Central Limit Theorem

CSE 312 Spring 24
Lecture 18

Why Learn Normals?

When we add together independent normal random variables, you get another normal random variable.

The sum of **any** independent random variables **approaches** a normal distribution.

Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables, with mean μ and variance σ^2. Let $Y_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$

As $n \to \infty$, the CDF of Y_n converges to the CDF of $\mathcal{N}(0, 1)$

Breaking down the theorem

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, with mean μ and variance σ^2. Let $Y_n = \frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$

As $n \to \infty$, the CDF of Y_n converges to the CDF of $\mathcal{N}(0,1)$

Proof of the CLT?

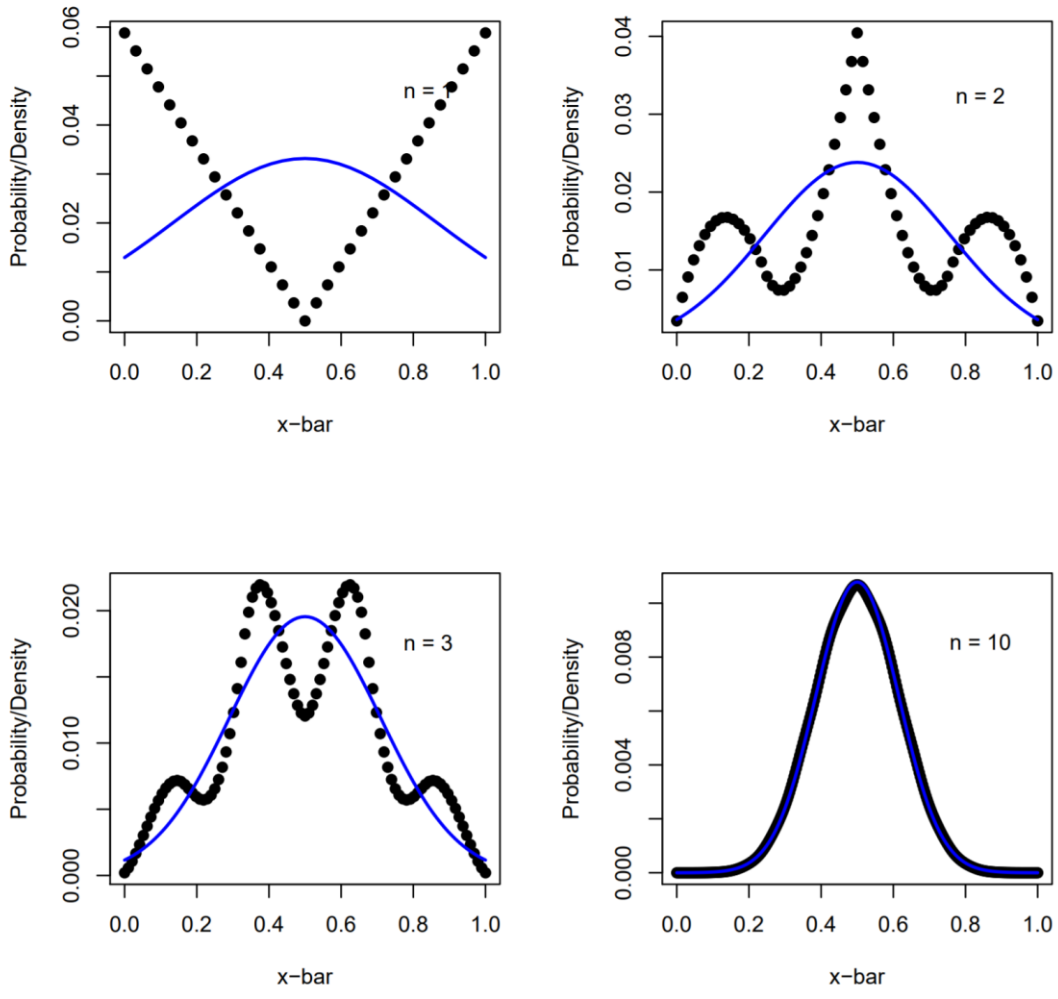
No.

How is the proof done?

