May 2020

Smoothed Quantiles for Claim Frequency Models, with Applications to Risk Measurement

Ponmalar Suruliraj Ratnam
University of Wisconsin-Milwaukee

Smoothed Quantiles for Claim Frequency Models, with Applications to Risk Measurement

by

Ponmalar Ratnam

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Mathematics

at

The University of Wisconsin-Milwaukee

May 2020

ABSTRACT

Smoothed Quantiles for Claim Frequency Models, with Applications to Risk Measurement

by

Ponmalar Ratnam
The University of Wisconsin-Milwaukee, 2020
Under the Supervision of Professor Vytaras Brazauskas

Statistical models for the claim severity and claim frequency variables are routinely constructed and utilized by actuaries. Typical applications of such models include identification of optimal deductibles for selected loss elimination ratios, pricing of contract layers, determining credibility factors, risk and economic capital measures, and evaluation of effects of inflation, market trends and other quantities arising in insurance. While the actuarial literature on the severity models is extensive and rapidly growing, that for the claim frequency models lags behind. One of the reasons for such a gap is that various actuarial metrics do not possess "nice" statistical properties for the discrete models whilst their counterparts for the continuous models do. The objectives of this dissertation to addressing the issue described above are the following:

1. Generalize the definitions of "smoothed quantiles" for samples and populations of claim counts to vectors of smoothed quantiles. This is motivated by the fact that multiple quantiles are needed for better understanding of insurance risks.

2. Investigate large- and small-sample properties of smoothed quantile estimators for vectors, when the underlying claim count distribution has finite support.

3. Extend the definition of smoothed quantiles for discrete distributions with infinite support, and study asymptotic and finite-sample properties of the associated estimators.

4. Illustrate the appropriateness and flexibility of such tools in solving risk measurement problems.

Smoothed quantiles are defined using the theory of fractional or imaginary order statistics, which was originated by Stigler (1977). To prove consistency and asymptotic normality of sample estimators of smoothed quantiles, we utilize the results of Wang and Hutson (2011) and generalize them to vectors of smoothed quantiles. Further, we thoroughly investigate extensions of this methodology to discrete populations with infinite support (e.g., Poisson and zero-inflated Poisson distributions). Furthermore, large- and small-sample properties of the newly designed estimators are investigated theoretically and through Monte Carlo simulations. Finally, applications of smoothed quantiles to risk measurement (e.g., estimation of distortion risk measures such as value-at-risk, conditional tail expectation, and proportional hazards transform) are discussed and illustrated using actual insurance data.

To

my son

and my family

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my special appreciation and thanks to my advisor Professor Vytaras Brazauskas for the continuous support of my research work and for his patience, motivation, and immense knowledge. After moving to industry for full time job, it would not be possible for me to continue my thesis without Dr. Brazauskas' support and motivation. I would also like to thank Professor Jay Beder, Professor Daniel Gervini, Professor Wei Wei and Professor David Spade for serving as my committee members.

Last but not the least, I would like to thank my family especially my son for all the support and encouragement.

Ponmalar Ratnam

Milwaukee, Wisconsin

March 2020

Chapter 1

Introduction

Insurance is a centuries-old data-driven industry with the main cash in-flow being premiums and main cash out-flow being claim payments. For any country, the insurance industry is of great importance because it is a form of economic remediation. It provides a means of reducing financial loss due to the consequences of risks by spreading or pooling the risk over a large number of policyholders, which results in large and complicated data sets. Actuarial science focuses on building and analyzing statistical and mathematical models for the financial sector data, with the objective to describe the process by which money flows in and out of an organization. It comprises diverse quantitative tools that help one make financial sense of the future in the insurance industry. These models help companies make vital decisions on risk measurement, reserve analysis, provisions for future liabilities, as well as contract pricing and pension planning.

Statistical models for the claim severity and claim frequency variables are routinely constructed and utilized by actuaries. Typical applications of such models include identification of optimal deductibles for selected loss elimination ratios, pricing of contract layers, determining credibility factors, risk and economic capital measures, and evaluation of effects of inflation, market trends and other quantities arising in insurance. The

actuarial literature on the claim frequency models is not as extensive as that for the claim severity models. This is due to the fact that statistical properties of various actuarial measures are relatively easy to prove for the continuous models but not for the discrete models.

1.1 Literature Review

Many attempts have been made in the actuarial literature to find the "best" (or at least in some sense better than existing ones) probabilistic model for the distribution of claim count data. Most of these models are parametric. For example: the Generalized Geometric and Negative Binomial distributions studied by Gossiaux and Lemaire (1981), Willmot (1987), and Besson (1992); the Poisson-Inverse Gaussian distribution discussed by Willmot (1987), Besson (1992), and Tremblay (1992); the Generalized Poisson-Pascal distribution was proposed by Consul (1989) and Islam and Consul (1992); and the Poisson-Goncharov distribution presented by Denuit (1997). Also, Yip and Yau (2005) and Boucher *et al.* (2007) have emphasized the use of parametric distributions other than Poisson to accommodate features of insurance count data that are inconsistent with the Poisson distribution assumption. In particular, these authors employed the negative binomial, zero-inflated and hurdle distributions. In this dissertation, we will develop new methodological tools that will be applicable to all of the distributions mentioned above. Our discussions and illustrations, however, will focus on two broad classes of discrete distributions, $(a, b, 0)$ and $(a, b, 1)$. These classes include the traditional discrete distributions such as binomial, Poisson, and negative binomial, their truncated and zero-inflated variants, as well as other discrete distributions. A short description of the two classes is provided in Chapter 2, and further details are available in Klugman *et al.* (2012, Chapter 6).

There is a large literature on risk measures, their estimation methods, hypothesis testing and risk-based decision making when the underlying loss variable is continuous (see, for example, Jones and Zitikis, 2003, 2005, 2007; Albrecht, 2004; Brazauskas and Kaiser, 2004; Tapiero, 2004; Kaiser and Brazauskas, 2006; Brazauskas *et al.*, 2007, 2008; Furman *et al.*, 2017; Samanthi *et al.*, 2017). When the loss variable is discrete or mixed, however, the definition of risk measures has to be broadened. See a comprehensive study by Acerbi and Tasche (2002) dedicated to two most popular risk measures: *value-at-risk* (VaR) and *conditional value-at-risk* (CVaR). The broadened definitions of those risk measures come at the expense of technically trickier statistical inference. While there were attempts to develop statistical inferential tools for VaR and CVaR based on count variables (see Göb, 2011), much more work needs to be done for these and other risk measures. The techniques presented in this dissertation will help alleviate the existing challenges and will facilitate a straighforward transition from the risk measurement literature of continuous loss variables to that of discrete.

1.2 Actuarial Applications

Discrete probability distributions play an important role in many different types of insurance problems. For example, to design insurance products that are financially manageable and can be priced competitively the insurance company needs to build models for the total payments. The building blocks of such models are random variables that describe the number of claims (N) and the amounts (X_j's) of those claims. Then they are combined into the aggregate loss (S_N) as follows:

$$S_N = \sum_{j=1}^{N} X_j,$$

where $N = 0, 1, 2, \ldots$ and $S_0 = 0$. There are two interpretations of this model – the *collective risk model* and the *individual risk model* – which are used for different insurance contracts and result in different modeling approaches.

The individual risk model is used to aggregate the losses (or payments) from a *fixed number* of contracts. A typical business situation where this model is used is a group life or health insurance policy that covers a group of $N = n$ employees. Each employee can have different coverage (e.g., life insurance benefit as a multiple of salary) and different levels of loss probabilities which, for example, depend on employee's age and health status. In summary, under the individual risk model, N is not random and the main source of uncertainty of total payments is the random amounts of losses.

A more accurate and flexible model (and with a much wider scope of applicability) can be constructed by modeling the distribution of N and the distribution of the X_j's separately. This is how the collective risk model is built. Klugman *et al.* (2012, p. 139) list seven distinct advantages for such a modeling approach. Here we quote three which emphasize the role of the claim count variable:

> (i) The expected number of claims changes as the number of insured policies changes. Growth in the volume of business needs to be accounted for in forecasting the number of claims in future years based on past years' data.
>
> (ii) The impact on claims frequencies of changing deductibles is better understood.
>
> (iii) The shape of the distribution of S depends on the shapes of both distributions of N and X. The understanding of the relative shapes is useful when modifying policy details. For example, if the severity distribution has a much heavier tail than the frequency distribution, the shape of the tail of the distribution of aggregate losses will be determined by the severity distribution and will be relatively insensitive to the choice of frequency distribution.

In addition to playing a significant role in modeling the aggregate losses for most

insurance contracts, the claim count distributions are frequently used in designing *bonus-malus* systems in automobile insurance. Here is a brief introduction into how claim counts appear in those systems.

The auto insurance markets are very competitive and the companies constantly compete for the best drivers available in the market. The first step in identifying the quality of a driver (e.g., good, average, bad) is regression-type modeling that helps to group the customers with similar risk characteristics. All policyholders belonging to the same class pay the same premium. To limit possible discriminatory practices, state regulators do not allow classification based on factors that are beyond the person's control (e.g., gender, age). Also, a number of important factors (e.g., alcohol consumption habits, swiftnesses of reflexes) are nearly impossible to measure. Naturally, the initial pricing system is imperfect, but it gets corrected over time by combining preliminary classification rates with individual experience. The combining of the two components is achieved by employing credibility theory, which defines the credibility premium as a convex combination of the observed experience and a priori rating (also known as "manual rate"):

$$\text{Credibility Premium} \;=\; Z \times \text{Observed Experience} \;+\; (1-Z) \times \text{Manual Rate},$$

where the weight Z, $0 \le Z \le 1$, is called the credibility factor. The observed experience in the above formula can be average loss amount or number of claims, or aggregate loss. In other words, it has a lot of flexibility. Moreover, credibility theory allows insurance companies to design rating systems that penalize drivers responsible for one or more accidents by charging them extra premium (also known as "maluses") and rewarding claim-free drivers by giving them discounts (also known as "bonuses"). Such systems are called "no-claim discounts", "experience rating", "merit rating", or "bonus-malus"

systems. Interestingly, most of the bonus-malus systems used by the insurance companies around the world rely on claim counts, not the claim amounts, as the base for discounts and penalties. To learn more about risk classification, credibility and bonus-malus systems, we may refer the reader to a comprehesive book by Denuit *et al.* (2007). The same book, on pages `xxi-xxii`, provides an explanation why bonus-malus systems use only claim counts:

> "The vast majority of bonus-malus systems in force around the world penalize the number of at-fault accidents reported to the company, and not their amounts. A severe accident involving bodily injuries is penalized in the same way as a fender-bender. The reason to base motor risk classification on just claim frequencies is the long delay to access the cost of bodily injury and other severe claims. Not incorporating claim sizes in bonus-malus systems and *a priori* risk classification requires an (implicit) assumption of independence between the random variables 'number of claims' and 'cost of a claim', as well as the belief that the latter does not depend on the driver's characteristics. This means that the actuarial practice considers that the cost of an accident is, for the most part, beyond the control of a driver: a cautious driver reduces the number of accidents, but for the most part cannot control the cost of these accidents (which is largely independent of the mistake that caused it)."

1.3 Statistical Problems

In pricing, reserving and other actuarial applications policyholder's risk profile is summarized using various metrics such as net premiums or risk measures. Many of those metrics can be defined in terms of distribution quantiles. The classical definition of quantile function is

$$Q_Y(u) = F_Y^{-1}(u) = \inf\{y:\ F_Y(y) \geq u\}, \quad \text{for } 0 \leq u \leq 1, \tag{1.1}$$

where $F_Y(y)$ denotes the cumulative distribution function, CDF, of the loss random variable Y. If random variable Y is continuous (and thus its CDF), then $Q_Y(u)$ satisfies

$$F_Y(Q_Y(u)) = \mathbf{P}\big[Y \leq Q_Y(u)\big] = u. \tag{1.2}$$

For discrete random variables, however, taking infimum in (1.1) is important as there is no guarantee there exists such y that satisfies (1.2). Hence, the quantile function may have jumps. To see this, consider a Poisson distribution with parameter $\lambda = 5$, denoted $P(\lambda = 5)$. We have:

Table 1.1: The first nine values of $P(\lambda = 5)$ cumulative distribution function.

y	0	1	2	3	4	5	6	7	8
$\mathbf{P}\big[Y \leq y\big]$	0.0067	0.0404	0.1247	0.2650	0.4405	0.6160	0.7622	0.8666	0.9319

Now, if we are interested in finding a 20% *value-at-risk* (which is the 80th percentile), then there is no value of loss y for which $\mathbf{P}\big[Y \leq y\big] = 0.80$. Thus we have to choose the smallest value for the loss that gives *at least* a 0.80 probability that the reserve is sufficient. In this illustration, the 20% value-at-risk measure is 7, but such quantile function discontinuities may result in pricing irregularities.

The sample estimator of the classical quantile function (1.1) is

$$\widehat{Q}_n(u) = \widehat{F}_n^{-1}(u) = \inf\big\{y : \widehat{F}_n(y) \geq u\big\} = y_{[nu]+1:n}, \tag{1.3}$$

where $\widehat{F}_n(y)$ is the empirical CDF, $y_{1:n} \leq y_{2:n} \leq \cdots \leq y_{n:n}$ are the order statistics from the sample $y_1, \ldots, y_n$, and $[\cdot]$ denotes the greatest integer part. As discussed earlier, the coarseness of discrete data, however, makes the classical estimator (1.3) inappropriate when the product nu results in non-integer values. One way to handle this problem is to interpolate between two order statistics with indices closest to nu, as it is done by

Klugman *et al.* (2012, Section 13.1):

$$\widehat{Q}_{\text{KPW}}(u) \;=\; (1-\delta)\, y_{j:n} + \delta\, y_{j+1:n} \;=\; y_{j:n} + \delta\,(y_{j+1:n} - y_{j:n}), \tag{1.4}$$

where $j = [(n+1)u]$, $\delta = (n+1)u - j$, and $0 < u < 1$.

To highlight the differences between the estimators defined by (1.3) and (1.4), in Table 1.2 we provide values of $\widehat{Q}_n(u)$ and $\widehat{Q}_{\text{KPW}}(u)$, with $u = 0.25, 0.50, 0.75$, for four sets of binary data. We see from the table that while the triplets of $\widehat{Q}_n$ and $\widehat{Q}_{\text{KPW}}$ estimates slightly differ for sets A and D, they are identical for B and C. The insufficient differentation between the methods happens because data smoothing in (1.4) is based on only two data points.

Table 1.2: Values of $\widehat{Q}_n(u)$ and $\widehat{Q}_{\text{KPW}}(u)$ for binary data sets of size 5.

Quantile Estimator	Level u	Data Set A $\{0, 0, 0, 0, 1\}$	Data Set B $\{0, 0, 0, 1, 1\}$	Data Set C $\{0, 0, 1, 1, 1\}$	Data Set D $\{0, 1, 1, 1, 1\}$
$\widehat{Q}_n(u)$	0.25	0	0	0	1
	0.50	0	0	1	1
	0.75	0	1	1	1
$\widehat{Q}_{\text{KPW}}(u)$	0.25	0	0	0	0.5
	0.50	0	0	1	1
	0.75	0.5	1	1	1

Parzen (1992, 2004) proposed a smoothing technique that uses all data of the sample, and constructed *mid-distribution* and *mid-quantile* functions. Further, Ma *et al.* (2011) derived the asymptotic properties of the sample quantile estimator based on the mid-distribution function. This work motivated Wang and Hutson (2011) to design a new and improved smooth quantile function for discrete data. It is based on the theory of fractional order statistics, which was initiated by Stigler (1977), and takes a similar form as the kernel quantile estimator of Harrell and Davis (1982).

In this dissertation, we will follow and generalize the methodology proposed by Wang

and Hutson (2011). As a preview, let us compare their estimator, denoted by $\widehat{Q}_{\text{WH}}$, with $\widehat{Q}_n$ and $\widehat{Q}_{\text{KPW}}$. For data sets A, B, C, D, Table 1.3 summarizes the calculation of $\widehat{Q}_{\text{WH}}(0.25)$, $\widehat{Q}_{\text{WH}}(0.50)$, and $\widehat{Q}_{\text{WH}}(0.75)$. Note the smooth transition of quantile estimates as u changes from 0.25 to 0.75. Moreover, the estimates are distinct and react mildly to the gradual changes in data composition.

Table 1.3: Values of $\widehat{Q}_{\text{WH}}(u)$ for binary data sets of size 5.

Quantile Estimator	Level u	Data Set A $\{0, 0, 0, 0, 1\}$	Data Set B $\{0, 0, 0, 1, 1\}$	Data Set C $\{0, 0, 1, 1, 1\}$	Data Set D $\{0, 1, 1, 1, 1\}$
$\widehat{Q}_{\text{WH}}(u)$	0.25	0.0178	0.0885	0.2338	0.4861
	0.50	0.1424	0.3735	0.6265	0.8576
	0.75	0.5139	0.7662	0.9115	0.9822

1.4 Plan of the Thesis

The main objective of this dissertation is to propose and thoroughly investigate a new methodology to smooth quantile funtions for discrete claim count distributions. We provide definitions of smoothed quantile functions for discrete data samples and populations, investigate large- and small- sample properties of the estimators, and apply them to risk measurement exercises. The dissertation is organized in the following manner.

In Chapter 2, we illustrate claim count models using $(a, b, 0)$ class and $(a, b, 1)$ class which is mainly used for zero inflated insurance data.

In Chapter 3, we give an overview of smoothed quantiles for discrete distributions as well as their asymptotic properties established by Wang and Hutson (2011). Further, we generalize the methodology to vectors of smoothed quantiles. We provide definition, establish asymptotic properties and investigate the statistical properties of the estimators using simulations. The results established in this chapter are valid for discrete distributions with finite support.

In Chapter 4 as an extension of Chapter 3, we investigate and evaluate the perfor-

mance of smoothed quantile functions for discrete distributions with infinite support, both theoretically and via simulations. We modify the smoothed quantile function by truncating the data support. We focus on the region where major proportion of probability mass across the whole sample space is covered.

In Chapter 5, applications of smoothed quantiles to risk measurement are demonstrated using actual insurance data. We evaluate the commonly used distortion risk measures and report several risk measure estimates of automobile data using this new methodology.

Finally, in Chapter 6, the conclusions are drawn, and future research venues are discussed. In particular, we will focus on proving asymptotic properties conjectured in Chapter 4 and on developing percentile-matching estimators.

Chapter 2

Claim Frequency Models

2.1 The $(a, b, 0)$ Class

To construct models for insurance claim counts one starts with the so-called $(a, b, 0)$ class which contains only three distributions – Poisson, binomial, and negative binomial. As defined by Klugman *et al.* (2012, Section 6.4), a random variable N belongs to the $(a, b, 0)$ class if its probability mass function, PMF, satisfies the following recursion:

$$\frac{p_k}{p_{k-1}} \;=\; a + \frac{b}{k}\,, \qquad k = 1, 2, 3, \ldots, \tag{2.1}$$

where $p_k = \mathbf{P}\big[N = k\big]$ and a and b are some real-valued constants. Since the probabilities must sum to 1, the probability at zero is obtained from the recursive formula (2.1) as follows:

$$1 \;=\; p_0 + \sum_{k=1}^{\infty} p_k \;=\; p_0 + \sum_{k=1}^{\infty} \left(a + \frac{b}{k}\right) p_{k-1} \;=\; p_0 + \sum_{k=1}^{\infty} \left(\prod_{j=1}^{k} \left(a + \frac{b}{j}\right)\right) p_0,$$

and therefore

$$p_0 = \left(1 + \sum_{k=1}^{\infty} \prod_{j=1}^{k} \left(a + \frac{b}{j}\right)\right)^{-1}. \tag{2.2}$$

2.1.1 Poisson Distribution

A random variable N_{P} has a Poisson distribution with parameter $\lambda > 0$ if its PMF is given by

$$\mathbf{P}\big[N_{\text{P}} = k\big] \;=\; \frac{\lambda^k\, e^{-\lambda}}{k!}, \qquad k = 0, 1, 2, \ldots .$$

To summarize this fact, we will write: $N_{\text{P}} \sim P(\lambda)$. The following verification of (2.1) and (2.2) shows that $P(\lambda)$ belongs to the $(a, b, 0)$ class of distributions:

$$\frac{p_k}{p_{k-1}} \;=\; \frac{\lambda^k\, e^{-\lambda}/k!}{\lambda^{k-1}\, e^{-\lambda}/(k-1)!} \;=\; \frac{\lambda}{k}\,, \qquad k = 1, 2, 3, \ldots .$$

Thus, for the Poisson distribution, we have $a = 0$, $b = \lambda$, and

$$p_0 \;= \left(1 + \sum_{k=1}^{\infty} \prod_{j=1}^{k} \left(0 + \frac{\lambda}{j}\right)\right)^{-1} \;=\; \left(1 + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!}\right)^{-1} \;=\; e^{-\lambda}.$$

Also, since the mean and variance of $P(\lambda)$ are equal ($\mathbf{E}\big[N_{\text{P}}\big] = \mathbf{Var}\big[N_{\text{P}}\big] = \lambda$), this distribution is appropriate for modeling equi-dispersed data.

2.1.2 Binomial Distribution

A random variable N_{B} has a binomial distribution with parameters $m \geq 1$ (integer) and $0 < q < 1$, denoted as $N_{\text{B}} \sim Bin(m,\, q)$, if its PMF is given by

$$\mathbf{P}\big[N_{\text{B}} = k\big] \;=\; \binom{m}{k} q^k\, (1-q)^{m-k}, \qquad k = 0, 1, 2, \ldots, m.$$

The following steps verify (2.1) and (2.2):

$$\frac{p_k}{p_{k-1}} \;=\; \frac{\binom{m}{k} q^k\, (1-q)^{m-k}}{\binom{m}{k-1} q^{k-1}\, (1-q)^{m-(k-1)}} \;=\; -\,\frac{q}{1-q} + \frac{m+1}{k}\,\frac{q}{1-q}\,, \qquad k = 1, 2, \ldots, m.$$

Thus, for the binomial distribution, $a = -q/(1-q)$, $b = (m+1)q/(1-q)$, and

$$\begin{aligned} p_0 &= \left(1 + \sum_{k=1}^{m} \prod_{j=1}^{k} \left(-\frac{q}{1-q} + \frac{(m+1)q/(1-q)}{j}\right)\right)^{-1} \\ &= \left(1 + \sum_{k=1}^{m} \left(\frac{q}{1-q}\right)^k \binom{m}{k}\right)^{-1} = \left(1 + (1-q)^{-m}\left(1 - (1-q)^m\right)\right)^{-1} \\ &= (1-q)^m. \end{aligned}$$

Also, since the mean of $Bin\,(m,\, q)$ is greater than its variance ($\mathbf{E}\big[N_{\text{B}}\big] = mq > mq(1-q) = \mathbf{Var}\big[N_{\text{B}}\big]$), this distribution is appropriate for modeling under-dispersed data.

