NoClaimCredit	Mean Claim	Total Claim
0	22,505	85,200,483
1	6,629	12,282,618

Table 14.5 : Summary of *Claim* by *NoClaimCredit*

You can obtain the polyserial or biserial correlation using the `polyserial()` or `biserial()` function from the `psych` library in `R`. Table 14.5 gives the summary of *Claim* by *NoClaimCredit* and the biserial correlation is illustrated using `R` code below. The $\hat{\rho_N} = -0.04$ which means that there is a negative correlation between *Claim* and *NoClaimCredit*.

R Code for Biserial Correlation

```
library(psych)
rhoN<-biserial(Claim,NoClaimCredit)
round(rhoN,2)

Output:
[1] -0.04
```

14.3 Introduction to Copulas

Copula functions are widely used in statistics and actuarial science literature for dependency modeling.

In this section, you learn how to:

- Describe a multivariate distribution function in terms of a copula function.

A *copula* is a multivariate distribution function with uniform marginals. Specifically, let $U_1, \ldots, U_p$ be p uniform random variables on $(0,1)$. Their distribution function

$$C(u_1, \ldots, u_p) = \Pr(U_1 \leq u_1, \ldots, U_p \leq u_p),$$

is a copula. We seek to use copulas in applications that are based on more than just uniformly distributed data. Thus, consider arbitrary marginal distribution functions $F_1(y_1)$,...,$F_p(y_p)$. Then, we can define a multivariate distribution function using the copula such that

$$F(y_1, \ldots, y_p) = C(F_1(y_1), \ldots, F_p(y_p)).$$

Here, F is a multivariate distribution function in this equation. Sklar (1959) showed that *any* multivariate distribution function F, can be written in the form of this equation, that is, using a copula representation.

Sklar also showed that, if the marginal distributions are continuous, then there is a unique copula representation. In this chapter we focus on copula modeling with continuous variables. For discrete case, readers can see (Joe, 2014) and (Genest and Nešlohva, 2007).

For bivariate case, $p = 2$, the distribution function of two random variables can be written by the bivariate copula function:

$$C(u_1, \, u_2) = \Pr(U_1 \leq u_1, \, U_2 \leq u_2),$$

$$F(y_1,\, y_2) = C(F_1(y_1),\, F_p(y_2)).$$

To give an example for bivariate copula, we can look at Frank's (1979) copula. The equation is

$$C(u_1,u_2) = \frac{1}{\theta}\ln\left(1+\frac{(\exp(\theta u_1)-1)(\exp(\theta u_2)-1)}{\exp(\theta)-1}\right).$$

This is a bivariate distribution function with its domain on the unit square $[0,1]^2$. Here θ is dependence parameter and the range of dependence is controlled by the parameter θ. Positive association increases as θ increases and this positive association can be summarized with Spearman's rho (ρ) and Kendall's tau (τ). Frank's copula is one of the commonly used copula functions in the copula literature. We will see other copula functions in Section 14.5.

14.4 Application Using Copulas

In this section, you learn how to:

- Discover dependence structure between random variables
- Model the dependence with a copula function

This section analyzes the insurance *losses* and *expenses* data with the statistical programming `R`. This data set was introduced in Frees and Valdez (1998) and is now readily available in the `copula` package. The model fitting process is started by marginal modeling of two variables (*loss* and *expense*). Then we model the joint distribution of these marginal outcomes.

14.4.1 Data Description

We start with getting a sample ($n = 1500$) from the whole data. We consider first two variables of the data; *losses* and *expenses*.

- *losses* : general liability claims from Insurance Services Office, Inc. (ISO)
- *expenses* : ALAE, specifically attributable to the settlement of individual claims (e.g. lawyer's fees, claims investigation expenses)

```
## Warning: package 'copula' was built under R version 3.3.3
```

To visualize the relationship between *losses* and *expenses* (ALAE), scatterplots in figure 14.2 are created on the real dollar scale and on the log scale.

R Code for Scatterplots

```
library(copula)
data(loss) # loss data
Lossdata <- loss
attach(Lossdata)
loss <- Lossdata$loss
par(mfrow=c(1, 2))
plot(loss,alae, cex=.5) # real dollar scale
plot(log(loss),log(alae),cex=.5) # log scale
par(mfrow=c(1, 2))
```

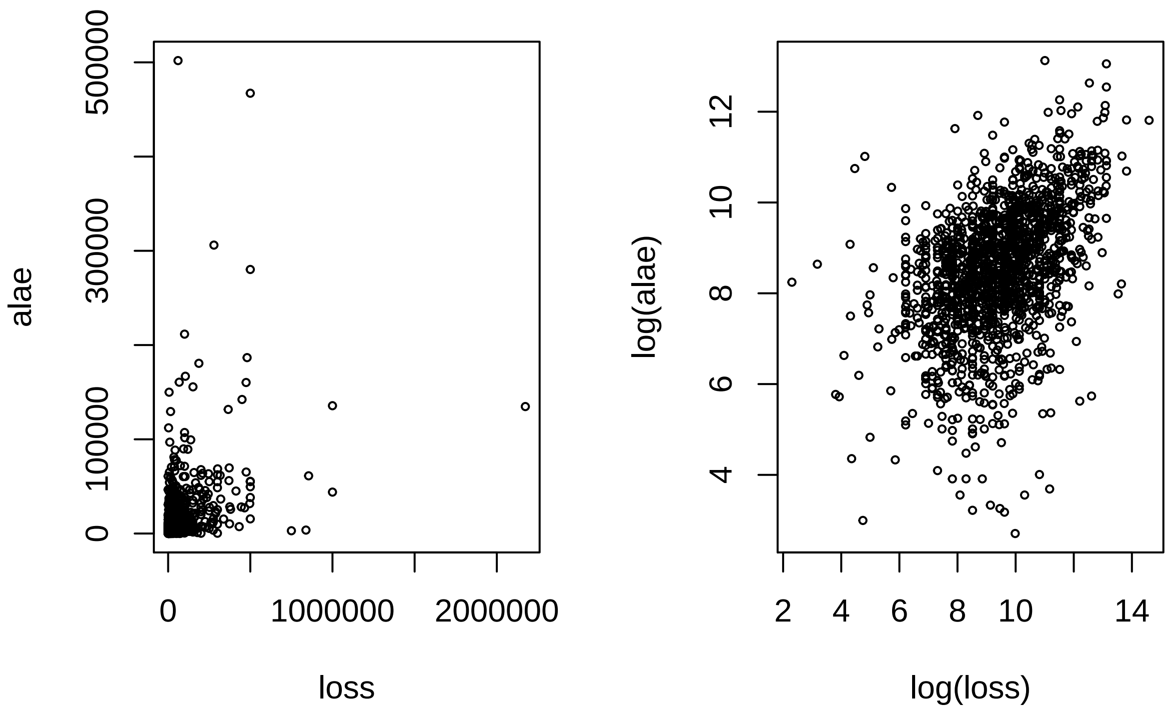


Figure 14.2: Scatter plot of Loss and ALAE

14.4.2 Marginal Models

We first examine the marginal distributions of *losses* and *expenses* before going through the joint modeling. The histograms show that both *losses* and *expenses* are right-skewed and fat-tailed.

For marginal distributions of losses and expenses, we consider a Pareto-type distribution, namely a Pareto type II with distribution function

$$F(y) = 1 - \left(1 + \frac{y}{\theta}\right)^{-\alpha},$$

where θ is the scale parameter and α is the shape parameter.

The marginal distributions of losses and expenses are fitted with maximum likelihood. Specifically, we use the *vglm* function from the `R VGAM` package. Firstly, we fit the marginal distribution of *expenses* .

R Code for Pareto Fitting

```
library(VGAM)

fit = vglm(alae ~ 1, paretoII(location=0, lscale="loge", lshape="loge")) # fit the model by vlgm function
coef(fit, matrix=TRUE) # extract fitted model coefficients, matrix=TRUE gives logarithm of estimated parame
Coef(fit)

Output:
               loge(scale) loge(shape)
 (Intercept)     9.624673   0.7988753

                  scale        shape
 (Intercept)  15133.603598     2.223039
```

We repeat this procedure to fit the marginal distribution of the *loss* variable. Because the loss data also seems right-skewed and heavy-tail data, we also model the marginal distribution with Pareto II distribution.

R Code for Pareto Fitting

```
fitloss = vglm(loss ~ 1, paretoII, trace=TRUE)
Coef(fit)
summary(fit)


Output:
       scale        shape
15133.603598     2.223039
```

To visualize the fitted distribution of *expenses* and *loss* variables, we use the estimated parameters and plot the corresponding distribution function and density function. For more details on marginal model selection, see Chapter 4.

14.4.3 Probability Integral Transformation

The *probability integral transformation* shows that any continuous variable can be mapped to a $U(0,1)$ random variable via its distribution function.

Given the fitted Pareto II distribution, the variable *expenses* is transformed to the variable u_1, which follows a uniform distribution on $[0,1]$:

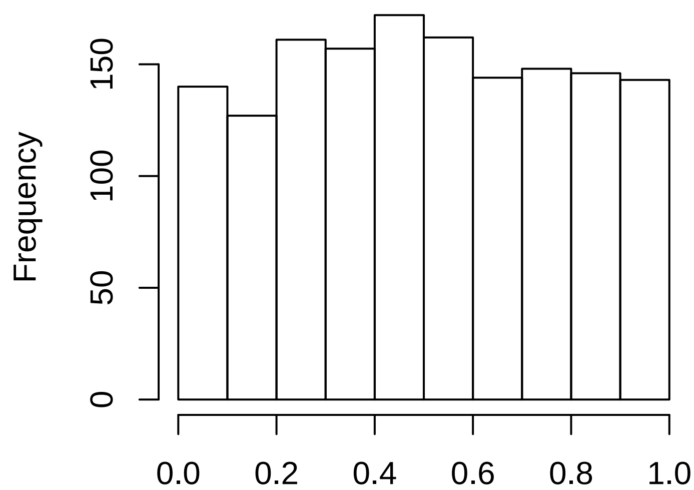


Figure 14.3: Histogram of Transformed Alae

$$u_1 = 1 - \left(1 + \frac{ALAE}{\hat{\theta}}\right)^{-\hat{\alpha}}.$$

After applying the probability integral transformation to *expenses* variable, we plot the histogram of *Transformed Alae* in Figure 14.3.

After fitting process,the variable *loss* is also transformed to the variable u_2, which follows a uniform distribution on $[0, 1]$. We plot the histogram of *Transformed Loss* . As an alternative, the variable *loss* is transformed to *normal scores* with the quantile function of standard normal distribution. As we see in Figure 14.4, normal scores of the variable *loss* are approximately marginally standard normal.

R Code for Histograms of Transformed Variables

```
u1 <- 1 - (1 + (alae/b))^(-s) # or u1 <- pparetoII(alae, location=0, scale=b, shape=s)
hist(u1, main = "", xlab = "Histogram of Transformed alae")

scaleloss <- Coef(fitloss)[1]
shapeloss <- Coef(fitloss)[2]
u2 <- 1 - (1 + (loss/scaleloss))^(-shapeloss)
par(mfrow = c(1, 2))
hist(u2, main = "", xlab = "Histogram of Transformed Loss")
hist(qnorm(u2), main = "", xlab = "Histogram of qnorm(Loss)")
```

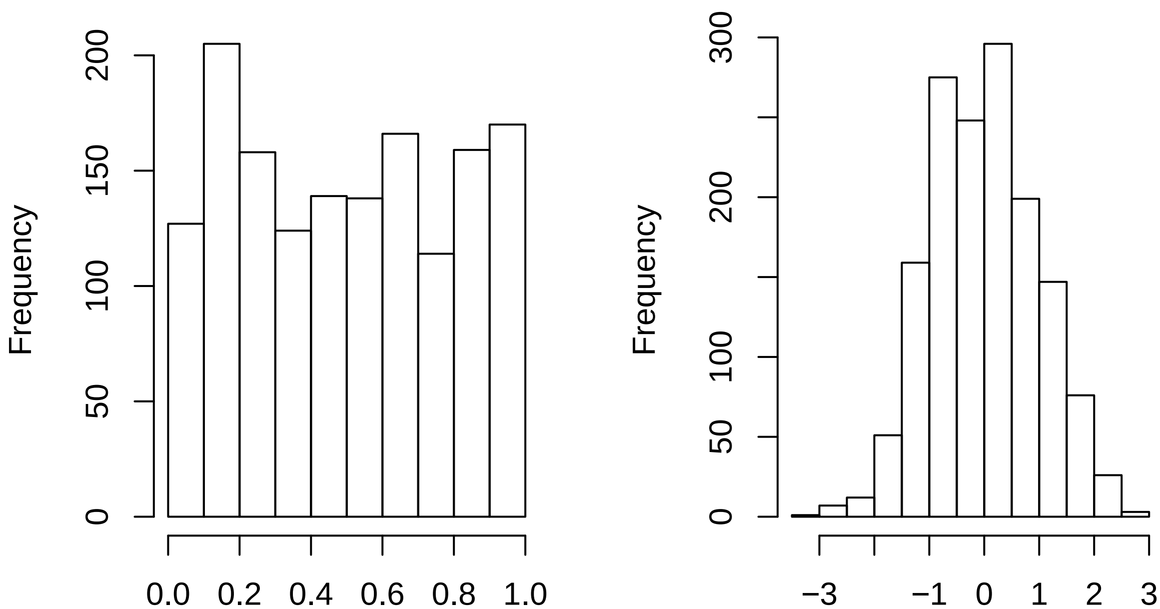


Figure 14.4: Histogram of Transformed Loss. The left-hand panel shows the distribution of probability integral transformed losses. The right-hand panel shows the distribution for the corresponding normal scores.

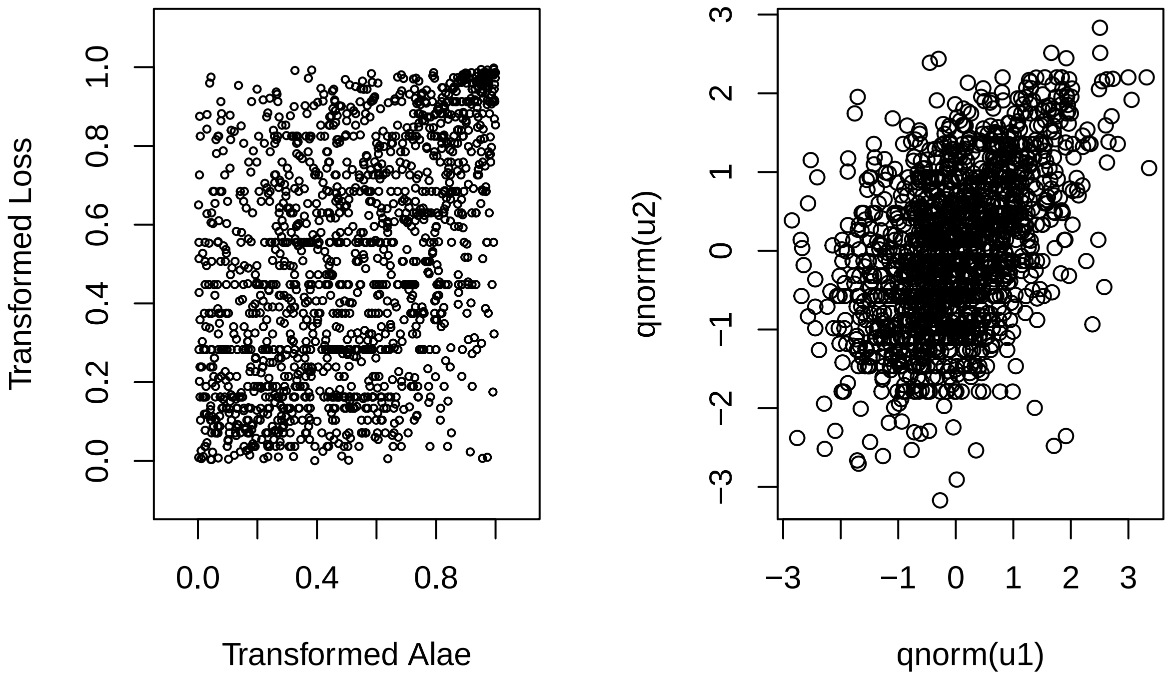


Figure 14.5: Left: Scatter plot for transformed variables. Right:Scatter plot for normal scores

14.4.4 Joint Modeling with Copula Function

Before jointly modeling losses and expenses, we draw the scatterplot of transformed variables (u_1, u_2) and the scatterplot of normal scores in Figure 14.5.

Then we calculate the Spearman's rho between these two uniform random variables.

R Code for Scatter Plots and Correlation

```
par(mfrow = c(1, 2))
plot(u1, u2, cex = 0.5, xlim = c(-0.1,1.1), ylim = c(-0.1,1.1),
     xlab = "Transformed Alae", ylab = "Transformed Loss")
plot(qnorm(u1), qnorm(u2))
cor(u1, u2, method = "spearman")

Output:

[1] 0.451872
```

Scatter plots and Spearman's rho correlation value (0.451) shows us there is a positive dependency between these two uniform random variables. It is more clear to see the relationship with normal scores in the second graph. To learn more details about normal scores and their applications in copula modeling, see (Joe, 2014).

(U_1, U_2), $(U_1 = F_1(ALAE)$ and $U_2 = F_2(LOSS))$, is fit to Frank's copula with maximum likelihood method.

R Code for Modeling with Frank Copula

```
uu = cbind(u1,u2)
frank.cop <- archmCopula("frank", param= c(5), dim = 2)
fit.ml <- fitCopula(frank.cop, uu, method="ml", start=c(0.4))
summary(fit.ml)


Output:

Call: fitCopula(copula, data = data, method = "ml", start = ..2)
Fit based on "maximum likelihood" and 1500 2-dimensional observations.
Copula: frankCopula
param
3.114
The maximized loglikelihood is 172.6
Convergence problems: code is 52 see ?optim.
Call: fitCopula(copula, data = data, method = "ml", start = ..2)
Fit based on "maximum likelihood" and 1500 2-dimensional observations.
Frank copula, dim. d = 2
      Estimate Std. Error
param    3.114         NA
The maximized loglikelihood is 172.6
Convergence problems: code is 52 see ?optim.
Number of loglikelihood evaluations:
function gradient
      45       45
```

The fitted model implies that losses and expenses are positively dependent and their dependence is significant.

