CIS 2033 Lectures 20 and 21, Spring 2017[1]

Instructor: David Dobor

April 13, November 13, 2016

[1] Supplemental reading: Bertsekas Textbook Chapter 7.

In this lecture, we develop the weak law of large numbers. Loosely speaking, the weak law of large numbers says that if we have a sequence of independent random variables with the same distribution, then their average, which is called the sample mean, approaches the expected value of the distribution. In this sense, it reinforces our interpretation of the expected value as some kind of overall average.

Introduction

The weak law of large numbers is the reason why polling works. By asking many people about the value of some attribute, and by taking the average of the responses, we can get a good estimate of the average over the entire population. On the mathematical side, in order to derive the weak law of large numbers, we will first need to develop some inequalities, namely the Markov and Chebyshev inequalities.

We will see that when discussing these inequalities, we will have to deal with a technical issue. The weak law of large numbers talks about the convergence of a random variable to a number. For this to make sense, we need to define an appropriate notion of convergence. We will introduce one such notion that goes under the name of *convergence in probability*. And we will see that in many respects, it is similar to the common notion of convergence of numbers.

Markov Inequality

The Markov inequality is a rather simple but quite useful and powerful fact about probability distributions. The basic idea behind the Markov inequality is the following. We may be interested in saying something about the probability of an extreme event. By extreme event, we mean that some random variable takes a very large value.

If we can calculate that probability exactly, then, of course, everything is fine. But suppose that we only have a little bit of information about the probability distribution at hand. For example, suppose that we only know the expected value associated with that distribution. Can we say something? Well, here's a statement, which is quite intuitive

If you have a non-negative random variable, and I tell you that the average or the expected value is rather small, then there should be only a very small probability that the random variable takes a

very large value. This is an intuitively plausible statement, and the Markov inequality makes that statement precise. Here is what it says.

If we have a random variable that's non-negative and you take any positive number a, the probability that the random variable exceeds that particular number is bounded by the ratio $E[X]/a$:

Markov inequality: If $X \geq 0$ and $a > 0$, then $\mathbf{P}(X \geq a) \leq \dfrac{\mathbf{E}[X]}{a}$

If the expected value of X is very small, then the probability of exceeding that value of a will also be small. Furthermore, if a is very large, the probability of exceeding that very large value drops down because this ratio becomes smaller. So that's what the Markov inequality says. Let us now proceed with a derivation.

Let's start with the formula for the expected value of X, and just to keep the argument concrete, let us assume that the random variable is continuous so that the expected value is given by an integral.

$$E[X] = \int_0^\infty x f(x) dx \tag{1}$$

The argument would be exactly the same as in the discrete case, but in the discrete case, we would be using a sum. Also, since the random variable is non-negative, this integral only ranges from 0 to infinity.

Now, we're interested, however, in values of X larger than or equal to a, and that tempts us to consider just the integral from a to infinity of the same quantity.

$$\int_a^\infty x f(x) dx \tag{2}$$

How do these two quantities compare to each other? Since we're integrating a non-negative quantity, if we're integrating over a smaller range, the resulting integral will be less than or equal to the expected value:

$$\int_a^\infty x f(x) dx \leq \int_0^\infty x f(x) dx = E[X] \tag{3}$$

Now, in equation (2), over the range of integration that we're considering x is at least as large as a. Therefore, the quantity that we're integrating from a to infinity is at least as large as a times the density of X. Furthermore, now we can take this a, which is a constant, pull it outside the integral.

$$\int_a^\infty x f(x) dx \geq a \int_a^\infty f(x) dx \tag{4}$$

The right hand side in (4) is the integral of the density from a to ∞, which is the probability that the random variable takes X a value larger than or equal to a, $P(X \geq a)$.

$$E[X] \geq \int_a^\infty x f(x) dx \geq a \int_a^\infty f(x) dx = aP(X \geq a)$$

So, sending a to left side, we get exactly the Markov inequality:

$$\boxed{\frac{E[X]}{a} \geq P(X \geq a)} \tag{5}$$

Now it is instructive to go through a *second derivation of the Markov inequality*. This derivation is essentially the same conceptually as the one that we just went through except that it is more abstract and does not require us to write down any explicit sums or integrals.

