Probability Theory and Statistics Notebook I

Zhang Ruiyang

Contents

Preface

Part I

Probability Theory

Chapter 1

Fundamentals

Chapter 2

Random Variables

Chapter 3

Common Distributions

Chapter 4

Limit Theorem

4.1 Common Inequalities and Convergence

Inequalities and types of convergence are essential for the development of law of large numbers, central limit theorem and many other identities in the study of Probability Theory and Statistics.

The importance of inequalities is that they enable us to derive bounds on probabilities when only the mean or both the mean and the variance of the probability distribution are known. Of course, if the actual distribution were known, then the desired probabilities could be computed exactly and we would not need to resort to bounds. However, this kind of case is, in reality, quite rare. Statistical theory is literally brimming with inequalities and identities.

Convergence is more commonly used in the study of theory of probability. It formalises the idea that a sequence of essentially random or unpredictable events can sometimes be expected to settle down into a behaviour that is essentially unchanging when items far enough into the sequence are studied.

–

We will be talking about Markov inequality as well as Chebyshev inequality in this notebook. There are other kinds of useful inequalities, but for the purpose of this notebook, we will mostly use these two. Other kinds of inequalities will be brought up separately when there is such need.

We will start with **Markov inequality**, which is named after Russian Mathematician Andrey Markov. Markov's name will appear again in Chapter 6 when we talk about Markov chains.

First, we have a nonnegative random variable X. Let us define an indicator function I.

For $a > 0$, let

$$I = \begin{cases} 1 \text{ if } X \geq a \\ 0 \text{ otherwise} \end{cases}$$

and note that, since $X > 0$,

$$I \leq \frac{X}{a}.$$

Taking the expectations of the preceding inequality yields

$$\mathrm{E}[I] = P(X \geq a) \geq \frac{\mathrm{E}[X]}{a}.$$

This is the Markov inequality. If X is a random variable that takes only nonnegative values, then for any value $a > 0$,

$$P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}.$$

We can extend it from solely the random variable to monotonically increasing functions. If φ is a monotonically increasing nonnegative function for the nonnegative real numbers, X is a random variable, $a \geq 0$ and $\varphi(a) > 0$, then

$$P(|X| \geq a) \leq \frac{\mathrm{E}[\varphi(|X|)]}{\varphi(a)}.$$

This gives us the extended Markov inequality.

We can also use higher moments of X as the monotonically increasing nonnegative function. This gives us

$$P(|X| \geq a) \leq \frac{|X|^n}{a^n}.$$

Another useful corollary of the Markov inequality is the **Chernoff bound** named after Herman Chernoff. The statement of Markov inequality involves $P(X \geq a)$, and if we raise both X and a to e^t, we get $P(e^{tX} \geq e^{ta})$ and the inequality becomes

$$P(e^{tX} \geq e^{ta}) \leq \frac{\mathrm{E}[e^{tX}]}{e^{ta}} = e^{-ta}\mathrm{E}[e^{tX}]$$

which is the statement of Chernoff bound for positive t.

–

Markov inequality has a special case, and that is known as the **Chebyshev inequality**, named after Russian Mathematician Pafnuty Chebyshev.[1] This is the most famous,

[1] Fun fact, Chebyshev is Markov's advisor at Saint Petersburg State University in Russia.

and perhaps most useful, probability inequality. Its usefulness comes from its wide applicability.

To construct it, let us first have a random variable X with finite mean μ and variance σ^2. We notice that $(X-\mu)^2$ is a nonnegative random variable. By applying Markov Inequality with $a = k^2$, we obtain

$$P((X-\mu)^2 \geq k^2) \leq \frac{\mathrm{E}[(X-\mu)^2]}{k^2}.$$

It is not hard to see that $(X-\mu)^2 \geq k^2$ and $|X-\mu| > k$ are two equivalent statements. So the above equation can be rewritten as

$$P(|X-\mu| > k) \leq \frac{\mathrm{E}[(X-\mu)^2]}{k^2} = \frac{\sigma^2}{k^2}.$$

So, if X is a random variable with finite mean μ and variance σ^2, then for any value $k > 0$, we get the Chebyshev inequality

$$P(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}.$$

Chebyshev inequality does not give very good bounds, but this is based on the fact that minimum conditions are required. We can get tighter bounds if we add in more conditions for different circumstances. Also, many other inequalities are similar in spirit to Chebychev's.