We use the fitted parameter to update the Frank's copula. The Spearman's correlation corresponding to the fitted copula parameter(3.114) is calculated with the `rho` function. In this case, the Spearman's correlation coefficient is 0.462, which is very close to the sample Spearman's correlation coefficient; 0.452.

R Code for Spearman's Correlation Using Frank's Copula

```
(param = fit.ml@estimate)
frank.cop <- archmCopula("frank", param= param, dim = 2)
rho(frank.cop)

Output :
[1] 0.4622722
```

To visualize the fitted Frank's copula, the distribution function and density function perspective plots are drawn in Figure 14.6.

R Code for Frank's Copula Plots

```
par(mar=c(3.2,3,.2,.2),mfrow=c(1,2))
persp(frank.cop, pCopula, theta=50, zlab="C(u,v)",
        xlab ="u", ylab="v", cex.lab=1.3)
persp(frank.cop, dCopula, theta=0, zlab="c(u,v)",
        xlab ="u", ylab="v", cex.lab=1.3)
```

Frank's copula models positive dependence for this data set, with $\theta = 3.114$. For Frank's copula, the dependence is related to values of θ. That is:

- $\theta = 0$: independent copula
- $\theta > 0$: positive dependence
- $\theta < 0$: negative dependence

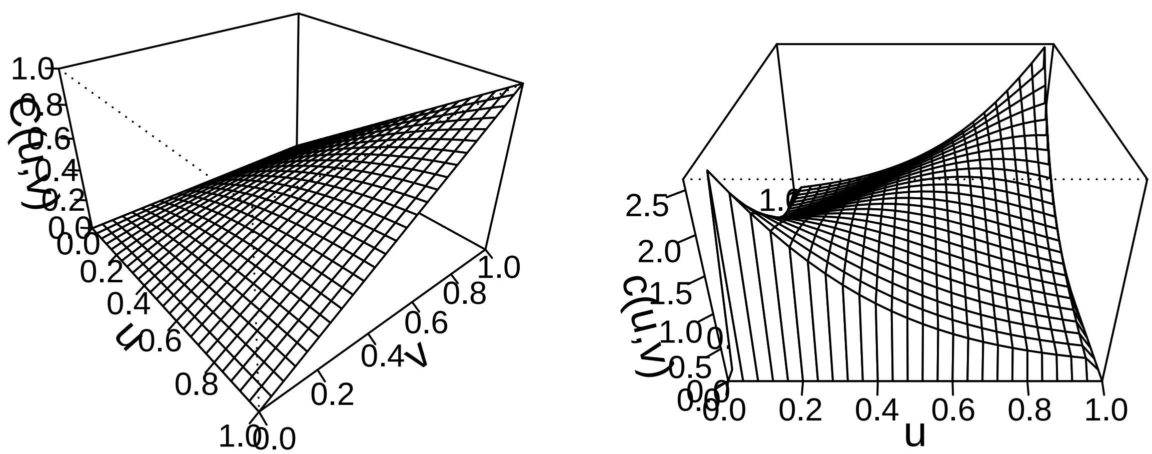


Figure 14.6: Left: Plot for distribution function for Franks Copula. Right:Plot for density function for Franks Copula

14.5 Types of Copulas

In this section, you learn how to:

- Define the basic families of the copula functions
- Calculate the association coefficients by the help of copula functions

---

There are several families of copulas have been described in the literature. Two main families of the copula families are the **Archimedian** and **Elliptical** copulas.

14.5.1 Elliptical Copulas

Elliptical copulas are constructed from elliptical distributions. This copula decompose (multivariate) elliptical distributions into their univariate elliptical marginal distributions by Sklar's theorem (Hofert et al., 2018).

Properties of elliptical copulas are typically obtained from the properties ofcorresponding elliptical distributions (Hofert et al., 2018).

For example, the normal distribution is a special type of elliptical distribution. To introduce the elliptical class of copulas, we start with the familiar multivariate normal distribution with probability density function

$$\phi_N(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}\mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z}\right).$$

Here, $\boldsymbol{\Sigma}$ is a correlation matrix, with ones on the diagonal. Let Φ and ϕ denote the standard normal distribution and density functions. We define the Gaussian (normal) copula density function as

$$c_N(u_1, \ldots, u_p) = \phi_N\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)\right) \prod_{j=1}^{p} \frac{1}{\phi(\Phi^{-1}(u_j))}.$$

As with other copulas, the domain is the unit cube $[0,1]^p$.

Specifically, a p-dimensional vector z has an *elliptical distribution* if the density can be written as

$$h_E(\mathbf{z}) = \frac{k_p}{\sqrt{\det \boldsymbol{\Sigma}}} g_p\left(\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right).$$

We will use elliptical distributions to generate copulas. Because copulas are concerned primarily with relationships, we may restrict our considerations to the case where $\mu = \mathbf{0}$ and $\boldsymbol{\Sigma}$ is a correlation matrix. With these restrictions, the marginal distributions of the multivariate elliptical copula are identical; we use H to refer to this marginal distribution function and h is the corresponding density. This marginal density is $h(z) = k_1 g_1(z^2/2)$.

We are now ready to define the *elliptical copula*, a function defined on the unit cube $[0,1]^p$ as

$$c_E(u_1, \ldots, u_p) = h_E\left(H^{-1}(u_1), \ldots, H^{-1}(u_p)\right) \prod_{j=1}^{p} \frac{1}{h(H^{-1}(u_j))}.$$

In the elliptical copula family, the function g_p is known as a *generator* in that it can be used to generate alternative distributions.

Distribution	*Generator* $g_p(x)$
Normal distribution	e^{-x}
t-distribution with r degrees of freedom	$(1+2x/r)^{-(p+r)/2}$
Cauchy	$(1+2x)^{-(p+1)/2}$
Logistic	$e^{-x}/(1+e^{-x})^2$
Exponential power	$\exp(-rx^s)$

Table 14.6 : Distribution and Generator Functions ($g_p(x)$) for Selected Elliptical Copulas

Most empirical work focuses on the normal copula and *t*-copula. That is, *t*-copulas are useful for modeling the dependency in the tails of bivariate distributions, especially in financial risk analysis applications.

The *t*-copulas with same association parameter in varying the degrees of freedom parameter show us different tail dependency structures. For more information on about *t*-copulas readers can see (Joe, 2014), (Hofert et al., 2018).

14.5.2 Archimedian Copulas

This class of copulas are constructed from a *generator* function,which is $g(\cdot)$ is a convex, decreasing function with domain [0,1] and range $[0, \infty)$ such that $g(0) = 0$. Use g^{-1} for the inverse function of g. Then the function

$$C_g(u_1, \ldots, u_p) = g^{-1}\left(g(u_1) + \cdots + g(u_p)\right)$$

is said to be an *Archimedean* copula. The function g is known as the *generator* of the copula C_g.

For bivariate case; $p = 2$, Archimedean copula function can be written by the function

$$C_g(u_1,\, u_2) = g^{-1}\left(g(u_1) + g(u_2)\right).$$

Some important special cases of Archimedean copulas are Frank copula, Clayton/Cook-Johnson copula, Gumbel/Hougaard copula. This copula classes are derived from different generator functions.

We can remember that we mentioned about Frank's copula with details in Section 14.3 and in Section 14.4. Here we will continue to express the equations for Clayton copula and Gumbel/Hougaard copula.

Clayton Copula

For $p = 2$, the Clayton copula is parameterized by $\theta \in [-1, \infty)$ is defined by

$$C_\theta^C(u) = \max\{u_1^{-\theta} + u_2^{-\theta} - 1, 0\}^{1/\theta}, \quad u \in [0,1]^2.$$

This is a bivariate distribution function of Clayton copula defined in unit square $[0,1]^2$. The range of dependence is controlled by the parameter θ as the same as Frank copula.

Gumbel-Hougaard copula

The Gumbel-Hougaarg copula is parametrized by $\theta \in [1, \infty)$ and defined by

$$C_\theta^{GH}(u) = \exp\left(-\left(\sum_{i=1}^{2}(-\log u_i)^\theta\right)^{1/\theta}\right), \quad u \in [0,1]^2.$$

Readers seeking deeper background on Archimedean copulas can see Joe (2014), Frees and Valdez (1998), and Genest and Mackay (1986).

14.5.3 Properties of Copulas

Bounds on Association

Like all multivariate distribution functions, copulas are bounded. The Fr$'e$chet-Hoeffding bounds are

$$\max(u_1 + \cdots + u_p + p - 1, 0) \leq \mathrm{C}(u_1, \ldots, u_p) \leq \min(u_1, \ldots, u_p).$$

To see the right-hand side of the equation, note that

$$\mathrm{C}(u_1, \ldots, u_p) = \Pr(U_1 \leq u_1, \ldots, U_p \leq u_p) \leq \Pr(U_j \leq u_j)$$

, for $j = 1, \ldots, p$. The bound is achieved when $U_1 = \cdots = U_p$. To see the left-hand side when $p = 2$, consider $U_2 = 1 - U_1$. In this case, if $1 - u_2 < u_1$ then $\Pr(U_1 \leq u_1, U_2 \leq u_2) = \Pr(1 - u_2 \leq U_1 < u_1) = u_1 + u_2 - 1$. (Nelson, 1997)

The product copula is $\mathrm{C}(u_1, u_2) = u_1 u_2$ is the result of assuming independence between random variables.

The lower bound is achieved when the two random variables are perfectly negatively related ($U_2 = 1 - U_1$) and the upper bound is achieved when they are perfectly positively related ($U_2 = U_1$).

We can see The Frechet-Hoeffding bounds for two random variables in the Figure 14.7.

R Code for Frechet-Hoeffding Bounds for Two Random Variables

```
library(copula)
n<-100
set.seed(1980)
U<-runif(n)
par(mfrow=c(1, 2))
plot(cbind(U,1-U), xlab=quote(U[1]), ylab=quote(U[2]),main="Perfect Negative Dependency") # W for p=2
plot (cbind(U,U), xlab=quote(U[1]),ylab=quote(U[2]),main="Perfect Positive Dependency")  #M for p=2
```

Measures of Association

Schweizer and Wolff (1981) established that the copula accounts for all the dependence between two random variables, Y_1 and Y_2, in the following sense. Consider m_1 and m_2, strictly increasing functions. Thus, the manner in which Y_1 and Y_2 "move together" is captured by the copula, regardless of the scale in which each variable is measured.

Schweizer and Wolff also showed the two standard nonparametric measures of association could be expressed solely in terms of the copula function. Spearman's correlation coefficient is given by

$$= 12 \int \int \{\mathrm{C}(u, v) - uv\} \, du dv.$$

Kendall's tau is given by

$$= 4 \int \int \mathrm{C}(u, v) d\mathrm{C}(u, v) - 1.$$

For these expressions, we assume that Y_1 and Y_2 have a jointly continuous distribution function. Further, the definition of Kendall's tau uses an independent copy of (Y_1, Y_2), labeled (Y_1^*, Y_2^*), to define the measure of

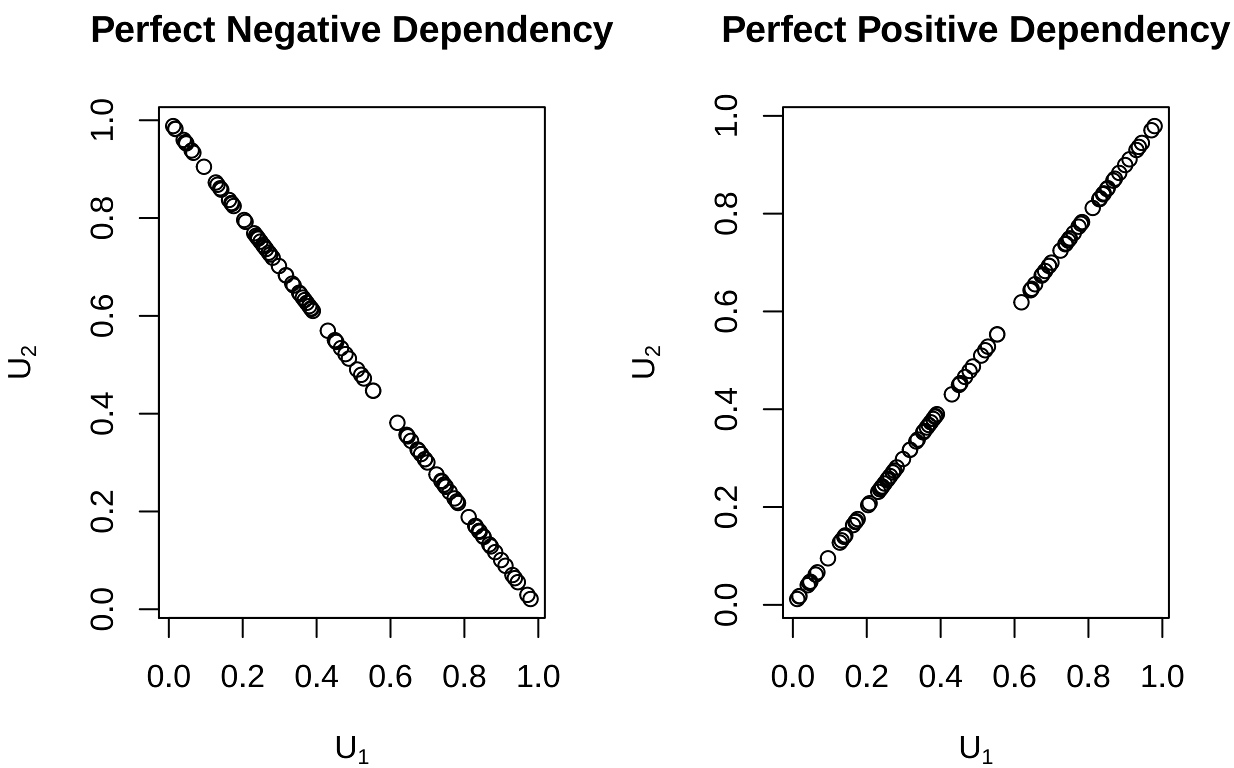


Figure 14.7: Perfect Positive and Perfect negative dependence plots

"concordance." the widely used Pearson correlation depends on the margins as well as the copula. Because it is affected by non-linear changes of scale.

Tail Dependency

There are some applications in which it is useful to distinguish by the part of the distribution in which the association is strongest. For example, in insurance it is helpful to understand association among the largest losses, that is, association in the right tails of the data.

To capture this type of dependency, we use the right-tail concentration function. The function is

$$R(z) = \frac{\Pr(U_1 > z, U_2 > z)}{1-z} = \Pr(U_1 > z | U_2 > z) = \frac{1 - 2z + \mathrm{C}(z,z)}{1-z}.$$

From this equation , $R(z)$ will equal to z under independence. Joe (1997) uses the term "upper tail dependence parameter" for $R = \lim_{z \rightarrow 1} R(z)$. Similarly, the left-tail concentration function is

$$L(z) = \frac{\Pr(U_1 \le z, U_2 \le z)}{z} = \Pr(U_1 \le z | U_2 \le z) = \frac{\mathrm{C}(z,z)}{1-z}.$$

Tail dependency concentration function captures the probability of two random variables both catching up extreme values.

We calculate the left and right tail concentration functions for four different types of copulas; Normal, Frank,Gumbel and t copula. After getting tail concentration functions for each copula, we show concentration function's values for these four copulas in Table 14.7. As in Venter (2002), we show $L(z)$ for $z \le 0.5$ and $R(z)$ for $z > 0.5$ in the tail dependence plot in Figure 14.8. We interpret the tail dependence plot, to mean that both the Frank and Normal copula exhibit no tail dependence whereas the t and the Gumbel may do so. The t copula is symmetric in its treatment of upper and lower tails.

