Polling and sampling.

Clement de Chaisemartin

UCSB

What pollsters do

- Pollsters want to answer difficult questions:

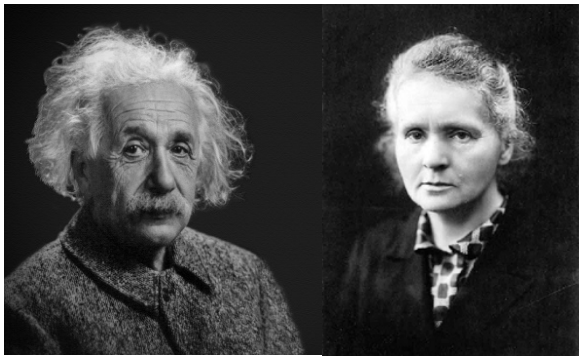
As of November 1st 2016, what % of the Pennsylvania electorate wants to vote for Hillary Clinton?

A brute force solution

- Interview all registered voters in Pennsylvania, and ask them if they want to vote for Clinton
- Issue: there are 8,448,674 registered voters in Pennsylvania!

A smarter solution

- Luckily, you have two very smart friends:



- Interview a sample of registered voters in Pennsylvania. Find that 53% of them intend to vote for Clinton.
- Then go to E-views, and make this very cryptic statement:

"We can be 95% confident that the share of registered voters in Pennsylvania that are willing to vote for Clinton is included between 50.8% and 55.2%. Moreover, we can be more than 99% confident that more than 50% of Pennsylvania voters are willing to vote for Clinton."

How many voters do we need to interview to make such a precise statement?

- After interviewing a sample of registered voters, your friends are able to make a very precise statement.
- Able to say that the share of voters that want to vote for Clinton has a very large probability (95%) to be included in a very narrow interval ([50.8%,55.2%]).
- Given that there are 8,448,674 registered voters in Pennsylvania, how many of them do you think that your friends had to interview to make such a precise statement? Discuss this question with your neighbor for 1 minute.

iClicker time

- Given that there are 8,448,674 registered voters in Pennsylvania, how many of them do you think that your friends had to interview to make such a precise statement?

a) 2,000,000
b) 200,000
c) 20,000
d) 2,000

2000 is enough!

- What we are going to see in these lectures is that interviewing 2000 Pennsylvania voters is enough to make such a precise statement, provided 4 assumptions are satisfied.

Roadmap

1. Lay-out 4 assumptions.
2. Show that if those 4 assumptions are satisfied, then what your friend is saying is indeed correct. Along the way, learn more general lessons about expectation estimation, hypothesis testing, and confidence intervals.
3. Critically assess those 4 assumptions. Relatedly, discuss potential explanations for why polls failed to predict the outcome of the last US presidential election. Briefly present job opportunities for econ majors in polling firms.

Part 1: the 4 magical assumptions

Assumption 1: your friend has access to some register including the contact details of all voters in Pennsylvania

- Plausible?
- Instead, to which large register with people's contact details do you think polling firms have access?
- For now, let's assume that actually, the firm has access to this ideal register.
- Assumption 1 is called: sampling from the whole population.

Assumption 2: your friend randomly drew the sample of 2000 voters he/she is going to interview out of this register.

- Plausible?
- Assumption 2 is called: random sampling.

Assumption 3: when he contacts them, the 2000 voters all answer your friend.

- Plausible?
- Assumption 3 is called: no nonrespondents.

Assumption 4: when they respond to your friend's question: "Are you willing to vote for Hillary Clinton?", the 2000 voters respond truthfully.

- Plausible?
- Assumption 4 is called: truthful responses.

Fundamental result of polling theory:

- If Assumptions 1 to 4 are satisfied, then your friend's answer to the question "What is the fraction of the Pennsylvania electorate that wants to vote for Hillary Clinton?" is correct!

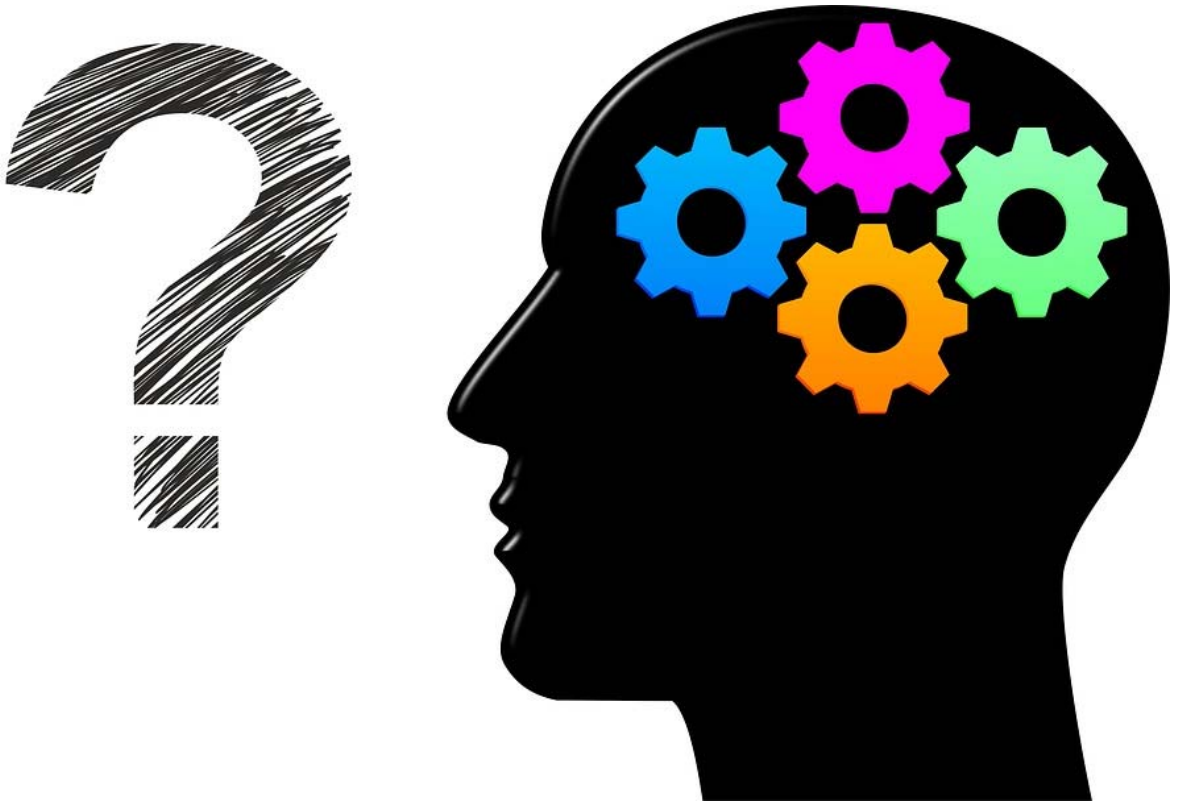
What you need to remember

- Polls/surveys use answers from a sample of a large population to infer the answers of people in the large population.
- Powerful method: instead of having to interview, say, the whole Pennsylvania population, we can learn useful things on that population by interviewing only 2000 Pennsylvania residents.
- However, polls/surveys can teach you something useful about the large population only if:
  - You can draw the sample of the poll/survey from the large population.
  - You draw the sample of the poll/survey randomly from the large population.
  - All the people you draw agree to answer the survey.
  - All of them answer your questions truthfully.

Roadmap

1. Lay-out 4 assumptions.
2. Show that if those 4 assumptions are satisfied, then what your friend is saying is indeed correct. Along the way, learn more general lessons about expectation estimation, hypothesis testing, and confidence intervals.
3. Critically assess those 4 assumptions. Relatedly, discuss potential explanations for why polls failed to predict the outcome of the last US presidential election. Present briefly job opportunities for econ majors in polling firms.

Part 2: why the 4 magical assumptions work!



Part 2.1: Defining our answer

First, a question needs an answer…

- Let y_1 be a number equal to 1 if the first voter in the Pennsylvania register wants to vote for Clinton, and to 0 otherwise. Similarly, for every k included between 1 and 8,448,674, let y_k be a number equal to 1 if person k wants to vote for Clinton, and to 0 otherwise.
- The y_ks are NOT random variables: whether person k wants to vote for Clinton or not is something deterministic, that person k knows. The y_ks are just numbers.
- Let $c = y_1 + y_2 + \cdots + y_{8,448,674}$, and let $p = \frac{c}{8,448,674}$.

What do c and p represent? Discuss this question with your neighbor for 1 minute.

iClicker time

- Let $c = y_1 + y_2 + \cdots + y_{8,448,674}$, and let $p = \frac{c}{8,448,674}$. What do c and p represent?

a) c is the sum of the wages of all Pennsylvania voters, and p is their average wage.

b) c is the number of Pennsylvania voters, and p is equal to 1.

c) c is the number of Pennsylvania voters who want to vote for Clinton, and p is the percentage of Pennsylvania voters who want to vote for Clinton.

… and p is our answer!

- $p = \frac{c}{8{,}448{,}674}$.
- $c = y_1 + y_2 + \cdots + y_{8{,}448{,}674}$ is the number of Pennsylvania voters that want to vote Clinton, so p is the proportion of Pennsylvania voters that want to vote for her.
- Everybody see that? If not, speak up now, or during sessions, this is key!

Now I remember, I am after p. Only 8,439,331 interviews left before I can compute it

Part 2.2: Is it enough to interview one voter?

Is it enough to interview one voter?

- In this section, we assume we randomly choose a number between 1 and 8,448,674, and then ask the voter with that number in our register if she wants to vote for Clinton.
- Is that person's answer enough for us to infer which percentage of the Pennsylvania electorate wants to vote for Clinton?
- To answer this question, we will study in great detail some properties of that person's answer.

Which values can that person's answer take?

- Assume we randomly choose 1 voter from the register. We interview that person, and ask her if she wants to vote for Clinton.
- Let Y_1 denote the answer of that person.
- What value can Y_1 take?

"Yes" or "No" which we respectively code as 1 and 0.

- What value can Y_1 take?
- Y_1 can be equal to "Yes", if the voter you drew from the register wants to vote for Clinton, and to "No" otherwise. Because it's easier to work with numbers than with words, we will say that $Y_1 = 1$ if the voter says yes, and $Y_1 = 0$ if the voter says no.

Is that person's answer deterministic or random?

- Assume we randomly choose 1 voter from the register , that we interview that voter, and ask her if she wants to vote for Clinton.
- Let Y_1 denote the answer of that voter.
- So far we have learned that Y_1 can either be equal to 0 or to 1.
- Is Y_1 a deterministic number or a random variable?

It is random, it depends on which voter we randomly draw from the register!

- Is Y_1 a deterministic number or a random variable?
- Y_1 is a random variable. Before we choose the voter we interview, we cannot know whether Y_1 will be equal to 1 or to 0. Assume that voter 32 in the register wants to vote for Clinton, but voter 37 is not willing to vote for Clinton. If you randomly choose to interview voter 32, $Y_1 = 1$. If you randomly choose to interview voter 37, $Y_1 = 0$. Y_1 is a random variable which depends on which voter you draw.

Reminder: definition of the expectation of a random variable.

- Let X be a random variable that can take K values $x_1, \dots, x_K$.
- $\mathbb{E}(X) = x_1 \mathbb{P}(X = x_1) + \cdots + x_K \mathbb{P}(X = x_K)$.
- Another way of writing exactly the same thing is

$$\mathbb{E}(X) = \sum_{k=1}^{K} x_k \mathbb{P}(X = x_k) .$$

- Expectation of random variable: weighted average of the possible values taken by that random variable weighted by their probabilities.

What is the expectation of that person's answer?

