Week 6

Sampling – When You Can’t or Won’t Have ALL the Data

Applied Data Science

Columbia University - Columbia Engineering

Course Agenda

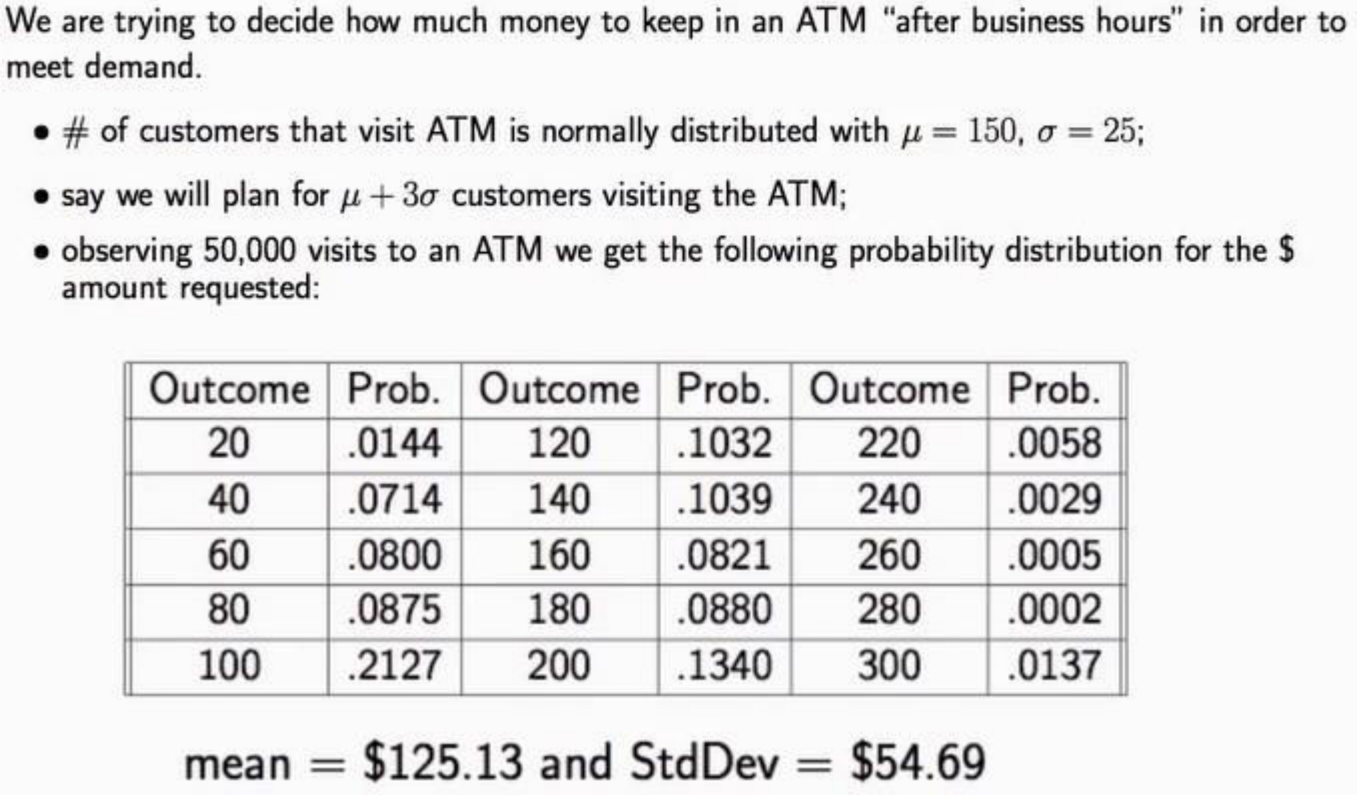
COLUMBIA | ENGINEERING
EXECUTIVE EDUCATION

- Week 1: Python Basics: How to Translate Procedures into Codes
- Week 2: Intermediate Python – Data structures for Your Analysis
- Week 3: Relational Databases – Where Big Data is Typically Stored
- Week 4: SQL – Ubiquitous Database Format/Language
- Week 5: Statistical Distributions – The Shape of Data
- **Week 6: Sampling – When You Can't or Won't Have ALL the Data**
- Week 7:Hypothesis Testing – Answering Questions about Your Data
- Week 8: Data Analysis and Visualization – Using Python's NumPy for Analysis
- Week 9: Data analysis and visualization – Using Python's Pandas for Data Wrangling
- Week 10: Text Mining – Automatic Understanding of Text
- Week 11: Machine learning – Basic Regression and Classification
- Week 12: Machine learning – Decision Trees and Clustering

Sample Mean and CLT

- Central Limit Theorem (CLT)
- Sample mean
- Accuracy of a sample estimate; sample size selection

Demand in ATM



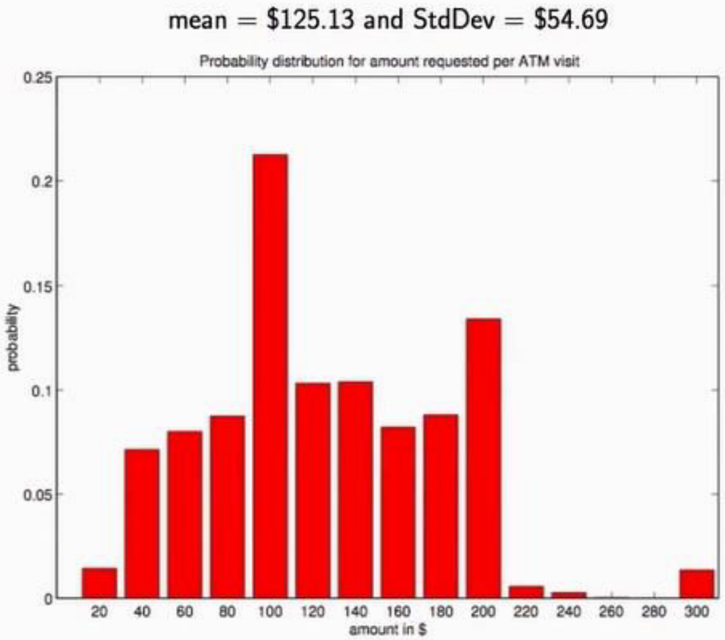
We are trying to decide how much money to keep in an ATM "after business hours" in order to meet demand.

