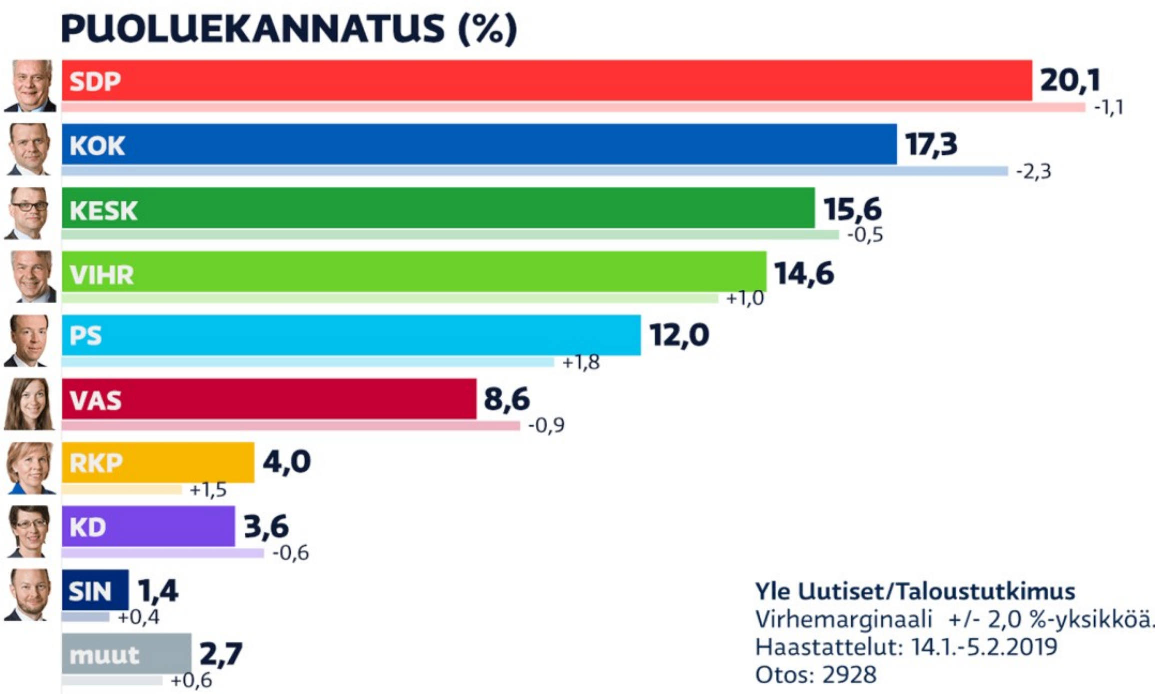
Learning outcomes

- Explain how and why polls are done by **sampling**
- Explain how sampling causes **error**
- Explain how error is characterized by **confidence intervals**
- **Calculate** a confidence interval in simple cases

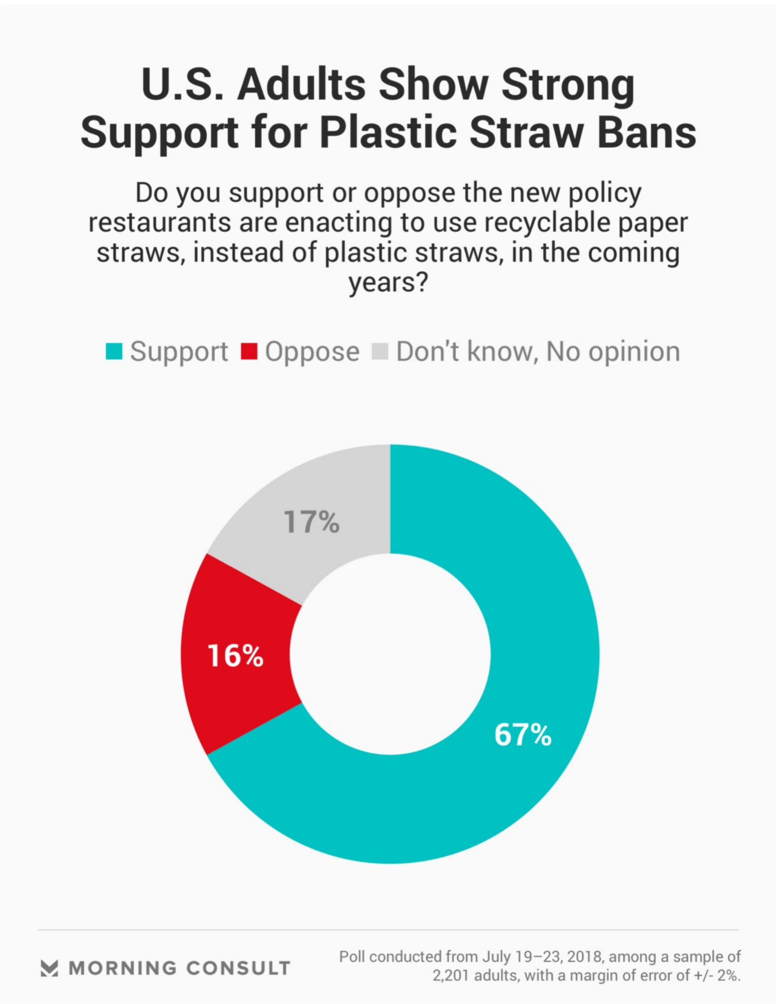
- **Choose** suitable sample size for desired accuracy
- **Recognize** situations where you need an advanced formula
- Critically **evaluate** CIs reported in polls
- Identify errors that are **not covered** by CI

