Level of confidence to confidence interval

Remember the fish.

If two samples have the same N, but
you make a 95% confidence interval for one
and
an 85% confidence interval for the other
which interval will be wider?

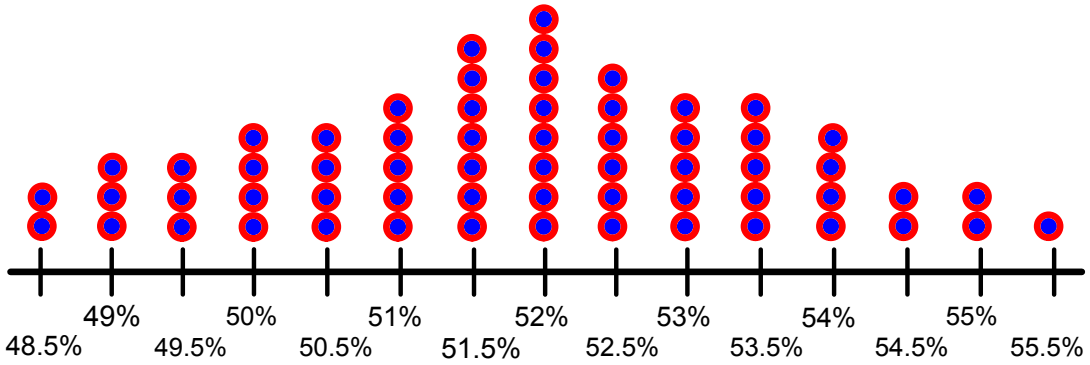
wider interval

narrower interval

To capture a higher % of the samples, you need a wider interval.

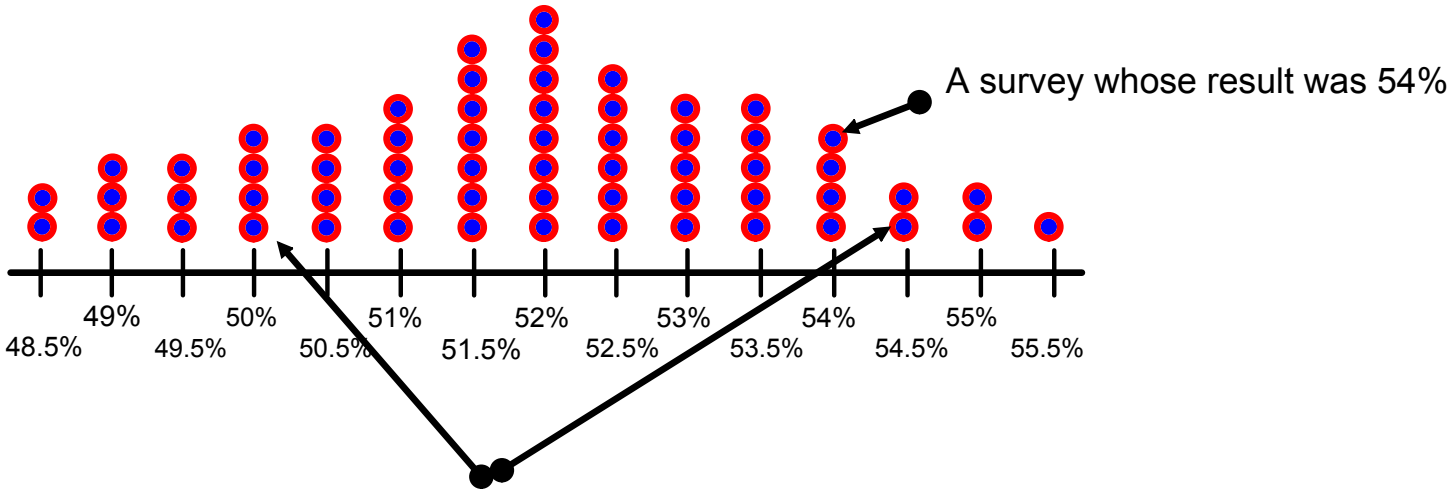
To capture a smaller % of the samples, you can have a narrower interval.

Let's say I took LOTS of surveys/samples



LOTS of surveys/samples

Assume the population parameter is 52%
Why do sample statistics vary from the population parameter?
(Why do your samples give you "wrong" results?



