2020

AGGREGATE LOSS MODEL WITH POISSON-TWEEDIE LOSS FREQUENCY

Si Chen
chen8470@mylaurier.ca

AGGREGATE LOSS MODEL WITH POISSON-TWEEDIE LOSS FREQUENCY

by

Si Chen

Bachelor of Mathematics, University of Waterloo, 2016

THESIS

Submitted to the Department of Mathematics

Faculty of Science

in partial fulfilment of the requirements for the

Master of Science in Mathematics

Wilfrid Laurier University

2020

Abstract

The aggregate loss model has applications in various areas such as financial risk management and actuarial science. The aggregate loss is the summation of all random losses occurred in a period, and it is governed by both the loss severity and the loss frequency. While the impact of the loss severity on aggregate loss is well studied, less focus is paid on the influence of loss frequency on aggregate loss, which motivates our study. In this thesis, we enrich the aggregate loss framework by introducing the Poisson-Tweedie distribution as a candidate for modelling loss frequency, prove the closedness of Poisson-Tweedie under binomial-thinning, investigate bias of parameter and quantile estimation through simulation and apply our proposed model on real data to demonstrate its advantage. The Poisson-Tweedie distribution family contains many of the commonly used distributions for modelling loss frequency, thus making loss frequency fitting more flexible and reduce the chance of model misspecification. Apart from this feature, the Poisson-Tweedie family is also convolution closed, which allows us to use the same distribution family to model frequency data over different time lengths. The proven closedness under binomial thinning implies that the frequency distribution remains in the same family of Poisson-Tweedie when the observations have a reporting threshold, simplifying the parameter estimation for loss frequency. Through simulation studies, we investigate and find the impact of misspecification of the loss frequency distribution to the aggregate loss quantile, as well as a non-negligible bias of the maximum likelihood estimator of the family index of Poisson-Tweedie. Finally, we have applied our proposed model to Transportation Security Administration (TSA) Claims data to demonstrate modelling capacity on real-world problems.

Acknowlegements

I would like to first thank my supervisor, Dr. Zilin Wang for sharing her knowledge of statistics and her strong support of my research during thesis preparation. Without her guidance and constant feedback, this thesis would not have been achievable.

Many thanks to my co-supervisor, Dr. Mary Kelly for imparting her expertise on insurance, providing additional references and enhancing my thesis.

I would also like to thank Dr. Roman Makarov and Dr. Chengguo Weng for their support and participation as examiners.

I greatly appreciate the support of Dr. Hongcan Lin and Dr. Dezhao Han for teaching C++, debugging R code and providing advice in actuarial science and operational risk.

Contents

Abstract **I**

Acknowlegements **II**

List of Abbreviations **VIII**

1 Introduction **1**

2 Aggregate Loss Models with Poisson-Tweedie Loss Frequency **6**

2.1 Aggregate Loss Model 7

2.2 Poisson-Tweedie Family for Loss Frequency 9

2.3 Binomial Thinning for Modelling Data With Reporting Threshold . . 11

2.4 Parameter Estimation 14

2.4.1 Data Structure 14

2.4.2 Estimation Method For Observations Without Reporting Threshold 15

2.4.3 Estimation Method for Observations with Reporting Threshold 16

2.4.4 Percentile of Aggregate Loss Distribution 19

2.4.5 MLE for Log-Normal Severity and Poisson-Tweedie Frequency 20

3 Simulation Study **22**

3.1 Percentile of Aggregate Loss Distribution Under Different Loss Frequency Distributions 22

3.2 Bias Investigation of Parameter Estimators for Loss Frequency 28

3.2.1 Observations Without Reporting Threshold 30
3.2.2 Observation With Reporting Threshold 33

4 Application **39**

4.1 Analysis of TSA Claims Data using Aggregate Loss Model with Poisson-Tweedie Frequency 39
4.1.1 Data Description 39
4.1.2 Estimation of Model Parameters 45
4.1.3 Quantile Estimation of Monthly Aggregate Loss 49
4.2 Analysis of Reporting Threshold for TSA Claims Data 50
4.2.1 Parameter Estimation of Monthly Data 51
4.2.2 Quantile Estimation of Monthly Aggregate Loss 55

5 Conclusion **59**

References **62**

Appendix A **65**

Derivation of Probability Generating Function Under Binomial Thinning . 65
Derivation of Poisson-Tweedie Algorithm Under Binomial Thinning 66

Appendix B **70**

Core Code for Simulation and Estimation 70

List of Tables

3.1 Tail Risk Statistics of Simulated Aggregate Loss 26
3.2 Summary Statistics of Simulated Parameter a 31
3.3 Summary Statistics of Simulated Frequency 32
3.4 Estimating Threshold 33
3.5 Naive Estimation of Poisson-Tweedie Parameter a 34
3.6 Statistics of Simulated Frequency Without Accounting for Reporting Threshold 35
3.7 Estimated Poisson-Tweedie Parameter a Accounting for Binomial Thinning 36
3.8 Summary Statistics of Simulated Frequency Accounting for Reporting Threshold 37

4.1 TSA Data Variable Description 40
4.2 Summary Statistics of TSA Claims Frequency (Number of Claims) . . 41
4.3 Summary Statistics of TSA Claims Severity (USD) 43
4.4 Fitted Frequency Statistics (Number of Claims) 45
4.5 $\hat{a}$ 95% Confidence Interval 46
4.6 Goodness-of-Fit of Monthly Distribution Fit 46
4.7 Fitted Statistics of TSA Claims Frequency 48
4.8 Fitted Statistics of Severity Data 48
4.9 Estimate of Removed data 51
4.10 Severity Parameter Estimation 51
4.11 Comparison of Monthly Frequency Parameter Estimation 53

4.12 Comparison of Monthly Frequency Summary Statistics 55
4.13 Summary Statitics of Estimated Monthly Aggregate Loss Under Different Reporting Thresholds and Estimation Methods 56

List of Figures

3.1 Aggregate Loss Percentiles (50th, 80th, 95th) 27

4.1 Periodic TSA Claim Frequency Scatter Plot 42
4.2 Periodic TSA Claim Frequency Histogram 43
4.3 TSA Claim Severity 44
4.4 Comparison of Estimated Monthly Frequency with Different Distributions 47
4.5 Estimated Monthly Loss of TSA Claims (Aggregated by Month) . . . 49
4.6 Comparison of Estimated Severity Distribution with Full and Incomplete Data 52
4.7 Comparison of Estimated Monthly Frequency Distribution with Full and Incomplete Data 54
4.8 Comparison of Estimated Aggregate Loss Distribution with Full and Incomplete Data 57

List of Abbreviations

Abbreviation	Description
MLE	Maximum Likelihood Estimation
VaR	Value at Risk
ES	Expected Shortfall
TSA	Transportation Security Administration
BCBS	Basel Committee on Banking Supervision

Chapter 1

Introduction

Losses and damages are events associated with costs that have various underlying causes. These losses occur individually from time to time. People who manage risk are often interested in total loss occurring in a fixed time-period; aggregating the losses by a certain time-period. For example, the aggregated yearly loss is typically referred to as yearly loss.

In a set period, the aggregate loss is defined as the sum of randomly occurred individual loss amounts (Klugman, Panjer, and Willmot 2012). The number of losses in this period is referred to as the loss frequency and the loss amount is the loss severity. Loss frequencies are non-negative integer random variables and loss severities are non-negative continuous random variables. The loss severity is assumed to be identically and independently distributed (i.i.d) within the given time-period whereas the loss frequency is identically and independently distributed across time-periods. Furthermore, loss severity and loss frequency are assumed to be independent (Shevchenko 2011); this independence assumption can simplify the estimation for the aggregate loss model as separate estimation procedures for loss frequency and loss severity respectively (Panjer 2014).

Aggregate loss is often used in the insurance and financial industries to

manage risks. Percentile based risk measures such as value at risk (VaR) and expected shortfall (ES) are often used to make decisions in risk management. For example, in banking, the aggregate loss model is used by the advanced measurement approaches (AMA) for operational risk, to estimate regulatory capital. The 99.9th percentile of the aggregate loss is used to calculate the regulatory capital in operational risk (Kerwer 2005). Regulatory capital is used to mitigate the damage of large losses to businesses. The AMA model in the New Basel Capital Accord (Basel II) is proposed by the Basel Committee on Banking Supervision (Horbenko, Ruckdeschel, and Bae 2011).

The aggregate loss is a non-linear function of a loss frequency and a loss severity distribution, it is usually difficult to derive a closed-form of the distribution of the aggregate loss. Hence, it is a challenge to estimate any measure that relies on the distribution of the aggregate loss such as risk measures. Usually, a direct numerical approach is used to estimate the percentile from a large number of simulated data.

In literature, loss severity has been intensively studied. Many pieces of literature list a series of commonly used distributions for modelling loss severity. Contrary to well-established research on the distribution of loss severity, less attention is paid to the loss frequency. Only a limited number of frequency distributions can be found in Shevchenko (2011). Poisson Inverse-Gaussian as a candidate for loss frequency can be found in Willmot (1987). It is noted that a single distribution may not be enough to fit the various count data well and may lead to misspecification, which consequently leads to poor estimation of the risk measures of the aggregate loss.

The limited choice of distribution of loss frequency in the existing analysis of the aggregate loss model motivates us to enlarge the set of candidates of the frequency distribution. We consider the Poisson-Tweedie family, which covers several of the commonly used loss frequency distributions (e.g., Poisson, Negative Binomial, Poisson Inverse-Gaussian). We expect this three-parameter distribution family to enhance model fitting for loss frequency in the aggregate loss model. Moreover, for

the Poisson-Tweedie family, it has a nice property of convolution closed with regards to family index parameter, which means that the frequencies of daily, weekly, monthly, quarterly and yearly loss are all within the Poisson-Tweedie family, implying no change of frequency modelling.

We verify that different frequency distributions contribute differently to percentile estimates of the aggregate loss through simulation with the same levels of frequency mean, variance and severity distribution respectively. As a result, a wrong specification of either loss frequency or loss severity will deteriorate the accuracy of the insurance premium or regulatory capital because the risk measures used to find the level of regulatory assets are based on the aggregate loss distribution. In this thesis, we find that our proposed aggregate loss model with Poisson-Tweedie frequency outperforms aggregate loss models with Poisson, Negative Binomial and Poisson Inverse-Gaussian frequency, when applied to a TSA claims dataset.

A special and important case in aggregate loss is the incomplete data due to reporting thresholds. For example, in banking, the Basel Committee on Banking Supervision (BCBS), a group of banking supervisory authorities, specifies that institutions must define minimum loss thresholds (Kerwer 2005). The reporting threshold is currently set at EUR 20,00o In Europe (EBA 2019), which means that loss events lower than EUR 20,000 are not reported to regulators. The consequence of the reporting threshold is that loss frequency and loss severity are no longer independent. As shown in our simulation study, naively ignoring the reporting threshold will lead to accuracy issues in estimation with the aggregate loss now. The major theoretical contribution of this thesis is to prove that Poisson-Tweedie is closed under binomial thinning, which makes frequency modelling convenient for incomplete data with a reporting threshold. In particular, the Poisson-Tweedie family index does not change under binomial thinning, which suggests all Poisson-Tweedie special cases (Poison, Negative-Binomial, Poisson Inverse-Gaussian, etc.) are also closed under binomial thinning.

For parameter estimation, we derive the maximum likelihood estimation (MLE) for the aggregate loss model concerning both complete and incomplete data. The biases of estimators have been investigated by a simulation study. We apply our proposed aggregate model to a real case and demonstrate the advantage of our model over aggregate loss models with Poisson, Negative Binomial and Poisson Inverse-Gaussian loss frequency.

In a simulation study, comparing the impact of loss frequency on the percentile of the aggregate loss distribution, we find that misspecifying the frequency distribution can lead to underestimation or overestimation of the percentile of the aggregate loss distribution. We also perform a simulation study to analyze bias in the estimation process, which reveals that when fitting loss frequency, the mean and variance can be captured well, but a non-negligible bias for the Poisson-Tweedie parameter a is observed. Consequently, further study is suggested to reduce this bias.

We apply the proposed model to Transportation Security Administration (TSA) Claims Data and find that the Poisson-Tweedie family fits better than common frequency distributions such as Poisson and Negative Binomial. We also fit the proposed model for incomplete data with hypothetical reporting thresholds (i.e., manually specifying a reporting threshold and removing observations less than the threshold). In these situations, the Poisson-Tweedie distributions, with estimated family parameters close to the previous fit for full data, are obtained. This further supports the applicability of the proposed model, according to the proven closedness property of Poisson-Tweedie under binomial thinning.

This thesis is organized as follows. In Chapter 2, we introduce the proposed model in detail and describe the maximum likelihood estimation of the aggregate loss, including the special case with a reporting threshold. In Chapter 3, we discuss and provide results of simulation methods for analyzing aggregate loss distribution percentile estimates and bias in parameter estimations. In Chapter 4, we apply Poisson-Tweedie distribution as the frequency estimation to Transportation Security

Administration Claims data. In Chapter 5, we summarize our findings and discuss future work.

Chapter 2

Aggregate Loss Models with Poisson-Tweedie Loss Frequency

The aggregate loss model has applications in various areas such as financial risk management and actuarial science. In this chapter, we establish the theoretical and computational foundation of the thesis. The Poisson-Tweedie family of distribution will be incorporated into the framework of the aggregate loss model because it can unify several widely used distributions for the loss frequency. We discover the closedness of Poisson-Tweedie under binomial thinning; this property is particularly useful for incomplete data caused by a reporting threshold.

The estimation of the aggregate loss will also be discussed. To calculate the risk measures used by financial institutions to manage risk and satisfy regulators, we seek to estimate the right side percentiles of the aggregate loss distribution.

2.1 Aggregate Loss Model

Within a single period, the aggregate loss L can be expressed as a random sum as follows:

$$L = \sum_{j=1}^{N} X_j, \tag{2.1}$$

where N is the total number of loss events observed in a certain period, X_j is the amount of loss for the jth event. Usually, to study the statistical properties of aggregate loss, N is referred to as frequency and is described by a non-negative discrete random variable. X_j, for all j, is considered as the loss severity and modelled by a non-negative continuous distribution. Loss frequency and loss severity are typically assumed to be independent.

The severity of loss, X_j, are assumed to be identically and independently distributed (i.i.d.) for $j = 1, 2, \ldots, N$ with density $f_X(x; \boldsymbol{\beta})$. The support of the loss severity is $[0, +\infty)$. The loss severity is well studied with distributions from the exponential family with positive support (Shevchenko 2011; Cummins et al. 1990; Jin, Provost, and Ren 2014; Griffiths and Mnif 2017).

Our research will utilize the log-normal distribution exclusively as the loss severity since the focus of this research is to examine alternate distributions for loss frequency. The log-normal distribution is one severity distribution commonly used in aggregate loss distribution (Papush, Patrik, and Podgaits 2001; Karam and Planchet 2012; Cummins et al. 1990; Griffiths and Mnif 2017). The log-normal has the form

$$f_X(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) \tag{2.2}$$

for $\mu \in (-\infty, +\infty)$ and $\sigma > 0$.

The number of claims, N, is assumed to have mass $f_N(n; \boldsymbol{\theta})$. Values of loss frequency are in the set $\{0, 1, 2, 3, \ldots\}$. Distributions available for frequency includes Poisson, Negative Binomial, Binomial, Geometric and Panjer class (Panjer 2006; Karam and Planchet 2012; Griffiths and Mnif 2017). Loss frequency has a limited

number of distributions available for modelling. Thus, we propose a new candidate, Poisson-Tweedie, for modelling loss frequency in the aggregate loss model (El-Shaarawi, Zhu, and Joe 2011; Bonat et al. 2016; Kokonendji, Demétrio, and Dossou Gbete 2004).

The idea of an extended Poisson and Tweedie family has been investigated in Smyth and Jørgensen (2002). Their research extends the Poisson and Gamma distributions with Tweedie's compound Poisson model for continuous random variables which allows for the direct estimation of the aggregate loss. Furthermore, in Kokonendji, Demétrio, and Dossou Gbete (2004), the Poisson-Tweedie distribution was introduced for discrete random variable to enlarge available distribution for count data. El-Shaarawi, Zhu, and Joe (2011) introduced a new parameterization and recursive algorithm for the probability mass function of the Poisson-Tweedie distribution. This parameterization and algorithmic probability mass function makes the study and analysis of distributions considered in our thesis convenient.

Under the assumption that the loss frequency N and the loss severity are independent within a period. The mean and variance of the aggregate loss can be determined as follows:

$$E(L) = E\left(\sum_{j=1}^{N} X_j\right) = E(N)E(X), \tag{2.3}$$

and

$$\begin{aligned} Var(L) &= Var\left(\sum_{j=1}^{N} X_j\right) \\ &= E\left(Var\left[\sum_{j=1}^{n} X_j | N = n\right]\right) + Var\left(E\left[\sum_{j=1}^{n} X_j | N = n\right]\right) \\ &= E(NVar[X]) + Var(NE[X]) \\ &= E(N)Var(X) + Var(N)E(X)^2. \end{aligned} \tag{2.4}$$

In practice, financial institutes are often interested in the right tail percentile of the aggregate loss. For example, value at risk (VaR) and expected shortfall (ES) are two commonly used risk measures based on the percentile of the distribution.

We define the qth percentile of a random variable X as $0.01 \times q = \Pr(X \leq x)$ where x is the value of the percentile for $0 \leq q \leq 100$.

