Anna Janicka

Probability Calculus 2021/2022
Lecture 11

1. Conditional Expectation as a Predictor

We have already signalled the importance of being able to predict one random variable with the use of another random variable in practical applications. Here we will explore the topic a little bit further, dropping the assumption of linear approximations we made until now.

Let us assume that a phenomenon may be described with the use of a two-dimensional random vector, (X, Y), but that one of the variables – Y – is hard to observe, or may be observed, but later on (in the future). Let us now assume that we wish to assess the value of Y, based on the observed values of X. As in the case of linear regression, we will be looking for the best possible approximation in terms of the mean square error; this time, however, we will not constrain the possible formulae to linear, but allow any Borel transformation of variable X.

Formally, we wish to find a Borel function $\varphi : \mathbb{R} \to \mathbb{R}$, such that φ minimizes

$$\mathbb{E}(Y - \varphi(X))^2.$$

It may be shown that in this case, the best possible approximation is the conditional expectation: $\varphi^*(x) = \mathbb{E}(Y|X = x)$; formally, we have:

Theorem 1. *Let $X, Y : \Omega \to \mathbb{R}$ be random variables such that $\mathbb{E}Y^2 < \infty$. Then, the function $\varphi^* : \mathbb{R} \to \mathbb{R}$, such that $\varphi^*(x) = \mathbb{E}(Y|X = x)$, satisfies:*

$$\mathbb{E}(Y - \varphi^*(X))^2 = \min\{\mathbb{E}(Y - \varphi(X))^2 : \varphi \textit{ is a Borel function} : \mathbb{R} \to \mathbb{R}\}.$$

2. Chebyshev Inequalities

In this section, we will explore a simple and easy to prove inequality, which has very sound theoretical implications. This inequality (and its derivatives) is extremely useful in that it allows to assess the probability of events of given types without having to refer to any knowledge about a given distribution, apart from basic information about the mean (variance, or other moments, depending on the version). The questions that may be answered with the use of this inequality revolve around obtaining an upper bound for the probability that a random variable exceeds a given value, or that the discrepancy between the random variable and its mean exceeds a given value. Such questions may easily arise in practical situations; for example, a gambler may be interested in a rule of thumb to determine if participating in a game is worthwhile or not (is the probability of loosing at least a given amount not too big?); a researcher may be interested in the probability that the error of measurements exceeds a given threshold, etc.

The basic version of the above-mentioned inequality, known as the **Chebyshev inequality** (sometimes referred to as the Markov inequality), may be formulated in the following way:

Theorem 2. *Let X be a nonnegative integrable random variable, and let $\varepsilon > 0$. We have:*

$$\mathbb{P}(X \geqslant \varepsilon) \leqslant \frac{\mathbb{E}X}{\varepsilon}.$$

The proof of this theorem is simple. We have

$$X \geqslant X\mathbf{1}_{\{X \geqslant \varepsilon\}} \geqslant \varepsilon\mathbf{1}_{\{X \geqslant \varepsilon\}},$$

so that

$$\mathbb{E}X \geqslant \mathbb{E}(\varepsilon\mathbf{1}_{\{X \geqslant \varepsilon\}}) = \varepsilon\mathbb{P}(X \geqslant \varepsilon),$$

which upon transformation gives the requested property.

Note that not in all cases the inequality gives meaningful assessments – the upper bound may be greater than 1 (or close to 1). However, as we have stated above, this simple inequality has very many important implications. A lot of them may be easily justified with one of the numerous transformations of the basic Chebyshev inequality. Note that since the inequality holds for any nonnegative random variable, we may substitute in place of X specific functions or transformations of a variable X (not necessarily nonnegative, given that the transformation is nonnegative); we may then obtain, for example, the following versions of the Chebyshev inequality:

Theorem 3. *Let X be a random variable.*

- **Markov Inequality:** *For any $p > 0$ such that $\mathbb{E}|X|^p$ exists, and any $\varepsilon > 0$,*

$$\mathbb{P}(|X| \geqslant \varepsilon) \leqslant \frac{\mathbb{E}|X|^p}{\varepsilon^p}.$$

- **Chebyshev-Bienaymé Inequality:** *For any $\varepsilon > 0$, if the random variable X^2 is integrable,*

$$\mathbb{P}(|X - \mathbb{E}X| \geqslant \varepsilon) \leqslant \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

- **Exponential Chebyshev Inequality:** *Let us assume that $\mathbb{E}e^{pX} < \infty$ for a given value $p > 0$. Then, for any $\lambda \in [0, p]$ and for any $\varepsilon > 0$,*

$$\mathbb{P}(X \geqslant \varepsilon) \leqslant \frac{\mathbb{E}e^{\lambda X}}{e^{\lambda\varepsilon}}.$$

These three inequalities may immediately be obtained from the basic Chebyshev inequality upon applying it to $|X|^p$, $(X - \mathbb{E}X)^2$ and $e^{\lambda X}$ in place of X, and ε^p, ε^2 and $e^{\lambda\varepsilon}$ in place of ε, respectively.

