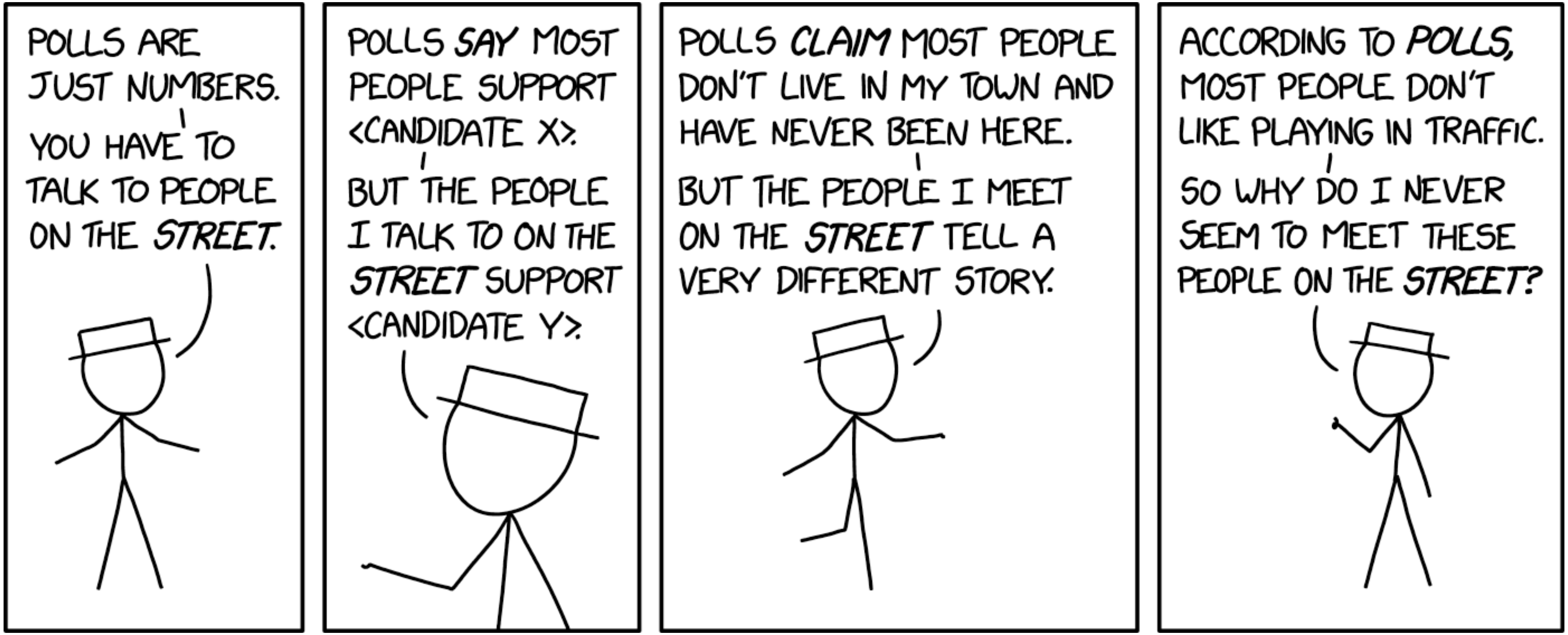


xkcd.com/2357

Application: Polling

CSE 312 23Su
Lecture 15

Goal for Today

43-49%

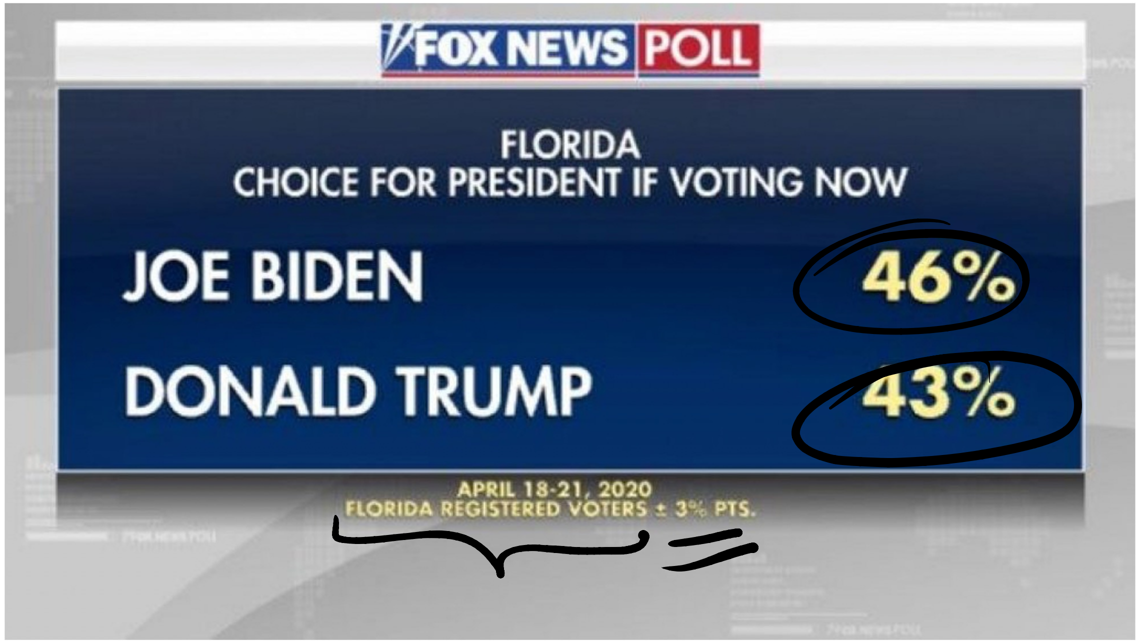
40-46%

In the real world, we often try to predict the outcomes of elections.

75-85%

You may often see statements like: "Candidate A had 80% of the votes, with a margin of error of 5%, 19 times out of 20".

95%



We want to know **where** these numbers are coming from and what people are really trying to convince of us.