- Assume we randomly choose 1 voter from the register , that we interview that voter, and ask her if she wants to vote for Clinton.
- Let Y_1 denote the answer of that voter.
- So far we have learned that Y_1 is a random variable that can either be equal to 0 or to 1.
- What is the expectation of Y_1, $\mathbb{E}(Y_1)$? Discuss that question with your neighbor for 1mn.

iClicker time

- What is the expectation of Y_1, $\mathbb{E}(Y_1)$?

a) $\mathbb{E}(Y_1) = \mathbb{P}(Y_1 = 0) + \mathbb{P}(Y_1 = 1)$

b) $\mathbb{E}(Y_1) = \mathrm{c}$

c) $\mathbb{E}(Y_1) = \mathbb{P}(Y_1 = 1)$

d) $\mathbb{E}(Y_1) = \mathbb{P}(Y_1 = 0)$

$\mathbb{E}(Y_1) = \mathbb{P}(Y_1 = 1)$, the probability we draw a voter who wants to vote for Clinton

- So far we have learned that Y_1 is a random variable that can either be equal to 0 or to 1.
- What is the expectation of Y_1, $\mathbb{E}(Y_1)$?
- $\mathbb{E}(Y_1) = 0 \times \mathbb{P}(Y_1 = 0) + 1 \times \mathbb{P}(Y_1 = 1) = \mathbb{P}(Y_1 = 1)$.
- Expectation of Y_1: weighted average of the possible values taken by Y_1 (0 and 1) weighted by their probabilities ($\mathbb{P}(Y_1 = 0)$ and$\mathbb{P}(Y_1 = 1)$).
- General result: the expectation of a binary random variable (random variable either equal to 0 or 1) is equal to the probability this random variable is equal to 1. You need to remember that.

What is the probability we draw a voter who wants to vote for Clinton?

- Assume we randomly choose 1 voter from the register , that we interview that voter, and ask her if she wants to vote for Clinton.
- Let Y_1 denote the answer of that voter.
- So far we have learned that Y_1 is a random variable, whose expectation is $\mathbb{P}(Y_1 = 1)$.
- What is the value of $\mathbb{P}(Y_1 = 1)$? Try to find the answer with your neighbor. Consider first a simple example where the register of voters only has 6 people, 3 of whom want to vote for Clinton. If you randomly draw one of those 6 voters, what is the probability that the one you draw wants to vote for Clinton?

iClicker time

- What is the value of $\mathbb{P}(Y_1 = 1)$?

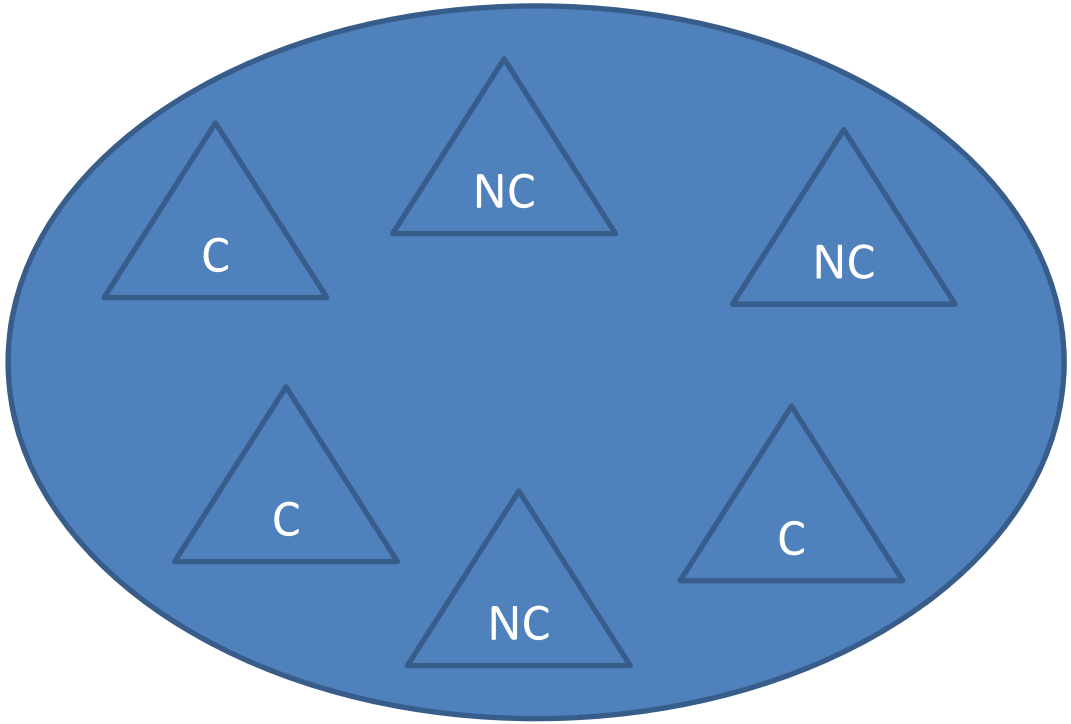
a) $\mathbb{P}(Y_1 = 1) = c$

b) $\mathbb{P}(Y_1 = 1) = p^2$

c) $\mathbb{P}(Y_1 = 1) = p$

d) $\mathbb{P}(Y_1 = 1) = 2p$

Answering this question with a simple example first.



- C="Clinton", NC="Not Clinton"
- In this example, what is the value of p? Of $\mathbb{P}(Y_1 = 1)$?

The probability we draw a voter who wants to vote for Clinton is equal to the % of voters who want to vote for Clinton!

- So far we have learned that Y_1 is a random variable, whose expectation is $\mathbb{P}(Y_1 = 1)$.
- What is the value of $\mathbb{P}(Y_1 = 1)$?
- $\mathbb{P}(Y_1 = 1) = p$! The voter we sample can be any of the $8{,}448{,}674$ voters in the register. $p\%$ of those voters are willing to vote Clinton. The probability that we sample a voter that wants to vote Clinton is p.

Reminder: estimators, unbiased estimators

- **Reminder: definition of an estimator.** An estimator is a function of the data we collect. Here, we collect Y_1, the answer of the voter we randomly draw to the question "do you want to vote for Clinton". Therefore, any function of Y_1 is an estimator: Y_1, $2Y_1$, Y_1^2…
- **Reminder: definition of unbiasedness.** Assume we would like to learn the value of a parameter. An unbiased estimator of that parameter is an estimator whose expectation is equal to that parameter.
- In our example, we would like to learn the value of p. Find an unbiased estimator of p. Discuss this question with your neighbor for 1 minute.

iClicker time

- Which of the following is an unbiased estimator of p?

a) Y_1

b) $2Y_1$

c) $\mathbb{E}(Y_1)$

Y_1 is an unbiased estimator of p

- So far we have learned that Y_1 is a random variable, whose expectation is p, the answer we seek.
- Y_1 is an unbiased estimator of p. Indeed, Y_1 is a function of the data we collect, Y_1. Moreover, $\mathbb{E}(Y_1) = p$.
- What that means is that when we average the value that Y_1 can take across all the voters we can randomly draw, that average is equal to p.

Are we likely to fall far from p when we use that person's answer to estimate it?

- Assume we randomly choose 1 voter from the register , that we interview that voter, and ask her if she wants to vote for Clinton.
- Let Y_1 denote the answer of that person.
- So far we have learned that Y_1 is an unbiased estimator of p.
- Are you likely to fall far from p when you use Y_1 to estimate it?

Yes, you will certainly fall far from p!

- So far we have learned that Y_1 is an unbiased estimator of p.
- Will you fall far from p when you use Y_1 to estimate p?
- Yes! Let's assume for a minute that $p = 0.5$. If you randomly draw a voter that wants to vote Clinton, $Y_1 = 1$, so you overestimate p by 0.5. If you randomly draw a voter that does not want to vote Clinton, $Y_1 = 0$, so you underestimate p by 0.5. Y_1 is an unbiased but imprecise estimator of p.
- Unbiasedness only means that your estimator is equally likely to over- or underestimate your target parameter. Does not make systematic mistakes in 1 direction. However, can still make large mistakes.

What's a good measure of the average mistake we make when using an estimator?

- Assume we randomly choose 1 voter from the register , that we interview that voter, and ask her if she wants to vote for Clinton.
- Let Y_1 denote the answer of that person.
- So far we have learned that Y_1 is an unbiased estimator of p, but using it can lead us to make large mistakes.
- How can we measure the average mistake we make when we use Y_1 to estimate p?

Reminder: the variance of a random variable.

- Let X be a random variable.
  $\mathrm{V}(X) = \mathbb{E}\big((X - \mathbb{E}(X))^2\big)$.
- Variance of a random variable: expectation of the square of the difference between X and its expectation.
- The standard deviation of X is
  $\mathrm{sd}(X) = \sqrt{\mathrm{V}(X)}$.

The variance of our estimator.

- So far we have learned that Y_1 is an unbiased estimator of p, but using it will lead us to make large mistakes.
- How can we measure the average mistake we make when we use Y_1 to estimate p?
- $\mathbb{E}(|Y_1-p|)$, or $\mathbb{E}(\ (Y_1-p)^2)$. First: average distance between our estimator and the target parameter. Second: average squared distance between our estimator and the target parameter. Second measure more commonly used, because it has nice properties.
- Notice that $\mathbb{E}\big((Y_1-p)^2\big) = \mathbb{E}\big((Y_1-E(Y_1))^2\big) = \mathrm{V}(Y_1)$.
- **$\mathbf{V}(Y_1)$ is a measure of the average mistake we make when using Y_1 to estimate p.**

A useful reminder: property 1 of the variance (P1Var)

- Let X be a random variable.
  $\mathrm{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.
- Variance of a random variable: expectation of the random variable squared minus the square of its expectation.
- Use this formula to compute $\mathrm{V}(Y_1)$. Discuss this question with your neighbor for 1 minute.

iClicker time

- Which answer is correct?

a) $\mathrm{V}(Y_1) = 1 - p$

b) $\mathrm{V}(Y_1) = p(1 - p)$

c) $\mathrm{V}(Y_1) = 1 - p^2$

d) $\mathrm{V}(Y_1) = 0$

e) $\mathrm{V}(Y_1) = 1$

The variance of our estimator is $p(1-p)$

- Can you compute $\mathrm{V}(Y_1)$?
- $\mathrm{V}(Y_1)$

$= \mathbb{E}\left(Y_1{}^2\right) - \mathbb{E}(Y_1)^2$

$= \mathbb{E}(Y_1) - \mathbb{E}(Y_1)^2$

$= p - p^2$

$= p(1-p)$

Why is this true?
Why is this true?
Why is this true?

What you need to remember

- If we randomly choose one Pennsylvanian voter and ask her if she wants to vote for Clinton:
  - That person's answer, Y_1, is an unbiased estimator of p, the percentage of Pennsylvania voters that want to vote for Clinton. Y_1 does not systematically over- or under-estimate p.
  - However, the variance of that estimator is $p(1-p)$, which is quite large for values of p close to 0.5 (in practice, we expect p to be close to 0.5).
  - Therefore Y_1 is not a precise estimator of p.

Part 2.3: Can we improve things by interviewing two voters?

Can we improve things by interviewing two voters?

- In this section, we are going to assume we interview 2 voters, and ask them if they want to vote for Clinton.
- By using the answers of these two people, can we build a better estimator p?

The 2 voters are chosen with replacement

- First, we randomly choose a number between 1 and 8,448,674. We interview the voter with that number in our register and ask her if she wants to vote for Clinton. Let Y_1 denote her answer.
- Second, we randomly choose another number between 1 and 8,448,674. We interview the voter with that number in our register and ask her if she wants to vote for Clinton. Let Y_2 denote her answer.
- **The voters are chosen with replacement**: if the first number we randomly chose was 12,235, we can still randomly choose 12,235 the second time. If that happens (very low probability), we do not interview voter 12,235 a 2nd time, we just let $Y_2 = Y_1$ (we already know whether that voter wants to vote for Clinton, no need to bother her again to ask her).