Different kinds of polls – same mathematics

Voting intent, multiple choices



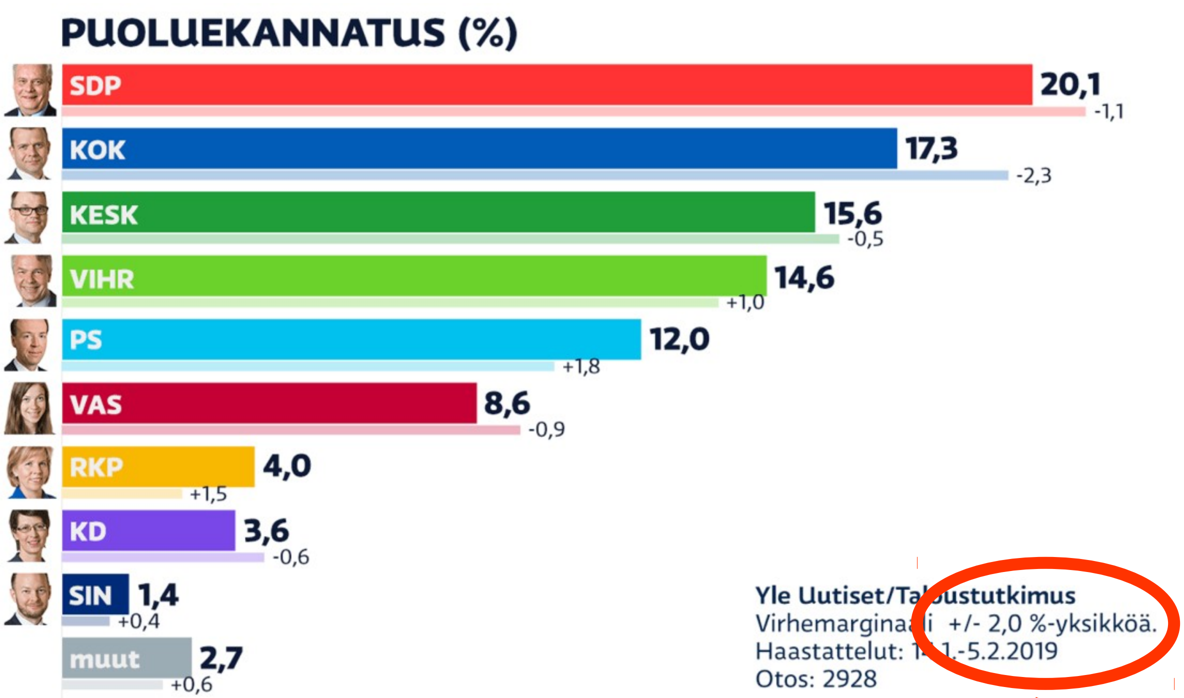
Binary opinion on an issue, 2+1 choices



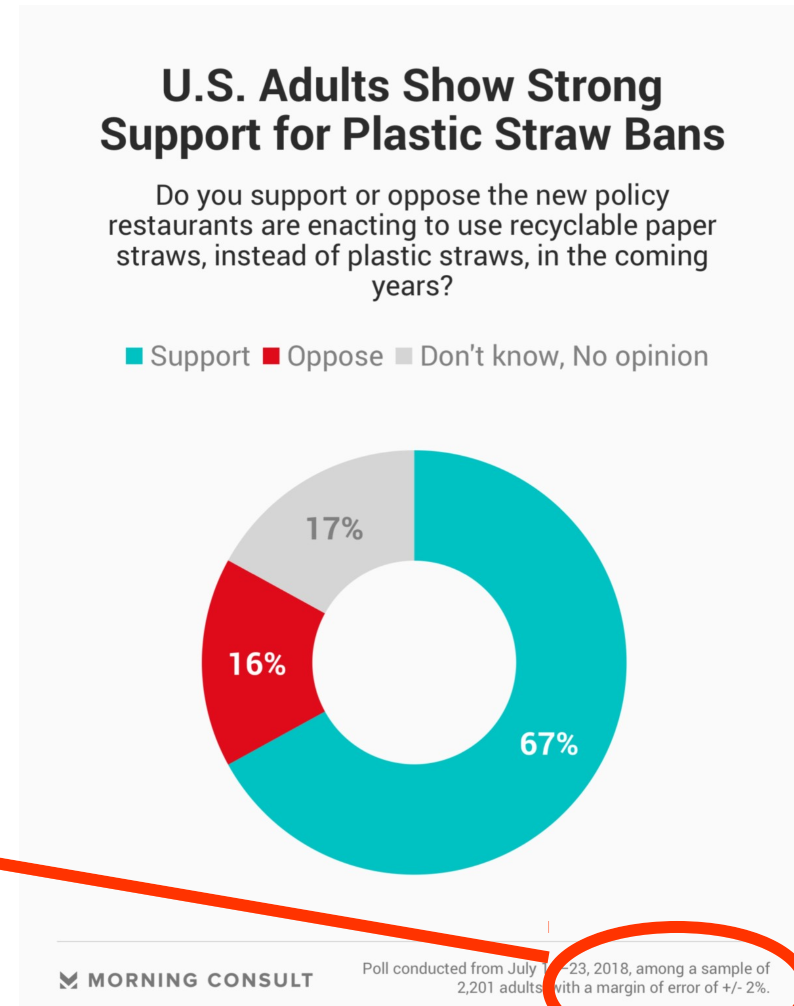
These samples (2928 and 2201 persons) are tiny portions of the whole population. How can we claim anything about the population’s opinions?

Different kinds of polls – same mathematics

Voting intent, multiple choices



