Exercise Walkthrough: Insurance Claims and Risk

Justin Lanfermann

25. June of 2025

Abstract

This document provides a step-by-step walkthrough for an exercise on insurance claims and risk management. We will apply core concepts from probability theory, such as the expectation of random variables, the Weak Law of Large Numbers, Chebyshev's inequality, and the Central Limit Theorem. Each step is explained in detail, referencing definitions and theorems from the "Discrete Probability Theory" script by Niki Kilbertus.

1 Problem Statement

An insurance company has $n = 10^4$ clients. For each client i, let X_i be the amount claimed in a year. Assume the claims $X_1, \dots, X_n$ are i.i.d.[1] RVRVs with mean $\mathbb{E}[X_i] = \mu = 500$ and variance $\text{var}[X_i] = \sigma^2 = 800^2 = 640,000$. The company charges each client a premium of Z Euros.

(i) Let $S_n = \sum_{i=1}^{n} X_i$ be the total amount claimed. Calculate the minimum premium Z the company needs to charge to make an expected profit of half a million per year.

(ii) Use the Weak Law of Large Numbers to describe the behavior of the average claim per client $\bar{X}_n = S_n/n$ as n becomes very large.

(iii) The company wants to be at least 95% sure that the total premiums collected will cover the total claims S_n. Use Chebyshev's inequality to assess if the current number of clients $n = 10^4$ is sufficient with a premium of $Z = 550$.

(iv) Now, use the CLT to approximate the probability that the total claims S_n exceed the total premiums $n \cdot Z$. Compare this approximation with the bound obtained from Chebyshev's inequality. Why might the CLT provide a different (and potentially more accurate) value? Which would you trust more?

2 Solution Walkthrough

2.1 Part (i): Minimum Premium for Expected Profit

2.1.1 Overview

The goal is to find the premium Z that results in an expected profit of $500,000$. We will first define the company's profit as a random variable and then compute its expectation using the linearity property of expectation.

2.1.2 Step-by-Step Derivation

1. **Define the Profit Function:** The total income for the company is the sum of all premiums, which is $n \cdot Z$. The total cost is the sum of all claims, $S_n = \sum_{i=1}^{n} X_i$. The profit, let's call it Π, is therefore:
$$\Pi = nZ - S_n$$

2. **Calculate the Expected Profit:** We want the expected profit to be $500,000$. We apply the expectation operator to our profit function:
$$\mathbb{E}[\Pi] = \mathbb{E}[nZ - S_n]$$

3. **Apply Linearity of Expectation:** According to **Proposition 2.4 (i)**[2], expectation is linear. This allows us to write:
$$\mathbb{E}[nZ - S_n] = \mathbb{E}[nZ] - \mathbb{E}[S_n]$$
Since nZ is a constant, its expectation is just itself. So, $\mathbb{E}[\Pi] = nZ - \mathbb{E}[S_n]$.

4. **Calculate the Expected Total Claim:** We also use linearity of expectation to find the expected value of the sum S_n:
$$\mathbb{E}[S_n] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$
Since the claims X_i are identically distributed with mean $\mu = 500$, we have:
$$\mathbb{E}[S_n] = \sum_{i=1}^{n} \mu = n\mu$$
With $n = 10^4$ and $\mu = 500$, we get $\mathbb{E}[S_n] = 10^4 \cdot 500 = 5,000,000$.

5. **Solve for the Premium Z:** Now we set up the final equation and solve for Z:
$$\begin{aligned} \mathbb{E}[\Pi] &= 500,000 \\ nZ - n\mu &= 500,000 \\ 10^4 Z - 5,000,000 &= 500,000 \\ 10^4 Z &= 5,500,000 \\ Z &= \frac{5,500,000}{10,000} = 550 \end{aligned}$$

So, the company must charge a premium of at least 550 per client.

2.1.3 Summary

We formulated the profit as the difference between total premiums and total claims. Using the **linearity of expectation**, we derived an equation linking the premium Z to the desired expected profit and solved it.

2.2 Part (ii): Weak Law of Large Numbers (WLLN)

2.2.1 Overview

The WLLN provides a theoretical guarantee that the average of a large number of i.i.d. random variables converges to the true mean of the distribution. We will state the law and interpret it in the context of the insurance claims.

2.2.2 Explanation

The **Weak Law of Large Numbers (WLLN, Theorem 2.61)**[3] states that for a sequence of pairwise uncorrelated, identically distributed random variables $X_1, X_2, \ldots$ with finite mean μ, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to μ. We write this as:

$$\bar{X}_n \xrightarrow{P} \mu$$

In our scenario, the claims X_i are i.i.d., which is a stronger condition than pairwise uncorrelated, and they have a finite mean $\mu = 500$. Therefore, the WLLN applies.

Interpretation: As the number of clients n grows very large, the average claim per client, $\bar{X}_n$, will get arbitrarily close to the expected claim amount, $\mu = 500$. This means that for a large insurance company, the average cost per client becomes very predictable, even though the claim for any single client is random. The randomness "averages out" over the large pool of clients. This principle is the foundation of the insurance business model.

