1

Mixed Poisson Models for Claim Numbers

1.1 Introduction

1.1.1 Poisson Modelling for the Number of Claims

In view of the economic importance of motor third party liability insurance in industrialized countries, many attempts have been made in the actuarial literature to find a probabilistic model for the distribution of the number of claims reported by insured drivers. This chapter aims to introduce the basic probability models for count data that will be applied in motor insurance. References to alternative models are gathered in the closing section to this chapter.

The Binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability q. Such a success/failure experiment is also called a Bernoulli experiment or Bernoulli trial. Two important distributions arise as approximations of Binomial distributions. If n is large enough and the skewness of the distribution is not too great (that is, q is not too close to 0 or 1), then the Binomial distribution is well approximated by the Normal distribution. When the number of observations n is large, and the success probability q is small, the corresponding Binomial distribution is well approximated by the Poisson distribution with mean $\lambda = nq$. The Poisson distribution is thus sometimes called the law of small numbers because it is the probability distribution of the number of occurrences of an event that happens rarely but has very many opportunities to happen. The parallel with traffic accidents is obvious.

The Poisson distribution was discovered by Siméon-Denis Poisson (1781–1840) and published in 1838 in his work entitled *Recherches sur la Probabilité des Jugements en Matières Criminelles et Matière Civile* (which could be translated as 'Research on the Probability of Judgments in Criminal and Civil Matters'). Typically, a Poisson random

Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems M. Denuit, X. Maréchal, S. Pitrebois and J.-F. Walhin

variable is a count of the number of events that occur in a certain time interval or spatial area. For example, the number of cars passing a fixed point in a five-minute interval, or the number of claims reported to an insurance company by an insured driver in a given period. A typical characteristic associated with the Poisson distribution is certainly equidispersion: the variance of the Poisson distribution is equal to its mean.

1.1.2 Heterogeneity and Mixed Poisson Model

The Poisson distribution plays a prominent role in modelling discrete count data, mainly because of its descriptive adequacy as a model when only randomness is present and the underlying population is homogeneous. Unfortunately, this is not a realistic assumption to make in modelling many real insurance data sets. Poisson mixtures are well-known counterparts to the simple Poisson distribution for the description of inhomogeneous populations. Of special interest are populations consisting of a finite number of homogeneous sub-populations. In these cases the probability distribution of the population can be regarded as a finite mixture of Poisson distributions.

The problem of unobserved heterogeneity arises because differences in driving behaviour among individuals cannot be observed by the actuary. One of the well-known consequences of unobserved heterogeneity in count data analysis is overdispersion: the variance of the count variable is larger than the mean. Apart from its implications for the low-order moment structure of the counts, unobserved heterogeneity has important implications for the probability structure of the ensuing mixture model. The phenomena of excesses of zeros as well as heavy upper tails in most insurance data can be seen as an implication of unobserved heterogeneity (Shaked's Two Crossings Theorem will make this clear). It is customary to allow for unobserved heterogeneity by superposing a random variable (called a random effect) on the mean parameter of the Poisson distribution, yielding a mixed Poisson model. In a mixed Poisson process, the annual expected claim frequency itself becomes random.

1.1.3 Maximum Likelihood Estimation

All the models implemented in this book are parametric, in the sense that the probabilities are known functions depending on a finite number of (real-valued) parameters. The Binomial, Poisson and Normal models are examples of parametric distributions. The first step in the analysis is to select a reasonable parametric model for the observations, and then to estimate the underlying parameters. The maximum likelihood estimator is the value of the parameter (or parameter vector) that makes the observed data most likely to have occurred given the data generating process assumed to have produced the observations. All we need to derive the maximum likelihood estimator is to formulate statistical models in the form of a likelihood function as a probability of getting the data at hand. The larger the likelihood, the better the model.

Maximum likelihood estimates have several desirable asymptotic properties: consistency, efficiency, asymptotic Normality, and invariance. The advantages of maximum likelihood estimation are that it fully uses all the information about the parameters contained in the data and that it is highly flexible. Most applied maximum likelihood problems lack closed-form solutions and so rely on numerical maximization of the likelihood function. The advent of fast computers has made this a minor issue in most cases. Hypothesis testing for maximum

likelihood parameter estimates is straightforward due to the asymptotic Normal distribution of maximum likelihood estimates and the Wald and likelihood ratio tests.

1.1.4 Agenda

Section 1.2 briefly reviews the basic probability concepts used throughout this chapter (and the entire book), for further reference. Notions including probability spaces, random variables and probability distributions are made precise in this introductory section.

In Section 1.3, we recall the main probabilistic tools to work with discrete distributions: probability mass function, distribution function, probability generating function, etc. Then, we review some basic counting distributions, including the Binomial and Poisson laws.

Section 1.4 is devoted to mixture models to account for unobserved heterogeneity. Mixed Poisson distributions are discussed, including Negative Binomial (or Poisson-Gamma), Poisson-Inverse Gaussian and Poisson-LogNormal models.

Section 1.5 presents the maximum likelihood estimation method. Large sample properties of the maximum likelihood estimators are discussed, and testing procedures are described. The large sample properties are particularly appealing to actuaries who usually deal with tens of thousands of observations in insurance portfolios.

Section 1.6 gives numerical illustrations on the basis of a Belgian motor third party liability insurance portfolio. The observed claim frequency distribution is fitted using the Poisson distribution and various mixed Poisson probability distributions, and the optimal model is selected on the basis of appropriate goodness-of-fit tests.

The final Section, 1.7, concludes Chapter 1 by providing suggestions for further reading and bibliographic notes about the models proposed in the actuarial literature for the annual number of claims.

1.2 Probabilistic Tools

1.2.1 Experiment and Universe

Many everyday statements for actuaries take the form 'the probability of A is p', where A is some event (such as 'the total losses exceed the threshold € 1 000 000' or 'the number of claims reported by a given policyholder is less than two') and p is a real number between zero and one. The occurrence or nonoccurrence of A depends upon the chain of circumstances under consideration. Such a particular chain is called an experiment in probability; the result of an experiment is called its outcome and the set of all outcomes (called the universe) is denoted by Ω.

The word 'experiment' is used here in a very general sense to describe virtually any process for which all possible outcomes can be specified in advance and for which the actual outcome will be one of those specified. The basic feature of an experiment is that its outcome is not definitely known by the actuary beforehand.

1.2.2 Random Events

Random events are subsets of the universe Ω associated with a given experiment. A random event is the mathematical formalization of an event described in words. It is random since

we cannot predict with certainty whether it will be realized or not during the experiment. For instance, if we are interested in the number of claims incurred by a policyholder of an automobile portfolio during one year, the experiment consists in observing the driving behaviour of this individual during an annual period, and the universe Ω is simply the set $\{0, 1, 2, \dots\}$ of the nonnegative integers. The random event $A =$ 'the policyholder reports at most one claim' is identified with the subset $\{0, 1\} \subset \Omega$.

As usual, we use $A \cup B$ and $A \cap B$ to represent the union and the intersection, respectively, of any two subsets A and B of Ω. The union of sets is defined to be the set that contains the points that belong to at least one of the sets. The intersection of sets is defined to be the set that contains the points that are common to all the sets. These set operations correspond to the 'or' and 'and' between sentences: $A \cup B$ is the event which is realized if A or B is realized and $A \cap B$ is the event realized if A and B are simultaneously realized during the experiment. We also define the difference between sets A and B, denoted as $A \setminus B$, as the set of elements in A but not in B. Finally, $\overline{A}$ is the complementary event of A, defined as $\Omega \setminus A$; it is the set of points of Ω that do not belong to A. This corresponds to the negation: $\overline{A}$ is realized if A is not realized during the experiment. In particular, $\overline{\Omega} = \emptyset$, where $\emptyset$ is the empty set.

1.2.3 Sigma-Algebra

One needs to specify a family $\mathcal{F}$ of events to which probabilities can be ascribed in a consistent manner. The family $\mathcal{F}$ has to be closed under standard operations on sets; indeed, given two events A and B in $\mathcal{F}$, we want $A \cup B$, $A \cap B$ and $\overline{A}$ to still be events (i.e. still belong to $\mathcal{F}$). Technically speaking, this will be the case if $\mathcal{F}$ is a sigma-algebra. Recall that a family $\mathcal{F}$ of subsets of the universe Ω is called a sigma-algebra if it fulfills the three following properties: (i) $\Omega \in \mathcal{F}$, (ii) $A \in \mathcal{F} \Rightarrow \overline{A} \in \mathcal{F}$, and (iii) $A_1, A_2, A_3, \dots \in \mathcal{F} \Rightarrow \bigcup_{i \geq 1} A_i \in \mathcal{F}$.

The three properties (i)-(iii) are very natural. Indeed, (i) means that Ω itself is an event (it is the event which is always realized). Property (ii) means that if A is an event, the complement of A is also an event. Finally, property (iii) means that the event consisting in the realization of at least one of the A_is is also an event.

1.2.4 Probability Measure

Once the universe Ω has been equipped with a sigma-algebra $\mathcal{F}$ of random events, a probability measure Pr can be defined on $\mathcal{F}$. The knowledge of Pr allows us to discuss the likelihood of the occurrence of events in $\mathcal{F}$. To be specific, Pr assigns to each random event A its probability $\Pr[A]$; $\Pr[A]$ is the likelihood of realization of A. Formally, a probability measure Pr maps $\mathcal{F}$ to $[0, 1]$, with $\Pr[\Omega] = 1$, and is such that given $A_1, A_2, A_3, \dots \in \mathcal{F}$ which are pairwise disjoint, i.e., such that $A_i \cap A_j = \emptyset$ if $i \neq j$,

$$\Pr\left[\bigcup_{i \geq 1} A_i\right] = \sum_{i \geq 1} \Pr[A_i];$$

this technical property is usually referred to as the sigma-additivity of Pr.

The properties assigned to Pr naturally follow from empirical evidence: if we were allowed to repeat an experiment a large number of times, keeping the initial conditions as equal

as possible, the proportion of times that an event A occurs would behave according to the definition of Pr. Note that $\Pr[A]$ is then the mathematical idealization of the proportion of times A occurs.

1.2.5 Independent Events

Independence is a crucial concept in probability theory. It aims to formalize the intuitive notion of 'not influencing each other' for random events: we would like to give a precise meaning to the fact that the realization of an event does not decrease nor increase the probability that the other event occurs. Formally, two events A and B are said to be independent if the probability of their intersection equals the product of their respective probabilities, that is, if $\Pr[A \cap B] = \Pr[A]\Pr[B]$.

This definition is extended to more than two events as follows. The events in a family $\mathcal{A}$ of events are independent if for every finite sequence $A_1, A_2, \ldots, A_k$ of events in $\mathcal{A}$,

$$\Pr\left[\bigcap_{i=1}^{k} A_i\right] = \prod_{i=1}^{k} \Pr[A_i]. \tag{1.1}$$

The concept of independence is very important in assigning probabilities to events. For instance, if two or more events are regarded as being physically independent, in the sense that the occurrence or nonoccurrence of some of them has no influence on the occurrence or nonoccurrence of the others, then this condition is translated into mathematical terms through the assignment of probabilities satisfying Equation (1.1).

1.2.6 Conditional Probability

Independence is the exception rather than the rule. In any given experiment, it is often necessary to consider the probability of an event A when additional information about the outcome of the experiment has been obtained from the occurrence of some other event B. This corresponds to intuitive statements of the form 'if B occurs then the probability of A is p', where B can be 'March is rainy' and A 'the claim frequency in motor insurance increases by 5 %'. This is called the conditional probability of A given B, and is formally defined as follows. If $\Pr[B] > 0$ then the conditional probability $\Pr[A|B]$ of A given B is defined to be

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}. \tag{1.2}$$

The definition of conditional probabilities through (1.2) is in line with empirical evidence. Repeating a given experiment a large number of times, $\Pr[A|B]$ is the mathematical idealization of the proportion of times A occurs in those experiments where B did occur, hence the ratio (1.2).

It is easily seen that A and B are independent if, and only if,

$$\Pr[A|B] = \Pr[A|\overline{B}] = \Pr[A]. \tag{1.3}$$

Note that this interpretation of independence is much more intuitive than the definition given above: indeed the identity expresses the natural idea that the realization or not of B does not increase nor decrease the probability that A occurs.

1.2.7 Random Variables and Random Vectors

Often, actuaries are not interested in an experiment itself but rather in some consequences of its random outcome. For instance, they are more concerned with the amounts the insurance company will have to pay than with the particular circumstances which give rise to the claims. Such consequences, when real-valued, may be thought of as functions mapping Ω into the real line $\mathbb{R}$.

Such functions are called random variables provided they satisfy certain desirable properties, precisely stated in the following definition: A random variable X is a measurable function mapping Ω to the real numbers, i.e., $X : \Omega \to \mathbb{R}$ is such that $X^{-1}((-\infty, x]) \in \mathcal{F}$ for any $x \in \mathbb{R}$, where $X^{-1}((-\infty, x]) = \{\omega \in \Omega | X(\omega) \leq x\}$. In other words, the measurability condition $X^{-1}((-\infty, x]) \in \mathcal{F}$ ensures that the actuary can make statements like 'X is less than or equal to x' and quantify their likelihood. Random variables are mathematical formalizations of random outcomes given by numerical values. An example of a random variable is the amount of a claim associated with the occurrence of an automobile accident.

A random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ is a collection of n univariate random variables, $X_1, X_2, \ldots, X_n$, say, defined on the same probability space $(\Omega, \mathcal{F}, \Pr)$. Random vectors are denoted by bold capital letters.

1.2.8 Distribution Functions

In many cases, neither the universe Ω nor the function X need to be given explicitly. The practitioner has only to know the probability law governing X or, in other words, its distribution. This means that he is interested in the probabilities that X takes values in appropriate subsets of the real line (mainly intervals).

To each random variable X is associated a function F_X called the distribution function of X, describing the stochastic behaviour of X. Of course, F_X does not indicate what is the actual outcome of X, but shows how the possible values for X are distributed (hence its name). More precisely, the distribution function of the random variable X, denoted as F_X, is defined as

$$F_X(x) = \Pr[X^{-1}((-\infty, x])] \equiv \Pr[X \leq x], \quad x \in \mathbb{R}.$$

In other words, $F_X(x)$ represents the probability that the random variable X assumes a value that is less than or equal to x. If X is the total amount of claims generated by some policyholder, $F_X(x)$ is the probability that this policyholder produces a total claim amount of at most $€x$. The distribution function F_X corresponds to an estimated physical probability distribution or a well-chosen subjective probability distribution.

Any distribution function F has the following properties: (i) F is nondecreasing, i.e. $F(x) \leq F(y)$ if $x < y$, (ii) $\lim_{x \searrow -\infty} F(x) = 0$, (iii) $\lim_{x \nearrow +\infty} F(x) = 1$, (iv) F is right-continuous,

i.e. $\lim_{h \searrow 0} F(x+h) = F(x)$, and (v) $\Pr[a < X \leq b] = F(b) - F(a)$, for any $a < b$. Henceforth, we denote as $F(\cdot-)$ the left limit of F, that is,

$$F(x-) = \lim_{\xi \nearrow x} F(\xi) = \Pr[X < x].$$

Suppose that $X_1, X_2, \ldots, X_n$ are n random variables defined on the same probability space $(\Omega, \mathcal{F}, \Pr)$. Their marginal distribution functions $F_1, F_2, \ldots, F_n$ contain all the information about their associated probabilities. But how can the actuary encapsulate information about their properties relative to each other? As explained above, the key idea is to think of $X_1, X_2, \ldots, X_n$ as being components of a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ taking values in $\mathbb{R}^n$ rather than being unrelated random variables each taking values in $\mathbb{R}$.

