T A R T U U N I V E R S I T Y

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Institute of Mathematical Statistics

Frazier Carsten

Overdispersed Models for Claim Count Distribution

Master's Thesis

Supervisor: Meelis Käärik, Ph.D

TARTU 2013

Contents

1 Introduction

Constructing models to predict future loss events is a fundamental duty of actuaries. However, large amounts of information are needed to derive such a model. When considering many similar data points (e.g., similar insurance policies or individual claims), it is reasonable to create a collective risk model, which deals with all of these policies/claims together, rather than treating each one separately. By forming a collective risk model, it is possible to assess the expected activity of each individual policy. This information can then be used to calculate premiums (see, e.g., Gray & Pitts, 2012).

There are several classical models that are commonly used to model the number of claims in a given time period. This thesis is primarily concerned with the Poisson model, but will also consider the Negative Binomial model and, to a lesser extent, the Binomial model. We will derive properties for each of these models, both in the case when all insurance policies cover the same time period, and when they cover different time periods.

The primary focus of this thesis is overdispersion, which occurs when the observed variance of the data in a model is greater than would be expected, given the model parameters. We consider several possible treatments for overdispersion, particularly those that apply to the Poisson model. First, we attempt to generalize the Poisson model by adding an overdispersion parameter (see, e.g., Käärik & Kaasik, 2012). Next, we search for ways to convert an overdispersed Poisson model to a Negative Binomial model. We will derive some basic properties (such as expectation, variance, and additivity properties) for all of the models mentioned above.

Finally, results of this thesis are explored in a practical sense, by attempting to fit computer-generated data into an overdispersed Poisson framework.

2 Classical Collective Risk Model

2.1 Properties

The Classical Collective Risk Model is a method of modeling insurance claims using grouped claims. In other words, a group of similar claims are all clustered together, rather than each being considered separately. This can be very useful when dealing with large amounts of similar data, such as claim sizes for car crashes or property damage.

Consider an insurance portfolio composed of some number of insurance policies, with each policy having a given distribution of loss sizes. Let N be the number of claims for this portfolio in a given time period (for example, one year). Let Z_j be the individual claim amounts, $j = 1, 2, ..., N$. For the moment, assume that individual claims are independent from one another, and that they are also independent of N. (Later, we will see what can happen when certain independence requirements are violated.) Then, the total claim amount S of the portfolio is (see, e.g., Gray & Pitts, 2012):

$$S = \sum_{j=1}^{N} Z_j$$

In general, the procedure for premium calculation using the collective risk model is as follows:

1. Group the policies of the portfolio into homogenous groups, or subportfolios. A subportfolio is homogenous when all of its policies have independent, identically distributed (i.i.d.) risk severities and frequencies.

2. Estimate the frequency (number of claims) and severity (average claim size) for each subportfolio.

3. For each subportfolio, estimate the total claim amount using the collective risk model. Proportionally divide the total expected claim amount between policies in the subportfolio.

Step (1) of this procedure (choosing a proper clustering algorithm for dividing the

portfolio into subportfolios) is outside of the scope of this thesis, so we will not discuss it. For our purposes, we need only to assume that a suitable algorithm exists, and that the portfolio is clustered accordingly.

For any given subportfolio, let S_* be the total claim amount and N_* the total number of claims. Let n be the number of policies in the subportfolio, and let N_i and Y_i, respectively, be the frequency (number of claims) and total claim amount for the ith policy ($i = 1, ..., n$). Let Z_{ij} be the jth claim from the ith policy ($j = 1, ..., N_i$). If certain indices are irrelevant to the current discussion, they may be omitted: N, Y may be written instead of N_i, Y_i, respectively, and Z_i or Z instead of Z_{ij}. If all of the above is assumed, then just like the portfolio as a whole, $S_* = \sum_{j=1}^{N_*} Z_j$. Furthermore,

$$ES = EN_* \cdot EZ, \tag{1}$$

$$VarS = EN_* \cdot VarZ + (EZ)^2 \cdot VarN_*. \tag{2}$$

Where EX and $VarX$ are, respectively, the expected value and variance of X. These properties of the collective risk model are well-known, but their proof is included in Appendix A.

These computations also hold for the individual policy level, so we see that:

$$EY_i = EN_i EZ, \tag{3}$$

$$VarY_i = EN_i VarZ + (EZ)^2 VarN_i. \tag{4}$$

Furthermore, we wish to find relations between the expectations and variances of N_* and N. We see that $N_* = \sum_{i=1}^{n} N_i$, so:

$$\begin{aligned} EN_* &= E\left(\sum_{i=1}^{n} N_i\right) = \sum_{i=1}^{n} E(N_i) \\ &= \sum_{i=1}^{n} EN = n \cdot EN. \end{aligned}$$

Since all N_i are independent and identically distributed, we see that:

$$\begin{aligned} VarN_* &= Var\left(\sum_{i=1}^{n} N_i\right) \\ &= \sum_{i=1}^{n} Var(N_i) = n \cdot VarN. \end{aligned}$$

If we assume that a portfolio can be divided into homogenous subportfolios, and that the total claim amount can be estimated for each subportfolio, then the pure premium for any given policy is the total pure premium of the subportfolio divided by the number of policies.

Note: We have been acting on the assumption that all of the policies have the same time duration. This is quite unlikely, so later in this paper, we will need to find a way to consider policies of different durations.

It is necessary to choose a distribution of claim frequency. The three classical distributions, which are the most commonly used and quoted, are the Poisson, Binomial, and Negative Binomial distributions (see Käärik & Kaasik, 2012).

Consider a random variable N.

- N follows the Poisson distribution ($N \sim Po(\lambda)$) if its probability mass function is given as:

  $$f(k) = \mathbf{P}(N = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

  Where $\lambda > 0$ is a given parameter (also called the *intensity*), and k is some nonnegative integer. Then, $f(k)$ gives the probability of k events occurring (in a specified time period depending on the model). For the Poisson distribution, $EN = VarN = \lambda$, so the Poisson distribution works well when the mean and variance are approximately equal.

  Intuitively, we may interpret the Poisson distribution as giving the probability that k independent, identically distributed, randomly-occurring events will take place in one unit of time, given that the expected (average) rate of occurrence is λ events per unit time. Under this interpretation, it is easy

to see why the Poisson distribution is so commonly used for insurance: it is a very simple way of forecasting random, independent loss events (car accidents, property damage, etc.), if the average rate of these events is known beforehand.

- N follows the Binomial distribution ($N \sim Bin(n, p)$) if its probability mass function is given as:

$$f(k) = \mathbf{P}(N = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $n \in \mathbb{N}$ and $p \in [0, 1]$ is a given probability. Note that for this distribution, k can only take values from 0 to n, all other values have probability 0. For the Binomial distribution. $EN = np, VarN = np(1-p)$. Since $p \in [0, 1]$, clearly $EN > VarN$.

Intuitively, the Binomial distribution models a process with only two possible outcomes being run multiple times. If the probability of success is p, then the Binomial distribution gives the probability of k successes out of a total of n trials. The simplest instance of a Binomial distribution involves flipping a coin n times, and counting the number of times that heads occur.

For insurance, the Binomial distribution can be useful for modeling the number of claims: if a portfolio has n i.i.d. policies, and each policy can incur at most one loss event with probability p, then clearly the number of claims will have a Binomial distribution.

- N follows the Negative Binomial distribution ($N \sim NB(r, p)$) if its probability mass function is given as:

$$f(k) = \mathbf{P}(N = k) = \binom{k+r-1}{k} (1-p)^r p^k,$$

where $r \in \mathbb{N}$ and $p \in [0, 1]$ is a given probability, as before. For the Negative Binomial distribution, $EN = \frac{pr}{1-p}, VarN = \frac{pr}{(1-p)^2}$. Therefore, $EN < VarN$.

Similar to the Binomial distribution, the Negative Binomial can be inter-

preted as modeling a process with only two outcomes that is run multiple times. If the probability of failure is $1-p$, and the process is repeated until r failures occur (and then stopped), then the Negative Binomial distribution gives the probability that k successes occur before the process is halted. A very simple example of the Negative Binomial distribution involves flipping a coin until it comes up tails a total of r times, and then counting the number of times that the coin displayed heads.

For most cases in insurance, $EN \leq VarN$, so the Binomial distribution is not often used. In general, we prefer to employ the Poisson distribution, due to its ease of use and wide range of applicability; however, it tends to fail when the mean and variance are notably different.

When a model has a variance that is greater than we would expect from the distribution that we use to describe it, we say that the model is overdispersed, or that it has overdispersion. For the Poisson model, clearly, overdispersion occurs when $VarN > EN$. Note that overdispersion does not necessarily mean that the entire model must be discarded; however, the distribution must be modified in some way. We will discuss overdispersion and possible treatments later in this paper; for now, it will not be considered.

2.2 Compound Poisson Model

Thus far we have begun to discuss ways of modeling the number of claims, but not the distribution of claim sizes. Let $N \sim Po(\lambda)$; that is, the random variable N has a Poisson distribution with parameter λ. Furthermore, let $Z_1, Z_2, \ldots$ be i.i.d. random variables with a given distribution, that are also independent of N. Consider $S = \sum_{i=1}^{N} Z_i$, the sum of a "Poisson-distributed number" of i.i.d. random variables. Then, S is said to have a Compound Poisson distribution.

At this point, it is useful for us to move from a fixed time model to one that can be used for varying periods of time. This is, in fact, quite simple to achieve with a Poisson model. We know that $EN = \lambda$, so λ can be considered as the average number of loss events occurring in one unit of time (say, one day or one

week). N is homogenous; that is, λ does not change over time. Because of this, we may consider the process $N(t)$, defined as the number of events occurring in time t, when the number of events occurring in each unit of time is an independent realization of N. $N(t)$ is then called a counting process with intensity λ (since $EN(1) = EN = \lambda$).

The Compound Poisson Model is extremely useful for the classical collective risk model, because it allows us to describe the number of claims in time t as $N(t)$, the individual claim sizes Z_i, and the total claim size S for a given portfolio.

