Math-UA.233: Theory of Probability
Lecture 23

Tim Austin

The Central Limit Theorem (Ross Secs 5.4.1 & 8.3)

Last time we met the weak law of large numbers:

If $X_1, X_2, \ldots$ are i.i.d. RVs, all with expectation equal to μ, then for any choice of error tolerance $\varepsilon > 0$ we have

$$\boxed{P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \longrightarrow 0 \quad \text{as } n \longrightarrow \infty.}$$

The averaged RV

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

is called the **sample mean**.

Simplest example: $X_1, \dots, X_n$ are Bernoulli(p) 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.*

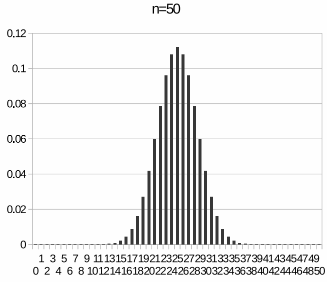
But how confident can we be of this approximation? Is a sample of a thousand large enough for the effect to be reliable?

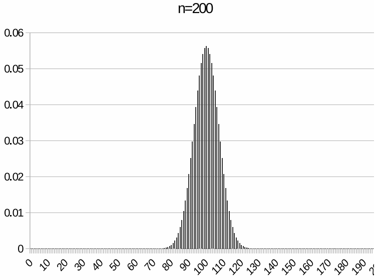
More precise version of the question:

> *Pick two error tolerances, $\varepsilon > 0$ and $\alpha > 0$. How large does n have to be so that*
>
> $$P\big(|\overline{X}_n - p| \geqslant \varepsilon\big) < \alpha \quad ?$$

(There were really *two* kinds of error involved all along: ε is how close you want $\overline{X}_n$ to be to p, and α is the small probability of error that you allow. We didn't give α a name before.)

Can also think of this question by looking again at those pictures of binomial PMFs:





Now our question is: effectively how wide *are* those spikes around the mean, as $n \longrightarrow \infty$?

Will answer this with a more refined approximation.

Let's focus on the Bernoulli$(1/2)$ case for now.

NOTATION: let

$$S_n = X_1 + \cdots + X_n.$$

Then $E[S_n] = n/2$.

(For this discussion, S_n is more convenient than $\overline{X}_n = S_n/n$.)

We want to approximate the 'shape' of the PMF

$$P(S_n = k)$$

when k/n is close to $1/2$, i.e. in a range
$n/2 - n\varepsilon < k < n/2 + n\varepsilon$.

We start from the formula

$$P(S_n = k) = \binom{n}{k} 2^{-n} = \frac{n!}{k!(n-k)!} 2^{-n}.$$

<u>IDEA 1:</u> This formula is a complicated product.
<u>Take logs in order to study a sum instead.</u>

Get:

$$\log P(S_n = k) = \log n! - \log k! - \log(n-k)! - n \log 2.$$

Now observe that

$$\log n! = \log(1 \times 2 \times \cdots \times n) = \log 1 + \log 2 + \log 3 + \cdots + \log n.$$

Similarly for $\log k!$ and $\log(n-k)!$.

<u>IDEA 2:</u> Approximate this sum by the integral

$$\int_1^n \log x \, dx \ , \quad \text{which equals} \quad n \log n - n + 1.$$

Also: drop the $+1$, as it's so small relative to the rest.

SPOILER: This approximation is *not* quite good enough to prove the result we want, but it comes close. Let's proceed, and discuss this correction at the end.

Inserting the integral approximation back into our original formula:

$$
\begin{aligned}
&\log P(S_n = k) \\
&= \log n! - \log k! - \log(n-k)! - n\log 2 \\
&\approx \int_1^n \log x\, dx - \int_1^k \log x\, dx - \int_1^{n-k} \log x\, dx - n\log 2 \\
&\approx [n\log n - n] - [k\log k - k] \\
&\qquad - [(n-k)\log(n-k) - (n-k)] - n\log 2 \\
&= n\log n \underbrace{-k\log k - (n-k)\log(n-k)}_{\text{the tricky parts, } \because \text{ depend on } k} - n\log 2.
\end{aligned}
$$

IDEA 3. We're only concerned with the range $n/2 - \varepsilon n < k < n/2 + \varepsilon n$ for some small ε. Let's define

$$\delta = k/n - 1/2,$$

so $-\varepsilon < \delta < \varepsilon$, and write our approximation in terms of δ.