As was the case for random variables, each random vector $\boldsymbol{X}$ possesses a distribution function $F_{\boldsymbol{X}}$ that describes its stochastic behaviour. The distribution function of the random vector $\boldsymbol{X}$, denoted as $F_{\boldsymbol{X}}$, is defined as

$$\begin{aligned} F_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n) &= \Pr\left[\boldsymbol{X}^{-1}((-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_n])\right] \\ &= \Pr[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n], \end{aligned}$$

$x_1, x_2, \ldots, x_n \in \mathbb{R}$. The value $F_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n)$ represents the probability that simultaneously X_1 assumes a value that is less than or equal to x_1, X_2 assumes a value that is less than or equal to x_2, $\ldots$, X_n assumes a value that is less than or equal to x_n; a more compact way to express this is

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \Pr[\boldsymbol{X} \leq \boldsymbol{x}], \quad \boldsymbol{x} \in \mathbb{R}^n.$$

Even if the distribution function $F_{\boldsymbol{X}}$ does not tell us which is the actual value of $\boldsymbol{X}$, it thoroughly describes the range of possible values for $\boldsymbol{X}$ and the probabilities assigned to each of them.

1.2.9 Independence for Random Variables

A fundamental concept in probability theory is the notion of independence. Roughly speaking, the random variables $X_1, X_2, \ldots, X_n$ are mutually independent when the behaviour of one of these random variables does not influence the others. Formally, the random variables $X_1, X_2, \ldots, X_n$ are mutually independent if, and only if, all the random events built with these random variables are independent. It can be shown that the random variables $X_1, X_2, \ldots, X_n$ are independent if, and only if,

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} F_{X_i}(x_i) \text{ holds for all } \boldsymbol{x} \in \mathbb{R}^n.$$

In other words, the joint distribution function of a random vector $\boldsymbol{X}$ with independent components is thus the product of the marginal distribution functions.

1.3 Poisson Distribution

1.3.1 Counting Random Variables

A discrete random variable X assumes only a finite (or countable) number of values. The most important subclass of nonnegative discrete random variables is the integer case, where each observation (outcome) is an integer (typically, the number of claims reported to the company). More precisely, a counting random variable N is valued in $\{0, 1, 2, \dots\}$. Its stochastic behaviour is characterized by the set of probabilities $\{p_k,\ k = 0, 1, \dots\}$ assigned to the nonnegative integers, where $p_k = \Pr[N = k]$. The (discrete) distribution of N associates with each possible integer value $k = 0, 1, 2, \dots$ the probability p_k that it will be the observed value. The distribution must satisfy the two conditions:

$$p_k \geq 0 \text{ for all } k \text{ and } \sum_{k=0}^{+\infty} p_k = 1,$$

i.e. the probabilities are all nonnegative real numbers lying between zero (impossibility) and unity (certainty), and their sum must be unity because it is certain that one or other of the values will be observed.

1.3.2 Probability Mass Function

In discrete distribution theory the p_ks are regarded as values of a mathematical function, i.e.

$$p_k = p(k|\boldsymbol{\xi}),\ k = 0, 1, 2, \dots, \tag{1.4}$$

where $p(\cdot|\boldsymbol{\xi})$ is a known function depending on a set of parameters $\boldsymbol{\xi}$. The function $p(\cdot|\boldsymbol{\xi})$ defined in (1.4) is usually called the probability mass function. Different functional forms lead to different discrete distributions. This is a parametric model.

The distribution function $F_N : \mathbb{R} \to [0, 1]$ of N gives for any real threshold x, the probability for N to be smaller than or equal to x. The distribution function F_N of N is related to the probability mass function via

$$F_N(x) = \sum_{k=0}^{\lfloor x \rfloor} p_k,\ x \in \mathbb{R}^+,$$

where p_k is given by Expression (1.4) and where $\lfloor x \rfloor$ denotes the largest integer n such that $n \leq x$ (it is thus the integer part of x). Considering (1.4), F_N also depends on $\boldsymbol{\xi}$.

1.3.3 Moments

There are various useful and important quantities associated with a probability distribution. They may be used to summarize features of the distribution. The most familiar and widely used are the moments, particularly the mean

$$\mathbb{E}[N] = \sum_{k=0}^{+\infty} k p_k,$$

which is given by the sum of the products of all the possible outcomes multiplied by their probability, and the variance

$$\mathbb{V}[N] = \mathbb{E}[(N - \mathbb{E}[N])^2] = \sum_{k=0}^{+\infty} \big(k - \mathbb{E}[N]\big)^2 p_k,$$

which is given by the sum of the products of the squared differences between all the outcomes and the mean, multiplied by their probability. Expanding the squared difference in the definition of the variance, it is easily seen that the variance can be reformulated as

$$\mathbb{V}[N] = \mathbb{E}\big[N^2 - 2N\mathbb{E}[N] + (\mathbb{E}[N])^2\big] = \mathbb{E}[N^2] - (\mathbb{E}[N])^2,$$

which provides a convenient way to compute the variance as the difference between the second moment $\mathbb{E}[N^2]$ and the square $(\mathbb{E}[N])^2$ of the first moment $\mathbb{E}[N]$. The mean and the variance are commonly denoted as μ and σ^2, respectively. Considering (1.4), both $\mathbb{E}[N]$ and $\mathbb{V}[N]$ are functions of $\boldsymbol{\xi}$, that is,

$$\mathbb{E}[N] = \mu(\boldsymbol{\xi}) \text{ and } \mathbb{V}[N] = \sigma^2(\boldsymbol{\xi}).$$

The mean is used as a measure of the location of the distribution: it is an average of the possible outcomes $0, 1, \ldots$ weighted by the corresponding probabilities $p_0, p_1, \ldots$ The variance is widely used as a measure of the spread of the distribution: it is a weighted average of the squared distances between the outcomes $0, 1, \ldots$ and the expected value $\mathbb{E}[N]$. Recall that $\mathbb{E}[\cdot]$ is a linear operator. From the properties of $\mathbb{E}[\cdot]$, it is easily seen that the variance $\mathbb{V}[\cdot]$ is shift-invariant and additive for independent random variables.

The degree of asymmetry of the distribution of a random variable N is measured by its skewness, denoted as $\gamma[N]$. The skewness is the third central moment of N, normalized by its variance raised to the power $3/2$ (in order to get a number without unit). Precisely, the skewness of N is given by

$$\gamma[N] = \frac{\mathbb{E}[(N - \mathbb{E}[N])^3]}{(\mathbb{V}[N])^{3/2}}.$$

For any random variable N with a symmetric distribution the skewness $\gamma[N]$ is zero. Positively skewed distributions tend to concentrate most of the probability mass on small values, but the remaining probability is stretched over a long range of larger values.

There are other related sets of constants, such as the cumulants, the factorial moments, the factorial cumulants, etc., which may be more convenient to use in some circumstances. For details about these constants, we refer the reader, e.g., to JOHNSON *ET AL.* (1992).

1.3.4 Probability Generating Function

In principle all the theoretical properties of the distribution can be derived from the probability mass function. There are, however, several other functions from which exactly the same information can be derived. This is because the functions are all one-to-one transformations

of each other, so each characterizes the distribution. One particularly useful function is the probability generating function, which is defined as

$$\varphi_N(z) = \mathbb{E}[z^N] = \sum_{k=0}^{+\infty} p_k z^k, \quad 0 < z < 1. \tag{1.5}$$

When p_k is given by (1.4), $\varphi_N(\cdot)$ depends on $\boldsymbol{\xi}$.

If any function that is known to be a probability generating function is expanded as a power series in z, then the coefficient of z^k must be p_k for the corresponding distribution. An alternative way of obtaining the probabilities is by repeated differentiation of φ_N with respect to z. Specifically,

$$\varphi_N(0) = \Pr[N = 0] \text{ and } \frac{d^k}{dt^k}\,\varphi_N(z)\Big|_{z=0} = k!\Pr[N = k] \text{ for } k = 1, 2, \ldots$$

1.3.5 Convolution Product

A key feature of probability generating functions is related to the computation of sums of independent discrete random variables. Considering two independent counting random variables N_1 and N_2, their sum is again a counting random variable and thus possesses a probability mass function as well as a probability generating function.

The probability mass function of $N_1 + N_2$ is obtained as follows: We obviously have that

$$\Pr[N_1 + N_2 = k] = \sum_{j=0}^{k} \Pr[N_1 = j, N_2 = k - j]$$

for any integer k. Since N_1 and N_2 are independent, their joint probability mass function factors to the product of the univariate probability mass functions. This simply comes from

$$\begin{aligned}
\Pr[N_1 = j, N_2 = k - j] &= \Pr[N_1 \le j, N_2 \le k - j] - \Pr[N_1 \le j, N_2 \le k - j - 1] \\
&\quad - \Pr[N_1 \le j - 1, N_2 \le k - j] + \Pr[N_1 \le j - 1, N_2 \le k - j - 1] \\
&= \Pr[N_1 \le j]\big(\Pr[N_2 \le k - j] - \Pr[N_2 \le k - j - 1]\big) \\
&\quad - \Pr[N_1 \le j - 1]\big(\Pr[N_2 \le k - j] - \Pr[N_2 \le k - j - 1]\big) \\
&= \Pr[N_2 = k - j]\big(\Pr[N_1 \le j] - \Pr[N_1 \le j - 1]\big) \\
&= \Pr[N_1 = j]\Pr[N_2 = k - j].
\end{aligned}$$

The probability mass function of $N_1 + N_2$ can thus be obtained from the discrete convolution formula

$$\Pr[N_1 + N_2 = k] = \sum_{j=0}^{k} \Pr[N_1 = j]\Pr[N_2 = k - j], \ k = 0, 1, \ldots$$

For large values of k, a direct application of the discrete convolution formula can be rather time-consuming.

The probability generating function of $N_1 + N_2$ is easily obtained from

$$\varphi_{N_1+N_2}(z) = \mathbb{E}[z^{N_1+N_2}] = \mathbb{E}[z^{N_1}]\mathbb{E}[z^{N_2}] = \varphi_{N_1}(z)\varphi_{N_2}(z)$$

since the mutual independence of N_1 and N_2 ensures that

$$\begin{aligned}\mathbb{E}[z^{N_1+N_2}] &= \sum_{k_1=0}^{+\infty}\sum_{k_2=0}^{+\infty} z^{k_1+k_2}\Pr[N_1 = k_1]\Pr[N_2 = k_2] \\ &= \sum_{k_1=0}^{+\infty} z^{k_1}\Pr[N_1 = k_1]\sum_{k_2=0}^{+\infty} z^{k_2}\Pr[N_2 = k_2] \\ &= \mathbb{E}[z^{N_1}]\mathbb{E}[z^{N_2}].\end{aligned}$$

Summing random variables thus corresponds to a convolution product for probability mass functions and to regular products for probability generating functions. An expansion of $\varphi_{N_1}\varphi_{N_2}(\cdot)$ as a series in powers of z then gives the probability mass function of $N_1 + N_2$, usually in a much easier way than computing the convolution product of the probability mass functions of N_1 and N_2.

1.3.6 From the Binomial to the Poisson Distribution

Bernoulli Distribution

The Bernoulli distribution is an extremely simple and basic distribution. It arises from what is known as a Bernoulli trial: a single observation is taken where the outcome is dichotomous, e.g., success or failure, alive or dead, male or female, 0 or 1. The probability of success is q. The probability of failure is $1 - q$.

If N is Bernoulli distributed with success probability q, which is denoted as $N \sim \mathcal{B}er(q)$, we have

$$p(k|q) = \begin{cases} 1-q \text{ if } k = 0 \\ q \text{ if } k = 1 \\ 0 \text{ otherwise.} \end{cases}$$

There is thus just one parameter: the success probability q. The mean is

$$\mathbb{E}[N] = 0 \times (1-q) + 1 \times q = q \tag{1.6}$$

and the variance is

$$\mathbb{V}[N] = \mathbb{E}[N^2] - q^2 = q - q^2 = q(1-q). \tag{1.7}$$

The probability generating function is

$$\varphi_N(z) = (1-q) \times z^0 + q \times z^1 = 1 - q + qz. \tag{1.8}$$

It is easily seen that $\varphi_N(0) = p(0|q)$ and $\varphi'_N(0|q) = p(1|q)$, as it should be.

Binomial Distribution

The Binomial distribution describes the outcome of a sequence of n independent Bernoulli trials, each with the same probability q of success. The probability that success is the outcome in exactly k of the trials is

$$p(k|n,q) = \binom{n}{k} q^k (1-q)^{n-k}, \quad k = 0, 1, \ldots, n, \tag{1.9}$$

and 0 otherwise. Formula (1.9) defines the Binomial distribution. There are now two parameters: the number of trials n (also called the exponent, or size) and the success probability q. Henceforth, we write $N \sim \mathcal{B}in(n,q)$ to indicate that N is Binomially distributed, with size n and success probability q.

Moments of the Binomial Distribution

The mean of $N \sim \mathcal{B}in(n,q)$ is

$$\begin{aligned}\mathbb{E}[N] &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} q^k (1-q)^{n-k} \\ &= nq \sum_{k=1}^{n} \Pr[M = k-1] = nq\end{aligned} \tag{1.10}$$

where $M \sim \mathcal{B}in(n-1,q)$. Furthermore, with M as defined before,

$$\begin{aligned}\mathbb{E}[N^2] &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} k q^k (1-q)^{n-k} \\ &= nq \sum_{k=1}^{n} k \Pr[M = k-1] \\ &= n(n-1)q^2 + nq\end{aligned}$$

so that the variance is

$$\mathbb{V}[N] = \mathbb{E}[N^2] - (nq)^2 = nq(1-q). \tag{1.11}$$

We immediately observe that the Binomial distribution is underdispersed, i.e. its variance is smaller than its mean : $\mathbb{V}[N] = nq(1-q) \leq \mathbb{E}[N] = nq$.

Probability Generating Function and Closure under Convolution for the Binomial Distribution

The probability generating function of $N \sim \mathcal{B}in(n,q)$ is

$$\varphi_N(z) = \sum_{k=0}^{n} \binom{n}{k} (qz)^k (1-q)^{n-k} = (1-q+qz)^n. \tag{1.12}$$

Note that Expression (1.12) is the Bernoulli probability generating function (1.8), raised to the nth power. This was expected since the Binomial random variable N can be seen as the

sum of n independent Bernoulli random variables with equal success probability q. This also explains why (1.10) is equal to n times (1.6) and why (1.11) is equal to n times (1.7) (in the latter case, since the variance is additive for independent random variables).

From (1.12), we also see that having independent random variables $N_1 \sim \mathcal{B}in(n_1, q)$ and $N_2 \sim \mathcal{B}in(n_2, q)$, the sum $N_1 + N_2$ is still Binomially distributed. This comes from the fact that the probability generating function of $N_1 + N_2$ is

$$\varphi_{N_1+N_2}(z) = \varphi_{N_1}(z)\varphi_{N_2}(z) = (1 - q + qz)^{n_1+n_2}$$

so that $N_1 + N_2 \sim \mathcal{B}in(n_1 + n_2, q)$. Note that this is not the case if the success probabilities differ.