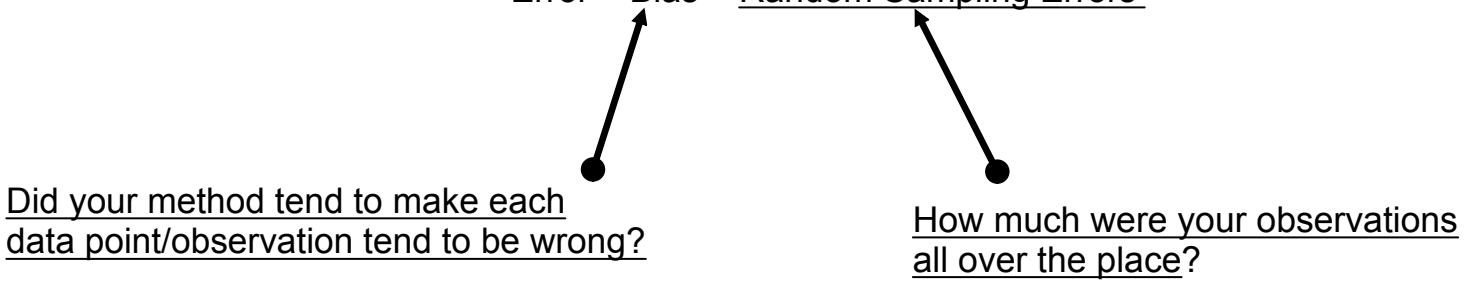
How can it happen that surveys get different results?
Shouldn't they all get 52%

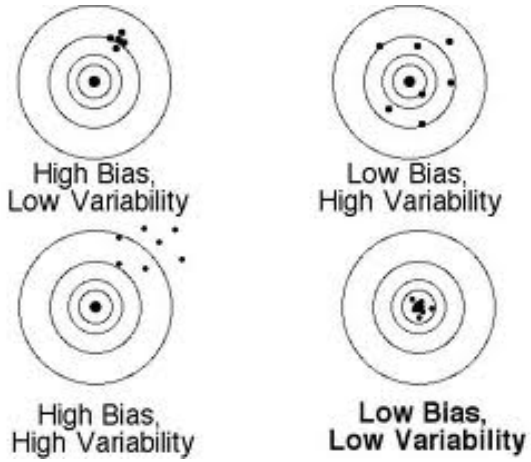
Error = Bias + random error

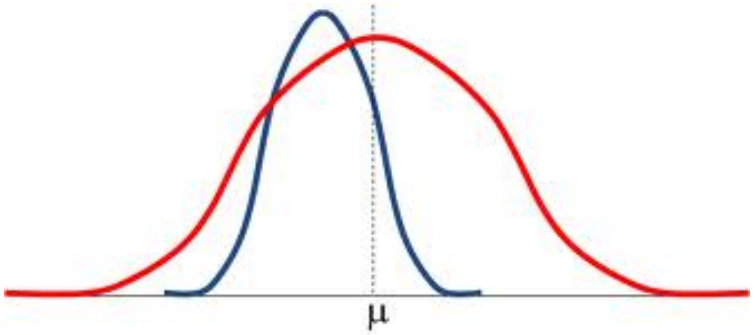
For a statistic calculated for sample:

(Such as the average or a proportion)

Error = Bias + Random Sampling Errors



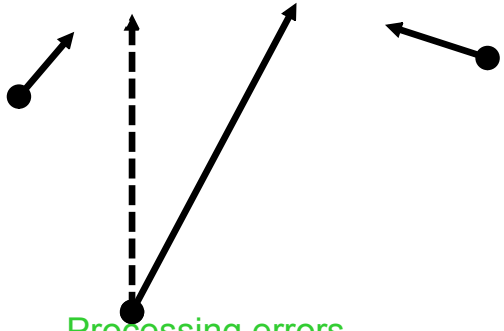

High Bias,
Low Variability
Low Bias,
High Variability
High Bias,
High Variability
Low Bias,
Low Variability


μ

Where we will end up

Why do sample statistics vary from the population parameter?
(Why do your samples give you "wrong" results?