–

We will be talking about three kinds of convergence here. They are convergence in distribution, convergence in probability, and almost surely convergence.

The weakest form of convergence is **convergence in distribution**. Other names of it include convergence in law and converge weakly. This is the weakest since all the other types of convergence can imply this statement. This form of convergence essentially means we increasingly expect to see the next outcome in a sequence of random experiments becoming better and better modelled by a given probability distribution. With that, it is not hard to notice the statement.

A sequence $X_1, X_2 \ldots$ of real valued random variables is said to converge in distribution to a random variable X if

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for every number $x \in \mathbb{R}$ at which F is continuous. Here, F_n and F are CDFs of random variable X_n and X respectively.

The next form of convergence is **convergence in probability**. This basically means the probability of an "unusual" outcome becomes smaller and smaller as the sequence progresses. The formal statement is below.

A sequence $X_1, X_2 \dots$ of real valued random variables is said to converge in probability to a random variable X if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

The last type of convergence we will be talking about here is that of **convergence almost surely**. It is sometimes denoted as convergence a.s. in short. This is similar to pointwise convergence in analysis, where the convergence happens for every pair of points. The formal statement of convergence almost surely is the following.

A sequence $X_1, X_2 \dots$ of real valued random variables is said to converge almost surely to a random variable X if

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

We should note that convergence almost surely implies convergence in probability, and convergence in probability implies convergence in distribution.

4.2 Law of Large Number

In a more intuitive manner, the law of large number (LLN) describes that when an experiment is repeated for many times, the average of these results will converge to the expectation value of the experiment.

Consider a sequence of independent and identically distributed (i.i.d.) random variables $X_1, X_2, \dots,$, each with mean μ and variance σ^2. The sample mean is $\frac{X_1+\cdots+X_n}{n}$, which gives us

$$\mathrm{E}[\frac{X_1 + \cdots + X_n}{n}] = \frac{n\mu}{n} = \mu$$

and

$$\mathrm{Var}(\frac{X_1 + \cdots + X_n}{n}) = \frac{\sigma^2 n}{n^2} = \frac{\sigma^2}{n}$$

since they are independent.

We apply Chebyshev inequality and get

$$P\Big\{\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| \geq \varepsilon\Big\} \leq \frac{\sigma^2}{n\varepsilon^2}.$$

This gives us the **weak law of large number**. Let $X_1, X_2, \dots,$ be a sequence of independent and identically distributed random variables with mean μ. For any $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P\Big\{\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| \geq \varepsilon\Big\} = 0$$

or equivalently,

$$\lim_{n\to\infty} P\Big\{\big|\frac{X_1+\cdots+X_n}{n}-\mu\big|<\varepsilon\Big\}=1.$$

The weak law of large numbers was originally proven by Jakob Bernoulli for the special case where the X_i are Bernoulli random variables. His statement and proof of this theorem were presented in his book *Ars Conjectandi*. Note that because Chebyshev's inequality was not known in Bernoulli's time, Bernoulli had to resort to a quite ingenious proof to establish the result. The general form of the weak law of large numbers presented just now was proved by the Soviet Mathematician Aleksandr Khinchin.

We have been dealing with identical and independent random variables earlier on. It is possible to remove the identical assumption while having a similar large number law.

Suppose we have a sequence of independent random variables $X_1, X_2, \ldots$, each X_j has mean μ_j and variance σ_j^2. Moreover, let us suppose there exists a constant $M<\infty$ such that for all j, $\sigma_j^2 \le M$.

