Journal of Money and Economy
Vol. 20, No. 1, Winter 2025
pp. 31-55

Original Research Article

Insurance Pricing Using Frequency-Severity Models: A Comprehensive Approach

Ghadir Mahdavi[*] **Zahra Asghari**[†]
Ali Souri[‡]

Received: 02 Jul 2025 Approved: 15 Sep 2025

This paper investigates insurance pricing using frequency-severity models focusing on the Potential Deviation Ratio (PDR) as a key measure in premium determination. The objective is to develop an accurate and actuarially sound pricing approach by modeling claim frequency and severity using appropriate statistical distributions. Specifically, we propose that the product of two fitted distributions—one for frequency and the other for severity—can be used to calculate the pure premium.

The methodology includes fitting various distributions to frequency and severity data from a major Iranian insurance company. To model the dependency structure, we employ the Clayton copula. This dependency structure is used in calculating the pure premium by multiplying frequency and severity samples based on the fitted distributions. Kolmogorov-Smirnov, Anderson-Darling, and Chi-Squared tests are applied to evaluate model performance.

The results demonstrate that using the pure premium based on their respective fitted distributions significantly enhances the accuracy of the model. This approach leads to more precise risk classification, improved premium setting, and ultimately, fairer pricing strategies. Additionally, the findings indicate that incorporating dependency structures between frequency and severity not only improves risk assessment but also contributes to greater financial stability for insurers.

Keywords: Insurance Pricing, Frequency-Severity Models, Severity Distribution, Copula.
JEL Classification*:* C13, G22, C51

[*] Associate Prof., Department of Insurance, ECO College of Insurance, Allameh Tabataba'i University, Tehran, Iran; mahdavi@atu.ac.ir (Corresponding author)
[†] Ph.D. Candidate, ECO College of Insurance, Allameh Tabataba'i University, Tehran, Iran; Zahra.asghary61@gmail.com
[‡] Associate Prof., Faculty of Economics, University of Tehran, Tehran, Iran; alisouri@ut.ir

[DOI: 10.66224/jme.20.1.25]

1 Introduction

One of the key activities in the insurance industry is the pricing of insurance services. Fair pricing is essential to maintain stability in insurance operations. Unlike other industries, insurance pricing is based on historical data under the assumption that past events can predict future risks. Insurance premiums must be sufficient to cover expected losses, underwriting expenses, and potential catastrophic events.

Two primary factors influence insurance pricing: (1) risk prediction in the face of future uncertainty, and (2) legal and regulatory constraints. As a result, determining a fair premium has long been a central issue in actuarial science. Moreover, ensuring an insurer's solvency to cover future claims and underwriting costs is a critical concern. From the policyholder's perspective, appropriate and fair pricing is vital in a competitive environment, while for insurers, it ensures financial stability and the ability to meet obligations with a specified level of confidence.

This study adopts the principle of equivalence, a fundamental concept in actuarial science, which requires the present value of premiums to equal the present value of expected losses over a given period. Through this method, both fair and total premiums can be calculated simultaneously, providing a measure of an insurer's financial strength, defined as its confidence level in covering future liabilities.

The key innovation of this study lies in its approach to calculating the Solvency Index (SI). Unlike traditional models that rely on capital adequacy ratios, this method defines financial strength objectively as the probability of meeting all policyholder claims. Thus, the Solvency Index is interpreted as the probability that an insurer can fulfill its obligations in full.

Accurate pricing models are essential to ensure the long-term viability of insurance companies while offering fair premiums to customers. One widely used actuarial approach is the frequency-severity model, which models the number of claims (frequency) and average claim size (severity) separately. This paper focuses on modeling total insurance losses as the product of frequency and severity, while also considering their dependency structure through the use of copulas.

Accurate insurance pricing and solvency assessment remain challenging due to the complex statistical nature of claim frequency and severity and their dependence. Traditional premium models often neglect this dependence and fail to provide a precise measure of an insurer's financial strength, limiting the reliability of pricing and solvency evaluations.

[DOI: 10.66224/jme.20.1.25]

This study aims to enhance insurance pricing accuracy and solvency evaluation by jointly modeling claim frequency and severity, taking into account their statistical properties and dependence structure. It introduces a comprehensive approach integrating distribution fitting, Potential Deviation Ratio (PDR) analysis, and copula modeling to address the shortcomings of conventional methods.

This paper is organized as follows. Section 2 reviews the relevant literature, covering theoretical foundations and empirical applications. Section 3 presents the methodology, including collective risk modeling and premium calculation based on the principle of equivalence. Section 4 describes the dataset and modeling procedures. Section 5 reports empirical results. Section 6 concludes with a summary of findings and implications.

2 Literature Review

This study is grounded in the theoretical foundations of frequency–severity modeling, which decomposes total insurance losses into two separate components: the number of claims (frequency) and the size of claims (severity).

The classical assumption of independence between frequency and severity simplifies modeling but may lead to biased estimates and underestimated risk. Therefore, this study adopts a more flexible approach by incorporating dependency through copulas, which allow the joint distribution of frequency and severity to be modeled separately from their marginals. Copula functions provide a robust mathematical framework for capturing tail dependencies and complex interactions between random variables, thus improving the accuracy of risk estimation.

Additionally, the solvency concept in insurance is reinterpreted in this paper as a probability measure: the likelihood that an insurer can meet all future obligations. This probabilistic view of solvency departs from traditional ratio-based methods and aligns with a more actuarially sound principle of equivalence. The Solvency Index (SI), as defined in this study, represents the probability that the aggregate loss does not exceed available capital, a formulation that requires precise modeling of total loss distributions derived from empirical frequency and severity data.

2.1 Theoretical Foundations: Generalized Linear Models (GLM) and Zero-Inflated Models

Generalized linear models (GLMs) are cornerstones in modern actuarial science, offering a unified and extensible framework for modeling various

[DOI: 10.66224/jme.20.1.25]

types of insurance data. As emphasized by De Jong and Heller (2008), GLMs are particularly well-suited for insurance applications due to their ability to incorporate different distributions from the exponential family and to handle the non-normal, skewed, and zero-inflated nature of claim data.

