Lecture 10 – The law of large numbers MATH-UA.0235 Probability and Statistics – Antoine Cerfon

Scientists have long understood that to obtain improved accuracy for the measurement of a physical quantity, a reasonably strategy is to repeat the measurement in an independent way a large number of times, and then average over all results obtained.

Likewise, we all intuitively see that if we want to find out whether a coin is fair or not, it is more reasonable to toss the coin a large number of times and compute the average number of heads, rather than just decide based on one toss.

1 Useful inequalities

1.1 Markov inequality

Let us consider a discrete random variable X taking values $a_1, a_2, \ldots$, with $\forall\, i \in \mathbb{N}^*,\ a_i \geq 0$. Let $a \in \mathbb{R}$ such that $a \geq 0$.

$$E[X] = \sum_i a_i p_X(a_i) \geq \sum_{i, a_i \geq a} a_i p_X(a_i) \geq \sum_{i, a_i \geq a} a p_X(a_i) = a \sum_{i, a_i \geq a} p_X(a_i) = aP(X \geq a)$$

We just showed that for any $a \in \mathbb{R}$ such that $a \geq 0$,

$$\underline{E[X] \geq aP(X \geq a)}$$

This inequality is known as *Markov's inequality.* It is a straightforward exercise to show that it also holds for continuous random variables, which I recommend as practice.

Intuitively, Markov's inequality says that if the mean of a random variable is small, then the probability that the random variable is larger than a given number is also small.

A standard application of Markov's inequality concerns income distribution. Specifically, let us say that the average income in a given country is E[X]=50,000\$. Then the probability of having an income of 200,000\$ is at most $\frac{1}{4}$ (it can be less than that depending on the distribution of X).

In this lecture, we use Markov's inequality to prove another inequality which we will use for the law of large numbers, and which is called *Chebyshev's inequality.*

1.2 Chebyshev's inequality

Let X be a discrete random variable, with expectation $E[X] = \mu$, and $a^2 \in \mathbb{R}$.

We note that $a^2 \geq 0$. By Markov's inequality, we then have

$$\begin{aligned} &E[(X-\mu)^2] \geq a^2 P((X-\mu)^2 \geq a^2) \\ \Leftrightarrow &\mathrm{Var}(X) \geq a^2 P(|X-\mu| \geq |a|) \end{aligned}$$

If $a \neq 0$, we can therefore write

$$\boxed{P(|X - E[X]| \geq |a|) \leq \frac{1}{a^2}\mathrm{Var}(X)} \tag{1}$$

This is what we call *Chebyshev's inequality.* It tells us that **most of the probability mass of a random variable is within a few standard deviations from its expectation**.

Indeed, let $\mathrm{Var}(X) = \sigma^2$, so that the standard deviation is σ. Consider $a = k\sigma$ for $k = 1, 2, 3, 4, 5, \ldots$. According to Chebyshev's inequality,

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2\sigma^2}\sigma^2 = \frac{1}{k^2}$$

The probability that X is more than 3 standard deviations away from the mean μ is less than $\frac{1}{9}$, the probability that it is more than 4 standard deviations away is less than $\frac{1}{16}$, and the probabilities keep getting smaller fast.

2 The law of large numbers

We now have all the tools needed to derive the central result of this section, namely the **law of large numbers**.

2.1 Averages vary less

Let X_1, X_2, $\ldots$, X_n be a sequence of **independent and identically distributed** random variables, each with expected value μ and variance σ^2. We say that the X_1, X_2, $\ldots$, X_n are **i.i.d.**, or **iid**, or **IID**. (In the context of the motivation given in the introdution of this section, you may view each X_i as the result of one of the independent experiments.)

Define now the random variable $\overline{X_n}$, called a **sample mean**, and given by

$$\boxed{\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}}$$

By the linearity of expectation,

$$E[\overline{X_n}] = E\left[\frac{X_1}{n}\right] + E\left[\frac{X_2}{n}\right] + \ldots + E\left[\frac{X_n}{n}\right] = \frac{1}{n}\left(E[X_1] + E[X_2] + \ldots + E[X_n]\right) = \frac{n\mu}{n} = \mu$$

We just showed that the sample mean $\overline{X_n}$ **has the same mean as each of the** X_i.

Furthermore, since the X_i are independent of one another, we can write

$$\begin{aligned}\mathrm{Var}(\overline{X_n}) = \mathrm{Var}\left(\frac{X_1}{n} + \frac{X_2}{n} + \ldots + \frac{X_n}{n}\right) &= \mathrm{Var}\left(\frac{X_1}{n}\right) + \mathrm{Var}\left(\frac{X_2}{n}\right) + \ldots + \mathrm{Var}\left(\frac{X_n}{n}\right) \\ &= \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)\right) \\ &= \frac{\sigma^2 + \sigma^2 + \ldots + \sigma^2}{n^2} = \frac{\sigma^2}{n}\end{aligned}$$

Thus,

$$\mathrm{Var}(\overline{X_n}) = \frac{\sigma^2}{n}$$

The standard deviation of the sample mean $\overline{X_n}$ is smaller by a factor of $\sqrt{n}$ as compared to the standard deviation of each of the X_i.

