Sampling

Who gets studied, and who doesn’t?

How can we know anything about large groups of people?

There are roughly 8 million New Yorkers (20+ million metro area), and over 300 million Americans.

Even if you want to know something about just CCNY students, there are 17,000 of you. That's a lot of interviewing.

Census

A study of the complete population of interest.

By constitutional law, the U.S. government must conduct a census every 10 years to determine population size for congressional representation (among other things).

So they give a survey to every single resident of the country (or they try to).

Similarly, the registrar at CCNY has information on every registered student, which can be considered a census of that population.

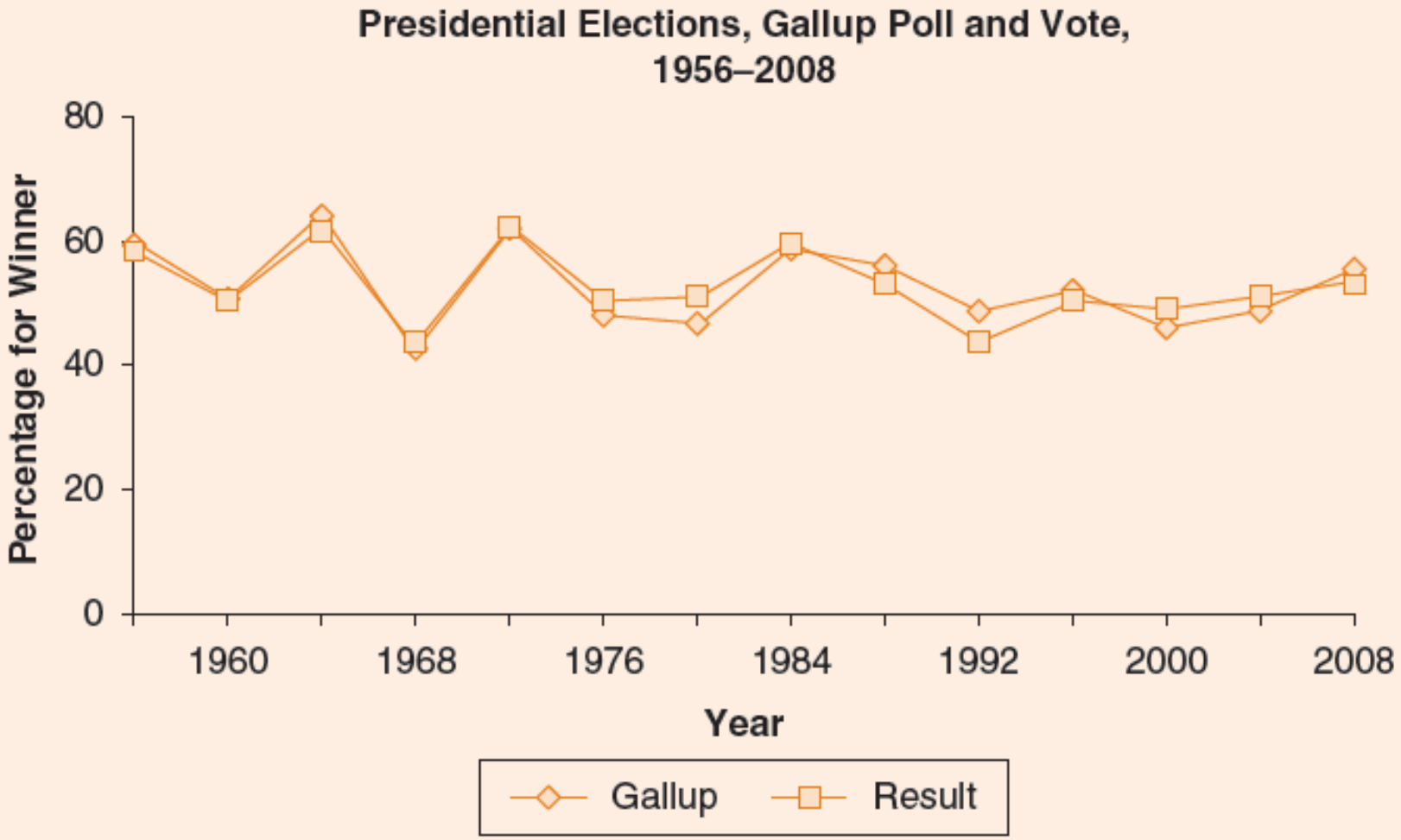
But Censuses are impractical for most research on big populations.