Let us define a new random variable Y as follows

$$Y = \begin{cases} 0 & \text{whenever } X < a, \\ a & \text{whenever } X \geq a \end{cases}$$

How is Y related to X? If X takes a value less than a, and since X is also nonnegative, then X is certianly going to be larger than Y, which is zero in this case. If $X \geq a$, Y will be a, so X will again be at least as large. So no matter what,

$$Y \leq X,$$

which also means that

$$E[Y] \leq E[X] \tag{6}$$

But now what is the expected value of Y? Since Y is either 0 or a, the expected value is

$$E[Y] = aP(X \geq a) \tag{7}$$

And by comparing the two sides of this inequality, what we have is exactly the Markov inequality.

$$\boxed{\frac{E[X]}{a} \geq P(X \geq a)} \tag{8}$$

So we just gave two proofs of Markov's inequality but the proofs may not necessarily make the inequality intuitive. Let us now go through some simple examples.

Example 1. Suppose we have a 100 people attending an event. Suppose we ask:

- Is it possible that at least 95% of the people are younger than the average person in the group?

Here by "average" we mean "of mean age" (if you consider their median age, the following conclusions would not hold).

The answer is yes: one older person can pull up the average a lot (just imagine 99 one year olds being babysat by a 100 year old person).

Now consider the question:

- Is it possible that at least 50% of the people are older than twice the average age in the group?

Well, no, because if you take just those 50% whose age is more than double the average age and compute their average, you've already exceeded the average. Or think of it in terms of totals. If the average is μ, the the total age is 100 μ. Now if you have 50 people whose age is more than 2 μ, sum of their ages is already greater than $50 \times 2\ \mu = 100\ \mu$, which is impossible.

Similarly, you can't have a third of the people in the group who are older than 3 times the average age in the group. And that's exactly what Markov's inequality says.

EXAMPLE 2. Suppose that X is exponentially distributed with parameter or equal to 1, i.e. $X \sim Exp(1)$ so that $E[X]$ is also going to be equal to 1. So this is then what Markov inequality says about $P(X \geq a)$:

$$P(X \geq a) \leq \frac{1}{a}$$

To put this result in perspective, note that we're trying to bound a probability. We know that the probability lies between 0 and 1. There's a true value for this probability which, in this particular example, because we have an exponential distribution, is equal to e^{-a}.

$$P(X \geq a) = e^{-a}$$

The Markov inequality gives us a bound. A bound will be considered good or strong or useful if that bound turns out to be quite close to the correct value of e^{-a}.

Unfortunately, in this example, this is not the case because the true value falls off exponentially with a, whereas the bound that we obtained falls off at a much slower rate of $1/a$.

This is ofter the case and for this reason one would like to have even better bounds than the Markov inequality, and this is one motivation for the Chebyshev inequality that we will be considering next.

Chebyshev's Inequality

Mathematically speaking, the Chebyshev inequality is just a simple application of the Markov inequality. However, it contains a somewhat different message.

Consider a random variable that has a certain mean and variance. What the Chebyshev inequality says is that if the variance is small, then the random variable is unlikely to fall too far off from the mean – if the variance is small, we have little randomness, and so X cannot be too far from the mean.

In more precise terms, we have the following inequality: the probability that the distance from the mean is larger than or equal to a certain number c is, at most, the variance divided by the square of c.

Chebyshev inequality: $\mathbf{P}\big(|X-\mu| \geq c\big) \leq \dfrac{\sigma^2}{c^2}$

So if the variance is small, the probability of falling far from the mean is also going to be small. And if the number c is large, so that we're talking about a large distance from the mean, then the probability of this event happening falls off at a rate at least $1/c^2$. By the way, I should add here that c is assumed to be a positive number: if c was negative, then the probability that we're looking at would be equal to 1 anyway, and there wouldn't be any point in trying obtain a bound for it.