Step 1: Prove that for all positive integers k, $\mathbb{E}\left[(Y_n)^k\right] \to \mathbb{E}[Z^k]$

Step 2: Prove that if $\mathbb{E}\left[(Y_n)^k\right] = \mathbb{E}\left[Z^k\right]$ for all k then $F_{Y_n}(z) = F_Z(z)$

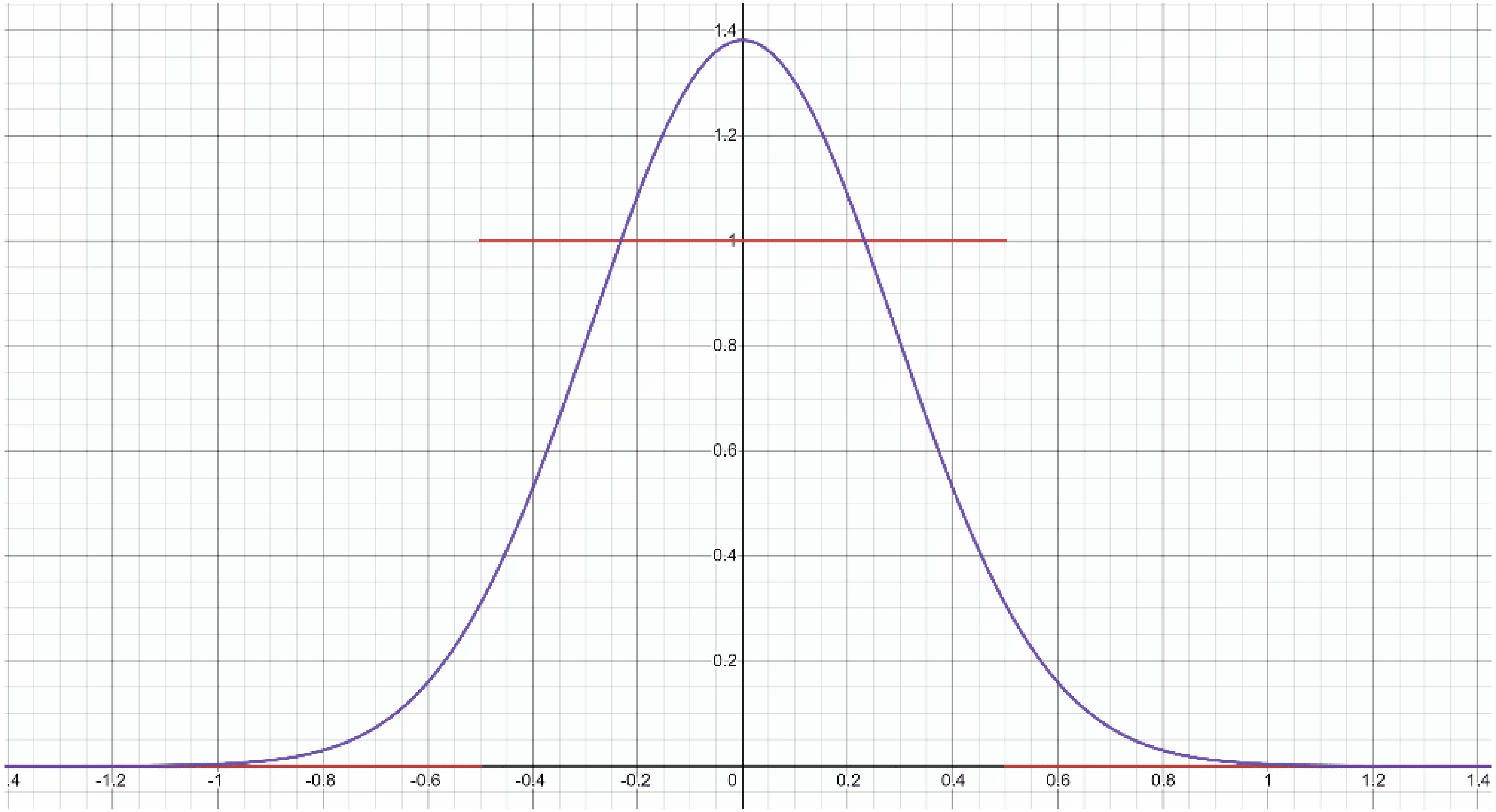
"Proof by example"



The dotted lines show an "empirical pmf" – a pmf estimated by running the experiment a large number of times.
The blue line is the normal rv that the CLT predicts.

