Recap

- (Ch 1-2) Looking at data and relationships
- (Ch 3-5) Probability

Today and the next few lectures

- (Ch 6-9) Statistical inference
  - (Ch 6) How to draw general conclusions from a sample of the population
  - (Ch 7) How to assess the significance of the evidence against a hypothesis
  - (Ch 9) How to infer a probability model from a dataset

Motivation: midterm grading example

- Suppose your instructor is going to grade ~~100~~ 150 midterms

- Being impatient to know how the class did, he will first grade a sample of 5 randomly selected exams
  - Suppose he gives the following scores $\{120, 130, 140, 140, 150\}$
  - So the realized **sample mean** is 136

- What does he now know about the **population mean**?
  - What is his best guess for the population mean? 136
  - How confident should he be in his best guess?

Would be more confident if he had graded say 10 exams

Motivation: election polling example

		DATES	POLLSTER	SAMPLE	RESULT				NET RESULT
U.S. House	• IL-12	SEP 26-27	DCCC Targeting Team*	574 LV	Kelly	44%	46%	Bost	Bost +2
	• IL-12	SEP 26-27	DCCC Targeting Team*	574 LV	Kelly	41%	42%	Bost	Bost +1
Governor	• Ill.	SEP 24-29	Southern Illinois University	715 LV	Pritzker	49%	27%	Rauner	Pritzker +22

Source: fivethirtyeight.com

- The Southern Illinois University poll tells us:
  - The sample consists of 715 likely voters
  - Pritzker's vote percentage has realized **sample mean** equal to 49%

- What do we know about the **population mean** of Pritzker's Sep 24-29 vote percentage, where the population is all likely voters in Illinois?

Population

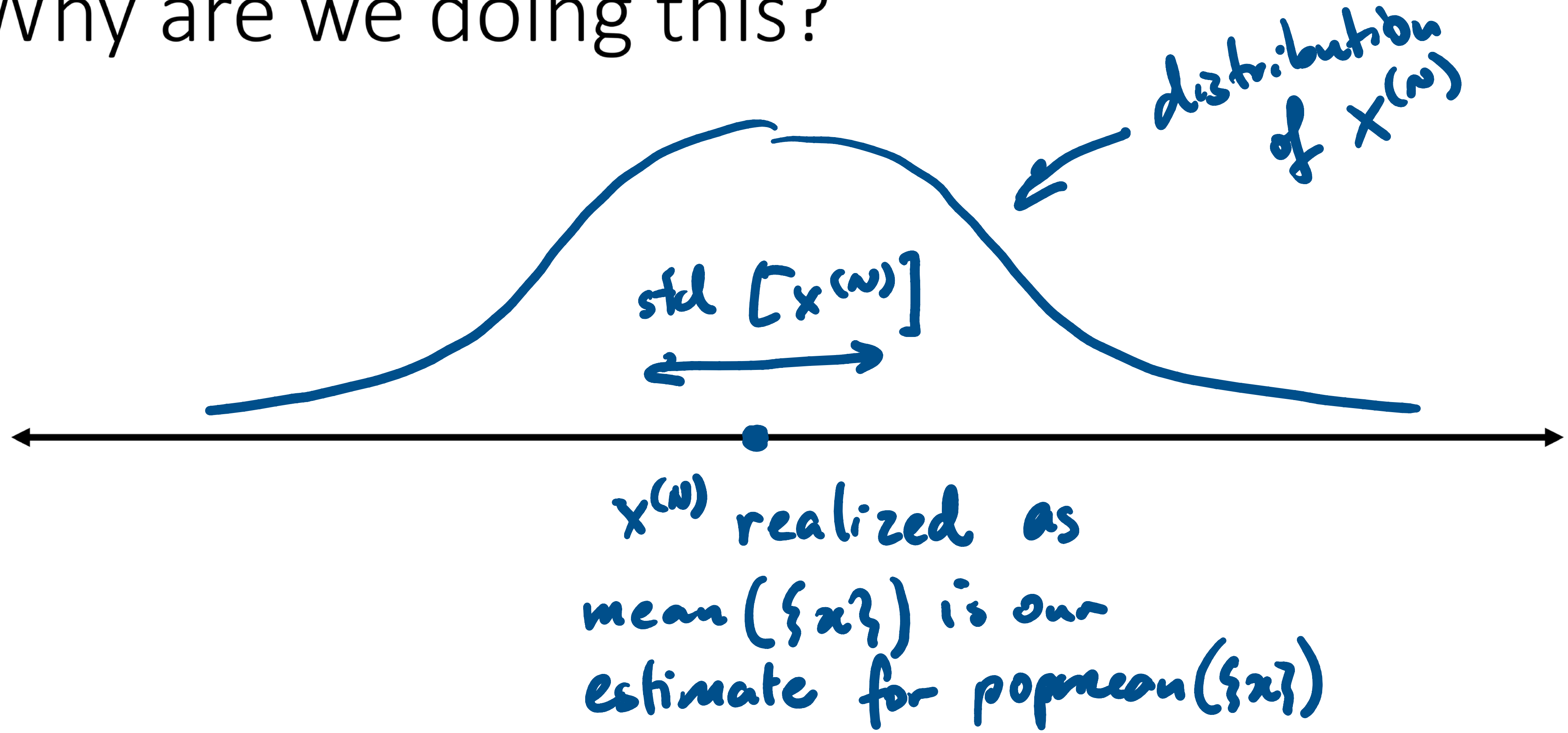
- The population is the entire dataset $\{x\}$
  - The population size is N_p
  - The population mean is $\text{popmean}(\{x\})$ and is a number (not a random variable)
  - The population standard deviation is $\text{popsd}(\{x\})$ and is a number
- We have formulas for $\text{popmean}(\{x\})$ and $\text{popsd}(\{x\})$ from Lec 1