Value at Risk (VaR) measures a potential loss for given normal market conditions for a given time frame. The relation between VaR and the aggregate loss L is

$$\Pr(L \leq VaR_\alpha) = \alpha \tag{2.5}$$

where L is the aggregate loss, VaR_α is the loss amount of the VaR statistic and $\alpha \in [0, 1]$ is the confidence level. The time-period length of the VaR statistics is the same as the time length of the aggregate loss. The loss amount VaR_α is then equivalent to the $\alpha \times 100$th percentile of the aggregate loss.

The expected shortfall is the average of losses greater than a percentile level. Similar to VaR, ES is also composed of a time-period, a confidence level α and a loss amount. It is used to measures the average loss if the loss exceeds the Value at Risk (VaR breach). We define ES as

$$E[L|L \geq VaR_\alpha]. \tag{2.6}$$

2.2 Poisson-Tweedie Family for Loss Frequency

Using the parameterization of El-Shaarawi, Zhu, and Joe (2011), we define Poisson-Tweedie distribution with parameters PT(a,b,c) which has mean $\mu_N = bc/(1-c)^{1-a}$ and variance $\sigma_N^2 = bc(1-ac)/(1-c)^{2-a}$. We selected this parameterization due to the convenience in studying various distributions covered by the Poisson-Tweedie family and the provided algorithm to calculate the probability mass function. The probability generating function, the power series representation of the probability

mass funciton, of Poisson-Tweedie is defined as

$$G_N(s) = \exp\left\{\frac{b}{a}\Big[(1-c)^a - (1-cs)^a\Big]\right\} \tag{2.7}$$

where $|s| \leq 1$. According to El-Shaarawi, Zhu, and Joe (2011), the three-parameter Poisson–Tweedie family $PT(a, b, c)$ has the probability mass function such that the probability mass p_{k+1} is a linear combination of probability mass $p_0, p_1, \cdots, p_k$, stated as follows:

$$\Pr(N = 0) = p_0 = \begin{cases} e^{b[(1-c)a-1]/a}, & a \neq 0 \\ (1-c)^b & a = 0, \end{cases}$$

$$\Pr(N = 1) = p_1 = bcp_0,$$

$$\Pr(N = k+1) = p_{k+1} = \frac{1}{k+1}\left(bcp_k + \sum_{j=1}^{k} jr_{k+1-j}p_j\right), \quad k = 1, 2, 3, \ldots, \tag{2.8}$$

where

$$r_1 = (1-a)c, \quad r_j + 1 = (\frac{j-1+a}{j+1})cr_j, \quad j = 1, 2, 3, \ldots, k-1,$$

and

$$-\infty < a \leq 1, \quad 0 < b < \infty \quad \text{and} \quad 0 < c \leq 1$$

El-Shaarawi, Zhu, and Joe (2011) defines parameter a of Poisson-Tweedie to be the family index, where the value of a determines to which distribution the Poisson-Tweedie family corresponds.

For example, the Poisson-Tweedie family includes Poisson ($a = 1$), Poisson Inverse-Gaussian ($a = 0.5$), Negative Binomial ($a = 0$), and Polya-Aeppli ($a = -1$). Thus, a Poisson-Tweedie family unifies these individual distribution families.

Parameters b and c are associated with the mean and variance, given by:

$$\mu = \frac{bc}{(1-c)^{1-a}} \tag{2.9}$$

and

$$\sigma^2 = \frac{bc(1-ac)}{(1-c)^{2-a}}. \tag{2.10}$$

Note that Panjer classes (both $(a, b, 0)$ and $(a, b, 1)$) also include Poisson and Negative Binomial distribution families as special cases. However, the probability mass p_{k+1} of the Panjer class depends on p_k, whereas p_{k+1} of Poisson-Tweedies depends on $p_0, p_1, \ldots, p_k$. Therefore, in general, the Poisson-Tweedie family and the Panjer class do not overlap.

2.3 Binomial Thinning for Modelling Data With Reporting Threshold

In practice, insurance and banking institutes often have claim policies and reporting practices such that losses less than the threshold are not reported. For example, a deductible is an amount a policyholder needs to pay before the insurance provider covers the additional costs. For the insurance provider, the claims with severity below the deductible threshold are not observed. European banks are subject to financial reporting standards. The regulatory committee Basel Committee on Banking Supervision (BCBS) assigns a threshold of EUR 20,000 for reported losses in operational risk for banks in Europe (EBA 2019). With a reporting threshold, small losses are not disclosed. Hence, we do not know how many events are missing or the size of the missing losses. This creates a complication as the loss severity X and the loss frequency N are no longer independent. Suppose that losses below the threshold H

are not reported. In insurance, the observed severity is defined as

$$X^* = \begin{cases} X - H & X \geq H \\ \text{Unreported} & X < H \end{cases}.$$

Without loss of generality, we can define the observed severity as

$$X_H = \begin{cases} X & X \geq H \\ \text{Unreported} & X < H \end{cases}. \tag{2.11}$$

It is easy to see that $X_H = X^* + H$.

Loss frequency under a reporting threshold H is then

$$N_H = \begin{cases} I_1 + \cdots + I_N & N > 0 \\ 0 & N = 0 \end{cases}, \tag{2.12}$$

where $I_1, \ldots, I_N$ are independent and identically Bernoulli random variables, defined as

$$I_j = \begin{cases} 1, & X \geq H \\ 0, & X < H \end{cases} \tag{2.13}$$

for $j = 1, 2, 3, \ldots$ and $p_H = \Pr(X \leq H) = \Pr(I_j = 0) = F_X(H; \boldsymbol{\beta})$ (i.e., the probability that the claim is less than the threshold). Typically H is given while p_H can be determined after severity fitting. The process to derive the above new frequency, N_H, which is a random sum of Bernoulli random variables, is called binomial thinning operation (Zhu 2002). Since the occurrence of a loss is reported depends on whether the loss is greater than the threshold, frequency is dependent on the loss severity parameters and the reporting threshold. The assumption that frequency and severity are independent is violated.

Under a reporting threshold, the domain of the severity is $H \leq X_H < \infty$

with the density of reported severity X_H

$$f_{X_H}(x;\boldsymbol{\beta}) = \frac{f_X(x;\boldsymbol{\beta})}{1-F_X(H;\boldsymbol{\beta})} = \frac{f_X(x;\boldsymbol{\beta})}{1-p_H}, \quad H \leq x < \infty$$

(Shevchenko 2011). The mass function of reported frequency, N_H, can be derived with the probability generating function (pgf) $G_{N_H}(s) = G_N(G_I(s))$ where $G_N(s)$ is the pgf of the frequency distribution and $G_I(s)$ is the pgf of the Bernoulli distribution mentioned before (Shevchenko 2011). When the distribution of frequency is closed under binomial thinning, the reported frequency, N_H, has the same form of the probability mass function (pmf) as the case without reporting threshold, thus, the pmf of N_H can be expressed as $f_N(n;\boldsymbol{\gamma})$ where $\boldsymbol{\gamma} = g(\boldsymbol{\theta}, \boldsymbol{\beta})$ is a vector transformation of the frequency parameters, the severity parameters and the threshold H.

The Poisson-Tweedie is closed under binomial-thinning, as proved in Theorm 1.

Theorem 1. *Assume* $N \sim PT(a,\ b,\ c)$, $\{I_1,\ I_2, \ldots\} \overset{i.i.d.}{\sim} Bernoulli(1-p_H)$. *Denote*

$$N_H = \begin{cases} I_1 + \cdots + I_N & N > 0 \\ 0 & N = 0 \end{cases},$$

which is binomial thinning of N. *Then* $N_H \sim PT\left(a,\ b(1-c\cdot p_H)^a,\ \frac{c(1-p_H)}{1-c\cdot p_H}\right)$.

Proof. Given the general probability generating function of frequency under binomial thinning

$$G_{N_H}(s) = G_N(G_I(s)),$$

and incorporating the Poisson-Tweedie probability generating function,

$$G_N(s) = \exp\left\{\frac{b}{a}\Big[(1-c)^a - (1-cs)^a\Big]\right\},$$

the probability generating of Poisson-Tweedie N_H is

$$\begin{aligned} G_{N_H}(s) &= G_N(G(s)) \\ &= G_N(p_H + (1-p_H)s) \\ &= \exp\left\{\frac{b}{a}\Big[(1-c)^a - (1-c(p_H+(1-p_H)s))^a\Big]\right\} \\ &= \exp\left\{\frac{b}{a}(1-c\cdot p_H)^a\Big[(1-\frac{c(1-p_H)}{1-c\cdot p_H})^a - (1-\frac{c(1-p_H)}{1-c\cdot p_H}s)^a\Big]\right\}. \end{aligned}$$

Thus N_H follows $PT\left(a,\ b(1-c\cdot p_H)^a,\ \frac{c(1-p_H)}{1-c\cdot p_H}\right)$. □

Therefore, the probability mass function of binomial thinned Poisson-Tweedie can be derived using the same recursive algorithm with different parameters.

2.4 Parameter Estimation

In this section, we show how MLE can be used to estimate parameters of the aggregate loss distribution and to simulate the aggregate loss distribution. From this, we discuss the use of MLE for typical statistics of data such as the mean, variance, parameters and percentile of the distribution.

Assume we observe T periods ($T = 1, 2, 3, ...$) of losses. N_i are the identically and independently distributed loss frequencies period $i = 1, 2, \ldots, T$ with mass $f_N(n; \boldsymbol{\theta})$. Further, $X_{i,j}$ are the identically and independently distributed loss severities in period i for $j = 1, 2, \ldots, N_i$ with density $f_X(x; \boldsymbol{\beta})$.

2.4.1 Data Structure

The data use for analysis of aggregate loss takes the following form:

Period	Loss Frequency	Loss Severity
1	n_1	$\mathbf{x}_1 = (x_{1,1},\, x_{1,2},\, x_{1,3},\ldots,\, x_{1,n_1})$
2	n_2	$\mathbf{x}_2 = (x_{2,1},\, x_{2,2},\, x_{2,3},\ldots,\, x_{2,n_2})$
3	n_3	$\mathbf{x}_3 = (x_{3,1},\, x_{3,2},\, x_{3,3},\ldots,\, x_{3,n_3})$
...	...	...
T	n_T	$\mathbf{x}_T = (x_{T,1},\, x_{T,2},\, x_{T,3},\ldots,\, x_{T,n_T})$

where $n_1, n_2, n_3, ..., n_T$ are the observed loss frequencies in each time-period and $x_{1,1}, ..., x_{T,n_T}$ are loss severities.

2.4.2 Estimation Method For Observations Without Reporting Threshold

We apply the MLE method for estimating parameters of interest. Recall the loss severity X has density $f_X(x;\boldsymbol{\beta})$ and the loss frequency N has mass $f_N(n;\boldsymbol{\theta})$. Therefore, the likelihood of the aggregate loss can be derived as follows:

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\beta};n_1,...,n_T,x_{1,1},...x_{T,n_T}) &= \prod_{i=1}^{T}\left(f_N(n_i;\boldsymbol{\theta})\prod_{j=1}^{n_i}f_X(x_{ij};\boldsymbol{\beta})\right) \\ &= \left(\prod_{i=1}^{T}f_N(n_i;\boldsymbol{\theta})\right)\left(\prod_{i=1}^{T}\prod_{j=1}^{n_i}f_X(x_{ij};\boldsymbol{\beta})\right). \end{aligned} \tag{2.14}$$

The log-likelihood is then

$$\begin{aligned} l = \log L(\boldsymbol{\theta},\boldsymbol{\beta};\mathbf{n},\mathbf{x}) &= \log\left[\left(\prod_{i=1}^{T}f_N(n_i;\boldsymbol{\theta})\right)\left(\prod_{i=1}^{T}\prod_{j=1}^{n_i}f_X(x_{ij};\boldsymbol{\beta})\right)\right] \\ &= \sum_{i=1}^{T}\log f_N(n_i;\boldsymbol{\theta}) + \sum_{i=1}^{T}\sum_{j=1}^{n_i}\log f_X(x_{ij};\boldsymbol{\beta}). \end{aligned} \tag{2.15}$$

Taking the partial derivative with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ yields the following two estimating equations:

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{T} \log f_N(n_i; \boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{T} \sum_{j=1}^{n_i} \log f_X(x_{ij}; \boldsymbol{\beta}) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{T} \sum_{j=1}^{n_i} \log f_X(x_{ij}; \boldsymbol{\beta}) \end{aligned} \tag{2.16}$$

and

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{T} \log f_N(n_i; \boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{T} \sum_{j=1}^{n_i} \log f_X(x_{ij}; \boldsymbol{\beta}) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{T} \log f_N(n_i; \boldsymbol{\theta}). \end{aligned} \tag{2.17}$$

Equating these partial derivatives to zero leads to the the following estimating equations:

$$\begin{cases} \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{T} \log f_N(n_i; \boldsymbol{\theta}) = 0 \\ \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{T} \sum_{j=1}^{n_i} \log f_X(x_{ij}; \boldsymbol{\beta}) = 0 \end{cases} . \tag{2.18}$$

The solutions, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, of these equations are the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively.

Because of the independence of N and X, we can apply MLE to frequency fitting and severity fitting individually.

2.4.3 Estimation Method for Observations with Reporting Threshold

We also consider MLE for parameter estimation under the circumstance with a reporting threshold and assume the loss frequency distribution is closed under binomial thinning. Recall that the distribution of the loss frequency under binomial thinning is then

$$f_{N_H}(n) = f_N(n; \boldsymbol{\gamma})$$

where $\gamma = g(\boldsymbol{\theta}, \boldsymbol{\beta})$, a function of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

The likelihood of the reported data under a reporting threshold is

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\beta}; n_1, ..., n_T, x_{1,1}, ...x_{T,n_T}) &= \prod_{i=1}^{T} \left(f_N(n_i; \boldsymbol{\gamma}) \prod_{j=1}^{n_i} f_{X_H}(x_{ij}; \boldsymbol{\beta}) \right) \\ &= \left(\prod_{i=1}^{T} f_N(n_i; \boldsymbol{\gamma}) \right) \left(\prod_{i=1}^{T} \prod_{j=1}^{n_i} f_{X_H}(x_{ij}; \boldsymbol{\beta}) \right). \end{aligned} \tag{2.19}$$

The log-likelihood is then

$$\begin{aligned} \log L(\boldsymbol{\theta}, \boldsymbol{\beta}; n_1, ..., n_T, x_{1,1}, ...x_{T,n_T}) &= \log \left[\left(\prod_{i=1}^{T} f_N(n_i; \boldsymbol{\gamma}) \right) \left(\prod_{i=1}^{T} \prod_{j=1}^{n_i} f_{X_H}(x_{ij}; \boldsymbol{\beta}) \right) \right] \\ &= \sum_{i=1}^{T} \log f_N(n_i; \boldsymbol{\gamma}) + \sum_{i=1}^{T} \sum_{j=1}^{n_i} \log f_{X_H}(x_{ij}; \boldsymbol{\beta}). \end{aligned} \tag{2.20}$$

Taking partial derivatives with regards to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ yields

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\sum_{i=1}^{T} \log f_N(n_i; \boldsymbol{\gamma}) \right) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\sum_{i=1}^{T} \log f_N(n_i; g(\boldsymbol{\theta}, \boldsymbol{\beta})) \right) \\ &= \left(\frac{\partial g(\boldsymbol{\theta}, \boldsymbol{\beta})}{\partial \boldsymbol{\theta}} \right) \left(\sum_{i=1}^{T} \frac{\partial}{\partial g(\boldsymbol{\theta}, \boldsymbol{\beta})} \log f_N(n_i; g(\boldsymbol{\theta}, \boldsymbol{\beta})) \right) \\ &= \left(\frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{\theta}} \right) \left(\sum_{i=1}^{T} \frac{\partial}{\partial \boldsymbol{\gamma}} \log f_N(n_i; \boldsymbol{\gamma}) \right) \end{aligned} \tag{2.21}$$

and, similarly,

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\sum_{i=1}^{T}\sum_{j=1}^{n_i}\log f_{X_H}(x_{ij};\boldsymbol{\beta})\right) \\ &\quad + \left(\frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{\beta}}\right)\frac{\partial}{\partial \boldsymbol{\gamma}}\left(\sum_{i=1}^{T}\log f_N(n_i;\boldsymbol{\gamma})\right) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\sum_{i=1}^{T}\sum_{j=1}^{n_i}\log \frac{f_X(x_{ij};\boldsymbol{\beta})}{1-F_X(H;\boldsymbol{\beta})}\right) \\ &\quad + \left(\frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{\beta}}\right)\frac{\partial}{\partial \boldsymbol{\gamma}}\left(\sum_{i=1}^{T}\log f_N(n_i;\boldsymbol{\gamma})\right) \end{aligned} \tag{2.22}$$

Equating these partial derivatives to zero leads to the following estimating equations

$$\begin{cases} \left(\frac{\partial g(\boldsymbol{\theta},\boldsymbol{\beta})}{\partial \boldsymbol{\theta}}\right)\frac{\partial}{\partial \boldsymbol{\gamma}}\left(\sum_{i=1}^{T}\log f_N(n_i;\boldsymbol{\gamma})\right) = 0 \\ \frac{\partial}{\partial \boldsymbol{\beta}}\left(\sum_{i=1}^{T}\sum_{j=1}^{n_i}\log \frac{f_X(x_{ij};\boldsymbol{\beta})}{1-F_X(H;\boldsymbol{\beta})}\right) + \left(\frac{\partial g(\boldsymbol{\theta},\boldsymbol{\beta},H)}{\partial \boldsymbol{\beta}}\right)\frac{\partial}{\partial \boldsymbol{\gamma}}\left(\sum_{i=1}^{T}\log f_N(n_i;\boldsymbol{\gamma})\right) = 0 \end{cases}. \tag{2.23}$$

Simplifying them further leads to the following estimating equations:

$$\begin{cases} \frac{\partial}{\partial \boldsymbol{\gamma}}\sum_{i=1}^{T}\log f_N(n_i;\boldsymbol{\gamma}) = 0 \\ \frac{\partial}{\partial \boldsymbol{\beta}}\sum_{i=1}^{T}\sum_{j=1}^{n_i}\log \frac{f_X(x_{ij};\boldsymbol{\beta})}{1-F_X(H;\boldsymbol{\beta})} = 0 \end{cases}. \tag{2.24}$$

Thus, we can find the MLE of $\boldsymbol{\beta}$ independently of the frequency distribution by maximizing

$$\sum_{i=1}^{T}\sum_{j=1}^{n_i}\log \frac{f_X(x_{ij};\boldsymbol{\beta})}{1-F_X(H;\boldsymbol{\beta})} \tag{2.25}$$

Using the estimate of severity distribution parameter, $\hat{\boldsymbol{\beta}}$, we find the MLE of $\boldsymbol{\theta}$ by maximizing

$$\sum_{i=1}^{T}\log f_N(n_i;\boldsymbol{\gamma}) \tag{2.26}$$

with respect to γ and then solving for $\hat{\boldsymbol{\theta}}$ in $\hat{\boldsymbol{\gamma}} = g(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$. It is also possible to find the MLE of $\boldsymbol{\theta}$ directly by maximizing

$$\sum_{i=1}^{T}\log f_N(n_i; g(\boldsymbol{\theta},\boldsymbol{\beta})|\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}). \tag{2.27}$$

From (2.26), we observe that the frequency distribution parameter estimation takes into account the severity distribution parameters, and thus, the loss severity is no longer independent of the loss frequency.