What is the probability distribution of Y_2?

- Assume we randomly choose with replacement 2 voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let Y_1 and Y_2 denote the answer of these two voters.
- What value can Y_2 take?
- Is Y_2 deterministic or random?
- What is the expectation of Y_2?
- What is its variance?

Y_1 and Y_2 follow the same distribution.

- $Y_2 = 1$ if second voter we draw wants to vote Clinton, $Y_2 = 0$ otherwise.
- Y_2 is a random variable. Before we draw the 2nd voter and interview her, we cannot know whether Y_2 will be equal to 1 or to 0.
- $\mathbb{E}(Y_2) = 0 \times \mathbb{P}(Y_2 = 0) + 1 \times \mathbb{P}(Y_2 = 1) = \mathbb{P}(Y_2 = 1) = p.$
- The 2nd voter we choose can be any of the $8{,}448{,}674$ voters in Pennsylvania. p% of voters are willing to vote Clinton. Therefore, the probability that 2nd voter we draw wants to vote Clinton is p.
- One can show that $\mathrm{V}(Y_2) = p(1\text{-}p)$.
- Y_1 and Y_2 have the same expectation, same variance, and same probability distribution. We say they are **identically distributed**.
- **That's because the lottery that gives rise to Y_1 (randomly choose one voter out of the 8,448,674 in the register) is exactly the same as that giving rise to Y_2.**

Reminder: definition of independent variables.

- Formal definition:
  - Let X be a random variable that can take K values $x_1, \dots, x_K$.
  - Let Z be a random variable that can take K values $z_1, \dots, z_K$.
  - X and Z are independent iif for any (x_k, z_k), $\mathbb{P}(X = x_K, Z = z_K) = \mathbb{P}(X = x_K)\mathbb{P}(Z = z_K)$.
- Example: when you draw two fair coins, what is the probability that each coin gives you a "heads"? What is the probability that the two coins both give you a "heads"?
- Informal definition: X and Y are independent if knowing the value of X does not affect the likelihood that Y takes specific values.
- For instance, if you know that the first coin gave you a heads, the probability that the second coin will give you a heads is still ½.
- On the other hand, if you know that it rained yesterday, that makes it more likely that it will also rain today.

Are Y_1 and Y_2 dependent or independent?

- Assume we randomly choose with replacement 2 voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let Y_1 and Y_2 denote the answer of these two voters.
- So far we have learned that Y_1 and Y_2 are identically distributed.
- Are Y_1 and Y_2 dependent or independent?

Y_1 and Y_2 are independent

- So far we have learned that Y_1 and Y_2 are identically distributed.
- Are Y_1 and Y_2 dependent or independent?
- Y_1 and Y_2 are independent. If the first voter we choose wants to vote Clinton, the second voter we choose is not more or less likely to want to vote Clinton.
- Y_1 and Y_2**: independent and identically distributed (iid).**
- Independence: the lottery that gives rise to Y_1 (randomly choose one voter out of the 8,448,674 in the register) is conducted independently from that giving rise to Y_2, just as when you independently toss two coins.

A useful reminder: properties 1 and 2 of the expectation

- P1Expectation: If X and Z are two random variables, then $\mathbb{E}(X + Z) = \mathbb{E}(X) + \mathbb{E}(Z)$.
- The expectation of the sum of two random variable is the sum of the expectation of these two random variables.
- P2Expectation: If X is a random variable, and a is a real number, then $\mathbb{E}(aX) = a\mathbb{E}(X)$.
- Watch out: P2expectation is only true if a is a real number. If X and Z are two random variables, usually $\mathbb{E}(XZ) \neq \mathbb{E}(X)\mathbb{E}(Z)$.

Can you find an unbiased estimator of p involving both Y_1 and Y_2?

- Assume we randomly choose with replacement 2 voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let Y_1 and Y_2 denote the answer of these two voters.
- So far we have learned that Y_1 and Y_2 are iid random variables whose expectation is p.
- Can you find an unbiased estimator of p involving both Y_1 and Y_2? Reminder: when we observed only Y_1, our unbiased estimator was Y_1. Discuss this question with your neighbor during 2mns.

iClicker time

- Which of the following estimators is an unbiased estimator of p?

a) $Y_1 - Y_2$

b) $Y_1 - 0.5Y_2$

c) $0.5Y_1 + 0.5Y_2$

$0.5Y_1 + 0.5Y_2$ is unbiased estimator of p

- Y_1 and Y_2: iid random variables whose expectation is p.
- Can you find an unbiased estimator of p involving both Y_1 and Y_2?
- $0.5Y_1 + 0.5Y_2$ is an unbiased estimator of p.
- $\mathbb{E}(0.5Y_1 + 0.5Y_2)$

$= \mathbb{E}(0.5Y_1) + \mathbb{E}(0.5Y_2)$ **Why is this true?**

$= 0.5\mathbb{E}(Y_1) + 0.5\mathbb{E}(Y_2)$ **Why is this true?**

$= 0.5p + 0.5p$ **Why is this true?**

$= p.$

- Can you find another unbiased estimator?

$2Y_1 - Y_2$ is other unbiased estimator of p

- Y_1 and Y_2 are iid random variables whose expectation is p.
- Can you find another unbiased estimator of p involving both Y_1 and Y_2?
- $2Y_1 - Y_2$ is an unbiased estimator of p.
- $\mathbb{E}(2Y_1 - Y_2)$

$= \mathbb{E}(2Y_1) + \mathbb{E}(-Y_2)$ **Why is this true?**

$= 2\mathbb{E}(Y_1) - \mathbb{E}(Y_2)$ **Why is this true?**

$= 2p - p$ **Why is this true?**

$= p.$

Any weighted sum of Y_1 and Y_2 is an unbiased estimator of p.

- Y_1 and Y_2: iid random variables whose expectation is p.
- For any real number x, $xY_1+(1-x)Y_2$ is an unbiased estimator of p:
- $\mathbb{E}(xY_1+(1-x)Y_2)$
$= \mathbb{E}(xY_1) + \mathbb{E}((1-x)Y_2)$
$= x\mathbb{E}(Y_1) + (1-x)\mathbb{E}(Y_2)$
$= xp + (1-x)p$
$= p.$

Why is this true?

Why is this true?

Why is this true?

- $xY_1 + (1-x)Y_2$: **linear unbiased estimator of p.** Linear function of Y_1 and Y_2, and unbiased.

Which criterion should we use to decide which unbiased estimator of p is the best?

- Assume we randomly choose with replacement 2 voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let Y_1 and Y_2 denote the answer of these two voters.
- So far we have learned that any weighted sum of Y_1 and Y_2 is an unbiased estimator of p.
- Which criterion should we use to assess which of all those unbiased estimators of p is the best? Discuss this question with your neighbor during 1mn.

iClicker time

- Which criterion should we use to assess which of all those unbiased estimators of p is the best?

a) The best estimator is the one with the highest expectation.

b) The best estimator is the one with the lowest variance.

c) The best estimator is the one with the lowest median.

We should choose the estimator with the lowest variance!

- So far we have learned that for any real number x, $xY_1+(1-x)Y_2$ is unbiased estimator of p.
- Which criterion should we use to decide which of all those estimators of p we should use?
- We should pick the estimator with the lowest variance! That's the estimator that will lead us to make the least mistakes when using it to proxy p.

A useful reminder: the covariance between two random variables.

- Let X and Z be two random variables.
- $cov(X, Z) = \mathbb{E}\big[\big(X - \mathbb{E}(X)\big)\big(Z - \mathbb{E}(\mathrm{Z})\big)\big]$.
- When $X > \mathbb{E}(X)$:
  - if $\mathrm{Z} > \mathbb{E}(\mathrm{Z})$ (X and Z "agree") then $\big(X - \mathbb{E}(X)\big)\big(\mathrm{Z} - \mathbb{E}(\mathrm{Z})\big) > 0$,
  - if $\mathrm{Z} < \mathbb{E}(\mathrm{Z})$ (X and Z "disagree") then $\big(X - \mathbb{E}(X)\big)\big(\mathrm{Z} - \mathbb{E}(\mathrm{Z})\big) < 0$.
- When $X < \mathbb{E}(X)$:
  - if $\mathrm{Z} < \mathbb{E}(\mathrm{Z})$ (X and Z "agree") then $\big(X - \mathbb{E}(X)\big)\big(\mathrm{Z} - \mathbb{E}(\mathrm{Z})\big) > 0$
  - if $\mathrm{Z} > \mathbb{E}(\mathrm{Z})$ (X and Z "disagree") then $\big(X - \mathbb{E}(X)\big)\big(\mathrm{Z} - \mathbb{E}(\mathrm{Z})\big) < 0$.
- When X and Z have a high probability to "agree" (both are above or below their expectation), then $\big(X - \mathbb{E}(X)\big)\big(\mathrm{Z} - \mathbb{E}(\mathrm{Z})\big)$ has a high probability of being positive, and then $cov(X, Z) = \mathbb{E}\big[\big(X - \mathbb{E}(X)\big)\big(\mathrm{Z} - \mathbb{E}(\mathrm{Z})\big)\big] > 0$.
- On the other hand, when X and Z have high probability to "disagree" (one is above its expectation, the other one is below it), covariance negative.

The covariance between rain and sunshine

- Let X and Z denote the amount of rain and the number of sunshine hours that we will have tomorrow in Santa Barbara.
- Do you think that $cov(X, Z) > 0$, $cov(X, Z) < 0$, or $cov(X, Z) = 0$? Discuss this during 1mn with your neighbor.

iClicker time

- Let X and Z denote the amount of rain and the number of sunshine hours that we will have tomorrow in Santa Barbara.
- Do you think that $cov(X,Z) > 0$, $cov(X,Z) < 0$, or $cov(X,Z) = 0$? Discuss this during 1mn with your neighbor.

a) $cov(X,Z) > 0$

b) $cov(X,Z) < 0$

c) $cov(X,Z) = 0$

Rain is negatively correlated with sunshine hours!

- Days with more rain than the average amount of rain per day also tend to have a lower number of sunshine hours than the average.
- If $X > \mathbb{E}(X)$ (more rain than the average), then it is likely that $Z < \mathbb{E}(Z)$ (less sunshine hours than the average).
- Conversely, if $X < \mathbb{E}(X)$ (less rain than the average), then it is likely that $Z > \mathbb{E}(Z)$ (more sunshine hours than the average).
- Therefore, $cov(X, Z) < 0$.
- **The covariance between two variables measures whether they move in a similar or opposite direction.**

A useful reminder: properties 1, 2, and 3 of the covariance

- P1Cov: $\text{cov}(X, Z) = \mathbb{E}(XZ) - \mathbb{E}(X)\mathbb{E}(Z)$.
- P1Cov is very useful to compute the covariance between two random variables. You just need to compute the expectation of each variable, and the expectation of their product.
- P2Cov: if X and Z are independent, $\text{cov}(X, Z) = 0$.
- P2Cov is intuitive: if X and Z are independent, they do not move in a similar or opposite direction, so $\text{cov}(X, Z) = 0$.
- P3Cov: $-sd(X)sd(Z) \leq \text{cov}(X, Z) \leq sd(X)sd(Z)$
- Let $\rho_{XZ} = \frac{\text{cov}(X,Z)}{sd(X)sd(Z)}$. ρ_{XZ} is called the correlation coefficient between X and Z.
- ρ_{XZ} must be included between two values. Which are those two values? Discuss this with your neighbour for 1mn.

iClicker time

- Let $\rho_{XZ} = \frac{\mathrm{cov}(X,Z)}{sd(X)sd(Z)}$. We have

a) $-1 \leq \rho_{XZ} \leq 1$

b) $0 \leq \rho_{XZ} \leq 1$

c) $0 \leq \rho_{XZ} \leq 2$

$-1 \leq \rho_{XZ} \leq 1$!