2.1.3 Negative Binomial Distribution

A random variable N_{NB} has a negative binomial distribution with parameters $r > 0$ and $\beta > 0$, denoted as $N_{\text{NB}} \sim NB\,(r,\, \beta)$, if its PMF is given by

$$\mathbf{P}\big[N_{\text{NB}} = k\big] = \binom{k+r-1}{k} \left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^k, \qquad k = 0, 1, 2, \ldots.$$

In the special case $r = 1$, we obtain the geometric distribution with the probability of success $1/(1+\beta)$. Equations (2.1) and (2.2) are verified as follows:

$$\frac{p_k}{p_{k-1}} = \frac{\binom{k+r-1}{k} \left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^k}{\binom{k+r-2}{k-1} \left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^{k-1}} = \frac{\beta}{1+\beta} + \frac{r-1}{k}\frac{\beta}{1+\beta}, \qquad k = 1, 2, 3, \ldots.$$

Thus, for the negative binomial distribution, $a = \beta/(1+\beta)$, $b = (r-1)\beta/(1+\beta)$, and

$$\begin{aligned} p_0 &= \left(1 + \sum_{k=1}^{\infty} \prod_{j=1}^{k} \left(\frac{\beta}{1+\beta} + \frac{(r-1)\beta/(1+\beta)}{j}\right)\right)^{-1} \\ &= \left(1 + \sum_{k=1}^{\infty} \left(\frac{\beta}{1+\beta}\right)^k \binom{k+r-1}{k}\right)^{-1} = \left(1 + (1+\beta)^r\left(1 - (1+\beta)^{-r}\right)\right)^{-1} \\ &= (1+\beta)^{-r}. \end{aligned}$$

Also, since the mean of $NB\,(r,\,\beta)$ is smaller than its variance ($\mathbf{E}\big[N_{\text{NB}}\big] = r\beta < r\beta(1+\beta) = \mathbf{Var}\big[N_{\text{NB}}\big]$), this distribution is appropriate for modeling over-dispersed data.

2.2 The $(a, b, 1)$ Class

Insurance data usually include a relatively large number of zeros (typical claim count data sets contain 80% or more zeros; see, e.g., Klugman *et al.*, 2012, Table 6.2). Zeros occur when no claims are reported by policyholders during the period under study. Introduction of deductibles and no claim discounts increases the proportion of zeros and leads to a small probability of occurrence of a loss (i.e., small p_k for $k \geq 1$). The scenario when p_0 is much larger than p_k, $k \geq 1$, cannot be properly accommodated by the members of the $(a, b, 0)$ class.

An adjustment of the probability at zero is done by modifying the $(a, b, 0)$ class as follows. First, we define a new class – the $(a, b, 1)$ class – which contains random variables whose PMF satisfies the recursive formula

$$\frac{p_k}{p_{k-1}} \;=\; a + \frac{b}{k}\,, \qquad k = 2, 3, 4, \ldots. \tag{2.3}$$

Note that the only difference between (2.1) and (2.3) is that the latter recursion begins at p_1 rather p_0. Then, we put an arbitrary amount of probability at zero, say c, and treat it as a parameter. This results in the following relationships between the probabilities of *zero-modified* (or *zero-inflated*) distribution, denoted by p_k^*, and the corresponding $(a, b, 0)$ distribution, denoted by p_k:

$$p_k^* = (1-c)\frac{p_k}{1-p_0} \quad \text{for} \quad k = 1, 2, 3, \ldots, \qquad \text{and} \quad p_0^* = c. \tag{2.4}$$

Note that the zero-inflated model (2.4) can be viewed as a mixture between a zero-

truncated member of the $(a, b, 0)$ class and a degenerate distribution that places all the probability at zero. It assigns a probability mass of c to the zeros and a mass of $(1-c)$ to the counting distribution defined on positive integers. The following relationships between the means and variances of variables N_* (from $(a, b, 1)$ class) and N (from $(a, b, 0)$ class) can be easily justified:

$$\mathbf{E}\big[N_*\big] \;=\; \frac{1-c}{1-p_0}\,\mathbf{E}\big[N\big] \quad \text{and} \quad \mathbf{Var}\big[N_*\big] \;=\; \frac{1-c}{1-p_0}\left\{\mathbf{Var}\big[N\big] + \frac{c-p_0}{1-p_0}\left(\mathbf{E}\big[N\big]\right)^2\right\}. \tag{2.5}$$

Also, when $c = 0$, zero-modified distributions are defined for $k = 1, 2, \ldots$, and are called *zero-truncated* distributions (see Klugman *et al.*, 2012, Section 6.6). Moreover, unlike the $(a, b, 0)$ class, the $(a, b, 1)$ class admits more than three standard distributions. For example, the negative binomial model within this class can be extended by replacing the condition $r > 0$ with $r > -1$ and $r \neq 0$. Zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), and several other zero-inflated distributions had been used for modeling automobile insurance claims by Yip and Yau (2005).

Chapter 3

Smoothed Quantiles

In this chapter, we provide definition, illustrations and asymptotic properties of $\widehat{Q}_{\text{WH}}$, extend them to vectors of quantile estimators, and conclude with a simulation study.

3.1 Smoothed Population Quantiles

3.1.1 Definition and Properties

Following Wang and Hutson (2011), let us consider a discrete random variable Y with CDF F_Y and PMF $p_j = \mathbf{P}\big[Y = y_{j:d}\big]$, where $y_{j:d}$ is the jth smallest *distinct* value that Y can take. Notice that $\sum_{j=1}^{d} p_j = 1$ and $1 < d < \infty$. Let us denote $F_j := F_Y(y_{j:d}) = \sum_{i=1}^{j} p_i$; also $F_0 \equiv 0$. Then the smoothed population quantile function for discrete random variable Y is defined as

$$Q_Y(u) \;=\; \sum_{j=1}^{d} \Big[B_{\alpha_u,\beta_u}(F_j) - B_{\alpha_u,\beta_u}(F_{j-1})\Big] y_{j:d} \;=:\; \sum_{j=1}^{d} w_{j(u)}\, y_{j:d}, \tag{3.1}$$

where B_{α_u,β_u} denotes the CDF of a beta random variable with parameters $\alpha_u = (d+1)u$ and $\beta_u = (d+1)(1-u)$. Note that the weights $w_{j(u)}$ are non-negative (because $F_j \geq F_{j-1}$

and B_{α_u,β_u} is an increasing function) and add up to one:

$$\begin{aligned}\sum_{j=1}^{d} w_{j(u)} &= \sum_{j=1}^{d} \big[B_{\alpha_u,\beta_u}(F_j) - B_{\alpha_u,\beta_u}(F_{j-1})\big] \\ &= B_{\alpha_u,\beta_u}(F_d) - B_{\alpha_u,\beta_u}(F_0) \;=\; B_{\alpha_u,\beta_u}(1) - B_{\alpha_u,\beta_u}(0) \;=\; 1.\end{aligned}$$

Also, the mean and variance of B, the random variable with CDF B_{α_u,β_u} used in (3.1), are given by

$$\mathbf{E}\big[B\big] \;=\; \frac{(d+1)u}{(d+1)u + (d+1)(1-u)} \;=\; u$$

and

$$\mathbf{Var}\big[B\big] \;=\; \frac{\big[(d+1)u\big]\big[(d+1)(1-u)\big]}{\big[(d+1)u + (d+1)(1-u)\big]^2\big[(d+1)u + (d+1)(1-u) + 1\big]} \;=\; \frac{u(1-u)}{d+2}.$$

The formulas of $\mathbf{E}\big[B\big]$ and $\mathbf{Var}\big[B\big]$ suggest that for discrete populations with large number of possible distinct values (i.e., when d is large), most significant contributions toward the value of $Q_Y(u)$ will be made by several F_j's clustered near the level u. This pattern is quite evident in Figure 3.1, where the density curves of B are plotted for various quantile levels u. On the horizontal axes, the F_j marks were computed for selected binomial distributions with the probability of "success" q equal to 0.7. Note that the weight $w_{j(u)}$ is the area under the density curve over the interval $[F_{j-1}; F_j]$.

Further, Wang and Hutson (2011) established the following three properties of the smoothed population quantile function defined by (3.1):

(*a*) $Q_Y(u)$ is a continuous and monotonically increasing function of u over $(0, 1)$.

(*b*) $Q_Y(u) \longrightarrow y_{1:d}$ as $u \to 0$.

(*c*) $Q_Y(u) \longrightarrow y_{d:d}$ as $u \to 1$.

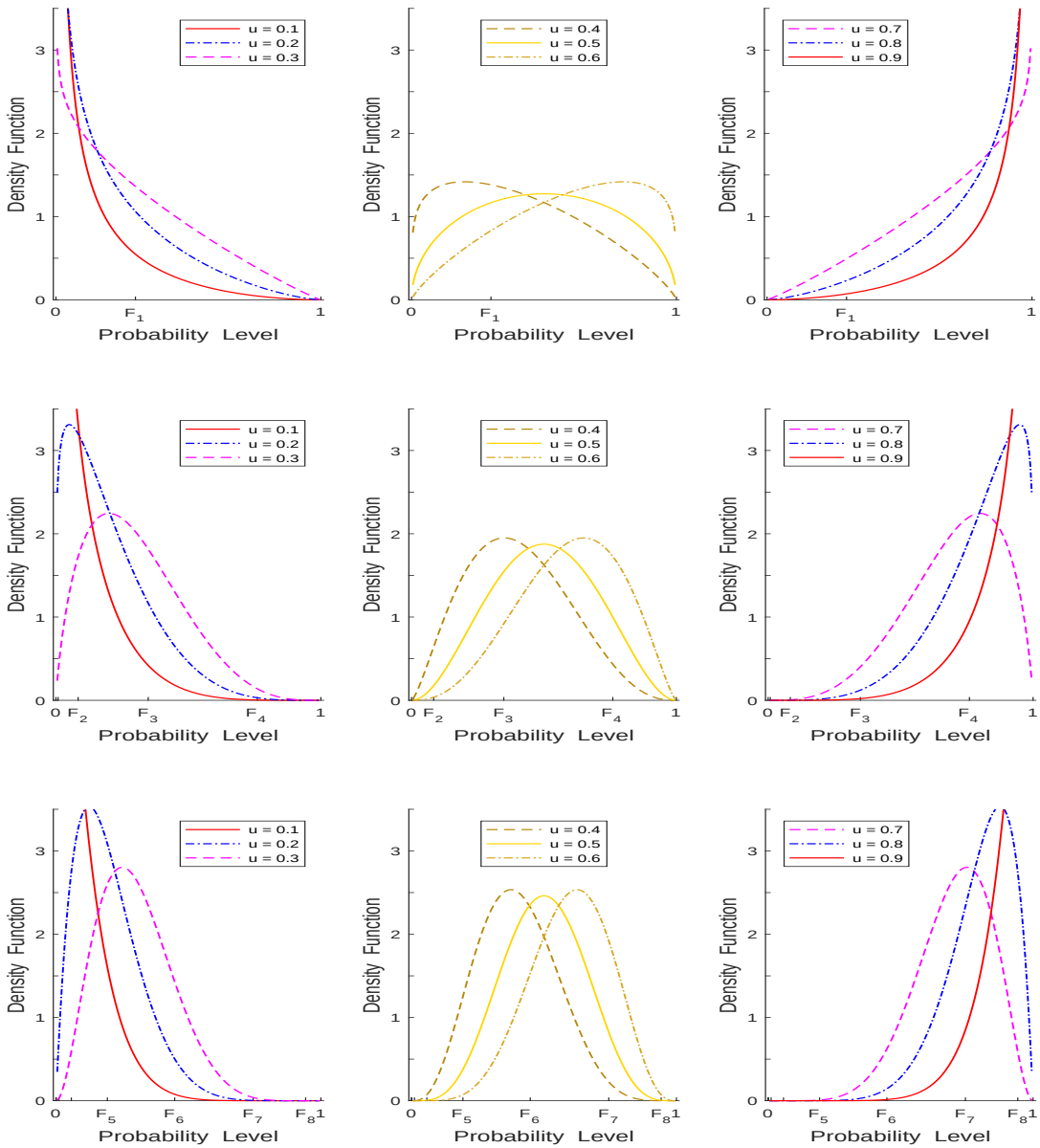

Figure 3.1: The density curves of B used to compute weights $w_{j(u)}$ for quantile levels $u = 0.1, \ldots, 0.9$. The weight $w_{j(u)}$ corresponds to the area under the curve over $[F_{j-1}; F_j]$, where $F_0 = 0$ and $F_1, \ldots, F_d$ are CDF's of $Bin\,(m = d - 1,\, q = 0.7)$ with $m = 1$ (top row), $m = 4$ (middle row), and $m = 8$ (bottom row).

Finally, equation (3.1) is easy to understand – the formula assigns weights to distinct data points which later get aggregated. However, for computational purposes and theoretical investigations, it will be easier to work with the $Q_Y(u)$ formula rewritten in terms of data spacings:

$$\begin{aligned} Q_Y(u) &= \sum_{j=1}^{d} \Big[B_{\alpha_u,\beta_u}(F_j) - B_{\alpha_u,\beta_u}(F_{j-1}) \Big] y_{j:d} \\ &= -y_{1:d} B_{\alpha_u,\beta_u}(F_0) + \sum_{j=1}^{d-1} \big(y_{j:d} - y_{j+1:d} \big) B_{\alpha_u,\beta_u}(F_j) + y_{d:d} B_{\alpha_u,\beta_u}(F_d) \\ &= \sum_{j=1}^{d-1} \big(y_{j:d} - y_{j+1:d} \big) B_{\alpha_u,\beta_u}(F_j) + y_{d:d}. \end{aligned} \tag{3.2}$$

3.1.2 Illustrations

In this section, we shall provide plots of $Q_Y(u)$ for selected binomial and zero-inflated binomial (ZIB) distributions. But before we do that let us first simplify equation (3.2) even further. Notice that for binomial distributions from the class $(a, b, 0)$ or $(a, b, 1)$ we have: $y_{1:d} = 0$, $y_{j+1:d} - y_{j:d} = 1$, and $y_{d:d} = d - 1$. This reduces (3.2) to

$$Q_Y(u) = (d-1) - \sum_{j=1}^{d-1} B_{\alpha_u,\beta_u}(F_j). \tag{3.3}$$

In Figure 3.2, we plot the quantile function $Q_Y(u)$, defined by (3.3), with F_j representing the CDF of binomial distribution for various q and $m = 1, 4, 8$. The minor "waves" visible for $m = 8$ with $q = 0.1$ and 0.9 are expected because as d grows the variance of the weights $w_{j(u)}$, which is equal to $u(1-u)/(d+2)$, decreases and the smoothed quantile function closer approximates the classical discrete quantile function (1.1).

In Figure 3.3, we plot the quantile function $Q_Y(u)$, defined by (3.3), with F_j representing the CDF of zero-inflated binomial distribution for $m = 20$ and various c and q.

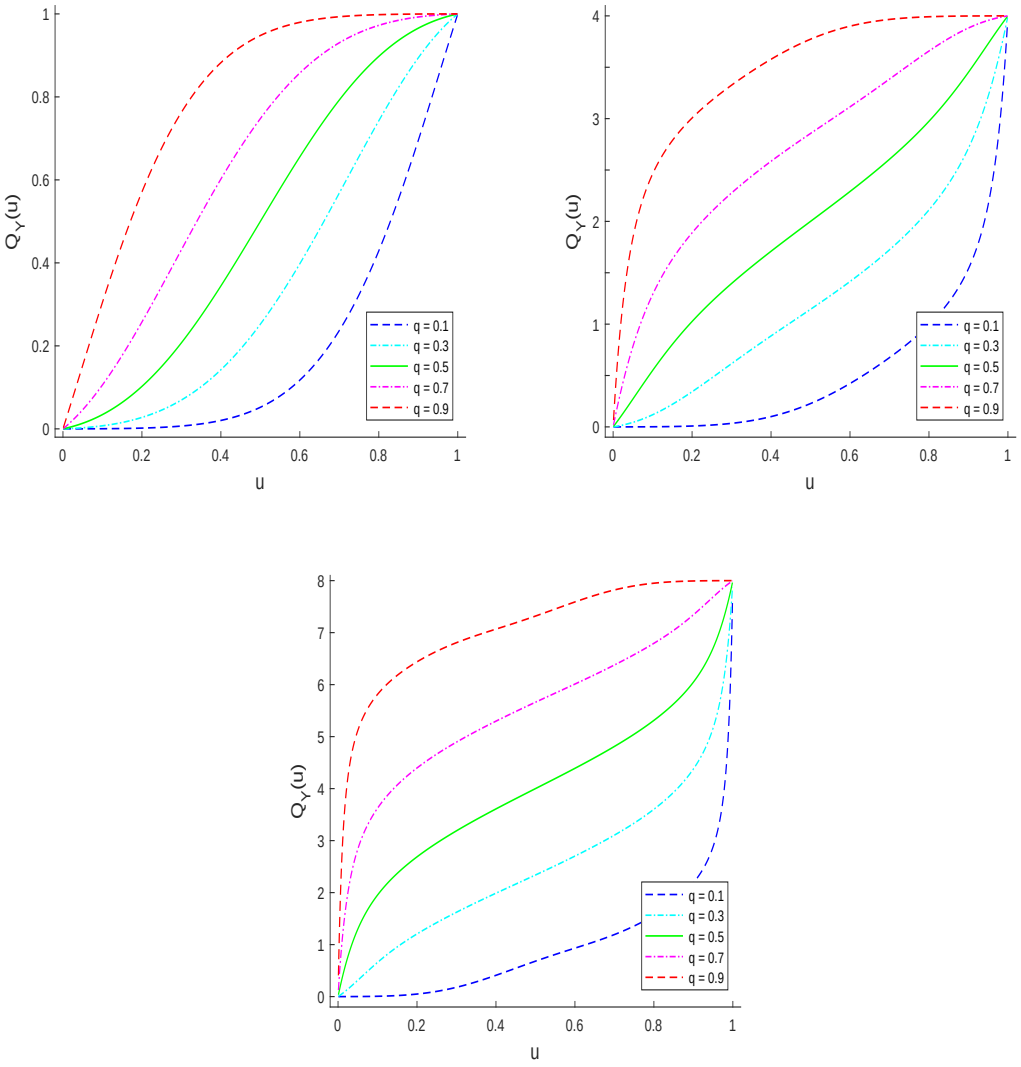

Figure 3.2: The quantile functions of $Bin\,(m,\,q)$ distributions with $q = 0.1, 0.3, 0.5, 0.7, 0.9$ and $m = 1$ (top left panel), $m = 4$ (top right panel), and $m = 8$ (bottom panel).

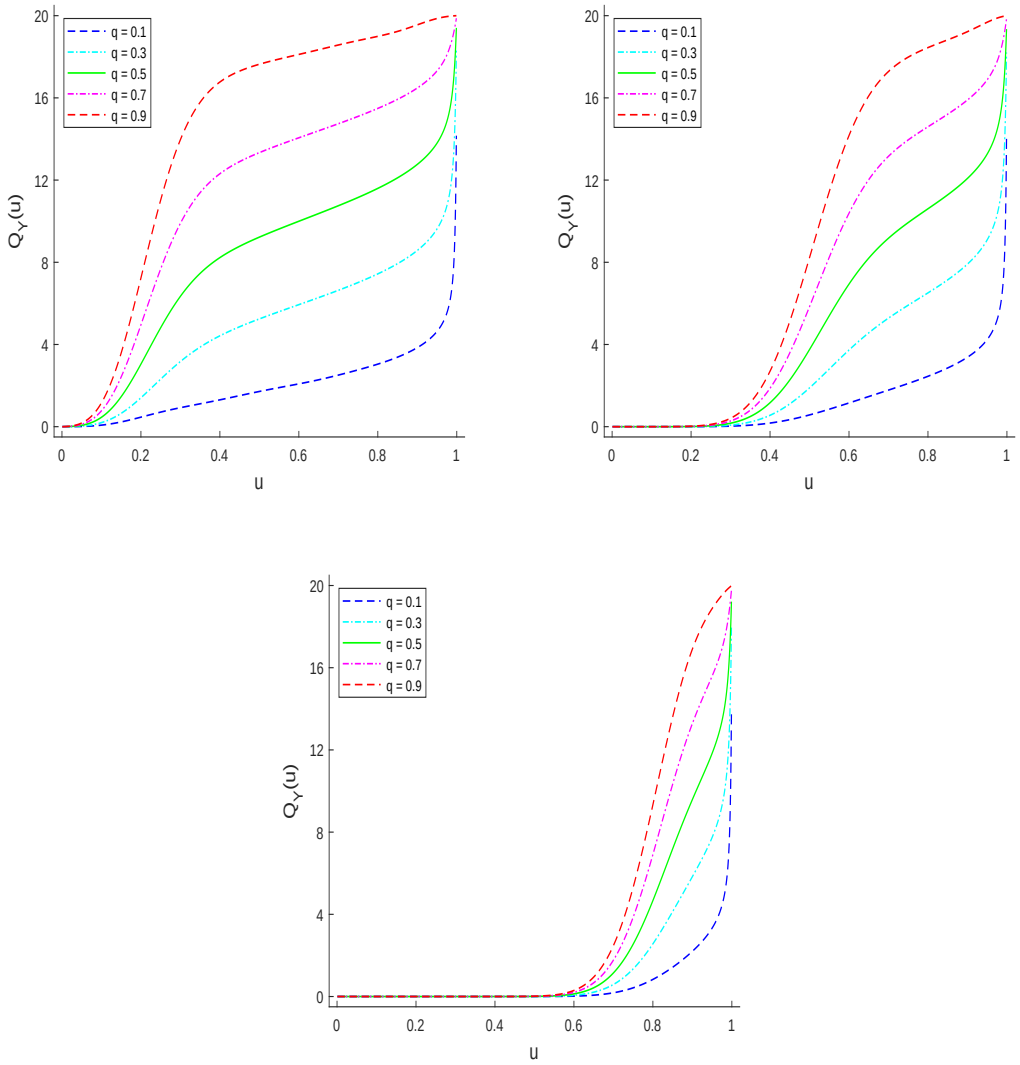

Figure 3.3: The quantile functions of $q = 0.1, 0.3, 0.5, 0.7, 0.9$ and $c = 0.2$ (top left panel), $c = 0.5$ (top right panel), and $c = 0.8$ (bottom panel).

The effect of excessive number of zeros on the quantile function is obvious.

3.2 Smoothed Sample Quantiles

In this section, we introduce the sample version of the smoothed population quantile function, provide a few numerical illustrations, and present the smoothed quantile estimator's asymptotic properties.

3.2.1 Definition

Consider a random sample $Y_1, Y_2, \ldots, Y_n$ from an unknown discrete distribution with CDF F_Y. Let $y_{1:d} < y_{2:d} < \cdots < y_{d:d}$ denote the distinct data values with corresponding frequencies $r_1, r_2, \ldots, r_d$. Then, the sample PMF is $\widehat{p}_i = r_i/n$ and the empirical CDF at $y_{j:d}$ is given by $\widehat{F}_j = \widehat{F}_Y(y_{j:d}) = \sum_{i=1}^{j} \widehat{p}_i = n^{-1} \sum_{i=1}^{j} r_i$. The sample estimator of the smoothed uth quantile for discrete data is defined by replacing F_j with $\widehat{F}_j$ in (3.1) (or equivalently in (3.2)). This leads to

$$\widehat{Q}_Y(u) = \sum_{j=1}^{d} \widehat{w}_{j(u)} y_{j:d} = \sum_{j=1}^{d} \Big[B_{\alpha_u,\beta_u}(\widehat{F}_j) - B_{\alpha_u,\beta_u}(\widehat{F}_{j-1}) \Big] y_{j:d} \tag{3.4}$$

$$= \sum_{j=1}^{d-1} \big(y_{j:d} - y_{j+1:d} \big) B_{\alpha_u,\beta_u}(\widehat{F}_j) + y_{d:d}, \tag{3.5}$$

where $\widehat{F}_0 = 0$ and B_{α_u,β_u} denotes the CDF of a beta random variable with parameters $\alpha_u = (d+1)u$ and $\beta_u = (d+1)(1-u)$.

