Loss Data Analytics

An open text authored by the Actuarial Community

2017-06-27

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.

Chapter 1

Introduction to Loss Data Analytics

Chapter Preview. This book introduces readers to methods of analyzing insurance data. Section 1 begins with a discussion of why the use of data is important in the insurance industry. Yes, this is obvious but we need to make a strong case for it as this is the whole premise of the book. Next, Section 2 provides an overview of the types of data that one encounters. There are many types from which to choose; your first step in the analysis of data is identify a broad class to help direct you to the appropriate tools and techniques. Section 3 gives a general overview of the purposes of analyzing insurance data which is reinforced in the Section 4 case study. Naturally, there is a huge gap between identifying the broad class of variables and what we learn from the data; 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.

- 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 1.3; 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 Variable Types

In this section, you learn how to:

- Describe different types of variables typically encountered in insurance practice
- Classify a variable into the appropriate category

Before discussing how to use insurance data to make decisions, it is helpful to first describe common features of data. In general, 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. We will encounter datasets that have many variables (high dimensional) and so it useful to begin by classifying them into different types. As will be seen, this classification is not strict; there is overlap among the types. Nonetheless, the classification summarized in Table 1.2 and explained in the remainder of this section provide a solid first step in framing a dataset.

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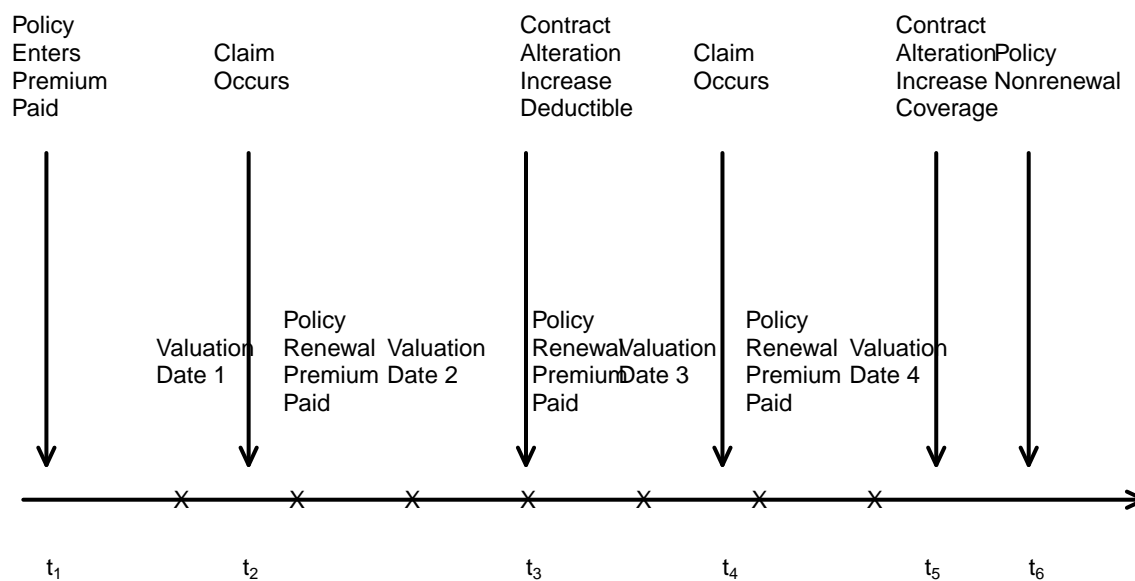
Figure 1.1: Timeline of a Typical Insurance Policy. Arrows mark the occurrences of random events. Each x marks the time of scheduled events that are typically non-random.

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

1.2.1   Qualitative Variables

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.

1.2.2   Quantitative Variables

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 $0, 1, 2, \ldots$. 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. For example, instead of recording a driver's age in year, it is common for insurers to group ages into three categories, (i) ages 16-24, representing young drives, (ii) ages 25-54, representing intermediate age drivers, and (iii) ages 55 and over, representing senior drivers.

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).

1.2.3   Multivariate Variables

Insurance data are 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*. To illustrate, with automobile claims may be limited or censored by 500,000, the upper limit that the insurer will pay. The loss amount may be in excess of 500000 but the insurer is only aware of its payout. To record censored claims, a binary variable

is used to indicate whether or not the claim is censored (limited) and a second variable is used to indicate the payout. In the same way, claims may be truncated by a deductible. Although there are many types of deductibles, in a common form the insurer pays the amount in excess of a deductible. To illustrate, suppose you have an auto policy with a 250 deductible. If you have a 1000 loss, then the insurer pays 750. If you have a 200 loss, then the insurer pays nothing. In principle, one would like to use a binary variable to indicate whether or not the claim has a deductible and a second variable is used to indicate the payout. As we will see, the tricky thing about deductibles is that for many sampling schemes, the insurer does not observe a claim if it the loss falls below the deductible amount. More on this topic later.

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.

1.3 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.

1.3.1   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.

1.3.2   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.

1.3.3 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 1.3.4.

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.

1.3.4 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.

1.4 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.

1.4.1 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 mor

Type									
Count	707	209	86	40	18	12	9	4	6	19
Proportion	0.637	0.188	0.077	0.036	0.016	0.011	0.008	0.004	0.005	0.017

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 purposes; the corresponding figure in the right-hand panel is much easier to interpret.

Table 1.2: 2010 Average Severity Distribution

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

R Code for Severity Distribution Table and Figures

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

1.4.2 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,
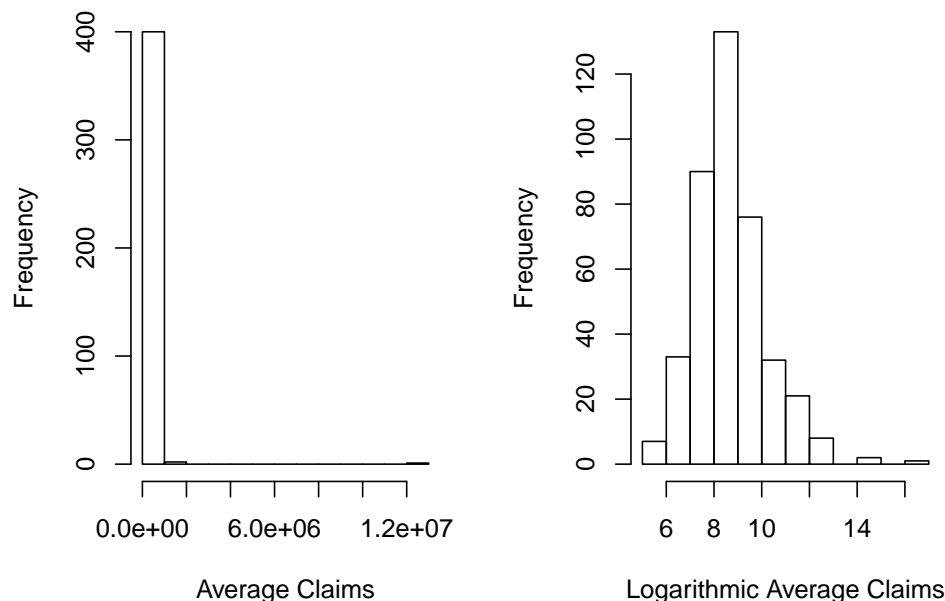
Figure 1.2: Distribution of Positive Average Severities

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

Year	Average Frequency	Average Severity	Average Coverage	Number of Policyholders
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=FALSE
library(doBy)
T1A <- summaryBy(Freq ~ Year, data = Insample,
   FUN = function(x) { c(m = mean(x), num=length(x)) } )
T1B <- summaryBy(yAvg    ~ Year, data = Insample,
   FUN = function(x) { c(m = mean(x), num=length(x)) } )
T1C <- summaryBy(BCcov   ~ Year, data = Insample,
   FUN = function(x) { c(m = mean(x), num=length(x)) } )
Table1In <- cbind(T1A[1],T1A[2],T1B[2],T1C[2],T1A[3])
names(Table1In) <- c("Year", "Average Frequency","Average Severity", "Average","Number of Policyholders")
Table1In
```

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=FALSE
t1<- summaryBy(Insample$Freq ~ 1, data = Insample,
   FUN = function(x) { c(ma=min(x), m1=median(x),m=mean(x),mb=max(x)) } )
names(t1) <- c("Minimum", "Median","Average", "Maximum")
t2 <- summaryBy(Insample$yAvg ~ 1, data = Insample,
   FUN = function(x) { c(ma=min(x), m1=median(x), m=mean(x),mb=max(x)) } )
names(t2) <- c("Minimum", "Median","Average", "Maximum")
t3 <- summaryBy(Deduct ~ 1, data = Insample,
   FUN = function(x) { c(ma=min(x), m1=median(x), m=mean(x),mb=max(x)) } )
names(t3) <- c("Minimum", "Median","Average", "Maximum")
t4 <- summaryBy(BCcov/1000 ~ 1, data = Insample,
   FUN = function(x) { c(ma=min(x), m1=median(x), m=mean(x),mb=max(x)) } )
names(t4) <- c("Minimum", "Median","Average", "Maximum")
Table2 <- rbind(t1,t2,t3,t4)
Table2a <- round(Table2,3)
Rowlable <- rbind("Claim Frequency","Claim Severity","Deductible","Coverage (000's)")
```

```
Table2aa <- cbind(Rowlable,as.matrix(Table2a))
Table2aa
```

Table 1.4.2 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.