- $\rho_{XZ} = \frac{\text{cov}(X,Z)}{sd(X)sd(Z)}$.
- P3Cov: $-sd(X)sd(Z) \leq \text{cov}(X,Z) \leq sd(X)sd(Z)$
- Therefore, $-1 \leq \rho_{XZ} \leq 1$.
- $\text{cov}(X,Z) > 0$ means that X and Z positively related (sunshine hours and temperature).
- $\text{cov}(X,Z) < 0$ means that X and Z positively related (sunshine hours and rain).
- But how can we assess the strength of the positive or negative relation between X and Z?
- By looking at ρ_{XZ}!
- $\rho_{XZ} = 1$: perfect positive correlation between X and Z: $Z = aX + b$ with $a > 0$.
- $\rho_{XZ} = -1$: perfect negative correlation between X and Z: $Z = aX + b$ with $a < 0$.
- $\rho_{XZ} = 0.7$: strong, but not perfect positive correlation between X and Z.

A useful reminder: property 2 of the variance

- P2Var: If X and Z are two random variables, and if a and b are two real numbers, then

$$\mathrm{V}(aX + bZ) = a^2\mathrm{V}(X) + b^2\mathrm{V}(Z) + 2ab \times cov(X, Z)$$

- Combining P2Var and P2Cov, we obtain that if X and Z are two **independent** random variables, and if a and b are two real numbers, then

$$\mathrm{V}(aX + bZ) = a^2\mathrm{V}(X) + b^2\mathrm{V}(Z)$$

Which value of x minimizes $\mathrm{V}(xY_1+(1-x)Y_2)$?

- After this little detour to learn about covariance and correlation coefficient, back to polling.
- Assume we randomly choose with replacement 2 voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let Y_1 and Y_2 denote the answer of these two voters.
- So far we have learned that any weighted sum of Y_1 and Y_2 is an unbiased estimator of p, and that we should choose the one with the lowest variance.
- Can you find the value of x such that $\mathrm{V}(xY_1+(1-x)Y_2)$ is minimized?

$x = 0.5$! Best estimator: average of Y_1 & Y_2

- Goal: minimize $\mathrm{V}(\, xY_1+(1-x)Y_2)$ wrt x.

- $\mathrm{V}(\, xY_1+(1-x)Y_2)$

$= x^2\mathrm{V}(Y_1) + (1-x)^2\mathrm{V}(Y_2)$

$= x^2p(1-p) + (1-x)^2p(1-p)$

$= p(1-p)(x^2 + (1-x)^2)$

$= p(1-p)(2x^2 - 2x + 1).$

Why is this true?

Why is this true?

- Differentiate wrt x:

$\partial_x \mathrm{V}(\, xY_1+(1-x)Y_2) = p(1-p)(4x-2).$

$\mathrm{V}(\, xY_1+(1-x)Y_2)$ minimized at $x = 0.5$!

- **$0.5Y_1 + 0.5Y_2$, average of Y_1 and Y_2, is best linear unbiased (BLU) estimator of p.**

Do we gain a lot by interviewing 2 voters instead of one?

- Assume we randomly choose with replacement 2 voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let Y_1 and Y_2 denote the answer of these two voters.
- So far we have learned that $0.5Y_1 + 0.5Y_2$ is the estimator of p we should use if we interview two voters.
- What is the variance of this estimator? How does it compare to the variance of Y_1, the estimator we would use if we only interviewed one voter?

Interviewing two voters instead of one divides the variance of our estimator by 2!

- So far we have learned that $0.5Y_1 + 0.5Y_2$ is the estimator of p we should use if we interview two voters.
- What is the variance of this estimator? How does it compare to the variance of Y_1, the estimator we would use if we had only interviewed one voter?
- $\mathrm{V}(0.5Y_1 + 0.5Y_2)$
$= 0.5^2\mathrm{V}(Y_1) + 0.5^2\mathrm{V}(Y_2)$
$= 0.5^2\mathrm{V}(Y_1) + 0.5^2\mathrm{V}(Y_1)$
$= 0.5\mathrm{V}(Y_1).$

Why is this true?

Why is this true?

- Interviewing 2 voters instead of 1 divides the variance of our estimator by 2!

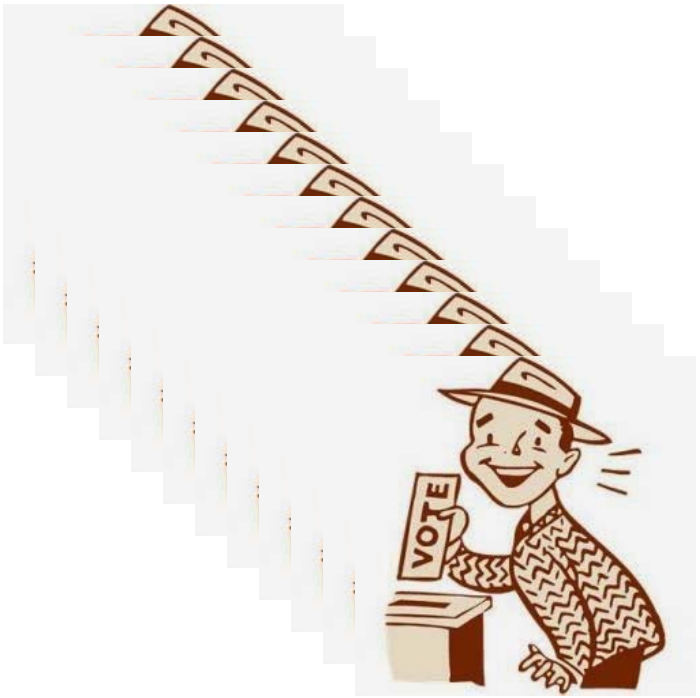
Why does variance of estimator diminishes when we go from 1 to 2 voters interviewed?

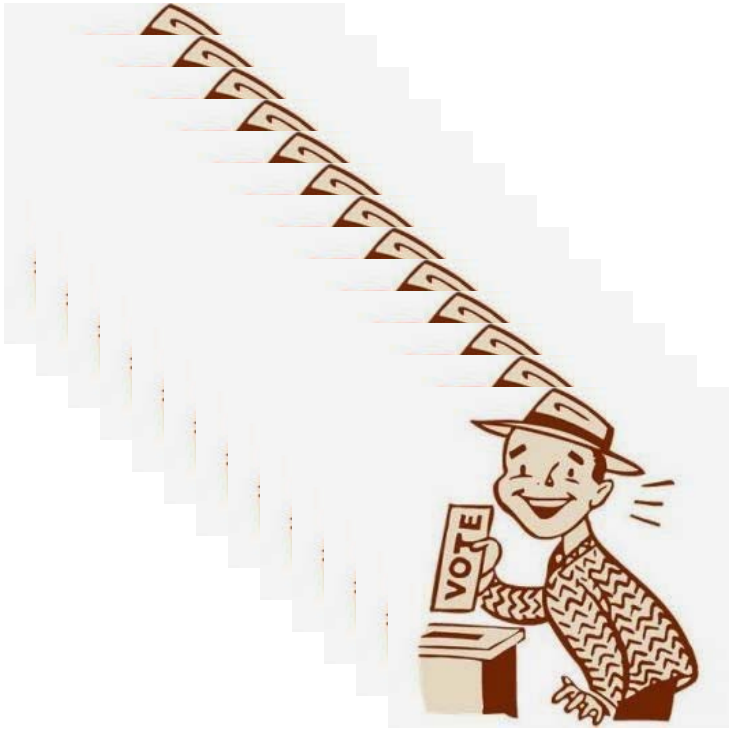
- Assume that $p = 0.5$.
- If we interview only one voter, our estimator is Y_1.
  - With probability 0.5, $Y_1 = 1$ (voter we randomly select = Clinton voter) => we overestimate p by 0.5.
  - With probability 0.5, $Y_1 = 0$ (voter we randomly select = Trump voter) => we underestimate p by 0.5.
- If we interview two voters, our estimator is $0.5Y_1 + 0.5Y_2$.
  - With probability 0.25, the 2 voters we randomly select = Clinton voters => $0.5Y_1 + 0.5Y_2 = 1$ => we overestimate p by 0.5.
  - With probability 0.25, the 2 voters we randomly select = Trump voters => $0.5Y_1 + 0.5Y_2 = 0$ => we underestimate p by 0.5.
  - With probability 0.5, we randomly select one Clinton voter and one Trump voter => $0.5Y_1 + 0.5Y_2 = 0.5$ => we do not over- or underestimate p.
- $0.5Y_1 + 0.5Y_2$ has a much lower probability of being far from p than Y_1.
- That's why its variance is much lower: variance = average mistake we make when use estimator as a proxy for p.

What you need to remember

- If we draw two voters from the register, and ask them if they want to vote for Clinton:
  - We can build many unbiased estimators of p using their two answers Y_1 and Y_2: for any real number x, $xY_1+(1-x)Y_2$ is an unbiased estimator of p.
  - We should choose the value of x which minimizes the variance of $xY_1+(1-x)Y_2$, our estimator of p.
  - This value is $x = 0.5$.
  - $0.5Y_1 + 0.5Y_2$ is the best linear unbiased estimator of p if we interview two voters.
  - The variance of $0.5Y_1 + 0.5Y_2$ is twice smaller than the variance of Y_1, the estimator we would have used if we had only interviewed one voter.
- Along the way, we have defined the covariance between two random variables, and the coefficient of correlation:
  - Covariance measures whether there is a positive or negative relation between two variables
  - Coefficient of correlation: normalized version of covariance, must be included between -1 and 1. 1: perfect positive correlation between the variables. -1 perfect negative correlation.

Part 2.3: Can we further improve things by interviewing many voters?





Can we improve things by interviewing many voters?

- In this section, we are going to assume we interview n voters, and ask them if they want to vote for Clinton.
- By using the answers of these n people, can we build a better estimator of the percentage of the Pennsylvania electorate that wants to vote for Clinton?

The voters are chosen with replacement

- First, we randomly choose a number between 1 and 8,448,674. We interview the voter with that number in our register and ask her if she wants to vote for Clinton. Let Y_1 denote her answer.
- Second, we randomly choose another number between 1 and 8,448,674. We interview the voter with that number in our register and ask her if she wants to vote for Clinton. Let Y_2 denote her answer.
- Third…
- … Finally, we randomly choose another number between 1 and 8,448,674. We interview the voter with that number in our register and ask her if she wants to vote for Clinton. Let Y_n denote her answer.
- **The voters are chosen with replacement**: if the first number we randomly chose was 12,235, we can still randomly choose 12,235 the second time.

What is the probability distribution of the random variables $Y_1, Y_2, \ldots, Y_n$?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- What value can $Y_1, Y_2, \ldots, Y_n$ take?
- What is the expectation of these variables?
- What is their variance?

$Y_1, Y_2, \ldots, Y_n$ follow the same distribution.

- For every number i included between 1 and n, $Y_i = 1$ if the ith voter we draw wants to vote Clinton, $Y_i = 0$ otherwise.
- $\mathbb{E}(Y_i) = 0 \times \mathbb{P}(Y_i = 0) + 1 \times \mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = 1) = p$.
- The ith voter we choose can be any of the $8{,}448{,}674$ voters in Pennsylvania. $p\%$ of voters are willing to vote Clinton. Therefore, the probability that the ith voter we choose wants to vote Clinton is p.
- One can show that $\mathrm{V}(Y_i) = p(1\text{-}\, p)$.
- All the Y_i have the same expectation, the same variance, and same probability distribution. We say they are **identically distributed**.

