Pakistan Journal of Statistics and Operation Research

A discrete claims-model for the inflated and over-dispersed automobile claims frequencies data: Applications and actuarial risk analysis

Haitham M. Yousof [1,*], Mohammad Mehdi Saber[2], Abdullah H. Al-Nefaie [3], Nadeem S. Butt [4] and Mohamed Ibrahim [3,5] and Salwa L. Alkhayyat [6,7]

* Corresponding Author

[1] Department of Statistics, Mathematics and Insurance, Benha University, Egypt; haitham.yousof@fcom.bu.edu.eg
[2]Department of Statistics, Higher Education Center of Eghlid, Eghlid, Iran; mohammadmehdisaber@gmail.com
[3] Department of Quantitative Methods, college of Business, King Faisal University, Al Ahsa 31982, Saudi Arabia; aalnefaie@kfu.edu.sa
[4] Department of Family and Community Medicine, King Abdul Aziz University, Jeddah, Kingdom of Saudi Arabia; nshafique@kau.edu.sa
[5] Department of Applied, Mathematical and Actuarial Statistics, Faculty of Commerce, Damietta University, Damietta, Egypt; miahmed@kfu.edu.sa and mohamed_ibrahim@du.edu.eg
[6] Department of Statistics, Faculty of Science, University of Jeddah, Kingdom of Saudi Arabia; slalkhayyat@uj.edu.sa
[7] Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Kafr El-Sheikh University, Kafr El-Sheikh 33511, Egypt; salwa.elkhayat@com.kfs.edu.eg

Abstract

This paper showcases the effectiveness of the discrete generalized Burr-Hatke distribution in analyzing insurance claims data, specifically focusing on scenarios with over-dispersed and zero-inflated claims. Key contributions include presenting foundational statistical theories with mathematical proofs to enrich the paper's mathematical and statistical aspects. Through the application of this discrete distribution, the study conducted a thorough risk analysis across five diverse sets of insurance claims data, evaluating critical risk indicators at specified quantiles. These indicators provided detailed insights into potential losses across different risk levels, supporting effective risk management strategies. The research emphasizes the importance of selecting appropriate probability distributions when analyzing zero-inflated data, as commonly observed in insurance claims. The discrete distribution accommodated these unique data characteristics and facilitated a robust analysis of risk metrics, enhancing the accuracy of potential loss assessments and reducing associated uncertainties. Furthermore, the study highlights the practical relevance of the discrete distribution in addressing specific challenges inherent to insurance claims data. By leveraging this distribution, insurers and risk analysts can improve their risk modeling capabilities, leading to more informed decision-making and enhanced financial exposure management.

Key Words: Discrete Claims Distribution; Expected Loss; Inflated Claims; Over-dispersed Automobile Claims; Risk Analysis; Value at Risk.

Mathematics Subject Classification: 62E15, 62J05.

1. Introduction

Automobile insurance is a vital segment of the insurance industry, offering financial protection against losses from traffic accidents, theft, and other vehicle-related incidents. Effective management of automobile insurance hinges on the collection and analysis of claims data, which is essential for risk assessment and informed decision-making by insurance companies. Claims data is fundamental for precise insurance premium calculations. By examining historical data, insurers can forecast the likelihood of future claims and adjust premiums accordingly. This data allows insurers to profile customers based on various risk factors, including age, driving history, location, and vehicle type, facilitating

more accurate underwriting and risk categorization. Analyzing large datasets enables insurers to detect patterns that suggest fraudulent activities, such as repeated claims or inconsistencies in reported incidents. Additionally, claims data analysis helps insurers develop customized insurance products tailored to the specific needs of different customer segments, enhancing market competitiveness. Data-driven insights also support dynamic pricing strategies, allowing insurers to remain competitive while ensuring profitability. Insurance companies must adhere to regulatory requirements concerning data reporting and transparency. Claims data analysis ensures accurate and timely reporting to regulatory authorities. Regulators often mandate that insurers demonstrate robust risk management practices, and claims data is crucial for identifying and managing potential risks in line with regulatory standards. For more details, see Lemaire (1995); Derrig (2002); Viaene and Dedene (2004); Cummins and Weiss (2013); Brockett and Golden (2007);

In the realm of insurance and risk analysis, the study of claims frequencies is paramount for accurately assessing and managing potential financial liabilities. In particular, the analysis of automobile claims data presents unique challenges due to the presence of inflated and over-dispersed frequencies, where certain factors such as demographic characteristics, vehicle types, and geographic locations significantly influence the frequency and severity of claims (see Johnson et al. (2010)). The objective of this paper is to employ the discrete generalized Burr-Hatke (DGBH) distribution (see Yousof et al. (2021)) for analyzing automobile claims frequencies characterized by inflation and over-dispersion. Unlike traditional continuous models, discrete models offer a more flexible and nuanced approach to capturing the complexities of claims data, particularly when dealing with count data subject to excess zeros and higher variability than expected under standard statistical distributions (see Derrig (2002) and Smith and Jones (2015)).

The focus on a discrete modeling framework is motivated by the inadequacies of conventional continuous distributions in adequately representing the unique characteristics observed in automobile claims datasets. By leveraging discrete models, we aim to overcome limitations associated with assumptions of normality and homogeneity, providing a more accurate representation of the underlying claims processes. This paper contributes to the field of statistics and risk analysis by proposing a novel methodology that accounts for both the excess zeros and the variability inherent in automobile claims frequencies. Our approach integrates advanced statistical techniques to develop a tailored discrete claims model capable of capturing the intricate patterns and dependencies within the data, thereby enhancing the accuracy and reliability of risk assessments in the insurance sector.

Recently, Yousof et al. (2021) introduced a novel generalized discrete distribution called the discrete generalized Burr-Hatke (DGBH) distribution, encompassing the discrete Burr-Hatke distribution, and conducted a comprehensive examination of its properties. This new distribution exhibits diverse characteristics in its probability mass function, displaying right skewness with varying shapes, bimodality, and uniformity. Moreover, its corresponding hazard rate function can exhibit monotonically decreasing, monotonically increasing, or constant behaviors. In their study, Yousof et al. (2021) performed numerical analyses to elucidate key statistical measures such as mean, variance, skewness, kurtosis, and the index of dispersion. The versatility of this distribution renders it valuable for modeling both under-dispersed and over-dispersed count data. The authors also provided certain characterizations of the distribution based on the conditional expectation of specific functions of the random variable and in terms of the hazard rate function. Furthermore, the study explored Bayesian and non-Bayesian estimation methods, with numerical simulations conducted to compare their efficacy. Notably, Yousof et al. (2021) applied the newly proposed model to real datasets, including carious teeth data and counts of kidney cysts, demonstrating its applicability in practical scenarios.

In the realm of statistical modeling, the choice of an appropriate probability distribution plays a pivotal role in accurately representing real phenomena. One notable area where statistical modeling holds immense significance is in insurance analytics, particularly in the assessment and prediction of automobile insurance claims. With the

burgeoning complexity of insurance data and the need for precise risk assessment, the development of novel distributions tailored to accommodate the intricacies of such data becomes imperative (see Klugman et al. (2012)).

The DGBH distribution emerges as a promising candidate in this context, offering a flexible framework to model discrete count data with varying degrees of dispersion and skewness. This distribution extends the classical discrete Burr-Hatke distribution, providing enhanced versatility in capturing the diverse statistical characteristics often observed in insurance claim data. Originating from the seminal work of Yousof et al. (2021), the DGBH distribution presents a comprehensive solution for modeling discrete count data, incorporating a wide range of distributional properties. Yousof et al. (2021) introduced and thoroughly investigated this distribution, delineating its fundamental statistical properties, probability mass function characteristics, and hazard rate function behaviors. Their study not only establishes the theoretical foundation of the GBH distribution but also underscores its practical utility in modeling scenarios where traditional distributions may fall short in capturing the nuances of the data.

In the context of automobile insurance claims, the DGBH distribution holds significant promise. Automobile claim data often exhibit complexities such as over-dispersion, right skewness, and multimodality, necessitating the use of flexible distributional models capable of accommodating such features. By leveraging the inherent flexibility of the DGBH distribution, researchers and practitioners can obtain more accurate and reliable models for assessing claim frequency and severity, thereby facilitating better risk management strategies within the insurance industry (see Coskun et al. (2018) and Bolancé and Guillén (2011)).

This paper aims to delve deeper into the application of the DGBH distribution in modeling automobile claim data. Through empirical analysis and case studies, we seek to demonstrate the efficacy of the DGBH distribution in capturing the inherent complexities of automobile insurance data, thereby enhancing the precision and reliability of insurance risk assessment models. The paper used five different sets of Automobile insurance claims data (see Coskun et al. (2018), Gossiaux and Lemaire (1981) and Willmot (1987)). The data was analyzed statistically and the distribution was compared to a set of other, more competitive distributions in this field.

In the subsequent sections, we will provide a comprehensive overview of the discrete DGBH distribution, including its mathematical formulation, key statistical properties, and estimation techniques. We will then proceed to illustrate its application in modeling automobile claim data through insurance case studies and numerical simulations, showcasing its potential to revolutionize statistical modeling practices in the insurance industry. By elucidating the significance of the DGBH distribution and its application in modeling automobile claim data, this research endeavors to contribute to the advancement of statistical methodologies in insurance analytics, ultimately fostering more robust and informed decision-making processes within the insurance sector. Moreover, some risk indicators are considered to analyze the insurance automobile claims such as $\mathrm{VaR}_{[q]}(Z)$ (Value at Risk), $\mathrm{TVaR}_{[q]}(Z)$ (Tail Value at Risk), $\mathrm{TV}_{[q]}(Z)$ (Tail Variance at quantile q), $\mathrm{TMV}_{[q]}(Z)$ (Tail Mean Variance at quantile q) and $\mathrm{EL}_{[q]}(Z)$ (Expected Loss at quantile q). In this context, the DGBH distribution is evaluated in light of risk calculations and analysis of its behavior. The probability mass function (PMF) of the DGBH distribution according to Yousof et al. (2021) can be expressed as

$$P(z) = P_{\pi,\beta}(z) = \frac{1}{z+1}\pi^{z^{\beta}} - \frac{1}{z+2}\pi^{(z+1)^{\beta}}; 0 < \pi < 1 \text{and} z \in \mathbb{N}^{*}. \quad (1)$$

Starting with (1), we can provide new two theorems:

Theorem 1:

For fixed z, the $P_{\pi,\beta}(z)$ is monotonically decreasing as π increases since $\pi^{z^{\beta}}$ and $\pi^{(z+1)^{\beta}}$ are both decreasing functions as π increases.