Let $S(t)$ be a compound Poisson process. Let $N(t)$ and Z_i be as defined above, $i = 1, \ldots, n$. Then,

$$S(t) = \sum_{i=1}^{N(t)} Z_i.$$

We want to derive some of the general properties of this model, starting with the mean and variance. However, to do this, we first need to demonstrate the additivity property of the Poisson distribution.

Proposition 2.1. *If $N_1, \ldots, N_n$ are independent Poisson random variables with parameters $\lambda_1, \ldots, \lambda_n$, respectively, then their sum $N = N_1 + \cdots + N_n$ is a Poisson random variable with parameter $\lambda = \lambda_1 + \cdots + \lambda_n$.*

This property is well-known, but its proof is included in Appendix A.

We will return to this additivity property several times during this paper. However, it is particularly useful for our current discussion of the counting process $N(t)$. First, we assume that t is a positive integer. (This is a reasonable assumption for our purposes, since we generally do not want to consider periods of time smaller than one day.) If $N_1, \ldots, N_t$ are i.i.d. and each identically equal to N, then $N(t) = N_1 + \cdots + N_t$ is still a Poisson process, with parameter $\lambda_1 + \cdots + \lambda_t = t\lambda$.

Let $EZ = \mu$. We are finally prepared to derive the expectation and variance of $S(t)$.

As we showed in equation (1), the expected value of $S(t)$ is:

$$E(S(t)) = E(N(t)) \cdot E(Z) = t\mu\lambda \tag{5}$$

Furthermore, from equation (2), the variance of $S(t)$ is:

$$\begin{aligned} Var(S(t)) &= EN \cdot VarZ + (EZ)^2 \cdot VarN = t\lambda \cdot VarZ + EZ^2 t\lambda \\ &= t\lambda(VarZ + (EZ)^2) = t\lambda \cdot E(Z^2). \end{aligned}$$

$$Var(S(t)) = t\lambda E(Z^2) \tag{6}$$

The Compound Poisson process is largely useful because it allows us to consider the total claim amount rather than looking at each claim separately, which is cumbersome when dealing with a large number of claims. Of course, the model requires all of the usual restrictions placed on the Classical Collective Risk Model.

The additivity property of the Poisson distribution also carries over to the Compound Poisson process: Let $S_1(t), \ldots, S_n(t)$ be independent Compound Poisson processes with intensities $\lambda_1, \ldots, \lambda_n$. Then $S(t) = S_1(t) + \cdots + S_n(t)$ is a Compound Poisson process with intensity $\lambda = \lambda_1 + \cdots + \lambda_n$.

Once again, we will use a proof by induction. For $n = 1$, the solution is trivial. For $n = 2$,

$$\begin{aligned} S(t) = S_1(t) + S_2(t) = \sum_{i=1}^{N_1(t)} Z_i + \sum_{j=1}^{N_2(t)} Z_j = \sum_{k=1}^{N_1(t)+N_2(t)} Z_k \\ = \sum_{k=1}^{N(t)} Z_k. \end{aligned}$$

As we have proven, $N(t)$ is a Poisson counting process with intensity $\lambda = \lambda_1 + \lambda_2$, so $S(t)$ is a compound Poisson process. The remainder of the inductive proof is straightforward and directly follows the proof of the additivity property for Poisson random variables.

The additive property helps us to simplify our model still further, by allowing us to group together Compound Poisson-modeled subportfolios that have the same claim size distribution.

3 Compound Poisson Model with Different Insurance Periods

It is highly unlikely that all of the insurance policies in a portfolio will have the same time duration. Therefore, we need to extend the model to take different time periods into account.

For a given portfolio, we define n_i as the number of claims in the ith policy, t_i as the corresponding insurance period (i.e., the length of time of the policy), and n_{i_j} as the number of claims of policy i in time unit j (e.g., one day or one year). Then, we can say that $n_i = \sum_{j=1}^{t_i} n_{i_j}$.

Assume that we are considering a homogenous subportfolio. Then, we can say that the claim numbers (per unit time) n_{i_j} are independent instances of the random variable N_{i_j} for all $i = 1, \ldots, n$ and $j = 1, \ldots, t_i$. (We may also write $N_{[i]}$ if the exact time j is irrelevant). Therefore, each n_i is an independent instance of the random variable $N_i = \sum_{j=1}^{t_i} N_{i_j}$. As we know, N_i is Poisson distributed if all N_{i_j} are also Poisson distributed. We consider equations (4) and (5), and apply them to this model, to get estimates for the mean and variance of the claim amount for risk i, $i = 1, \ldots, n$. Observe that $EN_i = t_i EN_{[i]}$, and since $EN_i = VarN_i$ and $EN_{[i]} = VarN_{[i]}$, then $VarN_i = t_i VarN_{[i]} = t_i VarN_{ij}$. Therefore,

$$EY_i = t_i EN_{[i]} EZ, \tag{7}$$

$$VarY_i = t_i EN_{[i]} VarZ + t_i (EZ)^2 VarN_{[i]}. \tag{8}$$

Consider how N_i is distributed if N_{i_j} has a Poisson, Binomial, or Negative Binomial distribution.

- If $N_{i_j} \sim Po(\lambda)$: In section 2.2, we proved that the sum of n Poisson random variables with parameters $\lambda_1, \ldots, \lambda_n$ is a Poisson random variable with parameter $\lambda = \lambda_1 + \cdots + \lambda_n$. Thus, since $N_i = \sum_{j=1}^{t_i} N_{i_j}$, we see that $N_i \sim Po(t_i\lambda)$.
- If $N_{i_j} \sim Bin(n, p)$: Consider two Binomial distributions, $X \sim Bin(n, p)$ and

$Y \sim Bin(m, p)$. We will show that $X + Y \sim Bin(m + n, p)$.

Using a convolution argument, we see that:

$$\begin{aligned}
\mathbf{P}(X + Y = z) &= \sum_{k=0}^{z} \mathbf{P}(X = k)\mathbf{P}(Y = z - k) \\
&= \sum_{k=0}^{z} \binom{n}{k} p^k (1-p)^{n-k} \binom{m}{z-k} p^{z-k} (1-p)^{m-z+k} \\
&= p^z (1-p)^{m+n-z} \sum_{k=0}^{z} \binom{n}{k} \binom{m}{z-k}
\end{aligned}$$

Vandermonde's Identity (Quaintance, 2010) tells us that this sum equals $\binom{m+n}{z}$, so:

$\mathbf{P}(X + Y = z) = \binom{m+n}{z} p^z (1-p)^{m+n-z}$.

Thus, $X + Y \sim Bin(m + n, p)$. A simple inductive argument will then show that, for n Binomial random variables $X_i \sim Bin(m_i, p)$, $i = 1, \ldots, n$, their sum $X_1 + \cdots + X_n \sim Bin(m_1 + \cdots + m_n, p)$.

Therefore, if N_{i_j} follows a Binomial distribution, $N_i \sim Bin(t_i n, p)$.

- If $N_{i_j} \sim NBin(r, p)$: Similar to the Binomial distribution, the sum of $NBin(r_1, p), \ldots, NBin(r_m, p)$ is $NBin(r_1 + \cdots + r_m, p)$. Once again, we will prove this using a convolution argument. Consider $X \sim NB(r, p)$, $Y \sim NB(s, p)$. Then,

$$\begin{aligned}
\mathbf{P}(X + Y = z) &= \sum_{k=0}^{z} \mathbf{P}(X = k)\mathbf{P}(Y = z - k) \\
&= \sum_{k=0}^{z} \binom{k+r-1}{k} (1-p)^r p^k \binom{z-k+s-1}{z-k} (1-p)^s p^{z-k} \\
&= (1-p)^{r+s} p^z \sum_{k=0}^{z} \binom{k+r-1}{k} \binom{z-k+s-1}{z-k}.
\end{aligned}$$

There is a combinatoric identity (Quaintance, 2010) stating that:

$$\sum_{k=0}^{z} \binom{a+k}{k} \binom{b+z-k}{z-k} = \binom{a+b+z+1}{z}$$

If we set $a = r-1$, $b = s-1$, then we see that:

$\mathbf{P}(X+Y=z) = (1-p)^{r+s} p^z \binom{z+r+s-1}{z}$, so $X+Y \sim NB(r+s,p)$.

Therefore, if N_{i_j} follows a Negative Binomial distribution, $N_i \sim NBin(t_i r, p)$.

Next, we will assume that for a given portfolio (or subportfolio), the frequency per time unit, N_{i_j}, is Poisson distributed: $N_{i_j} \sim Po(\lambda)$. Then during the insurance period for the ith policy, $N_i \sim Po(t_i\lambda)$, where t_i is the length of the insurance period. In order to find the score function for λ, we first need the log-likelihood function.

For $i = 1, \ldots, n$, let n_i be independent realizations of Poisson distributions N_i with corresponding parameters λt_i. Then, the likelihood function for the Compound Poisson distribution is:

$$\mathcal{L}(\lambda t_1, \cdots, \lambda t_n | n_1, \cdots, n_n) = \prod_{i=1}^{n} P(n_i | \lambda t_i) = \prod_{i=1}^{n} \frac{(\lambda t_i)^{n_i} e^{-\lambda t_i}}{n_i!}.$$

Then, the corresponding log-likelihood function is:

$$\begin{aligned} \ln \mathcal{L}(\lambda t_1, \cdots, \lambda t_n | n_1, \cdots, n_n) &= \sum_{i=1}^{n} \ln \left(\frac{(\lambda t_i)^{n_i} e^{-\lambda t_i}}{n_i!} \right) \\ &= \sum_{i=1}^{n} (n_i(\ln(t_i) + \ln(\lambda)) - \lambda t_i - \ln(n_i!)). \end{aligned}$$

The score function is the partial derivative of the log-likelihood function with respect to the parameter:

$$
\begin{aligned}
s(\lambda) &= \frac{\partial}{\partial \lambda} \sum_{i=1}^{n} (n_i(\ln(t_i) + \ln(\lambda)) - \lambda t_i - \ln(n_i!)) \\
&= \sum_{i=1}^{n} (\frac{n_i}{\lambda} - t_i) \\
&= \frac{1}{\lambda} \sum_{i=1}^{n} n_i - \sum_{i=1}^{n} t_i
\end{aligned}
$$

By setting $s(\lambda) = 0$ and solving for λ, we obtain the maximal value estimate for λ:

$$
\hat{\lambda} = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} t_i}. \tag{9}
$$

4 Overdispersed Models

4.1 Introduction

As we discussed earlier in section 2.1, the Poisson model requires that the mean and variance are equal. While this requirement is reasonable from a theoretical standpoint, it is highly restrictive in a practical sense; in general, the mean and variance of a distribution are not dependent on one another. However, a Poisson distribution is still highly desirable because of its ease of use and extendability characteristics, so we would like to find a way to apply the Poisson model when this restriction does not hold.