Table: 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
```

1.4.3 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 1.3; 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 (≈ 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.

1.5 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**

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

2.2.2 Probability Generating Function

- The **probability generating function** is

$$\text{P}(z) = \text{E } z^N = \text{E } \exp(N \ln z) = M_N(\ln z)$$
$$= \sum_{k=0}^{\infty} z^k p_k.$$

- By taking the mth derivative, we see that

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

the pgf "generates" the probabilities.

- Further, the pgf can be used to generate moments

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

and

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

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 $\text{E } N = \lambda$ which is the same as the variance, $\text{Var } N = \lambda$.

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 $(a, b, \mathbf{0})$ Class

- Recall the notation $p_k = \Pr(N = k)$.

- *Definition.* A count distribution is a member of the $(a, b, \mathbf{0})$ **class** if the probabilities p_k satisfy

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

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

    - 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 $(a, b, \mathbf{0})$ 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

$$\begin{cases} 1-q & \text{if } n_i = 0 \\ q & \text{if } n_i = 1 \end{cases}.$$

- 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

$$\begin{cases} \ln(1-q) & \text{if } n_i = 0 \\ \ln q & \text{if } n_i = 1 \end{cases}.$$

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

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

  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 **$(a, b, \mathbf{1})$ 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

$$p_k^M = c p_k^0$$

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 $N|\Lambda \sim \text{Poisson}(\Lambda)$ 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 (k)	Observed (m_k)	Fitted Counts using the 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

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

- 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

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

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

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

- From this, the expectation of the conditional variance is

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

- Further, note that the conditional expectation, $\mathrm{E}\ (h(X,Y)|X = x)$, is a function of x, say, $g(x)$.

- Now, $g(X)$ is a random variable with mean $\mathrm{E}\,h(X,Y)$ 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

$$\text{E } N = \text{E } \{\text{E } (N|T)\}$$
$$= \text{E } N_1 \times \alpha + \text{E } N_2 \times (1 - \alpha).$$

- From the law of total variation

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

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 $f_X(x)$. The k-th raw moment of X, denoted by μ'_k, is the expected value of the k-th power of X, provided it exists. The first raw moment μ'_1 is the mean of X usually denoted by μ. The formula for μ'_k is given as

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

The support of the random variable X is assumed to be nonnegative since actuarial phenomena are rarely negative.

The k-th central moment of X, denoted by μ_k, is the expected value of the k-th power of the deviation of X from its mean μ. The formula for μ_k is given as

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

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 (μ_3/σ^3) 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 (μ_4/σ^4) 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 $\mu_1' = E(X) = \alpha\theta$, $\mu_2' = E(X^2) = (\alpha+1)\alpha\theta^2$, $\mu_3' = E(X^3) = (\alpha+2)(\alpha+1)\alpha\theta^3$, and $Var(X) = \alpha\theta^2$.

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

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

$$F_X\left(\pi_p^-\right) \le p \le F(\pi_p),$$

for $0 \le p \le 1$.

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

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

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

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

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 $Ga\left(\alpha_i, \theta\right)$ random variables. Find the distribution of $S = \sum_{i=1}^n X_i$.

Solution

The moment generating function of S is

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

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

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

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

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

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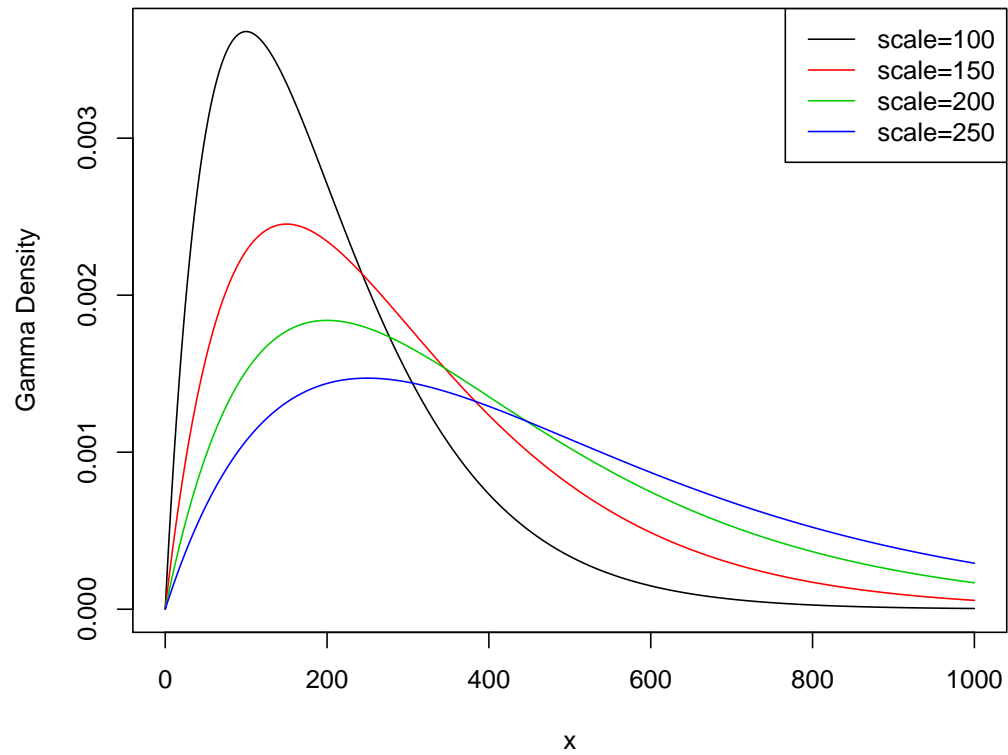
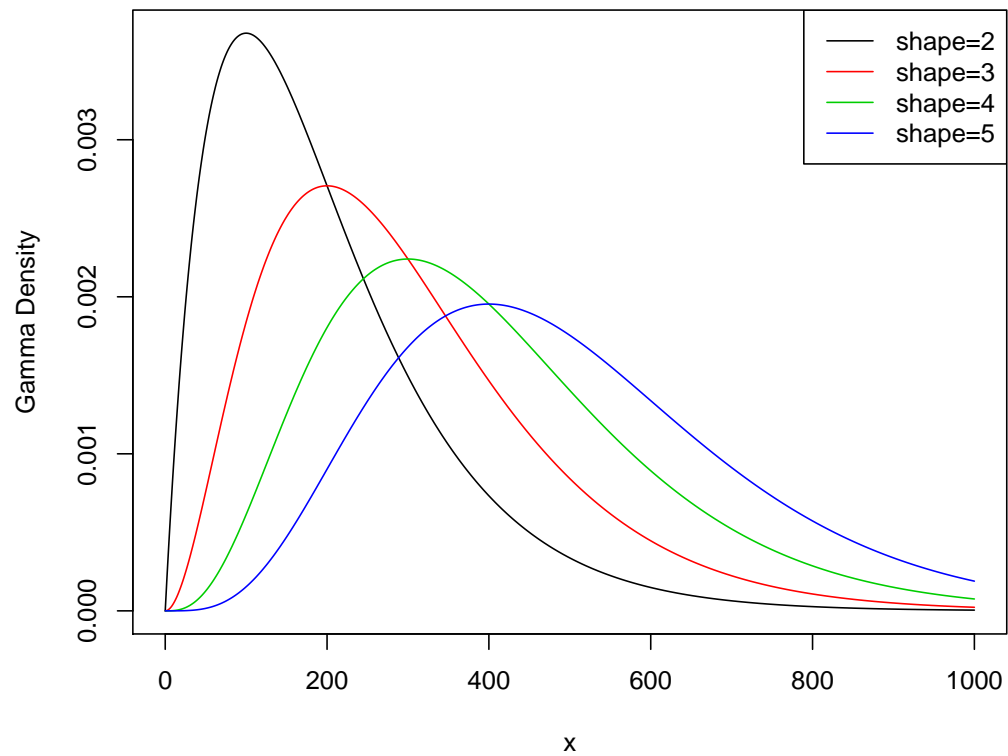Figure 3.1: Gamma Density, with shape=2 and Varying Scale

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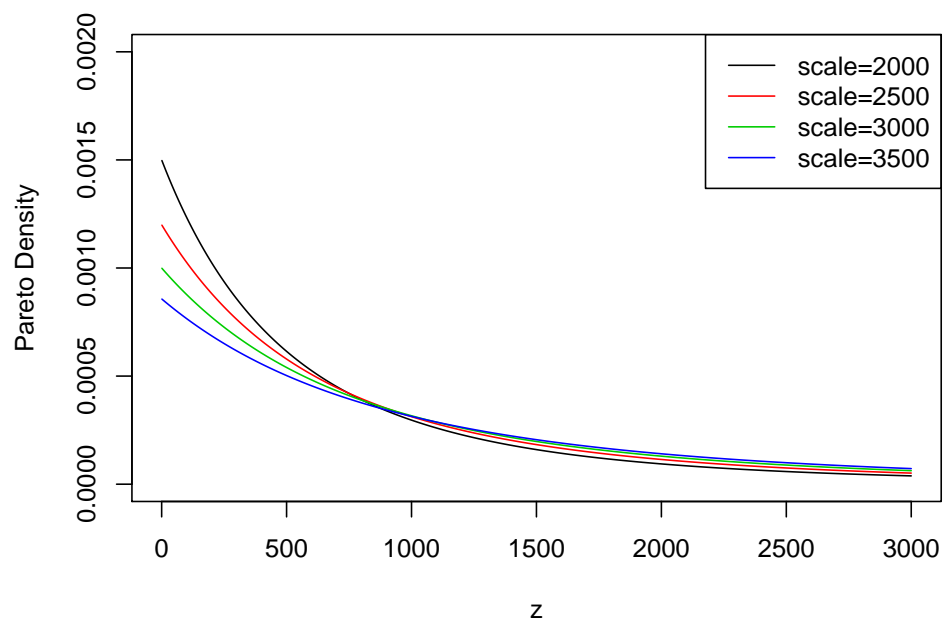
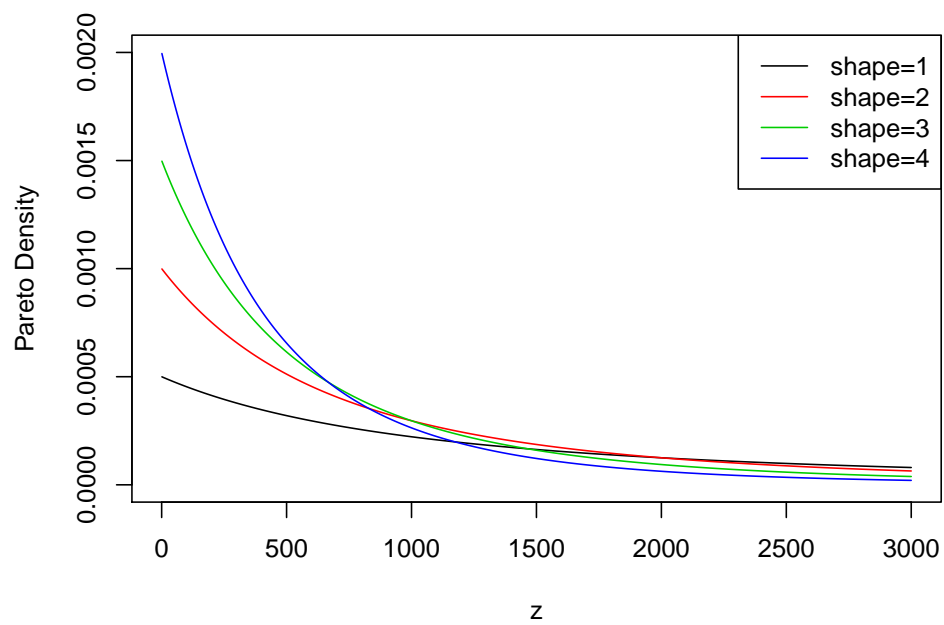Figure 3.3: Pareto Density, with shape=3 and Varying Scale

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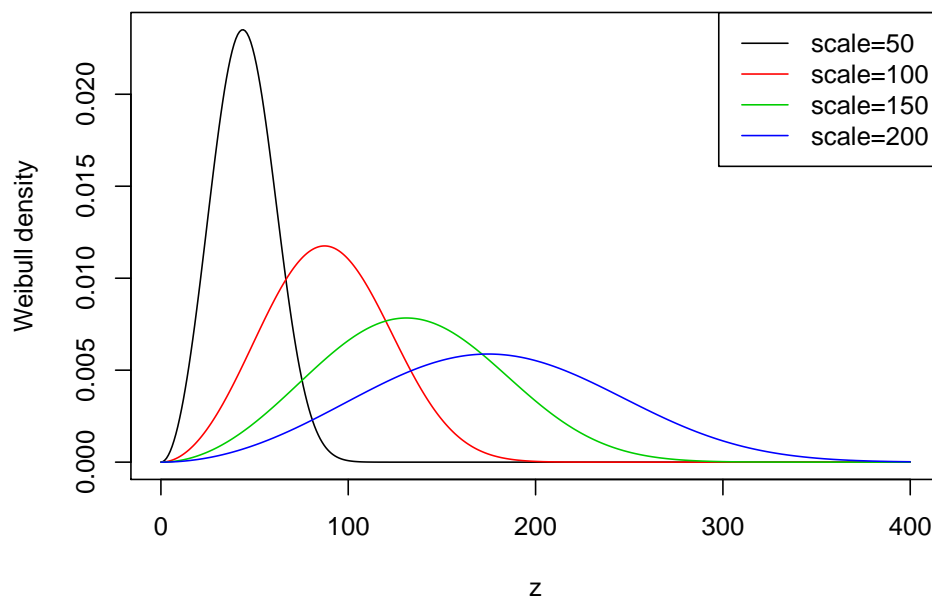
Figure 3.5: Weibull Density, with shape=3 and Varying Scale

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

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

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 densit
for(k in 2:length(shapeparam)){
  lines(z,dweibull(z,shape = shapeparam[k], scale = 100), col = k)}
legend("topright", c("shape=1.5", "shape=2", "shape=2.5", "shape=3"), lty=1, col = 1:4)
```

The distribution function of the Weibull distribution is given by

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

It can be easily seen that the shape parameter α describes the shape of the hazard function of the Weibull
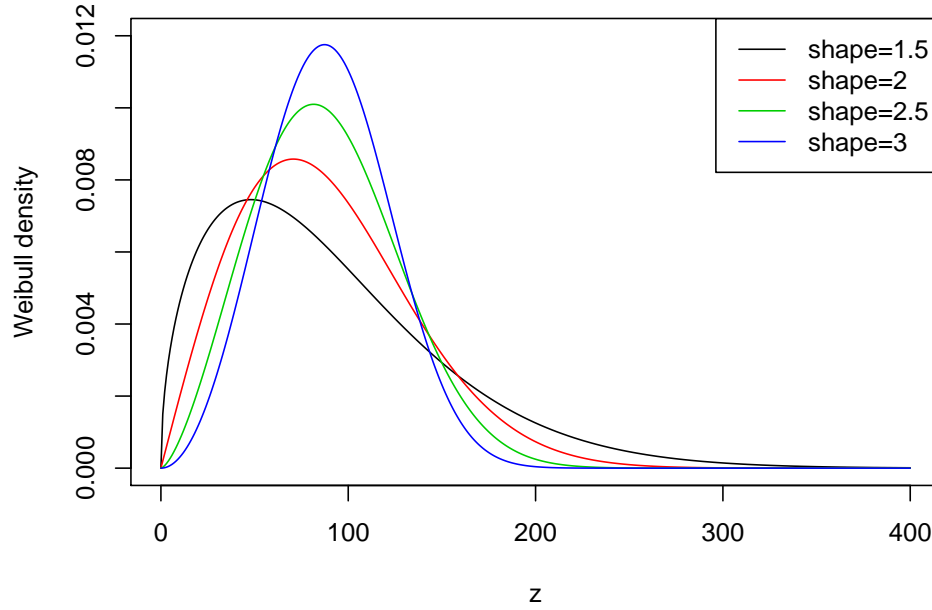
Figure 3.6: Weibull Density, with scale=100 and Varying Shape

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

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

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

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

Suppose that X belongs to a certain set of parametric distributions and define a rescaled version $Y = cX$, $c > 0$. If Y is in the same set of distributions then the distribution is said to be a scale distribution. When a member of a scale distribution is multiplied by a constant c ($c > 0$), the scale parameter for this scale distribution meets two conditions:

The parameter is changed by multiplying by c;

All other 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, $Y = 1.3X$.

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

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

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

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

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

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

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

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

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

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

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

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

3.3.4 Exponentiation

The normal distribution is a very popular model for a wide number of applications and when the sample size is large, it can serve as an approximate distribution for other models. If the random variable X has a normal distribution with mean μ and variance σ^2, then $Y = e^X$ has lognormal 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

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

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

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

Then,

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

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 $E(X^k) = a E(X_1^K) + (1-a) E(X_2^k)$, 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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

respectively.

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

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

In particular the mean and variance of X are given by

$$E(X) = E\left[E(X\,|\theta\;)\right]$$

and

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

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

$E(X) = E\left[E(X\,|q\;)\right] = E(100q) = 100E(q) = 60,$

$$Var(X) = E\left[Var(X\,|q\;)\right] + Var\left[E(X\,|q\;)\right] = E\left[100q(1-q)\right] + Var(100q)$$

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

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 $0 < \theta < \infty$.
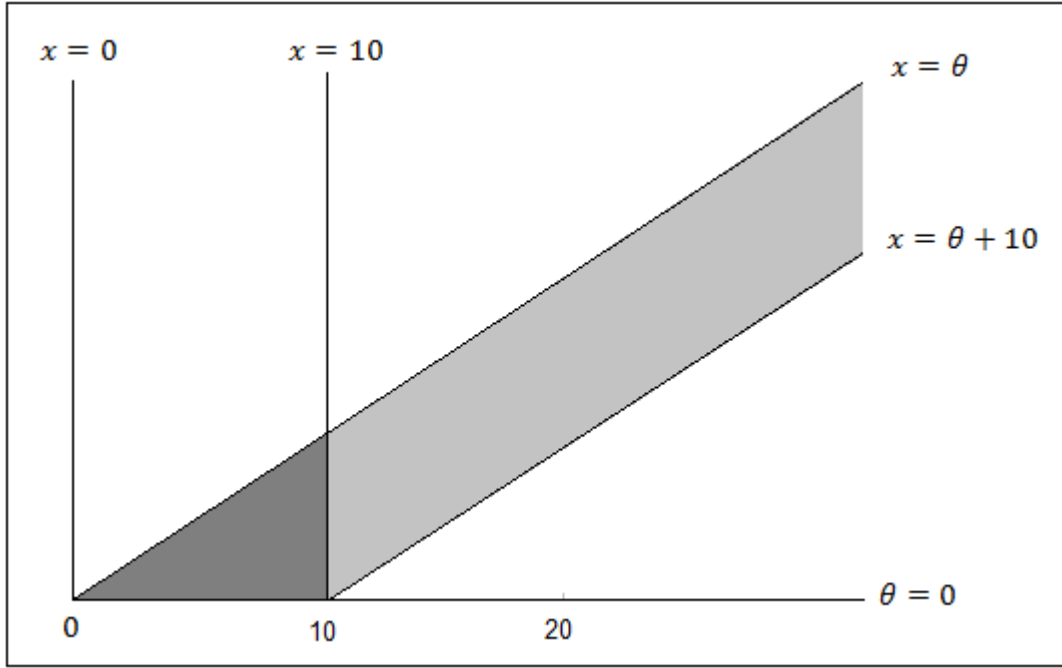
Figure 3.7:

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

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

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

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

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

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

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

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

On the other hand, the payment per 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

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

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 $X \leq d$) and the continuous part is spread over the interval $Y > 0$ (when $X > d$). For the discrete part, the probability that no payment is made is the probability that losses fall below the deductible; that is,

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

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

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

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

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

and

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

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

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

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

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

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

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

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

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,

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

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

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

Accordingly, the distribution function of Y is given by

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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 $E(X)$. If only an ordinary deductible is imposed, $\alpha=1$, $r=0$, $u=\infty$ and $E\left(Y^L\right)$ reduces to $E(X) - E(X \wedge d)$. If only a policy limit is imposed $\alpha=1$, $r=0$, $d=0$ and $E\left(Y^L\right)$ reduces to $E(X \wedge u)$.