Are $Y_1, Y_2, \ldots, Y_n$ dependent or independent?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- So far we have learned that $Y_1, Y_2, \ldots, Y_n$ are identically distributed.
- Are $Y_1, Y_2, \ldots, Y_n$ dependent or independent?

$Y_1, Y_2,\ldots, Y_n$ are independent

- Are $Y_1, Y_2,\ldots, Y_n$ dependent or independent?
- $Y_1, Y_2,\ldots, Y_n$ are independent. If the third voter we draw wants to vote Clinton, we are not more or less likely to draw a fourth voter that wants to vote Clinton.
- $Y_1, Y_2,\ldots, Y_n$**: independent and identically distributed (iid).**
- Independence: the lottery that gives rise to Y_3 (randomly choose one voter out of the 8,448,674 in the register) is conducted independently from that giving rise to Y_n, just as when you independently toss two coins.

Reminder: property 3 of expectation.

- P3Expectation: If $X_1, X_2, \ldots, X_n$ are n random variables, then $\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots \mathbb{E}(X_n)$.
- Another way of writing exactly the same thing: $\mathbb{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathbb{E}(X_i)$

Properties 1, 2, and 3 of summation.

- P1Sum: For any real number x,
  $\sum_{i=1}^{n} x = x + x + \cdots + x = nx$.
- Summation of n times the same number is equal to n times that number.
- P2Sum: For any sequence of real numbers $(x_1, x_2, \ldots, x_n)$ and for any real number q,
  $\sum_{i=1}^{n} qx_i = qx_1 + qx_2 + \cdots + qx_n = q(x_1 + x_2 + \cdots + x_n) = q\sum_{i=1}^{n} x_i$.
- P3Sum: For any sequences of real numbers $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$,
  $\sum_{i=1}^{n}(x_i + y_i) = x_1 + y_1 + x_2 + y_2 + \cdots + x_n + y_n = x_1 + x_2 + \cdots + x_n + y_1 + y_2 + \cdots + y_n = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$.
- Summation of the sum of two numbers is the sum of the summations of these numbers.

Property 1 of double summations.

- P2DoubleSum: For any sequences of real numbers $(x_1, x_2, \ldots, x_n)$ and $(z_1, z_2, \ldots, z_m)$,

$$\sum_{i=1}^{n} \sum_{\mathrm{j}=1}^{m} x_i z_j = \left(\sum_{i=1}^{n} x_i\right) \times \left(\sum_{\mathrm{j}=1}^{m} z_j\right).$$

Can you find an unbiased estimator of p involving Y_1 , Y_2,…, Y_n?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let Y_1, Y_2,…, Y_n denote the answers of these voters.
- So far we have learned that Y_1, Y_2,…, Y_n are iid random variables whose expectation is p.
- Can you find an unbiased estimator of p involving Y_1 , Y_2,…, Y_n? Reminder: when we observed only Y_1 and Y_2, the best unbiased estimator was $\frac{1}{2}(Y_1 + Y_2)$. Discuss this question with your neighbor during 2mns.

iClicker time

- Which of the following estimators is an unbiased estimator of p?

a) $\frac{1}{2}\sum_{i=1}^{n} Y_i$

b) $\frac{2}{n}\sum_{i=1}^{n} Y_i$

c) $\frac{1}{n}\sum_{i=1}^{n} Y_i$

$\frac{1}{n}\sum_{i=1}^{n} Y_i$ is unbiased estimator of p

- $Y_1, Y_2, \ldots, Y_n$: iid variables, expectation p. Unbiased estimator of p?
- $\frac{1}{n}\sum_{i=1}^{n} Y_i$ is an unbiased estimator of p.
- $\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$

$$= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} Y_i\right)$$ **Why is this true?**

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i)$$ **Why is this true?**

$$= \frac{1}{n}\sum_{i=1}^{n} p$$ **Why is this true?**

$$= \frac{1}{n}np$$ **Why is this true?**

$$= p$$

What is $\frac{1}{n}\sum_{i=1}^{n} Y_i$?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- So far we have learned that $\frac{1}{n}\sum_{i=1}^{n} Y_i$ is an unbiased estimator of p.
- What is $\frac{1}{n}\sum_{i=1}^{n} Y_i$ in plain English?

$\frac{1}{n}\sum_{i=1}^{n} Y_i$ is the % of voters who say they want to vote for Clinton in our sample.

- $\frac{1}{n}\sum_{i=1}^{n} Y_i$ is an unbiased estimator of p.
- What is $\frac{1}{n}\sum_{i=1}^{n} Y_i$ in plain English?
- $\frac{1}{n}\sum_{i=1}^{n} Y_i$ is just the % of voters that want to vote for Clinton, among those we interview. $\sum_{i=1}^{n} Y_i$ counts number of voters who want to vote Clinton. Divided by n, this gives the %.
- Intuitive: to estimate p, the % of voters that want to vote Clinton among **all** Pennsylvania voters, we use the % of voters that want to vote for Clinton among **the sample** of Pennsylvania voters we interview.
- **Random sampling** of our sample ensures it is **representative** of the entire population of voters in Pennsylvania.

A useful reminder: properties 3 and 4 of the variance

- P3Var: If X_1,…, X_n are n independent random variables, then $\mathrm{V}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{V}(X_i)$
- P4Var: If X is a random variable and a and b are real numbers, then $\mathrm{V}(aX + b) = a^2 V(X)$.

What is the variance of $\frac{1}{n}\sum_{i=1}^{n} Y_i$?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- So far we have learned that $\frac{1}{n}\sum_{i=1}^{n} Y_i$, the % of voters that want to vote for Clinton in our sample, is an unbiased estimator of p.
- What is the variance of $\frac{1}{n}\sum_{i=1}^{n} Y_i$? Try to find the answer with your neighbor, you have 2 mns.

iClicker time

- What is the variance of $\frac{1}{n}\sum_{i=1}^{n} Y_i$?

a) $\mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\mathrm{V}(Y_1)$

b) $\mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\mathrm{V}(Y_1)$

c) $\mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \mathrm{V}(Y_1)$

$$\mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)=\frac{1}{n}\mathrm{V}(Y_1)$$

- $\mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$

$= \frac{1}{n^2}\mathrm{V}\left(\sum_{i=1}^{n} Y_i\right)$ **Why is this true?**

$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}(Y_i)$ **Why is this true?**

$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}(Y_1)$ **Why is this true?**

$= \frac{1}{n^2} n\mathrm{V}(Y_1)$ **Why is this true?**

$= \frac{1}{n}\mathrm{V}(Y_1).$

- Many linear unbiased estimators of p. E.g.: $nY_1 - \sum_{i=2}^{n} Y_i$.
- However, none has a variance lower than $\frac{1}{n}\mathrm{V}(Y_1)$.

$\frac{1}{n}\sum_{i=1}^{n} Y_i$ is the BLU estimator of p, when we interview n voters

Can we estimate the variance of our estimator?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- So far we have learned that $\frac{1}{n}\sum_{i=1}^{n} Y_i$, % of voters that want to vote for Clinton in our sample, is an unbiased estimator of p. Its variance is $\frac{1}{n}\mathrm{V}(Y_1)$.
- Can we estimate the variance of $\frac{1}{n}\sum_{i=1}^{n} Y_i$? Would be useful to assess how precise our estimator is.

Reminder: estimating the variance of the average of random variables.

- Let $X_1, X_2, \ldots, X_n$ be n iid random variables.
- Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote their average.
- One has $\mathrm{V}(\bar{X}) = \frac{V(X_1)}{n}$. To estimate $\mathrm{V}(\bar{X})$, need to estimate $V(X_1)$.
- To estimate $V(X_1)$, we use $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
- Intuition:
  - To estimate $\mathbb{E}(X_1)$, we use $\frac{1}{n}\sum_{i=1}^{n} X_i$: we replace the expectation by the sample average of X_i.
  - $V(X_1) = \mathbb{E}\left((X_1 - \mathbb{E}(X_1))^2\right)$, so to estimate it we use the sample average of $(X_i - \bar{X})^2$.
- Accordingly, we use $\hat{V}(\bar{X}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ to estimate $\mathrm{V}(\bar{X})$.
- **Definition:** Let $X_1, X_2, \ldots, X_n$ be n iid random variables. $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the **sample variance** of those random variables.

We can use $\frac{\bar{Y}(1-\bar{Y})}{n}$ to estimate $\mathrm{V}(\bar{Y})$

- $Y_1, Y_2, \ldots, Y_n$ iid with expectation p. $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$: unbiased estimator of p, whose variance is $\frac{1}{n}\mathrm{V}(Y_1)$.
- How can we estimate $\mathrm{V}(\bar{Y})$?
- Following reminder, we use $\hat{V}(\bar{Y}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n}$.
- As the $Y_1, Y_2, \ldots, Y_n$ are binary, one can show $\hat{V}(\bar{Y}) = \frac{\bar{Y}(1-\bar{Y})}{n}$.
- Convenient: knowing $\bar{Y}$ and n is sufficient to compute $\hat{V}(\bar{Y})$.

What happens to the variance of our estimator when the number of voters we interview grows?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- So far we have learned that $\frac{1}{n}\sum_{i=1}^{n} Y_i$, the % of voters that want to vote for Clinton in our sample, is an unbiased estimator of p, whose variance is $\frac{1}{n}\mathrm{V}(Y_1)$.
- What happens to variance of $\frac{1}{n}\sum_{i=1}^{n} Y_i$ when n grows?

It goes to 0! With an infinite sample, our estimator becomes equal to p.

- $\frac{1}{n}\sum_{i=1}^{n} Y_i$: unbiased estimator of p, whose variance is $\frac{1}{n}\mathrm{V}(Y_1)$
- What happens to the variance of $\frac{1}{n}\sum_{i=1}^{n} Y_i$ when n grows?
- $\frac{1}{n+1}\mathrm{V}(Y_1) < \frac{1}{n}\mathrm{V}(Y_1)$. Each time we add one voter to sample, variance of estimator decreases.
- $\lim_{n\to+\infty} \frac{1}{n}\mathrm{V}(Y_1) = 0$. If we interview infinity of voters, variance of our estimator goes to 0.
- Variance of our estimator is the mistake it makes when estimating p.
- **=> with an infinite sample, our estimator no longer makes any mistake in estimating p, it becomes equal to p:**

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i = p\,.$$

- Law of large numbers.

Why does $\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i = p$?

- Assume that $p = 0.5$. 50% of Clinton voters and 50% of Trump voters in the population.
- If we interview two voters:
  - With probability 0.25, the 2 voters we randomly select = Clinton voters => our sample is very different from population.
  - With probability 0.25, the 2 voters we randomly select = Trump voters => sample again very different from population.
  - Overall, 0.5 probability of getting sample where percentage of Clinton voters is very different from that in population.
- If we interview 1000 voters:
  - probability of interviewing only Clinton voters = $0.5^{1000} \sim 0$.
  - probability of interviewing only Trump voters = $0.5^{1000} \sim 0$.
- With a sample of 1000 voters, much less likely to draw sample where percentage of Clinton voters is very different from that in population.
- When the sample size goes to infinity, 0 probability to draw sample where percentage of Clinton voters is different from that in population.

What you need to remember

- Assume we draw n voters from the register, and ask them if they want to vote for Clinton. Let Y_1, Y_2,…, Y_n denote their answers.
- The average of their answers, $\bar{Y}$, is an unbiased estimator of p, the % of Pennsylvania voters that want to vote for Clinton.
- Randomly drawing the sample of voters we interview ensures it is **representative** of the Pennsylvania electorate.
- $V(\bar{Y}) = \frac{1}{n}\mathrm{V}(Y_1)$. You need to remember how to prove that.
- $V(\bar{Y})$ goes to 0 when n grows.
- If we were to interview an infinity of voters, $\bar{Y}$ would become equal to p: law of large numbers.
- We can use $\hat{V}(\bar{Y}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}$ to estimate the variance of $\bar{Y}$. As the Y_1, Y_2,…, Y_n are binary, one can show $\hat{V}(\bar{Y}) = \frac{\bar{Y}(1-\bar{Y})}{n}$.