Limiting Form of the Binomial Distribution

When n becomes large, (1.9) may be approximated by a Normal distribution according to the De Moivre–Laplace theorem. The approximation is much improved when a continuity correction is applied. The Poisson distribution can be obtained as a limiting case of the Binomial distribution when n tends to infinity together with q becoming very small. Specifically, let us assume that $N_n \sim \mathcal{B}in(n, \lambda/n)$ and let n tend to $+\infty$. The probability mass at 0 then becomes

$$\Pr[N_n = 0] = \left(1 - \frac{\lambda}{n}\right)^n \to \exp(-\lambda), \text{ as } n \to +\infty.$$

To get the probability masses on the positive integers, let us compute the ratio

$$\frac{\Pr[N_n = k+1]}{\Pr[N_n = k]} = \frac{\frac{n-k}{k+1}\frac{\lambda}{n}}{1 - \frac{\lambda}{n}} \to \frac{\lambda}{k+1}, \text{ as } n \to +\infty,$$

from which we conclude

$$\lim_{n\to+\infty} \Pr[N_n = k] = \exp(-\lambda)\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Poisson Distribution

The Poisson random variable takes its values in $\{0, 1, \ldots\}$ and has probability mass function

$$p(k|\lambda) = \exp(-\lambda)\frac{\lambda^k}{k!}, \; k = 0, 1, \ldots. \tag{1.13}$$

Having a counting random variable N, we denote as $N \sim \mathcal{P}oi(\lambda)$ the fact that N is Poisson distributed with parameter λ. The Poisson distribution occupies a central position in discrete distribution theory analogous to that occupied by the Normal distribution in continuous distribution theory. It also has many practical applications.

The Poisson distribution describes events that occur randomly and independently in space or time. A classic example in physics is the number of radioactive particles recorded by a Geiger counter in a fixed time interval. This property of the Poisson distribution means that it can act as a reference standard when deviations from pure randomness are suspected. Although the Poisson distribution is often called the law of small numbers, there is no need

for $\lambda = nq$ to be small. It is the largeness of n and the smallness of $q = \lambda/n$ that are important. However most of the data sets analysed in the literature show a small frequency. This will be the case with motor data sets in insurance applications.

Moments of the Poisson Distribution

If $N \sim \mathcal{P}oi(\lambda)$, then its expected value is given by

$$\begin{aligned}\mathbb{E}[N] &= \sum_{k=1}^{+\infty} k \exp(-\lambda)\frac{\lambda^k}{k!} \\ &= \exp(-\lambda)\sum_{k=0}^{+\infty}\frac{\lambda^{k+1}}{k!} = \lambda. \end{aligned} \tag{1.14}$$

Moreover,

$$\begin{aligned}\mathbb{E}[N^2] &= \sum_{k=1}^{+\infty} k^2 \exp(-\lambda)\frac{\lambda^k}{k!} \\ &= \exp(-\lambda)\sum_{k=0}^{+\infty}(k+1)\frac{\lambda^{k+1}}{k!} = \lambda + \lambda^2,\end{aligned}$$

so that the variance of N is equal to

$$\mathbb{V}[N] = \mathbb{E}[N^2] - \lambda^2 = \lambda. \tag{1.15}$$

Considering Expressions (1.14) and (1.15), we see that both the mean and the variance of the Poisson distribution are equal to λ, a phenomenon termed as equidispersion.

The skewness of $N \sim \mathcal{P}oi(\lambda)$ is

$$\gamma[N] = \frac{1}{\sqrt{\lambda}}. \tag{1.16}$$

Clearly, $\gamma[N]$ decreases with λ. For small values of λ the distribution is very skewed (asymmetric) but as λ increases it becomes less skewed and is nearly symmetric by $\lambda = 15$.

Probability Generating Function and Closure Under Convolution for the Poisson Distribution

The probability generating function of the Poisson distribution has a very simple form. Coming back to the Equation (1.5) defining φ_N and replacing the p_ks with their expression (1.13) gives

$$\varphi_N(z) = \sum_{k=0}^{+\infty}\exp(-\lambda)\frac{(\lambda z)^k}{k!} = \exp\big(\lambda(z-1)\big). \tag{1.17}$$

This shows that the Poisson distribution is closed under convolution. Having independent random variables $N_1 \sim \mathcal{P}oi(\lambda_1)$ and $N_2 \sim \mathcal{P}oi(\lambda_2)$, the probability generating function of the sum $N_1 + N_2$ is

$$\varphi_{N_1+N_2}(z) = \varphi_{N_1}(z)\varphi_{N_2}(z) = \exp\big(\lambda_1(z-1)\big)\exp\big(\lambda_2(z-1)\big) = \exp\big((\lambda_1+\lambda_2)(z-1)\big)$$

so that $N_1 + N_2 \sim \mathcal{P}oi(\lambda_1 + \lambda_2)$.

The sum of two independent Poisson distributed random variables is also Poisson distributed, with parameter equal to the sum of the Poisson parameters. This property obviously extends to any number of terms, and the Poisson distribution is said to be closed under convolution (i.e. the convolution of Poisson distributions is still Poisson).

1.3.7 Poisson Process

Definition

Recall that a stochastic process is a collection of random variables $\{N(t), t \in \mathcal{T}\}$ indexed by a real-valued parameter t taking values in the index set $\mathcal{T}$. Usually, $\mathcal{T}$ represents a set of observation times. In this book, we will be interested in continuous-time stochastic processes where $\mathcal{T} = \mathbb{R}^+$.

A stochastic process $\{N(t), t \geq 0\}$ is said to be a counting process if $t \mapsto N(t)$ is right-continuous and $N(t) - N(t-)$ is 0 or 1. Intuitively speaking, $N(t)$ represents the total number of events that have occurred up to time t. Such a process enjoys the following properties: (i) $N(t) \geq 0$, (ii) $N(t)$ is integer valued, (iii) if $s < t$, then $N(s) \leq N(t)$, and (iv) for $s < t$, $N(t) - N(s)$ equals the number of events that have occurred in the interval $(s, t]$. By convention, $N(0) = 0$. The counting process $\{N(t), \ t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if

(i) the process has stationary increments, that is,

$$\Pr[N(t+\Delta) - N(t) = k] = \Pr[N(s+\Delta) - N(s) = k]$$

for any integer k, instants $s \leq t$ and increment $\Delta > 0$.

(ii) the process has independent increments, that is, for any integer $k > 0$ and instants $0 \leq t_0 < t_1 < t_2 < \cdots < t_k$, the random variables $N(t_1) - N(t_0)$, $N(t_2) - N(t_1), \ldots,$ $N(t_k) - N(t_{k-1})$ are mutually independent.

(iii) and

$$\Pr[N(h) = k] = \begin{cases} 1 - \lambda h + o(h) \text{ if } k = 0 \\ \lambda h + o(h) \text{ if } k = 1 \\ o(h) \text{ if } k \geq 2 \end{cases}$$

where $o(h)$ is a function of h that tends to 0 faster than the identity, that is, $\lim_{h \searrow 0} o(h)/h = 0$. Intuitively speaking, $o(h)$ is negligible when h becomes sufficiently small.

Assumption (i) implies that the probability of causing an accident is assumed to be the same for every day during any given period (we thus neglect the fact that meteorological conditions, safety conditions and other factors could vary over time). According to assumption (ii), the occurrence of an accident at one point in time is independent of all accidents that might have occurred before: reporting one accident does not increase nor decrease the probability of causing an accident in the future. This supports the fact that traffic accidents occur randomly in time. Assumption (iii) indicates that the probability that the policyholder files two or more

claims in a sufficiently small time interval is negligible when compared to the probability that he reports zero or only one claim.

Link with the Poisson Distribution

The Poisson process is intimately linked to the Poisson distribution, as precisely stated in the next result.

Property 1.1 *For any Poisson process, the number of events in any interval of length t is Poisson distributed with mean λt, that is, for all $s, t \geq 0$,*

$$\Pr[N(t+s) - N(s) = n] = \exp(-\lambda t)\frac{(\lambda t)^n}{n!}, \quad n = 0, 1, 2, \ldots$$

Proof Without loss of generality, we only have to prove that $N(t) \sim \mathcal{P}oi(\lambda t)$. For any integer k, let us denote $p_k(t) = \Pr[N(t) = k]$, $t \geq 0$. The announced result for $k = 0$ comes from

$$\begin{aligned} p_0(t+\Delta t) &= \Pr[N(t) = 0 \text{ and } N(t+\Delta t) - N(t) = 0] \\ &= \Pr[N(t) = 0]\Pr[N(t+\Delta t) - N(t) = 0] \\ &= p_0(t)p_0(\Delta t) \\ &= p_0(t)\big(1 - \lambda\Delta t + o(\Delta t)\big), \end{aligned}$$

where the joint probability factors into two terms since the increments of a Poisson process are independent random variables. This gives

$$\frac{p_0(t+\Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t) + \frac{o(\Delta t)}{\Delta t} p_0(t).$$

Taking the limit for $\Delta t \searrow 0$ yields

$$\frac{d}{dt} p_0(t) = -\lambda p_0(t).$$

This differential equation with the initial condition $p_0(0) = 1$ admits the solution

$$p_0(t) = \exp(-\lambda t), \tag{1.18}$$

which is in fact the $\mathcal{P}oi(\lambda t)$ probability mass function evaluated at the origin.

For $k \geq 1$, let us write

$$\begin{aligned} p_k(t+\Delta t) &= \Pr[N(t+\Delta t) = k] \\ &= \Pr[N(t+\Delta t) = k | N(t) = k]\Pr[N(t) = k] \\ &\quad + \Pr[N(t+\Delta t) = k | N(t) = k-1]\Pr[N(t) = k-1] \\ &\quad + \sum_{j=2}^{k} \Pr[N(t+\Delta t) = k | N(t) = k-j]\Pr[N(t) = k-j] \end{aligned}$$

$$
\begin{aligned}
&= \Pr[N(t+\Delta t) - N(t) = 0] \Pr[N(t) = k] \\
&\quad + \Pr[N(t+\Delta t) - N(t) = 1] \Pr[N(t) = k-1] \\
&\quad + \sum_{j=2}^{k} \Pr[N(t+\Delta t) - N(t) = j] \Pr[N(t) = k-j].
\end{aligned}
$$

Since the increments of a Poisson process are independent random variables, we can write

$$
\begin{aligned}
p_k(t+\Delta t) &= p_0(\Delta t) p_k(t) + p_1(\Delta t) p_{k-1}(t) + \sum_{j=2}^{k} p_j(\Delta t) p_{k-j}(t) \\
&= (1 - \lambda \Delta t) p_k(t) + \lambda \Delta t p_{k-1}(t) + o(\Delta t).
\end{aligned}
$$

This gives

$$
\frac{p_k(t+\Delta t) - p_k(t)}{\Delta t} = \lambda (p_{k-1}(t) - p_k(t)) + \frac{o(\Delta t)}{\Delta t}.
$$

Taking the limit for $\Delta t \searrow 0$ yields as above

$$
\frac{d}{dt} p_k(t) = \lambda (p_{k-1}(t) - p_k(t)), \quad k \geq 1. \tag{1.19}
$$

Multiplying by z^k each of the equation (1.19), and summing over k gives

$$
\sum_{k=0}^{+\infty} \left(\frac{d}{dt} p_k(t) \right) z^k = \lambda z \sum_{k=0}^{+\infty} p_k(t) z^k - \lambda \sum_{k=0}^{+\infty} p_k(t) z^k. \tag{1.20}
$$

Denoting as φ_t the probability generating function of $N(t)$, equation (1.20) becomes

$$
\frac{\partial}{\partial t} \varphi_t(z) = \lambda (z-1) \varphi_t(z). \tag{1.21}
$$

With the condition $\varphi_0(z) = 1$, Equation (1.21) has solution

$$
\varphi_t(z) = \exp(\lambda t (z-1)),
$$

where we recognize the $\mathcal{P}oi(\lambda t)$ probability generating function (1.17). This ends the proof. □

When the hypotheses behind a Poisson process are verified, the number $N(1)$ of claims hitting a policy during a period of length 1 is Poisson distributed with parameter λ. So, a counting process $\{N(t), \ t \geq 0\}$, starting from $N(0) = 0$, is a Poisson process with rate $\lambda > 0$ if

(i) The process has independent increments
(ii) The number of events in any interval of length t follows a Poisson distribution with mean λt (therefore it has stationary increments), i.e.

$$\Pr[N(t+s)-N(s)=k]=\exp(-\lambda t)\frac{(\lambda t)^k}{k!}, k=0,1,2,\ldots$$

Exposure-to-Risk

The Poisson process setting is useful when one wants to analyse policyholders that have been observed during periods of unequal lengths. Assume that the claims occur according to a Poisson process with rate λ. If the policyholder is covered by the company for a period of length d then the number N of claims reported to the company has probability mass function

$$\Pr[N=k]=\exp(-\lambda d)\frac{(\lambda d)^k}{k!}, \quad k=0,1,\ldots,$$

that is, $N\sim\mathcal{P}oi(\lambda d)$. In actuarial studies, d is referred to as the exposure-to-risk. We see that d simply multiplies the annual expected claim frequency λ in the Poisson model.

Time Between Accidents

The Poisson distribution arises for events occurring randomly and independently in time. Indeed, denote as $T_1, T_2, \ldots$ the times between two consecutive accidents. Assume further that these accidents occur according to a Poisson process with rate λ. Then, the T_ks are independent and identically distributed and

$$\Pr[T_k>t]=\Pr[T_1>t]=\Pr[N_t=0]=\exp(-\lambda t)$$

so that $T_1, T_2, \ldots$ have a common Negative Exponential distribution.

Note that in this case, the equality

$$\Pr[T_k>s+t|T_k>s]=\frac{\Pr[T_k>s+t]}{\Pr[T_k>s]}=\Pr[T_k>t]$$

holds for any s and $t\geq 0$. It is not difficult to see that this memoryless property is related to the fact that the increments of the process $\{N(t),\ t\geq 0\}$ are independent and stationary. Assuming that the claims occur according to a Poisson process is thus equivalent to assuming that the time between two consecutive claims has a Negative Exponential distribution.

Nonhomogeneous Poisson Process

A generalization of the Poisson process is obtained by letting the rate of the process vary with time. We then replace the constant rate λ by a function $t\mapsto\lambda(t)$ of time t and we define the nonhomogeneous Poisson process with rate $\lambda(\cdot)$. The Poisson process defined above (with a constant rate) is then termed as the homogeneous Poisson process. A counting process $\{N(t),\ t\geq 0\}$ starting from $N_0=0$ is said to be a nonhomogeneous Poisson process with rate $\lambda(\cdot)$, where $\lambda(t)\geq 0$ for all $t\geq 0$, if it satisfies the following conditions:

(i) The process $\{N(t),\ t \geq 0\}$ has independent increments, and
(ii)

$$\Pr[N(t+h) - N(t) = k] = \begin{cases} 1 - \lambda(t)h + o(h) \text{ if } k = 0 \\ \lambda(t)h + o(h) \text{ if } k = 1 \\ o(h) \text{ if } k \geq 2. \end{cases}$$

The only difference between the nonhomogeneous Poisson process and the homogeneous Poisson process is that the rate may vary with time, resulting in the loss of the stationary increment property.

For any nonhomogeneous Poisson process $\{N(t), t \geq 0\}$, the number of events in the interval $(s, t]$, $s \leq t$, is Poisson distributed with mean

$$m(s, t) = \int_s^t \lambda(u)du,$$

that is

$$\Pr[N(t) - N(s) = k] = \exp\big(-m(s, t)\big)\frac{\big(m(s, t)\big)^k}{k!}, k = 0, 1, \ldots$$

In the homogeneous case, we obviously have $m(s, t) = (t - s)\lambda$.