As we said before, if the variance of a model is greater than we expect, then it is overdispersed, or has overdispersion. In the case of the Poisson model, overdispersion occurs when $VarN > EN$.

Note: If the model has smaller variance than we expect, we say that it is underdispersed. Underdispersion is more unusual than overdispersion, and it is also much

less problematic for our purposes: data points that are clustered tightly together are easier to deal with than data points that are very spread out. Because of this, we will not discuss underdispersion.

Let us define the parameter φ, called the overdispersion parameter. For the Poisson distribution, φ is defined by $VarN = \varphi EN$; or alternatively, $\varphi = \frac{VarN}{EN}$. Note that, for the purposes of insurance, it is reasonable to assume that $EN \geq 0$. So if we assume that $VarN > EN$ as well, then clearly $\varphi > 1$. In this way, we can use a model similar to the classical Poisson model, but we estimate parameters using a quasi-likelihood framework. The quasi-likelihood function is similar to a log-likelihood function; however, the quasi-likelihood function is not based on any specific probability distribution. Other than this, we may use the same concepts that were employed for the Poisson model above.

4.2 Causes of Overdispersion

There are several possible causes of overdispersion. We will discuss a few of them here.

In a Poisson distribution (as well as many others), we assume that all events occur independently of one another. However, if events are positively correlated (i.e., one event increases the probability of further events), then this can lead to overdispersion. To begin our demonstration of this point, we will pause briefly in our discussion of the Poisson distribution, and instead consider the Binomial distribution.

Example: Consider a set of n Bernoulli random variables: $B_1, \ldots, B_n \sim Bernoulli(p)$ (where $0 \leq p \leq 1$). (See, e.g., Tutz, 2012.) Recall that a Bernoulli random variable takes a value of 1 (or "success") with probability p, and a value of 0 (or "failure") with probability $1 - p$. If all of these variables are independent, then clearly their sum will be a Binomial random variable: $B_1 + \cdots + B_n = B \sim Bin(n, p)$.

However, what would happen if we violate the requirement of independence? If

B_i are *not* necessarily all independent, then:

$$\begin{aligned} VarB &= Var\left(\sum_{i=1}^{n} B_i\right) \\ &= \sum_{i=1}^{n} VarB_i + \sum_{i\neq j} Cov(B_i, B_j), \end{aligned}$$

where $Cov(B_i, B_j)$ is the covariance of B_i and B_j. If we assume that there is some degree of positive correlation between these variables, then $\sum_{i\neq j} Cov(B_i, B_j) > 0$, so $VarB > Var\left(\sum_{i=1}^{n} B_i\right)$.

Furthermore, we can find φ exactly in this case, as the ratio of the actual variance to the theoretical variance (ie, the case where all B_is are independent):

$$\varphi = \frac{\sum_{i=1}^{n} VarB_i + \sum_{i\neq j} Cov(B_i, B_j)}{\sum_{i=1}^{n} VarB_i} = 1 + \frac{\sum_{i\neq j} Cov(B_i, B_j)}{np(1-p)}$$

Our consideration of the correlation of events for the Poisson model works in a similar way. We recall that the sum of independent Poisson random variables is still a Poisson variable, with parameter equal to the sum of the component parameters. Example: If $N_i \sim Po(\lambda)$ are i.i.d. Poisson distributed random variables for $i = 1, \ldots, t$, then $N(t) = N_1 + \cdots + N_t$ is a Poisson process and $N(t) \sim Po(t\lambda)$, as we discussed in Section 2.2. However, if we once again assume that the individual N_is have some degree of positive correlation, then:

$$\begin{aligned} VarN(t) &= \sum_{i=1}^{t} VarN_i + \sum_{i\neq j} Cov(N_i, N_j) \\ &= t\lambda + \sum_{i\neq j} Cov(N_i, N_j) > t\lambda. \end{aligned}$$

So, $N(t)$ is overdispersed. And once again, we see that the overdispersion param-

eter is:

$$\varphi = 1 + \frac{\sum_{i \neq j} Cov(N_i, N_j)}{t\lambda}$$

Practical example: If we use a Poisson model to predict the number of car crashes that occur in one week, then we must assume that the crashes all occur independently: one car crash does not increase (or decrease) the probability of further crashes. However, in practice, one car collision may directly lead to another, and so cause a "pileup" of multiple cars; in this case, the crashes certainly do not occur independently.

Another major cause of overdispersion is unobserved heterogeneity. Normally, when using the Poisson model (or most other models), we assume that the parameter is homogenous–that is, it is constant over time. However, the parameter may be subject to change. This will change the form of the distribution, and if this change is not observed and accounted for, it can skew the results and increase the variance of the model. In some cases, the parameter itself may be a random variable; we consider this possibility in more detail in section 5.

4.3 Overdispersion in the Natural Exponential Family

The natural exponential family is a particular class of probability distributions. All distributions in the natural exponential family have probability density functions (if continuous) or probability mass functions (if discrete) that can be expressed in the following form:

$$f(z; \theta, \varphi) = \exp\left(\frac{z\theta - b(\theta)}{\varphi} + C(z; \varphi)\right). \tag{10}$$

Where z is the test value, θ is a parameter known as the canonical parameter, φ is the overdispersion parameter, and $b(\theta)$, $C(z; \varphi)$ are some functions. The natural exponential family includes many different distributions, including the Poisson distribution. It is very useful to consider in our treatment of overdispersion, because it allows us to consider many different distributions at the same time.

As a primer for the exponential family, let us consider the case of the Poisson distribution. If $N \sim Po(\lambda)$ has probability mass function $f(z)$, then:

$$\begin{aligned} f(z) &= e^{\ln(f(z))} = \exp(\ln(\frac{\lambda^{-z}}{z!}e^{-\lambda})) \\ &= \exp(\ln(e^{-\lambda}) + \ln(\lambda^z) - \ln(z!)) \\ &= \exp(z\ln(\lambda) - \lambda - \ln(z!)). \end{aligned}$$

We want to choose θ, φ, b, C so that this is in the form of equation (10). If we take $\theta = \ln(\lambda)$, then $b(\theta) = \lambda = e^{\ln(\lambda)} = e^{\theta}$. Then, our remaining term must become $C(z;\varphi) = -\ln(z!)$. Finally, we see that $\varphi = 1$ (and obviously, $\varphi = 1$ for any such distribution if no overdispersion is present).

We want to show that the φ term in equation (10) is identically equal to the overdispersion parameter for the Poisson case.

Let Y be an overdispersed natural exponential family random variable with canonical parameter θ and overdispersion parameter φ. We will endeavor to find the moment generating function of Y.

$$\begin{aligned} M_Y(t) = E(e^{tY}) &= \int_Y e^{ty} \exp\left(\frac{\theta y - b(\theta)}{\varphi} + C(y;\varphi)\right) dy \\ &= \int_Y \exp\left(\frac{\theta y + t\varphi y - b(\theta)}{\varphi} + C(y;\varphi)\right) dy \\ &= \int_Y \exp\left(\frac{(\theta + t\varphi)y - b(\theta + t\varphi) + b(\theta + t\varphi) - b(\theta)}{\varphi} + C(y;\varphi)\right) dy \\ &= \exp\left(\frac{b(\theta + t\varphi) - b(\theta)}{\varphi}\right) \int_Y \exp\left(\frac{(\theta + t\varphi)y - b(\theta + t\varphi)}{\varphi} + C(y;\varphi)\right) dy. \end{aligned}$$

We see that the integrand is equal to $f(y;\theta + t\varphi, \varphi)$. Let us assume that t is small enough so that this is a probability density (or mass) function. This is a reasonable assumption for our purposes: when calculating moments, we only need to consider

the case $t = 0$, in which case $f(y; \theta + t\varphi, \varphi)$ reduces to $f(y; \theta, \varphi)$. Thus, the value of the integral over all possible values of Y must equal 1, so we see that:

$$M_Y(t) = e^{\frac{b(\theta+t\varphi)-b(\theta)}{\varphi}}.$$

From here, straightforward calculations will show us that $EY = b'(\theta)$, and $VarY = \varphi b''(\theta)$. At the start of this section, we stated that $VarY = \varphi EY$ for the Poisson distribution; now, we see a clear generalization to any distribution in the natural exponential family.