3.2.2 Numerical Examples

To get a better sense of how the weights $\widehat{w}_{j(u)}$ are assigned to particular data points, in Table 3.1 we revisit Table 2.2 and provide step-by-step calculations of the smoothed quartiles (i.e., $u = 0.25, 0.50, 0.75$) for data sets A, B, C, D. We note again the separation and gradual transition of the smooth estimates as u changes from 0.25 to 0.75. This is true for all data sets and is in contrast to the estimates based on the standard definition of discrete quantile function (see Table 2.1).

Table 3.1: Calculation of smoothed sample quartiles $\widehat{Q}_Y(0.25)$, $\widehat{Q}_Y(0.50)$, $\widehat{Q}_Y(0.75)$ for data sets A, B, C, D. Here $d = 2$, $\alpha_u = (d+1)u$, and $\beta_u = (d+1)(1-u)$.

u	j	$y_{j:d}$	$\widehat{F}_j$	$B_{\alpha_u,\beta_u}(\widehat{F}_j) - B_{\alpha_u,\beta_u}(\widehat{F}_{j-1}) = \widehat{w}_{j(u)}$	$\widehat{w}_{j(u)}\, y_{j:d}$	$\sum_{i=1}^{j} \widehat{w}_{i(u)}\, y_{i:d}$
Data Set A: $\{0,0,0,0,1\}$						
0.25	1	0	0.80	$0.9822 - 0 = 0.9822$	0	0
	2	1	1	$1 - 0.9822 = 0.0178$	0.0178	0.0178
0.50	1	0	0.80	$0.8576 - 0 = 0.8576$	0	0
	2	1	1	$1 - 0.8576 = 0.1424$	0.1424	0.1424
0.75	1	0	0.80	$0.4861 - 0 = 0.4861$	0	0
	2	1	1	$1 - 0.4861 = 0.5139$	0.5139	0.5139
Data Set B: $\{0,0,0,1,1\}$						
0.25	1	0	0.60	$0.9115 - 0 = 0.9115$	0	0
	2	1	1	$1 - 0.9115 = 0.0885$	0.0885	0.0885
0.50	1	0	0.60	$0.6265 - 0 = 0.6265$	0	0
	2	1	1	$1 - 0.6265 = 0.3735$	0.3735	0.3735
0.75	1	0	0.60	$0.2338 - 0 = 0.2338$	0	0
	2	1	1	$1 - 0.2338 = 0.7662$	0.7662	0.7662
Data Set C: $\{0,0,1,1,1\}$						
0.25	1	0	0.40	$0.7662 - 0 = 0.7662$	0	0
	2	1	1	$1 - 0.7662 = 0.2338$	0.2338	0.2338
0.50	1	0	0.40	$0.3735 - 0 = 0.3735$	0	0
	2	1	1	$1 - 0.3735 = 0.6265$	0.6265	0.6265
0.75	1	0	0.40	$0.0885 - 0 = 0.0885$	0	0
	2	1	1	$1 - 0.0885 = 0.9115$	0.9115	0.9115
Data Set D: $\{0,1,1,1,1\}$						
0.25	1	0	0.20	$0.5139 - 0 = 0.5139$	0	0
	2	1	1	$1 - 0.5139 = 0.4861$	0.4861	0.4861
0.50	1	0	0.20	$0.1424 - 0 = 0.1424$	0	0
	2	1	1	$1 - 0.1424 = 0.8576$	0.8576	0.8576
0.75	1	0	0.20	$0.0178 - 0 = 0.0178$	0	0
	2	1	1	$1 - 0.0178 = 0.9822$	0.9822	0.9822

3.2.3 Asymptotic Properties

As described by Serfling (1980, Section 2.3.3), the classical quantile estimator follows an asymptotically normal distribution if the underlying distribution that generated data is

smooth at that point. The lack of smoothness may result in estimators that are not normal. This does happen for discrete distributions (Genton *et al.*, 2006). However, the smoothed sample estimator $\widehat{Q}_Y(u)$, defined by (3.4), which estimates $Q_Y(u)$, defined by (3.1), is consistent and asymptotically normal. These properties are established in Theorem 4.1 of Wang and Hutson (2011) and restated below.

Theorem [Wang and Hutson, 2011]

Consider an i.i.d. sample of size n from a discrete distribution $F_V(\cdot)$ with finite support $v_{1:d} < v_{2:d} < \cdots < v_{d:d}$, $d < \infty$. Then as $n \to \infty$, we have the following results:

$$(i) \quad \widehat{Q}_V(u) \xrightarrow{P} Q_V(u),$$

$$(ii) \quad n^{1/2}\left(\widehat{Q}_V(u) - Q_V(u)\right) \sim \mathcal{AN}\big(0, \sigma^2\big),$$

where $\sigma^2 = K^2 l' D l$, l is a $d-1$ vector with jth $(1 \le j \le d-1)$ element $l_j = (v_{j:d} - v_{j+1:d}) F_j^{d'u-1}(1-F_j)^{d'(1-u)}$, D is a $(d-1)\times(d-1)$ matrix with ijth $(i \le j)$ element $D_{ij} = F_i(1-F_j)$ with $d' = d+1$, and $K = \frac{\Gamma(d')}{\Gamma(d'u)\Gamma(d'(1-u))}$ and $F_j = F_V(v_{j:d})$. Given real data, σ^2 can be estimated readily by substituting F_j with $\widehat{F}_j = \widehat{F}_V(v_{j:d})$.

3.3 Vectors of Quantiles

Distribution quantiles play a key role in defining risk measures, finding capital allocation proportions, and in many other actuarial applications. Usually those problems require simultaneous estimation of multiple quantiles, as well as subsequent joint statistical inference. Therefore, in this section we generalize the definitions and asymptotic properties of Section 3.2 to vectors of smoothed quantiles. These new results are then verified and augmented using finite-sample simulations.

3.3.1 Definition

Let us consider the same setups as in Sections 3.1.1 and 3.2.1. Using the smoothed uth quantile for discrete population, given by (3.1), and its sample estimator, given by (3.4), we now focus on vectors of such quantiles. That is, for $0 < u_1 < \cdots < u_l < 1$, the vector of smoothed population quantiles

$$\Big(Q_Y(u_1), \ldots, Q_Y(u_l)\Big) \tag{3.6}$$

with $Q_Y(u_i) = \sum_{j=1}^{d} \big[B_{\alpha_{u_i},\beta_{u_i}}(F_j) - B_{\alpha_{u_i},\beta_{u_i}}(F_{j-1})\big] y_{j:d}$ for $i = 1, \ldots, l$, will be estimated by

$$\Big(\widehat{Q}_Y(u_1), \ldots, \widehat{Q}_Y(u_l)\Big) \tag{3.7}$$

with $\widehat{Q}_Y(u_i) = \sum_{j=1}^{d} \big[B_{\alpha_{u_i},\beta_{u_i}}(\widehat{F}_j) - B_{\alpha_{u_i},\beta_{u_i}}(\widehat{F}_{j-1})\big] y_{j:d}$ for $i = 1, \ldots, l$. Here $B_{\alpha_{u_i},\beta_{u_i}}$ denotes the CDF of a beta random variable with parameters $\alpha_{u_i} = (d+1)u_i$ and $\beta_{u_i} = (d+1)(1-u_i)$.

3.3.2 Asymptotic Properties

We will demonstrate that the estimator (3.7) is a consistent estimator of (3.6) and it is asymptotically normal. These properties are established in the following theorem.

Theorem 3.1. *Consider an i.i.d. sample of size n from a discrete distribution F_Y with finite support $y_{1:d} < y_{2:d} < \cdots < y_{d:d}$ and $d < \infty$. Then, as $n \to \infty$, the following statements hold:*

$$(i) \quad \Big(\widehat{Q}_Y(u_1), \ldots, \widehat{Q}_Y(u_l)\Big) \xrightarrow{P} \Big(Q_Y(u_1), \ldots, Q_Y(u_l)\Big),$$

$$(ii) \quad \Big(\widehat{Q}_Y(u_1), \ldots, \widehat{Q}_Y(u_l)\Big) \sim \mathcal{AN}\left(\big(Q_Y(u_1), \ldots, Q_Y(u_l)\big), \frac{1}{n}\mathbf{HDH}'\right),$$

where $\mathbf{D} := \big[d_{ij}\big]_{(d-1)\times(d-1)}$ with $d_{ij} = d_{ji} = F_i(1-F_j)$, $i \le j$, and $\mathbf{H} := \big[h_{ij}\big]_{l\times(d-1)}$ with

$h_{ij} = (y_{j:d} - y_{j+1:d})\, b_{\alpha_{u_i},\beta_{u_i}}(F_j)$. *Here* $b_{\alpha_{u_i},\beta_{u_i}}$ *denotes the* PDF *of a beta random variable with parameters* $\alpha_{u_i} = (d+1)u_i$ *and* $\beta_{u_i} = (d+1)(1-u_i)$, *and* $F_j = F_Y(y_{j:d})$.

Proof: Firstly, for a multinomial experiment with m possible outcomes, let p_j denote the probability of occurrence of the jth outcome $\big(\sum_{j=1}^{m} p_j = 1\big)$. Based on a sample of n i.i.d. trials, p_j is estimated by the observed relative frequency, say $\widehat{p}_j = r_j/n$. Now recall the fact (Serfling, 1980, Section 2.7) that $(\widehat{p}_1, \ldots, \widehat{p}_{m-1})$ is a consistent estimator of $(p_1, \ldots, p_{m-1})$, and it is asymptotically normal:

$$\big(\widehat{p}_1, \ldots, \widehat{p}_{m-1}\big) \;\sim\; \mathcal{AN}\left(\big(p_1, \ldots, p_{m-1}\big),\, \frac{1}{n}\mathbf{\Sigma}\right),$$

where $\mathbf{\Sigma} := \big[\sigma_{ij}\big]_{(m-1)\times(m-1)}$ with $\sigma_{ij} = p_i(1-p_i)$ if $i = j$, and $= -p_i p_j$, if $i \neq j$.

Secondly, the data setup considered in this section can be interpreted as the above described multinomial experiment with $p_j = F_j - F_{j-1}$ and $\widehat{p}_j = \widehat{F}_j - \widehat{F}_{j-1}$ for $j = 1, \ldots, d$. Note that $F_0 = \widehat{F}_0 = 0$ and $F_d = \widehat{F}_d = 1$. Having the joint asymptotic normality result for the spacings $\widehat{F}_j - \widehat{F}_{j-1} = \widehat{p}_j$, we can apply the multivariate delta method (Serfling, 1980, Section 3.3) and derive joint asymptotically normal distribution for $\big(\widehat{F}_1, \ldots, \widehat{F}_{d-1}\big)$. That is, the inverse transformation of p_i's is $F_i = \sum_{j=1}^{i} p_j$, $i = 1, \ldots, d-1$, and its Jacobian – matrix $\mathbf{J}$ with the ijth entry $\partial F_i/\partial p_j$ – is the lower triangular matrix with the ijth entry equal to 1 for $i \leq j$ and 0 otherwise. Thus,

$$\big(\widehat{F}_1, \ldots, \widehat{F}_{d-1}\big) \;\sim\; \mathcal{AN}\left(\big(F_1, \ldots, F_{d-1}\big),\, \frac{1}{n}\mathbf{D}\right), \tag{3.8}$$

where

$$
\begin{aligned}
\mathbf{D} &= \mathbf{J}_{(\mathbf{d-1})\times(\mathbf{d-1})}\boldsymbol{\Sigma}_{(\mathbf{d-1})\times(\mathbf{d-1})}\mathbf{J}'_{(\mathbf{d-1})\times(\mathbf{d-1})} \\
&= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_{d-1} \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_{d-1} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{d-1}p_1 & -p_{d-1}p_2 & \dots & p_{d-1}(1-p_{d-1}) \end{bmatrix} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \\
&= \begin{bmatrix} \left(\sum_{i=1}^{1} p_i\right)\left(1-\sum_{i=1}^{1} p_i\right) & \left(\sum_{i=1}^{1} p_i\right)\left(1-\sum_{i=1}^{2} p_i\right) & \dots & \left(\sum_{i=1}^{1} p_i\right)\left(1-\sum_{i=1}^{d-1} p_i\right) \\ \left(1-\sum_{i=1}^{2} p_i\right)\left(\sum_{i=1}^{1} p_i\right) & \left(\sum_{i=1}^{2} p_i\right)\left(1-\sum_{i=1}^{2} p_i\right) & \dots & \left(\sum_{i=1}^{2} p_i\right)\left(1-\sum_{i=1}^{d-1} p_i\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(1-\sum_{i=1}^{d-1} p_i\right)\left(\sum_{i=1}^{1} p_i\right) & \left(1-\sum_{i=1}^{d-1} p_i\right)\left(\sum_{i=1}^{2} p_i\right) & \dots & \left(\sum_{i=1}^{d-1} p_i\right)\left(1-\sum_{i=1}^{d-1} p_i\right) \end{bmatrix} \\
&= \begin{bmatrix} F_1(1-F_1) & F_1(1-F_2) & \dots & F_1(1-F_{d-1}) \\ F_1(1-F_2) & F_2(1-F_2) & \dots & F_2(1-F_{d-1}) \\ \vdots & \vdots & \ddots & \vdots \\ F_1(1-F_{d-1}) & F_2(1-F_{d-1}) & \dots & F_{d-1}(1-F_{d-1}) \end{bmatrix}
\end{aligned}
$$

Thirdly, since the entries of the covariance-variance matrix in (3.8) diminish at the rate $1/n$, it follows from the multidimensional Chebyshev's inequality that $\left(\widehat{F}_1, \dots, \widehat{F}_{d-1}\right) \xrightarrow{P} \left(F_1, \dots, F_{d-1}\right)$. Next, from (3.5) we notice that

$$
\widehat{Q}_Y(u_i) = \sum_{j=1}^{d-1} \left(y_{j:d} - y_{j+1:d}\right) B_{\alpha_{u_i},\beta_{u_i}}(\widehat{F}_j) + y_{d:d}, \qquad i = 1, \dots, l,
$$

is a continuous transformation of $\left(\widehat{F}_1, \dots, \widehat{F}_{d-1}\right)$. Therefore, according to the continuous mapping theorem (Serfling, 1980, Section 1.7),

$$
\left(\widehat{Q}_Y(u_1), \dots, \widehat{Q}_Y(u_l)\right) \xrightarrow{P} \left(Q_Y(u_1), \dots, Q_Y(u_l)\right),
$$

which proves part (i) of the theorem.

Finally, to prove part (ii), we apply the multivariate delta method to (3.8). The Jacobian of transformations $Q_Y(u_i)$ (viewed as functions of $F_1, \ldots, F_{d-1}$) has the following ijth entry:

$$\begin{aligned} h_{ij} &= \frac{\partial Q_Y(u_i)}{\partial F_j} = \frac{\partial}{\partial F_j}\left[\sum_{j=1}^{d-1}\big(y_{j:d} - y_{j+1:d}\big)B_{\alpha_{u_i},\beta_{u_i}}(F_j) + y_{d:d}\right] \\ &= (y_{j:d} - y_{j+1:d})\, b_{\alpha_{u_i},\beta_{u_i}}(F_j), \end{aligned}$$

where $B_{\alpha_{u_i},\beta_{u_i}}$ and $b_{\alpha_{u_i},\beta_{u_i}}$ are the CDF and PDF, respectively, of a beta random variable with parameters $\alpha_{u_i} = (d+1)u_i$ and $\beta_{u_i} = (d+1)(1-u_i)$. This completes the proof. □

Note that the formulas and results established in Theorem 4.1 of Wang and Hutson (2011), which we presented in Section 3.2.3, can be readily inferred from Theorem 3.1 by choosing $l = 1$.

Next, in Table 3.2 we provide values of asymptotic means, covariance-variance ($\times n$) and correlation matrices of smoothed quartile estimators for selected binomial and zero-inflated binomial (ZIB) distributions. Several conclusions emerge from the table. Let us start with the Bernoulli case and prove that correlations among all quartile estimators are exactly 1. For Bernoulli, $d = 2$ and thus the matrix $\mathbf{D}$ has one entry: $F_1(1-F_1)$. For three quartiles ($u_1 = 0.25, u_2 = 0.50, u_3 = 0.75$), the matrix $\mathbf{H}$ has three entries: $\big(b_{\alpha_{u_1},\beta_{u_1}}(F_1), b_{\alpha_{u_2},\beta_{u_2}}(F_1), b_{\alpha_{u_3},\beta_{u_3}}(F_1)\big)'$. Denoting $b_i = b_{\alpha_{u_i},\beta_{u_i}}(F_1)$, we have

$$\mathbf{HDH}' = (b_1,b_2,b_3)'\big[F_1(1-F_1)\big](b_1,b_2,b_3) = F_1(1-F_1)\begin{bmatrix} b_1^2 & b_1b_2 & b_1b_3 \\ b_2b_1 & b_2^2 & b_2b_3 \\ b_3b_1 & b_3b_2 & b_3^2 \end{bmatrix} \tag{3.9}$$

And the ijth entry of the correlation matrix is $F_1(1-F_1)b_ib_j\left(F_1(1-F_1)b_i^2F_1(1-F_1)b_j^2\right)^{-1/2} =$

1. Farther, what is noticeable, and intuitively makes sense, is that asymptotic correlations are stronger between estimators of quartiles that are next to each other versus those that are further apart. For example, for binomial with $m > 1$ and ZIB with $c < 0.8$ distributions, the correlation entry $(1, 2)$ is significantly greater than $(1, 3)$. For ZIB, though, as the proportion of zeros gets larger, the estimates of quartiles approach zero and all correlations become almost 1. Also, as is known from large sample theory, under certain conditions on parameters, binomial distributions can be approximated by a normal distribution (which is symmetric). This explains why the means of quartiles are almost equally spaced and correlation entries $(1, 2)$ and $(2, 3)$ are nearly equal for binomial with $m > 1$.

Table 3.2: Asymptotic means, covariance-variance ($\times n$) and correlation matrices of smoothed quartile estimators $\widehat{Q}_Y(0.25), \widehat{Q}_Y(0.50), \widehat{Q}_Y(0.75)$ for binomial and ZIB distributions

Distribution	Means	$\mathbf{HDH}'$	Correlations
Bin $(m = d-1,\ q = 0.7)$			
$m = 1$	$\begin{bmatrix} 0.3434 \\ 0.7477 \\ 0.9548 \end{bmatrix}$	$\begin{bmatrix} 0.3262 & 0.3054 & 0.0915 \\ 0.3054 & 0.2860 & 0.0857 \\ 0.0915 & 0.0857 & 0.0257 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$
$m = 4$	$\begin{bmatrix} 2.0970 \\ 2.8557 \\ 3.5234 \end{bmatrix}$	$\begin{bmatrix} 1.5100 & 1.0074 & 0.5371 \\ 1.0074 & 1.0217 & 0.7916 \\ 0.5371 & 0.7916 & 0.9668 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.8110 & 0.4445 \\ 0.8110 & 1 & 0.7965 \\ 0.4445 & 0.7965 & 1 \end{bmatrix}$
$m = 8$	$\begin{bmatrix} 4.6637 \\ 5.6615 \\ 6.5771 \end{bmatrix}$	$\begin{bmatrix} 2.8883 & 1.8954 & 1.0608 \\ 1.8954 & 2.1927 & 1.6054 \\ 1.0608 & 1.6054 & 2.1432 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.7532 & 0.4264 \\ 0.7532 & 1 & 0.7406 \\ 0.4264 & 0.7406 & 1 \end{bmatrix}$
ZIB $(c,\ m = d-1 = 8,\ q = 0.7)$			
$c = 0.2$	$\begin{bmatrix} 2.4540 \\ 5.0900 \\ 6.3409 \end{bmatrix}$	$\begin{bmatrix} 30.3375 & 11.9030 & 3.9362 \\ 11.9030 & 6.3525 & 3.0037 \\ 3.9362 & 3.0037 & 2.8444 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.8574 & 0.4237 \\ 0.8574 & 1 & 0.7066 \\ 0.4237 & 0.7066 & 1 \end{bmatrix}$
$c = 0.5$	$\begin{bmatrix} 0.1743 \\ 2.2613 \\ 5.4760 \end{bmatrix}$	$\begin{bmatrix} 1.2535 & 6.7704 & 3.2850 \\ 6.7704 & 36.8979 & 18.5443 \\ 3.2850 & 18.5443 & 11.3601 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.9955 & 0.8705 \\ 0.9955 & 1 & 0.9058 \\ 0.8705 & 0.9058 & 1 \end{bmatrix}$
$c = 0.8$	$\begin{bmatrix} 0.0003 \\ 0.0821 \\ 2.0124 \end{bmatrix}$	$\begin{bmatrix} 0.0000 & 0.0032 & 0.0297 \\ 0.0032 & 0.4942 & 4.6106 \\ 0.0297 & 4.6106 & 43.5056 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.9994 & 0.9905 \\ 0.9994 & 1 & 0.9944 \\ 0.9905 & 0.9944 & 1 \end{bmatrix}$

3.3.3 Simulation Study

A Monte Carlo simulation study was conducted to verify and augment the asymptotic properties proved in Theorem 3.1. The study was performed for the following choices of simulation parameters:

- *Discrete distributions*:

  - Binomial, $Bin\,(m = d-1, q = 0.7)$: $m = 1,\ 4,\ 8.$
  - Zero-inflated binomial, $ZIB\,(c,\ m = d-1 = 8,\ q = 0.7)$: $c = 0.2,\ 0.5,\ 0.8.$

- *Sample size*: $n = 50,\ 100,\ 500.$
- *Estimated quantiles*, $(Q_Y(u_1), Q_Y(u_2), Q_Y(u_3))$: $u_1 = 0.25,\ u_2 = 0.50,\ u_3 = 0.75.$

From a specified discrete distribution, we generate 100,000 samples of a specified length n. For each sample we estimate the vector $\big(Q_Y(0.25), Q_Y(0.50), Q_Y(0.75)\big)$ according to (3.7) and then, based on those 100,000 estimates, compute the averages and variances of the vector coordinates, as well as sample covariances between the coordinates.

The results are summarized in Table 3.3, where the column $n = \infty$ corresponds to the asymptotic vectors and covariance-variance matrix entries which were derived in Section 3.3.2 and are included here as reference point. Overall, the table reveals that the sample estimates are very similar to the true quantities, including all entries of the covariance-variance matrices. Also, the convergence rate is fairly fast for binomial distributions, with samples as small as $n = 50$ practically matching the $n = \infty$ case, but requires much larger samples for ZIB distributions with $c \geq 0.5$.

Table 3.3: Estimated means and covariance-variance ($\times n$) matrices of smoothed quartile estimators $\widehat{Q}_Y(0.25), \widehat{Q}_Y(0.50), \widehat{Q}_Y(0.75)$ for selected binomial and ZIB models.

