Loss Data Analytics

An open text authored by the Actuarial Community

2018-04-08

Contents

Preface

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?

Although the intent is that this type of resource will eventually permeate throughout the actuarial curriculum, one has to start somewhere. 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 seek to recognize that big data (including social media and usage based insurance) are here and high speed computation s 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 Loss Data Analytics Project Site.

Contributor List

- **Zeinab Amin** American University in Cairo
- **Katrien Antonio**, KU Leuven
- **Jan Beirlant**, KU Leuven
- **Carolina Castro** - University of Buenos Aires
- **Gary Dean**, Ball State University
- **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** - University of Connecticut
- **Lisa Gao** is a doctoral student at the University of Wisconsin-Madison.
- **José Garrido**, Concordia University
- **Noriszura Ismail**, University Kebangsaan Malaysia
- **Joseph Kim**, Yonsei University
- **Shyamalkumar Nariankadu** - University of Iowa
- **Nii-Armah Okine** is a doctoral student at the University of Wisconsin-Madison.
- **Emine Selin Sarıdaş**, Mimar Sinan University
- **Peng Shi** - University of Wisconsin
- **Jianxi Su**, Purdue University
- **Tim Verdonck**, KU Leuven
- **Krupa Viswanathan** - Temple University

Acknowledgements

Reviewer Acknowledgment

- Hirokazu (Iwahiro) Iwasawa

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. Although obvious, the importance of data is critical - it is the whole premise of the book. Next, 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 these broads goals 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

In this section, you learn how to:

- Motivate the relevance of insurance
- Describe analytics
- Describe data generating events associated with the timeline of a typical insurance contract

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. Insurance may not be as entertaining as the sports industry nor as widely familiar as the agricultural industry but it does affect the financial livelihoods of many. By almost any measure, insurance is a major economy activity. On a global level, insurance premiums comprised about 6.3% of the world gross domestic product (GDP) in 2013, (Insurance Information Institute, 2015). To illustrate, premiums accounted for 17.6% of GDP in Taiwan (the highest in the study) and represented 7.5% 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. So, although not particulary entertaining nor widely familiar, insurance is an important piece of the economy and relevant to individual livelihoods.

Insurance is a data-driven industry. Like other major corporations, insurers use data when trying to decide how much to pay employees, how many employees to retain, how to market their services, how to forecast financial trends, and so on. Although each industry retains its own nuances, these represent general areas of activities that are not specific to the insurance industry. You will find that the data methods and tools introduced in this text relevant for these general areas.

Moreover, when introducing data methods, we will focus on losses that potentially arise from 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 will call these insurance claims or loss amounts. With this focus, we will be able to introduce generally applicable statistical tools in techniques in real-life situations where the tools can be used directly.

1.1.1 What is Analytics?

Insurance is a data-driven industry and analytics is a key to deriving 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 used separately, referring to distinct domains of applications. As an example of such distinctions, 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 applications in many scientific domains. For our purposes, we use the term analytics to refer to the process of using data to make decisions. This process involves gathering data, understanding models of uncertainty, making general inferences, and communicating results.

1.1.2 Short-term Insurance

This text will focus on short-term insurance contracts. By short-term, we mean contracts where the insurance coverage is typically provided for six months or a year. 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 insurance) or the contents and property of a building that is owned by you or a friend (known as homeowners insurance). Another easy example is automobile insurance. 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.

In the US, policies such as renters and homeowners are known as property insurance whereas a policy such as auto that covers medical damages to people is known as casualty insurance. In the rest of the world, these are both known as nonlife or general insurance, to distinguish them from life insurance.

Both life and nonlife insurances are important. To illustrate, (Insurance Information Institute, 2015) estimates that direct insurance premiums in the world for 2013 was 2,608,091 for life and 2,032,850 for nonlife; these figures are in millions of US dollars. As noted earlier, the total represents 6.3% of the world GDP. Put another way, life accounts for 56.2% of insurance premiums and 3.5% of world GDP, nonlife accounts for 43.8% of insurance premiums and 2.7% of world GDP. Both life and nonlife represent important economic activities and are worthy of study in their own right.

Yet, life insurance considerations differ from nonlife. In life insurance, 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. We think of this as a long-term contract.

Further, in life insurance, the benefit amount is often stipulated in the contract provisions. In contrast, most short-term contracts provide for reimbursement of insured losses which are unknown before the accident. (Of course, there are usually limits placed on the reimbursement amounts.) In a multi-year life insurance contract, the time value of money plays a prominent role. In contrast, in a short-term nonlife contract, the random amount of reimbursement takes priority.

In both life and nonlife 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 nonlife 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 than you may have seen 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., six months, 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.

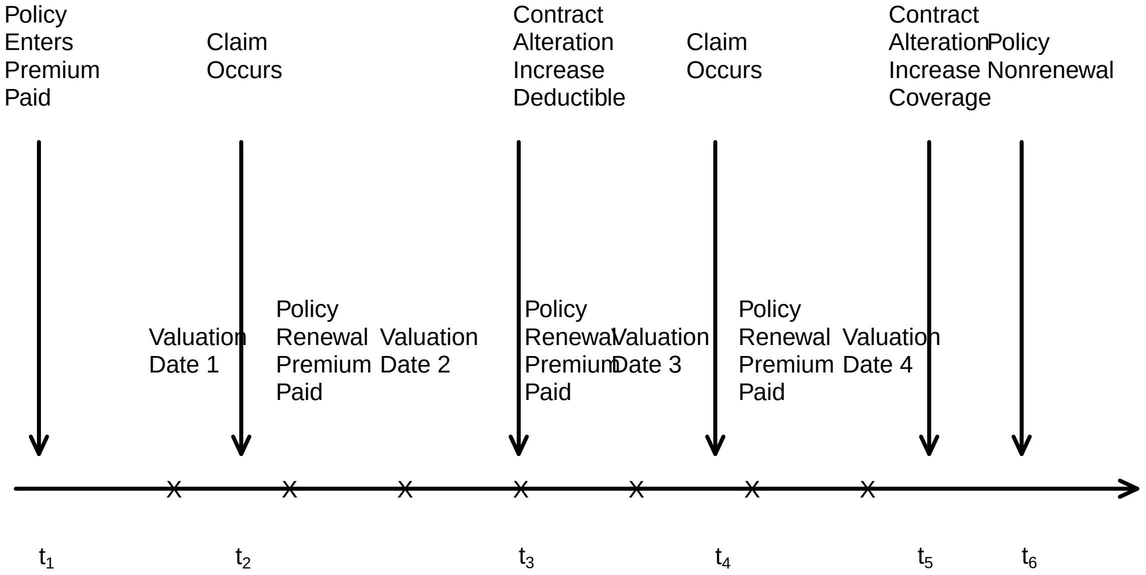

Policy Enters Premium Paid
Claim Occurs
Contract Alteration Increase Deductible
Claim Occurs
Contract Alteration Increase Coverage
Policy Nonrenewal
Valuation Date 1
Policy Renewal Premium Paid
Valuation Date 2
Policy Renewal Premium Paid
Valuation Date 3
Policy Renewal Premium Paid
Valuation Date 4
t1
t2
t3
t4
t5
t6

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.

- At the end of the period, the insurer reimburses the insured for a (possibly multivariate) random loss that we will denote as y.

This framework will be developed as we proceed but we first focus on integrating this framework with concerns about how the data may arise and what we can accomplish with this framework.

1.1.3 Insurance Processes

One way to describe the data arising from operations of a company that sells insurance products is to adopt a granular approach. In this micro oriented view, we can think specifically about what happens to a contract at various stages of its existence. Consider Figure 1.1 that traces a timeline of a typical insurance contract. Throughout the existence 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. Further, non-regular and unanticipated events also occur. To illustrate, times 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 the events when a policyholder wishes to alter certain contract features, such as the choice of a deductible or the amount of coverage. Moreover, 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.

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 and a method of organizing the data into variable types, the end goal is to use data to make decisions. Of course, we will need to learn more about methods of analyzing and extrapolating data but that is the purpose of the remaining chapters in the text. To begin, let us think about why we wish to do the analysis. To provide motivation, we take the insurer's viewpoint (not a person) and introduce ways of bringing money in, paying it out, managing costs, and making sure that we have enough money to meet obligations.

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 and other financial analysts 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 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 the estimates.

5. **Solvency and Capital Allocation**. Deciding on the requisite amount of capital and ways of allocating capital to alternative investment activities represent other 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. 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 is a critical component of solvency and capital allocation, other components including an economic framework and financial investments environment are also important. Because of the background needed to address these components, we will not address solvency and capital allocation issues further 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 good can be a perplexing problem. In manufacturing, the cost of a good is (relatively) known and provides a benchmark for assessing a market demand price. In other areas of financial services, market prices are available and provide the basis for a market-consistent pricing structure of products. In contrast, for many lines of insurance, the cost of a good is uncertain and market prices are unavailable. Expectations of the random cost is a reasonable place to start for a price, as this is the optimal price for a risk-neutral insurer. Thus, it has been traditional in insurance pricing to begin with the expected cost and to add to this so-called margins to account for the product's riskiness, expenses incurred in servicing the product, and a profit/surplus allowance for the insurance company.

For some lines of business, especially automobile and homeowners insurance, analytics has served to sharpen the market by making the calculation of the good's expectation more precise. The increasing availability of the internet among consumers has promoted transparency in pricing. Insurers seek to increase their market share by refining their risk classification systems and employing skimming the cream underwriting strategies. 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 homogenous categories and assigning policyholders to these categories, lies at the core of ratemaking. Policyholders within a class 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. To illustrate, an early contribution, Two Studies in Automobile Insurance Ratemaking, by (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, such as six months or a year, at which time 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. Renewal issues can also arise in life insurance, e.g., term (temporary) life insurance, although other contracts, such as life annuities, terminate upon the insured's death and so issues of renewability are irrelevant.

In 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, as 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 premiums are common in life insurance arrangements (where policyholders earned dividends in the U.S. and bonuses in the U.K.). 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 over the last three years. Claims history can reveal information that is 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 as the benefit amount is provided in the contract terms. However, in non-life areas such as property and casualty insurance, the process is much more complex. Think about even a relatively simple insured event such as automobile accident. Here, it is often helpful 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, and so forth. The process of determining coverage, legal liability, and settling claims is known as claims adjustment.

Insurance managers sometimes use the phrase claims leakage to mean dollars lost through claims management inefficiencies. There are many ways in which analytics can help manage the claims process, (Gorman and Swenson, 2013). Historically, the most important has been fraud detection. The claim adjusting process involves reducing information asymmetry (the claimant knows exactly what happened; the company knows some of what happened). Mitigating fraud is an important part of claims management process.

One can think about the management of claims severity 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 company 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 those 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 situations suitable for small claims handling processes and those for adjuster assignment to complex claims.
- **Adjustment decisions**. Once a complex claim has been identified and assigned to an adjuster, 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 important input to the insurer's loss reserves, Section 1.2.4.

In addition to the insured's reimbursement for insured losses, the insurer also needs to be concerned with another source of revenue outflow, expenses. Loss adjustment expenses are part of an insurer's cost of managing claims. Analytics can be used to reduce expenses directly related to claims handling (allocated) as well as general staff time for overseeing the claims processes (unallocated). 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 to data to manage their products. We have already discussed the need for analytics in underwriting, that is, risk classification at the initial acquisition stage. 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. In Chapter xx introduces common modifications including coinsurance, deductibles, and policy upper limits.

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. This leads to the need to hold a reservoir of wealth to meet future obligations in respect to obligations made. The size of this reservoir of wealth, and the importance of ensuring its adequacy in regard to liabilities already assumed, is a major concern for the insurance industry.

Setting aside money for unpaid claims is known as loss reserving; in some jurisdictions, reserves are also known as technical provisions. We saw in Figure 1.1 how future obligations arise naturally at a specific (valuation) date; 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 a loan 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 and lose market share.
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 of insurers.
4. In addition to the 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 concern 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 death 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. See, for example, (Bowers et al., 1986) or (Dickson et al., 2013) for introductions to reserving for life insurance.

1.3 Case Study: Wisconsin Property Fund

In this section, for a real case study such as the Wisconsin Property Fund, you learn how to:

- Describe how data generating events can produce data of interest to insurance analysts.
- Identify the type of each variable.
- Produce relevant summary statistics for each variable.
- Describe how these summary statistcs 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 property 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.

1.3.1 Fund Claims Variables

At a fundamental level, insurance companies accept premiums in exchange for promises to indemnify a policyholder upon the uncertain occurrence of an insured event. This indemnification is known as a claim. A positive amount, also known as the severity of the claim, is a key financial expenditure for an insurer. So, knowing only the claim amount summarizes the reimbursement to the policyholder.

Ignoring expenses, an insurer that examines only amounts paid would be indifferent to two claims of 100 when compared to one claim of 200, even though the number of claims differ. 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 written on a per occurrence basis, is routinely monitored by insurance regulators, and is often a key driven in the overall indemnification obligation of the insurer. We shall consider the two claims variables, the severity and frequency, as the two main outcome variables that we wish to understand, model, and manage.

To illustrate, in 2010 there were 1,110 policyholders in the property fund. Table 1.1 shows the distribution of the 1,377 claims. Almost two-thirds (0.637) of the policyholders did not have any claims and an additional 18.8% only had 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
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, one 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. The total severity divided by the number of claims is also known as the pure premium or loss cost.

Table 1.2 summarizes the sample distribution of average severities from the 403 policyholders; it shows that the average claim amount was 56,330 (all amounts are in US 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 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

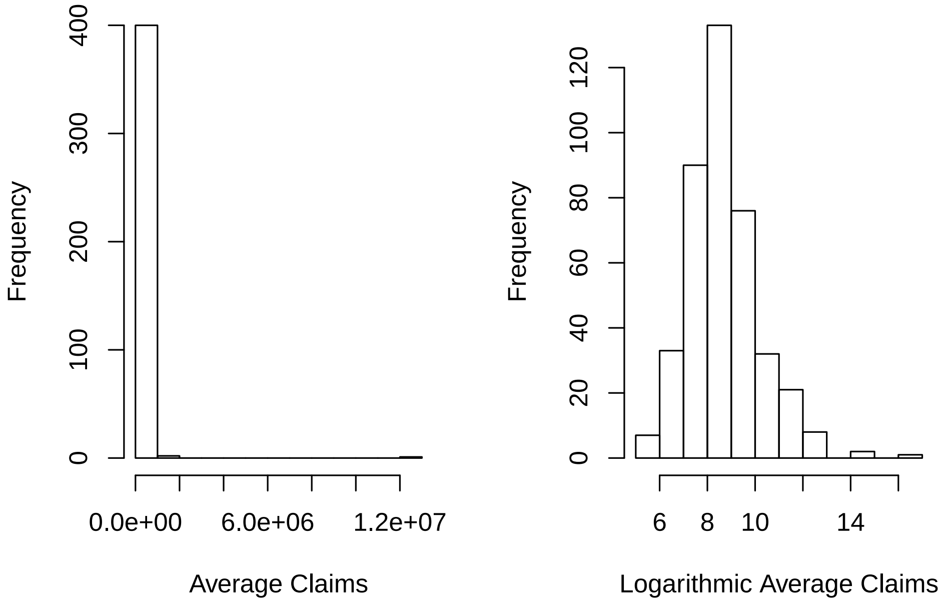


Figure 1.2: Distribution of Positive Average Severities

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=FALS
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 externally available 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 as they will be useful in setting insurance rates and premiums.

We earlier considered a sample of 1,110 observations which may seem like a lot. However, as we will seen 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, just want to show the reader how it works.

Specifically, Table 1.3 shows that we now consider policies over five years of data, years 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 all of the obligations are not 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 shows that the average claim varies over time, especially with the high 2010 value due to a single large claim. 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, it is our first rating variable. Other things being equal, we would expect that policyholders with larger coverage will have larger claims. We will make this vague idea much more precise as we proceed.

Table 1.3: Building and Contents Claims Summary

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 Building and Contents Claims Summary

```
Insample <- read.csv("Data/PropertyFundInsample.csv", header=T, na.strings=c("."), stringsAsFactors=FALS
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
```

For a different look at this five-year sample, 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=FALS
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. To handle the skewness, we henceforth focus on logarithmic transformations of coverage and deductibles. 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.

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

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
Fire5=0	2,508	0.502	13,935
Fire5=1	3,131	1.596	41,421
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 25,257 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 Category

Entity Type	Claim Frequency	Avg. Severity	Num. Policies	Claim Frequency	Avg. Severity	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 Category

Entity Type	Claim Frequency	Avg. Severity	Num. Policies	Claim Frequency	Avg. Severity	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, discrete quantitative variable, (logarithmic) deductibles, two binary rating variable, no claims credit and fire class, as well as 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 × 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 6 on premium calculation fundamentals. Selecting appropriate risks is introduced in Chapter 7 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 1.3 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 8 on experience rating.

Claims Management

Of course, the main story line of 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 suggestions opportunities for managing claims, the subject of Chapter 9.

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 were there are injuries to people, take much longer to develop. Chapter 10 introduces this concern and loss reserving, the discipline of determining how much the insurance company should retain to meet its obligations.

1.4 Further Resources and Contributors

This book introduces loss data analytic tools that are most relevant to actuaries and other financial risk analysts. Here are a few reference cited in the chapter.

- Bailey, Robert A. and J. Simon LeRoy (1960). “Two studies in automobile ratemaking,” Proceedings of the Casualty Actuarial Society Casualty Actuarial Society, Vol. XLVII.
- Bowers, Newton L., Hans U. Gerber, James C. Hickman, Donald A. Jones, and Cecil J. Nesbitt (1986). Actuarial Mathematics. Society of Actuaries Itasca, Ill.

- Dickson, David C. M., Mary Hardy, and Howard R. Waters (2013). Actuarial Mathematics for Life Contingent Risks. Cambridge University Press.
- Earnix (2013). “2013 Insurance Predictive Modeling Survey,” Earnix and Insurance Services Office, Inc. [Retrieved on May 10, 2016].
- Gorman, Mark and Stephen Swenson (2013). “Building believers: How to expand the use of predictive analytics in claims,” SAS, [Retrieved on May 10, 2016].
- Insurance Information Institute (2015). “International Insurance Fact Book. [Retrieved on May 10, 2016].
- Taylor, Gregory C. (2014). “Claims triangles/Loss reserves,” in Edward W. Frees, Glenn Meyers, and Richard A. Derrig eds. Predictive Modeling Applications in Actuarial Science, Cambridge. Cambridge University Press.

Contributor

- **Edward W. (Jed) Frees**, University of Wisconsin-Madison, is the principal author of the initital version of this chapter. Email: jfrees@bus.wisc.edu for chapter comments and suggested improvements.

Chapter 2

Frequency Distributions

These are overheads from a course that provides some structure for this chapter.

2.1 How Frequency Augments Severity Information

Basic Terminology

- **Claim** - indemnification upon the occurrence of an insured event
  - **Loss** - some authors use claim and loss interchangeably, others think of loss as the amount suffered by the insured whereas claim is the amount paid by the insurer
- **Frequency** - how often an insured event occurs, typically within a policy contract
- **Count** - In this chapter, we focus on count random variables that represent the number of claims, that is, how frequently an event occurs
- **Severity** - Amount, or size, of each payment for an insured event

The Importance of Frequency

- Insurers pay claims in monetary units, e.g., US dollars. So, why should they care about how frequently claims occur?
- Many ways to use claims modeling – easiest to motivate in terms of pricing for personal lines insurance
  - Recall from Chapter 1 that setting the price of an insurance good can be a perplexing problem.
  - In manufacturing, the cost of a good is (relatively) known
  - Other financial service areas, market prices are available
  - Insurance tradition: 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 insurance company.
- Think of the expected cost as the expected number of claims times the expected amount per claims, that is, expected frequency times severity.
- Claim amounts, or severities, will turn out to be relatively homogeneous for many lines of business and so we begin our investigations with frequency modeling.

Other Ways that Frequency Augments Severity Information

- **Contractual** - For example, deductibles and policy limits are often in terms of each occurrence of an insured event
- **Behaviorial** - 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 is related primarily to personal characteristics whereas the cost per user may be more related to characteristics of the healthcare provider (such as the physician).
- **Databases**. Many insurers keep separate data files that suggest developing separate frequency and severity models. This recording process makes it natural for insurers to model the frequency and severity as separate processes.
  - Policyholder file that 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.
  - 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.)
- **Regulatory and Administrative**
  - 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.

2.2 Basic Frequency Distributions

2.2.1 Foundations

- Claim count N has support on the non-negative integers $k = 0, 1, 2, \ldots$.
- The **probability mass function** is denoted as $\Pr(N = k) = p_k$
- We can summarize the distribution through its **moments**
  - The **mean**, or first moment, is

$$\mathrm{E}\ N = \mu_1 = \mu = \sum_{k=0}^{\infty} k p_k.$$

  - More generally, the rth moment is

$$\mathrm{E}\ N^r = \mu'_r = \sum_{k=0}^{\infty} k^r p_k.$$

  - The **variance** is

$$\mathrm{Var}\ N = \mathrm{E}\ (N - \mu)^2 = \mathrm{E}\ N^2 - \mu^2$$

- Also recall the **moment generating function**

$$M_N(t) = \mathrm{E}\ e^{tN} = \sum_{k=0}^{\infty} e^{tk} p_k.$$

2.2.2 Probability Generating Function

- The **probability generating function** is

$$\begin{aligned} \mathrm{P}(z) &= \mathrm{E}\ z^N = \mathrm{E}\ \exp\left(N \ln z\right) = M_N(\ln z) \\ &= \sum_{k=0}^{\infty} z^k p_k. \end{aligned}$$

- By taking the mth derivative, we see that

$$P^{(m)}(z)\Big|_{z=0} = \frac{\partial^m}{\partial z^m} P(z)|_{z=0} = p_m m!$$

  the pgf "generates" the probabilities.

- Further, the pgf can be used to generate moments

$$P^{(1)}(1) = \sum k p_k = \mathrm{E}\ N.$$

  and

$$P^{(2)}(1) = \mathrm{E}\ N(N-1).$$

2.2.3 Important Frequency Distributions

- The three important (in insurance) frequency distributions are:
  - Poisson
  - Negative binomial
  - Binomial
- They are important because:
  - They fit well many insurance data sets of interest
  - They provide the basis for more complex distributions that even better approximate real situations of interest to us

Poisson Distribution

- This distribution has parameter λ, probability mass function

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}$$

  and pgf

$$P(z) = M_N(\ln z) = \exp(\lambda(z-1))$$

- The expectation is E $N = \lambda$ which is the same as the variance, Var $N = \lambda$.

Negative Binomial Distribution

- This distribution has parameters (r, β), probability mass function (pmf)

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

  and probability generating function (pgf)

$$P(z) = (1-\beta(z-1))^{-r}$$

- The expectation is $\mathrm{E}\ N = r\beta$ and the variance is $\mathrm{Var}\ N = r\beta(1+\beta)$.
- When $\beta > 0$, we have $\mathrm{Var}\ N > \mathrm{E}\ N$. This distribution is said to be **overdispersed** (relative to the Poisson).

Binomial Distribution

- This distribution has parameters (m, q), probability mass function

$$p_k = \binom{m}{k} q^k (1-q)^{m-k}$$

  and pgf

$$P(z) = (1+q(z-1))^m$$

- The mean is $\mathrm{E}\ N = mq$ and the variance is $\mathrm{Var}\ N = mq(1-q)$.

2.3 The (a, b, 0) Class

- Recall the notation $p_k = \Pr(N = k)$.
- Definition. A count distribution is a member of the **$(a, b,$ 0) class** if the probabilities p_k satisfy

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k},$$

  for constants a, b and for $k=1,2,3, ...$.

  - There are only three distributions that are members of the $(a, b,$0) class. They are the Poisson $(a = 0)$, binomial$(a < 0)$, and negative binomial $(a > 0)$.
  - The recursive expression provides a computationally efficient way to generate probabilities.

The (a, b, 0) Class - Special Cases

- Example: Poisson Distribution.
  - Recall the pmf $p_k = \frac{\lambda^k}{k!} e^{-\lambda}$. Examining the ratio,

$$\frac{p_k}{p_{k-1}} = \frac{\lambda^k/k!}{\lambda^{k-1}/(k-1)!} \frac{e^{-\lambda}}{e^{-\lambda}} = \frac{\lambda}{k}$$

    Thus, the Poisson is a member of the $(a, b,$ 0) class with $a = 0$, $b = \lambda$, and initial starting value $p_0 = e^{-\lambda}$.

Other special cases (Please check)

- Example: Binomial Distribution. Use a similar technique to check that the binomial distribution is a member of the $(a, b, 0)$ class with $a = \frac{-q}{1-q}$, $b = \frac{(m+1)q}{1-q}$, and initial starting value $p_0 = (1-q)^m$.

Another special case of the $(a, b,$ 0) Class (Please check)

- Example: Negative Binomial Distribution. Check that the negative binomial distribution is a member of the $(a, b, 0)$ class with $a = \frac{\beta}{1+\beta}$, $b = \frac{(r-1)\beta}{1+\beta}$, and initial starting value $p_0 = (1+\beta)^{-r}$.

Exercise. A discrete probability distribution has the following properties

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

Determine the expected value of this discrete random variable (Ans: 9)

2.3.1 The (a, b, 0) Class - Example

Exercise. 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)

2.4 Estimating Frequency Distributions

Parameter estimation

- The customary method of estimation is **maximum likelihood**.
- To provide intuition, we outline the ideas in the context of Bernoulli distribution.
  - This is a special case of the binomial distribution with $m = 1$
  - For count distributions, either there is a claim $N = 1$ or not $N = 0$. The probability mass function is
  $$p_k = \Pr(N = k) = \begin{cases} 1 - q & \text{if } k = 0 \\ q & \text{if } k = 1 \end{cases}.$$
- The Statistical Inference Problem
  - Now suppose that we have a collection of independent random variables. The ith variable is denoted as N_i. Further assume they have the same Bernoulli distribution with parameter q.
  - In statistical inference, we assume that we observe a sample of such random variables. The observed value of the ith random variable is n_i. Assuming that the Bernoulli distribution is correct, we wish to say something about the probability parameter q.

Bernoulli Likelihoods

- Definition. The **likelihood** is the observed value of the mass function.
- For a single observation, the likelihood is

$$\left\{ \begin{array}{ll} 1-q & \text{if } n_i = 0 \\ q & \text{if } n_i = 1 \end{array} \right. .$$

- The objective of **maximum likelihood estimation (MLE)** is to find the parameter values that produce the largest likelihood.
  - Finding the maximum of the logarithmic function yields the same solution as finding the maximum of the corresponding function.
  - Because it is generally computationally simpler, we consider the logarithmic (log-) likelihood, written as

$$\left\{ \begin{array}{ll} \ln(1-q) & \text{if } n_i = 0 \\ \ln q & \text{if } n_i = 1 \end{array} \right. .$$

Bernoulli MLE

- More compactly, the log-likelihood of a single observation is

$$n_i \ln q + (1-n_i)\ln(1-q),$$

- Assuming independence, the log-likelihood of the data set is

$$L_{Bern}(q) = \sum_i \{n_i \ln q + (1-n_i)\ln(1-q)\}$$

  - The (log) likelihood is viewed as a function of the parameters, with the data held fixed.
  - In contrast, the joint probability mass function is viewed as a function of the realized data, with the parameters held fixed.
- The method of maximum likelihood means finding the values of q that maximize the log-likelihood.
- We began with the Bernoulli distribution in part because the log-likelihood is easy to maximize.
- Take a derivative of $L_{Bern}(q)$ to get

$$\frac{\partial}{\partial q} L_{Bern}(q) = \sum_i \left\{ n_i \frac{1}{q} - (1-n_i)\frac{1}{1-q} \right\}$$

  and solving the equation $\frac{\partial}{\partial q} L_{Bern}(q) = 0$ yields

$$\hat{q} = \frac{\sum_i n_i}{\text{sample size}}$$

  or, in words, the MLE $\hat{q}$ is the fraction of one's in the sample.
- Just to be complete, you should check, by taking derivatives, that when we solve $\frac{\partial}{\partial q} L_{Bern}(q) = 0$ we are maximizing the function $L_{Bern}(q)$, not minimizing it.