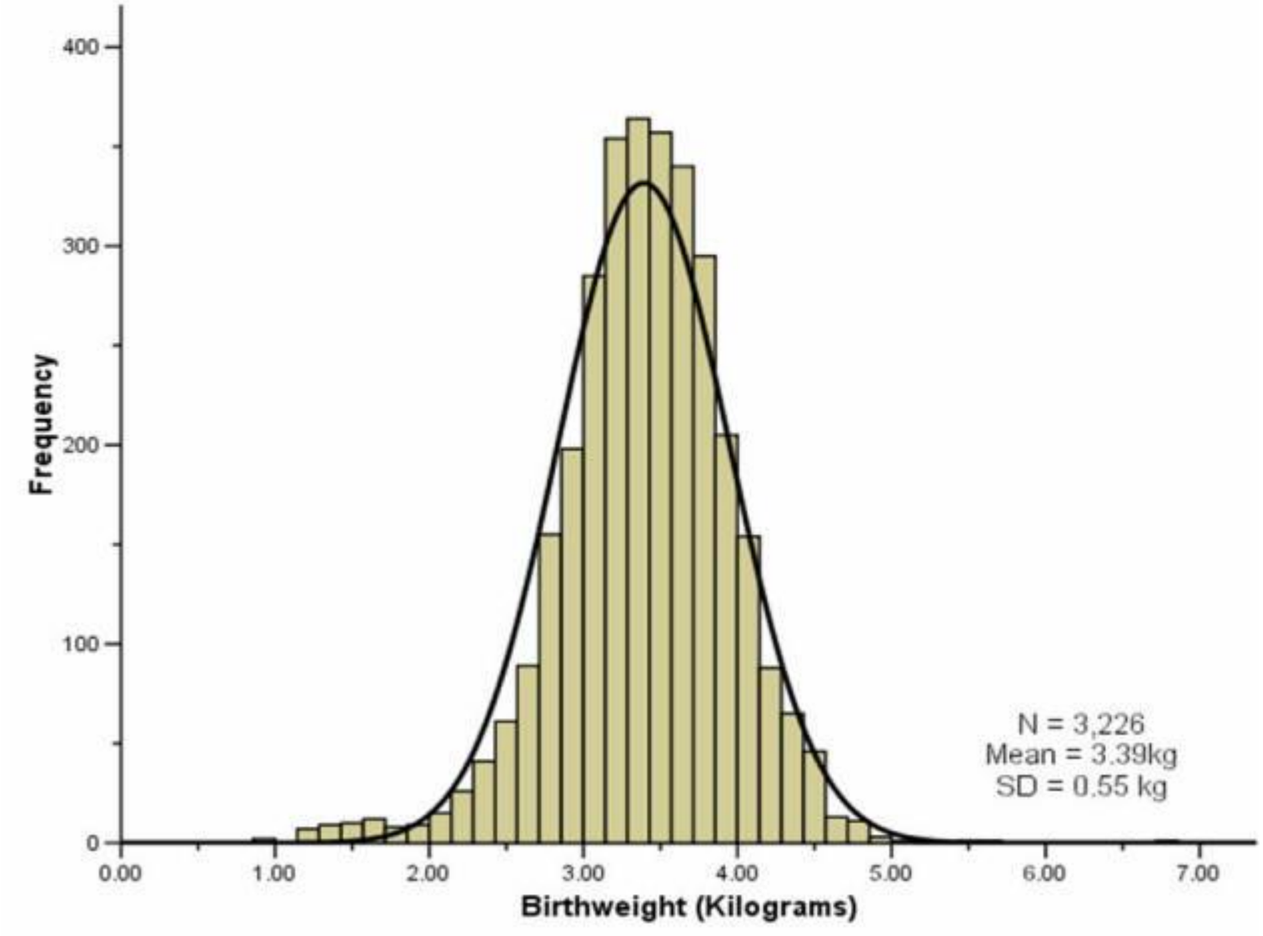
Shown are $n = 1,2,3,10$

"Proof by example" -- uniform



https://www.desmos.com/calculator/2n2m05a9km

"Proof by real-world"



birthweight

A lot of real-world bell-curves can be explained as:

1. The random variable comes from a combination of independent factors.

2. The CLT says the distribution will become like a bell curve.

Theory vs. Practice

The formal theorem statement is "in the limit"

You might not get exactly a normal distribution for any finite n (e.g. if you sum indicators, your random variable is always discrete and will be discontinuous for every finite n.

In practice, the approximations get very accurate very quickly (at least with a few tricks we'll see soon).

They won't be exact (unless the X_i are normals) but it's close enough to use even with relatively small n.

Using the Central Limit Theorem

Suppose you are managing a factory, that produces widgets. Each widget produced is defective (independently) with probability 5%.