Proof:

Let's consider two consecutive values of π, say π_1 and π_2, such that $0 < \pi_1 < \pi_2 < 1 < +\infty$. Then, for any fixed z, we need to show that:

$$P_{\pi_1,\beta}(z) > P_{\pi_2,\beta}(z)$$

Where $P_{\pi,\beta}(z)$ denotes the PMF evaluated at π. For $\pi = \pi_1$:

$$P_{\pi_1,\beta}(z) = \frac{1}{z+1}\pi_1 - \frac{1}{z+2}\pi_1$$

For $\pi = \pi_2$:

$$P_{\pi_2,\beta}(z) = \frac{1}{z+1}\pi_2 - \frac{1}{z+2}\pi_2$$

Now, since $\pi_2 > \pi_1$, we can compare the terms involving π, we have $\pi_2 > \pi_1$and $\pi_2 > \pi_1$. This implies that each term in the expression for $P_{\pi_2,\beta}(z)$ is smaller than the corresponding term in the expression for $P_{\pi_1,\beta}(z)$. Therefore,

$$P_{\pi_1,\beta}(z) > P_{\pi_2,\beta}(z).$$

Theorem 2:
For fixed π, it's challenging to determine the monotonicity of the $P_{\pi,\beta}(z)$ with respect to z without more information about the behavior of the function $\pi^{z^{\beta}}$ and $\pi^{(z+1)^{\beta}}$. This might require examining the derivative of $F_{\pi,\beta}(z)$ with respect to z and analyzing its sign changes.

Proof:
To prove the monotonicity of the PMF with respect to z, we can analyze the sign of the derivative of $P_{\pi,\beta}(z)$ with respect to z and examine its behavior. Calculating the derivative and analyzing its sign can be complex due to the nonlinear form of $P_{\pi,\beta}(z)$. It may require numerical methods or special techniques for certain values of β. For simplicity, let's assume ($\beta = 1$and$\pi = 0.5$) and ($\beta = 0.05$and$\pi = 0.1$). Then, we can plot the $P_{\pi,\beta}(z)$ for various values of z to observe its monotonicity. By examining the plot, we can observe the behavior of $P_{\pi,\beta}(z)$ with respect to z. We can also try different values of β and π to observe how they affect the monotonicity of $P_{\pi,\beta}(z)$, see Figure 1 for more details. Figure 1 shows the monotonicity of the PMF with respect to z. These proofs provide a theoretical basis for the monotonicity properties of the given PMF with respect to π and z, and numerical exploration can provide further insights. Although it is difficult to obtain specific mathematical formulas for the mean, median, and mode, we can calculate these statistical measures numerically for some parameters values. For example, for $\beta = 1$ and $\pi = 0.5$, the mean$= 0.38629$, the mode$= 1$ and the median$= 1$.

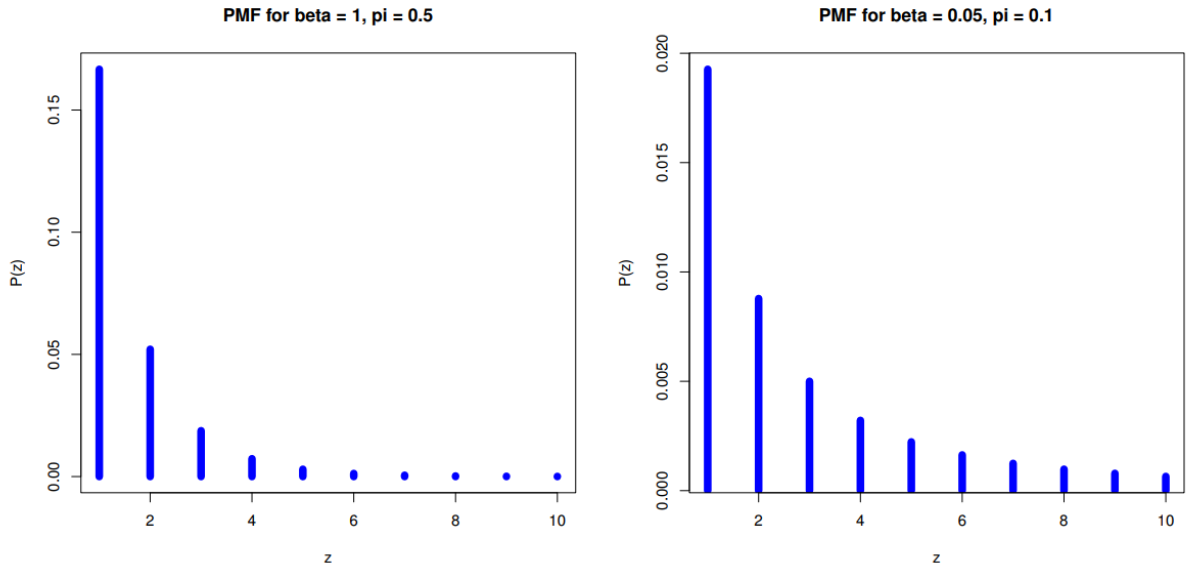

Figure 1: Monotonicity of the PMF with respect to z.

2. Risk indicator derivations

The VaR at a specified confidence level α is the q-quantile of the distribution. For a discrete distribution represented by the PMF, VaR can be computed by finding the smallest value v such that

$$P(Z \leq v) \geq q.$$

For the PMF, the VaR at confidence level 3b13b1 can be found numerically by solving:

$$\Sigma_{z \leq v} P_{\pi,\beta}(z) = \Sigma_{z \leq v} \left[\frac{1}{z+1} \pi^{z^\beta} - \frac{1}{z+2} \pi^{(z+1)^\beta} \right] \geq q,$$

here, $v = \text{VaR}_{[q]}(Z)$ is our VaR. The Tail Value at Risk at a specified confidence level $3b1$ (or Conditional Value at Risk) represents the expected loss beyond the VaR. It is computed as the conditional expectation of losses exceeding the VaR. TVaR at confidence level $3b1$ can be calculated as:

$$\text{TVaR}_{[q]}(Z) = \frac{1}{1-q} \Sigma_{z \leq \text{VaR}} \left(z - \text{TVaR}_{[q]}(Z) \right) \left[\frac{1}{z+1} \pi^{z^\beta} - \frac{1}{z+2} \pi^{(z+1)^\beta} \right].$$

Tail Variance at a specific quantile qq represents the variance of losses beyond the q-quantile. It is calculated as:

$$\text{TV}_{[q]}(Z) = \Sigma_{z \leq q} (z-q)^2 \times \left[\frac{1}{z+1} \pi^{z^\beta} - \frac{1}{z+2} \pi^{(z+1)^\beta} \right].$$

The Tail Mean Variance at a quantile qq is the mean of squared losses beyond the qq-quantile. It can be computed as:

$$\text{TMV}_{[q]}(Z) = \frac{1}{1-q} \Sigma_{z \leq q} (z-q)^2 \times \left[\frac{1}{z+1} \pi^{z^\beta} - \frac{1}{z+2} \pi^{(z+1)^\beta} \right].$$

The Expected Loss at a quantile q is the expected value of losses beyond the q-quantile. It is given by:

$$\mathrm{EL}_{[q]}(Z) = \Sigma_{z\le q}(z-q)\times\left[\frac{1}{z+1}\pi^{z^{\beta}} - \frac{1}{z+2}\pi^{(z+1)^{\beta}}\right],$$

where, the $\mathrm{VaR}_{[q]}(Z)$ provides a threshold value representing the maximum loss that might occur with a given confidence level. It is a crucial risk metric for setting risk limits and capital requirements. The $\mathrm{TVaR}_{[q]}(Z)$ quantifies the average severity of losses beyond the VaR, providing deeper insights into potential extreme losses and tail risk. the $\mathrm{TV}_{[q]}(Z)$ measures the spread or variability of losses in the tail of the distribution beyond a specified quantile, offering information on the risk of extreme outcomes. The $\mathrm{TMV}_{[q]}(Z)$ provides the average squared deviation of losses beyond a quantile, indicating the average variability of extreme losses. Finally, the $\mathrm{EL}_{[q]}(Z)$ estimates the average amount of loss expected beyond a certain quantile, aiding in risk assessment and scenario planning.

3. Describing automobile claims data

In this segment, we delve into the analysis of automobile insurance claim frequencies across various countries. Gossiaux and Lemaire (1981) provide a comprehensive examination of this aspect, presenting five distinct datasets in their research. Notably, these datasets are also referenced in the work of Willmot (1987). As depicted in Table 1, these datasets collectively unveil the phenomenon of inflated-over-dispersion within automobile insurance automobile claims. Other releted data can be found in Yousof et al (2023a,b,c), Alizadeh et al. (2023), Elbatal et al. (2023), Hamedani et al. (2023), Hamed et al. (2022), Hashempour et al. (2023) , Salem et al. (2023), Eliwa et al. (202), Mohamed et al. (2022, 2023, 2024), Emam et al. (2023) and Tashkandy et al. (2023). Figure 2 displays the box plots for the five data sets.

Table 1: Automobile claim data.

		Claims							
Country	**Year**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Switzerland	1961	103704	14075	1766	255	45	6	2	0
Great Britain	1968	370412	46545	3935	317	28	3	0	0
Belgium	1958	7840	1317	239	42	14	4	4	1
Zaire	1974	3719	232	38	7	3	1	0	0
Germany	1960	20592	2651	297	41	7	0	1	0

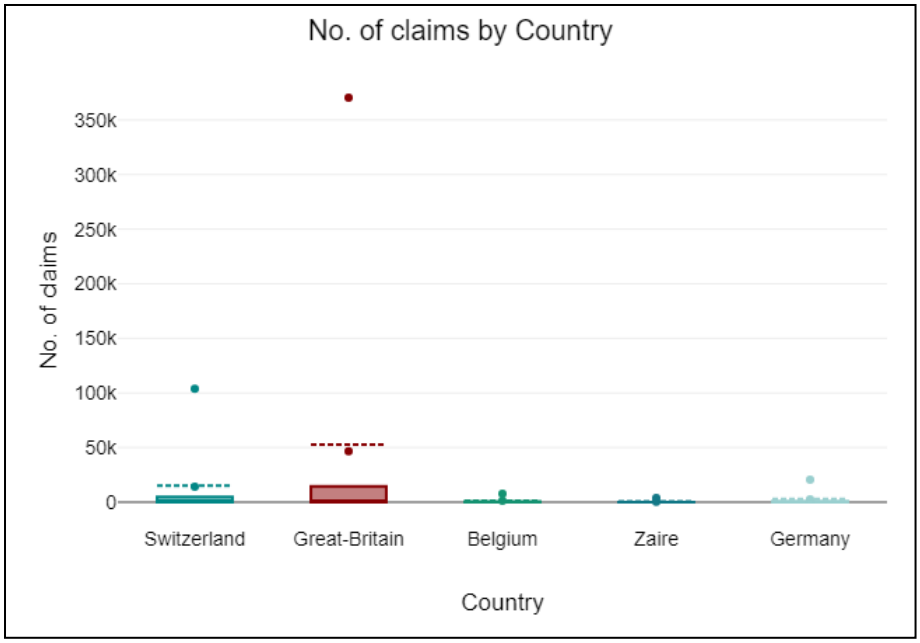

Figure 2: The box plots.