Our Poll

Inferential stats

We want to conduct a poll of CSE students, current and former, to determine whether they prefer cats or dogs.

It's not feasible for us to ask all those people. The **population** of interest is quite large.

It **is** feasible for us to poll a random **sample** (subset of the population).

Our question is: how many people should we poll to get an accurate idea of the population's preferences?

Sampling

There are N people in our population of CSE students, past and current.

We want a sample of size n to poll. This sample is uniformly randomly drawn.

0.05

The proportion of the population that prefers cats is p.

We don't know p – but we want to estimate its value, using our sample.

Polling Procedure

For each person i in our sample (1, …, n)

Pick a uniformly random person to call (probability $\frac{1}{N}$ of being chosen)

Ask them how they will vote.

categorical

$$X_i = \begin{cases} 1 & if\ person\ i\ prefers\ cats \\ 0 & otherwise \end{cases}$$

The proportion of people in our sample who love cats is our estimate for p.

$$\frac{|\text{prefer cats}|}{|\text{entire polled sample}|} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Polling Procedure

For each person i in our sample (1, …, n)

Pick a uniformly random person to call (probability $\frac{1}{N}$ of being chosen)

Ask them how they will vote.

$$X_i = \begin{cases} 1 & if\ person\ i\ prefers\ cats \\ 0 & otherwise \end{cases}$$

The proportion of people in our sample who love cats is our estimate for p.

$\bar{X}$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Margin of Error

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

How many people n do we sample so that $\bar{X}$ (our estimate of p) is within 5% of true p at least 98% of the time?

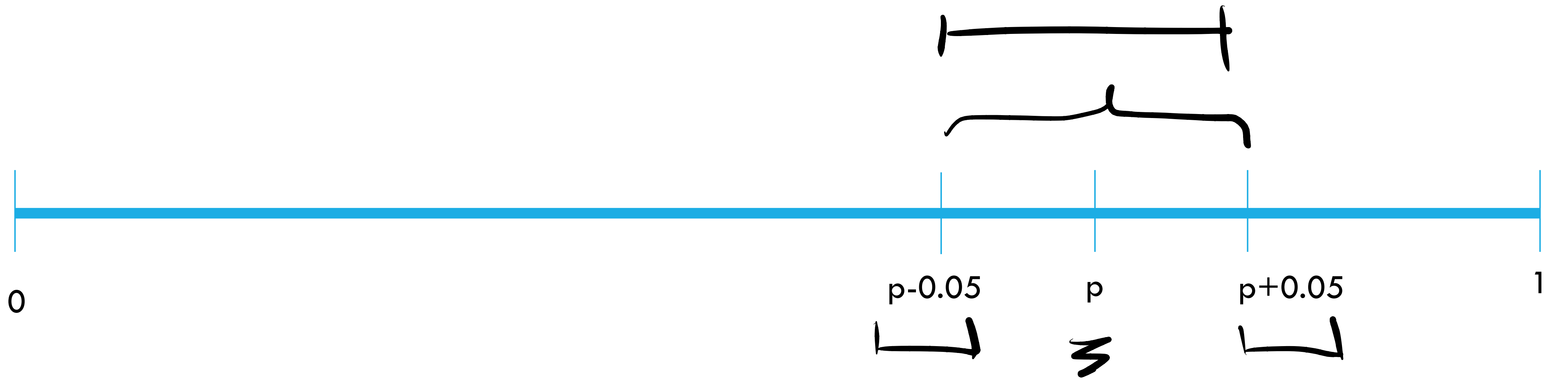
5% is our desired **margin of error**. Notice that $\bar{X}$ is a random variable! That means it has a probability distribution. It is not always going to be the same value. It has some variance.

The margin of error is an intuitive measurement of a poll's variance.

Margin of Error

How many people n do we sample so that $\bar{X}$ (our estimate of p) is within 5% of true p at least 98% of the time?

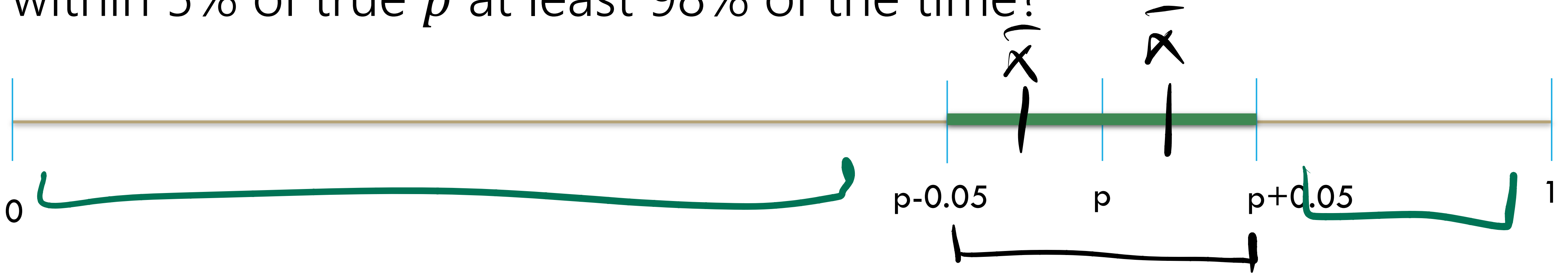
$\bar{x}$ $[p-0.05, p+0.05]$ 98% confidence interval.



Margin of Error

$$\mathbb{P}(|\bar{X} - p| > 0.05) \le 0.02$$

How many people n do we sample so that $\bar{X}$ (our estimate of p) is within 5% of true p at least 98% of the time?



What is the 'desirable event' in this question?

$$\mathbb{P}(p - 0.05 \le \bar{X} \le p + 0.05) \ge 0.98$$

Margin of Error

How many people n do we sample so that $\bar{X}$ (our estimate of p) is within 5% of true p at least 98% of the time?

What is the 'desirable event' in this question?

$$\mathbb{P}(p - 0.05 \le \bar{X} \le p + 0.05) \ge 0.98$$

For what values of n does the above equation hold?