- # of customers that visit ATM is normally distributed with $\mu = 150$, $\sigma = 25$;
- say we will plan for $\mu + 3\sigma$ customers visiting the ATM;
- observing 50,000 visits to an ATM we get the following probability distribution for the $ amount requested:

Outcome	Prob.	Outcome	Prob.	Outcome	Prob.
20	.0144	120	.1032	220	.0058
40	.0714	140	.1039	240	.0029
60	.0800	160	.0821	260	.0005
80	.0875	180	.0880	280	.0002
100	.2127	200	.1340	300	.0137

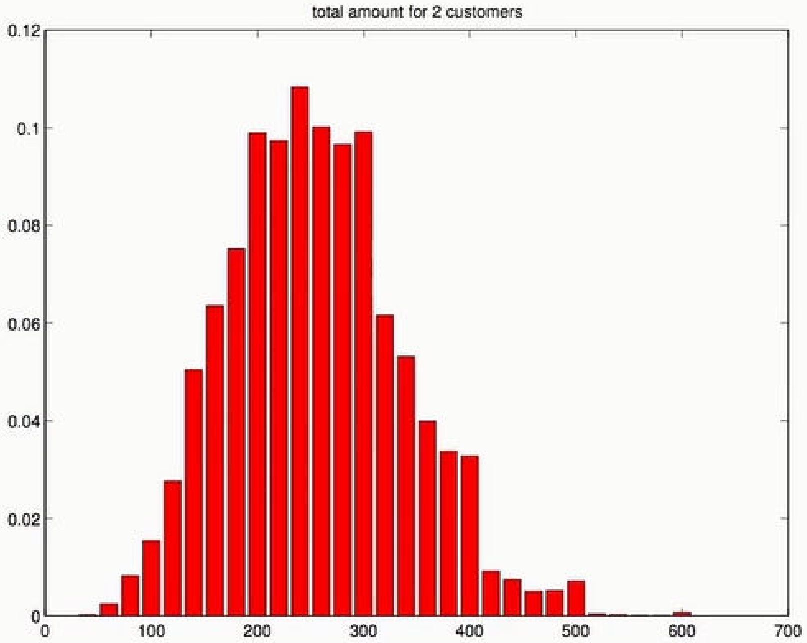
mean = \$125.13 and StdDev = \$54.69

ATM Demand

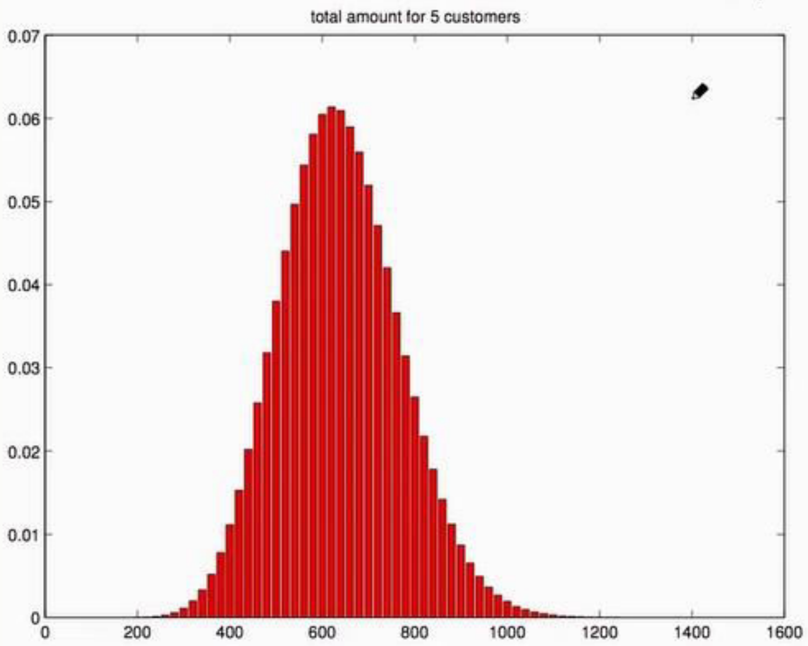


This does not seem to be normally distributed

2 customer: ATM Demand



5 customer: ATM Demand



Almost close to Normal Distribution

ATM Demand

CLT: *If $X_1, \ldots, X_n$ are random variables*

- *all having the same distribution with mean μ and variance σ^2,*

and

- *all independent,*

then

$$X_1 + X_2 + \cdots + X_n \approx N(n\mu, n\sigma^2).$$

In terms of the **sample mean** *we have that*

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \approx$$

Remarks:

- If the X_i's were normal then $\bar{X}$ would be normal.
- Need **independent & identical distributed**.
- The importance of the normal distribution is largely due to CLT.

$$\begin{tikzcd}
\textbf{Model} \arrow[r, shift left=2, "\text{probability}"] & \textbf{Data} \arrow[l, shift left=2, "\text{statistics}"]
\end{tikzcd}$$

Examples of models:

- demand for Parkas is $N(\mu, \sigma^2)$
- ATM withdrawal amounts: mean μ, stdev σ