This is why scientists rely on a large number of identical and independent experiments to improve the accuracy of their results.

2.2 The law of large numbers

We just showed that the sample mean of n i.i.d. random variables with expectation μ also had expectation μ, and varied much less about μ.

Furthermore, the larger n, the narrower the distribution of $\overline{X_n}$, which becomes arbitrarily narrow in the limit $n \to +\infty$. This can be shown explicitly using Chebyshev's inequality. Indeed, for any $\epsilon > 0$,

$$P(|\overline{X_n} - \mu| > \epsilon) \leq \frac{1}{\epsilon^2}\mathrm{Var}(\overline{X_n}) = \frac{\sigma^2}{n\epsilon^2}$$

The inequality above was obtained by direct application of Chebyshev's inequality.

$P(|\overline{X_n} - \mu| > \epsilon)$ is the probability that the distance to the mean is greater than ϵ, and since ϵ is a given, fixed number, we observe that

$$\lim_{n \to +\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

The law of large numbers can therefore be stated as follows:

If $\overline{X_n}$ is the average of n independent random variables each with expectation μ and variance σ^2, then for any $\epsilon > 0$ (which should be seen as a small number for the theorem to be truly meaningful),

$$\lim_{n \to +\infty} P(|\overline{X_n} - \mu| > \epsilon) = 0$$

Note that strictly speaking, this result is called the *weak law of large numbers.*

There exists a stronger version of the same result, called the *strong law of large numbers.* The intuitive interpretation of what the laws imply is the same in both cases. The difference between the two laws is in how we mathematically take the limit $n \to +\infty$, and in different ways for random variables to converge to a value. These mathematical subtleties are beyond the scope of this course.

2.3 Limitation of the law of large numbers

In deriving the law of large numbers, **we assumed that $E[X] = \mu$ existed and was finite**. This is not always the case: in Lecture 5, we showed that for the Cauchy distribution, the expectation was not defined, and in the textbook, it is shown that Pareto distributions can have $+\infty$ as expectation.

When $E[X]$ is not defined or not finite, the law of large numbers does not apply, as shown in page 187 of the textbook.

2.4 Practical illustration of Chebyshev's inequality and the law of large numbers

A pollster would like to predict the turnout in an upcoming election. It is obviously impossible to phone every single citizen of voting age to find out whether she or he will vote. So the pollster decides to use the amazing math she recently learned to obtain a reasonable estimate as follows.

- Phone n randomly picked citizens of voting age
- For each person i contacted, associate the Bernoulli random variable X_i defined according to:

$$X_i = \begin{cases} 1 \text{ if "Yes, I will vote on Tuesday"} \\ 0 \text{ if "No, I will not vote on Tuesday"} \end{cases}$$

- Then

$$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}$$

  is an estimate of the fraction of citizens of voting age who will vote on Tuesday.

- The pollster will make the approximation that the X_i are all independent, and identically distributed, with expected value μ, and variance $\sigma^2 = \mu(1-\mu)$, where where μ can be interpreted as the fraction of citizens who will vote.
- The question for the pollster is: how large should n be to be 95% certain that she will have an error of less than 1%?

  In this question, the first aspect – 95% certain – is what we call the **confidence interval**; the second aspect – an error of less than 1% – is what we call the **accuracy** of the estimation. Mathematically, this is written as:

$$P(|\overline{X_n} - \mu| \geq 0.01) \leq 0.05$$

- Chebyshev's inequality gives us

$$P(|\overline{X_n} - \mu| \geq 0.01) \leq \frac{\sigma^2}{n \cdot 10^{-4}}$$

- Since the pollster does not know μ – this is what she is looking for –, she makes a conservative estimate. Specifically, from elementary calculus, we know that

$$\forall \mu \in [0,1] \ , \ \mu(1-\mu) \leq \frac{1}{4}$$

Hence, she makes the conservative estimate that

$$\sigma^2 \leq \frac{1}{4}$$

This allows her to say

$$P(|\overline{X_n} - \mu| \geq 0.01) \leq \frac{1}{4 \cdot 10^{-4} n}$$

Thus, according to her calculation, to be 95% certain that she will have an error of less than 1%, n should be such that

$$\frac{1}{4 \cdot 10^{-4} n} \leq 0.05 \quad \Leftrightarrow \quad \underline{n \geq 50,000}$$