Within this framework, two common approaches for addressing zero-inflation are the frequency–severity model and the Tweedie compound model. The frequency–severity model, as discussed by Frees (2014), decomposes total claim costs into two components: claim frequency and severity. The frequency component models claim occurrence using Poisson or logit regression, while the severity component models claim size (conditional on occurrence) using gamma or inverse Gaussian regression. This flexible structure allows separate modeling of claim occurrence and magnitude, making it suitable for zero-inflated and skewed datasets. However, a key limitation is the assumption of independence between frequency and severity.

Another widely used approach for modeling zero-inflated insurance data is the Tweedie compound Poisson model, which integrates both claim count and claim severity into a single distributional framework. Jørgensen and de Souza (1994) addressed the challenge of pricing insurance contracts based on aggregated data by proposing the Tweedie model, where the number of claims follows a Poisson distribution and the claim sizes follow a Gamma distribution. The total claim cost per insured unit is thus modeled as a single random variable following a Tweedie distribution, effectively capturing both frequency and severity components within an exponential dispersion model framework.

Moreover, they highlighted that the structure of the Tweedie likelihood belongs to the linear exponential family, making it compatible with generalized linear modeling (GLM) techniques. To further improve predictive accuracy and better account for variability in claim amounts, they recommended modeling the dispersion parameter as a function of covariates—an extension implemented through double generalized linear models (DGLMs). This modeling strategy is especially useful in real-world scenarios where only aggregate claim amounts are available, and individual claim counts are unobservable.

2.2 Models Relaxing the Independence Assumption

Although traditional models assume independence between claim frequency and severity, empirical findings suggest these components are often correlated. Ignoring this dependence can lead to biased parameter estimates

[DOI: 10.66224/jme.20.1.25]

and poor predictions. Several studies have addressed this issue by proposing models that account for such dependencies.

Gschlößl and Czado (2007) developed a spatial Bayesian model for automobile insurance claims that jointly models claim frequency and claim size while allowing for dependence between them. The model incorporates covariates and spatial random effects to capture spatial dependencies. Claim counts follow a Poisson distribution and claim sizes a Gamma distribution. Parameters are estimated using Markov Chain Monte Carlo (MCMC), and model comparison is conducted with criteria such as the Deviance Information Criterion (DIC) and posterior predictive scoring rules. Applied to German car insurance data, the inclusion of spatial effects improved model fit and prediction accuracy, and significant dependence between claim frequency and severity was detected, offering meaningful actuarial insights.

Frees et al. (2011) develop statistical models to predict both the frequency and amount of health care expenditures. They use a two-part modeling approach where the number of claims (frequency) and the expenditure amount per claim (severity) are modeled separately but linked to capture dependencies. The study applies advanced regression techniques and incorporates demographic and health-related covariates to improve predictive accuracy. Using extensive health insurance data, the authors demonstrate that accounting for dependence between frequency and severity leads to better forecasting of total health care costs, aiding insurers and policymakers in resource allocation and risk management.

Erhardt and Czado (2012) address the challenge of modeling dependent yearly insurance claim totals across different coverage fields, which often include many zeros. Because the marginal distributions have a point mass at zero, the cumulative distribution functions are not uniform, complicating the use of copulas. They demonstrate how to express the joint probability function using copulas with discrete and continuous margins, applying a pair-copula construction to flexibly select suitable copulas for each pair of margins.

Czado et al. (2012) challenge the classical assumption of independence between claim frequency and severity in the compound Poisson framework by introducing a mixed copula model that captures the dependence between the number of claims and their average size. Using a Gaussian copula to model the joint distribution, they incorporate regression effects through generalized linear models (GLMs) for both frequency and severity components. Parameters are estimated via an adaptive version of the maximization by parts method. Their simulation results and empirical application to automobile insurance data demonstrate that the copula-based model offers improved

[DOI: 10.66224/jme.20.1.25]

accuracy over traditional models in capturing the joint behavior of claim counts and sizes.

Copulas provide a flexible tool for modeling dependencies between claim frequency and severity. Krämer et al. (2013) used parametric copulas to jointly model the two components. These models can capture complex dependence structures, improving predictive accuracy and risk evaluation.

Lee and Shi (2019) proposed a dependent frequency-severity modeling framework for longitudinal insurance claims that explicitly captures the temporal correlation between claim frequency and severity. Utilizing copula regression techniques, their approach models both the time-dependent behavior within each component and the dependence structure between frequency and severity. This methodology enhances the predictive accuracy of claim distributions, making it particularly useful for ratemaking and other actuarial forecasting applications.

Shi et al. (2015) developed a hurdle model framework for non-life insurance ratemaking, where the first stage models claim occurrence and the second stage models claim frequency and severity conditional on having at least one claim. To capture dependence between claim frequency and severity, they proposed two strategies: a conditional probability decomposition approach incorporating claim counts as a covariate in the severity regression, and a copula-based joint modeling approach that models the distribution of claim counts and sizes simultaneously. Extensive simulations showed both methods outperform traditional GLM-based approaches such as Tweedie compound Poisson and two-part GLMs, especially in predictive accuracy. Their models were also validated on a U.S. automobile insurance dataset, demonstrating superior performance over common industry benchmarks.

Shirkavand et al. (2019) introduced a novel method for calculating fair insurance premiums and assessing the financial solvency ratio of insurance companies based on the principle of equivalence. Their approach uses the PDR method, which derives the aggregate loss distribution from empirically estimated frequency and severity distributions rather than assuming normality. The study demonstrates that premiums and solvency margins calculated using the actual distributions differ significantly from those based on normal distribution assumptions, particularly at higher confidence levels. This highlights the importance of accurate modeling of claim data to avoid underestimating financial risk, which could otherwise lead to insolvency. Their methodology provides a scientifically grounded framework applicable across various insurance fields, for improved premium pricing and solvency assessment.

[DOI: 10.66224/jme.20.1.25]

2.3 Additional Methods for Overdispersion and Heavy Tails

Apart from the methods focusing on dependence, several other techniques have been proposed to address overdispersion and the fat-tailed nature of insurance claims data.

Boucher et al. (2007) compare various risk classification models for claim counts, focusing on zero-inflated mixed Poisson and hurdle models to address zero-inflation and overdispersion in insurance claim data. Using a dataset from an automobile insurance portfolio, they evaluate the models' ability to capture claiming behavior. Their analysis employs statistical tests such as Score, Hausman, and Vuong tests, along with information criteria, to assess model fit and compare alternatives. These models provide flexible approaches for modeling claim counts with excess zeros and variability beyond the standard Poisson assumption.