We use model to

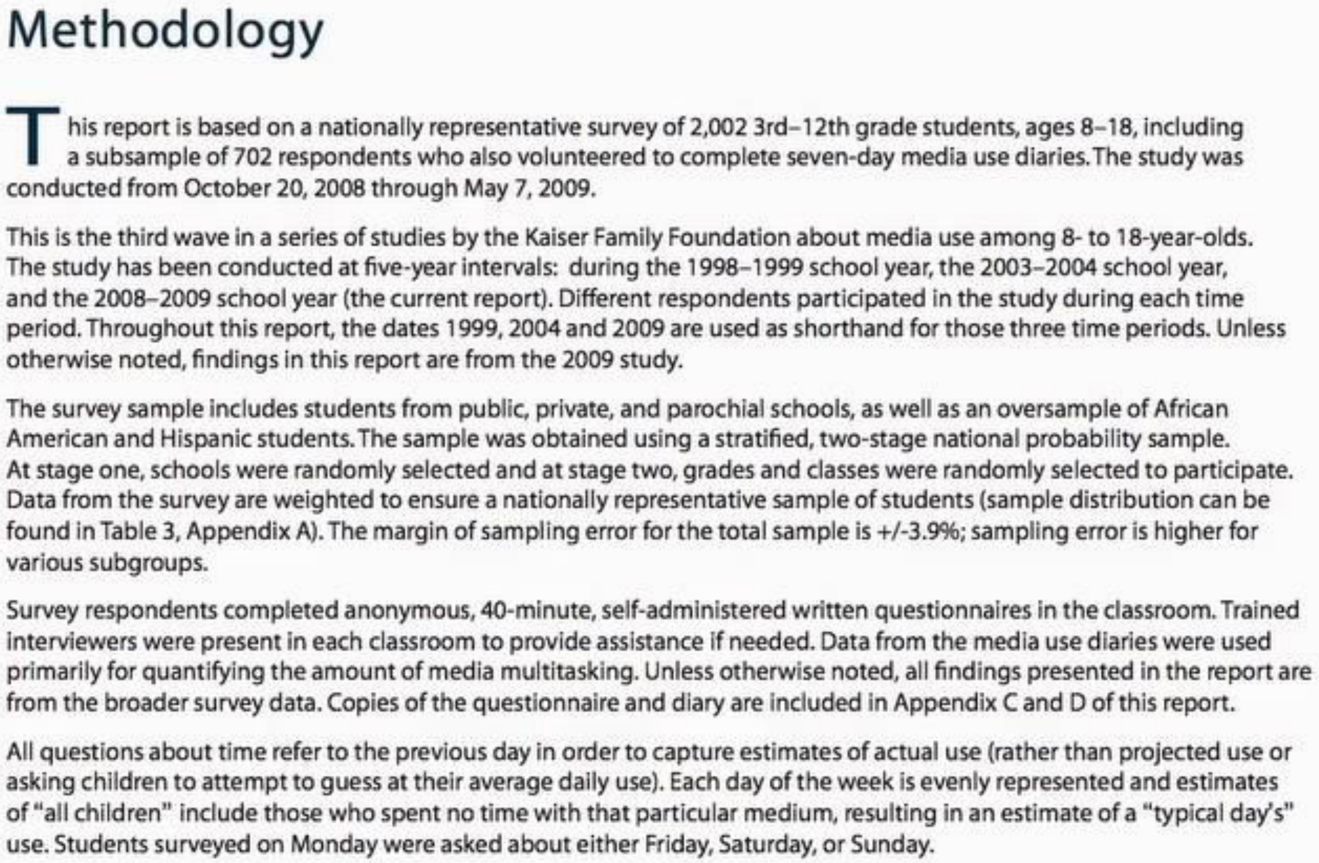
→ predict actual data; e.g., realized demand, actual stock returns

→ make managerial decisions; e.g., plan production capacity, optimize portfolio, etc.

Model depends on **parameters**; e.g., μ, $\sigma, \ldots$

…use data to estimate these parameters

Kaiser Family Study



Methodology

This report is based on a nationally representative survey of 2,002 3rd–12th grade students, ages 8–18, including a subsample of 702 respondents who also volunteered to complete seven-day media use diaries. The study was conducted from October 20, 2008 through May 7, 2009.

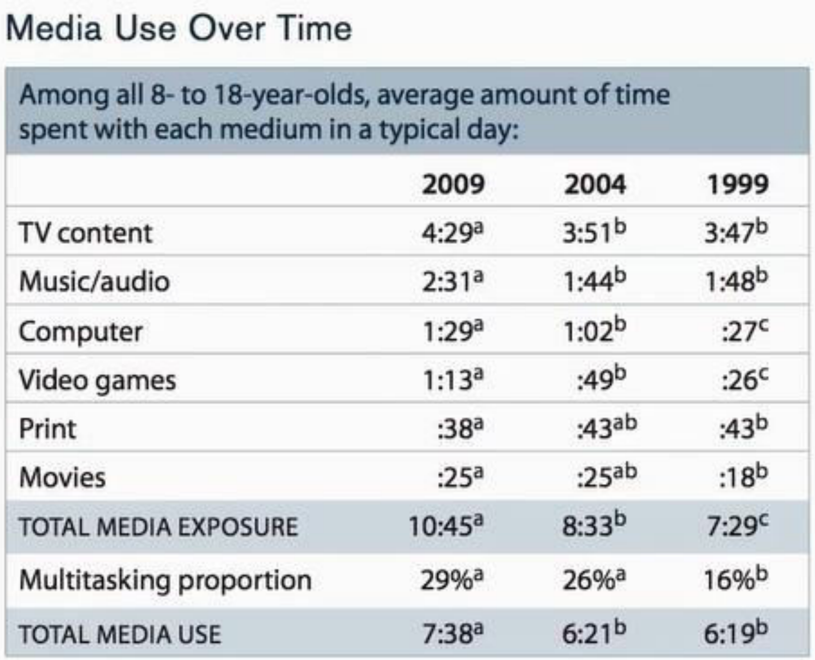
This is the third wave in a series of studies by the Kaiser Family Foundation about media use among 8- to 18-year-olds. The study has been conducted at five-year intervals: during the 1998–1999 school year, the 2003–2004 school year, and the 2008–2009 school year (the current report). Different respondents participated in the study during each time period. Throughout this report, the dates 1999, 2004 and 2009 are used as shorthand for those three time periods. Unless otherwise noted, findings in this report are from the 2009 study.

The survey sample includes students from public, private, and parochial schools, as well as an oversample of African American and Hispanic students. The sample was obtained using a stratified, two-stage national probability sample. At stage one, schools were randomly selected and at stage two, grades and classes were randomly selected to participate. Data from the survey are weighted to ensure a nationally representative sample of students (sample distribution can be found in Table 3, Appendix A). The margin of sampling error for the total sample is +/-3.9%; sampling error is higher for various subgroups.

Survey respondents completed anonymous, 40-minute, self-administered written questionnaires in the classroom. Trained interviewers were present in each classroom to provide assistance if needed. Data from the media use diaries were used primarily for quantifying the amount of media multitasking. Unless otherwise noted, all findings presented in the report are from the broader survey data. Copies of the questionnaire and diary are included in Appendix C and D of this report.

All questions about time refer to the previous day in order to capture estimates of actual use (rather than projected use or asking children to attempt to guess at their average daily use). Each day of the week is evenly represented and estimates of "all children" include those who spent no time with that particular medium, resulting in an estimate of a "typical day's" use. Students surveyed on Monday were asked about either Friday, Saturday, or Sunday.