Using the CLT

Can we approximate $\bar{X}$ as a normal distribution?

Distribution of Sample Mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Let $X_1, X_2, \dots, X_n$ be i.i.d with mean μ and variance σ^2. As $n \to \infty$:

$Y = \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$ [CLT, as $n \to \infty$]

$\bar{X} = \frac{1}{n} Y$

$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ [Scaling a normal]

Using the CLT

bootstrap

Can we approximate $\bar{X}$ as a normal distribution?

$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ where μ is the expected value of each Bernoulli, and σ^2 is the variance of each.

$$\underbrace{x_i = \begin{cases} 1 & \text{cats} \\ 0 & \text{dont} \end{cases}}$$

$$\mathbb{E}[X_i] = \mu$$

Using the CLT

$X_i = \begin{cases} 1 & \text{cats} \\ 0 & \text{dogs} \end{cases}$

Can we approximate $\bar{X}$ as a normal distribution?

$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ where μ is the expected value of each Bernoulli, and σ^2 is the variance of each.

$X_i \sim Ber(p) : \mathbb{E}[X] = p, Var(X) = p(1-p)$

$\bar{X} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$

Distribution of $\bar{X}$

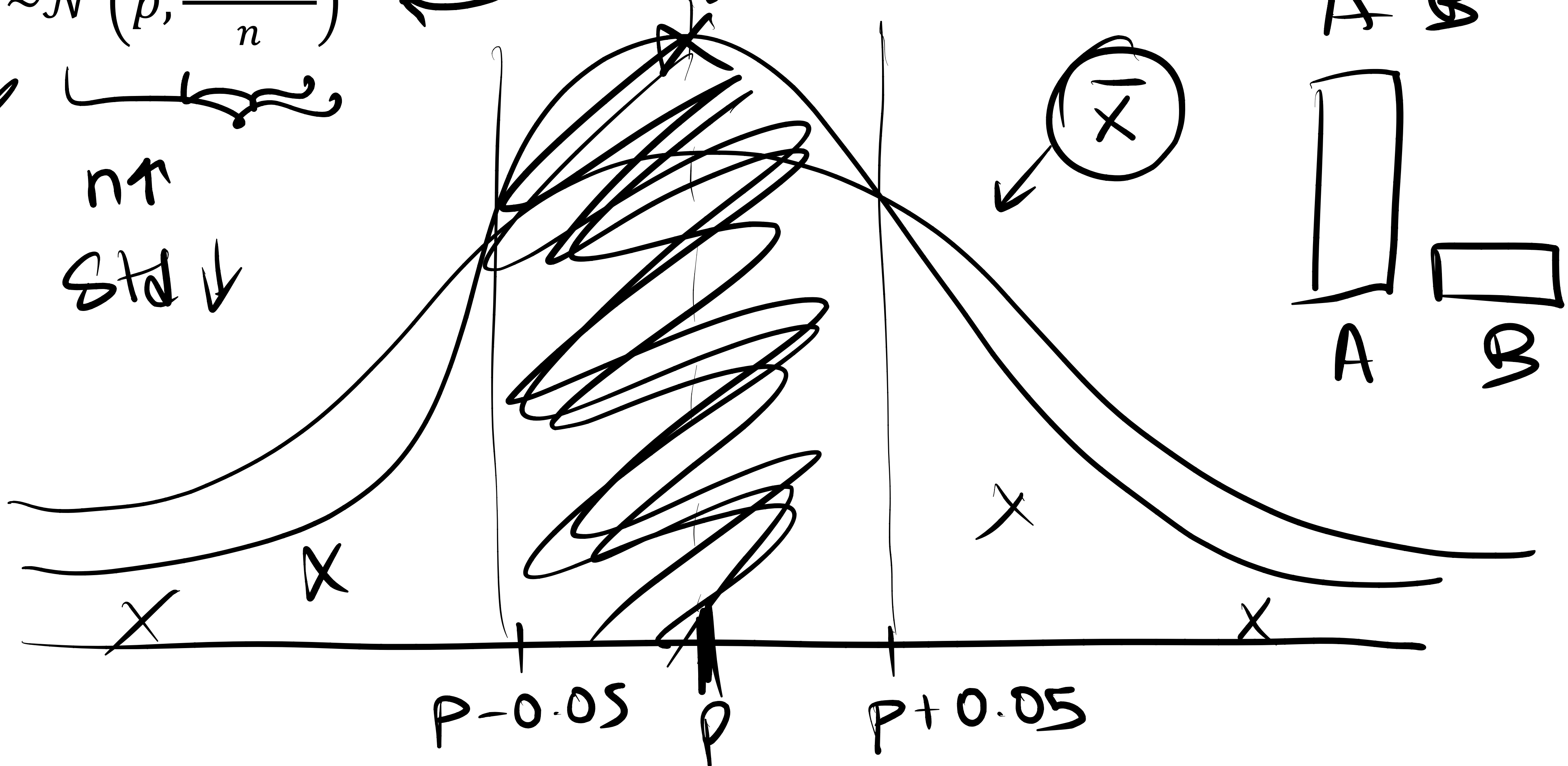
$$\bar{X} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

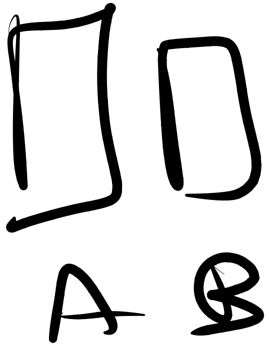
$n \to \infty$

$n \uparrow$

Std $\downarrow$

0 1





Using the CLT

- $\mathbb{P}(p - 0.05 \leq \bar{X} \leq p + 0.05)$

$$\bar{X} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

-

Using the CLT

$$\mathbb{P}(p - 0.05 \leq \bar{X} \leq p + 0.05)$$

$$= \mathbb{P}\left(\frac{p-0.05-p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{(p+0.05-p)}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