Markov inequality: If $X \geq 0$ and $a > 0$, then $\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}$

Proof of Chebyshev is straightforward. We will simply apply the Markov equality to the nonnegative random variable $(X-\mu)^2$ as follows:

$$P(|\ X-\mu\ | \geq c) = P((X-\mu)^2 \geq c^2) \leq \frac{E[(X-\mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

As an application of the Chebyshev inequality, let us look at the probability of the event that the distance from the mean is at least k

standard deviations, where k is some positive number, i.e., let us look at the quantity $P(\mid X - \mu \mid \geq k\,\sigma^2)$. We have

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

So what this is saying is that if you take, for example, $k = 3$, the probability that you fall three standard deviations away from the mean or more is going to be less than or equal to $1/9$. And this is true no matter what kind of distribution you have.

Example 3. Let us now revisit Example 2, where X is an exponential random variable and we are interested in $P(X \geq a)$. The Markov inequality gave us $P(X \geq a) \leq \frac{1}{a}$. And as we recall, the exact answer to this probability was $P(X \geq a) = e^{-a}$.

Let us see what we can get using the Chebyshev inequality. Now, $E[X] = 1$. Let us assume that $a > 1$, so that we're considering an event that we fall far away from the mean by a distance of at least $a - 1$.

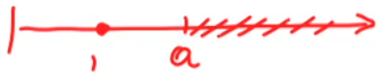


Figure 1: We are interested in bounding the probability that $X > a$, where $X \sim Exp(1)$ and $a > 1$.

First, we have that

$$P(X \geq a) = P(X - 1 \geq a - 1)$$

Now, it is also true that the event $X - 1 \geq a - 1$ is at most as big than the event $\mid X - 1 \mid \geq a - 1$. (This is because if $X - 1 \geq a - 1$ is true, then $\mid X - 1 \mid \geq a - 1$ will also be true.) So

$$P(X \geq a) = P(X - 1 \geq a - 1) \leq P(\mid X - 1 \mid \geq a - 1)$$

And now, we can apply the Chebyshev inequality. Taking into account that $Var(X) = 1$ we have

$$P(\mid X - 1 \mid \geq a - 1) \leq \frac{1}{(a-1)^2}$$

Notice that if a is a large number, the quantity on the right hand side behaves like $1/a^2$ which falls off much faster than $1/a$.

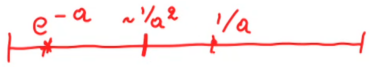


Figure 2: Chebyshev's inequality gives us a better bound than Markov's inequality for large a.

So at least for large a's, the Chebyshev bound is going to give us a smaller bound, which is more informative than what we obtained from the Markov inequality. In most cases, the Chebyshev inequality is, indeed, stronger and more informative than the Markov inequality. And one of the reasons is that it exploits more information about the distribution of the random variable X. That is it uses knowledge, not just about the mean of the random variable, but it also uses some information about the variance of the random variable.

The Week Law of Large Numbers

We now derive and discuss the weak law of large numbers. It is a rather simple result but plays a central role within probability theory. The setting is as follows. We start with a random variable X having a probability distribution that has a certain mean and variance, which we assume to be finite. We then draw independent random variables out of this distribution so that these X_i's are independent and identically distributed, i.i.d. for short.

- $X_1, X_2, \ldots$ i.i.d.; finite mean μ and variance σ^2

What's going on here is that we're carrying out a long experiment during which all of these random variables are drawn. Once we have drawn all of these random variables, we can calculate the average of the values that have been obtained, and this gives us the so-called sample mean.

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

Notice that

- the sample mean is a random variable because it is a function of random variables.
- the sample mean should be distinguished from the true mean, μ, which is the expected value of the X_i's. μ is a *number*, it is not random.

