Math-UA.233: Theory of Probability

Lecture 22

Tim Austin

Inequalities (Ross Secs 7.1, 7.2, 8.2)

So far we have spent a lot of the course learning how to compute *exactly* with random variables.

But there are also good reasons to study *estimates* and *inequalities* that concern probabilities and random variables:

- Simple reason: Sometimes we don't have enough information to compute exactly, so we look for an estimate based the information we do have.
- Deeper reason: Certain basic inequalities lead to 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

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

$$\boxed{E[X] \geqslant 0.}$$

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

Immediate consequences:

1. If $a < b$ are reals such that $a \leqslant X \leqslant b$ (that is: X always takes values between a and b), then

$$\boxed{a \leqslant E[X] \leqslant b.}$$

2. If X and Y are two RVs such that $X \geqslant Y$, then

$$\boxed{E[X] \geqslant E[Y].}$$

This property is called **monotonicity of expectation**.

Example (Ross E.g. 7.2d)

Use RVs to prove ***Boole's Inequality****: if* $A_1, \ldots A_n$ *are any events, then*

$$P\Big(A_1 \cup \cdots \cup A_n\Big) \leqslant P(A_1) + \cdots + P(A_n).$$

Example (Eccentric but fun; Ross E.g. 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 (Markov's inequality; Ross Prop 8.2.1)

If X is a non-negative RV, then for any $a > 0$ we have

$$\boxed{P(X \geqslant a) \leqslant \frac{E[X]}{a}.}$$

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

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 (Chebyshev's inequality; Ross Prop 8.2.2)

If X is any random variable with $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then for any $\kappa > 0$ we have

$$\boxed{P(|X - \mu| \geqslant \kappa) \leqslant \frac{\sigma^2}{\kappa^2}.}$$

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

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 $1/k^2$".

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

Example (Ross E.g. 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* $\mathrm{Var}(X) = 25$*, what can be said about the probability that this week's production will be between 40 and 60?*

Example (Ross E.g. 8.2a)

If X is Unif$(0, 10)$, then

$$P(|X - 5| > 4) = 0.2,$$

whereas Chebyshev gives

$$P(|X - 5| > 4) \leqslant \frac{(25/3)}{16} \approx 0.52.$$

So Chebyshev gives us a *guaranteed upper bound*, but in particular cases it need not be a *good estimate*!

The Law of Averages (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*
>
> $$\textit{(frequency of occurrence of } E\textit{)} \approx P(E).$$

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 the 'Law of Averages'.

Again:

> *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*
>
> $$\textit{(frequency of occurrence of } E\textit{)} \approx P(E).$$

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.

It's a good intuition, but the logic is a bit fishy. If we haven't yet defined probability, how do we define 'independence'? This is really a philosophical question, not mathematical.

But even after we've accepted the axioms of probability, the Law of Averages is still very important.

Look again:

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

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 Averages is a mathematical theorem. It justifies one of our most basic probabilistic intuitions. It is essential to the whole practice of statistics and sampling.

Next we are going to prove it.

Key tool: 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 trials of the experiment give independent copies of this random variable, say $X_1, X_2, \ldots$.

More formally:

Definition

*In general, a sequence of RVs $X_1, X_2, \ldots$ are **independent and identically distributed** ('**i.i.d.**') if*

(i) *they are independent, and*

(ii) *they all have the same distribution (i.e., the same CDF, or PMF if discrete, or PDF if continuous).*

For instance,

- if X_i indicates the i^{th} 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 $\text{Unif}(0, 1)$, or $\text{Poi}(\lambda)$, or $\text{Exp}(\lambda)$, or whatever.

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

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

EXAMPLE: If the X_is are Bernoulli trials with success probability p, then $\overline{X}_n$ is the fraction of successes among the first n trials. It is a binom(n, p) 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 (Weak Law of Large Numbers, 'WLLN' (Ross Thm 2.1))

In the situation above, for any $\varepsilon > 0$, *we have*

$$\boxed{P\left(\left|\overline{X}_n - \mu\right| \geqslant \varepsilon\right) \longrightarrow 0 \quad \text{as } n \longrightarrow \infty.}$$

We will prove this under an extra assumption: that $\mathrm{Var}(X_i)$ is finite for every i (not always true!). Like $E[X_i]$, this variance must be the same for every i. Call it σ^2. (The theorem is actually true without this assumption.)

