Contributeurs

MAS-II: Modern Actuarial Statistics II

aut., cre. Alec James van Rassel

Référence (manuels, YouTube, notes de cours) En ordre alphabétique :
Contributeurs

Contents

A

Prerequisites

Distributions

Context

We typically use 3 types of random variable to describe losses:

Frequency or number of losses	always discrete
Severity or amount of losses (payment)	usually continuous, can be discrete or mixed too
Aggregate or total loss from summing a number (Frequency) of Severity variables	same as the severity

Discrete Distributions

Context

Discrete random variables are usually counting (frequency) variables, meaning their possible values are $\{0, 1, 2, \dots\}$

Probability Mass Function (PMF)

N is a *discrete random variable* if it has a ***probability mass function*** p_k such that $\boxed{p_k = \Pr(N = k)}$

Definition	Domain	Condition
$p_k = \Pr(N = k)$	$p_k \in [0, 1]$	$\sum_k p_k = 1$

Poisson Distribution

Notation	Parameters	Domain
$N \sim \text{Poisson}(\lambda)$	$\lambda > 0$	$n = 0, 1, 2, \dots$

$\Pr(N = n)$	$= \dfrac{e^{-\lambda}\lambda^n}{n!}$
$E[N]$	$= \lambda$
$Var(N)$	$= \lambda$

- If N_1 and N_2 are *independent* Poisson r.v., then $N_1 + N_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.
- The $e^{-\lambda}$ term makes the probabilities sum to 1 as the Taylor series for e^{λ} is

$$e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^n}{n!} + \cdots$$

Binomial Distribution

Context

A binomial r.v. N has m *independent* trials each having a probability q of a loss where n is the total number of losses.

Notation	Parameters	Domain
$N \sim \text{Bin}(m,q)$	$q \in (0,1); m \in \mathbb{N}$	$n = 0,1,2,\ldots$

$\Pr(N=n)$	$= \binom{m}{n} q^n (1-q)^{m-n}$
$E[N]$	$= mq$
$Var(N)$	$= mq$

- If N_1 and N_2 are *independent* binomial r.v. with the ***same*** q then $N_1 + N_2 \sim \text{Bin}(m_1+m_2, q)$.
- The case where $m=1$ corresponds to a ***Bernoulli*** r.v.

Geometric Distribution

Context

A geometric r.v. N with mean β can be obtained by setting n as the number of years ***before*** the first loss. Given the geometric distribution is memoryless, each year *independently* has a loss with probability

$$\underbrace{\Pr(N=0)}_{\text{probability of a loss the first year}} = \frac{1}{1+\beta}.$$

Notation	Parameters	Domain
$N \sim \text{Geo}(\beta)$	$\beta > 0$	$n = 0,1,2,\ldots$

$\Pr(N=n)$	$= \left(\frac{\beta}{1+\beta}\right)^n \frac{1}{1+\beta}$
$\Pr(N \geq n)$	$= \left(\frac{\beta}{1+\beta}\right)^n$
$E[N]$	$= \beta$
$Var(N)$	$= \beta(1+\beta)$

- Like the exponential distribution, the geometric distribution is memoryless:

$$\Pr(N = d+n | N \geq d) = \Pr(N=n)$$
$$E[N-d | N \geq d] = E[N]$$

Negative Binomial Distribution

Context

A negative binomial r.v. N represents the number of years n with no loss *before* the r^{th} year with a loss. We obtain a negative binomial r.v. $N \sim \text{NBin}(r,\beta)$ by summing r iid geometric r.v., $N_1, N_2, \ldots, N_r$, all with the same mean β.

Notation	Parameters	Domain
$N \sim \text{NBin}(\beta)$	$r, \beta > 0$	$n = 0, 1, 2, \ldots$

$\Pr(N = n)$	$= \binom{r+n-1}{r-1} \left(\frac{\beta}{1+\beta}\right)^n \left(\frac{1}{1+\beta}\right)^r$
$\Pr(N \geq n)$	$= \left(\frac{\beta}{1+\beta}\right)^n$
$E[N]$	$= r\beta$
$Var(N)$	$= r\beta(1+\beta)$

- A geometric r.v. is a negative binomial r.v. with $r = 1$.

Distribution	Mean		Variance
Binomial	mq	$>$	$mq(1-q)$
Poisson	λ	$=$	λ
Geometric	β	$<$	$\beta(1+\beta)$
Negative Binomial	$r\beta$	$<$	$r\beta(1+\beta)$

Severity Distributions

Probability Density Function (PDF)

X is a *continuous random variable* if it has a ***probability density function*** $f(x)$ such that $\boxed{f(x)}$

Definition	Domain	Condition
$f(x) =$	$f(x) \geq 0$	$\int_{\mathbb{R}} f(x)dx = 1$

Joint Distributions

Conditional Distributions

Aggregate Distributions

Normal, Uniform, Pareto, Exponential, and Gamma

Normal Distribution

Contexte

La distribution Pareto est un mélange de deux distributions exponentielles originalement conçue pour étudier des distributions de revenus.