Binary opinion on an issue,
2+1 choices



Margin of error (MOE)

Very rough meaning:
"The numbers should not be off by more than that,
at least not too often"

Black box formula

Insert proportion
seen in sample

$$MOE = z^{*} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

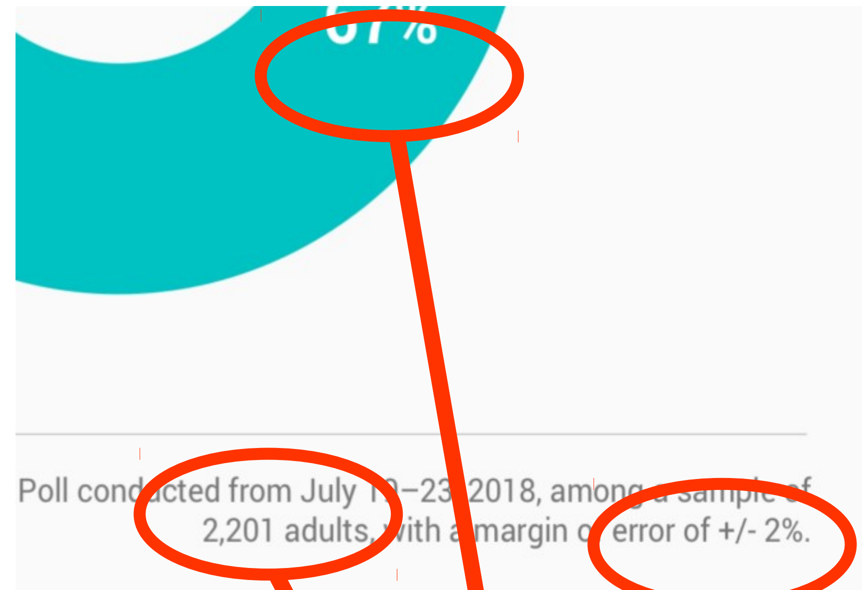
Use 1.96 for 95 %
confidence level

Insert sample size

If you master this,
you have exceeded
95% of the population!

