Math-UA.233: Theory of Probability

Lecture 24

Tim Austin

Previously...

Let $X_1, X_2, \ldots$ are i.i.d. RVs, all with $E[X_i] = \mu$. Let

$$S_n = X_1 + \cdots + X_n \quad \text{and} \quad \overline{X}_n = S_n/n.$$

The **weak law of large numbers**: for any $\varepsilon > 0$ we have

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = P\left(\mu n - \varepsilon n < S_n < \mu n + \varepsilon n\right) \longrightarrow 0 \quad \text{as } n \longrightarrow \infty.$$

The **central limit theorem**: if $\sigma^2 = \mathrm{Var}(X_i)$ is finite, then for any $a < b$ we have

$$P\left(a < \underbrace{\frac{S_n - n\mu}{\sigma\sqrt{n}}}_{\text{standard version}} < b\right) = P\left(\mu n + a\sigma\sqrt{n} < S_n < \mu n + b\sigma\sqrt{n}\right)$$

$$\longrightarrow \Phi(b) - \Phi(a) \quad \text{as } n \longrightarrow \infty.$$

A heuristic way to state these results together: as $n \longrightarrow \infty$, we have

$$S_n = \underbrace{\mu n}_{\text{from WLLN}} + \underbrace{\sigma Z \sqrt{n}}_{\text{from CLT}},$$

where Z is approximately an $\mathrm{N}(0,1)$ RV

(where 'approximately' means

$$P(a < Z < b) \approx \Phi(b) - \Phi(a).)$$

So:

- WLLN describes the "leading-order" behaviour of S_n — i.e., the term which is comparable to n;
- CLT describes the size and shape of the "fluctuations" — i.e. the term which is typically comparable to $\sqrt{n}$

The CLT is everywhere

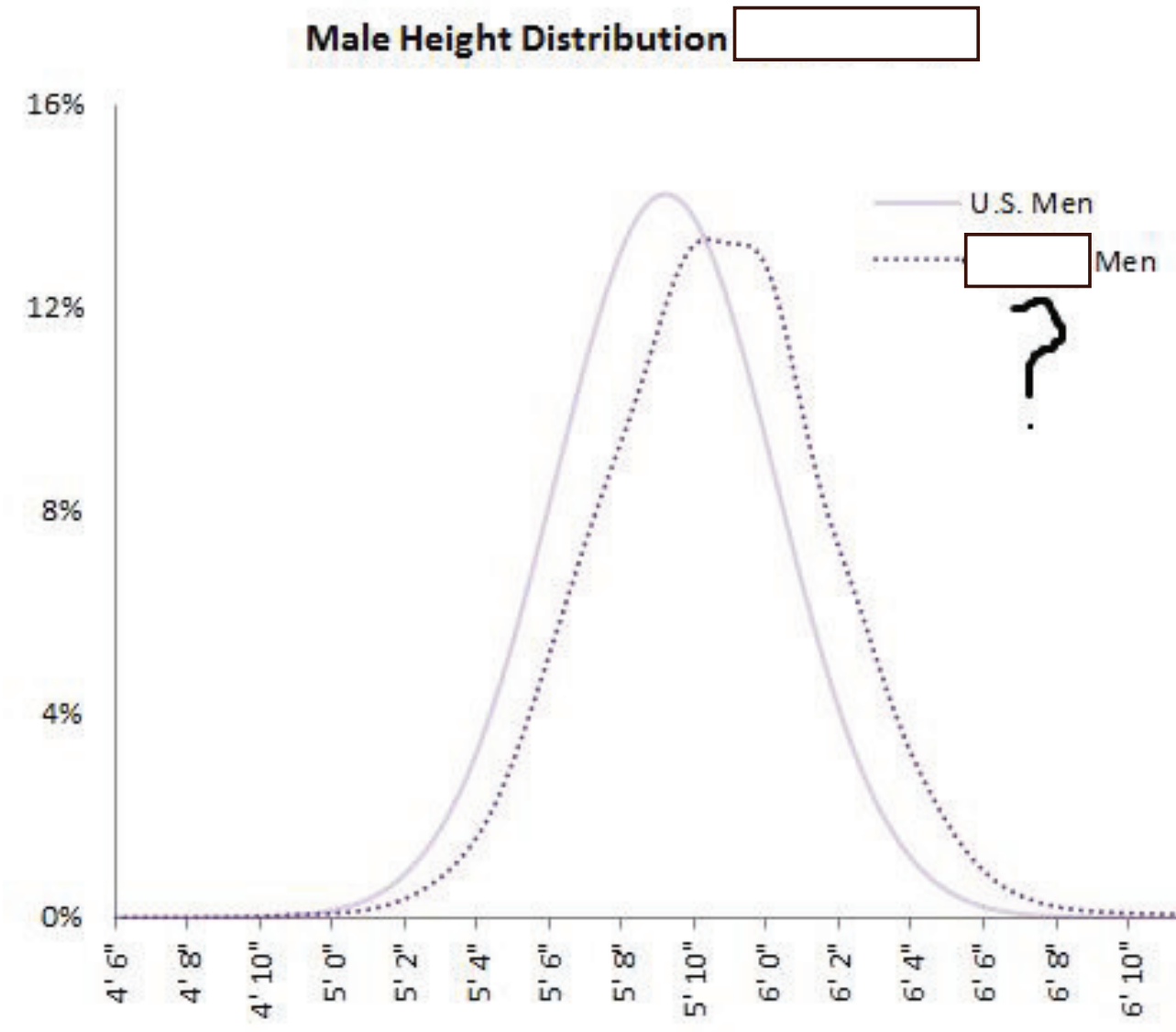
Because of the CLT, the normal distributions shows up *all over the place* in real-world data.