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

$$Y_{2006}^L = \begin{cases} 0 & X \le 100, \\ X - 100 & 100 < X \le 600, \\ 500 & X > 600. \end{cases}$$

$$Y_{2007}^L = \begin{cases} 0 & X \le 95.24, \\ 1.05X - 100 & 95.24 < X \le 571.43, \\ 500 & X > 571.43. \end{cases}$$

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

$$= 356.026.$$

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

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

$$= 361.659.$$

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

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

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 $i = 2005, \ 2006$

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

where $i = 2005, \ 2006$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $g\left(\hat{\theta}\right)$.

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

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

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

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

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

The maximum likelihood estimators of μ and σ, denoted by $\hat{\mu}$ and $\hat{\sigma}$, are the solutions to the equations

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

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

These yield the estimates

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

The second partial derivatives are

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

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 $\ln X$ is normally distributed with mean μ and variance σ^2.

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

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

and

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

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

$$g\left(\mu,\sigma\right) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

is

$$g\left(\hat{\mu},\hat{\sigma}\right) = \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) = 153,277.$$

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

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

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

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

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

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

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:

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

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

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

Thus,

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

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

and

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

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

$$\frac{12}{\hat{\theta}} - \frac{6\hat{\theta}}{\left(\hat{\theta}^2 + 4\right)} - \frac{10\hat{\theta}}{\left(\hat{\theta}^2 + 16\right)} = 0$$

or

$$12\left(\hat{\theta}^2 + 4\right)\left(\hat{\theta}^2 + 16\right) - 6\hat{\theta}^2\left(\hat{\theta}^2 + 16\right) - 10\hat{\theta}^2\left(\hat{\theta}^2 + 4\right) = -4\hat{\theta}^4 + 104\hat{\theta}^2 + 768 = 0,$$

which yields $\hat{\theta}^2 = 32$ and $\hat{\theta} = 5.7$.

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

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

For the single parameter Pareto the probability density and distribution functions are given by

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

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

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

$$\ln L(\alpha) = 8\ln\alpha - (\alpha+1)\sum_{i=1}^{8}\ln x_i + 10\alpha \ln 5 - 2\alpha \ln 25.$$

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

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

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

which yields

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

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 and Inference

4.1 Nonparametric Estimation Tools

4.1.1 Moments

Moment Estimators

- $X_1, \ldots, X_n$ is a random sample (with replacement) from F(.)

- Sometimes we say that $X_1, \ldots, X_n$ are identically and independently distributed (*iid*)

We will not assume a parametric form for the distribution function F() and so proceed with a *nonparametric* analysis.

- The kth (*raw*) moment is E $X^k = \mu'_k$.

- It is estimated by the corresponding statistic

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

- The kth (central) moment is E $(X - \mu)^k = \mu_k$.

- It is estimated by

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

Empirical Distribution Function

- Define the **empirical distribution function** to be

$$F_n(x) = \frac{\text{number of observations less than or equal to } x}{n}$$

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

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

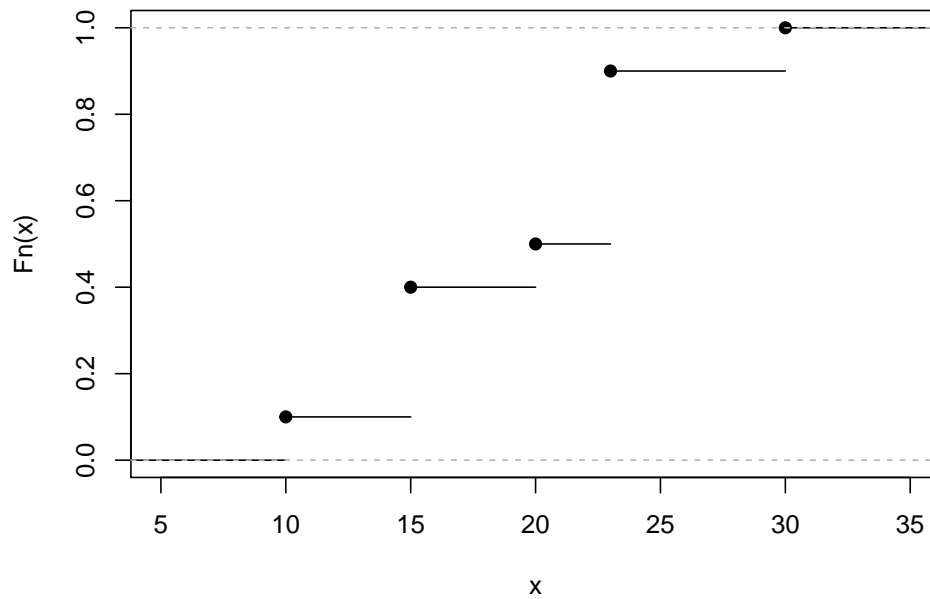- **Example** − **Toy**. Consider $n = 10$ observations as in Figure 4.1

Figure 4.1: Empirical Distribution Function of a Toy Example

i	1	2	3	4	5	6	7	8	9	10
X_i	10	15	15	15	20	23	23	23	23	30

- $\bar{x} = 19.7$ and that the estimate of the second central moment, the **sample variance**, is 34.45556.

```
##  [1] 10 15 15 15 20 23 23 23 23 30
```

R Code for Toy Example CDF

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

4.1.2   Quantiles

- Special Cases
  - The *median* is that point so that approximately half of a data set is below (or above) it.
  - The first *quartile* is that number so that approximately 25% of the data is below it.
  - A 100p *percentile* is that number so that $100 \times p$ percent of the data is below it.
- In general, for a given $0 < q < 1$, define the q**th quantile** q_F to be any number that satisfies

$$F(q_F-) \leq q \leq F(q_F).$$

  Here, the notation $F(x-)$ means to evaluate the function $F(\cdot)$ as a left-hand limit.
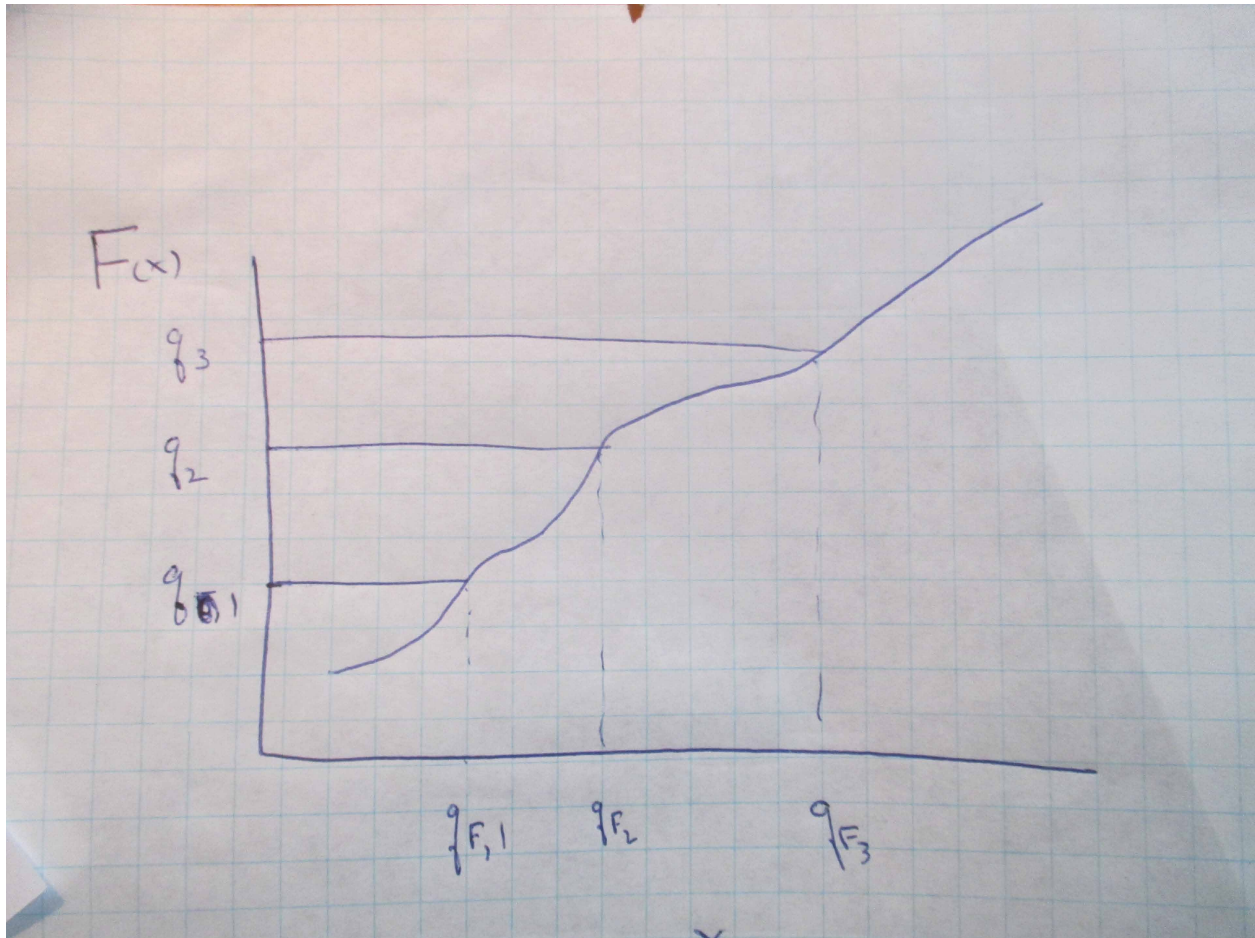- If $F(\cdot)$ is continuous at q_F, then $F(q_F-) = F(q_F)$

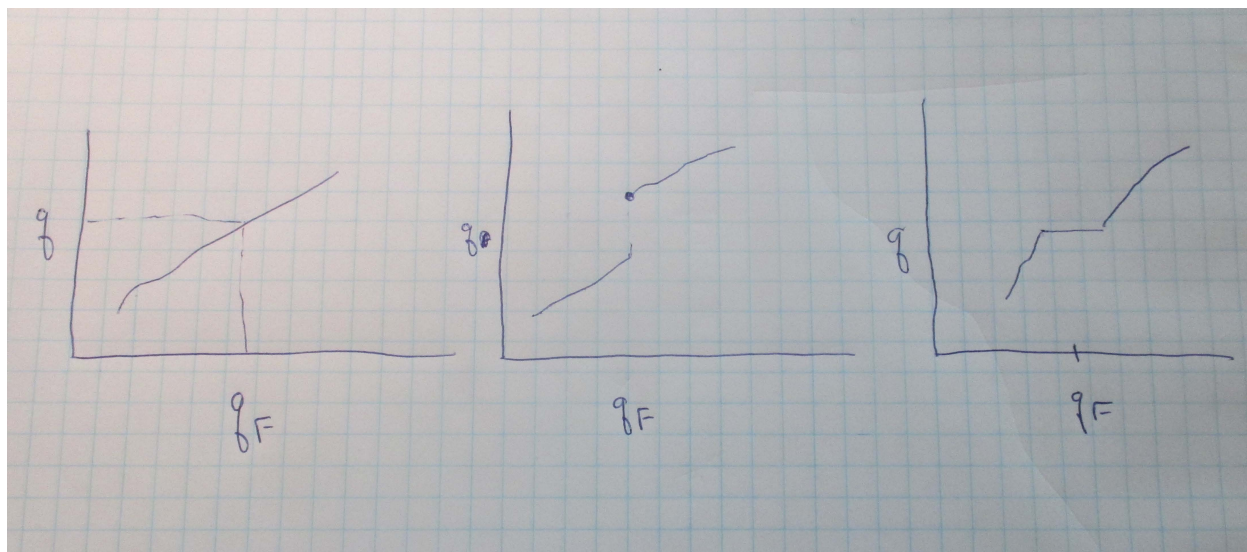Figure 4.2: Quantiles for a Continuous Distribution Function

Figure 4.3: Three Quantile Cases

Quantiles

- If F is smooth or there is a jump at q, the definition of the quantile q_F is unique
- if F is flat at q, then there a many definitions of q_F

Quantiles

Example – Toy. Consider $n = 10$ observations:

- The median might be defined to be any number between 20 and 23.
- Many software packages use the average 21.5.
- KPW defines the *smoothed empirical percentile* to be

$$\hat{\pi}_q = (1 - h)X_{(j)} + hX_{(j+1)}$$

where $j = [(n + 1)q]$ and, $h = (n + 1)q - j$, and $X_{(1)}, \ldots, X_{(n)}$ are the ordered values (the *order statistics*) corresponding to $X_1, \ldots, X_n$.

Example. Take $n = 10$ and $q = 0.5$. Then, $j = [(11)0.5] = [5.5] = 5$ and, $h = (11)(0.5) - 5 = 0.5$. With this

$$\hat{\pi}_{0.5} = (1 - 0.5)X_{(5)} + (0.5)X_{(6)} = 0.5(20) + (0.5)(23) = 21.5.$$

Take $n = 10$ and $q = 0.2$. Then, $j = [(11)0.2] = [2.2] = 2$ and, $h = (11)(0.2) - 2 = 0.2$. With this

$$\hat{\pi}_{0.2} = (1 - 0.2)X_{(2)} + (0.2)X_{(3)} = 0.2(15) + (0.8)(15) = 15.$$

4.1.3   Density Estimators

- When the random variable is discrete, estimate the probability mass function $f(x) = \Pr(X = x)$ is using

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

- Observations may be grouped in the sense that they fall into intervals of the form $[c_{j-1}, c_j)$, for $j = 1, \ldots, k$. The constants $\{c_0 < c_1 < \cdots < c_k\}$ form some partition of the domain of F(.).