	$n = 50$	$n = 100$	$n = 500$	$n = \infty$
Bin $(m = 1,\ q = 0.7)$				
$\widehat{\text{means}}$	$(0.35, 0.75, 0.95)$	$(0.35, 0.75, 0.95)$	$(0.34, 0.75, 0.95)$	$(0.343, 0.748, 0.955)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.33 & 0.30 & 0.09 \\ 0.30 & 0.28 & 0.09 \\ 0.09 & 0.09 & 0.03 \end{bmatrix}$	$\begin{bmatrix} 0.33 & 0.30 & 0.09 \\ 0.30 & 0.28 & 0.09 \\ 0.09 & 0.09 & 0.03 \end{bmatrix}$	$\begin{bmatrix} 0.33 & 0.31 & 0.09 \\ 0.31 & 0.29 & 0.09 \\ 0.09 & 0.09 & 0.03 \end{bmatrix}$	$\begin{bmatrix} 0.326 & 0.305 & 0.092 \\ 0.305 & 0.286 & 0.086 \\ 0.092 & 0.086 & 0.026 \end{bmatrix}$
Bin $(m = 4,\ q = 0.7)$				
$\widehat{\text{means}}$	$(2.10, 2.85, 3.51)$	$(2.10, 2.86, 3.52)$	$(2.10, 2.86, 3.52)$	$(2.097, 2.856, 3.523)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 1.50 & 1.00 & 0.54 \\ 1.00 & 1.04 & 0.79 \\ 0.54 & 0.79 & 0.95 \end{bmatrix}$	$\begin{bmatrix} 1.51 & 1.01 & 0.55 \\ 1.01 & 1.03 & 0.79 \\ 0.55 & 0.79 & 0.96 \end{bmatrix}$	$\begin{bmatrix} 1.51 & 1.01 & 0.54 \\ 1.01 & 1.03 & 0.80 \\ 0.54 & 0.80 & 0.97 \end{bmatrix}$	$\begin{bmatrix} 1.510 & 1.007 & 0.537 \\ 1.007 & 1.022 & 0.792 \\ 0.537 & 0.792 & 0.967 \end{bmatrix}$
Bin $(m = 8,\ q = 0.7)$				
$\widehat{\text{means}}$	$(4.67, 5.66, 6.57)$	$(4.67, 5.66, 6.57)$	$(4.66, 5.66, 6.58)$	$(4.664, 5.662, 6.577)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 2.86 & 1.90 & 1.07 \\ 1.90 & 2.17 & 1.62 \\ 1.07 & 1.62 & 2.12 \end{bmatrix}$	$\begin{bmatrix} 2.88 & 1.90 & 1.07 \\ 1.90 & 2.17 & 1.61 \\ 1.07 & 1.61 & 2.13 \end{bmatrix}$	$\begin{bmatrix} 2.89 & 1.90 & 1.06 \\ 1.90 & 2.19 & 1.61 \\ 1.06 & 1.61 & 2.14 \end{bmatrix}$	$\begin{bmatrix} 2.888 & 1.895 & 1.061 \\ 1.895 & 2.193 & 1.605 \\ 1.061 & 1.605 & 2.143 \end{bmatrix}$
ZIB $(c = 0.2,\ m = 8,\ q = 0.7)$				
$\widehat{\text{means}}$	$(2.50, 5.04, 6.32)$	$(2.48, 5.07, 6.33)$	$(2.46, 5.08, 6.34)$	$(2.454, 5.090, 6.341)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 27.31 & 11.82 & 3.89 \\ 11.82 & 7.06 & 3.17 \\ 3.89 & 3.17 & 2.85 \end{bmatrix}$	$\begin{bmatrix} 28.69 & 11.84 & 3.91 \\ 11.84 & 6.70 & 3.09 \\ 3.91 & 3.09 & 2.85 \end{bmatrix}$	$\begin{bmatrix} 29.98 & 11.89 & 3.93 \\ 11.89 & 6.42 & 3.02 \\ 3.93 & 3.02 & 2.85 \end{bmatrix}$	$\begin{bmatrix} 30.338 & 11.903 & 3.936 \\ 11.903 & 6.352 & 3.004 \\ 3.936 & 3.004 & 2.844 \end{bmatrix}$
ZIB $(c = 0.5,\ m = 8,\ q = 0.7)$				
$\widehat{\text{means}}$	$(0.23, 2.30, 5.39)$	$(0.20, 2.28, 5.44)$	$(0.18, 2.26, 5.47)$	$(0.174, 2.261, 5.476)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 2.12 & 7.62 & 3.65 \\ 7.62 & 32.81 & 18.11 \\ 3.65 & 18.11 & 12.77 \end{bmatrix}$	$\begin{bmatrix} 1.66 & 7.18 & 3.45 \\ 7.18 & 34.48 & 18.26 \\ 3.45 & 18.26 & 12.05 \end{bmatrix}$	$\begin{bmatrix} 1.33 & 6.86 & 3.32 \\ 6.86 & 36.41 & 18.52 \\ 3.32 & 18.52 & 11.53 \end{bmatrix}$	$\begin{bmatrix} 1.253 & 6.770 & 3.285 \\ 6.770 & 36.898 & 18.544 \\ 3.285 & 18.544 & 11.360 \end{bmatrix}$
ZIB $(c = 0.8,\ m = 8,\ q = 0.7)$				
$\widehat{\text{means}}$	$(0.00, 0.13, 2.04)$	$(0.00, 0.11, 2.03)$	$(0.00, 0.09, 2.02)$	$(0.000, 0.082, 2.012)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.00 & 0.02 & 0.08 \\ 0.02 & 1.05 & 5.62 \\ 0.08 & 5.62 & 38.30 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.01 & 0.05 \\ 0.01 & 0.76 & 5.17 \\ 0.05 & 5.17 & 40.77 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.03 \\ 0.00 & 0.54 & 4.72 \\ 0.03 & 4.72 & 42.90 \end{bmatrix}$	$\begin{bmatrix} 0.000 & 0.003 & 0.030 \\ 0.003 & 0.494 & 4.611 \\ 0.030 & 4.611 & 43.506 \end{bmatrix}$

NOTE: The entries for $n < \infty$ are the averages and sample covariances of estimated quartiles. Results are based on 100,000 simulated samples. Standard errors of these entries are ≤ 0.003.

Chapter 4

Methodological Extensions

The smoothed quantile function (3.1), its properties, and probabilistic behavior of its estimator are valid for discrete distributions with finite support, i.e., when $d < \infty$. For discrete distributions with infinite support, such as Poisson, negative binomial or their zero-inflated versions, $d = \infty$. Since such distributions are essential for modeling claim frequency, we need to extend the results of Chapter 3 to the case $d = \infty$. Thus, in this chapter we first investigate the proposal of Wang and Hutson (2011) on how to deal with such distributions, then make a new proposal and evaluate its performance, theoretically and via simulations.

4.1 Approximate Weights and Quantiles

For discrete distributions with infinitely countable support, Wang and Hutson (2011) argued that d can be viewed as a smoothing parameter which may be chosen arbitrarily (say, d_*). Thus, $Q_Y(u)$ formula can be approximated by replacing the weights $w_{j(u)}$, which are based on a beta random variable B, with those based on a normal distribution:

$$Q_*(u) \;=\; \sum_{j=1}^{d_*} \Big[\Phi\big((F_j-\mu_u)/\sigma_u\big) - \Phi\big((F_{j-1}-\mu_u)/\sigma_u\big)\Big] y_{j:d_*} \;=:\; \sum_{j=1}^{d_*} w^*_{j(u)}\, y_{j:d_*}, \quad (4.1)$$

where Φ is the CDF of the standard normal distribution, $\mu_u = \mathbf{E}\big[B\big] = u$ and $\sigma_u^2 = \mathbf{Var}\big[B\big] = u(1-u)/(d_*+2)$. Equivalently, (4.1) can be rewritten in terms of data spacings:

$$\begin{aligned} Q_*(u) &= \sum_{j=1}^{d_*} \Big[\Phi\big((F_j-\mu_u)/\sigma_u\big) - \Phi\big((F_{j-1}-\mu_u)/\sigma_u\big)\Big] y_{j:d_*} = -\, y_{1:d_*}\, \Phi\left(\frac{-\mu_u}{\sigma_u}\right) \\ &+ \sum_{j=1}^{d_*-1} \big(y_{j:d_*} - y_{j+1:d_*}\big)\Phi\left(\frac{F_j-\mu_u}{\sigma_u}\right) + y_{d_*:d_*}\, \Phi\left(\frac{F_{d_*}-\mu_u}{\sigma_u}\right). \end{aligned} \tag{4.2}$$

Moreover, noticing that for standard discrete distributions from the class $(a,b,0)$ or $(a,b,1)$ we have $y_{1:d_*} = 0$, $y_{j+1:d_*} - y_{j:d_*} = 1$, and $y_{d_*:d_*} = d_* - 1$, the formula (4.2) can be further reduced to

$$Q_*(u) = (d_*-1)\,\Phi\left(\frac{F_{d_*}-\mu_u}{\sigma_u}\right) - \sum_{j=1}^{d_*-1} \Phi\left(\frac{F_j-\mu_u}{\sigma_u}\right). \tag{4.3}$$

It was also mentioned that a reasonable choice of d_* could be the number of data points which occupy a major proportion of probability mass across the whole sample space, but no details provided.

In Figure 4.1, the density curves of $N\left(\mu_u\,, \sigma_u^2\right)$ are plotted for various quantile levels u. On the horizontal axes, the F_j marks were computed for Poisson distribution with $\lambda = 5$ and selected values of d_*. Of course, in theory d_* should be large, but note that even for $d_* = 10$ the F_j's for $j > 8$ are tightly clustered near 1, which makes the corresponding weights $w^*_{j(u)}$ practically equal to 0. In Figure 4.2, we plot the quantile function $Q_Y(u)$, defined by (3.1), with F_j representing the CDF of Poisson distribution for various choices of parameter λ, and truncated at d. In the left panel, each curve is overlaid with its normal distribution based approximation (4.1), with $d_* = d = 100$. The approximation works very well except for extreme right tail (e.g., $u \geq 0.90$), where it diverges from

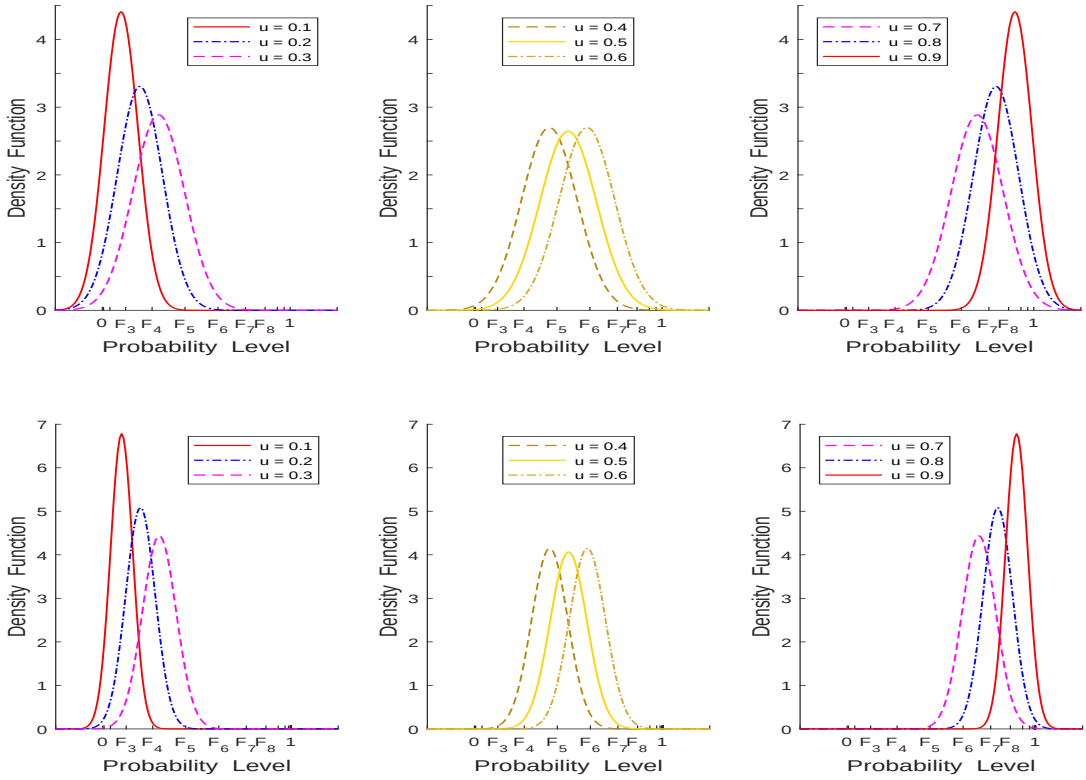

Figure 4.1: The density curves of $N(\mu_u, \sigma_u^2)$ used to compute weights $w^*_{j(u)}$ for quantile levels $u = 0.1, \ldots, 0.9$. The weight $w^*_{j(u)}$ corresponds to the area under the curve over $[F_{j-1}; F_j]$, where $F_0 = 0$ and $F_1, F_2, \ldots$ are CDF's of $P(\lambda = 5)$ with $d_* = 10$ (top row) and $d_* = 25$ (bottom row).

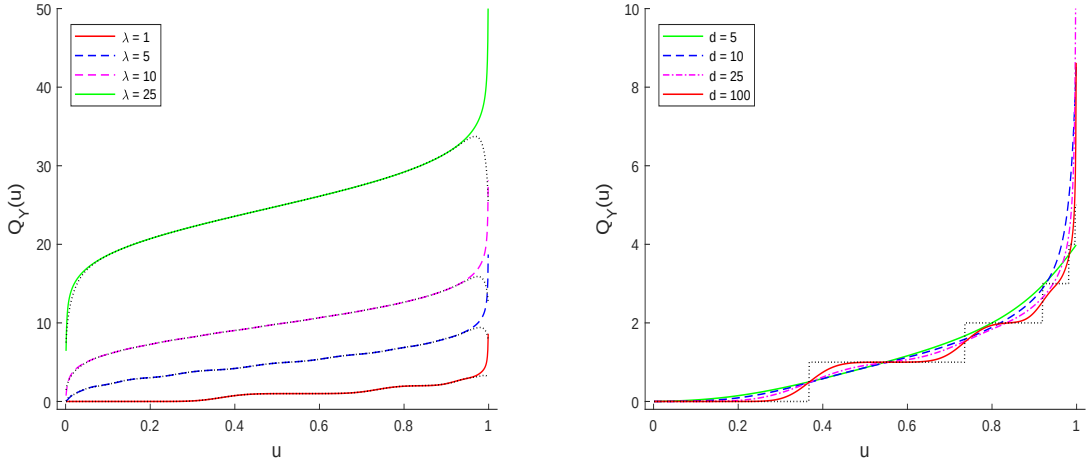

Figure 4.2: The quantile functions of $P(\lambda)$ distributions. *Left panel*: $\lambda = 1, 5, 10, 25$ and $d = 100$. The overlaid dotted curves are $Q_*(u)$ approximations with $d_* = 100$. *Right panel*: $\lambda = 1$ and $d = 5, 10, 25, 100$. The dotted step line corresponds to the classical discrete quantile function.

$Q_Y(u)$. This might have been anticipated from careful examination of the right column of Figure 4.1. There, the density curves are relatively high but the intervals $[F_{j-1}; F_j]$ are extremely narrow, which yields small weights (almost zero). Then, such weights multiplied by not-too-large $y_{j:d}$'s make neglible contributions toward $Q_Y(u)$ resulting in undervaluation of the right tail of the function. Also, the waves of the quantile function (see discussions of Figures 3.2 and 3.3) are now clearly visible for $\lambda = 1$ (left panel) and especially for $d = 100$ (right panel). Overall, it seems that normally-distributed weights provide no advantage over the formulas based on beta-distributed weights. Therefore, from now on we will focus on the original formulation of the $Q_Y(u)$ formula, i.e., with "beta" weights. Next, to better understand the effects of d on population and sample evaluations of $Q_Y(u)$, we performed a simulation study for selected Poisson models, $P(\lambda)$, and several choices of d. The study design is similar to the one of Section 3.3.3. The results are summarized in Table 4.1, where the means and covariance-variance matrices of $\widehat{Q}_Y(0.25), \widehat{Q}_Y(0.50), \widehat{Q}_Y(0.75)$ are reported for various sample sizes n. While at this

moment we do not know what is supposed to happen to those estimators as $n \to \infty$, let's see what are the parameters of the asymptotic distribution of Theorem 3.1 (truncated at the same d as the sample estimators). The parameter values are provided in Table 4.2, where we see that they are quite close to the simulated values for $n = 10^4$. However, there are significant differences between the theoretical and simulated values for smaller n. We also notice a relationship between the choice of d and λ. These observations will be more rigorously examined in Section 4.2.

Table 4.1: Estimated means and covariance-variance ($\times n$) matrices of smoothed quartile estimators $\widehat{Q}_Y(0.25), \widehat{Q}_Y(0.50), \widehat{Q}_Y(0.75)$ for selected Poisson models, $P(\lambda)$, and several d's.

λ	d		$n = 10^2$	$n = 10^3$	$n = 10^4$
1	20	$\widehat{\text{means}}$	$(0.13, 0.89, 1.60)$	$(0.11, 0.90, 1.60)$	$(0.11, 0.90, 1.60)$
		$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.84 & 0.79 & 0.68 \\ 0.79 & 1.10 & 1.07 \\ 0.68 & 1.07 & 3.14 \end{bmatrix}$	$\begin{bmatrix} 0.71 & 0.79 & 0.70 \\ 0.79 & 0.94 & 1.02 \\ 0.70 & 1.02 & 3.39 \end{bmatrix}$	$\begin{bmatrix} 0.69 & 0.78 & 0.70 \\ 0.78 & 0.90 & 1.00 \\ 0.70 & 1.00 & 3.38 \end{bmatrix}$
	100	$\widehat{\text{means}}$	$(0.03, 0.97, 1.59)$	$(0.01, 0.99, 1.63)$	$(0.01, 1.00, 1.64)$
		$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.62 & 0.10 & 0.61 \\ 0.10 & 0.48 & 0.63 \\ 0.61 & 0.63 & 8.07 \end{bmatrix}$	$\begin{bmatrix} 0.04 & 0.02 & 0.30 \\ 0.02 & 0.03 & 0.24 \\ 0.30 & 0.24 & 13.01 \end{bmatrix}$	$\begin{bmatrix} 0.02 & 0.02 & 0.25 \\ 0.02 & 0.01 & 0.19 \\ 0.25 & 0.19 & 13.68 \end{bmatrix}$
10	20	$\widehat{\text{means}}$	$(7.72, 9.85, 12.15)$	$(7.71, 9.84, 12.15)$	$(7.71, 9.84, 12.15)$
		$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 12.75 & 9.36 & 6.32 \\ 9.36 & 12.84 & 10.93 \\ 6.32 & 10.93 & 17.34 \end{bmatrix}$	$\begin{bmatrix} 12.85 & 9.33 & 6.15 \\ 9.33 & 12.77 & 10.79 \\ 6.15 & 10.79 & 17.29 \end{bmatrix}$	$\begin{bmatrix} 12.81 & 9.38 & 6.22 \\ 9.38 & 12.87 & 10.89 \\ 6.22 & 10.89 & 17.35 \end{bmatrix}$
	100	$\widehat{\text{means}}$	$(7.78, 9.84, 12.05)$	$(7.78, 9.84, 12.06)$	$(7.78, 9.85, 12.06)$
		$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 14.56 & 9.10 & 6.22 \\ 9.10 & 14.53 & 10.59 \\ 6.22 & 10.59 & 19.39 \end{bmatrix}$	$\begin{bmatrix} 14.61 & 8.76 & 5.76 \\ 8.76 & 13.47 & 9.78 \\ 5.76 & 9.78 & 18.27 \end{bmatrix}$	$\begin{bmatrix} 14.62 & 8.76 & 5.78 \\ 8.76 & 13.25 & 9.67 \\ 5.78 & 9.67 & 17.86 \end{bmatrix}$

NOTE: The entries are the averages and sample covariances of estimated quartiles. Results are based on 100,000 simulated samples. Standard errors of these entries are ≤ 0.003.

Table 4.2: Asymptotic means and covariance-variance ($\times n$) matrices of smoothed quartile estimators $\widehat{Q}_Y(0.25), \widehat{Q}_Y(0.50), \widehat{Q}_Y(0.75)$ for selected Poisson distributions, $P(\lambda)$.

	$\lambda = 1$		$\lambda = 10$	
	$d = 20$	$d = 100$	$d = 20$	$d = 100$
means	$(0.11, 0.90, 1.60)$	$(0.01, 1.00, 1.64)$	$(7.71, 9.84, 12.15)$	$(7.78, 9.85, 12.06)$
$\mathbf{HDH}'$	$\begin{bmatrix} 0.69 & 0.78 & 0.70 \\ 0.78 & 0.91 & 1.01 \\ 0.70 & 1.01 & 3.40 \end{bmatrix}$	$\begin{bmatrix} 0.02 & 0.02 & 0.25 \\ 0.02 & 0.01 & 0.18 \\ 0.25 & 0.18 & 13.84 \end{bmatrix}$	$\begin{bmatrix} 12.83 & 9.37 & 6.22 \\ 9.37 & 12.89 & 10.91 \\ 6.22 & 10.91 & 17.41 \end{bmatrix}$	$\begin{bmatrix} 14.64 & 8.73 & 5.76 \\ 8.73 & 13.20 & 9.66 \\ 5.76 & 9.66 & 17.87 \end{bmatrix}$

4.2 Truncated Weights and Quantiles

The observations made from the simulation study of Section 4.1 prompt us to look into the choice of d more carefully. In particular, for a discrete distribution with infinite support one has to choose not only d but, more importantly, the minimum and maximum distinct points, $y_{1:d}$ and $y_{d:d}$. Of course, for samples this is not a problem because observed data will always have finite support. Thus, in this section we propose to use $Q_Y(u)$ formulas with $y_{1:d}$, $y_{d:d}$, and d estimated from data. Properties of such estimators are investigated theoretically as well as via simulations.

4.2.1 Estimation

Suppose we observe a random sample $Y_1, \ldots, Y_n$ of i.i.d. discrete non-negative random variables with CDF F, mean $\mathbf{E}[Y] < \infty$ and variance $\mathbf{Var}[Y] < \infty$. A major proportion of probability mass across the whole sample space will be covered by the following intervals:

$$\left[\widehat{L}_k;\ \widehat{U}_k\right] \ := \ \left[\overline{Y} - k\sqrt{S^2};\ \overline{Y} + k\sqrt{S^2}\right], \tag{4.4}$$

where $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $S^2 = (n-1)^{-1}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2$, and $k \geq 2$ is a chosen constant. Since $\overline{Y} \xrightarrow{P} \mathbf{E}[Y]$ and $S^2 \xrightarrow{P} \mathbf{Var}[Y]$, intervals $\left[\widehat{L}_k;\ \widehat{U}_k\right]$ converge in probability to

$$\left[L_k;\ U_k\right] \ := \ \left[\mathbf{E}[Y] - k\sqrt{\mathbf{Var}[Y]};\ \mathbf{E}[Y] + k\sqrt{\mathbf{Var}[Y]}\right]. \tag{4.5}$$

According to Chebyshev's inequality, the probability that Y will fall in interval (4.5) is at least $1 - 1/k^2$. Saw *et al.* (1984) modified this inequality to accommodate estimated intervals such as (4.4). Their inequality was recently simplified by Kaban (2011):

$$\mathbf{P}\left\{\left|Y - \overline{Y}\right| \leq k\sqrt{S^2}\right\} \ \geq \ 1 - \frac{1}{n+1}\left[\frac{n+1}{n}\left(\frac{n-1}{k^2} + 1\right)\right], \tag{4.6}$$

Table 4.3: Probability bounds based on Chebyshev's and Kaban's inequalities for selected k and n.

k	Kaban's inequality (4.6)					Chebyshev's
	$n = 10$	$n = 25$	$n = 50$	$n = 75$	$n = 100$	inequality
2	0.727	0.731	0.745	0.750	0.743	0.750
3	0.818	0.885	0.882	0.882	0.881	0.889
4	0.909	0.923	0.922	0.934	0.931	0.938
5	0.909	0.923	0.941	0.947	0.950	0.960
7	0.909	0.962	0.961	0.974	0.970	0.980
10	0.909	0.962	0.980	0.987	0.980	0.990

where $[\cdot]$ denotes the greatest integer part.

In Table 4.3, we compute several probability bounds based on Chebyshev's inequality and on Kaban's inequality (4.6). Clearly, when $n > 50$ the two bounds are practically equal for most values of k. Also, for $k > 5$ improvements in the coverage probability are very small and perhaps not worth pursuing in practice, but this issue will be explored in detail in Sections 4.2.2 and 4.2.3.