Your factory will produce 1000 (possibly defective) widgets. You want to know what the chances are of having a “very bad day” where “very bad” means producing at most 940 non-defective widgets.
(In expectation, you produce 950 non-defective widgets)

What is the probability?

True Answer

Let $X \sim \text{Bin}(1000, .95)$

What is $\mathbb{P}(X \leq 940)$?

The cdf is ugly…and that's a big summation.

$\sum_{k=0}^{940} \binom{1000}{k} (.95)^k \cdot (.05)^{1000-k} \approx .08673$

What does the CLT give?

CLT setup

Bin(1000,.95) is the sum of a bunch of independent random variables (the indicators/Bernoullis we summed to get the binomial)

So, let's use the CLT instead

$\mathbb{E}[X_i] = p = .95.$

$\text{Var}(X_i) = p(1-p) = .0475$

$Y_{1000} = \frac{\sum_{i=1}^{1000} X_i - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}$ is approximately $\mathcal{N}(0,1)$.

With the CLT.

The event we're interested in is $\mathbb{P}(X \leq 940)$

$$\mathbb{P}(X \leq 940)$$

$$= \mathbb{P}\left(\frac{X - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}} \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$$

$$= \mathbb{P}(Y_{1000} \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}})$$

$$\approx \mathbb{P}\left(Y \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right) \textit{ by CLT}$$

$$= \Phi\left(\frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$$

$$\approx \Phi(-1.45) = 1 - \Phi(1.45)$$

$$\approx 1 - .92647 = .07353.$$

It's an approximation!

The true probability is

$$1 - \sum_{k=941}^{1000} \binom{1000}{k} (.95)^k \cdot (.05)^{1000-k} \approx .08673$$

The CLT estimate is off by about 1.3 percentage points.

We can get a better estimate if we fix a subtle issue with this approximation.

A problem

What's the probability that $\mathrm{X} = 950$? (exactly)

True value, we can get with binomial:

$\binom{1000}{950} \cdot (.95)^{950} \cdot (.05)^{50} \approx .05779$

What does the CLT say?

A problem

What's the probability that $\mathrm{X} = 950$? (exactly)

True value, we can get with binomial:

$\binom{1000}{950} \cdot (.95)^{950} \cdot (.05)^{50} \approx .05779$

What does the CLT say?

$= \mathbb{P}\left(\frac{X-1000\cdot.95}{\sqrt{1000\cdot.0475}} = \frac{950-1000\cdot.95}{\sqrt{1000\cdot.0475}}\right)$

$\approx \mathbb{P}(Y = 0)$

$= 0$

Uh oh.

Continuity Correction

The binomial distribution is discrete, but the normal is continuous.

Let's correct for that (called a "continuity correction")

Before we switch from the binomial to the normal, ask "what values of a continuous random variable would round to this event?"

Applying the continuity correction

$\mathbb{P}(X = 950)$

$= \mathbb{P}(949.5 \leq X < 950.5)$

Continuity correction.
This step really is an "exactly equal to"
The discrete rv X can't equal 950.2.

$= \mathbb{P}\left(\frac{949.5-950}{\sqrt{1000\cdot.0475}} \leq \frac{X-950}{\sqrt{1000\cdot.0475}} < \frac{950.5-950}{\sqrt{1000\cdot.0475}}\right)$

$\approx \mathbb{P}\left(\frac{949.5-950}{\sqrt{1000\cdot.0475}} \leq Y < \frac{950.5-950}{\sqrt{1000\cdot.0475}}\right)$ By CLT

$= \Phi\left(\frac{950.5-950}{\sqrt{1000\cdot.0475}}\right) - \Phi\left(\frac{949.5-950}{\sqrt{1000\cdot.0475}}\right)$

$\approx \Phi(0.07) - \Phi(-0.07) = \Phi(0.07) - (1 - \Phi(0.07))$

$\approx 0.5279 - (1 - 0.5279) = 0.0558$

Still an Approximation

$\binom{1000}{950} \cdot (.95)^{950} \cdot (.05)^{50} \approx .05779$ is the true value

The CLT approximates: 0.0558

Very close! But still not perfect.

Let’s fix that other one

Question was “what’s the probability of seeing at most 940 non-defective widgets?”

With the CLT.