Examples:

(1) Let us assume we wish to measure an unknown value μ (a physical value, for example), but that each measurement is laden with a random error. The natural model for this experiment is assuming that the subsequent measurements are independent random variables with mean μ, and variance not exceeding a constant c. From the Chebyshev-Bienaymé Inequality, we have that

$$\mathbb{P}(|X_i - \mu| \geqslant \varepsilon) \leqslant \frac{c}{\varepsilon^2}.$$

Note that if c is large (or ε relatively small), the information conveyed by the inequality may be useless. On the other hand, if we wish to approximate the unknown parameter μ with the mean of the measurements, the inequality proves very useful:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\varepsilon^2} = \frac{\sum_{i=1}^{n}\mathrm{Var}(X_i)}{n^2\varepsilon^2} \leqslant \frac{nc}{n^2\varepsilon^2} = \frac{c}{n\varepsilon^2}.$$

In this case, the limit of the upper bound as n increases to infinity is zero, which means that the approximation of μ, for a large number of repetitions of an experiment, becomes very good; moreover, the inequality allows to determine the number of repetitions necessary to obtain a required precision level with the required probability.

(2) Assume now that we wish to determine the unknown probability p of a single event (for example, the probability of success in a Bernoulli trial). Similarly to the example above, to determine this probability it will suffice to repeat (independently) a series of trials and calculate the empirical frequency. Formally, let X_i be a random variable equal to 1 if the required event occurred in the i-th trial, and 0 otherwise. We have that $\mathbb{E}X_i = p$, and $\mathrm{Var}X_i = p(1-p)$. Let us now denote the sum $X_1 + X_2 + \ldots + X_n$ by S_n; we have $\mathbb{E}S_n = np$ and $\mathrm{Var}S_n = np(1-p)$. From the Chebyshev-Bienaymé Inequality, we have that

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right) \leqslant \frac{p(1-p)}{n\varepsilon^2}.$$

This form of the upper bound is not too good in that it uses the unknown value of p. However, knowing that $p(1-p) \leqslant \frac{1}{4}$, we obtain the following assessment:

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right) \leqslant \frac{1}{4n\varepsilon^2}.$$

Therefore, if we wish to obtain an approximation of p on the basis of, say, the empirical frequency in 10000 repetitions of the experiment, the probability that the error we will make will exceed $\frac{1}{10}$ is not greater than $\frac{1}{400}$. In fact, it is much lower, which may be demonstrated with other tools, such as the Exponential Chebyshev Inequality, and its consequence: the Bernstein Inequality, which we will formulate later on.

(3) Another class of applications of the Chebyshev inequality comprise of situations where the parameters of distributions of random variables are known – but the probability of a given event of interest is, for one reason or other, difficult to calculate, and for our needs it suffices to find an assessment of this probability instead of a precise value. For example, let us assume that we toss a symmetric coin 20000 times, and we are interested in the probability that the number of heads obtained will deviate from the expected value of 10000 by more than 200, i.e. in $\mathbb{P}\left(|S_{20000} - 10000| \geqslant 200\right)$. We have:

$$\mathbb{P}\left(|S_{20000} - 10000| \geqslant 200\right) = \mathbb{P}\left(\left|\frac{S_{20000}}{10000} - \frac{1}{2}\right| \geqslant \frac{1}{100}\right) \leqslant \frac{1}{4 \cdot 20000 \cdot (0.01)^2} = \frac{1}{8}.$$

Again, this assessment may be improved considerably with other tools (other versions of the Chebyshev Inequality among them).

We will now formulate a more powerful inequality, which also may be derived (albeit in a slightly more complicated way) from the Chebyshev inequalities in the case of a Bernoulli scheme: the **Bernstein Inequality**.

Theorem 4. *Let S_n be a random variable from a binomial distribution with parameters n and p. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right) \leqslant 2e^{-2\varepsilon^2 n}.$$

If we wish to look at one-sided errors only, we have that

$$\mathbb{P}\left(\frac{S_n}{n} \geqslant p + \varepsilon\right) \leqslant e^{-2\varepsilon^2 n}$$

and

$$\mathbb{P}\left(\frac{S_n}{n} \leqslant p - \varepsilon\right) \leqslant e^{-2\varepsilon^2 n}.$$

We can now compare the assessments formulated on the base of the Chebyshev inequalities (described in the examples above) with those obtained with the Bernstein inequality:

(2) For large n, the upper bound of $\frac{1}{4n\varepsilon^2}$ from the Chebyshev Inequality is much larger than $2e^{-2\varepsilon^2 n}$.

(3) In the case of repetitive coin tossing, we have that

$$\mathbb{P}\left(|S_{20000} - 10000| \geqslant 200\right) = \mathbb{P}\left(\left|\frac{S_{20000}}{10000} - \frac{1}{2}\right| \geqslant \frac{1}{100}\right) \leqslant 2e^{-2\cdot(0.01)^2\cdot 20000} \approx 0.037,$$

which is significantly lower than the 0.125 obtained above.