To address the skewness and heavy-tailed behavior commonly observed in claim size distributions, Shi (2014) proposes several fat-tailed regression techniques. These methods aim to more accurately capture the distributional characteristics of claim severities, which often show pronounced right skewness and heavy tails, particularly in property and casualty insurance. Shi (2014) introduces four primary approaches: transformation models, exponential family models, generalized distribution models, and median regression. These techniques enhance modeling by providing a better representation of the entire claim size distribution, especially its tail behavior, which is essential for effective risk assessment and premium calculation in actuarial practice.

The connection between the theoretical framework and research methodology lies in directly translating each theoretical concept into corresponding empirical steps The collective risk model defines total claims as a combination of frequency and severity, which is reflected in the methodology by modeling frequency with discrete distributions and severity with continuous ones based on the data's expected statistical properties. The actuarial principle of equivalence underpins the premium calculation procedure, ensuring that modeled losses match premiums through explicit formulae. The solvency perspective in theory is operationalized via the Potential Deviation Ratio, calculated from the fitted models to measure risk probabilistically. Finally, the choice of goodness-of-fit tests is grounded in theoretical expectations about claim data behavior, ensuring that statistical validation aligns with the underlying actuarial concepts.

The main contribution of this paper lies in proposing a comprehensive framework for premium estimation that integrates empirical distribution

[DOI: 10.66224/jme.20.1.25]

fitting, dependency modeling using copulas, and PDR-based solvency evaluation. Unlike previous studies that model frequency and severity separately or assume independence, this study explicitly accounts for the dependency using the Clayton copula and derives the joint distribution of pure premium.

3 Methodology

3.1 Collective Risk Modeling

It is assumed that the total number of claims occurring during a given time period is denoted by N. In this case, the total amount of claims S is expressed as follows:

$$S = Y_1 + Y_2 + Y_3 + \ldots + Y_n = \Sigma Y_i \quad (1)$$

In this equation, the $Y_i's$ represent the individual claim amounts for each insurance policy. If we are at the beginning of the time period, neither the number of claims nor the individual claim amounts are known. Therefore, all of these unknowns must be modeled using random variables .The models derived for the total claim amount S are referred to as collective risk models because the entire portfolio is considered and modeled as a whole. Collective risk models are based on the law of large numbers, which allows insurance companies to benefit from the advantages of diversification.

The starting point for modeling S is a compound distribution. This compound distribution is based on a set of strong assumptions. The three standard assumptions are:

1) N is a discrete random variable that only includes non-negative integers.
2) $Y_1, Y_2, \ldots$ are identically distributed and independent of each other.
3) N and the $Y_i's$ are independent of each other.

If the three assumptions outlined above hold for the model of S, then S follows a compound distribution. This compound distribution is the basic model for collective risk modeling.

The characteristics of the compound distribution S are as follows:

$$E[S] = E[N] \times E[Y_1 \quad (2)$$
$$Var(S) = Var(N) \times E[Y_1]\text{^}2 + E[N] \times Var(Y_1) \quad (3)$$

Based on the above relationships, to obtain the collective risk distribution, one must first calculate the distribution of the number of claims and the

[DOI: 10.66224/jme.20.1.25]

distribution of claim amounts, and then combine them to derive the overall collective risk distribution (Wuthrich, 2024)

3.2 Calculating Premiums Based on the Principle of Equivalence

One of the core principles in insurance actuarial calculations is the principle of equivalence, which serves as the foundation for premium computations, especially in property and casualty insurance. Insurance companies must consider two key aspects when calculating premiums. On one hand, the insurer needs to set a premium that not only covers expected losses, administrative expenses, and the potential for catastrophic events but also provides a margin for the company's profit. On the other hand, in a competitive market, if an insurer sets premiums significantly higher than the fair rate, it risks losing customers (Werner & Modlin, 2010).

The principle of equivalence can be summarized as follows:

$$\text{Total Expected Losses} = \text{Sum of Collected Premiums} \tag{4}$$

Based on this principle, the total premiums collected from all policyholders within each line of insurance should approximately equal the total expected losses over a specified time period. Utilizing this principle, the formula for calculating the net premium is as follows:

$$\text{Net Premium} = \text{Average Expected Losses} \tag{5}$$

Breaking this down further:

$$\text{F.P} = \frac{Sum\ of\ \text{Expected Losses}}{\text{Total Policies or Exposures}} = \frac{\text{AL} \times \text{AF}}{\text{N}} \tag{6}$$

Where:
FP = Fair Premium
AL = Average Loss Severity
AF = Average Claim Frequency
N = Total Policies or Exposures

$$\text{AF} = \frac{\text{C}}{\text{N}} \tag{7}$$

$$\text{AL} = \frac{\text{L}}{\text{C}} \tag{8}$$

Where:
C = Total Number of Claims
L = Total Losses

[DOI: 10.66224/jme.20.1.25]

Thus, the net premium reflects the average expected losses spread across all insured individuals, effectively sharing the risk among policyholders. A premium calculated in this manner is also referred to as a fair premium, and it solely accounts for the cost of covering losses, excluding overhead costs.

3.3 Potential Deviation Ratio

Assuming the normality of frequency and severity distributions, and applying the concept of confidence intervals, PDR is defined. This ratio represents the amount or percentage that an insurance company should add to the expected frequency or severity in order to ensure, with a $(1-\alpha)$ confidence level, its ability to cover all claims. This ratio is calculated under the assumption that both frequency and severity follow a normal distribution as follows:

$$PDR_{(1-\alpha)\%} = \frac{\sigma_f Z_{(1-\alpha)}}{\bar{F}} \quad (9)$$

$$PDR_{(1-\alpha)\%} = \frac{\sigma_x Z_{(1-\alpha)}}{\bar{X}} \quad (10)$$

Where:

σ_f: The standard deviation of the frequency distribution, representing the variability or fluctuation in the frequency of claims.

$Z_{(1-\alpha)}$:The critical value from the standard normal distribution corresponding to a (1 - α)% confidence level.