Next, we want to consider the score function for a sum of overdispersed Poisson random variables. Let $Y_1, \cdots, Y_n$ be overdispersed Poisson variables with parameters $\lambda_1, \cdots, \lambda_n$ (or alternatively, canonical parameters $\theta_1 = \ln(\lambda_1), \cdots, \theta_n = \ln(\lambda_n)$), and common overdispersion parameter φ. Let $Y = Y_1 + \cdots + Y_n$. As stated earlier, we do not need an actual log-likelihood function; a quasi-likelihood function will suffice. The quasi-likelihood function will be the natural logarithm of the following:

$$\mathcal{L}(\lambda_1, \cdots, \lambda_n | y_1, \cdots, y_n) = \exp\left[\sum_{i=1}^{n}\left(\frac{y_i \ln(\lambda_i) - \lambda_i}{\varphi} + C(y_i; \varphi)\right)\right].$$

Taking the natural log of each side, we get the quasi-likelihood function with respect to some global parameter λ:

$$q\mathcal{L}(\lambda_1, \cdots, \lambda_n | y_1, \cdots, y_n) = \sum_{i=1}^{n}\left(\frac{y_i \ln \lambda_i - \lambda_i}{\varphi} + C(y_i; \varphi)\right).$$

To get the score function, we take the partial derivative with respect to λ:

$$
\begin{aligned}
s(\lambda) &= \frac{\partial}{\partial \lambda} \sum_{i=1}^{n} \left(\frac{y_i \ln \lambda_i - \lambda_i}{\varphi} + C(y_i; \varphi) \right) \\
&= \sum_{i=1}^{n} \frac{\partial \lambda_i}{\partial \lambda} \frac{\partial}{\partial \lambda_i} \left(\frac{y_i \ln \lambda_i - \lambda_i}{\varphi} \right) \\
&= \sum_{i=1}^{n} \frac{\partial \lambda_i}{\partial \lambda} \frac{y_i/\lambda_i - 1}{\varphi}.
\end{aligned}
$$

$C(y_i; \varphi)$ does not depend on λ, so this term vanishes from the formula. Finally, after simplifying, we get:

$$
s(\lambda) = \sum_{i=1}^{n} \frac{\partial \lambda_i}{\partial \lambda} \frac{y_i - \lambda_i}{\varphi \lambda_i}. \tag{11}
$$

Now, we assume that λ_i is defined by the global parameter λ and the time period length t_i as: $\lambda_i = \lambda t_i$. In this case, $\frac{\partial \lambda_i}{\partial \lambda} = t_i$. Plugging both of these assumptions into (11) gives us:

$$
s(\lambda) = \sum_{i=1}^{n} t_i \frac{y_i - \lambda t_i}{\varphi \lambda t_i} = \frac{1}{\varphi \lambda} \sum_{i=1}^{n} (y_i - \lambda t_i)
$$

Setting $s(\lambda) = 0$ and solving for λ, we find that the maximum likelihood estimate for λ is the same as in (9).

We also wish to find the maximum likelihood estimate for φ. By our definition, $\varphi = \frac{VarY}{EY}$; or more specifically, $\varphi = \frac{VarY_i}{EY_i}$. One particular expression for variance tells us that $VarY = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - EY_i)^2$. $EY_i = \hat{\lambda_i}$, which we can find using either equation (9) or (11). We may thus say that:

$$
\hat{\varphi} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda_i})^2}{\hat{\lambda_i}} \tag{12}
$$

Note: By convention, we use $\frac{1}{n-1}$ to compute variance when only a random sample

is known, and $\frac{1}{n}$ when the complete population is known. The data sets we will be using should be large enough that the difference here is insignificant, but we will use this form in order to be consistent.

Many real-world applications of the Poisson model will have some degree of overdispersion. However, determining how large overdispersion must be to require treatment is highly subjective. A model with $\varphi > 2$ would certainly need overdispersion treatment, but for smaller values (when φ is close to 1), the situation becomes more nebulous.

For the overdispersed Poisson model, formulae (7) and (8) become:

$$EY_i = \lambda t_i EZ, \tag{13}$$

$$VarY_i = \lambda t_i VarZ + \lambda\varphi t_i (EZ)^2 = \lambda t_i (VarZ + \varphi (EZ)^2). \tag{14}$$

Now, assume that the individual frequencies in our subportfolios are Poisson distributed, and use the overdispersed Poisson model. We would like to know how to model the total claim number for the sum of the subportfolios. If we assume that the severity distribution is independent of the overdispersion of the frequencies, then the following lemma shows that our findings for the closedness properties of the standard Poisson model also hold for the compound overdispersed Poisson model.

Lemma 4.1. *Let $N_1, \cdots, N_k$ be independent random variables that follow the overdispersed Poisson model. That is, $EN_i = \lambda_i$ and $VarN_i = \varphi_i EN_i$, $\varphi_i > 1$, for $i = 1, \cdots, k$. Then the sum of these variables $N_* = \sum_{i=1}^{k} N_i$ also follows the overdispersed Poisson model with parameter $\lambda_* = \sum_{i=1}^{k} \lambda_i$ and overdispersion parameter $\varphi_* = \frac{1}{\lambda_*} \sum_{i=1}^{k} \lambda_i \varphi_i$. If $\varphi_i = \varphi$ for all $i = 1, \cdots, k$, then $\varphi_* = \varphi$.*

Proof. We see that:

$$EN_* = E\left(\sum_{i=1}^{k} N_i\right) = \sum_{i=1}^{k} EN_i = \sum_{i=1}^{k} \lambda_i = \lambda_*.$$

For the variance, we want to show that there exists some $\varphi_* > 1$ such that $VarN_* = \varphi_* EN_*$.

$$VarN_* = Var\left(\sum_{i=1}^{k} N_i\right) = \sum_{i=1}^{k} VarN_i = \sum_{i=1}^{n} \varphi_i \lambda_i,$$

Since all N_is are independent. Now, we search for a value for φ_*.

$$VarN_* = \frac{\lambda_*}{\lambda_*} \sum_{i=1}^{k} \varphi_i \lambda_i = EN_* \frac{1}{\lambda_*} \sum_{i=1}^{k} \varphi_i \lambda_i.$$

Since $\varphi_i > 1$ for all i, $\sum_{i=1}^{k} \varphi_i \lambda_i > \sum_{i=1}^{k} \lambda_i$. Furthermore, since $\sum_{i=1}^{k} \lambda_i = \lambda_*$, we see that $\frac{1}{\lambda_*} \sum_{i=1}^{k} \varphi_i \lambda_i > 1$ always. Thus, this is an acceptable value for φ_*, and it is the value given in the lemma.

If $\varphi_i = \varphi$ for all i, then $\varphi_* = \frac{1}{\lambda_*} \sum_{i=1}^{k} \varphi \lambda_i = \varphi \frac{\lambda_*}{\lambda_*} = \varphi$. □

5 Handling Overdispersion in a More General Framework

The overdispersion model proposed above is simple to work with and easily applied, but it is not always sufficient. A stronger or more thorough treatment may be needed, particularly when the degree of overdispersion is especially large. Thus, we want to find other ways to generalize the claim process to handle overdispersion. However, we must keep in mind that any such model should, ideally, retain some of the Poisson model's closedness properties.

5.1 Mixed Poisson Model

Assume that the frequency $N_i \sim Po(\Lambda_i)$, where Λ_i is also a random variable, with $E\Lambda_i = \lambda_i$. Λ_i is known as a mixing variable, and we say that N_i follows a certain mixed Poisson distribution. (Clearly, if Λ_i is exactly equal to λ_i, then the construction becomes the standard Poisson model.) A suitable choice of Λ_i can take care of the overdispersion problem. We wish to determine whether such a model retains the additive property of the Poisson model, and if so, we want to find the conditions that are necessary for the property to hold. We can ask the same thing of the compound process.

To answer these questions, we choose a certain mixing variable Q_i such that $\Lambda_i = \lambda_i \cdot Q_i$, where Q_i is scaled such that $EQ_i = 1$. Under this setup, the conditional distribution of N_i becomes: $(N_i|Q_i = q_i) \sim Po(\lambda_i q_i)$. Once again, we want to find the expectation and variance of N_i.

The Law of Total Expectation tells us that $EN_i = E(E(N_i|Q_i))$. We see that $E(N_i|Q_i = q_i) = \lambda_i q_i$, so $E(N_i|Q_i) = \lambda_i Q_i$. Therefore,

$$EN_i = E(E(N_i|Q_i)) = E(\lambda_i Q_i) = E\lambda_i EQ_i = \lambda_i. \tag{15}$$

To find the variance, we consider the Law of Total Variance, which tells us that $VarN_i = E(Var(N_i|Q_i)) + Var(E(N_i|Q_i))$. We see that $Var(N_i|Q_i = q_i) = \lambda_i q_i$, so $Var(N_i|Q_i) = \lambda_i Q_i$. Using this, as well as the results above, we find that:

$$VarN_i = E(Var(N_i|Q_i)) + Var(E(N_i|Q_i)) = E(\lambda_i Q_i) + Var(\lambda_i Q_i) = \lambda_i + \lambda_i^2 Var(Q_i). \tag{16}$$

We see that $VarN_i \geq EN_i$. (Equality holds only for the degenerate cases $\lambda_i = 0$ or $Q_i \equiv 1$, the latter of which reduces to a simple Poisson distribution.) Thus, this model is suitable to handle overdispersion. Once again, we can find the expectation and variance of the i-th policy by substituting the above formulae into equations (3) and (4):

$$EY_i = \lambda_i EZ, \tag{17}$$

$$VarY_i = \lambda_i VarZ + (EZ)^2(\lambda_i + \lambda_i^2 VarQ_i). \tag{18}$$

Now, consider some mixed Poisson variables N_i, $i = 1, \cdots, k$, where k is arbitrary. It can be shown that their sum $N = \sum_{i=1}^{k} N_i$ is mixed Poisson-distributed if all N_i are either mutually independent, or are dependent on one another only through their mixing variables Q_i.

We should also consider compound mixed Poisson random variables. If such variables all have a common mixing variable Q (but are otherwise independent), then their sum is also a compound mixed Poisson variable, with mixing variable Q. However, in a more general framework, the sum of compound mixed Poisson variables is not necessarily a compound mixed Poisson variable.

It is also necessary to find a way to interpret this system. In practice, we might generate an unobserved realization of Λ_i, and then generate an observed realization of some Poisson random variable, which has the above realization of Λ_i as its parameter. In this way, it is reasonable to say that the parameters of the mixing distribution are found using auxiliary variables.