Summary findings:



Media Use Over Time

Among all 8- to 18-year-olds, average amount of time spent with each medium in a typical day:

	2009	2004	1999
TV content	4:29^{a}	3:51^{b}	3:47^{b}
Music/audio	2:31^{a}	1:44^{b}	1:48^{b}
Computer	1:29^{a}	1:02^{b}	:27^{c}
Video games	1:13^{a}	:49^{b}	:26^{c}
Print	:38^{a}	:43^{ab}	:43^{b}
Movies	:25^{a}	:25^{ab}	:18^{b}
TOTAL MEDIA EXPOSURE	10:45^{a}	8:33^{b}	7:29^{c}
Multitasking proportion	29%^{a}	26%^{a}	16%^{b}
TOTAL MEDIA USE	7:38^{a}	6:21^{b}	6:19^{b}

http://kaiserfamilyfoundation.files.wordpress.com/2013/04/8010.pdf

Kaiser Survey

Based on 702 respondents that completed 7-day media diaries:

- "average daily media usage for 8- to 18-year olds is 7hrs 38min (458 min)"
- is that a big enough sample to draw meaningful statistical conclusions?
- what is the "accuracy" of the above estimate?

P(Kaiser's estimate is off by more than 15min) $\leq$

Before study:

- Random sample: $X_1, X_2, \ldots, X_{702}$
- $X_i = \#$ min of media usage per day for the i^{th} respondent
- Assume that these are:
  - independent
  - identically distributed: same mean μ, variance σ^2.

Sample Mean

What is μ? True population Avg # min that kids spent in front of these media outlets

How do we estimate μ?

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}, \quad (n = 702 \text{ here}).$$

"tolerance"

Estimate will be off by more than 15 min if

$$P(\bar{X} > \mu + 15) + P(\bar{X} < \mu - 15)$$

Distribution of $\bar{X}$:

- $\mathbb{E}[\bar{X}] = \mu$

CLT: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

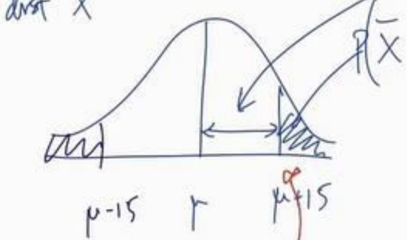
- $Var[\bar{X}] = \frac{\sigma^2}{n}$

$$\text{Stdev}(\bar{X}) = \text{Stderror}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Estimating the Probability of Error

Need to find $P(\bar{X} > \mu + 15)=$ measure this dist. in stdev $(\bar{X})$

dist $\bar{X}$



$$P(\bar{X} > \mu+15) = P\left(Z > \frac{15}{\sigma/\sqrt{702}}\right) \quad \sigma = ?$$

$$= P\left(Z > \frac{15}{180/\sqrt{702}}\right) = P(Z > 2.21) = 1.4\%$$

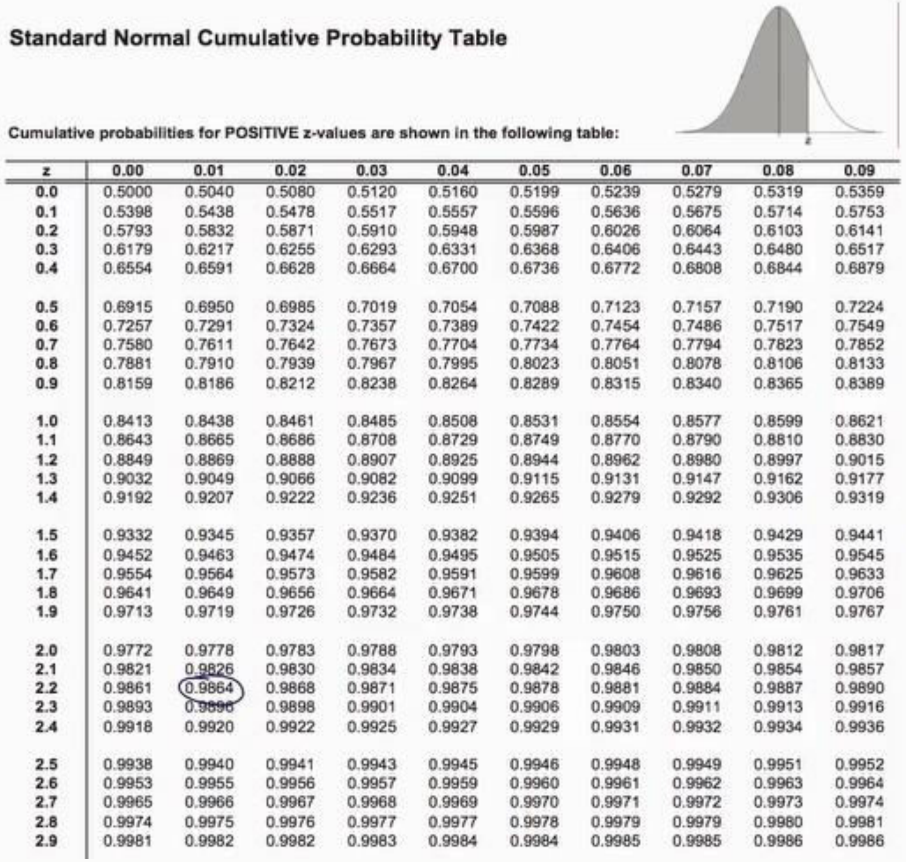
$P(|\text{error}| > 15) = 2 \cdot (1.4\%) = 2.8\%$ (low)

Need guess σ:

a) Guess reasonable range of responses: (0, 18 hrs) large upper bound

b) Most data within $\pm 3\sigma$ of mean $\Rightarrow$ Range $\approx 6\sigma$ wide

$$\Rightarrow \sigma \sim \frac{18}{6} = 3 \text{ hrs} = 180 \text{ min}$$



Standard Normal Cumulative Probability Table

Cumulative probabilities for POSITIVE z-values are shown in the following table:

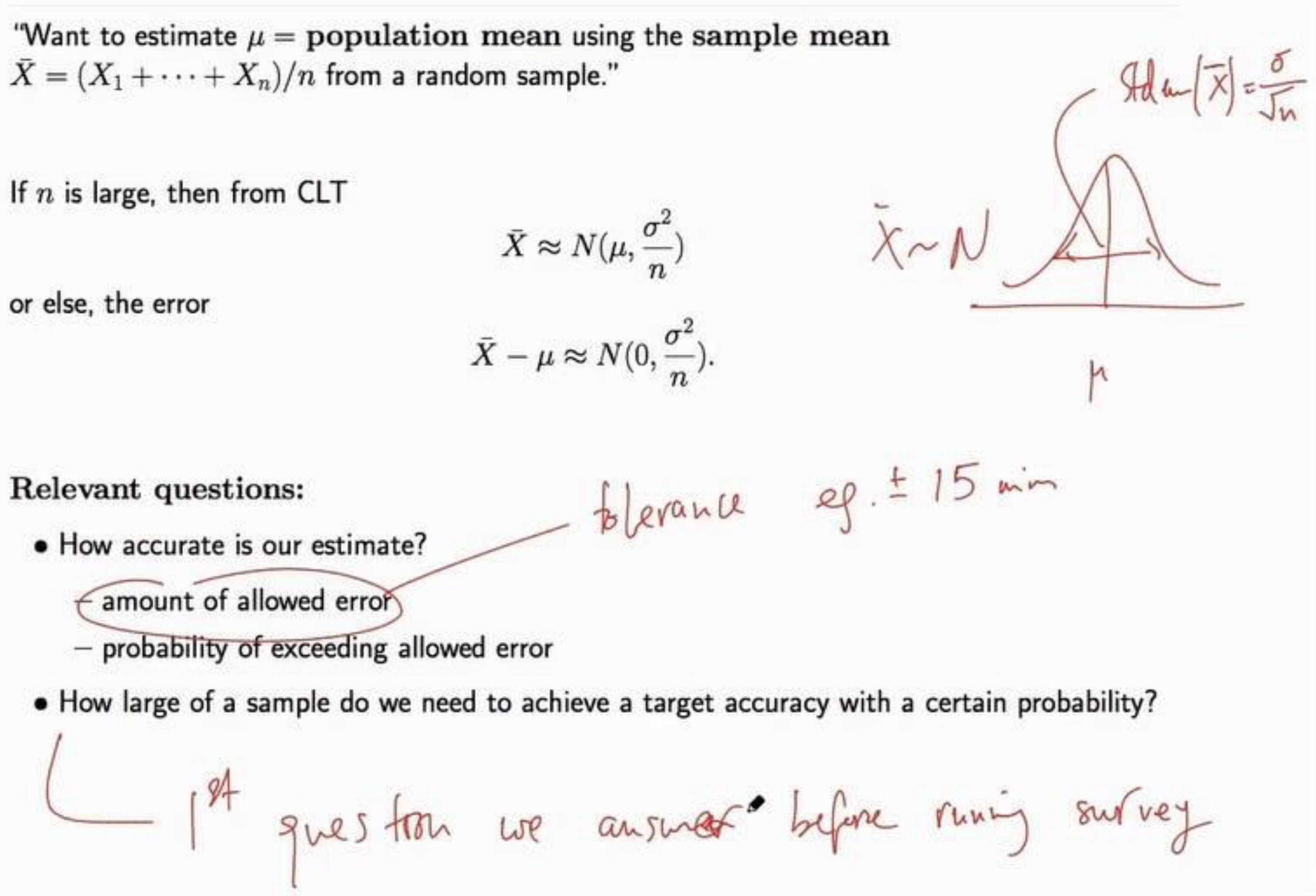
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

$$P(Z > 2.21)$$

$$= 1 - .9864$$

$$= 1.4\%$$

Recap of Kaiser Example



"Want to estimate $\mu =$ **population mean** using the **sample mean** $\bar{X} = (X_1 + \cdots + X_n)/n$ from a random sample."

If n is large, then from CLT

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$

or else, the error

$$\bar{X} - \mu \approx N(0, \frac{\sigma^2}{n}).$$

$\bar{X} \sim N$

$\text{Std dev}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

μ

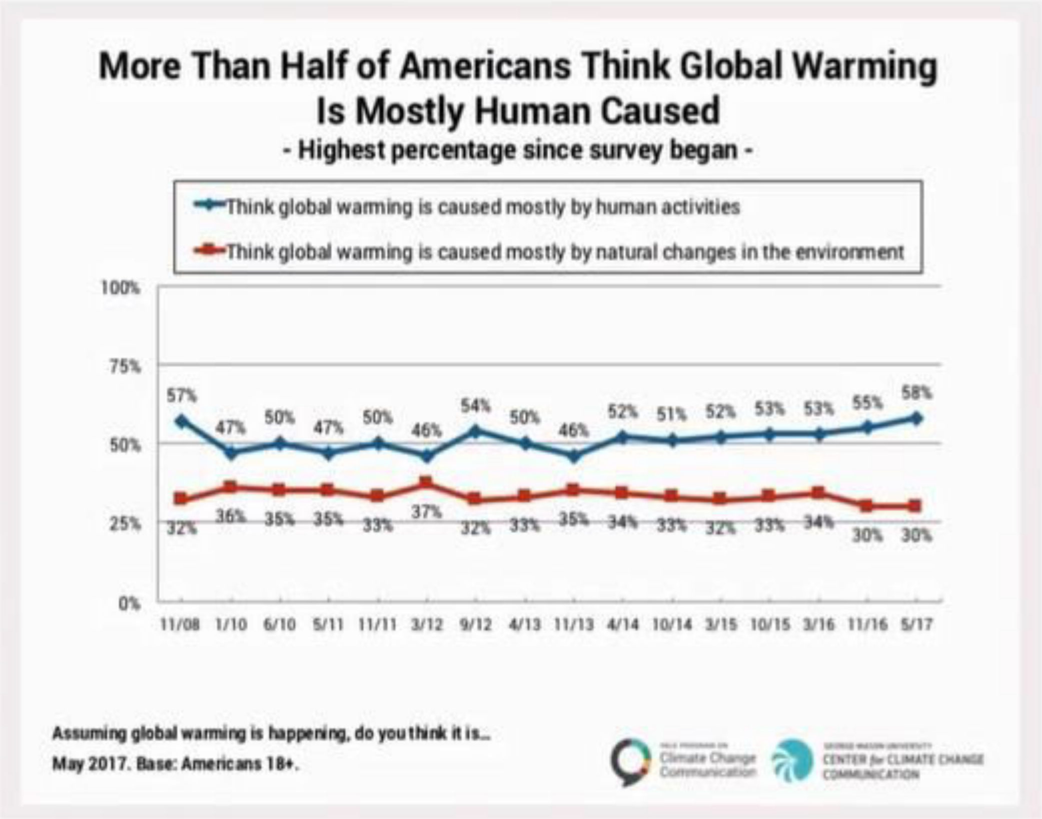
Relevant questions:

- How accurate is our estimate?
  - amount of allowed error
  - probability of exceeding allowed error
- How large of a sample do we need to achieve a target accuracy with a certain probability?

tolerance eg. ± 15 min

1st question we answer before running survey

Poll and Estimating Proportion



More Than Half of Americans Think Global Warming Is Mostly Human Caused

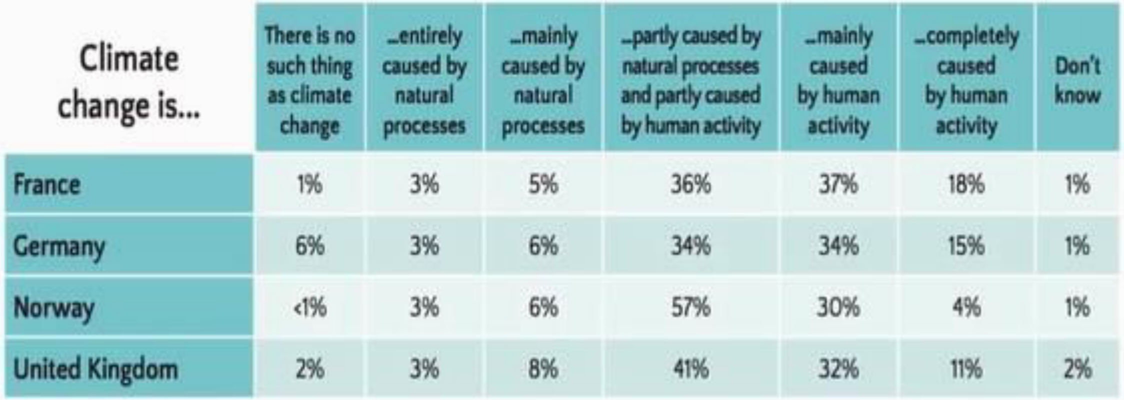
- Highest percentage since survey began -

Assuming global warming is happening, do you think it is...

May 2017. Base: Americans 18+.

European Perceptions on Climate Change report, Mar 2017
(1,000 respondents in each country)

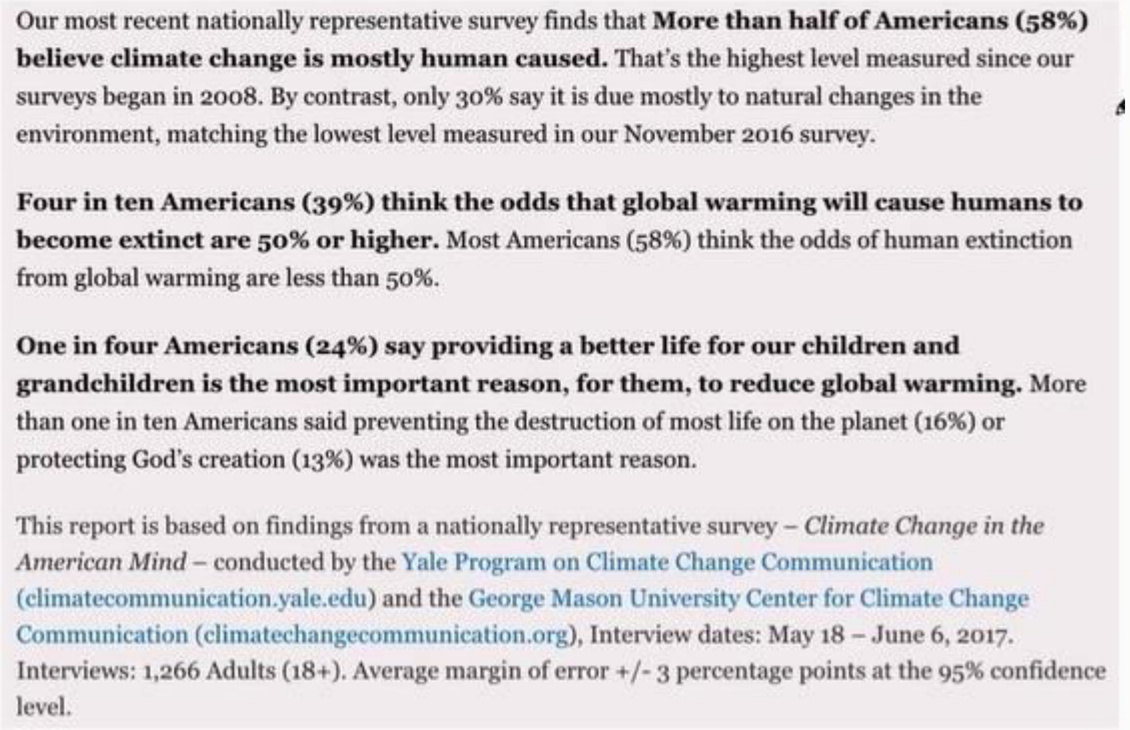
Table 4. Thinking about the causes of climate change, which, if any, of the following best describes your opinion? (Question 5)



Climate change is...	There is no such thing as climate change	...entirely caused by natural processes	...mainly caused by natural processes	...partly caused by natural processes and partly caused by human activity	...mainly caused by human activity	...completely caused by human activity	Don't know
France	1%	3%	5%	36%	37%	18%	1%
Germany	6%	3%	6%	34%	34%	15%	1%
Norway	<1%	3%	6%	57%	30%	4%	1%
United Kingdom	2%	3%	8%	41%	32%	11%	2%

There appears to be difference of perception on Climate change between European nations and US

Poll and Estimating Proportion



Our most recent nationally representative survey finds that **More than half of Americans (58%) believe climate change is mostly human caused.** That's the highest level measured since our surveys began in 2008. By contrast, only 30% say it is due mostly to natural changes in the environment, matching the lowest level measured in our November 2016 survey.