IDEA: Since the X_is are independent, we have

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

Now apply Chebyshev to $\overline{X}_n$.

More expansive statement of the WLLN:

- if we choose an 'error tolerance' $\varepsilon > 0$,
- and then wait for n to be large enough,
- then the 'probability of error'

$$P\left(\left|\overline{X}_n - \mu\right| \geqslant \varepsilon\right)$$

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 $\overline{X}_n$ is *guaranteed* to be close to p, only that this is very *likely*. If we're *very* unlucky, we might toss a fair coin 1000 times and get the outcome

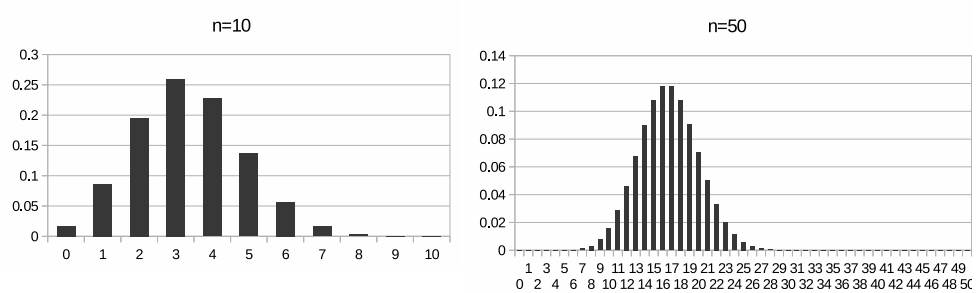
$$\underbrace{HHHH\ldots H}_{\text{fraction of heads } =1},$$

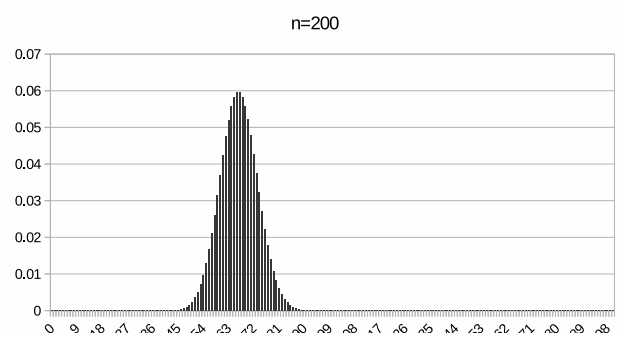
or maybe

$$\underbrace{HHTHHTHHTHHT\ldots HHTHHTH}_{\text{fraction of heads } =2/3}$$

— even though the true success probability is $1/2$.

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





In the next lecture we will meet a more refined estimate for how the PMF is spread around the mean, the Central Limit Theorem.

The Strong Law of Large Numbers (Ross Sec 8.4)

... an improvement of the Weak LLN.

Weak LLN:

- setting: fix a very large, finite number n of trials;
- conclusion: $\overline{X}_n$ is very likely to be close to its expectation μ.

Strong LLN:

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

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

will converge to μ as $n \longrightarrow \infty$: i.e. it eventually gets close to μ and then *stays close forever*.

Theorem (Strong Law of Large Numbers, 'SLLN' (Ross Thm 2.1))

In the situation above, we have

$$\boxed{P\left(\lim_{n\longrightarrow\infty} \overline{X}_n = \mu\right) = 1.}$$

So this convergence happens 'with probability equal to 1'. Equivalently, *non*-convergence happens with probability 0: it is 'infinitely unlikely'.

One can prove SLLN $\Longrightarrow$ WLLN (tricky, but not too bad).

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

STORY: if we consider our running sequence of sample means $\overline{X}_n$, then WLLN says that, for each individual large value of n, $\overline{X}_n$ is unlikely to 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, maybe $\overline{X}_n$ 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 than WLLN. See Ross Sec 8.4 for a sort-of proof.

Here I only care that you understand the statement.

Pictures to illustrate the story: from computer simulations of long sequences of i.i.d. RVs. Here are one with dice and another with coins.