2.4.4 Percentile of Aggregate Loss Distribution

Note that the aggregate loss defined in (2.1) is a random sum and its distribution usually has complicated or no closed-form depending on the distribution of loss frequency and loss severity. Thus, it is difficult to analytically evaluate the percentile of the aggregate loss distribution. This percentile is used to calculate risk measures such as value at risk, expected shortfall or other measures of interest. One of the methods used in banking and insurance is to estimate the percentile using simulated aggregate loss (Heckman 1983; Shevchenko 2011). The procedure to simulate the aggregate loss is as follows:

1. We simulate a loss frequency, N, based on estimated loss frequency parameters.
2. We simulate n loss severities $x_1, ..., x_n$, using the estimated severity parameters.
3. The simulated aggregate loss is calculated as $\sum_{i=1}^{n} x_i$.
4. Repeat steps 1 to 3 M times (based on the accuracy desired for a given statistic).
5. Empirically estimate the percentile by ordering the simulated aggregate losses from smallest to largest and find the value corresponding to the percentile position. Calculate VaR and ES with the estimated percentiles.

The number of periods M required for simulation depends on the accuracy needed and percentile to be estimated. A higher percentile (i.e., 95th percentile) requires more repetitions to simulate. One method to determine M is to increase the number of repetitions, M, until the desired risk measure converges (the difference of the risk measure between the number of repetitions becoming smaller than a defined value).

2.4.5 MLE for Log-Normal Severity and Poisson-Tweedie Frequency

Following from Section 2.4.2, we apply the estimation method when the claim severity X follows a log-normal distribution and the loss frequency N follows a Poisson-Tweedie distribution with parameters (a,b,c). Tthe probability generating function defined in Section 2.2. The loss severity follows a log-normal distribution with density specified by (2.2). When there is no reporting threshold, we can apply MLE to separately estimate the severity and frequency parameters. Given observed frequencies $n_1, n_2, \ldots, n_T$ and observed severities $x_{1,1}, x_{1,2}, \ldots, x_{T,n_T}$, parameter estimation involves the following two steps:

1. Estimate severity parameters

$$(\hat{\mu}, \hat{\sigma}) = \arg\max_{(\mu,\sigma)} \sum_{i=1}^{T} \sum_{j=1}^{n_T} \log f_X(x_{i,j}; \mu, \sigma). \tag{2.28}$$

2. Estimate frequency parameters

$$(\hat{a}, \hat{b}, \hat{c}) = \arg\max_{(a,b,c)} \sum_{i=1}^{T} \log f_N(n_i; a, b, c). \tag{2.29}$$

When a reporting threshold H exists, such that losses under the reporting threshold are not recorded, we can apply the estimation from Section 2.4.2. Theorem 1 states Poisson-Tweedie is closed under binomial thinning with parameters

$$\boldsymbol{\gamma} = g(a, b, c, \mu, \sigma) = (a,\ b\Big(1 - c \cdot p_H\Big)^a,\ \frac{c(1 - p_H)}{1 - c \cdot p_H}) \tag{2.30}$$

where

$$p_H = \Pr(X < H) = \int_{-\infty}^{H} f_X(s; \mu, \sigma) ds = F_X(H; \mu, \sigma). \tag{2.31}$$

The method of estimating parameters is:

1. Estimate severity parameters

$$(\hat{\mu}, \hat{\sigma}) = \underset{(\mu,\sigma)}{\arg\max} \sum_{i=1}^{T} \sum_{j=1}^{n_T} \log \frac{f_X(x_{i,j}; \mu, \sigma)}{1 - F_X(x_{i,j}; \mu, \sigma)}. \tag{2.32}$$

2. With the estimated parameters $\hat{\mu}$,$\hat{\sigma}$ estimate p_H by calculating

$$\hat{p}_H = F_X(H; \hat{\mu}, \hat{\sigma}). \tag{2.33}$$

3. Estimate frequency parameters

$$(\hat{a}, \hat{b}, \hat{c}) = \underset{(a,b,c)}{\arg\max} \sum_{i=1}^{T} \log f_N \left(n_i; a, b\Big(1 - c \cdot \hat{p}_H\Big)^a, \frac{c(1 - \hat{p}_H)}{1 - c \cdot \hat{p}_H} \right). \tag{2.34}$$

This MLE method of our proposed aggregate loss distribution parameters will be applied in the following simulation study in Chapter 3 and data application in Chapter 4.

Chapter 3

Simulation Study

In financial risk management, we wish to accurately predict losses since an underestimation would expose us to unnecessary financial burden and an overestimation typically leads to a loss of potential profit. In particular, we expect that due to the different characteristics between distributions, different loss frequency distributions will contribute differently to the estimation of the aggregate loss distribution percentiles. This implies that a misspecification of the loss frequency can lead to an inaccurate estimation of the aggregate loss distribution. We also wish to investigate bias in the MLE of the Poisson-Tweedie parameters. To this end, we employ simulation studies for the aggregate loss percentiles among different frequency distributions and for Poisson-Tweedie parameter estimation with both complete and incomplete data due to the reporting threshold.

3.1 Percentile of Aggregate Loss Distribution Under Different Loss Frequency Distributions

To study aggregate loss percentiles among different frequency distributions, we specify the same mean and variance for both frequency (number of claims) and severity (size of the claim) and then observe the impact of different frequency distributions on the

aggregate 95th percentile and the expected shortfall (above 95th percentile). The Poisson-Tweedie probability mass function algorithm is programmed according to (El-Shaarawi, Zhu, and Joe 2011). The Poisson-Tweedie mass under binomial thinning is also programmed according to Theorem 1 with the algorithm given in Appendix A. The algorithm for calculating the probability mass function of Poisson-Tweedie is tested against available known distributions including Poisson and Negative Binomial distributions. These steps are as follows

1. We use the Poisson-Tweedie distribution family to generate frequency random numbers of different distributions based on the family index a. We set different levels of mean and variance and determine Poisson-Tweedie parameters b and c based on the chosen mean and variance. We generate loss frequency n_i for T periods, that is, T Poisson-Tweedie random numbers.

2. Based on chosen severity distributions, we simulate $\sum_{i=1}^{T} n_i$ (simulated from step 1) loss severity random variables and aggregate them by period.

3. The percentile of the aggregate loss can be empirically approximated with a large number of periods. The number of periods T can be determined by increasing the number of periods until the specified risk measure converges (when the difference between the risk measure for increasing the number of periods becomes smaller than a defined number).

We choose Poisson (Poisson-Tweedie parameter $a = 1$), Negative Binomial (Poisson-Tweedie parameter $a = 0$) and Poisson Inverse-Gaussian (Poisson-Tweedie parameter $a = 0.5$) discrete distributions to study. For frequency, we chose means of 2, 10 and 30 number of claims per period. The variance is 5 times the mean except for Poisson distribution. The Poisson-Tweedie parameter equivalent of Poisson, Negative Binomial and Poisson Inverse-Gaussian will be set according to the relationship between parameters and frequency mean and variance (see Section 2.2).

We specified a Log-Normal distribution for the claim severity currency unit.

The parameter μ measures the mean log of the data and σ measures the standard deviation of the log of the data. We set the severity parameters at Log-Normal(7,0.1), Log-Normal(8,0.2), and Log-Normal(9,0.3) with claim size listed in Table 3.1. Real data may have a much higher mean and variance for both frequency and severity distributions. However, for our purpose, we chose these mean values to try to reduce the number of simulations needed. Since we are interested in the impact of loss frequency distributions, we minimize the impact of the loss severity distribution on the tail percentile of the aggregate loss distribution. The σ parameter of the log-normal distribution is chosen to minimize the impact the severity variance has on the aggregate loss so that we can use fewer simulations to obtain better accuracy. There is a total of 27 different combinations of frequency and severity distribution parameter levels. For each combination, we perform a Monte Carlo simulation with $T = 1000000$ periods of loss frequency and loss severities for each simulated loss frequency value. The number of periods is chosen such that at the 0.95 aggregate loss percentile, the convergence is within one percent tolerance, that is, if $T = 1000000$ periods output $loss_1$ for the chosen risk measure and $T = 1000001$ periods output $loss_2$, then $\frac{loss_2 - loss_1}{loss_1} < 0.01$. The resulting percentile estimates of the aggregate loss is compared between distributions.

From the results of the simulation in Table 3.1, we can observe that difference exists in tail estimates of the aggregate loss with different frequency distributions when the mean and the variance are kept at a fixed level. For example, at all levels of severity and loss frequency mean of 2 losses per period, we observe that Poisson (PT $a = 1$) has the smallest 0.95 percentile followed by Poisson Inverse-Gaussian (PIG) (PT $a = 0.5$) and Negative Binomial (PT $a = 0$) has the largest 0.95 percentile At a frequency mean of 10, Negative Binomial and PIG have very similar 0.95 percentile values while Poisson has the smallest 0.95 percentile value. At a frequency mean of 30, PIG has the 0.95 percentile greater than Negative Binomial. Poisson has the smallest 0.95 percentile at all levels compared with other distributions. This observation suggests that some underlying interaction may exist between the shape of

the severity distribution and the frequency distribution on how frequency distribution contributes to percentile estimates of the aggregate loss. For all combinations of loss severity and loss frequency parameters, the value of 95% expected shortfalls show that Poisson has the smallest 95% ES followed by Negative Binomial and PIG having the largest 95% ES. We also notice that as the level of frequency average increases, the relative difference between the percentile estimate of aggregate loss with different loss frequency distribution decreases. This suggests that Poisson Inverse-Gaussian has a thicker right tail than Negative Binomial.

Table 3.1: Tail Risk Statistics of Simulated Aggregate Loss

Sev. Parameter	Sev. Mean	Sev. SD	Freq. Dist.	Freq. Parameter	Freq. Mean	Freq. SD	0.95 Percentile	95th ES
Log-Norm(7,0.1)	1102.13	110.49	Poisson	PT(1,2,1)	2	1.414214	5167.54	6041.38
			NB	PT(0,0.5,0.8)		3.162278	9167.23	13564.58
			PIG	PT(0.5,0.75,0.89)		3.162278	8643.60	13758.27
			Poisson	PT(1,10,1)	10	3.162278	17089.86	18861.57
			NB	PT(0,2.5,0.8)		7.071068	26023.06	32150.42
			PIG	PT(0.5,3.75,0.89)		7.071068	26025.69	33473.44
			Poisson	PT(1,30,1)	30	5.477226	43347.18	46193.42
			NB	PT(0,7.5,0.8)		12.247449	57783.77	66324.21
			PIG	PT(0.5,11.25,0.89)		12.247449	58273.90	68190.74
Log-Norm(8,0.2)	3041.18	614.37	Poisson	PT(1,2,1)	2	1.414214	14139.89	16821.92
			NB	PT(0,0.5,0.8)		3.162278	25545.15	37477.62
			PIG	PT(0.5,0.75,0.89)		3.162278	23643.64	38056.28
			Poisson	PT(1,10,1)	10	3.162278	47463.30	52391.75
			NB	PT(0,2.5,0.8)		7.071068	71959.92	89108.17
			PIG	PT(0.5,3.75,0.89)		7.071068	71987.61	92717.04
			Poisson	PT(1,30,1)	30	5.477226	120054.57	128052.10
			NB	PT(0,7.5,0.8)		12.247449	159589.59	183054.16
			PIG	PT(0.5,11.25,0.89)		12.247449	160902.07	188249.00
Log-Norm(9,0.3)	8476.05	2601.12	Poisson	PT(1,2,1)	2	1.414214	40202.58	47997.21
			NB	PT(0,0.5,0.8)		3.162278	71372.17	104776.05
			PIG	PT(0.5,0.75,0.89)		3.162278	66164.66	105993.96
			Poisson	PT(1,10,1)	10	3.162278	133497.28	147907.74
			NB	PT(0,2.5,0.8)		7.071068	200926.27	248375.53
			PIG	PT(0.5,3.75,0.89)		7.071068	201795.63	259492.82
			Poisson	PT(1,30,1)	30	5.477226	336978.36	360090.22
			NB	PT(0,7.5,0.8)		12.247449	446163.66	511940.22
			PIG	PT(0.5,11.25,0.89)		12.247449	449077.76	524886.07

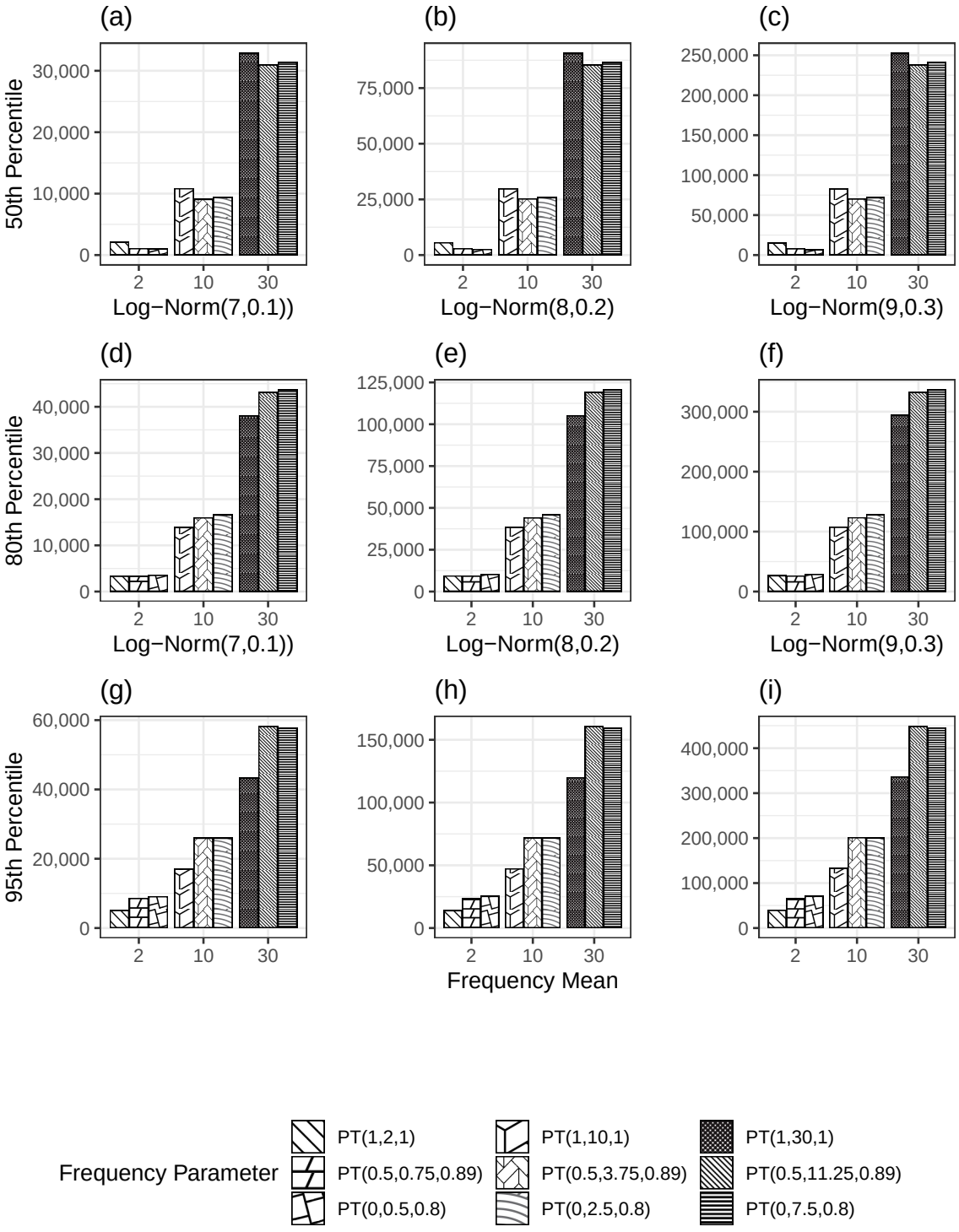

Figure 3.1: Aggregate Loss Percentiles (50th, 80th, 95th)

As shown in Figure 3.1, we see that at different percentile levels, the effects of frequency distribution on aggregate loss percentile value are also different. For example, with Log-Norm(8,0.2) claim severity and frequency mean of 30 claims per period show that the value of 50th percentile aggregate loss with Negative Binomial (PT(0,b,c)) frequency is greater than with Poisson (PT(1,b,c)), with Poisson-Inverse Gaussian (PT(0.5,b,c)) having the smallest estimate. At the 95th percentile aggregate loss estimate, the Negative Binomial frequency has the greatest estimate, followed by Poisson Inverse-Gaussian, with Poisson having the smallest estimate. This simulation shows that the frequency average has an impact on how different frequency distribution affects the aggregate loss percentile estimates. This phenomenon may be derived from the interaction between loss severity distribution shape and loss frequency on aggregate loss percentile estimates.