Based on probability bound values provided in Table 4.3, we now have clear understanding about what k in (4.4) and (4.5) will result in a "major" proportion of probability mass coverage. Thus, for $k \geq 1$, we propose the following sample and population definitions for d, $y_{1:d}$, and $y_{d:d}$:

$$\text{Sample:} \quad \widehat{d}_k = \widehat{y}_{\widehat{d}_k:\widehat{d}_k} - \widehat{y}_{1:\widehat{d}_k} + 1, \quad \widehat{y}_{1:\widehat{d}_k} = \max\left\{0, \left[\widehat{L}_k\right]\right\}, \quad \widehat{y}_{\widehat{d}_k:\widehat{d}_k} = \left[\widehat{U}_k\right] + 1, \tag{4.7}$$

$$\text{Population:} \quad d_k = y_{d_k:d_k} - y_{1:d_k} + 1, \quad y_{1:d_k} = \max\left\{0, \left[L_k\right]\right\}, \quad y_{d_k:d_k} = \left[U_k\right] + 1, \tag{4.8}$$

where $[\cdot]$ denotes the greatest integer part, and $\widehat{L}_k$, $\widehat{U}_k$ and L_k, U_k are defined in (4.4) and (4.5), respectively. Note that $\widehat{y}_{1:\widehat{d}_k}$ and $\widehat{y}_{\widehat{d}_k:\widehat{d}_k}$ are not necessarily *observed* distinct

minimum and maximum of the sample. Likewise, $y_{1:d_k}$ and $y_{d_k:d_k}$ are not necessarily distinct minimum and maximum of the population. By imposing these upper and lower bounds on possible values of data – in a sample and population – we defined a *truncated sample* and a *truncated population*.

Under this setting of truncated samples and populations, the definitions of CDF's $\widehat{F}_{j(k)} = \widehat{F}_Y(\widehat{y}_{j:\widehat{d}_k})$ and $F_{j(k)} = F_Y(y_{j:d_k})$ are the same as before (see Sections 3.1 and 3.2), but the boundary cases are not necessarily equal to 0 and 1, i.e., $\widehat{F}_{0(k)} \geq 0$, $F_{0(k)} \geq 0$ and $\widehat{F}_{\widehat{d}_k(k)} \leq 1$, $F_{d_k(k)} \leq 1$. The truncated sample CDF $\widehat{F}^*_{j(k)}$ is related to the standard (non-truncated) sample cdf $\widehat{F}_{j(k)}$ as follows:

$$\widehat{F}^*_{j(k)} \;=\; \widehat{\mathbf{P}}\left\{Y \leq \widehat{y}_{j:\widehat{d}_k} \,\middle|\, \widehat{y}_{0:\widehat{d}_k} < Y \leq \widehat{y}_{\widehat{d}_k:\widehat{d}_k}\right\} \;=\; \frac{\widehat{F}_{j(k)} - \widehat{F}_{0(k)}}{\widehat{F}_{\widehat{d}_k(k)} - \widehat{F}_{0(k)}} \tag{4.9}$$

This shows that $\widehat{F}^*_{j(k)}$ can be almost equal to $\widehat{F}_{j(k)}$, when $\widehat{F}_{0(k)} \approx 0$ and $\widehat{F}_{\widehat{d}_k(k)} \approx 1$. The equivalent relationship also holds for the population CDF's $F^*_{j(k)}$ and $F_{j(k)}$:

$$F^*_{j(k)} \;=\; \mathbf{P}\left\{Y \leq y_{j:d_k} \,\middle|\, y_{0:d_k} < Y \leq y_{d_k:d_k}\right\} \;=\; \frac{F_{j(k)} - F_{0(k)}}{F_{d_k(k)} - F_{0(k)}} \tag{4.10}$$

Using equation (4.9) in conjunction with (3.4), we define the smoothed uth quantile for (truncated) discrete sample as follows:

$$\widehat{Q}^{(k)}_*(u) \;=\; \sum_{j=1}^{\widehat{d}_k} \widehat{w}^{(k)}_{j(u)} \widehat{y}_{j:\widehat{d}_k} \;=\; \sum_{j=1}^{\widehat{d}_k} \left[B_{\widehat{\alpha}_u,\widehat{\beta}_u}(\widehat{F}^*_{j(k)}) - B_{\widehat{\alpha}_u,\widehat{\beta}_u}(\widehat{F}^*_{j(k)-1})\right] \widehat{y}_{j:\widehat{d}_k}, \tag{4.11}$$

where $B_{\widehat{\alpha}_u,\widehat{\beta}_u}$ denotes the beta variable CDF with parameters $\widehat{\alpha}_u = (\widehat{d}_k + 1)u$ and $\widehat{\beta}_u = (\widehat{d}_k + 1)(1-u)$. Likewise, using (4.10), we define the smoothed uth quantile for the

truncated discrete population:

$$Q_*^{(k)}(u) \;=\; \sum_{j=1}^{d_k} w_{j(u)}^{(k)} y_{j:d_k} \;=\; \sum_{j=1}^{d_k} \Big[B_{\alpha_u,\beta_u}(F^*_{j(k)}) - B_{\alpha_u,\beta_u}(F^*_{j(k)-1}) \Big] y_{j:d_k}, \tag{4.12}$$

where B_{α_u,β_u} denotes the beta variable CDF with parameters $\alpha_u = (d_k+1)u$ and $\beta_u = (d_k+1)(1-u)$.

Note that both quantile functions, $\widehat{Q}_*^{(k)}$ and $Q_*^{(k)}$, are directly related to their non-truncated versions, $\widehat{Q}_Y^{(k)}$ and $Q_Y^{(k)}$, respectively. Indeed, focusing on $Q_*^{(k)}$, we see that it can be interpreted as the inverse function of a smoothed version of F^* which is related to a similarly smoothed version of F_Y through (4.10). Inverting (4.10) for smoothed F^* and F_Y leads to the following formula relating $Q_*^{(k)}$ to $Q_Y^{(k)}$ (which is the inverse of smoothed F_Y):

$$Q_*^{(k)}(u) \;=\; Q_Y^{(k)}\Big(F_{0(k)} + u\big(F_{d_k(k)} - F_{0(k)}\big)\Big). \tag{4.13}$$

It is clear from (4.13) that $F_{0(k)} \approx 0$ and $F_{d_k(k)} \approx 1$ implies $Q_*^{(k)}(u) \;\approx\; Q_Y^{(k)}(u)$. This point is further illustrated in Figure 4.3, where the quantile functions $Q_Y^{(k)}\left(F_{0(k)} + u\big(F_{d_k(k)} - F_{0(k)}\big)\right)$ of the Poisson distribution $P(\lambda)$, with $\lambda = 1, 2, 5,$ and 10, are plotted. The smooth curves are constructed using data truncation intervals $\mathbf{E}[Y] \pm k\sqrt{\mathbf{Var}[Y]}$. As we can see from the figure, the curves with $k = 1, 2$ are truncated too severely and thus fail to cover most of the PMF support. On the other hand, for $k \geq 3$ the smooth curves capture both tails of the probability distribution fairly well.

4.2.2 Asymptotic Properties

Let us continue with the setup of Section 4.2.1, and for $0 < u_1 < \cdots < u_l < 1$ define the vector of smoothed sample quantiles

$$\left(\widehat{Q}_*^{(k)}(u_1), \ldots, \widehat{Q}_*^{(k)}(u_l)\right), \tag{4.14}$$

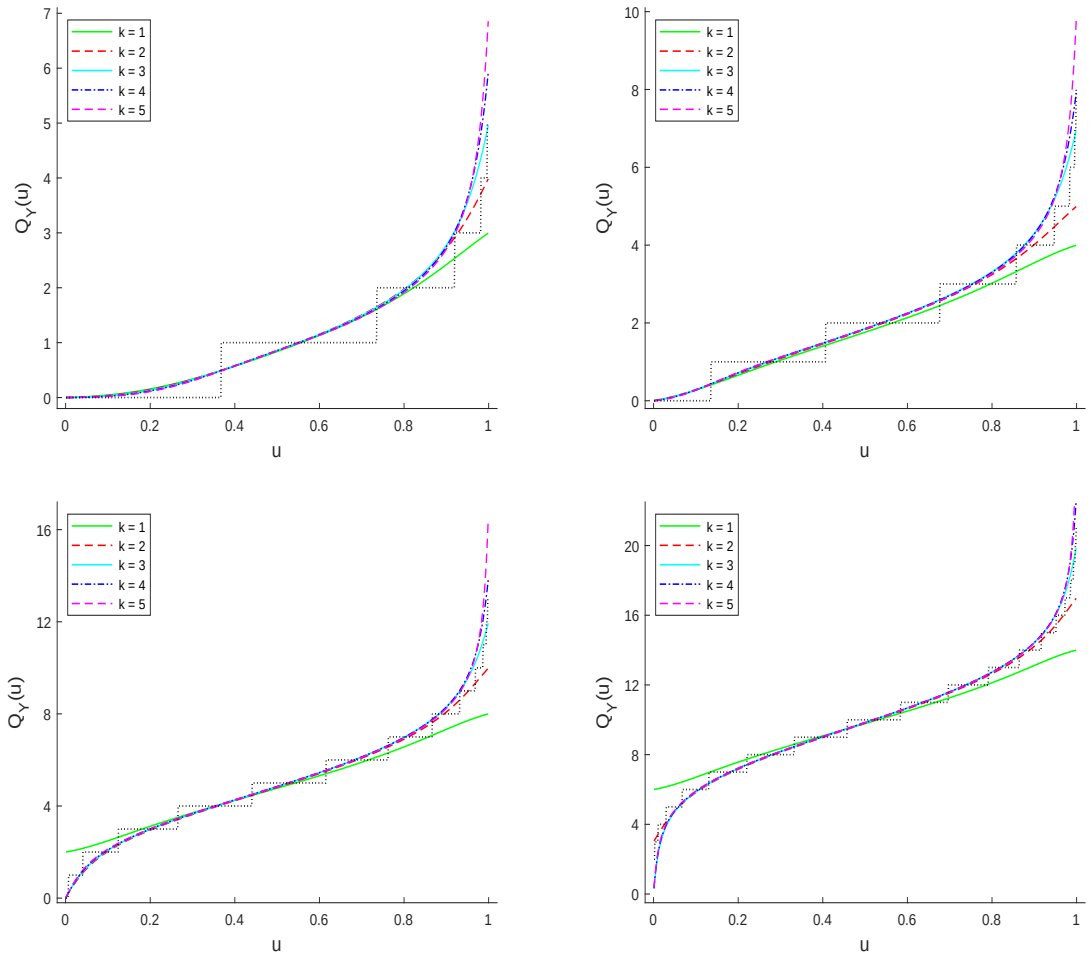

Figure 4.3: The quantile function $Q_Y^{(k)}(u)$ of $P(\lambda)$ distributions for data truncation intervals $\mathbf{E}[Y] \pm k\sqrt{\mathbf{Var}[Y]}$. *Top panel*: $\lambda = 1$ (left), $\lambda = 2$ (right). *Bottom panel*: $\lambda = 5$ (left), $\lambda = 10$ (right). The dotted step line corresponds to the classical discrete quantile function.

where each $\widehat{Q}_*^{(k)}(u_i)$ is given by (4.11). We conjecture that the vector of smoothed sample quantiles, given by (4.14), is asymptotically normal and a consistent estimator of the vector of smoothed population quantiles

$$\Big(Q_*^{(k)}(u_1), \ldots, Q_*^{(k)}(u_l)\Big), \tag{4.15}$$

where each $Q_*^{(k)}(u_i)$ is given by (4.12). These properties are summarized in the following conjecture.

Conjecture 4.1. *Consider an i.i.d. sample of size n from a discrete distribution F_Y with infinite support $y_{1:d} < y_{2:d} < y_{3:d} < \cdots$ and $d = \infty$. Suppose a truncated version of the sample, $y_{1:d_k} < y_{2:d_k} < \cdots < y_{d_k:d_k}$ with $d_k < \infty$, is constructed. Then, as $n \to \infty$, the following statements hold:*

$$(i) \quad \Big(\widehat{Q}_*^{(k)}(u_1), \ldots, \widehat{Q}_*^{(k)}(u_l)\Big) \xrightarrow{P} \Big(Q_*^{(k)}(u_1), \ldots, Q_*^{(k)}(u_l)\Big),$$

$$(ii) \quad \Big(\widehat{Q}_*^{(k)}(u_1), \ldots, \widehat{Q}_*^{(k)}(u_l)\Big) \sim \mathcal{AN}\left(\big(Q_*^{(k)}(u_1), \ldots, Q_*^{(k)}(u_l)\big), \frac{1}{n}\mathbf{HDH}'\right),$$

where $\mathbf{D} := \big[d_{ij}\big]_{(d_k-1)\times(d_k-1)}$ with $d_{ij} = d_{ji} = F^*_{i(k)}(1 - F^*_{j(k)})$, $i(k) \le j(k)$, and $\mathbf{H} := \big[h_{ij}\big]_{l\times(d_k-1)}$ with $h_{ij} = (y_{j:d_k} - y_{j+1:d_k})\, b_{\alpha_{u_i},\beta_{u_i}}(F^*_{j(k)})$. Here $b_{\alpha_{u_i},\beta_{u_i}}$ denotes the PDF *of a beta random variable with parameters $\alpha_{u_i} = (d_k+1)u_i$ and $\beta_{u_i} = (d_k+1)(1-u_i)$, and $F^*_{j(k)} = \mathbf{P}\left\{Y \le y_{j:d_k} \,\middle|\, y_{0:d_k} < Y \le y_{d_k:d_k}\right\} = \big(F_Y(y_{j:d_k}) - F_Y(y_{0:d_k})\big)/\big(F_Y(y_{d_k:d_k}) - F_Y(y_{0:d_k})\big)$.*

4.2.3 Simulation Study

Here we use simulations to cross-check the asymptotic properties specified in Conjecture 4.1. The study was performed for the following choices of distributions and simulation parameters:

- *Discrete distributions*:

  - Poisson, $P(\lambda)$: $\lambda = 1,\ 10$.
  - Zero-inflated Poisson, $ZIP(c,\ \lambda = 10)$: $c = 0.2,\ 0.8$.

- *Sample size*: $n = 50,\ 100,\ 500$.
- *Truncation intervals*, $\mathbf{E}[Y] \pm k\sqrt{\mathbf{Var}[Y]}$: $k = 1:5,\ 10$.
- *Estimated quantiles*, $(Q_Y(u_1), Q_Y(u_2), Q_Y(u_3))$: $u_1 = 0.25,\ u_2 = 0.50,\ u_3 = 0.75$.

From a specified discrete distribution, we generate 100,000 samples of a specified length n. For each sample we estimate the vector $\big(Q_Y(0.25), Q_Y(0.50), Q_Y(0.75)\big)$ according to (4.14) and then, based on those 100,000 estimates, compute the averages and variances of the vector coordinates, as well as sample covariances between the coordinates.

The simulation results are summarized in Tables 4.4-4.7, where the column $n = \infty$ corresponds to the asymptotic vectors and covariance-variance matrix entries which were specified in Conjecture 4.1 and are included here as reference point. We note right away that for $k = 1, 2$ the simulated and conjectured results differ from the other (more stable) cases. This outcome could be inferred from the graphs of Figure 4.3, where it is clear that $\widehat{Q}_Y^{(1)}(u)$ and $\widehat{Q}_Y^{(2)}(u)$ missed the tails of the distribution. For $k \geq 3$, the sample estimates are fairly close to the conjectured asymptotic quantities, including the entries of the covariance-variance matrices. They are also fairly similar across different k's. The convergence seems to be fast for $\lambda = 10$ and slower for $\lambda = 1$. For the zero-inflated distribution with $c = 0.2$, the convergence rates of simulated quantities are fast. In addition, for $c = 0.8$, there is a clear pattern: estimator means and most entries of the covariance-variance matrix converge to 0 (as they should because 80% of data are 0's), but the variance of $\widehat{Q}_Y^{(k)}(0.75)$ is large and keeps getting larger as k increases. Overall, the new approach based on truncated samples produces more stable results and faster

convergence than the original approach of choosing arbitrarily large d (see Tables 4.1-4.2). In summary, our recommendation is to construct data truncation bounds using $k = 3, 4, 5$.

Table 4.4: Estimated means and covariance-variance ($\times n$) matrices of smoothed quartile estimators $\widehat{Q}_Y^{(k)}(0.25), \widehat{Q}_Y^{(k)}(0.50), \widehat{Q}_Y^{(k)}(0.75)$ for $k = 1:5, 10$ and $\mathrm{P}(\lambda = 1)$.

	$n = 50$	$n = 100$	$n = 500$	$n = \infty$
$k = 1$				
$\widehat{\text{means}}$	$(0.25, 0.81, 1.56)$	$(0.24, 0.80, 1.56)$	$(0.23, 0.80, 1.55)$	$(0.247, 0.864, 1.731)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.43 & 0.70 & 0.68 \\ 0.70 & 1.38 & 1.74 \\ 0.68 & 1.74 & 2.99 \end{bmatrix}$	$\begin{bmatrix} 0.43 & 0.74 & 0.79 \\ 0.74 & 1.57 & 2.20 \\ 0.79 & 2.20 & 4.11 \end{bmatrix}$	$\begin{bmatrix} 0.47 & 0.99 & 1.46 \\ 0.99 & 2.71 & 5.15 \\ 1.46 & 5.15 & 11.69 \end{bmatrix}$	$\begin{bmatrix} 0.440 & 0.661 & 0.478 \\ 0.661 & 1.151 & 1.115 \\ 0.478 & 1.115 & 1.748 \end{bmatrix}$
$k = 2$				
$\widehat{\text{means}}$	$(0.25, 0.86, 1.70)$	$(0.24, 0.85, 1.70)$	$(0.24, 0.85, 1.70)$	$(0.232, 0.860, 1.731)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.48 & 0.69 & 0.53 \\ 0.69 & 1.25 & 1.34 \\ 0.53 & 1.34 & 2.32 \end{bmatrix}$	$\begin{bmatrix} 0.46 & 0.68 & 0.52 \\ 0.68 & 1.27 & 1.39 \\ 0.52 & 1.39 & 2.46 \end{bmatrix}$	$\begin{bmatrix} 0.44 & 0.67 & 0.48 \\ 0.67 & 1.40 & 1.69 \\ 0.48 & 1.69 & 3.18 \end{bmatrix}$	$\begin{bmatrix} 0.494 & 0.730 & 0.536 \\ 0.730 & 1.247 & 1.257 \\ 0.536 & 1.257 & 2.196 \end{bmatrix}$
$k = 3$				
$\widehat{\text{means}}$	$(0.24, 0.86, 1.72)$	$(0.23, 0.86, 1.72)$	$(0.23, 0.86, 1.71)$	$(0.219, 0.856, 1.709)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.53 & 0.72 & 0.50 \\ 0.72 & 1.26 & 1.25 \\ 0.50 & 1.25 & 2.12 \end{bmatrix}$	$\begin{bmatrix} 0.51 & 0.72 & 0.48 \\ 0.72 & 1.26 & 1.24 \\ 0.48 & 1.24 & 2.13 \end{bmatrix}$	$\begin{bmatrix} 0.50 & 0.70 & 0.42 \\ 0.70 & 1.27 & 1.23 \\ 0.41 & 1.23 & 2.03 \end{bmatrix}$	$\begin{bmatrix} 0.539 & 0.777 & 0.553 \\ 0.777 & 1.281 & 1.277 \\ 0.553 & 1.277 & 2.300 \end{bmatrix}$
$k = 4$				
$\widehat{\text{means}}$	$(0.23, 0.86, 1.70)$	$(0.22, 0.86, 1.70)$	$0.21, 0.86, 1.70)$	$(0.206, 0.855, 1.689)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.58 & 0.76 & 0.50 \\ 0.76 & 1.29 & 1.24 \\ 0.50 & 1.24 & 2.10 \end{bmatrix}$	$\begin{bmatrix} 0.56 & 0.76 & 0.49 \\ 0.76 & 1.28 & 1.23 \\ 0.49 & 1.23 & 2.11 \end{bmatrix}$	$\begin{bmatrix} 0.55 & 0.74 & 0.43 \\ 0.74 & 1.28 & 1.20 \\ 0.43 & 1.20 & 1.95 \end{bmatrix}$	$\begin{bmatrix} 0.578 & 0.812 & 0.563 \\ 0.812 & 1.291 & 1.267 \\ 0.563 & 1.267 & 2.338 \end{bmatrix}$
$k = 5$				
$\widehat{\text{means}}$	$(0.22, 0.85, 1.68)$	$(0.21, 0.86, 1.68)$	$(0.20, 0.86, 1.68)$	$(0.200, 0.856, 1.672)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.63 & 0.79 & 0.51 \\ 0.79 & 1.31 & 1.24 \\ 0.51 & 1.24 & 2.14 \end{bmatrix}$	$\begin{bmatrix} 0.60 & 0.79 & 0.50 \\ 0.79 & 1.29 & 1.23 \\ 0.50 & 1.23 & 2.15 \end{bmatrix}$	$\begin{bmatrix} 0.59 & 0.78 & 0.46 \\ 0.78 & 1.30 & 1.21 \\ 0.46 & 1.21 & 2.04 \end{bmatrix}$	$\begin{bmatrix} 0.611 & 0.839 & 0.572 \\ 0.839 & 1.287 & 1.250 \\ 0.572 & 1.252 & 2.377 \end{bmatrix}$
$k = 10$				
$\widehat{\text{means}}$	$(0.19, 0.86, 1.63)$	$(0.17, 0.86, 1.63)$	$(0.16, 0.87, 1.63)$	$(0.152, 0.871, 1.625)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.79 & 0.84 & 0.55 \\ 0.84 & 1.32 & 1.21 \\ 0.55 & 1.21 & 2.39 \end{bmatrix}$	$\begin{bmatrix} 0.75 & 0.84 & 0.55 \\ 0.84 & 1.26 & 1.19 \\ 0.55 & 1.19 & 2.46 \end{bmatrix}$	$\begin{bmatrix} 0.71 & 0.86 & 0.55 \\ 0.86 & 1.21 & 1.19 \\ 0.55 & 1.19 & 2.53 \end{bmatrix}$	$\begin{bmatrix} 0.704 & 0.879 & 0.632 \\ 0.879 & 1.161 & 1.150 \\ 0.632 & 1.150 & 2.704 \end{bmatrix}$

NOTE: The entries for $n < \infty$ are the averages and sample covariances of estimated quartiles. Results are based on 100,000 simulated samples. Standard errors of these entries are ≤ 0.003.

Table 4.5: Estimated means and covariance-variance ($\times n$) matrices of smoothed quartile estimators $\widehat{Q}_Y^{(k)}(0.25), \widehat{Q}_Y^{(k)}(0.50), \widehat{Q}_Y^{(k)}(0.75)$ for $k = 1:5, 10$ and $\mathrm{P}(\lambda = 10)$.