3. Convergence of Sequences of Random Variables

Due to the fact that random variables are functions rather than points, and that when dealing with random variables we always face the problem that instead of making a definite statement we can only say that something happens with some probability, the question of what happens with random variables when we look at infinite sequences and their limits is a complicated one. Different types of convergence of sequences of random variables may be defined. During this course, we will define only two of them: almost sure convergence and convergence in probability.

Definition 1. *A sequence $(X_n)_{n\geqslant 1}$ of random variables over Ω converges* **almost surely** *to X, if*

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

Equivalently, we may say that there exists a subset $\Omega' \subset \Omega$ such that $\mathbb{P}(\Omega') = 1$, such that for any $\omega \in \Omega'$, we have

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

Almost sure convergence is usually denoted by $X_n \xrightarrow{a.s.} X$.

An alternative formulation of the condition of almost sure convergence is the following:

$$\lim_{n\to\infty} \mathbb{P}(\sup_{k\geqslant n} |X_k - X| > \varepsilon) = 0.$$

Definition 2. *A sequence $(X_n)_{n\geqslant 1}$ of random variables over Ω converges* **in probability** *to X, if for any $\varepsilon > 0$, we have that*

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

Equivalently, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \leqslant \varepsilon) = 1.$$

Convergence in probability is usually denoted by $X_n \xrightarrow{\mathbb{P}} X$ or $\mathrm{plim}_{n\to\infty} X_n = X$.

Note that when the condition of almost sure convergence is defined in the alternative formulation, it becomes obvious that almost sure convergence of a sequence implies convergence in probability. The reverse does not hold, i.e. there exist sequences fulfilling the condition of convergence in probability, but such that the limit may be approached with "peaks" of discrepancies from the limit every now and then, which excludes almost sure convergence.

Note also that the limit of a sequence of random variables is a random variable; however, in many applications (for example, in the case of the sequences satisfying the assumptions of the Laws of Large Numbers, which we will discuss next), the limit random variable may be "degenerated" into a single point, i.e. a random variable which takes on a single value with probability 1.

The two types of convergence defined above have many of the properties of pointwise convergence, for example:

Theorem 5. *Let $(X_n)_{n\geqslant 1}$ and $(Y_n)_{n\geqslant 1}$ be sequences of random variables. If $(X_n)_{n\geqslant 1}$ converges to X and $(Y_n)_{n\geqslant 1}$ converges to Y almost surely (/in probability), then $X_n \pm Y_n \to X \pm Y$ and $X_n \cdot Y_n \to XY$ almost surely (/in probability).*

In the "standard" cases (when Y_n does not converge to a variable which takes on the value of 0 with non-negative probability), the limit of the ratio X_n/Y_n also converges to X/Y.

4. Weak Laws of Large Numbers

One of the more important applications of the Chebyshev inequalities are the Laws of Large Numbers. Under this term, we have several theorems describing the behavior of the series of sums of random variables, i.e. of the sequences

$$S_n = X_1 + X_2 + \ldots + X_n,$$

or rather the sequences of means:

$$\frac{S_n}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n},$$

for different types of sequences $(X_n)_{n\geqslant 1}$. Depending on whether the thesis of the theorem pertains to convergence in probability or almost surely, the laws are denoted either as Weak or Strong, respectively.

In one of the examples in the section above, when applying the Chebyshev Inequality to a Bernoulli scheme, we have already proven what may be denoted as the **Weak Law of Large Numbers for the Bernoulli Scheme**:

Theorem 6. *Let $X_1, X_2, \ldots$ be independent with distributions*

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

We then have that (S_n/n) converges in probability to p (i.e. a constant random variable equal to p); in other words, for any $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| > \varepsilon\right) = 0.$$

In fact, the reasoning may easily be extended to weaken the assumptions of the theorem, to obtain the **Weak Law of Large Numbers** (WLLN) for uncorrelated random variables (not necessarily with common distributions!):

Theorem 7. *Let $X_1, X_2, \ldots$ be uncorrelated random variables with a common upper bound to their variances. Then, the sequence $(X_n)_{n\geqslant 1}$ satisfies the weak law of large numbers:*

$$\frac{S_n - \mathbb{E}S_n}{n} \xrightarrow{\mathbb{P}} 0,$$

i.e. for any $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n - \mathbb{E}S_n}{n}\right| > \varepsilon\right) = 0.$$

Examples:

(1) We repeat tossing a symmetric coin. Let X_n be random variable equal to 1 if the result of the n-th toss is heads and 0 otherwise, for $n = 1, 2, \ldots$. Then, the sequence $\frac{X_1+\ldots+X_n}{n}$ converges in probability to $\frac{1}{2}$. This means that in an infinite sequence of coin tosses, we expect to see heads in half of the cases (and the chance that in the limit the proportion of heads will differ from $\frac{1}{2}$ by more than ϵ, for any $\epsilon > 0$, is equal to zero).

5. Strong Laws of Large Numbers

Now we will formulate two versions of the Strong Law of Large Numbers (SLLN), i.e. the counterparts to the WLLN 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 8. *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 9. *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.