Frequency Distributions MLE

- We can readily extend this procedure to all frequency distributions
- For notation, suppose that θ ("theta") is a parameter that describes a given frequency distribution $\Pr(N=k;\theta) = p_k(\theta)$
  - In later developments we will let θ be a vector but for the moment assume it to be a scalar.
- The log-likelihood of a a single observation is

$$\left\{ \begin{array}{ll} \ln p_0(\theta) & \text{if } n_i = 0 \\ \ln p_1(\theta) & \text{if } n_i = 1 \\ \vdots & \vdots \end{array} \right. .$$

  that can be written more compactly as

$$\sum_k I(n_i = k) \ln p_k(\theta).$$

  this uses the notation $I(\cdot)$ to be the indicator of a set (it returns one if the event is true and 0 otherwise).

- Assuming independence, the log-likelihood of the data set is

$$L(\theta) = \sum_i \left\{ \sum_k I(n_i = k) \ln p_k(\theta) \right\} = \left\{ \sum_k m_k \ln p_k(\theta) \right\}$$

  where we use the notation m_k to denote the number of observations that are observed having count k. Using notation, $m_k = \sum_i I(n_i = k)$.

- **Special Case**. Poisson. A simple exercise in calculus yields

$$\hat{\lambda} = \frac{\text{number of claims}}{\text{sample size}} = \frac{\sum_k k m_k}{\sum_k m_k}$$

  the average claim count.

2.5 Other Frequency Distributions

- Naturally, there are many other count distributions needed in practice
- For many insurance applications, one can work with one of our three basic distributions (binomial, Poisson, negative binomial) and allow the parameters to be a function of known explanatory variables.
  - This allows us to explain claim probabilities in terms of known (to the insurer) variables such as age, sex, geographic location (territory), and so forth.
  - This field of statistical study is known as **regression analysis** - it is an important topic that we will not pursue in this course.
- To extend our basic count distributions to alternatives needed in practice, we consider two approaches:
  - Zero truncation or modification
  - Mixing

2.5.1 Zero Truncation or Modification

- Why truncate or modify zero?
  - If we work with a database of claims, then there are no zero!
  - In personal lines (like auto), people may not want to report that first claim because they fear it will increase future insurance rates.
- Let's modify zero probabilities in terms of the $(a, b, 0)$ class
- Definition. A count distribution is a member of the **$(a, b,$ 1) class** if the probabilities p_k satisfy

  $$\frac{p_k}{p_{k-1}} = a + \frac{b}{k},$$

  for constants a, b and for $k = 2, 3, \ldots$.
- Note that this starts at $k = 2$, not $k = 1$. That is, the most important thing about this definition is that the recursion starts at p_1, not p_0.
- Thus, all distributions that are members of the $(a, b,\ 0)$ are members of the $(a, b,\ 1)$ class. Naturally, there are additional distributions that are members of this wider class.
- To see how this works, pick a specific distribution in the $(a, b,\ 0)$ class.
  - Consider p_k^0 to be a probability for this member of $(a, b, 0)$.
  - Let p_k^M be the corresponding probability for a member of $(a, b, 1)$, where the M stands for "modified".
  - Pick a new probability of a zero claim, p_0^M, and define

    $$c = \frac{1 - p_0^M}{1 - p_0^0}.$$

  - We then calculate the rest of the modified distribution as

    $$p_k^M = cp_k^0$$

Special Case: Poisson Truncated at Zero.

For this case, we assume that $p_0^M = 0$, so that the probability of $N = 0$ is zero, hence the name "truncated at zero."

- For this case, we use the letter T to denote probabilities instead of M, so we use p_k^T for probabilities. Thus,

  $$p_k^T = \begin{cases} 0 & k = 0 \\ \frac{1}{1-p_0^0} p_k^0 & k \geq 1 \end{cases}$$

Modified Poisson Example

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

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.

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	$p_3 = p_2\left(\frac{\lambda}{3}\right) = 0.180447$	$p_3^T = p_2^T\left(\frac{\lambda}{3}\right) = 0.208690$	$p_3^M = p_2^M\left(\frac{\lambda}{2}\right) = 0.083476$

Modified Poisson Exercise

Exercise: Course 3, May 2000, Exercise 37. 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.

2.6 Mixture Distributions

2.6.1 Mixtures of Finite Populations

- Suppose that our population consists of several subgroups, each having their own distribution
- We randomly draw an observation from the population, without knowing which subgroup that we are drawing from
- For example, suppose that N_1 represents claims form "good" drivers and N_2 represents claims from "bad" drivers. We draw
$$N = \begin{cases} N_1 & \text{with prob } \alpha \\ N_2 & \text{with prob } (1-\alpha). \end{cases}$$
- Here, α represents the probability of drawing a "good" driver.
- Our is said to be a "mixture" of two subgroups

Finite Population Mixture Example

Exercise. Exam "C" 170. 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

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.

2.6.2 Mixtures of Infinitely Many Populations

- We can extend the mixture idea to an infinite number of populations.
- To illustrate, suppose we have a population of drivers. The ith person has their own (personal) expected number of claims, λ_i.
- For some driver's, λ is small (good drivers), for others it is high (not so good drivers). There is a distribution of λ.
- A convenient distribution is to use a gamma distribution with parameters (α, θ).
- Then, one can check that

  $$N \sim \quad \text{Negative Binomial}(r = \alpha, \beta = \theta).$$

  See, for example, KPW, page 84.
- Mixture is very important in insurance applications, more on this later...

Negative Binomial as a Gamma Mixture of Poissons - Example

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

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

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

Now, one can directly use the negative binomial approach to get $r = \alpha = \frac{1}{2}$ and $\beta = \theta = 2$. Thus

$$\begin{aligned}\Pr(N = 1) = p_1 &= \binom{1 + r - 1}{1} (\frac{1}{(1+\beta)^r})(\frac{\beta}{1+\beta})^1 \\ &= \binom{1 + \frac{1}{2} - 1}{1} \frac{1}{(1+2)^{1/2}} (\frac{2}{1+2})^1 \\ &= \frac{1}{3^{3/2}} = 0.19245.\end{aligned}$$

2.7 Goodness of Fit

Example: Singapore Automobile Data

- A 1993 portfolio of $n = 7,483$ automobile insurance policies from a major insurance company in Singapore.
- The count variable is the number of automobile accidents per policyholder.
- There were on average 0.06989 accidents per person.

Table. Comparison of Observed to Fitted Counts Based on Singapore Automobile Data

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

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

Singapore Data: Adequacy of the Poisson Model

- With the Poisson distribution
  - The maximum likelihood estimator of λ is $\widehat{\lambda} = \overline{N}$.
  - Estimated probabilities, using $\widehat{\lambda}$, are denoted as $\widehat{p}_k$.
- For goodness of fit, consider Pearson's chi-square statistic

$$\sum_k \frac{(m_k - n\widehat{p}_k)^2}{n\widehat{p}_k}.$$

  - Assuming that the Poisson distribution is a correct model; this statistic has an asymptotic chi-square distribution
    - The degrees of freedom (*df*) equals the number of cells minus one minus the number of estimated parameters.
  - For the Singapore data, this is $df = 5 - 1 - 1 = 3$.
  - The statistic is 41.98; the basic Poisson model is inadequate.

Example. Course C/Exam 4. May 2001, 19.

During a one-year period, the number of accidents per day was distributed as follows:

Number of Accidents	0	1	2	3	4	5
Number 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 chi-square statistic.

2.8 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.

```
knitr::include_url("http://www.ssc.wisc.edu/~jfrees/loss-data-analytics/loss-data-analytics-problems/",
```

2.9 Technical Supplement: Iterated Expectations

Iterated Expectations

In some situations, we only observe a single outcome but can conceptualize an outcome as resulting from a two (or more) stage process. These are called **two-stage**, or "**hierarchical**," type situations. Some special cases include:

- problems where the parameters of the distribution are random variables,
- mixture problems, where stage 1 represents the type of subpopulation and stage 2 represents a random variable with a distribution that depends on population type
- an aggregate distribution, where stage 1 represents the number of events and stage two represents the amount per event.

In these situations, the law of iterated expectations can be useful. The law of total variation is a special case that is particularly helpful for variance calculations.

To apply these rules,

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 (ii), one for each stage 1 outcome. Then, combine these results using the iterated expectations or total variation rules.

Iterated Expectations

- Consider two random variables, X and Y, and a function $h(X,Y)$. Assuming expectations exists and are finite, a rule/theorem from probability states that

$$\mathrm{E}~h(X,Y) = \mathrm{E}~\left\{ \mathrm{E}~\left(h(X,Y)|X \right) \right\}.$$

- This result is known as the law of iterated expectations.
- Here, the random variables may be discrete, continuous, or a hybrid combination of the two.
- Similarly, the law of total variation is

$$\mathrm{Var}~h(X,Y) = \mathrm{E}~\left\{ \mathrm{Var}~\left(h(X,Y)|X \right) \right\} + \mathrm{Var}~\left\{ \mathrm{E}~\left(h(X,Y)|X \right) \right\},$$

  the expectation of the conditional variance plus the variance of the conditional expectation.

Discrete Iterated Expectations

- To illustrate, suppose that X and Y are both discrete random variables with joint probability

$$p(x,y) = \Pr(X=x, Y=y).$$

- Further, let $p(y|x) = \frac{p(x,y)}{\Pr(X=x)}$ be the conditional probability mass function.

- The conditional expectation is

$$\mathrm{E}\ (h(X,Y)|X=x)=\sum_y h(x,y)p(y|x)$$

- You can use the conditional expectation to get the unconditional expectation using

$$\begin{aligned}\mathrm{E}\ \{\mathrm{E}\ (h(X,Y)|X)\} &= \sum_x \left\{\sum_y h(x,y)p(y|x)\right\} \Pr(X=x)\\ &= \sum_x \sum_y h(x,y)p(y|x)\Pr(X=x)\\ &= \sum_x \sum_y h(x,y)p(x,y) = \mathrm{E}\ h(X,Y)\end{aligned}$$

- The proofs of the law of iterated expectations for the continuous and hybrid cases are similar.

Law of Total Variation

- To see this rule, first note that we can calculate a conditional variance as

$$\mathrm{Var}\ (h(X,Y)|X) = \mathrm{E}\ \big(h(X,Y)^2|X\big) - \{\mathrm{E}\ (h(X,Y)|X)\}^2 .$$

- From this, the expectation of the conditional variance is

$$\mathrm{E\ Var}\ (h(X,Y)|X) = \mathrm{E}\ \big(h(X,Y)^2\big) - \mathrm{E}\ \{\mathrm{E}\ (h(X,Y)|X)\}^2 .$$

- Further, note that the conditional expectation, E $(h(X,Y)|X=x)$, is a function of x, say, $g(x)$.
- Now, $g(X)$ is a random variable with mean E $h(X,Y)$ and variance

$$\begin{aligned}\mathrm{Var}\ \{\mathrm{E}\ (h(X,Y)|X)\} &= \mathrm{Var}\ g(X)\\ &= \mathrm{E}\ g(X)^2\ - (\mathrm{E}\ h(X,Y))^2\\ &= \mathrm{E}\ \{\mathrm{E}\ (h(X,Y)|X)\}^2 - (\mathrm{E}\ h(X,Y))^2\end{aligned}$$

- Adding the variance of the conditional expectation in equation to the expectation of conditional variance in equation gives the law of total variation.

Mixtures of Finite Populations: Example

- For example, suppose that N_1 represents claims form "good" drivers and N_2 represents claims from "bad" drivers. We draw

$$N = \begin{cases} N_1 & \text{with prob } \alpha \\ N_2 & \text{with prob } (1-\alpha). \end{cases}$$

- Here, α represents the probability of drawing a "good" driver.
- Let T be the type, so $T=1$ with prob α and $T=2$ with prob $1-\alpha$.
- From the law of iterated expectations, we have

$$\begin{aligned}\mathrm{E}\ N &= \mathrm{E}\ \{\mathrm{E}\ (N|T)\}\\ &= \mathrm{E}\ N_1 \times \alpha + \mathrm{E}\ N_2 \times (1-\alpha).\end{aligned}$$

- From the law of total variation

$$\mathrm{Var}\ N = \mathrm{E}\ \{\mathrm{Var}\ (N|T)\} + \mathrm{Var}\ \{\mathrm{E}\ (N|T)\} ,$$

Mixtures of Finite Populations: Example 2

- To be more concrete, suppose that N_j is Poisson with parameter λ_j. Then

$$\text{Var } N_j|T = \text{E } N_j|T = \begin{cases} \lambda_1 & T = 1 \\ \lambda_2 & T = 2 \end{cases}$$

- Thus

$$\text{E } \{\text{Var } (N|T)\} = \alpha\lambda_1 + (1-\alpha)\lambda_2$$

  and

$$\text{Var } \{\text{E } (N|T)\} = (\lambda_1 - \lambda_2)^2 \alpha(1-\alpha)$$

  (Recall: for a Bernoulli with outcomes a and b and prob α, the variance is $(b-a)^2\alpha(1-\alpha)$).

- Thus,

$$\text{Var } N = \alpha\lambda_1 + (1-\alpha)\lambda_2 + (\lambda_1 - \lambda_2)^2 \alpha(1-\alpha)$$

Chapter 3

Modeling Loss Severity

October 27, 2016

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. Mixture distributions are 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. In this chapter we 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.

3.1 Basic Distributional Quantities

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

3.1.1 Moments

Let X be a continuous random variable with probability density function $f_X\left(x\right)$. The k-th raw moment of X, denoted by $\mu_k^{\prime}$, is the expected value of the k-th power of X, provided it exists. The first raw moment $\mu_1^{\prime}$ is the mean of X usually denoted by μ. The formula for $\mu_k^{\prime}$ is given as

$$\mu_k^{\prime}=E\left(X^k\right)=\int_0^{\infty}x^k f_X\left(x\right)dx.$$

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

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

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 (SOA) X has a gamma distribution with mean 8 and skewness 1. Find the variance of X.

Solution

The probability density function of X is given by

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

for $x > 0$. For $\alpha > 0$,

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

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

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

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

3.1.2 Quantiles

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

$$F_X\left({\pi_p}^-\right) \leq p \leq F(\pi_p),$$

for $0 \leq p \leq 1$.

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

Solution

$F_X(x) = 1 - e^{-\theta x}$. Then, $F_X(\pi_{0.5}) = 1 - e^{-\theta\pi_{0.5}} = 0.5$. Thus, $1 - e^{-\theta/3} = 0.5$ and $\theta = 3\ln 2$.

3.1.3 The 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

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

for all t for which the expected value exists.

Example 3.3 (SOA) 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.

Solution

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

Then,

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

Thus, $b = 4$.

Example 3.4 Let $X_1,\ X_2,\ .,\ X_n$ be independent $\mathrm{Ga}(\alpha_i, \theta)$ random variables. Find the distribution of $S = \sum_{i=1}^n X_i$.

Solution

The moment generating function of S is

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

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

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, $Var(S) = \alpha\theta^2$.

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

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

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 evaluate it at $z = 1$, we get

$$E\left[X(X-1)\ldots(X-k+1)\right].$$

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 The Gamma Distribution

The gamma distribution is commonly used in modeling claim severity. 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 $\alpha > 0,\ \theta > 0$.

Figures 3.1 and 3.2 demonstrate the effect of the scale and shape parameters on the gamma density function.

R Code for Gamma Density Plots

```
# Varying Scale Gamma Densities
scaleparam <- seq(100,250,by=50)
shapeparam <- 2:5
x = seq(0,1000,by=1)
par(mar = c(4, 4, .1, .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
par(mar = c(4, 4, .1, .1))
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)
```

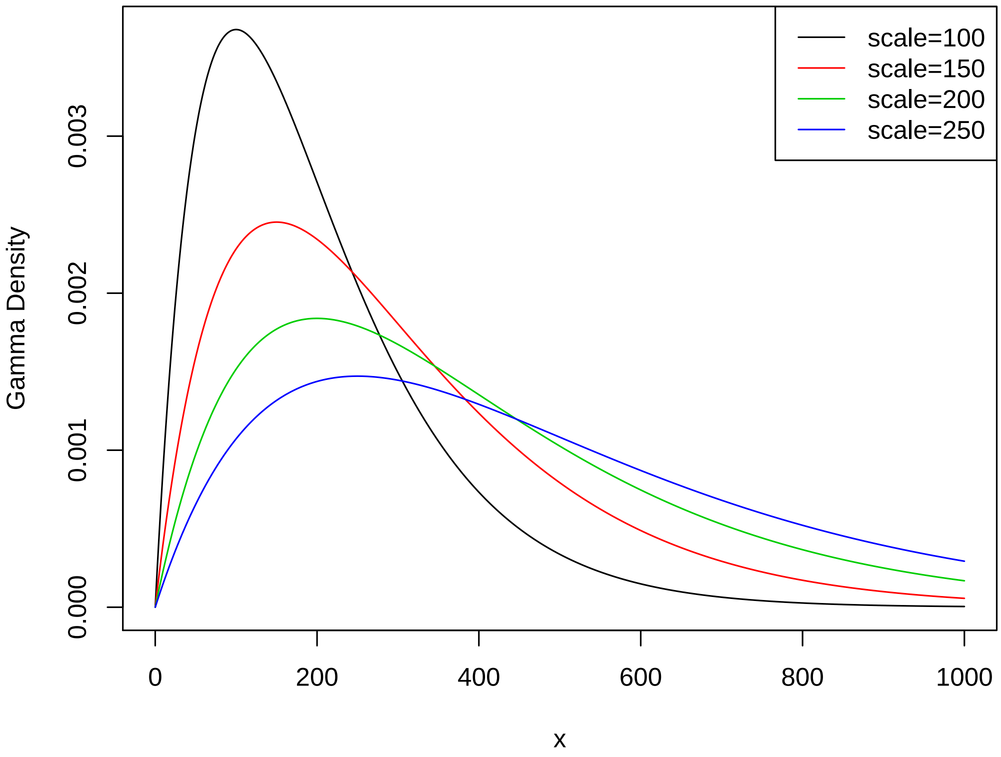


Figure 3.1: Gamma Density, with shape=2 and Varying Scale

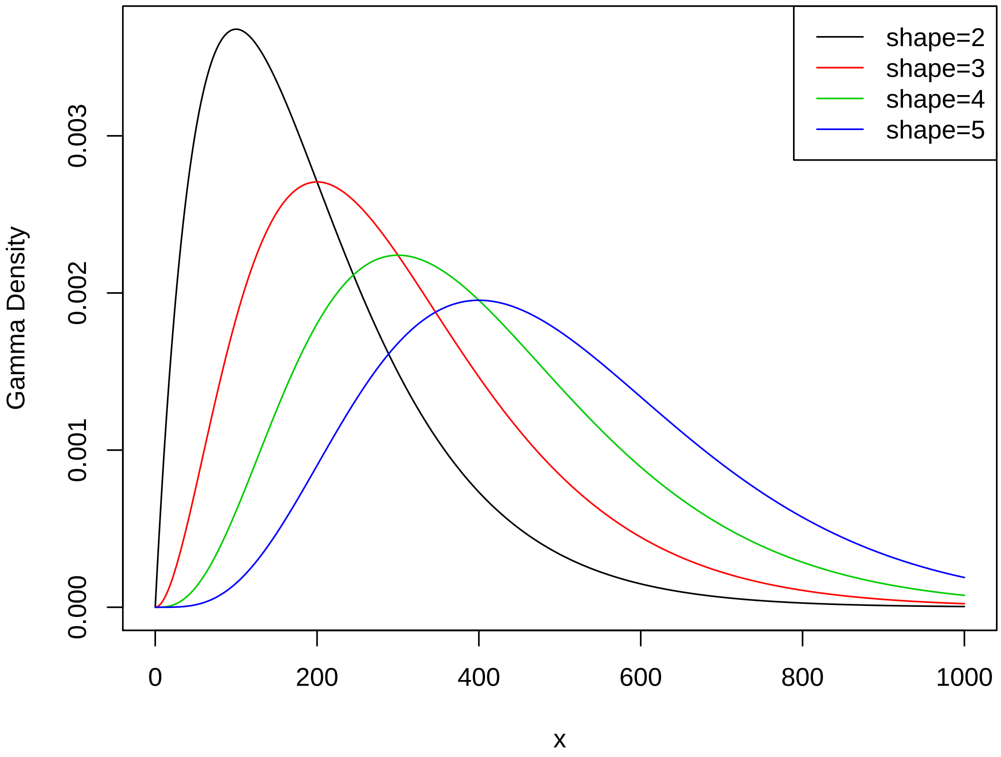


Figure 3.2: Gamma Density, with scale=100 and Varying Shape

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 3.5.2, 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 $\Gamma\left(\frac{\alpha;x}{\theta}\right)$, and defined as

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

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

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

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

The mean and variance are given by $E\left(X\right) = \alpha\theta$ and $Var\left(X\right) = \alpha\theta^2$, 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 The 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 modelled 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

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

Figures 3.3 and 3.4 demonstrate the effect of the scale and shape parameters on the Pareto density function.

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

R Code for Pareto Density Plots

```
# Varying Scale Pareto Densities
#install.packages("VGAM")
library(VGAM)
scaleparam <- seq(2000,3500,500)
shapeparam <- 1:4
z<- seq(1,3000,by=1)
fpareto <- dpareto(z, shape = 3, scale = scaleparam[1])
plot(z, fpareto, ylim=c(0,0.002),type = "l", ylab = "Pareto Density")
for(k in 2:length(shapeparam)){
  fpareto <- dpareto(z,shape = 3, scale = scaleparam[k])
  lines(z,fpareto, col = k) }
legend("topright", c("scale=2000", "scale=2500", "scale=3000", "scale=3500"), lty=1, col = 1:4)

# Varying Shape Pareto Densities
```

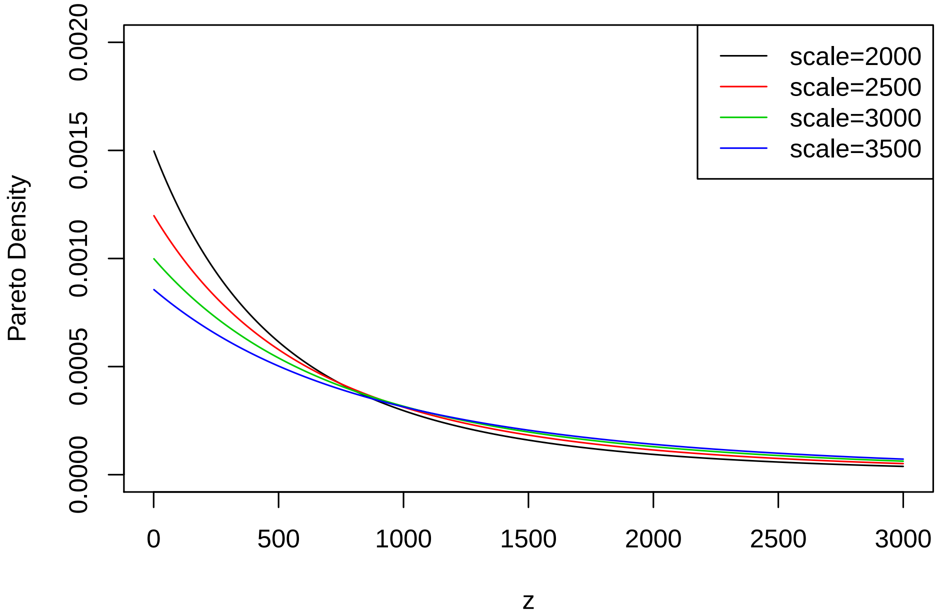


Figure 3.3: Pareto Density, with shape=3 and Varying Scale

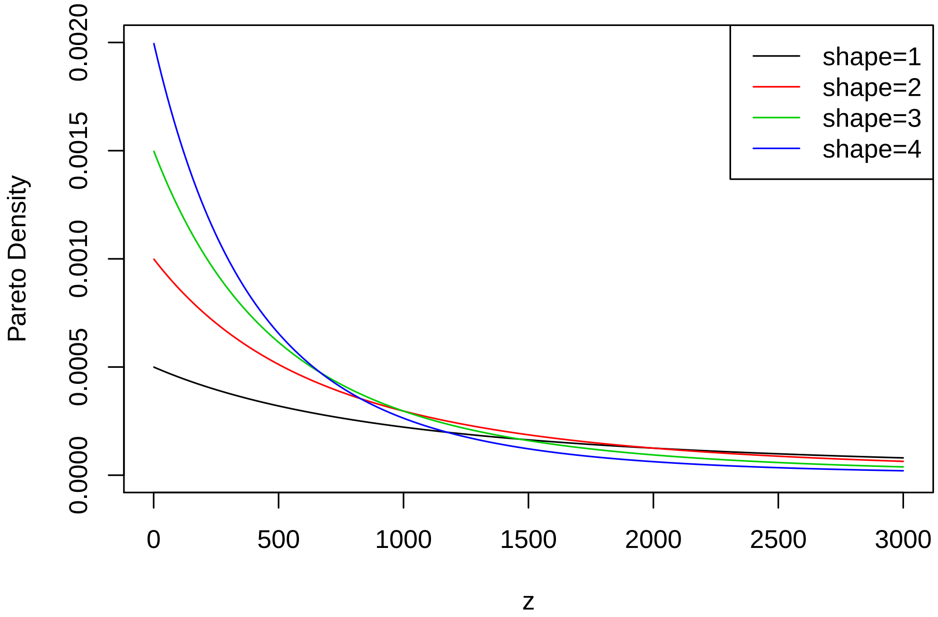


Figure 3.4: Pareto Density, with scale=2000 and Varying Shape

```
fpareto <- dpareto(z, shape = shapeparam[1], scale = 2000)
plot(z, fpareto, ylim=c(0,0.002),type = "l", ylab = "Pareto Density")
for(k in 2:length(shapeparam)){
  fpareto <- dpareto(z,shape = shapeparam[k], scale = 2000)
  lines(z,fpareto, col = k)}
legend("topright", c("shape=1", "shape=2", "shape=3", "shape=4"), lty=1, col = 1:4)
```

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

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

The mean and variance are given by

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

and

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

respectively.

Example 3.5 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.

Solution

$E\left(X\right) = \frac{\theta}{\alpha-1} = 40$ and $Var\left(X\right) = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} = 1800$. 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 $\theta = 680.72$.

