THEORY OF PROBABILITY

VLADIMIR KOBZAR

Lecture 20 - Conditional Expectation, Inequalities, Laws of Large Numbers, Central Limit Theorem

This lecture is based on the materials from the Courant Institute's Theory of Probability taught by Professor Austin in Spring 2016. All mistakes are mine.

Conditional Expectation (Ross, Secs 7.5 and 7.6. Suppose X and Y are discrete RVs and that y is a possible value of Y. Then for fixed y, the conditional PMF-values $p_{X|Y}(x|y) = P\{X = x|Y = y\}$ obey all the same rules as the unconditioned PMFs $p_X(x)$.

Definition 1. The conditional expectation of X given that $Y = y$:

(1) (discrete)

$$E[X|Y = y] = \sum_{x s.t. p_{X|Y}(x|y)>0} x \cdot p_{X|Y}(x|y)$$

(2) if X, Y ar jointly continuous with joint PDF f, then

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx$$

This plays the analogous role to $E[X]$ but in the situation where we have now learned that $Y = y$, so we condition on that information.

Example 1. (Ross, 7.5a) Suppose X and Y are two independent binom (n,p) RVs. Calculate $E[X|X + Y = m]$

Example 2. (Ross, 7.5b, parts are duplicative of Example 6.5b) Suppose that X, Y have joint PMF

$$f(x, y) = \begin{cases} \frac{e^{-x/y}e^{-y}}{y} & x, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find $E[X|Y = y]$.

Date: August 9, 2016.

Conditional expectation is extremely useful as a calculational tool.

The value $E[X|Y = y]$ depends on the value of y, say $g(y)$. So we get a new RV $g(Y)$: when Y takes the value y, this new RV takes the value $E[X|Y = y]$.

To denote this new RV, we simply write $E[X|Y]$.

Theorem 1. *(The Law of Total Expectation; Ross, Prop 7.5.1)*

$$E[E[X|Y]] = E[X]$$

You can apply this if X and Y are both discrete, jointly continuous, or in a mixed situation.

Example 3. (Ross 7.5l) Consider a random coin, whose bias U is a Unif$(0, 1)$ RV. Suppose we flip it n times. Let X be the number of heads obtained. Find the PMF of X.

IDEA: Apply the Law of Total Expectation to the indicator variables of the events $\{X = k\}$.

IN FACT, by using indicator variables, the Law of Total Expectation gives the general equation

$$P(E) = \int_{-\infty}^{\infty} P(E|Y = y) f_Y(y) dy$$

whenever Y is a continuous RV and E is an event. This is a 'continuous' version of the Law of Total Probability. See Ross Subsec 7.5.3.

Sometimes is is more informative to measure covariance relative to variance.

Definition 2. Let X and Y be positive, finite variances. Then their *correlation* is

$$corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

observe that $corr(aX, bY) = corr(X, Y)$ for any nonzero constants a, b. For example, if X and Y are two random distance measurements than their correlation doesn't depend on whether we use inches or centimeters.

Example 4. (Ross, 7.5f) The random vector (X, Y) is a bivariate standard normal with correlation $-1 < \rho < 1$ if it is jointly continuous with joint PDF

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2))$$

for $-\infty < x, y < \infty$. In Lecture 18, we showed that the marginal distributions of X and Y are standard normals, and in Lecture 19 we showed that the conditional distributions are $N(\rho y, 1-\rho^2)$.

By applying the law of total expectation to $E[XY]$ and the above-mentioned result from Lecture 19, we get

$$Cov(X,Y) = corr(X,Y) = \rho$$

Inequalities (Ross Secs 7.1, 7.2, 8.2). So far we have spent a lot of course learning how to compute exactly with random variables. (The Poisson approximation is perhaps an exception.)

But there are also reasons to study estimates and inequalities concerning probabilities and random variables.

- Sometimes we don't have enough information to compute a probability or expectation exactly, so we work out a range of possible values which are permitted given the information we do have.
- Certain basic inequalities are "responsible" for the Limit Theorems, which describe the asymptotic behaviour of large collections of RVs as the size of the collection tends to ∞.

The most basic inequality:

Proposition 2. *Let X be a RV such that $X \geqslant 0$: this means that the value taken by X is always non-negative, for every outcome of the experiment. Then*

$$E[X] \geqslant 0$$

REASON: $E[X]$s is a weighted average of the values taken by X.

Immediate consequences:

(1) If $a < b$ are reals such that $a \leqslant X \leqslant b$, then

$$a \leqslant E[X] \leqslant b$$

(2) (monotonicity of expectation) if X and Y are two RVs such that $X \geqslant Y$, then
$$E[X] \geqslant E[Y]$$

Example 5. (Boole's Inequality, Ross, 7.2d): If $P(A_1 \cup ... \cup A_n) \leqslant P(A_1) + ... + P(A_n)$.

Example 6. (Ross, 7.2r): A grove of 52 trees is arranged in a circle. If 15 chipmunks live in these trees, show that there is a group of 7 consecutive trees that together house at least 3 chipmunks.