Notation	Parameters	Domain
$X \sim \text{Pareto}(\alpha, \theta)$	$\alpha, \theta > 0$	$x \geq 0$

$f(x)$	$= \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}$
$F(x)$	$= 1 - \left(\frac{\theta}{x+\theta}\right)^\alpha$

- Si $X \sim \text{Pareto}(\alpha, \theta)$ alors $\boxed{Y = (X - d \mid X > d) \sim \text{Pareto}(\alpha, \theta + d)}$.

Uniform Distribution

Notation	Parameters	Domain
$X \sim \text{Beta}(a, b, \theta)$	$a, b > 0$ et $\theta \geq 0$	$x \in [0, \theta]$

$f(x)$	$= \frac{\theta}{\text{B}(a,b)} \left(\frac{x}{\theta}\right)^{a-1} \left(1 - \frac{x}{\theta}\right)^{b-1}$

- $X \sim \text{Beta}(a = 1, b = 1, \theta) \sim \text{Unif}(0, \theta)$.
- Si $X \sim \text{Unif}(a, b)$ alors $\boxed{(X \mid X > d) \sim \text{Unif}(d, b)}$ et $\boxed{(X - d \mid X > d) \sim \text{Unif}(0, b - d)}$.

Pareto Distribution

Contexte

La distribution Pareto est un mélange de deux distributions exponentielles originalement conçue pour étudier des distributions de revenus.

Notation	Parameters	Domain
$X \sim \text{Pareto}(\alpha, \theta)$	$\alpha, \theta > 0$	$x \geq 0$

$f(x)$	$= \dfrac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}$
$F(x)$	$= 1 - \left(\dfrac{\theta}{x+\theta}\right)^\alpha$

- Si $X \sim \text{Pareto}(\alpha, \theta)$ alors $\boxed{Y = (X - d \mid X > d) \sim \text{Pareto}(\alpha, \theta + d)}$.

Exponential Distribution

Gamma Distribution

Notation	Parameters	Domain
$X \sim \text{Gamma}(\alpha, \theta)$	$\alpha, \theta > 0$	$x \geq 0$

$f(x)$	$= \dfrac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}$

- On appelle θ la moyenne et $\lambda = \frac{1}{\theta}$ le paramètre de fréquence (" *rate* ").
- Soit n v.a. indépendantes $\boxed{X_i \sim \text{Gamma}(\alpha_i, \theta)}$ alors $\boxed{\sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \theta)}$.
- Soit n v.a. indépendantes $\boxed{X_i \sim \text{Exp}(\lambda_i)}$ alors $\boxed{Y = \min(X_1, \ldots, X_n) \sim \text{Exp}(\frac{1}{\sum_{i=1}^n \lambda_i})}$.
- Si $X \sim \text{Exp}(\theta)$ alors $\boxed{(X - d \mid X > d) \sim \text{Exp}(\theta)}$.

Statistics

Mode

Context

The mode is the value that occurs the most often. A non-mathematical example of the concept is looking at the most used letter in the English alphabet. The letter E is the most used letter in the dictionary and as such is the mode of the English language.

In mathematical terms, the mode is the point which maximises the PMF/PDF.

Finding the mode of a continuous r.v. can be done by calculating the derivative of the PDF and finding the point where it equals 0. If the distribution is

- **unimodal**, i.e. it has a hump, then $\boxed{\text{mode} = x \text{ s.t. } f'(x) = 0}$.
- strictly increasing or decreasing, the mode will be one of the 2 extremes.
  - For example, the exponential distribution is strictly decreasing and its mode is always 0.

For discrete variables, there are some ways to simplify it's calculation:

- Using the table function on the calculator and seeing where the probabilities peak.
- Using the algebraic approach of looking at p_k/p_{k-1}.
  - $p_k > p_{k-1}$ iff $p_k/p_{k-1} > 1$.
  - The mode is the largest k s.t. $p_k > p_{k-1}$.

Note In the exam, it's best to use the calculator approach.

B

Introduction to Credibility

Basic Framework of Credibility

Context

The ***limitation fluctuation credibility*** approach, or ***classical credibility*** approach, calculates an updated prediction (U) of the **loss measure** as a weighted (Z) average of recent claim experience (D) and a rate (M) specified in the manual. Thus, we calculate the *premium* paid by the *risk group* as $\boxed{U = ZD + (1-Z)M}$.

Notation

M Predicted loss based on the "***manual***".

D Observe**d** losses based on the recent experience of the risk group.

Z Weight assigned to the recent experience D called the ***credibility factor*** with $\boxed{Z \in [0,1]}$.

U **U**pdated prediction of the premium.