5.2 Negative Binomial Model

As we discussed earlier, the Negative Binomial distribution may be considered if the observed variance exceeds the expected value. It is one of the classical models used to describe claim number in a portfolio or subportfolio. We will now discuss the topic of our primary interest, which we have until now considered only for the Poisson distribution and its variants: is the sum of negative binomial variables or compound negative binomial variables also negative binomial or compound negative binomial, respectively? (See, e.g., Hilbe, 2007)

In general, the sum of negative binomial distributions $NBin(r_i, p_i)$, $i = 1, \cdots, k$ (k arbitrary) is not a negative binomial distribution. If all parameters p_i are equal (ie, if $p_i \equiv p$), then the sum will, in fact, be a negative binomial distribution, as we showed this in section 3. However, this does not necessarily hold when the p parameters are different. Consider $X \sim NB(r_1, p_1)$, $Y \sim NB(r_2, p_2)$, where

$p_1 \neq p_2$ (but r_1, r_2 may or may not be unequal). If $X + Y \sim NB(\dot{)}$, then it must have probability density function:

$$P(X+Y=z) = \binom{z + f(r_1, r_2) - 1}{z} g(p_1, p_2)^{f(r_1,r_2)} (1 - g(p_1, p_2))^z,$$

For some functions f and g. Once again, we consider the convolution of these distributions.

$$\begin{aligned} P(X+Y=z) &= \sum_{k=0}^{z} P(X=k)P(Y=z-k) \\ &= \sum_{k=0}^{z} \binom{k+r_1-1}{k} p_1^{r_1}(1-p_1)^k \binom{z-k+r_2-1}{z-k} p_2^{r_2}(1-p_2)^{z-k} \\ &= (p_1)^{r_1}(p_2)^{r_2} \sum_{k=0}^{z} \binom{k+r_1-1}{k} \binom{z-k+r_2-1}{z-k} (1-p_1)^k (1-p_2)^{z-k}. \end{aligned}$$

Observe that r_1 and r_2 no longer appear in the exponents in the summation. Therefore, $g(p_1, p_2)^{f(r_1,r_2)}$ must be in the form $p_1^{r_1+c_1} p_2^{r_2+c_2}$, where c_1, c_2 are constants that do not rely on our parameters. We may rewrite this term as $(p_1^{(r_1+c_1)/(r_2+c_2)} p_2)^{r_2+c_2}$, giving us our formulae for $f(r_1, r_2)$ and $g(p_1, p_2)$.

Note that the $(1-p_1)^k(1-p_2)^{z-k}$ term in the summation resembles a binomial expansion. Thus, its contribution to the final formula should be similar to $((1-p_1)+(1-p_2))^z = (2-p_1-p_2)^z$, which is not in the form of our $g(p_1, p_2)$. Therefore, $X + Y$ cannot take a Negative Binomial distribution when the probability parameters are different. A simple inductive argument will show that the same holds for the sum of an arbitrary number of Negative Binomial RVs, when at least two of the probability parameters are unequal.

Next, we shall consider a compound Negative Binomial distribution. As with the compound Poisson distribution, let $N \sim NB(r, p)$, and $Z_1, Z_2, \ldots$ be i.i.d. random variables with some given distribution, that are also independent of N. Then, we will say that $S = \sum_{i=1}^{N} Z_i$ has a compound Negative Binomial distribution.

As with the compound Poisson model, we can describe the number of claims in time t as $N(t)$. We know that $EN = \frac{pr}{1-p}$, so $EN(t) = t\frac{pr}{1-p}$. Using equations (1) and (2), we find that $ES(t) = EN(t)EZ = t\frac{pr}{1-p}\lambda$ and $VarS(t) = t\lambda E(Z^2)$, where $\lambda = EZ$.

We also want to consider the sum of compound Negative Binomial distributions. Let S_1, S_2 be compound Negative Binomial distributions, having corresponding claim number distributions $N_1 \sim NB(r_1, p_1)$, $N_2 \sim NB(r_2, p_2)$. Let S_1, S_2 have i.i.d. claim sizes. If $p_1 = p_2 = p$, then we see that $S = S_1 + S_2$ will have a corresponding claim number distribution $N = N_1 + N_2 \sim NB(r_1 + r_2, p)$. We can easily extend this finding inductively: if $S_1, \cdots, S_n$ are compound Negative Binomial distributions with i.i.d. individual claim sizes and claim numbers $N_i \sim NB(r_i, p)$, $i = 1, \cdots, n$, then $S = S_1 + \cdots + S_n$ will be a compound Negative Binomial distribution with corresponding $N \sim NB(r_1 + \cdots + r_n, p)$. Of course, as we found above, the summation does not hold if not all p parameters are equal.

Alternatively to the above, let us consider the cumulant-generating function (cgf) $\psi_X(t)$, defined as the natural log of $M_X(t)$. The cgf is useful to consider here because of its additivity properties. If X and Y are independent random variables with cgfs $\psi_X(t)$ and $\psi_Y(t)$, the cgf of $X + Y$ is $\psi_{X+Y}(t) = \psi_X(t) + \psi_Y(t)$.

For $N \sim NB(r, p)$, the moment-generating function $M_N(t) = \left(\frac{(1-p)e^t}{1-pe^t}\right)^r$. Let us consider $N_1 \sim NB(r_1, p_1)$ and $N_2 \sim NB(r_2, p_2)$. These will have cumulant-generating functions $\psi_{N_1}(t) = \ln\left[\left(\frac{(1-p_1)e^t}{1-p_1e^t}\right)^{r_1}\right]$ and $\psi_{N_2}(t) = \ln\left[\left(\frac{(1-p_2)e^t}{1-p_2e^t}\right)^{r_2}\right]$, respectively. Then, their sum will be:

$$\begin{aligned}\psi_{N_1}(t) + \psi_{N_2}(t) &= \psi_{N_1+N_2}(t) \\ &= \ln\left[\left(\frac{(1-p_1)e^t}{1-p_1e^t}\right)^{r_1}\left(\frac{(1-p_2)e^t}{1-p_2e^t}\right)^{r_2}\right].\end{aligned}$$

By exponentiating, we can get the moment-generating function for $N_1 + N_2$. In general, this expression cannot be simplified into a form that is consistent with the moment generating function (MGF) for a Negative Binomial distribution. Next, we want to consider special cases.

If $p_1 = p_2 = p$, then this formula simplifies to:

$$\psi_{N_1+N_2}(t) = \ln\left[\frac{(1-p)^{r_1+r_2}e^{t(r_1+r_2)}}{(1-pe^t)^{r_1+r_2}}\right] = \ln\left[\left(\frac{(1-p)e^t}{1-pe^t}\right)^{r_1+r_2}\right].$$

Thus, we clearly see that $(N_1+N_2) \sim NB(r_1+r_2, p)$. A simple inductive argument shows that if $N_i \sim NB(r_i, p)$, then $(N_1 + \cdots + N_n) \sim NB(r_1 + \cdots + r_n, p)$.

If $r_1 = r_2 = r$, then $\psi_{N_1+N_2}(t) = \ln\left[\left(\frac{(1-p_1)(1-p_2)e^{2t}}{(1-p_1e^t)(1-p_2e^t)}\right)^r\right]$. Once again, we cannot simplify this into a form corresponding to an MGF for a Negative Binomial distribution. In general, the sum of Negative Binomial random variables is a Negative Binomial only if all p parameters are equal.

Furthermore, we will consider a compound Negative Binomial model. This model is constructed in a way similar to the compound Poisson model: let $Z_1, \cdots, Z_N$ be i.i.d. random variables with a given distribution (for our purposes, these represent individual claim sizes), and let $N \sim NB(r, p)$ (this represents the number of claims). Then, we say that $S = \sum_{i=1}^{N} Z_i$ has a compound Negative Binomial distribution. We recall that the MGF of S (see formula (21) in Appendix A) is:

$$M_S(t) = M_N(\ln M_Z(t)) = \left(\frac{(1-p)e^{\ln M_Z(t)}}{1-pe^{\ln M_Z(t)}}\right)^r = \left(\frac{(1-p)M_Z(t)}{1-pM_Z(t)}\right)^r.$$

We see that the form for S is the same as the form for N, except that the e^t terms have been replaced with $M_Z(t)$. Therefore, it is clear that all of the additivity properties we just proved for the Negative Binomial model also hold for the compound Negative Binomial model.

We may represent the Negative Binomial distribution as a Poisson mixture with a Gamma-distributed mixing variable. Because of this, all of the properties we discussed for the mixed Poisson model apply to the Negative Binomial model. In particular, if we mix the Poisson distribution with $\Lambda_i \sim \Gamma(\alpha_i, \frac{\alpha_i}{\lambda_i})$ (or, written differently, with normalized mixing variable $Q_i \sim \Gamma(\alpha_i, \alpha_i)$), then we produce a

Negative Binomial random variable $N_i \sim NB(\alpha_i, \frac{\lambda_i}{\alpha_i+\lambda_i})$. Using formulae (17) and (18), we see that this distribution has expectation and variance:

$$EN_i = \lambda_i \tag{19}$$

$$VarN_i = \lambda_i + \frac{\lambda_i^2}{\alpha_i} \tag{20}$$

6 Practical Applications of the Overdispersed Poisson Model

Testing the findings of this paper using actual insurance data is infeasible, due to the difficulty in attaining this data. However, it is also possible to use semi-random computer generated data sets to study our results. In some ways, this method is preferable, since we already know the theoretical distribution and its parameters.

Simulation 1: Parameters of the overdispersed Poisson model

Assumptions

One way to simulate the overdispersed Poisson model is to use a Negative Binomial model. Recall that, for an overdispersed Poisson variable N, $EN = \lambda$, and $VarN = \varphi EN$. For a Negative Binomial variable N, $EN = \frac{pr}{1-p}$, and $VarN = \frac{pr}{(1-p)^2}$. Elementary algebra shows that an overdispersed Poisson model with parameter λ and overdispersion parameter φ can be modeled as a Negative Binomial random variable, with parameters $p = \frac{1}{\varphi}$, $r = \frac{\lambda p}{1-p}$.

Test description

Using a language such as R, we may write a function to generate overdispersed Poisson random numbers for given λ, φ. This function is included in Appendix B. By generating a sufficiently large dataset and finding its mean and variance, we can approximate the initial parameters, and test these parameters against the actual

values. We will do this using several different values of φ, generating $n = 10000$ random numbers for each trial.