The 95-th percentile, $\pi_{0.95}$, satisfies the equation

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

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

3.2.3 The 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.

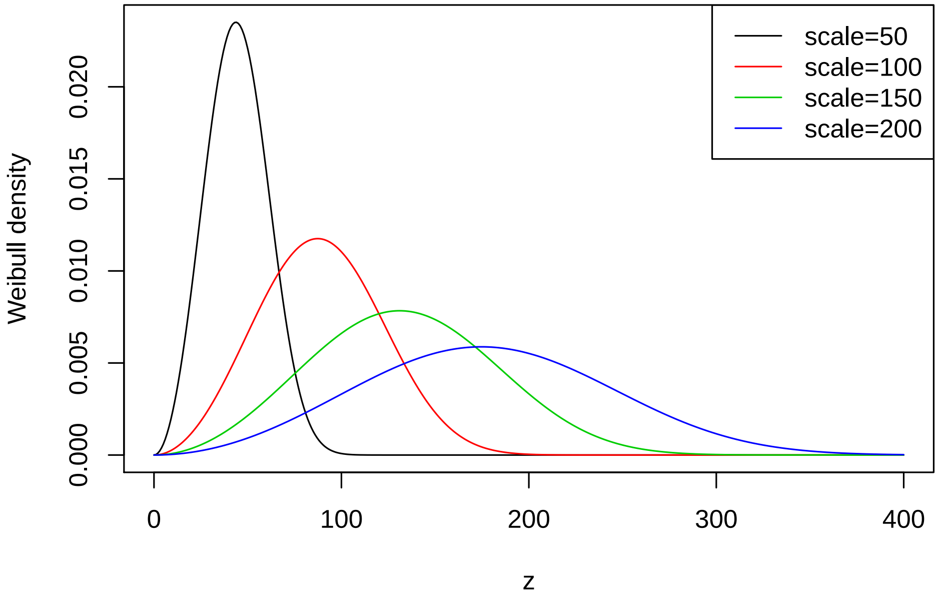


Figure 3.5: Weibull Density, with shape=3 and Varying Scale

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

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

Figures 3.5 and 3.6 demonstrate the effects of the scale and shape parameters on the Weibull density function.

R Code for Weibull Density Plots

```
# 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 d
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

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

It can be easily seen that the shape parameter α describes the shape of the hazard function of the Weibull

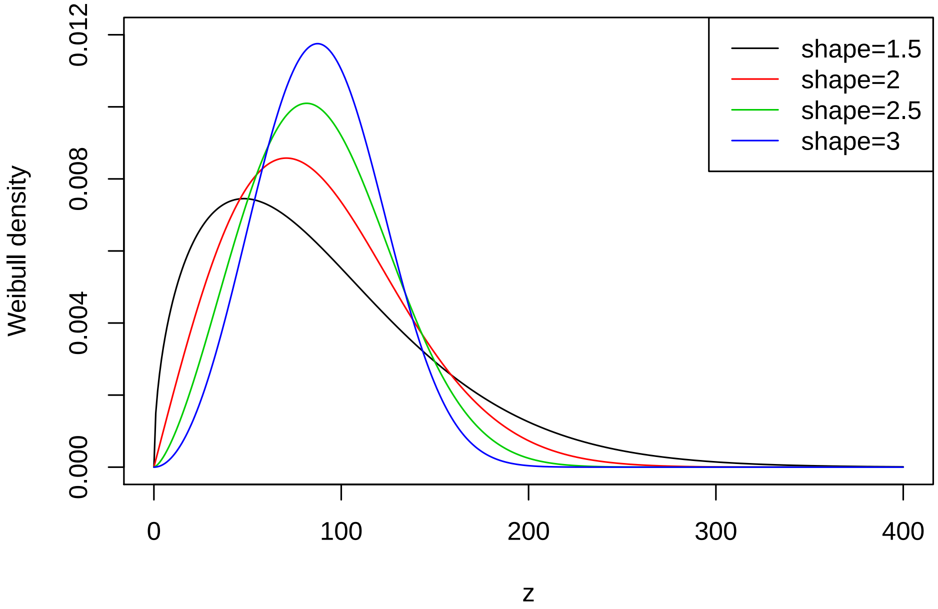


Figure 3.6: Weibull Density, with scale=100 and Varying Shape

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.

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

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

The mean and variance are given by

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

and

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

respectively.

Example 3.6 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 this distribution.

Solution

Let X be the lifetime of AIDS patients (in months)

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

Let Y be the number of patients who die within one year of diagnosis. Then, $Y \sim Bin\left(10,\ 0.254\right)$ and $\Pr\left(Y \leq 2\right) = 0.514$. Let $\pi_{0.99}$ denote the 99-th percentile of this distribution. Then,

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

and $\pi_{0.99} = 118.99$.

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

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

$a, b, \alpha, \beta > 0$, 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

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

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

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

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

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

Suppose that X belongs to a certain set of parametric 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 parameter remain unchanged.

Example 3.7 (SOA) 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?

Solution

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

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

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

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

Then, the probability density function of interest $f_Y(y)$ is

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

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

Example 3.8 Demonstrate that the gamma distribution is a scale distribution. Solution

Let $X \sim Ga(\alpha, \theta)$ and $Y = cX$, then

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

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

3.3.3 Raising to a Power

In the previous section we have 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

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

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

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

and

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

Example 3.9 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. Solution

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

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

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 3.10 (SOA) X has a uniform distribution on the interval $(0,\ c)$. $Y = e^X$. Find the distribution of Y. Solution

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

Then,

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

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 $E\left(X^k\right) = aE\left(X_1^K\right) + (1-a) E\left(X_2^k\right)$, and

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

respectively.

Example 3.11 (SOA) The distribution of the random variable X is an equally weighted mixture of two Poisson distributions with parameters λ_1 and λ_2 respectively. The mean and variance of X are 4 and 13, respectively. Determine $\Pr(X > 2)$. Solution

$$E\left(X\right)=0.5\lambda_1+0.5\lambda_2=4$$

$$E\left(X^2\right)=0.5\left(\lambda_1+\lambda_1^2\right)+0.5\left(\lambda_2+\lambda_2^2\right)=13+16$$

Simplifying the two equations we get $\lambda_1+\lambda_2=8$ and $\lambda_1^2+\lambda_2^2=50$. Then, the parameters of the two Poisson distributions are 1 and 7.

$$\Pr\left(X>2\right)=0.5\Pr\left(X_1>2\right)+0.5\Pr\left(X_2>2\right)=0.05$$

***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 $f_{X_i}\left(x\right)$. The probability distribution of X is given by

$$f_X\left(x\right)=\sum_{i=1}^{k}p_if_{X_i}\left(x\right),$$

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

$$\begin{aligned}F_X\left(x\right)&=\sum_{i=1}^{k}p_iF_{X_i}\left(x\right),\\E\left(X^r\right)&=\sum_{i=1}^{k}p_iE\left(X_i^r\right),\text{and}\\M_X\left(t\right)&=\sum_{i=1}^{k}p_iM_{X_i}\left(t\right),\end{aligned}$$

respectively.

Example 3.12 (SOA) 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. Solution

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

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

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

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

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

Then, Z is a mixture of X_1, X_2, X_3 and X_4, with mixing weights ca, $c\left(1-a\right)$, $\left(1-c\right)b$ and $\left(1-c\right)\left(1-b\right)$.

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

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

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

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

respectively.

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

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

In particular the mean and variance of X are given by

$$E(X) = E[E(X|\theta)]$$

and

$$Var(X) = E[Var(X|\theta)] + Var[E(X|\theta)].$$

Example 3.13 (SOA) X has a binomial distribution with a mean of $100q$ and a variance of $100q(1-q)$ and q has a beta distribution with parameters $a = 3$ and $b = 2$. Find the unconditional mean and variance of X.
Solution

$E(q) = \frac{a}{a+b} = \frac{3}{5}$ and $E(q^2) = \frac{a(a+1)}{(a+b)(a+b+1)} = \frac{2}{5}$.

$E(X) = E[E(X|q)] = E(100q) = 100E(q) = 60,$

$$Var(X) = E[Var(X|q)] + Var[E(X|q)] = E[100q(1-q)] + Var(100q)$$

$= 100E(q) - 100E(q^2) + 100^2 V(q) = 420.$

Exercise 3.14 (SOA) Claim sizes, X, are uniform on for each policyholder. varies by policyholder according to an exponential distribution with mean 5. Find the unconditional distribution, mean and variance of X.
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 $0 < \theta < \infty$.

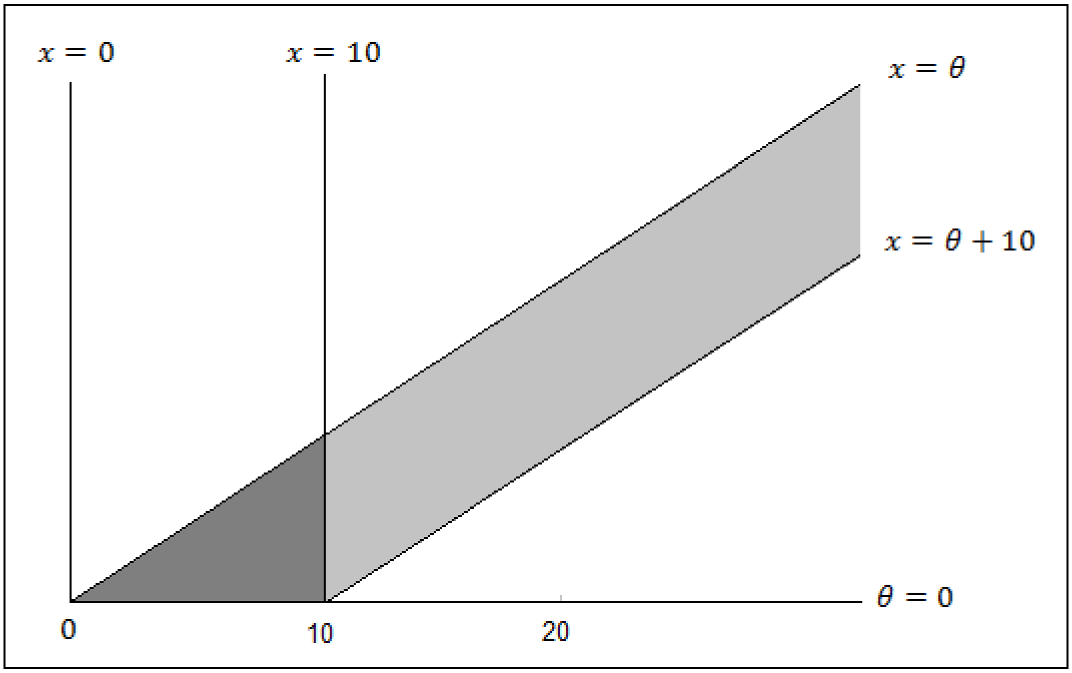


Figure 3.7:

The conditional mean and variance of X are given by

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

and

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

respectively.

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

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

and

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

The unconditional distribution of X is

$$f_X\left(x \right) = \int f_X\left(x | \theta \right) \; g\left(\theta \right) d\theta.$$

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

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. 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 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

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

and

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

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

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

and

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

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

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

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

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

and

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

respectively.

Example 3.15 (SOA) 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.

Solution

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

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

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

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

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

The solution can be simplified if we 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. 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

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

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

Example 3.16 (SOA) For an insurance:

Losses have a density function

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

The insurance has an ordinary deductible of 4 per loss.

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

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

Solution

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

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

Example 3.17 (SOA) 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.

Solution

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

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

3.4.2 Policy Limits

Under a limited policy, the insurer is responsible for covering the actual loss X up to the limit of its coverage. This fixed limit of 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.

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

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

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

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

Example 3.18 (SOA) 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

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

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

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

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

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

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

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

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

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

A growth factor $\left(1+r\right)$ may be applied to X resulting in an inflated loss random variable $\left(1+r\right)X$ (the prespecified d and u remain unchanged). The resulting per loss variable can be written as

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

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

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

and

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

respectively.

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

Example 3.19 (SOA) 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. Solution

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

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

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

$=356.026.$

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

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

$=361.659.$

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

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

There is an increase of 1.1% from 2006 to 2007.

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, when the loss exceeds the deductible, the insurer is not required to pay until the insured has paid the fixed 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. The reinsurer is not required to pay under the reinsurance contract until the insurer has paid a loss to its original insured. The 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, expertize and proficient complex claim file handling of the larger reinsurance companies.

Example 3.20 (SOA) 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 $\frac{R_{2006}}{P_{2006}}$. 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 $i=2005,\ 2006$

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

where $i=2005,\ 2006$

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

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

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

$$P_{2005}=1.2E\left(Y_{2005}\right)=3000$$

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

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

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

$$P_{2006}=1.2E\left(Y_{2006}\right)=3702.857$$

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

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

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

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

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

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

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

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

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). The principles for using maximum likelihood to estimate model parameters were introduced in Chapter **xxx**. 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 3.21 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 $\Pr\left(X \leq 9,000\right).$

Solution

The probability density function is

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

where $x > 0$. 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 loglikelihood 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 $\ln L\left(\theta\right)$ is given by

$$\frac{d^2 \ln L\left(\theta\right)}{d\theta^2} = \frac{-4}{\theta^2}.$$

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.

Taking reciprocal of negative expectation of the second derivative of $\ln L\left(\theta\right)$, 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.

The 95% confidence interval for θ is given by

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

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 3.22 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.

Solution

The probability density function is

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

where $x > 0$. The likelihood function, $L(\mu, \sigma)$, is the product of the pdf for each data point.

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

The loglikelihood function, $\ln L(\mu, \sigma)$, is the sum of the individual logarithms.

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

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

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{6}(lnx_i - \hat{\mu}) = 0.$$

$$\frac{-6}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{6}(\ln x_i - \hat{\mu})^2 = 0.$$

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

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

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.

$E\left(\frac{\partial^2 \text{lnL}(\mu,\sigma)}{\partial\mu^2}\right) = E\left(\frac{-6}{\sigma^2}\right) = \frac{-6}{\sigma^2},$

$E\left(\frac{\partial^2 \text{lnL}(\mu,\sigma)}{\partial\mu\partial\sigma}\right) = \frac{-2}{\sigma^3}\sum_{i=1}^{6}E(\ln x_i - \mu) = \frac{-2}{\sigma^3}\sum_{i=1}^{6}[E(\ln x_i) - \mu]{=}\frac{-2}{\sigma^3}\sum_{i=1}^{6}(\mu-\mu) = 0,$

and

$E\left(\frac{\partial^2 \text{lnL}(\mu,\sigma)}{\partial\sigma^2}\right) = \frac{6}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{6}E(\ln x_i - \mu)^2 = \frac{6}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{6}V(\ln x_i) = \frac{6}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{6}\sigma^2 = \frac{-12}{\sigma^2}.$

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

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

.

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

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 $g\left(\hat{\mu},\hat{\sigma}\right)$.

$\frac{\partial g(\mu,\sigma)}{\partial\mu} = exp\left(\mu+\frac{\sigma^2}{2}\right)$ 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 $g\left(\hat{\mu},\hat{\sigma}\right)$ is given by

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

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

71,374,380,000

The 95% confidence interval for $\exp\left(\mu+\frac{\sigma^2}{2}\right)$ is given by

$153,277 \pm 1.96\sqrt{71,374,380,000} = (-370,356,\ 676,910)$.

Since the mean of the lognormal distribution cannot be negative, we should replace the negative lower limit in the previous interval by a zero.

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

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

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 3.23 (SOA) For a group of policies, you are given that losses follow the distribution function $F\left(x\right) = 1-\frac{\theta}{x}$, 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.

Solution

The contribution of each of the 9 observations in the first interval to the likelihood function is the probability of $X \le 10$; that is, $\Pr\left(X \le 10\right) = F\left(10\right)$. Similarly, the contributions of each of 6 and 5 observations in the second and third intervals are $\Pr\left(10 < X \le 25\right) = F\left(25\right) - F(10)$ and $P\left(X > 25\right) = 1 - F(25)$, respectively. The likelihood function is thus given by

$$L\left(\theta\right) = \left[F\left(10\right)\right]^9\left[F\left(25\right) - F(10)\right]^6\left[1 - F(25)\right]^5$$

$$= \left(1 - \frac{\theta}{10}\right)^9 \left(\frac{\theta}{10} - \frac{\theta}{25}\right)^6 \left(\frac{\theta}{25}\right)^5$$

$$= \left(\frac{10-\theta}{10}\right)^9 \left(\frac{15\theta}{250}\right)^6 \left(\frac{\theta}{25}\right)^5.$$

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 distinguishing feature of data gathering mechanism is censoring. While for some event of interest (losses, claims, lifetimes, etc.) the complete data maybe available, for others only partial information is available; information 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 3.24 (SOA) The random variable 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. Solution

The contributions of the two observations 2 and 4 are $f_X(2)$ and $f_X(4)$ respectively. The contribution of the third observation, which is only known to exceed 4 is $S_X(4)$. The likelihood function is thus given by

$$L(\theta) = f_X(2) f_X(4) S_X(4).$$

The probability density function of X is given by

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

Thus,

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

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

and

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

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 3.25 (SOA) 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.

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

$$L(\alpha) = \frac{\prod_{i=1}^{8} f_X(x_i)}{\left[S_X(5)\right]^8}\left[\frac{S_X(25)}{S_X(5)}\right]^2.$$

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

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

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

In describing losses, actuaries fit appropriate parametric distribution models for the frequency and severity of loss. This involves finding appropriate statistical distributions that could efficiently model the data in hand. After fitting a distribution model to a data set, the model should be validated. Model validation is a crucial step in the model building sequence. It assesses how well these statistical distributions fit the data in hand and how well can we expect this model to perform in the future. If the selected model does not fit the data, another distribution is to be chosen. If more than one model seems to be a good fit for the data, we then have to make the choice on which model to use. It should be noted though that the same data should not serve for both purposes (fitting and validating the model). Additional data should be used to assess the performance of the model. There are many statistical tools for model validation. Alternative goodness of fit tests used to determine whether sample data are consistent with the candidate model, will be presented in a separate chapter.

Further Readings and References

- Cummins, J. D. and Derrig, R. A. 1991. Managing the Insolvency Risk of Insurance Companies, Springer Science+ Business Media, LLC.
- Frees, E. W. and Valdez, E. A. 2008. Hierarchical insurance claims modeling, Journal of the American Statistical Association, 103, 1457-1469.
- Klugman, S. A., Panjer, H. H. and Willmot, G. E. 2008. Loss Models from Data to Decisions, Wiley.
- Kreer, M., Kizilers, A., Thomas, A. W. and Eg?dio dos Reis, A. D. 2015. Goodness-of-fit tests and applications for left-truncated Weibull distributions to non-life insurance, European Actuarial Journal, 5, 139-163.
- McDonald, J. B. 1984. Some generalized functions for the size distribution of income, Econometrica 52, 647-663.
- McDonald, J. B. and Xu, Y. J. 1995. A generalization of the beta distribution with applications, Journal of Econometrics 66, 133-52.
- Tevet, D. 2016. Applying generalized linear models to insurance data: Frequency/severity versus premium modeling in: Frees, E. W., Derrig, A. R. and Meyers G. (Eds.) Predictive Modeling Applications in Actuarial Science Vol. II Case Studies in Insurance. Cambridge University Press.
- Venter, G. 1983. Transformed beta and gamma distributions and aggregate losses. Proceedings of the Casualty Actuarial Society 70: 156-193.

Contributors

- **Zeinab Amin**, The American University in Cairo, is the principal author of this chapter. Date: October 27, 2016. 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 .

3.7 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.

```
knitr::include_url("http://www.ssc.wisc.edu/~jfrees/loss-data-analytics/chapter-3-modeling-loss-severity
```

Chapter 4

Model Selection, Validation, and Inference

Chapter Preview. Chapters 2 and 3 have described how to fit parametric models to frequency and severity data, respectively. This chapter describes selection of models. To compare alternative parametric models, it is helpful to introduce models that 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 is a hybrid combination of discrete and continuous data. Further, Section 4.3 introduces for alternative sampling schemes, included grouped, censored and truncated data. 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

Consider $X_1, \ldots, X_n$, a **random sample** (with replacement) from an unknown underlying population distribution $F(\cdot)$. As independent draws from the same distribution, we say that $X_1, \ldots, X_n$ are independently and identically distributed (iid) random variables. Now say we have a data sample, $x_1, \ldots, x_n$, which represents a realization of $X_1, \ldots, X_n$. Note that $x_1, \ldots, x_n$ is non-random; it is simply a particular set of data values, i.e. an observation of the random variables $X_1, \ldots, X_n$. Using this sample, we will try to estimate the population distribution function $F(\cdot)$. We first proceed with a **nonparametric** analysis, in which we do not assume or rely on any explicit parametric distributional forms for $F(\cdot)$.

4.1.1 Nonparametric Estimation

The population distribution $F(\cdot)$ can be summarized in various ways. These include moments, the distribution function $F(\cdot)$ itself, the quantiles or percentiles associated with the distribution, and the corresponding mass or density function $f(\cdot)$. Summary statistics based on the sample, $X_1, \ldots, X_n$, are known as **nonparametric estimators** of the corresponding summary measures of the distribution. We will examine moment estimators, distribution function estimators, quantile estimators, and density estimators, as well as their statistical properties such as expected value and variance. Using our data observations $x_1, \ldots, x_n$, we can put numerical values to these estimators and compute **nonparametric estimates**.

Moment Estimators

The **k-th moment**, $\mathrm{E}\left[X^k\right] = \mu'_k$, is our first example of a population summary measure. It is estimated with the corresponding sample statistic

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k .$$

In typical applications, k is a positive integer, although it need not be. For the first moment ($k = 1$), the prime symbol ($'$) and the 1 subscript are usually dropped, using $\mu = \mu'_1$ to denote the **mean**. The corresponding sample estimator for μ is called the **sample mean**, denoted with a bar on top of the random variable:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

Sometimes, μ'_k is called the k-th raw moment to distinguish it from the **k-th central moment**, $\mathrm{E}\left[(X-\mu)^k\right] = \mu_k$, which is estimated as

$$\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^k .$$

The second central moment ($k = 2$) is an important case for which we typically assign a new symbol, $\sigma^2 = \mathrm{E}\left[(X-\mu)^2\right]$, known as the **variance**. The corresponding sample estimator for σ^2 is called the **sample variance**.

Empirical Distribution Function

To estimate the distribution function nonparametrically, we define the **empirical distribution function** to be

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

Here, the notation $I(\cdot)$ is the indicator function; it returns 1 if the event $(\cdot)$ is true and 0 otherwise.

Example – 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 34.45556. The corresponding empirical distribution function is

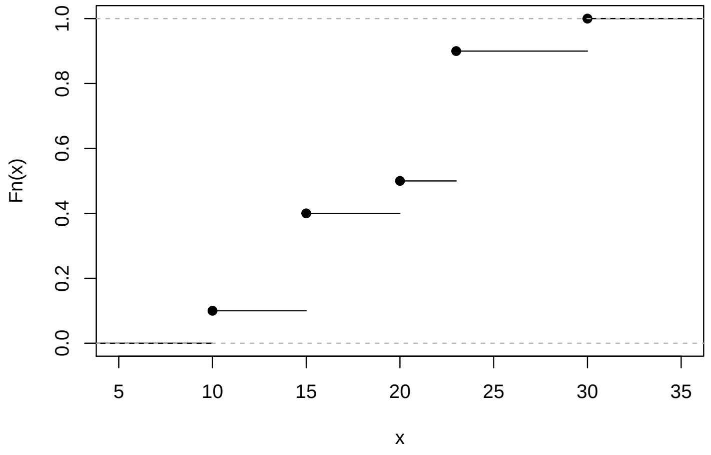


Figure 4.1: Empirical Distribution Function of a Toy Example

$$F_n(x) = \begin{cases} 0 & \text{for } x < 10 \\ 0.1 & \text{for } 10 \leq x < 15 \\ 0.4 & \text{for } 15 \leq x < 20 \\ 0.5 & \text{for } 20 \leq x < 23 \\ 0.9 & \text{for } 23 \leq x < 30 \\ 1 & \text{for } x \geq 30, \end{cases}$$

which is shown in the following graph in Figure 4.1.

Show R Code