[Statistics made up.]

Black box formula



Insert proportion seen in sample

$$MOE = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Use 1.96 for 95 % confidence level

Insert sample size

$$\boxed{\begin{gathered} z^* = 1.96 \\ \hat{p} = 0.67 \\ n = 2201 \\ MOE = 0.020 \end{gathered}}$$

In the **sample**, proportion of supporters is 67 %. This we know. In the **population**, we can now say that the proportion of supporters is **(67±2) %**, at 95% confidence. Is this magic?

Back to basics: How does sampling work?

Let us first understand the **direct problem.**

- What is the **process** that creates the numbers we observe?
- Where is the **randomness**?
- Is there a familiar **distribution** involved?

We do this on the blackboard.

Then hopefully we can understand the inverse or **inference problem**: What do the observations tell us?

(From the blackboard)

- From a large population (***N***), we took a small sample (***n***).
- In the population we had ***Np* persons** of the type we are interested in (e.g. supporters of party A).
- In the sample we will have some number ***K***, which has binomial distribution Bin(*n, k*).
- Naturally we would like to say that the **sample proportion**

$$\hat{p} = K/n$$

is a good estimate for *p*. Can we do that?

Properties of the estimator

- At least the **expected value** is just right,

$$\mathrm{E}(\hat{p}) = \mathrm{E}(K/n) = \mathrm{E}(K)/n = np/n = p$$

because *K* is binomially distributed with expectation *np.*

- But how far will it the estimator be from its expectation? The **standard deviation**

$$\mathrm{D}(\hat{p}) = \mathrm{D}\left(\frac{K}{n}\right) = \frac{\mathrm{D}(K)}{n} = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$$

gives a good understanding of this.

Simulate taking a sample

```
% Simulate an opinion poll from a huge population
p = 0.14;         % True proportion of supporters (in population)
n = 1000;         % Sample size

k = binornd(n,p); % Take a sample of n persons
phat = k/n;       % Observed proportion (in sample)

fprintf('Sample has K=%d: ', k);
fprintf(' estimate phat=%.3f,', phat);
fprintf(' error = %+.3f\n', phat-p);
```

It is not very complicated!

Simulate taking a sample

```
% Simulate an opinion poll from a huge population
p = 0.14;         % True proportion of supporters (in population)
n = 1000;         % Sample size

k = binornd(n,p); % Take a sample of n persons
phat = k/n;       % Observed proportion (in sample)

fprintf('Sample has K=%d: ', k);
fprintf(' estimate phat=%.3f,', phat);
fprintf(' error = %+.3f\n', phat-p);
```

```
>> onepoll
Sample has K=135:  estimate phat=0.135, error = -0.005
```

Simulate taking a sample

```
% Simulate an opinion poll from a huge population
p = 0.14;           % True proportion of supporters (in population)
n = 1000;           % Sample size

k = binornd(n,p); % Take a sample of n persons
phat = k/n;         % Observed proportion (in sample)

fprintf('Sample has K=%d: ', k);
fprintf(' estimate phat=%.3f,', phat);
fprintf(' error = %+.3f\n', phat-p);
```

```
>> onepoll
Sample has K=135:  estimate phat=0.135, error = -0.005
```

Oh well.
We were “expecting” E(K)=140 supporters, but we got only K=135.
We were off by 5. Equivalently, our proportion is off by 5/1000.