The event we're interested in is $\mathbb{P}(X \leq 940)$

$$\mathbb{P}(X \leq 940)$$

$$= \mathbb{P}\left(\frac{X - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}} \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$$

$$\approx \mathbb{P}\left(Y \leq \frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right) \textit{ By CLT}$$

$$= \Phi\left(\frac{940 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$$

$$\approx \Phi(-1.45) = 1 - \Phi(1.45)$$

$$\approx 1 - .92647 = .07353.$$

$$\mathbb{P}(X \leq 940.5)$$

$$= \mathbb{P}\left(\frac{X - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}} \leq \frac{940.5 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$$

$$\approx \mathbb{P}\left(Y \leq \frac{940.5 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right) \textit{ By CLT}$$

$$= \Phi\left(\frac{940.5 - 1000 \cdot .95}{\sqrt{1000 \cdot .0475}}\right)$$

$$\approx \Phi(-1.38) = 1 - \Phi(1.38)$$

$$\approx 1 - .91621 = .08379.$$

True answer: .08673

Approximating a continuous distribution

You buy lightbulbs that burn out according to an exponential distribution with parameter of $\lambda = 1.8$ lightbulbs per year.

You buy a 10 pack of (independent) light bulbs. What is the probability that your 10-pack lasts at least 5 years?

Let X_i be the time it takes for lightbulb i to burn out.

Let X be the total time. Estimate $\mathbb{P}(X \geq 5)$.

Where’s the continuity correction?

There’s no correction to make – it was already continuous!!

$\mathbb{P}(X \geq 5)$

$= \mathbb{P}\left(\frac{X-10/1.8}{\sqrt{10/1.8^2}} \geq \frac{5-10/1.8}{\sqrt{10/1.8^2}}\right)$

$\approx \mathbb{P}\left(Y \geq \frac{5-10/1.8}{\sqrt{10/1.8^2}}\right)$ By CLT

$\approx \mathbb{P}(Y \geq -0.32)$

$= 1 - \Phi(-0.32) = \Phi(0.32)$

$\approx .62552$

True value (needs a distribution not in our zoo) is ≈ 0.58741

Outline of CLT steps

1. Write event you are interested in, in terms of sum of random variables.

2. Apply continuity correction if RVs are discrete.

3. Normalize RV to have mean 0 and standard deviation 1.

4. Replace RV with $\mathcal{N}(0,1)$.

5. Write event in terms of Φ

6. Look up in table.

Polling

Suppose you know that 60% of CSE students support you in your run for SAC. If you draw a sample of 30 students, what is the probability that you don't get a majority of their votes.

How are you sampling?

Method 1: Get a uniformly random subset of size 30.

Method 2: Independently draw 30 people with replacement.

Which do we use?

Polling

Method 1 is what's accurate to what is actually done…

…but we're going to use the math from Method 2.

Why?
Hypergometric variable formulas are rough, and for increasing population size they're very close to binomial.

And we're going to approximate with the CLT anyway, so…the added inaccuracy isn't a dealbreaker.

If we need other calculations, independence will make any of them easier.

Polling

Let X_i be the indicator for "person i in the sample supports you."

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{30}$ is the fraction who support you.

We're interested in the event $\mathbb{P}(\bar{X} \leq .5)$.

What is $\mathbb{E}[\bar{X}]$? What is $\text{Var}(\bar{X})$?

$\mathbb{E}[\bar{X}] = \frac{1}{30}\mathbb{E}[\sum X_i] = \frac{.6 \cdot 30}{30} = \frac{3}{5}$.

$\text{Var}(\bar{X}) = \frac{1}{30^2}\text{Var}(\sum X_i) = \frac{1}{30} \cdot .6 \cdot .4 = \frac{1}{125}$.

Using the CLT

$\mathbb{P}(\bar{X} \leq .5)$

$= \mathbb{P}\left(\frac{\bar{X}-.6}{1/\sqrt{125}} \leq \frac{.5-.6}{1/\sqrt{125}}\right)$

$\approx \mathbb{P}\left(Y \leq \frac{.5-.6}{1/\sqrt{125}}\right)$ where $Y \sim \mathcal{N}(0,1)$

$\approx \mathbb{P}(Y \leq -1.12)$

$= \Phi(-1.12) = 1 - \Phi(1.12) \approx 1 - 0.86864 = 0.13136$

Confidence Intervals

A "confidence interval" tells you the probability (how confident you should be) that your random variable fell in a certain range (interval)

Usually "close to its expected value"

$$\mathbb{P}(|X-\mu| > \varepsilon) \leq \delta$$

If your RV has expectation equal to the value you're searching for (like our polling example) you get a probability of being "close enough" to the target value.