- Then, use

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

  where n_j is the number of observations (X_i) that fall into the interval $[c_{j-1}, c_j)$.

- Another way to write this is

$$f_n(x) = \frac{1}{n(c_j - c_{j-1})} \sum_{i=1}^{n} I(c_{j-1} < X_i \leq c_j).$$

Uniform Kernel Density Estimator

- Let $b > 0$, known as a bandwidth,

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

- The estimator is the average over n *iid* realizations of a random variable with mean

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

  as $b \to 0$. That is, $f_n(x)$ is an asymptotically unbiased estimator of $f(x)$.

Kernel Density Estimator

- More generally, define the **kernel density estimator**

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

  where k is a probability density function centered about 0.

- Special Cases

  - uniform kernel, $k(y) = \frac{1}{2} I(-1 < y \leq 1)$.

  - triangular kernel, $k(y) = (1 - |y|) \times I(|y| \leq 1)$

  - Epanechnikov kernel, $k(y) = \frac{3}{4}(1 - y^2) \times I(|y| \leq 1)$, and

  - Gaussian kernel $k(y) = \phi(y)$, where $\phi(\cdot)$ is the standard normal density function.

Kernel Density Estimator of a Distribution Function

- The kernel density estimator of a **distribution function** is

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right).$$

  where K is a probability distribution function associated with the kernel density k.

- To illustrate, for the uniform kernel, we have $k(y) = \frac{1}{2} I(-1 < y \le 1)$ so

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

4.2 Nonparametric Estimation Tools For Model Selection

4.2.1 Graphical Comparisions

Comparing 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.

PP **Plot**

- 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.

KPW also recommends plotting the difference $D(x) = F_n(x) - F^*(x)$ versus x. Here, $F^*(x)$ is the fitted model distribution function.

QQ **Plot**

- 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.

- The lower-right hand panel suggests that the Pareto distribution does a good job with large observations but provides a poorer fit for small observations.
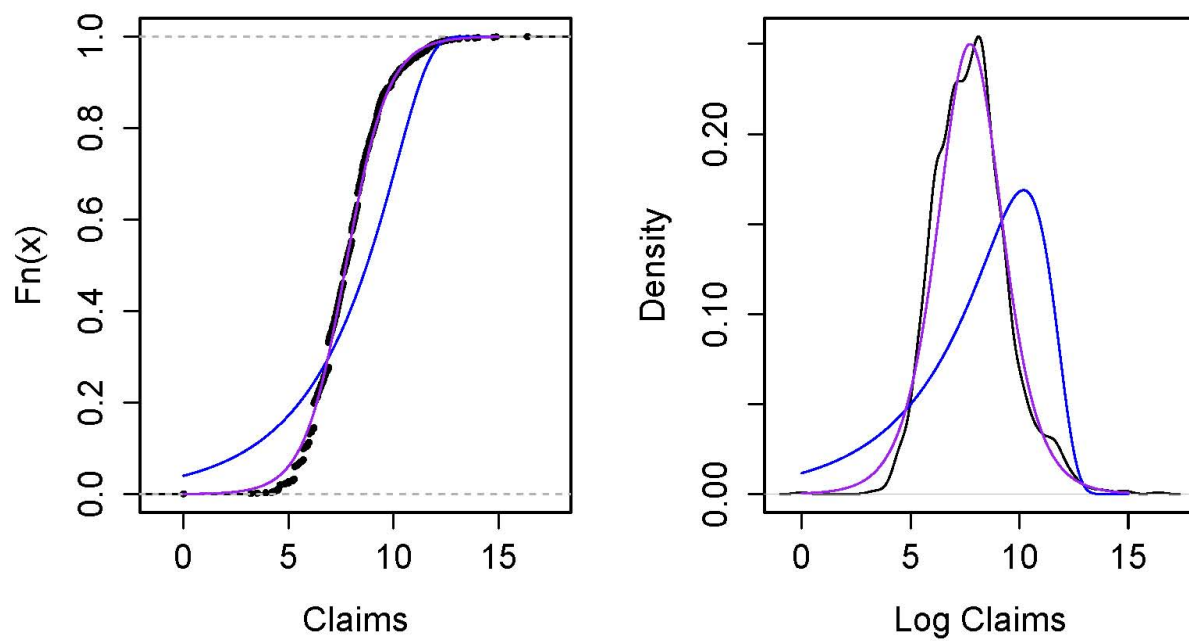
Figure 4.4: Nonparametric Versus Fitted Parametric Distribution and Density Functions
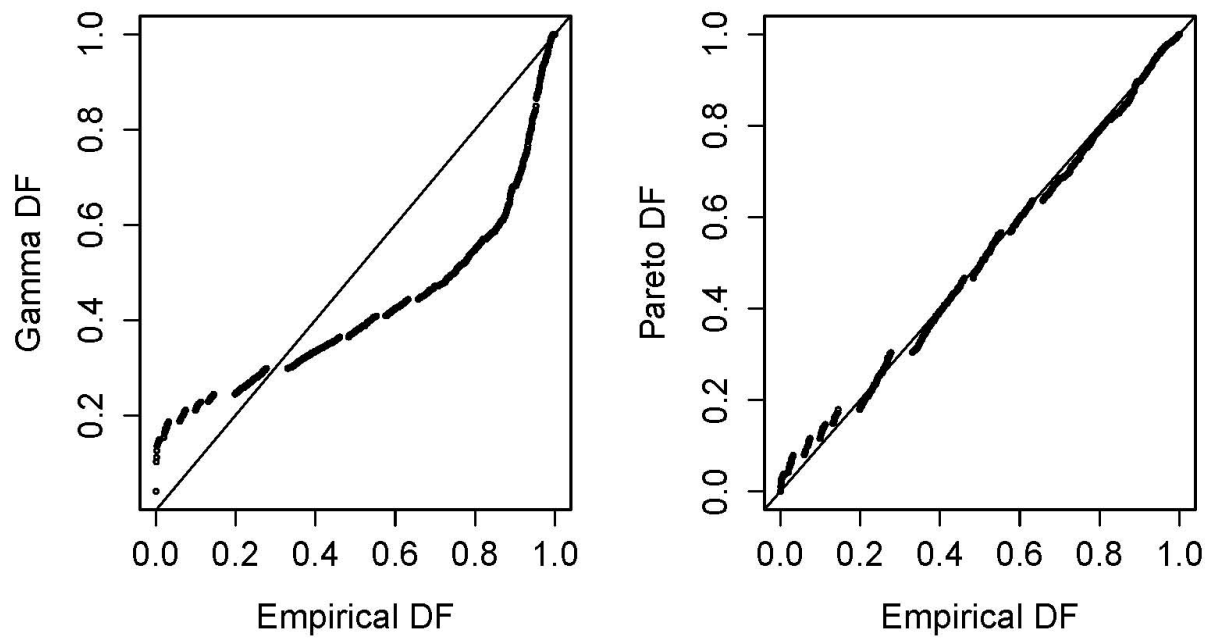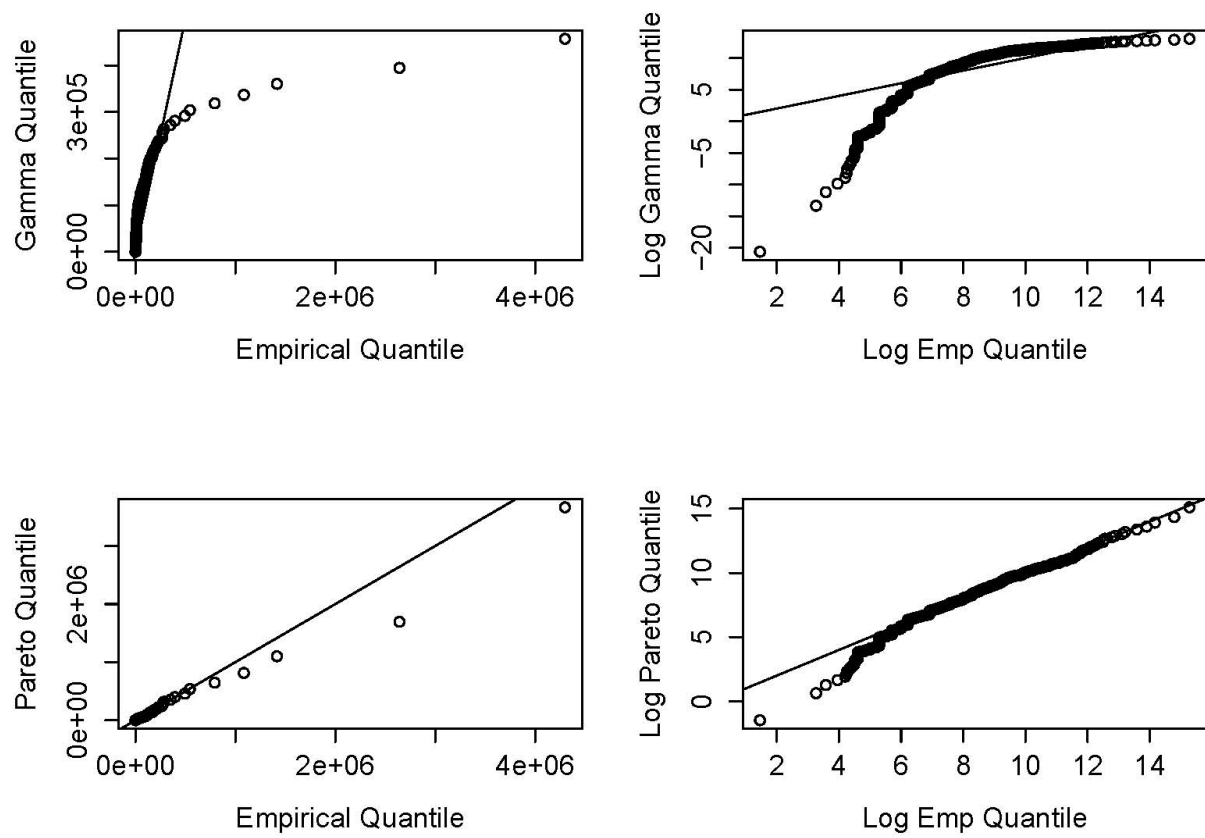

Figure 4.5: Probability-Probability (pp) Plots.

Figure 4.6: Quantile-Quantile (qq) Plots

4.2.2 Statistical Comparisions

Three Goodness of Fit Statistics

Statistic	Definition	Computational Expression
Kolmogorov $-Smirnov$	$sup_x|F_n(x) - F(x)|$	$max(D^+ - D^-)$ $D^+ = \max_{i=1,...,n}\left(\frac{i}{n} - F_i\right)$ $D^- = \max_{i=1,...,n}\left(F_i - \frac{i-1}{n}\right)$
Cramer $-vonMises$	$n\int (F_n(x) - F(x))^2 dx$	$\frac{1}{12n} + \sum_{i=1}^n \left(F_i - (2i-1)/n\right)^2$
Anderson $-Darling$	$n\int \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))}dx$	$-\frac{1}{n}\sum_{i=1}^n (2i-1)\log\left(F_i(1-F_{n+1-i})\right)^2$

4.3 Nonparametric Estimation using Modified Data

4.3.1 Grouped Data

Grouped Data

- Observations may be grouped in the sense that they fall into intervals of the form $[c_{j-1}, c_j)$, for $j = 1, \ldots, k$.

- The constants $\{c_0 < c_1 < \cdots < c_k\}$ form some partition of the domain of F(.).

- Define the empirical distribution function at the boundaries is defined in the usual way:

$$F_n(c_j) = \frac{\text{number of observations } \le c_j}{n}$$

- For other values of x, one could use the

  Ogive: connect values of the boundaries with a straight line.

- For another way of smoothing, recall the kernel density estimator of the distribution function

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n K\left(\frac{x - X_i}{b}\right).$$

- For densities, use

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

4.3.2 Censored Data

Censored Data

- Censoring occurs when we observe only a limited value of an observation.

- Suppose that X represents a loss due to an insured event and that u is a known censoring point.

- If observations are censored from the **right** (or from above), then we observe

$$Y = \min(X, u).$$

- – In this case, u may represent the upper limit of coverage for an insurer. The loss exceeds the amount u but the insurer does not have in its records the amount of the actual loss.

- If observations are censored from the **left** (or from below), then we observe

$$Y = \max(X, u).$$

- – Let u represents the upper limit of coverage but now $Y - u$ represents the amount that a *reinsurer* is responsible for. If the loss $X < u$, then $Y = 0$, no loss for the reinsurer. If the loss $X \geq u$, then $Y = X - u$ represents the reinsurer's retained claims.

Kaplan-Meier Product Limit Estimator

- Let $t_1 < \cdots < t_c$ be distinct points at which an event of interest occurs, or non-censored losses, and let s_j be the number of events at time point t_j .

- Further, the corresponding risk set is the number of observations that are active at an instant just prior to t_j . Using notation, the risk set is $R_j = \sum_{i=1}^{n} I(x_i \geq t_j)$.

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

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

- Greenwood (1926) derived the formula for the estimated variance

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

4.3.3   Truncated Data

Truncated 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 truncated from the **left** (or below) at d when the amount observed is

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

- – In this case, d may represent the deductible associated with an insurance coverage. If the insured loss is less than the deductible, then the insurer does not observe the loss. If the loss exceeds the deductible, then the excess $X - d$ is the claim that the insurer covers.

- Observations may also truncated from the **right** (or above) at d when the amount observed is

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

- – Classic examples of truncation from the right include X as a measure of distance of a star. When the distance exceeds a certain level d, the star is no longer observable.