In terms of δ, we have

$$k = (1/2 + \delta)n \quad \text{and so} \quad n - k = (1/2 - \delta)n.$$

Thus:

$$\begin{aligned} k \log k &= (1/2 + \delta)n\big[\log(1/2 + \delta) + \log n\big] \quad \text{and} \\ (n-k)\log(n-k) &= (1/2 - \delta)n\big[\log(1/2 - \delta) + \log n\big]. \end{aligned}$$

Adding these together:

$$
\begin{aligned}
& k \log k + (n-k) \log(n-k) \\
& = (1/2+\delta) n \big[\log(1/2+\delta) + \log n \big] \\
& \qquad + (1/2-\delta) n \big[\log(1/2-\delta) + \log n \big] \\
& = n \log n + \big[(1/2+\delta) \log(1/2+\delta) + (1/2-\delta) \log(1/2-\delta) \big] n.
\end{aligned}
$$

Putting this back into our probability approximation:

$$
\begin{aligned}
& \log P(S_n = k) \\
& \approx n \log n - k \log k - (n-k) \log(n-k) - n \log 2 - 1 \\
& = -\big[(1/2+\delta) \log(1/2+\delta) + (1/2-\delta) \log(1/2-\delta) \big] n \\
& \qquad - n \log 2.
\end{aligned}
$$

IDEA 4. Look at the function

$$D(\delta) = (1/2 + \delta)\log(1/2 + \delta) + (1/2 - \delta)\log(1/2 - \delta).$$

Since we only care about small δ, let's approximate this with its Taylor expansion around $\delta = 0$.

You can get this by substituting from the Taylor expansion of log itself around $1/2$.

The result:

$$\begin{aligned} D(\delta) &\approx \log\frac{1}{2} + 0 \times \delta + 2 \times \delta^2 + \text{(higher order in } \delta\text{)}. \\ &= -\log 2 + 2\delta^2 + \text{(higher order in } \delta\text{)}. \end{aligned}$$

(The function $D(\delta)$, and some closely related things, appear in the analysis of many important probability questions. It is a variant of the **Shannon entropy function**.)

Putting this back into our probability approximation:

$$\begin{aligned}\log P(S_n = k) &\approx -D(\delta)n - n\log 2 \\ &\approx n\log 2 - 2n\delta^2 - n\log 2 \\ &= 2n\delta^2.\end{aligned}$$

Therefore, finally:

$$P(S_n = k) \approx e^{-2n\delta^2}$$
$$= \exp\Big[- \frac{(k - n/2)^2}{n/2} \Big]$$

<u>CONCLUSION:</u> The PMF of S_n is roughly the PDF of a normal RV with parameters

$$\mu = n/2 = E[S_n] \quad \text{and} \quad \sigma^2 = n/4 = \mathrm{Var}(S_n).$$

CAVEATS AND CORRECTION: The approximation

$$P(S_n = k) \approx \exp\Big[-\frac{(k - n/2)^2}{n/2} \Big]$$

cannot be quite right: for the right-hand side to be a normal PDF, it should also have a constant factor of

$$\frac{1}{\sqrt{2\pi}\sigma} = \sqrt{\frac{2}{n\pi}}.$$

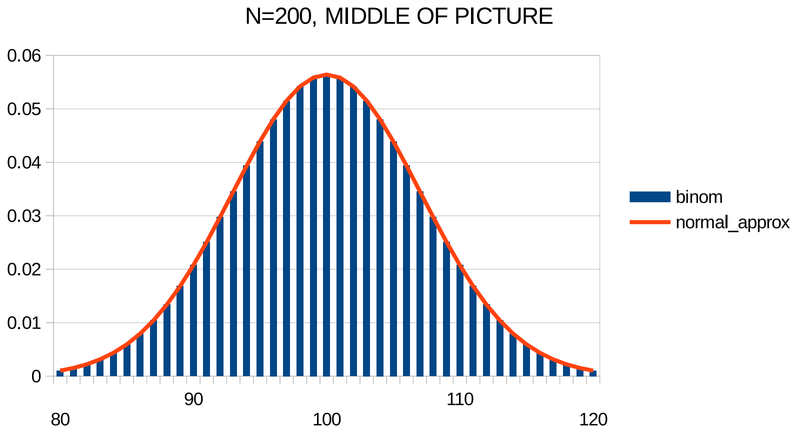
This error is because that integral approximation we used isn't quite good enough.