The sample mean is the simplest and most natural way for trying to estimate the true mean, and the weak law of large numbers will provide some support to this notion. Let us now look at the properties of the sample mean. First, let us calculate its expectation.

$$\boxed{\text{Calculate } E[M_n].}$$

By the way, $E[M_n]$ involves two different kinds of averaging. The sample mean averages over the values observed during one long experiment, whereas the expectation averages over all possible outcomes of this experiment. The expectation is some kind of theoretical average because we do not get to observe all the possible outcomes of this experiment, but the sample mean is something that we actually calculate on the basis of our observations.

In any case, by linearity,

$$E[M_n] = \frac{E[X_1 + X_2 + \ldots + X_n]}{n} = \frac{E[X_1] + E[X_2] + \ldots + E[X_n]}{n} = \frac{n\mu}{n}$$
$$= \mu.$$

So the theoretical average, the expected value of the sample mean, is equal to the true mean.

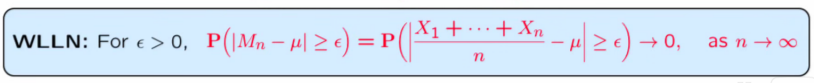
Let us now calculate the variance of the sample mean. The variance of a random variable divided by a number is the variance of that random variable divided by the square of that number. Now, since the X_i's are independent, the variance is the sum of the variances, and therefore,

$$Var(M_n) = Var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)$$
$$= \frac{Var(X_1 + X_2 + \ldots + X_n)}{n^2} = \frac{n\sigma^2}{n^2}$$
$$= \frac{\sigma^2}{n}.$$

We're now in a position to apply the Chebyshev inequality. The Chebyshev inequality tells us that the distance of a random variable from its mean, being larger than a certain number, has a probability that's bounded above as follows:

$$P(|\ M_n - \mu\ |) \geq \epsilon) \leq \frac{Var(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

And now, if we consider epsilon as a fixed number and let n go to infinity, then what we obtain is a limiting value of 0. So the probability of falling far from the mean diminishes to zero as we draw more and more samples. That's exactly what the weak law of large numbers tells us. If we fix any particular ϵ, which is a positive constant, the probability that the sample mean falls away from the true mean by more than ϵ becomes smaller and smaller and converges to 0 as n goes to infinity.



Let us now interpret the weak law of large numbers. As I already hinted, we have to think in terms of one long experiment, and during that experiment, we draw several independent random variables.

Each of these random variables is drawn from the same distribution. One way of thinking about those random variables is that each one of them is equal to the mean, the true mean, plus some measurement noise, which is a term that has zero expected value, with all of these noises are independent:

- One experiment
- Many measurements $X_i = \mu + W_i$.
- W_i: measurement noise; $E[W_i] = 0$; independent W_i.

So we have a collection of noisy measurements, and then we take those measurements and form the average of them. What the weak law of large numbers tells us is that the sample mean is unlikely to be far off from the true mean. And by far off, we mean at least ϵ distance away. So the sample mean is, in some ways, a good way of estimating the true mean. If n is large enough, then we have high confidence that the sample mean gives us a value that's close to the true mean.

As a special case let us consider a probabilistic model in which we repeat independently many times the same experiment. There's a certain event A associated with that experiment that has a certain probability, and each time that we carry out the experiment, we use an indicator variable to indicate whether the outcome was inside the event or outside the event. So X_i is 1 if A occurs, and it is 0 otherwise.

- Many independent repetitions of the same experiment
- Event A with $p = P(A)$.
- X_i: indicator of event A.

In this case, $E[X_i] = p$. In this particular example, the sample mean just counts how many times the event A occurred out of the n experiments that we carried out, so it's the frequency with which the event A has occurred. And we call it the *empirical frequency* of event A. What the weak law of large numbers tells us is that the empirical frequency will be close to the probability of that event. In this sense, it reinforces or justifies the interpretation of probabilities as frequencies.

Polling

//TODO