Confidence Intervals

Using the CLT, we estimated the probability of "missing low"

There's a few drawbacks though

1. Using the CLT we get an estimate, not a guarantee---what if the CLT estimate is underestimating the probability of failure?

2. We needed to know the true value to do that computation---if we knew the true value, we wouldn't run the poll!

Some algebra tricks can handle problem 2, but 1 really asks for a new tool; we'll see concentration inequalities next week.

Application: Idealized Polling

This is a *very* detailed example to try to understand confidence intervals better. You may find it helpful to read on your own; we'll discuss more aspects of computations like these when we get to confidence intervals next week.

Polling

Our end goal is to answer the question “how many people do I need to poll to get an accurate sense of how the population is going to vote?”

That’s a weird question (it’ll require “going backwards” in the algebra) so first we’ll “go forwards” (given the poll size how accurate will we be?) to see what’s happening more clearly.

Polling

Suppose you know that 60% of CSE students support you in your run for SAC. If you draw a sample of 30 students, what is the probability that you don't get a majority of their votes.

How are you sampling?

Method 1: Get a uniformly random subset of size 30.

Method 2: Independently draw 30 people with replacement.

Which do we use?

Polling

Method 1 is what's accurate to what is actually done…

…but we're going to use the math from Method 2.

Why?
Hypergometric variable formulas are rough, and for increasing population size they're very close to binomial.

And we're going to approximate with the CLT anyway, so…the added inaccuracy isn't a dealbreaker.

If we need other calculations, independence will make any of them easier.

Polling

Let X_i be the indicator for "person i in the sample supports you."

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{30}$ is the fraction who support you.

We're interested in the event $\mathbb{P}\left(\bar{X} \leq .5\right)$.

What is $\mathbb{E}\left[\bar{X}\right]$? What is $\mathrm{Var}\left(\bar{X}\right)$?

Polling

Let X_i be the indicator for "person i in the sample supports you."

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{30}$ is the fraction who support you.

We're interested in the event $\mathbb{P}(\bar{X} \leq .5)$.

What is $\mathbb{E}[\bar{X}]$? What is $\text{Var}(\bar{X})$?

$\mathbb{E}[\bar{X}] = \frac{1}{30}\mathbb{E}[\sum X_i] = \frac{.6 \cdot 30}{30} = \frac{3}{5}$.

$\text{Var}(\bar{X}) = \frac{1}{30^2}\text{Var}(\sum X_i) = \frac{1}{30} \cdot .6 \cdot .4 = \frac{1}{125}$.

Using the CLT

$\mathbb{P}(\bar{X} \leq .5)$

$= \mathbb{P}\left(\frac{\bar{X}-.6}{1/\sqrt{125}} \leq \frac{.5-.6}{1/\sqrt{125}}\right)$

$\approx \mathbb{P}\left(Y \leq \frac{.5-.6}{1/\sqrt{125}}\right)$ where $Y \sim \mathcal{N}(0,1)$

$\approx \mathbb{P}(Y \leq -1.12)$

$= \Phi(-1.12) = 1 - \Phi(1.12) \approx 1 - 0.86864 = 0.13136$

Hey! Where's the continuity correction?

If this were just a question about $n = 30$, we would have used one. But for preparing for the next calculation it made sense to skip it.

What is $\bar{X}$?

It's the *average* of a bunch of indicators.

So the support is:

$\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \dots, \frac{n-1}{n}, \frac{n}{n}$.

Instead of .5, we'd use $.5 + \frac{1}{2n}$. Which makes the algebra much worse.

And for real polling applications, n is going to be quite big anyway where $\frac{1}{2n}$ is not going to make a substantial difference.

Hey! You didn’t tell us how many students were in CSE!

The accuracy of a poll is dependent on the number of people you sample, not the size of the population.*

Weird right?

This isn’t a trick of the fact that we used the CLT. The same is true if we calculated exactly with a binomial.

*at least for this idealized scenario, where the answer is a simple “yes” or “no” and you can get a uniformly random person. Those things become less likely as populations get bigger.

The Reverse Question

Polls are made by sampling n people from a population. They are then reported with "52% of likely voters would vote in favor of proposal if held today (margin of error +/- 3%)"

You are going to run your own poll. And you want a better "margin of error" – you want 2% how many people do you need to poll?