Indeed, whenever a random quantity is an *accumulation of small independent contributions* — that is, a sum of small independent RVs — the CLT promises an approximately normal distribution.

- Example: velocity of a randomly-chosen particle in a gas. That velocity has been *added up* over many small, roughly independent collisions.
- Example: height distribution for people of a given sex, age and ethnicity. The real-world data does look fairly normal here, but finding the right explanation seems tricky.

Question for today: What is this?

Answer at the end of class



Today: more practice, and sketch of some areas of application.

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of t

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992

Example (Ross E.g. 8.3a)

Pooja wants to measure the distance to a star. She knows her radio telescope makes measurement errors because of clouds. So she makes several measurements on different days and takes an average. The results are $X_1, \ldots, X_n$: i.i.d. RVs measured in light-years. Previous experience with the telescope tells her that the variance is 4 (light-years)^2. How many measurements does she need to be fairly sure that the average is within 0.5 light-years of the true distance d?

The probability in question: $P(-0.5 \leqslant \overline{X}_n - d \leqslant 0.5)$.

PROCEDURE: standardize the question *carefully*. Decide how close to 1 you want to this probability to be, then find n from the Φ-table.

NOTE ABOUT PRECEDING EXAMPLE: $n = 62$ is typically large enough to apply CLT to assert 95% confidence.

But this is *not a guarantee*. There can be tricky RVs for which the CLT is *not* a good approximation when $n = 62$: see last question on hwk 12.

The best *guarantee* you can get from the information in the question is by going back to Chebyshev:

$$E[\overline{X}_n] = d, \quad \mathrm{Var}(\overline{X}_n) = \sigma^2/n = 4/n,$$

so

$$P(|\overline{X}_n - d| > 0.5) \leqslant \frac{4}{n \times (0.5)^2} = 16/n.$$

This *guarantees* 95% confidence once $n \geqslant 320$.

Application: sampling, estimators and confidence

The basic problem of statistics:

> *We have some real-world phenomenon described by a RV whose distribution we don't know. By sampling from it many times independently, we try to figure out that distribution (or at least its mean, variance, ...).*

Examples:

- What proportion of the electorate plan to vote for Alice rather than Bob? (RV in this example: indicator that a randomly-chosen voter will vote for Alice.)
- What is the distribution of lifespans among a certain breed of sheep? (Relevant for insuring sheep.)
- What is the expected tensile strength of a yard of rebar from a given steel mill?

Again: the goal is to sample many times independently from some unknown distribution in order to find out about it.

- Reason this can work *at all*: WLLN.
- Way to estimate quality of the resulting information: CLT.

Example (Ross E.g. 5.4j)

*52% of NYC residents are in favour of creating 'local' and 'tourist' lanes on certain busy sidewalks (*not a real statistic). Suppose we sample n residents at random. What is the approximate probability that more than half the sample are in favour when*

(a) $n = 11$,

(b) $n = 101$,

(c) $n = 1001$*?*

Part (d): How large must n be so that this probability exceeds 95%?

NOTE: To understand this completely, there are *two* approximation steps.

(Local/tourist lanes, cont.)