Sample

- The sample is a **random** subset of the population $\{x\}$, where the sampling is done with replacement
  - The sample size is N, assumed to be much less than population size N_p $N \ll N_p$
  - The sample mean is $X^{(N)}$ and is a **random variable**

- Coming up, we will obtain expressions for
  - The expected value of the sample mean $E[X^{(N)}]$
  - The standard deviation of the sample mean $\text{std}[X^{(N)}]$ (also known as **standard error**)

Why are we doing this?



A smaller value of $\mathrm{std}[X^{(N)}]$ produces a narrower distribution, which would give us more confidence in the estimate of $\mathrm{popmean}(\{x\})$

Expected value of the sample mean …

- The sample mean is the average of IID samples

$$X^{(N)} = \frac{1}{N}(X_1 + X_2 + \cdots + X_N)$$

- By linearity of expectation and the fact that the sample mean $X^{(1)}$ of 1 sample is the sample itself

$$E\left[X^{(N)}\right] = \frac{1}{N}\left(E[X^{(1)}] + E[X^{(1)}] + \cdots + E[X^{(1)}]\right) = E[X^{(1)}]$$

… is the population mean

- Since each sample is sampled uniformly from the population

$$E[X^{(1)}] = \text{popmean}(\{x\})$$

- Therefore

$$E[X^{(N)}] = \text{popmean}(\{x\})$$

- We say that $X^{(N)}$ is an unbiased estimator for $\text{popmean}(\{x\})$

- We actually proved something very similar in Lec 9 in the proof of WLLN using slightly different notation

What does this mean?