We could learn p by interviewing an infinite random sample of voters. Do we want to do that?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- So far we have learned that $\frac{1}{n}\sum_{i=1}^{n} Y_i$, the % of voters that want to vote for Clinton in our sample, is an unbiased estimator of p, and will become equal to p if we interview infinity of voters.
- Do we want to do that?

No!

- $\frac{1}{n}\sum_{i=1}^{n} Y_i$ becomes exactly equal to p if we interview an infinity of voters.
- Do we want to do that?
- Nope, we are lazy and we have a life, unlike this guy!

- Your smart friend's response relies on an even smarter tool than the law of large numbers: the central limit theorem.

Reminder: The central limit theorem.

- Let $X_1, X_2, \ldots, X_n$ be n iid random variables.
- Let m denote their expectation.
- Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote their average.
- Let $\hat{V}\left(\bar{X}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ denote the estimator of $\mathrm{V}(\bar{X})$.
- CLT: if n is larger than 100, then $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}}$ approximately follows a normal distribution with expectation 0 and variance 1.
- $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}}$: difference between average of the variables and their expectation, divided by estimator of standard deviation of $\bar{X}$.

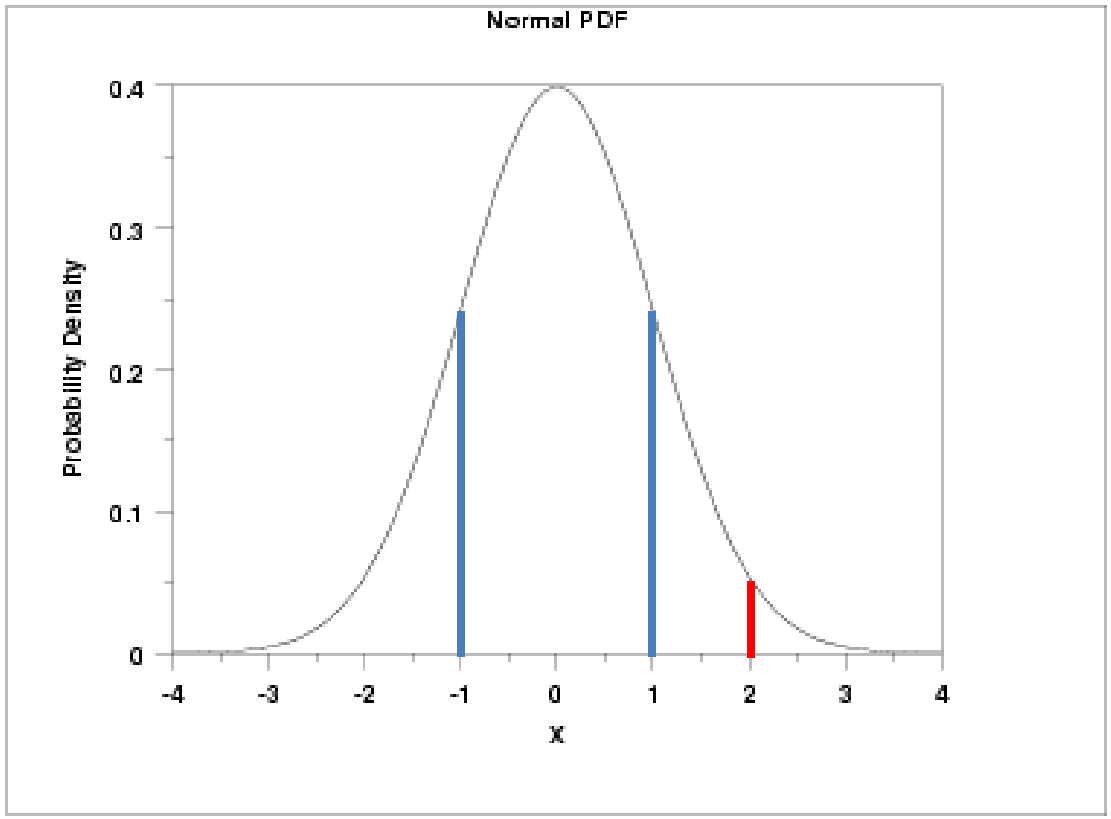
Reminder: the N(0,1) distribution.

- A random variable X follows a normal distribution with expectation 0 and variance 1 (N(0,1)) if for any real number x,

$$P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

- Its density function is $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

A graph of the density of a N(0,1) variable



- Area below curve and between blue lines: probability that X included between -1 and 1: quite large.
- Area below curve and to the right of the red line: probability that X greater than 2. Very small.
- **A N(0,1) random variable is more likely to be close to than far from 0.**

3 facts to remember about N(0,1) variables

- F1-N(0,1): A N(0,1) variable has a 90% probability of being included between -1.64 and 1.64.
- If you draw a N(0,1), it is possible but unlikely (10% chance) that you get a result below -1.64 or above 1.64.
- F2-N(0,1): A N(0,1) variable has a 95% probability of being included between -1.96 and 1.96.
- If you draw a N(0,1), it is possible but very unlikely (5% chance) that you get a result below -1.96 or above 1.96.
- F3-N(0,1): A N(0,1) variable has a 99% probability of being included between -2.57 and 2.57.
- If you draw a N(0,1), it is possible but very very unlikely (1% chance) that you get a result below -2.57 or above 2.57.

What is the distribution of $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ denote average of their answers.
- Let $\hat{V}(\bar{Y}) = \frac{\bar{Y}(1-\bar{Y})}{n}$ be the estimator of the variance of $\bar{Y}$.
- If $n \geq 100$, what is the distribution of $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$? Discuss this question with your neighbor during 2mns.

iClicker time

- If $n \geq 100$, what is the (approximate) distribution of $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$?

a) $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$ approximately follows a normal distribution with expectation 0 and variance 2

b) $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$ approximately follows a normal distribution with expectation 0 and variance 1

c) $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$ approximately follows a binomial distribution with parameters n and p.

If $n \geq 100$, $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$ follows the N(0,1) distribution

- What is the distribution of $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$?
- $Y_1, Y_2, \ldots, Y_n$ are iid, and their expectation is equal to p.
- Therefore, if $n \geq 100$ we can apply the CLT: $\frac{\bar{Y}-p}{\sqrt{\hat{V}(\bar{Y})}}$ follows N(0,1) distribution.
- Because the random variables $Y_1, Y_2, \ldots, Y_n$ are binary, $\hat{V}(\bar{Y}) = \frac{\bar{Y}(1-\bar{Y})}{n}$.
- Therefore, $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$ follows the N(0,1) distribution.

If we interview 2000 voters and find that $\bar{Y} = 0.53$, is it plausible that $p = 0.50$?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- So far we have learned that $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$ follows N(0,1) distribution.
- If we randomly choose 2000 voters and find that 1060 want to vote for Clinton, thus implying that $\bar{Y} = 0.53$, is it plausible that p, the percentage of the Pennsylvania electorate that wants to vote for Clinton, is equal to 0.50? Discuss this question with your neighbor during 2mns. Hint: plug the values of $\bar{Y}$, n, and p into $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$, and use the fact this quantity follows a N(0,1) distribution.

iClicker time

- If we randomly choose 2000 voters and find that 1060 want to vote for Clinton, thus implying that $\bar{Y} = 0.53$, is it plausible that p, the percentage of the Pennsylvania electorate that wants to vote for Clinton, is equal to 0.50?

a) Yes, this is plausible.

b) No, this is not plausible.

$p = 0.50$ incompatible with $\bar{Y} = 0.53$ if we interview 2000 voters.

- If we interview 2000 voters and find $\bar{Y} = 0.53$, plausible that $p = 0.50$?
- $n = 2000$ and $\bar{Y} = 0.53$. CLT says $\frac{0.53-p}{\sqrt{\frac{0.53(1-0.53)}{2000}}}$ follows N(0,1).
- Plugging $p = 0.50$ into this expression yields 2.688125.
- A N(0,1) random variable has very low probability of being that large.
- => implausible that $p = 0.50$, incompatible with the data we observe.
- We might be wrong: maybe share of Pennsylvania voters that want to vote for Clinton is equal to 0.50, but out of bad luck 1060 voters out of the 2000 we drew want to vote for Clinton. That's possible, exactly as it's possible to get 1060 heads when you flip a fair coin 2000 times.
- However, the CLT theorem tells us that such an unlucky draw, where the sample differs so much from the Pennsylvania electorate, is as likely to happen as drawing a N(0,1) random variable equal to 2.688125: very unlikely.

If we interview 200 voters and find that $\bar{Y} = 0.53$, is it plausible that $p = 0.50$?

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- So far we have learned that $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$ follows N(0,1) distribution.
- If we randomly choose 200 voters and find that 106 want to vote for Clinton, thus implying that $\bar{Y} = 0.53$, is it plausible that p, the percentage of the Pennsylvania electorate that wants to vote for Clinton, is equal to 0.50? Discuss this question with your neighbor during 2mns. Hint: plug the values of $\bar{Y}$, n, and p into $\frac{\bar{Y}-p}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$, and use the fact this quantity should follow a N(0,1) distribution.

iClicker time

- If we randomly choose 200 voters and find that 106 want to vote for Clinton, thus implying that $\bar{Y} = 0.53$, is it plausible that p, the percentage of the Pennsylvania electorate that wants to vote for Clinton, is equal to 0.50?

a) Yes, this is plausible.

b) No, this is not plausible.

$p = 0.50$ compatible with $\bar{Y} = 0.53$ if we interview 200 voters.

- If we interview 200 voters and find $\bar{Y} = 0.53$, plausible that $p = 0.50$?
- $n = 200$ and $\bar{Y} = 0.53$. CLT says $\frac{0.53-p}{\sqrt{\frac{0.53(1-0.53)}{200}}}$ follows N(0,1).
- Plugging $p = 0.50$ into this expression yields 0.85006.
- A N(0,1) random variable might very well be equal to 0.85006, not rare value.
- Therefore, not implausible that p is equal to 0.50. Compatible with the data we observe.
- Intuition: if you draw 2000 times a coin and get 1060 heads, you can be pretty sure the coin is not fair, biased towards heads. If you draw it 200 times and get 106 heads, the coin might be fair. Same here, except that you draw voters instead of tossing a coin, and heads=voting Clinton.
- So far: intuitive introduction to the theory of statistical tests. Now we formally review this theory. Then we will come back to our example.

Reminder: the t-test with 5% level

- Let $X_1, X_2, \ldots, X_n$ be n iid random variables with expectation m.
- Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and let $\hat{V}(\bar{X}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$.
- CLT: if $n \geq 100$, $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}}$ approximately follows a normal distribution with expectation 0 and variance 1.
- For any real number x, we can use this to test the **null hypothesis** $m = x$, while controlling the probability of **type 1 error**: rejecting the null hypothesis while it is true.
- If we want to have 5% chances of wrongly rejecting $m = x$, test is:

Reject $m = x$ if $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} > 1.96$ or $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} < -1.96$.

Otherwise, do not reject $m = x$.

- $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}}$ is called the t-statistic, and the test above is called a t-test with 5% **level** (probability of type 1 error)

If we use a 5%-level t-test, what are the chances we make a type 1 error?

- Our test is the following:

Reject $m = x$ if $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} > 1.96$ or $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} < -1.96$.

Otherwise, do not reject $m = x$.

- If we use this test, what are the chances we reject $m = x$ while actually it is true that $m = x$?
- If $m = x$, then $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} = \frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}}$. Then, we will wrongly

reject $m = x$ if $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}} > 1.96$ or if $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}} < -1.96$.

