Anna Janicka

Probability Calculus 2019/2020
Lecture 12

1. Strong Laws of Large Numbers

During the previous lecture, we have seen how the Chebyshev inequality allowed to formulate and prove the Weak Laws of Large Numbers, dealing with the conditions for the convergence in probability of the sequences of means of sequences of random variables. Here we will formulate two versions of the Strong Law of Large Numbers (SLLN), i.e. the counterparts which deal with convergence almost surely.

The first theorem describes the case of the Bernoulli Scheme (**Strong Law of Large Numbers for the Bernoulli Scheme**):

Theorem 1. *Let $X_1, X_2, \ldots$ be a sequence of independent random variables, such that*

$$\mathbb{P}(X_n = 1) = p = 1 - \mathbb{P}(X_n = 0), \quad n = 1, 2, \ldots.$$

Then, the sequence (S_n/n) converges almost surely to p; in other words, there exists an event Ω' of measure 1 such that for any $\omega \in \Omega'$, we have

$$\lim_{n\to\infty} \frac{S_n(\omega)}{n} = p.$$

A very important implication of the above theorem is that the intuitive definition of probability as a limit of empirical frequencies does indeed lead to the correct understanding of probability.

The second theorem is more general, and deals with independent random variables of identical distributions (**Kolmogorov's Strong Law of Large Numbers**):

Theorem 2. *Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed integrable random variables. Then,*

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}X_1.$$

This version of the theorem underlines the fact that empirical averages are a good approximation of the true mean of a distribution. We do not know, however, how good this approximation is for a given value of n – from the theorem itself we do not know anything about the rate of convergence of the sequences.

2. Applications of Strong Laws of Large Numbers in Statistics

In most real-life applications, the researcher does not know the exact distribution of a random variable; rather, his aim is precisely to find the basic characteristics of a variable based on observations only. We have already hinted above that the SLLN is a tool which allows to assess the validity of considering empirical sample means when aiming at a description of an unknown distribution: if $X_1, X_2, \ldots$ is a sequence of independent integrable random variables of identical distributions, we have that

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}X_1.$$

This observation may be extended further; if $X_1, X_2, \ldots$ is a sequence of independent random variables of identical distributions, whose squares are integrable, we have that (also on the basis of the SLLN, applied to the sequence of squares):

$$S^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2 \xrightarrow[n\to\infty]{a.s.} \mathrm{Var}X_1.$$

In other words, the sample variance (defined as above) is a good approximation of the true distribution variance.

The SLLN allow to say even more. Assume that the sequence $X_1, X_2, \ldots, X_n$ of independent identically distributed random variables represents a sample from a distribution (perhaps unknown) of size n. We may define an empirical distribution for this sample:

$$\mu_n(A) = \frac{1_A(X_1) + 1_A(X_2) + \ldots + 1_A(X_n)}{n}.$$

From the SLLN, we have that for any event $A \subseteq \Omega$:

$$\mu_n(A) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}1_A(X_1) = \mathbb{P}(X_1 \in A),$$

which means that the true distribution of the variables X_n is a limit of the empirical distributions. In many cases, however, it is not convenient to speak in terms of distributions (which are formulated in terms of probabilities of different events); it is more convenient to talk about cumulative distribution functions (which also identify a distribution unequivocally). A cumulative distribution function for the empirical distribution associated with a sample of size n (which is also called the empirical CDF of the sample) may be defined as

$$F_n(t) = \frac{1_{\{X_1 \leqslant t\}} + 1_{\{X_2 \leqslant t\}} + \ldots + 1_{\{X_n \leqslant t\}}}{n}.$$

From the SLLN, we have that for any $t \in \mathbb{R}$

$$F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t).$$

However, an even stronger result may be proven: uniform convergence. This result is referred to as the **Glivenko–Cantelli Theorem**, which is of primary importance in statistics:

Theorem 3. *Let $X_1, X_2, \ldots$ be independent random variables from a distribution with a CDF F. Then,*

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0.$$

3. Limit Theorems

We have stated above that the SLLN do not say much about the rate of convergence of sequences of averages. An extremely important answer to this question is the **Central Limit Theorem** (CLT). The classical version of the CLT describes the size and the distributional form of the fluctuations around the theoretical mean during this convergence:

Theorem 4. *Let $X_1, X_2, \ldots$ be identically distributed independent random variables, such that $\mathbb{E}X_1^2 < \infty$. If by $m = \mathbb{E}X_1$ we denote the mean, and by $\sigma^2 = \mathrm{Var}X_1$ the variance of this distribution, then for any $t \in \mathbb{R}$, we have that*

$$\mathbb{P}\left(\frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}} \leqslant t\right) \xrightarrow{n\to\infty} \Phi(t),$$

where

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx$$

is the CDF of the standard normal distribution.