- The sample mean is an unbiased estimate of the population mean as long as the samples have been drawn independently and with equal probability from the population

- Examples of poor sampling
  - To estimate the average height of people in Champaign, should I sample at the local daycare or perhaps try at the basketball team practice?
  - To poll likely voters, should we call them only at landline phone numbers? But it is costly for pollsters to call cellphones since robocalling them is illegal

Standard deviation of the sample mean

- We can also rewrite another result from Lec 9 as

$$\mathrm{var}[X^{(N)}] = \frac{\mathrm{popvar}(\{x\})}{N}$$

- So the **standard error** (standard deviation of the sample mean) is

$$\mathrm{stderr}(\{x\}) = \mathrm{std}[X^{(N)}] = \frac{\mathrm{popsd}(\{x\})}{\sqrt{N}}$$

- But we don't know the value of the population standard deviation

Estimating the population standard deviation

- We will try to use the realized sample to estimate $\text{popsd}(\{x\})$

realized mean

- Does this work?

$$\sqrt{\frac{1}{N}\sum_{x_i \epsilon \text{sample}} (x_i - \text{mean}(\{x_i\}))^2}$$

- No! This formula turns out to underestimate $\text{popsd}(\{x\})$ on average

- To make it an unbiased estimator, we must multiply it by $\sqrt{\frac{N}{N-1}}$

Unbiased estimate of population std. dev.

- The unbiased estimate of $\mathrm{popsd}(\{x\})$ is defined as

$$\mathrm{stdunbiased}(\{x\}) = \sqrt{\frac{1}{N-1} \sum_{x_i \epsilon \mathrm{sample}} (x_i - \mathrm{mean}(\{x_i\}))^2}$$

- So the standard error is estimated as

$$\mathrm{stderr}(\{x\}) = \mathrm{std}[X^{(N)}] = \frac{\mathrm{popsd}(\{x\})}{\sqrt{N}} = \frac{\mathrm{stdunbiased}(\{x\})}{\sqrt{N}}$$

Standard error: midterm grading example

The realized sample of scores is $\{120, 130, 140, 140, 150\}$ with $N = 5$

$$\text{mean}(\{120, 130, 140, 140, 150\}) = 136$$

$$\text{stdunbiased}(\{x\}) = \sqrt{\frac{1}{5-1}\left((120-136)^2 + (130-136)^2 + 2(140-136)^2 + (150-136)^2\right)} \simeq 11.4$$

$$\text{stderr}(\{x\}) = \frac{11.4}{\sqrt{5}} \simeq 5.1$$

So we estimate the population mean as 136 with standard error 5.1

Standard error: election polling example

		DATES	POLLSTER	SAMPLE	RESULT				NET RESULT
U.S. House	IL-12	SEP 26-27	DCCC Targeting Team*	574 LV	Kelly	44%	46%	Bost	Bost +2
	IL-12	SEP 26-27	DCCC Targeting Team*	574 LV	Kelly	41%	42%	Bost	Bost +1
Governor	Ill.	SEP 24-29	Southern Illinois University	715 LV	Pritzker	49%	27%	Rauner	Pritzker +22