Right-Censored, Left-Truncated Empirical Distribution Functions

- Procedure from **KPW**. Notation:
    - For each observation i, let d_i be the lower truncation limit (0 if no truncation)
    - Let u_i be the upper censoring limit ($=\infty$ if no censoring)
    - The recorded value is x_i in the case of no censoring, u_i if there is censoring.
    - For notation, let $t_1 < \cdots < t_k$ be k unique observations of x_i that are uncensored.
    - Define s_j to be the number of x_i's at t_j.
    - Define the risk set

$$R_j = \sum_{i=1}^n I(x_i \geq t_j) + \sum_{i=1}^n I(u_i \geq t_j) - \sum_{i=1}^n I(d_i \geq t_j)$$

- The product-limit estimator of the distribution function is

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

- The Nelson-Aalen estimator of the distribution function is

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

4.4 Topics in Parametric Estimation

4.4.1 Starting Values

- Maximum likelihood is a desirable estimation technique because
    - It employs data efficiently (enjoys certain optimality properties)
    - It can be used in a variety of data sampling schemes (e.g.,*iid*, grouped, censored, regression, and so forth)
- However, maximum likelihood is a recursive estimation procedure that requires starting values to begin the recursion
- Two alternative estimation techniques are:
    - Method of moments
    - Percentile matching
- These are non-recursive techniques. Easy to implement and explain. Although less efficient than maximum likelihood, they can be employed to provide starting values for maximum likelihood.

Method of Moments

- Idea: Approximate the moments using a parametric distribution to the empirical (nonparametric) moments

- **Example - Property Fund.** For the 2010 property fund, there are $n = 1,377$ individual claims (in thousands of dollars) with

$$m_1 = \frac{1}{n} \sum_{i=1}^{n} X_i = 26.62259 \quad \text{and} \quad m_2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 = 136154.6.$$

- Gamma Distribution

  - From theory, $\mu_1 = \alpha\theta$ and $\mu_2' = \alpha(\alpha + 1)\theta^2$.

  - Equating the two yields the method of moments estimators, easy algebra shows that

$$\alpha = \frac{\mu_1^2}{\mu_2' - 2\mu_1^2} \quad \text{and} \quad \theta = \frac{\mu_2' - \mu_1^2}{\mu_1}.$$

  - The method of moment estimators are

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

  - In contrast, the maximum likelihood values turn out to be $\hat{\alpha}_{MLE} = 0.2905959$ and $\hat{\theta}_{MLE} = 91.61378$

  - Big discrepancies between the two estimation procedures, suggesting that the gamma model fits poorly.

- **Example - Property Fund.** Recall the nonparametric estimates

$$m_1 = \frac{1}{n} \sum_{i=1}^{n} X_i = 26.62259 \quad \text{and} \quad m_2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 = 136154.6.$$

- Pareto Distribution

  - From theory, $\mu_1 = \theta/(\alpha - 1)$ and $\mu_2' = 2\theta^2/((\alpha - 1)(\alpha - 2))$.

  - Easy algebra shows
$$\alpha = 1 + \frac{\mu_2'}{\mu_2' - 2\mu_1^2} \quad \text{and} \quad \theta = (\alpha - 1)\mu_1.$$

  - The method of moment estimators are

$$\hat{\alpha} = 1 + \frac{136154.6}{136154.6 - 2 * 26,62259^2} = 2.01052$$
$$\hat{\theta} = (2.01052 - 1) \cdot 26.62259 = 26.9027$$

  - The maximum likelihood values turn out to be $\hat{\alpha}_{MLE} = 0.9990936$ and $\hat{\theta}_{MLE} = 2.2821147$.

  - Interesting that $\hat{\alpha}_{MLE} < 1$; for the Pareto distribution; recall that $\alpha < 1$ means that the mean is infinite.

  - Indicates that the property claims data set is a long tail distribution.

Percentile Matching

- Under percentile matching, one approximates the parametric distribution using the empirical (non-parametric) quantiles, or percentiles.

- **Example - Property Fund.** The 25th percentile (the first quartile) turns out to be 0.78853 and the 95th percentile is 50.98293 (both in thousands of dollars).

- Pareto Distribution

  - The Pareto distribution is particularly intuitively pleasing because of the closed-form solution for the quantiles.

  - The distribution function is $F(x) = 1 - (\theta/(x+\theta))^\alpha$.

  - 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, $0 < q < 1$.

  - With two equations

  $$0.78853 = \theta \left((1 - .25)^{-1/\alpha} - 1 \right) \quad \text{and} \quad 50.98293 = \theta \left((1 - .95)^{-1/\alpha} - 1 \right)$$

  and two unknowns, the solution is

  $$\hat{\alpha} = 0.9412076 \quad \text{and} \quad \hat{\theta} = 2.205617.$$

  - A numerical routine was required for these solutions - no analytic solution available.

  - Recall that the maximum likelihood values are $\hat{\alpha}_{MLE} = 0.9990936$ and $\hat{\theta}_{MLE} = 2.2821147$.

  - The percentile matching provides a better approximation for the Pareto distribution than did the method of moments.

4.4.2 Grouped Data

Parametric Estimation Using Grouped Data

- Observations may be grouped in the sense that they fall into intervals of the form $(c_{j-1}, c_j]$, for $j = 1, \ldots, k$.

- The constants $\{c_0 < c_1 < \cdots < c_k\}$ form some partition of the domain of F(.).

- Define n_j to be the number of observations that fall in the jth interval, $(c_{j-1}, c_j]$.

- The probability of an observation X falling in the jth interval is

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

- The probability of an observation X falling in the jth interval is

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

- The corresponding mass function is

$$f(x) = \begin{cases} F(c_1) - F(c_0) & \text{if } x \in (c_0, c_1] \\ \vdots & \vdots \\ F(c_k) - F(c_{k-1}) & \text{if } x \in (c_{k-1}, c_k] \end{cases}$$

$$= \prod_{j=1}^{k} \{F(c_j) - F(c_{j-1})\}^{I(x \in (c_{j-1}, c_j])}$$

- The likelihood is

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

- The log-likelihood is

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

4.4.3   Parametric Estimation Using Censored Data

Censored Data Likelihood

- Censoring occurs when we observe only a limited value of an observation.
- Suppose that X represents a loss due to an insured event and that u is a known censoring point.
- If observations are censored from the **right** (or from above), then we observe we observe $Y = \min(X, u)$ and $\delta_u = \mathrm{I}(X \geq u)$.
- If censoring occurs so that $\delta_u = 1$, then $X \geq u$ and the likelihood is $\Pr(X \geq u) = 1 - \mathrm{F}(u)$.
- If censoring does not occur so that $\delta_u = 0$, then $X < C_U$ and the likelihood is $\mathrm{f}(y)$.
- Summarizing, we have

$$Likelihood = \begin{cases} \mathrm{f}(y) & \text{if } \delta = 0 \\ 1 - \mathrm{F}(u) & \text{if } \delta = 1 \end{cases}$$

$$= (\mathrm{f}(y))^{1-\delta} (1 - \mathrm{F}(u))^{\delta}.$$

The second right-hand expression allows us to present the likelihood more compactly.

- For a single observation, we have

$$Likelihood = \begin{cases} \mathrm{f}(y) & \text{if } \delta = 0 \\ 1 - \mathrm{F}(u) & \text{if } \delta = 1 \end{cases}$$

$$= (\mathrm{f}(y))^{1-\delta} (1 - \mathrm{F}(u))^{\delta}.$$

- Consider a random sample of size n,

$$\{(y_1, \delta_1), \ldots, (y_n, \delta_n)\}$$

with potential censoring times $\{ u_1, \ldots, u_n \}$.
- The likelihood is

$$\prod_{i=1}^{n} (\mathrm{f}(y_i))^{1-\delta_i} (1 - \mathrm{F}(u_i))^{\delta_i} = \prod_{\delta_i=0} \mathrm{f}(y_i) \prod_{\delta_i=1} \{1 - \mathrm{F}(u_i)\},$$

No observed value under

left–truncation

No observed value under

right–truncation

No exact value under

interval–censoring

No exact value under

left–censoring

No exact value under

right–censoring

$$\xrightarrow{\hspace{2cm}\vdash\hspace{1cm}\vdash\hspace{1cm}\vdash\hspace{2cm}}$$

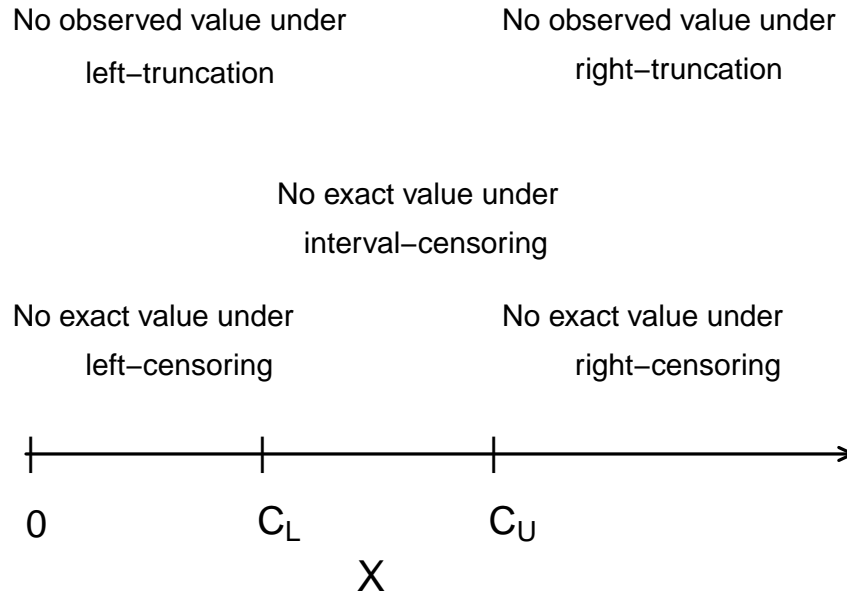0 $\qquad$ C_L $\qquad$ C_U

X

Figure 4.7: Censoring and Truncation

- Here, the notation $\prod_{\delta_i=0}$ means take the product over uncensored observations, and similarly for $\prod_{\delta_i=1}$.

- The log-likelihood is

$$L(\theta) = \sum_{i=1}^{n} \left\{ (1 - \delta_i) \ln \mathrm{f}(y_i) + \delta_i \ln \left(1 - \mathrm{F}(u_i)\right) \right\}$$

4.4.4 Censored and Truncated Data

- Let X denote the outcome and let C_L and C_U be two constants.

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 $X < C_L$

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.

- For each subject, the beginning of the arrow represents that the the subject was recruited and the
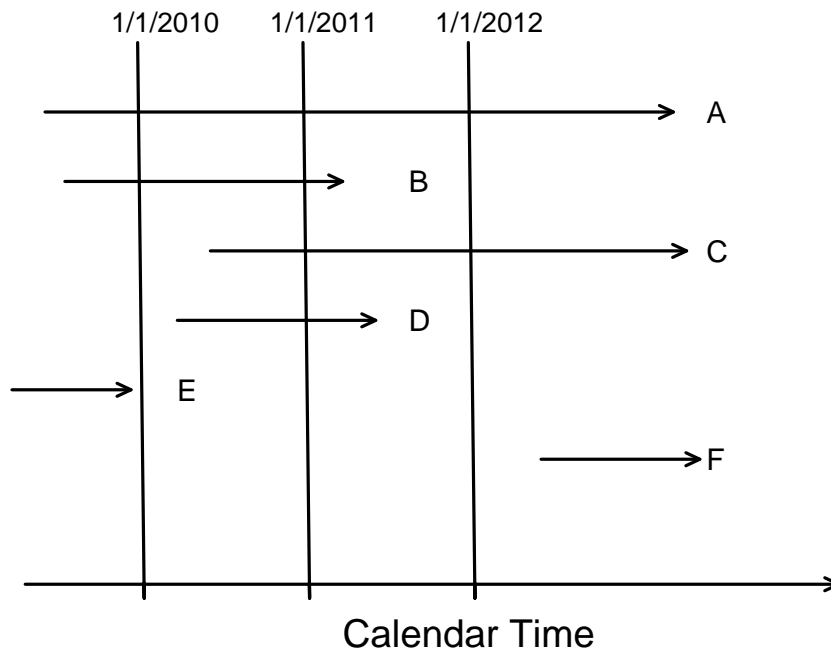
Figure 4.8: Subjects on Test in a Mortality Study

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.

4.4.5   Parametric Estimation Using Censored and Truncated Data

- Truncated data are handled in likelihood inference via conditional probabilities.

- Adjust the likelihood contribution by dividing by the probability that the variable was observed.

- Summarizing, we have the following contributions to the likelihood 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

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

  where $\prod_E$ is the product over observations with *E*xact values, and similarly for *R*ight-, *L*eft- and *I*nterval-censoring.

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

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

- Similarly for other combinations.

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 logarithmic likelihood is

$$\ln Likelihood = \sum_E \left(\ln f(x_i) - \ln(1 - F(C_{Li})) \right) - \sum_R \left(\ln(1 - F(C_{Ui})) - \ln(1 - F(C_{Li})) \right)$$
$$= \sum_E (-\ln \theta - (x_i - C_{Li})/\theta) - \sum_R (C_{Ui} - C_{Li})/\theta.$$

- To simplify the notation, define $\delta_i = I(X_i \geq C_{Ui})$ to be a binary variable that indicates right-censoring.

- Let $X_i^{**} = \min(X_i, C_{Ui}) - C_{Li}$ be the amount that the observed variable exceeds the lower truncation limit.

- With this, the logarithmic likelihood is

$$\ln Likelihood = -\sum_{i=1}^{n} \left((1 - \delta_i) \ln \theta + \frac{x_i^{**}}{\theta} \right).$$

- 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.

4.5   Bayesian Inference

4.5.1   Bayesian Model

- In the **frequentist interpretation**, one treats the vector of parameters $\boldsymbol{\theta}$ as fixed yet unknown, whereas the outcomes X are realizations of random variables.

- With Bayesian statistical models, one views both the model parameters and the data as random variables.

  - Use probability tools to reflect this uncertainty about the parameters $\boldsymbol{\theta}$.

- For notation, we will think about $\boldsymbol{\theta}$ as a random vector and let $\pi(\boldsymbol{\theta})$ denote the distribution of possible outcomes.

Bayesian Inference Strengths

There are several advantages of the Bayesian approach.

1. One can describe the entire distribution of parameters conditional on the data. This allows one, for example, to provide probability statements regarding the likelihood of parameters.

2. This approach allows analysts to blend information known from other sources with the data in a coherent manner. This topic is developed in detail in the credibility chapter.

3. The Bayesian approach provides for a unified approach for estimating parameters. Some non-Bayesian methods, such as least squares, required a approach to estimating variance components. In contrast, in Bayesian methods, all parameters can be treated in a similar fashion. Convenient for explaining results to consumers of the data analysis.