- We know from the CLT that $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}}$ follows a normal distribution with expectation 0 and variance 1.
- What is probability that $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}} > 1.96$ or $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}} < -1.96$?

5%!

- $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}}$ follows a N(0,1) distribution.
- What is probability that $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}} > 1.96$ or $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}} < -1.96$?
- F2-N(0,1): a N(0,1) random variable has a 95% probability of being included between -1.96 and 1.96.
- Therefore, the probability that a N(0,1) is either greater than 1.96 or lower than -1.96 is 5%!
- Now, we have a 5% level test of $m = x$. With your neighbor, try to find a 10% level test of $m = x$. Hint: take a look at F1-N(0,1).

iClicker time

- Which of the following is a 10% level test of $m = x$?

a) Reject $m = x$ if $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} > 1.64$ or $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} < -1.64$.

Otherwise, do not reject $m = x$.

b) Reject $m = x$ if $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} > 2.57$ or $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} < -2.57$.

Otherwise, do not reject $m = x$.

T-tests with 10% and 1% level

- If we want to have 10% chances of wrongly rejecting $m = x$, our test is:

Reject $m = x$ if $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} > 1.64$ or if $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} < -1.64$.

Otherwise, do not reject $m = x$.

- If we want to have 1% chances of wrongly rejecting $m = x$, our test is:

Reject $m = x$ if $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} > 2.57$ or if $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} < -2.57$.

Otherwise, do not reject $m = x$.

Reminder: a 95% confidence interval for m

- Let $X_1, X_2, \ldots, X_n$ be n iid random variables with expectation m.
- A 95% confidence interval (CI) is an interval such that m has 95% chances of belonging to this interval.
- Simple formula: $\left[\overline{X} - 1.96\sqrt{\hat{V}(\overline{X})},\ \overline{X} + 1.96\sqrt{\hat{V}(\overline{X})}\right]$.
- Confidence interval: $\overline{X}$, our estimator of m, +/- a statistical margin of error: $1.96\sqrt{\hat{V}(\overline{X})}$.
- $\hat{V}(\overline{X}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$ goes to 0 when n grows => margin of error vanishes when sample gets large.
- Now, we have 95% CI for m. With your neighbor, try to find 90% CI for m.

iClicker time

- Which of the following is a 90% confidence interval for m?

a) $\left[\bar{X} - 2.57\sqrt{\hat{V}(\bar{X})},\ \bar{X} + 2.57\sqrt{\hat{V}(\bar{X})}\right]$

b) $\left[\bar{X} - 1.64\sqrt{\hat{V}(\bar{X})},\ \bar{X} + 1.64\sqrt{\hat{V}(\bar{X})}\right]$

90 and 99% CIs for m

- 90% CI: $\left[\bar{X} - 1.64\sqrt{\hat{V}(\bar{X})},\ \bar{X} + 1.64\sqrt{\hat{V}(\bar{X})}\right]$.
- 99% CI: $\left[\bar{X} - 2.57\sqrt{\hat{V}(\bar{X})},\ \bar{X} + 2.57\sqrt{\hat{V}(\bar{X})}\right]$.
- We now have 95%, 90% and 99% CIs for m. Which is the widest? Which is the tightest? Intuition?

The more certain we want to be, the more margin of error we need to allow.

- The widest is the 99% CI: width =

$$\bar{X} + 2.57\sqrt{\hat{V}(\bar{X})} - \left(\bar{X} - 2.57\sqrt{\hat{V}(\bar{X})}\right) = 2 \times 2.57\sqrt{\hat{V}(\bar{X})}.$$

- The tightest is the 90% one: width = $2 \times 1.64\sqrt{\hat{V}(\bar{X})}$.
- Intuition: if we want to be sure with 99% certainty that m belongs to some interval around $\bar{X}$, we need to allow for a larger margin of error around $\bar{X}$ than if we only want 90% certainty.

A 5% level t-test that 50% of voters want to vote Clinton

- Assume we randomly choose with replacement n voters from the register, that we interview them, and ask them if they want to vote for Clinton.
- Let $Y_1, Y_2, \ldots, Y_n$ denote the answers of these voters.
- So far: $Y_1, Y_2, \ldots, Y_n$ are iid, and that their expectation is p. Because $Y_1, Y_2, \ldots, Y_n$ are binary, $\hat{V}(\bar{Y}) = \frac{\bar{Y}(1-\bar{Y})}{n}$
- From our review of theory of statistical tests, it follows that the following is a 5%-level test of $p = x$:

Reject $p = x$ if $\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} > 1.96$ or $\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} < -1.96$.

Otherwise, do not reject $p = x$.

- If $n = 2000$, $\bar{Y} = 0.53$, do you reject $p = 0.50$ at 5% level? Try to find the answer with your neighbor during 1 minute.

iClicker time

- If $n = 2000$, $\bar{Y} = 0.53$, do you reject $p = 0.50$ at the 5% level?

a) Yes

b) No

Yes!

- From our review of theory of statistical tests, it follows that the following is a 5%-level test of $p = x$:

Reject $p = x$ if $\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} > 1.96$ or $\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} < -1.96$.

Otherwise, do not reject $p = x$.

- If $n = 2000$ and $\bar{Y} = 0.53$, can you reject $p = 0.50$ at the 5% level?
- Yes! Let $n = 2000$, $\bar{Y} = 0.53$, and $x = 0.50$. Then,

$\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} = 2.688125$. This number is larger than 1.96.

- If $n = 2000$, $\bar{Y} = 0.53$, do you reject $p = 0.50$ at 1% level? Try to find the answer with your neighbor during 1 minute.

iClicker time

- If $n = 2000$, $\bar{Y} = 0.53$, do you reject $p = 0.50$ at 1% level?

a) Yes

b) No

Yes!

- From our review of theory of statistical tests, it follows that the following is a 1%-level test of $p = x$:

Reject $p = x$ if $\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} > 2.57$ or $\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} < -2.57$.

Otherwise, do not reject $p = x$.

- If $n = 2000$ and $\bar{Y} = 0.53$, can you reject $p = 0.50$ at the 1% level?
- Yes! Let $n = 2000$, $\bar{Y} = 0.53$, and $x = 0.50$. Then,

$\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} = 2.688125$. This number is larger than 2.57.

- Find the 95% confidence interval of p. Try to find the answer with your neighbor during 1 minute.

iClicker time

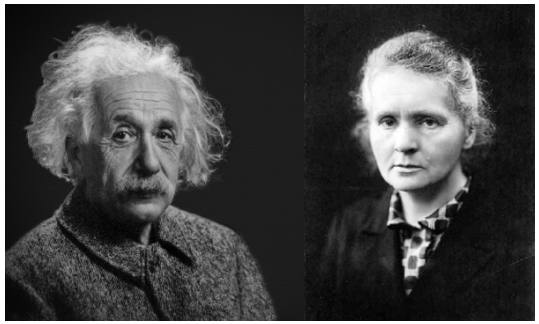
- If $n = 2000$, $\bar{Y} = 0.53$, the 95% confidence interval for p is equal to:

a) [0.508,0.552]

b) [0.372,0.971]

The confidence interval of p is [0.508,0.552].

- From our review of theory of confidence intervals, general formula for 95% confidence interval of p is
- $\left[\bar{Y} - 1.96\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}, \bar{Y} + 1.96\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right]$
- Here, $n = 2000$, $\bar{Y} = 0.53$, so this interval is [0.508,0.552].
- Given that 53% of our sample of 2000 randomly chosen Pennsylvania voters wants to vote for Clinton, we can be 95% confident that the share of Pennsylvania voters that want to vote for Clinton is included by 50.8% and 55.2%.
- Your smart friends were right!



Interpreting what a confidence interval means in the polling example.

- When we randomly draw 2000 voters with replacement from the Pennsylvania electorate, there are $8{,}448{,}674^{2000}$ possible outcomes of this random draw.
- When we say that we are 95% confident that p belongs to $\left[\bar{Y} - 1.96\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}, \bar{Y} + 1.96\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right]$, that means that for 95% of these $8{,}448{,}674^{2000}$ samples we can draw, p belongs to that interval.
- $1.96\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}$ is our statistical margin of error.
- Maybe $\bar{Y}$ is not exactly equal to p, but for 95% of all the random samples of 2000 voters we can draw, $\bar{Y}$ will not be more than $1.96\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}$ away from p.
- If you conduct 1000 polls (in each of the 50 states and at 20 different points in time) and that after each poll you say that proportion of voters that want to vote for Clinton in that state and at that point in time is included in the confidence interval of the poll, that statement will be right for 950 of those polls, but it will be wrong for 50 polls.

What you need to remember (1/2)

- Assume we draw n voters from the register, and ask them if they want to vote for Clinton. Let $Y_1, Y_2, \ldots, Y_n$ denote their answers.
- To estimate p, we use $\bar{Y}$. $V(\bar{Y}) = \frac{1}{n}\mathrm{V}(Y_1)$. We use $\hat{V}(\bar{Y}) = \frac{\bar{Y}(1-\bar{Y})}{n}$ to estimate $V(\bar{Y})$.
- The following is a 5%-level test of $p = x$:

Reject $p = x$ if $\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} > 1.96$ or $\frac{\bar{Y}-x}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} < -1.96$.

Otherwise, do not reject $p = x$.

- We can 95% confident that p lies in the following interval: $\left[\bar{Y} - 1.96\sqrt{\hat{V}(\bar{Y})},\ \bar{Y} + 1.96\sqrt{\hat{V}(\bar{Y})}\right]$.

What you need to remember (2/2)

- All results we saw hold whenever we observe sample of iid variables. Poll = a special case where we observe iid sample.
- Assume we observe $X_1, X_2, \ldots, X_n$, n iid random variables. Let m denote their expectation. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote their average.
- $\bar{X}$: unbiased estimator of m. Linear unbiased estimator of m with lowest variance.
- $V(\bar{X}) = \frac{V(X_1)}{n}$. $\lim_{n\to+\infty} \frac{1}{n} V(\bar{X}) = 0$.
- With infinite sample, $\bar{X}$ becomes equal to m: $\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = m$.
- Let $\hat{V}(\bar{X}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ be estimator of $V(\bar{X})$.
- CLT: if n is larger than 100, $\frac{\bar{X}-m}{\sqrt{\hat{V}(\bar{X})}}$ approximately follows N(0,1) distribution.
- A 5%-level test of $m = x$ is:

Reject if $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} > 1.96$ or $\frac{\bar{X}-x}{\sqrt{\hat{V}(\bar{X})}} < -1.96$.

Otherwise, do not reject.

- A 95% confidence interval for m is $\left[\bar{X} - 1.96\sqrt{\hat{V}(\bar{X})},\ \bar{X} + 1.96\sqrt{\hat{V}(\bar{X})}\right]$.

Some practical details

- In practice, polling firms choose voters without replacement:
  - randomly choose a number between 1 and 8,448,674. Interview voter with that number and ask if she wants to vote for Clinton.
  - then, randomly choose another number between 1 and 8,448,674, excluding the first number they drew. Interview the voter with that number and ask if she wants to vote for Clinton.
  - Etc.
- If we do this, all the results we saw remain true, but harder to derive.
- Only 1 change: now $\mathrm{V}(\bar{Y})$ is $\frac{\bar{Y}(1-\bar{Y})}{n}\left(1-\frac{n}{8{,}448{,}674}\right)$, instead of $\frac{\bar{Y}(1-\bar{Y})}{n}$.
- $1-\frac{n}{8{,}448{,}674} \leq 1$: drawing voters without replacement reduces variance of our estimator (that's why pollsters draw voters without replacement)
- When we draw small sample from very large population, makes almost no difference: e.g. $1-\frac{2000}{8{,}448{,}674}=0.999845.$
- In practice how do you draw a random sample without replacement from a large population? You assign to each unit a random number, and you select, say, the 2000 units with the lowest value of that random number.