	$n = 50$	$n = 100$	$n = 500$	$n = \infty$
$k = 1$				
$\widehat{\text{means}}$	$(7.58, 9.49, 11.40)$	$(7.56, 9.48, 11.41)$	$(7.54, 9.50, 11.46)$	$(7.695, 9.850, 12.161)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 13.29 & 10.82 & 8.35 \\ 10.82 & 13.68 & 13.04 \\ 8.35 & 13.04 & 17.40 \end{bmatrix}$	$\begin{bmatrix} 13.75 & 11.18 & 8.79 \\ 11.18 & 14.09 & 13.64 \\ 8.79 & 13.64 & 18.39 \end{bmatrix}$	$\begin{bmatrix} 16.26 & 13.52 & 12.03 \\ 13.52 & 17.30 & 18.91 \\ 12.03 & 18.91 & 27.62 \end{bmatrix}$	$\begin{bmatrix} 6.768 & 5.447 & 2.811 \\ 5.447 & 8.259 & 5.903 \\ 2.811 & 5.903 & 8.092 \end{bmatrix}$
$k = 2$				
$\widehat{\text{means}}$	$(7.67, 9.80, 12.05)$	$(7.66, 9.79, 12.04)$	$(7.65, 9.78, 12.05)$	$(7.689, 9.848, 12.176)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 12.38 & 9.61 & 6.92 \\ 9.61 & 12.95 & 11.92 \\ 6.92 & 11.92 & 18.35 \end{bmatrix}$	$\begin{bmatrix} 12.47 & 9.67 & 7.00 \\ 9.67 & 13.06 & 12.03 \\ 7.00 & 12.03 & 18.47 \end{bmatrix}$	$\begin{bmatrix} 12.56 & 9.58 & 6.68 \\ 9.58 & 12.87 & 11.64 \\ 6.68 & 11.64 & 18.00 \end{bmatrix}$	$\begin{bmatrix} 12.053 & 8.985 & 5.774 \\ 8.985 & 12.082 & 10.166 \\ 5.774 & 10.166 & 15.603 \end{bmatrix}$
$k = 3$				
$\widehat{\text{means}}$	$(7.73, 9.85, 12.13)$	$(7.72, 9.84, 12.13)$	$(7.71, 9.84, 12.13)$	$(7.712, 9.844, 12.145)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 12.66 & 9.35 & 6.40 \\ 9.35 & 12.84 & 10.91 \\ 6.40 & 10.91 & 17.25 \end{bmatrix}$	$\begin{bmatrix} 12.76 & 9.41 & 6.48 \\ 9.41 & 12.95 & 11.03 \\ 6.48 & 11.03 & 17.45 \end{bmatrix}$	$\begin{bmatrix} 12.85 & 9.36 & 6.25 \\ 9.36 & 12.89 & 10.88 \\ 6.25 & 10.88 & 17.43 \end{bmatrix}$	$\begin{bmatrix} 12.865 & 9.339 & 6.166 \\ 9.339 & 12.908 & 10.826 \\ 6.166 & 10.826 & 17.327 \end{bmatrix}$
$k = 4$				
$\widehat{\text{means}}$	$(7.74, 9.85, 12.13)$	$(7.73, 9.84, 12.13)$	$(7.72, 9.84, 12.13)$	$(7.721, 9.842, 12.132)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 12.90 & 9.32 & 6.30 \\ 9.32 & 12.97 & 10.76 \\ 6.30 & 10.76 & 17.26 \end{bmatrix}$	$\begin{bmatrix} 12.98 & 9.37 & 6.37 \\ 9.37 & 13.05 & 10.84 \\ 6.37 & 10.84 & 17.37 \end{bmatrix}$	$\begin{bmatrix} 13.05 & 9.33 & 6.16 \\ 9.33 & 12.98 & 10.70 \\ 6.16 & 10.70 & 17.33 \end{bmatrix}$	$\begin{bmatrix} 13.057 & 9.355 & 6.192 \\ 9.355 & 13.105 & 10.873 \\ 6.192 & 10.873 & 17.635 \end{bmatrix}$
$k = 5$				
$\widehat{\text{means}}$	$(7.75, 9.85, 12.11)$	$(7.74, 9.84, 12.12)$	$(7.73, 9.84, 12.12)$	$(7.727, 9.841, 12.121)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 13.03 & 9.30 & 6.29 \\ 9.30 & 13.11 & 10.75 \\ 6.29 & 10.75 & 17.43 \end{bmatrix}$	$\begin{bmatrix} 13.12 & 9.35 & 6.36 \\ 9.35 & 13.19 & 10.82 \\ 6.36 & 10.82 & 17.54 \end{bmatrix}$	$\begin{bmatrix} 13.19 & 9.31 & 6.15 \\ 9.31 & 13.12 & 10.66 \\ 6.15 & 10.66 & 17.47 \end{bmatrix}$	$\begin{bmatrix} 13.206 & 9.343 & 6.185 \\ 9.343 & 13.244 & 10.854 \\ 6.185 & 10.854 & 17.797 \end{bmatrix}$
$k = 10$				
$\widehat{\text{means}}$	$(7.76, 9.84, 12.08)$	$(7.76, 9.84, 12.08)$	$(7.75, 9.84, 12.09)$	$(7.747, 9.838, 12.090)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 13.55 & 9.21 & 6.25 \\ 9.21 & 13.61 & 10.67 \\ 6.25 & 10.67 & 18.09 \end{bmatrix}$	$\begin{bmatrix} 13.63 & 9.26 & 6.32 \\ 9.26 & 13.69 & 10.74 \\ 6.32 & 10.74 & 18.20 \end{bmatrix}$	$\begin{bmatrix} 13.72 & 9.22 & 6.09 \\ 9.22 & 13.61 & 10.56 \\ 6.09 & 10.56 & 18.11 \end{bmatrix}$	$\begin{bmatrix} 13.790 & 9.270 & 6.149 \\ 9.270 & 13.730 & 10.736 \\ 6.149 & 10.736 & 18.382 \end{bmatrix}$

NOTE: The entries for $n < \infty$ are the averages and sample covariances of estimated quartiles. Results are based on 100,000 simulated samples. Standard errors of these entries are ≤ 0.003.

Table 4.6: Estimated means and covariance-variance ($\times n$) matrices of smoothed quartile estimators $\widehat{Q}_Y^{(k)}(0.25), \widehat{Q}_Y^{(k)}(0.50), \widehat{Q}_Y^{(k)}(0.75)$ for $k = 1:5, 10$ and selected ZIP models.

	$n = 50$	$n = 100$	$n = 500$	$n = \infty$
ZIP$(c = 0.2,\ \lambda = 10)$, $k = 1$				
$\widehat{\text{means}}$	$(5.09, 8.42, 10.75)$	$(4.84, 8.44, 10.76)$	$(4.72, 8.43, 10.74)$	$(4.979, 8.809, 11.482)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 76.25 & 31.38 & 19.91 \\ 31.38 & 28.88 & 21.60 \\ 19.91 & 21.60 & 25.05 \end{bmatrix}$	$\begin{bmatrix} 113.31 & 34.28 & 23.54 \\ 34.28 & 25.72 & 20.78 \\ 23.54 & 20.78 & 25.34 \end{bmatrix}$	$\begin{bmatrix} 206.33 & 51.17 & 40.78 \\ 51.17 & 30.23 & 28.79 \\ 40.78 & 28.79 & 38.16 \end{bmatrix}$	$\begin{bmatrix} 13.670 & 4.739 & 1.812 \\ 4.739 & 8.974 & 4.950 \\ 1.812 & 4.950 & 7.057 \end{bmatrix}$
ZIP$(c = 0.2,\ \lambda = 10)$, $k = 2$				
$\widehat{\text{means}}$	$(4.12, 8.71, 11.47)$	$(4.08, 8.73, 11.48)$	$(4.04, 8.75, 11.48)$	$(4.072, 8.780, 11.525)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 113.10 & 40.74 & 20.07 \\ 40.74 & 26.40 & 17.30 \\ 20.07 & 17.30 & 21.78 \end{bmatrix}$	$\begin{bmatrix} 123.96 & 40.94 & 20.92 \\ 40.94 & 24.96 & 16.95 \\ 20.92 & 16.95 & 21.76 \end{bmatrix}$	$\begin{bmatrix} 136.30 & 41.72 & 21.96 \\ 41.72 & 24.20 & 16.92 \\ 21.96 & 16.92 & 22.06 \end{bmatrix}$	$\begin{bmatrix} 137.548 & 40.395 & 20.548 \\ 40.395 & 22.840 & 15.500 \\ 20.548 & 15.500 & 20.416 \end{bmatrix}$
ZIP$(c = 0.2,\ \lambda = 10)$, $k = 3$				
$\widehat{\text{means}}$	$(4.18, 8.75, 11.49)$	$(4.16, 8.78, 11.50)$	$(4.14, 8.79, 11.50)$	$(4.139, 8.798, 11.505)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 124.70 & 40.23 & 20.52 \\ 40.23 & 25.05 & 16.33 \\ 20.52 & 16.33 & 21.29 \end{bmatrix}$	$\begin{bmatrix} 139.55 & 40.71 & 21.56 \\ 40.71 & 23.72 & 15.95 \\ 21.56 & 15.95 & 21.21 \end{bmatrix}$	$\begin{bmatrix} 156.50 & 41.76 & 22.65 \\ 41.76 & 22.99 & 15.77 \\ 22.65 & 15.77 & 21.30 \end{bmatrix}$	$\begin{bmatrix} 161.278 & 41.856 & 22.349 \\ 41.856 & 22.741 & 15.641 \\ 22.349 & 15.641 & 21.402 \end{bmatrix}$
ZIP$(c = 0.2,\ \lambda = 10)$, $k = 4$				
$\widehat{\text{means}}$	$(4.21, 8.77, 11.47)$	$(4.20, 8.79, 11.48)$	$(4.20, 8.80, 11.49)$	$(4.203, 8.808, 11.490)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 135.39 & 40.40 & 21.17 \\ 40.40 & 24.60 & 16.09 \\ 21.17 & 16.09 & 21.59 \end{bmatrix}$	$\begin{bmatrix} 153.43 & 41.18 & 22.41 \\ 41.18 & 23.39 & 15.73 \\ 22.41 & 15.73 & 21.51 \end{bmatrix}$	$\begin{bmatrix} 174.60 & 42.69 & 23.75 \\ 42.69 & 22.80 & 15.58 \\ 23.75 & 15.58 & 21.60 \end{bmatrix}$	$\begin{bmatrix} 181.814 & 43.007 & 23.459 \\ 43.007 & 22.630 & 15.477 \\ 23.459 & 15.477 & 21.699 \end{bmatrix}$
ZIP$(c = 0.2,\ \lambda = 10)$, $k = 5$				
$\widehat{\text{means}}$	$(4.24, 8.78, 11.46)$	$(4.24, 8.80, 11.47)$	$(4.25, 8.81, 11.48)$	$(4.262, 8.815, 11.478)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 144.49 & 40.63 & 21.69 \\ 40.63 & 24.35 & 15.92 \\ 21.69 & 15.92 & 21.85 \end{bmatrix}$	$\begin{bmatrix} 165.44 & 41.71 & 23.09 \\ 41.71 & 23.25 & 15.58 \\ 23.09 & 15.58 & 21.76 \end{bmatrix}$	$\begin{bmatrix} 190.44 & 43.63 & 24.65 \\ 43.63 & 22.76 & 15.45 \\ 24.65 & 15.45 & 21.86 \end{bmatrix}$	$\begin{bmatrix} 199.881 & 44.135 & 24.388 \\ 44.135 & 22.638 & 15.359 \\ 24.388 & 15.359 & 21.943 \end{bmatrix}$
ZIP$(c = 0.2,\ \lambda = 10)$, $k = 10$				
$\widehat{\text{means}}$	$(4.33, 8.80, 11.44)$	$(4.38, 8.81, 11.44)$	$(4.46, 8.82, 11.45)$	$(4.487, 8.828, 11.451)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 175.60 & 41.87 & 23.25 \\ 41.87 & 24.12 & 15.48 \\ 23.25 & 15.48 & 22.72 \end{bmatrix}$	$\begin{bmatrix} 207.58 & 44.09 & 25.22 \\ 44.09 & 23.35 & 15.22 \\ 25.22 & 15.22 & 22.62 \end{bmatrix}$	$\begin{bmatrix} 246.07 & 47.25 & 27.44 \\ 47.25 & 23.04 & 15.15 \\ 27.44 & 15.15 & 22.78 \end{bmatrix}$	$\begin{bmatrix} 260.238 & 48.091 & 27.266 \\ 48.091 & 22.937 & 15.069 \\ 27.266 & 15.069 & 22.813 \end{bmatrix}$

NOTE: The entries for $n < \infty$ are the averages and sample covariances of estimated quartiles. Results are based on 100,000 simulated samples. Standard errors of these entries are ≤ 0.003.

Table 4.7: Estimated means and covariance-variance ($\times n$) matrices of smoothed quartile estimators $\widehat{Q}_Y^{(k)}(0.25), \widehat{Q}_Y^{(k)}(0.50), \widehat{Q}_Y^{(k)}(0.75)$ for $k = 1:5, 10$ and selected ZIP models.

	$n = 50$	$n = 100$	$n = 500$	$n = \infty$
ZIP$(c = 0.8,\ \lambda = 10)$, $k = 1$				
$\widehat{\text{means}}$	$(0.00, 0.01, 0.35)$	$(0.00, 0.00, 0.26)$	$(0.00, 0.00, 0.18)$	$(0.000, 0.005, 2.092)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.03 & 0.50 \\ 0.00 & 0.50 & 14.94 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.16 \\ 0.00 & 0.16 & 9.86 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.03 \\ 0.00 & 0.03 & 5.74 \end{bmatrix}$	$\begin{bmatrix} 0.000 & 0.000 & 0.000 \\ 0.000 & 0.006 & 0.612 \\ 0.000 & 0.612 & 59.62 \end{bmatrix}$
ZIP$(c = 0.8,\ \lambda = 10)$, $k = 2$				
$\widehat{\text{means}}$	$(0.00, 0.03, 1.65)$	$(0.00, 0.03, 1.65)$	$(0.00, 0.01, 1.46)$	$(0.000, 0.022, 2.642)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.00 & 0.00 & 0.02 \\ 0.00 & 0.56 & 6.79 \\ 0.02 & 6.79 & 131.84 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.01 \\ 0.00 & 0.19 & 4.31 \\ 0.01 & 4.31 & 133.53 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.05 & 2.38 \\ 0.00 & 2.38 & 135.41 \end{bmatrix}$	$\begin{bmatrix} 0.000 & 0.000 & 0.001 \\ 0.000 & 0.078 & 3.035 \\ 0.001 & 3.035 & 120.506 \end{bmatrix}$
ZIP$(c = 0.8,\ \lambda = 10)$, $k = 3$				
$\widehat{\text{means}}$	$(0.00, 0.045, 2.69)$	$(0.00, 0.03, 2.54)$	$(0.00, 0.02, 2.41)$	$(0.000, 0.017, 2.593)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.39 & 5.59 \\ 0.00 & 5.59 & 147.78 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.14 & 3.96 \\ 0.00 & 3.96 & 159.42 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.05 & 2.69 \\ 0.00 & 2.69 & 169.49 \end{bmatrix}$	$\begin{bmatrix} 0.000 & 0.000 & 0.001 \\ 0.000 & 0.058 & 3.023 \\ 0.001 & 3.023 & 160.369 \end{bmatrix}$
ZIP$(c = 0.8,\ \lambda = 10)$, $k = 4$				
$\widehat{\text{means}}$	$(0.00, 0.02, 2.63)$	$(0.00, 0.01, 2.49)$	$(0.00, 0.01, 2.36)$	$(0.000, 0.007, 2.362)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.20 & 3.52 \\ 0.00 & 3.52 & 140.96 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.05 & 2.10 \\ 0.00 & 2.10 & 148.96 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.01 & 1.19 \\ 0.00 & 1.19 & 154.01 \end{bmatrix}$	$\begin{bmatrix} 0.000 & 0.000 & 0.000 \\ 0.000 & 0.011 & 1.186 \\ 0.000 & 1.186 & 154.009 \end{bmatrix}$
ZIP$(c = 0.8,\ \lambda = 10)$, $k = 5$				
$\widehat{\text{means}}$	$(0.00, 0.01, 2.47)$	$(0.00, 0.01, 2.31)$	$(0.00, 0.00, 2.15)$	$(0.000, 0.002, 2.094)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.11 & 2.41 \\ 0.00 & 2.41 & 149.12 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.02 & 1.21 \\ 0.00 & 1.21 & 156.85 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.54 \\ 0.00 & 0.54 & 160.20 \end{bmatrix}$	$\begin{bmatrix} 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.567 \\ 0.000 & 0.567 & 188.483 \end{bmatrix}$
ZIP$(c = 0.8,\ \lambda = 10)$, $k = 10$				
$\widehat{\text{means}}$	$(0.00, 0.00, 2.00)$	$(0.00, 0.00, 1.75)$	$(0.00, 0.00, 1.48)$	$(0.000, 0.000, 1.399)$
$\widehat{\mathbf{HDH}'}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.02 & 0.54 \\ 0.00 & 0.54 & 179.04 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.11 \\ 0.00 & 0.11 & 186.96 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.01 \\ 0.00 & 0.01 & 180.08 \end{bmatrix}$	$\begin{bmatrix} 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.006 \\ 0.000 & 0.006 & 198.653 \end{bmatrix}$

NOTE: The entries for $n < \infty$ are the averages and sample covariances of estimated quartiles. Results are based on 100,000 simulated samples. Standard errors of these entries are ≤ 0.003.

Chapter 5

Risk Measurement

In this chapter, we use the newly developed methodology to evaluate the riskiness of the automobile accident data, which is taken from Klugman *et al.* (2012, Table 6.2). In Section 5.1, we introduce a class of commonly used risk measures – distortion risk measures – and provide three examples. Then, in Section 5.2, we estimate the smoothed quantile function for the given data set and evaluate a few selected risk measures.

5.1 Risk Measures

For the purposes of risk estimation, it is worth noting that many risk measures used in the current practice can be defined as the expectation of loss with respect to distorted probabilities. Specifically, for a random variable $X \geq 0$ with CDF F, a risk measure R is defined as

$$R[F] \;=\; \int_0^\infty g(1 - F(x))\, dx, \tag{5.1}$$

where the *distortion function* $g(\cdot)$ is an increasing function with $g(0) = 0$ and $g(1) = 1$. In addition, if g is differentiable, then integration by parts in (5.1) leads to

$$R[F] \;=\; \int_0^1 F^{-1}(u)\psi(u)\, du, \tag{5.2}$$

where $\psi(u) = g'(1-u)$ and F^{-1} is the quantile function of variable X. Now if in (5.2) we replace F^{-1} with its (empirical or parametric) estimator, then we will have an estimator of $R[F]$. For instance, the empirical estimator $R[\widehat{F}_n]$ is derived by replacing F with its empirical counterpart $\widehat{F}_n$. This estimator belongs to the class of L-statistics, asymptotic properties of which are well established. The following are examples of commonly used *distortion risk measures*.

Example 5.1 [VaR, value-at-risk]

The VaR measure on a portfolio of risks (i.e., potential losses) is the maximum loss one might expect over a given period of time, at a given level of confidence (say, β). In mathematical terms, this measure is defined as the $(1-\beta)$-level quantile of the distribution function F:

$$\text{VaR}_\beta[F] \;=\; F^{-1}(1-\beta). \tag{5.3}$$

VaR can be expressed as (5.1) by choosing $g(u) = 0$ for $0 \le u < \beta$, and $= 1$ for $\beta \le u \le 1$. This risk measure, however, has some axiomatic flaws – it is not *coherent* as it does not satisfy the sub-additivity property (see Artzner *et al.*, 1999). □

Example 5.2 [CTE, conditional tail expectation]

The CTE measure (also known as Tail-VaR or expected shortfall) is the conditional expectation of a loss variable given that it exceeds a specified quantile, VaR_β. It measures the expected maximum loss in the $100\beta\%$ worst cases, over a given period of time:

$$\text{CTE}_\beta[F] \;=\; \frac{1}{\beta}\int_{1-\beta}^{1} F^{-1}(u)\,du. \tag{5.4}$$

CTE is a coherent risk measure, and it can be expressed as (5.2) by choosing $\psi(u) = 0$ for $0 \le u \le 1-\beta$, and $= 1/\beta$ for $1-\beta < u \le 1$. □

Example 5.3 [PHT, proportional hazards transform]

The name of the PHT measure is motivated by the fact that the hazard function of the distorted distribution is proportional to the hazard function of F. The measure is defined as

$$\text{PHT}_r[F] \;=\; r\int_0^1 F^{-1}(u)(1-u)^{r-1}\,du, \tag{5.5}$$

where constant r $(0 < r \leq 1)$ represents the degree of distortion. Note that small r corresponds to high distortion, and $\text{PHT}_r[F]$ for $r = 1$ is the expected value of X. PHT is a coherent risk measure, and it can be expressed as (5.2) by choosing $\psi(u) = r(1-u)^{r-1}$. □

5.2 Numerical Example

The automobile accident data represent the risk profile of 9,461 insurance policies; the data set is provided in Table 5.1. As can be seen from the table, more than 80% of policies reported no accident, which is very typical for an insurance portfolio, and less than 20% reported at least one claim. As one would expect, when the number of accidents increases, the number of policies decreases. Moreover, only one policy reported 7 accidents and none had 8 or more.

Table 5.1: *Automobile Data*: The number of accidents under the policy.

Number of accidents	0	1	2	3	4	5	6	7	≥ 8	Total
Number of policies	7,840	1,317	239	42	14	4	4	1	0	9,461

Since the data set has eight distinct data points, we can estimate its smoothed quantile function using (3.4) or (3.5), with $d = 8$ and $y_{1:d} = 0, y_{2:d} = 1, \ldots, y_{d:d} = 7$. The smoothed quantile function $\widehat{Q}(u)$ and the classical discrete quantile function $\widehat{F}_n^{-1}$ are depicted in Figure 5.1. The additional curves plotted there represent the product $\widehat{Q}(u)\psi(u)$, with $\psi(u)$ taken from Examples 5.2-5.3.

The smoothed quantile function is continuous and thus the product $\widehat{Q}(u)\psi(u)$ is well-

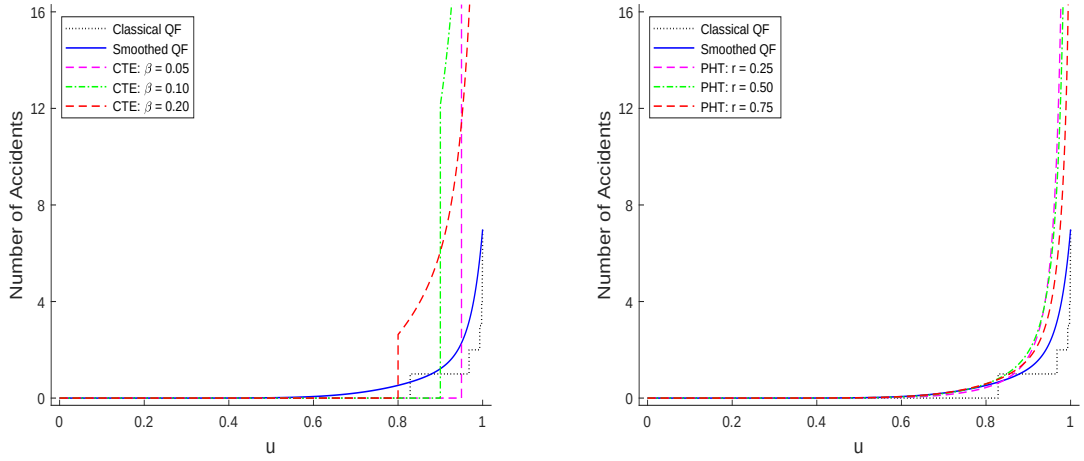

Figure 5.1: The classical and smoothed quantile functions for Automobile Data. The other curves represent the smoothed quantile function multiplied by the risk measure weights $\psi(u)$. *Left panel*: $\text{CTE}_\beta[\widehat{F}]$ with $\beta = 0.05, 0.10, 0.20$. *Right panel*: $\text{PHT}_r[\widehat{F}]$ with $r = 0.25, 0.50, 0.75$.

defined and can be integrated using basic numerical procedures (e.g., trapezoidal rule). The risk measure value is simply the area under the corresponding curve, either (5.4) or (5.5). In Table 5.2, we report several risk measure estimates for Automobile Data.