$\bar{F}$: The expected or average frequency of claims.

σ_x: The standard deviation of the severity distribution, representing the variability or fluctuation in the severity of claims.

$\bar{X}$: The expected or average severity of claims.

Equation (9) provides the PDR for claim frequency, while Equation (10) corresponds to claim severity. The PDR value quantifies the additional proportion by which the expected frequency or severity must be adjusted to ensure, with a confidence level of $(1-\alpha)$, that the insurer can meet all claim obligations. This adjustment accounts for the possibility that actual outcomes may exceed historical averages.

In practice, claim frequency and severity distributions often deviate from normality. Insurance data, particularly for losses, frequently exhibit heavy-tailed characteristics. In such cases, the PDR should be calculated using the actual fitted loss distribution rather than assuming normality.

For a given distribution, the PDR at a specific confidence level (e.g., 90%) is defined as the percentage deviation between the value of the random variable (e.g., claim frequency or severity) at the corresponding quantile of

[DOI: 10.66224/jme.20.1.25]

the fitted cumulative distribution function (CDF) and the mean of the distribution.

$$PDR = \frac{Value_{P_{(1-\alpha)}} - Value_{P_{mean}}}{Value_{P_{mean}}} = \frac{k\sigma}{\mu} \quad (11)$$

Here's the breakdown of the symbols in the equation:

$Value_{P_{(1-\alpha)}}$: The value of the distribution at the (1 - α) confidence level, which corresponds to a critical value based on the chosen confidence level (such as 95% or 99%).

$Value_{P_{mean}}$: The mean (or expected) value of the distribution (either frequency or severity), based on historical data or other estimations.

k: A constant factor that links the standard deviation to the deviation ratio (it could be related to the confidence level or other factors depending on the context).

σ: The standard deviation of the distribution (either frequency or severity), representing the amount of variation or dispersion from the mean.

μ: The mean (or expected) value of the distribution (either frequency or severity), representing the average value based on historical data.

The fraction in Equation (11), representing the PDR, quantifies the relative deviation of the distribution's $(1-\alpha)$-quantile from its mean. This ratio reflects the proportion by which the value at a specified confidence level exceeds the expected value of the distribution.

Under general distributional assumptions, Chebyshev's inequality provides a useful bound for this deviation. Specifically, it guarantees that for any random variable with finite mean and variance, the probability of observing a value that deviates from the mean by more than k standard deviations is at most $1/k^2$. This result applies to all probability distributions, regardless of their shape, and offers a conservative estimate of the potential deviation. It ensures that most values lie within a bounded range around the mean, which is particularly helpful when the exact distributional form of the data (e.g., claim frequency or severity) is unknown or exhibits heavy tails.

4 Scope of the Research

The input data in this study consists of total claim amounts, claim counts, and the number of policies related to Collision insurance, sourced from one of the largest insurance companies in Iran. The original dataset contains 52,000 individual claim records collected over a six-year period from 2016 to 2021 (corresponding to the Iranian years 1395 to 1400). These records were

[DOI: 10.66224/jme.20.1.25]

categorized based on relevant underwriting risk factors such as the policyholder's age, vehicle usage type, and other characteristics. After aggregation based on these variables, the data were grouped into 134 homogeneous classes used for statistical modeling.

Discrete and continuous statistical distributions have been employed to fit the distribution to the variables of claim severity and frequency. Modeling and statistical calculations were conducted using Easyfit software. Kolmogorov-Smirnov, Anderson-Darling, and Chi-square tests were performed to assess the goodness of fit of the mentioned distributions on the data. Parameter estimation for the fitted distributions was performed using methods including maximum likelihood estimation, the method of moments, and least squares.

The initial step involved selecting appropriate probability distribution models that capture the essential characteristics of the claim data, such as the presence of zero values and distribution skewness. Preliminary analyses guided the choice of models, including options like Poisson, Negative Binomial, and various continuous distributions, chosen for their applicability in analyzing claim frequency and severity in insurance contexts.

After selecting the models, the parameters for each distribution were estimated using statistical methods like Maximum Likelihood Estimation (MLE) or the Method of Moments, ensuring accuracy as these parameters directly impact the model's predictive power and alignment with observed data. Next, the goodness of fit for each model was assessed through statistical tests, including the Kolmogorov-Smirnov and Anderson-Darling tests, and visual methods, such as Q-Q and P-P plots, to verify the models' alignment with empirical data. Finally, specific selection criteria were applied, such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which balance fit quality and model simplicity, allowing the identification of the most accurate and parsimonious model for representing the claim data.

5 Results

The empirical analysis follows a clear sequence: data exploration and descriptive statistics, distribution fitting and goodness-of-fit testing for frequency and severity, dependency modeling via copulas, and finally using the proposed Solvency Index.

Using EasyFit software, several probability distributions were fitted to the claim frequency data. The results of the goodness-of-fit tests revealed that the Dagum and Normal distributions offered the best fit to the observed data. The Weibull distribution also demonstrated a relatively good fit. In contrast, the

[DOI: 10.66224/jme.20.1.25]

Inverse Gaussian distribution exhibited a poor fit, particularly in the distribution tails. Table 1 presents the results of the goodness-of-fit tests, summarizing the statistical criteria used to assess and compare the performance of each fitted distribution.

Table 1
Goodness-of-Fit Test Results for Claim Frequency Data at a 95% Confidence Level

Distribution	Kolmogorov Smirnov			Anderson Darling		Chi-Squared		
	Statistic	P-Value	Critical Value	Statistic	Critical Value	Statistic	P-Value	Critical Value
Dagum	0.04505	0.92282	0.11396	0.33726	2.5018	3.4204	0.84358	14.067
Normal	0.04675	0.90095	0.11396	0.46996	2.5018	3.3533	0.85052	14.067
Weibull	0.0633	0.5972	0.11396	0.63033	2.5018	6.7983	0.45018	14.067
Inv. Gaussian	0.1111	0.05554	0.11396	4.7655	2.5018	23.184	0.00158	14.067

Source: Authors' calculations. Dagum and Normal distributions show high p-values and low test statistics, indicating a good fit. In contrast, the Inverse Gaussian shows poor fit due to low p-values and high statistics, particularly in the tails.