1.4 Mixed Poisson Distributions

1.4.1 Expectations of General Random Variables

Mixed Poisson distributions involve expectations of Poisson probabilities with a random parameter. Therefore, we need to be able to compute expectations with respect to general distribution functions.

Continuous probability distributions are widely used in probability and statistics when the underlying random phenomenon is measured on a continuous scale. If the distribution function is a continuous function, the associated probability distribution is called a continuous distribution. Note that in this case,

$$\Pr[X = x] = \lim_{h \searrow 0} \Pr[x < X \leq x + h] = \lim_{h \searrow 0} F(x + h) - F(x) = 0$$

for every real x. If X has a continuous probability distribution, then $\Pr[X = x] = 0$ for any real x.

In this book, we often consider distribution functions possessing a derivative, $f(x) = dF(x)/dx$. The function f is called the probability density function. Then one can integrate the probability density function to recover the distribution function, that is,

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

One can calculate the probability of an event by integrating the probability density function; for example, if the event is an interval $[a, b]$, then

$$\Pr[a \leq X \leq b] = F(b) - F(a) = \int_a^b f(x)dx.$$

The interpretation of the probability density function is that

$$\Pr[x \le X \le x+h] \approx f(x)h \text{ for small } h > 0.$$

That is, the probability that a random variable, with an absolutely continuous probability distribution, takes a value in a small interval of length h is given by the probability density function times the length of the interval.

A general type of distribution function is a combination of the discrete and (absolutely) continuous cases, being continuous apart from a countable set of exception points $x_1, x_2, x_3, \ldots$ with positive probabilities of occurrence, causing jumps in the distribution function at these points. Such a distribution function F_X can be represented as

$$F_X(x) = (1-p)F_X^{(c)}(x) + pF_X^{(d)}(x),\ x \in \mathbb{R}, \tag{1.22}$$

for some $p \in [0,1]$, where $F_X^{(c)}$ is a continuous distribution function and $F_X^{(d)}$ is a discrete distribution function with support $\{d_1, d_2, \ldots\}$.

Let us assume that F_X is of the form (1.22) with

$$pF_X^{(d)}(t) = \sum_{d_n \le t} \Big(F_X(d_n) - F_X(d_n-)\Big) = \sum_{d_n \le t} \Pr[X = d_n],$$

where $\{d_1, d_2, \ldots\}$ denotes the set of discontinuity points and

$$(1-p)F_X^{(c)}(t) = F_X(t) - pF_X^{(d)}(t) = \int_{-\infty}^{t} f_X^{(c)}(x)dx.$$

Then,

$$\begin{aligned}\mathbb{E}[X] &= \sum_{n\ge 1} d_n\Big(F_X(d_n) - F_X(d_n-)\Big) + \int_{-\infty}^{+\infty} x f_X^{(c)}(x)dx \\ &= \int_{-\infty}^{+\infty} x dF_X(x),\end{aligned} \tag{1.23}$$

where the differential of F_X, denoted as dF_X, is defined as

$$dF_X(x) = \begin{cases} F_X(d_n) - F_X(d_n-), & \text{if } x = d_n, \\ f_X^{(c)}(x)dx, & \text{otherwise.}\end{cases}$$

This unified notation allows us to avoid tedious repetitions of statements like 'the proof is given for continuous random variables; the discrete case is similar'. A very readable introduction to differentials and Riemann–Stieltjes integrals can be found in Carter & Van Brunt (2000).

1.4.2 Heterogeneity and Mixture Models

Definition

Mixture models are a discrete or continuous weighted combination of distributions aimed at representing a heterogeneous population comprised of several (two or more) distinct sub-populations. Such models are typically used when a heterogeneous population of sampling

units consists of several sub-populations within each of which a relatively simpler model applies. The source of heterogeneity could be gender, age, geographical area, etc.

Discrete Mixtures

In order to define a mixture model mathematically, suppose the distribution of N can be represented by a probability mass function of the form

$$\Pr[N = k] = p(k|\boldsymbol{\psi}) = q_1 p_1(k|\xi_1) + \cdots + q_\nu p_\nu(k|\xi_\nu) \tag{1.24}$$

where $\boldsymbol{\psi} = (\boldsymbol{q}^T, \boldsymbol{\xi}^T)^T$, $\boldsymbol{q}^T = (q_1, \ldots, q_\nu)$, $\boldsymbol{\xi}^T = (\xi_1, \ldots, \xi_\nu)$. The model is usually referred to as a discrete (or finite) mixture model. Here ξ_j is a (vector) parameter characterizing the probability mass function $p_j(\cdot|\xi_j)$ and the q_js are mixing weights.

Example 1.1 A particular example of finite mixture is the zero-inflated distribution. It has been observed empirically that counting distributions often show excess of zeros against the Poisson distribution. In order to accommodate this feature, a combination of the original distribution $\{p_k,\ k = 0, 1, \ldots\}$ (be it Poisson or not) together with the degenerate distribution with all probability concentrated at the origin, gives a finite mixture with

$$\begin{aligned}\Pr[N = 0] &= \omega + (1 - \omega)p_0 \\ \Pr[N = k] &= (1 - \omega)p_k,\ k = 1, 2, \ldots\end{aligned}$$

A mixture of this kind is usually referred to as zero-inflated, zero-modified or as a distribution with added zeros.

Model (1.24) allows each component probability mass function to belong to a different parametric family. In most applications, a common parametric family is assumed and thus the mixture model takes the following form

$$p(k|\boldsymbol{\psi}) = q_1 p(k|\xi_1) + \cdots + q_\nu p(k|\xi_\nu) \tag{1.25}$$

which we assume to hold in the sequel. The mixing weight $\boldsymbol{q}$ can be regarded as a discrete probability function over $\boldsymbol{\xi}$, describing the variation in the choice of $\boldsymbol{\xi}$ across the population of interest.

This class of mixture models includes mixtures of Poisson distributions. Such a mixture is adequate to model count data (number of claims reported to an insurance company, number of accidents caused by an insured driver, etc.) where the components of the mixture are Poisson distributions with mean ξ_j. In that respect, (1.25) means that there are ν categories of policyholders, with annual expected claim frequencies $\xi_1, \xi_2, \ldots, \xi_\nu$, respectively. The proportion of the portfolio in the different categories is $q_1, q_2, \ldots, q_\nu$, respectively. Considering a given policyholder, the actuary does not know to which category he belongs, but the probability that he comes from category j is q_j. The probability mass function of the number of claims reported by this insured driver is thus a weighted average of the probability mass functions associated with the k categories.

Continuous Mixtures

Multiplying the number ν of categories in (1.25) often leads to a dramatic increase in the number of parameters (the q_js and the ξ_js). For large ν, it is therefore preferable to switch to a continuous mixture, where the sum in (1.25) is replaced with an integral with respect to some simple parametric continuous probability density function.

Specifically, if we allow ξ to be continuous with probability density function $g(\cdot)$, the finite mixture model suggested above is replaced by the probability mass function

$$p(k) = \int p(k|\xi)g(\xi)d\xi,$$

which is often referred to as a mixture distribution. When $g(\cdot)$ is modelled without parametric assumptions, the probability mass function $p(\cdot)$ is a semiparametric mixture model. Often in actuarial science, $g(\cdot)$ is taken from some parametric family, so that the resulting probability mass function is also parametric.

Mixed Poisson Model for the Number of Claims

The Poisson distribution often poorly fits observations made in a portfolio of policyholders. This is in fact due to the heterogeneity that is present in the portfolio: driving abilities vary from individual to individual. Therefore it is natural to multiply the mean frequency λ of the Poisson distribution by a positive random effect Θ. The frequency will vary within the portfolio according to the nonobservable random variable Θ. Obviously we will choose Θ such that $\mathbb{E}[\Theta] = 1$ because we want to obtain, on average, the frequency of the portfolio. Conditional on Θ, we then have

$$\Pr[N = k|\Theta = \theta] = p(k|\lambda\theta) = \exp(-\lambda\theta)\frac{(\lambda\theta)^k}{k!}, \ k = 0, 1, \ldots, \tag{1.26}$$

where $p(\cdot|\lambda\theta)$ is the Poisson probability mass function, with mean $\lambda\theta$. The interpretation we give to this model is that not all policyholders in the portfolio have an identical frequency λ. Some of them have a higher frequency ($\lambda\theta$ with $\theta \geq 1$), others have a lower frequency ($\lambda\theta$ with $\theta \leq 1$). Thus we use a random effect to model this empirical observation.

The annual number of accidents caused by a randomly selected policyholder of the portfolio is then distributed according to a mixed Poisson law. In this case, the probability that a randomly selected policyholder reports k claims to the company is obtained by averaging the conditional probabilities (1.26) with respect to Θ. In general, Θ is not discrete nor continuous but of mixed type. The probability mass function associated with mixed Poisson models is defined as

$$\Pr[N = k] = \mathbb{E}\big[p(k|\lambda\Theta)\big] = \int_0^\infty \exp(-\lambda\theta)\frac{(\lambda\theta)^k}{k!}dF_\Theta(\theta) \tag{1.27}$$

where F_Θ denotes the distribution function of Θ, assumed to fulfill $F_\Theta(0) = 0$. The mixing distribution described by F_Θ represents the heterogeneity of the portfolio of interest; dF_Θ is often called the structure function. It is worth mentioning that the mixed Poisson model (1.27) is an accident-proneness model: it assumes that a policyholder's mean claim frequency does not change over time but allows some insured persons to have higher mean claim frequencies than others. We will say that N is mixed Poisson distributed with parameter λ and risk level Θ, denoted as $N \sim \mathcal{MP}oi(\lambda, \Theta)$ when it has probability mass function (1.27).

Remark 1.1 Note that a better notation would have been $\mathcal{MP}oi(\lambda, F_\Theta)$ instead of $\mathcal{MP}oi(\lambda, \Theta)$ since only the distribution function of Θ matters to define the associated Poisson mixture. We have nevertheless opted for $\mathcal{MP}oi(\lambda, \Theta)$ for simplicity.

Note that the condition $\mathbb{E}[\Theta] = 1$ ensures that when $N \sim \mathcal{MP}oi(\lambda, \Theta)$

$$\begin{aligned}\mathbb{E}[N] &= \int_0^\infty \sum_{k=0}^{+\infty} k \exp(-\lambda\theta) \frac{(\lambda\theta)^k}{k!} dF_\Theta(\theta) \\ &= \lambda \mathbb{E}[\Theta] = \lambda,\end{aligned}$$

or, more briefly,

$$\mathbb{E}[N] = \mathbb{E}\Big[\mathbb{E}[N|\Theta]\Big] = \mathbb{E}[\lambda\Theta] = \lambda. \tag{1.28}$$

In (1.28), $\mathbb{E}[\cdot|\Theta]$ means that we take an expected value considering Θ as a constant. We then average with respect to all the random components, except Θ. Consequently, $\mathbb{E}[\cdot|\Theta]$ is a function of Θ. Given Θ, N is Poisson distributed with mean $\lambda\Theta$ so that $\mathbb{E}[N|\Theta] = \lambda\Theta$. The mean of N is finally obtained by averaging $\mathbb{E}[N|\Theta]$ with respect to Θ. The expectation of N given in (1.28) is thus the same as the expectation of a $\mathcal{P}oi(\lambda)$ distributed random variable. Taking the heterogeneity into account by switching from the $\mathcal{P}oi(\lambda)$ to the $\mathcal{MP}oi(\lambda, \Theta)$ distribution has no effect on the expected claim number.

1.4.3 Mixed Poisson Process

The Poisson processes are suitable models for many real counting phenomena but they are insufficient in some cases because of the deterministic character of their intensity function. The doubly stochastic Poisson process (or Cox process) is a generalization of the Poisson process when the rate of occurrence is influenced by an external process such that the rate becomes a random process. So, the rate, instead of being constant (homogeneous Poisson process) or a deterministic function of time (nonhomogeneous Poisson process) becomes itself a stochastic process. The only restriction on the rate process is that it has to be nonnegative. Mixed Poisson distributions are linked to mixed Poisson processes in the same way that the Poisson distribution is associated with the standard Poisson process.

Specifically, let us assume that given $\Theta = \theta$, $\{N(t), t \geq 0\}$ is a homogeneous Poisson process with rate $\lambda\theta$. Then $\{N(t), \ t \geq 0\}$ is a mixed Poisson process, and for any $s, t \geq 0$, the probability that k events occur during the time interval $(s, t]$ is

$$\begin{aligned}\Pr[N(t+s) - N(s) = k] &= \int_0^\infty \Pr[N(t+s) - N(s) = k | \Theta = \theta] dF_\Theta(\theta) \\ &= \int_0^\infty \exp(-\lambda\theta t) \frac{(\lambda\theta t)^k}{k!} dF_\Theta(\theta),\end{aligned}$$

that is, $N(t+s) - N(s) \sim \mathcal{MP}oi(\lambda t, \Theta)$. Note that, in contrast to the Poisson process, mixed Poisson processes have dependent increments. Hence, past number of claims reveal future number of claims in this setting (in contrast to the Poisson case).

1.4.4 Properties of Mixed Poisson Distributions

Moments and Overdispersion

If $N \sim \mathcal{MP}oi(\lambda, \Theta)$ then its second moment is

$$\mathbb{E}[N^2] = \int_0^{+\infty} (\lambda\theta + \lambda^2\theta^2) dF_\Theta(\theta) = \lambda\mathbb{E}[\Theta] + \lambda^2\mathbb{E}[\Theta^2]$$

so that

$$\begin{aligned}\mathbb{V}[N] &= \lambda\mathbb{E}[\Theta] + \lambda^2\mathbb{E}[\Theta^2] - \lambda^2\big(\mathbb{E}[\Theta]\big)^2 \\ &= \lambda + \lambda^2\mathbb{V}[\Theta]. \end{aligned} \tag{1.29}$$

It is then easily seen that the variance of N exceeds its mean, that is,

$$\mathbb{V}[N] = \lambda + \lambda^2\mathbb{V}[\Theta] \geq \lambda = \mathbb{E}[N]. \tag{1.30}$$

Therefore, unless Θ is degenerated in 1, we observe that mixed Poisson distributions are overdispersed: the variance exceeds the mean. The skewness can be expressed as

$$\gamma[N] = \frac{1}{(\mathbb{V}[N])^{3/2}} \left(3\mathbb{V}[N] - 2\mathbb{E}[N] + \frac{\gamma[\Theta]}{\sqrt{\mathbb{V}[\Theta]}} \frac{(\mathbb{V}[N] - \mathbb{E}[N])^2}{\mathbb{E}[N]} \right). \tag{1.31}$$

Shaked's Two Crossings Theorem

Recall that $\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X])$ for any random variable X and convex function ϕ. This inequality, known as the Jensen inequality, ensures that if $N \sim \mathcal{MP}oi(\lambda, \Theta)$ then

$$\Pr[N = 0] = \int_0^{\infty} \exp(-\lambda\theta) dF_\Theta(\theta) \geq \exp\left(-\int_0^{\infty} \lambda\theta dF_\Theta(\theta) \right) = \exp(-\lambda),$$

showing that mixed Poisson distributions have an excess of zeros compared to Poisson distributions with the same mean. This is in line with empirical studies, where actuaries often observe more policyholders producing 0 claims than the number predicted by the Poisson model.