```
(xExample <- c(10,rep(15,3),20,rep(23,4),30))
PercentilesxExample <- ecdf(xExample)
plot(PercentilesxExample, main="",xlab="x")
```

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 from a continuous variable, 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.

To be precise, for a given $0 < q < 1$, define the q**th quantile** q_F to be any number that satisfies

$$F(q_F-) \leq q \leq F(q_F) \tag{4.1}$$

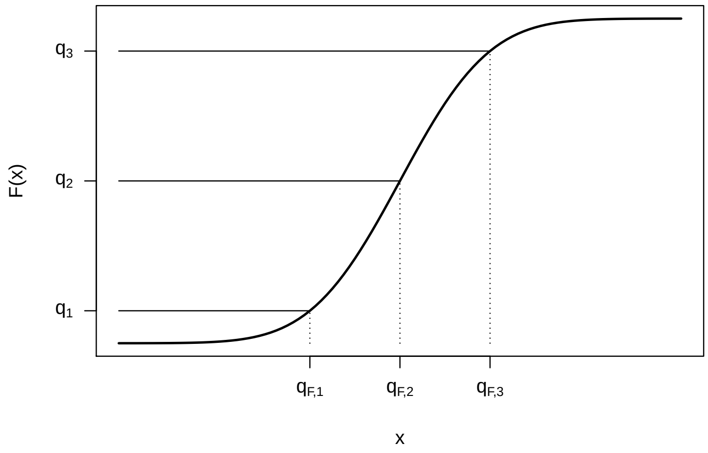


Figure 4.2: Continuous Quantile Case

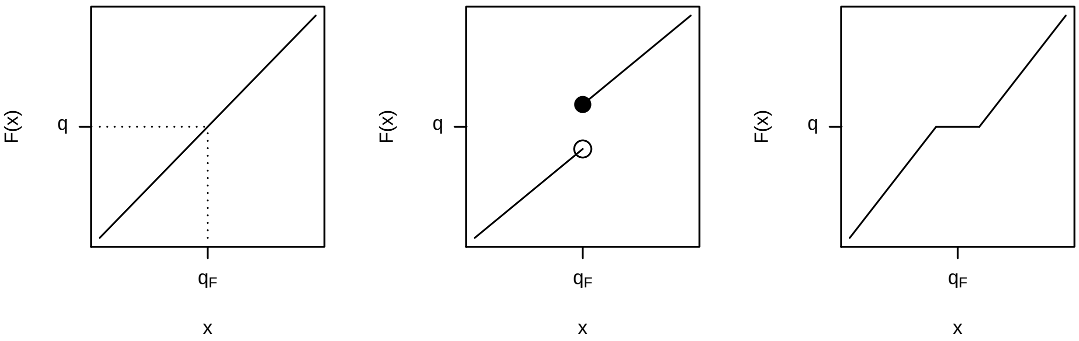


Figure 4.3: Figure 3: Three Quantile Cases

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 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 can not be uniquely determined for the q shown as there is a range of q_F's satisfying equation (4.1).

Example – 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 (the **order statistics**) corresponding to $X_1, \ldots, X_n$. Note that this is a linear interpolation between $X_{(j)}$ and $X_{(j+1)}$.

Example – 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

When the random variable is discrete, estimating the probability mass function $f(x) = \Pr(X = x)$ is straightforward. We simply use the empirical average, defined to be

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x).$$

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 $[c_{j-1}, c_j)$.

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 realizations of a random variable with mean

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

as $b \rightarrow 0$. 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** 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 the table below.

Table 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 – 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.

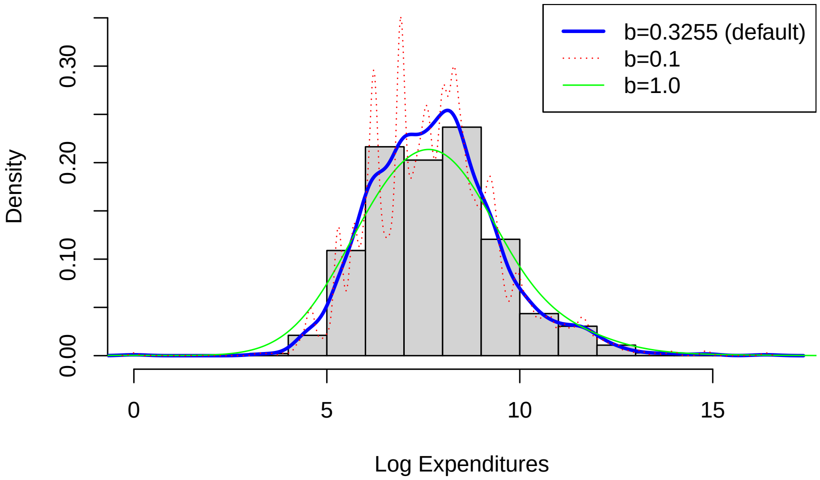


Figure 4.4: Figure 4: Histogram of Logarithmic Property Claims with Superimposed Kernel Density Estimators

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

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

Exercise – Exam C Question 3. You study five lives to estimate the time from the onset of a disease to death. The times to death are:

$$2 \quad 3 \quad 3 \quad 3 \quad 7$$

Using a triangular kernel with bandwith 2, calculate the density function estimate at 2.5.

Show Solution

Solution: For the kernel density estimate, we have

$$f_n(x) = \frac{1}{nb}\sum_{i=1}^{n} w\left(\frac{x - X_i}{b}\right),$$

where $n = 5$, $b = 2$, and $x = 2.5$. For the triangular kernel, $w(x) = (1 - |x|) \times I(|x| \leq 1)$. Thus,

X_i	$\frac{x-X_i}{b}$	$w\left(\frac{x-X_i}{b}\right)$
2	$\frac{2.5-2}{2} = \frac{1}{4}$	$(1-\frac{1}{4})(1) = \frac{3}{4}$
3	$\frac{2.5-3}{2} = \frac{-1}{4}$	$\left(1 - \left\lvert\frac{-1}{4}\right\rvert\right)(1) = \frac{3}{4}$
3		
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

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.

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 on these figures, the Pareto distribution is the clear choice for the analyst.

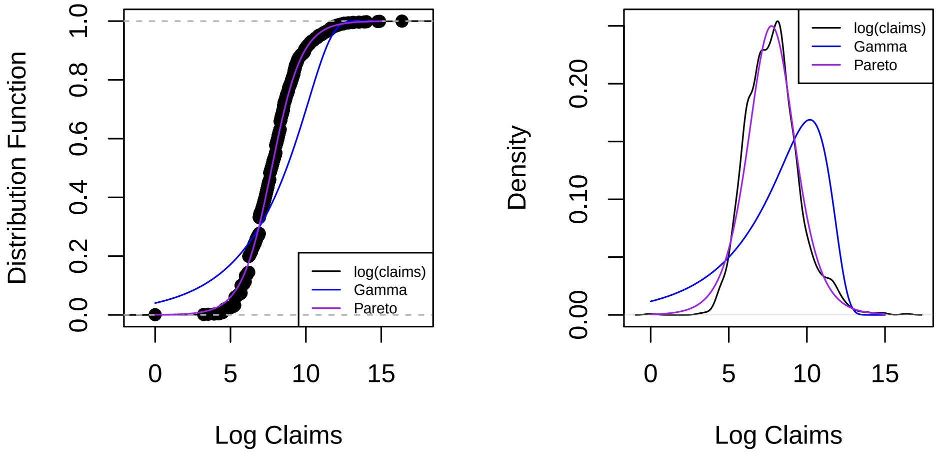


Figure 4.5: Figure 5: Nonparametric Versus Fitted Parametric Distribution and Density Functions. The left-hand panel compares distribution functions, with the dots corresponding to the empirical distribution, the thick blue curve corresponding to the fitted gamma and the light purple curve corresponding to the fitted Pareto. The right hand panel compares these three distributions summarized using probability density functions.

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. 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 they 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 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.

Exercise – Exam C Question 59. The graph below shows a *pp* plot of a fitted distribution compared to

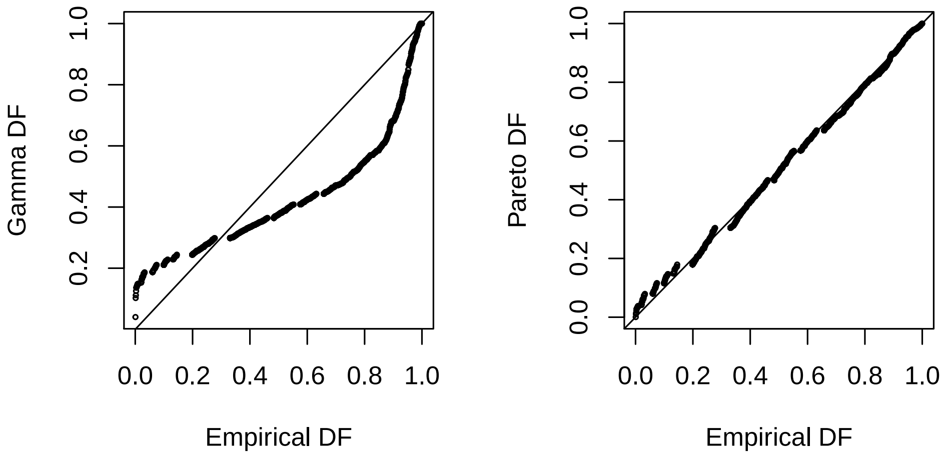


Figure 4.6: Figure 6: Probability-Probability (*pp*) Plots. The horizontal axes gives the empirical distribution function at each observation. In the left-hand panel, the corresponding distribution function for the gamma is shown in the vertical axis. The right-hand panel shows the fitted Pareto distribution. Lines of $y = x$ are superimposed.

a sample.

Comment on the two distributions with respect to left tail, right tail, and median probabilities.

Show 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 2 `\@ref(tab:GoFstats)` provides three commonly used goodness of fit statistics. Here, F_n is the empirical distribution and F is the fitted distribution.

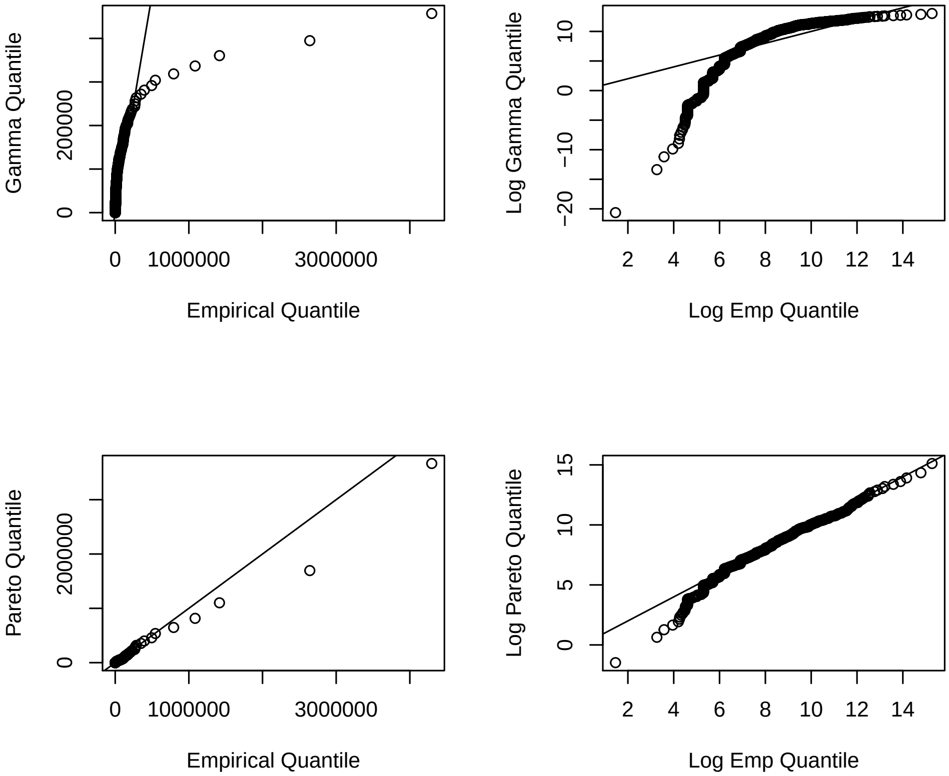


Figure 4.7: Figure 7: Quantile-Quantile (qq) Plots. The horizontal axes gives the empirical quantiles at each observation. The right-hand panels they are graphed on a logarithmic basis. The vertical axis gives the quantiles from the fitted distributions; Gamma quantiles are in the upper panels, Pareto quantiles are in the lower panels.

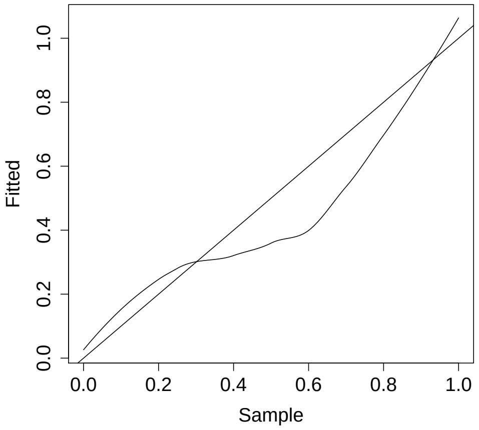


Table 2: Three Goodness of Fit Statistics

Statistic	Definition	Computational Expre
Kolmogorov-Smirnov	$\max_x \lvert F_n(x) - F(x)\rvert$	$\max(D^+, D^-)$ whe $D^+ = \max_{i=1,\ldots,n} \left\lvert \frac{i}{n}\right.$ $D^- = \max_{i=1,\ldots,n} \left\lvert F_i\right.$
Cramer-von Mises	$n\int (F_n(x) - F(x))^2 f(x)dx$	$\frac{1}{12n} + \sum_{i=1}^n \left(F_i - (2i -\right.$
Anderson-Darling	$n\int \frac{(F_n(x)-F(x))^2}{F(x)(1-F(x))} f(x)dx$	$-n - \frac{1}{n}\sum_{i=1}^n (2i-1)\log\left(F_i(\right.$

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).

Exercise – Exam C Question 40 (modified). A sample of claim payments is:

29 64 90 135 182

Compare the empirical claims distribution to an exponential distribution with mean 100 by calculating the value of the Kolmogorov-Smirnov test statistic.

Show 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. 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 – 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

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 for the infinite moment situation, 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.

Show Example

Example – 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, $0 < q < 1$.

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.

Exercise – Exam C Question 1. You are given:

(i) Losses follow a loglogistic distribution with cumulative distribution function:

$$F(x) = \frac{(x/\theta)^{\gamma}}{1+(x/\theta)^{\gamma}}$$

(ii) The sample of losses is:

10 35 80 86 90 120 158 180 200 210 1500

Calculate the estimate of θ by percentile matching, using the 40th and 80th empirically smoothed percentile estimates.

Show 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 *gamma*:

$$0.4 = \frac{(89.2/\theta)^\gamma}{1 + (89.2/\theta)^\gamma} \quad \text{and} \quad 0.8 = \frac{(206/\theta)^\gamma}{1 + (206 + \theta)^\gamma}$$

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 $4^{1/2.1407} = 206/\theta \Rightarrow \theta = 107.8$

4.2 Model Validation

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
- Describe the Gini statistic for model selection

This section revisits 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 summarizing the process of selecting a model based on the dataset at hand, we will focus on the process of validating the selected model by applying it to a different dataset.

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. 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. The second step in the iteration is based on the assumptions of the specified model. These assumptions must be consistent with the data to make valid use of the model. The third step, **diagnostic checking**, is also known as data and model criticism; 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.

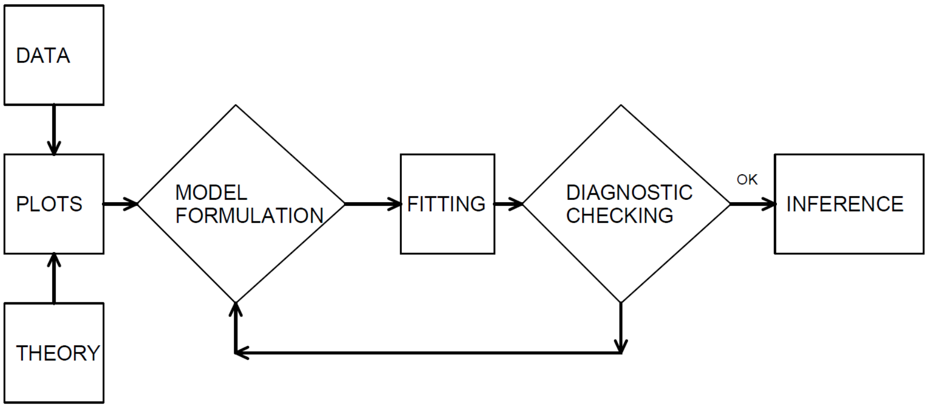


Figure 4.8: The iterative model specification process.

4.2.2 Summarizing Model Selection

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.

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.

For comparing model fits, if one model is a subset of another, then a **likelihood ratio test** may be employed. Generally, models are not proper subsets of one another so overall goodness of fit statistics, summarized in Section 1.2 `\ref{S:NonparametricModelSelection}`, are useful for model comparison. For discrete data, a **chi-square goodness of fit statistic** is generally preferred as it is more intuitive and simpler to explain.

Information criteria, such as Akaike's Information Criterion (**AIC**) and the Schwarz Bayesian Criterion (**BIC**) are widely cited because they can be readily generalized to multivariate settings.

Finally, a likelihood statistic that we have not yet considered is **Vuong's test**. This statistic is gaining popularity among analysts because it is a likelihood based statistic that has the ability to compare models that are non-nested.

4.2.3 Out of Sample Validation

Model validation is the process of confirming that the proposed model is appropriate, especially in light of the purposes of the investigation. An important criticism of the model selection process is that it can be

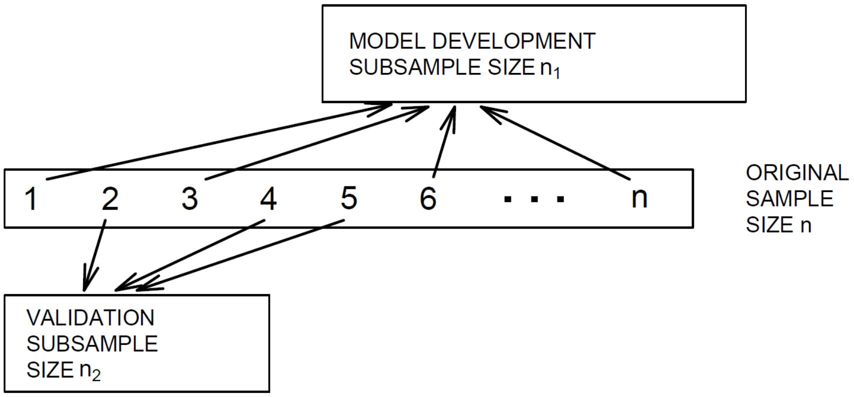


Figure 4.9: Model Validation. A data set of size n is randomly split into two subsamples.

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.

We can respond to this criticism by using a technique called **out-of-sample validation**. The ideal situation is to have available two sets of data, one for model development and one for model validation. We initially develop one or several models on a first data set. The models developed from the first set of data are called our candidate models. Then, the relative performance of the candidate models could be measured on a second set of data. In this way, the data used to validate the model is unaffected by the procedures used to formulate the model.

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 two subsamples. We call these the **model development subsample** and **validation subsample**, 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. As a rule of thumb, for data sets with 100 or fewer observations, use about 25-35% of the sample for out-of-sample validation. For data sets with 500 or more observations, use 50% of the sample for out-of-sample validation.

Because of these criticisms, several variants of the basic out-of-sample validation process are used by analysts. Although there is no theoretically best procedure, it is widely agreed that model validation is an important part of confirming the usefulness and appropriateness of a model.

4.2.4 Gini Statistic

The Classic Lorenz Curve

In welfare economics, it is common to compare distributions via the **Lorenz curve**, developed by Max Otto Lorenz (**?**). 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**. 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.

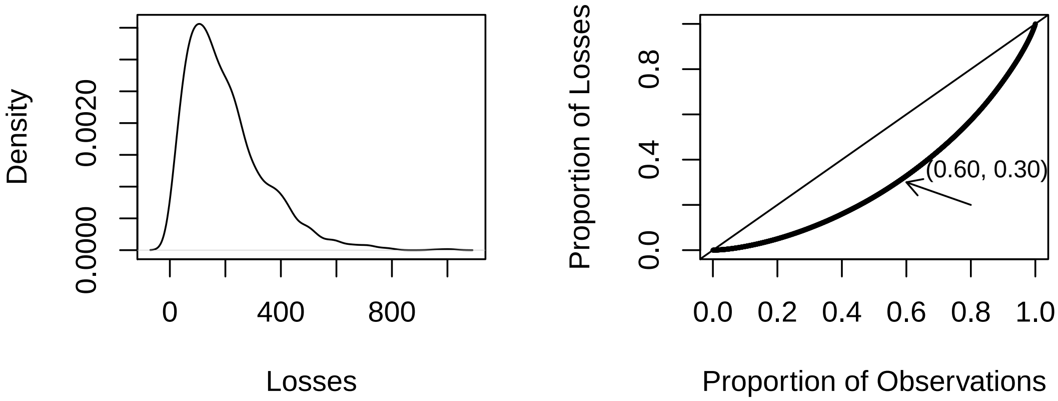


Figure 4.10: Distribution of insurance losses.

Example – Classic Lorenz Curve. For an insurance example, Figure 4.10 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, 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.

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

$$\hat{F}_P(s) = \frac{\sum_{i=1}^n P(\mathbf{x}_i) \mathrm{I}(R_i \leq s)}{\sum_{i=1}^n P(\mathbf{x}_i)}, \tag{4.4}$$

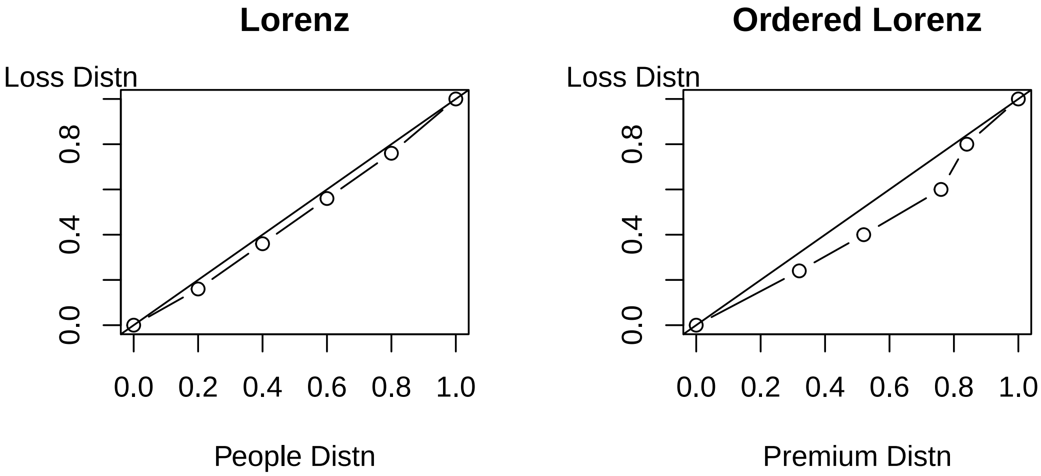


Figure 4.11: Lorenz versus Ordered Lorenz Curve

and the loss distribution is

$$\hat{F}_L(s) = \frac{\sum_{i=1}^n y_i \mathrm{I}(R_i \leq s)}{\sum_{i=1}^n y_i}, \tag{4.5}$$

where $\mathrm{I}(\cdot)$ 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 4.11 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

$$\begin{aligned}\widehat{Gini} &= 2\sum_{j=0}^{n-1}(a_{j+1} - a_j)\left\{\frac{a_{j+1} + a_j}{2} - \frac{b_{j+1} + b_j}{2}\right\} \\ &= 1 - \sum_{j=0}^{n-1}(a_{j+1} - a_j)(b_{j+1} + b_j). \qquad (4.6)\end{aligned}$$

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%. □

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. Use an estimated competing model and validation sample explanatory variables 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.

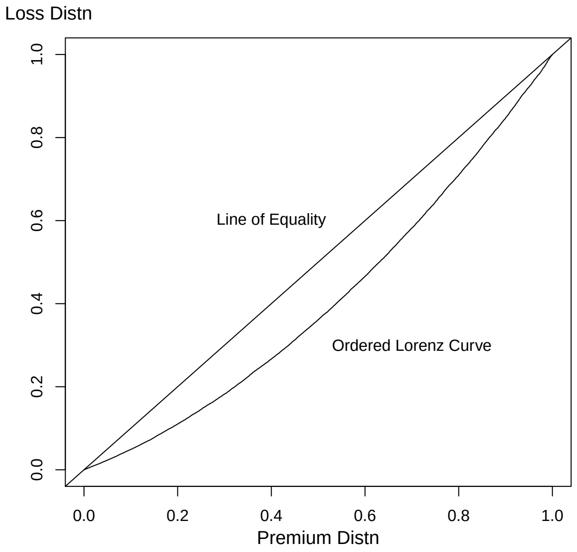
Suppose that we have experience from 25 states. For each state, we have available 500 observations that can be used to predict future losses. For this illustration, we have generated losses using a gamma distrbution with common shape parameter equal to 5 and a scale parameter that varies by state, from a low of 20 to 66.

Determine the ordered Lorenz curve and the corresponding Gini statistic to compare the two rate procedures.

Show Example Solution

For our base premium, we simply use the maximum likelihood estimate assuming a common distribution among all states. For the gamma distribution, this turns out to be simply the average which for our simulation 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. Because this illustration uses means that vary by states, we anticipate this alternative rating procedure to be preferred to the community rating procedure. (Recall for the gamma distribution that the mean equals the shape times the scale or, 5 times the scale parameter, for our example.)

Out of sample claims were generated from the same gamma distribution, with 200 observations for each state. In the following, we have the ordered Lorenz curve.



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

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. 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 an ordering that produces to a large Gini index. Thus, 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."

The insurance version of the Gini statistic was developed by (**?**) and (**?**) where you can find formulas for the standard errors and other additional background information.

4.3 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. Chapter 3 introduced the concept of observations that are "modified" due to two common types of limitations: **censoring** and **truncation**. This section will address parametric estmation 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, but not the exact value. This is referred to as **grouped** or **interval-censored** data. For example, we may be looking at two successive years of annual employee records. People employed in the first year but not the second have left sometime during the year. With an exact departure date (individual data), we could compute the amount of time that they were with the firm. Without the departure date (grouped data), we only know that they departed sometime during a year-long interval.

Formalizing this idea, suppose there are k groups or intervals delimited by boundaries $c_0 < c_1 < \cdots < c_k$. For each observation, we only observe the interval into which it fell (e.g. (c_{j-1}, c_j)), not the exact value. Thus, we only know the number of observations in each interval. The constants $\{c_0 < c_1 < \cdots < c_k\}$ form some partition of the domain of $F(\cdot)$. Then the probability of an observation X_i falling in the jth interval is

$$\Pr\left(X_i \in (c_{j-1}, c_j]\right) = F(c_j) - F(c_{j-1}).$$

The corresponding probability mass function for an observation is

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

Now, define n_j to be the number of observations that fall in the jth interval, $(c_{j-1}, c_j]$. Thus, the likelihood function (with respect to the parameter(s) θ) is

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

And the log-likelihood function is

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

Maximizing the likelihood function (or equivalently, maximizing the log-likelihood function) would then produce the maximum likelihood estimates for grouped data.

Censored Data

Censoring occurs when we observe 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. Let C_U denote the censoring time, such as $C_U = 5$. With right-censored observations, we observe X if it is below censoring point C_U; otherwise if X is higher than the censoring point, we only observe the censored C_U. Therefore, we record $X_U^* = \min(X, C_U)$. We also observe whether or not censoring has occurred. Let $\delta_U = \mathrm{I}(X \geq C_U)$ be a binary variable that is 1 if censoring occurs, $y \geq C_U$, and 0 otherwise.

For example, C_U may represent the upper limit of coverage of an insurance policy. 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 only observe X if X is above censoring point (e.g. time or loss amount) C_L; otherwise we observe C_L. Thus, we record $X_L^* = \max(X, C_L)$ along with the censoring indicator $\delta_L = \mathrm{I}(X \leq C_L)$.

For example, suppose a reinsurer will cover insurer losses greater than C_L. Let $Y = X_L^* - C_L$ represent the amount that the reinsurer is responsible for. If the policyholder loss $X < C_L$, then the insurer will pay the entire claim and $Y = 0$, no loss for the reinsurer. If the loss $X \geq C_L$, then $Y = X - C_L$ represents the reinsurer's retained claims. If a loss occurs, the reinsurer knows the actual amount if it exceeds the limit C_L, otherwise it only knows that it had a loss of 0.

As another example of a left-censored observation, suppose we are conducting a study and interviewing a person about an event in the past. The subject may recall that the event occurred before C_L, but not the exact date.

Truncated Data

We just saw that censored observations are still available 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 tfhe amount is

$$Y = \begin{cases} \text{we do not observe } X & X < C_L \\ X - C_L & X \geq C_L. \end{cases}$$

In other words, if X is less than the threshold C_L, then it is not observed. FOr example, C_L may represent the deductible associated with an insurance coverage. If the insured loss is less than the deductible, then the insurer does not observe or record the loss at all. If the loss exceeds the deductible, then the excess $X - C_L$ is the claim that the insurer covers.

Similarly for **right-truncated** data, if X exceeds a threshold C_U, then it is not observed. In this case, the amount is

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

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 4.12 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).

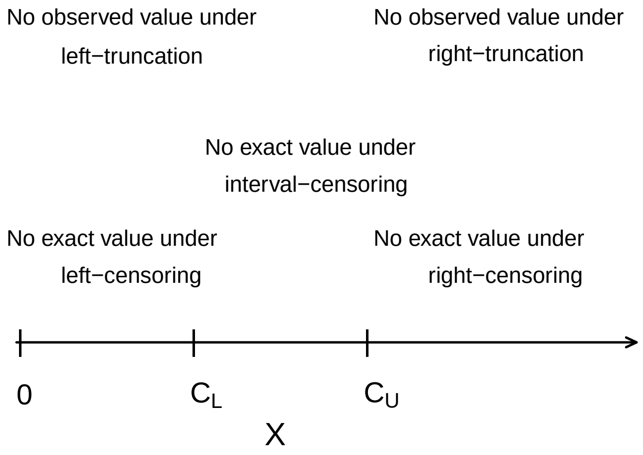


Figure 4.12: Figure 8: Censoring and Truncation

Show 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 4.13 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.
- **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	Censoring Information
right censoring	$X_U^* = \min(X, C_U)$	$\delta_U = \mathrm{I}(X \geq C_U)$
left censoring	$X_L^* = \max(y, C_L)$	$\delta_L = \mathrm{I}(X \leq C_L)$
interval censoring		
right truncation	X	observe X if $X < C_U$
left truncation	X	observe X if $X < C_L$

Parametric Estimation using Censored and Truncated Data

For simplicity, we assume fixed censoring times 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 $\delta_U = \mathrm{I}(X \geq C_U)$. If censoring occurs so that $\delta_U = 1$, then $X \geq C_U$ and the likelihood is $\Pr(X \geq C_U) = 1 - F(C_U)$. If censoring does not occur so that $\delta_U = 0$, then $X < C_U$ and the likelihood is $f(x)$. Summarizing, we have the likelihood of a single observation as

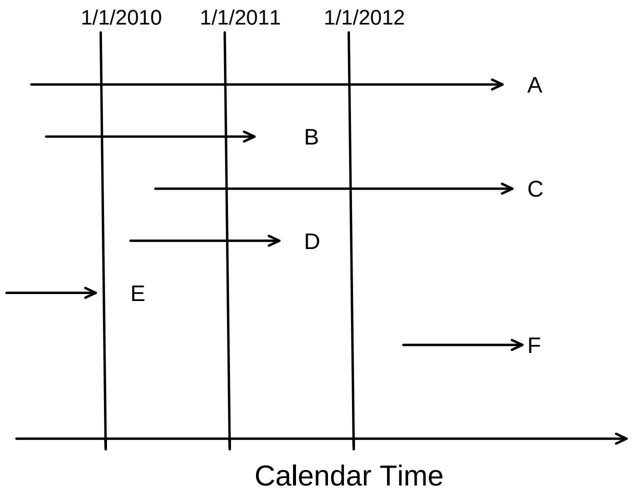


Figure 4.13: Figure 9: Timeline for Several Subjects on Test in a Mortality Study

On the other hand, truncated data are handled in likelihood inference via conditional probabilities. Specifically, we adjust the likelihood contribution by dividing by the probability that the variable was observed. To summarize, we have the following contributions to the likelihood function for six types of outcomes:

Outcome	Likelihood Contribution
exact value	$f(x)$
right-censoring	$1-F(C_U)$
left-censoring	$F(C_L)$
right-truncation	$f(x)/F(C_U)$
left-truncation	$f(x)/(1-F(C_L))$
interval-censoring	$F(C_U)-F(C_L)$

For known outcomes and censored data, the likelihood is

$$\mathcal{L}(\theta)=\prod_E f(x_i)\prod_R\{1-F(C_{Ui})\}\prod_L F(C_{Li})\prod_I(F(C_{Ui})-F(C_{Li})),$$

where "$\prod_E$" is the product over observations with Exact values, and similarly for Right-, Left- and Interval-censoring.

For right-censored and left-truncated data, the likelihood is

$$\mathcal{L}(\theta)=\prod_E\frac{f(x_i)}{1-F(C_{Li})}\prod_R\frac{1-F(C_{Ui})}{1-F(C_{Li})},$$

and similarly for other combinations. To get further insights, consider the following.

Show Example

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 $\delta_i = \mathrm{I}(X_i \geq C_{Ui})$ 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

$$
L(\theta) = -\sum_{i=1}^{n} ((1-\delta_i)\ln\theta + \frac{x_i^{**}}{\theta}) \tag{4.7}
$$

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.

Exercise – Exam C Question 44. 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 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 $\hat{\theta} = \frac{-1000}{\ln(20/33)} = 1996.90$.

Exercise – Exam C Question 152. 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 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

$$L'(\theta) = -3\theta^{-1} + 700\theta^{-2} = 0 \ \Rightarrow\ \hat{\theta} = \frac{700}{3} = 233.33$$

Course 4: Fall 2000, Question 22. 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 Solution

Solution: The likelihood function is

$$\begin{aligned}\mathcal{L}(\theta) &= f(20)f(30)f(45)[1-F(50)]^2 \\ &= \frac{2(20/\theta)^2e^{-(20/\theta)^2}}{20}\frac{2(30/\theta)^2e^{-(30/\theta)^2}}{30}\frac{2(45/\theta)^2e^{-(45/\theta)^2}}{45}(e^{-(50/\theta)^2})^2 \\ &\propto \frac{1}{\theta^6}e^{-8325/\theta^2}\end{aligned}$$

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

$$f_n(x) = F_n'(x) = \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} \quad \text{for } c_{j-1} \leq x < c_j.$$

Exercise – Exam C Question 195. 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 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 likelihood methods with nonparametric methods that do not rely on a parametric form of the distribution. The product-limit estimator due to (**?**) is a well-known estimator of the distribution in the presence of censoring.

To begin, first note that the empirical distribution function $F_n(x)$ is an **unbiased** estimator of the distribution function $F(x)$ (in the "usual" case in the absence of censoring). This is because $F_n(x)$ is the average of indicator variables that are also unbiased, that is, $\mathrm{E}\ I(X \leq x) = \Pr(X \leq x) = F(x)$. 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, $\mathrm{E}\ I(X^* > x) = 1 - F(x) = S(x)$.

Now consider two random variables that have different censoring limits. For illustration, suppose that we observe $X_1^* = \min(X_1, 5)$ and $X_2^* = \min(X_2, 10)$ where X_1 and X_2 are independent draws from the same distribution. For $x \leq 5$, the empirical distribution function $F_2(x)$ is an unbiased estimator of $F(x)$. However, for $5 < x \leq 10$, the first observation cannot be used for the distribution function because of the censoring limitation. Instead, the strategy developed by (**?**) is to use $S_n(5)$ as an estimator of $S(5)$ and then to use the second observation to estimate the conditional survivor function $\Pr(X > x | X > 5) = \frac{S(x)}{S(5)}$. Specifically, for $5 < x \leq 10$, the estimator of the survival function is

$$\hat{S}(x) = S_2(5) \times I(X_2^* > x).$$

Extending this idea, for each observation i, let u_i be the upper censoring limit ($= \infty$ if no censoring). Thus, the recorded value is x_i in the case of no censoring and u_i if there is censoring. Let $t_1 < \cdots < t_k$ be k distinct points at which an uncensored loss occurs, and let s_j be the number of uncensored losses x_i's at t_j. The corresponding **risk set** is the number of observations that are active (not censored) at a value less than t_j, denoted as $R_j = \sum_{i=1}^n I(x_i \geq t_j) + \sum_{i=1}^n I(u_i \geq t_j)$.

With this notation, the **product-limit estimator** of the distribution function is

$$\hat{F}(x) = \begin{cases} 0 & x < t_1 \\ 1 - \prod_{j:t_j \leq x} \left(1 - \frac{s_j}{R_j}\right) & x \geq t_1. \end{cases} \tag{4.8}$$

As usual, the corresponding estimate of the survival function is $\hat{S}(x) = 1 - \hat{F}(x)$.

Exercise – Exam C Question 252. 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 Solution

Solution: There are four event times (non-censored observations). For each time t_j, we can calcuate 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 (4.8).

(**?**) derived the formula for the estimated variance of the product-limit estimator to be

$$\widehat{Var}(\hat{F}(x)) = (1 - \hat{F}(x))^2 \sum_{j:t_j \leq x} \frac{s_j}{R_j(R_j - s_j)}.$$

`R`'s `survfit` method takes a survival data object and creates a new object containing the Kaplan-Meier estimate of the survival function along with confidence intervals. The Kaplan-Meier method (`type='kaplan-meier'`) is used by default to construct an estimate of the survival curve. The resulting discrete survival function has point masses at the observed event times (discharge dates) t_j, where the probability of an event given survival to that duration is estimated as the number of observed events at the duration s_j divided by the number of subjects exposed or 'at-risk' just prior to the event duration R_j.

Two alternate types of estimation are also available for the `survfit` method. The alternative (`type='fh2'`) handles ties, in essence, by assuming that multiple events at the same duration occur in some arbitrary order. Another alternative (`type='fleming-harrington'`) uses the Nelson-Äalen (see (**?**)) estimate of the **cumulative hazard function** to obtain an estimate of the survival function. The estimated cumulative hazard $\hat{H}(x)$ starts at zero and is incremented at each observed event duration t_j by the number of events s_j divided by the number at risk R_j. With the same notation as above, the **Nelson-Äalen** estimator of the distribution function is

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

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

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

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

Exercise – Exam C Question 135. 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 Solution

Solution: Recall the risk set $R_j = \sum_{i=1}^n \{I(x_i \geq t_j) + I(u_i \geq t_j) - I(d_i \geq t_j)\}$. Then

j	t_j	s_j	R_j	$\hat{S}(t_j)$
1	0.9	1	$10-3=7$	$1-\frac{1}{7}=\frac{6}{7}$
2	1.5	1	$8-2=6$	$\frac{6}{7}\left(1-\frac{1}{6}\right)=\frac{5}{7}$
3	1.7	1	$5-0=5$	$\frac{5}{7}\left(1-\frac{1}{5}\right)=\frac{4}{7}$
4	2.1	2	3	$\frac{4}{7}\left(1-\frac{2}{3}\right)=\frac{4}{21}$

The Kaplan-Meier estimate is therefore $\hat{S}(1.6) = \frac{5}{7}$.

Exercise – Exam C Question 252. - 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 Solution

Solution: As before, there are four event times (non-censored observations). For each time t_j, we can calcuate the number of events s_j and the risk set R_j as the following:

j	t_j	s_j	R_j
1	4	2	10
2	8	1	5
3	12	1	2
4	15	1	1

The Nelson-Äalen estimate of $S(11)$ is $\hat{S}_{NA}(11) = e^{-\hat{H}(11)} = e^{-0.4} = 0.67$, since

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

From earlier work, the Kaplan-Meier estimate of $S(11)$ is $\hat{S}(11) = 0.64$. Then Greenwood's estimate of the variance of the product-limit estimate of $S(11)$ is

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

4.4 Bayesian Inference

In this section, you learn how to:

- Describe the Bayes model as an alternative to the frequentist approach and summarize the five components of this modeling approach
- Describe the Bayesian decision framework and its role in determining Bayesian predictions
- Determine posterior predictions

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.

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, 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. Third, 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. Fourth, Bayesian analysis is particularly useful for forecasting future responses.

4.4.1 Bayesian Model

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 $\boldsymbol{\theta}$ 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.

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

$$f(x) = f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

This is analogous to a frequentist mixture distribution.

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.

We can summarize the distribution using a confidence interval type statement.

Definition. $[a,b]$ is said to be a $100(1-\alpha)\%$ **credibility interval** for $\boldsymbol{\theta}$ if

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

Exercise – Exam C Question 157. 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 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

$$P(0.7 < q < 0.8|1,0) = \int_{0.7}^{0.8} \frac{q^4-q^5}{0.014069}dq = 0.5572$$

Exercise – Exam C Question 43. 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 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

$$P(\Theta > 2|3) = \int_2^\infty 32(3+\theta)^{-3}d\theta = -16(3+\theta)^{-2}\Big|_2^\infty = \frac{16}{25} = 0.64$$

4.4.2 Decision Analysis

In classical decision analysis, 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 that value that minimizes the expected loss E $[l(\hat{\theta}, \theta)]$.

Some important special cases include:

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

For new data y, the predictive distribution is

$$f(y|x) = \int f(y|\theta)\pi(\theta|x)d\theta.$$

With this, the **Bayesian prediction** of y is

$$\begin{aligned}\mathrm{E}(y|x) &= \int yf(y|x)dy = \int y\left(\int f(y|\theta)\pi(\theta|x)d\theta\right)dy \\ &= \int \mathrm{E}(y|\theta)\pi(\theta|x)d\theta.\end{aligned}$$

Exercise – Exam C Question 190. 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 estimate (Bühlmann credibility estimate) of the number of claims in Year 2.

Show Solution

Solution: Note that $E(\theta) = 0.05(0.8) + 0.3(0.2) = 0.1$ and $E(\theta^2) = 0.05^2(0.8) + 0.3^2(0.2) = 0.02$

We also have $\mu(\theta) = 0(2\theta) + 1(\theta) + 2(1 - 3\theta) = 2 - 5\theta$ and $v(\theta) = 0^2(2\theta) + 1^2(\theta) + 2^2(1 - 3\theta) - (2 - 5\theta)^2 = 9\theta - 25\theta^2$.

Thus

$$\begin{aligned}
\mu &= E(2-5\theta) = 2-5(0.1) = 1.5\\
v &= EVPV = E(9\theta - 25\theta^2) = 9(0.1) - 25(0.02) = 0.4\\
a &= VHM = Var(2-5\theta) = 25Var(\theta) = 25(0.02-0.1^2) = 0.25\\
\Rightarrow k &= \frac{v}{a} = \frac{0.4}{0.25} = 1.6\\
\Rightarrow Z &= \frac{1}{1+1.6} = \frac{5}{13}
\end{aligned}$$

Therefore, $P = \frac{5}{13}2 + \frac{8}{13}1.5 = 1.6923$.

Exercise – Exam C Question 11. 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 Solution

Solution: We are given the prior distribution of θ as $P(\theta = 1) = P(\theta = 3) = \frac{1}{2}$, the conditional distribution $f(x|\theta)$, and the fact that we observed $X_1 = 5$. The goal is to find the predictive probability $P(X_2 > 8|X_1 = 5)$.

The posterior probabilities are

$$\begin{aligned}
P(\theta = 1|X_1 = 5) &= \frac{f(5|\theta=1)P(\theta=1)}{f(5|\theta=1)P(\theta=1) + f(5|\theta=3)P(\theta=3)}\\
&= \frac{\frac{1}{36}(\frac{1}{2})}{\frac{1}{36}(\frac{1}{2}) + \frac{3}{64}(\frac{1}{2})} = \frac{\frac{1}{72}}{\frac{1}{72} + \frac{3}{128}} = \frac{16}{43}
\end{aligned}$$

$$\begin{aligned}
P(\theta = 3|X_1 = 5) &= \frac{f(5|\theta=3)P(\theta=3)}{f(5|\theta=1)P(\theta=1) + f(5|\theta=3)P(\theta=3)}\\
&= 1 - P(\theta = 1|X_1 = 5) = \frac{27}{43}
\end{aligned}$$

Note that the conditional probability that losses exceed 8 is

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

The predictive probability is therefore

$$\begin{aligned}
P(X_2 > 8|X_1 = 5) &= P(X_2 > 8|\theta = 1)P(\theta = 1|X_1 = 5) + P(X_2 > 8|\theta = 3)P(\theta = 3|X_1 = 5)\\
&= \frac{1}{8+1}\left(\frac{16}{43}\right) + \frac{3}{8+3}\left(\frac{27}{43}\right) = 0.2126
\end{aligned}$$

Exercise – Exam C Question 15. 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 Solution

Solution: The posterior probability density is

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

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

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

Exercise – Exam C Question 29. 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 Solution

Solution: The probabilities are from a binomial distribution with 6 trials in which 3 successes were observed.

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

The probability of observing three successes is

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

The three posterior probabilities are

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

The posterior probability of a claim is then

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

4.4.3 Posterior Distribution

How can we calculate the posterior distribution $\pi(\boldsymbol{\theta}|x) = \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(x)}$?

- **By hand:** we can do this in special cases
- **Simulation:** use modern computational techniques such as Markov Chain Monte Carlo (MCMC) simulation
- **Normal approximation:** !!! Theorem 12.39 of **KPW** provides a justification
- **Conjugate distributions:** classical approach. 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. We focus on this approach.

To relate the prior and posterior distributions of the parameters, we have

$$
\begin{array}{ccc}
\pi(\boldsymbol{\theta}|x) & = & \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(x)} \\
 & \propto & f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\
\text{Posterior} & \text{is proportional to} & \text{likelihood} \times \text{prior}
\end{array}
$$

For **conjugate distributions**, the posterior and the prior come from the same family of distributions.

Show Example

Example – Poisson-Gamma Assume a Poisson(λ) model distribution so that

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Assume λ follows a gamma(α, θ) prior distribution so that

$$\pi(\lambda) = \frac{(\lambda/\theta)^{\alpha} \exp(-\lambda/\theta)}{\lambda \Gamma(\alpha)}.$$

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 $\alpha_{new} = \sum_i x_i + \alpha$ and $\theta_{new} = 1/(n+1/\theta)$. Thus, the gamma distribution is a conjugate prior for the Poisson model distribution.

Exercise – Exam C Question 215. 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 credibility estimate 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 credibility estimate for the number of claims in Year 3 is 0.20.

Calculate θ.

Show Solution

Solution: Since the conditional distribution of the number of claims per policyholder, $E(X|\lambda) = Var(X|\lambda) = \lambda$

Thus,

$$\begin{aligned}
\mu = v &= E(\lambda) = \alpha\theta \\
a &= Var(\lambda) = \alpha\theta^2 \\
k &= \frac{v}{a} = \frac{1}{\theta} \\
\Rightarrow Z &= \frac{n}{n + 1/\theta} = \frac{n\theta}{n\theta + 1}
\end{aligned}$$

Using the credibility estimates given,

$$\begin{aligned}
0.15 &= \frac{\theta}{\theta+1}(1) + \frac{1}{\theta+1}\mu = \frac{\theta+\mu}{\theta+1} \\
0.20 &= \frac{2\theta}{2\theta+1}(2) + \frac{1}{2\theta+1}\mu = \frac{4\theta+\mu}{2\theta+1}
\end{aligned}$$

From the first equation, $0.15\theta + 0.15 = \theta + \mu \ \Rightarrow\ \mu = 0.15 - 0.85\theta$.

Then the second equation becomes $0.4\theta + 0.2 = 4\theta + 0.15 - 0.85\theta \ \Rightarrow\ \theta = 0.01818$

4.5 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.

Contributors

- **Lisa Gao** and **Edward W. (Jed) Frees**, University of Wisconsin-Madison, are the principal authors of the initital version of this chapter. Email: jfrees@bus.wisc.edu for chapter comments and suggested improvements.

Chapter 5

Aggregate Loss Models

This is a placeholder file

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	$B_1 =$	$\text{mod } (3 \times 1 + 2) = 5$	$U_1 = \frac{5}{15}$
2	$B_2 =$	$\text{mod } (3 \times 5 + 2) = 2$	$U_2 = \frac{2}{15}$
3	$B_3 =$	$\text{mod } (3 \times 2 + 2) = 8$	$U_3 = \frac{8}{15}$
4	$B_4 =$	$\text{mod } (3 \times 8 + 2) = 11$	$U_4 = \frac{11}{15}$

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

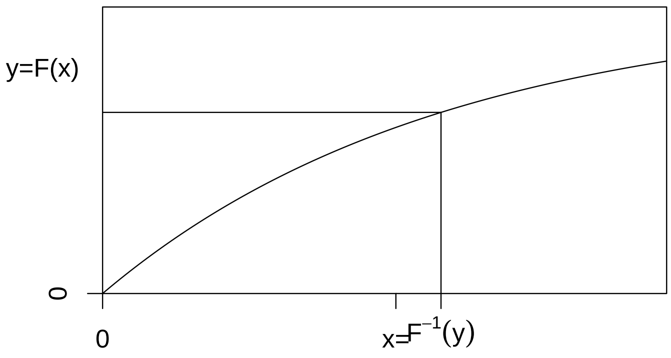


Figure 6.1: Inverse of a Distribution Function

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

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

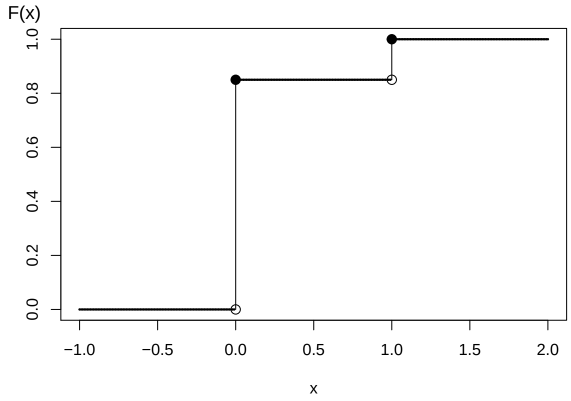


Figure 6.2: Distribution Function of a Binary Random Variable

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:

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

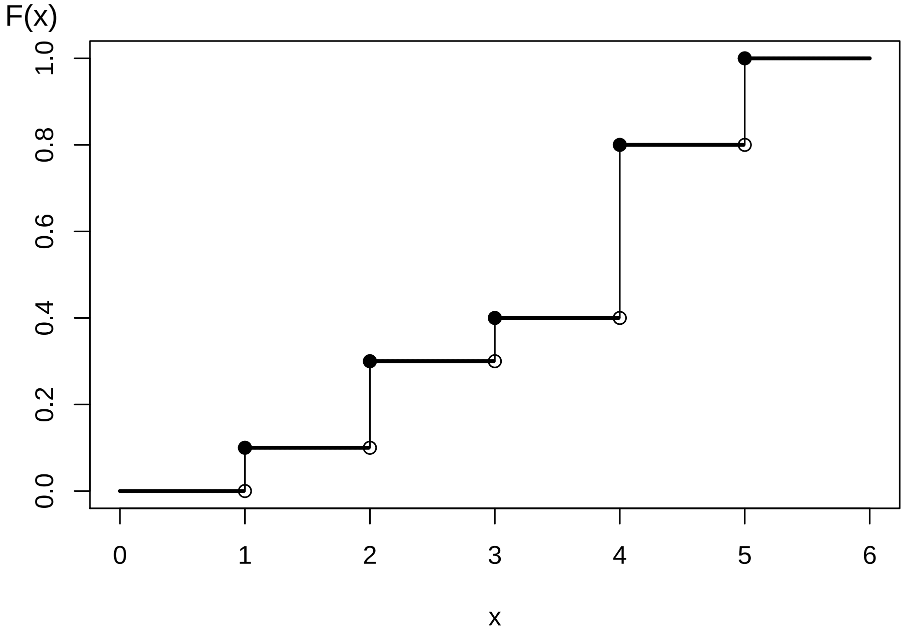


Figure 6.3: Distribution Function of a Discrete Random Variable

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

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

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

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

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

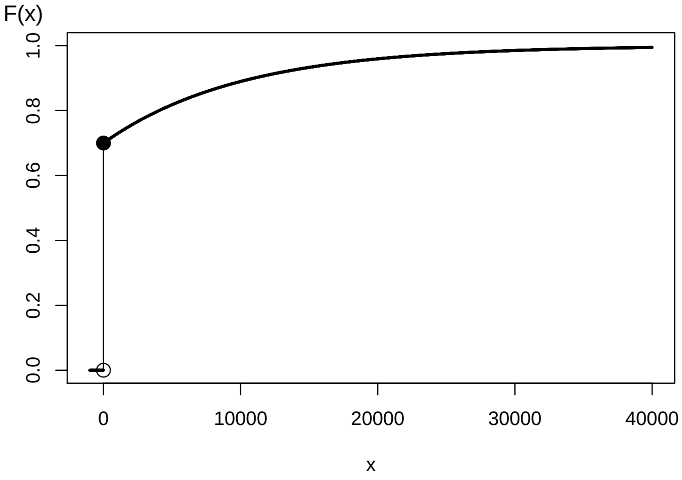


Figure 6.4: Distribution Function of a Hybrid Random Variable

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 Calcuations Fundamentals

This is a placeholder file

Chapter 8

Risk Classification

This is a placeholder file

8.1 Introduction

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 to the benefit outgo. Ignoring the frictional expenses associated with the administrative cost and the profit margin, the net premium thus should be equal to the expected loss occurring from the risk that is transferred from the policyholder to the insurer.

If all policyholders in the insurance pool have identical risk profiles, the insurer may simply charge the same premium for each policyholder because all of them would 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 of individual policyholders can enhance the ability of calculating a fair premium as they can be used to estimate or predict the expected losses more accurately at the individual level. Indeed, if the insurer do not differentiate the risk characteristics of individual policyholders and simply charges the same premium to all individuals based on the average characteristic of 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 industry 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 natually an average mortality loss accounting for both smokers and non-smokers; the average is a weighted one using the proportion of smokers and non-smokers. Thus it is easy to see that that a smoker would have a good incentive to purchase insurance from EquitabAll than 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. The result of this tendency for the EquitabAll's insurance portfolio is that there will be more smokers and less non-smokers in the pool, which leads to larger-than-expected mortality 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 the cycle continues over time, EquitabAll would gradually retain more smokers in its portfolio with the premium continually raised, and this vicious cycle eventually leads 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 each policyholder and the long term sustainability of an insurer.

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 process). All policyholders sharing identical risk factors thus are assigned to the same risk class, and are considered homogeneous; the insurer consequently charge them the same premium.

An important task in any risk classification is to construct a quantitative model that can determine the expected loss given various rating factors for a policyholder. The standard approach is to adopt a statistical regression model which produces the expected loss as the output when the relevant risk factors are given as the inputs. We introduce and discuss the Poisson regression, which can be used when the loss is a count variable, as a prominent example of an insurance pricing tool under risk classification schemes.

Chapter 9

Experience Rating using Credibility Theory

This is a placeholder file

Chapter 10

Portfolio Management including Reinsurance

10.0.1 Overview:

Define S to be (random) obligations that arise from a collection (portfolio) of insurance contracts

- We are particularly interested in probabilities of large outcomes and so formalize the notion of a heavy-tail distribution
- How much in assets does an insurer need to retain to meet obligations arising from the random S? A study of risk measures helps to address this question
- As with policyholders, insurers also seek mechanisms in order to spread risks. A company that sells insurance to an insurance company is known as a reinsurer

10.1 Tails of Distributions

In 1998 freezing rains fell on eastern Ontario, south-western Quebec and lasted for six days. The event doubled the amount of precipitation in the area experienced in any prior ice storm, and resulted in a catastrophe that produced excess of 840,000 cases of insurance claims. This number is 20% more than that of the claims caused by the Hurricane Andrew - one of the largest natural disasters in the history of North America. After all, the catastrophe caused approximately 1.44 billion Canadian dollars insurance settlements which is the highest loss burden in the history of Canada (Lecomte et al., 1998). More examples of similar catastrophic events that caused extremal insurance losses are Hurricanes Harvey and Sandy, the 2011 Japanese earthquake and tsunami, and so forth.

In the context of insurance, a few heavy 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 of course daunting. For instance, periods of financial stress may appear with higher frequency than the actuaries expect, and insurance losses may occur with worse severity. Therefore, studying the 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 (r.v.’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. Speaking plainly, a r.v. is said to be heavier-tailed if higher probabilities are assigned to larger values. Unwelcome outcomes are more likely to occur for an insurance portfolio that is described by a loss r.v. possessing heavier (right) tail. Tail weight can be an absolute or a relative concept. Specifically, for the former, we may consider a r.v. 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.

In the statistics and probability literature, there are several quantitative approaches have been proposed to compare and classify tail weight. Among most of these approaches, the survival functions serve as the building block. In what follows, we are going to introduce two simple yet useful tail classification methods, in which the basic idea is to study the quantities that are closely related to the survival function of X.

Classification Based on Moments

One possible 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/loss r.v. X to be positive. At the outset, let us recall that the $k-$th raw moment of X, for $k \in \mathcal{R}_+$, can be computed via

$$\mu_k^{'} = k \int_0^\infty x^{k-1} S(x) dx,$$

where $S(\cdot)$ denotes the survival function of X. It is a simple matter to see 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 the associated r.v. possesses. Hence the maximal order of finite moment, denoted by $k^* := \sup\{k \in \mathcal{R}_+ | \mu_k^{'} < \infty\}$, can be considered as an indicator of tail weight. This observation leads us to moment-based tail weight classification, which is defined formally next.

Definition 1. For 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 the $k^* = a \in (0, \infty)$, 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 $k_1^* \leq k_2^*$.

It is noteworthy that the first part of the aforementioned definition is an absolute concept of tail weight, while the second part is a relative concept that compares the weight of (right) tails between two distributions. Next, we are going to present a few examples that illustrate the applications of the moment-based method. Some of these examples are borrowed from Klugman et al., 2012.

Example 1. Let $X \sim Gamma(\alpha, \theta)$, with $\alpha > 0$ and $\theta > 0$, then for all $k \in \mathcal{R}_+$,

$$\begin{aligned}
\mu_k^{'} &= \int_0^\infty x^k \frac{x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha} dx \\
&= \int_0^\infty (y\theta)^k \frac{(y\theta)^{\alpha-1} e^{-y}}{\Gamma(\alpha)\theta^\alpha} \theta dy \\
&= \frac{\theta^k}{\Gamma(\alpha)} \Gamma(\alpha + k) < \infty.
\end{aligned}$$

Since all the positive moments exist, i.e., $k^* = \infty$, in accordance with the moment-based classification method in Definition 1, the gamma distribution is light-tailed.

Example 2. Let $X \sim \mathit{Weibull}(\theta, \tau)$, with $\theta > 0$ and $\tau > 0$, then for all $k \in \mathcal{R}_+$,

$$\begin{aligned}\mu_k^{'} &= \int_0^\infty x^k \frac{\tau x^{\tau-1}}{\theta^\tau} e^{-(x/\theta)^\tau} dx \\ &= \int_0^\infty \frac{y^{k/\tau}}{\theta^\tau} e^{-y/\theta^\tau} dy \\ &= \theta^k \Gamma(1 + k/\tau) < \infty.\end{aligned}$$

Again, due to the existence of all the positive moments, the Weibull distribution is light-tailed.

We notice in passing that the gamma and Weibull distributions have been used quite intensively in actuarial practice nowadays. 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 seem two examples of using the moment-based method to analyze light-tailed distributions. We document a heavy-tailed example in what follows.

Example 3. Let $X \sim \mathit{Pareto}(\alpha, \theta)$, with $\alpha > 0$ and $\theta > 0$, then for $k \in \mathcal{R}_+$

$$\begin{aligned}\mu_k^{'} &= \int_0^\infty x^k \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} dx \\ &= \alpha\theta^\alpha \int_\theta^\infty (y-\theta)^k y^{-(\alpha+1)} dy.\end{aligned}$$

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 $\mu_k^{'}$ is finite if and only if g_k is finite. Hence we can conclude that the raw moments of Pareto r.v.'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 moments depends only on the shape parameter α and it is an increasing function of α. In other words, based on the moments method, the tail weight of the Pareto r.v.'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 are going to conclude this current section by 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 straightforward to derive and/or the identification of the maximal order of finite moment can be very challenging. Second, the moment-based method does not well comply with main body of the well established heavy tail theory in literature. Specifically, the existence of moment generating functions (MGF's) is arguably the most popular method for classifying heavy tail verse light tail within the community of academic actuaries. However, for some r.v's such as the log normal r.v.'s, their MGF's do not exist even that all the positive moments are finite. In these cases, applications of the moment-based and the MFG-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.

Comparison Based on Limiting Tail Behavior

In order to resolve the shortfalls of the moment-based method discussed in the previous section, an alternative approach for comparing tail weight is to directly study the limiting behavior of the survival functions.

Definition 2. For two r.v.'s X and Y, and 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 $\gamma = c \in \mathcal{R}_+$;
- X has a lighter right tail than Y if $\gamma = 0$.

Example 4. Let $X \sim \mathit{Pareto}(\alpha, \theta)$ and $Y \sim \mathit{Weibull}(\tau, \theta)$, for $\alpha > 0$, $\tau > 0$, and $\theta > 0$, we have

$$\begin{aligned}
\lim_{t\to\infty} \frac{S_X(t)}{S_Y(t)} &= \lim_{t\to\infty} \frac{(1+t/\theta)^{-\alpha}}{\exp\{-(t/\theta)^\tau\}} \\
&= \lim_{t\to\infty} \frac{\exp\{t/\theta^\tau\}}{(1+t^{1/\tau}/\theta)^\alpha} \\
&= \lim_{t\to\infty} \frac{\sum_{i=0}^\infty \left(\frac{t}{\theta^\tau}\right)^i / i!}{(1+t^{1/\tau}/\theta)^\alpha} \\
&= \lim_{t\to\infty} \sum_{i=0}^\infty \left(t^{-i/\alpha} + \frac{t^{(1/\tau - i/\alpha)}}{\theta}\right)^{-\alpha} / \theta^{\tau i} i! \\
&= \infty.
\end{aligned}$$

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, the survival functions do not admit explicite expressions. In such cases, 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 5. Let $X \sim \mathit{Pareto}(\alpha, \theta)$ and $Y \sim \mathit{Gamma}(\alpha, \theta)$, for $\alpha > 0$ and $\theta > 0$, we have

$$\begin{aligned}
\lim_{t\to\infty} \frac{f_X(t)}{f_Y(t)} &= \lim_{t\to\infty} \frac{\alpha\theta^\alpha (t+\theta)^{-\alpha-1}}{t^{\tau-1} e^{-t/\lambda} \lambda^{-\tau} \Gamma(\tau)^{-1}} \\
&= c \lim_{t\to\infty} \frac{e^{t/\lambda}}{(t+\theta)^{\alpha+1} t^{\tau-1}} \\
&= \infty,
\end{aligned}$$

as exponentials go to infinity faster than polynomials.

10.2 Measures of Risk

- A **risk measure** is a mapping from the r.v. representing the loss associated with the risks to the real line.

- A risk measure gives a single number that is intended to quantify the risk.
  - For example, the standard deviation is a risk measure.
- Notation: $\rho(X)$.
- We briefly mention:
  - **VaR**: Value at Risk;
  - **TVaR**: Tail Value at Risk.

Value at Risk

- Say $F_X(x)$ represents the cdf of outcomes over a fixed period of time, e.g. one year, of a portfolio of risks.
- We consider positive values of X as losses.
- **Definition 3.11**: let X denote a loss r.v., then the **Value-at-Risk** of X at the $100p\%$ level, denoted $VaR_p(X)$ or π_p, is the $100p$ percentile (or quantile) of the distribution of X.
- E.g. for continuous distributions we have

$$P(X > \pi_p) = 1 - p.$$

- VaR has become the standard risk measure used to evaluate exposure to risk.
- **VaR** is the **amount of capital** required to ensure, with a **high degree of certainty**, that the **enterprise does not become technically insolvent**.
- Which degree of certainty?
  - 95%?
  - in Solvency II 99.5% (or: ruin probability of 1 in 200).
- **VaR is not subadditive**.
  - Subadditivity of a risk measure $\rho(.)$ requires

$$\rho(X+Y) \leq \rho(X) + \rho(Y).$$

  - Intuition behind subadditivity: combining risks is less riskier than holding them separately.
- **Example:** let X and Y be i.i.d. r.v.'s which are Bern(0.02) distributed.
  - Then, $P(X \leq 0) = 0.98$ and $P(Y \leq 0) = 0.98$. Thus, $F_X^{-1}(0.975) = F_Y^{-1}(0.975) = 0$.
  - For the sum, $X+Y$, we have $P[X+Y=0] = 0.98 \cdot 0.98 = 0.9604$. Thus, $F_{X+Y}^{-1}(0.975) > 0$.
  - VaR is not subadditive, since VaR$(X+Y)$ in this case is larger than VaR(X) + VaR(Y).
- Another **drawback of VaR**:
  - it is a single quantile risk measure of a predetermined level p;
  - no information about the thickness of the upper tail of the distribution function from VaR$_p$ on;
  - whereas stakeholders are interested in both frequency and severity of default.
- Therefore: study other risk measures, e.g. **Tail Value at Risk** (TVaR).

Tail Value at Risk

- **Definition 3.12:** let X denote a loss r.v., then the Tail Value at Risk of X at the $100p\%$ security level, TVaR(p), is the **expected loss given that the loss exceeds the** $100p$ **percentile** (or: quantile) of the distribution of X.
- We have (assume continuous distribution)

$$\begin{aligned}\text{TVaR}_p(X) &= E(X|X>\pi_p)\\ &= \frac{\int_{\pi_p}^{\infty} x\cdot f(x)dx}{1-F(\pi_p)}.\end{aligned}$$

- We can rewrite this as **the usual definition of TVaR**

$$\begin{aligned}\text{TVaR}_p(X) &= \frac{\int_{\pi_p}^{\infty} x dF_X(x)}{1-p}\\ &= \frac{\int_p^1 \text{VaR}_u(X)du}{1-p},\end{aligned}$$

  using the substitution $F_X(x)=u$ and thus $x=F_X^{-1}(u)$.
- From the definition

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

  we understand
  - TVaR is the **arithmetic average** of the quantiles of X, from level p on;
  - TVaR is averaging high level VaR;
  - TVaR **tells us much more about the tail** of the distribution than does VaR alone.
- Finally, TVaR can also be written as

$$\begin{aligned}\text{TVaR}_p(X) &= E(X|X>\pi_p)\\ &= \frac{\int_{\pi_p}^{\infty} xf(x)dx}{1-p}\\ &= \pi_p + \frac{\int_{\pi_p}^{\infty}(x-\pi_p)f(x)dx}{1-p}\\ &= \text{VaR}_p(X)+e(\pi_p),\end{aligned}$$

  with $e(\pi_p)$ the mean excess loss function evaluated at the $100p$th percentile.
- We can understand these connections as follows. (Assume continuous r.v.'s)
- The relation

$$\text{CTE}_p(X) = \text{TVaR}_{F_X(\pi_p)}(X),$$

  then follows immediately by combining the other two expressions.
- TVaR is a coherent risk measure, see e.g. Foundations of Risk Measurement course.
- Thus, $\text{TVaR}(X+Y)\leq\text{TVaR}(X)+\text{TVaR}(Y)$.
- When using this risk measure, we never encounter a situation where combining risks is viewed as being riskier than keeping them separate.

- **KPW Example 3.18** (Tail comparisons) Consider three loss distributions for an insurance company. Losses for the next year are estimated to be on average 100 million with standard deviation 223.607 million. You are interested in finding high quantiles of the distribution of losses. Using the normal, Pareto, and Weibull distributions, obtain the VaR at the 90%, 99%, and 99.99% security levels.

- **Solution**

- Normal distribution has a lighter tail than the others, and thus smaller quantiles.

- Pareto and Weibull with $\tau < 1$ have heavy tails, and thus relatively larger extreme quantiles.

- **Example 3.18** (Tail comparisons) Consider three loss distributions for an insurance company. Losses for the next year are estimated to be on average 100 million with standard deviation 223.607 million. You are interested in finding high quantiles of the distribution of losses. Using the normal, Pareto, and Weibull distributions, obtain the VaR at the 99%, 99.9%, and 99.99% security levels.