Using the CLT

$$\mathbb{P}(p - 0.05 \leq \bar{X} \leq p + 0.05)$$

$$= \mathbb{P}\left(\frac{\cancel{p} - 0.05 - \cancel{p}}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{(\cancel{p} + 0.05 - \cancel{p})}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$= \mathbb{P}\left(\frac{-0.05}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{0.05}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

Using the CLT

$$\mathbb{P}(p - 0.05 \leq \bar{X} \leq p + 0.05)$$

$$= \mathbb{P}\left(\frac{p-0.05-p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{(p+0.05-p)}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$= \mathbb{P}\left(\frac{-0.05}{\sqrt{\frac{p(1-p)}{n}}} \leq \underline{\frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}}} \leq \frac{0.05}{\sqrt{\frac{p(1-p)}{n}}}\right) \star$$

$$\approx \mathbb{P}\left(\underline{\frac{-0.05}{\sqrt{\frac{p(1-p)}{n}}}} \leq \underset{Z}{Z} \leq \underline{\frac{0.05}{\sqrt{\frac{p(1-p)}{n}}}}\right) \text{(CLT)} \star$$

$Z \sim N(0,1)$

Using the CLT

$$\mathbb{P}(p - 0.05 \le \bar{X} \le p + 0.05)$$

$$= \mathbb{P}\left(\frac{p-0.05-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{(p+0.05-p)}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$= \mathbb{P}\left(\frac{-0.05}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{0.05}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$\approx \mathbb{P}\left(\frac{-0.05}{\sqrt{\frac{p(1-p)}{n}}} \le Z \le \frac{0.05}{\sqrt{\frac{p(1-p)}{n}}}\right) \text{(CLT)}$$

$$= \mathbb{P}\left(\frac{-0.05\sqrt{n}}{\sqrt{p(1-p)}} \le Z \le \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

Apply the CLT

$\mathbb{P}(p - 0.05 \le \bar{X} \le p + 0.05) \ge 0.98$

Is well approximated by $\mathbb{P}\left(\frac{-0.05\sqrt{n}}{\sqrt{p(1-p)}} \le Z \le \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) \ge 0.98$ 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 0.98. Easy!

Apply the CLT

$\mathbb{P}(p - 0.05 \leq \bar{X} \leq p + 0.05) \geq 0.98$

Is well approximated by $\mathbb{P}\left(\frac{-0.05\sqrt{n}}{\sqrt{p(1-p)}} \leq Z \leq \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) \geq 0.98$ 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 0.98. Easy!

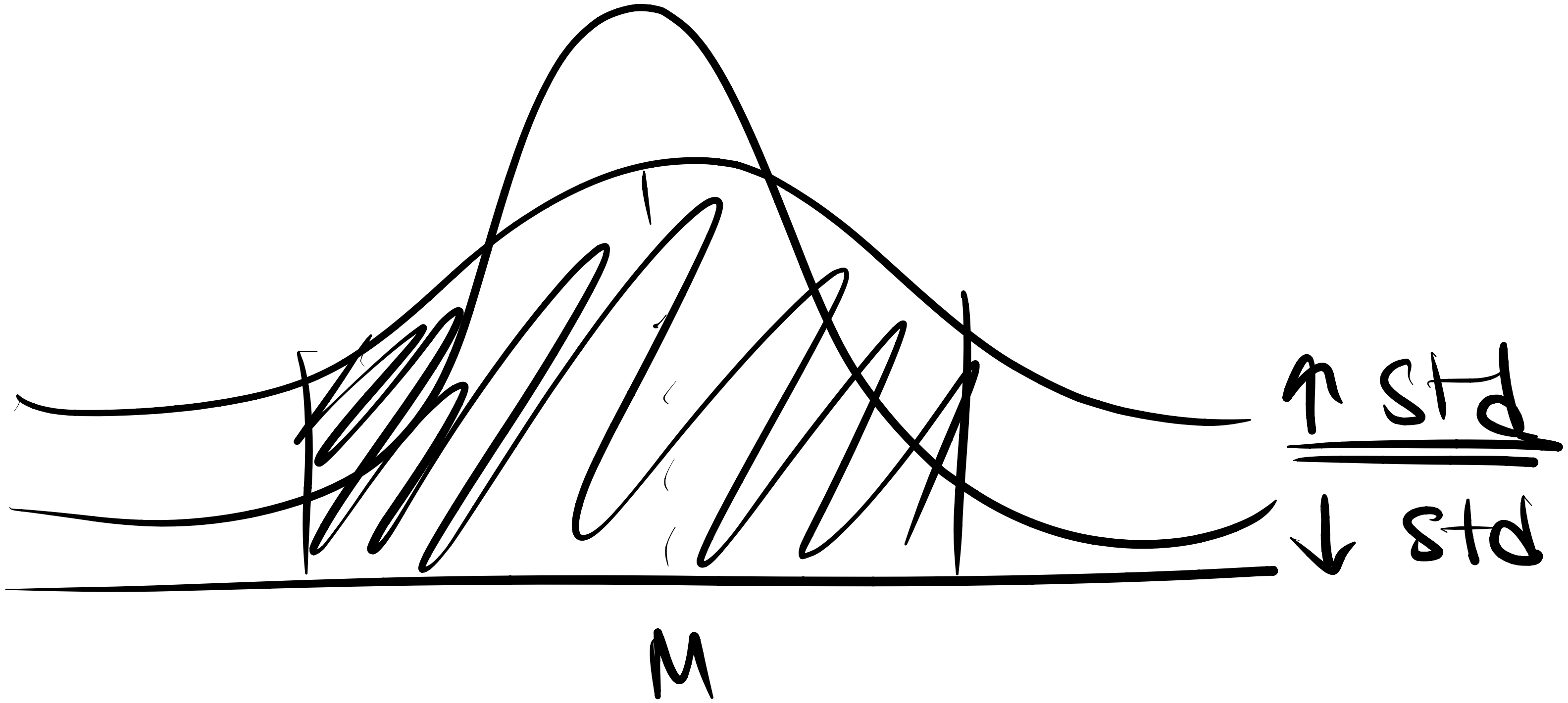
WAIT, what's $\sqrt{p(1-p)}$? We don't know p. That's *why* we're doing the poll in the first place.

