Please download the activity slide! ☺

Application: Polling

CSE 312 Summer 21
Lecture 17

Announcements

Review Summary 2 due tonight!

HW 6 has been released!

One programming question and an analysis of the code.

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.

Replace $\frac{X-\mu}{\sigma}$ with Y or Z

3. Normalize RV to have mean $\mathbf{0}$ and standard deviation $\mathbf{1}$.
4. Replace RV with $\mathcal{N}(\mathbf{0},\mathbf{1})$.

$Y \sim \mathcal{N}(0,1)$ or $Z \sim \mathcal{N}(0,1)$

5. Write event in terms of $\mathbf{\Phi}$
6. Look up in table.

Application: Idealized Polling

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 UW students support you in your run for ASUW. 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. Hypergeometric

Method 2: Independently draw 30 people with replacement. Binomial

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?
Hypergeometric 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

\# people who did not support you.

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

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

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

Fill out the poll everywhere so Kushal knows how long to explain
Go to pollev.com/cse312su21

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 $\text{Var}(\bar{X})$?

$E[\bar{X}] = E\left[\frac{\sum X_i}{30}\right] = \frac{1}{30} \sum E[X_i]$

$=$

$$\mathbb{E}[\bar{X}] = \frac{1}{30}\mathbb{E}\left[\sum X_i\right] = \frac{0.6 \cdot 30}{30} = \frac{3}{5}.$$

$$\text{Var}(\bar{X}) = \frac{1}{30^2}\text{Var}\left(\sum X_i\right) = \frac{1}{30} \cdot 0.6 \cdot 0.4 = \frac{1}{125}$$

$\frac{p(1-p)}{n}$

std dev $(\bar{X}) = \frac{1}{\sqrt{125}}$

Using the CLT

$$\mathbb{P}(\bar{X} \leq 0.5)$$

$$= \mathbb{P}\left(\frac{\bar{X}-0.6}{1/\sqrt{125}} \leq \frac{0.5-0.6}{1/\sqrt{125}}\right)$$

$$\approx \mathbb{P}\left(Y \leq \frac{0.5-0.6}{1/\sqrt{125}}\right) \text{ where } Y \sim \mathcal{N}(0,1)$$

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

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

approx

13.136%

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 0.5, we'd use $0.5 + \frac{1}{2n}$. 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~~! UW

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.

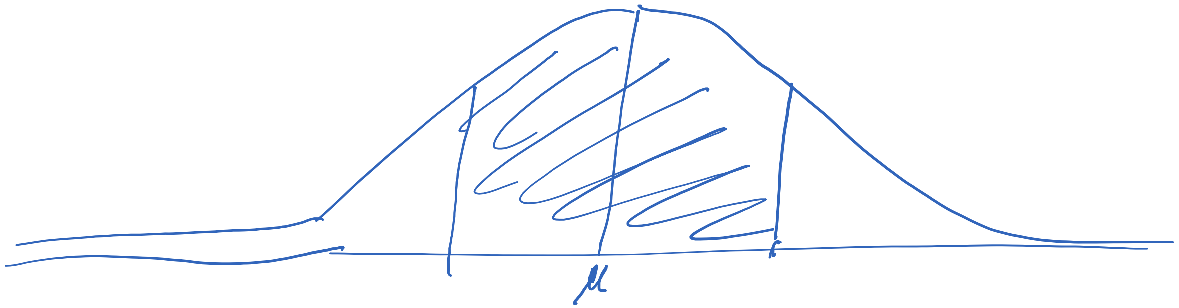
The Reverse Question

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 $\hat{p}$: $\frac{\sum X_i}{n}$

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

What is the

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

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

$\bar{X} = \frac{\sum X_i}{n}$

Poll Setup

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

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

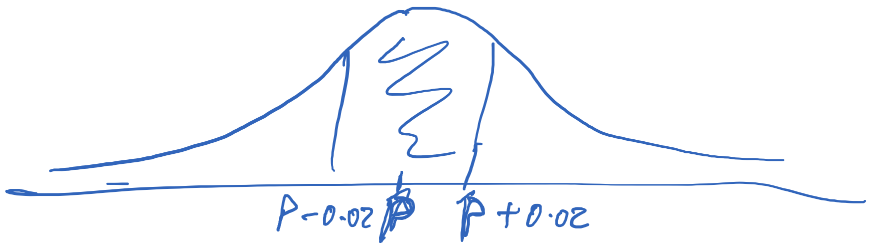
Let p be the true fraction of people who support the proposal.

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:

$$\mathbb{P}(p - 0.02 \leq \hat{p} \leq p + 0.02) \geq 0.95$$

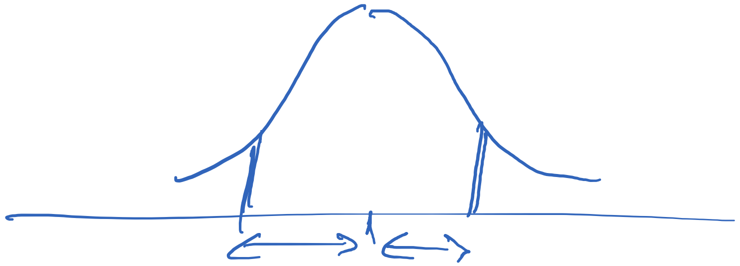
95% confidence interval

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 deviation.

$$\mathbb{P}\left(\frac{p-.02-p}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{p+.02-p}{\sqrt{p(1-p)/n}}\right) \geq 0.95$$