```
> qnorm(c(0.9,0.99,0.999),mu,sigma)
[1] 386.5639 620.1877 790.9976
> qpareto(c(0.9,0.99,0.999),alpha,s)
[1]  226.7830  796.4362 2227.3411
> qweibull(c(0.9,0.99,0.999),tau,theta)
[1]  265.0949 1060.3796 2385.8541
```

- We learn from Example 3.18 that results vary widely depending on the choice of distribution.

- Thus, the selection of an **appropriate loss model** is highly important.

- To obtain numerical values of VaR or TVaR:

  - estimate from the data directly;

  - or use distributional formulas, and plug in parameter estimates.

- When estimating VaR directly from the data:

  - use R to get quantile from the empirical distribution;

  - R has 9 ways to estimate a VaR at level p from a sample of size n, differing in the way the interpolation between order statistics close to np .

- When estimating TVaR directly from the data:

  - take average of all observations that exceed the threshold (i.e.π_p);

- **Caution:** we need a large number of observations (and a large number of observations $> \pi_p$) in order to get reliable estimates.

- When not may observations in excess of the threshold are available:

  - construct a loss model;

  - calculate values of VaR and TVaR directly from the fitted distribution.

- For example

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

  see Appendix A for those expressions.

10.3 Reinsurance

Recall that reinsurance is simply insurance purchased by an insurer. Insurance purchased by non-insurers is sometimes known as primary insurance 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 contract is 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. A reinsurance contract is also known as a treaty. Non-proportional agreements are simply everything else. As examples of non-proportional agreements, this chapter focuses on stop-loss and excess of loss contracts. For all types of agreements, we split the total risk S into the portion taken on by the reinsurer, $Y_{reinsurer}$, and that retained by the insurer, $Y_{insurer}$, that is, $S = Y_{insurer} + Y_{reinsurer}$.

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 $Y_{insurer} = cS$ is identical to a coinsurance adjustment in personal insurance. For stop-loss reinsurance, the transformation $Y_{reinsurer} = \max(0, S-M)$ is the same as an insurer's payment with a deductible M and $Y_{insurer} = \min(S,M) = S \wedge M$ 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 constrast, 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 chapter describes the foundational and most basic of reinsurance treaties: Section 10.1 for proportional and Section 10.2 for non-proportional. Section 10.3 gives a flavor of more complex contracts.

10.3.1 Proportional Reinsurance

The simplest example of a proportional treaty is called quota share.

- 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.

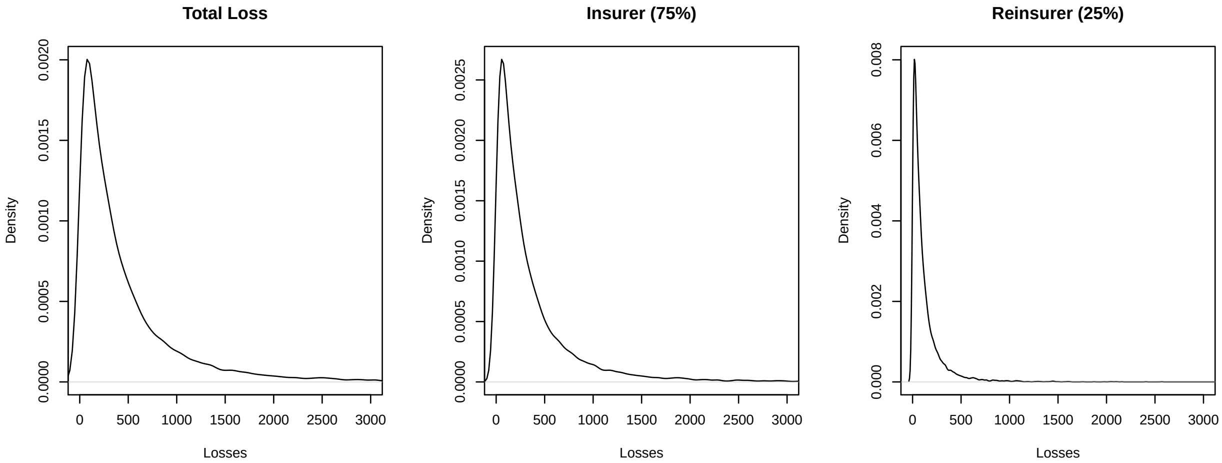
The amounts paid by the direct insurer and the reinsurer are summarized as

$$Y_{insurer} = cS \quad \text{and} \quad Y_{reinsurer} = (1-c)S.$$

Note that $Y_{insurer} + Y_{reinsurer} = S$.

Example. Distribution of Losses under Quota Share

To develop 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.



Click Here to see the R Code