Four in ten Americans (39%) think the odds that global warming will cause humans to become extinct are 50% or higher. Most Americans (58%) think the odds of human extinction from global warming are less than 50%.

One in four Americans (24%) say providing a better life for our children and grandchildren is the most important reason, for them, to reduce global warming. More than one in ten Americans said preventing the destruction of most life on the planet (16%) or protecting God's creation (13%) was the most important reason.

This report is based on findings from a nationally representative survey – *Climate Change in the American Mind* – conducted by the Yale Program on Climate Change Communication (climatecommunication.yale.edu) and the George Mason University Center for Climate Change Communication (climatechangecommunication.org), Interview dates: May 18 – June 6, 2017. Interviews: 1,266 Adults (18+). Average margin of error +/- 3 percentage points at the 95% confidence level.

- What fraction of Americans believe that climate change is mostly human caused? . . . Conduct a poll.

  - Use a sample of 1,266 adults

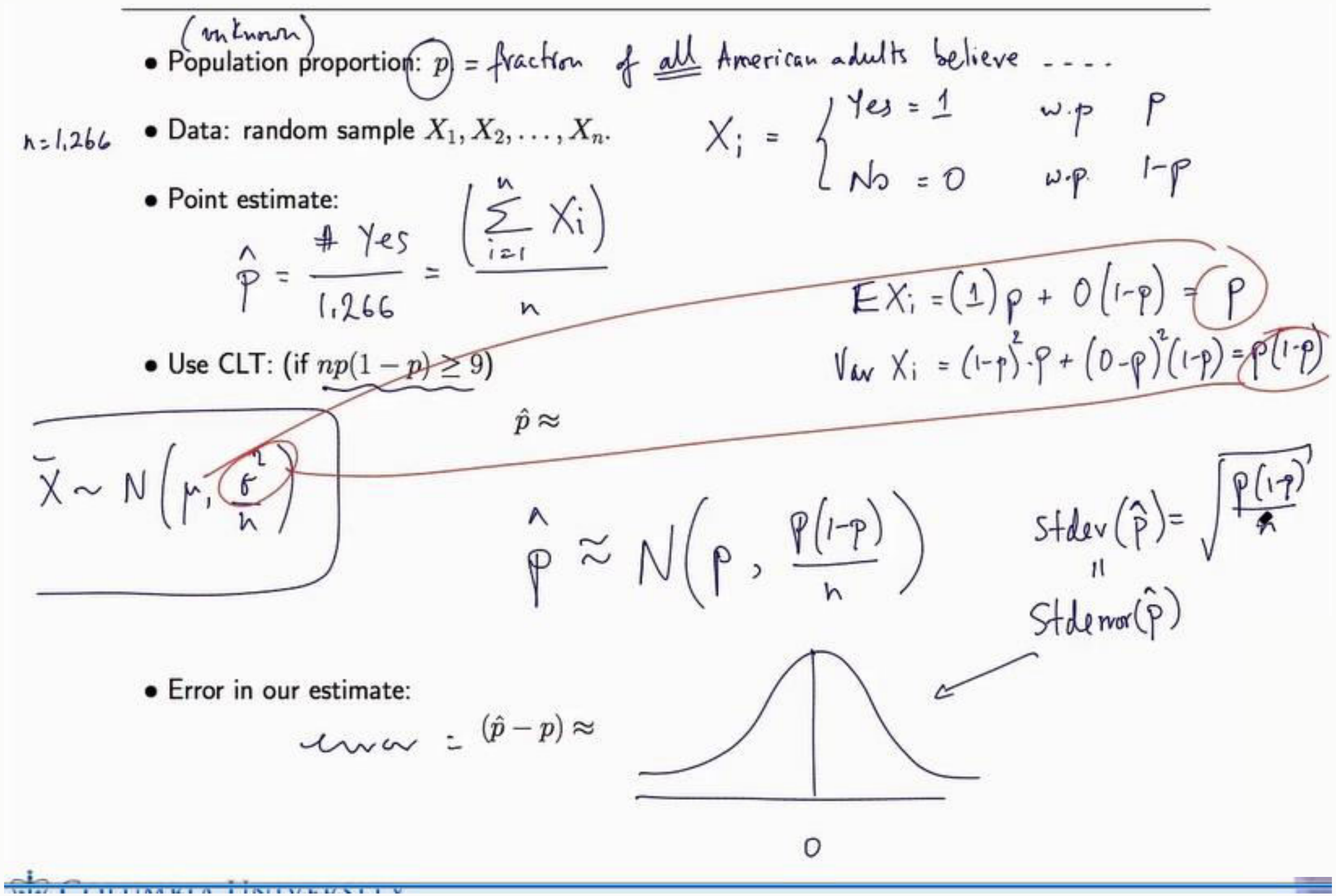
  - Sample estimates:

    mostly human caused, 58%
    mostly natural changes in environment, 30%

- What is the interpretation of the "margin of sampling error of plus or minus three percentage points"?

- Is a sample of 1,266 sufficiently large?

Poll and Estimating Proportion



- Population proportion: $\overset{\text{(unknown)}}{p}$ = fraction of all American adults believe ....

n=1,266

- Data: random sample $X_1, X_2, \ldots, X_n$.

$$X_i = \begin{cases} \text{Yes} = 1 & \text{w.p } p \\ \text{No} = 0 & \text{w.p. } 1-p \end{cases}$$

- Point estimate:

$$\hat{p} = \frac{\#\text{ Yes}}{1{,}266} = \frac{\left(\sum_{i=1}^{n} X_i\right)}{n}$$

$$EX_i = (1)p + 0(1-p) = p$$

$$\text{Var } X_i = (1-p)^2 \cdot p + (0-p)^2(1-p) = p(1-p)$$

- Use CLT: (if $np(1-p) \geq 9$)

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\hat{p} \approx$

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

$$\text{Stdev}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

$$\| \quad \text{Stderror}(\hat{p})$$

- Error in our estimate:

error $= (\hat{p} - p) \approx$

Accuracy of the Poll

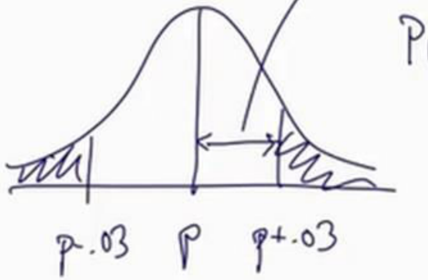
- Poll: $\boxed{1{,}266}$ adults
- Accuracy must be specified as $\mathbb{P}$(error exceeds margin of error) = . . .

measure dist. in stdev.