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

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.

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



Worst value of p

Calculus time!

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

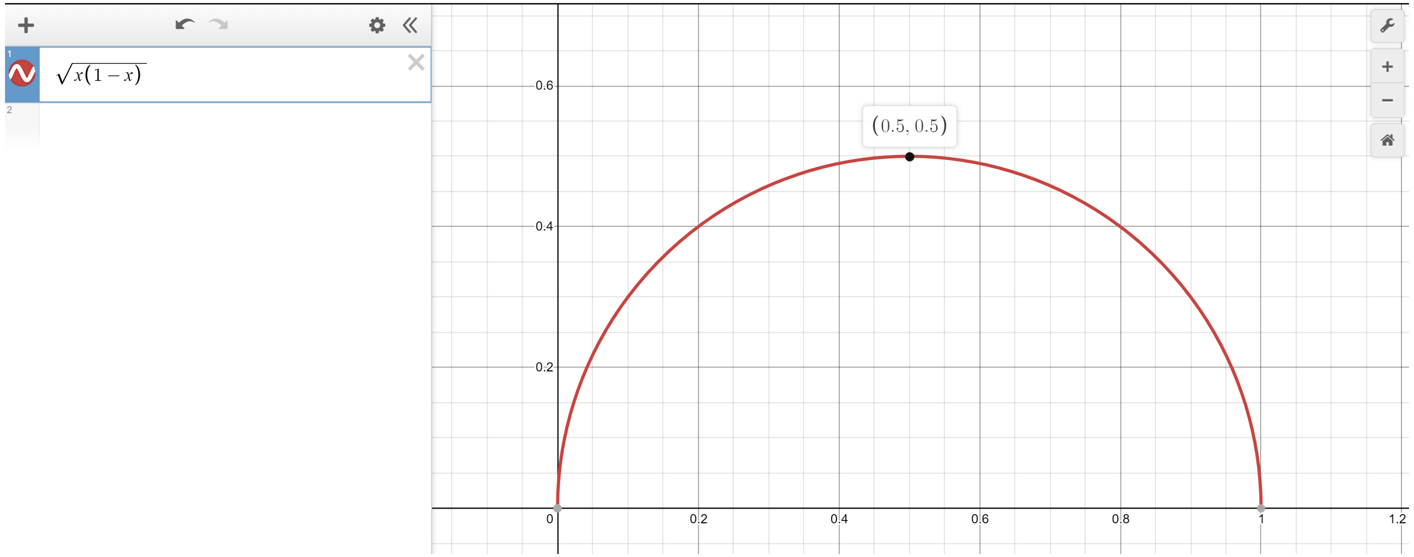
$\frac{1}{\sqrt{p-p^2}}(1 - 2p) = 0$

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

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

Or just plot it.

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

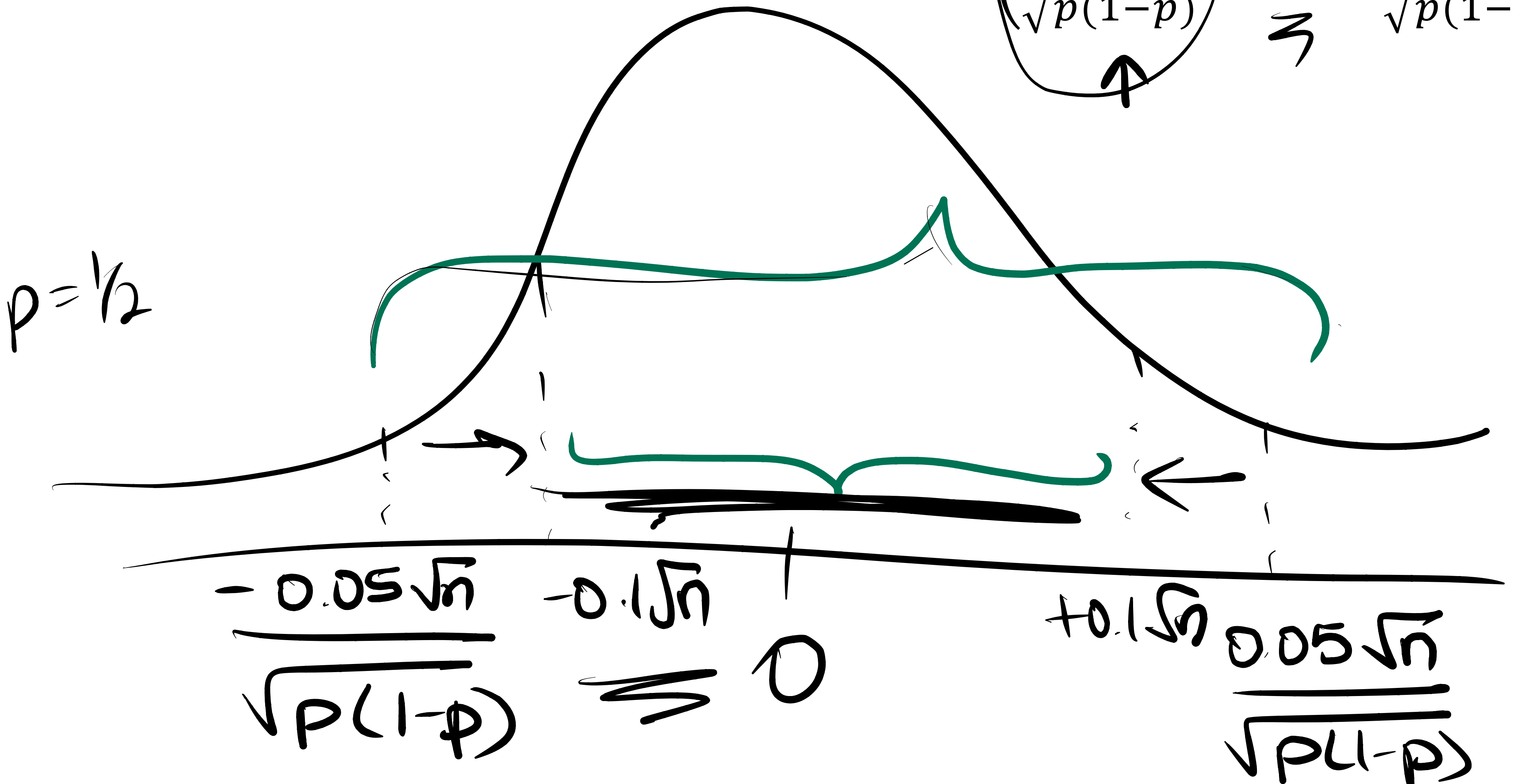


$p(1-p)$

$p = 1/2$

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

We want to minimize the interval in $\mathbb{P}\left(\frac{-0.05\sqrt{n}}{\sqrt{p(1-p)}} \le Z \le \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right)$.