Given that D(K) = $\sqrt{np(1-p)}$ = 11, we should not be too surprised.
This was a simulation, but **same happens when you really sample.**

Simulate 20 polls

```
Poll # 1: Sample has K=126:  estimate phat=0.126, error = -0.014
Poll # 2: Sample has K=142:  estimate phat=0.142, error = +0.002
Poll # 3: Sample has K=146:  estimate phat=0.146, error = +0.006
Poll # 4: Sample has K=145:  estimate phat=0.145, error = +0.005
Poll # 5: Sample has K=135:  estimate phat=0.135, error = -0.005
Poll # 6: Sample has K=127:  estimate phat=0.127, error = -0.013
Poll # 7: Sample has K=169:  estimate phat=0.169, error = +0.029
Poll # 8: Sample has K=143:  estimate phat=0.143, error = +0.003
Poll # 9: Sample has K=145:  estimate phat=0.145, error = +0.005
Poll #10: Sample has K=155:  estimate phat=0.155, error = +0.015
Poll #11: Sample has K=138:  estimate phat=0.138, error = -0.002
Poll #12: Sample has K=145:  estimate phat=0.145, error = +0.005
Poll #13: Sample has K=149:  estimate phat=0.149, error = +0.009
Poll #14: Sample has K=136:  estimate phat=0.136, error = -0.004
Poll #15: Sample has K=142:  estimate phat=0.142, error = +0.002
Poll #16: Sample has K=149:  estimate phat=0.149, error = +0.009
Poll #17: Sample has K=147:  estimate phat=0.147, error = +0.007
Poll #18: Sample has K=122:  estimate phat=0.122, error = -0.018
Poll #19: Sample has K=168:  estimate phat=0.168, error = +0.028
Poll #20: Sample has K=138:  estimate phat=0.138, error = -0.002
```

Let's be very cheap

- Approximate the binomial with a normal, with same mean and same standard deviation

$$K \sim N\left(np,\ np\left(1-p\right)\right)$$

- The error of our estimate is

$$\hat{p}-p = \left(K/n\right)-p$$

- So the error is normally distributed with **zero mean** and standard deviation

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

And that gives us our basic formula

$$MOE = 1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

The estimate will be **within ±MOE** of the true proportion 95% of the time, when we do this kind of sampling.

That is because a normally distributed variable is within ±1.96·sd of its own mean, 95% of the time.

We will say that

$$\left[\hat{p} - MOE,\ \ \hat{p} + MOE\right]$$

is the 95% **confidence interval** for *p*.

Improved code: Compute CI

```
% Simulate an opinion poll from a huge population
p = 0.14;         % True proportion of supporters (in population)
n = 1000;         % Sample size

k = binornd(n,p); % Take a sample of n persons
phat = k/n;       % Observed proportion (in sample)

moe = 1.96 * sqrt(phat*(1-phat)/n);
ci  = [phat-moe, phat+moe];

fprintf('Sample has K=%d: ', k);


fprintf(' estimate phat=%.3f,', phat);
fprintf(' CI = [%.3f, %.3f]  ', ci(1), ci(2));

if ci(1)<=p && p<=ci(2)
    fprintf('OK\n');
else
    fprintf('FAIL\n');
end
```

Simulate 20 polls

Recall that the true proportion in our population is 0.140.

```
Poll # 1: Sample has K=126:  estimate phat=0.126, CI = [0.105, 0.147]  OK
Poll # 2: Sample has K=142:  estimate phat=0.142, CI = [0.120, 0.164]  OK
Poll # 3: Sample has K=146:  estimate phat=0.146, CI = [0.124, 0.168]  OK
Poll # 4: Sample has K=145:  estimate phat=0.145, CI = [0.123, 0.167]  OK
Poll # 5: Sample has K=135:  estimate phat=0.135, CI = [0.114, 0.156]  OK
Poll # 6: Sample has K=127:  estimate phat=0.127, CI = [0.106, 0.148]  OK
Poll # 7: Sample has K=169:  estimate phat=0.169, CI = [0.146, 0.192]  FAIL
Poll # 8: Sample has K=143:  estimate phat=0.143, CI = [0.121, 0.165]  OK
Poll # 9: Sample has K=145:  estimate phat=0.145, CI = [0.123, 0.167]  OK
Poll #10: Sample has K=155:  estimate phat=0.155, CI = [0.133, 0.177]  OK
Poll #11: Sample has K=138:  estimate phat=0.138, CI = [0.117, 0.159]  OK
Poll #12: Sample has K=145:  estimate phat=0.145, CI = [0.123, 0.167]  OK
Poll #13: Sample has K=149:  estimate phat=0.149, CI = [0.127, 0.171]  OK
Poll #14: Sample has K=136:  estimate phat=0.136, CI = [0.115, 0.157]  OK
Poll #15: Sample has K=142:  estimate phat=0.142, CI = [0.120, 0.164]  OK
Poll #16: Sample has K=149:  estimate phat=0.149, CI = [0.127, 0.171]  OK
Poll #17: Sample has K=147:  estimate phat=0.147, CI = [0.125, 0.169]  OK
Poll #18: Sample has K=122:  estimate phat=0.122, CI = [0.102, 0.142]  OK
Poll #19: Sample has K=168:  estimate phat=0.168, CI = [0.145, 0.191]  FAIL
Poll #20: Sample has K=138:  estimate phat=0.138, CI = [0.117, 0.159]  OK
```