The following result that has been proved by Shaked (1980) reinforces this straightforward conclusion.

Property 1.2 *Let N be mixed Poisson distributed with mean $\mathbb{E}[N] = \lambda$. Then there exist two integers $0 \leq k_0 < k_1$ such that*

$$\begin{aligned} \Pr[N = k] &\geq \exp(-\lambda)\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots, k_0, \\ \Pr[N = k] &\leq \exp(-\lambda)\frac{\lambda^k}{k!}, \quad k = k_0 + 1, \ldots, k_1, \\ \Pr[N = k] &\geq \exp(-\lambda)\frac{\lambda^k}{k!}, \quad k \geq k_1 + 1. \end{aligned}$$

Shaked's Two Crossings Theorem tells us (i) that the mixed Poisson distribution has an excess of zeros compared to the Poisson distribution with the same mean and (ii) that the mixed Poisson distribution has a thicker right tail than the Poisson distribution with the same mean.

Probability Generating Function

The probability generating function of Poisson mixtures is closely related to the moment generating function of the underlying random effect. Moment generating functions are a widely used tool in many statistics texts, and also in actuarial mathematics. They serve to prove statements about convolutions of distributions, and also about limits. Recall that the moment generating function of the nonnegative random variable X, denoted as M_X, is given by

$$M_X(t) = \mathbb{E}[\exp(tX)], \quad t > 0.$$

It is interesting to mention that M_X characterizes the probability distribution of X, i.e. the information contained in F_X and M_X is equivalent.

If there exists $h > 0$ such that $M_X(t)$ exists and is finite for $0 < t < h$ then the Taylor expansion of the exponential function yields

$$M_X(t) = 1 + \sum_{n=1}^{+\infty} \frac{t^n}{n!} \mathbb{E}[X^n] \text{ for } 0 < t < h. \tag{1.32}$$

It is well known that if any moment of a distribution is infinite, the moment generating function does not exist. However, it is conceivable that there might exist distributions with moments of all orders and, yet, the moment generating function does not exist in any neighbourhood around 0. In fact, the LogNormal distribution is one such example.

Just as the probability generating function was interesting for analyzing sums of independent counting random variables, the moment generating function is a powerful tool to deal with sums of independent continuous random variables. Specifically, if X_1 and X_2 are independent random variables with respective moment generating functions $M_1(\cdot)$ and $M_2(\cdot)$, then the sum $X_1 + X_2$ has a moment generating function that is just the product $(M_1 M_2)(\cdot)$ of $M_1(\cdot)$ and $M_2(\cdot)$.

Now, consider $N \sim \mathcal{MP}oi(\lambda, \Theta)$. We have

$$\begin{aligned}
\varphi_N(z) &= \int_0^{+\infty} \exp(-\lambda\theta) \sum_{k=0}^{+\infty} \frac{(z\lambda\theta)^k}{k!} dF_\Theta(\theta) \\
&= \int_0^{+\infty} \exp\big(\lambda\theta(z-1)\big) dF_\Theta(\theta) \\
&= M_\Theta\big(\lambda(z-1)\big),
\end{aligned}$$

or, equivalently,

$$M_\Theta(t) = \varphi_N\left(1 + \frac{t}{\lambda}\right). \tag{1.33}$$

From (1.33), we see that the knowledge of the mixed Poisson distribution $\mathcal{MP}oi(\lambda, \Theta)$ is equivalent to the knowledge of F_Θ. The mixed Poisson distributions are thus identifiable, that is, having $N_1 \sim \mathcal{MP}oi(\lambda, \Theta_1)$ and $N_2 \sim \mathcal{MP}oi(\lambda, \Theta_2)$ then N_1 and N_2 are identically distributed if, and only if, Θ_1 and Θ_2 are identically distributed.

1.4.5 Negative Binomial Distribution

Gamma Distribution

Recall that a random variable X is distributed according to the two-parameter Gamma distribution, which will henceforth be denoted as $X \sim \mathcal{G}am(\alpha, \beta)$, if its probability density function is given by

$$f(x) = \frac{x^{\alpha-1}\beta^\alpha \exp(-\beta x)}{\Gamma(\alpha)}, \quad x > 0. \tag{1.34}$$

Note that when $\alpha = 1$, the Gamma distribution reduces to the Negative Exponential one (which is denoted as $X \sim \mathcal{E}xp(\beta)$) with probability density function

$$f(x) = \beta \exp(-\beta x), \quad x > 0.$$

The distribution function F of X can be expressed in terms of the incomplete Gamma function. Specifically, if $X \sim \mathcal{G}am(\alpha, \beta)$, then $F(x) = \Gamma(\alpha, \beta x)$.

Probability Mass Function

The Negative Binomial distribution is a widely used alternative to the Poisson distribution for handling count data when the variance is appreciably greater than the mean (this condition, known as overdispersion, is frequently met in practice, as discussed above).

There are several models that lead to the Negative Binomial distribution. A classic example arises from the theory of accident proneness which was developed after Greenwood & Yule (1920). This theory assumes that the number of accidents suffered by an individual is Poisson distributed, but that the Poisson mean (interpreted as the individual's accident proneness) varies between individuals in the population under study. If the Poisson mean is assumed to be Gamma distributed, then the Negative Binomial is the resultant overall distribution of accidents per individual.

Specifically, completing (1.26)–(1.27) with $\Theta \sim \mathcal{G}am(a, a)$, that is, with probability density function

$$f_\Theta(\theta) = \frac{1}{\Gamma(a)} a^a \theta^{a-1} \exp(-a\theta), \quad \theta > 0, \tag{1.35}$$

yields the Negative Binomial probability mass function

$$\begin{aligned}\Pr[N = k] &= \frac{(a+k-1)\cdots a}{k!}\left(\frac{a}{a+\lambda d}\right)^a \left(\frac{\lambda d}{a+\lambda d}\right)^k \\ &= \frac{\Gamma(a+k)}{\Gamma(a)k!}\left(\frac{a}{a+\lambda d}\right)^a \left(\frac{\lambda d}{a+\lambda d}\right)^k, \quad k = 0, 1, 2, \ldots\end{aligned}$$

where λ is the annual expected claim number and d is the length of the observation period (the exposure-to-risk). The probability mass function can be expressed using the generalized binomial coefficient:

$$\begin{aligned}\Pr[N=k] &= \frac{\Gamma(a+k)}{\Gamma(a)\Gamma(k+1)}\left(\frac{a}{a+\lambda d}\right)^a\left(\frac{\lambda d}{a+\lambda d}\right)^k \\ &= \binom{a+k-1}{k}\left(\frac{a}{a+\lambda d}\right)^a\left(\frac{\lambda d}{a+\lambda d}\right)^k,\ k=0,1,2,\ldots\end{aligned}$$

Henceforth, we write $N \sim \mathcal{NB}in(a, \lambda d)$ to indicate that N obeys the Negative Binomial distribution with parameters a and λd. This model has been applied to retail purchasing, absenteeism, doctor's consultations, amongst many others.

Moments

If $X \sim \mathcal{G}am(\alpha, \beta)$, its mean is $\mathbb{E}[X] = \alpha/\beta$ and its variance is $\mathbb{V}[X] = \alpha/\beta^2$. If $N \sim \mathcal{NB}in(a, \lambda d)$ then the mean is $\mathbb{E}[N] = \lambda d$ and the variance is $\mathbb{V}[N] = \lambda d + (\lambda d)^2/a$ according to (1.29). It can be shown that $\gamma[\Theta]/\sqrt{\mathbb{V}[\Theta]} = 2$, in (1.31) for the Negative Binomial distribution.

Probability Generating Function

If $X \sim \mathcal{G}am(\alpha, \beta)$, its moment generating function is

$$M(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha} \text{ if } t < \beta. \tag{1.36}$$

The probability generating function of $N \sim \mathcal{NB}in(a, \lambda d)$ is

$$\varphi_N(z) = \left(\frac{a}{a - \lambda d(z-1)}\right)^a. \tag{1.37}$$

This result comes from (1.33) together with (1.36).

True and Apparent Contagion

Apparent contagion arises from the recognition that sampled individuals come from a heterogeneous population in which individuals have a constant but different propensity to experience accidents. A given individual may have a high (or low) propensity for accidents but occurrence of an accident does not make it more (or less) likely that another accident will occur. However, aggregation across heterogeneous individuals may generate a misleading statistical finding which suggests that occurrence of an accident increases the probability of another accident; the observed but persistent heterogeneity can be misinterpreted as due to a strong serial dependence.

True contagion refers to dependence between the occurrences of successive events. The occurrence of an event, such as an accident or illness, may change the probability of subsequent occurrences of similar events. True positive contagion implies that the occurrence of an event shortens the expected waiting time to the next occurrence of the event.

The alleged phenomenon of accident proneness can be interpreted in terms of true contagion as suggesting that an individual who has experienced an accident is more likely

to experience another accident. In a longitudinal setting, actual and future outcomes are directly influenced by past values, and this causes a substantial change over time in the corresponding distribution.

Since with event count data we only observe the total number of events at the end of the period, contagion, like heterogeneity, is an unobserved, within-observation process. For research problems where both heterogeneity and contagion are plausible, the different underlying processes are not distinguishable with aggregate count data because they both lead to the same probability distribution for the counts. One can still use this distribution to derive fully efficient and consistent estimates, but this analysis will only be suggestive of the underlying process.

Poisson Limiting Form

The Negative Binomial distribution has a Poisson limiting form if $\mathbb{V}[\Theta] = \frac{1}{a} \to 0$. This result can be recovered from the sequence of the probability generating functions, noting that

$$\lim_{a \nearrow \infty} \left(\frac{a}{a - \lambda d(1-z)} \right)^a = \lim_{a \nearrow \infty} \left(1 - \frac{\lambda d}{a}(1-z) \right)^{-a} = \exp(-\lambda d(1-z))$$

that is seen to converge to the probability generating function of the Poisson distribution with parameter λd.

Derivation as a Compound Poisson Distribution

A different type of heterogeneity occurs when there is clustering. If it is assumed that the number of clusters is Poisson distributed, but the number of individuals in a cluster is distributed according to the Logarithmic distribution, then the overall distribution is Negative Binomial. In an actuarial context, this amounts to recognizing that several vehicles can be involved in the same accident, each of the insured drivers filing a claim. Therefore, a single accident may generate several claims. If the number of claims per accident follows a Logarithmic distribution, and the number of accidents over the time interval of interest follows a Poisson distribution, then the total number of claims for the time interval can be modelled with the Negative Binomial distribution.

Let us formally establish this result. Recall that the random variable M has a Logarithmic distribution if

$$\Pr[M = k] = \frac{\theta^k}{-k \ln(1-\theta)}, \quad k = 1, 2, \ldots$$

where $0 < \theta < 1$. The probability generating function of M is given by

$$\varphi_M(z) = \frac{\ln(1-\theta z)}{\ln(1-\theta)}.$$

Now, let $M_1, M_2, \ldots$ be a sequence of independent Logarithmically distributed random variables with the same parameter θ, and let $K \sim \mathcal{P}oi(\mu)$. Define

$$N = M_1 + \cdots + M_K.$$

The random variable N just defined has a compound Poisson distribution. The probability generating function of a compound distribution is given by

$$\begin{aligned}\varphi_N(z) &= \mathbb{E}[z^{M_1+\cdots+M_K}] \\ &= \sum_{k=0}^{+\infty} \Pr[K=k]\mathbb{E}[z^{M_1+\cdots+M_k}] \\ &= \sum_{k=0}^{+\infty} \Pr[K=k]\big(\varphi_M(z)\big)^k \\ &= \varphi_K\big(\varphi_M(z)\big). \end{aligned} \tag{1.38}$$

Note that formula (1.38) is true in general for compound distributions. Replacing φ_K and φ_M with their expressions gives the probability generating function of N

$$\begin{aligned}\varphi_N(z) &= \exp\Big(-\mu(1-\varphi_M(z))\Big) \\ &= \left(\frac{1-\theta}{1-\theta z}\right)^{-\mu/\ln(1-\theta)}. \end{aligned}$$

It can be checked that the probability generating function φ_N corresponds to the probability generating function (1.37) of a Negative Binomial distribution with $d=1$, $a=-\mu/\ln(1-\theta)$ and $\lambda=-\theta\mu/((1-\theta)\ln(1-\theta))$.

1.4.6 Poisson-Inverse Gaussian Distribution

There is no reason to restrict ourselves to the Gamma distribution for Θ, except perhaps mathematical convenience. In fact, any distribution with support in the half positive real line is a candidate to model the stochastic behaviour of Θ. Here, we discuss the Inverse Gaussian distribution.

Inverse Gaussian Distribution

The Inverse Gaussian distribution is an ideal candidate for modelling positive, right-skewed data. Recall that a random variable X is distributed according to the Inverse Gaussian distribution, which will be henceforth denoted as $X \sim \mathcal{IG}au(\mu,\beta)$, if its probability density function is given by

$$f(x) = \frac{\mu}{\sqrt{2\pi\beta x^3}} \exp\left(-\frac{1}{2\beta x}(x-\mu)^2\right), \quad x>0. \tag{1.39}$$

If $X \sim \mathcal{IG}au(\mu,\beta)$ then the mean is $\mathbb{E}[X]=\mu$ and the variance is $\mathbb{V}[X]=\mu\beta$. The moment generating function is given by

$$M(t) = \int_0^{+\infty} \frac{\mu}{\sqrt{2\pi\beta x^3}} \exp\left(-\frac{1}{2\beta x}(x-\mu)^2 + tx\right) dx$$

$$= \exp\left(\frac{\mu}{\beta}\right) \int_0^{+\infty} \frac{\mu}{\sqrt{2\pi\beta x^3}} \exp\left(-\frac{1}{2\beta x}\left(x^2(1-2\beta t)+\mu^2\right)\right) dx.$$

Making the change of variable $\xi = x\sqrt{1-2\beta t}$ yields

$$M(t) = \exp\left(\frac{\mu}{\beta}\right) \int_0^{+\infty} \frac{\mu}{\sqrt{2\pi \frac{\beta}{\sqrt{1-2\beta t}} \xi^3}} \exp\left(-\frac{1}{2\frac{\beta}{\sqrt{1-2\beta t}}\xi}(\xi^2+\mu^2)\right) d\xi$$

$$= \exp\left(\frac{\mu}{\beta}\left(1-\sqrt{1-2\beta t}\right)\right). \tag{1.40}$$

For the last three decades, the Inverse Gaussian distribution has gained attention in describing and analyzing right-skewed data. The main appeal of Inverse Gaussian models lies in the fact that they can accommodate a variety of shapes, from highly skewed to almost Normal. Moreover, they share many elegant and convenient properties with Gaussian models. In applied probability, the Inverse Gaussian distribution arises as the distribution of the first passage time to an absorbing barrier located at a unit distance from the origin in a Wiener process.

Poisson-Inverse Gaussian Distribution

Let us now complete (1.26)–(1.27) with $\Theta \sim \mathcal{IG}au(1,\tau)$, that is,

$$f_\Theta(\theta) = \frac{1}{\sqrt{2\pi\tau\theta^3}} \exp\left(-\frac{1}{2\tau\theta}(\theta-1)^2\right), \quad \theta > 0. \tag{1.41}$$

The probability mass function is given by

$$\Pr[N=k] = \int_0^{\infty} \exp(-\lambda d\theta) \frac{(\lambda d\theta)^k}{k!} \frac{1}{\sqrt{2\pi\tau\theta^3}} \exp\left(-\frac{1}{2\tau\theta}(\theta-1)^2\right) d\theta. \tag{1.42}$$

The probability mass function can be expressed using modified Bessel functions of the second kind. Bessel functions have some useful properties that can be used to compute the Poisson-Inverse Gaussian probabilities and to find the maximum likelihood estimators, for instance.