Using the CLT

$$\star = \mathbb{P}\left(\frac{-0.05\sqrt{n}}{\sqrt{p(1-p)}} \le Z \le \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) \text{ (from previous slides)}$$

$p(1-p)$ max

$\dfrac{0.05\sqrt{n}}{\sqrt{p(1-p)}}$ min

$$= \mathbb{P}\left(\frac{-0.05\sqrt{n}}{\sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)}} \le Z \le \frac{0.05\sqrt{n}}{\sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)}}\right)$$

$p = 1/2$ maximizes.

$$\ge \mathbb{P}(-0.1\sqrt{n} \le Z \le 0.1\sqrt{n}) \;\star$$

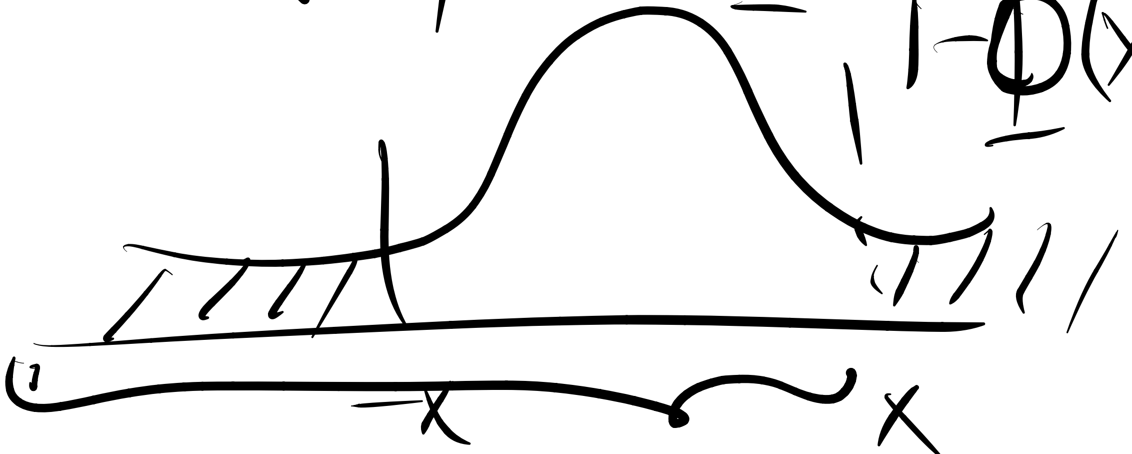
$\} \; -\dfrac{0.05\sqrt{n}}{\sqrt{1/4}}$

$$= \Phi(0.1\sqrt{n}) - \Phi(-0.1\sqrt{n})$$

$$= \Phi(0.1\sqrt{n}) - (1 - \Phi(0.1\sqrt{n}))$$

$$= 2\Phi(0.1\sqrt{n}) - 1$$

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



$$\frac{0.05\sqrt{n}}{p(1-p)} \uparrow \quad \downarrow \quad p(1-p) \text{ max}$$

Finishing it off!

$$2\Phi(0.1\sqrt{n}) - 1 \geq 0.98$$

$$\Phi(0.1\sqrt{n}) \geq \underline{0.99}$$

$$\frac{0.98+1}{2}$$

$$0.1\sqrt{n} \geq \Phi^{-1}(0.99)$$

$$0.1\sqrt{n} \geq 2.33$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999

Finishing it off!

$$2\Phi(0.1\sqrt{n}) - 1 \geq 0.98$$

$$\Phi(0.1\sqrt{n}) \geq 0.99$$

$$0.1\sqrt{n} \geq 2.33$$

$$\sqrt{n} \geq \frac{2.33}{0.1}$$

$$\boxed{n \geq \left(\frac{2.33}{0.1}\right)^2}$$

$n \geq 542.89$ => we should poll 543 people!

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999

So, I Lied to You...

(A few times)

Sampling Details

How are you sampling?

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

Method 2: Independently draw n people with replacement.

Which do we use?

N

$\downarrow$

n

PN ? success!

Sampling Details

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

…but we used 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.

$\Omega_{\bar{X}}$

$\frac{1}{n}\sum_{i=1}^{n} X_i$

$\frac{1}{2}, \frac{1}{n}$

$\left\{ \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} \right\}$

$$\mathbb{P}(p - 0.05 \le \bar{X} \le p + 0.05)$$

$$\mathbb{P}\left(p - 0.05 - \frac{1}{2n} \le \bar{X} \le p + 0.05 + \frac{1}{2n}\right)$$

Hey! Where's the continuity correction?

What is $\bar{X}$?

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

So the support is:

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

Hey! Where's the continuity correction?

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 $\mathbb{P}(p - 0.05 \leq \bar{X} \leq p + 0.05)$, we would look at $\mathbb{P}\left(p - 0.05 - \frac{1}{2n} \leq \bar{X} \leq p + 0.05 + \frac{1}{2n}\right)$

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.

$n \to \infty$

*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.

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.