It can be improved to a famous fact called **Stirling's approximation**

$$n! \approx \sqrt{2\pi n}\frac{n^n}{e^n} \qquad \text{(i.e. } \tfrac{\text{LHS}}{\text{RHS}} \longrightarrow 1 \text{ as } n \longrightarrow \infty\text{).}$$

Using this instead, then you also get the right constant in front.

Picture with $n = 200$, and with accompanying normal curve:



The previous story generalizes easily to binom(n, p) for $p \neq 1/2$. Now the WLLN tells us to focus on

$$\frac{k}{n} = p + \delta \quad \text{for some very small } \delta.$$

The same ideas as above lead to

$$P(S_n = k) \approx \frac{1}{\sqrt{2\pi n}} \times \frac{1}{\sqrt{p(1-p)}} \times \exp\Big(- \frac{(k-np)^2}{2np(1-p)} \Big).$$

(For details of this proof, see Sec VII.2 of Feller's *An Introduction to Probability Theory and its Applications, Vol 1*, or the Wikipedia article for “De Moivre–Laplace Theorem”.)

More compact formula: for Bernoulli(p), we have $\sigma^2 = \mathrm{Var}(X_i) = p(1-p)$, so

$$\boxed{P(S_n = k) \approx \frac{1}{\sqrt{2\pi n}\sigma} \exp\Big(-\frac{(k-np)^2}{2n\sigma^2}\Big).}$$

This is called the *De Moivre–Laplace Limit Theorem* or the *normal approximation to the binomial*.

WARNING: Do not confuse with the Poisson approximation. They are good under *different conditions*:

- Poisson approximation: $n \longrightarrow \infty$, $p \longrightarrow 0$ so that $np = \lambda$ stays constant;
- De Moivre–Laplace: $n \longrightarrow \infty$, *p is fixed*.

Example (Similar to Ross E.g. 5.4g)

Let a fair coin be flipped 40 times, and let X be the number of heads obtained. Find the probability that $X = 20$, and compare with the normal approximation.

ANS: From binomial formula:

$$P(X = 20) = \binom{40}{20} 2^{-40} \approx 0.1254$$

From normal approximation: observe that $n = 40$ and $k = 20 = n/2$, so we get

$$P(X = 20) \approx \frac{2}{\sqrt{2\pi \times 40}} e^{0} \approx 0.1262.$$

Again:

$$\boxed{P(S_n = k) \approx \frac{1}{\sqrt{2\pi n}\sigma} \exp\Big(-\frac{(k-np)^2}{2n\sigma^2}\Big).}$$

This is great, but it's still not quite the theorem we want.

It says these two sides are comparable, i.e. their ratio is close to one. But in fact both sides are usually *really small*: if n is large, then S_n has tiny probability of hitting any exact value k.

More useful: estimate the probability that S_n lands in a certain *range*.

To obtain that estimate, it's easiest to translate S_n so that its expectation is zero (i.e., subtract np), and re-scale so that its variance is one (i.e. divide by $\mathrm{SD}(S_n) = \sigma\sqrt{n}$). That is, we need to *standardize*. This means we will consider the *standard version* of S_n:

$$\frac{S_n - np}{\sigma\sqrt{n}}, \quad \text{where } \sigma = \sqrt{p(1-p)}.$$

(Be careful not to confuse this with $\overline{X}_n$, which is not translated, and is divided by n rather than $\sigma\sqrt{n}$.)

Now for any real values $a < b$, we have

$$P\Big(a < \frac{S_n - np}{\sigma\sqrt{n}} < b\Big) = P\big(np + a\sigma\sqrt{n} < S_n < np + b\sigma\sqrt{n}\big)$$
$$= \sum_{np+a\sigma\sqrt{n}<k<np+b\sigma\sqrt{n}} P(S_n = k).$$