First, in truth, we will sample n people *without replacement* from the total number, say N, of NYC residents. That's $\binom{N}{n}$ equally likely selections of the n people. So exact answer is:

$$\sum_{k=(n+1)/2}^{n} \frac{\binom{(0.52)N}{k}\binom{(0.48)N}{n-k}}{\binom{N}{n}}$$

(because k, the number in favour in the sample, is hypergeometric$(N, (0.52)N, n)$).

This is much too complicated. But N is enormously bigger than n, so let's instead analyze sampling *with* replacement. That is, we allow the possibility of sampling the same person twice — the actual probability of this is negligible. (This was the "binomial approx to hypergeometric")

(Local/tourist lanes, cont.)

So now we have n Bernoulli(0.52) RVs $X_1, \ldots, X_n$.

X_i indicates the event that the i^{th} person sampled is in favour.

New answer:

$$\sum_{k=(n+1)/2}^{n} \binom{n}{k} (0.52)^k (0.48)^{n-k}.$$

Still too complicated. So now we apply CLT to this.

(Local/tourist lanes, cont.)

Let $S_n = X_1 + \cdots + X_n$. We want $P(S_n > n/2)$.

Standardize: this is the same as

$$P\Big(\frac{S_n - (0.52)n}{\sqrt{(0.52)(0.48)n}} > \frac{(0.5)n - (0.52)n}{\sqrt{(0.52)(0.48)n}}\Big)$$
$$= P\Big(\frac{S_n - (0.52)n}{\sqrt{(0.52)(0.48)n}} > -(0.04)\sqrt{n}\Big) \approx \Phi((0.04)\sqrt{n}).$$

Now Φ-table gives approx answers:

(a) $n = 11$: ANS equals $\Phi(0.1328) = 0.5528$;

(b) $n = 101$: ANS equals $\Phi(0.4020) = 0.6562$;

(c) $n = 1001$: ANS equals $\Phi(1.2665) = 0.8973$.

(d) For prob at least 0.95, need $(0.04)\sqrt{n} \geqslant 1.645$, hence $n \geqslant 1692$.

Real-life examples are usually not like the last one: usually we would not *know* the 52% figure in advance, so we could not use it to compute our choice of n!

(And anyway, with modern resources, its rarely hard to get a very large sample — mistakes are made because it's hard to get an *unbiased* sample.)

A more realistic question:

> *Given your sample, how confident can you be in $\overline{X}_n$ as an estimate for the mean?*

In general, suppose $X_1, \ldots, X_n$ are i.i.d. samples from a distribution with mean μ and variance σ^2. Suppose we don't know μ or σ^2, but we do know an upper bound, say σ_{bound}, for σ.

Example

$\text{Var}(\text{Bernoulli}(p)) = p(1-p)$, *which* $\leqslant 1/4$ *for every possible* p.

Now

$$P\Big(-\frac{2\sigma_{\text{bound}}}{\sqrt{n}} < \overline{X}_n - \mu < \frac{2\sigma_{\text{bound}}}{\sqrt{n}}\Big) \geqslant P\Big(-\frac{2\sigma}{\sqrt{n}} < \overline{X}_n - \mu < \frac{2\sigma}{\sqrt{n}}\Big)$$

$$= P\Big(-2 < \underbrace{\frac{S_n - n\mu}{\sigma\sqrt{n}}}_{\text{standard version}} < 2\Big),$$

and by CLT this is

$$\approx \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 \approx 0.9544 > 0.95.$$

Again: if we assume CLT approximation is good, then

$$P\Big(-\frac{2\sigma_{\text{bound}}}{\sqrt{n}} < \overline{X}_n - \mu < \frac{2\sigma_{\text{bound}}}{\sqrt{n}} \Big) > 0.95.$$

So we are 95% sure that μ satisfies

$$\overline{X}_n - \frac{2\sigma_{\text{bound}}}{\sqrt{n}} < \mu < \overline{X}_n + \frac{2\sigma_{\text{bound}}}{\sqrt{n}}.$$

This reasoning works even though we don't know what μ is!

This is called the **95% confidence interval** for μ, based on the data we have.

In the NYC local/tourist lanes example, we found that once $n \geqslant 1692$, the 95% confidence interval has width less than 4%. If the true mean is 52%, then with $>$ 95% the confidence interval will lie entirely above 50%.