```
set.seed(2018)
theta = 1000
alpha = 3
nSim = 10000
library(actuar)
S <-  rpareto(nSim, shape = alpha, scale = theta)

par(mfrow=c(1,3))
plot(density(S), xlim=c(0,3*theta), main="Total Loss", xlab="Losses")
plot(density(0.75*S), xlim=c(0,3*theta), main="Insurer (75%)", xlab="Losses")
plot(density(0.25*S), xlim=c(0,3*theta), 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 S such that

$$Y_{insurer} = g(S) \quad \text{and} \quad Y_{reinsurer} = S - g(S),$$

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, $Var((1-c)S) \leq Var(g(S))$ for all $g(.)$.

Click Here to see the Justification of the Proposition

Proof of the Proposition. With $Y_{reinsurer} = S - Y_{insurer}$ and the law of total variation

$$\begin{aligned} Var(Y_{reinsurer}) &= Var(S - Y_{insurer}) \\ &= Var(S) + Var(Y_{insurer}) - 2Cov(S, Y_{insurer}) \\ &= Var(S) + Q - 2Corr(S, Y_{insurer}) \times \sqrt{Q}\sqrt{Var(S)} \end{aligned}$$

In this expression, we see that Q and $Var(S)$ do not change with the choice of g. Thus, we can minimize $Var(Y_{reinsurer})$ by maximizing the correlation $Corr(S, Y_{insurer})$. If we use a quota share reinsurance agreement, then $Corr(S, Y_{insurer}) = Corr(S, (1-c)S) = 1$, 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 $S = X_1 + \cdots + X_n$. 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 uncertainty of the retained risks as measured by the variance.

Click Here to see 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 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

$$\frac{\partial}{\partial c_i} L = 2c_i\ Var\ X_i - \lambda\ E\ X_i = 0$$

so that

$$c_i = \frac{\lambda}{2} \frac{E\ X_i}{Var\ X_i}.$$

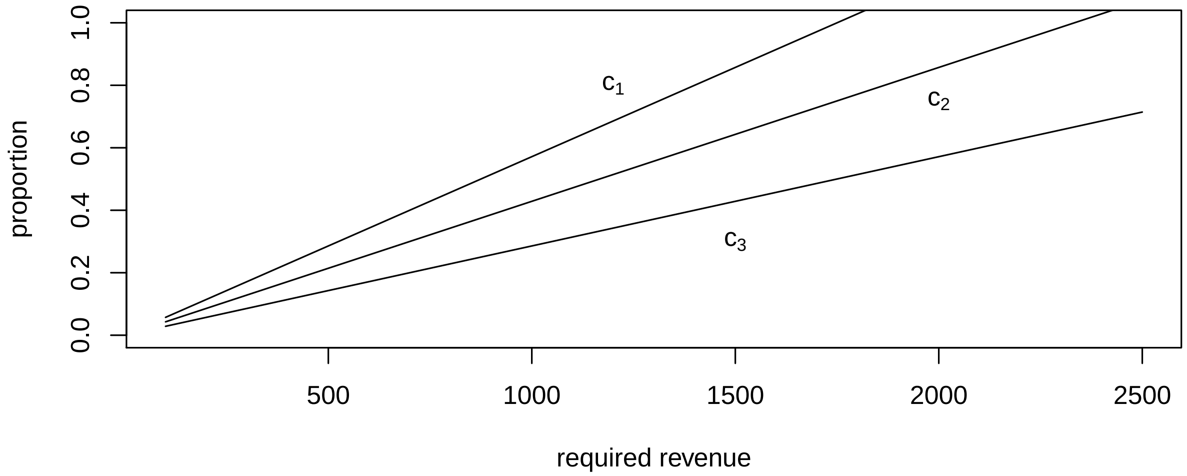
From the math, it turns out that the constant for the ith risk, c_i is proportional to $\frac{E\ X_i}{Var\ X_i}$. This is intuitively appealing. Other things being equal, a higher revenue as measured by $E\ X_i$ means a higher value of c_i. In the same way, a higher value of uncertainty as measured by $Var\ X_i$ 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.

Click Here to see an Example with Three Pareto Risks

Example. Consider three risks that have a Pareto distribution. The graph, and supporting code, give values of c_1, c_2, and c_3 for a required revenue K. Note that these values increase linearly with K.

```
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
  ratio <- mu/var
  ratio
}
c1 <- propnfct(alpha1, theta1)
c2 <- propnfct(alpha2, theta2)
c3 <- propnfct(alpha3, theta3)
summeans = mpareto(shape=alpha1, scale=theta1, order=1)+
           mpareto(shape=alpha2, scale=theta2, order=1)+
           mpareto(shape=alpha3, scale=theta3, order=1)
temp = c1*mpareto(shape=alpha1, scale=theta1, order=1)+
       c2*mpareto(shape=alpha2, scale=theta2, order=1)+
       c3*mpareto(shape=alpha3, scale=theta3, order=1)
KVec = seq(100,summeans,length.out=20)
c1Vec <- c2Vec <-c3Vec <- 0*KVec
for (j in 1:20) {
  c1Vec[j] = c1 * KVec[j]/temp
  c2Vec[j] = c2 * KVec[j]/temp
  c3Vec[j] = c3 * KVec[j]/temp
  }
plot(KVec,c1Vec, type="l", ylab="proportion", xlab="required revenue", ylim=c(0,1))
lines(KVec,c2Vec)
lines(KVec,c3Vec)
text(1200,.8,expression(c[1]))
text(2000,.75,expression(c[2]))
text(1500,.3,expression(c[3]))
```



10.3.2 Non-Proportional Reinsurance

The Optimality of Stop Loss Insurance

Under this arrangement, the insurer sets a retention level $M(>0)$ and pays in full total claims for which $S \leq M$. Thus, the insurer retains an amount M of the risk. Further, for claims for which $S > M$, the direct insurer pays M and the reinsurer pays the remaining amount $S - M$. Summarizing, the amounts paid by the direct insurer and the reinsurer are

$$Y_{insurer} = \begin{cases} S & \text{for } S \leq M \\ M & \text{for } S > M \end{cases} \quad = \min(S, M) = S \wedge M$$

and

$$Y_{reinsurer} = \begin{cases} 0 & \text{for } S \leq M \\ S - M & \text{for } S > M \end{cases} \quad = \max(0, S - M).$$

As before, note that $Y_{insurer} + Y_{reinsurer} = S$.

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 $Y_{insurer} = g(S)$ and $Y_{reinsurer} = S - g(S)$ 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, $Var(S \wedge M) \leq Var(g(S))$ for all $g(.)$.

Click Here to see the Justification of the Proposition

Proof of the Proposition. Add and subtract a constant M and expand the square to get

$$\begin{aligned} Var\ g(S) &= E(g(S)-K)^2 = E(g(S)-M+M-K)^2 \\ &= E(g(S)-M)^2 + (M-K)^2 + 2E(g(S)-M)(M-K) \\ &= E(g(S)-M)^2 - (M-K)^2, \end{aligned}$$

because $E\ g(S) = K$.

Now, for any retention function, we have $g(S) \leq S$, that is, the insurer's retained claims are less than or equal to total claims. Using the notation $g_{SL}(S) = S \wedge M$ for stop loss insurance, we have

$$\begin{aligned} M - g_{SL}(S) &= M - (S \wedge M) \\ &= (M-S) \wedge 0 \\ &\leq (M - g(S)) \wedge 0. \end{aligned}$$

Squaring each side yields

$$(M - g_{SL}(S))^2 \leq (M - g(S))^2 \wedge 0 \leq (M - g(S))^2.$$

Returning to our expression for the variance, we have

$$\begin{aligned} Var\ g_{SL}(S) &= E(g_{SL}(S)-M)^2 - (M-K)^2 \\ &\leq E(g_{SL}(S)-M)^2 - (M-K)^2 = Var\ g(S), \end{aligned}$$

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 coverage. Under this contract, we assume that the total risk S can be thought of as composed as n separate risks $X_1, \ldots, X_n$ and that each of these risks are subject to 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, $Y_{reinsurer} = S - Y_{insurer}$. 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 uncertainty of the retained risks (as measured by the variance).

Click Here to see 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

$$E\ S \wedge M = \int_0^M\ (1 - F(S))dx$$

and

$$E\ (S \wedge M)^2 = 2\int_0^M\ x(1 - F(x))dx$$

Taking a partial derivative with respect to λ and setting this equal 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, ….

Click Here to see an Example with Three Pareto Risks

Example. Consider three risks that have a Pareto distribution, each having a different set of parameters (so they are independent but non-identical). We first optimize the Lagrangian using the `R` package `alabama` for Augmented Lagrangian Adaptive Barrier Minimization Algorithm. Then, we show 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. Finally, we graphically compare the distribution of total risks to that retained by the insurer and by the reinsurer.

```
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))
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)
```

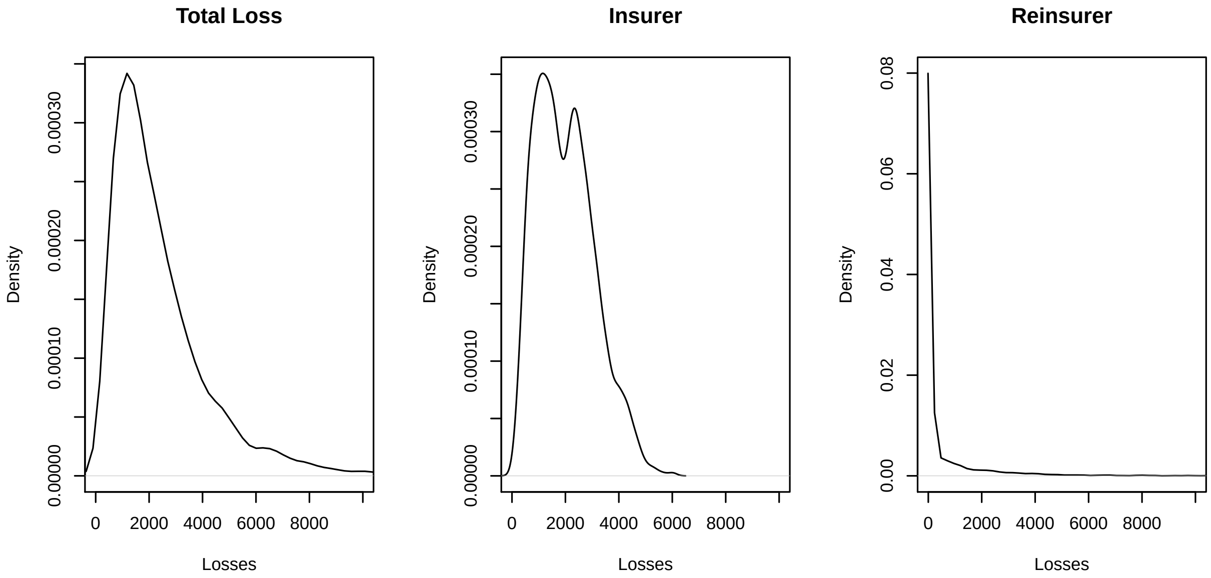
```
[1] 1344.133
```

```
M3star -levpareto(M3star,shape=alpha3, scale=theta3,order=1)
```

```
[1] 1344.133
```

```
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; 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 let the given limit be 4 lines (\$400,000). Then, if S is the loss, the reinsurer's portion is $\min(400000, (S - 100000)_+)$.

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 $k = 4$, it could be an insurer and three different reinsurers.

Example

- 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

$$Y_j = \min(S, M_j) - \min(S, M_{j-1}).$$

With the expression $Y_j = \min(S, M_j) - \min(S, M_{j-1})$, 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 $S = Y_1 + Y_2 + \cdots + Y_k$. As emphasized in the following example, we also remark that the parties need not be different.

Example - 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 $Y_1 = \min(S, 100)$ and $Y_3 = S - \min(S, 100000) = \max(0, S - 100000)$.

For additional reading, wee the Wisconsin Property Fund site for more info on layers of reinsurance, https://sites.google.com/a/wisc.edu/local-government-property-insurance-fund/home/reinsurance.

Portfolio Management Example

Many other variations of the foundational contracts are possible. For one more illustration, consider the following.

Example. You are the Chief Risk Officer of a telecommunications firm. Your firm has several property and liabililty 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

$$S = X_1 + X_2 + X_3 + X_4.$$

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 $Y_{retained} = S - Y_{insurer} = \min(X_1, M_1) + \min(X_2, M_2)$. 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.

R Code for Example Solution

In preparation, here is the code needed to set the parameters.