In addition, we wanted to test the effectiveness of several different models: namely, the overdispersed Poisson, Negative Binomial, and standard Poisson, using different overdispersion parameters. Note that $\varphi = 1$ reduces to the standard Poisson; and furthermore, it is undefined for the Negative Binomial model. Thus, we used the values $\varphi = 1.01, 2, 3, 5, 10$.

Results, Interpretation, Explanation

For the Overdispersed Poisson model:

φ	1.01	2	3	5	10
$\hat{\lambda}$	0.9860	0.9800	1.0061	1.0240	1.0294
$\hat{\varphi}$	0.9999012	2.0224471	3.0271020	4.9452445	10.3942157

Table 1: Estimated parameters $\hat{\lambda}$, $\hat{\varphi}$ for overdispersed Poisson model

The estimates for $\hat{\lambda}$ remain at a fairly steady level of accuracy for all cases, with a margin of error of $2 - 3\%$. The absolute size of the error of $\hat{\varphi}$ increases as φ increases, but the relative difference stays about the same.

For the standard Poisson model:

φ	1.01	2	3	5	10
$\hat{\lambda}$	1.0160	0.9924	1.0010	1.0171	1.0012
$\hat{\varphi}$	1.0085103	1.0151582	0.9799172	0.9903729	1.0020964

Table 2: Estimated parameters $\hat{\lambda}$, $\hat{\varphi}$ for standard Poisson model

Estimates of $\hat{\lambda}$ are all consistent, although they may be slightly more accurate than for the overdispersed Poisson. However, not surprisingly, this model completely fails to account for overdispersion parameters. The model appears to be reasonably close when $\varphi = 1.01$, but all values after that are unacceptable.

For the Negative Binomial model:

Simulation 2: Additivity property of the overdispersed Poisson model

φ	1.01	2	3	5	10
$\hat{\lambda}$	1.0105	1.0125	0.9630	0.9769	0.9640
$\hat{\varphi}$	1.015430	1.914605	2.926389	4.874237	9.544631

Table 3: Estimated parameters $\hat{\lambda}$, $\hat{\varphi}$ for Negative Binomial model

Assumptions

We want to test Lemma 4.1, the additivity property of the overdispersed Poisson model. We assume that $\lambda_1, \ldots, \lambda_k$ are independent and follow the overdispersed Poisson model, with parameter pairs $(\lambda_1, \varphi_1), \ldots, (\lambda_k, \varphi_k)$.

Test description

We will test two different sets of overdispersed Poisson variables. First, we will test the variables with parameter pairs $(1, 2), (2, 2), (3, 3), (4, 3)$. According to Lemma 4.1, the sum of these variables should have $\lambda = 10$ and $\varphi = 2.7$.

Second, we will test the variables with parameters $(7, 7), (2, 5), (8, 3), (9, 6)$. We expect that $\lambda = 26$ and $\varphi = 5.2692$.

For each of these tests, we will generate 10000 instances, and estimate their parameters.

Results, Interpretation

For the first test, we found $\hat{\lambda} = 9.9022$, $\hat{\varphi} = 2.707289$. $\hat{\lambda}$ is accurate to within 0.1% error, and $\hat{\varphi}$ is accurate to within 5% error, both of which are acceptable.

In a second test, we got $\hat{\lambda} = 26.08$, $\hat{\varphi} = 5.3050$. $\hat{\lambda}$ was slightly less accurate than the first test, but $\hat{\varphi}$ was somewhat more accurate. For sufficiently large n, it seems that we can expect the sum of variables following the overdispersed Poisson model to have the additivity properties described in Lemma 4.1.

Simulation 3: Accuracy of the parameter values in the overdispersed Poisson model

Assumptions

Choosing a value of n (the number of random numbers generated) and estimating

parameters based on that makes it difficult to calibrate the degree of accuracy of the parameter values. We would like to find a way to calculate λ and φ such that both the mean and variance are within an acceptable margin of error. One of the simplest ways to do this is to generate successively more values, and adjust the mean and variance estimates with each additional batch of random numbers. In this case, n is the total number of values generated. Two different codes for this goal are included in Appendix B.

Test description

We want to see how many values need to be generated in order to get the estimated expected value and variance to within acceptable limits. We will set $\lambda = 1$, $\varphi = 2$, and then test how many values need to be generated for the estimates to have maximum error 10%, 1%, and 0.1%. 50 trials will be run in each case, and the mean, variance, and minimum and maximum values of n will be recorded.

Results

For podcheck1:

Max error	10%	1%	0.1%
Mean of n	476	19074	731268
Var(n)	$4.53 * 10^5$	$3.65 * 10^9$	$3.84 * 10^{12}$
Min(n)	100	300	3400
Max(n)	3900	411700	$1.316 * 10^7$

Table 4: Descriptive statistics for number of values generated with function podcheck1

For podcheck2:

Max error	10%	1%	0.1%
Mean of n	416	26402	601526
Var(n)	$2.07 * 10^5$	$5.37 * 10^9$	$1.70 * 10^{12}$
Min(n)	100	900	2900
Max(n)	2900	502200	6341100

Table 5: Descriptive statistics for number of values generated with function podcheck2

Interpretation, Explanation

There does not appear to be any significant difference in the efficiency of these two codes. The most striking thing about these results is the enormous variability in n. With acceptable error 10%, the minimum n was 100 for both functions, which is the number of values generated in each cycle, and thus the smallest possible n. For error 10%, the maximum and minimum value vary by an order of magnitude; for 1% and 0.1%, the maximum and minimum vary by at least $2-3$ orders of magnitude. The minimum n for error 0.1% is comparable to the maximum n for error 10% in both cases. Until this process is better understood, it appears that this is a rather poor way of estimating necessary sample sizes, due to the extreme degree of variation.

Simulation 4: Compound overdispersed Poisson model

Assumptions

Finally, we would like to test the validity of the compound overdispersed Poisson model. Using equations (13) and (14), we can predict EY_i and $VarY_i$, the expected value and variance of the ith policy in a portfolio. We can then write a code that computes these values experimentally, using a given distribution Z of insurance claim sizes. In this way, we can see how the accuracy of the model increases with the number of policies n.

For these tests, we will not consider policies with different durations, since when EZ and $VarZ$ are known and λ, φ are fixed, then EY_i and $VarY_i$ are linearly proportional to t_i. Therefore, the mean and variance will vary across policies in a way that is strictly predictable.

Test description

First, we will choose a model for insurance claim amounts. We will consider the Uniform, Lognormal, Normal, and Gamma distributions, and use two different sets of parameters for each. All parameters will be chosen so that the average claim amount $EZ = 1000$, in order to make comparison between models simpler.

Note: While the Normal distribution can take negative values, we will choose our

parameters so that this possibility is sufficiently improbable as to be ignored.

For each model, we will use $m = 50$ runs per test. We will consider $\varphi = 1.01$, 2, 3, and $n = 100$, 1000, 10000. In all cases, we will take the rate parameter $\lambda = 1$. For each test, we will find:

- The estimate of the policy mean and variance (denoted $\widehat{EY}$, $\widehat{VarY}$),
- The average absolute error of the expected value of the policy claim amount $\frac{1}{m}\sum_{i=1}^{m}|\widehat{EY} - EY|$ (denoted AD)
- The average absolute error of the above in the cases when $\widehat{EY} < EY$ and the number of trials for which this is the case (denoted AD_- and m_- respectively), and the same for the cases when $\widehat{EY} > EY$ (denoted AD_+ and m_+, respectively).

Results

Test 1: $Z \sim U(500, 1500)$

n	φ	$\widehat{EY}$	$\widehat{VarY}$	AD	AD_-	m_-	AD_+	m_+
100	1.01	992.556	1079596.301	24.739	26.819	30	21.618	20
	2	1006.172	2109961.582	22.920	22.036	19	23.461	31
	3	1002.576	3103795.872	26.965	25.405	24	28.404	26
1000	1.01	999.177	1091868.621	6.916	7.442	26	6.347	24
	2	999.566	2081527.638	6.086	6.269	26	5.888	24
	3	999.604	3080988.066	6.230	5.917	28	6.629	22
10000	1.01	999.946	1093331.042	2.550	2.504	26	2.599	24
	2	1000.341	2084822.417	2.278	2.105	23	2.425	27
	3	1000.355	3085682.546	2.241	2.050	23	2.403	27

Table 6: Compound overdispersed Poisson model with uniformly distributed claims.

Test 2: $Z \sim U(900, 1100)$

n	φ	$\widehat{EY}$	$\widehat{VarY}$	AD	AD_-	m_-	AD_+	m_+
100	1.01	1001.16	1015694.19	4.19	3.30	23	4.95	27
	2	1000.30	2004600.30	5.49	5.19	25	5.79	25
	3	999.62	3001115.52	4.38	4.11	29	4.76	21
1000	1.01	999.85	1013021.94	1.48	1.51	27	1.44	23
	2	1000.09	2003714.31	1.42	1.33	25	1.52	25
	3	999.88	3002615.93	1.45	1.31	30	1.67	20
10000	1.01	999.98	1013293.26	0.46	0.47	26	0.46	24
	2	1000.03	2003437.12	0.48	0.44	26	0.53	24
	3	1000.02	3003424.55	0.45	0.42	26	0.48	24

Table 7: Compound overdispersed Poisson model with uniformly distributed claims.

Test 3: $Z \sim LN(1, 3.43737)$

For the Lognormal distribution with parameters mu, $sigma$, $EZ = e^{\mu + \frac{\sigma^2}{2}}$. Thus, setting $EZ = 1000$ and choosing a value of μ, we can find the corresponding σ.

n	φ	$\widehat{EY}$	$\widehat{VarY}$	AD	AD_-	m_-	AD_+	m_+
100	1.01	423.36	19183009.52	684.37	716.48	44	448.86	6
	2	1466.22	3234237516.86	1640.45	667.18	44	8777.77	6
	3	988.08	559588677.70	1165.54	717.96	41	3204.51	9
1000	1.01	701.86	397267858.15	495.06	521.84	38	410.26	12
	2	1055.38	1189416265.59	661.09	432.65	35	1194.12	15
	3	963.65	1090452568.81	748.12	530.04	37	1368.81	13
10000	1.01	833.39	1303942019.75	340.75	309.36	41	483.73	9
	2	886.35	4245883236.55	400.74	313.65	41	797.48	9
	3	913.72	5567767010.65	371.12	265.92	43	1017.31	7

Table 8: Compound overdispersed Poisson model with lognormally distributed claims.

