Case Study: Heavy-Tailed Distribution and Reinsurance Rate-making

October 21, 2016

The purpose of this case study is to give a brief introduction to a heavy-tailed distribution and its distinct behaviors in contrast with familiar light-tailed distributions in standard texts. You will learn about QQ-plot, which is a popular tool for checking goodness-of-fit for a particular statistical model. You will also work on a real-life application of heavy-tailed distributions in reinsurance rate-making.

Reinsurance is a very important component of the global financial market. It allows insurers to take on risks that they would otherwise not be able to. Did you know that NASA buys insurance contracts for every rocket it launches and every satellite and probe it sends to the outer space? These equipments are so expensive that typical insurers would not be able to cover on their own. Therefore, they can go to the reinsurance market, slide up the coverage and transfer partial coverage to reinsurers that exceed their financial capabilities. By the end of this case study, you will be able to learn basic principles of pricing an reinsurance contract.

Learning Objectives:

- Visualize the concepts in the Central Limit Theorem;
- Identify cases where the Central Limit Theorem does not apply;
- Reinforce the concept of cumulative distribution function;
- Understand why and how QQ-plot works for the assessment of goodness-of-fit;
- Reinforce the concepts of conditional probability and conditional expectation;
- Apply basic integration technique to compute mean excess function;
- Learn about behaviors of a heavy-tailed distribution;
- Learn how to use order statistics to estimate quantiles and mean excess function;
- Develop intuition behind point estimators.

1 Background

1.1 Central limit theorem

This section is to provide visualization of central limit theorem which you should already be familiar with. We provide examples on both discrete random variable and continuous random variable.

Example 1.1. (Bernoulli random variables) Suppose we intend to test the fairness of a coin, i.e. whether the coin has equal chance of landing on a head or a tail. We can do so by counting the number of heads in a sequence of coin tosses. The number of heads in each toss is a Bernoulli random variable, denoted by X_1. Let p be the probability of a head and $q = 1 - p$ be the probability of a tail. Then, its probability mass function is given by

$$\mathbb{P}(X_1 = x) = \begin{cases} p, & x = 1, \\ q, & x = 0. \end{cases}$$

We let X_k be the number of heads in the k-th coin toss, $k = 1, 2, \cdots, n$. Then we count the total number of heads after n coin tosses.

$$S_n := \sum_{k=1}^{n} X_k.$$

Then it is easy to show that S_n is a binomial random variable with parameters n and p and its probability mass function is given by

$$\mathbb{P}(S_n = x) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, \cdots, n.$$

For example, suppose that we have an unfair coin with $p = 0.3$. Figure 1 shows the probability mass functions of the number of heads, S_n, where the number of coin tosses $n = 1, 2, 3, 10, 20, 50$. Since $p < 0.5$, we are more likely to see a smaller number of heads than that of tails in any given n tosses. In general, the probability mass function of S_n tends to skew towards to the right. However, as one can see in the later graphs in Figure 1, the probability mass function becomes more and more symmetric[1] as n gets bigger and bigger. This phenomenon is present for any $p \in (0, 1)$, no matter how extreme is p. Why is this happening? The answer is the **Central Limit Theorem**, which we have already learned in class.

Let us consider the sample average

$$\overline{X}_n := \frac{1}{n} S_n$$

[1] Note, however, this is not to suggest that the coin becomes fair.

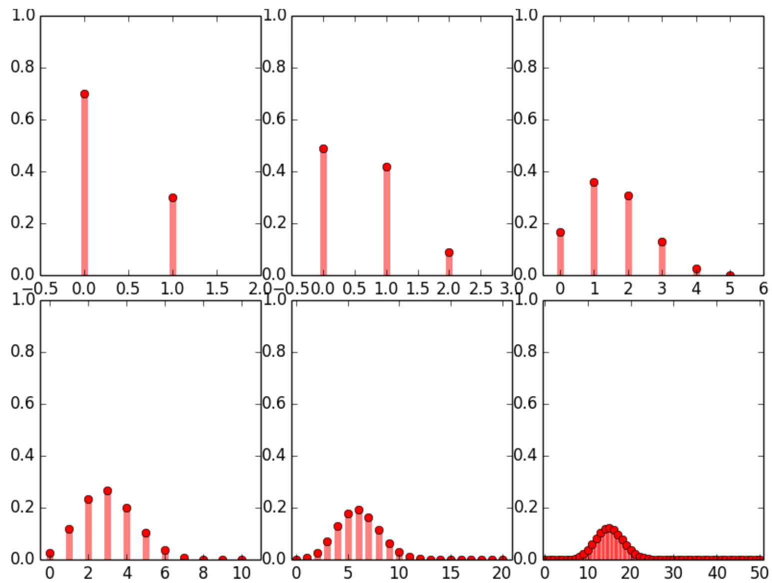

Figure 1: Probability mass function of S_n

Since the expectation of the sample average

$$\mathbb{E}(\overline{X}_n) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(X_k) = p,$$

the sample mean provides an unbiased estimator of the unknown parameter of fairness p.