Copula	Lower	Upper
Frank	0	0
Gumbel	0	0.74
Normal	0	0
t	0.10	0.10

Table 14.7 : Tail concentration function values for different copulas

R Code for Tail Copula Functions for Different Copulas

```
library(copula)
U1 = seq(0,0.5, by=0.002)
U2 = seq(0.5,1, by=0.002)
U = rbind(U1, U2)
TailFunction <- function(Tailcop) {
  lowertail <- pCopula(cbind(U1,U1), Tailcop)/U1
  uppertail <- (1-2*U2 +pCopula(cbind(U2,U2), Tailcop))/(1-U2)
  jointtail <- rbind(lowertail,uppertail)
}
Tailcop1 <- archmCopula(family = "frank", param= c(0.05), dim = 2)
Tailcop2 <- archmCopula(family = "gumbel",param = 3)
Tailcop3 <- ellipCopula("normal", param = c(0.25),dim = 2, dispstr = "un")
Tailcop4 <- ellipCopula("t", param = c(0.25),dim = 2, dispstr = "un", df=5)
jointtail1 <- TailFunction(Tailcop1)
```

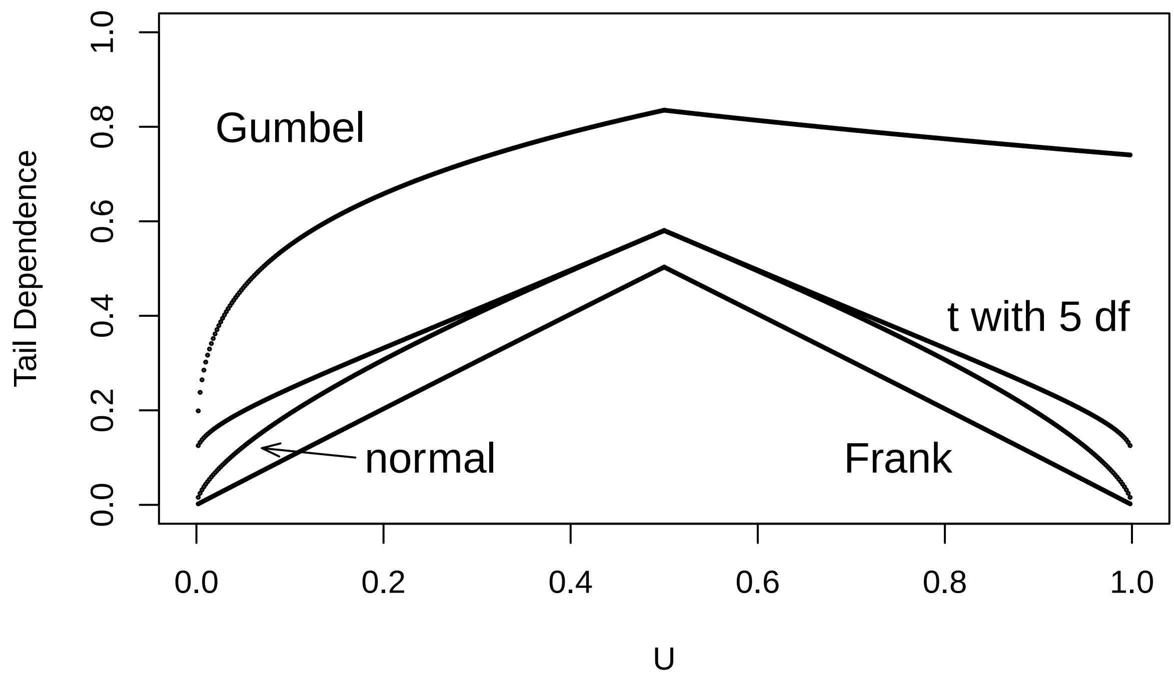


Figure 14.8: Tail dependence plots

```
jointtail2 <- TailFunction(Tailcop2)
jointtail3 <- TailFunction(Tailcop3)
jointtail4 <- TailFunction(Tailcop4)
tailIndex(Tailcop1)
tailIndex(Tailcop2)
tailIndex(Tailcop3)
tailIndex(Tailcop4)
```

R Code for Tail Dependence Plots for Different Copulas

```
plot(U,jointtail1, cex=.2, xlim=c(0,1),ylab="Tail Dependence", ylim=c(0,1))
lines(U,jointtail2, type="p",lty=1, cex=.2)
lines(U,jointtail3, type="p",lty=1, cex=.2)
lines(U,jointtail4, type="p",lty=1, cex=.2)
text(0.75, 0.1, "Frank", cex=1.3)        #1
text(0.1, 0.8, "Gumbel", cex=1.3)        #2
text(0.25, 0.1, "normal", cex=1.3)       #3
arrows(.17, 0.1, .07, 0.12,code=2, angle=20, length=0.1)
text(0.9, 0.4, "t with 5 df", cex=1.3)   #4
```

14.6 Why is Dependence Modeling Important?

Dependence Modeling is important because it enables us to understand the dependence structure by defining the relationship between variables in a dataset. In insurance, ignoring dependence modeling may not impact pricing but could lead to misestimation of required capital to cover losses. For instance, from Section 14.4 , it is seen that there was a positive relationship between *Loss* and *Expense*. This means that, if there is a large loss then we expect expenses to be large as well and ignoring this relationship could lead to misestimation of reserves.

To illustrate the importance of dependence modeling, we refer you back to Portfolio Management example in Chapter 6 that assumed that the property and liability risks are independent. Here, we incorporate dependence by allowing the 4 lines of business to depend on one another through a Gaussian copula. In Table 14.8, we show that dependence affects the portfolio quantiles (VaR_q), although not the expect value. For instance , the $VaR_{0.99}$ for total risk which is the amount of capital required to ensure, with a 99% degree of certainty that the firm does not become technically insolvent is higher when we incorporate dependence. This leads to less capital being allocated when dependence is ignored and can cause unexpected solvency problems.

Independent	Expected Value	$VaR_{0.9}$	$VaR_{0.95}$	$VaR_{0.99}$
Retained	269	300	300	300
Insurer	2,274	4,400	6,173	11,859
Total	2,543	4,675	6,464	12,159
Gaussian Copula	**Expected Value**	$VaR_{0.9}$	$VaR_{0.95}$	$VaR_{0.99}$
Retained	269	300	300	300
Insurer	2,340	4,988	7,339	14,905
Total	2,609	5,288	7,639	15,205

Table 14.8 : Results for portfolio expected value and quantiles (VaR_q)

R Code for Simulation Using Gaussian Copula

```
# For the gamma distributions, use
alpha1 <- 2;      theta1 <- 100
alpha2 <- 2;      theta2 <- 200
# For the Pareto distributions, use
alpha3 <- 2;      theta3 <- 1000
alpha4 <- 3;      theta4 <- 2000
# Deductibles
d1     <- 100
d2     <- 200


# Simulate the risks
nSim <- 10000  #number of simulations
set.seed(2017) #set seed to reproduce work
X1 <- rgamma(nSim,alpha1,scale = theta1)
X2 <- rgamma(nSim,alpha2,scale = theta2)
# For the Pareto Distribution, use
library(VGAM)
X3 <- rparetoII(nSim,scale=theta3,shape=alpha3)
X4 <- rparetoII(nSim,scale=theta4,shape=alpha4)
# Portfolio Risks
S         <- X1 + X2 + X3 + X4
```

```
Sretained <- pmin(X1,d1) + pmin(X2,d2)
Sinsurer  <- S - Sretained

# Expected Claim Amounts
ExpVec <- t(as.matrix(c(mean(Sretained),mean(Sinsurer),mean(S))))
colnames(ExpVec) <- c("Retained", "Insurer","Total")
round(ExpVec,digits=2)

# Quantiles
quantMat <- rbind(
  quantile(Sretained, probs=c(0.80, 0.90, 0.95, 0.99)),
  quantile(Sinsurer,  probs=c(0.80, 0.90, 0.95, 0.99)),
  quantile(S        ,  probs=c(0.80, 0.90, 0.95, 0.99)))
rownames(quantMat) <- c("Retained", "Insurer","Total")
round(quantMat,digits=2)

plot(density(S), main="Density of Total Portfolio Risk S", xlab="S")

### Normal Copula ##
library(VGAM)
library(copula)
library(GB2)
library(statmod)
library(numDeriv)
set.seed(2017)
parm<-c(0.5,0.5,0.5,0.5,0.5,0.5)
nc <- normalCopula(parm, dim = 4, dispstr = "un")
mcc <- mvdc(nc, margins = c("gamma", "gamma","paretoII","paretoII"),
            paramMargins = list(list(scale = theta1, shape=alpha1),
                                list(scale = theta2, shape=alpha2),
                                list(scale = theta3, shape=alpha3),
                                list(scale = theta4, shape=alpha4)))
X <- rMvdc(nSim, mvdc = mcc)

X1<-X[,1]
X2<-X[,2]
X3<-X[,3]
X4<-X[,4]

# Portfolio Risks
S         <- X1 + X2 + X3 + X4
Sretained <- pmin(X1,d1) + pmin(X2,d2)
Sinsurer  <- S - Sretained

# Expected Claim Amounts
ExpVec <- t(as.matrix(c(mean(Sretained),mean(Sinsurer),mean(S))))
colnames(ExpVec) <- c("Retained", "Insurer","Total")
round(ExpVec,digits=2)

# Quantiles
quantMat <- rbind(
  quantile(Sretained, probs=c(0.80, 0.90, 0.95, 0.99)),
  quantile(Sinsurer,  probs=c(0.80, 0.90, 0.95, 0.99)),
  quantile(S        ,  probs=c(0.80, 0.90, 0.95, 0.99)))
```

```
rownames(quantMat) <- c("Retained", "Insurer","Total")
round(quantMat,digits=2)

plot(density(S), main="Density of Total Portfolio Risk S", xlab="S")
```

14.7 Further Resources and Contributors

Contributors

- **Edward W. (Jed) Frees** and **Nii-Armah Okine**, University of Wisconsin-Madison, and **Emine Selin Sarıdaş**, Mimar Sinan University, are the principal authors of the initial version of this chapter. Email: jfrees@bus.wisc.edu for chapter comments and suggested improvements.

Technical Supplement A. Other Classic Measures of Scalar Associations

A.1. Blomqvist's Beta

Blomqvist (1950) developed a measure of dependence now known as *Blomqvist's beta*, also called the *median concordance coefficient* and the *medial correlation coefficient*. Using distribution functions, this parameter can be expressed as

$$\beta = 4F\left(F_X^{-1}(1/2), F_Y^{-1}(1/2)\right) - 1.$$

That is, first evaluate each marginal at its median ($F_X^{-1}(1/2)$ and $F_Y^{-1}(1/2)$, respectively). Then, evaluate the bivariate distribution function at the two medians. After rescaling (multiplying by 4 and subtracting 1), the coefficient turns out to have a range of $[-1, 1]$, where 0 occurs under independence.

Like Spearman's rho and Kendall's tau, an estimator based on ranks is easy to provide. First write $\beta = 4C(1/2, 1/2) - 1 = 2\Pr((U_1 - 1/2)(U_2 - 1/2)) - 1$ where U_1, U_2 are uniform random variables. Then, define

$$\hat{\beta} = \frac{2}{n}\sum_{i=1}^{n} I\left((R(X_i) - \frac{n+1}{2})(R(Y_i) - \frac{n+1}{2}) \geq 0\right) - 1.$$

See, for example, (Joe, 2014), page 57 or (Hougaard, 2000), page 135, for more details.

Because Blomqvist's parameter is based on the center of the distribution, it is particularly useful when data are censored; in this case, information in extreme parts of the distribution are not always reliable. How does this affect a choice of association measures? First, recall that association measures are based on a bivariate distribution function. So, if one has knowledge of a good approximation of the distribution function, then calculation of an association measure is straightforward in principle. Second, for censored data, bivariate extensions of the univariate Kaplan-Meier distribution function estimator are available. For example, the version introduced in (Dabrowska, 1988) is appealing. However, because of instances when large masses of data appear at the upper range of the data, this and other estimators of the bivariate distribution function are unreliable. This means that, summary measures of the estimated distribution function based on Spearman's rho or Kendall's tau can be unreliable. For this situation, Blomqvist's beta appears to be a better choice as it focuses on the center of the distribution. (Hougaard, 2000), Chapter 14, provides additional discussion.

You can obtain the Blomqvist's beta, using the `betan()` function from the `copula` library in `R`. From below, $\beta = 0.3$ between the *Coverage* rating variable in millions of dollars and *Claim* amount variable in dollars.

R Code for Blomqvist's Beta

```
### Blomqvist's beta correlation between Claim and Coverage ###
library(copula)
n<-length(Claim)
U<-cbind(((n+1)/n*pobs(Claim)),((n+1)/n*pobs(Coverage)))
beta<-betan(U, scaling=FALSE)
round(beta,2)

Output:
[1]  0.3

### Blomqvist's beta correlation between Claim and log(Coverage) ###
n<-length(Claim)
Fx<-cbind(((n+1)/n*pobs(Claim)),((n+1)/n*pobs(log(Coverage))))
beta<-betan(Fx, scaling=FALSE)
round(beta,2)

Output:
[1]  0.3
```

In addition,to show that the Blomqvist's beta is invariate under strictly increasing transformations , $\beta = 0.3$ between the *Coverage* rating variable in logarithmic millions of dollars and *Claim* amount variable in dollars.

A.2. Nonparametric Approach Using Spearman Correlation with Tied Ranks

For the first variable, the average rank of observations in the sth row is

$$r_{1s} = n_{m_1*} + \cdots + n_{s-1,*} + \frac{1}{2}\left(1 + n_{s*}\right)$$

and similarly $r_{2t} = \frac{1}{2}\left[(n_{*m_1} + \cdots + n_{*,s-1} + 1) + (n_{*m_1} + \ldots + n_{*s})\right]$. With this, we have Spearman's rho with tied rank is

$$\hat{\rho}_S = \frac{\sum_{s=m_1}^{m_2}\sum_{t=m_1}^{m_2} n_{st}(r_{1s} - \bar{r})(r_{2t} - \bar{r})}{\left[\sum_{s=m_1}^{m_2} n_{s*}(r_{1s} - \bar{r})^2 \sum_{t=m_1}^{m_2} n_{*t}(r_{2t} - \bar{r})^2\right]^2}$$

where the average rank is $\bar{r} = (n+1)/2$.

Click to Show Proof for Special Case: Binary Data.

Special Case: Binary Data. Here, $m_1 = 0$ and $m_2 = 1$. For the first variable ranks, we have $r_{10} = (1+n_{0+})/2$ and $r_{11} = (n_{0+} + 1 + n)/2$. Thus, $r_{10} - \bar{r} = (n_{0+} - n)/2$ and $r_{11} - \bar{r} = n_{0+}/2$. This means that we have $\sum_{s=0}^{1} n_{s+}(r_{1s} - \bar{r})^2 = n(n - n_{0+})n_{0+}/4$ and similarly for the second variable. For the numerator, we have

$$
\begin{aligned}
\sum_{s=0}^{1}\sum_{t=0}^{1} & n_{st}(r_{1s}-\bar{r})(r_{2t}-\bar{r}) \\
&= n_{00}\frac{n_{0+}-n}{2}\frac{n_{+0}-n}{2}+n_{01}\frac{n_{0+}-n}{2}\frac{n_{+0}}{2}+n_{10}\frac{n_{0+}}{2}\frac{n_{+0}-n}{2}+n_{11}\frac{n_{0+}}{2}\frac{n_{+0}}{2} \\
&= \frac{1}{4}(n_{00}(n_{0+}-n)(n_{+0}-n)+(n_{0+}-n_{00})(n_{0+}-n)n_{+0} \\
&\quad +(n_{+0}-n_{00})n_{0+}(n_{+0}-n)+(n-n_{+0}-n_{0+}+n_{00})n_{0+}n_{+0}) \\
&= \frac{1}{4}(n_{00}n^2-n_{0+}(n_{0+}-n)n_{+0} \\
&\quad +n_{+0}n_{0+}(n_{+0}-n)+(n-n_{+0}-n_{0+})n_{0+}n_{+0}) \\
&= \frac{1}{4}(n_{00}n^2-n_{0+}n_{+0}(n_{0+}-n+n_{+0}-n+n-n_{+0}-n_{0+}) \\
&= \frac{n}{4}(nn_{00}-n_{0+}n_{+0}).
\end{aligned}
$$

This yields

$$
\begin{aligned}
\hat{\rho}_S &= \frac{n(nn_{00}-n_{0+}n_{+0})}{4\sqrt{(n(n-n_{0+})n_{0+}/4)(n(n-n_{+0})n_{+0}/4)}} \\
&= \frac{nn_{00}-n_{0+}n_{+0}}{\sqrt{n_{0+}n_{+0}(n-n_{0+})(n-n_{+0})}} \\
&= \frac{n_{00}-n(1-\hat{\pi}_X)(1-\hat{\pi}_Y)}{\sqrt{\hat{\pi}_X(1-\hat{\pi}_X)\hat{\pi}_Y(1-\hat{\pi}_Y)}}
\end{aligned}
$$

where $\hat{\pi}_X=(n-n_{0+})/n$ and similarly for $\hat{\pi}_Y$. Note that this is same form as the Pearson measure. From this, we see that the joint count n_{00} drives this association measure.

You can obtain the ties-corrected Spearman correlation statistic r_S using the `cor()` function in `R` and selecting the `spearman` method. From below $\hat{\rho}_S=-0.09$

R Code for Ties-corrected Spearman Correlation

```
rs_ties<-cor(AlarmCredit,NoClaimCredit, method = c("spearman"))
round(rs_ties,2)

Output:
[1] -0.09
```

Chapter 15

Appendix A: Review of Statistical Inference

Chapter preview. The appendix gives an overview of concepts and methods related to statistical inference on the population of interest, using a random sample of observations from the population. In the appendix, Section 15.1 introduces the basic concepts related to the population and the sample used for making the inference. Section 15.2 presents the commonly used methods for point estimation of population characteristics. Section 15.3 demonstrates interval estimation that takes into consideration the uncertainty in the estimation, due to use of a random sample from the population. Section 15.4 introduces the concept of hypothesis testing for the purpose of variable and model selection.

15.1 Basic Concepts

In this section, you learn the following concepts related to statistical inference.

- Random sampling from a population that can be summarized using a list of items or individuals within the population
- Sampling distributions that characterize the distributions of possible outcomes for a statistic calculated from a random sample
- The central limit theorem that guides the distribution of the mean of a random sample from the population

Statistical inference is the process of making conclusions on the characteristics of a large set of items/individuals (i.e., the **population**), using a representative set of data (e.g., a **random sample**) from a list of items or individuals from the population that can be sampled. While the process has a broad spectrum of applications in various areas including science, engineering, health, social, and economic fields, statistical inference is important to insurance companies that use data from their existing policy holders in order to make inference on the characteristics (e.g., risk profiles) of a specific segment of target customers (i.e., the population) whom the insurance companies do not directly observe.

Show An Empirical Example Using the Wisconsin Property Fund

Example – Wisconsin Property Fund. Assume there are 1,377 *individual* claims from the 2010 experience.