We were cheap

We **approximated** many times:

- sample without replacement ≈ binomial sample
- binomial distribution ≈ normal distribution
- standard deviation estimated from sample

All of these are reasonable under mild assumptions.

Discuss with your neighbor when they might fail.

See exercises and reading material for improved formulas.

Why not take a bigger sample?

$$MOE = 1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

Look at the formula.

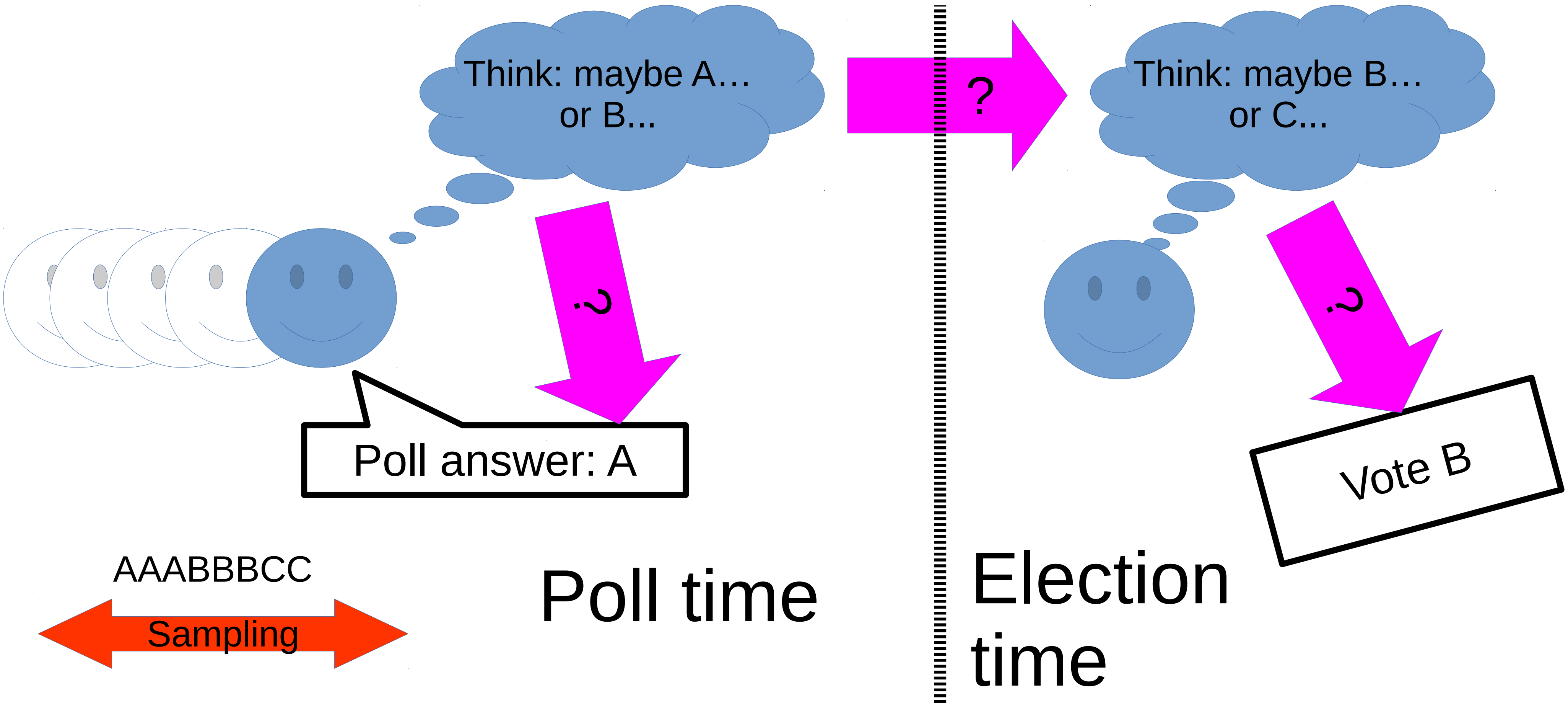
We can decrease the margin of error by increasing n.