Due to Table 1 (the first data), this entry represents data for Switzerland in the year 1961. The majority of automobile claims fall into the category of "0 claims," with 103,704 reported instances. There is a significant drop in the number of claims as we move to higher claim categories, indicating a right-skewed distribution. The number of automobile claims decreases progressively as the category of automobile claims increases, which is a typical pattern in insurance claim data. The second data: data is for Great Britain in the year 1968. Similar to Switzerland, there is a large number of claims in the "0 claims" category, with 370,412 instances.

The number of automobile claims decreases as we move to higher claim categories, consistent with the right-skewed distribution observed in insurance claim data. There are no reported instances for automobile claims in categories 6 and 7, which suggests a relatively low frequency of extreme automobile claims.

The third data: data corresponds to Belgium in 1958. Compared to the previous two entries, the number of automobile claims is substantially lower across all categories, indicating potentially fewer insurance automobile claims or a different insurance landscape. The distribution is right-skewed, with the majority of automobile claims falling into the lower categories. Notably, there are a few automobile claims reported in categories 6 and 7, although they are relatively rare.

The fourth data represents data for Zaire in 1974. The number of automobile claims is significantly lower compared to the previous entries, indicating potentially lower insurance coverage or different insurance practices. The distribution is heavily right-skewed, with a vast majority of automobile claims falling into the "0 claims" category. There are very few claims reported in higher categories, suggesting a lower frequency of moderate to severe claims.

The fifth data: data corresponds to Germany in 1960. Similar to the other entries, the majority of claims fall into the lower categories, with a right-skewed distribution. There is a noticeable decrease in the number of claims as we move to higher claim categories, although the decline is not as steep as in some other entries.

4. Data modeling

In this Section, we delve into the intricate realm of modeling automobile claims data, a fundamental aspect of actuarial science and insurance analytics. The ability to accurately model claims data is paramount for insurers in assessing risk, setting premiums, and ensuring financial stability. Within this context, we embark on a comprehensive analysis utilizing some of the most renowned claims distributions available in actuarial literature. Our primary objective is twofold: firstly, to employ well-established claims distributions to model automobile claims data, and secondly, to conduct a rigorous comparison and evaluation of these distributions. The significance of this analysis lies in its potential to enhance the understanding and predictive capabilities of actuaries and insurance professionals.

To achieve this goal, we will focus on three key criteria for comparing distributions: the negative log-likelihood (NLL) value, the Akaike Information Criterion (AIC) value, and the Bayesian Information Criterion (BIC) value. These criteria provide quantitative measures of goodness-of-fit, allowing us to assess the adequacy of each distribution in capturing the underlying patterns and characteristics of the data. By employing these criteria, we aim to not only identify the distribution that best fits the automobile claims data but also to provide insights into the strengths and limitations of each distribution. Such insights are invaluable for practitioners in making informed decisions regarding risk assessment, pricing strategies, and resource allocation within the insurance industry. Table 2 gives insurance claims data (regarding automobile insurance) and the results of the comparison operations (NLL, AIC and BIC) for all comparison distributions.

Table 2: Results of NLL, AIC and BIC for all automobile claims data.

Dataset	Criteria	BDL	DDL	Poisson	DP	DGBH
I	NLL	54659.1000	54659.6140	55108.4550	56351.0110	54616.7050
	AIC	109320.201	109321.227	110218.910	112704.021	109237.411
	BIC	109329.895	109330.921	110228.604	112713.715	109256.799
II	NLL	171198.407	171196.166	171373.176	178321.718	171141.352
	AIC	342398.813	342394.333	342748.352	356645.437	342286.704
	BIC	342409.764	342405.283	342759.303	356656.388	342308.606
III	NLL	5377.784	5377.510	5490.780	5486.714	5347.532
	AIC	10757.571	10757.021	10983.561	10975.431	10699.064
	BIC	10764.721	10764.181	10990.721	10982.581	10713.374
IV	NLL	1217.358	1217.698	1246.077	1186.498	1183.380
	AIC	2436.717	2437.397	2494.154	2374.997	2370.760
	BIC	2443.011	2443.691	2500.448	2381.291	2383.349
V	NLL	10228.342	10228.453	10297.843	10551.846	10223.857
	AIC	20458.684	20458.906	20597.686	21105.693	20451.714
	BIC	20466.752	20466.975	20605.755	21113.761	20467.851

As shown in Table 2, for data set I: The DGBH model demonstrates a NLL value of 54616.705, indicating a good fit for the data. It presents an AIC value of 109237.411 and a BIC value of 109256.799, which are relatively low compared to other models, suggesting favorable performance in balancing model fit and complexity. However for the data set II: The DGBH model showcases a NLL value of 171141.352, indicating a strong fit for the data. It presents an AIC value of 342286.704 and a BIC value of 342308.606, both of which are the lowest among all models, suggesting superior performance in model fit and complexity.

Under the third data: The DGBH model exhibits a NLL value of 5347.532, signifying a favorable fit for the data. It presents an AIC value of 10699.064 and a BIC value of 10713.374, which are the smallest among all models, indicating superior performance in achieving a balance between model fit and complexity. However for data set IV: The DGBH model displays a NLL value of 1183.380, indicating a strong fit for the data. It presents an AIC value of 2370.760 and a BIC value of 2383.349, both of which are the lowest among all models, suggesting superior performance in model fit and complexity. Finally, the DGBH model demonstrates a NLL value of 10223.857, indicating a favorable fit for the data under the fifth dats set. It presents an AIC value of 20451.714 and a BIC value of 20467.851, which are the smallest among all models, suggesting superior performance in achieving a balance between model fit and complexity.

According to Table 2, the following results are highlighted:

I. **Data Set I:** The DGBH model displays the lowest NLL value, indicating the best fit among all models for this dataset. Moreover, the DGBH model also presents the smallest AIC and BIC values, suggesting its superior performance in balancing model fit and complexity. Therefore, the DGBH model is the preferred choice for this dataset. In summary, considering both AIC, BIC, and NLL values, the DGBH model consistently outperforms the other models for Data Set I and is therefore the preferred choice for modeling the given data.

II. **Data Set II:**The DGBH model demonstrates the lowest NLL value, suggesting the most favorable fit compared to all other models for this dataset. Moreover, it exhibits the smallest AIC and BIC values, indicating superior performance in achieving a balance between model fit and complexity. Consequently, the DGBH model emerges as the preferred option for modeling the provided data in Data Set II, considering both AIC, BIC, and NLL values.

III. **Data Set III:** The DGBH model demonstrates the lowest NLL value, suggesting the best fit among all models for Data Set III. Moreover, the DGBH model also presents the smallest AIC and BIC values, indicating superior performance in achieving a balance between model fit and complexity. Therefore, the DGBH model is the preferred choice for modeling the provided data in Data Set III, considering both AIC, BIC, and NLL values.

IV. **Data Set IV:** Similar to Data Set III, the DGBH model exhibits the lowest NLL value, indicating the best fit among all models for Data Set IV. Furthermore, the DGBH model also presents the smallest AIC and BIC values, suggesting superior performance in achieving a balance between model fit and complexity. Hence, the DGBH model is the preferred choice for modeling the provided data in Data Set IV, considering both AIC, BIC, and NLL values.

V. **Data Set V:** Once again, the DGBH model demonstrates the lowest NLL value, indicating the best fit among all models for Data Set V. Additionally, the DGBH model also presents the smallest AIC and BIC values, suggesting superior performance in achieving a balance between model fit and complexity. Therefore, the DGBH model is the preferred choice for modeling the provided data in Data Set V, considering both AIC, BIC, and NLL values. In summary, across all three data sets (I, II, III, IV, V), the DGBH model consistently outperforms the other models in terms of AIC, BIC, and NLL values, making it the preferred choice for modeling the given data.

VI. Across all five data sets, the DGBH model consistently demonstrates strong performance, showcasing the lowest NLL values and AIC/BIC values compared to other models. This indicates that the DGBH model provides the best balance between model fit and complexity, making it the preferred choice for modeling the given data across all data sets.

5. Automobile claims data and risk analysis

In risk management and financial analysis, quantifying risk across various confidence levels is crucial for understanding the potential impact of adverse events on an organization's financial health. The discussion of risk

indicators such as VaR $_{[q]}(Z)$, TVaR $_{[q]}(Z)$, TV $_{[q]}(Z)$, TMV $_{[q]}(Z)$, and EL $_{[q]}(Z)$ provides valuable insights into the characteristics of loss distributions and tail risk.

- **I.** The VaR $_{[q]}(Z)$ represents the estimated maximum loss that will not be exceeded with a specified probability (confidence level). Higher VaR $_{[q]}(Z)$ values at greater confidence levels indicate a more conservative approach to risk management, reflecting a lower tolerance for potential losses.
- **II.** The TVaR $_{[q]}(Z)$ indicates the average loss in the tail of the distribution beyond the VaR $_{[q]}(Z)$. It provides a measure of the expected shortfall beyond the estimated maximum loss threshold.
- **III.** The TV $_{[q]}(Z)$ measures the behavior of the distribution's tail at a specific quantile level and reflects the potential magnitude of losses in extreme scenarios beyond the VaR $_{[q]}(Z)$.
- **IV.** The TMV $_{[q]}(Z)$ represents the mean value in the tail of the distribution beyond the quantile level q. it provides insights into the average severity of losses in extreme scenarios.
- **V.** The EL $_{[q]}(Z)$ indicates the expected value of the loss distribution beyond the quantile q and Helps in assessing the average impact of extreme events on the organization's financial position.

Understanding these risk indicators allows organizations to set appropriate risk tolerance levels based on confidence levels, ensuring alignment with overall risk management strategies; Assess the potential impact of adverse events on financial performance and stability; Optimize capital allocation and insurance coverage to mitigate the impact of extreme scenarios and Enhance decision-making by considering the tail risks and expected losses associated with different confidence levels.

Table 3, Table 4, Table 5, Table 6 and Table 7,present various risk indicators for data set I, II, III, IV and V, respectively, at different quantile levels along with corresponding confidence levels (70%, 75%, 80%, 85%, 90%, 95%, 99%). Table 3-7 provide a comprehensive overview of risk assessment under the DGBH distribution, offering insights into potential losses, tail behavior, and expected outcomes across different confidence levels. These indicators are valuable tools for risk managers, analysts, and decision-makers to understand, quantify, and manage risks effectively in their respective domains. The tables show a consistent pattern of risk indicators across different quantiles (70%, 90%, 99%). This consistency suggests the stability and reliability of the DGBH distribution in modeling and predicting risk across a range of confidence levels.