	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Standard Deviation
Claims	1	788	2,250	26,620	6,171	12,920,000	368,030
Logarithmic Claims	0	6.670	7.719	7.804	8.728	16.370	1.683

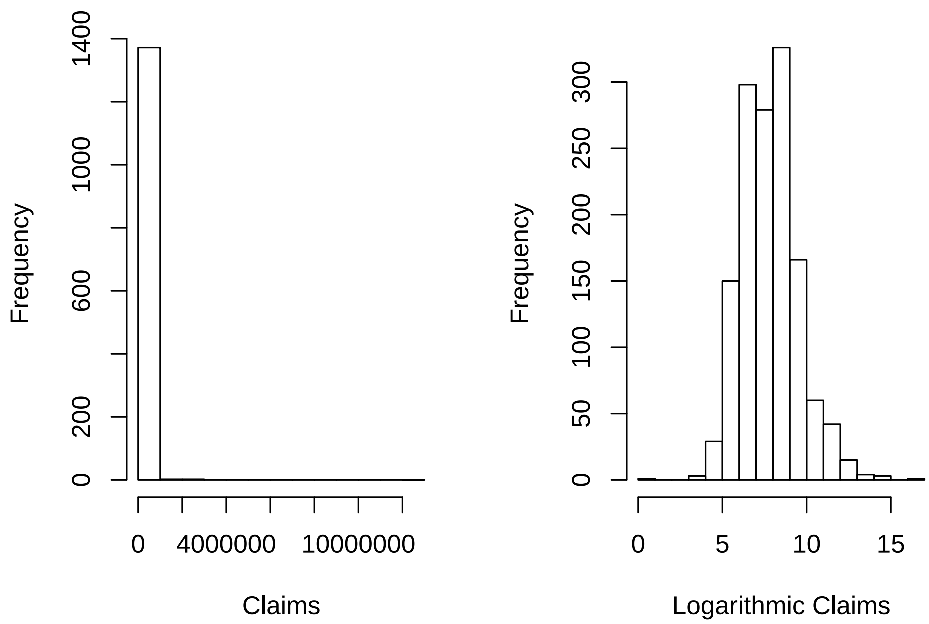


Figure 15.1: Distribution of Claims

```
ClaimLev <- read.csv("Data/CLAIMLEVEL.csv", header=TRUE)
ClaimLevBC10<-subset(ClaimLev,Year==2010);
cat("Sample size: ", nrow(ClaimLevBC10), "\n")
par(mfrow=c(1, 2))
hist(ClaimLevBC10$Claim, main="", xlab="Claims")
hist(log(ClaimLevBC10$Claim), main="", xlab="Logarithmic Claims")
```

```
## Sample size:  1377
```

Using the 2010 claim experience (the sample), the Wisconsin Property Fund may be interested in assessing the severity of all claims that could potentially occur, such as 2010, 2011, and so forth (the population). This process is important in the contexts of ratemaking or claim predictive modeling. In order for such inference to be valid, we need to assume that

- the set of 2010 claims is a *random sample* that is representative of the population,
- the *sampling distribution* of the average claim amount can be estimated, so that we can quantify the bias and uncertainty in the esitmation due to use of a finite sample.

15.1.1 Random Sampling

In statistics, a sampling **error** occurs when the **sampling frame**, the list from which the sample is drawn, is not an adequate approximation of the population of interest. A sample must be a representative subset of a population, or universe, of interest. If the sample is not representative, taking a larger sample does not eliminate bias, as the same mistake is repeated over again and again. Thus, we introduce the concept for random sampling that gives rise to a simple **random sample** that is representative of the population.

We assume that the random variable X represents a draw from a population with a distribution function

$F(\cdot)$ with mean $\mathrm{E}[X] = \mu$ and variance $\mathrm{Var}[X] = \mathrm{E}[(X-\mu)^2]$, where $E(\cdot)$ denotes the expectation of a random variable. In **random sampling**, we make a total of n such draws represented by $X_1, \ldots, X_n$, each unrelated to one another (i.e., *statistically independent*). We refer to $X_1, \ldots, X_n$ as a **random sample** (*with replacement*) from $F(\cdot)$, taking either a parametric or nonparametric form. Alternatively, we may say that $X_1, \ldots, X_n$ are identically and independently distributed (*iid*) with distribution function $F(\cdot)$.

15.1.2 Sampling Distribution

Using the random sample $X_1, \ldots, X_n$, we are interested in making a conclusion on a specific attribute of the population distribution $F(\cdot)$. For example, we may be interested in making an inference on the population mean, denoted μ. It is natural to think of the **sample mean**, $\bar{X} = \sum_{i=1}^{n} X_i$, as an estimate of the population mean μ. We call the sample mean as a **statistic** calculated from the random sample $X_1, \ldots, X_n$. Other commonly used summary statistics include sample standard deviation and sample quantiles.

When using a statistic (e.g., the sample mean $\bar{X}$) to make statistical inference on the population attribute (e.g., population mean μ), the quality of inference is determined by the bias and uncertainty in the estimation, owing to the use of a sample in place of the population. Hence, it is important to study the distribution of a statistic that quantifies the bias and variability of the statistic. In particular, the distribution of the sample mean, $\bar{X}$ (or any other statistic), is called the **sampling distribution**. The sampling distribution depends on the sampling process, the statistic, the sample size n and the population distribution $F(\cdot)$. The central limit theorem gives the large-sample (sampling) distribution of the sample mean under certain conditions.

15.1.3 Central Limit Theorem

In statistics, there are variations of the central limit theorem (CLT) ensuring that, under certain conditions, the sample mean will approach the population mean with its sampling distribution approaching the normal distribution as the sample size goes to infinity. We give the Lindeberg–Levy CLT that establishes the asymptotic sampling distribution of the sample mean $\bar{X}$ calculated using a random sample from a universe population having a distribution $F(\cdot)$.

Lindeberg–Levy CLT. Let $X_1, \ldots, X_n$ be a random sample from a population distribution $F(\cdot)$ with mean μ and variance $\sigma^2 < \infty$. The difference between the sample mean $\bar{X}$ and μ, when multiplied by $\sqrt{n}$, converges in distribution to a normal distribution as the sample size goes to infinity. That is,

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma).$$

Note that the CLT does not require a parametric form for $F(\cdot)$. Based on the CLT, we may perform statistical inference on the population mean (we *infer*, not *deduce*). The types of inference we may perform include **estimation** of the population, **hypothesis testing** on whether a null statement is true, and **prediction** of future samples from the population.

15.2 Point Estimation and Properties

In this section, you learn how to

- estimate population parameters using method of moments estimation
- estimate population parameters based on maximum likelihood estimation

The population distribution function $F(\cdot)$ can usually be characterized by a limited (finite) number of terms called **parameters**, in which case we refer to the distribution as a **parametric distribution**. In contrast, in **nonparametric** analysis, the attributes of the sampling distribution are not limited to a small number of parameters.

For obtaining the population characteristics, there are different attributes related to the population distribution $F(\cdot)$. Such measures include the mean, median, percentiles (i.e., 95th percentile), and standard deviation. Because these summary measures do not depend on a specific parametric reference, they are **nonparametric** summary measures.

In **parametric** analysis, on the other hand, we may assume specific families of distributions with specific parameters. For example, people usually think of logarithm of claim amounts to be normally distributed with mean μ and standard deviation σ. That is, we assume that the claims have a *lognormal* distribution with parameters μ and σ. Alternatively, insurance companies commonly assume that claim severity follows a gamma distribution with a shape parameter α and a scale parameter θ. Here, the normal, lognormal, and gamma distributions are examples of parametric distributions. In the above examples, the quantities of μ, σ, α, and θ are known as *parameters*. For a given parametric distribution family, the distribution is uniquely determined by the values of the parameters.

One often uses θ to denote a summary attribute of the population. In parametric models, θ can be a parameter or a function of parameters from a distribution such as the normal mean and variance parameters. In nonparametric analysis, it can take a form of a nonparametric summary such as the population mean or standard deviation. Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be a function of the sample that provides a proxy, or an **estimate**, of θ. It is referred to as a **statistic**, a function of the sample $X_1, \ldots, X_n$.

Show Wisconsin Property Fund Example - Continued

Example – Wisconsin Property Fund. The sample mean 7.804 and the sample standard deviation 1.683 can be either deemed as nonparametric estimates of the population mean and standard deviation, or as parametric estimates of μ and σ of the normal distribution concerning the logarithmic claims. Using results from the lognormal distribution, we may estimate the expected claim, the lognormal mean, as 10,106.8 ($= \exp(7.804 + 1.683^2/2)$).

For the Wisconsin Property Fund data, we may denote $\hat{\mu} = 7.804$ and $\hat{\sigma} = 1.683$, with the hat notation denoting an **estimate** of the parameter based on the sample. In particular, such an estimate is referred to as a **point estimate**, a single approximation of the corresponding parameter. For point estimation, we introduce the two commonly used methods called the method of moments estimation and maximum likelihood estimation.

15.2.1 Method of Moments Estimation

Before defining the method of moments estimation, we define the the concept of **moments**. Moments are population attributes that characterize the distribution function $F(\cdot)$. Given a random draw X from $F(\cdot)$, the expectation $\mu_k = \mathrm{E}[X^k]$ is called the k**th moment** of X, $k = 1, 2, 3, \ldots$ For example, the population mean μ is the *first* moment. Furthermore, the expectation $\mathrm{E}[(X-\mu)^k]$ is called a k**th central moment**. Thus, the variance is the second central moment.

Using the random sample $X_1, \ldots, X_n$, we may construct the corresponding sample moment, $\hat{\mu}_k = (1/n)\sum_{i=1}^n X_i^k$, for estimating the population attribute μ_k. For example, we have used the sample mean $\bar{X}$ as an estimator for the population mean μ. Similarly, the second central moment can be estimated as $(1/n)\sum_{i=1}^n (X_i - \bar{X})^2$. Without assuming a parametric form for $F(\cdot)$, the sample moments constitute nonparametric estimates of the corresponding population attributes. Such an estimator based on matching of the corresponding sample and population moments is called a **method of moments estimator** (MME).

While the MME works naturally in a nonparametric model, it can be used to estimate parameters when a specific parametric family of distribution is assumed for $F(\cdot)$. Denote by $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$ the vector of parameters corresponding to a parametric distribution $F(\cdot)$. Given a distribution family, we commonly know the relationships between the parameters and the moments. In particular, we know the specific forms of the functions $h_1(\cdot), h_2(\cdot), \cdots, h_m(\cdot)$ such that $\mu_1 = h_1(\boldsymbol{\theta})$, $\mu_2 = h_2(\boldsymbol{\theta})$, $\cdots$, $\mu_m = h_m(\boldsymbol{\theta})$. Given the MME $\hat{\mu}_1, \ldots, \hat{\mu}_m$ from the random sample, the MME of the parameters $\hat{\theta}_1, \cdots, \hat{\theta}_m$ can be obtained by solving the equations of

$$\hat{\mu}_1 = h_1(\hat{\theta}_1, \cdots, \hat{\theta}_m);$$
$$\hat{\mu}_2 = h_2(\hat{\theta}_1, \cdots, \hat{\theta}_m);$$
$$\cdots$$
$$\hat{\mu}_m = h_m(\hat{\theta}_1, \cdots, \hat{\theta}_m).$$

Show Wisconsin Property Fund Example - Continued

Example – Wisconsin Property Fund. Assume that the claims follow a lognormal distribution, so that logarithmic claims follow a normal distribution. Specifically, assume $\ln(X)$ has a normal distribution with mean μ and variance σ^2, denoted as $\ln(X) \sim N(\mu, \sigma^2)$. It is straightforward that the MME $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = \sqrt{(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2}$. For the Wisconsin Property Fund example, the method of moments estimates are $\hat{\mu} = 7.804$ and $\hat{\sigma} = 1.683$.

15.2.2 Maximum Likelihood Estimation

When $F(\cdot)$ takes a parametric form, the maximum likelihood method is widely used for estimating the population parameters $\boldsymbol{\theta}$. Maximum likelihood estimation is based on the likelihood function, a function of the parameters given the observed sample. Denote by $f(x_i|\boldsymbol{\theta})$ the probability function of X_i evaluated at $X_i = x_i$ $(i = 1, 2, \cdots, n)$; it is the probability mass function in the case of a discrete X and the probability density function in the case of a continuous X. Assuming independence, the **likelihood function** of $\boldsymbol{\theta}$ associated with the observation $(X_1, X_2, \cdots, X_n) = (x_1, x_2, \cdots, x_n) = \mathbf{x}$ can be written as

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

with the corresponding **log-likelihood function** given by

$$l(\boldsymbol{\theta}|\mathbf{x}) = \ln(L(\boldsymbol{\theta}|\mathbf{x})) = \sum_{i=1}^{n} \ln f(x_i|\boldsymbol{\theta}).$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is the set of values of $\boldsymbol{\theta}$ that maximize the likelihood function (log-likelihood function), given the observed sample. That is, the MLE $\hat{\boldsymbol{\theta}}$ can be written as

$$\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta} \in \Theta} l(\boldsymbol{\theta}|\mathbf{x}),$$

where Θ is the parameter space of $\boldsymbol{\theta}$, and $\mathrm{argmax}_{\boldsymbol{\theta} \in \Theta} l(\boldsymbol{\theta}|\mathbf{x})$ is defined as the value of $\boldsymbol{\theta}$ at which the function $l(\boldsymbol{\theta}|\mathbf{x})$ reachs its maximum.

Given the analytical form of the likelihood function, the MLE can be obtained by taking the first derivative of the log-likelihood function with respect to $\boldsymbol{\theta}$, and setting the values of the partial derivatives to zero. That is, the MLE are the solutions of the equations of

$$\frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_1} = 0;$$
$$\frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_2} = 0;$$
$$\cdots$$
$$\frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_m} = 0,$$

provided that the second partial derivatives are negative.

For parametric models, the MLE of the parameters can be obtained either analytically (e.g., in the case of normal distributions and linear estimators), or numerically through iterative algorithms such as the Newton-Raphson method and its adaptive versions (e.g., in the case of generalized linear models with a non-normal response variable).

Normal distribution. Assume $(X_1, X_2, \cdots, X_n)$ to be a random sample from the normal distribution $N(\mu, \sigma^2)$. With an observed sample $(X_1, X_2, \cdots, X_n) = (x_1, x_2, \cdots, x_n)$, we can write the likelihood function of μ, σ^2 as

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

with the corresponding log-likelihood function given by

$$l(\mu, \sigma^2) = -\frac{n}{2}[\ln(2\pi) + \ln(\sigma^2)] - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 .$$

By solving

$$\frac{\partial l(\hat{\mu}, \sigma^2)}{\partial \hat{\mu}} = 0,$$

we obtain $\hat{\mu} = \bar{x} = (1/n)\sum_{i=1}^{n} x_i$. It is straightforward to verify that $\frac{\partial l^2(\hat{\mu},\sigma^2)}{\partial \hat{\mu}^2} |_{\hat{\mu}=\bar{x}} < 0$. Since this works for arbitrary x, $\hat{\mu} = \bar{X}$ is the MLE of μ. Similarly, by solving

$$\frac{\partial l(\mu, \hat{\sigma}^2)}{\partial \hat{\sigma}^2} = 0,$$

we obtain $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(x_i - \mu)^2$. Further replacing μ by $\hat{\mu}$, we derive the MLE of σ^2 as $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Hence, the sample mean $\bar{X}$ and $\hat{\sigma}^2$ are both the *MME* and *MLE* for the mean μ and variance σ^2, under a normal population distribution $F(\cdot)$. More details regarding the properties of the likelihood function, and the derivation of MLE under parametric distributions other than the normal distribution are given in Appendix Chapter 16.

15.3 Interval Estimation

In this section, you learn how to

- derive the exact sampling distribution of the MLE of the normal mean
- obtain the large-sample approximation of the sampling distribution using the large sample properties of the MLE
- construct a confidence interval of a parameter based on the large sample properties of the MLE

Now that we have introduced the MME and MLE, we may perform the first type of statistical inference, **interval estimation** that quantifies the uncertainty resulting from the use of a finite sample. By deriving the sampling distribution of MLE, we can estimate an interval (a confidence interval) for the parameter. Under the frequentist approach (e.g., that based on maximum likelihood estimation), the confidence intervals generated from the same random sampling frame will cover the true value the majority of times (e.g., 95% of the times), if we repeat the sampling process and re-calculate the interval over and over again. Such a process requires the derivation of the sampling distribution for the MLE.

15.3.1 Exact Distribution for Normal Sample Mean

Due to the **additivity** property of the normal distribution (i.e., a sum of normal random variables that follows a multivariate normal distribution still follows a normal distribution) and that the normal distribution belongs to the **location–scale family** (i.e., a location and/or scale transformation of a normal random variable has a normal distribution), the sample mean $\bar{X}$ of a random sample from a normal $F(\cdot)$ has a normal sampling distribution for any finite n. Given $X_i \sim^{iid} N(\mu, \sigma^2)$, $i = 1, \ldots, n$, the MLE of μ has an exact distribution

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Hence, the sample mean is an unbiased estimator of μ. In addition, the uncertainty in the estimation can be quantified by its variance σ^2/n, that decreases with the sample size n. When the sample size goes to infinity, the sample mean will approach a single mass at the true value.

15.3.2 Large-sample Properties of MLE

For the MLE of the mean parameter and any other parameters of other parametric distribution families, however, we usually cannot derive an exact sampling distribution for finite samples. Fortunately, when the sample size is sufficiently large, MLEs can be approximated by a normal distribution. Due to the general maximum likelihood theory, the MLE has some nice large-sample properties.

- The MLE $\hat{\theta}$ of a parameter θ, is a **consistent** estimator. That is, $\hat{\theta}$ converges in probability to the true value θ, as the sample size n goes to infinity.
- The MLE has the **asymptotic normality** property, meaning that the estimator will converge in distribution to a normal distribution centered around the true value, when the sample size goes to infinity. Namely,
$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N\left(0, V\right), \quad \text{as} \quad n \rightarrow \infty,$$
where V is the inverse of the Fisher Information. Hence, the MLE $\hat{\theta}$ approximately follows a normal distribution with mean θ and variance V/n, when the sample size is large.
- The MLE is **efficient**, meaning that it has the smallest asymptotic variance V, commonly referred to as the **Cramer–Rao lower bound**. In particular, the Cramer–Rao lower bound is the inverse of the Fisher information defined as $\mathcal{I}(\theta) = -\mathrm{E}(\partial^2 \ln f(X;\theta)/\partial\theta^2)$. Hence, $\mathrm{Var}(\hat{\theta})$ can be estimated based on the observed Fisher information that can be written as $-\sum_{i=1}^n \partial^2 \ln f(X_i;\theta)/\partial\theta^2$.

For many parametric distributions, the Fisher information may be derived analytically for the MLE of parameters. For more sophisticated parametric models, the Fisher information can be evaluated numerically using numerical integration for continuous distributions, or numerical summation for discrete distributions.