Suppose we took a sample of size *n*=1000,
but we want to **decrease *MOE* by a factor of ten.**

What is our new sample size?

Oh, and the other errors…

we only covered sampling error!



Do now: Explain the pink “error arrows” to your neighbor, and discuss. (1 min)

Further examples

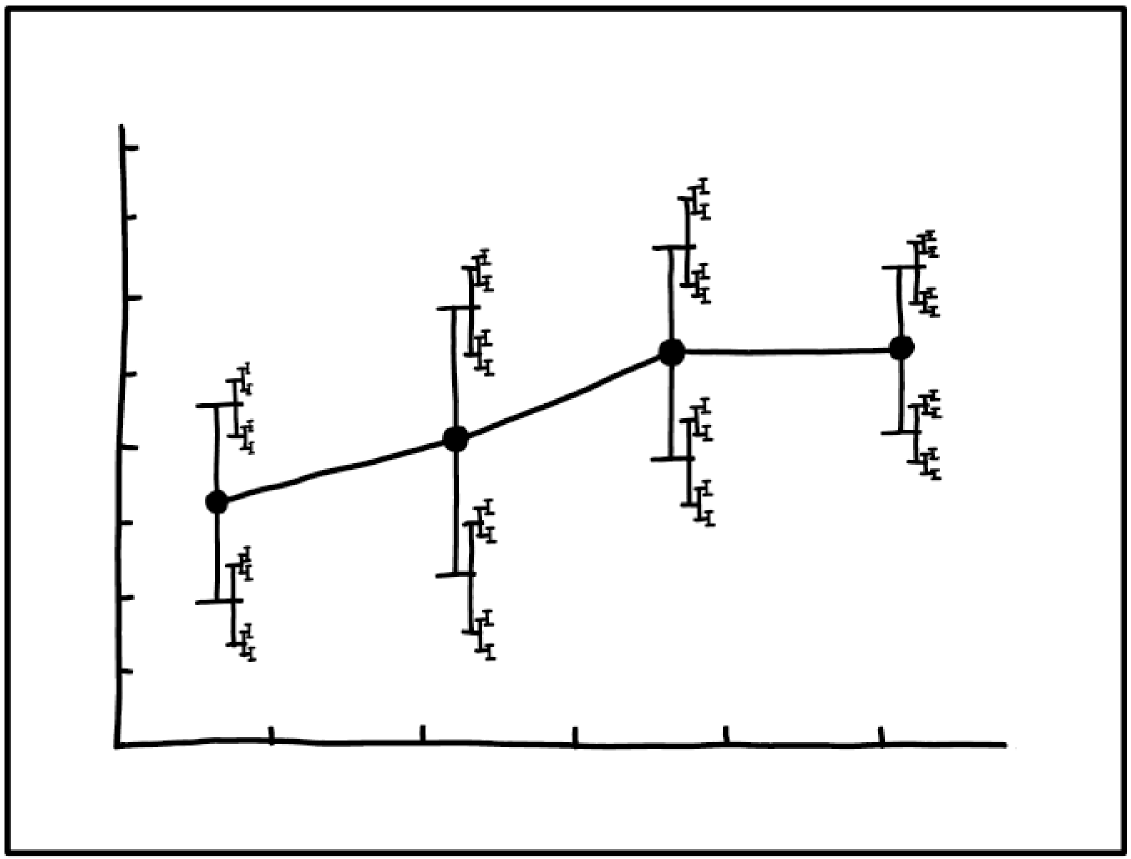
1. The population is N=1000 and **the sample is *n*=*N* .** What does our basic formula say? Explain why the result does/doesn't make sense. Then read about finite sample correction and repeat using that.

2. We have a large population and a sample of n=100.
We observe **0 supporters** of party X. Calculate the CI,

(a) by our basic formula,
(b) exactly, by considering the binomial distribution, and
(c) by a computer simulation.

Compare and explain the results.

Questions?



I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.