Table 1 presents the results of the goodness-of-fit tests for the claim frequency data at the 95% confidence level. The tests employed include the Kolmogorov-Smirnov, Anderson–Darling, and Chi-Squared tests, each of which evaluates how well a given distribution aligns with the observed data. A brief overview of these tests is provided below:

Kolmogorov-Smirnov (K-S) Test: This test compares the empirical distribution function of the sample data to the cumulative distribution function (CDF) of the fitted distribution. The test statistic represents the maximum absolute difference between these two functions. A high p-value (typically above 0.05) suggests that the distribution provides a good fit to the data.

Anderson–Darling (A-D) Test: An extension of the K-S test, the A-D test assigns more weight to the tails of the distribution. The test statistic indicates the degree of deviation between the sample and the theoretical distribution, while the critical value provides a benchmark for determining the goodness of fit. A lower test statistic and a higher p-value imply a better fit.

Chi-Squared Test: This test compares the observed and expected frequencies in pre-defined intervals (bins). The test statistic measures the overall discrepancy between the two sets of frequencies. A p-value greater than 0.05 typically indicates that the differences are not statistically significant, thereby supporting the adequacy of the fitted distribution.

Similarly, Table 2 presents the results of the goodness-of-fit tests applied to the claim severity data. Using EasyFit software, various distributions were

[DOI: 10.66224/jme.20.1.25]

fitted, and the statistical results summarized in Table 2 facilitate the comparison of the adequacy of each fitted model.

Table 2
Goodness-of-Fit Test Results for Claim Severity Data at a 95% Confidence Level

Distribution	Kolmogorov Smirnov			Anderson Darling		Chi-Squared		
	Statistic	P-Value	Critical Value	Statistic	Critical Value	Statistic	P-Value	Critical Value
log normal	0.06244	0.61458	0.11396	0.51566	2.5018	5.0503	0.65383	14.067
Log-Logistic	0.06979	0.47258	0.11396	0.62709	2.5018	7.8816	0.34315	14.067
Inv. Gaussian	0.09525	0.14246	0.11396	1.1161	2.5018	11.121	0.1341	14.067
Normal	0.36857	1.6968E-17	0.11396	32.99	2.5018	58.267	2.7702E-11	14.067

Source: Authors' calculations. Log-Normal, Log-Logistic, and Inverse Gaussian distributions show a good fit to the claim severity data, while the Normal distribution fails to adequately represent the data.

At this stage, the net or fair premium was calculated using the statistical distributions fitted to both claim frequency and claim severity. The principle of equivalence was applied, under which the premium is set equal to the expected cost of claims. This ensures that the insurer's expected income from premiums matches its expected liability from claims, excluding any profit or loading components.

The calculation relies on the mean values of the fitted distributions. Specifically, the expected number of claims (frequency) is multiplied by the expected cost per claim (severity) to obtain the expected total claims cost per policyholder. This product represents the fair premium, which is the amount that the insurer must charge to cover expected losses without incurring a financial gain or deficit.

The results of this process—showing the fair premiums under different scenarios, based on the selected distributions for frequency and severity—are summarized in Table 3.

In summary, the net premium reflects the minimum premium necessary to cover the expected cost of claims, and it is determined by the mean values derived from the best-fitting distributions for frequency and severity.

[DOI: 10.66224/jme.20.1.25]

Table 3
Net or Fair Premium

Claim severity distributions / Claim frequency distributions	log normal	Log-Logistic	Inv. Gaussian	Normal
Dagum	27,563,238	59,962,951	31,811,331	31,811,331
Normal	27,408,295	59,625,878	31,632,508	31,632,508
Weibull	27,236,826	59,252,851	31,434,611	31,434,611
Inv. Gaussian	27,408,295	59,625,878	31,632,508	31,632,508

Source: Authors' calculations. This table presents fair premium estimates based on various claim frequency and severity distributions used in the analysis.

In the next step, the Potential Deviation Ratio (PDR) is calculated using Equation 13. For claim severity, the PDR is evaluated at confidence levels higher than those corresponding to the mean for each fitted distribution. This adjustment is necessary since the previously calculated fair premiums were based on the mean values of the respective distributions. However, the mean occurs at different confidence levels depending on the distribution type.

For instance, in a normal distribution, the mean corresponds to the 50% confidence level, meaning there is an equal probability of values falling above or below the mean. In contrast, for a fitted log-normal distribution, the mean corresponds to approximately the 78% confidence level. This difference arises from the unique shape and skewness of each distribution and is critical when assessing claim severity in insurance applications.

To illustrate, assume claim severity is modeled using a log-normal distribution. Given the positive skewness of the log-normal distribution, the mean value lies above the median and corresponds to a confidence level of approximately 78%. This means that there is a 78% probability that a randomly selected claim severity will be below the mean value. Consequently, when calculating the PDR, it is essential to consider these distribution-specific confidence levels to accurately reflect the risk and ensure appropriate premium loading for extreme claim outcomes.

Table 4
Claim Severity Values Corresponding to Different Confidence Levels

P	0.78	0.9	0.95	0.99
log normal	206,590,000	453,420,000	795,250,000	2,281,500,000

Source: Authors' calculations. This table shows claim severity values at different confidence levels based on the log-normal distribution. As the confidence level increases from 78% to 99%, the corresponding claim severity values rise from approximately 206,590,000 to 2,281,500,000.

[DOI: 10.66224/jme.20.1.25]

severity data. Specifically, for the fitted Log-normal distribution, the mean claim severity of 206,590,000 corresponds to a confidence level of 78%. This implies that if the insurer bases the net premium calculation on this mean value under the principle of equivalence, the resulting solvency level will only be 78%, meaning there is a 78% probability that future claims can be covered.

To improve this solvency level, the claim severity values corresponding to higher confidence levels have been computed using the cumulative distribution function of the fitted model. For instance, to increase the solvency level from 78% to 90%, 95%, and 99%, the adjusted claim severity amounts should be 453,420,000, 795,250,000, and 2,281,500,000 respectively in the net premium calculation.