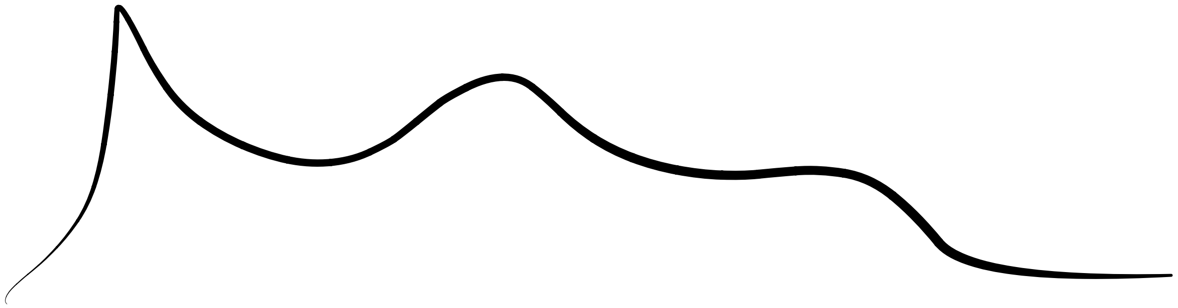
But using a binomial would be even harder.

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

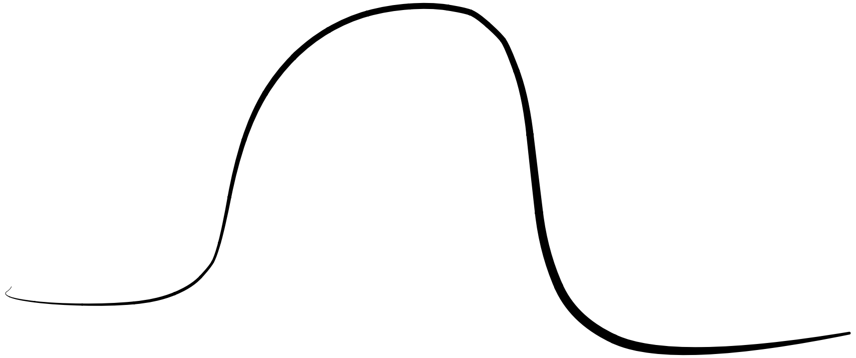


Use the CLT when:

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.



Real-World Polling

Inferential Statistics

Inferential statistics is a field where we derive statistics based on **samples** in order to draw conclusions about a **population**.

Consider these two statements:

1. 56% of those surveyed love cats.
2. 56% of CSE 312 alumni love cats.

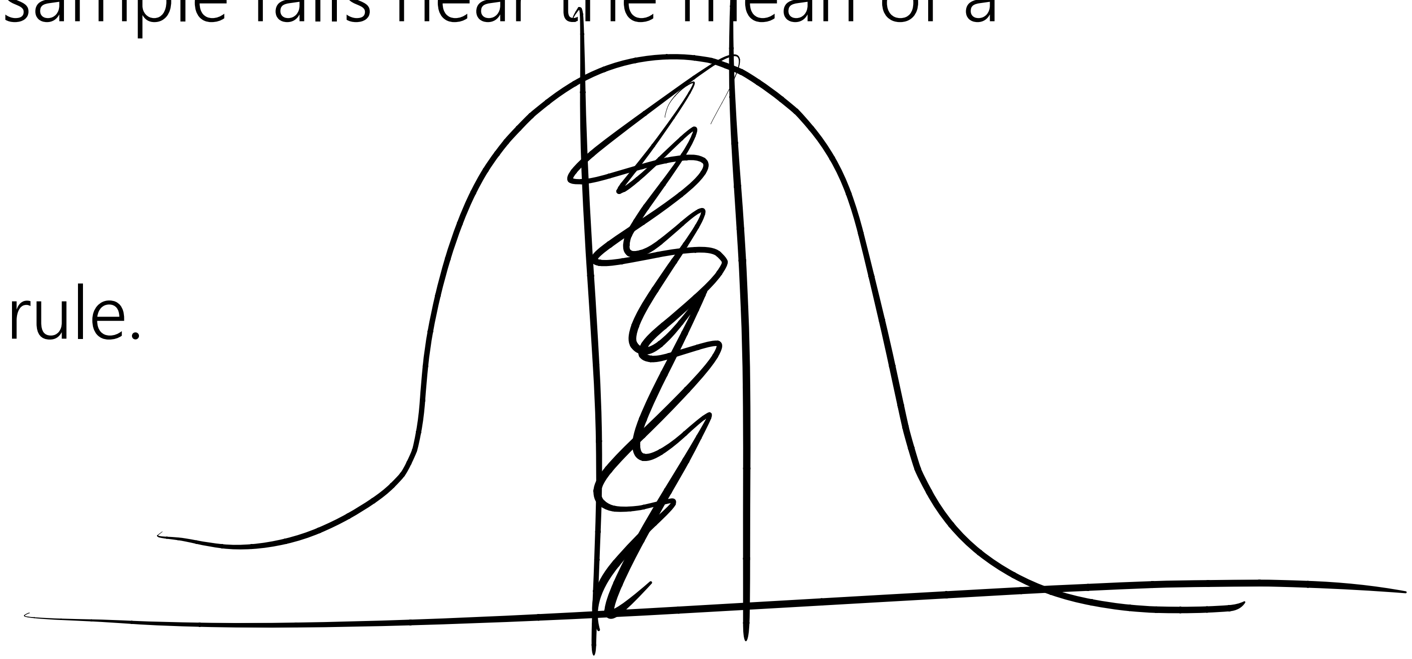
The second is an **inference** we make about our population, based on our sample.

The CLT

We can make that inferential leap because of the CLT: if I keep taking random samples from a population, the sample mean will conform to a normal distribution.

And the mean from a random sample falls near the mean of a population.

We even know the 68-95-99.7 rule.