Error = Bias + random error



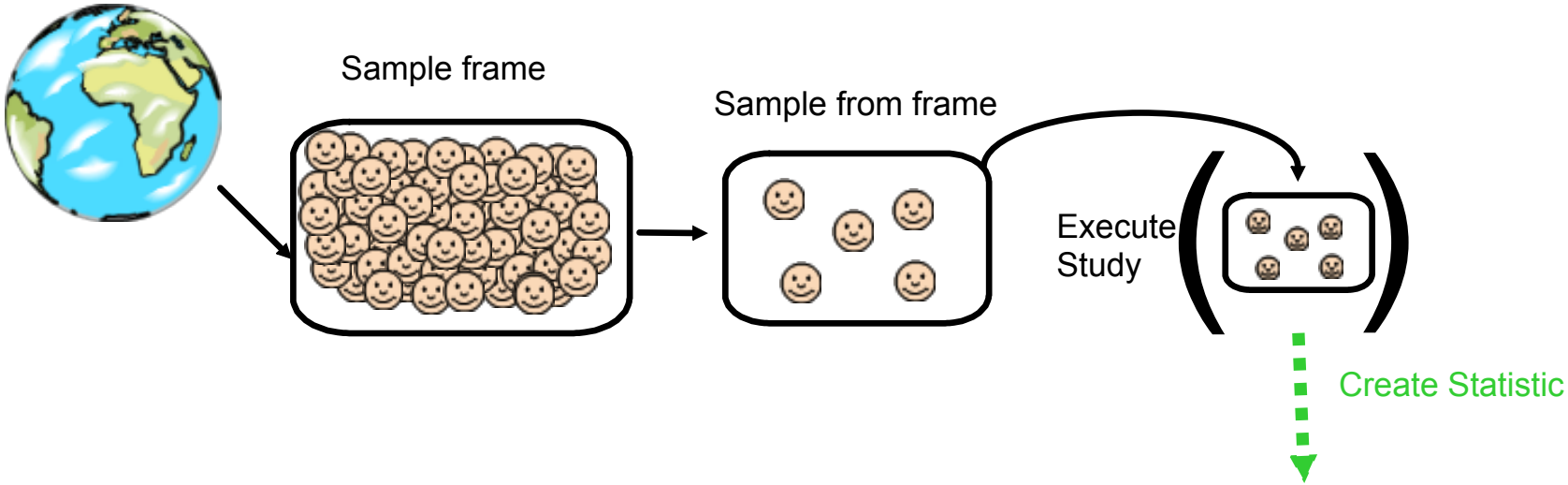
Mistakes picking your sample

Undercoverage
Convenience sample
Voluntary response

Ohter design errors
Nonresponse
Response error lies
Wording Bias

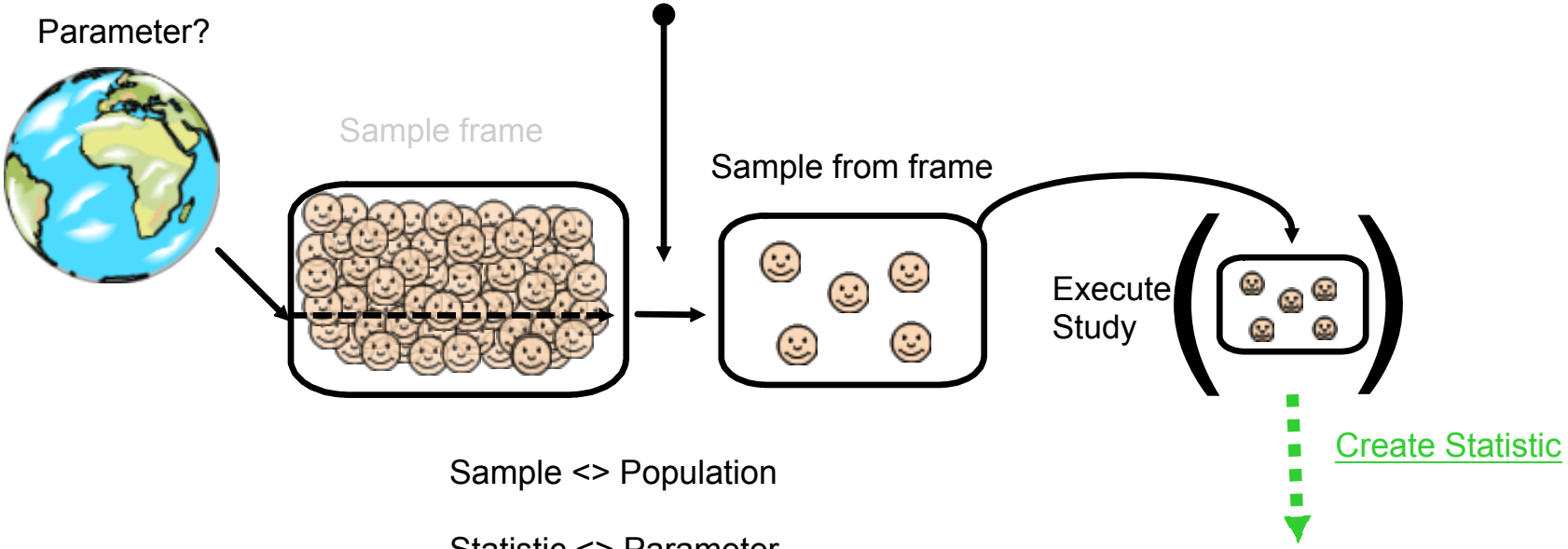
Processing errors

Random sampling errors
(Response error mistakes)



Stuff happens along the way.

Random sampling errors
- Without bias
- Reduced by greater n
- Measured by the margin of error.
- Only error in a perfect sampling process