4. Bayesian analysis is particularly useful for forecasting future responses.

Bayesian Model

- **Prior Distribution.** $\pi(\boldsymbol{\theta})$ is called the *prior distribution*.

  - Typically, it is a regular distribution and so integrates to one.

  - We may be very uncertain (or have no clue) about the distribution of $\boldsymbol{\theta}$; the Bayesian machinery allows this 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, not surprisingly, known as the *joint distribution* and denoted as $f(x, \boldsymbol{\theta}) = f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

- **Marginal Outcome Distribution.** The distribution of outcomes can be expressed as

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

This is analogous to a frequentist mixture distribution.

- **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

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

Two Examples

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 $0.7 < q < 0.8$.

Exam C Question 43. You are given:
(i) The prior distribution of the parameter Θ has probability density function:

$$\pi(\theta) = 1/\theta^2, \quad 1 < \theta < \infty$$

(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.

4.5.2   Bayesian Inference - 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 $\mathrm{E}\, l(\hat{\theta}, \theta)$.

- Some important special cases include:

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

- 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

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

Posterior Distribution

How to calculate the posterior distribution?

- **By hand** - can do this in special cases

- **Simulation** - uses modern computational techniques. **KPW** (Section 12.4.4) mentions 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

$$\pi(\boldsymbol{\theta}|x) = \frac{f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(x)}$$
$$\propto f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

Posterioris proportional tolikelihood × prior

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

Special Case: Poisson Gamma Conjugate Family

- 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

$$\pi(\lambda|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\lambda)$$
$$= C\lambda^{\sum_i x_i + \alpha - 1} \exp(-\lambda(n + 1/\theta))$$

where C is a constant.

- We recognize this to be a gamma distribution with new parameters $\alpha_{new} = \sum_i x_i + \alpha$ and $\theta_{new} = 1/(n + 1/\theta)$.

4.6   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-model-selec
```

Chapter 5

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.

5.1 Generating Independent Uniform Observations

We begin with a historically prominent method.

Linear Congruential Generator. To generate a sequence of random numbers, start with B_0, a starting value that is known as a "seed." Update it using the recursive relationship

$$B_{n+1} = aB_n + c \text{ modulo } m, \quad n = 0, 1, 2, \ldots.$$

This algorithm is called a *linear congruential generator*. The case of $c = 0$ is called a *multiplicative* congruential generator; it is particularly useful for really fast computations.

For illustrative values of a and m, Microsoft's Visual Basic uses $m = 2^{24}$, $a = 1,140,671,485$, and $c = 12,820,163$ (see http://support.microsoft.com/kb/231847). This is the engine underlying the random number generation in Microsoft's Excel program.

The sequence used by the analyst is defined as $U_n = B_n/m$. The analyst may interpret the sequence $\{U_i\}$ to be (approximately) identically and independently uniformly distributed on the interval (0,1). To illustrate the algorithm, consider the following.

Example. Take $m = 15$, $a = 3$, $c = 2$ and $B_0 = 1$. Then we have:

step n	B_n	U_n
0	$B_0 = 1$	
1	$B_1 = \mod(3 \times 1 + 2) = 5$	$U_1 = \frac{5}{15}$
2	$B_2 = \mod(3 \times 5 + 2) = 2$	$U_2 = \frac{2}{15}$
3	$B_3 = \mod(3 \times 2 + 2) = 8$	$U_3 = \frac{8}{15}$
4	$B_4 = \mod(3 \times 8 + 2) = 11$	$U_4 = \frac{11}{15}$

Sometimes computer generated random results are known as *pseudo*-random numbers to reflect the fact that they are machine generated and can be replicated. That is, despite the fact that $\{U_i\}$ appears to be i.i.d, it can be reproduced by using the same seed number (and the same algorithm). The ability to replicate results can be a tremendous tool as you use simulation while trying to uncover patterns in a business process.

The linear congruential generator is just one method of producing pseudo-random outcomes. It is easy to understand and is (still) widely used. The linear congruential generator does have limitations, including the fact that it is possible to detect long-run patterns over time in the sequences generated (recall that we can interpret "independence" to mean a total lack of functional patterns). Not surprisingly, advanced techniques have been developed that address some of this method's drawbacks.

5.2   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*

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

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

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

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:

$$y = F(x) \Leftrightarrow y = 1 - e^{-x/\theta}$$
$$\Leftrightarrow -\theta \ln(1-y) = x = F^{-1}(y).$$

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

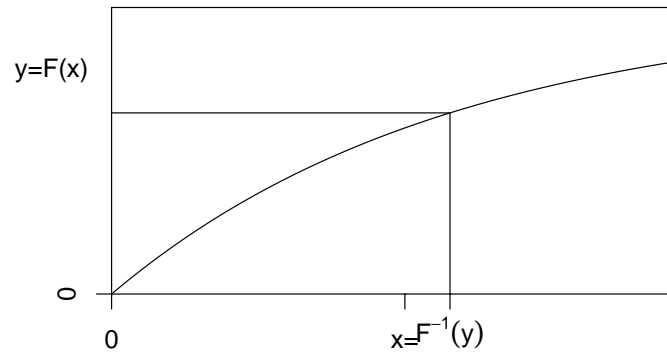| U	0.26321364	0.196884752	0.897884218
$X = -10\ln(1-U)$	1.32658423	0.952221285	9.909071325

Figure 5.1: 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:

$$y = F(x) \Leftrightarrow 1 - y = \left(\frac{\theta}{x+\theta}\right)^{\alpha}$$
$$\Leftrightarrow (1-y)^{-1/\alpha} = \frac{x+\theta}{\theta} = \frac{x}{\theta} + 1$$
$$\Leftrightarrow \theta\left((1-y)^{-1/\alpha} - 1\right) = x = F^{-1}(y).$$

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 $\Pr(U \leq y) = y$, for $0 \leq y \leq 1$. Thus,

$$\Pr(X \leq x) = \Pr(F^{-1}(U) \leq x)$$
$$= \Pr(F(F^{-1}(U)) \leq F(x))$$
$$= \Pr(U \leq F(x)) = F(x)$$

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

$$F^{-1}(y) = \begin{cases} 0 & 0 < y \leq 0.85 \\ 1 & 0.85 < y \leq 1.0. \end{cases}$$
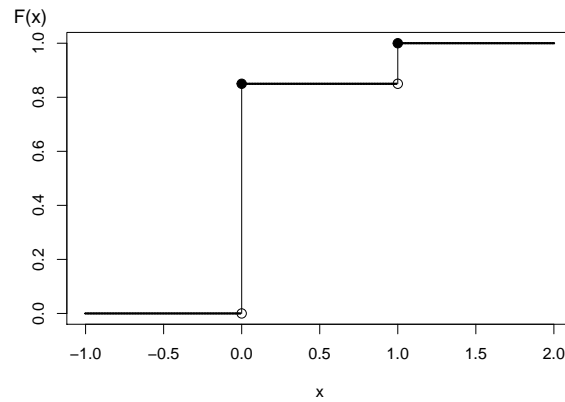
Figure 5.2: Distribution Function of a Binary Random Variable

Thus, with the inverse transform we may define

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

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

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

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:

$$X = \begin{cases} \text{whole life} & 0 < U \le 0.1 \\ \text{endowment} & 0.1 < U \le 0.3 \\ \text{term life} & 0.3 < U \le 0.4 \\ \text{universal life} & 0.4 < U \le 0.8 \\ \text{variable life} & 0.8 < U \le 1.0. \end{cases}$$

Another analyst may use an alternative procedure such as:

$$X = \begin{cases} \text{whole life} & 0.9 < U < 1.0 \\ \text{endowment} & 0.7 \le U < 0.9 \\ \text{term life} & 0.6 \le U < 0.7 \\ \text{universal life} & 0.2 \le U < 0.6 \\ \end{cases}$$
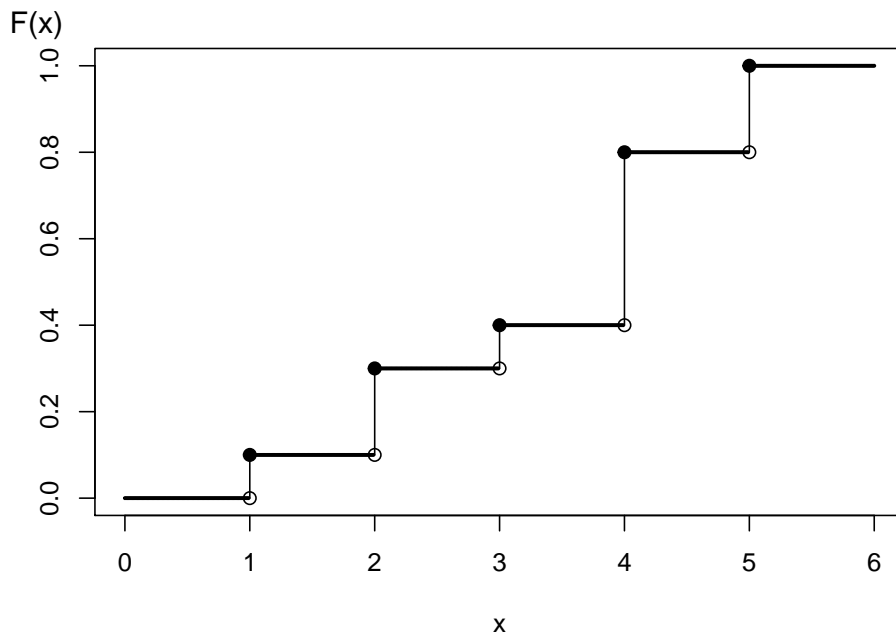
Figure 5.3: Distribution Function of a Discrete Random Variable

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

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.

5.3 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

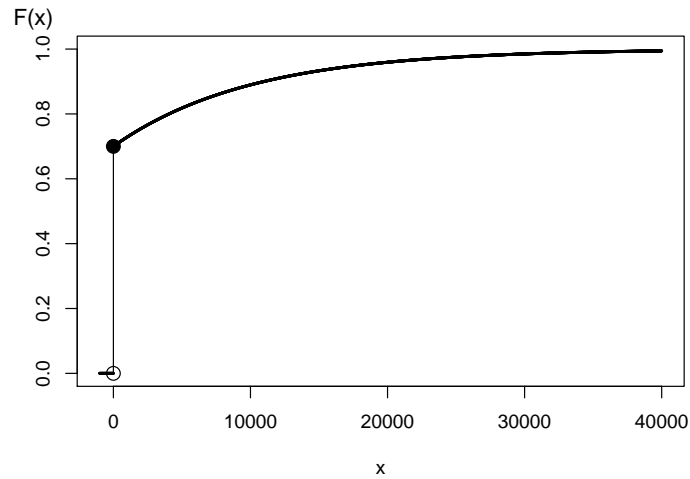$$\overline{h}_R = \frac{1}{R} \sum_{i=1}^{R} h(X_i)$$

F(x)

Figure 5.4: 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 6

Portfolio Management including Reinsurance

6.0.1 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

6.1 Tails of Distributions

- The **tail** of a distribution (more specifically: the **right tail**) is the portion of the distribution corresponding to large values of the r.v.

- Understanding large possible loss values is important because they have the greatest impact on the total of losses.

- R.v.'s that tend to assign higher probabilities to larger values are said to be **heavier tailed**.

- When choosing models, tail weight can help narrow choices or can confirm a choice for a model.

Classification Based on Moments

- One way of classifying distributions:

- are all moments finite, or not?

- The **finiteness** of **all positive moments** indicates a (relatively) light right tail.

- The **finiteness** of only positive moments **up to a certain value** indicates a heavy right tail.

- **KPW Example 3.9**: demonstrate that for the gamma distribution all positive moments are finite but for the Pareto distribution they are not.

- For the gamma distribution

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

- **KPW Example 3.9**: demonstrate that for the gamma distribution all positive moments exist but for the Pareto distribution they do not.

- For the Pareto distribution

$$\begin{aligned}
\mu_k' &= \int_0^\infty x^k \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}dx \\
&= \int_\theta^\infty (y-\theta)^k \frac{\alpha\theta^\alpha}{y^{\alpha+1}}dy \\
&= \alpha\cdot\theta^\alpha \int_\theta^\infty \sum_{j=0}^k \binom{k}{j} y^{j-\alpha-1}(-\theta)^{k-j}dy,
\end{aligned}$$

for integer values of k.

- This integral is finite only if $\int_\theta^\infty y^{j-\alpha-1}dy = \frac{y^{j-\alpha}}{j-\alpha}\big|_\theta^\infty$ is finite.

- Finiteness occurs when $j-\alpha < 0$ for $j=1,\ldots,k$. Or, equivalently, $k < \alpha$.

- Pareto is said to have a heavy tail, and gamma has a light tail.

Comparison Based on Limiting Tail Behavior

- Consider two distributions with the same mean.

- If ratio of $S_1(.)$ and $S_2(.)$ diverges to infinity, then distribution 1 has a heavier tail than distribution 2.

- Thus, we examine

$$\begin{aligned}
\lim_{x\to\infty} \frac{S_1(x)}{S_2(x)} &= \lim_{x\to\infty} \frac{S_1'(x)}{S_2'(x)} \\
&= \lim_{x\to\infty} \frac{-f_1(x)}{-f_2(x)} = \lim_{x\to\infty} \frac{f_1(x)}{f_2(x)}.
\end{aligned}$$

- **KPW Example 3.10**: demonstrate that Pareto distribution has a heavier tail than the gamma distribution using the limit of the ratio of their density functions.

- We consider

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

- Exponentials go to infinity faster than polynomials, thus the limit is infinity.

6.2 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)

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

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

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

  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

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

  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, TVaR$(X + Y) \leq$ TVaR$(X) +$ 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

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

see Appendix A for those expressions.

6.3   Reinsurance

- A major difference between reinsurance and primary insurance is that a reinsurance program is generally tailored more closely to the buyer

- There are two major types of reinsurance

  - Proportional

  - Excess of Loss

- A proportional treaty is an agreement between a reinsurer and a ceding company (the reinsured) in which the reinsurer assumes a given percent of losses and premium.

6.3.1   Proportional Reinsurance

- A proportional treaty is an agreement between a reinsurer and a ceding company (the reinsured) in which the reinsurer assumes a given percent of losses and premium.

- 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 defined as follows:

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

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

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

6.3.2 Surplus Share Proportional Treaty

- Another proportional treaty is known as *Surplus Share*; these are common in property business.

- 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 X is the loss, the reinsurer's portion is $\min(400000, (X - 100000)_+)$.

6.3.3 Excess of Loss Reinsurance

- Under this arrangement, the direct insurer sets a retention level $M(> 0)$ and pays in full any claim for which $X \leq M$.

- The direct insurer retains an amount M of the risk.

- For claims for which $X > M$, the direct insurer pays M and the reinsurer pays the remaining amount $X - M$.

- The amounts paid by the direct insurer and the reinsurer are defined as follows.

$$Y_{insurer} = cX \qquad Y_{reinsurer} = (1 - c)X$$

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

6.3.4 Relations with Personal Insurance

- We have already seen the needed tools to handle reinsurance in the context of personal insurance

  - For a proportional reinsurance, the transformation $Y_{insurer} = c X$ is the same as a coinsurance adjustment in personal insurance

  - For excess of loss reinsurance, the transformation $Y_{reinsurer} = \max(0, X - M)$ is the same as an insurer's payment with a deductible M and $Y_{insurer} = \min(X, M) = X \wedge M$ is equivalent to what a policyholder pays with deductible M.

- Reinsurance applications suggest introducing *layers of coverage*, a (small) mathematical extension.

6.3.5 Layers of Coverage

- 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.

- In general, we consider k parties. If $k = 4$, it could be an insurer and three different reinsurers.

- Consider a simple 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, 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

- Mathematically, partition the positive real line into k intervals using the cut-points $0 = c_0 < c_1 < \cdots < c_{k-1} < c_k = \infty$.

    - The jth interval is $(c_{j-1}, c_j]$.

- Let Y_j be the amount of risk shared by the jth party

- To illustrate, if a loss x is such that $c_{j-1} < x \leq c_j$, then

$$
\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_j \\ Y_{j+1} \\ \vdots \\ Y_k \end{pmatrix} = \begin{pmatrix} c_1 - c_0 \\ c_2 - c_1 \\ \vdots \\ x - c_{j-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}
$$

- More succinctly, we can write

$$Y_j = \min(X, c_j) - \min(X, c_{j-1})$$

- With the expression $Y_j = min(X, c_j) - min(X, c_{j-1})$, we see that the jth party is responsible for claims in the interval $(c_{j-1}, c_j]$

- Note that $X = Y_1 + Y_2 + \cdots + Y_k$

- The parties need not be different.

    - For 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 $c_1 = 100$, $c_2 = 100000$.

    - The policyholder is responsible for $Y_1 = \min(X, 100)$ and $Y_3 = X - \min(X, 100000) = \max(0, X - 100000)$.

- See 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

Chapter 7

Technical Supplement: Statistical Inference

7.1 Overview of Statistical Inference

- A set of data (a **sample**) has been collected that is considered representative of a larger set (the **population**). This relationship is known as the **sampling frame**.

- Often, we can describe the distribution of the population in terms of a limited (finite) number of terms called **parameters**. These are referred to as *parametric distributions*. With **nonparametric** analysis, we do not limit ourselves to only a few parameters.

- The **statistical inference** goal is to say something about the (larger) population based on the observed sample (we "*infer*," not "*deduce*"). There are three types of statements:

    1. **Estimation**

    2. **Hypothesis Testing**

    3. **Prediction**

Wisconsin Property Fund

- Discuss ideas of statistical inference in the context of a sample from the Wisconsin Property Fund

- Specifically, consider 1,377 *individual* claims from 2010 experience (slightly different from the analysis of 403 average claims in Chapter 1)

	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Standard Deviation
Claims	1	788	2,250	26,620	6,171	12,920,000	36
Logarithmic Claims	0	6.670	7.719	7.804	8.728	16.370	