Further studies may include investigating the interaction between loss severity distribution shape and loss frequency on the percentile estimates of the aggregate loss, as well as the impact of different loss frequency on percentile estimates of the aggregate loss under very large means of loss frequency.

3.2 Bias Investigation of Parameter Estimators for Loss Frequency

We investigate the bias of maximum likelihood estimators of the frequency parameter for observations without threshold (complete data) and with a threshold (incomplete data) respectively.

We apply the Monte Carlo simulation to assess whether bias exists in estimating frequency parameters for both complete data and data and incomplete data using the maximum likelihood method. We focus on the shape parameter of the Poisson-Tweedie distribution (parameter a) and the estimated frequency mean and variance. Using simulation, we obtain information on the bias of estimates. For the

investigation of bias in MLE estimation of Poisson-Tweedie using the algorithm from El-Shaarawi, Zhu, and Joe (2011):

1. Simulate T periods of loss frequency based on given Poisson-Tweedie parameters.
2. Estimate Poisson-Tweedie parameters with the simulated sample according to (2.29).
3. Repeat steps 1 and 2, K times to obtain a set of estimated parameters. K is chosen such that the average estimated parameters converge when repetitions increase (when the difference of the parameters for increasing repetitions becomes smaller than a defined value).

For the investigation of bias in MLE estimation of Poisson-Tweedie under binomial thinning:

4. Simulate loss severity based on the simulated frequency from step 1 and the given Log-Normal parameter.
5. Apply a reporting threshold by removing loss severity based on the chosen threshold H and calculate the loss frequency under binomial thinning by adding up the number of losses in each period.
6. Perform naive MLE estimation of Poisson-Tweedie parameters according to (2.29). Naive estimation to perform estimation without accounting for the reporting threshold and estimate assuming the reporting threshold does not exist.
7. Account for the reporting threshold and estimate parameters according to (2.32), (2.33) and (2.34)
8. Repeat step 4 to 7 K times, the same number as in step 3 for convenience.

For frequency, we set the shape parameter a of the Poisson-Tweedie distribution at certain intervals between -2 to 1 (a=-2, -1.5, -1, -0.5, 0, 0.2, 0.4, 0.5, 0.6, 0.8 and 1). We choose a mean of 50 claims per period and a variance of 250 to observe overdispersed models (except for $a = 1$ as this corresponds to Poisson distribution) to keep the number of required calculations relatively small. Here we specify a severity

distribution of Log-Norm(8,3) and choose the number of periods to be 100. This chosen parameter has a higher loss severity variance and standard deviation since we are only interested in frequency parameters in this simulation. The loss severity mean is 268,337.30 and the standard deviation is 24,153,462. We specify reporting thresholds at severity levels of 100, 250, 500, 750, and 1000. We repeat this process $K = 100000$ times for each combination of frequency parameter, total periods and threshold level. This number of repetitions gives stable average parameter estimation at more than 3 decimal places. The estimates using simulated data (i.e. parameter estimates, mean and variance of estimated frequency, etc.) are compared with the true values to determine if any bias

3.2.1 Observations Without Reporting Threshold

Base on our simulation, we believe that bias does exist in some aspects of parameter estimation. The number of repetitions, $K = 100000$, gives us an average Poisson-Tweedie parameter a estimate tolerance of less than 3 decimal places.

Table 3.2: Summary Statistics of Simulated Parameter a

Actual a	$E(\hat{a})$	$SD(\hat{a})$	Bias Estimate - Actual	Bias Relative %
-2.00	-1.20	1.98	0.80	39.99
-1.50	-1.12	1.93	0.38	25.39
-1.00	-1.00	1.86	0.00	-0.48
-0.50	-0.82	1.72	-0.32	-64.33
0.00	-0.49	1.43	-0.49	N/A
0.20	-0.28	1.21	-0.48	-241.50
0.40	0.00	0.88	-0.40	-99.95
0.50	0.17	0.66	-0.33	-65.22
0.60	0.37	0.43	-0.23	-39.06
0.80	0.74	0.06	-0.06	-7.85
1.00	0.82	0.06	-0.18	-18.24
(1)	(2)	(3)	(4)	(5)
			(2)-(1)	$\frac{(4)}{\lvert(1)\rvert} * 100\%$

From Table 3.2, we observe that bias exists in estimating the Poisson-Tweedie parameter a. We also find that the higher the value of the parameter a, the lower the empirical variance (the variance of the 100,000 sample estimates). The low empirical variance of the parameter estimate at actual parameter $a = 0.8$ and $a = 1$ suggest that the bias is consistent at those levels. The results also suggest there could be a pattern or some association between the theoretical value and the bias of the Poisson-Tweedie parameter a. We observe underestimation when theoretical $a > -1$ and overestimation when theoretical $a < -1$.

- When $a = 0$ (i.e., Negative-Binomial), the bias could reach -0.5. Such underestimation would cause the fitted frequency distribution to have a lighter tail

than it would be, and further tend to underestimate the tail percentile of the aggregate loss distribution.

- When $0 < a \leq 1$, the bias is negative. Similarly, both frequency and aggregate loss distributions tend to underestimate tails.
- When $-1 < a \leq 0$, the bias is still negative and we would underestimate aggregate loss tail percentiles.
- when $a < -1$, the bias is positive and we would overestimate aggregate loss tail percentiles.

Table 3.3 suggests that the MLE Poisson-Tweedie estimate of the mean and variance of the loss frequency does not have any bias. The difference between sample and estimate is less than 2% for both mean and variance. The Poisson-Tweedie parameter $a = 1$ corresponds with the Poisson distribution which has equal variance property, thus we observe that the sample variance is equal to the sample mean.

Table 3.3: Summary Statistics of Simulated Frequency

Theoretical Parameter a	Sample Mean	Estimated Mean	Sample Variance	Estimated Variance
-2.0	49.995	49.973	249.970	251.297
-1.5	49.995	49.974	249.969	251.084
-1.0	49.995	49.976	249.972	250.815
-0.5	49.995	49.979	249.964	250.426
0.0	49.995	49.984	249.964	249.916
0.2	49.995	49.987	249.959	249.700
0.4	49.995	49.990	249.953	249.589
0.5	49.996	49.992	249.947	249.663
0.6	49.996	49.993	249.939	249.943
0.8	49.996	49.995	249.876	252.838
1.0	49.998	49.996	49.995	52.585

In general, we find that while estimating frequency using Poisson-Tweedie and MLE, the estimated mean and variance do not seem to have any bias. However, there is a difference between the estimated parameter a and the theoretical value of

parameter a. Based on the results in Section 3.1, the percentile-based estimates would also be different from the true value.

3.2.2 Observation With Reporting Threshold

The missing percentage of data, p_H, is estimated according to Section 2.4.3. The theoretical threshold is calculated from the true parameters, the sample threshold is calculated from simulated data and the estimated threshold is calculated from the estimated parameters.

Table 3.4: Estimating Threshold

Threshold Level H	Theoretical Threshold %	Sample Threshold %	Estimated Threshold %	Estimated Threshold SE (%)
100	12.89	12.89	12.90	0.86
250	20.44	20.44	20.46	1.36
500	27.59	27.59	27.61	1.88
750	32.28	32.28	32.30	2.24
1000	35.79	35.79	35.81	2.51

After estimating the missing percentage, Table 3.4 shows the estimate to be extremely close to the sample and the theoretical missing percentage. We do notice that the estimated threshold standard error increases as the threshold increase. As more information is removed, we can expect the uncertainty to also increase.

In frequency estimation, if we do not take into consideration the binomial thinning effect, increasing the threshold of data removal will decrease the estimated Poisson-Tweedie parameter a (Table 3.5). This effect is added on top of the bias in estimating full data. The empirical variance of the parameter estimate also increases as the threshold increases.

Table 3.5: Naive Estimation of Poisson-Tweedie Parameter a

Theoretical Threshold	12.89 %	20.44 %	27.59 %	32.28 %	35.79 %
Theoretical a	Bias of Estimated a (Variance of Estimated a)	Bias of Estimated a (Variance of Estimated a)	Bias of Estimated a (Variance of Estimated a)	Bias of Estimated a (Variance of Estimated a)	Bias of Estimated a (Variance of Estimated a)
-2.0	0.75 (2.2)	0.72 (2.38)	0.69 (2.53)	0.68 (2.6)	0.67 (2.7)
-1.5	0.33 (2.14)	0.3 (2.31)	0.27 (2.47)	0.26 (2.55)	0.25 (2.63)
-1.0	-0.05 (2.06)	-0.08 (2.22)	-0.12 (2.37)	-0.13 (2.45)	-0.14 (2.52)
-0.5	-0.37 (1.91)	-0.39 (2.05)	-0.43 (2.21)	-0.44 (2.27)	-0.46 (2.34)
0.0	-0.54 (1.59)	-0.57 (1.72)	-0.6 (1.85)	-0.62 (1.91)	-0.63 (1.98)
0.2	-0.52 (1.35)	-0.55 (1.47)	-0.59 (1.59)	-0.61 (1.65)	-0.62 (1.71)
0.4	-0.44 (1)	-0.47 (1.1)	-0.5 (1.2)	-0.52 (1.26)	-0.54 (1.31)
0.5	-0.36 (0.76)	-0.39 (0.85)	-0.42 (0.95)	-0.44 (1)	-0.46 (1.05)
0.6	-0.26 (0.51)	-0.29 (0.57)	-0.31 (0.65)	-0.33 (0.7)	-0.35 (0.74)
0.8	-0.07 (0.08)	-0.08 (0.1)	-0.1 (0.12)	-0.11 (0.14)	-0.11 (0.16)
1.0	-0.18 (0.06)	-0.18 (0.06)	-0.18 (0.07)	-0.21 (0.07)	-0.2 (0.07)

Table 3.6: Statistics of Simulated Frequency Without Accounting for Reporting Threshold

H	100	250	500	750	1000
Theoretical Threshold	12.89 %	20.44 %	27.59 %	32.28 %	35.79 %
Theoretical a	Estimated Mean (Estimated Variance)	Estimated Mean (Estimated Variance)	Estimated Mean (Estimated Variance)	Estimated Mean (Estimated Variance)	Estimated Mean (Estimated Variance)
Sample Mean	43.55	39.78	36.2	33.86	32.1
(Sample Variance)	(195.26)	(166.34)	(141.01)	(125.54)	(114.52)
(Sample Poisson Variance)	(43.53)	(39.76)	(36.19)	(33.84)	(32.09)
-2	43.53	39.76	36.19	33.84	32.09
	(196.37)	(167.33)	(141.88)	(126.33)	(115.26)
-1.5	43.53	39.76	36.19	33.84	32.09
	(196.21)	(167.19)	(141.76)	(126.24)	(115.17)
-1	43.53	39.76	36.19	33.84	32.09
	(196)	(167.01)	(141.61)	(126.11)	(115.06)
-0.5	43.54	39.77	36.19	33.85	32.09
	(195.69)	(166.74)	(141.39)	(125.9)	(114.87)
0	43.54	39.77	36.19	33.85	32.09
	(195.3)	(166.41)	(141.11)	(125.66)	(114.65)
0.2	43.54	39.77	36.19	33.85	32.09
	(195.12)	(166.26)	(140.97)	(125.53)	(114.53)
0.4	43.55	39.77	36.2	33.85	32.1
	(195.02)	(166.15)	(140.88)	(125.44)	(114.44)
0.5	43.55	39.77	36.2	33.85	32.1
	(195.06)	(166.18)	(140.9)	(125.44)	(114.45)
0.6	43.55	39.78	36.2	33.86	32.1
	(195.25)	(166.33)	(141.01)	(125.54)	(114.53)
0.8	43.55	39.78	36.2	33.86	32.1
	(197.45)	(168.16)	(142.52)	(126.86)	(115.72)
1	43.55	39.78	36.2	33.86	32.1
	(45.81)	(41.84)	(38.09)	(35.65)	(33.79)

Without considering binomial thinning, Table 3.6 shows that we estimate the sample frequency mean and variance instead of the theoretical mean and variance. In our simulation, we underestimate both mean and variance.

Table 3.7: Estimated Poisson-Tweedie Parameter a Accounting for Binomial Thinning

Theoretical Threshold	12.89 %	20.44 %	27.59 %	32.28 %	35.79 %
Theoretical a	Bias of Estimated a (Variance of Estimated a)	Bias of Estimated a (Variance of Estimated a)	Bias of Estimated a (Variance of Estimated a)	Bias of Estimated a (Variance of Estimated a)	Bias of Estimated a (Variance of Estimated a)
-2.0	0.95 (1.7)	1 (1.65)	0.91 (1.85)	0.98 (1.8)	0.98 (1.78)
-1.5	0.52 (1.66)	0.57 (1.61)	0.49 (1.8)	0.55 (1.76)	0.54 (1.75)
-1.0	0.12 (1.59)	0.16 (1.55)	0.08 (1.74)	0.14 (1.68)	0.13 (1.68)
-0.5	-0.22 (1.48)	-0.18 (1.44)	-0.26 (1.63)	-0.21 (1.57)	-0.22 (1.57)
0.0	-0.43 (1.24)	-0.41 (1.22)	-0.48 (1.4)	-0.44 (1.34)	-0.46 (1.37)
0.2	-0.44 (1.07)	-0.43 (1.06)	-0.49 (1.22)	-0.47 (1.18)	-0.49 (1.21)
0.4	-0.39 (0.8)	-0.39 (0.81)	-0.44 (0.95)	-0.42 (0.92)	-0.45 (0.96)
0.5	-0.32 (0.62)	-0.33 (0.64)	-0.38 (0.76)	-0.37 (0.73)	-0.39 (0.79)
0.6	-0.24 (0.42)	-0.25 (0.44)	-0.29 (0.54)	-0.29 (0.52)	-0.31 (0.58)
0.8	-0.07 (0.07)	-0.08 (0.09)	-0.09 (0.11)	-0.1 (0.12)	-0.11 (0.14)
1.0	-0.34 (0.06)	-0.47 (0.02)	-0.49 (0.01)	-0.49 (0.01)	-0.49 (0.01)

Table 3.7 shows that while there is no obvious association between the threshold level and the estimated parameter a. The parameter a bias effect of estimating data with threshold using this method is similar to estimating data without threshold. The empirical variance of the parameter estimate also increases as the threshold increases. It is interesting to note that for some values of parameter a, the variance decreases at 32.28% threshold. Comparing Table 3.5 and Table 3.7, we find that for some values of a, the bias of the naive estimation is smaller than the bias of the estimation accounting for the reporting threshold. This is due to the fact that naive estimation adds another bias to the bias of the parameter estimation. For these values of a, the effects of the biases may cancel out and give the impression that the bias is lower.

Table 3.8: Summary Statistics of Simulated Frequency Accounting for Reporting Threshold

Theoretical Threshold	12.89 %	20.44 %	27.59 %	32.28 %	35.79 %
Theoretical a	Estimated Mean (Estimated Variance)	Estimated Mean (Estimated Variance)	Estimated Mean (Estimated Variance)	Estimated Mean (Estimated Variance)	Estimated Mean (Estimated Variance)
Sample Mean	43.55	39.78	36.2	33.86	32.1
(Sample Variance)	(195.26)	(166.34)	(141.01)	(125.54)	(114.52)
(Sample Poisson Variance)	(43.53)	(39.76)	(36.19)	(33.84)	(32.09)
-2	49.98	50	50.02	50.04	50.06
	(251.6)	(251.86)	(252.13)	(252.47)	(252.87)
-1.5	49.98	50	50.02	50.04	50.06
	(251.37)	(251.65)	(251.92)	(252.27)	(252.65)
-1	49.99	50	50.02	50.05	50.07
	(251.11)	(251.36)	(251.63)	(252)	(252.38)
-0.5	49.99	50	50.03	50.05	50.07
	(250.69)	(250.94)	(251.2)	(251.56)	(251.92)
0	49.99	50.01	50.03	50.05	50.07
	(250.16)	(250.4)	(250.67)	(251.03)	(251.38)
0.2	50	50.01	50.03	50.06	50.08
	(249.92)	(250.15)	(250.4)	(250.74)	(251.09)
0.4	50	50.02	50.04	50.06	50.08
	(249.77)	(249.97)	(250.21)	(250.54)	(250.87)
0.5	50	50.02	50.04	50.06	50.08
	(249.81)	(250)	(250.24)	(250.56)	(250.88)
0.6	50	50.02	50.04	50.06	50.08
	(250.06)	(250.21)	(250.45)	(250.76)	(251.08)
0.8	50.01	50.02	50.04	50.07	50.09
	(252.95)	(253.1)	(253.33)	(253.67)	(253.96)
1	50.01	50.02	50.04	50.07	50.09
	(53.81)	(55.06)	(55.78)	(56.24)	(56.64)

Table 3.8 shows that when we take into account the binomial thinning of frequency, we estimate the theoretical mean and variance instead of the sample mean and variance. Except for Poisson (a=1) estimated variance, all results are within 2% of true value. The Poisson variance overestimates the true variance up to more than 10%.