2.3 Part (iii): Sufficiency Check with Chebyshev's Inequality

2.3.1 Overview

We want to determine if, with a premium of $Z = 550$, the company can be 95% certain of not making a loss. This means we want to check if the probability of a loss, $P(S_n > nZ)$, is less than or equal to 5%. We will use **Chebyshev's inequality** to find an upper bound for this probability.

2.3.2 Step-by-Step Derivation

1. **Formalize the Goal:** We want to check if $P(S_n \leq nZ) \geq 0.95$, which is equivalent to checking if the probability of a loss is at most 5%:

$$P(S_n > nZ) \leq 0.05$$

2. **Recall Chebyshev's Inequality:** From **Theorem 2.40**[4], for a random variable Y with finite mean and variance, and any $\epsilon > 0$:

$$P(|Y - \mathbb{E}[Y]| \geq \epsilon) \leq \frac{\text{var}[Y]}{\epsilon^2}$$

3. **Calculate Mean and Variance of S_n:** We use S_n as our random variable Y. We already found $\mathbb{E}[S_n] = n\mu = 5,000,000$. Now, we find its variance. Since the X_i are independent (a property of i.i.d.[1]), the variance of the sum is the sum of the variances (**Proposition 2.8 (iv)**):

$$\begin{aligned}\text{var}[S_n] = \text{var}\left[\sum_{i=1}^{n} X_i\right] &= \sum_{i=1}^{n} \text{var}[X_i] = n\sigma^2 \\ &= 10^4 \cdot (800)^2 = 10^4 \cdot 640,000 = 6,400,000,000\end{aligned}$$

4. **Relate the Event to Chebyshev's Form:** Our event is $S_n > nZ$. We need to express it in the form $|S_n - \mathbb{E}[S_n]| \geq \epsilon$. First, let's calculate the threshold $nZ = 10^4 \cdot 550 = 5,500,000$. The event is $S_n > 5,500,000$. We can rewrite this in terms of the deviation from the mean:

$$S_n - \mathbb{E}[S_n] > 5,500,000 - 5,000,000 \implies S_n - \mathbb{E}[S_n] > 500,000$$

Let $\epsilon = 500,000$. The event $\{S_n - \mathbb{E}[S_n] > \epsilon\}$ is a subset of the event $\{|S_n - \mathbb{E}[S_n]| \geq \epsilon\}$. Therefore:

$$P(S_n > nZ) = P(S_n - \mathbb{E}[S_n] > \epsilon) \leq P(|S_n - \mathbb{E}[S_n]| \geq \epsilon)$$

5. **Apply the Inequality:** Now we can plug our values into Chebyshev's inequality:

$$P(S_n > 5,500,000) \leq \frac{\text{var}[S_n]}{\epsilon^2} = \frac{6,400,000,000}{(500,000)^2} = \frac{6.4 \times 10^9}{2.5 \times 10^{11}} = 0.0256$$

6. **Conclusion:** The upper bound for the probability of a loss is 0.0256, or 2.56%. Since $2.56\% \leq 5\%$, Chebyshev's inequality confirms that the company is at least 95% sure to cover its claims.

2.3.3 Summary

We used Chebyshev's inequality to find an upper bound on the probability of the total claims exceeding the total premiums. The calculated bound of 2.56% is within the company's 5% risk tolerance, so based on this general-purpose tool, the number of clients is sufficient.

2.4 Part (iv): Approximation with the Central Limit Theorem (CLT)

2.4.1 Overview

The CLT provides a much more precise approximation of the distribution of a sum of i.i.d. random variables than the bound from Chebyshev. We will use it to estimate the same loss probability, $P(S_n > nZ)$, and compare the results.

2.4.2 Step-by-Step Derivation

1. **Recall the Central Limit Theorem (CLT):** From **Theorem 2.64**[5], for a sum S_n of a large number of i.i.d. random variables with mean μ and variance σ^2, the standardized sum converges in distribution to a standard normal distribution:

$$Z_{std} = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\text{var}[S_n]}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

2. **Standardize the Event:** We want to approximate $P(S_n > nZ)$. We perform the same algebraic manipulations on both sides of the inequality to express it in terms of the standardized variable Z_{std}:

$$\begin{aligned} P(S_n > nZ) &= P(S_n - n\mu > nZ - n\mu) \\ &= P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} > \frac{nZ - n\mu}{\sigma\sqrt{n}}\right) \\ &\approx P\left(Z_{std} > \frac{n(Z-\mu)}{\sigma\sqrt{n}}\right) \\ &= P\left(Z_{std} > \frac{\sqrt{n}(Z-\mu)}{\sigma}\right) \end{aligned}$$

3. **Calculate the Value of the Standardized Threshold:** Let's plug in the numbers: $n = 10^4$, $Z = 550$, $\mu = 500$, $\sigma = 800$.

$$\frac{\sqrt{10^4}(550 - 500)}{800} = \frac{100 \cdot 50}{800} = \frac{5000}{800} = 6.25$$

So we need to calculate $P(Z_{std} > 6.25)$.