Alternatively, these adjustments can be expressed using the Potential Deviation Ratio (PDR), calculated as follows:

$$PDR = \frac{453{,}420{,}000 - 206{,}590{,}000}{206{,}590{,}000} = 119\ \% \qquad (12)$$

By adding a 119% PDR to the fair premium based on the mean severity, the insurer effectively raises its solvency confidence from 78% to 90%. Table 5 summarizes the PDR values required to enhance the financial solvency level beyond that associated with the mean claim severity.

Table 5
PDR for Different Claim Severity Distributions

	0.9	**0.95**	**0.99**
log normal	119%	285%	1004%
Log-Logistic	-7%	79%	663%
Inv. Gaussian	122%	344%	532%
Normal	378%	485%	685%

Source: Authors' calculations. Potential Deviation Ratios (PDR) required to raise the Financial Solvency Index above specified confidence levels (90%, 95%, and 99%) are presented for different claim severity distributions.

For claim frequency, the Dagum, Normal, and Weibull distributions provide the best fit to the data. For each of these distributions, the Potential Deviation Ratio (PDR) is calculated to achieve a solvency index higher than the confidence level corresponding to the mean value. These calculations are performed similarly to those for claim severity. The results are summarized in Table 6.

[DOI: 10.66224/jme.20.1.25]

Table 6
PDR for Different Claim Frequency Distributions

SI / Claim Frequency Distributions	0.9	0.95	0.99
Dagum	51%	66%	103%
Normal	54%	69%	98%
Weibull	58%	75%	108%
Inv. Gaussian	56%	80%	135%

Source: Authors' calculations. Potential Deviation Ratios (PDR) required to reach solvency index thresholds at 90%, 95%, and 99% confidence levels for selected claim frequency distributions.

The effect of increasing the number of insurance policies on the Potential Deviation Ratio (PDR) for claim frequency is examined based on the law of large numbers. According to this law, as the number of observations (or policies) increases, the observed mean of claim frequency converges to the true mean, thereby reducing the potential deviation. Table 7 illustrates this effect for the Dagum distribution, showing how the PDR changes as the number of policies increases by factors of 4, 9, 25, and 100.

For a small number of policies, the PDR is relatively high. For example, at a confidence level of 0.9, the PDR is 51% in the base case, which decreases to 25% when the number of policies increases by a factor of 4, to 17% at 9 times, and further drops to 5% when increased by 100 times. Similarly, at higher confidence levels of 0.95 and 0.99, the PDR decreases as the number of policies grows, demonstrating how larger sample sizes reduce the potential deviation from the expected mean.

Table 7
PDR with an Increase in the Number of Insurance Policies

SI / PDR	0.9	0.95	0.99
PDR	51%	66%	103%
PDR(4)	25%	33%	51%
PDR(9)	17%	22%	34%
PDR(25)	10%	13%	21%
PDR(100)	5%	7%	10%

Source: Authors' calculations. Potential Deviation Ratios (PDR) decrease as the number of insurance policies increases, illustrating reduced deviation from the mean at higher sample sizes across confidence levels of 90%, 95%, and 99%.

Assuming the independence of claim frequency and severity, if an insurance company aims to achieve the desired solvency margin while

[DOI: 10.66224/jme.20.1.25]

considering the risk of both variables, it must set a higher target based on the potential deviation from the calculated mean of each variable for the same solvency margin. This approach ensures that the risk associated with an increase in either variable is covered, thereby maintaining financial stability.

For example, assuming a Dagum distribution for claim frequency, the PDR from the mean required to achieve a 90% solvency index is 51%. On the other hand, assuming a log-normal distribution for claim severity, the PDR from the mean to achieve the same solvency margin is 119%.

To ensure coverage of future claims at a 90% confidence level while accounting for the risk of both variables, the insurance company must adjust the calculated fair premium accordingly. Based on the Dagum distribution for claim frequency and the log-normal distribution for claim severity, a premium of 27,563,238 Rials should be increased by 119% to 60,495,296 Rials. However, as demonstrated, an increase in the number of policies reduces the PDR from the mean required to achieve the 90% solvency index for claim frequency.

Table 8 presents the net premium required to achieve a 90% solvency margin, considering different frequency and severity distributions. The frequency distributions include Dagum, Normal, Weibull, and Inverse Gaussian, each modeling claim occurrence in different ways. On the other hand, the severity distributions include log-normal, log-logistic, Inverse Gaussian, and Normal, which model the size of claims.

Table 8
Net Premium to Achieve a 90% S.I

	log normal	**Log-Logistic**	**Inv. Gaussian**	**Normal**
Dagum	60,495,296.40	90,303,969.90	70,541,822.40	151,912,012.00
Normal	60,155,231.40	91,903,940.70	70,145,282.40	151,058,062.00
Weibull	59,778,892.80	93,405,036.90	69,706,444.80	150,113,024.00
Inv. Gaussian	60,155,231.40	92,847,743.70	70,145,282.40	151,058,062.00

Source: Authors' calculations. Net premiums required to achieve a 90% solvency index based on various claim frequency and severity distributions are presented.

The net premium is calculated so that the insurance company can confidently meet potential claims, even in the face of unexpected events and the inherent variability in both claim frequency and severity, while maintaining its 90% solvency margin. This solvency margin ensures the insurer holds reserves equal to 90% of its liabilities, enabling it to remain financially secure under challenging circumstances.

[DOI: 10.66224/jme.20.1.25]

In competitive market conditions, particularly in a monopolistic market, the insurance company can adjust its premium by the difference between the fair premium values (calculated in Table 3, not shown here) and the net premiums required to achieve the 90% solvency margin (from Table 8). This adjustment allows the insurer to increase it's premium to cover potential deviations from expected losses and maintain financial stability.

Ultimately, this analysis helps the insurance company ensure financial stability, enabling it to meet claims even in unfavorable conditions and safeguard against unexpected risks.

In a perfectly competitive market, the insurance company must maintain the total required premium increase to ensure it meets the desired solvency index. This solvency index is calculated based on the PDR, which reflects the potential risks and uncertainties in the market. To cover these risks, the insurer is required to set aside the reserve amounts listed in Table 9 for each policy underwritten. These reserves guarantee the company's ability to meet its financial obligations, even in the event of unexpected losses or claims.