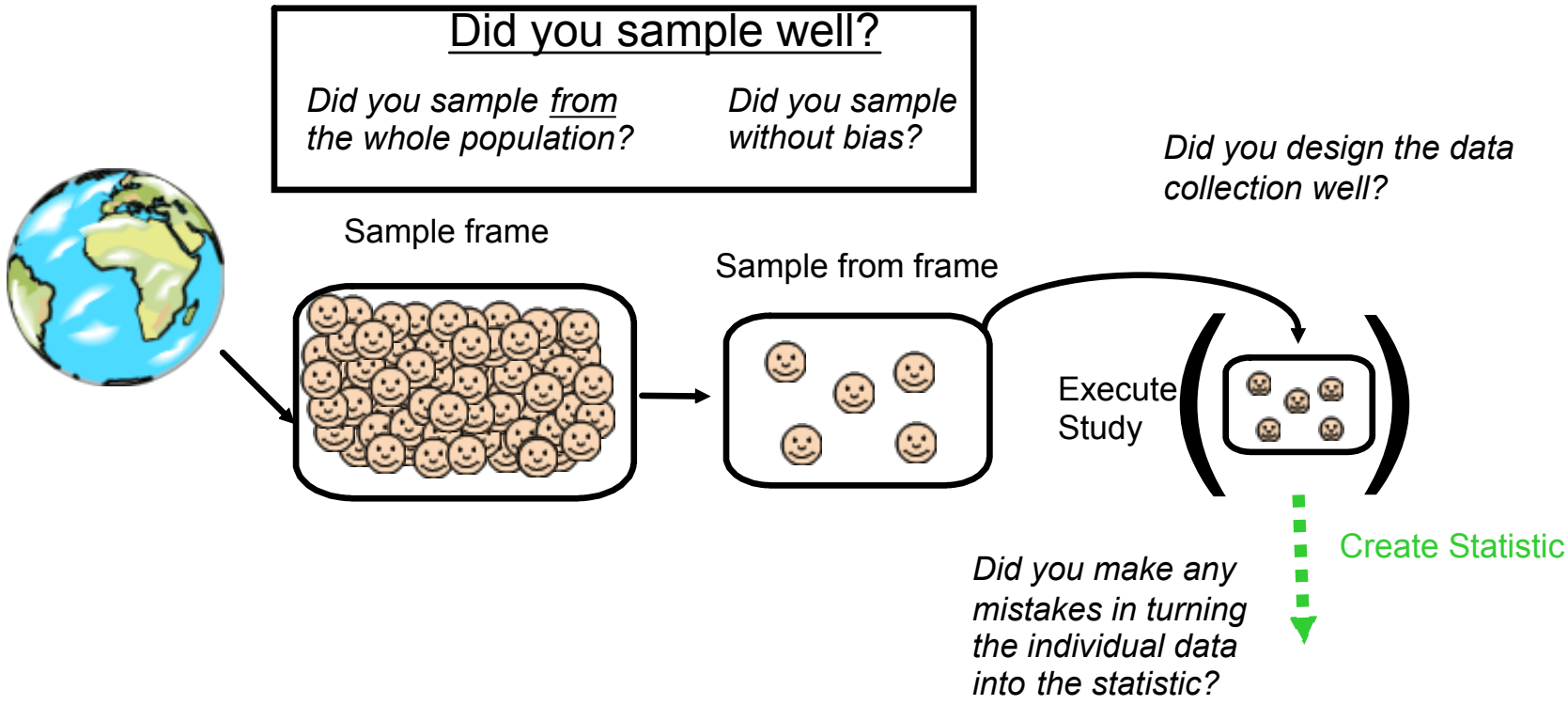


Sample <> Population

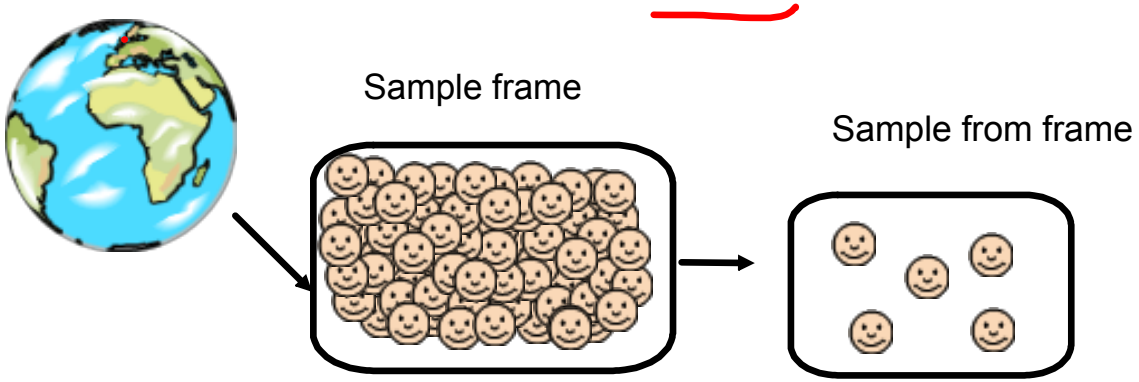
Statistic <> Parameter

but equally likely high or low if no bias

Did the study have errors with systematic effect?



Did you sample from the whole population?



Usually cannot match the population

Phone book
School rolls
Soc. Sec. #
Taxpayer rolls

IF there is anything systematic about what is missing then error arises

A SAMPLING ERROR