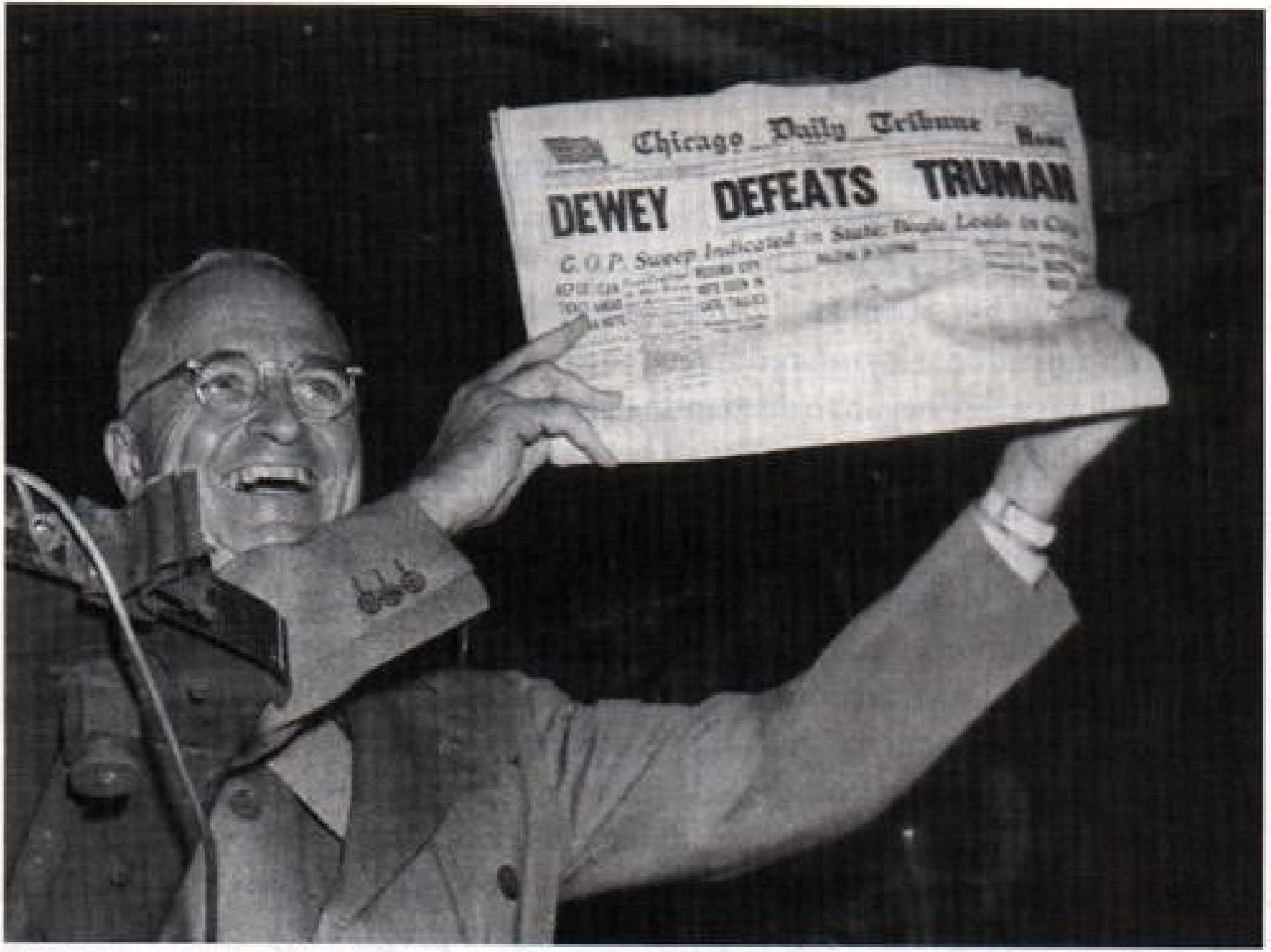
Sampling allows you to make generalizations to the entire population from just a part of it.