15.3.3 Confidence Interval

Given that the MLE $\hat{\theta}$ has either an exact or an approximate normal distribution with mean θ and variance $\mathrm{Var}(\hat{\theta})$, we may take the square root of the variance and plug-in the estimate to define $se(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$. A **standard error** is an estimated standard deviation that quantifies the uncertainty in the estimation resulting from the use of a finite sample. Under some regularity conditions governing the population distribution, we may establish that the statistic

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})}$$

converges in distribution to a Student-t distribution with degrees of freedom (a parameter of the distribution) $n - p$, where p is the number of parameters in the model other than the variance. For example, for the normal distribution case, we have $p = 1$ for the parameter μ; for a linear regression model with an independent variable, we have $p = 2$ for the parameters of the intercept and the independent variable. Denote by $t_{n-p}(1-\alpha/2)$ the $100 \times (1-\alpha/2)$-th percentile of the Student-t distribution that satisfies $\Pr\left[t < t_{n-p}\left(1-\alpha/2\right)\right] = 1-\alpha/2$. We have,

$$\Pr\left[-t_{n-p}\left(1-\frac{\alpha}{2}\right) < \frac{\hat{\theta}-\theta}{se(\hat{\theta})} < t_{n-p}\left(1-\frac{\alpha}{2}\right)\right] = 1-\alpha,$$

from which we can derive a **confidence interval** for θ. From the above equation we can derive a pair of statistics, $\hat{\theta}_1$ and $\hat{\theta}_2$, that provide an interval of the form $[\hat{\theta}_1, \hat{\theta}_2]$. This interval is a $1-\alpha$ confidence interval for θ such that $\Pr\left(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\right) = 1-\alpha$, where the probability $1-\alpha$ is referred to as the **confidence level**. Note that the above confidence interval is not valid for small samples, except for the case of the normal mean.

Normal distribution. For the normal population mean μ, the MLE has an exact sampling distribution $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$, in which we can estimate $se(\hat{\theta})$ by $\hat{\sigma}/\sqrt{n}$. Based on the **Cochran's theorem**, the resulting statistic has an exact Student-t distribution with degrees of freedom $n-1$. Hence, we can derive the lower and upper bounds of the confidence interval as

$$\hat{\mu}_1 = \hat{\mu} - t_{n-1}\left(1-\frac{\alpha}{2}\right)\frac{\hat{\sigma}}{\sqrt{n}}$$

and

$$\hat{\mu}_2 = \hat{\mu} + t_{n-1}\left(1-\frac{\alpha}{2}\right)\frac{\hat{\sigma}}{\sqrt{n}}.$$

When $\alpha = 0.05$, $t_{n-1}(1-\alpha/2) \approx 1.96$ for large values of n. Based on the Cochran's theorem, the confidence interval is valid regardless of the sample size.

Show Wisconsin Property Fund Example - Continued

Example – Wisconsin Property Fund. For the lognormal claim model, (7.715235, 7.893208) is a 95% confidence interval for μ.

More details regarding interval estimation based the MLE of other parameters and distribution families are given in Appendix Chapter 17.

15.4 Hypothesis Testing

In this section, you learn how to

- understand the basic concepts in hypothesis testing including the level of significance and the power of a test
- perform hypothesis testing such as a Student-t test based on the properties of the MLE
- construct a likelihood ratio test for a single parameter or multiple parameters from the same statistical model
- use information criteria such as the Akaike's information criterion or the Bayesian information criterion to perform model selection

For the parameter(s) $\boldsymbol{\theta}$ from a parametric distribution, an alternative type of statistical inference is called **hypothesis tesing** that verifies whether a hypothesis regarding the parameter(s) is true, under a given probability called the **level of significance** α (e.g., 5%). In hypothesis testing, we reject the null hypothesis, a restrictive statement concerning the parameter(s), if the probability of observing a random sample as extremal as the observed one is smaller than α, if the null hypothesis were true.

15.4.1 Basic Concepts

In a statistical test, we are usually interested in testing whether a statement regarding some parameter(s), a **null hypothesis** (denoted H_0), is true given the observed data. The null hypothesis can take a general form $H_0 : \theta \in \Theta_0$, where Θ_0 is a subset of the parameter space Θ of θ that may contain multiple parameters. For the case with a single parameter θ, the null hypothesis usually takes either the form $H_0 : \theta = \theta_0$ or

$H_0 : \theta \leq \theta_0$. The opposite of the null hypothesis is called the **alternative hypothesis** that can be written as $H_a : \theta \neq \theta_0$ or $H_a : \theta > \theta_0$. The statistical test on $H_0 : \theta = \theta_0$ is called a **two-sided** as the alternative hypothesis contains two ineqalities of $H_a : \theta < \theta_0$ or $\theta > \theta_0$. In contrast, the statistical test on either $H_0 : \theta \leq \theta_0$ or $H_0 : \theta \geq \theta_0$ is called a **one-sided** test.

A statistical test is usually constructed based on a statistic T and its exact or large-sample distribution. The test typically rejects a two-sided test when either $T > c_1$ or $T < c_2$, where the two constants c_1 and c_2 are obtained based on the sampling distribution of T at a probability level α called the **level of significance**. In particular, the level of significance α satisfies

$$\alpha = \Pr(\text{reject } H_0 | H_0 \text{ is true}),$$

meaning that if the null hypothesis were true, we would reject the null hypothesis only 5% of the times, if we repeat the sampling process and perform the test over and over again.

Thus, the level of significance is the probability of making a **type I error** (error of the first kind), the error of incorrectly rejecting a true null hypothesis. For this reason, the level of significance α is also referred to as the type I error rate. Another type of error we may make in hypothesis testing is the **type II error** (error of the second kind), the error of incorrectly accepting a false null hypothesis. Similarly, we can define the **type II error rate** as the probability of not rejecting (accepting) a null hypothesis given that it is not true. That is, the type II error rate is given by

$$\Pr(\text{accept } H_0 | H_0 \text{ is false}).$$

Another important quantity concerning the quality of the statistical test is called the **power** of the test β, defined as the probability of rejecting a false null hypothesis. The mathematical definition of the power is

$$\beta = \Pr(\text{reject } H_0 | H_0 \text{ is false}).$$

Note that the power of the test is typically calculated based on a specific alternative value of $\theta = \theta_a$, given a specific sampling distribution and a given sample size. In real experimental studies, people usually calculate the required sample size in order to choose a sample size that will ensure a large chance of obtaining a statistically significant test (i.e., with a prespecified statistical power such as 85%).

15.4.2 Student-t test based on MLE

Based on the results from Section 15.3.1, we can define a Student t test for testing $H_0 : \theta = \theta_0$. In particular, we define the test statistic as

$$t\text{-stat} = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})},$$

which has a large-sample distribution of a Student-t distribution with degrees of freedom $n - p$, when the null hypothesis is true (i.e., when $\theta = \theta_0$).

For a given **level of significance** α, say 5%, we reject the null hypothesis if the event t-stat $< -t_{n-p}(1 - \alpha/2)$ or t-stat $> t_{n-p}(1 - \alpha/2)$ occurs (the **rejection region**). Under the null hypothesis H_0, we have

$$\Pr\left[t\text{-stat} < -t_{n-p}\left(1 - \frac{\alpha}{2}\right)\right] = \Pr\left[t\text{-stat} > t_{n-p}\left(1 - \frac{\alpha}{2}\right)\right] = \frac{\alpha}{2}.$$

In addition to the concept of rejection region, we may reject the test based on the p**-value** defined as $2\Pr(T > |t\text{-stat}|)$ for the aforementioned two-sided test, where the random variable $T \sim T_{n-p}$. We reject the null hypothesis if p-value is smaller than and equal to α. For a given sample, a p-value is defined to be the smallest significance level for which the null hypothesis would be rejected.

Similarly, we can construct a one-sided test for the null hypothesis $H_0 : \theta \leq \theta_0$ (or $H_0 : \theta \geq \theta_0$). Using the same test statistic, we reject the null hypothesis when t-stat $> t_{n-p}(1 - \alpha)$ (or t-stat $< -t_{n-p}(1 - \alpha)$ for the test on $H_0 : \theta \geq \theta_0$). The corresponding p-value is defined as $\Pr(T > |t\text{-stat}|)$ (or $\Pr(T < |t\text{-stat}|)$ for the test on $H_0 : \theta \geq \theta_0$). Note that the test is not valid for small samples, except for the case of the test on the normal mean.

One-sample t Test for Normal Mean. For the test on the normal mean of the form $H_0 : \mu = \mu_0$, $H_0 : \mu \leq \mu_0$ or $H_0 : \mu \geq \mu_0$, we can define the test statistic as

$$t\text{-stat} = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

for which we have an exact sampling distribution t-stat $\sim T_{n-1}$ from the Cochran's theorem, with T_{n-1} denoting a Student-t distribution with degrees of freedom $n - 1$. According to the Cochran's theorem, the test is valid for both small and large samples.

Show Wisconsin Property Fund Example - Continued

Example – Wisconsin Property Fund. Assume that mean logarithmic claims have historically been approximately by $\mu_0 = \ln(5000) = 8.517$. We might want to use the 2010 data to assess whether the mean of the distribution has changed (a two-sided test), or whether it has increased (a one-sided test). Given the actual 2010 average $\hat{\mu} = 7.804$, we may use the one-sample t test to assess whether this is a significant departure from $\mu_0 = 8.517$ (i.e., in testing $H_0 : \mu = 8.517$). The test statistic t-stat $= (8.517 - 7.804)/(1.683/\sqrt{1377}) = 15.72 > t_{1376}\,(0.975)$. Hence, we reject the two-sided test at $\alpha = 5\%$. Similarly, we will reject the one-sided test at $\alpha = 5\%$.

Show Wisconsin Property Fund Example - Continued

Example – Wisconsin Property Fund. For numerical stability and extensions to regression applications, statistical packages often work with transformed versions of parameters. The following estimates are from the **R** package **VGAM** (the function). More details on the MLE of other distribution families are given in Appendix Chapter 17.

Distribution	Parameter Estimate	Standard Error	t-stat
Gamma	10.190	0.050	203.831
	-1.236	0.030	-41.180
Lognormal	7.804	0.045	172.089
	0.520	0.019	27.303
Pareto	7.733	0.093	82.853
	-0.001	0.054	-0.016
GB2	2.831	1.000	2.832
	1.203	0.292	4.120
	6.329	0.390	16.220
	1.295	0.219	5.910

15.4.3 Likelihood Ratio Test

In the previous subsection, we have introduced the Student-t test on a single parameter, based on the properties of the MLE. In this section, we define an alternative test called the **likelihood ratio test** (LRT). The LRT may be used to test multiple parameters from the same statistical model.

Given the likelihood function $L(\theta|\mathbf{x})$ and $\Theta_0 \subset \Theta$, the likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ against $H_a : \theta \notin \Theta_0$ is given by

$$L = \frac{\sup_{\theta\in\Theta_0} L(\theta|\mathbf{x})}{\sup_{\theta\in\Theta} L(\theta|\mathbf{x})},$$

and that for testing $H_0 : \theta = \theta_0$ versis $H_a : \theta \neq \theta_0$ is

$$L = \frac{L(\theta_0|\mathbf{x})}{\sup_{\theta\in\Theta} L(\theta|\mathbf{x})}.$$

The LRT rejects the null hypothesis when $L < c$, with the threshold depending on the level of significance α, the sample size n, and the number of parameters in θ. Based on the **Neyman–Pearson Lemma**, the LRT is the **uniformly most powerful** (UMP) test for testing $H_0 : \theta = \theta_0$ versis $H_a : \theta = \theta_a$. That is, it provides the largest power β for a given α and a given alternative value θ_a.

Based on the **Wilks's Theorem**, the likelihood ratio test statistic $-2\ln(L)$ converges in distribution to a Chi-square distribution with the degree of freedom being the difference between the dimensionality of the parameter spaces Θ and Θ_0, when the sample size goes to infinity and when the null model is nested within the alternative model. That is, when the null model is a special case of the alternative model containing a restricted sample space, we may approximate c by $\chi^2_{p_1-p_2}(1-\alpha)$, the $100\times(1-\alpha)$ th percentile of the Chi-square distribution, with $p_1 - p_2$ being the degrees of freedom, and p_1 and p_2 being the numbers of parameters in the alternative and null models, respectively. Note that the LRT is also a large-sample test that will not be valid for small samples.

15.4.4 Information Criteria

In real-life applications, the LRT has been commonly used for comparing two nested models. The LRT approach as a model selection tool, however, has two major drawbacks: 1) It typically requires the null model to be nested within the alternative model; 2) models selected from the LRT tends to provide in-sample over-fitting, leading to poor out-of-sample prediction. In order to overcome these issues, model selection based on information criteria, applicable to non-nested models while taking into consideration the model complexity, is more widely used for model selection. Here, we introduce the two most widely used criteria, the Akaike's information criterion and the Bayesian information criterion.

In particular, the **Akaike's information criterion** (AIC) is defined as

$$AIC = -2\ln L(\hat{\boldsymbol{\theta}}) + 2p,$$

where $\hat{\boldsymbol{\theta}}$ denotes the MLE of $\boldsymbol{\theta}$, and p is the number of parameters in the model. The additional term $2p$ represents a penalty for the complexity of the model. That is, with the same maximized likelihood function, the AIC favors model with less parameters. We note that the AIC does not consider the impact from the sample size n.

Alternatively, people use the **Bayesian information criterion** (BIC) that takes into consideration the sample size. The BIC is defined as

$$BIC = -2\ln L(\hat{\boldsymbol{\theta}}) + p\ \ln(n).$$

We observe that the BIC generally puts a higher weight on the number of parameters. With the same maximized likelihood function, the BIC will suggest a more parsimonious model than the AIC.

Show Wisconsin Property Fund Example - Continued

Example – Wisconsin Property Fund. Both the AIC and BIC statistics suggest that the $GB2$ is the best fitting model whereas gamma is the worst.

Distribution	AIC	BIC
Gamma	28,305.2	28,315.6
Lognormal	26,837.7	26,848.2
Pareto	26,813.3	26,823.7
GB2	26,768.1	26,789.0

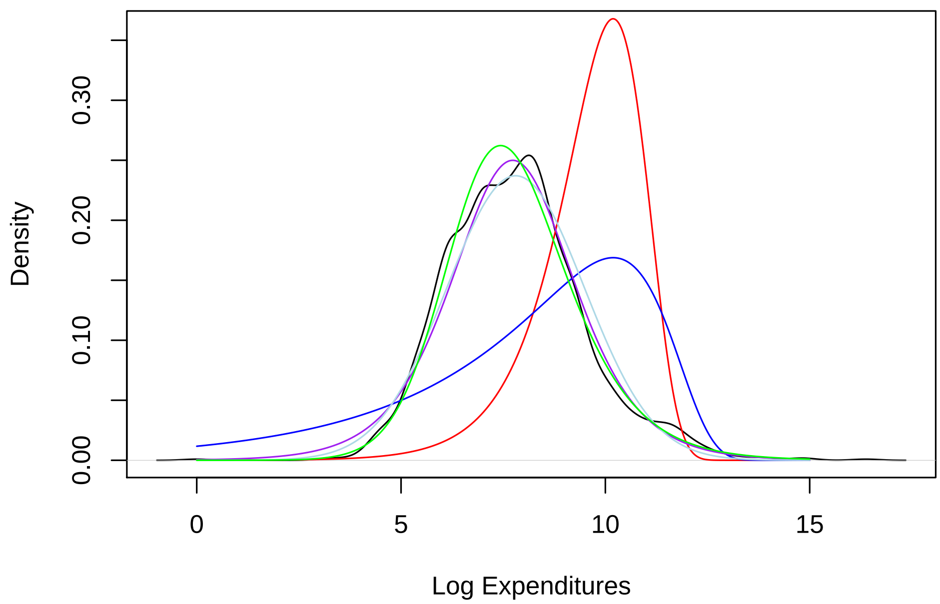


Figure 15.2: Fitted Claims Distribution

In this graph,

- black represents actual (smoothed) logarithmic claims
- Best approximated by green which is fitted GB2
- Pareto (purple) and Lognormal (lightblue) are also pretty good
- Worst are the exponential (in red) and gamma (in dark blue)

```
## Sample size:  6258
```

Show R Code

R Code for Fitted Claims Distributions

```
# R Code to fit several claims distributions
ClaimLev <- read.csv("Data/CLAIMLEVEL.csv", header=TRUE); nrow(ClaimLev)
ClaimData<-subset(ClaimLev,Year==2010);
#Use "VGAM" library for estimation of parameters
library(VGAM)
fit.LN <- vglm(Claim ~ 1, family=lognormal, data = ClaimData)
fit.gamma <- vglm(Claim ~ 1, family=gamma2, data = ClaimData)
  theta.gamma<-exp(coef(fit.gamma)[1])/exp(coef(fit.gamma)[2])
  alpha.gamma<-exp(coef(fit.gamma)[2])
fit.exp <- vglm(Claim ~ 1, exponential, data = ClaimData)
fit.pareto <- vglm(Claim ~ 1, paretoII, loc=0, data = ClaimData)

###################################################
#  Inference assuming a GB2 Distribution - this is more complicated
# The likelihood functon of GB2 distribution (negative for optimization)
```

```
likgb2 <- function(param) {
  a1 <- param[1]
  a2 <- param[2]
  mu <- param[3]
  sigma <- param[4]
  yt <- (log(ClaimData$Claim)-mu)/sigma
  logexpyt<-ifelse(yt>23,yt,log(1+exp(yt)))
 logdens <- a1*yt - log(sigma) - log(beta(a1,a2)) - (a1+a2)*logexpyt -log(ClaimData$Claim)
  return(-sum(logdens))
}
#  "optim" is a general purpose minimization function
gb2bop <- optim(c(1,1,0,1),likgb2,method=c("L-BFGS-B"),
                lower=c(0.01,0.01,-500,0.01),upper=c(500,500,500,500),hessian=TRUE)
###################################################
# Plotting the fit using densities (on a logarithmic scale)
plot(density(log(ClaimData$Claim)), ylim=c(0,0.36),main="", xlab="Log Expenditures")
x <- seq(0,15,by=0.01)
fexp_ex = dgamma(exp(x), scale = exp(-coef(fit.exp)), shape = 1)*exp(x)
lines(x,fexp_ex, col="red")
fgamma_ex = dgamma(exp(x), shape = alpha.gamma, scale=theta.gamma)*exp(x)
lines(x,fgamma_ex,col="blue")
fpareto_ex = dparetoII(exp(x),loc=0,shape = exp(coef(fit.pareto)[2]), scale = exp(coef(fit.pareto)[1]))*ex
lines(x,fpareto_ex,col="purple")
flnorm_ex = dlnorm(exp(x), mean = coef(fit.LN)[1], sd = exp(coef(fit.LN)[2]))*exp(x)
lines(x,flnorm_ex, col="lightblue")
# density for GB II
gb2density <- function(x){
  a1 <- gb2bop$par[1]
  a2 <- gb2bop$par[2]
  mu <- gb2bop$par[3]
  sigma <- gb2bop$par[4]
  xt <- (log(x)-mu)/sigma
  logexpxt<-ifelse(xt>23,yt,log(1+exp(xt)))
  logdens <- a1*xt - log(sigma) - log(beta(a1,a2)) - (a1+a2)*logexpxt -log(x)
  exp(logdens)
}
fGB2_ex = gb2density(exp(x))*exp(x)
lines(x,fGB2_ex, col="green")
```

Contributors

- **Lei (Larry) Hua**, Northern Illinois University, and **Edward W. (Jed) Frees**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter. Email: lhua@niu.edu or jfrees@bus.wisc.edu for chapter comments and suggested improvements.