Undercoverage occurs when any group in the population is left out or is likely to be underrepresented in the sample.

Biasing
Expect Stat > Param
or
Expect Stat < Param

Did the study have errors with systematic effect?

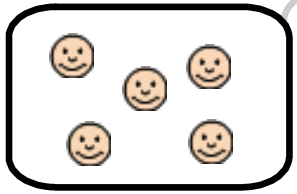
Did you sample well?

Did you sample <u>from</u> the whole population?

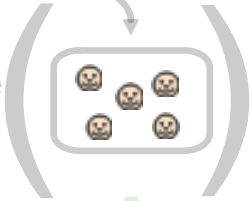
Did you sample without bias?

Sample frame

Sample from frame



Execute Study

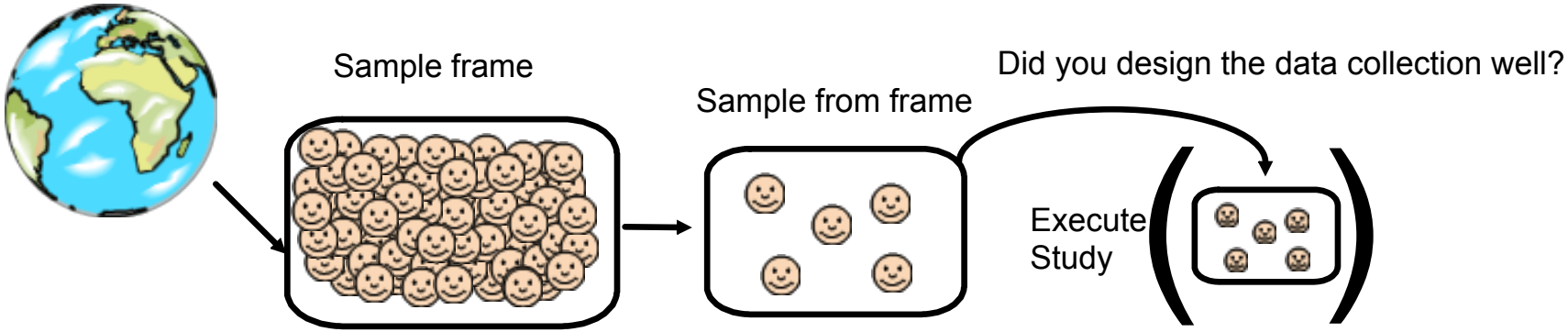


Create Statistic

This is what we discussed before.