±3%

$P(|\text{error}| > .03) \approx$ **4%**

Dist. $\hat{p}$



$$P(|\text{error}| > .03) = 2 \cdot \mathbb{P}(\hat{p} > p + .03)$$

$$= 2 \cdot \mathbb{P}\left(Z > \frac{.03}{\sqrt{\frac{p(1-p)}{1266}}}\right)$$

p is still in this expression (unknown)

Guess reasonable value for p:

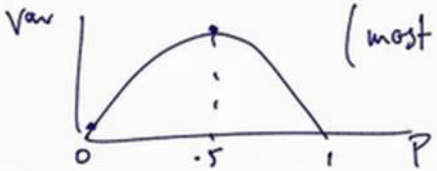
1. Use $\hat{p} = .58$

$$= 2 \cdot \mathbb{P}\left(Z > \frac{.03}{\sqrt{\frac{\frac{1}{2}\,\frac{1}{2}}{1266}}}\right) = 2 \cdot \mathbb{P}(Z > 2.13)$$

$$\approx .04$$

2. Use conservative estimate for p:

$\boxed{p = 1/2}$



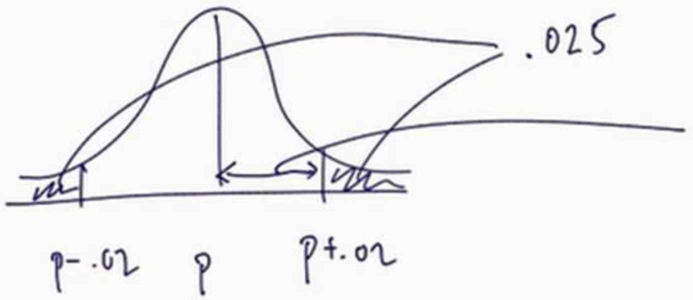
(most random) - Always use $p = 1/2$ before you collect data & when the poll has several questions

How Large of a Poll Do We Need?

1% n ~ 4 times larger

- Suppose we want to be within 2% with probability 95%.
- How many people should we poll?

In order to have a tail prob. of .025



$$.02 = (1.96)\ \text{Stderror}(\hat{p})$$

$$= (1.96)\sqrt{\frac{p(1-p)}{n}}$$

use conserv. estimate

$$= (1.96)\cdot\sqrt{\frac{\frac{1}{2}\,\frac{1}{2}}{n}}$$

Sample size for poll

$$n \sim (1.96)^2\,\frac{1}{4}\,\frac{1}{(\text{margin of error})^2}$$

$$\Longrightarrow n = (1.96)^2\,\frac{1}{4}\left(\frac{1}{.02}\right)^2 \approx 2400$$

Summary of Sampling Proportion

- Use **data** from a **random sample** to estimate desired parameter:
  - population mean μ
  - population proportion p
- 2 issues are important in quantifying how good is our estimate?
  - margin of allowed error
  - probability of exceeding this error
- 2 questions are interesting:
  - given data sample and error specification what is

    $$\mathbb{P}(\text{exceeding allowed error}) = ?$$

  - how much data is sufficient to guarantee desired accuracy for an estimate?

Our approach:

- Use CLT to approximate error $\bar{X} - \mu$ or $\hat{p} - p$ with a normal r.v..
- Use sample to find an estimate for $\mathrm{Var}[\bar{X}]$ or $\mathrm{Var}[\hat{p}]$.

Confidence Level

We would like to complement our **point estimate** $\bar{X}$ or $\hat{p}$ **with an interval**

data → estimate

"We are 95% confident that the true parameter p lies in the interval between this number and that number"

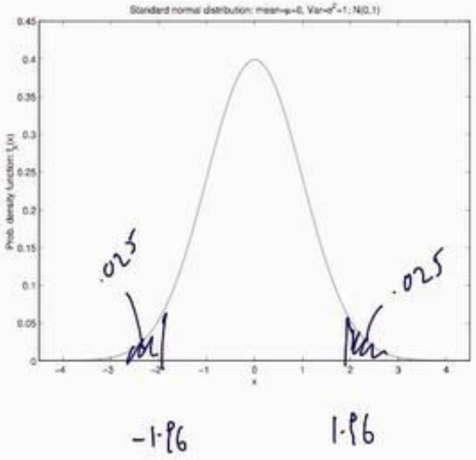
Approach:

- All CIs are computed in the same way:

$$\begin{aligned} \text{CI} &= \overset{.58}{(\text{point estimate})} \pm (\text{margin of error}) \\ &= \text{point estimate} \pm \text{multiplier} \times \text{Stderror(estimator)} \end{aligned}$$

- Stderror is the Stdev of the estimator, e.g., $\bar{X}$ or $\hat{p}$

- How can we construct a confidence interval for the true population parameter p of the form $\hat{p} \pm \ldots$?
- Using that $\hat{p} \approx N(p, p(1-p)/n)$, find z such that $\mathbb{P}$(being within z stdev's) $= 0.95$



To be 95% conf. ⟺ 2.5% on each tail

$z = 1.96$

So, we will be 95% confident that p lies in:

$$\text{CI} = \hat{p} \pm 1.96 \times \text{Stderror}(\hat{p})$$

$$= .58 \pm (1.96) \underbrace{\sqrt{\frac{.58(1-.58)}{1266}}}_{\substack{\parallel \\ \text{stderror of our estimate}}}$$

CIs at Different Class Levels

$$\begin{aligned} \text{CI} &= \text{point estimate} \pm \text{margin of error} \\ &= \text{point estimate} \pm \text{multiplier} \times \text{Stderror(estimator)} \end{aligned}$$

What would be the 90% CI for the population proportion p?

Confidence level	$\alpha = (1 - \text{CI})$	Multiplier	CI
90%	.10		
95%	.05	1.96	$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
99%	.01		

- Notation: let z_w be the point such that $\mathbf{P}(Z \geq z_w) = w$ (i.e., area to the right of z_w is w)

In general CIs are expressed in terms of α:
for a level-α CI, multiplier $= z_{\alpha/2}$

GLOBAL IVY EDUCATION

www.emeritus.org