Test 4: $Z \sim LN(4, 2.411537)$

n	φ	$\widehat{EY}$	$\widehat{VarY}$	AD	AD_-	m_-	AD_+	m_+
100	1.01	974.29	58719500.68	488.75	401.92	32	643.12	18
	2	831.02	29138802.73	443.19	364.38	42	856.92	8
	3	906.70	55869217.62	502.84	413.98	36	731.32	14
1000	1.01	1060.47	272357960.27	321.04	186.12	35	635.87	15
	2	1107.62	528098341.34	374.33	215.08	31	634.15	19
	3	960.29	66877682.60	200.15	181.71	33	235.96	17
10000	1.01	1005.22	227688061.45	95.68	90.46	25	100.91	25
	2	957.81	103523985.07	94.53	89.94	38	109.06	12
	3	986.52	155224773.48	87.15	81.15	31	96.94	19

Table 9: Compound overdispersed Poisson model with lognormally distributed claims.

Test 5: $Z \sim N(1000, 100)$

n	φ	$\widehat{EY}$	$\widehat{VarY}$	AD	AD_-	m_-	AD_+	m_+
100	1.01	999.06	1018183.71	7.18	7.51	27	6.78	23
	2	999.15	2006699.08	8.07	7.68	29	8.60	21
	3	1000.39	3012376.17	7.53	6.38	28	9.00	22
1000	1.01	1000.56	1021120.10	2.84	2.59	22	3.04	28
	2	1000.19	2010835.66	2.48	2.39	24	2.57	26
	3	999.86	3009189.38	2.57	2.81	24	2.34	26
10000	1.01	999.84	1019688.96	0.78	0.73	32	0.85	18
	2	999.85	2009410.02	0.74	0.79	28	0.68	22
	3	1000.09	3010548.16	0.92	0.86	24	0.97	26

Table 10: Compound overdispersed Poisson model with normally distributed claims.

Test 6: $Z \sim N(1000, 200)$

n	φ	$\widehat{EY}$	$\widehat{VarY}$	AD	AD_-	m_-	AD_+	m_+
100	1.01	1001.81	1054584.49	15.31	13.50	25	17.13	25
	2	998.46	2033977.59	15.72	15.41	28	16.12	22
	3	1003.65	3063658.92	15.75	13.15	23	17.97	27
1000	1.01	1001.96	1054178.32	4.01	3.65	14	4.15	36
	2	999.12	2036389.35	4.30	4.79	27	3.72	23
	3	999.67	3037805.07	4.77	5.09	25	4.44	25
10000	1.01	1000.23	1050569.46	1.45	1.32	23	1.55	27
	2	1000.38	2041642.57	1.62	1.18	26	2.09	24
	3	1000.38	3042373.00	1.81	1.69	21	1.89	29

Table 11: Compound overdispersed Poisson model with normally distributed claims.

Test 7: $Z \sim \Gamma(20, 50)$

n	φ	$\widehat{EY}$	$\widehat{VarY}$	AD	AD_-	m_-	AD_+	m_+
100	1.01	999.55	1058903.10	14.12	16.56	22	12.21	28
	2	998.96	2044765.42	14.95	15.37	26	14.50	24
	3	1000.43	3053696.86	17.32	17.59	24	17.07	26
1000	1.01	1001.77	1064096.47	5.60	4.35	22	6.58	28
	2	998.95	2046230.25	4.55	4.37	32	4.87	18
	3	999.08	3044202.52	6.11	6.75	26	5.41	24
10000	1.01	1000.17	1060289.89	1.67	1.63	23	1.71	27
	2	1000.05	2050205.38	1.67	1.68	24	1.66	26
	3	1000.54	3053333.79	1.66	1.33	21	1.90	29

Table 12: Compound overdispersed Poisson model with Gamma distributed claims.

Test 8: $Z \sim \Gamma(40, 25)$

n	φ	$\widehat{EY}$	$\widehat{VarY}$	AD	AD_-	m_-	AD_+	m_+
100	1.01	998.61	1032810.55	12.69	13.04	27	12.29	23
	2	1000.60	2027749.67	11.23	9.84	27	12.86	23
	3	996.21	3002873.15	14.73	14.46	32	15.19	18
1000	1.01	999.70	1034620.35	3.71	4.36	23	3.16	27
	2	999.44	2022680.20	3.85	3.55	31	4.33	19
	3	998.78	3017705.92	3.53	4.10	29	2.80	21
10000	1.01	999.54	1034041.47	1.53	1.50	33	1.59	17
	2	1000.24	2025927.39	1.49	1.36	23	1.60	27
	3	999.65	3023012.65	1.36	1.52	28	1.15	22

Table 13: Compound overdispersed Poisson model with Gamma distributed claims.

Interpretation, Explanation

For the Uniform distribution $Z \sim U(a, b)$, $\widehat{EY}$ was close to the theoretical value $EY = 1000$, with a maximum error of less than 1%. The accuracy of $\widehat{EY}$ improved as n increased. For the variance, $VarZ = \frac{(b-a)^2}{12}$, so $VarY = \frac{(a-b)^2}{12} + 1000^2 * \varphi$. As such, we expected variances 1093333, 2083333, 3083333 for $\varphi = 1.01$, 2, 3, for the first set of parameters, and 1013333, 2003333, 3003333 for the second set. The calculated values were very close to these, increasing in accuracy as n increased.

The absolute error terms decreased roughly by a factor of 3 when n was increased by a factor of 10. This seems to indicate a practical lower bound for absolute error sizes based on portfolio size.

The results were similar when testing the Gamma- and Normally-distributed claims. The only troublesome case was for the Lognormally-distributed claims. $\widehat{EY}$ and $\widehat{VarY}$ varied wildly across tests, and accuracy improved only slightly as n increased. Average absolute error was extremely large, even for the largest values of n. Such high variability of the results can be explained by the huge variance of the proposed Lognormal models. In conclusion, while the compound overdispersed Poisson model appears to have held quite well for the other distributions tested, the Lognormal distribution with given parameters is not a viable choice.

Ülehajuvusega mudelid kahjude arvu jaotuse kirjeldamiseks

Magistritöö (30 EAP)

Frazier Carsten

Kokkuvõte

Käesoleva lõputöö peaeesmärk on uurida ülehajuvusega mudeleid kindlustusportfelli kogukahju hindamiseks. Töö on jagatud kuueks osaks. Esimeses osas tutvustame lühidalt töö uurimisvaldkonda ja probleeme. Töö teises osas tuletame ja defineerime kõigepealt klassikalise kollektiivmudeli ning tuletame meelde tema olulisemad omadused. Seejärel keskendume Poissoni liitjaotuse mudelile, mis oma mitmete heade omaduste tõttu on üks sagedamini kasutatavaid mudeleid.

Töö kolmandas osas üldistame Poissoni liitjaotuse mudelit, lubades (erinevalt klassikalisest definitsioonist) poliisidele ka erineva pikkusega kindlustusperioode. Näitame ka, kuidas avalduvad suurima tõepära hinnangud kirjeldatud mudeli parameetritele.

Neljas osa keskendub ülehajuvuse probleemile. Alustame ülehajuvuse definitsiooniga ja võimalike tekkepõhjustega, uurime ülehajuvust tõenäosusjaotuse eksponentsiaalses peres ning näitame, kuidas hinnata vastava mudeli parameetreid. Lõpuks näitame, et klassikalise Poissoni mudeli aditiivsuse omadused jäävad kehtima ka ülehajuvusega Poissoni mudeli korral.

Viies osa tutvustab lühidalt üldisi võimalusi ülehajuvuse käsitlemiseks. Tutvustame Poissoni segamudeleid, millest tuntuim on negatiivse binoomjaotuse mudel, ja analüüsime, kas ja kuivõrd Poissoni mudeli head omadused nende mudelite korral kehtima jäävad.

Magistritöö kuuendas osas uurime teoreetilises osas defineeritud mudelite käitumist simuleeritud andmete peal. Esiteks pakume välja ühe võimaluse, kuidas negatiivse binoomjaotuse abil genereerida juhuslikke väärtusi etteantud parameetritega ülehajuvusega Poissoni mudelist. Teises simulatsioonis vaatleme ülehajuvusega Poissoni mudeli aditiivsuse omadust. Seejärel uurime, milline on piisav valimi suurus ülehajuvusega Poissoni mudeli parameetrite arvutamiseks etteantud täpsusega. Lõpetuseks analüüsime ülehajuvusega Poissoni liitjaotusega mudeli käitumist erinevate üksikkahjude jaotuste korral.

Töö lisades on toodud mõnede lemmade tõestused ja töö praktilises osas kasutatud programmide tekstid.

References

1. Gray, R.J. & Pitts, S.M. (2012) *Risk Modelling in General Insurance. From Principles to Practice.* Cambridge University Press.

2. Hilbe, Joseph M. (2007) *Negative Binomial Regression.* Cambridge University Press, New York.

3. Käärik, M. & Kaasik, A. (2012) On Premium Estimation Using the C&RT/Poisson Model and its Extensions. *Lithuanian Journal of Statistics*, **51**(1), 5-13.

4. Quaintance, J. (2010) Combinatorial Identities: Table I: Intermediate Techniques for Summing Finite Series. (1.78), (6.1)

5. Tutz, G. (2012) Regression for Categorical Data. New York: Cambridge University Press. pp. 133-134.

Appendices

A Proofs

Proof of equations (1) and (2): The function

$$M_S(t) = E(e^{tS}),$$

Is called the Moment Generating Function (MGF) of random variable S. The moments of S are determined by:

$$E(S^n) = M_S^{(n)}(0),$$

Where $M_S^{(n)}$ is the nth derivative of the function. First, we need to find a way to express $M_S(t)$ in terms of M_N and M_Z.