Let's think about idealized polling – pretend we're really getting a uniformly random person.

Margin of Error

Wait…what's a "margin of error"

The result of the poll is a random variable – it has a distribution.

You'd like to know something about its variance (Did you poll everyone in the entire country? Just 3 people? How much variance is there in the poll?)

A "margin of error" is an intuitive measurement of the variance of the poll. "If I performed this poll repeatedly, 95% of the time, we're within true +/- the margin of error."

Our Goal

Set a target – I want my margin of error to be 2%. That is, at least 95% of the time, your poll's estimate of the fraction of people in favor will be within 2 percentage points of the true value.

So…how many people are you going to need to interview?

Poll Setup

Let X_i be the indicator that the i^{th} person you interview supports the proposal.

Your random variable is $\hat{p}$: $\sum X_i/n$

Let p be the true fraction of people who support the proposal.

What is the

$\mathbb{E}[\hat{p}] =$

$\text{Var}(\hat{p}) =$

Poll Setup

Let X_i be the indicator that the i^{th} person you interview supports the proposal.

Your random variable is $\hat{p}$: $\sum X_i / n$

Let p be the true fraction of people who support the proposal.

What is the

$$\mathbb{E}[\hat{p}] = \frac{1}{n} \cdot \mathbb{E}[\sum X_i] = \frac{pn}{n} = p$$

$$\text{Var}(\hat{p}) = \frac{1}{n^2} \text{Var}(X_i) = \frac{p(1-p)}{n}$$

Using the CLT

What are we looking for? Well we have a margin of error:

$$\mathbb{P}(p - .02 \leq \hat{p} \leq p + .02) \geq .95$$

That says we're within the 2% margin of error at least 95% of the time.

What is that probability? Well let's setup to use the CLT. Subtract the expectation and divide by the standard devation.

$$\mathbb{P}\left(\frac{p-.02-p}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{p+.02-p}{\sqrt{p(1-p)/n}}\right) \geq .95$$

Apply the CLT

$$\mathbb{P}\left(\frac{p-.02-p}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{p+.02-p}{\sqrt{p(1-p)/n}}\right) \geq .95$$

Is well approximated by $\mathbb{P}\left(\frac{-\sqrt{n}\cdot .02}{\sqrt{p(1-p)}} \leq Z \leq \frac{\sqrt{n}\cdot .02}{\sqrt{p(1-p)}}\right) \geq .95$ for $Z \sim \mathcal{N}(0,1)$

So as n changes, the probability changes. So choose the smallest n for which the probability is at least .95

WAIT, what's $\sqrt{p(1-p)}$? We don't know p. That's *why* we're doing the poll in the first place.

Handling $\sqrt{p(1-p)}$

Justification 1: If we make a mistake, we want it to be making n bigger. (since we're trying to say "take n at least this big, and you'll be safe").

The bigger the standard deviation, the bigger n will need to be to control it. So assume the biggest possible standard deviation.

Justification 2:

As $\sqrt{p(1-p)}$ gets bigger, the interval gets smaller (it's in the denominator), so assuming the biggest value of $\sqrt{p(1-p)}$ gives us the most restricted interval. So no matter what the true interval is we have a subset of it. And if our probability is at least $.95$ then the true probability is at least $.95$.

What's the maximum of $\sqrt{p(1-p)}$?

Worst value of p

Calculus time!