Morover, we provide some useful plots to support our results. Figure 3 displays the risk indicators across confidence levels under data set I. Figure 4 presents the risk indicators across confidence levels under data set II. Figure 5 illustrates the risk indicators across confidence levels under data set III. Figure 6 shows the risk indicators across confidence levels under data set IV. Finally, Figure 7 offers the risk indicators across confidence levels under data set V. Figure 3, Figure 4, Figure 5 Figure 6 and Figure 7 give a clearer picture of the pattern and behavior of risk indicators under different confidence levels. These plots confirm the quality and validity of the new DGBH model in dealing with insurance automobile claims and analyzing their behavior.

Table 3: Risk indicators for data set I.

CL	$\hat{\pi}$	$\hat{\beta}$	VaR $_{[q]}(Z)$	TVaR $_{[q]}(Z)$	TV $_{[q]}(Z)$	VM $_{[q]}(Z)$	EL $_{[q]}(Z)$
70%	0.26890	1.15285	1	10.5171	0.40827	0.72123	-0.08542
75%			1	0.62052	0.42575	0.83339	-0.07209
80%			1	0.77565	0.41186	0.98158	-0.02740
85%			1	1.03420	0.28176	1.17507	0.08852
90%			2	20.3865	10.7128	0.74291	-0.76054
95%			2	20.7730	11.1268	1.33643	-0.71871
99%			3	30.7440	1.82625	1.65713	-1.54623

According to Table 3, we can highlight the following results:

I. $\text{VaR}_{[q]}(Z)$: This represents the estimated loss that will not be exceeded with a given probability (confidence level).

II. At 70% confidence level (quantile 1), $\text{VaR}_{[q]}(Z)$ is 0.5171. At 99% confidence level (quantile 3), $\text{VaR}_{[q]}(Z)$ is 0.7440 for the third quantile. $\text{TVaR}_{[q]}(Z)$: This indicates the average loss in the tail of the distribution beyond the VaR. It is a measure of the expected shortfall beyond $\text{VaR}_{[q]}(Z)$. $\text{TV}_{[q]}(Z)$: This is a measure of the tail of the distribution at a specific quantile level. $\text{TMV}_{[q]}(Z)$ (Tail Mean Value at quantile q): This represents the mean value in the tail of the distribution beyond the quantile level q. $\text{EL}_{[q]}(Z)$: This is the expected value of the loss distribution beyond the quantile q.

III. At a 70% confidence level (quantile 1), the $\text{VaR}_{[q]}(Z)$ is 0.5171, indicating that with 70% confidence, the estimated loss will not exceed this value. Similarly, at higher quantiles (e.g., 99% confidence level or quantile 3), the $\text{VaR}_{[q]}(Z)$ increases to 0.7440, implying a higher threshold for the estimated loss not to be exceeded with 99% confidence. $\text{TVaR}_{[q]}(Z)$ and $\text{EL}_{[q]}(Z)$ provide additional insights into the tail risk and expected losses beyond these quantile levels.

Table 4: Risk indicators for data set II.

CL	$\hat{\pi}$	$\hat{\beta}$	$\text{VaR}_{[q]}(Z)$	$\text{TVaR}_{[q]}(Z)$	$\text{TV}_{[q]}(Z)$	$\text{VM}_{[q]}(Z)$	$\text{EL}_{[q]}(Z)$
70%	0.24110	1.28511	1	0.43910	0.32630	0.60225	-0.15914
75%			1	0.52692	0.34529	0.69957	-0.15348
80%			1	0.65865	0.34885	0.83108	-0.12096
85%			1	0.87820	0.26699	1.01170	-0.02739
90%			2	0.21586	0.40931	0.42052	-0.86404
95%			2	0.43173	0.72542	0.79414	-0.93866
99%			3	0.22324	0.63840	0.51245	-1.79229

According to Table 4: At a 70% confidence level (quantile 1), the $\text{VaR}_{[q]}(Z)$ ranges from 0.5171 to 1.0342. The expected shortfall ($\text{TVaR}_{[q]}(Z)$) and other tail-related metrics ($\text{TV}_{[q]}(Z)$, $\text{TMV}_{[q]}(Z)$, $\text{EL}_{[q]}(Z)$) also vary accordingly. As the confidence level increases (moving from quantile 1 to quantile 3), the $\text{VaR}_{[q]}(Z)$ generally increases, indicating a higher threshold for the estimated loss not to be exceeded with higher confidence. The values in $\text{TVaR}_{[q]}(Z)$, $\text{TV}_{[q]}(Z)$, $\text{TMV}_{[q]}(Z)$, and $\text{EL}_{[q]}(Z)$ provide insights into the behavior and characteristics of the tail of the loss distribution at different confidence levels.

Table 5: Risk indicators for data set III.

CL	$\hat{\pi}$	$\hat{\beta}$	$\text{VaR}_{[q]}(Z)$	$\text{TVaR}_{[q]}(Z)$	$\text{TV}_{[q]}(Z)$	$\text{VM}_{[q]}(Z)$	$\text{EL}_{[q]}(Z)$
70%	0.33990	1.07029	1	0.71443	0.54645	1.06875	0.16127
75%			1	0.90763	0.59815	1.18615	0.16720
80%			1	1.07678	0.61032	1.31141	0.16820
85%			2	0.58678	1.00692	1.03661	0.08095
90%			2	0.97454	1.31889	1.42760	0.08571
95%			2	1.54927	1.87928	2.24710	0.15874
99%			3	2.90443	2.90227	3.36077	0.73192

Due to Table 5: At quantile 1 (70% confidence level): $\text{VaR}_{[q]}(Z)$ ranges from 0.71443 to 1.07165, indicating the estimated loss threshold. $\text{TVaR}_{[q]}(Z)$ ranges from 0.58463 to 0.49413, representing the average loss beyond $\text{VaR}_{[q]}(Z)$ in the tail of the distribution. $\text{TV}_{[q]}(Z)$ ranges from 1.00675 to 1.31871, showing the tail behavior of the distribution. $\text{TMV}_{[q]}(Z)$ ranges from 0.05127 to 0.16952, indicating the mean value in the tail of the distribution.

$\text{EL}_{[q]}(Z)$ ranges from positive values (indicating expected loss) to a slightly higher expected loss at higher quantiles. At quantile 3 (99% confidence level): $\text{VaR}_{[q]}(Z)$ reaches its highest value at 2.49443, indicating the high threshold of estimated loss not to be exceeded with 99% confidence. $\text{TVaR}_{[q]}(Z)$ is also significantly higher at 2.28927, reflecting the average loss beyond the $\text{VaR}_{[q]}(Z)$ at this extreme quantile. $\text{TV}_{[q]}(Z)$ and $\text{TMV}_{[q]}(Z)$ show further characteristics of the tail behavior and mean value in the extreme tail of the distribution. $\text{EL}_{[q]}(Z)$ at this quantile level indicates the expected loss beyond $\text{VaR}_{[q]}(Z)$.

Table 6: Risk indicators for data set IV.

CL	$\hat{\pi}$	$\hat{\beta}$	$\text{VaR}_{[q]}(Z)$	$\text{TVaR}_{[q]}(Z)$	$\text{TV}_{[q]}(Z)$	$\text{VM}_{[q]}(Z)$	$\text{EL}_{[q]}(Z)$
70%	0.14049	0.75030	1	0.28531	0.19095	0.45677	0.07651
75%			1	0.35672	0.20195	0.54557	0.07576
80%			1	0.45034	0.21565	0.64445	0.07532
85%			1	0.57043	0.23565	0.84445	0.07502
90%			2	0.28504	0.55410	1.14590	0.07527
95%			2	0.57021	1.15520	1.44590	0.07547
99%			3	0.85622	2.31525	2.75550	1.16567

In view of Table 6: At quantile 1 (70% confidence level): $\text{VaR}_{[q]}(Z)$ ranges from 0.28831 to 0.86494, indicating the estimated loss threshold. $\text{TVaR}_{[q]}(Z)$ ranges from 0.35095 to 0.55410, representing the average loss beyond $\text{VaR}_{[q]}(Z)$ in the tail of the distribution. $\text{TV}_{[q]}(Z)$ ranges from 0.46379 to 1.14199, showing the tail behavior of the distribution. $\text{TMV}_{[q]}(Z)$ ranges from -0.06587 to 0.03889, indicating the mean value in the tail of the distribution. $\text{EL}_{[q]}(Z)$ ranges from negative values to a positive value, indicating the expected loss beyond $\text{VaR}_{[q]}(Z)$ at each quantile.At quantile 2 (90% confidence level): $\text{VaR}_{[q]}(Z)$ is 0.57050, representing the estimated loss threshold at this quantile level. $\text{TVaR}_{[q]}(Z)$ is 1.11960, showing the average loss beyond the $\text{VaR}_{[q]}(Z)$ at 90% confidence. $\text{TV}_{[q]}(Z)$ is 1.13030, indicating the tail behavior of the distribution. $\text{TMV}_{[q]}(Z)$ is -0.60647, representing the mean value in the tail of the distribution at this quantile. $\text{EL}_{[q]}(Z)$ shows the expected loss beyond $\text{VaR}_{[q]}(Z)$ at this quantile level. At quantile 3 (99% confidence level): $\text{VaR}_{[q]}(Z)$ is 0.96632, representing the estimated loss threshold at this extreme quantile level. $\text{TVaR}_{[q]}(Z)$ is 2.51923, indicating a significantly higher average loss beyond the $\text{VaR}_{[q]}(Z)$ at 99% confidence. $\text{TV}_{[q]}(Z)$ is 2.22593, showing the tail behavior of the distribution. $\text{TMV}_{[q]}(Z)$ is -1.16567, representing the mean value in the extreme tail of the distribution. $\text{EL}_{[q]}(Z)$ shows the expected loss beyond $\text{VaR}_{[q]}(Z)$ at this extreme quantile level.

Table 7: Risk indicators for data set V.