In this simulation study, we find that the frequency mean and variance can be estimated with negligible or no bias. However, the Poisson-Tweedie family index parameter a seems to have a bias. When theoretical parameter a is less than -1, the estimate tends to overestimate the true value. This would cause aggregate loss right

tail percentiles to be overestimated. When theoretical parameter a is greater than -1, the estimate tends to underestimate the true value which would cause aggregate loss tail percentiles to be underestimated. For data with reporting threshold, directly estimating the frequency will underestimate the mean, variance and parameter a. Higher levels of the threshold have a greater impact on the underestimation. The method provided in Section 2.4.3 can estimate the missing data percent, frequency mean and variance and severity parameters without bias. Higher threshold levels will contribute to higher variance (calculated with simulated sample). Bias is still present in Poisson-Tweedie MLE estimation of parameter a.

Chapter 4

Application

The Transportation Security Administration (TSA) is a United States agency focused on air traffic security in the United States. It was created in response to the September 11, 2000 attack as a centralized organization that provides security for United States transportation systems. The TSA claims data includes claims against TSA for injury, loss or damage of property during passenger's screening process. We apply our proposed model on this dataset to illustrate the use of Poisson-Tweedie as the loss frequency distribution. The data does not seem to have a reporting threshold, thus we apply thresholds at 10, 20, and 30 USD to study parameter estimation with real incomplete data.

4.1 Analysis of TSA Claims Data using Aggregate Loss Model with Poisson-Tweedie Frequency

4.1.1 Data Description

TSA data for the years of 2002 to the end of 2015 was obtained from the Department of Homeland Security website (https://www.dhs.gov/tsa-claims-data).

Table 4.1: TSA Data Variable Description

Variable Name	Variable Description
Claim.Number	claim identification number
Date.Received	date that the claim is recieved by TSA
Claim.Type	The type of damage (i.e., passenger injury, property damage)
Item.Category	The category of the damaged object (i.e., electornics, clothing)
Claim.Amount	the dollar amout requested for compensation in USD
Close.Amount	the dollar amout given for compensation in USD
Disposition	the status of the claim, includes "Approved", "Claim entered", "Canceled", "Closed as a contractor claim", "Denied", "In litigation", "In review", "Insufficient" or "Settled"

The data contains “claim number”, “date received”, “claim type”, “item category”, “close amount” and “disposition”. Data from 2002 to 2006 also have “claim amount”. Our variables of interest are defined in Table 4.1. In total there are 286,952 observations from 2002 to 2015. The variables of interest for this analysis are the date received, close amount, claim amount, and the disposition. Date received is the date that the claim is received by TSA. Close amount is the final amount TSA pays out to claimants. The claim amount is the amount requested by the claimant. Disposition is a categorical variable that indicates the outcome of the claim, whether the claim is settled, approved in full or denied. In this study, we use the date that the claim is received and the payout of the claim. Only claims that are settled or approved in full were used. We use the closed amount for loss severity. If the closed amount is missing, then the claim amount is used. Observations with missing payments are removed. After applying these filters, the resulting number of observations is 81,065. We then

aggregated claims by day, week, month and quarter.

Time plots for the entire time frame indicate an unexplained spike in frequency before 2005 and clear seasonality in 2013 to 2015. But, the data from 2008 to 2012 in Figure 4.1 shows a relatively stable period in terms of frequency which is suitable for analysis, even though we observe a slight dip in frequency between the end of 2009 and early 2010[1]. Hence, we only use the data from 2008 to 2012. Our cleaned data has 15882 observations.

We apply the aggregate loss model defined in (2.1) to the TSA claim data where the frequency is calculated by counting the number of claims within each time-period and the severity is given by the close amount.

The frequency summary statstics are shown in Table 4.2.

Table 4.2: Summary Statistics of TSA Claims Frequency (Number of Claims)

	Min.	25th Percentile	Median	Mean	75th Percentile	Max.	SD	Variance
Daily	0	0.00	8.0	8.69	14.00	48	8.11	65.70
Weekly	6	50.00	61.0	60.62	71.00	111	15.08	227.36
Monthly	74	233.25	265.5	264.70	303.75	396	55.81	3114.79
Quarterly	500	679.50	834.0	794.10	922.75	967	134.85	18184.94

[1]We suspect the dip in loss frequency in 2009 may be related to the lagged effect of the 2008 US Housing crisis which negatively affected the US economy

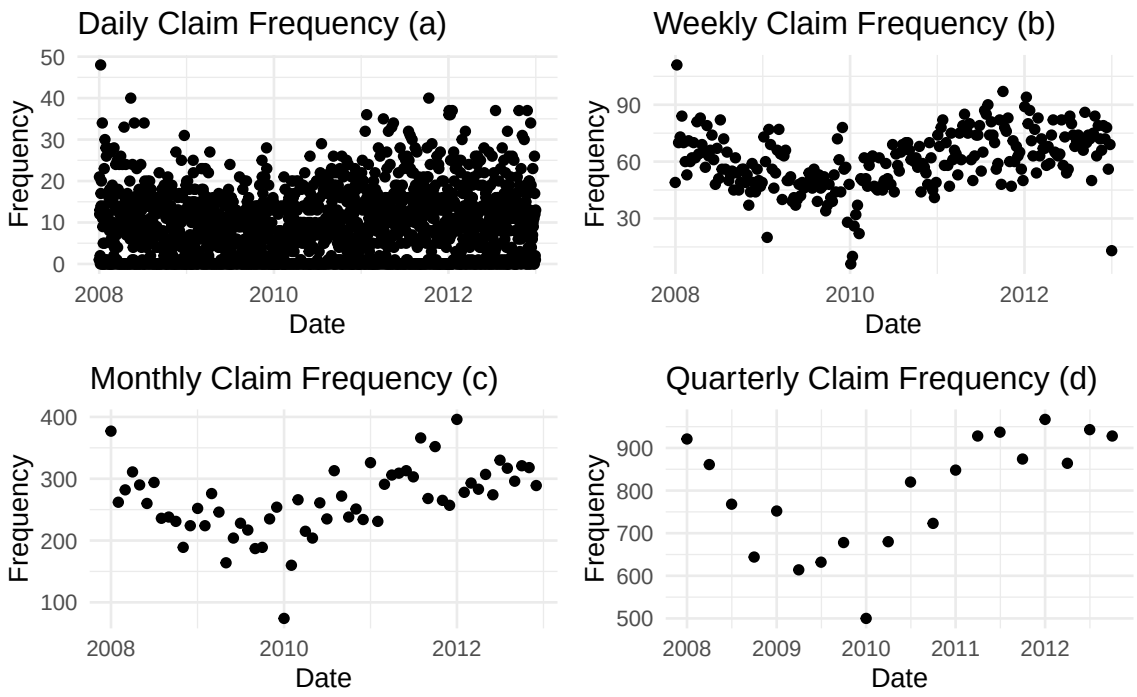

Figure 4.1: Periodic TSA Claim Frequency Scatter Plot

Figure 4.1 shows the scatter plot for different frequency period lengths. We observed a dip at around 2009 to 2010.

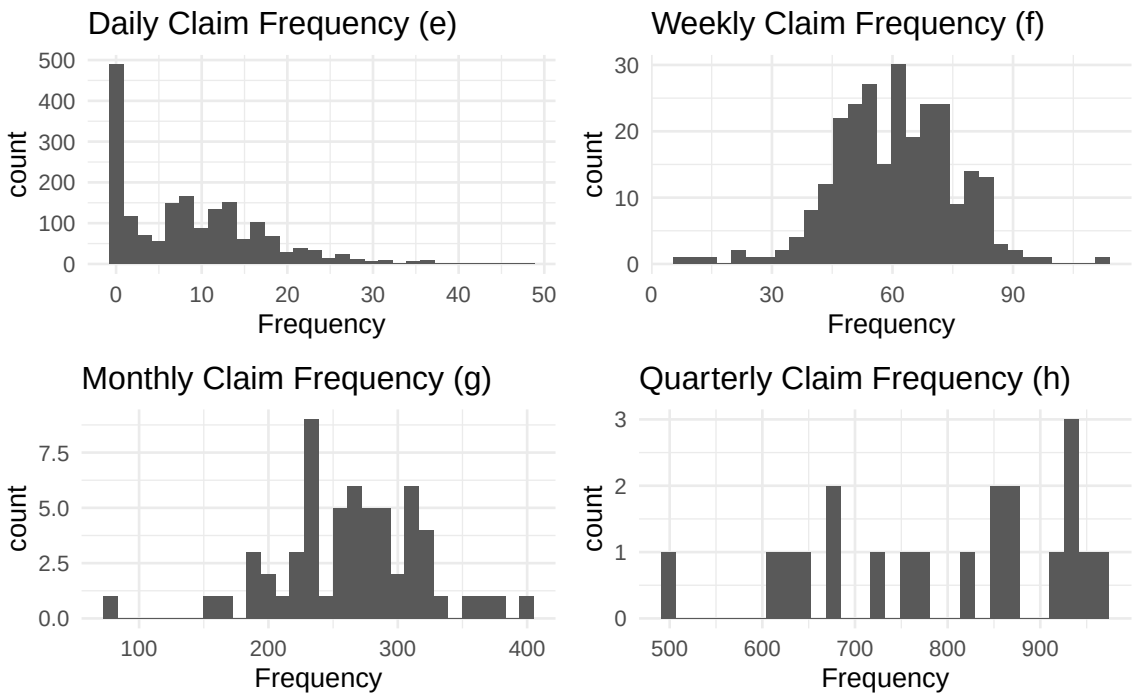

Figure 4.2: Periodic TSA Claim Frequency Histogram

We will mainly focus on monthly data since daily data are highly zero-inflated (a very large number of periods with no losses as shown in Figure 4.1(e)), weekly data have high autocorrelation at lag 4 (monthly correlation) and quarterly data lack a sufficient number of observations. We use a Poisson-Tweedie distribution PT(a,b,c) to model the number of claims for monthly data.

And now we will discuss loss severity summary statistics listed in Table 4.3.

Table 4.3: Summary Statistics of TSA Claims Severity (USD)

Min.	25th Percentile	Median	Mean	75th Percentile	Max.	SD
1	40	99.99	243.31	246.15	25000	598.39

The minimum claim amount over the given time-period is 1 USD with 3

observations. There are 502 observations with a claim amount of less than 10 USD. These are typically for locks, travel accessories, food, currencies that are lost or damaged. There does not seem to be any reporting threshold and the data source does not indicate any threshold exists.

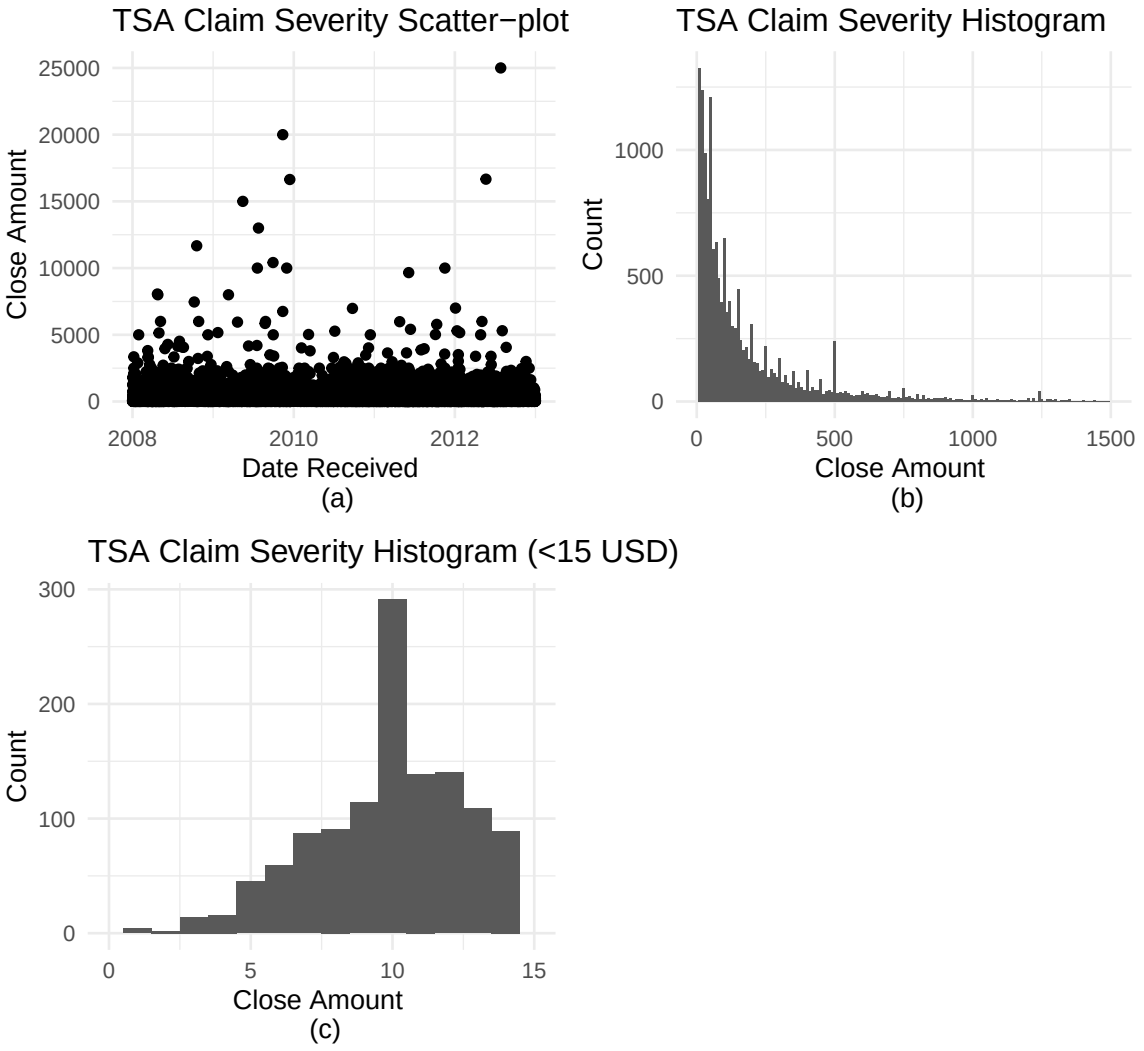

Figure 4.3: TSA Claim Severity

The scatter plot for the closed claim amount (the observed severity in figure 4.3(a)) shows that claim severity has not changed between 2008 and 2012. Figure 4.3(b) shows that the severity is positively skewed, thus we select one of the commonly used

positive skewed distribution Log-Normal(μ,σ) to model the claim amount (Papush, Patrik, and Podgaits 2001; Karam and Planchet 2012; Cummins et al. 1990). Also in the same histogram, we see sharp spikes at every 50 USD increment amounts with a very significant spike at the 500 USD amount. Further analysis shows that at the 500 USD amount, the losses are from damaged or lost personal electronic devices and pieces of jewelry. At the 100 USD amount, the common lost or damaged items are luggage, cosmetics and clothes. We suspect these spikes are related to how common certain items get damaged or lost and their perceived value.

We then apply the Maximum Likelihood Estimation method outlined in (2.28) and 2.29 to estimate parameters, moments and quantiles of aggregate losses.

4.1.2 Estimation of Model Parameters

We estimate the model parameters for the Poisson-Tweedie and Log-Normal distributions by fitting the data with maximum likelihood estimation described in section 2.4.5. We are interested in the family index, parameter a in PT(a,b,c) and the estimated mean and variance. Parameters b and c depend on the parameter a, mean and variance, thus it is more convenient to compare parameter a, mean and variance.

Table 4.4: Fitted Frequency Statistics (Number of Claims)

	$\hat{a}$ (SE)	Frequency Mean	Frequency SE	Frequency Variance
Monthly	-1.14 (0.57)	264.21	58.53	3426.18

In Table 4.4, we provide all the estimates for Poisson-Tweedie parameters in terms of the family index parameter a, the estimated parameters are not close to any commonly used distributions such as Poisson (a=1), Poisson Inverse-Gaussian (a=0.5) and Negative-Binomial (a=0). Thus, Poisson-Tweedie may be more appropriate for

this data than Poisson and Poisson-Inverse Gaussian. The frequency mean, variance and standard error estimated from fitted parameters are very close the sample mean, variance and standard deviations in Table 4.2.

In Table 4.5 we calculate the 95% confidence interval of parameter a using the standard error of the fit.