Table 5.2: Selected risk measure estimates for Automobile Data.

$\widehat{\text{VaR}}_\beta[\widehat{F}]$			$\widehat{\text{CTE}}_\beta[\widehat{F}]$			$\widehat{\text{PHT}}_r[\widehat{F}]$		
$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.20$	$\beta = 0.05$	$\beta = 0.10$	$\beta = 0.20$	$r = 0.25$	$r = 0.50$	$r = 0.75$
2.286	1.216	0.527	3.997	2.822	1.818	2.401	1.361	0.731

This risk measurement exercise illustrates and quantifes the obvious: the deeper one goes into the distribution tail (i.e., when β gets smaller for VaR_β and CTE_β or r gets smaller for PHT_r), the riskier it gets. In practice, the choice of the risk measure and associated tail parameters (β and r) would be determined from the risk appetite statement of the company. (Under the current insurance industry regulations, company's risk appetite has to be specified by the top company executives and approved by the board of directors.) To help with that determination, one should assess uncertainty of the risk measure estimates. Using the available results established in this dissertation (Theorem

3.1), we can specify a joint asymptotically normal distribution of the VaR_β estimates:

$$\left(\widehat{\text{VaR}}_{0.05}[\widehat{F}],\, \widehat{\text{VaR}}_{0.10}[\widehat{F}],\, \widehat{\text{VaR}}_{0.20}[\widehat{F}]\right) \;\sim\; \mathcal{AN}\left(\left(2.286,\, 1.216,\, 0.527\right),\, \frac{1}{9{,}461}\,\widehat{\mathbf{HDH}}'\right),$$

where

$$\widehat{\mathbf{HDH}}' \;=\; \begin{bmatrix} 51.783 & 16.232 & 3.459 \\ 16.232 & 7.276 & 2.684 \\ 3.459 & 2.684 & 1.960 \end{bmatrix}.$$

Now it follows that (approximate) 95% confidence intervals for the three VaR measures are:

$$\begin{aligned}
\text{VaR}_{0.05}[F]: &\quad 2.286 \pm 1.96\sqrt{\frac{51.783}{9461}} \;=\; 2.286 \pm 0.145; \quad \text{or} \;\; [2.141;\, 2.431]. \\
\text{VaR}_{0.10}[F]: &\quad 1.216 \pm 1.96\sqrt{\frac{7.276}{9461}} \;=\; 1.216 \pm 0.054; \quad \text{or} \;\; [1.162;\, 1.270]. \\
\text{VaR}_{0.20}[F]: &\quad 0.527 \pm 1.96\sqrt{\frac{1.960}{9461}} \;=\; 0.527 \pm 0.028; \quad \text{or} \;\; [0.499;\, 0.555].
\end{aligned}$$

Finally, note that evaluation of CTE_β, $\beta = 1$, or PHT_r, $r = 1$, yields the mean of *smoothed* data. It is equal to 0.415 and is almost twice as large as the "regular" mean (computed directly from the data in Table 5.1), which is equal 0.214. This discrepancy could be anticipated from Figure 5.1, because the smoothed quantile function is almost uniformly above the classical quantile function.

Chapter 6

Final Remarks

6.1 Summary

In this dissertation, we have studied smoothing of quantiles for discrete distributions, with both finite and countable supports. Properties of sample quantile estimators were investigated theoretically as well as via simulations. Definitions, properties and illustrations of smoothed quantile functions were provided. These functions have also been applied to risk measurement exercises.

The first main contribution of the dissertation is introduction and development of the vectors of smoothed quantile estimators. We established the asymptotic properties for the vector of smoothed quantile estimators and investigated their small-sample properties using Monte Carlo simulations. The simulation study revealed convergence of sample estimates to the true quantities as the sample size increased.

The second main contribution is extension of smoothed quantiles for discrete distributions with infinite support. The properties and estimators of smoothed quantile functions established by Wang and Hutson (2011) are valid for discrete distributions with finite support. In this dissertation we extensively studied the choice of the count of distinct points d, the minimum and maximum distinct points, $y_{1:d}$ and $y_{d:d}$ when discrete distributions have

infinite (countable) support. We proposed the use of sample and population smoothed quantile estimators using truncated data, based on d, $y_{1:d}$ and $y_{d:d}$, which were calculated using Chebychev's inequality bounds. We conjectured the asymptotic properties of the new estimators and used simulations to check the conjectured claims.

Finally, we used the newly developed methodology for smoothed quantiles to evaluate the riskiness of the automobile accident data, and reported several point estimates of VaR, CTE, and PHT measures. For VaR measures, the 95% confidence intervals were also constructed.

6.2 Future Work

Most immediate future research stemming from this dissertation will pursue two problems: (i) proof of Conjecture 4.1; and (ii) development of percentile-matching estimators for discrete data.

6.2.1 Proof of Conjecture 4.1

While our simulation study of Section 4.2.3 provides guidance on what can be expected from the new smoothed quantile estimators, a rigorous theoretical study is needed to establish asymptotic properties of those estimators. The key quantities used in contruction of estimators (4.7) are sample mean, $\overline{Y}$, and variance, S^2. Their consistency and joint asymptotic normality are well known and would carry through to their transformations such as $\widehat{L}_k = \overline{Y} - k\sqrt{S^2}$ and $\widehat{U}_k = \overline{Y} + k\sqrt{S^2}$. The main challenge in proving the conjecture is that estimators $\widehat{d}_k$, $\widehat{y}_{1:\widehat{d}_k}$ and $\widehat{y}_{\widehat{d}_k:\widehat{d}_k}$ require taking the greatest integer part of $\widehat{L}_k$ and $\widehat{U}_k$, which results in discontinuities when the respective limits of $\widehat{L}_k$ and $\widehat{U}_k$, $L_k = \mathbf{E}[Y] - k\sqrt{\mathbf{Var}[Y]}$ and $U_k = \mathbf{E}[Y] + k\sqrt{\mathbf{Var}[Y]}$, are integers. This topic and related theoretical investigations will be pursued in future studies.

6.2.2 Percentile Matching

In this research line, we will use the smoothed quantile definitions for population and sample and introduce percentile-matching (PM) estimators for estimating parameters of discrete distributions. The idea behind this method of estimation is identical to that of the method of moments – create a system of equations by matching *smoothed* sample and model percentiles (instead of moments) and then solve it with respect to unknown parameters. The same approach has also been employed in designing the method of trimmed moments, MTM (Brazauskas et al., 2009, and Brazauskas, 2009) and method of winsorized moments, MWM (Zhao *et al.*, 2018a,b) estimators for continous distributions. Given the effectiveness of MTMs and MWMs, we anticipate that PM estimators will be more robust against model misspecification than the widely used maximum likelihood estimators. For more details on PM estimation, see Klugman *et al.* (2012, Section 13.1).

Bibliography

[1] Acerbi, C. and Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance*, **26**(7), 1487–1503.

[2] Albrecht, P. (2004). Risk measures. In *Encyclopedia of Actuarial Science* (B. Sundt and J. Teugels, eds.), volume **3**, 1493–1501; Wiley, London.

[3] Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, **9**(3), 203–228.

[4] Besson, J.-L. and Partrat, Ch. (1992). Trend et systèmes de bonus-malus. *ASTIN Bulletin*, **22**(1), 11–31.

[5] Boucher, J.-P., Denuit, M., and Guillén, M. (2007). Risk classification for claim counts: a comparative analysis of various zero-inflated mixed Poisson and hurdle models. *North American Actuarial Journal*, **11**(4), 110–131.

[6] Brazauskas, V. (2009). Robust and efficient fitting of loss models: diagnostic tools and insights. *North American Actuarial Journal*, **13**(3), 1–14.

[7] Brazauskas, V., Jones, B., Puri, M., and Zitikis, R. (2007). Nested L-statistics and their use in comparing the riskiness of portfolios. *Scandinavian Actuarial Journal*, **107**(3), 162–179.

[8] Brazauskas, V., Jones, B., Puri, M., and Zitikis, R. (2008). Estimating conditional tail expectations with actuarial applications in view. *Journal of Statistical Planning and Inference*, **138**(11), 3590–3604.

[9] Brazauskas, V., Jones, B., and Zitikis, R. (2009). Robust fitting of claim severity distributions and the method of trimmed moments. *Journal of Statistical Planning and Inference*, **139**(6), 2028–2043.

[10] Brazauskas, V. and Kaiser, T. (2004). Discussion of "Empirical estimation of risk measures and related quantities" by Jones and Zitikis. *North American Actuarial Journal*, **8**(3), 114–117.

[11] Consul, P.C. (1989). *Generalized Poisson Distributions*. Dekker, New York.

[12] Consul, P. and Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics: Theory and Methods*, **21**(1), 89–109.

[13] Denuit, M. (1997). A new distribution of Poisson-type for the number of claims. *ASTIN Bulletin*, **27**(2), 229–242.

[14] Denuit, M., Maréchal, X., Pitrebois, S., and Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Wiley, Chichester.

[15] Furman, E., Wang, R., and Zitikis, R. (2017). Gini-type measures of risk and variability: Gini shortfall, capital allocations, and heavy-tailed risks. *Journal of Banking and Finance*, **83**, 70–84.

[16] Genton, M.G., Ma, Y., and Parzen, E. (2006). Discussion of "Sur une limitation très générale de la dispersion de la médiane" by M. Fréchet, 1940. *Journal de la Société Française de Statistique*, **147**(2), 51–60.

[17] Göb, R. (2011). Estimating value-at-risk and conditional value-at-risk for count variables. *Quality and Reliability Engineering International*, **27**, 659–672.

[18] Gossiaux, A. and Lemaire, J. (1981). Méthodes d'ajustement de distributions de sinistres. *Bulletin of the Swiss Association of Actuaries*, **81**, 87–95.

[19] Harrell, F.E. and Davis, C.E. (1982). A new distribution-free quantile estimator. *Biometrika*, **69**(3), 635–640.

[20] Islam, MN and Consul, PC (1992). A probabilistic model for automobile claims. *Bulletin of the Swiss Association of Actuaries*, 85–93.

[21] Jones, B.L. and Zitikis, R. (2003). Empirical estimation of risk measures and related quantities. *North American Actuarial Journal*, **7**(4), 44–54.

[22] Jones, B.L. and Zitikis, R. (2005). Testing for the order of risk measures: an application of L-statistics in actuarial science. *Metron*, **63**(2), 193–211.

[23] Jones, B.L. and Zitikis, R. (2007). Risk measures, distortion parameters, and their empirical estimation. *Insurance: Mathematics and Economics*, **41**(2), 279–297.

[24] Kabán, A. (2012). Non-parametric detection of meaningless distances in high dimensional data. *Statistics and Computing*, **22**(2), 375–385.

[25] Kaiser, T. and Brazauskas, V. (2006). Interval estimation of actuarial risk measures. *North American Actuarial Journal*, **10**(4), 249–268.

[26] Klugman, S.A., Panjer, H.H., and Willmot, G.E. (2012). *Loss Models: From Data to Decisions*, 4th edition. Wiley, New York.

[27] Ma, Y., Genton, M.G., and Parzen, E. (2011). Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics*, **63**(2), 227–243.

[28] Parzen, E. (1992). Unification of statistical methods for continuous and discrete data. In *Computing Science and Statistics*, 235–242. Springer.

[29] Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science*, **19**(4), 652–662.

[30] Samanthi, R., Wei, W., and Brazauskas, V. (2017). Comparing the riskiness of dependent portfolios via nested L-statistics. *Annals of Actuarial Science*, **11**(2), 237–252.

[31] Saw, J.G., Yang, M.C.K., and Mo, T.C. (1984). Chebyshev inequality with estimated mean and variance. *The American Statistician*, **38**(2), 130–132.

[32] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

[33] Stigler, S.M. (1977). Fractional order statistics, with applications. *Journal of the American Statistical Association*, **72**(359), 544–550.

[34] Tapiero, C.S. (2004). Risk management: An interdisciplinary framework. In *Encyclopedia of Actuarial Science* (B. Sundt and J. Teugels, eds.), volume **3**, 1483–1493; Wiley, London.

[35] Tremblay, L. (1992). Using the Poisson inverse gaussian in bonus-malus systems. *ASTIN Bulletin*, **22**(1), 97–106.

[36] Wang, D. and Hutson, A.D. (2011). A fractional order statistic towards defining a smooth quantile function for discrete data. *Journal of Statistical Planning and Inference*, **141**(9), 3142–3150.

[37] Willmot, G.(1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal*, **1987**(3-4), 113–127.

[38] Yip, K.C. and Yau, K.K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, **36**(2), 153–163.

[39] Zhao, Q., Brazauskas, V., and Ghorai, J. (2018a). Robust and efficient fitting of severity models and the method of Winsorized moments. *ASTIN Bulletin*, **48**(1), 275–309.

[40] Zhao, Q., Brazauskas, V., and Ghorai, J. (2018b). Small-sample performance of the MTM and MWM estimators for the parameters of log-location-scale families. *Journal of Statistical Computation and Simulation*, **88**(4), 808–824.

Appendix A

R code: Asymptotic and Estimated Means and Covariance-Variance Matrices for Binomial Distribution

```
#Q_Binomial.R file
#Generates numbers for tables 3.2 and 3.3 for Binomial Distribution
rm(list=ls())
#########Binomial parameters###################################
q=0.7
n=4
#####################initialize variables####################
v=c(0:n)
d=length(v)
#d=n+1
# v.minus1 and v minus d
v.minus_1=v[2:length(v)]
v.minus_d=v[1:length(v)-1]

seed_set=1;
sample_size=500;n_sim=100000
u=c(1/4,1/2,3/4)

Q=vector(length=length(u))
Q_cap=matrix(nrow=n_sim,ncol=length(u))
mean_Q_cap=vector(length=length(u))

tau=matrix(nrow=length(u),ncol=d-1)
tau_theo=matrix(nrow=length(u),ncol=d-1)
Q_var=matrix(nrow=length(u),ncol=length(u))
Q_var_theo=matrix(nrow=length(u),ncol=length(u))
mean_Q_var_sim=matrix(0,nrow=length(u),ncol=length(u))
mean_Q_cap=vector(length=length(u))

#theory:calculate Q, HDH' and correlations #######

#F cumultiv distribution function for binomial
F=as.vector(pbinom((0:n),n,q))
F_minus_d=F[1:length(F)-1]

#D_theo is D in HDH'for theory
```

```
D_theo=matrix((1-F_minus_d),nrow=length(F_minus_d),
              ncol=length(F_minus_d),
              byrow=TRUE)
D_theo=F_minus_d*D_theo
D_theo[lower.tri(D_theo)] <- t(D_theo)[lower.tri(D_theo)]

for (ui in seq_along(u)){
  a=(d+1)*u[ui]
  b=(d+1)*(1-u[ui])
  beta_dist_th=pbeta(F,a,b)
  limit=beta_dist_th[1]
  for (i in 2:length(beta_dist_th))
    {limit[i]=beta_dist_th[i]-beta_dist_th[i-1]}
  #calculate smooth quantile
  Q[ui]=round(sum(limit*v),5)
  #theory variance:tau_theo is same as H in HDH'
  beta_density=dbeta(F_minus_d,(d+1)*u[ui], (d+1)*(1-u[ui]))
  tau_theo[ui,]=beta_density*(v.minus_d-v.minus_1)
}
HDH_theo=tau_theo%*%D_theo%*%t(tau_theo)
#correlation matrix
Corr1=cov2cor(HDH_theo)

###############simulation loop starts#################
###binomial simulated data is based on the
###parameters n, q  set initially.
###As d=n+1, d is same for theory and simulation
for(j in 1:n_sim){
  seed_set=seed_set+1
  set.seed(seed_set)
  ###generate binomial data
  seq1=sort(rbinom(sample_size,n,q))
  fn=ecdf(seq1)
  fn_seq1=fn(unique(seq1))
  uniq_seq1<-unique(seq1)
  fn_v<-rep(0,d)

  ###find Fn for missing v in simulation
   for (i in 1:d){
    for (m in seq_along(uniq_seq1)){
      if(v[i]==uniq_seq1[m]){fn_v[i]<-fn_seq1[m]}
    }
  }
  #extract non zero fn_v if the first few fn_v are zero
  fn_v_mod<-fn_v
  if (length(which(fn_v == 0))!=0){
    if (min(which(fn_v == 0))==1){
      nzero_loc<-min(which(fn_v != 0))
      fn_v_mod<-fn_v[nzero_loc:d]
    }
  }
```

```
  ## for missing v,replace its F  with F for v-1
  while(length(which(fn_v_mod == 0))!=0){
    loc <- which(fn_v_mod == 0)
    for(m in seq_along(loc))
    {loc_zero<-loc[m]
    fn_v_mod[loc_zero]<-fn_v_mod[loc_zero-1]}
  }
  if (length(which(fn_v == 0))!=0){
    if (min(which(fn_v == 0))==1){
      nzero_loc<-min(which(fn_v != 0))
      app_0<-rep(0,nzero_loc-1)
      fn_v_mod<-c( app_0,fn_v_mod)
    }
  }
  fn_v<-fn_v_mod
  v_minus_1=v[2:d]
  v_minus_d=v[1:d-1]
  fn_v_minus_1=fn_v[2:length(fn_v)]
  fn_v_minus_d=fn_v[1:length(fn_v)-1]

  #loop for different u
  for (ui in seq_along(u)){
    a=(d+1)*u[ui]
    b=(d+1)*(1-u[ui])
    ##calculate Qcap using simulated data
    beta_dist=pbeta(fn_v,a,b)
    lim=beta_dist[1]
    for (i in 2:length(beta_dist))
      {lim[i]=beta_dist[i]-beta_dist[i-1]}
    Q_cap[j,ui]=sum(lim*v)
  }
} #simulation loop ends
HDH_cap=cov(Q_cap)*sample_size
mean_Q_cap=colMeans(Q_cap)

###############output ################################
Q_result=rbind(Q,mean_Q_cap)
colnames(Q_result)<-u
Q_result
HDH_theo
HDH_cap
#########################################################
```

Appendix B

R code: Asymptotic and Estimated Means and Covariance-Variance Matrices for ZIB Distribution

```
#Q_ZIB.R file
#Generates numbers for tables 3.2 and 3.3 for ZIB Distribution
rm(list=ls())
library(VGAM)
#########ZIB parameters###################################
q=0.7
n=8
c=0.8
######################initialize variables################
v=c(0:n)
d=length(v)
#d=n+1
v.minus_d=v[2:length(v)]
v.minus_1=v[1:length(v)-1]
p0=pbinom(0,n,q)
pi=(c-p0)/(1-p0)
seed_set=1;
sample_size=100;n_sim=100000
#u=1/15;u=0.5
u=c(1/4,1/2,3/4)

Q=vector(length=length(u))
Q_cap=matrix(nrow=n_sim,ncol=length(u))
mean_Q_cap=vector(length=length(u))

tau=matrix(nrow=length(u),ncol=d-1)
tau_theo=matrix(nrow=length(u),ncol=d-1)
Q_var=matrix(nrow=length(u),ncol=length(u))
Q_var_theo=matrix(nrow=length(u),ncol=length(u))
mean_Q_var_sim=matrix(0,nrow=length(u),ncol=length(u))
mean_Q_cap=vector(length=length(u))


#theory:calculate Q, HDH' and correlations ######

#F cumultiv distribution function for ZIB
```

```
F=as.vector(pzibinom((0:n),n,q,pi))
F_minus_d=F[1:length(F)-1]

#D_theo is D in HDH'for theory
D_theo=matrix((1-F_minus_d),nrow=length(F_minus_d),
              ncol=length(F_minus_d),byrow=TRUE)
D_theo=F_minus_d*D_theo
D_theo[lower.tri(D_theo)] <- t(D_theo)[lower.tri(D_theo)]

for (ui in seq_along(u)){
  a=(d+1)*u[ui]
  b=(d+1)*(1-u[ui])
  beta_dist_th=pbeta(F,a,b)
  limit=beta_dist_th[1]
  #calculate smooth quantile
  for (i in 2:length(beta_dist_th))
   {limit[i]=beta_dist_th[i]-beta_dist_th[i-1]}
  Q[ui]=round(sum(limit*v),5)
  #theory variance:tau_theo is same as H in HDH'
  beta_density=dbeta(F_minus_d,(d+1)*u[ui], (d+1)*(1-u[ui]))
  tau_theo[ui,]=beta_density*(v.minus_d-v.minus_1)
  }

HDH_theo=tau_theo%*%D_theo%*%t(tau_theo)
#correlation matrix
Corr1=cov2cor(HDH_theo)

################simulation loop starts#################
###ZIB simulated data is based on the
###parameters n, q  set initially.
###As d=n+1, d is same for theory and simulation
for(j in 1:n_sim){
  seed_set=seed_set+1
  set.seed(seed_set)
  ###generate ZIB data
  seq1=sort(rzibinom(sample_size,n,q,pi))
  fn=ecdf(seq1)
  fn_seq1=fn(unique(seq1))
  uniq_seq1<-unique(seq1)
  fn_v<-rep(0,d)

  #find Fn for missing v in simulation
   for (i in 1:d){
    for (m in seq_along(uniq_seq1)){
      if(v[i]==uniq_seq1[m]){fn_v[i]<-fn_seq1[m]}
    }
  }
  #extract non zero fn_v if the first few fn_v are zero
  fn_v_mod<-fn_v
  if (length(which(fn_v == 0))!=0){
    if (min(which(fn_v == 0))==1){
```

```
      nzero_loc<-min(which(fn_v != 0))
      fn_v_mod<-fn_v[nzero_loc:d]
    }
  }
  ## for missing v,replace its F  with F for v-1
  while(length(which(fn_v_mod == 0))!=0){
    loc <- which(fn_v_mod == 0)
    for(m in seq_along(loc))
    {loc_zero<-loc[m]
    fn_v_mod[loc_zero]<-fn_v_mod[loc_zero-1]}
  }
  if (length(which(fn_v == 0))!=0){
    if (min(which(fn_v == 0))==1){
      nzero_loc<-min(which(fn_v != 0))
      app_0<-rep(0,nzero_loc-1)
      fn_v_mod<-c( app_0,fn_v_mod)
    }
  }
  fn_v<-fn_v_mod
  v_minus_1=v[2:d]
  v_minus_d=v[1:d-1]
  fn_v_minus_1=fn_v[2:length(fn_v)]
  fn_v_minus_d=fn_v[1:length(fn_v)-1]

  #loop for different u
  for (ui in seq_along(u)){
    a=(d+1)*u[ui]
    b=(d+1)*(1-u[ui])
    ##calculate Qcap using simulated data
    beta_dist=pbeta(fn_v,a,b)
    lim=beta_dist[1]
    for (i in 2:length(beta_dist))
      {lim[i]=beta_dist[i]-beta_dist[i-1]}
    Q_cap[j,ui]=sum(lim*v)
  }
}#simulation loop ends
HDH_cap=cov(Q_cap)*sample_size
mean_Q_cap=colMeans(Q_cap)

###############output ##############################
result=rbind(Q,mean_Q_cap)
colnames(result)<-u
result
HDH_theo
HDH_cap
#####################################################
```