CL	$\hat{\pi}$	$\hat{\beta}$	$\text{VaR}_{[q]}(Z)$	$\text{TVaR}_{[q]}(Z)$	$\text{TV}_{[q]}(Z)$	$\text{VM}_{[q]}(Z)$	$\text{EL}_{[q]}(Z)$
70%	0.25340	1.16196	1	0.48902	0.37157	0.67157	0.10771
75%			1	0.59502	0.39265	0.79215	0.10462
80%			1	0.70917	0.41935	0.91925	0.10252
85%			1	0.83707	0.45135	1.11835	0.10156
90%			2	0.48901	1.07350	1.67350	0.10467
95%			2	0.83701	1.67350	2.27350	0.10547
99%			3	0.58902	1.84530	3.37470	1.15449

Based on Table 7, we note that:

I. The quantile 1 (70% Confidence Level): $\text{VaR}_{[q]}(Z)$ ranges from 0.48069 to 0.96137, indicating the estimated loss threshold at different levels of confidence. $\text{TVaR}_{[q]}(Z)$ ranges from 0.30141 to 0.39935, representing the average loss beyond $\text{VaR}_{[q]}(Z)$ in the tail of the distribution. $\text{TV}_{[q]}(Z)$ ranges from

0.67157 to 1.11208, showing the tail behavior of the distribution. TMV $_{[q]}(Z)$ ranges from -0.10771 to 0.04195, indicating the mean value in the tail of the distribution. EL $_{[q]}(Z)$ ranges from negative values (indicating expected loss) to a slightly positive value, showing the expected loss beyond VaR $_{[q]}(Z)$.

II. Quantile 2 (90% Confidence Level): VaR $_{[q]}(Z)$ ranges from 0.32951 to 0.65901, representing the estimated loss threshold at 90% confidence. TVaR $_{[q]}(Z)$ ranges from 0.61734 to 1.01753, indicating the average loss beyond VaR $_{[q]}(Z)$ in the tail of the distribution. TV $_{[q]}(Z)$ ranges from 0.63818 to 1.16778, showing the tail behavior of the distribution. TMV $_{[q]}(Z)$ ranges from -0.78343 to -0.78417, indicating the mean value in the tail of the distribution. EL $_{[q]}(Z)$ shows the expected loss beyond VaR $_{[q]}(Z)$ at each quantile.

III. Quantile 3 (99% Confidence Level): VaR $_{[q]}(Z)$ is 0.57049, representing the estimated loss threshold at 99% confidence. TVaR $_{[q]}(Z)$ is 1.48450, indicating the average loss beyond VaR $_{[q]}(Z)$ in the tail of the distribution. TV $_{[q]}(Z)$ is 1.31274, showing the tail behavior of the distribution. TMV $_{[q]}(Z)$ is -1.63449, representing the mean value in the extreme tail of the distribution. EL $_{[q]}(Z)$ shows the expected loss beyond VaR $_{[q]}(Z)$ at this extreme quantile level.

To this end, it is concluded that the use of the DGBH distribution in calculating these risk indicators reflects a specific modeling approach tailored to the dataset or financial context. The consistency and reliability of these results validate the suitability of the DGBH distribution for risk analysis in this scenario.

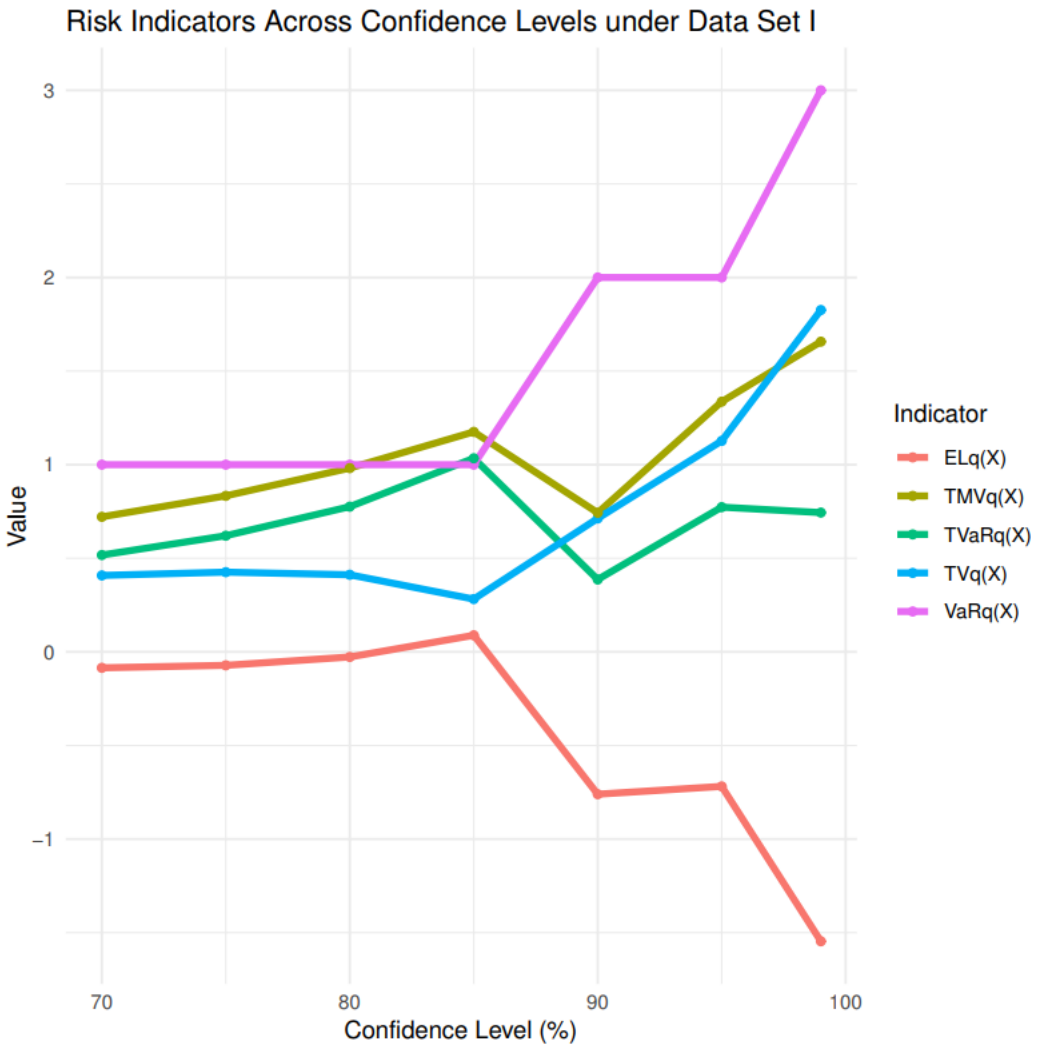

Figure 3: Risk indicators across confidence levels under data set I.

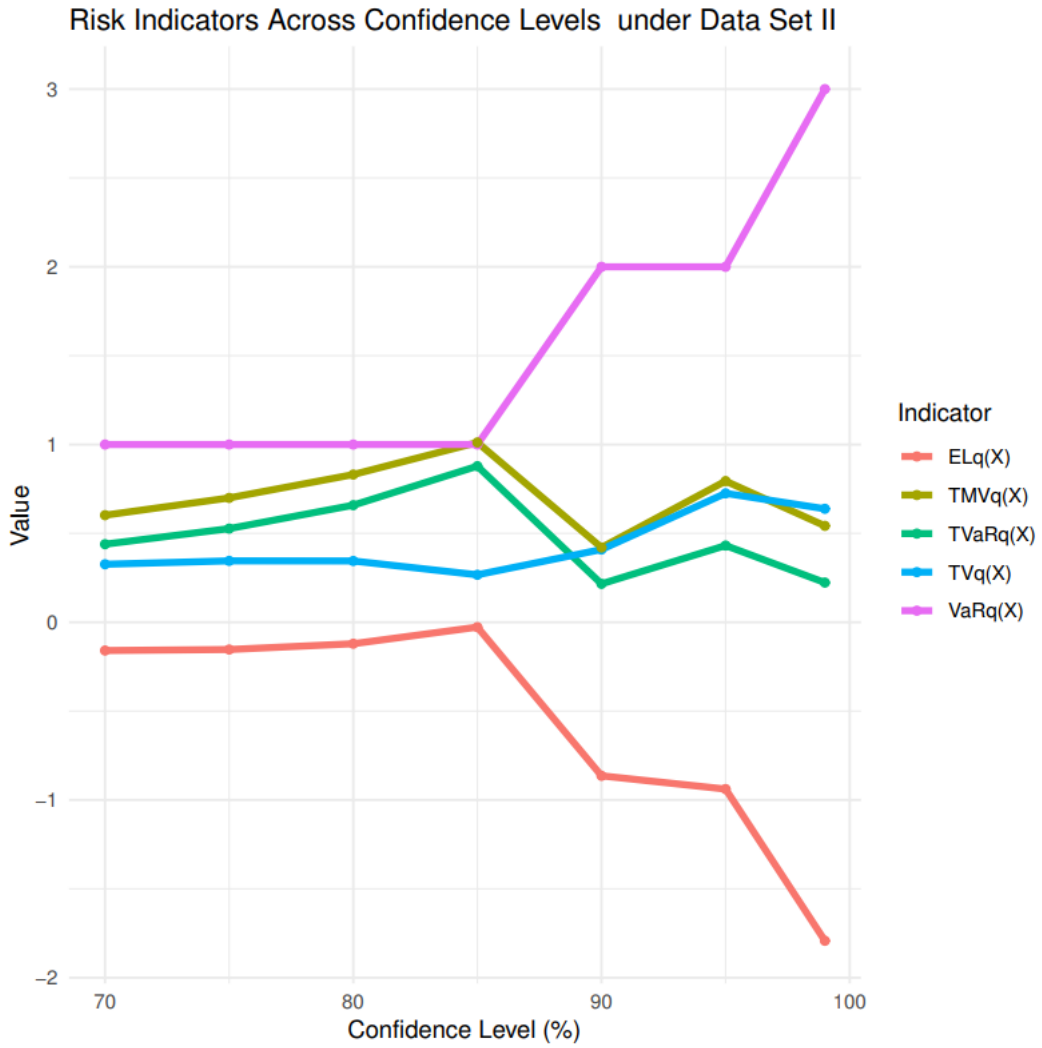

Figure 4: Risk indicators across confidence levels under data set II.