The theorem may easily be extended to versions with lower limits for the standardized sums: for any $s, t \in \mathbb{R}$ such that $s < t$ we have

$$\mathbb{P}\left(s \leqslant \frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}}\right) \xrightarrow{n\to\infty} 1 - \Phi(s),$$

and

$$\mathbb{P}\left(s \leqslant \frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}} \leqslant t\right) \xrightarrow{n\to\infty} \Phi(t) - \Phi(s).$$

Note that any of the inequalities above may be changed to strict without any change in the limits on the right hand side. What should also be noted is that although we have formulated

the CLT for identically distributed random variables, this is not a necessary condition; the CLT also holds for sequences of non-identical random variables, provided that they comply with certain conditions (for example the Lyapunov condition).

Note that the CLT provides an answer to the question of the prevalence of the normal probability distribution in the real-world (for example the appearance of the "Bell Curve" in density estimates): many quantities and characteristics may be thought of as a (balanced) sum of a large number of random factors.

A specific case of the CLT is the **De Moivre - Laplace Theorem**, which deals with the case of a Bernoulli Scheme:

Theorem 5. *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables, such that*

$$\mathbb{P}(X_n = 1) = p = 1 - \mathbb{P}(X_n = 0).$$

Then, we have that for any $s < t$,

$$\mathbb{P}\left(s \leqslant \frac{X_1 + X_2 + \ldots + X_n - np}{\sqrt{np(1-p)}} \leqslant t\right) \xrightarrow{n\to\infty} \Phi(t) - \Phi(s).$$

As before, any of the inequalities above may be changed to strict without consequences for the formula on the right-hand side.

We will now formulate some examples to show the usefulness of the CLT.

(1) In many problems, we assume that the probability that a newborn /an individual will be male or female is equal to $\frac{1}{2}$. Under such assumptions, if we were to answer the question of what is the probability that out of 10000 newborns, the number of girls will exceed the number of boys – the answer would be $\frac{1}{2}$ (approximately). However, say that in reality the probability that a newborn will be a boy is equal to 0.517. What is the answer to the question now? Let $X_i = \mathbf{1}_{i\text{-th newborn is a boy}}$. We have that $\mathbb{E}X_i = 0.517$ and $\mathrm{Var}X_i = 0.517 \cdot 0.483$; therefore,

$$\mathbb{P}(X_1 + X_2 + \ldots + X_{10000} < 5000) = \mathbb{P}(X_1 + X_2 + \ldots + X_{10000} - 10000 \cdot 0.517 < 5000 - 5170)$$

$$= \mathbb{P}\left(\frac{X_1 + X_2 + \ldots + X_{10000} - 5170}{\sqrt{10000 \cdot 0.517 \cdot 0.483}} < \frac{-170}{\sqrt{10000 \cdot 0.517 \cdot 0.483}}\right) \approx \Phi\left(\frac{-170}{\sqrt{10000 \cdot 0.517 \cdot 0.483}}\right).$$

Due to the fact that the standard normal distribution is symmetric around 0, we can transform the above using the property $\Phi(t) + \Phi(-t) = 1$ to

$$= 1 - \Phi\left(\frac{170}{\sqrt{10000 \cdot 0.517 \cdot 0.483}}\right) \approx 1 - \Phi(3.40) \approx 0.0004.$$

This means that for large n, contrary to the small sample situation, using an approximation of $p = \frac{1}{2}$ instead of $p = 0.517$ may lead to major errors.

(2) Previous experience suggests that approximately 70% of students who pass matriculation finally enroll at a given faculty. A faculty has the right to determine the exam threshold. How many students should be initially accepted, if the faculty wants to approximate that with probability of at least 0.9, the number who eventually enroll does not exceed 200?

Assume that initially N individuals pass matriculation. Let $X_i = 1_{i\text{-th student will enroll}}$, for $i = 1, 2, \ldots, N$. Let us assume that X_i are independent. Their distribution is given by:

$$\mathbb{P}(X_i = 1) = 0.7 = 1 - \mathbb{P}(X_i = 0).$$

Thus, we have that $m = \mathbb{E}X_1 = 0.7$, $\sigma = \sqrt{\mathrm{Var}X_1} = \sqrt{0.7 \cdot 0.3} \approx 0.46$. We are interested in the event

$$\{X_1 + X_2 + \ldots + X_N \leqslant 200\},$$

which may be transformed to

$$\left\{\frac{X_1+X_2+\ldots+X_N-0.7N}{\sigma\sqrt{N}} \leqslant \frac{200-0.7N}{0.46\sqrt{N}}\right\}.$$

Using the de Moivre-Laplace theorem, we approximate the probability of the above event by

$$\Phi\left(\frac{200-0.7N}{0.46\sqrt{N}}\right).$$

For which N will the above probability be equal to at least 0.9? We may search in the standard normal cumulative distribution tables to find that $\Phi(1.29) \approx 0.90147$, therefore it will suffice to take N such that $\frac{200-0.7N}{0.46\sqrt{N}}$ is as close as possible to 1.29 (or smaller). The solution is $N \leqslant 271.74$, so we should have $N \leqslant 271$. A similar reasoning will allow us to find the minimum number of students who must pass matriculation in order for the number of enrolled not to fall under a given threshold (with a given probability).