Let us centralise all the random variables to get $X_j^0 = X_j - \mu_j$ and $S_n^0 = \sum_{j=1}^n X_j - \mu_j$. It is not hard to see that $\mathrm{E}[S_n^0]=0$ as expectation equals to mean. In addition, due to independence, we have

$$\mathrm{E}[(S_n^0)^2] = \mathrm{Var}(S_n^0) = \sum_{j=1}^n \mathrm{Var}(X_j) = \sum_{j=1}^n \sigma_j^2.$$

From the assumptions, we have $\sigma_j^2 \le M$. This, along with the above equation, implies $\mathrm{E}[(S_n^0)^2] \le nM$ and then $\mathrm{E}[(\frac{S_n^0}{n})^2] \le \frac{M}{n}$.

Recall from last section the extended Markov inequality for higher moment. We will apply the inequality for second moment and get

$$\lim_{n\to\infty} P\Big(\Big|\frac{S_n^0}{n}\Big| \ge c\Big) \le \lim_{n\to\infty} \frac{\mathrm{E}[(\frac{S_n^0}{n})^2]}{c^2} \le \lim_{n\to\infty} \frac{M}{nc^2} = 0.$$

By changing S_n^0 to $\sum_{j=1}^n X_j$ and taking the complement on both sides, we get

$$\lim_{n\to\infty} P\Big(\Big|\frac{X_1+\cdots+X_n}{n} - \frac{\mu_1+\cdots+\mu_n}{n}\Big| < c\Big) = 1,$$

which is the **extended weak law of large number**.

There is a stronger version of the weak law, called the **strong law of large number**. The proof of this law will be omitted in this notebook, and it will be placed in *Probability Theory and Statistics Notebook II*. The statement of the law is the following.

Let $X_1, X_2, \ldots$, be a sequence of i.i.d. random variables with mean μ. We have

$$P\Big(\lim_{n\to\infty} \frac{X_1+\cdots+X_n}{n} = \mu\Big) = 1.$$

If we recall from the previous section the different types of convergence, we can notice that the strong law converges almost surely while the weak law converges in probability. The difference is quite subtle, and it cannot be fully explained without the usage of measure theory. Briefly speaking, converge in probability means that the chance for the random variable to not behave like X decreases to limit 0 as n increases to limit ∞; converge almost surely means that the chance for the random variable to not behave like X is 0 after a certain value of n. Of course, this is only a rough explanation that does not involve measure 0 and other important things. The readers are encouraged to keep this question in mind and come back to it when they are doing measure-theoretic Probability Theory.

4.3 Central Limit Theorem

In some situations, when independent random variables are added, their properly normalised sum tends toward a normal distribution even if the original variables themselves are not normally distributed. This is due to **central limit theorem**. Readers should already have a brief sense of the power of this theorem from the derivation of normal distribution we included in the previous Chapter.

Central limit theorem was first established as a way to approximate normal distribution from binomial distribution by de Moivre and Laplace, in the form of the de Moivre-Laplace central limit theorem we mentioned earlier on in Chapter 3.4.

The theorem was generalised in 1901 by the Russian Mathematician Aleksandr Liapounov by introducing its characteristic functions. It was then extended by Markov for the case of independent variables.

Central limit theorem was originally called 'limit theorem', and it was added the prefix 'central' due to the central position it has in the study of Probability Theory by George Pólya.

The statement of the theorem will be stated without proof here as it involves more advanced Mathematics than the level of this notebook. It will, like many other more advanced concepts, be included in the next notebook *Probability Theory and Statistics Notebook II*.

Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables, each having mean μ and variance σ^2. Then the distribution of

$$\frac{X_1 + \cdots X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is, for $-\infty < a < \infty$,

$$\lim_{n\to\infty} P(\frac{X_1 + \cdots X_n - n\mu}{\sigma\sqrt{n}} \leq a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx.$$