Table 9
Required Values to Increase the Fair Premium

Claim severity distributions / **Claim frequency distributions**	**log normal**	**Log-Logistic**	**Inv. Gaussian**	**Normal**
Dagum	32,932,058.60	30,341,019.30	38,730,491.80	120,100,681.40
Normal	32,746,936.10	32,278,062.60	38,512,774.30	119,425,553.90
Weibull	32,542,067.20	34,152,185.70	38,271,833.60	118,678,412.80
Inv. Gaussian	32,746,936.10	33,221,865.60	38,512,774.30	119,425,553.90

Source: Authors' calculations. Total reserve amounts per policy required to achieve the solvency index are presented for different combinations of claim frequency and severity distributions. The highest reserve is for the Dagum–Normal combination (120,100,681.40), and the lowest for Dagum–Log-Logistic (30,341,019.30).

In a semi-competitive market, the insurance company enjoys greater flexibility. It can add a portion of the total required premium increase to the fair premium, which is calculated based on the PDR, reflecting expected losses and associated risk factors. The company then retains the remaining portion of the required premium increase as reserves. For instance, the company may choose to add 30% of the amounts listed in Table 9 to the fair premiums while keeping the remaining 70% as reserves. This approach enables the insurer to offer more competitive pricing while maintaining adequate reserves to ensure solvency and financial stability.

[DOI: 10.66224/jme.20.1.25]

At this stage of the project, our focus shifts to integrating two distributions. Because these distributions are dependent, copulas are employed to model their joint behavior. Specifically, we examine the product distribution derived from the two best-fitting distributions identified for the frequency and severity data, which are the Log-normal and Dagum distributions, respectively, based on goodness-of-fit results. The Clayton copula, which best fits the data, is used to simulate the product distribution, generating synthetic data that represents the combined behavior of frequency and severity processes. These simulation results offer valuable insights into the joint distribution of the two variables.

By using copulas, we explicitly account for the dependence between claim frequency and severity, rather than assuming independence. The Clayton copula captures positive dependence, allowing for a more accurate joint modeling of these variables. This is critical, as assuming independence could lead to misleading conclusions regarding the combined distribution and thus the risk assessment.

The goodness-of-fit tests—Kolmogorov-Smirnov, Anderson-Darling, and Chi-squared—were applied to evaluate how well various statistical distributions fit the simulated pure premium data. Table 10 presents these results, indicating that the Log Normal (3-parameter) and Log Gamma distributions provide the best fit, supported by low test statistics and high p-values. For example, the Kolmogorov-Smirnov p-value is 0.41611 for the Log Normal (3p) distribution and 0.1581 for the Log Gamma distribution, both exceeding the conventional significance threshold of 0.05, indicating a good fit. Furthermore, their Chi-squared statistics fall below the critical values, reinforcing their suitability.

In contrast, the Burr and Normal distributions fit the data poorly. The Burr distribution shows very low p-values in the Kolmogorov-Smirnov and Chi-squared tests (0.01129 and 4.34×10^{-7}, respectively), indicating significant deviation. Similarly, the Normal distribution yields extreme test values, with a Kolmogorov-Smirnov p-value of zero and large statistics in the Anderson-Darling and Chi-squared tests, confirming an inadequate fit.

[DOI: 10.66224/jme.20.1.25]

Table 10
Goodness-of-Fit Test Results for Simulated pure premium Data at a 95% Confidence Level

Distribution	Kolmogorov Smirnov			Anderson Darling		Chi-Squared		
	Statistic	P-Value	Critical Value	Statistic	Critical Value	Statistic	P-Value	Critical Value
Log Normal(3p)	0.01246	0.41611	0.1921	0.65977	2.5018	9.0533	0.69837	21.026
Log Gamma	0.0159	0.1581	0.01921	1.4709	2.5018	7.977	0.78692	21.026
Burr	0.02272	0.01129	0.01921	8.2684	2.5018	52.873	4.3401e^-7	21.026
Normal	0.38118	0	0.01921	1131	2.5018	2547.8	0	21.026

Source: Authors' calculations. Goodness-of-fit test results at a 95% confidence level for various distributions fitted to simulated pure premium data. Log Normal (3p) and Log Gamma show the best fit, while Burr and Normal distributions perform poorly.

Following the goodness-of-fit tests, the next step involves calculating the Potential Deviation Ratio (PDR) using Equation 11. This ratio is computed at confidence levels exceeding the mean for each fitted distribution. It is important to emphasize that the previously calculated fair premiums were based on the mean values of the respective distributions; however, the mean corresponds to different confidence levels depending on the distribution. For instance, in the Normal distribution, the mean aligns with the 50% confidence level, whereas in the Log Normal (3-parameter) distribution, the mean corresponds to approximately the 79% confidence level.

This variation in the confidence level associated with the mean is crucial for determining fair premiums, as it reflects the inherent risk and uncertainty within each distribution. Consequently, the PDR is calculated for the fair premium derived from the Log Normal (3p) distribution, with results presented in Table 11. This table details the deviation at various confidence levels, ensuring that the premiums better reflect the variability of the distribution and capture the true risk more accurately.

Table 11
Simulated pure premium data Corresponding to Different Confidence Levels

	0.79	0.9	0.95	0.99
log normal(3p)	5,498,500	11,756,000	21,163,000	63,843,000

Source: Authors' calculations. Potential Deviation Ratios (PDR) for the Log Normal (3p) distribution fitted to simulated pure premium data, shown at increasing confidence levels from 79% (mean) to 99%, reflecting the reserves needed for higher risk.

For fair premium, Log normal(3p) distribution provide the best fit to the data. For this distribution, the PDR is calculated to achieve a solvency index exceeding the confidence level associated with the mean value. The

[DOI: 10.66224/jme.20.1.25]

calculations are similar to those performed for claim severity and frequency, and the results are presented in Table 12

Table 12
PDR for simulated pure premium data Distribution

	0.9	**0.95**	**0.99**
PDR	114%	285%	1061%

Source: Authors' calculations. Potential Deviation Ratios (PDR) for the Log Normal (3p) distribution fitted to simulated pure premium data, showing significant increases at confidence levels above the mean (79%), reaching up to 1061% at 99%.

The effect of increasing the number of insurance policies on the Potential Deviation Ratio (PDR) for both claim frequency and severity is examined based on the law of large numbers. According to this law, as the number of observations (or policies) increases, the sample mean of the claim frequency converges to the true mean, thereby reducing the potential deviation. Table 13 illustrates this effect for the Log Normal (3-parameter) distribution, demonstrating how the PDR decreases as the number of policies grows by factors of 4, 9, 25, and 100.