Moments and Probability Generating Function

Considering (1.28) and (1.29), we have

$$\mathbb{E}[N] = \lambda \text{ and } \mathbb{V}[N] = \lambda + \lambda^2\tau.$$

It can be shown that $\gamma[\Theta]/\sqrt{\mathbb{V}[\Theta]} = 3$ in (1.31) for the Poisson-Inverse Gaussian distribution. Therefore the skewness of a Poisson-Inverse Gaussian distribution exceeds the skewness of the Negative Binomial distribution having the same mean and the same variance.

Setting $\mu = 1$ and $\beta = \tau$, the probability generating function of N can be obtained from (1.33) together with (1.40), which gives

$$\varphi_N(z) = \exp\left(\frac{1}{\tau}\left(1-\sqrt{1-2\tau\lambda(z-1)}\right)\right).$$

Computation of the Probability Mass Function
The probability mass at the origin is

$$\varphi_N(0) = \Pr[N=0] = \exp\left(\frac{1}{\tau}\left(1-\sqrt{1+2\tau\lambda}\right)\right).$$

Now, taking the derivatives of φ_N with respect to t, and evaluating it at 0 gives the probability mass function for positive integers. Specifically,

$$\begin{aligned}\varphi_N'(0) &= \Pr[N=1]\\ &= \frac{\lambda}{\sqrt{1-2\tau\lambda(z-1)}}\varphi_N(z)\Big|_{z=0}\\ &= \frac{\lambda}{\sqrt{1+2\tau\lambda}}\Pr[N=0]\end{aligned}$$

and

$$\begin{aligned}\varphi_N'(0) &= 2\Pr[N=2]\\ &= \frac{\lambda^2\tau}{\left(1-2\tau\lambda(z-1)\right)^{3/2}}\varphi_N(z)\Big|_{z=0} + \frac{\lambda}{\sqrt{1-2\tau\lambda(z-1)}}\varphi_N'(z)\Big|_{z=0}\\ &= \frac{\lambda^2\tau}{\left(1+2\tau\lambda\right)^{3/2}}\Pr[N=0] + \frac{\lambda}{\sqrt{1+2\tau\lambda}}\Pr[N=1]\\ &= \frac{\lambda^2\tau}{\left(1+2\tau\lambda\right)^{3/2}}\frac{\sqrt{1+2\tau\lambda}}{\lambda}\Pr[N=1] + \left(\frac{\lambda}{\sqrt{1+2\tau\lambda}}\right)^2\Pr[N=0]\\ &= \frac{\lambda\tau}{1+2\tau\lambda}\Pr[N=1] + \frac{\lambda^2}{1+2\tau\lambda}\Pr[N=0].\end{aligned}$$

In general, we have the following recursive formula

$$\begin{aligned}\Pr[N=n] =& \frac{2\lambda\tau}{1+2\lambda\tau}\left(1-\frac{3}{2n}\right)\Pr[N=n-1]\\ &+ \frac{\lambda^2}{(1+2\lambda\tau)n(n-1)}\Pr[N=n-2]\end{aligned} \tag{1.43}$$

valid for $n=2,3,4,\ldots$, which allows us to compute the probability mass function of the Poisson-Inverse Gaussian distribution. The formal proof of (1.43) is based on properties of the modified Bessel function.

1.4.7 Poisson-LogNormal Distribution

In addition to the Gamma and Inverse Gaussian distributions to model Θ, the LogNormal distribution is often used in biostatistical studies.

LogNormal Distribution

Recall that a random variable X is Normally distributed with mean μ and variance σ^2, denoted as $X \sim \mathcal{N}or(\mu, \sigma^2)$, if its distribution function is

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-y^2/2)\, dy. \tag{1.44}$$

Now, a random variable X is LogNormally distributed with parameters μ and σ (notation $X \sim \mathcal{LN}or(\mu, \sigma^2)$) if $\ln X$ is Normally distributed with mean μ and variance σ^2, that is, if its probability density function is given by

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

If $X \sim \mathcal{LN}or(\mu, \sigma^2)$, then its mean is

$$\mathbb{E}[X] = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

and its variance

$$\mathbb{V}[X] = \exp\big(2\mu + \sigma^2\big)\big(\exp(\sigma^2) - 1\big).$$

Poisson-LogNormal Distribution

Taking $\mu = -\sigma^2/2$ (to ensure that $\mathbb{E}[\Theta] = 1$), the probability density function of $\Theta \sim \mathcal{LN}or(-\sigma^2/2, \sigma^2)$ is

$$f_\Theta(\theta) = \frac{1}{\theta\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln\theta + \sigma^2/2)^2}{2\sigma^2}\right), \quad \theta > 0. \tag{1.45}$$

The probability mass function of the Poisson-LogNormal distribution is given by

$$\Pr[N = k] = \frac{1}{\sigma\sqrt{2\pi}} \frac{(\lambda d)^k}{k!} \int_0^\infty \exp(-\lambda d\theta)\theta^{k-1} \exp\left(-\frac{(\ln\theta + \sigma^2/2)^2}{2\sigma^2}\right) d\theta.$$

Coming back to (1.28) and (1.29), we easily see that

$$\mathbb{E}[N] = \lambda \text{ and } \mathbb{V}[N] = \lambda + \lambda^2\big(\exp(\sigma^2) - 1\big).$$

It can be shown that $\gamma[\Theta]/\sqrt{\mathbb{V}[\Theta]} = 2 + \exp(\sigma^2)$ in (1.31) for the Poisson-LogNormal distribution. Therefore the skewness of a Poisson-LogNormal distribution exceeds the skewness of the Poisson-Inverse Gaussian distribution having the same mean and the same variance.

1.5 Statistical Inference for Discrete Distributions

1.5.1 Maximum Likelihood Estimators

Maximum likelihood is a method of estimation and inference for parametric models. The maximum likelihood estimator is the value of the parameter (or parameter vector) that makes the observed data most likely to have occurred given the data generating process assumed to have produced the variable of interest.

The likelihood of a sample of observations is defined as the joint density of the data, with the parameters taken as variable and the data as fixed (multiplied by any arbitrary constant or function of the data but not of the parameters). Specifically, let $N_1, N_2, \ldots, N_n$ be a set of independent and identically distributed outcomes with probability mass function $p(\cdot|\boldsymbol{\xi})$ where $\boldsymbol{\xi}$ is a vector of parameters. The likelihood function is the probability of observing the data $N_1 = k_1, \ldots, N_n = k_n$, that is,

$$\mathcal{L}(\boldsymbol{\xi}) = \prod_{i=1}^{n} p(k_i|\boldsymbol{\xi}).$$

The key idea for estimation in likelihood problems is that the most reasonable estimate is the value of the parameter vector that would make the observed data most likely to occur. The implicit assumption is of course that the data at hand are reliable. More formally we seek a value of $\boldsymbol{\xi}$ that maximizes $\mathcal{L}(\boldsymbol{\xi})$. The maximum likelihood estimator of $\boldsymbol{\xi}$ is the random variable $\widehat{\boldsymbol{\xi}}$ for which the likelihood is maximum, that is

$$\mathcal{L}(\widehat{\boldsymbol{\xi}}) \geq \mathcal{L}(\boldsymbol{\xi}) \text{ for all } \boldsymbol{\xi}.$$

It is usually simpler mathematically to find the maximum of the logarithm of the likelihood

$$L(\boldsymbol{\xi}) = \ln \mathcal{L}(\boldsymbol{\xi}) = \sum_{i=1}^{n} \ln p(k_i|\boldsymbol{\xi})$$

rather than the likelihood itself. The function $L(\boldsymbol{\xi})$ is usually referred to as the log-likelihood. Because the logarithm is a monotonic transformation, the log-likelihood will be maximized at the same parameter value that maximizes the likelihood (although the shape of the log-likelihood is different from that of the likelihood).

When working with counting variables, it is often easier to use the observed frequencies

$$f_k = \#\{\text{observations equal to } k\}, \quad k = 0, 1, 2, \ldots \tag{1.46}$$

In other words, f_k is the number of times that the value k has been observed in the sample. Denoting the largest observation as

$$k_{\max} = \max_{i=1,\ldots,n} k_i,$$

the log-likelihood becomes

$$L(\boldsymbol{\xi}) = \sum_{k=0}^{k_{\max}} f_k \ln p(k|\boldsymbol{\xi}).$$

We may solve analytically for the maximum likelihood estimator. To maximize any regular function, we find the value of the parameters that makes the first derivatives of the function with respect to the parameters equal to zero. The first derivative of the log-likelihood is called Fisher's score, and is denoted by

$$U_j(\boldsymbol{\xi}) = \frac{\partial}{\partial \xi_j} L(\boldsymbol{\xi}),\ j = 1, \ldots, \dim(\boldsymbol{\xi}). \tag{1.47}$$

Then one can find the maximum likelihood estimator by setting the score to zero, i.e. by solving the system of equations

$$U_j(\boldsymbol{\xi}) = 0,\ j = 1, \ldots, \dim(\boldsymbol{\xi}).$$

We also check the second derivatives to ensure that this is a maximum.

Example 1.2 Assume that policyholder i has been observed during a period d_i and produced k_i claims. Assuming that the annual number of claims is Poisson distributed with mean λ (here, λ is the annual expected claim frequency, so that policyholder i is expected to produce $d_i\lambda$ claims during the observation period, under the condition of the Poisson process); the log-likelihood is

$$\begin{aligned} L(\lambda) &= \sum_{i=1}^{n} \ln\left(\exp(-\lambda d_i)\frac{(\lambda d_i)^{k_i}}{k_i!}\right) \\ &= -\lambda \sum_{i=1}^{n} d_i + \sum_{i=1}^{n} k_i\big(\ln\lambda + \ln d_i\big) - \sum_{i=1}^{n} \ln k_i! \\ &= -\lambda \sum_{i=1}^{n} d_i + \ln\lambda \sum_{i=1}^{n} k_i + \text{constant}. \end{aligned}$$

Setting the first derivative of $L(\lambda)$ with respect to λ equal to 0 gives

$$-\sum_{i=1}^{n} d_i + \frac{1}{\lambda}\sum_{i=1}^{n} k_i = 0 \Rightarrow \widehat{\lambda} = \frac{\sum_{i=1}^{n} k_i}{\sum_{i=1}^{n} d_i}.$$

The second derivative is

$$-\frac{1}{\lambda^2}\sum_{i=1}^{n} k_i < 0 \text{ for any } \lambda$$

so that $\widehat{\lambda}$ indeed corresponds to the maximum of $L(\lambda)$. The estimated annual expected claim frequency $\widehat{\lambda}$ is thus obtained as the ratio of the total number of claims to the total exposure-to-risk. It is important to note here that the total number of claims is not divided by the number of policies, because of unequal risk exposure.

Example 1.3 Assume, as above, that policyholder i has been observed during a period d_i and produced k_i claims. Assuming that the annual number of claims filed by policyholder i is Negative Binomially distributed with mean λd_i, the log-likelihood is

$$L(a,\lambda)=\sum_{i=1}^{n}\sum_{j=0}^{k_i-1}\ln(a+j)+na\ln a-\sum_{i=1}^{n}(a+k_i)\ln(a+\lambda d_i)+\ln\lambda\sum_{i=1}^{n}k_i+\text{constant}.$$

The maximum likelihood estimators for a and λ solve

$$\frac{\partial}{\partial a}L(a,\lambda)=\sum_{i=1}^{n}\sum_{j=0}^{k_i-1}\frac{1}{a+j}+n\ln a+n-\sum_{i=1}^{n}\ln(a+\lambda d_i)-\sum_{i=1}^{n}\frac{a+k_i}{a+\lambda d_i}=0$$

$$\frac{\partial}{\partial\lambda}L(a,\lambda)=-\sum_{i=1}^{n}d_i\frac{a+k_i}{a+\lambda d_i}+\frac{1}{\lambda}\sum_{i=1}^{n}k_i=0.$$

These equations do not possess explicit solutions, and must be solved numerically. A convenient choice is to use the Newton–Raphson algorithm (see Section 1.5.3). Initial values for the parameters are obtained by the method of moments. Specifically, the moment estimator of λ is simply

$$\hat{\lambda}=\frac{\sum_{i=1}^{n}k_i}{\sum_{i=1}^{n}d_i},$$

which is the maximum likelihood estimate of λ in the homogeneous Poisson case. For the variance, we start from $\mathbb{V}[N_i]=\mathbb{E}[N_i]+(\lambda d_i)^2\tau$ where $\tau=\mathbb{V}[\Theta_i]$. The empirical analogue is given by

$$\hat{\tau}=\frac{\sum_{i=1}^{n}\left((k_i-\lambda d_i)^2-\lambda d_i\right)}{\sum_{i=1}^{n}(\lambda d_i)^2}$$

from which we easily deduce an estimator for a in the Negative Binomial case.

1.5.2 Properties of the Maximum Likelihood Estimators

Maximum likelihood estimators enjoy a number of convenient properties that are discussed below. It is important to note that these are asymptotic properties, i.e. properties that hold only as the sample size becomes infinitely large. It is impossible to say in general at what point a sample is large enough for these properties to apply, but the majority of actuarial applications involve large data sets so that actuaries generally trust in the large sample properties of the maximum likelihood estimators.

Consistency

First, maximum likelihood estimators are consistent. There are several definitions of consistency, but an intuitive version is that as the sample size gets large the estimator is increasingly likely to fall within a small region around the true value of the parameter. This

is called convergence in probability and is defined more formally as follows: A consistent estimator T_j for some parameter ξ_j computed from a sample of size n is one for which

$$\lim_{n\nearrow\infty} \Pr[|T_j - \xi_j| \geq c] = 0$$

for all positive c. This will henceforth be denoted as $T_j \to_{\text{proba}} \xi_j$ as $n \nearrow +\infty$. A consistent estimator is thus an estimator that converges to the population parameter as the sample size goes to infinity. Consistency is an asymptotic property.

Asymptotic Normality

Any estimator will vary across repeated samples. We must be able to calculate this variability in order to express our uncertainty about a parameter value and to make statistical inferences about the parameters. This variability is measured by the variance-covariance matrix of the estimators. This matrix provides the variances for each parameter on the main diagonal while the off-diagonal elements estimate the covariances between all pairs of parameters.

The asymptotic variance-covariance matrix $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\xi}}}$ for maximum likelihood estimators $\widehat{\boldsymbol{\xi}}$ is the inverse of what is called the Fisher information matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\xi})$. Element ij of $\boldsymbol{\mathcal{I}}(\boldsymbol{\xi})$ is given by

$$\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2}{\partial\xi_i\partial\xi_j}\ln L(\boldsymbol{\xi})\right] &= -n\mathbb{E}\left[\frac{\partial^2}{\partial\xi_i\partial\xi_j}\ln p(N_1|\boldsymbol{\xi})\right] \\
&= n\mathbb{E}\left[\frac{\partial}{\partial\xi_i}\ln p(N_1|\boldsymbol{\xi})\frac{\partial}{\partial\xi_j}\ln p(N_1|\boldsymbol{\xi})\right] \\
&= n\sum_{k=0}^{\infty} p(k|\boldsymbol{\xi})\frac{\partial}{\partial\xi_i}\ln p(k|\boldsymbol{\xi})\frac{\partial}{\partial\xi_j}\ln p(k|\boldsymbol{\xi}).
\end{aligned}$$

Thus, $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\xi}}} = \left(\boldsymbol{\mathcal{I}}(\boldsymbol{\xi})\right)^{-1}$.