Here is a slightly more subtle consequence of the monotonicity of expectation.

Proposition 3. *(Markov's inequality; Ross Prop 8.2.1) If X is a non-negative RV, then for any $a > 0$ we have*
$$P\{X \geqslant a\} \leqslant \frac{E[X]}{a}$$

IDEA: let I be the indicator variable of the event $\{X \geqslant a\}$ and notice that $X \geqslant a \cdot I$.

So Markov's inequality gives us an upper estimate on the probability that X takes a value above some threshold. But more often we want to estimate the probability that X takes a value far away from its expectation.

Proposition 4. *(Chebyshev's inequality; Ross Prop 8.2.2). If X is any random variable with $E[X] = \mu$ and $Var(X) = \sigma^2$, then for any $k > 0$, we have then for any $\kappa > 0$ we have*
$$P\{|X - \mu| \geqslant \kappa\} \leqslant \frac{\sigma^2}{k^2}$$

IDEA: Apply Markov to $|X - \mu|^2$.

Observe: Markov requires $X \geqslant 0$, but Chebyshev does not. If we let $\kappa = k\sigma$ for some positive integer k, then Chebyshev becomes
$$P\{|X - \mu| \geqslant k\sigma\} \leqslant \frac{1}{k^2}$$

SLOGAN: 'The probability that X takes a value at least k standard deviations ($= \sigma$) away from the mean ($= \mu$) is at most $\frac{1}{k^2}$.'

This finally gives a precise mathematical statement to justify the idea that "the variance/standard deviation indicates how spread out a RV is".

Example 7. (Ross 8.2a) Suppose that the number of items produced in a factory during a week is a RV X with mean 50.

(a) What can be said about the probability that this week's production will exceed 75?
(b) If $\text{Var}(X) = 25$, what can be said about the probability that this week's production will be between 40 and 60?

Example 8. If X is Unif$(0, 10)$, then

$$P\{X - 5\} > 4\} = 0.2$$

whereas Chebyshev gives

$$P\{X - 5\} > 4\} \leqslant \frac{25}{3(16)} \approx .52$$

$(Var(X) = 10^2/12 = 25/3)$

So Chebyshev gives us a guaranteed upper bound, but no-one is promising that it;s always a good estimate!

The Laws of Large Numbers (Ross Secs 8.2, 8.4). One of our basic intuitions about probability is this: If we perform an experiment independently many times, and E is an event that can happen for each performance of the experiment, then in the long-run average

$$\text{frequency of occurrence of } E \approx P(E).$$

For instance, if 37% (not a real statistic) of US citizens have visible dandruff, and we randomly select a few thousand citizens (a large number, but much less than US population), then we expect about 37% of those sampled to have visible dandruff.

This is a 'Law of Large Numbers'.

In fact, one possible route to the axioms of probability is to define $P(E)$ to be this long-run frequency. This is the 'frequency interpretation' of probability values.

The right-hand side is a number. But the left-hand side is a random variable: it depends on the exact sequence of outcomes from our independent trials.

So this is saying that, under these long-run average conditions, this 'frequency random variable' settles down, in some approximate sense, to the fixed value P(E).

In this form, the Law of Large Numbers is a mathematical theorem.

It is essential to the whole practice of statistics and sampling.

Key tool to proving it: Chebyshev's inequality.

First, it's valuable to make the situation a bit more general. Instead of an event E, assume our basic experiment has a random variable X. Independent repeats of the experiment give independent copies of this random variable, say $X_1, X_2, ...$

In general, a sequence of RVs $X_1, X_2, ...$ are independent and identically distributed ('i.i.d.') if (i) they are independent, and (ii) they all have the same distribution. For instance,

- if X_i indicates the ith repeat of the event E, and $P(E) = p$, then the X_i's are Bernoulli trials with parameter p;
- OR, they could all be Unif(0, 1), or Poi(λ), or Exp(λ), etc.

To be formal about condition (ii), we should say that the X_i's all have the same CDF (or PMF if discrete, or PDF if continuous).

Let $X_1, X_2, ...$ be i.i.d. RVs. For a positive integer n, define their *sample mean* to be

$$\overline{X}_n = \frac{X_1 + ... + X_n}{n}$$

EXAMPLE: If the X_is are Bernoulli trials with success probability p, then X_n is the fraction of successes among the first n trials. It is a binom(p, n) RV, re-scaled by dividing by n.

Observe: 'identically distributed' implies that $E[X_i]$ is the same for every i, if it exists. Assume it does and call it μ.

Theorem 5. *(Weak Law of Large Numbers, 'WLLN', Ross Thm 8.2.1). In the situation above, for any $\epsilon > 0$, we have*

$$P\{|\overline{X_n} - \mu| \geqslant \epsilon\} \to 0 \text{ as } n \to \infty$$

We will prove this subject to the extra assumption that every X_i has a well-defined and finite variance (some RVs don't!). Again, this must be the same for every i. Call it σ^2. The theorem is actually true without this assumption.