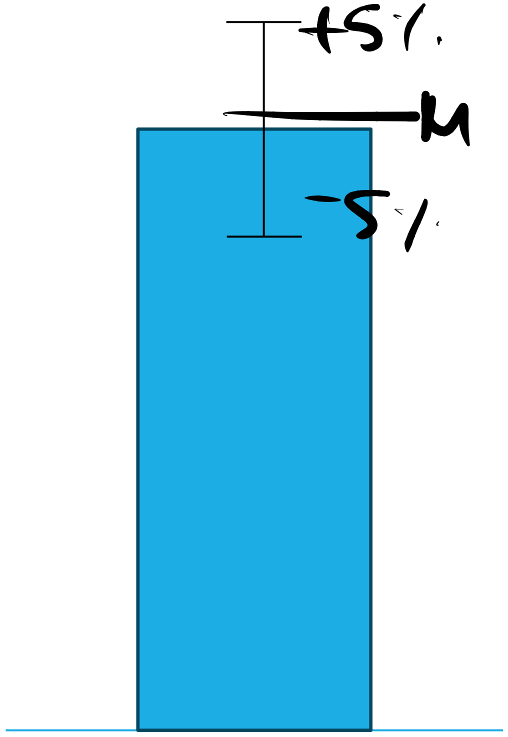
What, really, is the margin of error? ±5%

The **standard error** describes how accurate our estimate based on the sample is.

$Std(\bar{X})$

This is calculated using the formula: $\sqrt{\frac{\sigma^2}{n}}$.

$\frac{\sigma^2}{n}$



Ideally, we want to minimize this error.

↓ σ^2

- Minimize standard deviation (noise).

↑ n

- Get larger samples. ↓ 2x ↑ 4x samples

±3

We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results.

By NATE COHN SEPT. 20, 2016

How four pollsters, and The Upshot, interpreted 867 poll responses:

https://www.nytimes.com/interactive/2016/09/20/upshot/the-error-the-polling-world-rarely-talks-about.html?mtrref=bookdown.org&gwh=C0A5B2E02F2251AFB77C74B0943DFE1A&gwt

What is a representative sample?

If I'm interested in trying to predict the outcome of a U.S. presidential election, who should make up my sample?

U.S. citizens?

What is a representative sample?

If I'm interested in trying to predict the outcome of a U.S. presidential election, who should make up my sample?

U.S. citizens?

Registered voters?

How do we ensure the demographics of registered voters in the population are reflected in a sample's <u>responses</u>? (race, age, education, region…)

What is a representative sample?

If I’m interested in trying to predict the outcome of a U.S. presidential election, who should make up my sample?

U.S. citizens?

Registered voters?

Are registered voters the same as likely voters?

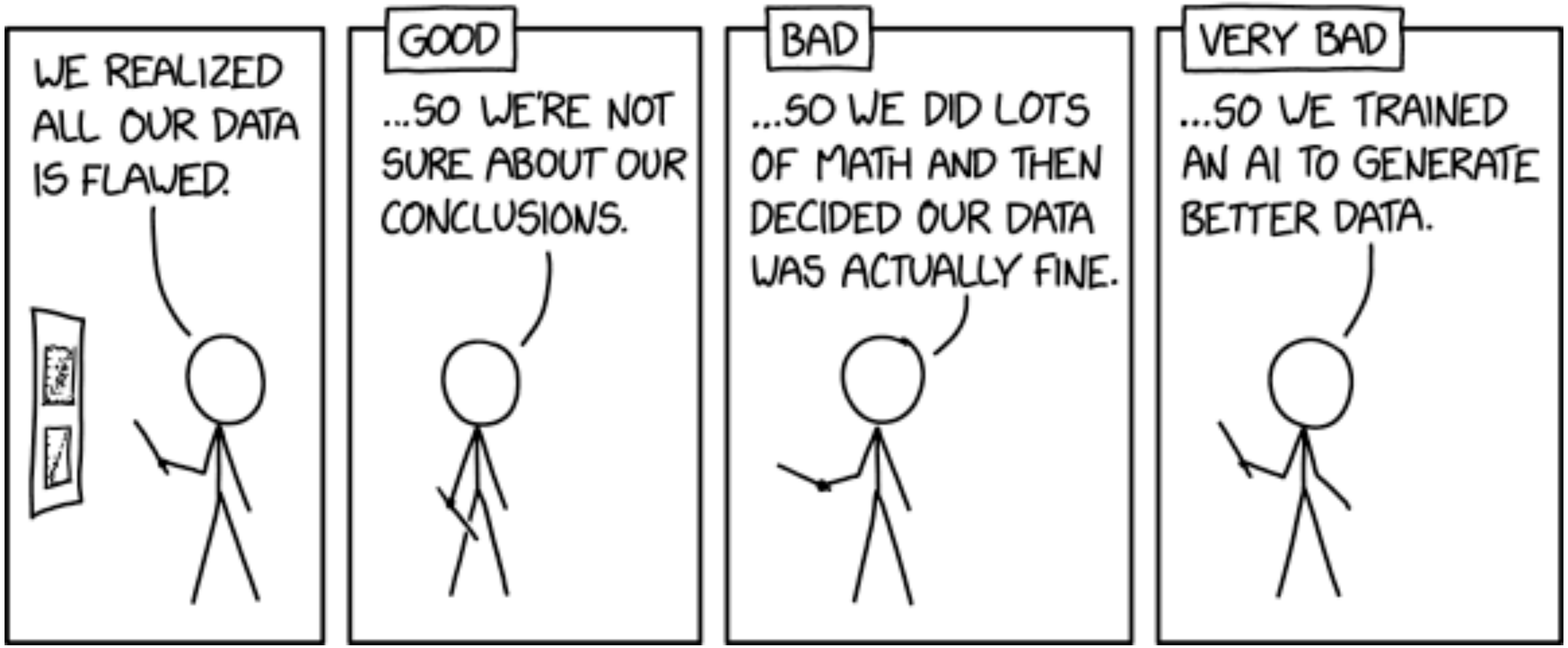
What is a representative sample?

If I’m interested in trying to predict the outcome of a U.S. presidential election, who should make up my sample?

More recent changes: Weighting surveys based on whom respondents recall voting for in a previous election (for partisan imbalances)

Garbage in, garbage out

+ Snapshots vs predictions



https://xkcd.com/2494/