```
ClaimLev <- read.csv("Data/CLAIMLEVEL.csv", header=TRUE)
ClaimLevBC10<-subset(ClaimLev,Year==2010); nrow(ClaimLevBC10)
```
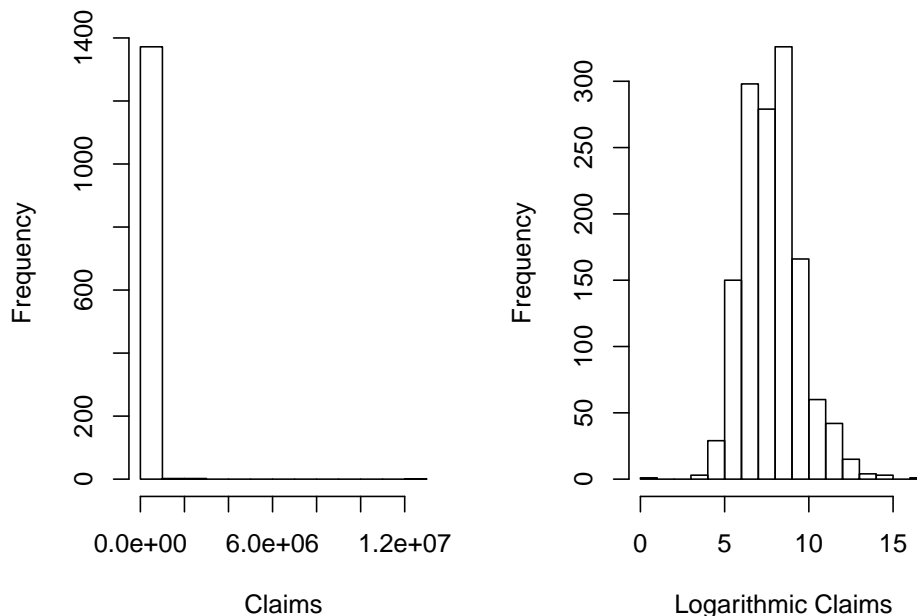
```
## [1] 1377
```

Figure 7.1: Distribution of Claims