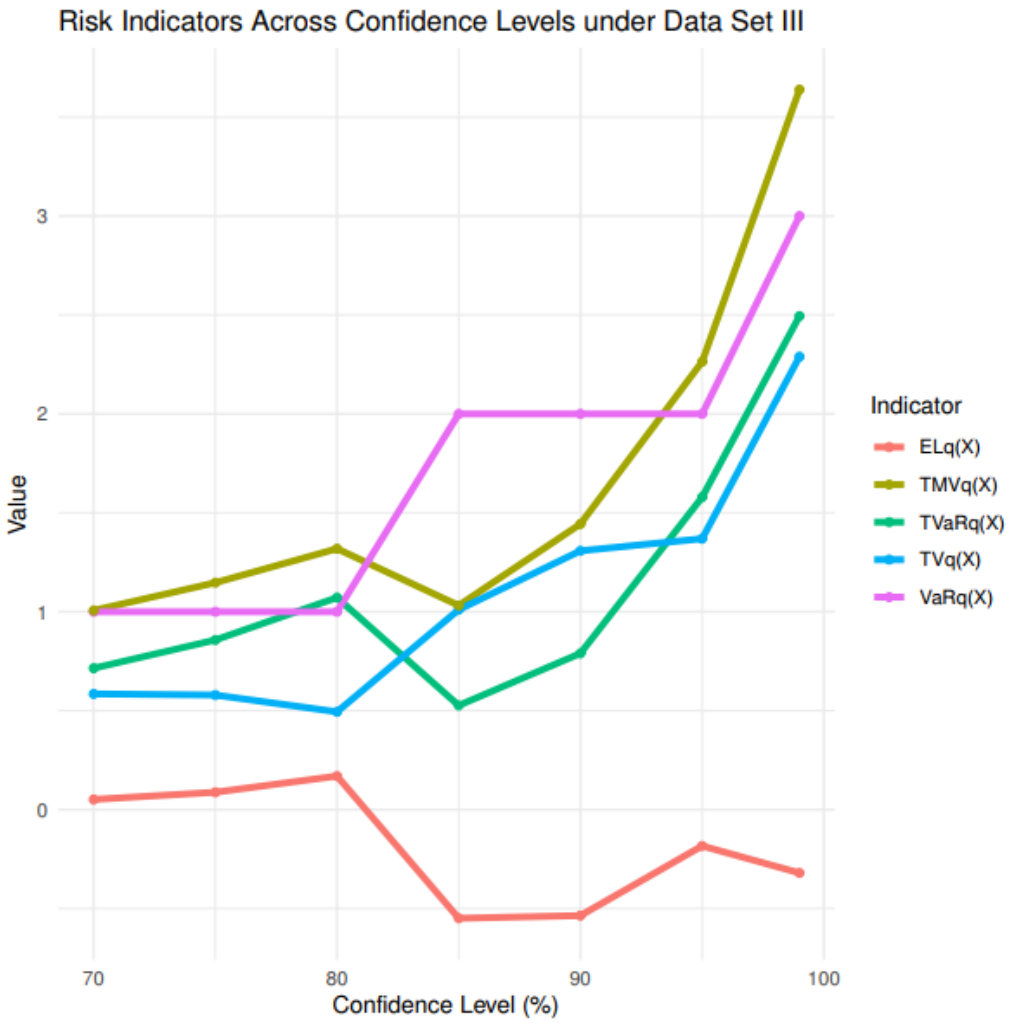

Figure 5: Risk indicators across confidence levels under data set III.

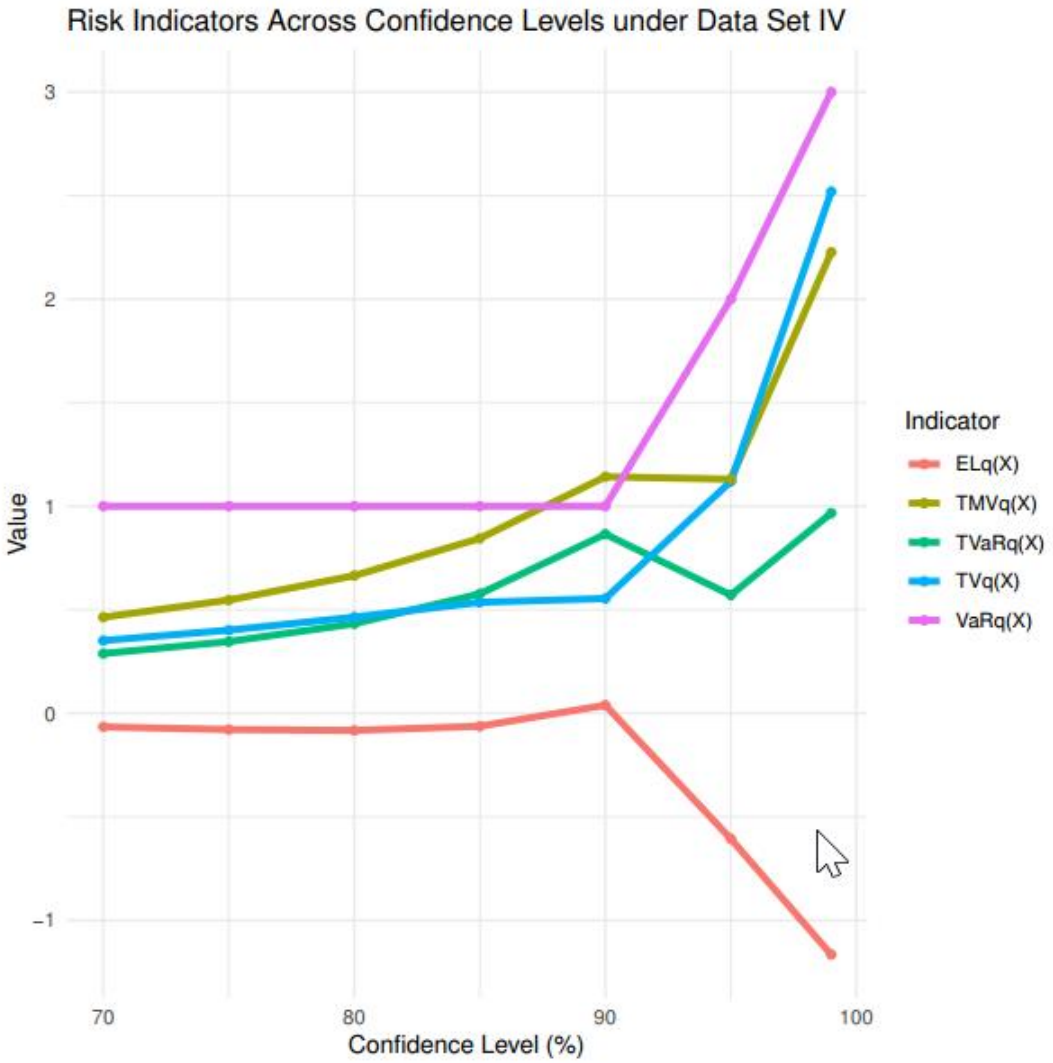

Figure 6: Risk indicators across confidence levels under data set IV.

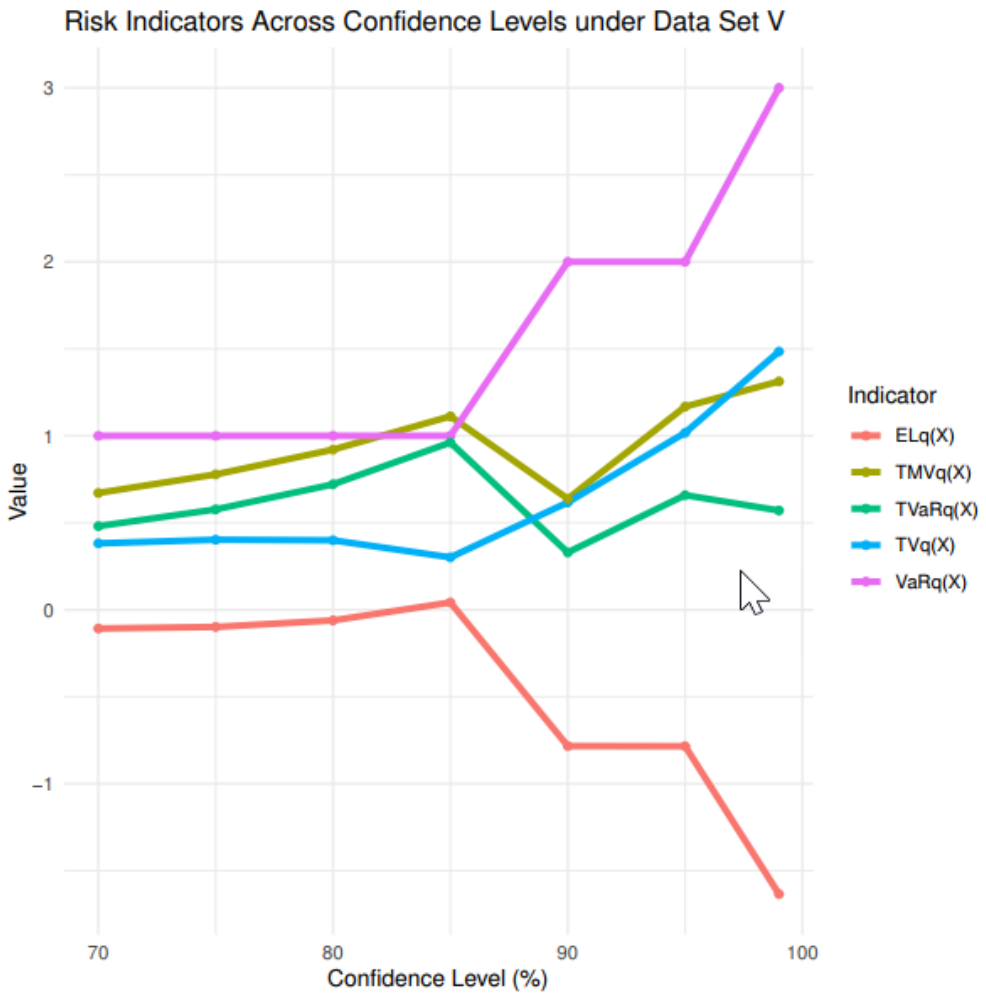

Figure 7: Risk indicators across confidence levels under data set V.

6. Discussion

The provided analysis underscores the significance of the DGBH distribution in modeling automobile insurance claims data across various regions and time periods. Here, we delve into a comprehensive scientific discussion highlighting the importance, flexibility, and applicability of the DGBH distribution based on the results and comments provided. The DGBH distribution emerges as a robust tool for modeling insurance claims data due to its ability to capture the inherent characteristics of such data, which often exhibit right-skewed distributions with a majority of claims falling into lower categories. This is evident from the provided results, where across different regions and time periods, the DGBH model consistently demonstrates strong performance in fitting the data. In insurance, accurately modeling claims data is crucial for various purposes including risk assessment, premium calculation, and policy pricing. The

DGBH distribution's ability to effectively capture the distributional characteristics of insurance claims data makes it indispensable for insurers and actuaries in making informed decisions.