Substituting from previous approximation, this becomes

$$\frac{1}{\sqrt{2\pi n}\sigma} \sum_{np+a\sigma\sqrt{n}<k<np+b\sigma\sqrt{n}} \exp\Big(-\frac{(k-np)^2}{2n\sigma^2}\Big)$$
$$= \frac{1}{\sqrt{2\pi n}\sigma} \sum_{a\sigma\sqrt{n}<\ell<b\sigma\sqrt{n}} \exp\Big(-\frac{1}{2}\Big(\frac{\ell}{\sigma\sqrt{n}}\Big)^2\Big).$$

Again:

$$P\Big(a < \frac{S_n - np}{\sigma\sqrt{n}} < b\Big) \approx \frac{1}{\sqrt{2\pi n}\sigma} \sum_{a\sigma\sqrt{n}<\ell<b\sigma\sqrt{n}} \exp\Big(-\frac{1}{2}\Big(\frac{\ell}{\sigma\sqrt{n}}\Big)^2\Big).$$

LAST KEY IDEA: <u>This is a Riemann sum.</u> We have dissected the interval $[a, b]$ into smaller intervals of length $1/\sigma\sqrt{n}$, and we are summing the function

$$\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big)$$

over those smaller steps. As $n \longrightarrow \infty$, the step-size goes to zero, and the sum converges to the integral:

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,dx = \Phi(b) - \Phi(a).$$

(Recall: Φ is the CDF of $N(0, 1)$.)

Summary:

Theorem

If S_n is binom(n, p), then its standard version satisfies

$$P\Big(a < \frac{S_n - np}{\sigma\sqrt{n}} < b\Big) \longrightarrow \Phi(b) - \Phi(a) \quad \textit{as } n \longrightarrow \infty.$$

(This version is what Ross calls the ‘De Moivre–Laplace Theorem’: see p194)

Example (Ross E.g. 5.4h)

The ideal size of a first-year class at The College is 150 students. On average, only 30% of the applicants accepted actually attend. So they make offers to 450 students. Estimate the probability that more that 150 new students attend.

PROCEDURE: Step 1: Let $X_1, \ldots, X_{450}$ indicate the attendance of each student. Modeling assumption: these are i.i.d. Bernoulli(0.3). Find

$$E[X_i] = 0.3, \quad \mathrm{Var}(X_i) = (0.3)(0.7).$$

Step 2: Standardize $S_{450} = X_1 + \cdots + X_{450}$ (i.e., write the question in terms of the standard version).

Step 3: Use the table of values for Φ.

ANS: $1 - \Phi(1.59) \approx 0.0559$.

Finally, the real miracle:

Theorem (Central Limit Theorem; Ross Theorem 8.3.1)

Let $X_1, \ldots, X_n$ be i.i.d. with $\mu = E[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i)$ both finite. Then the limiting distribution of $(S_n - n\mu)/\sigma\sqrt{n}$ is $N(0,1)$, in the following sense:

$$P\left(a < \frac{S_n - n\mu}{\sigma\sqrt{n}} < b\right) \longrightarrow \Phi(b) - \Phi(a) \quad \text{as } n \longrightarrow \infty.$$

That is, the same phenomenon holds for *any* i.i.d. sequence of RVs with finite variance! No matter *what* the distribution!

THIS IS THE ONE YOU NEED TO REMEMBER, AND LEARN TO USE.

Rule of thumb: for most 'common' examples of the X_is, the effect is already pretty good once $n \geqslant 10$.

Example (Ross E.g. 8.3c)

If 10 fair dice are rolled, find the approximate probability that the sum obtained is between 30 through 40.

PROCEDURE: Step 1: Find

$$E[X_i] = 3.5, \quad \mathrm{Var}(X_i) = 35/12.$$

Step 2: Standardize $S_{10} = X_1 + \cdots + X_{10}$ (i.e., write the question in terms of the standard version).

Step 3: Use the table of values for Φ.

ANS: $2\Phi(1.0184) - 1 \approx 0.692$.

Example (Ross E.g. 8.3d)

Let X_i, $i = 1, 2, \ldots, 10$, be independent Unif$(0, 1)$ RVs. Calculate an approximation to $P\left(\sum_{i=1}^{10} X_i > 6\right)$.

PROCEDURE: Step 1: $E[X_i] = 1/2$, $\mathrm{Var}(X_i) = 1/12$.

Step 2: Standardize $S_{10} = X_1 + \cdots + X_{10}$.

Step 3: Use the table of values for Φ.

ANS: $1 - \Phi(\sqrt{1.2}) \approx 0.1367$.