Source: fivethirtyeight.com

Number of sampled voters who selected Pritzker

$= 715(0.49) \approx 350$

Number of sampled voters who did not select Pritzker

$= 715(0.51) \approx 365$

Standard error: election polling example

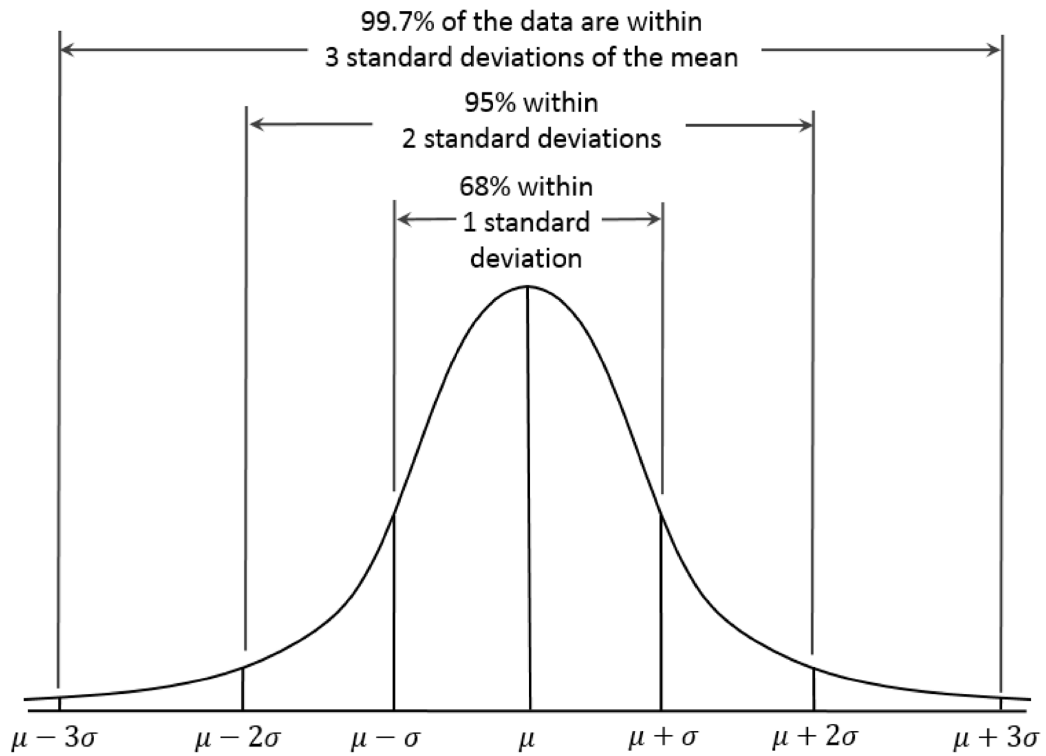
$$\text{stdunbiased}(\{x\}) = \sqrt{\frac{1}{715-1}\left(350(1-0.49)^2 + 365(0-0.49)^2\right)} \simeq 0.50$$

$$\text{stderr}(\{x\}) = \frac{0.50}{\sqrt{715}} \simeq 0.019 = 1.9\%$$

So we estimate the population mean as 49% with standard error 1.9%

Interpreting the standard error when $N \geq 30$

If the sample size $N \geq 30$, the Central Limit Theorem tells us that we can approximate the distribution of the sample mean $X^{(N)}$ as a normal distribution with $\mu = \text{popmean}(\{x\})$ and $\sigma = \text{stderr}(\{x\})$



Confidence intervals when $N \geq 30$

- For about 68% of samples

$$\text{mean}(\{x\}) - \text{stderr}(\{x\}) \leq \text{popmean}(\{x\}) \leq \text{mean}(\{x\}) + \text{stderr}(\{x\})$$

- For about 95% of samples

$$\text{mean}(\{x\}) - (2)\text{stderr}(\{x\}) \leq \text{popmean}(\{x\}) \leq \text{mean}(\{x\}) + (2)\text{stderr}(\{x\})$$

- For about 99% of samples

$$\text{mean}(\{x\}) - (3)\text{stderr}(\{x\}) \leq \text{popmean}(\{x\}) \leq \text{mean}(\{x\}) + (3)\text{stderr}(\{x\})$$

Confidence interval: election polling example

		DATES	POLLSTER	SAMPLE	RESULT				NET RESULT
U.S. House	IL-12	SEP 26-27	DCCC Targeting Team*	574 LV	Kelly	44%	46%	Bost	Bost +2
	IL-12	SEP 26-27	DCCC Targeting Team*	574 LV	Kelly	41%	42%	Bost	Bost +1
Governor	Ill.	SEP 24-29	Southern Illinois University	715 LV	Pritzker	49%	27%	Rauner	Pritzker +22

N=715 >> 30

Source: fivethirtyeight.com

- We estimated the population mean as 49% with standard error 1.9%
- The 99% confidence interval for Pritzker's vote percentage is

$$[49\% - (3)1.9\%, 49\% + (3)1.9\%] = [43.3\%, 54.7\%]$$

Confidence intervals when $N < 30$

- If the sample size $N < 30$, we should not use the normal distribution to create confidence intervals

- Instead we use Student's t-distribution with its parameter (called degrees of freedom) set to $N - 1$

- In the midterm grading example, we should use Student's t-distribution with degrees of freedom set to 4 since $N = 5$