Chapter 16

Appendix B: Iterated Expectations

This appendix introduces the laws related to iterated expectations. In particular, Section 16.1 introduces the concepts of conditional distribution and conditional expectation. Section 16.2 introduces the Law of Iterated Expectations and the Law of Total Variance.

In some situations, we only observe a single outcome but can conceptualize an outcome as resulting from a two (or more) stage process. Such types of statistical models are called **two-stage**, or **hierarchical** models. Some special cases of hierarchical models include:

- models where the parameters of the distribution are random variables;
- mixture distribution, where Stage 1 represents the draw of a sub-population and Stage 2 represents a random variable from a distribution that is determined by the sub-population drew in Stage 1;
- an aggregate distribution, where Stage 1 represents the draw of the number of events and Stage two represents the loss amount occurred per event.

In these situations, the process gives rise to a conditional distribution of a random variable (the Stage 2 outcome) given the other (the Stage 1 outcome). The Law of Iterated Expectations can be useful for obtaining the unconditional expectation or variance of a random variable in such cases.

16.1 Conditional Distribution and Conditional Expectation

In this section, you learn

- the concepts related to the conditional distribution of a random variable given another
- how to define the conditional expectation and variance based on the conditional distribution function

The iterated expectations are the laws regarding calculation of the expectation and variance of a random variable using a conditional distribution of the variable given another variable. Hence, we first introduce the concepts related to the conditional distribution, and the calculation of the conditional expectation and variable based on a given conditional distribution.

16.1.1 Conditional Distribution

Here we introduce the concept of conditional distribution respectively for discrete and continuous random variables.

Discrete Case

Suppose that X and Y are both discrete random variables, meaning that they can take a finite or countable number of possible values with a positive probability. The **joint probability (mass) function** of (X, Y) is defined as

$$p(x, y) = \Pr[X = x, Y = y]$$

.

When X and Y are **independent** (the value of X does not depend on that of Y), we have

$$p(x, y) = p(x)p(y),$$

with $p(x) = \Pr[X = x]$ and $p(y) = \Pr[Y = y]$ being the **marginal probability function** of X and Y, respectively.

Given the joint probability function, we may obtain the marginal probability functions of Y as

$$p(y) = \sum_x p(x, y),$$

where the summation is over all possible values of x, and the marginal probability function of X can be obtained in a similar manner.

The **conditional probability (mass) function** of $(Y|X)$ is defined as

$$p(y|x) = \Pr[Y = y|X = x] = \frac{p(x, y)}{\Pr[X = x]},$$

where we may obtain the conditional probability function of $(X|Y)$ in a similar manner. In particular, the above conditional probability represents the probability of the event $Y = y$ given the event $X = x$. Hence, even in cases where $\Pr[X = x] = 0$, the function may be given as a particular form, in real applications.

Continuous Case

For continuous random variables X and Y, we may define their joint probability (density) function based on the joint cumulative distribution function. The **joint cumulative distribution function** of (X, Y) is defined as

$$F(x, y) = \Pr[X \leq x, Y \leq y].$$

When X and Y are *independent*, we have

$$F(x, y) = F(x)F(y),$$

with $F(x) = \Pr[X \leq x]$ and $F(y) = \Pr[Y \leq y]$ being the **cumulative distribution function** (cdf) of X and Y, respectively. The random variable X is referred to as a **continuous** random variable if its cdf is continuous on x.

When the cdf $F(x)$ is continuous on x, then we define $f(x) = \partial F(x)/\partial x$ as the **(marginal) probability density function** (pdf) of X. Similarly, if the joint cdf $F(x, y)$ is continuous on both x and y, we define

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

as the **joint probability density function** of (X, Y), in which case we refer to the random variables as **jointly continuous**.

When X and Y are *independent*, we have

$$f(x, y) = f(x)f(y).$$

Given the joint density function, we may obtain the marginal density function of Y as

$$f(y) = \int_x f(x, y)\, dx,$$

where the integral is over all possible values of x, and the marginal probability function of X can be obtained in a similar manner.

Based on the joint pdf and the marginal pdf, we define the **conditional probability density function** of $(Y|X)$ as

$$f(y|x) = \frac{f(x, y)}{f(x)},$$

where we may obtain the conditional probability function of $(X|Y)$ in a similar manner. Here, the conditional density function is the density function of y given $X = x$. Hence, even in cases where $\Pr[X = x] = 0$ or when $f(x)$ is not defined, the function may be given in a particular form in real applications.

16.1.2 Conditional Expectation and Conditional Variance

Now we define the conditional expectation and variance based on the conditional distribution defined in the previous subsection.

Discrete Case

For a discrete random variable Y, its **expectation** is defined as $\mathrm{E}[Y] = \sum_y y\, p(y)$ if its value is finite, and its **variance** is defined as $\mathrm{Var}[Y] = \mathrm{E}\{(Y - \mathrm{E}[Y])^2\} = \sum_y y^2\, p(y) - \{\mathrm{E}[Y]\}^2$ if its value is finite.

For a discrete random variable Y, the **conditional expectation** of the random variable Y given the event $X = x$ is defined as

$$\mathrm{E}[Y|X = x] = \sum_y y\, p(y|x),$$

where X does not have to be a discrete variable, as far as the conditional probability function $p(y|x)$ is given.

Note that the conditional expectation $\mathrm{E}[Y|X = x]$ is a fixed number. When we replace x with X on the right hand side of the above equation, we can define the expectation of Y given the random variable X as

$$\mathrm{E}[Y|X] = \sum_y y\, p(y|X),$$

which is still a *random variable*, and the randomness comes from X.

In a similar manner, we can define the **conditional variance** of the random variable Y given the event $X = x$ as

$$\mathrm{Var}[Y|X = x] = \mathrm{E}[Y^2|X = x] - \{\mathrm{E}[Y|X = x]\}^2 = \sum_y y^2\, p(y|x) - \{\mathrm{E}[Y|X = x]\}^2.$$

The variance of Y given X, $\mathrm{Var}[Y|X]$ can be defined by replacing x by X in the above equation, and $\mathrm{Var}[Y|X]$ is still a random variable and the randomness comes from X.

Continuous Case

For a continuous random variable Y, its **expectation** is defined as $\mathrm{E}[Y] = \int_y y\, f(y)dy$ if the integral exists, and its **variance** is defined as $\mathrm{Var}[Y] = \mathrm{E}\{(X - \mathrm{E}[Y])^2\} = \int_y y^2\, f(y)dy - \{\mathrm{E}[Y]\}^2$ if its value is finite.

For jointly continuous random variables X and Y, the **conditional expectation** of the random variable Y given $X = x$ is defined as

$$\mathrm{E}[Y|X = x] = \int_y y\, f(y|x)dy.$$

where X does not have to be a continuous variable, as far as the conditional probability function $f(y|x)$ is given.

Similarly, the conditional expectation $\mathrm{E}[Y|X = x]$ is a fixed number. When we replace x with X on the right-hand side of the above equation, we can define the expectation of Y given the random variable X as

$$\mathrm{E}[Y|X] = \int_y y\, p(y|X)\, dy,$$

which is still a *random variable*, and the randomness comes from X.

In a similar manner, we can define the **conditional variance** of the random variable Y given the event $X = x$ as

$$\mathrm{Var}[Y|X = x] = \mathrm{E}[Y^2|X = x] - \{\mathrm{E}[Y|X = x]\}^2 = \int_y y^2\, f(y|x)\, dy - \{\mathrm{E}[Y|X = x]\}^2.$$

The variance of Y given X, $\mathrm{Var}[Y|X]$ can then be defined by replacing x by X in the above equation, and similarly $\mathrm{Var}[Y|X]$ is also a random variable and the randomness comes from X.

16.2 Iterated Expectations and Total Variance

In this section, you learn

- the Law of Iterated Expectations for calculating the expectation of a random variable based on its conditional distribution given another random variable
- the Law of Total Variance for calculating the variance of a random variable based on its conditional distribution given another random variable
- how to calculate the expectation and variance based on an example of a two-stage model

16.2.1 Law of Iterated Expectations

Consider two random variables X and Y, and $h(X, Y)$, a random variable depending on the function h, X and Y.

Assuming all the expectations exist and are finite, the **Law of Iterated Expectations** states that

$$\mathrm{E}[h(X, Y)] = \mathrm{E}\{\mathrm{E}[h(X, Y)|X]\},$$

where the first (inside) expectation is taken with respect to the random variable Y and the second (outside) expectation is taken with respect to X.

For the Law of Iterated Expectations, the random variables may be discrete, continuous, or a hybrid combination of the two. We use the example of discrete variables of X and Y to illustrate the calculation of the unconditional expectation using the Law of Iterated Expectations. For continuous random variables, we only need to replace the summation with the integral, as illustrated earlier in the appendix.

Given $p(y|x)$ the joint pmf of X and Y, the conditional expectation of $h(X,Y)$ given the event $X = x$ is defined as

$$\mathrm{E}\left[h(X,Y)|X=x\right] = \sum_y h(x,y)p(y|x),$$

and the conditional expectation of $h(X,Y)$ given X being a *random variable* can be written as

$$\mathrm{E}\left[h(X,Y)|X\right] = \sum_y h(X,y)p(y|X).$$

The unconditional expectation of $h(X,Y)$ can then be obtained by taking the expectation of $\mathrm{E}\left[h(X,Y)|X\right]$ with respect to the random variable X. That is, we can obtain $\mathrm{E}[h(X,Y)]$ as

$$\begin{aligned}\mathrm{E}\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\} &= \sum_x \left\{\sum_y h(x,y)p(y|x)\right\} p(x) \\ &= \sum_x \sum_y h(x,y)p(y|x)p(x) \\ &= \sum_x \sum_y h(x,y)p(x,y) = \mathrm{E}[h(X,Y)]\end{aligned}\quad .$$

The Law of Iterated Expectations for the continuous and hybrid cases can be proved in a similar manner, by replacing the corresponding summation(s) by integral(s).

16.2.2 Law of Total Variance

Assuming that all the variances exist and are finite, the **Law of Total Variance** states that

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

where the first (inside) expectation/variance is taken with respect to the random variable Y and the second (outside) expectation/variance is taken with respect to X. Thus, the unconditional variance equals to the expectation of the conditional variance plus the variance of the conditional expectation.

Show Technical Detail

In order to verify this rule, first note that we can calculate a conditional variance as

$$\mathrm{Var}\left[h(X,Y)|X\right] = \mathrm{E}[h(X,Y)^2|X] - \left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2.$$

From this, the expectation of the conditional variance is

$$\begin{aligned}\mathrm{E}\{\mathrm{Var}\left[h(X,Y)|X\right]\} &= \mathrm{E}\left\{\mathrm{E}\left[h(X,Y)^2|X\right]\right\} - \mathrm{E}\left(\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2\right) \\ &= \mathrm{E}\left[h(X,Y)^2\right] - \mathrm{E}\left(\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2\right).\end{aligned} \tag{16.1}$$

Further, note that the conditional expectation, $\mathrm{E}\left[h(X,Y)|X\right]$, is a function of X, denoted $g(X)$. Thus, $g(X)$ is a random variable with mean $\mathrm{E}[h(X,Y)]$ and variance

$$\begin{aligned}\mathrm{Var}\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\} &= \mathrm{Var}[g(X)] \\ &= \mathrm{E}[g(X)^2] - \{\mathrm{E}[g(X)]\}^2 \\ &= \mathrm{E}\left(\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2\right) - \{\mathrm{E}[h(X,Y)]\}^2.\end{aligned} \tag{16.2}$$

Thus, adding Equations (16.1) and (16.2) leads to the unconditional variance $\mathrm{Var}\left[h(X,Y)\right]$.

16.2.3 Application

To apply the Law of Iterated Expectations and the Law of Total Variance, we generally adopt the following procedure.

1. Identify the random variable that is being conditioned upon, typically a stage 1 outcome (that is not observed).
2. Conditional on the stage 1 outcome, calculate summary measures such as a mean, variance, and the like.
3. There are several results of the step 2, one for each stage 1 outcome. Then, combine these results using the iterated expectations or total variance rules.

Mixtures of Finite Populations. Suppose that the random variable N_1 represents a realization of the number of claims in a policy year from the population of good drivers and N_2 represents that from the population of bad drivers. For a specific driver, there is a probability α that (s)he is a good driver. For a specific draw N, we have

$$N = \begin{cases} N_1, & \text{if (s)he is a good driver;} \\ N_2, & \text{otherwise.} \end{cases}$$

Let T be the indicator whether (s)he is a good driver, with $T = 1$ representing that the driver is a good driver with $\Pr[T = 1] = \alpha$ and $T = 2$ representing that the driver is a bad driver with $\Pr[T = 2] = 1 - \alpha$.

From the Law of Iterated Expectations, we can obtain the expected number of claims as

$$\mathrm{E}[N] = \mathrm{E}\left\{\mathrm{E}\left[N|T\right]\right\} = \mathrm{E}[N_1] \times \alpha + \mathrm{E}[N_2] \times (1-\alpha).$$

From the Law of Total Variance, we can obtain the variance of N as

$$\mathrm{Var}[N] = \mathrm{E}\left\{\mathrm{Var}\left[N|T\right]\right\} + \mathrm{Var}\left\{\mathrm{E}\left[N|T\right]\right\}.$$

To be more concrete, suppose that N_j follows a Poisson distribution with the mean λ_j, $j = 1, 2$. Then we have

$$\mathrm{Var}[N|T = j] = \mathrm{E}[N|T = j] = \lambda_j, \quad j = 1, 2.$$

Thus, we can derive the expectation of the conditional variance as

$$\mathrm{E}\left\{\mathrm{Var}\left[N|T\right]\right\} = \alpha\lambda_1 + (1-\alpha)\lambda_2$$

and the variance of the conditional expectation as

$$\mathrm{Var}\left\{\mathrm{E}\left[N|T\right]\right\} = (\lambda_1 - \lambda_2)^2\alpha(1-\alpha).$$

Note that the later is the variance for a Bernoulli with outcomes λ_1 and λ_2, and the binomial probability α.

Based on the Law of Total Variance, the unconditional variance of N is given by

$$\mathrm{Var}[N] = \alpha\lambda_1 + (1-\alpha)\lambda_2 + (\lambda_1 - \lambda_2)^2\alpha(1-\alpha).$$

16.3 Conjugate Distributions

As described in Section 4.4.1, for conjugate distributions the posterior and the prior come from the same family of distributions. In insurance applications, this broadly occurs in a "family of distribution familes" known as the linear exponential family which we introduce first.

16.3.1 Linear Exponential Family

Definition. The distribution of the *linear exponential family* is

$$f(x;\gamma,\theta) = \exp\left(\frac{x\gamma - b(\gamma)}{\theta} + S(x,\theta)\right). \tag{16.3}$$

Here, x is a dependent variable and γ is the parameter of interest. The quantity θ is a scale parameter. The term $b(\gamma)$ depends only on the parameter γ, not the dependent variable. The statistic $S(x,\theta)$ is a function of the dependent variable and the scale parameter, not the parameter γ.

The dependent variable x may be discrete, continuous or a hybrid combination of the two. Thus, $f(.)$ may be interpreted to be a density or mass function, depending on the application. The following table provides several examples, including the normal, binomial and Poisson distributions.

Selected Distributions of the Linear Exponential Family

Distribution	Parameters	Density or Mass Function	Components
General	$\gamma,\ \theta$	$\exp\left(\frac{x\gamma-b(\gamma)}{\theta}+S(x,\theta)\right)$	$\gamma,\ \theta, b(\gamma), S(x,\theta)$
Normal	μ,σ^2	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mu,\sigma^2,\frac{\gamma^2}{2},-\left(\frac{x^2}{2\theta}+\frac{\ln(2\pi\theta)}{2}\right)$
Binomial	π	$\binom{n}{x}\pi^x(1-\pi)^{n-x}$	$\ln\left(\frac{\pi}{1-\pi}\right), 1, n\ln(1+e^\gamma),$ $\ln\binom{n}{x}$
Poisson	λ	$\frac{\lambda^x}{x!}\exp(-\lambda)$	$\ln\lambda, 1, e^\gamma, -\ln(x!)$
Negative Binomial*	r,p	$\frac{\Gamma(x+r)}{x!\Gamma(r)}p^r(1-p)^x$	$\ln(1-p), 1, -r\ln(1-e^\gamma),$ $\ln\left[\frac{\Gamma(x+r)}{x!\Gamma(r)}\right]$
Gamma	α,θ	$\frac{1}{\Gamma(\alpha)\theta^\alpha}x^{\alpha-1}\exp(-x/\theta)$	$-\frac{1}{\alpha\gamma},\frac{1}{\alpha},-\ln(-\gamma),-\theta^{-1}\ln\theta$ $-\ln\left(\Gamma(\theta^{-1})\right)+(\theta^{-1}-1)\ln x$

*This assumes that the parameter r is fixed but need not be an integer.

The Tweedie (see Section **??**) and inverse Gaussian distributions are also members of the linear exponential family. The linear exponential family of distribution families is extensively used as the basis of generalized linear models as described in, for example, Frees (2009b).