Set $\frac{d}{dp}\sqrt{p - p^2} = 0$

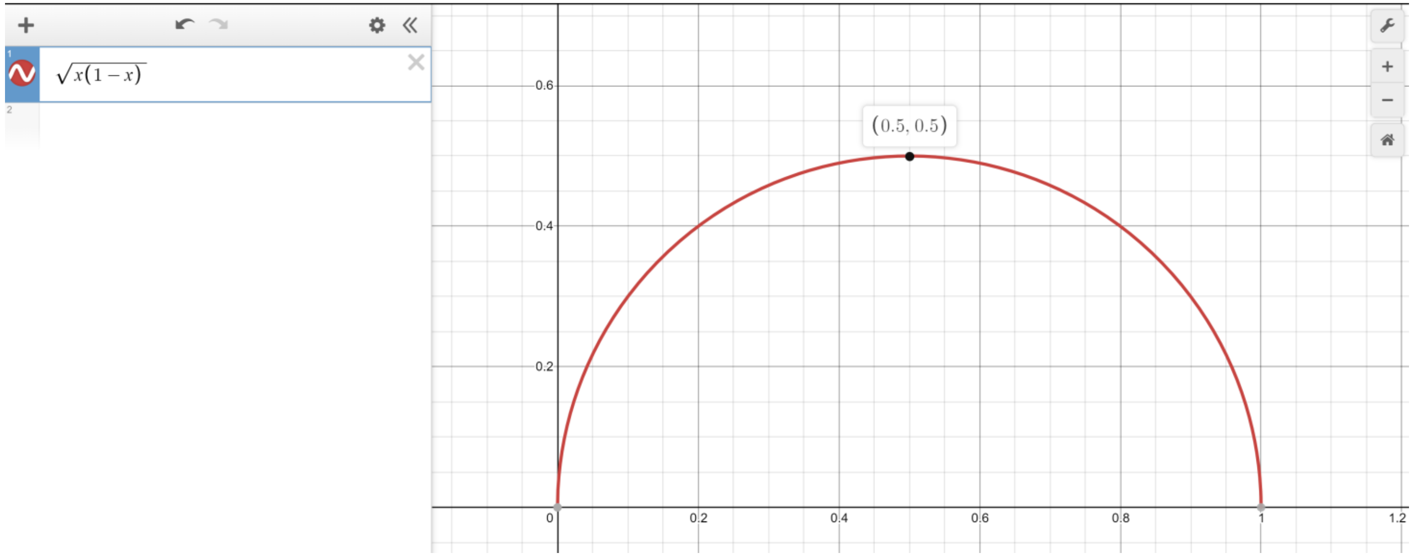
$\frac{1}{\sqrt{p-p^2}}(1 - 2p) = 0$

$1 - 2p = 0 \rightarrow p = 1/2$

Second derivative test will confirm $p = \frac{1}{2}$ is a maximizer

Or just plot it.

$\sqrt{\frac{1}{2}\left(1 - \frac{1}{2}\right)} = \sqrt{1/4}.$



Doing the algebra

$$\mathbb{P}\left(\frac{p-.02-p}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{p+.02-p}{\sqrt{p(1-p)/n}}\right)$$

$$\approx \mathbb{P}\left(\frac{-\sqrt{n}\cdot.02}{\sqrt{p(1-p)}} \leq Z \leq \frac{\sqrt{n}\cdot.02}{\sqrt{p(1-p)}}\right) \text{ by CLT; } Z\sim\mathcal{N}(0,1)$$

$$\geq \mathbb{P}\left(\frac{-\sqrt{n}\cdot.02}{\sqrt{1/4}} \leq Z \leq \frac{\sqrt{n}\cdot.02}{\sqrt{1/4}}\right)$$

$$= \mathbb{P}\left(-.04\sqrt{n} \leq Z \leq .04\sqrt{n}\right)$$

$$= \Phi\left(.04\sqrt{n}\right) - \left(1 - \Phi\left(.04\sqrt{n}\right)\right) = 2\Phi\left(.04\sqrt{n}\right) - 1$$

$$2\Phi\left(.04\sqrt{n}\right) - 1 \geq .95 \rightarrow \Phi\left(.04\sqrt{n}\right) \geq \frac{1.95}{2}$$

Using the Φ-table

$\Phi(.04\sqrt{n}) \geq .975$

Φ-table says:

$.04\sqrt{n} \geq 1.96$

$\sqrt{n} \geq 49$

$n \geq 2401$. gives 95% confidence interval of +/- 2%.

I.e. 95% of the time, our poll gets a value within 2% of the true value.

CLT Wrap-up

It's not ideal that we had an approximation symbol in the middle (that "≥" isn't really a guarantee at this point, it's an approximation)
Observation 1: with our current tools, we wouldn't get an answer in a reasonable amount of time.

But using a binomial would be even harder.

As n changes, the distribution of a binomial changes. Wolfram alpha isn't even enough here (unless you have 2+ hours to spare to guess and check values). You need a computer program to get the exact value.

You're computer scientists! You can write that program. But it takes time.

Observation 2: if you need an absolute guarantee, you won't get one. The tool you want is a "concentration inequality/tail bound." We'll see those next week.

CLT Wrap-up

Use the CLT when:

1. The random variable you're interested in is the sum of independent random variables.

2. The random variable you're interested in does not have an easily accessible or easy to use pmf/pdf (or the question you're asking doesn't lend it self to easily using the pmf/pdf)

3. You only need an approximate answer, and the sum is of at least a moderate number of random variables.