- Do an SRS
- Don't do
  - <u>Convenience sample</u>
  - <u>Voluntary response</u>
  - Other

Nonsampling error, after the sampling...



1) Studies of humans must always give the person a choice.

In a survey opting out = Nonresponse

Can create systematic error/bias

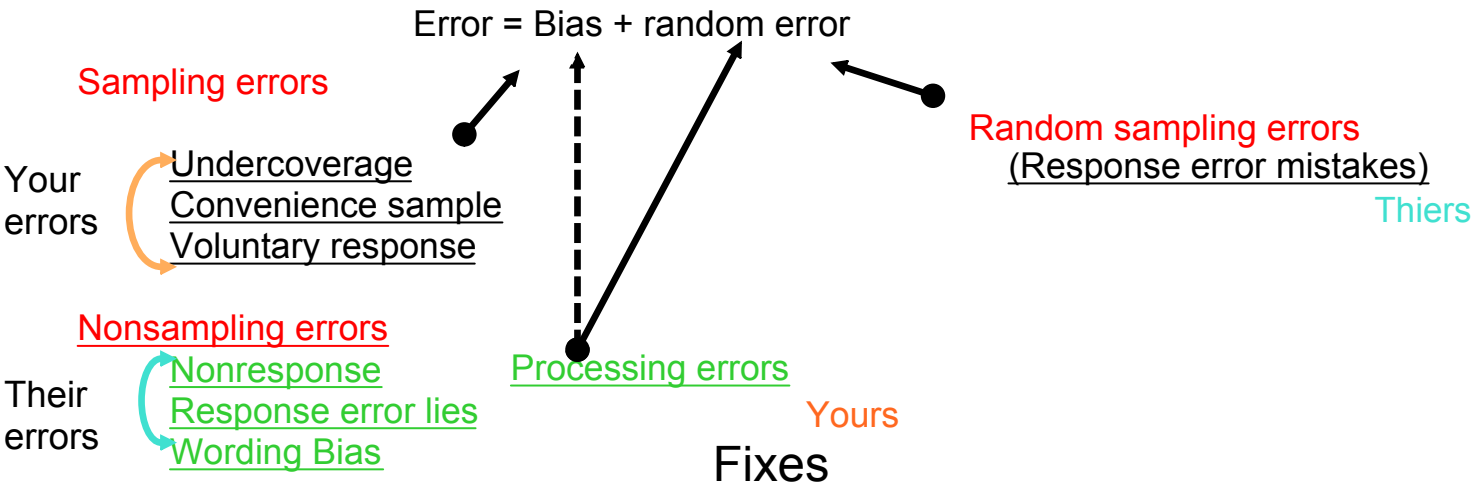
2) Response error
- Mistakes => random
- Lies => probably not

3) Design bias
- Wording
- Other

Targets for today:

- Review fixes: Design better or get a greater sample size
- Random errors are measured by margin of error
- Clarify wording problems:
    Can generate non-response.
    Can create bias
    Can just be bad
- Calculate response rate
- Learn stratified random sampling

Why do statistics vary the parameter?



Fixes

Fixes? Invest more money spend more time

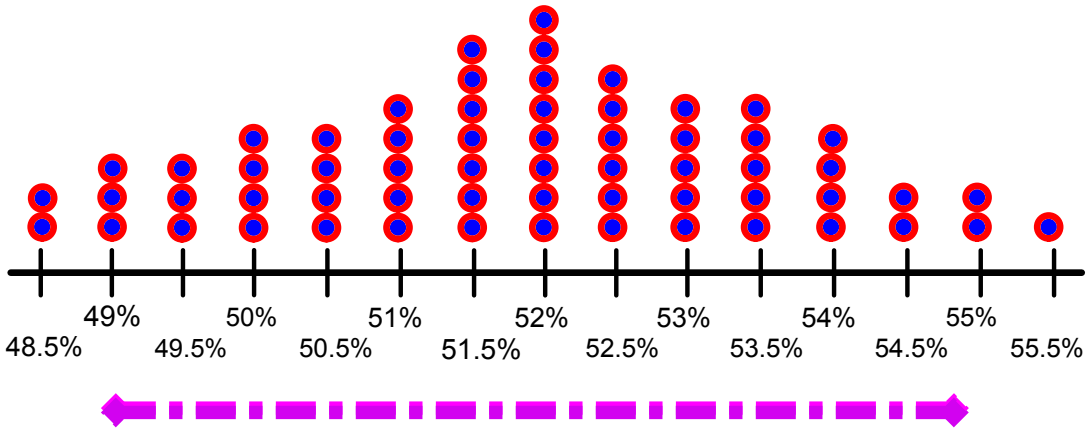
Your errors:

- Better design
- Invest in process

Their errors:

- Better design
- Better training
- More contact
- More information
- Research based adjustments to calculations

Let's say I took LOTS of surveys/samples



95% confident parameter is in here

Margin of error measures the random error.

$\frac{1}{\sqrt{n}}$ is the margin of error

Margin of error measures
- Random error
- Does not measure design problems

Error = Bias + random error

Your margin of error covers random error.

95% confidence, given only random errors, that....

95% confident parameter is in here

UNLESS....

When you say 95% confident, no you should be slightly less confident =>
someone may have goofed up non-randomly.

Digression: Calculating response rate.

$$\frac{\text{Number of Responses}}{\text{Sample Size}} = \text{Response rate}$$

Calculate the response rate:

The city planner wants to know how many commuters would switch from driving to using a new light rail. There are has 320,600 people commute to the city everyday. 233,119 commute by car. The planner takes an SRS of 10,000 people from the lists of registered car owners in all the communities from which people commute. 10,000 surveys are mailed. The following table accounts for the 10,000

206 report that the individual does not commute to the city
113 report that the individual commutes, but not by car
545 report that the individual would, at least, seriously consider light rail
388 report that the individual would not seriously consider light rail
12 were spoiled (physically marred or filled out in some unclear way).
98 were returned by the post office as undeliverable
8,638 were not returned
10,000 Total

Bias from non-response?
Pop of int?
Sample frame?
Sample?
Issues?

Badly designed questions:
- Imprecise language
- Unclear instructions for response
- Ignoring cultural differences
- Erudite diction
- Complex
- Biased word choice

Other

Wording bias

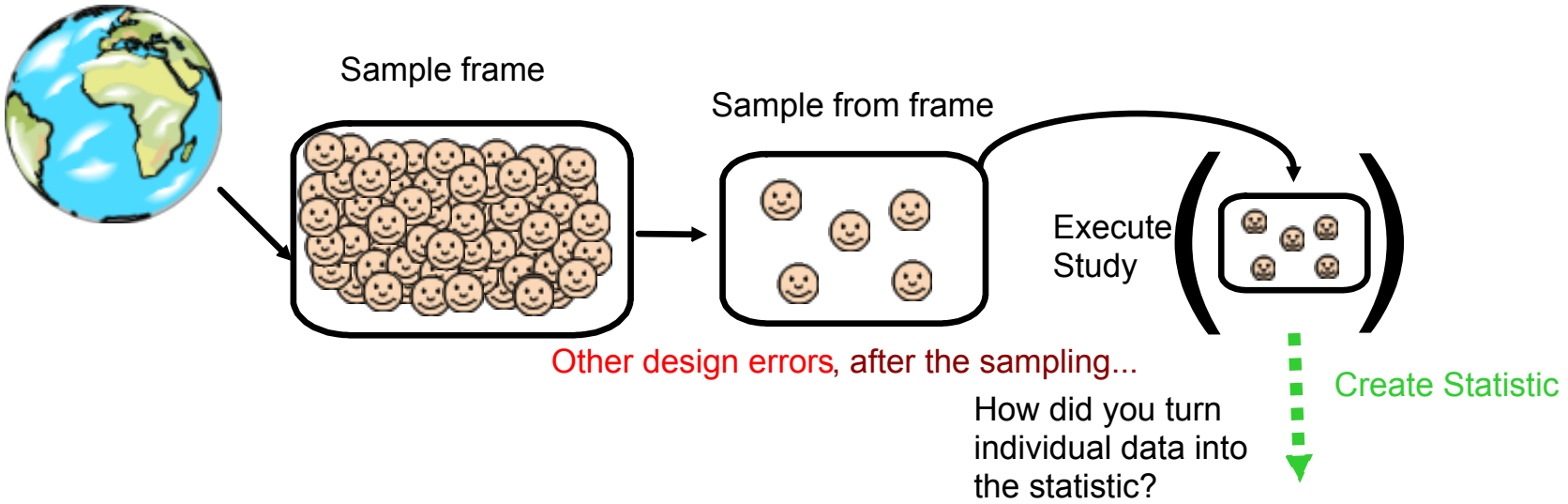
Do you favor continued disproportionate tax for the people with higher incomes?