We know that $S = \sum_{j=1}^{N} Z_j$. The Law of Total Expectation tells us that $E(e^{tS}) = E(E(e^{tS}|N))$, where $E(A|B)$ is the conditional expectation of A with respect to B. Using our definition of S in terms of Z_j, we see that:

$$\begin{aligned} E(E(e^{tS}|N)) &= E(E(e^{tZ})^N) \\ &= E(M_Z(t)^N) \\ &= E(e^{\ln M_Z(t) N}) \\ &= M_N(\ln M_Z(t)). \end{aligned}$$

Finally,

$$M_S(t) = M_N(\ln M_Z(t)). \tag{21}$$

To derive equation (1), we take the first derivative of the $M_S(t)$ using the chain rule: $M_S^{(1)}(t) = M_N^{(1)}(\ln M_Z(t)) \cdot \frac{M_Z^{(1)}(t)}{M_Z(t)}$. We consider the value of the derivative at

$t = 0$. Note that $M_X(0) = E(X^0) = E(1) = 1$, so:

$$\begin{aligned} M_S^{(1)}(0) &= M_N^1(\ln 1) \cdot \frac{M_Z^{(1)}(0)}{1} \\ &= M_N^{(1)}(0) \cdot M_Z^{(1)}(0) \\ &= EN \cdot EZ. \end{aligned}$$

To derive (2), we take the second derivative of $M_S(t)$. This becomes:

$$M_S^{(2)}(t) = M_N^{(2)}(\ln M_Z(t)) \cdot \frac{M_Z^{(1)}(t)^2}{M_Z(t)^2} + M_N^{(1)}(\ln M_Z(t)) \cdot \frac{M_Z(t) \cdot M_Z^{(2)}(t) - M_Z^{(1)}(t)^2}{M_Z(t)^2}$$

Taking $t = 0$ once again, we find that $E(S^2) = E(N^2) \cdot (EZ)^2 + EN \cdot VarZ$. Using the formula for variance $VarX = E(X^2) - (EX)^2$, we find that:

$$VarS = EN \cdot VarZ + (EZ)^2 \cdot VarN.$$

Proof of Proposition 2.1: We will prove the proposition by induction. For $n = 1$, the solution is trivial: $N = N_1$ is Poisson distributed with parameter $\lambda = \lambda_1$. For $n = 2$, consider two independent Poisson random variables N_1 and N_2, with parameters λ_1, λ_2, and let $N = N_1 + N_2$. We construct an argument based on convolutions: If $N = z$, then we can say that $N_1 = k$ and $N_2 = z - k$, for some appropriately chosen value(s) of k. Since N_1, N_2 must be nonnegative integers, then clearly k may take any integer value from 0 to z. Thus, the convolution of

these distributions is:

$$\begin{aligned}
\mathbf{P}(N = z) &= \mathbf{P}(N_1 + N_2 = z) \\
&= \sum_{k=0}^{z} \mathbf{P}(N_1 = k, N_2 = z - k) \\
&= \sum_{k=0}^{z} \mathbf{P}(N_1 = k)\mathbf{P}(N_2 = z - k) \\
&= \sum_{k=0}^{z} \frac{e^{-\lambda_1}(\lambda_1)^k}{k!} \cdot \frac{e^{-\lambda_2}(\lambda_2)^{z-k}}{(z-k)!} \\
&= e^{-(\lambda_1+\lambda_2)} \cdot \sum_{k=0}^{z} \frac{\lambda_1^k \lambda_2^{z-k}}{k!(z-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{z!} \cdot \sum_{k=0}^{z} \frac{\lambda_1^k \lambda_2^{z-k} z!}{k!(z-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{z!} \cdot \sum_{k=0}^{z} \binom{z}{k} \lambda_1^k \lambda_2^{z-k}
\end{aligned}$$

We recognize the sum as the binomial expansion of $(\lambda_1 + \lambda_2)^z$, so at last we get:

$$\mathbf{P}(N = z) = \frac{((\lambda_1 + \lambda_2))^z e^{-(\lambda_1+\lambda_2)}}{z!}$$

This is clearly a Poisson random variable with intensity $\lambda_1 + \lambda_2$. Thus, the sum of two independent Poisson random variables is still a Poisson random variable, with parameter equal to the sum of the component parameters. As an inductive assumption for $n = m$, suppose that $N_1, \ldots, N_m$ are independent Poisson random variables with parameters $\lambda_1, \ldots, \lambda_m$, and that $N = N_1 + \cdots + N_m$ is a Poisson random variable with parameter $\lambda_1 + \cdots + \lambda_m$. Then, for $n = m + 1$, $N' = N_1 + \cdots + N_m + N_{m+1} = N + N_{m+1}$ (due to our inductive assumption). This is clearly still a Poisson random variable, with parameter $\lambda' = \lambda_1 + \cdots + \lambda_{m+1}$, because we already proved the case for two such variables.

B Program codes

All programs have been written in R.

Function for generating n overdispersed Poisson variables using Negative Binomial random numbers:

```
rpois.od=function(n,lambda,phi){
        p=1/phi
        alpha=lambda*p/(1-p)
        #We determined these values for p and alpha algebraically.
        r=rnbinom(n,prob=p,size=alpha)
        return(r)
}
```

We use this function frequently in the practical section of this thesis.

Functions for generating mean and variance to within a given accuracy:

```
podcheck1=function(lambda,phi,meanerror,varerror){
        n0=100 #How many values we compute each cycle.
        n=0
        meanerr=1 #We set this number to force the cycle to start.
        sum_Z=0
        sum_Z2=0
        while(abs(meanerr)>meanerror){
                varerr=1 #We set this number to force the cycle to start.
                while(abs(varerr)>varerror){
                        n=n+n0 #The total number of values generated.
                        Z=rpois.od(n0,lambda,phi)
                        sum_Z=sum_Z+sum(Z)
                        sum_Z2=sum_Z2+sum(Z**2)
                        variance=sum_Z2/(n-1)-(sum_Z/n)**2
                        varerr=phi*lambda-variance
                }
                meanerr=lambda-sum_Z/n
```

```
        }
        phi=variance/(sum_Z/n)
        return(c(sum_Z/n,variance,phi,meanerr,varerr,n))
}
```

This function generates 100 overdispersed Poisson values at a time. First, it generates enough values to get the error of the variance to within an acceptable margin of error, and then generates additional values to do the same for the mean. It then checks whether the variance is still within acceptable limits given the additional values generated. If both mean and variance are acceptable, it returns these values, as well as φ, the final error for mean and variance, and the total number of values generated.

```
podcheck2=function(lambda,phi,meanerror,varerror){
        n0=100 #How many values we compute each cycle
        n=0
        varerr=1 #We set this number to force the cycle to start.
        sum_Z=0
        sum_Z2=0
        while(abs(varerr)>varerror){
                meanerr=1 #We set this number to force
                #the cycle to start
                while(abs(meanerr)>meanerror){
                        n=n+n0 #The total number of values generated.
                        Z=rpois.od(n0,lambda,phi)
                        sum_Z=sum_Z+sum(Z)
                        sum_Z2=sum_Z2+sum(Z**2)
                        meanerr=lambda-sum_Z/n
                }
                variance=sum_Z2/(n-1)-(sum_Z/n)**2
                varerr=phi*lambda-variance
        }
        phi=variance/(sum_Z/n)
        return(c(sum_Z/n,variance,phi,meanerr,varerr,n))
}
```

This code works the same way as the code above; however, it first calculates the mean, and then the variance.

Lastly, the following code generated the values used in Simulation 4.

```
compound_pod=function(n,phi,randtype,param1,param2){
        lambda=1#We can adjust the time period so that this is true.
        sumclaimmean=0
        #The sum of the average portfolio claim sizes
        sumabsdiff=0
        #The sum of the absolute value of the difference
        #between actual and expected portfolio claim sizes
        sumabsless=0
        #The sum of the above absolute values when actual
        #was less than expected
        nless=0
        #The number of cases when actual was less than expected
        sumabsmore=0
        #The sum of the above absolute values when actual
        #was more than expected
        nmore=0
        #The number of cases when actual was more than expected
        sumclaimvar=0
        #The sum of the portfolio claim size variance
        EY=1000
        #The theoretical average portfolio claim size, since
        #EZ=1000 and EY=EZ*lambda=1000*1
        m=50 #Number of trials
        for(i in 1:m){
                claimnos=rpois.od(n,lambda,phi)
                #The number of claims in each of the n policies.
                claims=randtype(n=sum(claimnos),param1,param2)
                #The size of each claim.
                claimmean=lambda*mean(claims)
```

```
                sumabsdiff=sumabsdiff+abs(lambda*mean(claims)-EY)
                if(lambda*mean(claims)>EY){
                        sumabsmore=sumabsmore+abs(lambda*mean(claims)-EY)
                        nmore=nmore+1
                }
                if(lambda*mean(claims)<EY){
                        sumabsless=sumabsless
                        +abs(lambda*mean(claims)-EY)
                        nless=nless+1
                }
                sumclaimmean=sumclaimmean+lambda*mean(claims)
                sumclaimvar=sumclaimvar+lambda*(var(claims)
                +phi*(mean(claims)**2))
        }
        portmean=sumclaimmean/m
        portvar=sumclaimvar/m
        #Policies are independent, so the sum of their
        #variances is the portfolio variance.
        return(c(portmean,portvar,sumabsdiff/m,sumabsless/nless,
        nless,sumabsmore/nmore,nmore))
}
```

Generating the data for a given test (e.g., test 1) was done as follows:

```
n=c(100,1000,10000)
phi=c(1.01,2,3)
param1=500
param2=1500
for(i in 1:length(n)){
        for(j in 1:length(phi)){
                print(round(compound_pod(n[i],phi[j],runif,
                param1,param2),digits=2))
        }
}
```

Non-exclusive licence to reproduce thesis and make thesis public

I, Frazier Carsten

(date of birth: November 6, 1987),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Overdispersed Models for Claim Count Distribution,

supervised by Meelis Käärik,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu/Tallinn/Narva/Pärnu/Viljandi, **20.05.2013**