```r
par(mfrow=c(1, 2))
hist(ClaimLevBC10$Claim, main="", xlab="Claims")
hist(log(ClaimLevBC10$Claim), main="", xlab="Logarithmic Claims")
```

Sampling Frame

- In statistics, a sampling frame **error** occurs when the sampling frame, the list from which the sample is drawn, is not an adequate approximation of the population of interest.

- For the property fund example, the sample consists of all 2010 claims

  - The population might be all claims that could have potentially occurred in 2010.

  - Or, it might be all claims that could potentially occur, such as in 2010, 2011, and so forth

- 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; you simply repeat the same mistake over again and again.

- A sample should be a representative subset of a population, or "universe," of interest.

- Formally

  - We assume that the random variable X represents a draw from a population with distribution function F(.)

  - We make several such draws (n), each unrelated to one another (statistically independent)

  - Sometimes we say that $X_1, \ldots, X_n$ is a random sample (with replacement) from F(.)

– Sometimes we say that $X_1, \ldots, X_n$ are identically and independently distributed (*iid*)

Describing the Population

- We think of the random variable X as a draw from the population with distribution function F(.)

- There are several ways to summarize F(.). We might consider the mean, standard deviation, 95th percentile, and so on.

    – Because these summary stats do not depend on a specific parametric reference, they are **nonparametric** summary measures.

- In contrast, we can think of logarithmic claims as normally distributed with mean μ and standard deviation σ, that is, claims have a *lognormal* distribution

- We will also look at the gamma distribution, with parameters α and θ, as a claims model

    – The normal, lognormal, and gamma are examples of **parametric** distributions.

    – The quantities μ, σ, α, and θ are known as *parameters*. When we know the parameters of a distribution family, then we have knowledge of the entire distribution.

7.2   Estimation and Prediction

Estimation

- Use θ to denote a summary of the population.

    – Parametric - It can be a parameter from a distribution such as μ or σ.

    – Nonparametric - It can also be a nonparametric summary such as the mean or standard deviation.

- Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be a function of the sample that provides proxy, or **estimate**, of θ. It is a function of the sample $X_1, \ldots, X_n$.

- In our property fund case,

    – 7.804 is a (nonparametric) estimate of the population expected logarithmic claim and 1.683 is an estimate of the corresponding standard deviation.

    – These are (parametric) estimates of the normal distribution for logarithmic claims

    – The estimate of the expected claim using the lognormal distribution is 10,106.8 (=exp(7.804 + $1.683^2/2$)).

Lognormal Distribution and Estimation

- Assume that claims follow a lognormal distribution, so that logarithmic claims follow the familiar normal distribution.

- Specifically, assume $\ln X$ has a normal distribution with mean μ and variance σ^2, sometimes denoted as $X \sim N(\mu, \sigma^2)$.

- For the property data, estimates are $\hat{\mu} = 7.804$ and $\hat{\sigma} = 1.683$. The "hat" notation is common. These are said to be **point estimates**, a single approximation of the corresponding parameter.

- Under general maximum likelihood theory (that we will do in a little bit), these estimates typically have a normal distribution for large samples.

  – Using notation, $\hat{\theta}$ has an approximate normal distribution with mean θ and variance, say, $\text{Var}(\hat{\theta})$.

  – Take the square root of the variance and plug-in the estimate to define $se(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$. A **standard error** is an estimated standard deviation.

  – The next step in the mathematical statistics theory is to establish that $(\hat{\theta} - \theta)/se(\hat{\theta})$ has a t-distribution with "degrees of freedom" (a parameter of the distribution) equal to the sample size minus the dimension of θ.

- Assume that claims follow a lognormal distribution, so that logarithmic claims follow the familiar normal distribution.

- Under general maximum likelihood theory

  – $\hat{\theta}$ has an approximate normal distribution with mean θ and variance, say, $\text{Var}(\hat{\theta})$.

  – Take the square root of the variance and plug-in the estimate to define $se(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$. A **standard error** is an estimated standard deviation.

  – $(\hat{\theta} - \theta)/se(\hat{\theta})$ has a t-distribution with "degrees of freedom" (a parameter of the distribution) equal to the sample size minus the dimension of θ.

  – As an application, we can invert this result to get a **confidence interval** for θ.

- A pair of statistics, $\hat{\theta}_1$ and $\hat{\theta}_2$, provide an interval of the form $[\hat{\theta}_1, \hat{\theta}_2]$ This interval is a $1 - \alpha$ confidence interval for θ if $\Pr\left(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\right) \geq 1 - \alpha$.

- For example, $\hat{\theta}_1 = \hat{\mu} - (t - value)\hat{\sigma}/\sqrt{n}$ and $\hat{\theta}_2 = \hat{\mu} + (t - value)\hat{\sigma}/\sqrt{n}$ provide a confidence interval for $\theta = \mu$. When $\alpha = 0.05$, $t - value \approx 1.96$.

- For the property fund, (7.715235, 7.893208) is a 95% confidence interval for μ.

Lognormal Distribution and Hypothesis Testing

An important statistical inference procedure involves verifying ideas about parameters.

- To illustrate, in the property fund, assume that mean logarithmic claims have historically been approximately been $\mu_0 = log(5000) = 8.517$. I might want to use 2010 data to see whether the mean of the distribution has changed. I also might want to test whether it has increased.

- The actual 2010 average was $\hat{\mu} = 7.804$. Is this a significant departure from $\mu_0 = 8.517$?

- One way to think about it is in terms of standard errors. The deviation is $(8.517 - 7.804)/(1.683/\sqrt{1377}) = 15.72$ standard errors. This is highly unlikely assuming an approximate normal distribution.

- One hypothesis testing procedure begin with the calculation the test statistic $t - stat = (\hat{\theta} - \theta_0)/se(\hat{\theta})$. Here, θ_0 is an assumed value of the parameter.

- Then, one rejects the hypothesized value if the test statistic $t - stat$ is "unusual." To gauge "unusual," use the same t-distribution as introduced for confidence intervals.

- If you only want to know about a difference, this is known as a "two-sided" test; use the same $t - value$ as the case for confidence intervals.

- If you want to investigate whether there has been an increase (or decrease), then use a "one-sided" test.

- Another useful concept in hypothesis testing is the p-value, which is short hand for probability value. For a data set, a p-value is defined to be the smallest significance level for which the null hypothesis would be rejected.

Property Fund – Other Distributions

- For numerical stability and extensions to regression applications, statistical packages often work with transformed version of parameters

- The following estimates are from the **R** package **VGAM** (the function)

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

7.3 Maximum Likelihood Theory

7.3.1 Likelihood Function

- Let $f(\cdot; \boldsymbol{\theta})$ be the probability mass function if X is discrete or the probability density function if it is continuous.

- The likelihood is a function of the parameters ($\boldsymbol{\theta}$) with the data ($\mathbf{x}$) fixed rather than a function of the data with the parameters fixed.

- Define the *log-likelihood function*,

$$L(\boldsymbol{\theta}) = L(\mathbf{x}; \boldsymbol{\theta}) = \ln f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i; \boldsymbol{\theta}),$$

evaluated at a realization $\mathbf{x}$.

- In the case of independence, the joint density function can be expressed as a product of the marginal density functions and, by taking logarithms, we can work with sums.

Example. Pareto Distribution

- Suppose that $X_1, \ldots, X_n$ represent a random sample from a single-parameter Pareto with cumulative distribution function:
$$F(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500.$$

- In this case, the single parameter is $\theta = \alpha$.

- The corresponding probability density function is $f(x) = 500^{\alpha} \alpha x^{-\alpha-1}$ and the logarithmic likelihood is
$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln f(x_i; \alpha) = n\alpha \ln 500 + n \ln \alpha - (\alpha + 1) \sum_{i=1}^{n} \ln x_i.$$

Properties of Likelihood Functions

- One basic property of likelihood functions is:

$$\mathrm{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}}L(\boldsymbol{\theta})\right)=\mathbf{0}$$

- The derivative of the log-likelihood function, $\partial L(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, is called the *score function*.

- To see this,

$$\mathrm{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}}L(\boldsymbol{\theta})\right)=\mathrm{E}\left(\frac{\frac{\partial}{\partial\boldsymbol{\theta}}\mathrm{f}(\mathbf{x};\boldsymbol{\theta})}{\mathrm{f}(\mathbf{x};\boldsymbol{\theta})}\right)=\int\frac{\partial}{\partial\boldsymbol{\theta}}\mathrm{f}(\mathbf{x};\boldsymbol{\theta})d\mathbf{y}$$

$$=\frac{\partial}{\partial\boldsymbol{\theta}}\int\mathrm{f}(\mathbf{x};\boldsymbol{\theta})d\mathbf{y}=\frac{\partial}{\partial\boldsymbol{\theta}}1=\mathbf{0}.$$

- Another basic property is:

$$\mathrm{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}L(\boldsymbol{\theta})\right)+\mathrm{E}\left(\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)=\mathbf{0}.$$

- With this, we can define the *information matrix*

$$\mathbf{I}(\boldsymbol{\theta})=\mathrm{E}\left(\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)=-\mathrm{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}L(\boldsymbol{\theta})\right).$$

- In general

$$\frac{\partial}{\partial\boldsymbol{\theta}}L(\boldsymbol{\theta})=\frac{\partial}{\partial\boldsymbol{\theta}}\ln\prod_{i=1}^{n}\mathrm{f}(x_i;\boldsymbol{\theta})=\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\ln\mathrm{f}(x_i;\boldsymbol{\theta}).$$

  has a large sample **normal distribution** with mean $\mathbf{0}$ and variance $\mathbf{I}(\boldsymbol{\theta})$.

Maximum Likelihood Estimators

- The value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_{MLE}$, that maximizes $\mathrm{f}(\mathbf{x};\boldsymbol{\theta})$ is called the *maximum likelihood estimator*.

- Maximum likelihood estimators are values of the parameters $\boldsymbol{\theta}$ that are "most likely" to have been produced by the data.

- Because $\ln(\cdot)$ is a one-to-one function, we can also determine $\boldsymbol{\theta}_{MLE}$ by maximizing the log-likelihood function, $L(\boldsymbol{\theta})$.

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:

$$\mathrm{F}(x)=1-\left(\frac{500}{x}\right)^{\alpha},\quad x>500.$$

Calculate the maximum likelihood estimate of the parameter α.

Instructor Notes

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:

$$\mathrm{F}(x)=1-\left(\frac{500}{x}\right)^{\alpha},\quad x>500.$$

Calculate the maximum likelihood estimate of the parameter α.

Solution. With $n = 5$, the logarithmic likelihood is

$$L(\alpha) = \sum_{i=1}^{5} \ln \mathrm{f}(x_i; \alpha) = 5\alpha \ln 500 + 5 \ln \alpha - (\alpha + 1) \sum_{i=1}^{5} \ln x_i.$$

Solving for the root of the score function yields

$$\frac{\partial}{\partial \alpha} L(\alpha) = 5 \ln 500 + 5/\alpha - \sum_{i=1}^{5} \ln x_i =_{set} 0 \Rightarrow \alpha_{MLE} = \frac{5}{\sum_{i=1}^{5} \ln x_i - 5 \ln 500} = 2.453.$$

Asymptotic Normality of Maximum Likelihood Estimators

- Under broad conditions, $\boldsymbol{\theta}_{MLE}$ has a large sample normal distribution with mean $\boldsymbol{\theta}$ and variance $(\mathbf{I}(\boldsymbol{\theta}))^{-1}$.

- $2\left(L(\boldsymbol{\theta}_{MLE}) - L(\boldsymbol{\theta})\right)$ has a chi-square distribution with degrees of freedom equal to the dimension of $\boldsymbol{\theta}$.

- These are critical results upon which much of estimation and hypothesis testing is based.

  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, $\hat{\theta}_1, \hat{\theta}_2$.

Instructor Notes

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, $\hat{\theta}_1, \hat{\theta}_2$.

Solution. The matrix of second derivatives is

$$\begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} L & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} L \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} L & \frac{\partial^2}{\partial \theta_1^2} L \end{pmatrix} = \begin{pmatrix} -5 & -3 \\ -3 & -2 \end{pmatrix}$$

Thus, the information matrix is:

$$\mathbf{I}(\theta_1, \theta_2) = -\mathrm{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} L(\boldsymbol{\theta})\right) = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$$

and

$$\mathbf{I}^{-1}(\theta_1, \theta_2) = \frac{1}{5(2) - 3(3)} \begin{pmatrix} 2 & -3 \\ -3 & 5 \end{pmatrix} = \begin{pmatrix} 2 & -3 \\ -3 & 5 \end{pmatrix}.$$

Maximum Likelihood Estimation (MLE)

- Why use maximum likelihood estimation?

    - General purpose tool - works in many situations (data can be censored, truncated, include covariates, time-dependent, and so forth)

    - It is "optimal," the best, in the sense that it has the smallest variance among the class of all unbiased estimators. (Caveat: for large sample sizes).

- A drawback: Generally, maximum likelihood estimators are computed iteratively, no closed-form solution.

    - For example, you may recall a "Newton-Raphson" iterative algorithm from calculus

    - Iterative algorithms require starting values. For some problems, the choice of a close starting value is critical.

MLE and Statistical Significance

One important type inference is to say whether a parameter estimate is "statistically significant"

- We learned earlier that $\boldsymbol{\theta}_{MLE}$ has a large sample normal distribution with mean $\boldsymbol{\theta}$ and variance $(\mathbf{I}(\boldsymbol{\theta}))^{-1}$.

- Look to the jth element of $\boldsymbol{\theta}_{MLE}$, say $\theta_{MLE,j}$.

- Define $se(\theta_{MLE,j})$, the standard error (estimated standard deviation) to be square root of the j diagonal element of $(\mathbf{I}(\boldsymbol{\theta})_{MLE})^{-1}$.

- To assess the hypothesis that θ_j is 0, we look at the rescaled estimate $t(\theta_{MLE,j}) = \theta_{MLE,j}/se(\theta_{MLE,j})$. It is said to be a t-statistic or t-ratio.

- Under this hypothesis, it has a t-distribution with degrees of freedom equal to the sample size minus the dimension of $\boldsymbol{\theta}_{MLE}$.

- For most actuarial applications, the t-distribution is very close to the (standard) normal distribution. Thus, sometimes this ratio is also known a z-statistic or "z-score."

Assessing Statistical Significance

- If the t-statistic $t(\theta_{MLE,j})$ exceeds a cut-off (in absolute value), then the jth variable is said to be "statistically significant."

    - For example, if we use a 5% significance level, then the cut-off is 1.96 using a normal distribution approximation.

    - More generally, using a $100\alpha\%$ significance level, then the cut-off is a $100(1-\alpha/2)\%$ quantile from a t-distribution using degrees of freedom equal to the sample size minus the dimension of $\boldsymbol{\theta}_{MLE}$.

- Another useful concept in hypothesis testing is the p-value, shorthand for probability value.

    - For a data set, a p-value is defined as the smallest significance level for which the null hypothesis would be rejected.

    - The p-value is a useful summary statistic for the data analyst to report because it allows the reader to understand the strength of the deviation from the null hypothesis.

MLE and Model Validation

Another important type inference is to select a model from two choices, where one choice is a subset of the other

- Suppose that we have a (large) model and determine the maximum likelihood estimator, $\boldsymbol{\theta}_{MLE}$.

- Now assume that p elements in $\boldsymbol{\theta}$ are equal to zero and determine the maximum likelihood estimator over the remaining set. Call this estimator $\boldsymbol{\theta}_{Reduced}$

- The statistic, $LRT = 2\left(L(\boldsymbol{\theta}_{MLE}) - L(\boldsymbol{\theta}_{Reduced})\right)$, is called the likelihood ratio (a difference of the logs is the log of the ratio. Hence, the term "ratio.")

- Under the hypothesis that the reduce model is correct, the likelihood ratio has a chi-square distribution with degrees of freedom equal to p, the number of variables set equal to zero.

- This allows us to judge which of the two models is correct. If the statistic LRT is large relative to the chi-square distribution, then we reject the simpler, reduced, model in favor of the larger one.

7.3.2 Information Criteria

- These statistics can be used when comparing several alternative models that are not necessarily nested. One picks the model that minimizes the criterion.

- *Akaike's Information Criterion*

$$AIC = -2 \times L(\boldsymbol{\theta}_{MLE}) + 2 \times (number\ of\ parameters)$$

   - The additional term $2 \times$ (number of parameters) is a penalty for the complexity of the model.

   - Other things equal, a more complex model means more parameters, resulting in a larger value of the criterion.

- *Bayesian Information Criterion*, defined as

$$BIC = -2 \times L(\boldsymbol{\theta}_{MLE}) + (number\ of\ parameters) \times \ln(number\ of\ observations)$$

   - This measure gives greater weight to the number of parameters.

   - Other things being equal, BIC will suggest a more parsimonious model than AIC.

Property Fund Information Criteria

- 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

Property Fund Fitted Distributions

- In this graph, black represents actual (smoothed) logarithmic claims
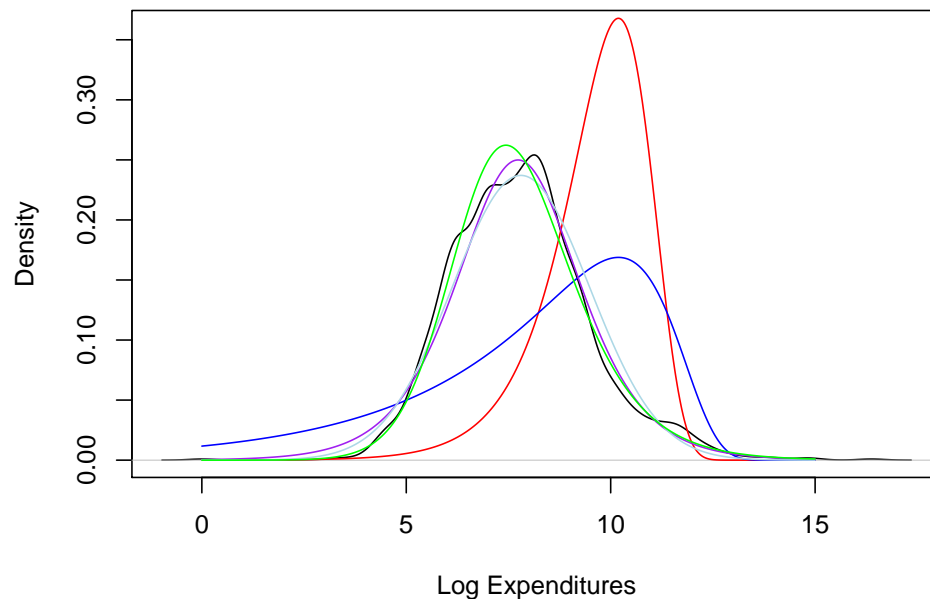
Figure 7.2: Fitted Claims Distribution

- 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)

[1] 6258

R Code for Fitted Claims Distributions

```
# R Code to fit several claims distributions
ClaimLev <- read.csv("Data/CLAIMLEVEL.csv", header=TRUE); nrow(ClaimLev)
ClaimData<-subset(ClaimLev,Year==2010);
#Use "VGAM" library for estimation of parameters
library(VGAM)
fit.LN <- vglm(Claim ~ 1, family=lognormal, data = ClaimData)
fit.gamma <- vglm(Claim ~ 1, family=gamma2, data = ClaimData)
  theta.gamma<-exp(coef(fit.gamma)[1])/exp(coef(fit.gamma)[2])
  alpha.gamma<-exp(coef(fit.gamma)[2])
fit.exp <- vglm(Claim ~ 1, exponential, data = ClaimData)
fit.pareto <- vglm(Claim ~ 1, paretoII, loc=0, data = ClaimData)


####################################################
#  Inference assuming a GB2 Distribution - this is more complicated
# The likelihood functon of GB2 distribution (negative for optimization)
likgb2 <- function(param) {
  a1 <- param[1]
  a2 <- param[2]
  mu <- param[3]
```

```
  sigma <- param[4]
  yt <- (log(ClaimData$Claim)-mu)/sigma
  logexpyt<-ifelse(yt>23,yt,log(1+exp(yt)))
  logdens <- a1*yt - log(sigma) - log(beta(a1,a2)) - (a1+a2)*logexpyt -log(ClaimData$Claim)
  return(-sum(logdens))
}
#  "optim" is a general purpose minimization function
gb2bop <- optim(c(1,1,0,1),likgb2,method=c("L-BFGS-B"),
                lower=c(0.01,0.01,-500,0.01),upper=c(500,500,500,500),hessian=TRUE)
####################################################
# Plotting the fit using densities (on a logarithmic scale)
plot(density(log(ClaimData$Claim)), ylim=c(0,0.36),main="", xlab="Log Expenditures")
x <- seq(0,15,by=0.01)
fexp_ex = dgamma(exp(x), scale = exp(-coef(fit.exp)), shape = 1)*exp(x)
lines(x,fexp_ex, col="red")
fgamma_ex = dgamma(exp(x), shape = alpha.gamma, scale=theta.gamma)*exp(x)
lines(x,fgamma_ex,col="blue")
fpareto_ex <- dparetoII(exp(x),loc=0,shape = exp(coef(fit.pareto)[2]), scale = exp(coef(fit.pareto)[1]))*ex
lines(x,fpareto_ex,col="purple")
flnorm_ex = dlnorm(exp(x), mean = coef(fit.LN)[1], sd = exp(coef(fit.LN)[2]))*exp(x)
lines(x,flnorm_ex, col="lightblue")
# density for GB II
gb2density <- function(x){
  a1 <- gb2bop$par[1]
  a2 <- gb2bop$par[2]
  mu <- gb2bop$par[3]
  sigma <- gb2bop$par[4]
  xt <- (log(x)-mu)/sigma
  logexpxt<-ifelse(xt>23,yt,log(1+exp(xt)))
  logdens <- a1*xt - log(sigma) - log(beta(a1,a2)) - (a1+a2)*logexpxt -log(x)
  exp(logdens)
}
fGB2_ex = gb2density(exp(x))*exp(x)
lines(x,fGB2_ex, col="green")
```

Bibliography

Bailey, R. A. and LeRoy, J. S. (1960). Two studies in automobile ratemaking. *Proceedings of the Casualty Actuarial Society Casualty Actuarial Society*, XLVII(I).

Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A., and Nesbitt, C. J. (1986). *Actuarial Mathematics*. Society of Actuaries Itasca, Ill.

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].

Gorman, M. and Swenson, S. (2013). Building believers: How to expand the use of predictive analytics in claims. SAS. [Retrieved on August 17, 2014].

Insurance Information Institute (2015). International insurance fact book. Insurance Information Institute. [Retrieved on May 10, 2016].

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica: journal of the Econometric Society*, pages 647–663.

Venter, G. (1983). Transformed beta and gamma distributions and aggregate losses. In *Proceedings of the Casualty Actuarial Society*, volume 70, pages 289–308.