Apply the CLT



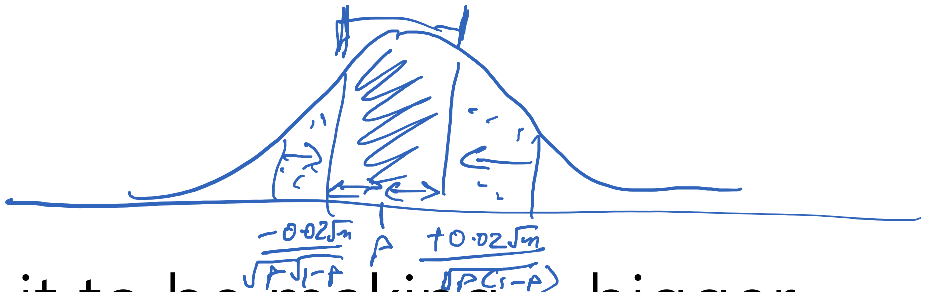
$$\mathbb{P}\left(\frac{p-0.02-p}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{p+0.02-p}{\sqrt{p(1-p)/n}}\right) \geq 0.95$$

Is well approximated by $\mathbb{P}\left(\frac{-\sqrt{n}\cdot 0.02}{\sqrt{p(1-p)}} \leq Z \leq \frac{\sqrt{n}\cdot 0.02}{\sqrt{p(1-p)}}\right) \geq 0.95$ 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 0.95

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 0.95 then the true probability is at least 0.95.

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

Worst value of p

Calculus time!

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

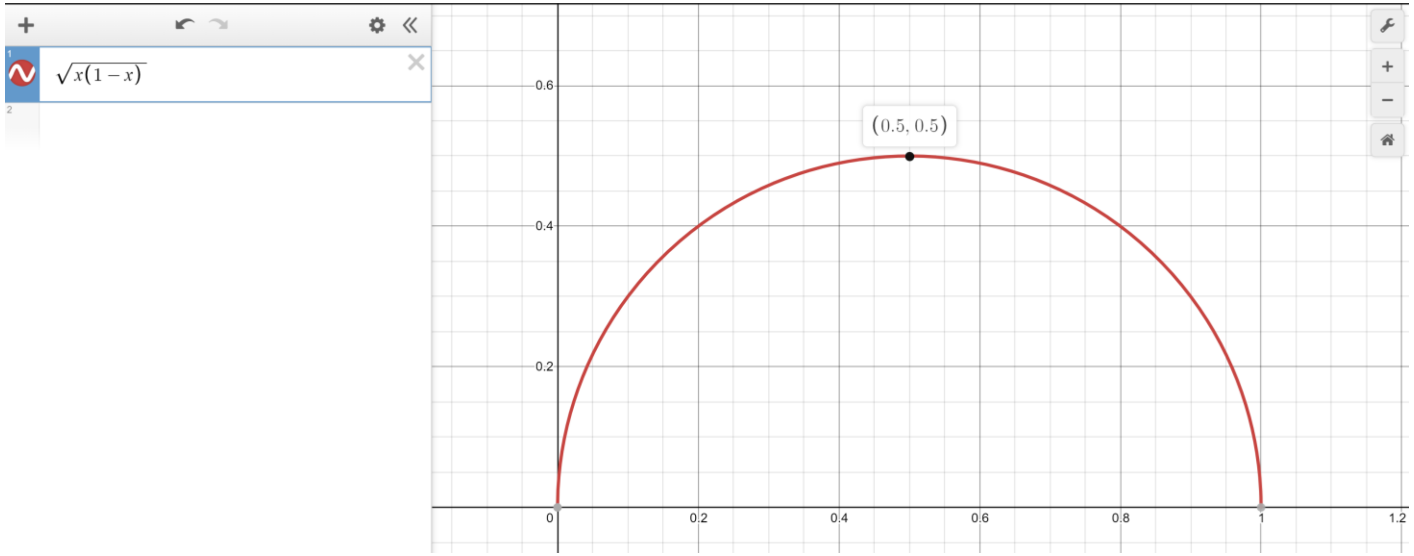
$\frac{1}{\sqrt{p-p^2}}(1 - 2p) = 0$

$1 - 2p = 0 \rightarrow p = \frac{1}{2}$

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

Or just plot it.

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



Doing the algebra

$$\mathbb{P}\left(\frac{p-0.02-p}{\sqrt{p(1-p)/n}} \le \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \le \frac{p+0.02-p}{\sqrt{p(1-p)/n}}\right)$$

$$\approx \mathbb{P}\left(\frac{-\sqrt{n}\cdot 0.02}{\sqrt{p(1-p)}} \le Z \le \frac{\sqrt{n}\cdot 0.02}{\sqrt{p(1-p)}}\right) \text{ by CLT; } Z \sim \mathcal{N}(0,1)$$

$$\ge \mathbb{P}\left(\frac{-\sqrt{n}\cdot 0.02}{\sqrt{1/4}} \le Z \le \frac{\sqrt{n}\cdot 0.02}{\sqrt{1/4}}\right)$$

$$= \mathbb{P}\left(-0.04\sqrt{n} \le Z \le 0.04\sqrt{n}\right)$$

$$= \Phi\left(0.04\sqrt{n}\right) - \left(1 - \Phi\left(0.04\sqrt{n}\right)\right) = 2\Phi\left(0.04\sqrt{n}\right) - 1$$

$$2\Phi\left(0.04\sqrt{n}\right) - 1 \ge 0.95 \rightarrow \Phi\left(0.04\sqrt{n}\right) \ge \frac{1.95}{2}$$

Using the Φ-table

$\Phi(0.04\sqrt{n}) \geq 0.975$

$0.04\sqrt{n} \geq \Phi^{-1}(0.975)$

Φ-table says:

= $0.04\sqrt{n} \geq 1.96$

$\sqrt{n} \geq 49$

$n \geq 2401$ 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 "$\geq$" 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 itself 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.

relative