(3) Let us assume that we take a sum of 400 numbers, each of them rounded up to 10^{-2}. Assume that the rounding errors are independent random variables with uniform distribution over $[-10^{-2}, 10^{-2}]$. What is the probability that the total error exceeds 0.1?

Let X_i be the error of rounding the i-th number. We have $m = \mathbb{E}X_1 = 0$, $\sigma = \sqrt{\frac{4\cdot 10^{-4}}{12}} \approx 0.006$, so

$$\mathbb{P}(X_1+X_2+\ldots+X_{400} > 0.1) = \mathbb{P}\left(\frac{X_1+X_2+\ldots+X_{400}-400\cdot 0}{0.006\sqrt{400}} > \frac{0.1}{0.12}\right)$$

$$\approx 1-\Phi(\frac{0.1}{0.12}) \approx 0.202,$$

based on the CLT.

(4) **Confidence Intervals.** Another important example of the application of the CLT is the construction of confidence intervals. Let us assume that $X_1, X_2, \ldots, X_n$ is a sample from a known class of distributions, but with an unknown parameter θ – for example, we toss a coin multiple times, but we do not know if the coin is unbiased or not. We know that the average number of heads obtained approximates the true probability of obtaining a head. But this average, for finite samples, is almost surely not the precise result (and would change if we added another trial). Therefore, we should not pay too much attention to the exact result. It would be better to describe the true probability by means of an interval, rather than a point approximation. We will say that the interval (θ_1, θ_2) is a confidence interval at a confidence level $1-\alpha$ for the parameter θ, if

$$\mathbb{P}(\theta \in (\theta_1, \theta_2)) \geqslant 1-\alpha.$$

θ_1 and θ_2 are random variables (functions of $X_1, X_2, \ldots, X_n$). Obviously, our aim is to assure that this interval is the narrowest possible.

Let us now return to the tossing coin experiment. Let $X_1, X_2, \ldots, X_n$ be a random sample from a two-point distribution, such that

$$\mathbb{P}(X_i = 1) = p = 1-\mathbb{P}(X_i = 0)$$

(p is unknown). Based on this sample, we wish to find the confidence interval for p at a confidence level 0.9, i.e. an interval (p_1, p_2) such that

$$\mathbb{P}(p_1 < p < p_2) \geqslant 0.9.$$

We already know that a good candidate for the approximate of the distribution mean (in our case – the value of p) is the sample average $\overline{X}$. If we know that a standardized average will behave similarly to the standard normal distribution, which is symmetric

around the mean and whose density has one maximum at the mean, we may infer that the narrowest possible interval will be obtained by taking

$$p_1 = \overline{X} - \varepsilon \quad \text{and} \quad p_2 = \overline{X} + \varepsilon,$$

for a value $\varepsilon > 0$ which we should determine. In other words, we are searching for ε such that

$$\mathbb{P}(-\varepsilon < \overline{X} - p < \varepsilon) \geqslant 0.9.$$

Transforming the formula to obtain the form from the CLT, we multiply by $\sqrt{n}$ and divide by $\sqrt{p(1-p)}$ to obtain

$$\mathbb{P}\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} < \frac{X_1 + X_2 + \ldots + X_n - np}{\sqrt{np(1-p)}} < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \geqslant 0.9.$$

From the CLT, we have that the above is approximately equal to

$$\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) = 2\Phi(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}) - 1.$$

We have that $\Phi(1.64) \approx 0.95$ (so that $2\Phi(1.64) - 1 \approx 0.9$); therefore, we will need $\varepsilon = \frac{1.64p(1-p)}{\sqrt{n}}$ (or larger, if we want the probability to exceed 0.9). Since we do not know anything about the true value of p, we must assume the least favorable case; this is $p(1-p) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Thus, we should take $\varepsilon = \frac{1.64}{4\sqrt{n}}$ – this value will provide the narrowest possible confidence interval for a confidence level of 0.9. For example, for a sample of size 900, we would obtain the following 90% confidence interval for p:

$$(\overline{X} - 0.014, \overline{X} + 0.014).$$