Table 4.5: $\hat{a}$ 95% Confidence Interval

	$\hat{a}$ 95% Lower Bound	$\hat{a}$ 95% Upper Bound
Monthly	-2.26	-0.03

We estimate 95% confidence interval of parameter a using $\hat{a} \pm C_{rit} * \frac{\text{SE}(\hat{a})}{\sqrt{T}}$ where $\hat{a}$ is the estimate value of parameter a, SE($\hat{a}$) is the standard error of estimated parameter a, T is the total number of periods used in the estimation and C_{rit} is the 95% critical value based on the assumed distribution of the estimated parameter (Normal for $T \geq 50$ and Student t-distribution for $T \leq 50$ based on asymptotic normality of MLE). Poisson, Negative Binomial and PIG distribution values of a are not withing the the 95% confidence interval of the parameter a. This reinforces our assumption that these distributions will not be a good fit for this data set.

Table 4.6: Goodness-of-Fit of Monthly Distribution Fit

	Poisson-Tweedie	Negative Binomial	Poisson Inverse-Gaussian	Poisson
Negative Log-Likelihood	328.87	330.44	334.24	595.25
AIC	663.74	664.88	672.49	1192.50
BIC	670.02	669.07	676.68	1194.59

From Table 4.6, we find that fitting frequency data with Poisson-Tweedie distribution results in the smallest negative log-likelihood for monthly data, and the AIC and BIC show that Poisson Tweedie and Negative Binomial are the best fits.

As mentioned earlier, we will focus our study on monthly data.

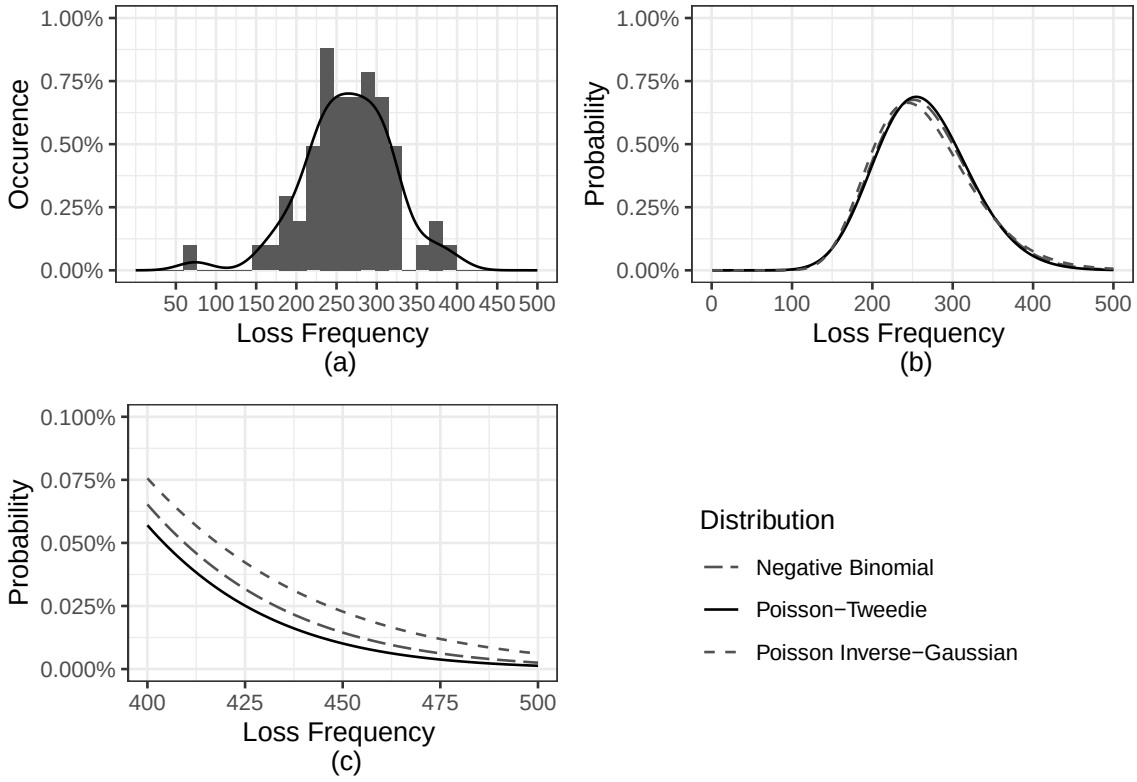

Figure 4.4: Comparison of Estimated Monthly Frequency with Different Distributions

Selecting monthly data for further analysis, we observe that the histogram of the sample loss frequency in Figure 4.4(a) seems to be somewhat symmetric. The sample dispersion (sample variance over sample mean) is 11.76, indicating that Poisson distribution is not a good fit. From Figure 4.4(b) and (c), we observe that the fitted Negative Binomial and Poisson Inverse-Gaussian are slightly more right-skewed than Poisson-Tweedie. Based on the results from Table 4.6, this implies that fitting with Poisson-Tweedie is more effective when the data is symmetric.

Table 4.7: Fitted Statistics of TSA Claims Frequency

	Sample Mean	Sample SE	Sample Variance	$\hat{\bar{N}}$	$SE(\hat{N})$	$var(\hat{N})$	95% $VaR(\hat{N})$
Monthly	264.7	55.81	3114.79	264.21	58.53	3426.18	366

From Table 4.7, we observe that the estimated mean and variance matches the sample mean and variance and implies that our proposed method can estimate the first two moments accurately. The 95% value at risk of the frequency means that 95% of the time there will be less than 366 claims per month. This may help management in allocating manpower to deal with the claims.

We consider estimating the aggregate loss parameters with Log-Normal distribution as mentioned Section 2.1. We also consider the Lomax distribution which is a special case of the Pareto Type II distribution with density

$$f_X(x;\alpha,\lambda) = \frac{\alpha\lambda^\alpha}{(x+\lambda)^{\alpha+1}}$$

for $x \geq 0$, $\alpha > 0$ and $\lambda > 0$.

Additional we look at the Gamma distribution defined as

$$f_X(x;k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^\alpha \Gamma(\alpha)}$$

for $x > 0$, and $a, \theta > 0$. The results are listed in Table 4.8.

Table 4.8: Fitted Statistics of Severity Data

	First Parameter Estimate	Second Parameter Estimate	AIC	BIC
Log-Normal(μ, σ)	4.59	1.31	199579.2	231339.2
Lomax(α, λ)	2.01	247.35	200517.6	232277.6
Gamma(k, θ)	0.38	1471.65	210687.7	242447.7

Based on our goodness-of-fit, we select the log-normal distribution for further analysis as this distribution has the lowest AIC and BIC, which implies it has the best performance.

4.1.3 Quantile Estimation of Monthly Aggregate Loss

The aggregate loss distribution of monthly data is estimated using the fitted distributions Poisson-Tweedie(-1.37,3.68,0.9) and Log-Normal(4.59,1.31). We apply Monte-Carlo simulation to create 100,000 months to approximate the aggregate loss distribution as shown in Figure 4.5.

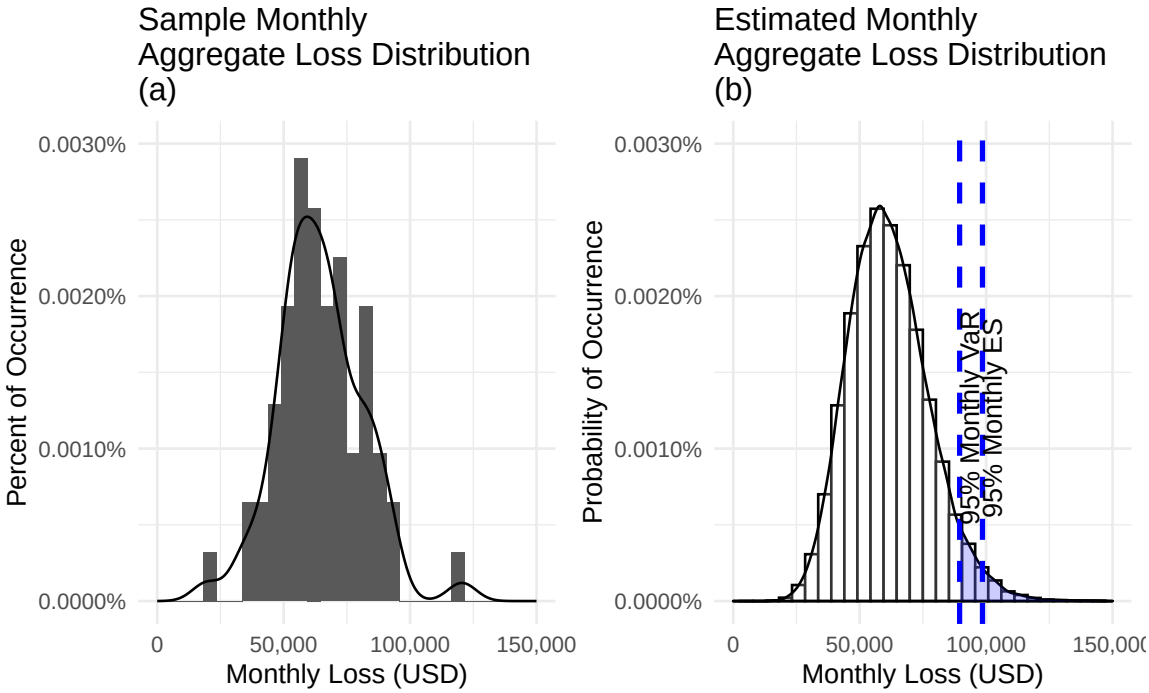

Figure 4.5: Estimated Monthly Loss of TSA Claims (Aggregated by Month)

We find that the estimated monthly aggregate loss distribution has mean of 61,616.43 and standard error of 15,864.73. The estimated 0.95 monthly Value at Risk (VaR) is 89,533.42 USD and the estimated 0.95 monthly Expected Shortfall (ES) is 98,570.69 USD. We also estimated VaR and ES directly with kernel density

estimation and obtained VaR of 90,140.09 and ES of 96,838.23 which is very close to the estimate of our proposed method. The kernel density estimate for VaR is higher than our proposed method and the kernel density estimate for ES is lower. However, with only 60 observations issues with boundaries that may affect the tail estimates and affect the validity of the kernel estimates. Here, we demonstrated the application of our proposed model with Poisson-Tweedie loss frequency on real data.

4.2 Analysis of Reporting Threshold for TSA Claims Data

As mentioned earlier, the TSA claim data is the full data without any truncation or reporting threshold. We are also able to analyze the effect of incomplete observations with this data set by introducing reporting thresholds. We can choose \$10, \$20 and \$30 as the claim threshold to simulate data with incomplete data since typically only the threshold is known and the percent of reporting threshold is unknown. These thresholds corresponds with 3.16%, 11.26%, and 18.88% data removed respectively. This means that the injured party, in this case, travellers that incurred personal or property damage due to TSA, cannot make a claim under the threshold amount and we do not observe any claims under the threshold amount. In our application, we remove claims under the specified amount to create the reporting threshold. Here, we will examine the performance of the estimation method in 2.4.5 for analyzing incomplete data. In particular we would like to use the method in 2.32 to estimate the underlying severity distribution first by maximizing the objective function:

$$(\hat{\mu}, \hat{\sigma}) = \arg\max_{(\mu,\sigma)} \sum_{i=1}^{T} \sum_{j=1}^{n_T} \log \frac{f_X(x_j; \mu, \sigma)}{1 - F_X(x_j; \mu, \sigma)}$$

where $f_X(x; \boldsymbol{\beta}))$ is the assumed severity distribution, $p_H = P(X < h) = F_X(H; \boldsymbol{\beta})$ and $\sum_{i=1}^{T} n_i$ is the total number of observed claims with reporting threshold.

Note p_H is usually unknown in our model and needs to be estimated. In this

data set, however, we can compute the actual observations removed by the chosen reporting threshold.

4.2.1 Parameter Estimation of Monthly Data

While estimating p_H using the method in (2.33), the percent of data missing, we find that this value is higher by around 2% than the actual removed amount with results in Table 4.9. This means that we would expect to obtain higher mean, variance and right tail estimates.

Table 4.9: Estimate of Removed data

	Actual p_H % From Sample Estimation	Estimated $\hat{p_H}$ % From Parameter Estimation
Threshold at 10 USD	3.161	5.762
Threshold at 20 USD	11.264	13.646
Threshold at 30 USD	18.877	20.189

Table 4.10: Severity Parameter Estimation

	$\hat{\mu}$	$\hat{\sigma}$
Full Data	4.59	1.31
Naive Estimation		
10 USD Threshold	4.68	1.25
20 USD Threshold	4.87	1.12
30 USD Threshold	5.02	1.04
Accounting for Incomplete Data		
10 USD Threshold	4.50	1.40
20 USD Threshold	4.51	1.38
30 USD Threshold	4.54	1.37

From Table 4.10 we observe that while ignoring the reporting threshold, we observe that $\hat{\mu}$ is higher than full data estimate while $\hat{\sigma}$ is lower than full data estimate when there is a reporting threshold. However, when we account for the reporting threshold, the opposite occurs. This effect can be seen in Figure 4.6. While ignoring the reporting threshold, we obtain lower loss severity mean and lighter right tail. While accounting for the reporting threshold, the resulting distribution has a higher mean and heavier right tail. The difference between full data estimate and naive estimate is intensified when the reporting threshold increases.

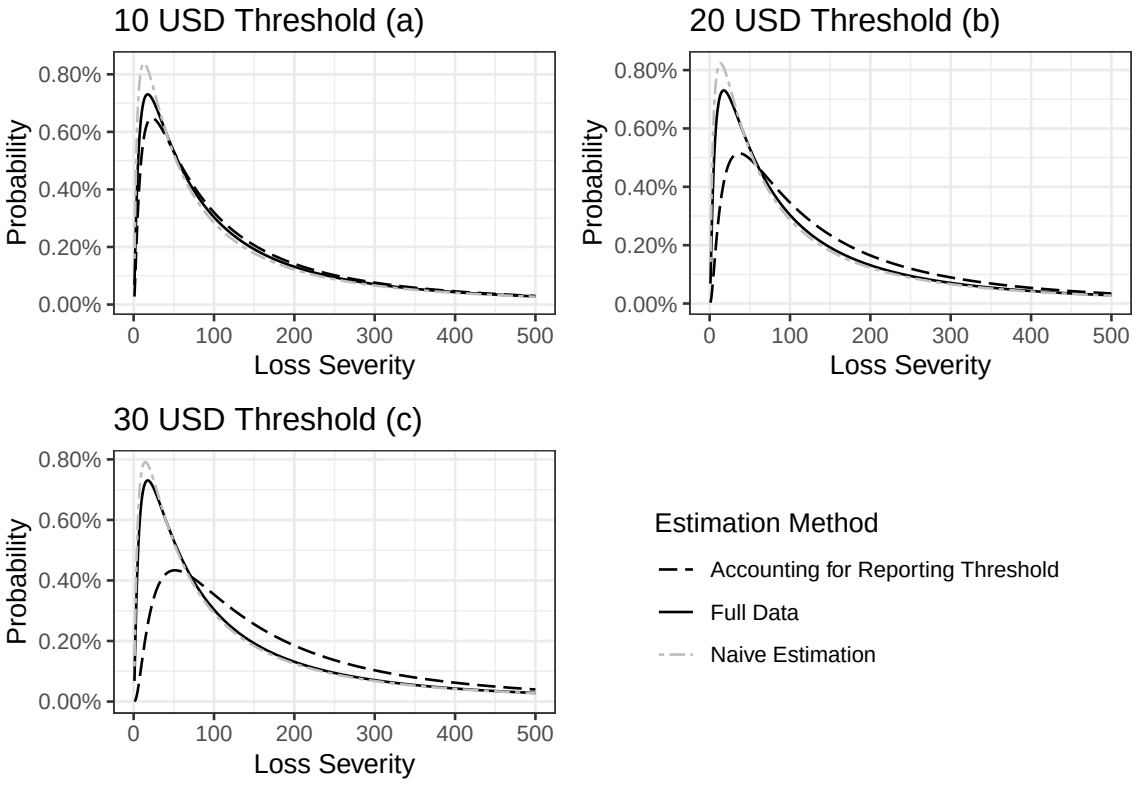

Figure 4.6: Comparison of Estimated Severity Distribution with Full and Incomplete Data

We proved that Poisson-Tweedie is closed under the binomial thinning process. When the underlying distribution is PT(a,b,c), under Binomial-Thinning, the resulting distribution is $PT(a, b\Big(1 - c \cdot p_H\Big)^a, \frac{c(1-p_H)}{1-c \cdot p_H})$, where $p_H = P(X < h) = F_X(h)$ is the

percent of removed claims. We can apply this estimation method and compare it with naive estimation.

Table 4.11: Comparison of Monthly Frequency Parameter Estimation

	Parameter a	Parameter b	Parameter c
Full Data	-1.14	5.48	0.85
Naive Estimation			
10 USD Threshold	-1.52	4.58	0.81
20 USD Threshold	-1.79	4.05	0.79
30 USD Threshold	-1.47	4.92	0.80
Accounting for Incomplete Data			
10 USD Threshold	-1.22	5.33	0.84
20 USD Threshold	-1.28	5.01	0.84
30 USD Threshold	-1.23	4.94	0.85

From Table 4.11, we find that we underestimate $\hat{a}$ with and without accounting for the reporting threshold. The naive estimate greatly underestimates parameter a and the method which accounts for the reporting threshold greatly improves the estimate.