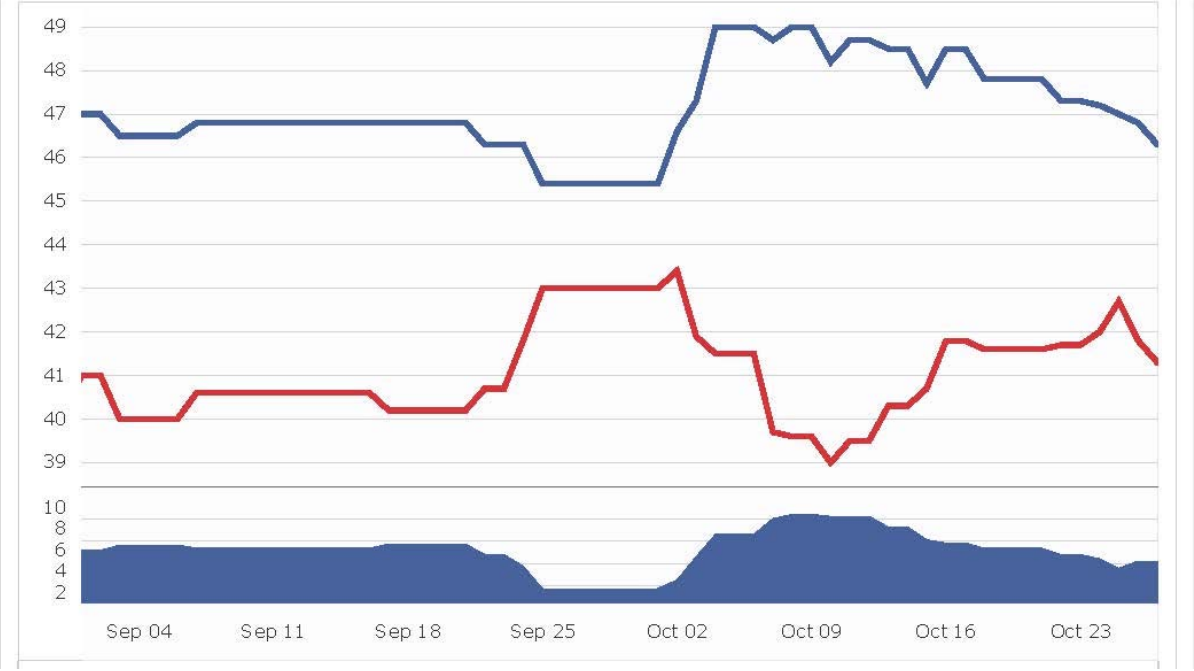
Roadmap

1. Lay-out 4 assumptions.
2. Show that if those 4 assumptions are satisfied, then what your friend is saying is indeed correct. Along the way, learn more general lessons about expectation estimation, hypothesis testing, and confidence intervals.
3. Critically assess those 4 assumptions. Relatedly, discuss potential explanations for why polls failed to predict the outcome of the last US presidential election. Present briefly job opportunities for econ majors in polling firms.

Part 3: Where we critically assess the 4 magical assumptions, and try to explain why polls failed to predict outcome of last US presidential election

Pennsylvania polls were very favorable to Clinton, but she lost the state.

- Average of all polls in Pennsylvania. Clinton in Blue.
- Average of polls in week before election: 46.8% Clinton, 44.7% Trump.
- Real result: 47.5% Clinton, 48.2% Trump.
- 48.2% outside of 95% confidence interval for % of voters voting Trump one could compute using surveys before the election.

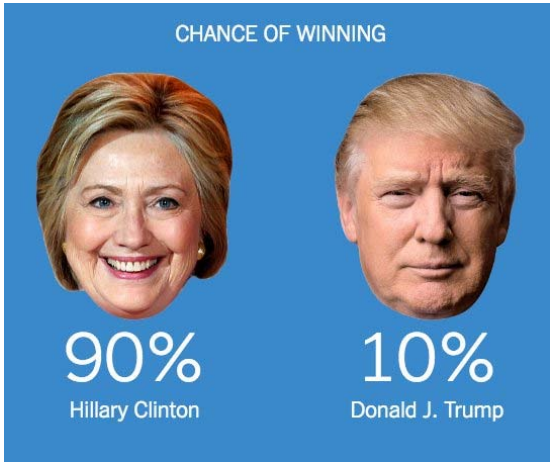


Polls even more favorable in Wisconsin & Michigan, but Clinton also lost those states.

- Wisconsin:
  - Average of polls in week before election: 46.8% Clinton, 40.3% Trump.
  - Real result: 46.5% Clinton, 47.2% Trump.
  - 47.2% outside of 95% confidence interval for the % of voters voting Trump one could compute using surveys before the election.
- Michigan:
  - Average of polls in week before election: 47.0% Clinton, 43.4% Trump.
  - Real result: 47.0% Clinton, 47.3% Trump.
  - 47.3% outside of 95% confidence interval for the % of voters voting Trump one could compute using surveys before the election.

Based on these polls, high degree of confidence Clinton would win.

- NY Times forecast, couple of days before the election: Clinton has 90% chance of winning.



- Those forecasts might have discouraged democrat voters from voting: useless, Clinton is going to win anyway.

One of our 4 magical assumptions must have failed. Which one?

- % of electorate that voted for Trump in Pennsylvania, Wisconsin, and Michigan was not in the 95% confidence interval of Trump vote share based on surveys in these 3 states.
- => polls were wrong, much beyond their statistical margin of error.
- => one of our 4 magical assumptions must have failed.
- Which one?
- Now we review the 4 magical assumptions, we discuss if they are plausible in general and in the specific context of the 2016 US presidential election.

Assumption 1: your friend has access to some register including the contact details of all voters in Pennsylvania

- This assumption is never true. Instead polling firms typically use the phone book, call people and ask them if they are going to vote in the next election. If not, interview ends there. If yes, then they ask who they want to vote for.
- => their sample is representative of voters in the phone book, not of all voters.
- Is this an issue?

Polls sample is representative of voters in phonebook, not of all voters.

- 30 years ago: no big deal. Almost everybody was in the phone book. Nowadays: much less true. For instance, many people no longer have a landline.
- People in the phonebook might be different from people who are not, and might vote differently. One might for instance suspect that people who are not in the phonebook are younger, and therefore more liberal.
- If that's the case polls that use the phonebook are only representative of voters in the phonebook, not of the entire electorate.
- Therefore, polling firms now use emails data bases to draw their samples. Not clear that these data bases representative of electorate.
- Is the failure of Assumption 1 likely to explain why polls failed to see Trump would win?

But this is unlikely to explain why polls failed to see Trump would win.

- Is the failure of Assumption 1 likely to explain why polls failed to see Trump would win?
- Maybe, but not super plausible: the share of people not in the phonebook was already quite high in the 2012 presidential election, and yet polls had done a good job forecasting Obama would win.

Assumption 2: your friend drew randomly the sample of 2000 voters he/she is going to interview out of this register.

- Most polling companies randomly draw their sample from, the phonebook, or large data base of emails.
- When you read a poll, very important to see whether this methodology was used. If not, the results of the poll should be interpreted with a lot of caution: sample is not representative of any larger population.
- Even serious polls using random sampling failed to see Trump would win => problem does not come from failure of Assumption 2.

Assumption 3: when he contacts them, the 2000 voters all answer to your friend.

- Assumption 3 not plausible: many people do not answer to polls.
- In your analysis, you can only use data from people who have responded. => your sample is representative of voters who respond to polls.
- This is an issue if people who do not respond to polls vote differently from people who respond to polls.

Respondents might vote differently from non respondents, but hard to say in which direction.

- People who do not respond to polls are less likely, equally likely, or more likely to be liberals than people who respond to polls?
- We know that people who do not respond to polls are richer, and younger than people who respond.
- Hard to say whether more or less likely to be liberals than respondents: the fact they are younger makes them more likely to be liberals. The fact they are richer makes them more likely to be conservatives.
- In any case, respondents and non respondents are unlikely to be representative of each other. Big issue for polls.
- Unfortunately, even serious pollsters do not report their non-response rates: http://apps.washingtonpost.com/g/page/politics/post-abc-tracking-poll-oct-27-30-2016/2118/
- Any idea to increase response rates to polls?

One could pay people to participate in polls.

- Any idea to increase the response rates to polls?
- Economist's answer: you should pay respondents.
- Issue: might have differential effects according to people's income.
- Larger effect among poorer people: many accept to respond. Smaller effect among richer people: few accept to respond.
- Might then lead to more biased sample of respondents: poorer people more over-represented among respondents than if we had not given incentive.
- Issue less likely to be present if financial incentive large. Would mean that polls more costly, but maybe that's the price to pay to get representative polls.
- Is the failure of Assumption 3 likely to explain why polls failed to see Trump would win?

Failure of Assumption 3 unlikely to explain why polls failed to see Trump would win.

- Is the failure of Assumption 3 likely to explain why polls failed to see Trump would win?
- Maybe, but not super plausible: the share of people not responding to polls was already very high in the 2012 presidential election, and yet polls had done a pretty good job at forecasting Obama would win.

Assumption 4: when they respond to your friend's question: "Are you willing to vote for Hillary Clinton?", voters respond truthfully.

- Plausible in general? Yes, no reason why you would lie to pollster.
- However, is Assumption 4 still plausible in the context of the 2016 US presidential election?

Social stigma attached to voting for Trump => voters might have responded less truthfully

- Stigma associated with voting for Trump => maybe people who wanted to vote for Trump did not dare to say it to pollsters.
- What's striking in Pennsylvania, Wisconsin, and Michigan is that polls were spot on at predicting Clinton's vote share, but underestimated Trump's.
- Some people who were planning to vote for Trump did not dare to tell pollsters, and said instead they were undecided, or planning on voting for Gary Johnson.
- Can you think of a way to account for this?

Use discrepancy between results and polls in the last election to adjust current polls.

- Can you think of a way to account for this?
- The last-election correction method:
  - Compute the following adjustment factor: share of people who voted for the “frowned upon” candidate in last election / share of survey respondents who said they would vote for that candidate before last election.
  - multiply the share of respondents who say they want to vote for “frowned upon” candidate in the current election by that factor.
  - If polls underestimate the “frowned upon” candidate equally in the current and in the last election, this method will work.

Polling is an interesting and important industry, and some polling firms hire Econ majors

- Polling is a difficult, and therefore interesting job.
- Polls matter: last US election.
- If you want to learn more about polling, check Nate Silver's blog. You can start with this article: https://fivethirtyeight.com/features/the-state-of-the-polls-2016/ , where he explains how he rates polls.
- Some polling firms hire econ majors. E.g., the Public Policy Institute of California recently had an opening for a research associate position, and the required qualifications were: "Bachelors or Master's degree in economics, health, political science, public policy, sociology or a related field, or equivalent experience."
- Lists of main polling firms in the US available here: https://projects.fivethirtyeight.com/pollster-ratings/ or here: https://en.wikipedia.org/wiki/List_of_polling_organizations#United_States

What you need to remember

- Polls largely underestimated Trump's vote share in Pennsylvania, Michigan, and Wisconsin in the 2016 US presidential election.
- => violation of one of the 4 magical assumptions must explain that.
- Pollsters cannot draw their sample from a register with all voters. Instead, use the phonebook. Unlikely to explain why polls failed in 2016: was already the case in 2012.
- Most pollsters randomly draw their sample => assumption 2 satisfied.
- Not all sampled voters answer to polls. Unlikely to explain why polls failed in 2016: was already the case in 2012.
- Normally, one would expect that voters answer truthfully to polls. However, Trump was a "frowned upon" candidate: social stigma attached to saying you vote for him => maybe sampled voters did not dare to say they would vote for him.