```r
# 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
S         <- X1 + X2 + X3 + X4
Yretained <- pmin(X1,M1) + pmin(X2,M2)
Yinsurer  <- S - Yretained
```

(a) Here is the code for the expected claim amounts.

```
# Expected Claim Amounts
ExpVec <- t(as.matrix(c(mean(Yretained),mean(Yinsurer),mean(S))))
colnames(ExpVec) <- c("Retained", "Insurer","Total")
round(ExpVec,digits=2)
```

```
     Retained Insurer   Total
[1,]   269.05 2274.41 2543.46
```

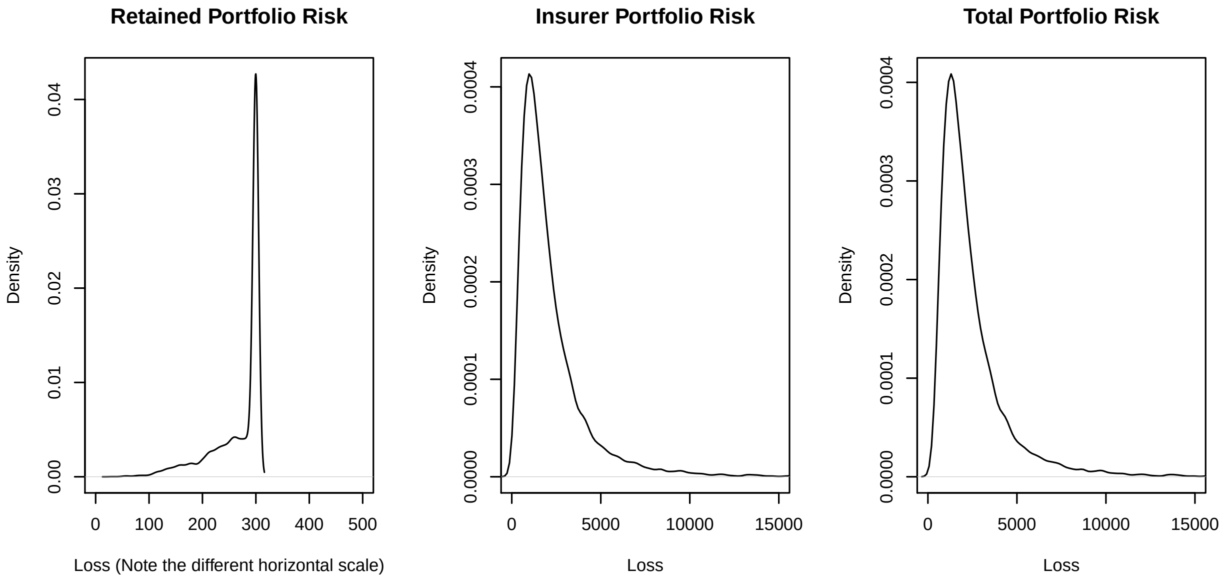
(b) Here is the code for the quantiles.

```
# 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(S        ,  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
plot(density(Yinsurer), xlim=c(0,15000), main="Insurer Portfolio Risk", xlab="Loss")
plot(density(S), xlim=c(0,15000), main="Total Portfolio Risk", xlab="Loss")
```


Retained Portfolio Risk
Density
0.00
0.01
0.02
0.03
0.04
0
100
200
300
400
500
Loss (Note the different horizontal scale)
Insurer Portfolio Risk
Density
0.0000
0.0001
0.0002
0.0003
0.0004
0
5000
10000
15000
Loss
Total Portfolio Risk
Density
0.0000
0.0001
0.0002
0.0003
0.0004
0
5000
10000
15000
Loss

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 (Tukey, 1962):

```
procedures for analyzing data,  techniques for interpreting the results of such procedures, ways of plan
```

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.
- **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.

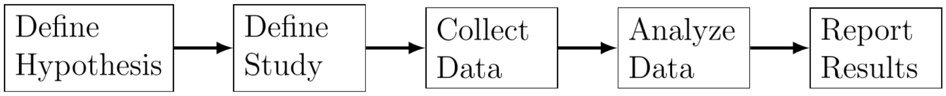


Figure 13.1: The process of a typical study in behavioral and social sciences.

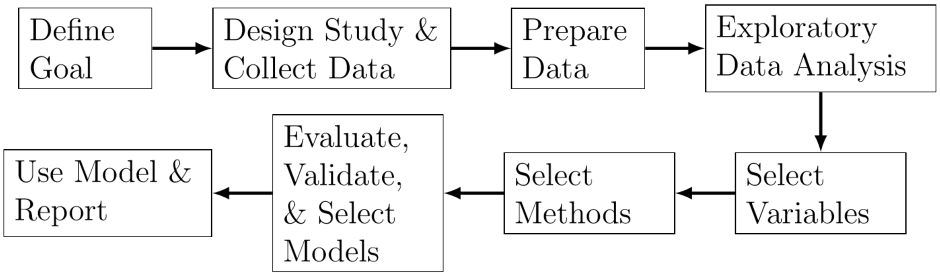

Define Goal
Design Study & Collect Data
Prepare Data
Exploratory Data Analysis
Select Variables
Select Methods
Evaluate, Validate, & Select Models
Use Model & Report

Figure 13.2: The process of statistical modeling.

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 1.1 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 1.1: 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 1.2 gives a list of common names for different types of variables (Frees, 2009). 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 1.2: 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 1.3 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 1.3: 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 1.4: Some commonly used descriptive statistics.

Table 1.4 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 $i = 1, 2, \dots, n$, 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, 2009):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n,$$

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) $x_{i1}, x_{i2}, \dots, x_{ik}$ are nonstochastic variables.

(b) Var $(y_i) = \sigma^2$, where Var (y_i) denotes the variance of y_i.

(c) $y_1, y_2, \dots, y_n$ 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) $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ 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 $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})'$ 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) $x_{i1}, x_{i2}, \dots, x_{in}$ are nonstochastic variables.

(b) $y_1, y_2, \dots, y_n$ 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, 2009).

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 1.5: Some R functions for data analysis.

Table 1.5 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

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 motivates the importance of understanding what type of variable you are working with and Section 14.2 provides an overview of the types of variables. Section 14.3 then elaborates basic measures for modeling the dependence between variables.

Section 14.4 introduces a novel approach to modeling dependence using Copulas which is reinforced with practical illustrations in Section 14.5. The types of Copula families and basic properties of Copula functions is explained Section 14.6. The chapter concludes by explaining why the study of dependence modeling is important in Section 14.7.

14.1 Introduction

Consider a pair of random variables (Coverage,Claim) 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 but as seen in later sections, the EntityType variable belongs to a different class of variables. Hence, modeling the dependence between Claim and Coverage may require a different technique from that of Claim and EntityType.

14.2 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 datasets 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 14.2 and explained in the remainder of this section provide a solid first step in framing a data set.

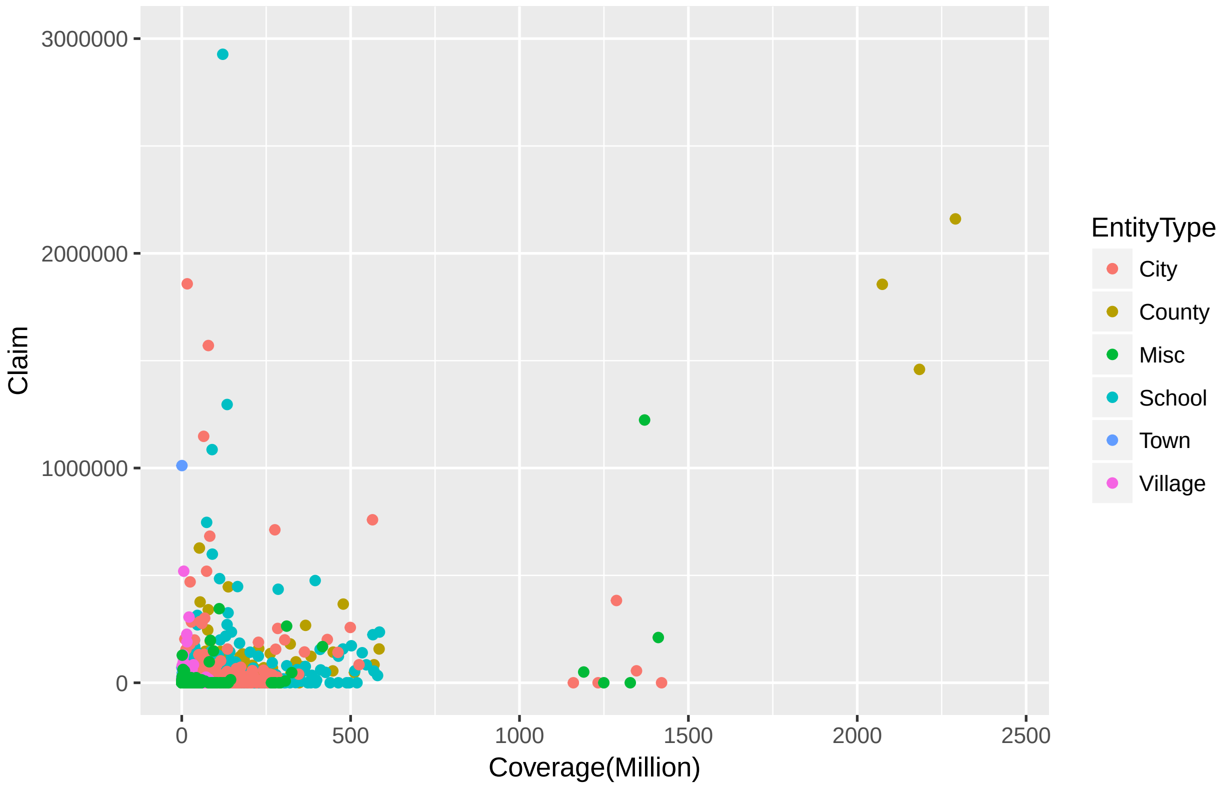


Figure 14.1: Scatter plot of (Coverage,Claim) from LGPIF data

Table : Variable Types

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

14.2.1 Qualitative Variables

In this sub-section, you learn how to:

- Describe binary variable and code it
- Classify qualitative variables as nominal or ordinal

Let us start with the simplest type, a binary variable. As suggested by its name, a binary variable is one with only two possible values. Although not necessary, the two values are commonly taken to be a 0 and a 1. Binary variables are typically used to indicate whether or not an entity possesses an attribute. For example, we might code a variable in a dataset to be a 1 if an insured is female and a 0 if male. (An insured is a person who is covered under an insurance agreement).

More generally, 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 in which area of the country in which an insured resides, you might use a 1 for the northern part, 2 for southern, and 3 for everything else. A binary variable is a special type of categorical variable where there are only two categories. 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 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.

14.2.2 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. For example, when examining a policyholder's choice of deductibles, it may be that values of 0, 250, 500, and 1000 are the only possible outcomes. Like a 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 0, 1, 2, ... For example, we will be particularly interested in the number of claims arising from a policy during a given period. This is known as the claim frequency.

Given that we will develop ways to analyze discrete variables, do we really need separate methods for dealing with continuous variables? After all, one can argue that few things in the physical world are truly continuous. For example, each currency has a smallest unit that is not subdivided further. (In the US, you cannot pay for anything smaller than one cent.) Nonetheless, models using continuous variables serve as excellent approximations to real-world discrete outcomes, in part due to their simplicity. It will be well worth our time and effort to develop models and analyze continuous and discrete variables differently.

Having said that, 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).

14.2.3 Multivariate Variables

In this sub-section, you learn how to:

- Differentiate between univariate and multivariate data
- Code missing variables
- Describe aggregate claims

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. If there are many variables, such data are also known as high dimensional.

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. By considering only one measurement, variables are scalars and, as described, can be thought broadly as either qualitative or quantitative.

In contrast, for some variables, it makes little sense to only look a 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. Historically, some statistical packages used a -99 to report when a variable, such as policyholder's age, was not available or not reported. This led to many unsuspecting analysts providing strange statistics when summarizing a set of 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. In this situation, an insurer has potentially zero, one, two, or more claims, within a policy period. Each claim has its own level (possibly mediated by deductibles and upper limits) and there are an uncertain, or random, number of each claims for each individual. This is a case where the the dimension of the multivariate variable is not known in advance.

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.

14.3 Classic Measures of Scalar Associations

In this section, you learn how to:

- Estimate correlation using Pearson method
- Use rank based measures like Spearman, Kendall and Blomqvist's Beta 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

14.3.1 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

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.

14.3.2 Rank Based Measures

Spearman

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 $R(X)$. 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).

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

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

$$\tau = \frac{2}{n(n-1)}\sum_{i<j} sgn(X_j - X_i)sgn(Y_j - Y_i) = \frac{2}{n(n-1)}\sum_{i<j} sgn(R(X_j) - R(X_i))sgn(R(Y_j) - R(Y_i)).$$

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.

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.

```
Attaching package: 'copula'
```

```
The following objects are masked from 'package:VGAM':

    log1mexp, log1pexp, rlog
```

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.

14.3.3 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 Fr’{e}chet-H”{o}effding 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 Fr'{e}chet-H"{o}effding 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

$$\pi_{ij}^{new} \quad = \quad a_i b_j \pi_{ij}$$

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 Fr'{e}chet-H"{o}effding bounds of the tetrachoric correlation, another measures of association, see also, (Joe, 2014), page 48.

Table : 2 × 2 table of counts for Fire5 and NoClaimCredit

	*Fire*5		
NoClaimCredit	0	1	*Total*
0	1611	2175	3786
1	897	956	1853
Total	2508	3131	5639

From Table 14.3.3, $OR(\pi_{11}) = \frac{1611(956)}{897(2175)} = 0.79$. 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 NoClaimCredit 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.

Table : Two-way table of counts for EntityType and NoClaimCredit

	NoClaimCredit	
EntityType	0	1
City	644	149
County	310	18
Misc	336	273
School	1103	494
Town	492	479
Village	901	440

Table 14.3.3 gives the 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 from the LGPIF data.

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 14.2.1, 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

$$\hat{\rho_N} = \text{argmax}_\rho \sum_{s=m_1}^{m_2} \sum_{t=m_1}^{m_2} n_{st} \log \left\{ \Phi_2(\hat{\xi}_{1s}, \hat{\xi}_{2t}; \rho) - \Phi_2(\hat{\xi}_{1,s-1}, \hat{\xi}_{2t}; \rho) - \Phi_2(\hat{\xi}_{1s}, \hat{\xi}_{2,t-1}; \rho) + \Phi_2(\hat{\xi}_{1,s-1}, \hat{\xi}_{2,t-1}; \rho) \right\}$$

It is called a tetrachoric correlation for binary variables.

Table : Two-way table of counts for AlarmCredit and NoClaimCredit

	NoClaimCredit	
AlarmCredit	0	1
1	1669	942
2	121	118
3	195	132
4	1801	661

You can obtain the polychoric or tetrachoric correlation using the `polychoric()` or `tetrachoric()` function from the `psych` library in `R`. From Table 14.3.3, the polychoric correlation is illustrated using AlarmCredit and NoClaimCredit variable from the LGPIF data. $\hat{\rho_N} = -0.14$, 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
```

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

$$\begin{aligned}
\sum_{s=0}^{1}\sum_{t=0}^{1} & \quad n_{st}(r_{1s} - \bar{r})(r_{2t} - \bar{r}) \\
= & \quad n_{00}\frac{n_{0+} - n}{2}\frac{n_{+0} - n}{2} + n_{01}\frac{n_{0+} - n}{2}\frac{n_{+0}}{2} + n_{10}\frac{n_{0+}}{2}\frac{n_{+0} - n}{2} + n_{11}\frac{n_{0+}}{2}\frac{n_{+0}}{2} \\
= & \quad \frac{1}{4}(n_{00}(n_{0+} - n)(n_{+0} - n) + (n_{0+} - n_{00})(n_{0+} - n)n_{+0} \\
& \qquad + (n_{+0} - n_{00})n_{0+}(n_{+0} - n) + (n - n_{+0} - n_{0+} + n_{00})n_{0+}n_{+0}) \\
= & \quad \frac{1}{4}(n_{00}n^2 - n_{0+}(n_{0+} - n)n_{+0} \\
& \qquad + n_{+0}n_{0+}(n_{+0} - n) + (n - n_{+0} - n_{0+})n_{0+}n_{+0}) \\
= & \quad \frac{1}{4}(n_{00}n^2 - n_{0+}n_{+0}(n_{0+} - n + n_{+0} - n + n - n_{+0} - n_{0+}) \\
= & \quad \frac{n}{4}(nn_{00} - n_{0+}n_{+0}).
\end{aligned}$$

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
```

Interval Variables

As described in Section 14.2.2, 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

$$\hat{\rho_N} = \text{argmax}_\rho \sum_{i=1}^{n} \log \left\{ \phi(z_{i1}) \left[\Phi(\frac{\hat{\zeta}_{2,y_{i2}} - \rho z_{i1}}{(1-\rho^2)^{1/2}}) - \Phi(\frac{\hat{\zeta}_{2,y_{i2-1}} - \rho z_{i1}}{(1-\rho^2)^{1/2}}) \right] \right\}$$

The biserial correlation is defined similarly, when one variable is continuous and the other binary.

Table : Claim by NoClaimCredit

NoClaimCredit	Mean	Total
	Claim	Claim
0	22,505	85,200,483
1	6,629	12,282,618

You can obtain the polyserial or biserial correlation using the `polyserial()` or `biserial()` function from the `psych` library in `R`. Table 14.3.3 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
```

14.4 Introduction to Copula

Copula functions are widely used in statistics and actuarial science literature for dependency modeling.

In this section, you learn how to:

- Describe the 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

$$F(y_1, \ldots, y_p) = C(F_1(y_1), \ldots, F_p(y_p))$$

.

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 are just focusing copula modeling with continuous variables. For discrete case, readers can see (Joe, 2014);(Genest and Neslohva, 2007).

For bivariate case; $d = 2$, 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 copula is one of the commonly used copula functions in the copula literature. We will learn more details about copula types in Section 14.6.

14.5 Application Using Copulas

In this section, you learn how to:

- Figure out dependency structure between random variables
- Model the dependent data with a copula function

This section analyzes the insurance losses and expenses data with the statistical programming "R". This data set is included in `copula` package. Model fitting process is started by marginal modeling of two variables (*loss* and *expense*). Then we model the joint distribution of these marginal outcomes.

14.5.1 Data Description

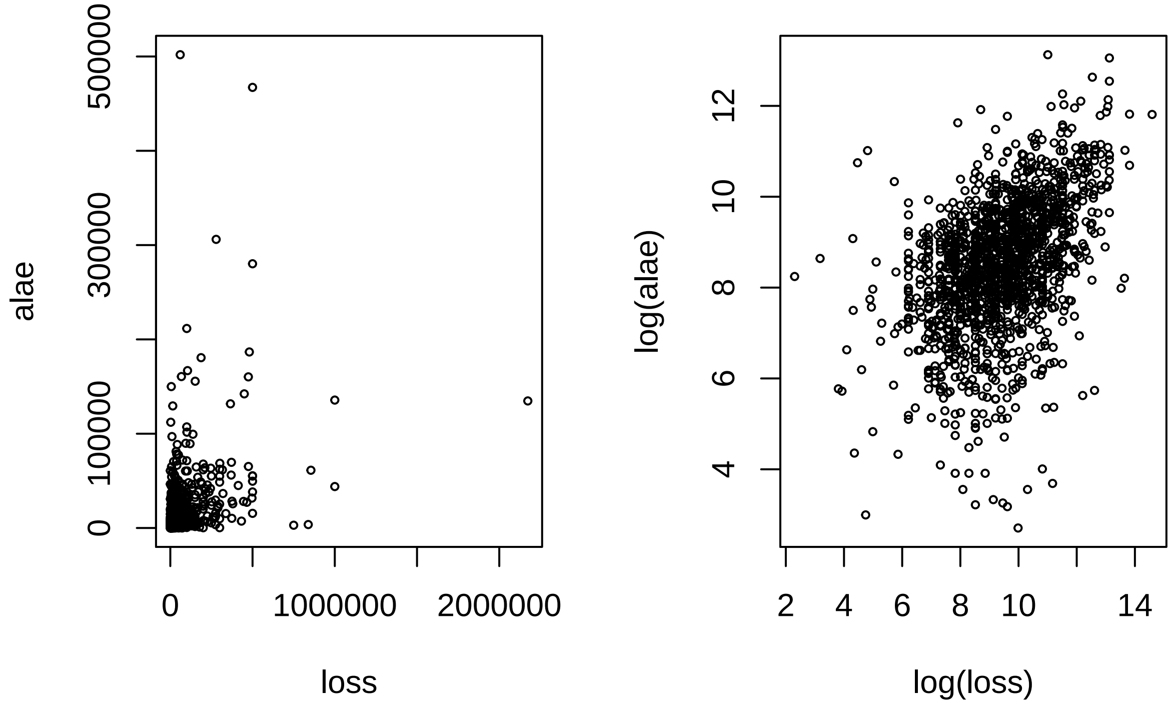
We start with getting a sample ($n = 1500$) from whole data. We consider first two variables of the data; losses and expenses .

- 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)

```
library(copula)
data(loss) # loss data
Lossdata <- loss # rename dataframe to distinguish from the loss variable
attach(Lossdata) # attach data to freely use any variable without referencing its parent dataframe
loss <- Lossdata$loss
```

To visualize the relationship between losses and expenses (ALAE), scatterplots are created on the real dollar scale and on the log scale.

```
par(mfrow=c(1, 2))
plot(loss,alae, cex=.5) # real dollar scale
plot(log(loss),log(alae),cex=.5) # 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))
```

14.5.2 Marginal Models

We need to check the histograms 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 regression. *vglm* function is used for the estimation from `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 functi
coef(fit, matrix=TRUE) # extract fitted model coefficients, matrix=TRUE gives logarithm of estimated pa
Coef(fit)

Output:
               loge(scale) loge(shape)
 (Intercept)     9.624673   0.7988753

                  scale        shape
 (Intercept)  15133.603598     2.223039
```

We repeat the procedure above to fit the marginal distribution of the loss variable. Because the loss data also seems right-skewed and heavy-tail data, we 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.To see more details of marginal model selection,readers can remember the Chapter 4 .

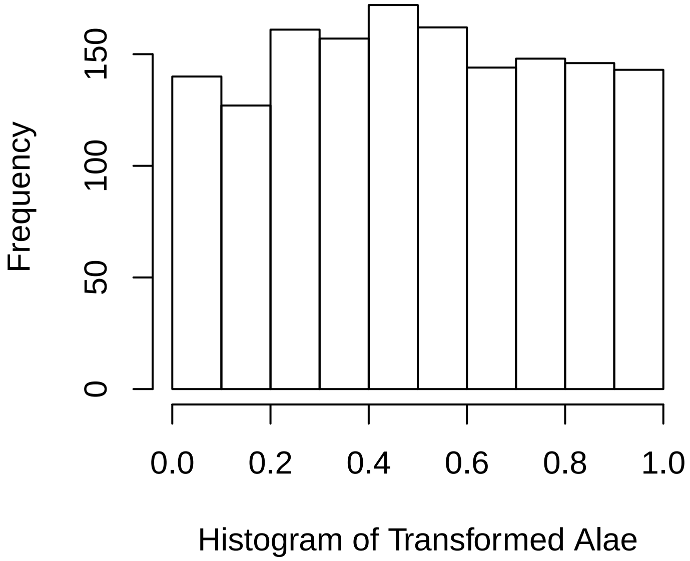
14.5.3 Probability Integral Transformation

Probability integral transformation shows us any continuous variable can be mapped to a $U(0,1)$ random variable via its distribution function.

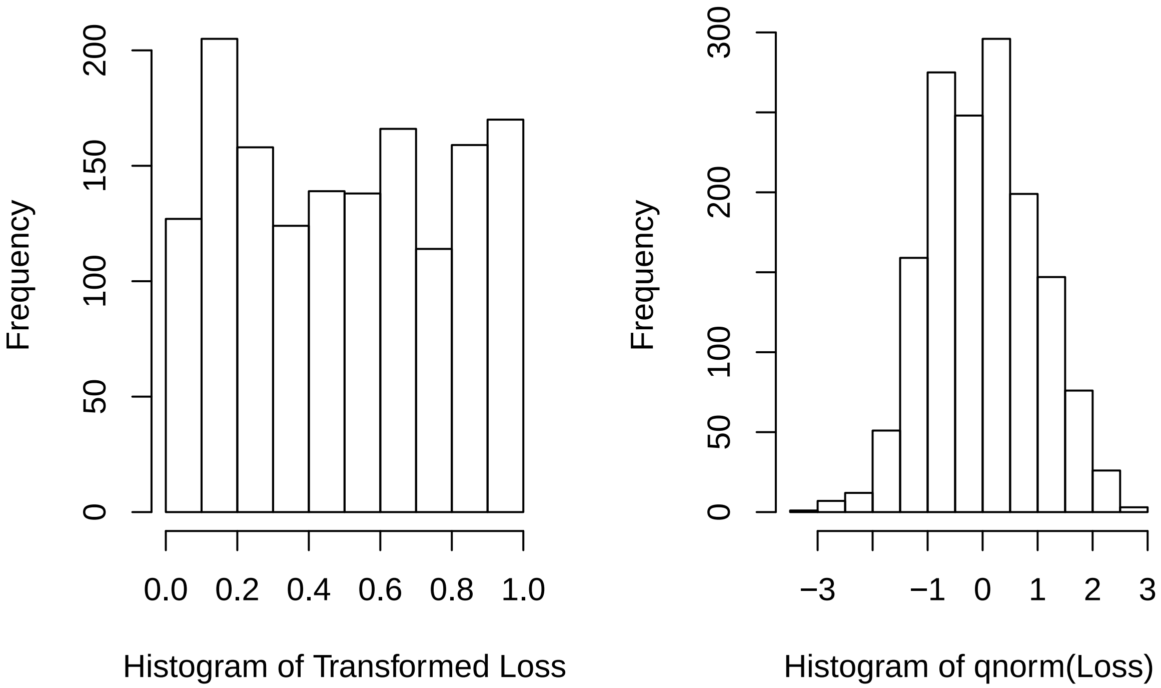
Given the fitted Pareto II distribution, the variable expenses is transformed to the variable u_1, which follows a uniform distribution on $[0,1]$:

$$u_1 = 1 - \left(1 + \frac{ALAE}{\hat{\theta}}\right)^{-\hat{\alpha}}.$$

After applying the probablity integral transformation to expenses variable, we plot the histogram of Transformed Alae



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 trasformed to *normal scores* with the quantile function of standard normal distribution. As we see in the second histogram, normal scores of the variable loss are approximately marginally standard normal.



R Code for The Histogram of Transformed Alae

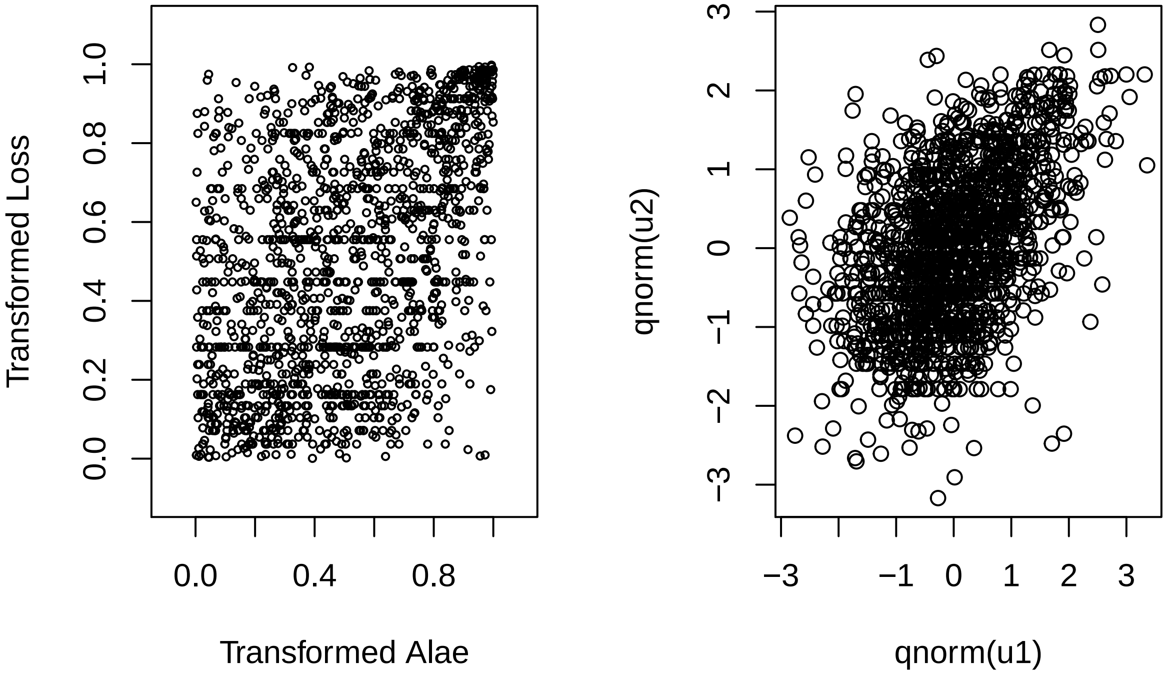
```
u1 = 1 - (1 + (alae/b))^(-s) # or u1=pparetoII(alae, location=0, scale=b, shape=s)
hist(u1, main="", xlab="Histogram of Transformed alae")