Do you feel government interference in the free market improves the economy?

Which candidate in the current senatorial race do believe has more valuable experience. Senator James H. McIntyre or Willy Kingsley?

Did the study have errors with systematic effect?

- Did you sample from the whole population?
- Did you sample without bias?
- Did the individuals do their part?
- How did you turn individual data into the statistic?



Processing errors:

Assume unbiased (?)

Why do statistics vary from the parameter? Error = Bias + random error

Your margin of error covers random error.

95% confidence, given only random errors, that....

95% confident parameter is in here

UNLESS....

When you say 95% confident, no you should be slightly
less => Someone goofed up non-randomly.

Let say you bet Pat $50 that over 30% of high school students would agree with the following statement:

"Freshman girls shouldn't waste their time dating Freshmen boys"

You and Pat agreed that Fran, a stat student, was 100% honest and could do a fair survey based on an SRS. He was going to poll 10 students.

Your Prediction:	Girls	Boys
Freshmen	Agree	Disagree
Sophomore	Disagree	Agree
Junior	Disagree	Agree
Senior	Disagree	Agree

You think the table works so you are pretty confident.

But what really unlucky thing could happen in the SRS?

By dumb luck, the SRS could easily pick more from the groups you are pretty sure disagree.

"Freshman girls shouldn't waste their time dating Freshmen boys"

	Girls	Boys
Freshmen	Agree	Disagree
Sophomore	Disagree	Agree
Junior	Disagree	Agree
Senior	Disagree	Agree

What instructions do you want to give Fran in conducting the survey?

Do an SRS BY GROUP in the table.
Pick, say 3, in each group.

A stratified random sample

When you can identify how bad luck
in conducting an SRS
could give you a predictably different statistic
from the population parameter

an SRS may not be appropriate

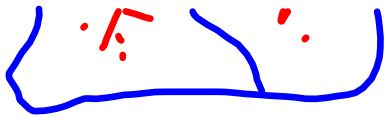
When some feature of a population divides it into groups
where each group
may have a predictably different statistic
from the population parameter

a stratified random sample may be appropriate

Why might you stratify in each following case? (Think bad luck)
What would your strata be?

"Only FLHS seniors should get parking passes" in FLHS
Polling national support for Barack Obama.
Asking if there still exists a "glass ceiling" for women in the workplace.
Surveying Fairfield about proposed zoning law changes.
Studying support for Red Sox or Yankees in CT.

Hmm...
Size of strata..
The same number in each.
Or match their % of the whole.
Or more complex.



Issues:

1) "stratified random sample may be appropriate"

Can involve complex issues.

With a big enough sample chance are the SRS will have the right distribution.

BUT the sample may need to be so large:
- You can't afford it
- You might as well do a census

2) You may be able to name strata you want to make, but you can only find out who belongs to what strata during the survey.

Chicken vs. Egg

Stratified random samples is just one approach to more sophisticated sampling.

Probability samples: Any sampling process based on random selection of individuals.

Another is cluster samples (part convenience).

1) Is there an issue with this study?
2) Identify the issue and its impact on the study. Use the correct terminology.
3) Measured by the margin of error or not?

---

1) Robin wanted to study people's attitudes towards volunteering. She hired people to go door to door Monday to Friday 9:00 to 5:00 and survey people about volunteering.

2) Pat wanted to study attitudes towards the Occupy Wall Street Movement. He asked, "Do you support the protesters against corporate greed now on wall street?"

3) Pam want to understand people's level of fitness. She asked, "How much of the day are you active?"

4) Jim wanted to understand people's investment choices. He asked, "Given your current income level, future spending requirements, and demonstrable level of risk aversion, do you believe that equity instruments are an appropriate part of your investment portfolio or do you favor another, specific type of investment?"

The relationship of the width of the confidence interval to confidence level

Design errors

Undercoverage
Convenience sample
Voluntary response

Nonresponse response rate
Response error
Wording Bias

Processing errors

Random sampling errors

Stratified random sample
- What?
- Why?
- How?
- Strata