16.3.2 Conjugate Distributions

Now assume that the parameter γ is random with distribution $\pi(\gamma,\tau)$, where τ is a vector of parameters that describe the distribution of γ. In Bayesian models, the distribution π is known as the prior and reflects our belief or information about γ. The likelihood $f(x|\gamma)$ is a probability conditional on γ. The distribution of γ with knowledge of the random variables, $\pi(\gamma,\tau|x)$, is called the posterior distribution. For a given likelihood distribution, priors and posteriors that come from the same parametric family are known as conjugate families of distributions.

For a linear exponential likelihood, there exists a natural conjugate family. Specifically, consider a likelihood of the form $f(x|\gamma) = \exp\{(x\gamma - b(\gamma))/\theta\}\exp\{S(x,\theta)\}$. For this likelihood, define the prior distribution

$$\pi(\gamma,\tau) = C\exp\{\gamma a_1(\tau) - b(\gamma)a_2(\tau)\},$$

where C is a normalizing constant. Here, $a_1(\tau) = a_1$ and $a_2(\tau) = a_2$ are functions of the parameters τ although we simplify the notation by dropping explicit dependence on τ. The joint distribution of x and γ is given by $f(x,\gamma) = f(x|\gamma)\pi(\gamma,\tau)$. Using Bayes Theorem, the posterior distribution is

$$\pi(\gamma,\tau|x) = C_1 \exp\left\{\gamma\left(a_1 + \frac{x}{\theta}\right) - b(\gamma)\left(a_2 + \frac{1}{\theta}\right)\right\},$$

where C_1 is a normalizing constant. Thus, we see that $\pi(\gamma,\tau|x)$ has the same form as $\pi(\gamma,\tau)$.

Special case. Poisson-Gamma Model. Consider a Poisson likelihood so that $b(\gamma) = e^{\gamma}$ and scale parameter (θ) equals one. Thus, we have

$$\pi(\gamma,\tau) = C\exp\{\gamma a_1 - a_2 e^{\gamma}\} = C\ \left(e^{\gamma}\right)^{a_1} \exp\left(-a_2 e^{\gamma}\right).$$

From the table of exponential family distributions, we recognize this to be a gamma distribution. That is, we have that the prior distribution of $\lambda = e^{\gamma}$ is a gamma distribution with parameters $\alpha_{prior} = a_1 + 1$ and $\theta^{-1}_{prior} = a_2$. The posterior distribution is a gamma distribution with parameters $\alpha_{post} = a_1 + x + 1 = \alpha_{prior} + x$ and $\theta^{-1}_{post} = a_2 + 1 = \theta^{-1}_{prior} + 1$. This is consistent with our Section **??** result.

Special case. Normal-Normal Model. Consider a normal likelihood so that $b(\gamma) = \gamma^2/2$ and the scale parameter is σ^2. Thus, we have

$$\pi(\gamma,\tau) = C\exp\left\{\gamma a_1 - \frac{\gamma^2}{2}a_2\right\} = C_1(\tau)\exp\left\{-\frac{a_2}{2}\left(\gamma - \frac{a_1}{a_2}\right)^2\right\},$$

The prior distribution of γ is normal with mean a_1/a_2 and variance a_2^{-1}. The posterior distribution of γ given x is normal with mean $(a_1 + x/\sigma^2)/(a_2 + \sigma^{-2})$ and variance $(a_2 + \sigma^{-2})^{-1}$.

Special case. Beta-Binomial Model. Consider a binomial likelihood so that $b(\gamma) = n\ln(1 + e^{\gamma})$ and scale parameter equals one. Thus, we have

$$\pi(\gamma,\tau) = C\exp\{\gamma a_1 - n a_2 \ln(1 + e^{\gamma})\} = C\ \left(\frac{e^{\gamma}}{1+e^{\gamma}}\right)^{a_1}\left(1 - \frac{e^{\gamma}}{1+e^{\gamma}}\right)^{-na_2+a_1}.$$

This is a beta distribution. As in the other cases, prior parameters a_1 and a_2 are updated to become posterior parameters $a_1 + x$ and $a_2 + 1$.

Contributors

- **Lei (Larry) Hua**, Northern Illinois University, and **Edward W. (Jed) Frees**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter. Email: lhua@niu.edu or jfrees@bus.wisc.edu for chapter comments and suggested improvements.

Chapter 17

Appendix C: Maximum Likelihood Theory

Chapter preview. Appendix Chapter 15 introduced the maximum likelihood theory regarding estimation of parameters from a parametric family. This appendix gives more specific examples and expands some of the concepts. Section 17.1 reviews the definition of the likelihood function, and introduces its properties. Section 17.2 reviews the maximum likelihood estimators, and extends their large-sample properties to the case where there are multiple parameters in the model. Section 17.3 reviews statistical inference based on maximum likelihood estimators, with specific examples on cases with multiple parameters.

17.1 Likelihood Function

In this section, you learn

- the definitions of the likelihood function and the log-likelihood function
- the properties of the likelihood function.

From Appendix 15, the likelihood function is a function of parameters given the observed data. Here, we review the concepts of the likelihood function, and introduces its properties that are bases for maximum likelihood inference.

17.1.1 Likelihood and Log-likelihood Functions

Here, we give a brief review of the likelihood function and the log-likelihood function from Appendix 15. Let $f(\cdot|\boldsymbol{\theta})$ be the probability function of X, the probability mass function (pmf) if X is discrete or the probability density function (pdf) if it is continuous. The likelihood is a function of the parameters ($\boldsymbol{\theta}$) given the data ($\mathbf{x}$). Hence, it is a function of the parameters with the data being fixed, rather than a function of the data with the parameters being fixed. The vector of data $\mathbf{x}$ is usually a realization of a *random sample* as defined in Appendix 15.

Given a realized of a random sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ of size n, the **likelihood function** is defined as

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

with the corresponding **log-likelihood function** given by

$$l(\boldsymbol{\theta}|\mathbf{x}) = \ln L(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i|\boldsymbol{\theta}),$$

where $f(\mathbf{x}|\boldsymbol{\theta})$ denotes the joint probability function of $\mathbf{x}$. The log-likelihood function leads to an additive structure that is easy to work with.

In Appendix 15, we have used the normal distribution to illustrate concepts of the likelihood function and the log-likelihood function. Here, we derive the likelihood and corresponding log-likelihood functions when the population distribution is from the Pareto distribution family.

Show Example

Example – Pareto Distribution. Suppose that $X_1, \ldots, X_n$ represents a random sample from a single-parameter Pareto distribution with the **cumulative distribution function** given by

$$F(x) = \Pr(X_i \le x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500,$$

where the parameter $\theta = \alpha$.

The corresponding probability density function is $f(x) = 500^{\alpha} \alpha x^{-\alpha-1}$ and the log-likelihood function can be derived as

$$l(\boldsymbol{\alpha}|\mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i; \alpha) = n\alpha \ln 500 + n \ln \alpha - (\alpha+1) \sum_{i=1}^{n} \ln x_i.$$

17.1.2 Properties of Likelihood Functions

In mathematical statistics, the first derivative of the log-likelihood function with respect to the parameters, $u(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}|\mathbf{x})/\partial \boldsymbol{\theta}$, is referred to as the **score function**, or the **score vector** when there are multiple parameters in $\boldsymbol{\theta}$. The score function or score vector can be written as

$$u(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ln f(x_i; \boldsymbol{\theta}),$$

where $u(\boldsymbol{\theta}) = (u_1(\boldsymbol{\theta}), u_2(\boldsymbol{\theta}), \cdots, u_p(\boldsymbol{\theta}))$ when $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ contains $p > 2$ parameters, with the element $u_k(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}|\mathbf{x})/\partial \theta_k$ being the partial derivative with respect to θ_k $(k = 1, 2, \cdots, p)$.

The likelihood function has the following properties:

- One basic property of the likelihood function is that the expectation of the score function with respect to $\mathbf{x}$ is 0. That is,

$$\mathrm{E}[u(\boldsymbol{\theta})] = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{x})\right] = \mathbf{0}$$

To illustrate this, we have

$$\begin{aligned} \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{x})\right] &= \mathrm{E}\left[\frac{\frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta})}{f(\mathbf{x}; \boldsymbol{\theta})}\right] = \int \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y} = \frac{\partial}{\partial \boldsymbol{\theta}} 1 = \mathbf{0}. \end{aligned}$$

- Denote by $\partial^2 l(\boldsymbol{\theta}|\mathbf{x})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}' = \partial^2 l(\boldsymbol{\theta}|\mathbf{x})/\partial \boldsymbol{\theta}^2$ the second derivative of the log-likelihood function when $\boldsymbol{\theta}$ is a single parameter, or by $\partial^2 l(\boldsymbol{\theta}|\mathbf{x})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}' = (h_{jk}) = (\partial^2 l(\boldsymbol{\theta}|\mathbf{x})/\partial x_j \partial x_k)$ the hessian matrix of the log-likelihood function when it contains multiple parameters. Denote $[\partial l(\boldsymbol{\theta}|\mathbf{x}) \partial \boldsymbol{\theta}][\partial l(\boldsymbol{\theta}|\mathbf{x}) \partial \boldsymbol{\theta}'] = u^2(\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ is a single parameter, or let $[\partial l(\boldsymbol{\theta}|\mathbf{x}) \partial \boldsymbol{\theta}][\partial l(\boldsymbol{\theta}|\mathbf{x}) \partial \boldsymbol{\theta}'] = (uu_{jk})$ be a $p \times p$ matrix when $\boldsymbol{\theta}$ contains a total of p parameters, with each element $uu_{jk} = u_j(\boldsymbol{\theta})u_k(\boldsymbol{\theta})$ and $u_j(\boldsymbol{\theta})$ being the kth element of the score vector as defined earlier. Another basic property of the likelihood function is that sum of the expectation of the hessian matrix and the expectation of the kronecker product of the score vector and its transpose is $\mathbf{0}$. That is,

$$\mathrm{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l(\boldsymbol{\theta}|\mathbf{x})\right) + \mathrm{E}\left(\frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}}\frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}'}\right) = \mathbf{0}.$$

- Define the **Fisher information matrix** as

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}\left(\frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}}\frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}'}\right) = -\mathrm{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l(\boldsymbol{\theta}|\mathbf{x})\right).$$

As the sample size n goes to infinity, the score function (vector) converges in distribution to a **normal distribution** (or **multivariate normal distribution** when $\boldsymbol{\theta}$ contains multiple parameters) with mean $\mathbf{0}$ and variance (or covariance matrix in the multivariate case) given by $\mathcal{I}(\boldsymbol{\theta})$.

17.2 Maximum Likelihood Estimators

In this section, you learn

- the definition and derivation of the maximum likelihood estimator (MLE) for parameters from a specific distribution family
- the properties of maximum likelihood estimators that ensure valid large-sample inference of the parameters
- why using the MLE-based method, and what caution that needs to be taken.

In statistics, maximum likelihood estimators are values of the parameters $\boldsymbol{\theta}$ that are most likely to have been produced by the data.

17.2.1 Definition and Derivation of MLE

Based on the definition given in Appendix 15, the value of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}_{MLE}$, that maximizes the likelihood function, is called the *maximum likelihood estimator* (MLE) of $\boldsymbol{\theta}$.

Because the log function $\ln(\cdot)$ is a one-to-one function, we can also determine $\hat{\boldsymbol{\theta}}_{MLE}$ by maximizing the log-likelihood function, $l(\boldsymbol{\theta}|\mathbf{x})$. That is, the MLE is defined as

$$\hat{\boldsymbol{\theta}}_{MLE} = \mathrm{argmax}_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta}|\mathbf{x}).$$

Given the analytical form of the likelihood function, the MLE can be obtained by taking the first derivative of the log-likelihood function with respect to $\boldsymbol{\theta}$, and setting the values of the partial derivatives to zero. That is, the MLE are the solutions of the equations of

$$\frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial\hat{\boldsymbol{\theta}}} = \mathbf{0}.$$

Show Example

Example. Course C/Exam 4. May 2000, 21. You are given the following five observations: 521, 658, 702, 819, 1217. You use the single-parameter Pareto with cumulative distribution function:

$$F(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500.$$

Calculate the maximum likelihood estimate of the parameter α.

Show Solution

Solution. With $n = 5$, the log-likelihood function is

$$l(\alpha|\mathbf{x}) = \sum_{i=1}^{5} \ln f(x_i; \alpha) = 5\alpha \ln 500 + 5 \ln \alpha - (\alpha + 1) \sum_{i=1}^{5} \ln x_i.$$

Solving for the root of the score function yields

$$\frac{\partial}{\partial \alpha} l(\alpha|\mathbf{x}) = 5 \ln 500 + 5/\alpha - \sum_{i=1}^{5} \ln x_i =_{set} 0 \Rightarrow \hat{\alpha}_{MLE} = \frac{5}{\sum_{i=1}^{5} \ln x_i - 5 \ln 500} = 2.453.$$

17.2.2 Asymptotic Properties of MLE

From Appendix 15, the MLE has some nice large-sample properties, under certain regularity conditions. We presented the results for a single parameter in Appendix 15, but results are true for the case when $\boldsymbol{\theta}$ contains multiple parameters. In particular, we have the following results, in a general case when $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_p)$.

- The MLE of a parameter $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{MLE}$, is a **consistent** estimator. That is, the MLE $\hat{\boldsymbol{\theta}}_{MLE}$ converges in probability to the true value $\boldsymbol{\theta}$, as the sample size n goes to infinity.
- The MLE has the **asymptotic normality** property, meaning that the estimator will converge in distribution to a multivariate normal distribution centered around the true value, when the sample size goes to infinity. Namely,

  $$\sqrt{n}(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}) \rightarrow N(\mathbf{0}, \boldsymbol{V}), \quad \text{as} \quad n \rightarrow \infty,$$

  where $\boldsymbol{V}$ denotes the asymptotic variance (or covariance matrix) of the estimator. Hence, the MLE $\hat{\boldsymbol{\theta}}_{MLE}$ has an approximate normal distribution with mean $\boldsymbol{\theta}$ and variance (covariance matrix when $p > 1$) $\boldsymbol{V}/n$, when the sample size is large.
- The MLE is **efficient**, meaning that it has the smallest asymptotic variance $\boldsymbol{V}$, commonly referred to as the **Cramer–Rao lower bound**. In particular, the Cramer–Rao lower bound is the inverse of the Fisher information (matrix) $\mathcal{I}(\boldsymbol{\theta})$ defined earlier in this appendix. Hence, $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{MLE})$ can be estimated based on the observed Fisher information.

Based on the above results, we may perform statistical inference based on the procedures defined in Appendix 15.

Show Example

Example. Course C/Exam 4. Nov 2000, 13. A sample of ten observations comes from a parametric family $f(x,; \theta_1, \theta_2)$ with log-likelihood function

$$l(\theta_1, \theta_2) = \sum_{i=1}^{10} f(x_i; \theta_1, \theta_2) = -2.5\theta_1^2 - 3\theta_1\theta_2 - \theta_2^2 + 5\theta_1 + 2\theta_2 + k,$$

where k is a constant. Determine the estimated covariance matrix of the maximum likelihood estimator, $\hat{\theta}_1, \hat{\theta}_2$.

Show Solution

Solution. Denoting $l = l(\theta_1, \theta_2)$, the hessian matrix of second derivatives is

$$\left(\begin{array}{cc} \frac{\partial^2}{\partial \theta_1^2} l & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} l \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} l & \frac{\partial^2}{\partial \theta_1^2} l \end{array} \right) = \left(\begin{array}{cc} -5 & -3 \\ -3 & -2 \end{array} \right)$$

Thus, the information matrix is:

$$\mathcal{I}(\theta_1, \theta_2) = -\mathrm{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} l(\boldsymbol{\theta}|\mathbf{x}) \right) = \left(\begin{array}{cc} 5 & 3 \\ 3 & 2 \end{array} \right)$$

and

$$\mathcal{I}^{-1}(\theta_1, \theta_2) = \frac{1}{5(2) - 3(3)} \left(\begin{array}{cc} 2 & -3 \\ -3 & 5 \end{array} \right) = \left(\begin{array}{cc} 2 & -3 \\ -3 & 5 \end{array} \right).$$

17.2.3 Use of Maximum Likelihood Estimation

The method of maximum likelihood has many advantages over alternative methods such as the method of moment method introduced in Appendix 15.

- It is a general tool that works in many situations. For example, we may be able to write out the closed-form likelihood function for censored and truncated data. Maximum likelihood estimation can be used for regression models including covariates, such as survival regression, generalized linear models and mixed models, that may include covariates that are time-dependent.
- From the efficiency of the MLE, it is optimal, the best, in the sense that it has the smallest variance among the class of all unbiased estimators for large sample sizes.
- From the results on the asymptotic normality of the MLE, we can obtain a large-sample distribution for the estimator, allowing users to assess the variability in the estimation and perform statistical inference on the parameters. The approach is less computationally extensive than re-sampling methods that require a large of fittings of the model.

Despite its numerous advantages, MLE has its drawback in cases such as generalized linear models when it does not have a closed analytical form. In such cases, maximum likelihood estimators are computed iteratively using numerical optimization methods. For example, we may use the Newton-Raphson iterative algorithm or its variations for obtaining the MLE. Iterative algorithms require starting values. For some problems, the choice of a close starting value is critical, particularly in cases where the likelihood function has local minimums or maximums. Hence, there may be a convergence issue when the starting value is far from the maximum. Hence, it is important to start from different values across the parameter space, and compare the maximized likelihood or log-likelihood to make sure the algorithms have converged to a global maximum.

17.3 Statistical Inference Based on Maximum Likelihood Estimation

In this section, you learn how to

- perform hypothesis testing based on MLE for cases where there are multiple parameters in $\boldsymbol{\theta}$
- perform likelihood ratio test for cases where there are multiple parameters in $\boldsymbol{\theta}$

In Appendix 15, we have introduced maximum likelihood-based methods for statistical inference when $\boldsymbol{\theta}$ contains a single parameter. Here, we will extend the results to cases where there are multiple parameters in $\boldsymbol{\theta}$.

17.3.1 Hypothesis Testing

In Appendix 15, we defined hypothesis testing concerning the null hypothesis, a statement on the parameter(s) of a distribution or model. One important type of inference is to assess whether a parameter estimate is statistically significant, meaning whether the value of the parameter is zero or not.