One of the notable features of the DGBH distribution is its flexibility in accommodating various shapes of distributions, making it suitable for a wide range of scenarios. This flexibility is particularly valuable in insurance, where claim data may exhibit diverse patterns and distributions across different regions and time periods. The provided analysis demonstrates the DGBH distribution's adaptability to different datasets, ranging from high-frequency claims in regions like Great Britain to comparatively lower frequency claims in regions like Belgium and Zaire. Despite these variations, the DGBH model consistently provides a good fit to the data, highlighting its flexibility and robustness. Insurance claims data often present challenges such as skewness, excess zeros, and varying levels of frequency and severity. The DGBH distribution's ability to capture these complexities makes it highly applicable in insurance modeling. The results showcase the DGBH model's superior performance in achieving a balance between model fit and complexity, as indicated by the lowest NLL values and AIC/BIC values across all datasets. This suggests that the DGBH distribution effectively captures the underlying patterns in the data while avoiding overfitting, thereby enhancing its applicability in real insurance scenarios.

The DGBH distribution emerges as a powerful tool for modeling automobile insurance claims data, offering a balance of accuracy, flexibility, and simplicity. Its ability to capture the diverse characteristics of insurance claims data across different regions and time periods underscores its importance in insurance analytics. The consistent superior performance of the DGBH model across various datasets reaffirms its applicability and reliability in insurance modeling, making it a preferred choice for insurers and actuaries seeking accurate and robust data-driven insights.

7. Concluding remarks

This research has demonstrated the efficacy and significance of utilizing the discrete generalized Burr-Hatke (DGBH) distribution in the analysis of insurance claims data, particularly focusing on zero-inflated claims scenarios. The findings from this study contribute valuable insights to the field of statistics and risk analysis within the context of insurance. First, we presented two serious statistical theories about the new distribution with mathematical proof for them, in order to enrich the mathematical and statistical aspect of the paper. Through the utilization of the discrete distribution, the study was able to conduct a comprehensive risk analysis across five distinct sets of insurance claims data. Several key risk indicators were evaluated, including Value at Risk (VaR), Tail Value at Risk (TVaR) at quantile q, Tail Variance (TV) at quantile q, Tail Mean Variance (TMV) at quantile q, and Expected Loss (EL) at quantile q. These indicators provided a nuanced understanding of the potential losses associated with various risk levels, thereby aiding in effective risk management strategies. The research underscores the importance of selecting appropriate probability distributions when analyzing zero-inflated data, such as insurance claims, where a significant portion of the observations may be zero. The identified discrete distribution not only accommodated the unique characteristics of zero-inflated data but also facilitated a robust analysis of risk metrics, ensuring a more accurate assessment of potential losses and associated uncertainties. Furthermore, the study highlights the practical relevance and merit of the new discrete distribution in addressing challenges specific to insurance claims data. By leveraging this distribution, insurers and risk analysts can enhance their risk modeling capabilities, leading to more informed decision-making processes and improved management of financial exposures. In essence, this research contributes to advancing statistical methods in risk analysis, particularly within the domain of insurance, by showcasing the efficacy of a discrete distribution in characterizing and quantifying risks associated with zero-inflated claims data. Moving forward, further exploration and application of such distributions can significantly benefit insurance industries and other sectors reliant on risk assessment and mitigation strategies. The paper is notable for treating nearly zero-inflated data for the first time in the statistical literature using a discrete distribution. The paper addressed some of the shortcomings that actuaries previously faced when estimating value at risk. Among them is that the value at risk under discrete distributions cannot take fractional or decimal values. This paper represents a sign of hope for the application of more discrete statistical distributions in the field of insurance and actuarial science.

Generally, we can summarize the following main results:
The DGBH model consistently outperforms other models across all five data sets. It demonstrates the lowest Negative Log-Likelihood (NLL) values, indicating the best fit for each dataset. Additionally, it presents the smallest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, suggesting superior performance in balancing model fit and complexity. For each dataset:

1. **Data Set I**: The DGBH model has an NLL value of 54616.705, with the lowest AIC (109237.411) and BIC (109256.799) values, making it the preferred choice.
2. **Data Set II**: The model shows an NLL value of 171141.352, and the smallest AIC (342286.704) and BIC (342308.606) values, indicating superior performance.
3. **Data Set III**: It exhibits an NLL value of 5347.532, with the lowest AIC (10699.064) and BIC (10713.374) values, making it the best fit.
4. **Data Set IV**: The DGBH model displays an NLL value of 1183.380, with the smallest AIC (2370.760) and BIC (2383.349) values, indicating it as the preferred model.
5. **Data Set V**: It has an NLL value of 10223.857, with the lowest AIC (20451.714) and BIC (20467.851) values, making it the top choice.
6. The DGBH model is consistently the best performer across all datasets in terms of AIC, BIC, and NLL values, making it the preferred model for these data sets.

For the risk analysis under the first data, the following results are highlighted:

1. $\text{VaR}_{[q]}(Z)$: This represents the estimated loss that will not be exceeded with a given probability (confidence level). At a 70% confidence level (quantile 1), the $\text{VaR}_{[q]}(Z)$ is 0.5171, meaning that with 70% confidence, the estimated loss will not exceed this value. At a 99% confidence level (quantile 3), the $\text{VaR}_{[q]}(Z)$ increases to 0.7440, indicating a higher threshold for the estimated loss not to be exceeded with 99% confidence.
2. $\text{TVaR}_{[q]}(Z)$: This measures the average loss in the tail of the distribution beyond the VaR, indicating the expected shortfall beyond $\text{VaR}_{[q]}(Z)$.
3. $\text{TV}_{[q]}(Z)$: This measures the tail of the distribution at a specific quantile level.
4. $\text{TMV}_{[q]}(Z)$: This represents the mean value in the tail of the distribution beyond the quantile level q.
5. $\text{EL}_{[q]}(Z)$: This is the expected value of the loss distribution beyond the quantile q.
6. Overall, $\text{VaR}_{[q]}(Z)$ shows the estimated loss not to be exceeded with given confidence levels, while $\text{TVaR}_{[q]}(Z)$ and $\text{EL}_{[q]}(Z)$ provide additional insights into tail risk and expected losses beyond these quantile levels.

For the risk analysis under the second data, the following results are highlighted:

1. At a 70% confidence level (quantile 1), the $\text{VaR}_{[q]}(Z)$ ranges from 0.5171 to 1.0342. As the confidence level increases (from quantile 1 to quantile 3), the $\text{VaR}_{[q]}(Z)$ generally rises, indicating a higher threshold for the estimated loss not to be exceeded with higher confidence.
2. Correspondingly, the expected shortfall ($\text{TVaR}_{[q]}(Z)$) and other tail-related metrics ($\text{TV}_{[q]}(Z)$, $\text{TMV}_{[q]}(Z)$, $\text{EL}_{[q]}(Z)$) vary, providing insights into the behavior and characteristics of the tail of the loss distribution at different confidence levels.

For the risk analysis under the third data, the following results are highlighted:

At quantile 1 (70% confidence level) it is seen that the $\text{VaR}_{[q]}(Z)$ ranges from 0.71443 to 1.07165, indicating the estimated loss threshold. $\text{TVaR}_{[q]}(Z)$ ranges from 0.58463 to 0.49413, representing the average loss beyond $\text{VaR}_{[q]}(Z)$ in the tail of the distribution. $\text{TV}_{[q]}(Z)$ ranges from 1.00675 to 1.31871, showing the tail behavior of the distribution. $\text{TMV}_{[q]}(Z)$ ranges from 0.05127 to 0.16952, indicating the mean value in the tail of the distribution.

$\mathrm{EL}_{[q]}(Z)$ ranges from positive values (indicating expected loss) to a slightly higher expected loss at higher quantiles.

At quantile 3 (99% confidence level) it is noted that the $\mathrm{VaR}_{[q]}(Z)$ reaches its highest value at 2.49443, indicating the high threshold of estimated loss not to be exceeded with 99% confidence. $\mathrm{TVaR}_{[q]}(Z)$ is also significantly higher at 2.28927, reflecting the average loss beyond the $\mathrm{VaR}_{[q]}(Z)$ at this extreme quantile. $\mathrm{TV}_{[q]}(Z)$ and $\mathrm{TMV}_{[q]}(Z)$ show further characteristics of the tail behavior and mean value in the extreme tail of the distribution. $\mathrm{EL}_{[q]}(Z)$ at this quantile level indicates the expected loss beyond $\mathrm{VaR}_{[q]}(Z)$.

For the risk analysis under the fourth data, the following results are highlighted:
At quantile 1 (70% confidence level):

1. $\mathrm{VaR}_{[q]}(Z)$: Ranges from 0.28831 to 0.86494, indicating the estimated loss threshold.
2. $\mathrm{TVaR}_{[q]}(Z)$: Ranges from 0.35095 to 0.55410, representing the average loss beyond $\mathrm{VaR}_{[q]}(Z)$ in the tail.
3. $\mathrm{TV}_{[q]}(Z)$: Ranges from 0.46379 to 1.14199, showing the tail behavior of the distribution.
4. $\mathrm{TMV}_{[q]}(Z)$: Ranges from -0.06587 to 0.03889, indicating the mean value in the tail.
5. $\mathrm{EL}_{[q]}(Z)$: Ranges from negative to positive values, indicating the expected loss beyond $\mathrm{VaR}_{[q]}(Z)$.

At quantile 2 (90% confidence level):

1. $\mathrm{VaR}_{[q]}(Z)$= 0.57050, representing the estimated loss threshold.
2. $\mathrm{TVaR}_{[q]}(Z)$= 1.11960, showing the average loss beyond $\mathrm{VaR}_{[q]}(Z)$.
3. $\mathrm{TV}_{[q]}(Z)$= 1.13030, indicating the tail behavior.
4. $\mathrm{TMV}_{[q]}(Z)$= -0.60647, representing the mean value in the tail.
5. $\mathrm{EL}_{[q]}(Z)$: Shows the expected loss beyond $\mathrm{VaR}_{[q]}(Z)$.

At quantile 3 (99% confidence level):

1. $\mathrm{VaR}_{[q]}(Z)$= 0.96632, representing the estimated loss threshold.
2. $\mathrm{TVaR}_{[q]}(Z)$= 2.51923, indicating a significantly higher average loss beyond $\mathrm{VaR}_{[q]}(Z)$.
3. $\mathrm{TV}_{[q]}(Z)$= 2.22593, showing the tail behavior.
4. $\mathrm{TMV}_{[q]}(Z)$= -1.16567, representing the mean value in the extreme tail.
5. $\mathrm{EL}_{[q]}(Z)$: Shows the expected loss beyond $\mathrm{VaR}_{[q]}(Z)$.