An insight into why this makes sense is that the second derivatives measure the rate of change in the first derivatives, which in turn determines the value of the maximum likelihood estimate. If the first derivatives are changing rapidly near the maximum, then the peak of the likelihood is sharply defined and the maximum is easy to see. In this case, the second derivatives will be large and their inverse small, indicating a small variance of the estimated parameters. If on the other hand the second derivatives are small, then the likelihood function is relatively flat near the maximum and so the parameters are less precisely estimated. The inverse of the second derivatives will produce a large value for the variance of the estimates, indicating low precision of the estimates.

The distribution of $\widehat{\boldsymbol{\xi}}$ is usually difficult to obtain. Therefore we resort to the following asymptotic theory: Under mild regularity conditions (including that the true value of the parameter $\boldsymbol{\xi}$ must be interior to the parameter space, that the log-likelihood function must be thrice differentiable, and that the third derivatives must be bounded) that are usually fulfilled, the maximum likelihood estimator $\widehat{\boldsymbol{\xi}}$ has approximately in large samples a multivariate Normal distribution with mean equal to the true parameter and variance-covariance matrix given by the inverse of the information matrix.

Recall that having a $n \times n$ positive definite matrix $\boldsymbol{M}$ and a real vector $\boldsymbol{\mu}$, the random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ is said to have the multivariate Normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{M}$ if its probability density function is of the form

$$f_X(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{M})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{M}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{1.48}$$

Henceforth, we indicate that the random vector $\boldsymbol{X}$ has the multivariate Normal distribution with probability density function (1.48) as $\boldsymbol{X} \sim \mathcal{N}or(\boldsymbol{\mu}, \boldsymbol{M})$. A convenient characterization of the multivariate Normal distribution is as follows: $\boldsymbol{X} \sim \mathcal{N}or(\boldsymbol{\mu}, \boldsymbol{M})$ if, and only if, any random variable of the form $\sum_{i=1}^n \alpha_i X_i$ with $\boldsymbol{\alpha} \in \mathbb{R}^n$, has the univariate Normal distribution.

Coming back to the properties of the maximum likelihood estimator $\widehat{\boldsymbol{\xi}}$, we have that

$$\widehat{\boldsymbol{\xi}} \text{ is approximately } \mathcal{N}or(\boldsymbol{\xi}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\xi}}}) \text{ distributed,} \tag{1.49}$$

that is, the distribution function of $\widehat{\boldsymbol{\xi}}$ can be approximated by integrating the Normal probability density function

$$f_{\boldsymbol{\xi}}(\boldsymbol{S}) = \frac{1}{\sqrt{(2\pi)^{\dim(\boldsymbol{\xi})} \det(\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\xi}}})}} \exp\left(-\frac{1}{2}(\boldsymbol{S}-\boldsymbol{\xi})^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\xi}}}^{-1}(\boldsymbol{S}-\boldsymbol{\xi})\right), \quad \boldsymbol{S} \in \mathbb{R}^{\dim(\boldsymbol{\xi})}.$$

Attribute (1.49) says that maximum likelihood estimators converge in distribution to a Normal with mean equal to the population value of the parameter and variance-covariance matrix equal to the inverse of the information matrix. This means that regardless of the distribution of the variable of interest the maximum likelihood estimator of the parameters will have a multivariate Normal distribution. Thus, a variable may be Poisson distributed, but the maximum likelihood estimate of the Poisson mean will be asymptotically Normally distributed, and likewise for any distribution. Note however that in the Poisson case, the exact distribution of the maximum likelihood estimator of the parameter derived in Example 1.2 can easily be derived from the stability of the Poisson family under convolution.

Invariance

A natural question is how the parameterization of a likelihood affects the resulting inference. Maximum likelihood has the property that any transformation of a parameter can be estimated by the same transformation of the maximum likelihood estimate of that parameter. This provides substantial flexibility in how we parameterize our models while guaranteeing that we will get the same result if we start with a different parameterization.

The invariance property can be stated formally as follows: If $\boldsymbol{\gamma} = t(\boldsymbol{\xi})$, where $t(\cdot)$ is a one-to-one transformation, then the maximum likelihood estimator of $\boldsymbol{\gamma}$ is $t(\widehat{\boldsymbol{\xi}})$. In particular, the maximum likelihood estimator of $p(k|\boldsymbol{\xi})$ is simply $p(k|\widehat{\boldsymbol{\xi}})$, that is,

$$\widehat{p(k|\boldsymbol{\xi})} = p(k|\widehat{\boldsymbol{\xi}}).$$

1.5.3 Computing the Maximum Likelihood Estimators with the Newton–Raphson Algorithm

Calculation of the maximum likelihood estimators often requires iterative procedures. Let $\boldsymbol{H}$ denote the Hessian (or matrix of second derivatives) of the log-likelihood function, with elements

$$\begin{aligned} H_{ij}(\boldsymbol{\xi}) &= \frac{\partial^2}{\partial\xi_i\partial\xi_j}L(\boldsymbol{\xi}) \\ &= \frac{\partial}{\partial\xi_i}U_j(\boldsymbol{\xi}) \\ &= -\sum_{k=0}^{k_{\max}} f_k \frac{\partial^2}{\partial\xi_i\partial\xi_j}\ln p(k|\boldsymbol{\xi}) \end{aligned} \tag{1.50}$$

for $i, j = 1, \ldots, \dim(\boldsymbol{\xi})$. For $\boldsymbol{\xi}^\star$ close enough to $\widehat{\boldsymbol{\xi}}$, a first-order Taylor expansion gives

$$0 = \boldsymbol{U}(\widehat{\boldsymbol{\xi}}) \approx \boldsymbol{U}(\boldsymbol{\xi}^\star) + \boldsymbol{H}(\boldsymbol{\xi}^\star)\left(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^\star\right)$$

yielding

$$\widehat{\boldsymbol{\xi}} \approx \boldsymbol{\xi}^\star - \boldsymbol{H}^{-1}(\boldsymbol{\xi}^\star)\boldsymbol{U}(\boldsymbol{\xi}^\star).$$

Starting from an appropriate initial value $\boldsymbol{\xi}^{(0)}$, the Newton–Raphson algorithm is based on the recurrence relation

$$\widehat{\boldsymbol{\xi}}^{(r+1)} = \widehat{\boldsymbol{\xi}}^{(r)} - \boldsymbol{H}^{-1}\left(\widehat{\boldsymbol{\xi}}^{(r)}\right)\boldsymbol{U}\left(\widehat{\boldsymbol{\xi}}^{(r)}\right). \tag{1.51}$$

This result provides the basis for an iterative approach for computing the maximum likelihood estimator known as the Newton–Raphson technique. Given a trial value, we use (1.51) to obtain an improved estimate and repeat the process until the elements of the vector of first derivatives are sufficiently close to zero.

This procedure tends to converge quickly if the log-likelihood is well-behaved in a neighbourhood of the maximum and if the starting value is reasonably close to the maximum likelihood estimator.

Remark 1.2 (Fisher Scoring) Noting that $\boldsymbol{\mathcal{I}}(\boldsymbol{\xi}) = -\mathbb{E}[\boldsymbol{H}(\boldsymbol{\xi})]$, an alternative procedure is to replace minus the Hessian by its expected value, i.e. minus the Fisher information matrix. The resulting procedure takes as an improved estimate

$$\widehat{\boldsymbol{\xi}}^{(r+1)} \approx \widehat{\boldsymbol{\xi}}^{(r)} + \boldsymbol{\mathcal{I}}^{-1}\left(\widehat{\boldsymbol{\xi}}^{(r)}\right)\boldsymbol{U}\left(\widehat{\boldsymbol{\xi}}^{(r)}\right)$$

and is known as Fisher Scoring.

1.5.4 Hypothesis Tests

Sample Distribution of Individual Parameters

Standard hypothesis tests about parameters in maximum likelihood models are handled quite easily, thanks to the asymptotic Normal distribution of the maximum likelihood estimator. Specifically, we use the fact that

$$\frac{\widehat{\xi}_j - \xi_j}{\sigma_{\widehat{\xi}_j}} \text{ is approximately } \mathcal{N}or(0,1)$$

where the standard deviation $\sigma_{\widehat{\xi}_j}$ of $\widehat{\xi}_j$ is the square root of the jth diagonal element of

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\xi}}} = (\boldsymbol{\mathcal{I}}(\boldsymbol{\xi}))^{-1}.$$

Such tests will be useful in Chapter 2 to select the relevant risk factors.

In practice, $\sigma_{\widehat{\xi}_j}$ often involves unknown parameters $\boldsymbol{\xi}$ so that it is estimated by the jth element $\widehat{\sigma}_{\widehat{\xi}_j}$ of

$$\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\xi}}} = (\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\xi}}))^{-1}.$$

In such a case, $(\widehat{\xi}_j - \xi_j)/\widehat{\sigma}_{\widehat{\xi}_j}$ is approximately Student's distributed with $n-1$ degrees of freedom. This is the familiar z-score for a standard Normal variable developed in all introductory statistics classes. The Normality of maximum likelihood estimates means that our testing of hypotheses about the parameters is as simple as calculating the z-score and finding the associated p-value from a table or by calling a software function.

The hypothesis test is based on Student's t-distribution. However, because the maximum likelihood properties are all asymptotic we are unable to address the finite sample distribution. Asymptotically, the Student's t-distribution converges to the Normal as the degrees for freedom grow, so that using $\mathcal{N}or(0,1)$ p-values in the maximum likelihood test is the same as the t-test as long as the number of cases is large enough. Specifically,

$$\frac{\widehat{\xi}_j - \xi_j}{\widehat{\sigma}_{\widehat{\xi}_j}} \text{ is approximately } \mathcal{N}or(0,1) \tag{1.52}$$

if the sample size n is large enough.

In addition to the test of hypotheses about a single parameter, there are three classical tests that encompass hypotheses about sets of parameters as well as one parameter at a time: the likelihood ratio, Wald, and Lagrange multiplier tests. All are asymptotically equivalent, but they differ in the ease of implementation depending on the particular case. Here, we will present Wald, likelihood ratio and Vuong tests, as well as the Score test.

Likelihood Ratio Test

This test is based on a comparison of maximized likelihoods for nested models. Specifically, the null hypothesis H_0 corresponds to a constrained model with $\dim(\boldsymbol{\xi}) - j$ parameters, whereas the alternative H_1 corresponds to the full model with $\dim(\boldsymbol{\xi})$ parameters. Most of

the time, the test is performed with $j = 1$, so that we compare the full model to a simpler one with one parameter less.

Let $\widetilde{\boldsymbol{\xi}}$ be the maximum likelihood estimator under H_0, and let $\widehat{\boldsymbol{\xi}}$ be the maximum likelihood estimator under H_1. The likelihood ratio test is based on the ratio of the likelihoods between a full and a restricted (or reduced) nested model with fewer parameters. The restricted model must be nested within (i.e., be a subset of) the full model. The likelihood ratio test statistic is

$$T = 2\ln\frac{\mathcal{L}(\widehat{\boldsymbol{\xi}})}{\mathcal{L}(\widetilde{\boldsymbol{\xi}})} = 2\Big(L(\widehat{\boldsymbol{\xi}}) - L(\widetilde{\boldsymbol{\xi}})\Big).$$

The evidence against H_0 will be strong when T is large.

The Chi-square distribution plays a prominent role in likelihood ratio tests. Recall that the Gamma distribution with $\alpha = \nu/2$ and $\beta = 1/2$ for some positive integer ν is known as the Chi-square distribution with ν degrees of freedom (which is denoted as χ^2_ν), with associated probability density function

$$f(x) = \frac{x^{\nu/2-1}\exp(-x/2)}{\Gamma\left(\frac{\nu}{2}\right)2^{\nu/2}}, \quad x > 0.$$

If $X \sim \chi^2_\nu$ then its mean is ν, and its variance 2ν. It is useful to recall that the χ^2_ν distribution is closely related to the Normal distribution. Specifically, the χ^2_ν arises as the distribution of the sum of ν independent squared $\mathcal{N}or(0, 1)$ random variables.

Under H_0, the test statistic T is approximatively Chi-square distributed with degrees of freedom equal to the number of parameters in the full model minus the number of parameters in the restricted model (that is, with j degrees of freedom) when the sample size n is sufficiently large (and additional mild regularity conditions are fulfilled). Note that the likelihood ratio test requires us to perform two maximum likelihood estimations, one under H_0 and another one under H_1. When the largest model H_1 is misspecified (that is, the data have not been generated by this probability model), the likelihood ratio statistic is no longer necessarily Chi-square distributed under H_0.

Unfortunately, there are cases where regularity conditions do not hold for T to be approximately χ^2_j distributed under H_0. In particular this happens when a constrained parameter is on the boundary of the parameter space, e.g., testing Poisson versus Negative Binomial. Here Poisson is a particular case of Negative Binomial when the latter has a parameter on its boundary space. In this case, the limiting distribution of the statistic T becomes a mixture of Chi-square distributions. We refer the reader to TITTERINGTON *ET AL.* (1985) for more details about these situations.

Wald Tests

The Wald test provides an alternative to the likelihood ratio test that requires the estimation of only the full model, not the restricted model. The logic of the Wald test is that if the restrictions are correct then the unrestricted parameter estimates should be close to the value hypothesized under the restricted model.

The Wald test is based on the distribution of a quadratic form of the weighted sum of squared Normal deviates, a form that is known to be Chi-square distributed. Specifically, using (1.49), we can test $H_0 : \boldsymbol{\xi} = \boldsymbol{\xi}_0$ versus $H_1 : \boldsymbol{\xi} \neq \boldsymbol{\xi}_0$ with the statistic

$$W = (\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0)\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\xi}})(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0)$$

which is approximately $\chi^2_{\dim(\boldsymbol{\xi})}$ distributed under H_0, in large samples. The test statistic can be interpreted as a measure of the distance between the maximum likelihood estimator $\widehat{\boldsymbol{\xi}}$ and the hypothesized value $\boldsymbol{\xi}_0$. The Wald test leads to the rejection of H_0 in favor of H_1 if $\widehat{\boldsymbol{\xi}}$ is too far from $\boldsymbol{\xi}_0$. Note that the Wald test suffers from the same problems as likelihood ratio tests when $\boldsymbol{\xi}_0$ lies on the boundary of the parametric space.

Sometimes the calculation of the expected information is difficult, and we may use the observed information instead.

Score Tests

Using the asymptotic theory, we have that $\boldsymbol{U}(\boldsymbol{\xi})$ is approximately $\mathcal{N}or(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\xi}))$ distributed. Therefore we can test $H_0 : \boldsymbol{\xi} = \boldsymbol{\xi}_0$ versus $H_1 : \boldsymbol{\xi} \neq \boldsymbol{\xi}_0$ with the statistic

$$Q = \boldsymbol{U}(\boldsymbol{\xi}_0)\boldsymbol{I}^{-1}(\boldsymbol{\xi}_0)\boldsymbol{U}(\boldsymbol{\xi}_0)$$

which is approximately $\chi^2_{\dim(\boldsymbol{\xi})}$ distributed under H_0, in large samples.

The advantage of the score test is that the calculation of the maximum likelihood estimator $\widehat{\boldsymbol{\xi}}$ is bypassed. Moreover, it remains applicable even if $\boldsymbol{\xi}_0$ lies on the boundary of the parametric space.