Appendix C

R code: Asymptotic and Estimated Means and Covariance-Variance Matrices for Poisson Distribution

```
#Q_Poisson.R file
#Generates numbers for tables 4.1 and 4.2 for Poisson Distribution
rm(list=ls())
#########Poisson parameters######
lambda=1

######################initialize variables#############
max_v=19
v=c(0:max_v)
d=length(v)
# v.minus1 and v minus d
v.minus_1=v[2:length(v)]
v.minus_d=v[1:length(v)-1]

seed_set=1;
sample_size=100;n_sim=100000
u=c(1/4,1/2,3/4)

Q=vector(length=length(u))
Q_cap=matrix(nrow=n_sim,ncol=length(u))
mean_Q_cap=vector(length=length(u))

tau=matrix(nrow=length(u),ncol=d-1)
tau_theo=matrix(nrow=length(u),ncol=d-1)
Q_var=matrix(nrow=length(u),ncol=length(u))
Q_var_theo=matrix(nrow=length(u),ncol=length(u))
mean_Q_var_sim=matrix(0,nrow=length(u),ncol=length(u))
mean_Q_cap=vector(length=length(u))

#theory:calculate Q, and HDH'############

#F cumultiv distribution function for Poisson
F=as.vector(ppois(v,lambda))
F_minus_d=F[1:length(F)-1]

#D_theo is D in HDH'for theory
```

```
D_theo=matrix((1-F_minus_d),nrow=length(F_minus_d),
              ncol=length(F_minus_d),byrow=TRUE)
D_theo=F_minus_d*D_theo
D_theo[lower.tri(D_theo)] <- t(D_theo)[lower.tri(D_theo)]
for (ui in seq_along(u)){
  a=(d+1)*u[ui]
  b=(d+1)*(1-u[ui])
  beta_dist_th=pbeta(F,a,b)
  limit=beta_dist_th[1]
  #calculate smooth quantile
  for (i in 2:length(beta_dist_th))
    {limit[i]=beta_dist_th[i]-beta_dist_th[i-1]}
  Q[ui]=round(sum(limit*v),5)
  #theory variance:tau_theo is same as H in HDH'
  beta_density=dbeta(F_minus_d,(d+1)*u[ui], (d+1)*(1-u[ui]))
  tau_theo[ui,]=beta_density*(v.minus_d-v.minus_1)
}

HDH_theo=tau_theo%*%D_theo%*%t(tau_theo)
#correlation matrix
Corr1=cov2cor(HDH_theo)

################simulation loop starts###################
for(j in 1:n_sim){
  seed_set=seed_set+1
  set.seed(seed_set)
  ###generate poisson data
  seq1=sort(rpois(sample_size, lambda))
  fn=ecdf(seq1)
  fn_seq1=fn(unique(seq1))
  uniq_seq1<-unique(seq1)
  fn_v<-rep(0,d)
  ###find Fn for missing v in simulation
  for (i in 1:d){
    for (m in seq_along(uniq_seq1)){
      if(v[i]==uniq_seq1[m]){fn_v[i]<-fn_seq1[m]}
    }
  }
  fn_v_mod<-fn_v
  #extract non zero fn_v if the first few fn_v are zero
  if (min(which(fn_v == 0))==1){
    nzero_loc<-min(which(fn_v != 0))
    fn_v_mod<-fn_v[nzero_loc:d]
  }
  ## for missing v,replace its F with F for v-1
  while (length(which(fn_v_mod == 0))!=0){
    loc <- which(fn_v_mod == 0)
    for(m in seq_along(loc))
    {loc_zero<-loc[m]
     fn_v_mod[loc_zero]<-fn_v_mod[loc_zero-1]}
     }
```

```
  if (min(which(fn_v == 0))==1){
    nzero_loc<-min(which(fn_v != 0))
    app_0<-rep(0,nzero_loc-1)
    fn_v_mod<-c( app_0,fn_v_mod)
  }
  fn_v<-fn_v_mod
  v_minus_1=v[2:d]
  v_minus_d=v[1:d-1]
  fn_v_minus_1=fn_v[2:length(fn_v)]
  fn_v_minus_d=fn_v[1:length(fn_v)-1]

  #loop for different u
  for (ui in seq_along(u)){
    a=(d+1)*u[ui]
    b=(d+1)*(1-u[ui])
    ##calculate Qcap using simulated data
    beta_dist=pbeta(fn_v,a,b)
    lim=beta_dist[1]
    for (i in 2:length(beta_dist))
      {lim[i]=beta_dist[i]-beta_dist[i-1]}
    Q_cap[j,ui]=sum(lim*v)
  }
} #simulation loop ends
HDH_cap=cov(Q_cap)*sample_size
mean_Q_cap=colMeans(Q_cap)

###############output ################################
Q_result=rbind(Q,mean_Q_cap)
colnames(Q_result)<-u

Q_result
HDH_theo
HDH_cap
#########################################################
```

Appendix D

R code: Asymptotic and Estimated Means and Covariance-Variance Matrices for Poisson Distribution Using Truncated Data

```
#Q_Poi_Chebychev.R file
#Generates numbers for tables 4.4 and 4.5
rm(list=ls())
#########Poisson parameters###################################
library(VGAM)
lambda=1;

#####################initialize variables###################
k=3
sample_size=500;n_sim=100000
seed_set=1;
ustar=c(1/4,1/2,3/4)

####################Apply Chebychev rule to theory########
M_theory=lambda
S_theory=sqrt(lambda)

#setting Lk(theory_min) Uk(theory_max) as in equation 4.5
theory_min=max(0,floor(M_theory-k*S_theory))
theory_max=floor(M_theory+k*S_theory)+1
v<-seq(theory_min,theory_max)

#F cumultiv distribution function for Poisson
F_initial <- as.vector(ppois(v,lambda))
#Resetting cumulative distribution as in equation 4.10
F<-(F_initial-ppois((theory_min-1),lambda))/
    (ppois(theory_max,lambda)-ppois((theory_min-1),lambda))
u <- (ustar-ppois((theory_min-1),lambda))/
       (ppois(theory_max,lambda)-ppois((theory_min-1),lambda))

#setting d as in equation 4.8
d = theory_max-theory_min+1

#v_theory=seq(min(uniq_seq1),max(uniq_seq1))
v.minus_1=v[2:length(v)]
v.minus_d=v[1:length(v)-1]
```

```
Q=vector(length=length(u))
Q_cap=matrix(nrow=n_sim,ncol=length(u))
mean_Q_cap=vector(length=length(u))

tau_theo=matrix(nrow=length(u),ncol=d-1)
Q_var=matrix(nrow=length(u),ncol=length(u))
Q_var_theo=matrix(nrow=length(u),ncol=length(u))
mean_Q_var_sim=matrix(0,nrow=length(u),ncol=length(u))
mean_Q_cap=vector(length=length(u))

###theory:calculate Q, and HDH'############################
F_minus_d=F[1:length(F)-1]
#D_theo is D in HDH'for theory
D_theo=matrix((1-F_minus_d),nrow=length(F_minus_d),
              ncol=length(F_minus_d),byrow=TRUE)
D_theo=F_minus_d*D_theo
D_theo[lower.tri(D_theo)] <- t(D_theo)[lower.tri(D_theo)]
for (ui in seq_along(u)){
  a=(d+1)*u[ui]
  b=(d+1)*(1-u[ui])
  beta_dist_th=pbeta(F,a,b)
  limit=beta_dist_th[1]
  #calculate smooth quantile
  for (i in 2:length(beta_dist_th))
    {limit[i]=beta_dist_th[i]-beta_dist_th[i-1]}
  Q[ui]=round(sum(limit*v),5)
  #theory variance:tau_theo is same as H in HDH'
  beta_density=dbeta(F_minus_d,(d+1)*u[ui], (d+1)*(1-u[ui]))
  tau_theo[ui,]=beta_density*(v.minus_d-v.minus_1)
}
HDH_theo=tau_theo%*%D_theo%*%t(tau_theo)
#correlation matrix
Corr1=cov2cor(HDH_theo)

###################simulation loop starts###############
for(j in 1:n_sim){
  seed_set=seed_set+1
  set.seed(seed_set)
  ###generate poisson data
  seq1=sort(rpois(sample_size,lambda))
  uniq_seq1<-unique(seq1)
  fn1=ecdf(seq1)
  fn_seq1=fn1(unique(seq1))
  fn_seq2=fn_seq1
  uniq_seq2<-uniq_seq1
  d_sample= max(uniq_seq1)-min(uniq_seq1)+1

  fn_v<-rep(0,d_sample)
  v_sample=seq(min(uniq_seq1),max(uniq_seq1))
  #setdiff(v_sample,uniq_seq1)
```

```
  #find Fn for missing v in simulation
  for (i in 1:d_sample){
    for (m in seq_along(uniq_seq1)){
      if(v_sample[i]==uniq_seq1[m]){fn_v[i]<-fn_seq1[m]}
    }
  }
  fn_v_mod<-fn_v
  #extract non zero fn_v if the first few fn_v are zero
  if (length(which(fn_v == 0))!=0){
    if (min(which(fn_v == 0))==1){
      nzero_loc<-min(which(fn_v != 0))
      fn_v_mod<-fn_v[nzero_loc:d_sample]
    }
   }
  ## for missing v,replace its F with F for v-1
  while (length(which(fn_v_mod == 0))!=0){
    loc <- which(fn_v_mod == 0)
    for(m in seq_along(loc))
    {loc_zero<-loc[m]
    fn_v_mod[loc_zero]<-fn_v_mod[loc_zero-1]}
  }
  if (length(which(fn_v == 0))!=0){
    if (min(which(fn_v == 0))==1){
      nzero_loc<-min(which(fn_v != 0))
      app_0<-rep(0,nzero_loc-1)
      fn_v_mod<-c( app_0,fn_v_mod)
    }}
  fn_v<-fn_v_mod

##################Apply chebhychev to simulation#####################
  M=mean(seq1)
  S=sd(seq1)
  #setting Lkcap(mod_min) Ukcap(mod_max) as in equation 4.4
  mod_min=max(0,floor(M-k*S))
  mod_max=floor(M+k*S)+1
  #Setting dcap(d_sample2) as in equation 4.7
  d_sample2=mod_max-mod_min+1

 #Tuning v_sample and fnv before resetting
  if (min(v_sample) <= mod_min){
    if (mod_min == 0 || mod_min == min(v_sample)) {
      f0=0} else{f0=fn_v[which(v_sample == (mod_min-1))]
    }
    fn_v <- fn_v[min(which(v_sample >= mod_min)):length(fn_v)]
    v_sample <- v_sample[min(which(v_sample >= mod_min)):length(v_sample)]

  }  else {
    f0=0
    seq2<- seq(mod_min,(min(v_sample)-1))
    v_sample<-c(seq2,v_sample)
    fn_v<-c(rep(0,max(seq2)-min(seq2)+1),fn_v)
```

```
  }

  if (max(v_sample) >= mod_max){
    fn_v <- fn_v[1:max(which(v_sample <= mod_max))]
    v_sample <- v_sample[1:max(which(v_sample <= mod_max))]
  }else {
    seq2<- seq(max(v_sample)+1,mod_max)
    v_sample<-c(v_sample,seq2)
    fn_v<-c(fn_v,rep(1,max(seq2)-min(seq2)+1))
  }
  #Resetting cumulative distribution as in equation 4.9
  fn_v <- (fn_v-f0)/(fn_v[length(fn_v)]-f0)

  v_minus_1=v_sample[2:length(v_sample)]
  v_minus_d=v_sample[1:(length(v_sample)-1)]
  fn_v_minus_1=fn_v[2:length(fn_v)]
  fn_v_minus_d=fn_v[1:(length(fn_v)-1)]

  #loop for different u
  u_sample <- (ustar-f0)/(fn_v[length(fn_v)]-f0)
  #print(u_sample)
  for (ui in seq_along(u_sample)){
    a=(d_sample2+1)*u_sample[ui]
    b=(d_sample2+1)*(1-u_sample[ui])

    ##calculate Qcap using simulated data
    beta_dist=pbeta(fn_v,a,b)
    lim=beta_dist[1]
    if (length(v_sample) == 1){Q_cap[j,ui]=lim*v_sample}else {
      for (i in 2:length(beta_dist))
        {lim[i]=beta_dist[i]-beta_dist[i-1]}
      Q_cap[j,ui]=sum(lim*v_sample)}
   } #ui loop ends
} #sim loop ends
HDH_cap=cov(Q_cap)*sample_size
mean_Q_cap=colMeans(Q_cap)

###############output ########################################
Q_result=rbind(Q,mean_Q_cap)
colnames(Q_result)<-ustar

Q_result
HDH_theo
HDH_cap
#################################################################
```

Appendix E

R code: Asymptotic and Estimated Means and Covariance-Variance Matrices for ZIP Distribution Using Truncated Data

```
#Q_ZIP_Chebychev.R file
#Generates numbers for tables 4.6 and 4.7
rm(list=ls())
#########ZIP parameters###################################
library(VGAM)
lambda=10;
c=0.2

######################initialize variables###################
k=2
sample_size=50;n_sim=100000
seed_set=1;
ustar=c(1/4,1/2,3/4)

#####################Apply Chebychev rule to theory########
p0=ppois(0,lambda)
#mean of ZIP
M_theory=((1-c)/(1-p0))*lambda
#std dev of ZIP
S_theory=sqrt(((1-c)/(1-p0))*(lambda+((c-p0)/(1-p0))*lambda^2))

#setting Lk(theory_min) Uk(theory_max) as in equation 4.5
theory_min=max(0,floor(M_theory-k*S_theory))
theory_max=floor(M_theory+k*S_theory)+1
v<-seq(theory_min,theory_max)

pi=(c-p0)/(1-p0)
#F cumultiv distribution function for ZIP
F_initial=as.vector(pzipois(v,lambda,pi))
#Resetting cumulative distribution as in equation 4.10
F<-(F_initial-pzipois((theory_min-1),lambda,pi))/
    (pzipois(theory_max,lambda,pi)-pzipois((theory_min-1),lambda,pi))
u <- (ustar-pzipois((theory_min-1),lambda,pi))/
       (pzipois(theory_max,lambda,pi)-pzipois((theory_min-1),lambda,pi))

#setting d as in equation 4.8
```

```
d = theory_max-theory_min+1

v.minus_1=v[2:length(v)]
v.minus_d=v[1:length(v)-1]

Q=vector(length=length(u))
Q_cap=matrix(nrow=n_sim,ncol=length(u))
mean_Q_cap=vector(length=length(u))


tau_theo=matrix(nrow=length(u),ncol=d-1)
Q_var=matrix(nrow=length(u),ncol=length(u))
Q_var_theo=matrix(nrow=length(u),ncol=length(u))
mean_Q_var_sim=matrix(0,nrow=length(u),ncol=length(u))
mean_Q_cap=vector(length=length(u))

###theory:calculate Q, and HDH'############################
F_minus_d=F[1:length(F)-1]
#D_theo is D in HDH'for theory
D_theo=matrix((1-F_minus_d),nrow=length(F_minus_d),
                ncol=length(F_minus_d),byrow=TRUE)
D_theo=F_minus_d*D_theo
D_theo[lower.tri(D_theo)] <- t(D_theo)[lower.tri(D_theo)]
for (ui in seq_along(u)){
  a=(d+1)*u[ui]
  b=(d+1)*(1-u[ui])
  beta_dist_th=pbeta(F,a,b)
  limit=beta_dist_th[1]
  #calculate smooth quantile
  for (i in 2:length(beta_dist_th))
    {limit[i]=beta_dist_th[i]-beta_dist_th[i-1]}
  Q[ui]=round(sum(limit*v),5)
  #theory variance:tau_theo is same as H in HDH'
  beta_density=dbeta(F_minus_d,(d+1)*u[ui], (d+1)*(1-u[ui]))
  tau_theo[ui,]=beta_density*(v.minus_d-v.minus_1)
}
HDH_theo=tau_theo%*%D_theo%*%t(tau_theo)
#correlation matrix
Corr1=cov2cor(HDH_theo)

###################simulation loop starts###############
for(j in 1:n_sim){
  #print(c("simulation",j));
  seed_set=seed_set+1
  set.seed(seed_set)
  ###generate zip data
  seq1=sort(rzipois(sample_size,lambda,pi))
  uniq_seq1<-unique(seq1)
  fn1=ecdf(seq1)
  fn_seq1=fn1(unique(seq1))
  fn_seq2=fn_seq1
```

```
  uniq_seq2<-uniq_seq1
  d_sample= max(uniq_seq1)-min(uniq_seq1)+1

  fn_v<-rep(0,d_sample)
  v_sample=seq(min(uniq_seq1),max(uniq_seq1))

  #find Fn for missing v in simulation
  for (i in 1:d_sample){
    for (m in seq_along(uniq_seq1)){
      if(v_sample[i]==uniq_seq1[m]){fn_v[i]<-fn_seq1[m]}
    }
  }
  fn_v_mod<-fn_v
  #extract non zero fn_v if the first few fn_v are zero
  if (length(which(fn_v == 0))!=0){
    if (min(which(fn_v == 0))==1){
      nzero_loc<-min(which(fn_v != 0))
      fn_v_mod<-fn_v[nzero_loc:d_sample]
    }}
  ## for missing v,replace its F with F for v-1
  while (length(which(fn_v_mod == 0))!=0){
    loc <- which(fn_v_mod == 0)
    for(m in seq_along(loc))
    {loc_zero<-loc[m]
    fn_v_mod[loc_zero]<-fn_v_mod[loc_zero-1]}
  }
  if (length(which(fn_v == 0))!=0){
    if (min(which(fn_v == 0))==1){
      nzero_loc<-min(which(fn_v != 0))
      app_0<-rep(0,nzero_loc-1)
      fn_v_mod<-c( app_0,fn_v_mod)
    }}
  fn_v<-fn_v_mod

##################Apply chebhychev to simulation#####################
  M=mean(seq1)
  S=sd(seq1)
  #setting Lkcap(mod_min) Ukcap(mod_max) as in equation 4.4
  mod_min=max(0,floor(M-k*S))
  mod_max=floor(M+k*S)+1
  #Setting dcap(d_sample2) as in equation 4.7
  d_sample2=mod_max-mod_min+1

  #Tuning v_sample and fnv before resetting
  if (min(v_sample) <= mod_min){
    if (mod_min == 0 || mod_min == min(v_sample))
      {f0=0} else{f0=fn_v[which(v_sample == (mod_min-1))]}
    fn_v <- fn_v[min(which(v_sample >= mod_min)):length(fn_v)]
    v_sample <- v_sample[min(which(v_sample >= mod_min)):length(v_sample)]

  }  else {
```

```
    f0=0
    seq2<- seq(mod_min,(min(v_sample)-1))
    v_sample<-c(seq2,v_sample)
    fn_v<-c(rep(0,max(seq2)-min(seq2)+1),fn_v)
  }

  if (max(v_sample) >= mod_max){
    fn_v <- fn_v[1:max(which(v_sample <= mod_max))]
    v_sample <- v_sample[1:max(which(v_sample <= mod_max))]
  }else {
    seq2<- seq(max(v_sample)+1,mod_max)
    v_sample<-c(v_sample,seq2)
    fn_v<-c(fn_v,rep(1,max(seq2)-min(seq2)+1))
  }

  #Resetting cumulative distribution as in equation 4.9
  fn_v <- (fn_v-f0)/(fn_v[length(fn_v)]-f0)

  v_minus_1=v_sample[2:length(v_sample)]
  v_minus_d=v_sample[1:(length(v_sample)-1)]
  fn_v_minus_1=fn_v[2:length(fn_v)]
  fn_v_minus_d=fn_v[1:(length(fn_v)-1)]

  u_sample <- (ustar-f0)/(fn_v[length(fn_v)]-f0)
  u_sample<-replace(u_sample, (u_sample < 0 | u_sample > 1), 0)
  #loop for different u
  for (ui in seq_along(u_sample)){
    a=(d_sample2+1)*u_sample[ui]
    b=(d_sample2+1)*(1-u_sample[ui])
    ##calculate Qcap using simulated data
    beta_dist=pbeta(fn_v,a,b)
    lim=beta_dist[1]
    if (length(v_sample) == 1){Q_cap[j,ui]=lim*v_sample}else {
      for (i in 2:length(beta_dist))
        {lim[i]=beta_dist[i]-beta_dist[i-1]}
      Q_cap[j,ui]=sum(lim*v_sample)}
   } #ui loop ends
} #sim loop ends
HDH_cap=cov(Q_cap)*sample_size
mean_Q_cap=colMeans(Q_cap)

###############output ##############################################
Q_result=rbind(Q,mean_Q_cap)
colnames(Q_result)<-ustar

Q_result
Q_result
HDH_theo
HDH_cap
#####################################################################
```

Curriculum Vitae
Ponmalar Ratnam

RESEARCH INTERESTS

Claim Frequency Models; Robust Statistics; Statistical Inference.

EDUCATION

Ph.D., Mathematics (concentration in Statistics & Actuarial Science),
University of Wisconsin - Milwaukee, (2020).
Thesis Title: "Smoothed Quantiles for Claim Frequency Models, with Applications to Risk Measurement." Advisor: Prof. Vytaras Brazauskas.
M.S., Mathematical Statistics, University of Wisconsin - Milwaukee, (2005).
Project Title: "Covariance function estimation by ridge tensor product splines."
Advisor: Prof. Daniel Gervini.
M.S., Mathematics, Statistics and Computer Science,
Marquette University, (2003).

TEACHING AND INDUSTRY EXPERIENCE

Actuarial Team Leader (2018 - Current), Excellus BlueCross BlueShield, Rochester, NY.

- Collaborate with BlueCross BlueShield Association(BCBSA) in reviewing Medicare and Commercial risk adjustment models and policies implemented by Centers of Medicare and Medicaid Services (CMS).
- Attend webinars and work forums conducted by CMS/BCBSA and communicate the updates to executives and team.
- Mentor and train direct reports.

Manager (2016 - 2018), Cognizant Solutions
Clients: Kaiser Permanente, Xerox Corp

- Played a key role in data analytics which helped in drawing insights from data.
- Translated business needs into business goals and measures of success.
- Built predictive models for preventing chronic disease progression and image recognition.

Data Scientist (2013 – 2016), Systech Solutions
Client: Walmart E-commerce

- Designed and implemented predictive models and advance data mining techniques for Big Data.
- Instrumental in training and developing a team for predictive model building.

Teaching Assistant & Instructor (2009 - 2013)

– Taught Calculus, Algebra, and Statistics classes at University of Wisconsin - Milwaukee and Marquette University.

Senior Actuarial Analyst (2007 - 2009), Actuarial Assistant (2006 - 2007), Actuarial Intern (2005), Assurant Health, Milwaukee,WI.

– Built predictive models using multivariate and logistic regression models to price different health insurance products and to determine expected return on revenue.

– Built forecast models to predict premium, claims, income and expenses for short term health insurance product. Determined forecasting factors based on updates from the management, business trends, seasonality changes.

PROFESSIONAL ACTIVITIES AND MEMBERSHIP

- American Statistical Association (ASA), 2009-present.
- Referee for Metron –International journal of statistics, (2011).

PUBLICATION

Brazauskas, V., Dornheim, H., and Ratnam, P. (2014). "Credibility and regression modeling". In *Predictive Modeling Applications in Actuarial Science, Volume I: Predictive Modeling Techniques* (E. Frees, R. Derrig, G. Meyers, eds.), 217–235; Cambridge University Press.

CONFERENCE PRESENTATIONS

"Smoothed Quantiles, Value-at-Risk, and the Method of Percentile-Matching for Claim Count Data." 48th Actuarial Research Conference, Philadelphia, PA, (2013).

HONORS, AWARDS AND FELLOWSHIPS

- SOA Travel Award, Society of Actuaries, Schaumburg, IL, (2013).
- Graduate School Travel Award, University of Wisconsin-Milwaukee, WI, (2013).
- SOA Predictive Modeling Symposium, Chicago, IL, (2010).
- SAMSI Genomes to Global Health Workshop, Research Triangle Park, NC, (2004).