4. **Evaluate using the Standard Normal CDF:** Let $\Phi(z)$ be the cumulative distribution function (CDF) of the standard normal distribution[6]. Then $P(Z_{std} > z) = 1 - P(Z_{std} \leq z) = 1 - \Phi(z)$.

$$P(S_n > nZ) \approx P(Z_{std} > 6.25) = 1 - \Phi(6.25)$$

Using a calculator (like WolframAlpha), we find that $\Phi(6.25)$ is extremely close to 1. The probability is:

$$1 - \Phi(6.25) \approx 1 - 0.99999999979... \approx 2.02 \times 10^{-10}$$

2.4.3 Comparison and Conclusion

- **Chebyshev's Bound:** $P(\text{loss}) \leq 0.0256$ (or 2.56%).
- **CLT Approximation:** $P(\text{loss}) \approx 2.02 \times 10^{-10}$ (practically zero).

Why are they different? Chebyshev's inequality is a universal, worst-case bound. It holds for *any* distribution with a given finite mean and variance, no matter how strangely shaped. Because it makes so few assumptions, its bound is often very loose (pessimistic).

The Central Limit Theorem, however, uses much more information. It leverages the fact that we are summing a *large number of independent and identically distributed* variables. The CLT tells us that the shape of the resulting distribution of the sum will be approximately a Normal (Gaussian) distribution, regardless of the shape of the original distribution of a single claim X_i.

Which to trust more? For a large number of clients like $n = 10^4$, the CLT approximation is far more reliable and accurate. The rule of thumb for the CLT to be a good approximation is often cited as $n \geq 30$, and we are well beyond that. The CLT gives a realistic estimate of the risk, whereas Chebyshev gives a mathematically guaranteed but practically oversized upper limit. A business would make decisions based on the CLT's estimate, as it provides a much more accurate picture of the actual risk involved. The probability of a loss is, for all practical purposes, zero.

In-depth Explanations

[1] I.i.d. Random Variables (Definition 1.74): "Independent and Identically Distributed".

- **Identically Distributed:** All random variables X_i in the sequence follow the same probability distribution. In this case, every client's claim is assumed to come from a distribution with $\mu = 500$ and $\sigma = 800$.
- **Independent:** The outcome of one random variable does not influence the outcome of another. In this case, the claim amount of one client is unrelated to the claim amount of any other client. This is crucial for calculating the variance of the sum: $\text{var}[S_n] = \sum \text{var}[X_i]$ only holds if the variables are uncorrelated (and independence implies uncorrelatedness).

[2] Linearity of Expectation (Proposition 2.4 (i)): For any two random variables X and Y and any constants $a, b \in \mathbb{R}$, the expectation operator has the property:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Crucially, this property holds regardless of whether X and Y are independent. It is one of the most powerful and frequently used properties of expectation.

[3] Weak Law of Large Numbers (Theorem 2.61): The WLLN formalizes the intuition of "averaging out". It states that the sample mean $\bar{X}_n$ converges *in probability* to the true mean μ. Convergence in probability (**Definition 2.56 (i)**) means that for any arbitrarily small tolerance $\epsilon > 0$, the probability that the sample mean differs from the true mean by more than ϵ approaches zero as the sample size n goes to infinity.

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

[4] Chebyshev's Inequality (Theorem 2.40): This theorem provides a universal (but often loose) bound on the probability that a random variable deviates from its mean by a certain amount. For a random variable Y with mean $\mathbb{E}[Y]$ and variance $\text{var}[Y]$, and for any $\epsilon > 0$:

$$P(|Y - \mathbb{E}[Y]| \geq \epsilon) \leq \frac{\text{var}[Y]}{\epsilon^2}$$

Its strength is its generality; it requires no knowledge of the distribution of Y other than its mean and variance.

[5] Central Limit Theorem (Theorem 2.64): The CLT is one of the cornerstone results of probability. It states that the sum (or average) of a large number of i.i.d. random variables, when properly standardized, will be approximately normally distributed, regardless of the underlying distribution of the individual variables. This explains why the normal (Gaussian) distribution appears so frequently in nature and statistics. It describes convergence *in distribution* (**Definition 2.56 (iii)**), meaning the CDF of the standardized sum approaches the CDF of a standard normal variable.

[6] Standard Normal CDF $\Phi(z)$: The function $\Phi(z)$ denotes the Cumulative Distribution Function (CDF) of a standard normal random variable $Z_{std} \sim \mathcal{N}(0, 1)$. It gives the probability that the variable will take a value less than or equal to z.

$$\Phi(z) = P(Z_{std} \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Graphically, it represents the area under the standard "bell curve" to the left of the point z.