Vuong Test

The Chi-square approximation to the distribution of the likelihood ratio test statistic is valid only for testing restrictions on the parameters of a statistical model (i.e., H_0 and H_1 are nested hypotheses). With non-nested models, we cannot make use of likelihood ratio tests for model comparison. In this case, information criteria like AIC or (S)BIC are useful, as well as the Vuong test for non-nested models. Recall that the AIC (Akaike Information Criteria), is given by

$$\text{AIC} = -2L(\widehat{\boldsymbol{\xi}}) + 2\dim(\boldsymbol{\xi}),$$

and the BIC (Bayesian Information Criteria), is given by

$$\text{BIC} = -2L(\widehat{\boldsymbol{\xi}}) + \ln(n)\dim(\boldsymbol{\xi}).$$

Both criteria are equal to minus two times the maximum log-likelihood, penalized by a function of the number of observations and sample size.

Vuong (1989) proposed a likelihood ratio-based statistic for testing the null hypothesis that the competing models are equally close to the true data generating process against the alternative that one model is closer. Consider two statistical models given by the probability mass functions $p(\cdot|\boldsymbol{\xi})$ and $q(\cdot|\boldsymbol{\zeta})$ with $\dim(\boldsymbol{\xi}) = \dim(\boldsymbol{\zeta})$, and define the likelihood ratio statistic for the model $p(\cdot|\boldsymbol{\xi})$ against $q(\cdot|\boldsymbol{\zeta})$ as

$$LR(\widehat{\boldsymbol{\xi}}_n, \widehat{\boldsymbol{\zeta}}_n) = \sum_{k=0}^{k_{\max}} f_k \ln \frac{p(k|\widehat{\boldsymbol{\xi}}_n)}{q(k|\widehat{\boldsymbol{\zeta}}_n)}$$

where $\widehat{\boldsymbol{\xi}}_n$ and $\widehat{\boldsymbol{\zeta}}_n$ are the maximum likelihood estimators in each model based on the sample $\{k_1, \ldots, k_n\}$ and f_k is defined as in (1.46).

If both models are strictly non-nested (so that standard likelihood ratio tests do not apply) then under H_0

$$\frac{LR(\widehat{\boldsymbol{\xi}}_n, \widehat{\boldsymbol{\zeta}}_n)}{\widehat{\omega}_n \sqrt{n}} \text{ is approximately } \mathcal{N}or(0, 1) \text{ distributed,}$$

where

$$\widehat{\omega}_n = \frac{1}{n} \sum_{i=1}^{n} \left(\ln \frac{p(k_i|\widehat{\boldsymbol{\xi}}_n)}{q(k_i|\widehat{\boldsymbol{\zeta}}_n)} \right)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} \ln \frac{p(k_i|\widehat{\boldsymbol{\xi}}_n)}{q(k_i|\widehat{\boldsymbol{\zeta}}_n)} \right)^2.$$

This provides a very simple test for model selection. Specifically, the actuary chooses a critical value z_ϵ from the $\mathcal{N}or(0, 1)$ distribution for some significance level ϵ. If the value of the test statistic is higher than z_ϵ then he rejects the null hypothesis that the models are equivalent in favour of $p(\cdot|\boldsymbol{\xi})$ being better than $q(\cdot|\boldsymbol{\zeta})$. If the test statistic is smaller than $-z_\epsilon$ then he rejects the null hypothesis in favour of $q(\cdot|\boldsymbol{\zeta})$ being better than $p(\cdot|\boldsymbol{\xi})$. Finally, if the test statistic is between $-z_\epsilon$ and z_ϵ then we cannot discriminate between the two competing models given the data.

The test statistic can be adjusted if the competing models do not have the same number of parameters, i.e. $\dim(\boldsymbol{\xi}) \neq \dim(\boldsymbol{\zeta})$ (which is not the case in this chapter).

1.6 Numerical Illustration

Here, we consider a Belgian motor third party liability insurance portfolio observed during the year 1997 (henceforth referred to as Portfolio A). The observed claim distribution is given in Table 1.1. A thorough description of this portfolio is deferred to Section 2.2.

We see from Table 1.1 that the total exposure is not equal to the number of policies due to the fact that some policies have not been in force during the full observation period (12 months). Some of them have been cancelled before the end of the observation period. Others have been written after the start of the observation period.

Let us now fit the observations to the Poisson, the Negative Binomial, the Poisson-Inverse Gaussian and the Poisson-LogNormal distributions. The results are summarized below:

Table 1.1 Observed claim distribution in Portfolio A.

Number of claims	Number of policies	Total exposure (in years)
0	12962	10 545.94
1	1369	1 187.13
2	157	134.66
3	14	11.08
4	3	2.52
Total	14505	11 881.35

Poisson the maximum likelihood estimate of the Poisson mean is $\widehat{\lambda} = 0.1462$. The 95 % confidence interval for λ is (0.1395;0.1532). The log-likelihood of the Poisson model is -5579.339.

Negative Binomial the maximum likelihood estimate of the mean is $\widehat{\lambda} = 0.1474$ and the dispersion parameter $\widehat{a} = 0.889$. The variance of the random effect is estimated as $\widehat{\mathbb{V}[\Theta]} = 1/\widehat{a} = 1.1253$. The respective 95 % confidence intervals are (0.1402;0.1551) for λ and (0.8144;1.4361) for $\mathbb{V}[\Theta]$. The log-likelihood of the Negative Binomial model is -5534.36, which is better than the Poisson log-likelihood.

Poisson-Inverse Gaussian the maximum likelihood estimation of the mean is $\widehat{\lambda} = 0.1475$, and the variance of the random effect is estimated to $\widehat{\mathbb{V}[\Theta]} = \widehat{\tau} = 1.1770$. The respective 95 % confidence intervals are (0.1402;0.1552) for λ and (0.8258;1.5282) for $\mathbb{V}[\Theta]$. The log-likelihood of the Poisson-Inverse Gaussian model is -5534.28, which is better than the Poisson log-likelihood and almost equivalent to the Negative Binomial log-likelihood.

Poisson-LogNormal the maximum likelihood estimation of the mean is $\widehat{\lambda} = 0.1476$, and $\widehat{\sigma}^2 = 0.7964$. The variance of the random effect is estimated to $\widehat{\mathbb{V}[\Theta]} = 1.2175$. The respective 95 % confidence intervals are (0.1403;0.1553) for λ and (0.6170;0.9758) for σ^2. The log-likelihood of the Poisson-LogNormal model is -5534.44, which is better than the Poisson log-likelihood and almost equivalent to the Negative Binomial and Poisson-Inverse Gaussian log-likelihoods.

The results have been obtained with the help of the SAS® procedure GENMOD for the Poisson and Negative Binomial distributions (details will be given in the next chapter) and by a direct maximization of the log-likelihood using the Newton–Raphson procedure (coded in the SAS® environment IML) in the Poisson-Inverse Gaussian and Poisson-LogNormal cases.

It is interesting to note that the values of $\widehat{\lambda}$ are different in the Poisson and mixed Poisson models. If all the risk exposures were equal then these values would have been the same in all cases.

Let us now compare the Poisson fit to Portfolio A with each of the mixed Poisson fits. To this end, we use a likelihood ratio test, with an adjusted Chi-square approximation (since the Poisson case is at the border of the mixed Poisson family). Comparing the Poisson fit to any of the three mixed Poisson models leads to a clear rejection of the former one:

Poisson against Negative Binomial likelihood ratio test statistic of 89.95, with a p-value less than 10^{-10}.

Poisson against Poisson-Inverse Gaussian likelihood ratio test statistic of 90.12, with a p-value less than 10^{-10}.

Poisson against Poisson-LogNormal likelihood ratio test statistic of 89.80, with a p-value less than 10^{-10}.

The rejection of the Poisson assumption in favour of a mixed Poisson model is interpreted as a sign that the portfolio is composed of different types of drivers (i.e. the portfolio is heterogeneous).

Now, comparing the three mixed Poisson models with the Vuong test gives:

Negative Binomial against Poisson-Inverse Gaussian Vuong test statistic equal to -0.1086, with p-value 91.36 %.

Poisson-LogNormal against Negative Binomial Vuong test statistic equal to -0.0435, with p-value 96.54 %.

Poisson-LogNormal against Poisson-Inverse Gaussian Vuong test statistic equal to -0.3254, with p-value 74.48 %.

We cannot discriminate between the three competing models given the data, and they all fit the model equally well.

Remark 1.3 (Chi-Square Goodness-of-Fit Tests) In many papers appearing in the actuarial literature devoted to the analysis of claim numbers, as well as in most empirical studies, Chi-square goodness-of-fit tests are performed to select the optimal model. However, this approach neglects the exposures-to-risk (acting as if all the policies were in the portfolio for the whole year). We do not rely on Chi-square goodness-of-fit tests here since they do not allow for unequal risk exposures. Note that the vast majority of papers appearing in the actuarial literature disregard risk exposures (and proceed as if all the risk exposures were equal to 1).

1.7 Further Reading and Bibliographic Notes

1.7.1 Mixed Poisson Distributions

Mixed Poisson distributions are often used to model insurance claim numbers. The statistical analysis of counting random variables is described in much detail in JOHNSON *ET AL.* (1992). An excellent introduction to statistical inference is provided by FRANKLIN (2005). In the actuarial literature, KLUGMAN *ET AL.* (2004) provide a good account of statistical inference applied to insurance data sets, and in particular the analysis of counting random variables. Generating functions are described in KENDALL & STUART (1977) and FELLER (1971).

The axiomatic approach for which the (mixed) Poisson distribution is the counting distribution for a (mixed) Poisson process is presented in GRANDELL (1997). Mixture models are discussed in LINDSAY (1995). See also TITTERINGTON *ET AL.* (1985). Let us mention the work by KARLIS (2005), who applied the EM algorithm for maximum likelihood estimation in mixed Poisson models.

1.7.2 Survey of Empirical Studies Devoted to Claim Frequencies

KESTEMONT & PARIS (1985), using mixtures of Poisson processes, defined a large class of probability distributions and developed an efficient method for estimating their parameters. For the six data sets in GOSSIAUX & LEMAIRE (1981), they proposed a law depending on three parameters and they always obtained extremely good fits. As particular cases of the laws introduced in KESTEMONT & PARIS (1985), we find the ordinary Poisson distribution, the Poisson-Inverse Gaussian distribution, and the Negative Binomial distribution.

TREMBLAY (1992) used the Poisson-Inverse Gaussian distribution. WILLMOT (1987) compared the Poisson-Inverse Gaussian distribution to the Negative Binomial one and concluded that the fits are superior with the Poisson-Inverse Gaussian in all the six cases

studied by Gossiaux & Lemaire (1981). See also the paper by Besson & Partrat (1990). Ruohonen (1987) considered a model for the claim number process. This model is a mixed Poisson process with a three-parameter Gamma distribution as the structure function and is compared with the two-parameter Gamma model giving the Negative Binomial distribution. He fitted his model to some data that can be found in the actuarial literature and the results were satisfying. Panjer (1987) proposed the Generalized Poisson-Pascal distribution (in fact, the Hofmann distribution), which includes three parameters, for the modelling of the number of automobile claims. The fits obtained were satisfactory, too. Note that the Pólya-Aeppli, the Poisson-Inverse Gaussian and the Negative Binomial are special cases of this distribution. Consul (1990) tried to fit the same six data sets by the Generalized Poisson distribution. Although the Generalized Poisson law is not rejected by a Chi-square test, the fits obtained by Kestemont & Paris (1985), for instance, are always better. Furthermore, Elvers (1991) reported that the Generalized Poisson distribution did not fit the data observed in a motor third party liability insurance portfolio very well. Islam & Consul (1992) suggested the Consul distribution as a probabilistic model for the distribution of the number of claims in automobile insurance. These authors approximated the chance mechanism which produces vehicle accidents by a branching process. They fit the model to the data sets used by Panjer (1987) and by Gossiaux & Lemaire (1981). Note that this model deals only with cars in accidents. Consequently, the zero-class has to be excluded. The fitted values seem good. However, this has to be considered cautiously, due to the comments by Sharif & Panjer (1993) who found serious flaws embedded in the fitting of the Consul model. In particular, the very restricted parameter space and some theoretical problems in the derivation of the maximum likelihood estimators. They refer to other simple probability models, such as the Generalized Poisson-Pascal or the Poisson-Inverse Gaussian, whose fits were found quite satisfying.

Denuit (1997) demonstrated that the Poisson-Goncharov distribution introduced by Lefèvre & Picard (1996) provides an appropriate probability model to describe the annual number of claims incurred by an insured motorist. Estimation methods were proposed, and the Poisson-Goncharov distribution successfully fitted the six observed claims distributions in Gossiaux & Lemaire (1981), as well as other insurance data sets.

1.7.3 Semiparametric Approach

Traditionally, actuaries have assumed that the distribution of θ values among all drivers is well approximated by a parametric distribution, be it Gamma, Inverse Gaussian or LogNormal. However, there is no particular reason to believe that F_Θ belongs to some specified parametric family of distributions. Therefore, it seems interesting to resort to a nonparametric estimator for F_Θ.

There have been several attempts to estimate the structure function in a mixed Poisson model nonparametrically. Most of them include the annual claim frequency λ in the random effect, and thus work with $\widetilde{\Theta} = \lambda\Theta$. Assuming that $\widetilde{\Theta}$ has a finite number of support points and that its probability distribution is uniquely determined by its moments, Tucker (1963) suitably precised by Lindsay (1989a,b) suggested estimation of the support points of $\widetilde{\Theta}$ and the corresponding probability masses by solving a moment system. This estimator was then smoothed by Carrière (1993b) using a mixture of LogNormal distribution functions where all the parameters are estimated by a method of moments.

In a seminal paper, Simar (1976) gave a detailed description of the nonparametric maximum likelihood estimator of $F_{\widetilde{\Theta}}$, as well as an algorithm for its computation. The nonparametric maximum likelihood estimator has a discrete distribution, and Simar (1976) obtained an upper bound for the size of its support.

Walhin & Paris (1999) showed that, although the nonparametric maximum likelihood estimator is powerful for evaluation of functionals of claim counts, it is not suitable for ratemaking, because it is purely discrete. For this reason, Denuit & Lambert (2001) proposed a smoothed version of the nonparametric maximum likelihood estimator. This approach is somewhat similar to the route followed by Carrière (1993b), who proposed to smooth the Tucker-Lindsay moment estimator with a LogNormal kernel.

Young (1997) applied nonparametric density estimation techniques to estimate $F_{\widetilde{\Theta}}$. Because the actuary only observes claim numbers and not the conditional mean, an estimation of the underlying risk parameter relating to the ith policy of the portfolio is the average claim number $\overline{x}_i$ (i.e. the total number of claims generated by this policy divided by the length of the exposure period). Therefore, given a kernel K, Young (1997) suggested estimating $dF_{\Theta}(\theta)$ by

$$d\widehat{F}_{\widetilde{\Theta}}(t) = \sum_{i=1}^{n} \frac{w_i}{h_i} K\left(\frac{t-\overline{x}_i}{h_i}\right)$$

in which h_i is a positive parameter called the bandwidth and w_i is a weight (taken to be the number of years the ith policy is in force divided by the total number of policy-years for the collective). Young (1997) suggested using the Epanechnikov kernel and determined the h_is in order to minimize the mean integrated squared error (by reference to a Normal prior).