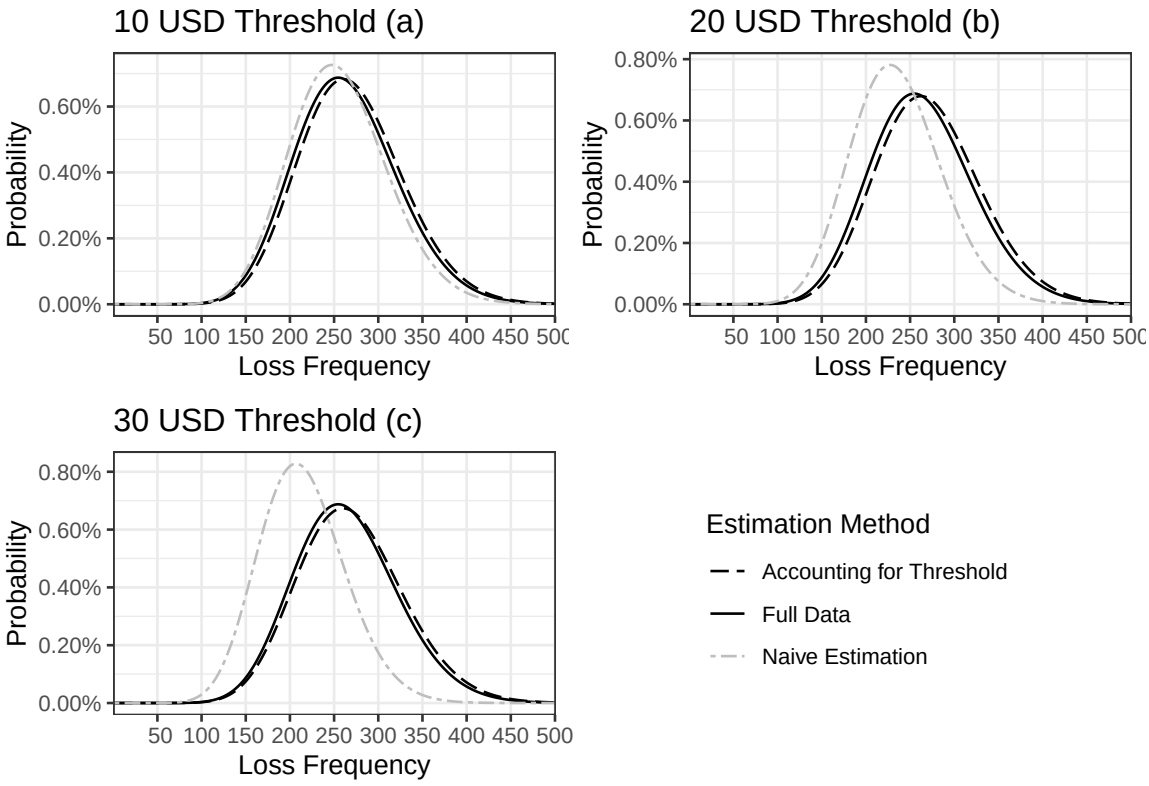

Figure 4.7: Comparison of Estimated Monthly Frequency Distribution with Full and Incomplete Data

Table 4.12: Comparison of Monthly Frequency Summary Statistics

	Mean	Variance
Full Data	264.70	3114.79
Sample		
10 USD Threshold	256.33	2907.85
20 USD Threshold	234.88	2533.39
30 USD Threshold	214.73	2143.11
Naive Estimation		
10 USD Threshold	256.48	3072.12
20 USD Threshold	235.21	2645.38
30 USD Threshold	214.63	2359.99
Accounting for Incomplete Data		
10 USD Threshold	271.95	3500.17
20 USD Threshold	271.56	3501.05
30 USD Threshold	268.38	3544.46

From Table 4.12 we find that with naive estimation, our model underestimates the frequency mean and variance which makes sense because the sample mean and variance of the binomial thinned frequency are smaller than the mean and variance of the full frequency data. When we account for the reporting threshold, the estimated mean is fairly close to the sample mean and we estimate a larger variance than full data estimation. This gives us a fatter tail and a higher estimate of right tail quantile estimates.

4.2.2 Quantile Estimation of Monthly Aggregate Loss

We apply Monte-Carlo simulation with frequency parameters from Table 4.11 and severity parameters from Table 4.9 to estimate distributions of incomplete data at

various levels. The aggregate losses are approximated with 100,000 periods for all levels of reporting threshold with and without accounting for the threshold.

Table 4.13: Summary Statitics of Estimated Monthly Aggregate Loss Under Different Reporting Thresholds and Estimation Methods

	Min.	25th Percentile	Median	Mean	75th Percentile	Max.	SE($\hat{L}$)
Full Data	15,239.37	50,368.88	60,446.78	61,616.43	71,464.31	230,022.65	15,864.73
Naive Estimation							
10 USD Threshold	11,466.16	49,235.00	58,671.93	59,715.66	69,051.55	147,106.48	14,841.40
20 USD Threshold	10,694.76	47,547.73	56,452.26	57,277.85	66,092.45	145,880.75	13,830.34
30 USD Threshold	12,829.48	46,358.03	55,196.33	56,082.03	64,842.38	132,546.82	13,780.36
Accounting for Reporting Threshold							
10 USD Threshold	13,345.88	53,322.64	64,019.59	65,476.18	75,996.46	214,906.84	17,216.74
20 USD Threshold	12,330.39	52,701.15	63,331.72	64,631.35	74,992.83	224,264.45	16,887.44
30 USD Threshold	10,878.22	52,140.00	62,699.48	64,022.39	74,429.95	218,821.43	16,801.29

From Table 4.13 we find that using naive estimation, all quantile levels, standard error and variance are lower than estimates for full data. The higher the level threshold, the greater the difference between estimates using incomplete data and using full data. While taking into account for reporting threshold, our estimates for quantiles larger than 50%, mean and variance with incomplete data are greater than the estimates with full data. Part of this may be due to our overestimation of p_H. The effect of this overestimation is then amplified through subsequent parameter estimations which cause this overestimation. These estimates accounting for reporting threshold does not seem to be affected by the level of the threshold.

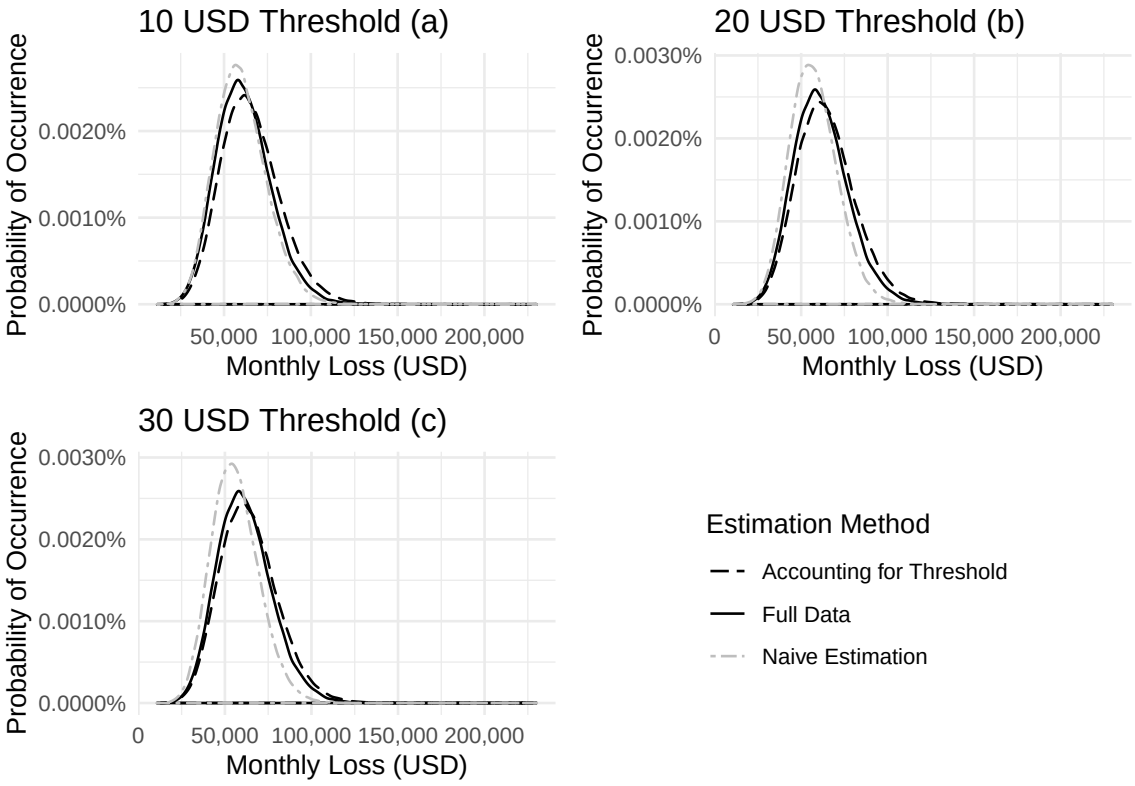

Figure 4.8: Comparison of Estimated Aggregate Loss Distribution with Full and Incomplete Data

Figure 4.8 shows the difference between aggregate loss distribution estimation with full data, naive estimation with incomplete data and estimation which accounts for incomplete data. The estimated distributions of the aggregate loss accounting for the reporting threshold have higher mean and heaver right tail than the distribution for full data. The opposite is true for estimated distributions ignoring the threshold. The intensity of this effect increases as the threshold increases. Overall while accounting for the reporting threshold, the shape of the aggregate loss distribution is close to estimating with full data and stable for different levels of reporting threshold, that is, the shape of the aggregate loss distribution does not change with the threshold level.

Figure 4.8 shows the difference between aggregate loss distribution estimation with full data, naive estimation with incomplete data and estimation which accounts

for incomplete data. The estimated distributions of the aggregate loss accounting for the reporting threshold have higher mean and heaver right tail than the distribution for full data. The opposite is true for estimated distributions ignoring the threshold. The intensity of this effect increases as the threshold increases. Overall while accounting for the reporting threshold, the shape of the aggregate loss distribution is close to estimating with full data and stable for different levels of reporting threshold, that is, the shape of the aggregate loss distribution does not change with the threshold level.

Chapter 5

Conclusion

The aggregate loss model with the Poisson-Tweedie frequency family extends existing candidates of loss frequency, and such extension would reduce the chance of frequency misspecification. For this proposed aggregate loss model we use MLE to estimate parameters and conduct simulations to investigate various concerns. For observations with a reporting threshold, where observations smaller than the threshold are not reported, the loss frequency experiences a binomial thinning process. Another benefit of using Poisson-Tweedie frequency in aggregate loss is that this distribution family is closed under binomial thinning.

From our limited simulation, we observe that different frequency distribution contributes differently to the percentile estimate of the aggregate loss. The effect of the frequency distribution on the aggregate loss is different at each percentile level and for each frequency mean level. The severity parameters have some interaction with how frequency distribution contributes to the aggregate loss, however, more investigation is required to determine the exact effect.

During the simulation, we found that the Poisson-Tweedie algorithm tends to underestimate parameter a when the frequency average is low (around 4). We suspect that this may be due to many values of the generated random variable are 0 during

the simulation and lower values of the Poisson-Tweedie parameter a corresponds with higher zero-inflated distributions. These two properties may have caused the optimization algorithm to overestimate how zero-inflated the data is.

In the limited simulation study of parameter estimation with Log-Normal severity and Poisson-Tweedie frequency, we find that the first two moments of frequency can be estimated without bias. However, the Poisson-Tweedie family index parameter a seems to have a bias. When theoretical parameter a is less than -1, overestimation occurs. This would cause aggregate loss right tail percentiles to be overestimated. When theoretical parameter a is greater than -1, underestimation occurs which causes the aggregate loss right percentiles to be underestimated. With the reporting threshold, directly estimating the frequency will underestimate the mean, variance and parameter a. Higher levels of the reporting threshold have a greater impact on the underestimation. The method in section 2.4.3 can estimate the missing data percent, frequency mean and severity parameters without bias. Higher threshold levels will contribute to higher variance (calculated with simulated sample). Bias is still present in Poisson-Tweedie MLE estimation of parameter a. More simulation can be performed to determine if any further bias or technical issues exist.

The reporting threshold levels in the simulation are specified so that unreported data are less than 40%. We do so because that higher threshold levels will introduce more uncertainty and increase variance. For very high thresholds, only rare events are observed. This creates problems for estimation due to a lack of observations.

While estimating data, we found a technical issue with this algorithm. When the frequency mean is around 12,000 and the dispersion index (variance divided by mean) is around 1200, some levels of the Poisson-Tweedie parameters results in the estimate of $P(N = 0)$ is rounded to 0 which cause all subsequent estimates $P(N = 1)$, $P(N = 2)$, $P(N = 3)$, $\cdots$ to also be 0. Therefore, alternative approaches to calculating the probability of Poisson-Tweedie mass is in future consideration.

The application of our aggregate loss model with Poisson-Tweedie frequency

on the TSA Claims data suggest the frequency distribution is not the commonly used Poisson, Negative Binomial or Poisson Inverse-Gaussian distribution at daily, weekly, monthly and quarterly periods. Furthermore, we introduced data incompleteness by removing claims under specified severity thresholds. Directly estimating the distribution using MLE shows bias and the frequency mean and variance is underestimated. Our result using the method from (2.34) gives conservative estimates (similar mean and greater variance) at all levels of missing data for monthly period loss frequency. The estimated frequency parameters accounting for the reporting threshold is closer to the full data estimated parameters than naive estimation.

Due to our limited simulation study, we were not able to fully capture relationships between loss severity, loss frequency and the aggregate loss. Future research can include studying the effect of different loss frequency distribution on aggregate loss percentile estimates at high levels of frequency mean. This relationship was not completely captured as the maximum frequency mean we had chosen was 30. Additional frequency parameters can be used to help determine any interaction between severity distribution and how different aggregate loss distributions impact the aggregate loss percentile estimates. For parameter estimation, future studies may include performing simulation studies with additional parameter levels to further determine the behaviour of the parameter a bias.

We find that the aggregate loss model with the Poisson-Tweedie frequency family can be applied to real-world data to estimate aggregate percentile statistics of interest. We believe our proposed model can aid users in banking and insurance by providing more flexibility in estimating aggregate loss model parameters.

References

Bonat, W., B. Jørgensen, C. Kokonendji, J. Hinde, and C. Demétrio. 2016. "Extended Poisson–Tweedie: Properties and Regression Models for Count Data." *Statistical Modelling: An International Journal* 18 (August). https://doi.org/10.1177/1471082X17715718.

Cummins, J., G. Dionne, J. Mcdonald, and B.M. Pritchett. 1990. "Application of Gb2 Family of Distribution in Modeling Insurance Losses Processes." *Insurance: Mathematics and Economics* 9 (December): 257–72.

EBA. 2019. "Policy Advice on the BASEL III Reforms: Operational Risk." https://eba.europa.eu/eba-advises-the-european-commission-on-the-implementation-of-the-final-basel-iii-framework.

El-Shaarawi, A. H., R. Zhu, and H. Joe. 2011. "Modelling Species Abundance Using the Poisson–Tweedie Family." *Environmetrics* 22 (2): 152–64.

Griffiths, R., and W. Mnif. 2017. "Various Approximations of the Total Aggregate Loss Quantile Function with Application to Operational Risk." *The Journal of Operational Risk* 12 (June). https://doi.org/10.21314/JOP.2017.191.

Heckman, G., P.and Meyers. 1983. "The Calculation of Aggregate Loss Distributions from Claim Severity and Claim Count Distributions." In *Proceedings of the Casualty Actuarial Society*, 70:22–73.

Horbenko, N., P. Ruckdeschel, and T. Bae. 2011. "Robust Estimation of Operational

Risk." *Journal of Operational Risk* 6 (June): 3–30.

Jin, T., S. Provost, and J. Ren. 2014. "Moment-Based Density Approximations for Aggregate Losses." *Scandinavian Actuarial Journal*, August. https://doi.org/10.1080/03461238.2014.921640.

Karam, E., and F. Planchet. 2012. "Operational Risks in Financial Sectors." *Advances in Decision Sciences* 2012 (December). https://doi.org/10.1155/2012/385387.

Kerwer, D. 2005. "Rules That Many Use: Standards and Global Regulation." *Governance-an International Journal of Policy Administration and Institutions - GOVERNANCE-INT J POLICY ADM I* 18 (October): 611–32.

Klugman, S.A., H.H. Panjer, and G.E. Willmot. 2012. *Loss Models: From Data to Decisions*. Wiley Series in Probability and Statistics. Wiley. https://books.google.ca/books?id=8aBQP/_CGGBMC.

Kokonendji, C., C. Demétrio, and S. Dossou Gbete. 2004. "Overdispersion and Poisson-Tweedie Exponential Dispersion Models." *VIII Journées Zaragoza-Pau de Mathématiques Appliquées et de Statistiques* 31 (January): 365–74.

Panjer, H. H. 2014. "Aggregate Loss Modeling." In *Wiley StatsRef: Statistics Reference Online*. American Cancer Society. https://doi.org/10.1002/9781118445112.stat04301.

Panjer, H.H. 2006. *Operational Risk: Modeling Analytics*. Wiley Series in Probability and Statistics. Wiley.