Exercise 1.1. The Central Limit Theorem tells us that the estimator $\overline{X}_n$ is asymptotically normal. In particular, we can construct the following random variable, Y_n, such that

$$Y_n := \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{pq}} \longrightarrow \mathcal{N}(0,1),$$

where $\mathcal{N}(\mu, \sigma^2)$ is a normal random variable with mean μ and variance σ^2. Explain why the estimator $\overline{X}_n$ behaves roughly like $\mathcal{N}\left(p, \frac{pq}{n}\right)$.

[**HINT:** First, note that Y_n is approximately $\mathcal{N}(0,1)$. Now, reformulate the given equation to express $\overline{X}_n$ as a function of Y_n; what kind of function do you get? Combine these two facts to determine the distribution of $\overline{X}_n$ by computing $\mathbb{E}(\overline{X}_n)$ and $\mathrm{Var}(\overline{X}_n)$.]

As the sample size n gets big, the variance is so small that the sample average gives very good estimate of the actual parameter p. That is why in practice we use the value of $\overline{X}_n$ as an estimate, despite the fact that it is in fact a random variable.

Exercise 1.2. What is the exact distribution of $\overline{X}_n$?

[**HINT:** First, determine the set of all values of X_n that can occur. Then, note that $\overline{X}_n$ as a function of S_n, and apply the probability mass function of S_n to derive the probability mass function of $\overline{X}_n$. As a reminder, your probability mass function should give $\mathbb{P}(\overline{X}_n = k)$ for all possible outcomes of $\overline{X}_n$.]

Let us verify numerically the conclusion of Central Limit Theorem. Similar to what you showed in Exercise 1.2, one can show that the exact probability mass function of Y_n is given by

$$\mathbb{P}(Y_n = y) = \binom{n}{h} p^h q^{n-h}, \qquad h := \sqrt{npq}y + np,$$

where $y = (k - np)/\sqrt{npq}$ for $k = 0, 1, \cdots, n$.

We can draw graphs of the probability mass functions and see how they converge to a normal distribution as n increases. Figure 2 below is an illustration of the central limiting theorem. The blue bars visually depict how a point mass function of a binomial random variable behave over an interval. The red dashed lines indicate the normal density function. From left to right, top to bottom we have the densities for binomial random variables with sample size n=1,2,5,20,100,1000 respectively, with probability of success being once again 30%.

Example 1.2. (Exponential random variables) Recall that the probability density function of an exponential random variable is given by

$$f(t) = \lambda e^{-\lambda t}, \qquad t \geq 0$$

where $\mathbb{E}(X_i) = 1/\lambda$. If we redefine $\overline{X}_n$, Y_n, and S_n according to this new random variable, then the Central Limit Theorem tells us that

$$Y_n := \sqrt{n}(\lambda \overline{X}_n - 1) \longrightarrow \mathcal{N}(0, 1).$$

Let us compute this analytically.

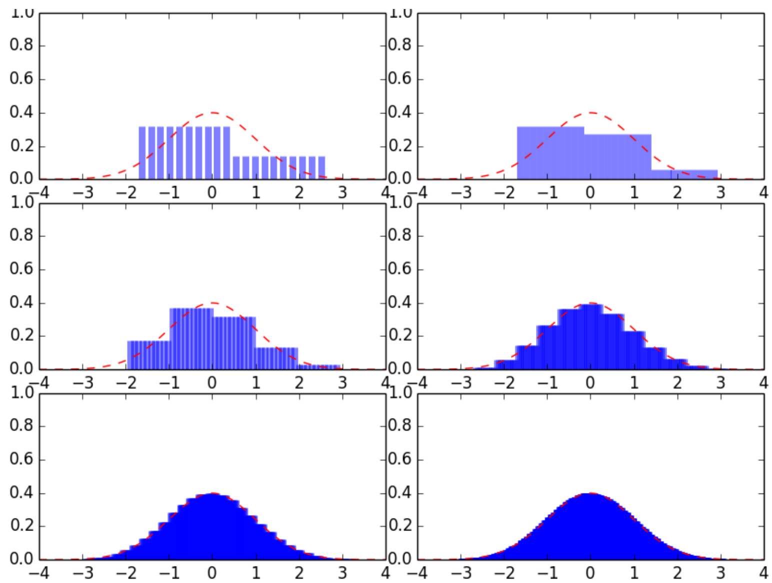

Figure 2: Binomial probability mass converging to normal density

Exercise 1.3. It can be shown that the probability density function of S_n is given by

$$f_{S_n}(x) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x}, \qquad x \geq 0 \tag{1}$$

Using this density function, show that the probability density function of Y_n is

$$f_{Y_n}(y) = \frac{n^{n-1/2}}{(n-1)!} \left(1 + \frac{y}{\sqrt{n}}\right)^{n-1} \exp\left\{-n\left(1 + \frac{y}{\sqrt{n}}\right)\right\}, \qquad y > -\sqrt{n} \tag{2}$$

[**HINT:** First, write Y_n as a function of S_n. Then, use this function to determine both (a) the set of possible outcomes for Y_n based on the set of potential outcomes for S_n, and (b) the probability density function of Y_n over these outcomes, based on the probability density function of S_n.]