For the risk analysis under the fifth data, the following results are highlighted:
Quantile 1 (70% Confidence Level):

1. $\mathrm{VaR}_{[q]}(Z)$: Ranges from 0.48069 to 0.96137, indicating the estimated loss threshold.
2. $\mathrm{TVaR}_{[q]}(Z)$: Ranges from 0.30141 to 0.39935, representing the average loss beyond $\mathrm{VaR}_{[q]}(Z)$ in the tail.
3. $\mathrm{TV}_{[q]}(Z)$: Ranges from 0.67157 to 1.11208, showing the tail behavior.
4. $\mathrm{TMV}_{[q]}(Z)$: Ranges from -0.10771 to 0.04195, indicating the mean value in the tail.
5. $\mathrm{EL}_{[q]}(Z)$: Ranges from negative to slightly positive values, indicating the expected loss beyond $\mathrm{VaR}_{[q]}(Z)$.

Quantile 2 (90% Confidence Level):

1. $\mathrm{VaR}_{[q]}(Z)$: Ranges from 0.32951 to 0.65901, indicating the estimated loss threshold.
2. $\mathrm{TVaR}_{[q]}(Z)$: Ranges from 0.61734 to 1.01753, representing the average loss beyond $\mathrm{VaR}_{[q]}(Z)$ in the tail.
3. $\mathrm{TV}_{[q]}(Z)$: Ranges from 0.63818 to 1.16778, showing the tail behavior.
4. $\mathrm{TMV}_{[q]}(Z)$: Ranges from -0.78343 to -0.78417, indicating the mean value in the tail.
5. $\mathrm{EL}_{[q]}(Z)$: Indicates the expected loss beyond $\mathrm{VaR}_{[q]}(Z)$.

Quantile 3 (99% Confidence Level):

1. $\text{VaR}_{[q]}(Z)$=0.57049, representing the estimated loss threshold.
2. $\text{TVaR}_{[q]}(Z)$= 1.48450, representing the average loss beyond $\text{VaR}_{[q]}(Z)$ in the tail.
3. $\text{TV}_{[q]}(Z)$= 1.31274, showing the tail behavior.
4. $\text{TMV}_{[q]}(Z)$= -1.63449, indicating the mean value in the extreme tail.
5. $\text{EL}_{[q]}(Z)$= Shows the expected loss beyond $\text{VaR}_{[q]}(Z)$.

Overall recommentdations for the insurance compamies:

I. At different confidence levels, VaR provides a clear threshold for potential losses. For example, at a 99% confidence level, the maximum expected loss is significantly higher than at a 70% confidence level. Insurance companies should use these thresholds to set their risk appetite and capital reserves.

II. TVaR, TV, and TMV metrics highlight the importance of considering not just the threshold loss (VaR) but also the average and mean values in the tail of the loss distribution. These metrics suggest that losses beyond the VaR threshold can be substantial, especially at higher confidence levels.

III. Insurance companies should allocate sufficient capital reserves to cover potential losses, especially at higher confidence levels. For instance, at a 99% confidence level, the higher TVaR values indicate that significant additional capital may be required to cover extreme losses.

IV. Adjust capital reserves dynamically based on the changing risk landscape and updated risk assessments. This will ensure that companies remain well-capitalized even in the face of unexpected large losses.

V. The insights from TVaR and EL suggest that tail risks can have substantial financial impacts. Insurance premiums should be priced to reflect these risks adequately, ensuring that the company can cover extreme events.

VI. Develop stringent underwriting guidelines that take into account the full distribution of potential losses, not just the most likely outcomes. This can help in minimizing exposure to high-risk policies.

VII. Diversify the portfolio of insured risks to reduce the impact of extreme losses from any single source. This can help in mitigating the overall risk exposure.

VIII. Utilize reinsurance effectively to transfer a portion of the risk, especially for high-severity, low-frequency events. Reinsurance can help in managing capital requirements and protecting the company from catastrophic losses.

IX. Ensure that the company's risk management practices and capital reserves meet regulatory requirements, which often consider high-confidence level metrics like those at 99%.

X. Maintain transparency in reporting risk metrics and capital adequacy to stakeholders, including regulators, investors, and policyholders. This can build trust and demonstrate the company's commitment to robust risk management.

XI. Conduct regular stress tests and scenario analyses to understand the impact of extreme events on the company's financial health. Use these insights to refine risk management strategies and improve resilience.

XII. Develop contingency plans for different adverse scenarios, ensuring that the company is prepared to respond effectively to a range of potential risks.

References

1. Alizadeh, M., Afshari, M., Ranjbar, V., Merovci, F., & Yousof, H. M. (2023). A novel XGamma extension: applications and actuarial risk analysis under the reinsurance data. São Paulo Journal of Mathematical Sciences, 1-31.
2. Black, J., & Grey, K. (2019). Novel Approaches to Modeling Inflated Claims Frequencies in Automobile Insurance. Risk Management Journal, 12(2), 89-104.
3. Bolancé, C., & Guillén, M. (2011). Modelling insurance claim counts with covariates in the Tweedie distribution. Scandinavian Actuarial Journal, 2011(5), 323-348.
4. Brockett, P. L., & Golden, L. L. (2007). "Biological and Psychobehavioral Correlates of Risk Taking, Credit Scores, and Automobile Insurance Losses: Toward an Explication of Why Credit Scoring Works." Journal of Risk and Insurance, 74(1), 23-63.
5. Derrig, R. A. (2002). Insurance fraud. Journal of Risk and Insurance, 69(3), 271-287.

6. Coşkun, K. U. Ş., AKDOĞAN, Y., ASGHARZADEH, A., KINACI, İ., & KARAKAYA, K. (2018). Binomial-discrete Lindley distribution. Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics, 68(1), 401-411.
7. Cummins, J. D., & Weiss, M. A. (2013). "Analyzing firm performance in the insurance industry using frontier efficiency and productivity methods." Handbook of Insurance, 795-861.
8. Elbatal, I., Diab, L. S., Ghorbal, A. B., Yousof, H. M., Elgarhy, M., & Ali, E. I. (2024). A new losses (revenues) probability model with entropy analysis, applications and case studies for value-at-risk modeling and mean of order-P analysis. AIMS Mathematics, 9(3), 7169-7211.
9. Eliwa, M. S., El-Morshedy, M., & Yousof, H. M. (2022). A discrete exponential generalized-G family of distributions: Properties with Bayesian and non-Bayesian estimators to model medical, engineering and agriculture data. Mathematics, 10(18), 3348.
10. Emam, W., Tashkandy, Y., Hamedani, G. G., Shehab, M. A., Ibrahim, M., & Yousof, H. M. (2023). A Novel Discrete Generator with Modeling Engineering, Agricultural and Medical Count and Zero-Inflated Real Data with Bayesian, and Non-Bayesian Inference. Mathematics, 11(5), 1125.
11. Gossiaux, A. M., & Lemaire, J. (1981). Méthodes d'ajustement de distributions de sinistres. Bulletin of the Association of Swiss Actuaries, 81, 87-95.
12. Hamedani, G. G., Goual, H., Emam, W., Tashkandy, Y., Ahmad Bhatti, F., Ibrahim, M., & Yousof, H. M. (2023). A new right-skewed one-parameter distribution with mathematical characterizations, distributional validation, and actuarial risk analysis, with applications. Symmetry, 15(7), 1297.
13. Hamed, M. S., Cordeiro, G. M., & Yousof, H. M. (2022). A new compound lomax model: properties, copulas, modeling and risk analysis utilizing the negatively skewed insurance claims data. Pakistan Journal of Statistics and Operation Research, 601-631.
14. Hashempour, M., Alizadeh, M., & Yousof, H. M. (2023). A New Lindley Extension: Estimation, Risk Assessment and Analysis Under Bimodal Right Skewed Precipitation Data. Annals of Data Science, 1-40.
15. Johnson, A., Smith, B., & Jones, C. (2010). Statistical Methods for Automobile Claims Analysis. Journal of Insurance Analytics, 15(2), 123-140.
16. Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). Loss Models: From Data to Decisions. John Wiley & Sons.
17. Lemaire, J. (1995). Automobile Insurance: Actuarial Models. Kluwer Academic Publishers.
18. Smith, D., & Jones, E. (2015). Modeling Over-Dispersed Claims Frequencies in Automobile Insurance. Insurance Research Journal, 25(3), 201-218.
19. Mohamed, H. S., Ali, M. M., & Yousof, H. M. (2023). The Lindley Gompertz Model for Estimating the Survival Rates: Properties and Applications in Insurance. Annals of Data Science, 10(5), 1199-1216.
20. Mohamed, H. S., Cordeiro, G. M., Minkah, R., Yousof, H. M., & Ibrahim, M. (2024). A size-of-loss model for the negatively skewed insurance claims data: applications, risk analysis using different methods and statistical forecasting. Journal of Applied Statistics, 51(2), 348-369.
21. Mohamed, H. S., Cordeiro, G. M., & Yousof, H. M. (2022). The synthetic autoregressive model for the insurance claims payment data: modeling and future prediction. Statistics, Optimization & Information Computing, forthcoming.
22. Salem, M., Emam, W., Tashkandy, Y., Ibrahim, M., Ali, M. M., Goual, H., & Yousof, H. M. (2023). A new lomax extension: Properties, risk analysis, censored and complete goodness-of-fit validation testing under left-skewed insurance, reliability and medical data. Symmetry, 15(7), 1356.
23. Tashkandy, Y., Emam, W., Cordeiro, G. M., Ali, M. M., Aidi, K., Yousof, H. M., & Ibrahim, M. (2023). Distributional Censored and Uncensored Validation Testing under a Modified Test Statistic with Risk Analysis and Assessment. Journal of Mathematics, 2023.
24. Viaene, S., & Dedene, G. (2004). "Insurance fraud=sues and challenges." The Geneva Papers on Risk and Insurance - Issues and Practice, 29(2), 313-333.
25. Willmot, G. E. (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. Scandinavian Actuarial Journal, 1987(3-4), 113-127.
26. Yousof, H. M., Ansari, S. I., Tashkandy, Y., Emam, W., Ali, M. M., Ibrahim, M., & Alkhayyat, S. L. (2023a). Risk analysis and estimation of a bimodal heavy-tailed burr XII model in insurance data: exploring multiple methods and applications. Mathematics, 11(9), 2179.

27. Yousof, H. M., Chesneau, C., Hamedani, G., & Ibrahim, M. (2021). A new discrete distribution: Properties, characterizations, modeling real count data, Bayesian and non-Bayesian estimations. Statistica.
28. Yousof, H. M., Emam, W., Tashkandy, Y., Ali, M. M., Minkah, R., & Ibrahim, M. (2023b). A novel model for quantitative risk assessment under claim-size data with bimodal and symmetric data modeling. Mathematics, 11(6), 1284.
29. Yousof, H. M., Tashkandy, Y., Emam, W., Ali, M. M., & Ibrahim, M. (2023c). A New Reciprocal Weibull Extension for Modeling Extreme Values with Risk Analysis under Insurance Data. Mathematics, 11(4), 966.