We have learned earlier that the MLE $\hat{\boldsymbol{\theta}}_{MLE}$ has a large-sample normal distribution with mean $\boldsymbol{\theta}$ and the variance covariance matrix $\mathcal{I}^{-1}(\boldsymbol{\theta})$. Based on the multivariate normal distribution, the jth element of $\hat{\boldsymbol{\theta}}_{MLE}$, say $\hat{\theta}_{MLE,j}$, has a large-sample univariate normal distribution.

Define $se(\hat{\theta}_{MLE,j})$, the standard error (estimated standard deviation) to be the square root of the jth diagonal element of $\mathcal{I}^{-1}(\boldsymbol{\theta})_{MLE}$. To assess the null hypothesis that $\theta_j = \theta_0$, we define the t-statistic or t-ratio to be $t(\hat{\theta}_{MLE,j}) = (\hat{\theta}_{MLE,j} - \theta_0)/se(\hat{\theta}_{MLE,j})$.

Under the null hypothesis, it has a Student-t distribution with degrees of freedom equal to $n - p$, with p being the dimension of $\boldsymbol{\theta}$.

For most actuarial applications, we have a large sample size n, so the t-distribution is very close to the (standard) normal distribution. In the case when n is very large or when the standard error is known, the t-statistic can be referred to as a z-statistic or z-score.

Based on the results from Appendix 15, if the t-statistic $t(\hat{\theta}_{MLE,j})$ exceeds a cut-off (in absolute value), then the test for the j parameter θ_j is said to be statistically significant. If θ_j is the regression coefficient of the j th independent variable, then we say that the jth variable is statistically significant.

For example, if we use a 5% significance level, then the cut-off value is 1.96 using a normal distribution approximation for cases with a large sample size. More generally, using a $100\alpha\%$ significance level, then the cut-off is a $100(1-\alpha/2)\%$ quantile from a Student-t distribution with the degree of freedom being $n - p$.

Another useful concept in hypothesis testing is the p-value, shorthand for probability value. From the mathematical definition in Appendix 15, a p-value is defined as the smallest significance level for which the null hypothesis would be rejected. Hence, the p-value is a useful summary statistic for the data analyst to report because it allows the reader to understand the strength of statistical evidence concerning the deviation from the null hypothesis.

17.3.2 MLE and Model Validation

In addition to hypothesis testing and interval estimation introduced in Appendix 15 and the previous subsection, another important type of inference is selection of a model from two choices, where one choice is a special case of the other with certain parameters being restricted. For such two models with one being nested in the other, we have introduced the likelihood ratio test (LRT) in Appendix 15. Here, we will briefly review the process of performing a LRT based on a specific example of two alternative models.

Suppose that we have a (large) model under which we derive the maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_{MLE}$. Now assume that some of the p elements in $\boldsymbol{\theta}$ are equal to zero and determine the maximum likelihood estimator over the remaining set, with the resulting estimator denoted $\hat{\boldsymbol{\theta}}_{Reduced}$.

Based on the definition in Appendix 15, the statistic, $LRT = 2\left(l(\hat{\boldsymbol{\theta}}_{MLE}) - l(\hat{\boldsymbol{\theta}}_{Reduced})\right)$, is called the likelihood ratio statistic. Under the null hypothesis that the reduced model is correct, the likelihood ratio has a Chi-square distribution with degrees of freedom equal to d, the number of variables set to zero.

Such a test allows us to judge which of the two models is more likely to be correct, given the observed data. If the statistic LRT is large relative to the critical value from the chi-square distribution, then we reject the reduced model in favor of the larger one. Details regarding the critical value and alternative methods based on information criteria are given in Appendix 15.

Contributors

- **Lei (Larry) Hua**, Northern Illinois University, and **Edward W. (Jed) Frees**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter. Email: lhua@niu.edu or jfrees@bus.wisc.edu for chapter comments and suggested improvements.

Bibliography

Aalen, Odd (1978). “Nonparametric inference for a family of counting processes,” *The Annals of Statistics*, Vol. 6, pp. 701–726.

Abbott, Dean (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*, Hoboken, NJ. Wiley.

Abdullah, Mohammad F. and Kamsuriah Ahmad (2013). “The mapping process of unstructured data to structured data,” in *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*, pp. 151–155.

Actuarial Standards Board (2018). “Actuarial Standards of Practice,” American Academy of Actuaries, URL: http://www.actuarialstandardsboard.org/standards-of-practice/, [Retrieved on Oct 3, 2018].

Aggarwal, Charu C. (2015). *Data Mining: The Textbook*, New York, NY. Springer.

Agresti, Alan (1996). *An Introduction to Categorical Data Analysis*, Vol. 135. Wiley New York.

Albers, Michael J. (2017). *Introduction to Quantitative Data Analysis in the Behavioral and Social Sciences*, Hoboken, NJ. John Wiley & Sons, Inc.

Bahnemann, David (2015). “Distributions for Actuaries,” *CAS Monograph Series*, URL: https://www.casact.org/pubs/monographs/papers/02-Bahnemann.pdf.

Bailey, Robert A. and J. Simon LeRoy (1960). “Two studies in automobile ratemaking,” *Proceedings of the Casualty Actuarial Society Casualty Actuarial Society*, Vol. XLVII.

Bandyopadhyay, Prasanta S. and Malcolm R. Forster eds. (2011). *Philosophy of Statistics*, Handbook of the Philosophy of Science 7. North Holland.

Billingsley, Patrick (2008). *Probability and measure*. John Wiley & Sons.

Bishop, Christopher M. (2007). *Pattern Recognition and Machine Learning*, New York, NY. Springer.

Bishop, Yvonne M., Stephen E. Fienberg, and Paul W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge [etc.]: MIT.

Blomqvist, Nils (1950). “On a measure of dependence between two random variables,” *The Annals of Mathematical Statistic*, pp. 593–600.

Bluman, Allan (2012). *Elementary Statistics: A Step By Step Approach*, New York, NY. McGraw-Hill.

Bowers, Newton L., Hans U. Gerber, James C. Hickman, Donald A. Jones, and Cecil J. Nesbitt (1986). *Actuarial Mathematics*. Society of Actuaries Itasca, Ill.

Box, George EP (1980). “Sampling and Bayes’ inference in scientific modelling and robustness,” *Journal of the Royal Statistical Society. Series A (General)*, pp. 383–430.

Breiman, Leo (2001). “Statistical Modeling: The Two Cultures,” *Statistical Science*, Vol. 16, pp. 199–231.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen (1984). *Classification and Regression Trees*, Raton Boca, FL. Chapman and Hall/CRC.

Buhlmann, Hans (1967). "The Complement of Credibility," pp. 199–207.

Buhlmann, Hans and Alois Gisler (2005). *A Course in Credibility Theory and its Applications.* ACTEX Publications.

Buttrey, Samuel E. and Lyn R. Whitaker (2017). *A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R*, Hoboken, NJ. Wiley.

Chen, Min, Shiwen Mao, Yin Zhang, and Victor CM Leung (2014). *Big Data: Related Technologies, Challenges and Future Prospects*, New York, NY. Springer.

Clark, David R (1996). "Basics of reinsurance pricing," *CAS Study Note*, pp. 41–43, URL: https://www.soa.org/files/edu/edu-2014-exam-at-study-note-basics-rein.pdf.

Clarke, Bertrand, Ernest Fokoue, and Hao Helen Zhang (2009). *Principles and theory for data mining and machine learning*, New York, NY. Springer-Verlag.

Cummins, J. David and Richard A. Derrig (2012). *Managing the Insolvency Risk of Insurance Companies: Proceedings of the Second International Conference on Insurance Solvency*, Vol. 12. Springer Science & Business Media.

Dabrowska, Dorota M. (1988). "Kaplan-meier estimate on the plane," *The Annals of Statistics*, pp. 1475–1489.

Daroczi, Gergely (2015). *Mastering Data Analysis with R*, Birmingham, UK. Packt Publishing.

De Jong, Piet and Gillian Z. Heller (2008). *Generalized Linear Models for Insurance Data.* Cambridge University Press Cambridge.

Dickson, David C. M., Mary Hardy, and Howard R. Waters (2013). *Actuarial Mathematics for Life Contingent Risks.* Cambridge University Press.

Dobson, Annette J and Adrian Barnett (2008). *An Introduction to Generalized Linear Models.* CRC press.

Earnix (2013). "2013 Insurance Predictive Modeling Survey," Earnix and Insurance Services Office, Inc. URL: http://earnix.com/2013-insurance-predictive-modeling-survey/3594/, [Retrieved on July 7, 2014].

Faraway, Julian J (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Vol. 124. CRC press.

Fechner, G. T (1897). "Kollektivmasslehre," *Wilhelm Englemann, Leipzig.*

Forte, Rui Miguel (2015). *Mastering Predictive Analytics with R*, Birmingham, UK. Packt Publishing.

Frees, Edward W (2009a). *Regression modeling with actuarial and financial applications.* Cambridge University Press.

Frees, Edward W. (2009b). *Regression Modeling with Actuarial and Financial Applications.* Cambridge University Press.

——— (2009c). *Regression Modeling with Actuarial and Financial Applications.* Cambridge University Press.

Frees, Edward W, Gee Lee, and Lu Yang (2016). "Multivariate frequency-severity regression models in insurance," *Risks*, Vol. 4, p. 4.

Frees, Edward W., Glenn Meyers, and A. David Cummings (2011). "Summarizing insurance scores using a Gini index," *Journal of the American Statistical Association*, Vol. 106, pp. 1085–1098.

——— (2014). "Insurance ratemaking and a Gini index," *Journal of Risk and Insurance*, Vol. 81, pp. 335–366.

Frees, Edward W. and Emiliano A. Valdez (1998). "Understanding relationships using copulas," *North American Actuarial Journal*, Vol. 2, pp. 1–25.

——— (2008). "Hierarchical insurance claims modeling," *Journal of the American Statistical Association*, Vol. 103, pp. 1457–1469.

Gan, Guojun (2011). *Data Clustering in C++: An Object-Oriented Approach*, Data Mining and Knowledge Discovery Series, Boca Raton, FL, USA. Chapman & Hall/CRC Press, DOI: http://dx.doi.org/10.1201/b10814.

Gan, Guojun, Chaoqun Ma, and Jianhong Wu (2007). *Data Clustering: Theory, Algorithms, and Applications*, Philadelphia, PA. SIAM Press, DOI: http://dx.doi.org/10.1137/1.9780898718348.

Gelman, Andrew (2004). "Exploratory Data Analysis for Complex Models," *Journal of Computational and Graphical Statistics*, Vol. 13, pp. 755–779.

Genest, Christian and Josh Mackay (1986). "The joy of copulas: Bivariate distributions with uniform marginals," *The American Statistician*, Vol. 40, pp. 280–283.

Genest, Christian and Johanna Nešlohva (2007). "A primer on copulas for count data," *Journal of the Royal Statistical Society*, pp. 475–515.

Good, I. J. (1983). "The Philosophy of Exploratory Data Analysis," *Philosophy of Science*, Vol. 50, pp. 283–295.

Gorman, Mark and Stephen Swenson (2013). "Building believers: How to expand the use of predictive analytics in claims," SAS, URL: http://www.sas.com/resources/whitepaper/wp_59831.pdf, [Retrieved on August 17, 2014].

Greenwood, Major (1926). "The errors of sampling of the survivorship tables," in *Reports on Public Health and Statistical Subjects*, Vol. 33. London: Her Majesty's Stationary Office.

Hardy, Mary R. (2006). "An introduction to risk measures for actuarial applications.."

Hartman, Brian (2016). "Bayesian Computational Methods," *Predictive Modeling Applications in Actuarial Science: Volume 2, Case Studies in Insurance*.

Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan (2015). "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, Vol. 47, pp. 98 – 115.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*. Wiley.

Hofert, Marius, Ivan Kojadinovic, Martin Mächler, and Jun Yan (2018). *Elements of Copula Modeling with R*. Springer.

Hougaard, P (2000). *Analysis of Multivariate Survival Data*. Springer New York.

Hox, Joop J. and Hennie R. Boeije (2005). "Data collection, primary versus secondary," in *Encyclopedia of social measurement*. Elsevier, pp. 593 – 599.

Igual, Laura and Santi Segu (2017). *Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications*, New York, NY. Springer.

Inmon, W.H. and Dan Linstedt (2014). *Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault*, Cambridge, MA. Morgan Kaufmann.

Insurance Information Institute (2016). "International Insurance Fact Book," Insurance Information Institute, URL: http://www.iii.org/sites/default/files/docs/pdf/international_insurance_factbook_2016.pdf, [Retrieved on Sept 9, 2018].

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An introduction to statistical learning*, Vol. 112. Springer.

Janert, Philipp K. (2010). *Data Analysis with Open Source Tools*, Sebastopol, CA. O'Reilly Media.

Joe, Harry (2014). *Dependence Modeling with Copulas*. CRC Press.

de Jong, Piet and Gillian Z. Heller (2008). *Generalized linear models for insurance data*, Cambridge, UK. Cambridge University Press.

Judd, Charles M., Gary H. McClelland, and Carey S. Ryan (2017). *Data Analysis. A Model Comparison Approach to Regression, ANOVA and beyond*, New York, NY. Routledge, 3rd edition.

Kaplan, Edward L. and Paul Meier (1958). "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, Vol. 53, pp. 457–481.

Kendall, Maurice G (1938). "A new measure of rank correlation," *Biometrika*, pp. 81–93.

Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot (2012). *Loss Models: From Data to Decisions*. John Wiley & Sons.

Kreer, Markus, Ayşe Kızılersü, Anthony W Thomas, and Alfredo D Egídio dos Reis (2015). "Goodness-of-fit tests and applications for left-truncated Weibull distributions to non-life insurance," *European Actuarial Journal*, Vol. 5, pp. 139–163.

Kubat, Miroslav (2017). *An Introduction to Machine Learning*, New York, NY. Springer, 2nd edition.

Lee Rodgers, J and W. A Nicewander (1998). "Thirteen ways to look at the correlation coeffeicient," *The American Statistician*, Vol. 42, pp. 59–66.

Levin, Bruce, James Reeds et al. (1977). "Compound multinomial likelihood functions are unimodal: Proof of a conjecture of IJ Good," *The Annals of Statistics*, Vol. 5, pp. 79–87.

Lorenz, Max O. (1905). "Methods of measuring the concentration of wealth," *Publications of the American statistical association*, Vol. 9, pp. 209–219.

Mailund, Thomas (2017). *Beginning Data Science in R: Data Analysis, Visualization, and Modelling for the Data Scientist*. Apress.

McCullagh, Peter and John A. Nelder (1989). *Generalized linear models*, Vol. 37. CRC press.

McDonald, James B (1984). "Some generalized functions for the size distribution of income," *Econometrica: journal of the Econometric Society*, pp. 647–663.

McDonald, James B and Yexiao J Xu (1995). "A generalization of the beta distribution with applications," *Journal of Econometrics*, Vol. 66, pp. 133–152.

Miles, Matthew, Michael Hberman, and Johnny Sdana (2014). *Qualitative Data Analysis: A Methods Sourcebook*, Thousand Oaks, CA. Sage, 3rd edition.

Mirkin, Boris (2011). *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*, London, UK. Springer.

Mitchell, Tom M. (1997). *Machine Learning*. McGraw-Hill.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2012). *Foundations of Machine Learning*, Cambridge, MA. MIT Press.

NAIC Glossary (2018). "Glossary of Insurance Terms," National Association of Insurance Commissioners, URL: https://www.naic.org/consumer_glossary.htm, [Retrieved on Sept 11, 2018].

Nelson, Roger B. (1997). *An Introduction to Copulas.* Lecture Notes in Statistics 139.

Ohlsson, Esbjörn and Björn Johansson (2010). *Non-life Insurance Pricing with Generalized Linear Models*, Vol. 21. Springer.

O'Leary, D. E. (2013). "Artificial Intelligence and Big Data," *IEEE Intelligent Systems*, Vol. 28, pp. 96–99.

Olkin, Ingram, A John Petkau, and James V Zidek (1981). "A comparison of n estimators for the binomial distribution," *Journal of the American Statistical Association*, Vol. 76, pp. 637–642.

Olson, Jack E. (2003). *Data Quality: The Accuracy Dimension*, San Francisco, CA. Morgan Kaufmann.

Picard, Richard R. and Kenneth N. Berk (1990). "Data splitting," *The American Statistician*, Vol. 44, pp. 140–147.

Pries, Kim H. and Robert Dunnigan (2015). *Big Data Analytics: A Practical Guide for Managers*, Boca Raton, FL. CRC Press.

Samuel, A. L. (1959). "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, Vol. 3, pp. 210–229.

Shmueli, Galit (2010). "To Explain or to Predict?" *Statistical Science*, Vol. 25, pp. 289–310.

Snee, Ronald D. (1977). "Validation of regression models: methods and examples," *Technometrics*, Vol. 19, pp. 415–428.

Spearman, C (1904). "The proof and measurement of association between two things," *The American Journal of Psychology*, Vol. 15, pp. 72–101.

Tevet, Dan (2016). "Applying Generalized Linear Models to Insurance Data," *Predictive Modeling Applications in Actuarial Science: Volume 2, Case Studies in Insurance*, p. 39.

Tse, Yiu-Kuen (2009). *Nonlife Actuarial Models: Theory, Methods and Evaluation.* Cambridge University Press.

Tukey, John W. (1962). "The Future of Data Analysis," *The Annals of Mathematical Statistics*, Vol. 33, pp. 1–67.

Venter, Gary (1983). "Transformed beta and gamma distributions and aggregate losses," in *Proceedings of the Casualty Actuarial Society*, Vol. 70, pp. 289–308.

Venter, Gary G. (2002). "Tails of copulas," in *Proceedings of the Casualty Actuarial Society*, Vol. 89, pp. 68–113.

Yule, G. Udny (1900). "On the association of attributes in statistics: with illustrations from the material of the childhood society," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pp. 257–319.

——— (1912). "On the methods of measuring association between two attributes," *Journal of the Royal Statistical Society*, pp. 579–652.