Figure 3 is a visual illustration of the probability density function of Y_n for various choices of n. We can see how the probability density function for Y_n converges to the standard normal density function. Again the red dashed lines is the standard normal density function while the blue lines are the densities for Y_n, given above, for $n = 1, 2, 3, 5, 10, 100$.

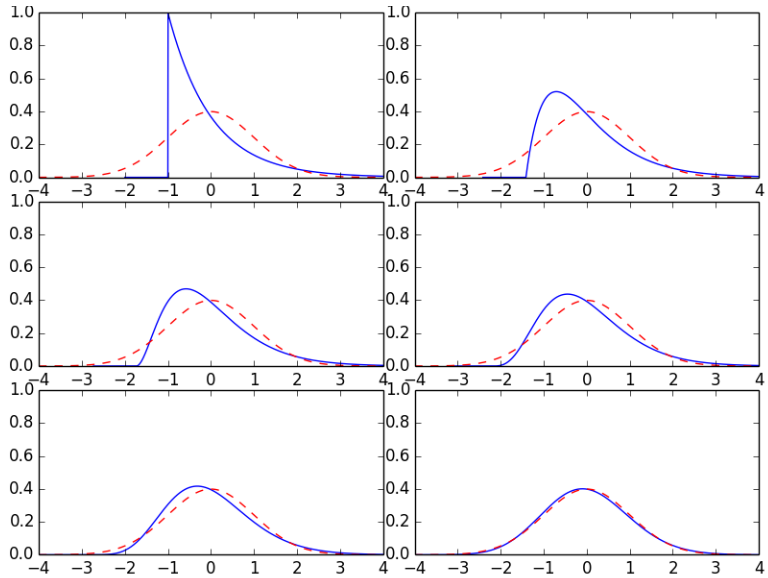

Figure 3: Exponential densities converging to normal density

Exercise 1.4. Use *matplotlib* in Python to create plots showing how $f_{Y_n}(y)$ converges to the normal distribution as n increases. Your plots should contain an outline of the normal distribution like Figure 3. Use $n = 1, 2, 10$.

In general, we can conclude that *Central Limit Theorem* tells us that the distribution of an average tends to be Normal, even when the distribution from which the average is computed is non-Normal. However, there are certain cases in which the average deviate from Normal behavior. When does the Central Limit Theorem not hold?

1.2 Heavy-tailed Distributions

Let us introduce heavy-tailed distributions which are probability distributions with a heavier tail than the exponential. We will see how the extremes produced by heavy-tailed distributions will corrupt the average so that an asymptotic behavior different from the Normal behavior is obtained. Formally, a random variable X is said to have a **heavy-tailed** if

$$\lim_{x\to\infty} \frac{e^{-\lambda x}}{\overline{F}(x)} = 0, \qquad \text{for all } \lambda > 0 \tag{3}$$

where $\overline{F}(x) := \mathbb{P}(X > x)$ (here, $\overline{F}(x)$ is often referred to as a **survival function**).

An example of a heavy-tailed distribution is the Pareto distribution. Consider the *strict Pareto* random variable whose density is given by

$$f(x) = \alpha x^{-\alpha-1}, \qquad x > 1$$

where α is a positive number, called the *Pareto index*. The Pareto distribution is very important in reinsurance so we will study it closely.

Exercise 1.5. Show that the Pareto may not have finite mean or variance by calculating the mean and variance of a strict Pareto random variable. Are there any values of α for which either does not take a finite value?

[**HINT:** Examine the case when $\alpha \in (0, 1]$ and $\alpha > 1$]

Exercise 1.6. The Pareto distribution is closely related to the exponential. Given X is a strict Pareto random variable, show that $Y = \ln(X)$ is exponentially distributed with mean $\frac{1}{\alpha}$.

Exercise 1.7. Using the definition of heavy-tailed from Equation 3, show the following:

(a) A strict Pareto random variable is heavy-tailed.

(b) An exponential random variable with rate α is not heavy-tailed.

[**HINT**: You may need to apply L'Hospitôl's rule when taking the limit for (a). For part (b), consider selecting $\lambda < \alpha$.]

References

[1] Hill, B.M. (1975). A simple approach to inference about the tail of a distribution. *Annals of Statistics*. **3**: 1163–1174.