Papush, D. E., G. S. Patrik, and F. Podgaits. 2001. "Approximations of the Aggregate Loss Distribution." Las Vegas, Nevada: 2001 Winter Forum, Ratemaking Discussion Papers; Data Management/Data Quality/Data Technology Call Papers.

Shevchenko, P.V. 2011. *Modelling Operational Risk Using Bayesian Inference*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-15923-7.

Smyth, Gordon K., and Bent Jørgensen. 2002. "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling." *ASTIN Bulletin* 32 (1): 143–57. https://doi.org/10.2143/AST.32.1.1020.

Willmot, G. 1987. "The Poisson-Inverse Gaussian Distribution as an Alternative to the Negative Binomial." *Scandinavian Actuarial Journal* 1987 (July): 113–27. https://doi.org/10.1080/03461238.1987.10413823.

Zhu, R. 2002. "On Continuous-Time Generalized Ar(1) Processes: Models, Statistical Inference, and Applications to Non-Normal Times Series." PhD thesis, University of British Columbia.

Appendix A

Derivation of Probability Generating Function Under Binomial Thinning

The probability generating function of Bernoulli random variable I with $p = 1 - p_H$ is given by

$$G_I(s) = p_H + (1 - p_H)s.$$

We clarify the proof in Shevchenko (2011), page 191. The probability generating function of N_H is then

$$\begin{aligned} G_{N_H}(s) &= \sum_{k=0}^{\infty} \Pr[N_H = k] s^k \\ &= \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} \Pr[I_1 + I_2 + \cdots + I_n = k | N = n] \Pr[N = n] \right) s^k \\ &= \Pr[N = 0] + \Pr[I_1 = 0 | N = 1] \Pr[N = 1] + \cdots \\ &\qquad + \Pr[I_1 = 1 | N = 1] \Pr[N = 1] s + \cdots \end{aligned}$$

$$
\begin{aligned}
&= \Pr[N=0] \\
&\quad + (\Pr[I_1=0|N=1]\Pr[N=1] + \Pr[I_1=1|N=1]\Pr[N=1]s) \\
&\quad + \dots \\
&= \sum_{n=0}^{\infty} \Pr[N=n] \left(\sum_{k=0}^{n} \binom{n}{k} \cdot (1-p_H)^k p_H^{n-k} s^k \right) \\
&= \sum_{n=0}^{\infty} \Pr[N=n] (p_H + (1-p_H)s)^n \\
&= \sum_{n=0}^{\infty} \Pr[N=n] (G_I(s))^n \\
&= G_N(G_I(s)),
\end{aligned}
$$

where $G_N(s)$ is the probability generating function of the count distribution.

Derivation of Poisson-Tweedie Algorithm Under Binomial Thinning

Following El-Shaarawi, Zhu, and Joe (2011), we find the probability mass function for the reported frequency N_H.

Let

$$B_0 = \frac{b}{a}(1 - c \cdot p_H)^a, \quad B_1 = \frac{c(1-p_H)}{1 - c \cdot p_H}.$$

Then

$$G_{N_H}(s) = \exp\{B_0((1-B_1)^a - (1-B_1 s)^a)\}.$$

Taking partial derivative of $\log G_{N_H}(s)$ with respect to s gives

$$
\begin{aligned}
\log(G_{N_H}(s)) &= B_0((1-B_1)^a - (1-B_1 s)^a) \\
\frac{\partial \log(G_{N_H}(s))}{\partial s} &= -B_0 a (1-B_1)^{a-1}(-B_1) = a B_0 B_1 (1-B_1 s)^{a-1}.
\end{aligned}
$$

Apply the chain rule,

$$\frac{\partial \log(G_{N_H}(s))}{\partial s} = \frac{\partial G_{N_H}(s)}{\partial s}\frac{1}{G_{N_H}(s))} = \frac{G'_{N_H}(s)}{G_{N_H}(s)},$$

to obtain the following equation:

$$\frac{G'_{N_H}(s)}{G_{N_H}(s)} = aB_0B_1(1-B_1s)^{a-1}.$$

That is

$$aB_0B_1G_{N_H}(s) = (1-B_1s)^{a-1}G'_{N_H}(s).$$

Let

$$A_0 = aB_0B_1, \quad G_{N_H}(s) \ = p_0 + p_1s + p_2s^2 + p_3s^3 + \cdots .$$

Then we have

$$\begin{aligned} G'_{N_H}(s) &= p_1 + 2p_2s + 3p_3s^2 + \cdots \\ aB_0B_1G_{N_H}(s) &= A_0G_{N_H}(s) = A_0[p_0 + p_1s + p_2s^2 + p_3s^3 + \cdots] \\ &= A_0p_0 + A_0p_1sA_0 + p_2s^2 + A_0p_3s^3 + \cdots . \end{aligned}$$

Also

$$\begin{aligned} (1-B_1s)^{a-1} &= 1 - (1-a)B_1s - \frac{(1-a)a}{2!}B_1^2s^2 - \frac{(1-a)a(1+a)}{3!}B_1^3s^3 - \cdots \\ &= 1 - r_1s - r_2s^2 - r_3s^3 - \cdots , \end{aligned}$$

where

$$r_1 = (1-a)B_1, \quad r_{j+1} = (\frac{j-1+a}{j+1})B_1r_j, \quad j = 1, 2, \ldots .$$

Then

$$
\begin{aligned}
(1-B_1s)^{a-1}G'_{N_H}(s) &= [1-r_1s-r_2s^2-r_3s^3-\cdots]\cdot[p_1+2p_2s+3p_3s^2+\cdots] \\
&= (p_1+2p_2s+3p_3s^2+\cdots) \\
&\quad -r_1s(p_1+2p_2s+3p_3s^2+\cdots) \\
&\quad -r_2s^2(p_1+2p_2s+3p_3s^2+\cdots) \\
&\quad -\cdots \\
&= p_1+(2p_2s-r_1p_1s)+(3P_3s^2-r_12p_2s^2-r_2p_1s^2)+\cdots
\end{aligned}
$$

Equating $aB_0B_1G_{N_H}(s) = (1-B_1s)^{a-1}G'_{N_H}(s)$, we obtain

$$A_0p_0+A_0p_1sA_0+p_2s^2+A_0p_3s^3+\cdots = p_1+(2p_2s-r_1p_1s)+(3P_3s^2-r_12p_2s^2-r_2p_1s^2)+\cdots .$$

From $G_{N_H}(S)$, we have

$$p_0 = G_{N_H}(0) = \exp\{B_0[1-B_1]^a-1\} \quad \text{when} \quad a\neq 0.$$

When $a=0$, applying the L'Hospital rule

$$
\begin{aligned}
&\lim_{a\to 0}\frac{b}{a}(1-c\cdot p_H)[(1-B_1)^a-(1-B_1s)^a)] \\
&\quad = (1-c\cdot p_H)b\lim_{a\to 0}\frac{(1-B_1)^a-(1-B_1s)^a}{a} \\
&\quad = (1-c\cdot p_H)b\lim_{a\to 0}\frac{(1-B_1)^a\log(1-B_1)-(1-B_1s)^a\log(1-B_1s)}{1} \\
&\quad = (1-c\cdot p_H)b[\log(1-B_1)-\log(1-B_1s)].
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\lim_{a\to 0}G_{N_H}(s) &= \exp\{(1-c\cdot p_H)b[\log(1-B_1)-\log(1-B_1s)]\} \\
&= (\frac{1-B_1}{1-B_1s})^{(1-c\cdot p_H)b},
\end{aligned}
$$

and

$$
p_0 = \begin{cases} \exp\{B_0[1-B_1]^a-1\} & a\neq 0, \\ (1-B_1)^{(1-c\cdot p_H)b} & a=0. \end{cases}
$$

Equating the terms of s in $A_0 G_{N_H}(s) = (1 - B_1 s)^{a-1} G'_{N_H}(s)$, we obtain $p_1 = A_0 p_0$ and

$$
\begin{aligned}
A_0 p_k s^k &= ((k+1)p_{k+1} - r_1 p_{k-1} - r_2 p_{k-2} - \cdots)s^k \\
A_0 p_k &= (k+1)p_{k+1} - \sum_{j=1}^{k} j r_{k+1-j} p_j \\
p_{k+1} &= \frac{1}{k+1}(A_0 p_k + \sum_{j=1}^{k} j r_{k+1-j} p_j), \quad k = 1, 2, 3, \ldots .
\end{aligned}
$$

We obtain a recursive algorithm for estimating the probability of Poisson-Tweedie under binomial thinning.

Appendix B

Core Code for Simulation and Estimation

We program the core algorithm for Poisson-Tweedie probability mass function in C++ below.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
NumericVector dPTzero(const int & x,
                      const double & a,
                      const double & b,
                      const double & c){
  if (x < 0) stop("Error: x must be non-negative integer");
  double pzero= pow((1-c) , b);
  if (a != 0) pzero=exp(b*( pow ((1-c) , a)-1)/a);
  NumericVector p (x+1);
  p[0]=pzero;
  if (x > 0){
    double pinit=b*c*pzero;
    p[1] = pinit;
    if (x > 1) {
      NumericVector r( x );
      r[0]=(1-a)*c;
      for(int k=1; k<x; k++){
        r[k]=(((double)k+1)-2+a)/((double)k+1)*c*r[k-1];
      }
      double temp;
      for( int i=1; i<x; i++ ){
```

```
        temp = 0;
        for( int j=0;j<i;j++ ){
          temp = temp + ( (double)j + 1 )*r[i-1-j]*p[j+1];
        }
        p[i+1] = ( 1/( (double)i + 1)*(b*c*p[i] + temp) );
      }
    }
  }
  return p;
}
```

The core algorithm is loaded into R and called by wrapper functions. Functions for density, probability, quantile and random number generation are programmed in R.

```
#density
dPT=function(x=0,a=0.5,b=1,c=0.5){
  p=dPTzero(max(x),a,b,c)
  return(p[x+1])
}

# probability
pPT=function(q,a=0.5,b=1,c=0.5){
  if(min(q)<0){
    return("error: q must be non-negative integer")
  }
  else{
    p=dPTzero(max(q),a,b,c)
    out=cumsum(p)
    return(out[q+1])
  }
}

#quantile

qPT=function(p,a=0.5,b=1,c=0.5){
  if(p<0 || p>1){
    return("error: p must be a probability")
  }
  else{
    mu=b*c/(1-c)^(1-a)
```

```
    k.upper=ceiling(mu/(1-max(p)))
    p.dens=dPTzero(k.upper,a,b,c)
    p.cumu=cumsum(p.dens)
    out.rand=NULL
    for(i in 1:length(p)){
      out.rand=c(out.rand,sum(p.cumu<=p[i]))
    }
    return(out.rand)
  }
}

# random number generator
rPT=function(n,a=0.5,b=1,c=0.5){
    rand.unif=runif(n)
    r.max=max(rand.unif)
    p0=0
    j=0
    while(r.max>p0){
        p.dens=dPTzero(j,a,b,c)
        p0=sum(p.dens)
        j=j+1
    }
    p.cumu=cumsum(p.dens)
    countcompare=function(x){
        return(sum(p.cumu<=x))
    }
    out.rand=sapply(rand.unif,countcompare)
    return(out.rand)
}
```

Similar as above, we create density and probability functions for Poisson-Tweedie under binomial thinning given earlier in Appendix A.

```
# core algorithm for PT under Binomial Thinning
dPTrunc.0=function(x=0,a=0.5,b=1,c=0.5,fl=0.3){
  b.trunc=b*(1-c*fl)^a
  c.trunc=(c*(1-fl))/(1-c*fl)
  dPTzero(x,a,b.trunc,c.trunc)
}

# density
dPTrunc=function(x=0,a=0.5,b=1,c=0.5,fl=0.3){
```

```
  p=dPTrunc.0(max(x),a,b,c,fl)
  return(p[x+1])
}

#probability
pPTrunc=function(q,a=0.5,b=1,c=0.5,fl=0.3){
  if(min(q)<0){
    return("error: q must be non-negative integer")
  }
  else{
    p=dPTrunc.0(max(q),a,b,c,fl)
    out=cumsum(p)
    return(out[q+1])
  }
}
```

Regarding the aggregate loss model with Log-Normal severity and Poisson-Tweedie frequency, the code for data simulation, model fitting and quantile estimation are listed below.

```
# set parameters
#periods
m=100
#PT parameter
a=1
b=2
c=1
#Log-Norm Parameter
mu=8
sigma=3
#reporting threshold
h=1000
# simulate frequency
sim.freq <- rPT(m, a,b,c)
# simulate severity
sim.severity <- rlnorm(sum(sim.freq),
                       meanlog = mu,
                       sdlog = sigma)
# separate severty into periods
split.severity <- split(sim.severity,
                        rep(1:length(sim.freq),
```

```
                                sim.freq))
# remove data based on threshold
listremovespecial=function(x,h){
if(length(x)>1){
  if(length(x[x>h])>0){
    return(x[x>h])
  }
  else{return(NULL)}
}
if(x>h){
  return(x)
}
else{
  return(NULL)
}
}
# truncate data by threshold h
# to simulate data with reporting threshold
split.cut=lapply(split.severity, listremovespecial,h=h)
cut.freq <-rep(0, m)
cut.freq[as.numeric(names(split.cut))]=lengths(split.cut)
cut.severity=sim.severity[sim.severity>h]

# estimating severity of full data
x.full=sim.severity
fn.sev <- function(theta) {
  if(theta[1]<0 || theta[2]<0){Inf}
  else{
    -sum(log(dlnorm(x, theta[1], theta[2])))
  }
}
init.sev=c(mean(log(x.full)), sd(log(x.full)))
model.sev=optim(init.sev, fn.sev, hessian=TRUE)

# estimating frequency of full data
ni.full=sim.freq

fn.freq <- function(theta) {
  if(theta[1]>1 || theta[2]<=0 || theta[3]<=0 || theta[3]>1){Inf}
  else
  {-sum(log(dPT(ni.full,theta[1],theta[2],theta[3])))}
}

mean.freq=mean(ni.full)
```

```
D.freq=var(ni.full)/mean.freq
init.a=0.5
init.c=(D.freq-1)/(D.freq-init.a)
init.c=max(0.1,min(1,init.c))
init.b=mean.cut2*(1-init.c)^(1-init.a)/init.c
init.b=max(0.1,init.b)
init.freq=c(init.a,init.b,init.c)

model.freq=optim(par = init.freq,
                 fn = fn.freq,
                 gr = NULL,
                 method = "Nelder-Mead",
                 hessian = TRUE
)

#estimating data with reporting threshold
#estimating severity
x.cut=cut.severity

fn.sev <- function(theta) {
if(theta[1]<0 || theta[2]<0){Inf}
else{
  -sum(log(dlnorm(x.cut,theta[1],theta[2])))+
    length(x.cut)*(log(1-plnorm(h,theta[1],theta[2])))
}
}

init.sev.cut=c(mean(log(x.cut)),sd(log(x.cut)))
model.cut.sev=optim(init.sev.cut, fn.sev, hessian=TRUE)
ph=plnorm(h,model.sev.cut$par[1],model.sev.cut$par[2])

#estimating frequency with reporting threshold
ni.cut=cut.freq
fn.cut.freq <- function(theta) {
  if(theta[1]>1 || theta[2]<=0 || theta[3]<=0 || theta[3]>1){Inf}
  else
  {-sum(log(dPTrunc(ni.cut,theta[1],theta[2],theta[3],ph)))}
}

init.a=0.5
init.c=(D.cut2-1)/(D.cut2-init.a)
init.c=max(0.1,min(1,init.c))
init.b=mean.cut2*(1-init.c)^(1-init.a)/init.c
init.b=max(0.1,init.b)
```

```
init.cut.freq=c(init.a,init.b,init.c)

model.cut.freq=optim(par = init.cut.freq,
                          fn = fn.cut.freq,
                          gr = NULL,
                          method = "Nelder-Mead",
                          hessian = TRUE)

# estimating quantile of full data aggregate loss
# using 100000 periods
est.freq <- rPT(100000, model.freq$par[1],
                 model.freq$par[2],
                 model.freq$par[3])
# aggregate sum function
aggsum=function(x,mu,sigma){
  if(x==0){return(0)}
  else{
    return(sum(rlnorm(x,
                      meanlog = mu,
                      sdlog = sigma)))
  }
}
# simulate severity and then aggregate
est.agg=sapply(est.freq,aggsum,
               mu=model.sev$par[1],
               sigma=model.sev$par[2])
# 0.95 VaR
var95=quantile(est.agg,0.95)
# 0.95 ES
es95=mean(quantile(est.agg,seq(0.95, 1, 0.000005)))
```

Loops for simulation study can be run in parallel to save time. The following code demonstrates parallel computing in R.

```
#set total number of repetitions

total=100000

library(doParallel)

# get number of cores, set number depending on hardware
ncores = 4
```

```
# registers the number of cores for parallel processing
registerDoParallel(cores=ncores)
# this how many cores are available, and how many you have requested.
print(ncores)
# you can compare with the number of actual workers
getDoParWorkers()
# file to save data
conn <- file("saveddata.csv", "w")
wtab <- function(conn, d) {
  write.table(d, conn, sep = ",",
              row.names = FALSE,
              col.names = FALSE)
  conn
}
#Parallel Loop, load packages to parallel process as needed
foreach(k=1:total,
        .packages="",
        .init=conn,
        .combine='wtab') %dopar% {
  #run code inside loop
  #return code output in matrix row format and saves output to file
  #by row
  return(codeoutput)
}
```