Realistic procedure when polling for a win/lose vote:

- Choose an initial n which you guess is large enough;
- Sample n people as randomly as you can to avoid bias;
- Compute the sample mean $\overline{X}_n$ from the sample, and obtain the confidence interval

$$\left[\overline{X}_n - \frac{1}{\sqrt{n}}, \overline{X}_n + \frac{1}{\sqrt{n}}\right]$$

(because we can use $\sigma_{\text{bound}} = 1/2$ for Bernoulli examples)

- If the confidence interval lies entirely above [below] 50%, predict win [lose] with confidence $> 95\%$. Otherwise, say "I don't know", and try to get money for a bigger sample.

Example (Ross E.g. 8.3e)

An instructor has 50 exams that will be graded in sequence. The times requires to grade the 50 exams are independent, with a common distribution that has mean 20 minutes and standard deviation 4 (minutes). Approximate the probability that the instructor will grade at least 25 of the exams in the first 450 minutes of work.

ANS: $1 - \Phi(2.5) \approx 0.006$

ANOTHER HANDY TRICK:

Suppose each X_i takes *integer values*, let

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

and you want to apply the CLT to approximate $P(a \leqslant S_n \leqslant b)$ for some *integers* a, b.

Then the event $\{a \leqslant S_n \leqslant b\}$ is the same as the event $\{a - 0.5 < S_n < b + 0.5\}$, because S_n must be an integer.

But when you apply the CLT, you usually get a slightly better approximation using the interval $(a - 0.5, b + 0.5)$ than $[a, b]$. This is especially true if a and b are close together.

This trick is called the **continuity correction**. It applies *only* in case S_n is integer-valued.

Example (Pishro-Nik, pp384-5)

Suppose Y is binom$(20, 1/2)$. Compute/estimate $P(8 \leqslant Y \leqslant 10)$.

- Exact:

$$\left[\binom{20}{8} + \binom{20}{9} + \binom{20}{10}\right]\left(\frac{1}{2}\right)^{20} \approx 0.4565$$

- Using CLT without continuity correction:

$$P\left(\frac{8 - 20 \times \frac{1}{2}}{\sqrt{20} \times \frac{1}{2}} \leqslant \frac{Y - 20 \times \frac{1}{2}}{\sqrt{20} \times \frac{1}{2}} \leqslant \frac{10 - 20 \times \frac{1}{2}}{\sqrt{20} \times \frac{1}{2}}\right) \approx 0.3145$$

- Using CLT with continuity correction:

$$P\left(\frac{7.5 - 20 \times \frac{1}{2}}{\sqrt{20} \times \frac{1}{2}} < \frac{Y - 20 \times \frac{1}{2}}{\sqrt{20} \times \frac{1}{2}} < \frac{10.5 - 20 \times \frac{1}{2}}{\sqrt{20} \times \frac{1}{2}}\right) \approx 0.4567$$

Example (Pishro-Nik, solved problem 7.1.2)

Let $X_1, \ldots, X_{25}$ be i.i.d. with possible values ± 1 and PMF

$$p(1) = 0.6, \quad p(-1) = 0.4.$$

Let $Y = X_1 + \cdots + X_n$. Estimate $P(4 \leqslant Y \leqslant 6)$.

Good example where the continuity correction is appropriate.

Calcs: $E[X_1] = \mu = 1/5$, $E[X_1^2] = 1$, $\mathrm{Var}(X_1) = \sigma^2 = 24/25$.

ANS: $2\Phi(0.3062) - 1 \approx 0.2405$

(See Pishro-Nik solved probs for section 7.1: more examples.)

Example (Ross E.g. 8.3b)

The number of students who enroll in a psychology course is a Poisson RV with mean 100. The administration will split the course into two sections if enrollment as at least 120. Approximate the probability of this happening.

This is a "hidden" CLT question. Key thing to realize: $\text{Poi}(100)$ is equal to a sum of 100 independent $\text{Poi}(1)$'s.

Since the number of students is an integer, we can (and should) start by applying the continuity correction.

Then, follow the procedure: formulate the question; standardize it; go to the Φ-table.

Here's an example of 'hypothesis testing' — another topic from statistics.

Example (Based on Ross E.g. 5.4i)

Dr Ajax claims that his diet reduces cholesterol in 80% of people. Dr Sour wants to test this, so he puts 100 people on the diet. He will endorse the diet if 65 people or more experience reduced cholesterol. Assume that normally cholesterol fluctuates up or down with equal probability.

(a) *Find the probability of Dr Sour's endorsement if we assume that diet actually does nothing.*

(b) *If Prof Fiddle initially gives Dr Ajax's diet only a 5% probability of effectiveness, how should she change that estimate once she hears of Dr Sour's endorsement?*

METHOD for (b): CLT approximation + Bayes' formula.

Question for today: What is this?