Terminology

Risk group block of insurance policies, covered for a period of time upon payment of a *premium*.

Claim frequency The number of claims denoted N.

Claim severity The amount of the i^{th} claim denoted X_i.

Aggregate loss The total loss denoted S where $S = X_1 + X_2 + \ldots + X_N$.

Pure premium The pure premium denoted P where $P = S/E$ with E denoting the number of exposure units.

Exam tips

Typical questions about this involve being given 3 of $M, D, Z,$ and U then finding the missing one.

Context

With $\min\{D, M\} \leq U \leq \max\{D, M\}$, we can see that the credibility factor determines the relative importance of the claim experience of the risk group D relative to the manual rate M.

If $Z = 1$, we obtain *Full Credibility* where the predicted premium depends only on the data ($U = D$). It follows that with $Z < 1$, we obtain *Partial Credibility* as the weighted average of both D and M.

Full Credibility

Contexte

The classical credibility approach determines the ***minimum*** *data size* required for the experience data (D) to be given ***full credibility***. The minimum data size, or ***standard for full credibility***, depends on the **loss measure**.

Claim Frequency

The claim frequency random variable N has mean μ_N and variance σ_N^2.
If we assume $N \approx \mathcal{N}(\mu_N, \sigma_N^2)$, then the probability of observing claim frequency **within** k **of the mean** is $\boxed{\Pr(\mu_N - k\mu_N \leq N \leq \mu_N + k\mu_N) = 2\Phi\left(\frac{k\mu_N}{\sigma_N}\right) - 1}$.

We often assume that the claim frequency $\boxed{N \sim \text{Pois}(\lambda_N)}$ and then apply the normal approximation to find the standard for full credibility for claim frequency $\boxed{\lambda_F}$. First, we impose that the probability of the claim being with k of the mean must be at least $1-\alpha$. Then, we rewrite $\boxed{\frac{k\mu_N}{\sigma_N} = k\sqrt{\lambda_N}}$ and set $\boxed{\lambda_N \geq \left(\frac{z_{1-\alpha/2}}{k}\right)^2}$ where $\boxed{\lambda_F = \left(\frac{z_{1-\alpha/2}}{k}\right)^2}$.

Claim Severity

We assume that the loss amounts $X_1, X_2, \ldots, X_N$ are independent and identically distributed random variables with mean μ_X and variance σ_X^2. Full credibility is attributed to $\boxed{D = \bar{X}}$ if $\boxed{2\Phi\left(\frac{k\mu_X}{\sigma_N/\sqrt{N}}\right) - 1 \geq 1 - \alpha}$.

Similarly to claim frequency, we apply the normal approximation with $\boxed{\bar{X} \approx \mathcal{N}\left(\mu_X, \sigma_X^2/N\right)}$. Then, we find $\boxed{N \geq \left(\frac{z_{1-\alpha/2}}{k}\right)^2 \cdot \left(\frac{\sigma_X}{\mu_X}\right)^2 = \lambda_F CV_X^2}$ where the ***standard for full credibility for claim severity*** is $\lambda_F CV_X^2$.

Aggregate Loss

For the aggregate loss $S = X_1 + X_2 + \ldots + X_N$, we have $\boxed{\mu_S = \mu_N \mu_X}$ and $\boxed{\sigma_S^2 = \mu_N \sigma_X^2 + \mu_X^2 \sigma_N^2}$.

With the same normality assumptions for the Poisson distributed N, we find $\boxed{\lambda_N \geq \left(\frac{z_{1-\alpha/2}}{k}\right)^2 \cdot \left(\frac{\mu_X^2 + \sigma_X^2}{\mu_X^2}\right) = \lambda_F(1 + CV_X^2)}$ where the ***standard for full credibility for claim severity*** is $\lambda_F(1 + CV_X^2)$.

Note The conditions are the same for the ***Pure Premium*** as for the aggregate loss.

Partial Credibility

The ***credibility factor*** for :

Claim Frequency is $\boxed{Z = \sqrt{\frac{\lambda_N}{\lambda_F}}}$.

Claim Severity is $\boxed{Z = \sqrt{\frac{N}{\lambda_F CV_X^2}}}$.

Aggregate Loss and Pure Premium is $\boxed{Z = \sqrt{\frac{\lambda_N}{\lambda_F(1 + CV_X^2)}}}$

Bühlmann Credibility

Context

Buhlmann's approach, a.k.a. the greatest accuracy approach or the least squares approach, estimates the future loss measure X_n

Basic framework

Variance components

Credibility factors

Bayesian Credibility

Basic framework

Premium

Conjugate distributions

Nonparametric empirical Bayes method

C

Linear Mixed Models

D

Bayesian Analysis and Markov Chain Monte Carlo

E

Statistical Learning

K-Nearest Neighbors

Decision Trees

Principal Components Analysis (PCA)

Clustering