Proof. : Since the X_is are independent, we have

$$E[\overline{X}_n] = \mu \text{ (fixed) and } Var(\overline{X}_n) = \frac{\sigma^2}{n} \text{ (which} \to 0).$$

Then by Chebyshev

$$P\{|\overline{X}_n - \mu| \geqslant \epsilon\} \leqslant \frac{\sigma^2}{n\epsilon^2}$$

□

More expansive statement of the WLLN:
if we choose an 'error tolerance' $\epsilon > 0$, and then wait for n to be large enough, then the 'probability of error'

$$P\{|\overline{X}_n - \mu| \geqslant \epsilon\}$$

will be very small.

BE CAREFUL:

- How long you have to wait (i.e., how large n has to be) depends on how good an approximation you want (i.e., how small you choose ϵ). The proof above gives an explicit estimate for how long we have to wait, given ϵ.
- The WLLN does not say that X_n is guaranteed to be close to p, only that this is very likely. Of course, if we?re very unlucky, we might toss a fair coin but still get the outcome

  $HHHH...H$, or maybe $HHTHHTHHTHHT...HHT$.

  For these very unlikely outcomes, the sample mean takes the values 1 and 2/3 respectively, far away from the true mean, which is 1/2.

Another way of visualizing the Bernoulli-trials case: once n is large, the binom(p, n), PMF puts almost all of its mass into a narrow window around the mean np.

The Strong Law (Ross Sec 8.4): The WLLN has a companion, the Strong LLN (SLLN).

WLLN:

- setting: fix a sufficiently large, finite number n of trials;
- conclusion: for that n, X_n is very likely to be close to its expectation μ.

SLLN:

- setting: consider a truly infinite sequence of trials;
- conclusion: the running sequence of sample means

$$\overline{X}_1 = X_1,\ \overline{X}_2 = \frac{X_1 + X_2}{2},\ \overline{X}_n = \frac{X_1 + ... + X_n}{n}, ...$$

is essentially guaranteed to converge to μ as $n \to \infty$:, i.e., it eventually gets close to μ and then stays close forever.

Theorem 6. *(Strong Law of Large Numbers, 'SLLN', Ross Thm 8.4.1). In the situation above, we have*

$$P\{\lim_{n\to\infty} \overline{X}_n = \mu\} = 1$$

So 'essentially guaranteed' means 'with probability equal to 1'. Alternatively, the event that this convergence fails has probability 0; it is 'infinitely unlikely'.

One can prove SLLN → WLLN (with some work).

But there' s no direct implication WLLN → SLLN: the SLLN is really a stronger statement.

STORY: if we consider our running sequence of sample means X_n, then WLLN says that, for each individual large value of n, X_n is unlikely be far away from μ. But that's an infinite sequence of unlikely events. Even though their individual probabilities are small, we can still imagine that one of them occurs very occasionally. That is, it could be that X_n mostly stays close to μ, but as n increases it very occasionally makes a large deviation away from μ. SLLN says this doesn't happen. Proof is more difficult that WLLN. See Ross Sec 8.4 for a proof under restrictive assumptions.

The Central Limit Theorem, Ross, Secs 5.4.1 & 8.3. Simplest example: $X_1, ..., X_n$ are Bernoulli RVs. Then $\overline{X}_n$ is the fraction of successes from n independent trials, each with success probability p. In this case $\mu = p$.

The WLLN says: when n is large, $\overline{X}_n$ takes a value close to p with high probability.

If 37% of US citizens have visible dandruff, and we randomly select a thousand citizens (a large number, but much less than the US population), then we expect about 37% of those sampled to have visible

dandruff.

But how confident can we be of this approximation? Is a sample of a thousand large enough for the effect to be reliable?

More precise version of the question: Pick two error tolerances, $\epsilon > 0$ and $\alpha > 0$. How large does n have to be so that

$$P\{|\overline{X}_n - p| \geqslant \epsilon\} < \alpha$$

(There were really two kinds of error tolerance involved all along: ϵ is how close you want X_n to be to p, and α is the small probability of error that you allow. We just didn?t give α a name before.)

Can also think of this question by looking again at pictures of binomial PMFs: Now our question is: effectively how wide are spikes around the mean as $n \to \infty$

Theorem 7. *(The Central Limit Theorem, Ross 8.3.1) Let $X_1, ..., X_n$ be i.i.d with $\mu = E[X_i]$ and $\sigma^2 = Var(X_i)$ both finite. Let $S_n = X_1 + ... + X_n$. Then the limiting distribution of $(S_n - n\mu)/\sigma\sqrt{n}$ is $N(0, 1)$, in the following sense.*

$$P\{a < \frac{S_n - n\mu}{\sigma\sqrt{n}} < b\} \to \Phi(b) - \Phi(a) as n \to \infty$$

This result holds for any i.i.d sequence of RVs with finite variance!! For proof, we follow Ross, 7.7 and 8.3.

References

[1] Austin, *Theory of Probability lecture notes*, https://cims.nyu.edu/ tim/TofP
[2] Bernstein, *Theory of Probability lecture notes*, http://www.cims.nyu.edu/ brettb/probSum2015/index.html
[3] Ross, *A First Course in Probability* (9th ed., 2014)