Table 13
PDR with an Increase in the Number of Insurance Policies

SI / **PDR**	**0.9**	**0.95**	**0.99**
PDR	114%	285%	1061%
PDR(4)	57%	143%	531%
PDR(9)	38%	95%	354%
PDR(25)	23%	57%	212%
PDR(100)	11%	29%	106%

Source: Authors' calculations. Potential Deviation Ratios (PDR) decrease as the number of insurance policies increases, illustrating reduced deviation from the mean at confidence levels of 90%, 95%, and 99%.

6 Conclusion

Based on the analysis of the data, the best-fitting distributions for claim severity and frequency were identified, and the corresponding fair premiums were calculated.

For claim severity, the Log Normal distribution demonstrated the best fit, supported by high p-values in both the Kolmogorov-Smirnov (0.61458) and Chi-squared tests (0.6538), indicating close coincidence with the observed

[DOI: 10.66224/jme.20.1.25]

data. For claim frequency, the Dagum distribution was found to be most appropriate, with a Kolmogorov-Smirnov p-value of 0.9228, confirming a good fit.

Using the mean values of these best-fitting distributions, the fair premium was calculated as the product of the mean severity and frequency. Specifically, the mean severity (Log Normal) was 206,590,000, and the mean frequency (Dagum) was 0.13342, resulting in a fair premium of 27,563,238 million rials.

Next, the Potential Deviation Ratio (PDR) at the 90% confidence level was computed for both distributions. The PDR for severity at this level was 119%, while for frequency it was 51%. To ensure coverage of future claims at the 90% confidence level while accounting for the risks associated with both variables, the insurer must adjust the calculated fair premium accordingly. Based on the Dagum and Log Normal distributions, the premium of 27,563,238 million rials should be increased by 119%, resulting in an adjusted premium of 60,495,296 million rials.

Subsequently, the Clayton Copula was employed to model the dependence structure between claim severity and frequency. By combining the fitted Log Normal and Dagum distributions with the Clayton Copula, the fair premium was recalculated to be 11,756,000 million rials at the 90% confidence level.

Comparing the two approaches—the product of means method (yielding 60,495,296 million rials) and the copula-based method (yielding 11,756,000 million rials)—reveals that the copula approach produces a significantly lower premium. This difference underscores the importance of accounting for the dependency between claim frequency and severity. The Clayton Copula captures the positive dependence between these variables, leading to a more accurate and refined estimate of the fair premium.

The results of this study are consistent with previous research highlighting the critical role of modeling dependence between claim frequency and severity in insurance pricing.

Gschlößl and Czado (2007) developed a spatial Bayesian model that jointly models claim frequency and severity, detecting significant dependence and improving predictive accuracy. Similarly, Czado et al. (2012) proposed a mixed copula model capturing the dependence structure between the number of claims and their average size, which enhanced actuarial insight and model performance.

Krämer et al. (2013) utilized parametric copulas to jointly model claim frequency and severity, showing improved risk evaluation. Lee and Shi (2019) extended this by incorporating temporal correlation in a longitudinal

[DOI: 10.66224/jme.20.1.25]

framework using copula regression, which further enhanced predictive accuracy.

Consistent with these studies, our findings demonstrate that accounting for dependence via the Clayton Copula leads to a significantly different and more accurate estimation of the fair premium compared to assuming independence. This confirms the practical importance of incorporating frequency-severity dependence in actuarial modeling for more realistic premium pricing and solvency assessment.

References

Boucher, J.-P., Denuit, M., & 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*. https://doi.org/10.1080/10920277.2007.10597487

Czado, C., Kastenmeier, R., Brechmann, E. C., & Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal, 2012(4), 278–305*. https://doi.org/10.1080/03461238.2010.546147

De Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511755408

Erhardt, V., & Czado, C. (2012). Modeling dependent yearly claim totals including zero claims in private health insurance. *Scandinavian Actuarial Journal, 2012(2), 106–129*. https://doi.org/10.1080/03461238.2010.489762

Frees, E. W., Gao, J., & Rosenberg, M. A. (2011). Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal, 15(3), 377–392*. https://doi.org/10.1080/10920277.2011.10597626

Frees, E. W. (2014). Frequency and severity models. In E. W. Frees, G. Meyers, & R. A. Derrig (Eds.), *Predictive modeling applications in actuarial science, 1*. Cambridge University Press. https://doi.org/10.1017/CBO9781139342674

Gschlößl, S., & Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal, 2007(3), 202–225*. https://doi.org/10.1080/03461230701414764

Jørgensen, B., & Paes De Souza, M. C. (1994). Fitting Tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal, 1994(1), 69–93*. https://doi.org/10.1080/03461238.1994.10413930

Krämer, N., Brechmann, E. C., Silvestrini, D., & Czado, C. (2013). Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics, 53(3), 829–839*. https://doi.org/10.1016/j.insmatheco.2013.09.003

Lee, G. Y., & Shi, P. (2019). A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics, 87, 115–129*. https://doi.org/10.1016/j.insmatheco.2019.04.004

Shirkavand, S., Mahdavi Kalishami, G. H., & Pazoki, N. (2019). Insurance products ratemaking and insurance company financial solvency ratio calculation via

[DOI: 10.66224/jme.20.1.25]

potential deviation ratio method. *Financial Research Journal, 21(2), 165–186*. [In Persian] https://doi.org/10.22059/frj.2019.270699.1006769

Shi, P., Feng, X., & Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics*, *64*, *417–428*. https://doi.org/10.1016/j.insmatheco.2015.07.006

Shi, P. (2014). Fat-tailed regression models. In E. W. Frees, R. A. Derrig, & G. Meyers (Eds.), *Predictive modeling applications in actuarial science* (pp. 236–259). Cambridge University Press.

Werner, G., & Modlin, C. (2010). *Basic ratemaking*. In Casualty Actuarial Society (4th ed.).

Wuthrich, M. V. (2024). *Non-life insurance: mathematics & statistics*. Available at SSRN 2319328.

[DOI: 10.66224/jme.20.1.25]