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)")
```

14.5.4 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.

Then we calculate the Spearman's rho correlation 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, readers can see (Joe, 2014).

(U_1, U_2), $(U_1 = F_1(ALAE)$ and $U_2 = F_2(LOSS))$, is fitted to Frank's copula with Maximumlikelihood 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
param = fit.ml@estimate # fitted copula parameter
frank.cop <- archmCopula("frank", param= param, dim = 2)
#rho(frank.cop) # Spearman's correlation
```

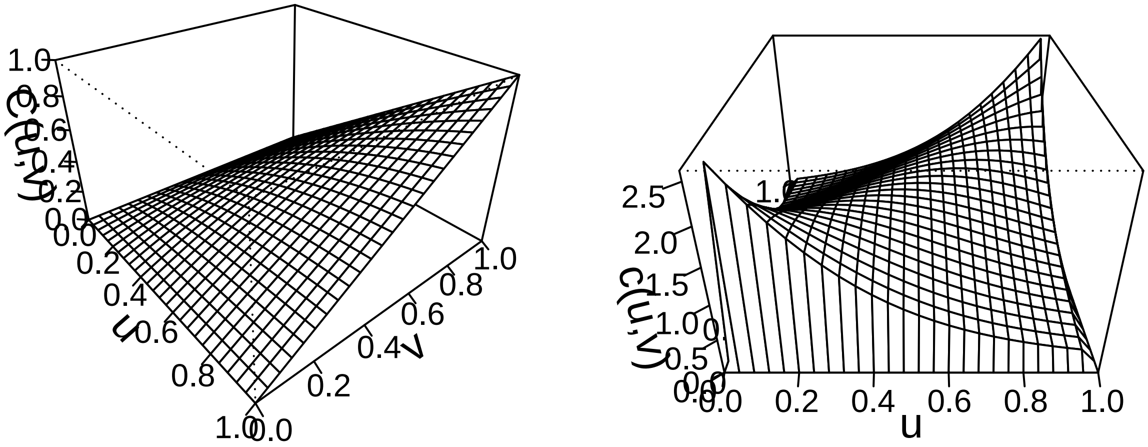
R Code for Spearman's Correlation Using Frank's Copula

```
(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.

```r
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)
```



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 dependency for this data set, with $\theta = 3.114$. Frank's copula can model not only positive dependence between the marginals, but also independence and negative dependence, according to different values of θ:

- $\theta = 0$: independent copula
- $\theta > 0$: positive dependence
- $\theta < 0$: negative dependence

14.6 Types of Copulas

We will go on reviewing copula functions with more details.

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 faimilies are Archimedian copulas and Elliptical copulas.

14.6.1 Elliptical Copulas

Elliptical copulas are constructed from elliptical distributions.This copula decompose (multivariate) elliptical distributions into their univariate elliptical marginal distrubitions by Sklar's theorem (Hofert et al., 2017).

Properties of elliptical copulas are typically obtained from the properties of corresponding elliptical distributions(Hofert et al., 2017).

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 $[0,1]^p$ by

$$c_E(u_1,\ldots,u_p) = h_E\left(H^{-1}(u_1),\ldots,H^{-1}(u_p)\right)\prod_{j=1}^{p}\frac{1}{h(H^{-1}(u_j))}.$$

In the Elliptical copula family, the function g_p is known as a "generator" in that it can be used to generate alternative distributions.

Table : (#tab:Generator Functions) Distribution and Generator Functions ($\mathrm{g}_p(x)$) for Selected Elliptical Copulas:

	Generator
Distribution	$\mathrm{g}_p(x)$
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	$\exp(-rx^s)$

Most empirical work focuses on the normal copula and t-copula. t-copulas are useful for modeling the dependency in the tails of bivariate distrubitions,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., 2017).

14.6.2 Archimedian Copulas

This class of copulas are constructed from a *generator* function,which is g(·) 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

$$\mathrm{C_g}(u_1, \ldots, u_p) = \mathrm{g}^{-1}\left(\mathrm{g}(u_1) + \cdots + \mathrm{g}(u_p)\right)$$

is said to be an Archimedean copula. The function "g" is known as the generator of the copula $\mathrm{C_g}$.

For bivariate case; $d = 2$, Archimedean copula function can be written by the function

$$\mathrm{C_g}(u_1,\, u_2) = \mathrm{g}^{-1}\left(\mathrm{g}(u_1) + \mathrm{g}(u_2)\right).$$

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 14.4 and in Section 14.5. Here we will continue to express the equations for Clayton copula and Gumbel/Hougaard copula.

Clayton Copula

For $d = 2$, 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); (Genest and Mackay, 1986).

14.6.3 Properties of Copulas

Bounds on Association

Like all multivariate distribution functions, copulas are bounded. The Fr$'$echet-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

$$\mathrm{C}(u_1, \ldots, u_p) = \Pr(U_1 \leq u_1, \ldots, U_p \leq u_p) \leq \Pr(U_j \leq u_j)$$

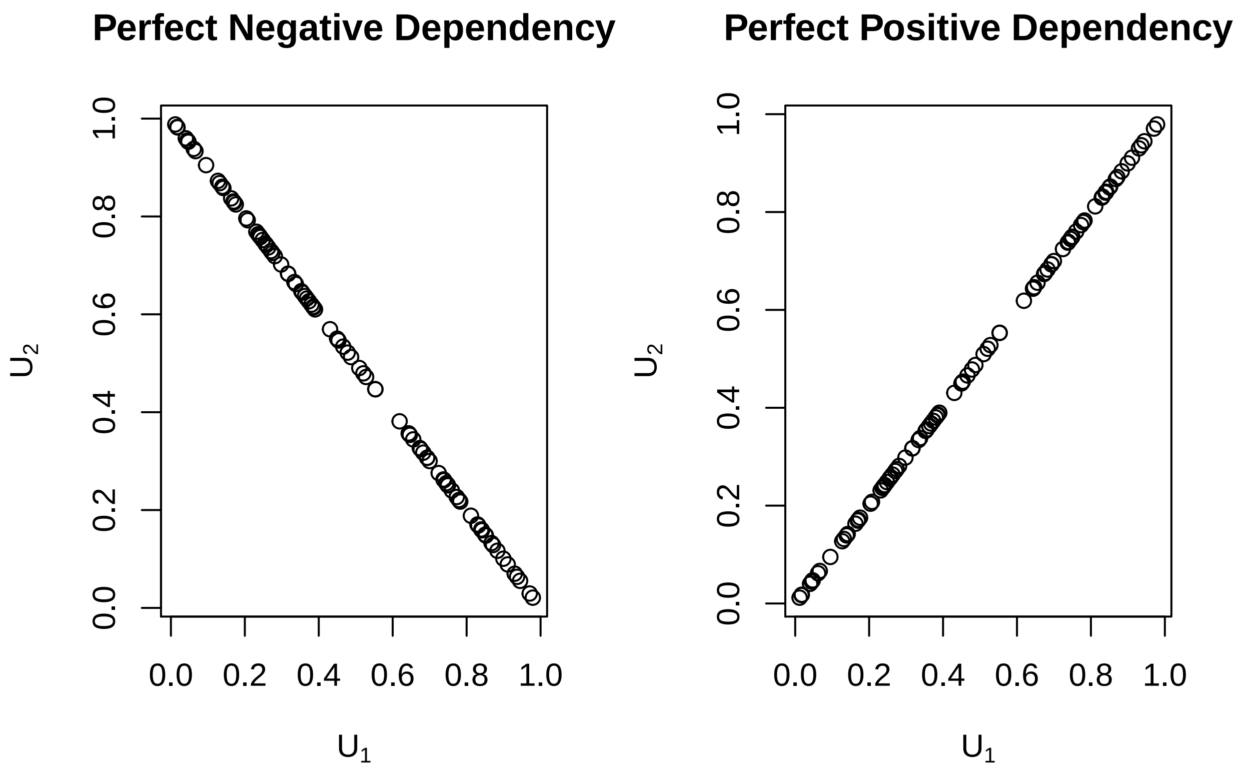
, for $j = 1, \ldots, p$. The bound is achieved when $U_1 = \cdots = U_p$. The bound is achieved when $U_1 = \cdots = U_p$. To see the left-hand side when $p = 2$, consider $U_2 = 1 - U_1$. In this case, if $1 - u_2 < u_1$ then $\Pr(U_1 \leq u_1, U_2 \leq u_2) = \Pr(1 - u_2 \leq U_1 < u_1) = u_1 + u_2 - 1$. (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 graphs.

```r
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 d=2
plot (cbind(U,U), xlab=quote(U[1]),ylab=quote(U[2]),main="Perfect Positive Dependency")  #M for d=2
```



R Code for Frechet-Hoeffding Bounds for Two Random Variables

```r
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 d=2
plot (cbind(U,U), xlab=quote(U[1]),ylab=quote(U[2]),main="Perfect Positive Dependency")  #M for d=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

$$= 12 \int \int \{\mathrm{C}(u,v) - uv\} \, dudv.$$

And 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 "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 $R = \lim_{z \to 1} R(z)$. Similarly, the left-tail concentration function is

$$L(z) = \frac{\Pr(U_1 \leq z, U_2 \leq z)}{z} = \Pr(U_1 \leq z | U_2 \leq z) = \frac{\mathrm{C}(z,z)}{1 - z}.$$

14.7 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 14.5 , 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 14.7, 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 that can cause unexpected solvency problems.

Table : Claim by NoClaimCredit

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

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")
```

Chapter 15

Appendix A: Review of Statistical Inference

Chapter preview. The supplement 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 supplement, 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 Example

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

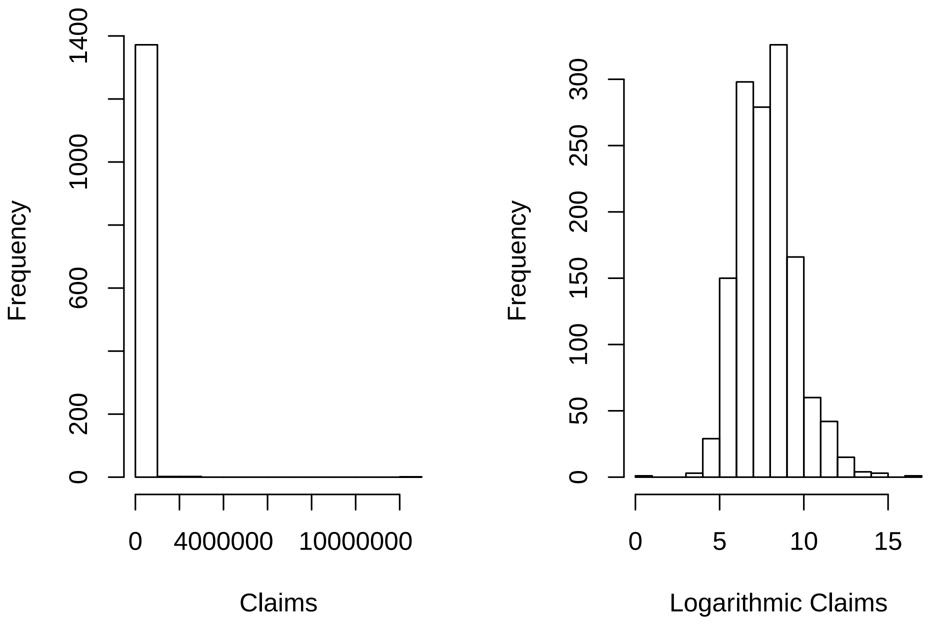


Figure 15.1: Distribution of Claims

```r
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 Example

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 **kth moment** of X, $k = 1, 2, 3, \cdots$. For example, the population mean μ is the first moment. Furthermore, the expectation $\mathrm{E}[(X-\mu)^k]$ is called a **kth 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

$$\begin{aligned}\hat{\mu}_1 &= h_1(\hat{\theta}_1, \cdots, \hat{\theta}_m);\\ \hat{\mu}_2 &= h_2(\hat{\theta}_1, \cdots, \hat{\theta}_m);\\ &\cdots\\ \hat{\mu}_m &= h_m(\hat{\theta}_1, \cdots, \hat{\theta}_m).\end{aligned}$$

Show Example

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 $X_i = x_i$ $(i = 1, 2, \cdots, n)$, the probability mass function in the case of a discrete X and the probability density function in the case of a continuous X. Then 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

$$\begin{aligned}\frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_1} &= 0;\\ \frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_2} &= 0;\\ &\cdots\\ \frac{\partial l(\hat{\boldsymbol{\theta}}|\mathbf{x})}{\partial \hat{\theta}_m} &= 0,\end{aligned}$$

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 Supplement 17.

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 $X_i \sim^{iid} N(\mu, \sigma^2)$, $i = 1, \dots, n$, 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,

  $$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, V), \quad \text{as} \quad n \to \infty,$$

  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 Example

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 Supplement 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 Example

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 Example

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 Supplement 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 Example

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

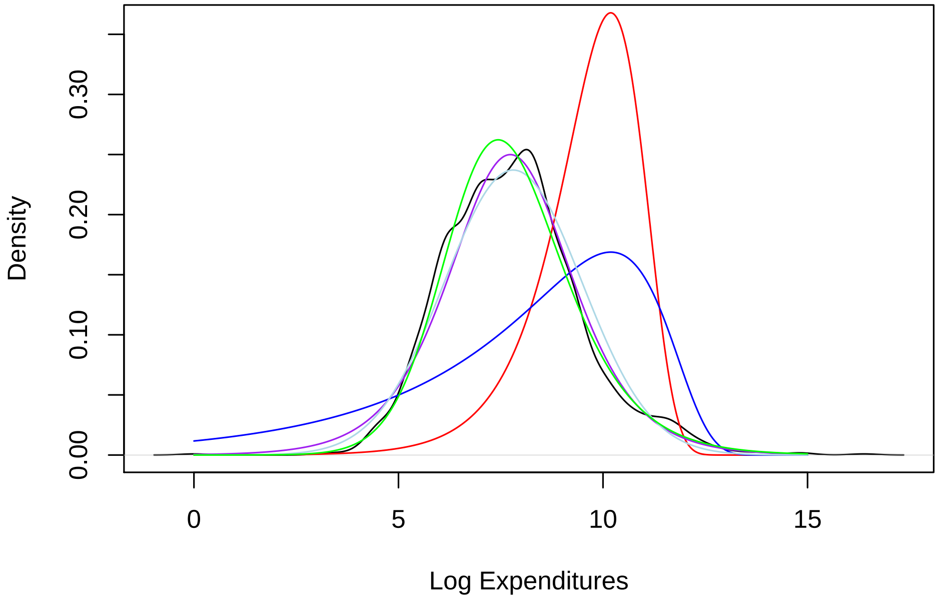


Figure 15.2: Fitted Claims Distribution

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$Claim)), ylim=c(0,0.36),main="", xlab="Log Expenditures")
x <- seq(0,15,by=0.01)
fexp_ex = dgamma(exp(x), scale = exp(-coef(fit.exp)), shape = 1)*exp(x)
lines(x,fexp_ex, col="red")
fgamma_ex = dgamma(exp(x), shape = alpha.gamma, scale=theta.gamma)*exp(x)
lines(x,fgamma_ex,col="blue")
fpareto_ex = dparetoII(exp(x),loc=0,shape = exp(coef(fit.pareto)[2]), scale = exp(coef(fit.pareto)[1]))*
lines(x,fpareto_ex,col="purple")
flnorm_ex = dlnorm(exp(x), mean = coef(fit.LN)[1], sd = exp(coef(fit.LN)[2]))*exp(x)
lines(x,flnorm_ex, col="lightblue")
# density for GB II
gb2density <- function(x){
  a1 <- gb2bop$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")
```

Chapter 16

Appendix B: Iterated Expectations

This supplement 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

$$F(x, y) = \Pr[X \le x, Y \le y].$$

When X and Y are independent, we have

$$F(x, y) = F(x)F(y),$$

with $F(x) = \Pr[X \le x]$ and $F(y) = \Pr[Y \le y]$ 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

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

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 supplement.

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

$$\begin{aligned}\mathrm{E}\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\} &= \sum_x \left\{\sum_y h(x,y)p(y|x)\right\} p(x) \\ &= \sum_x \sum_y h(x,y)p(y|x)p(x) \qquad\qquad . \\ &= \sum_x \sum_y h(x,y)p(x,y) = \mathrm{E}[h(X,Y)]\end{aligned}$$

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

$$\begin{aligned}\mathrm{E}\{\mathrm{Var}\left[h(X,Y)|X\right]\} &= \mathrm{E}\left\{\mathrm{E}\left[h(X,Y)^2|X\right]\right\} - \mathrm{E}\left(\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2\right) \\ &= \mathrm{E}\left[h(X,Y)^2\right] - \mathrm{E}\left(\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2\right).\end{aligned}$$

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

$$\begin{aligned}\mathrm{Var}\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\} &= \mathrm{Var}[g(X)] \\ &= \mathrm{E}[g(X)^2] \ - \{\mathrm{E}[g(X)]\}^2 \\ &= \mathrm{E}\left(\left\{\mathrm{E}\left[h(X,Y)|X\right]\right\}^2\right) - \{\mathrm{E}[h(X,Y)]\}^2.\end{aligned}$$

Thus, adding Equations (16.2.2) and (16.2.2) 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

$$\mathrm{Var}[N] = \alpha\lambda_1 + (1 - \alpha)\lambda_2 + (\lambda_1 - \lambda_2)^2\alpha(1 - \alpha).$$

Chapter 17

Appendix C: Maximum Likelihood Theory

Chapter preview. Supplement 15 introduced the maximum likelihood theory regarding estimation of parameters from a parametric family. This supplement 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 Supplement 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 Supplement 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 Supplement 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 Supplement 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

$$F(x) = \Pr(X_i \le x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \qquad x > 500,$$

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 Supplement 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 Supplement 15, the MLE has some nice large-sample properties, under certain regularity conditions. We presented the results for a single parameter in Supplement 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,

  $$\sqrt{n}(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}) \rightarrow N\left(\mathbf{0}, \boldsymbol{V}\right), \quad \text{as} \quad n \rightarrow \infty,$$

  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 supplement. Hence, 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 Supplement 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 Supplement 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 Supplement 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 Supplement 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 Supplement 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 Supplement 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 Supplement 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 Supplement 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 Supplement 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 Supplement 15.

Bibliography

Abbott, D. (2014). Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Wiley, Hoboken, NJ.

Abdullah, M. F. and Ahmad, K. (2013). The mapping process of unstructured data to structured data. In 2013 International Conference on Research and Innovation in Information Systems (ICRIIS), pages 151–155.

Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer, New York, NY.

Albers, M. J. (2017). Introduction to Quantitative Data Analysis in the Behavioral and Social Sciences. John Wiley & Sons, Inc., Hoboken, NJ.

Bailey, R. A. and LeRoy, J. S. (1960). Two studies in automobile ratemaking. Proceedings of the Casualty Actuarial Society Casualty Actuarial Society, XLVII(I).

Bandyopadhyay, P. S. and Forster, M. R., editors (2011). Philosophy of Statistics. Handbook of the Philosophy of Science 7. North Holland.

Bishop, C. M. (2007). Pattern Recognition and Machine Learning. Springer, New York, NY.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge [etc.]: MIT.

Blomqvist, N. (1950). On a measure of dependence between two random variables. The Annals of Mathematical Statistic, pages 593–600.

Bluman, A. (2012). Elementary Statistics: A Step By Step Approach. McGraw-Hill, New York, NY.

Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A., and Nesbitt, C. J. (1986). Actuarial Mathematics. Society of Actuaries Itasca, Ill.

Breiman, L. (2001). Statistical modeling: The two cultures. Statistical Science, 16(3):199–231.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). Classification and Regression Trees. Chapman and Hall/CRC, Raton Boca, FL.

Buttrey, S. E. and Whitaker, L. R. (2017). A Data Scientist s Guide to Acquiring, Cleaning, and Managing Data in R. Wiley, Hoboken, NJ.

Chen, M., Mao, S., Zhang, Y., and Leung, V. C. (2014). Big Data: Related Technologies, Challenges and Future Prospects. Springer, New York, NY.

Clarke, B., Fokoue, E., and Zhang, H. H. (2009). Principles and theory for data mining and machine learning. Springer-Verlag, New York, NY.

Dabrowska, D. M. (1988). Kaplan-meier estimate on the plane. The Annals of Statistics, pages 1475–1489.

Daroczi, G. (2015). Mastering Data Analysis with R. Packt Publishing, Birmingham, UK.

de Jong, P. and Heller, G. Z. (2008). Generalized linear models for insurance data. Cambridge University Press, Cambridge, UK.

Dickson, D. C. M., Hardy, M., and Waters, H. R. (2013). Actuarial Mathematics for Life Contingent Risks. Cambridge University Press.

Earnix (2013). 2013 insurance predictive modeling survey. Earnix and Insurance Services Office, Inc. [Retrieved on July 7, 2014].

Fechner, G. T. (1897). Kollektivmasslehre. Wilhelm Englemann, Leipzig.

Forte, R. M. (2015). Mastering Predictive Analytics with R. Packt Publishing, Birmingham, UK.

Frees, E. and Valdez, E. (1998). Understanding relationships using copulas. North American Actuarial Journal, 2(01):1–25.

Frees, E. W. (2009). Regression Modeling with Actuarial and Financial Applications. Cambridge University Press.

Gan, G. (2011). Data Clustering in C++: An Object-Oriented Approach. Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC Press, Boca Raton, FL, USA.

Gan, G., Ma, C., and Wu, J. (2007). Data Clustering: Theory, Algorithms, and Applications. SIAM Press, Philadelphia, PA.

Gelman, A. (2004). Exploratory data analysis for complex models. Journal of Computational and Graphical Statistics, 13(4):755–779.

Genest, C. and Mackay, J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. The American Statistician, 40:280–283.

Genest, C. and Neslohva, J. (2007). A primer on copulas for count data. Journal of the Royal Statistical Society, pages 475–515.

Good, I. J. (1983). The philosophy of exploratory data analysis. Philosophy of Science, 50(2):283–295.

Gorman, M. and Swenson, S. (2013). Building believers: How to expand the use of predictive analytics in claims. SAS. [Retrieved on August 17, 2014].

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., and Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47:98 − 115.

Hettmansperger, T. P. (1984). Statistical inference based on ranks.

Hofert, M., Kojadinovic, I., Machler, M., and Yan, J. (2017). Elements of Copula Modeling with R. Springer.

Hougaard, P. (2000). Analysis of Multivariate Survival Data. Springer New York.

Hox, J. J. and Boeije, H. R. (2005). Data collection, primary versus secondary. In Encyclopedia of social measurement, pages 593 − 599. Elsevier.

Igual, L. and Segu, S. (2017). Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications. Springer, New York, NY.

Inmon, W. and Linstedt, D. (2014). Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault. Morgan Kaufmann, Cambridge, MA.

Insurance Information Institute (2015). International insurance fact book. Insurance Information Institute. [Retrieved on May 10, 2016].

Janert, P. K. (2010). Data Analysis with Open Source Tools. O’Reilly Media, Sebastopol, CA.

Joe, H. (2014). Dependence Modeling with Copulas. CRC Press.

Judd, C. M., McClelland, G. H., and Ryan, C. S. (2017). Data Analysis. A Model Comparison Approach to Regression, ANOVA and beyond. Routledge, New York, NY, 3rd edition.

Kendall, M. G. (1938). A new measure of rank correlation. Biometrika, pages 81–93.

Kubat, M. (2017). An Introduction to Machine Learning. Springer, New York, NY, 2nd edition.

Lee Rodgers, J. and Nicewander, W. A. (1998). Thirteen ways to look at the correlation coeffeicient. The American Statistician, 42(01):59–66.

Mailund, T. (2017). Beginning Data Science in R: Data Analysis, Visualization, and Modelling for the Data Scientist. Apress.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. Econometrica: journal of the Econometric Society, pages 647–663.

Miles, M., Hberman, M., and Sdana, J. (2014). Qualitative Data Analysis: A Methods Sourcebook. Sage, Thousand Oaks, CA, 3rd edition.

Mirkin, B. (2011). Core Concepts in Data Analysis: Summarization, Correlation and Visualization. Springer, London, UK.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). Foundations of Machine Learning. MIT Press, Cambridge, MA.

Nelson, R. B. (1997). An Introduction to Copulas. Lecture Notes in Statistics 139.

O'Leary, D. E. (2013). Artificial intelligence and big data. IEEE Intelligent Systems, 28(2):96–99.

Olson, J. E. (2003). Data Quality: The Accuracy Dimension. Morgan Kaufmann, San Francisco, CA.

Pries, K. H. and Dunnigan, R. (2015). Big Data Analytics: A Practical Guide for Managers. CRC Press, Boca Raton, FL.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3):210–229.

Shmueli, G. (2010). To explain or to predict? Statistical Science, 25(3):289–310.

Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(01):72–101.

Tukey, J. W. (1962). The future of data analysis. The Annals of Mathematical Statistics, 33(1):1–67.

Venter, G. (1983). Transformed beta and gamma distributions and aggregate losses. In Proceedings of the Casualty Actuarial Society, volume 70, pages 289–308.

Yule, G. U. (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, pages 257–319.

Yule, G. U. (1912). On the methods of measuring association between two attributes. Journal of the Royal Statistical Society, pages 579–652.