By talking to just 1,000 Americans, you can get a good idea about the entire country… if those 1,000 Americans are a **representative sample.**

Need for small samples

EXHIBIT 5.3 Election Outcomes: Predicted and Actual




Chicago Daily Tribune
DEWEY DEFEATS TRUMAN

Probability Sampling

Everyone in the population has a chance to be in the sample, and you know what that chance is.

(A census is a probability sample where everyone has 100% chance of inclusion)

A simple random sample is the purest form of this: Everyone has an equal chance of being in the sample.

Example

Every CCNY student flips a coin, and you only talk to those who get heads. They all have a 50% chance of being in the sample, which will be roughly 8500 students.

Why this works

Sampling Frame

Is your “list” or source of members of the population.

If this is incomplete in any way, that could introduce bias.

So phone numbers in the U.S. in 1948 was an incomplete and biased sampling frame.

Sampling Bias

Anything that makes the sample unrepresentative.

If one characteristic, attribute, or group is more likely to get in the sample than their true proportion of the population, then that's sampling bias.

Non-response Bias

Imagine if I wanted to find out how busy CCNY students were. So I got a list of their phone numbers and called a random sample of them, but only 50% agreed to answer my survey.

What potential bias should I be concerned about?

Other Examples

Systematic Random Sampling

Get a list of your population and pick every Nth member.

For instance, every 170th CCNY student from the roster would be a 100 person sample.

But you have to be careful about **periodicity** in the data.

For instance, if you pick each 10th student in a class listed by seating order, what might be the problem with that?

Cluster Sampling

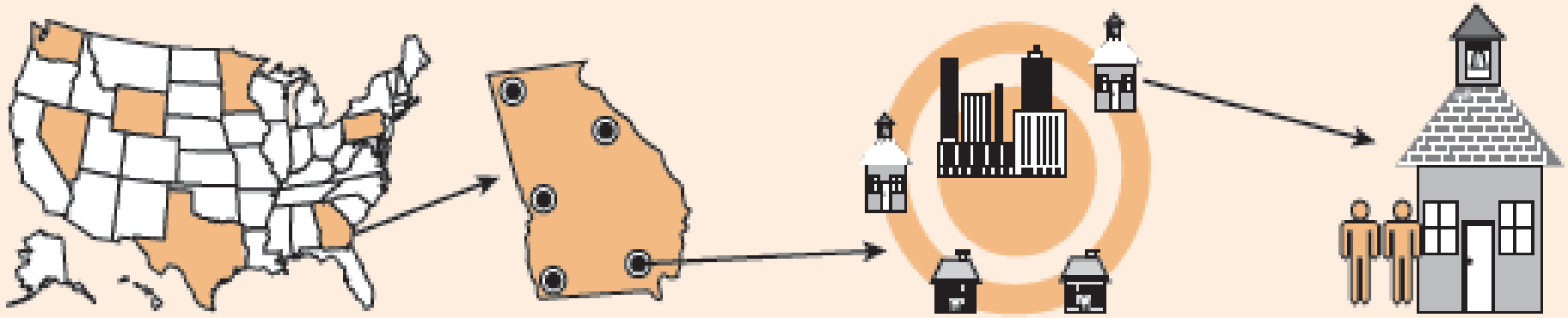
Sometimes its easier to sample within samples.

Students within schools.

Voters within cities within states.

Employees within Divisions.

EXHIBIT 5.5 Multistage Cluster Sampling



Stage 1: Randomly select states

Stage 2: Randomly select cities, towns, and counties within those states

Stage 3: Randomly select schools within those cities and towns

Stage 4: Randomly select students within each school

Clustering

Its only a probability sample if you randomly select each level of the sampling.

Nielsen Sampling

Daily Metering: 25,000 households
Set-top meters
People Meters

Sweeps Diaries: 1.6 million handwritten paper diaries of TV viewing (only during Sweeps Weeks)

Address based sampling using Census defined areas

Stratified Random Sampling

Sampling by attributes or characteristics.

Disproportionate Stratified Sampling is useful if you want to make sure you get a large enough number or responses from a minority group within the population.

Its only random if sampling is random within each strata, and you know the population sizes of each strata.

Examples

Comparing Hetero and Same-Sex Couples

Studying school children in the U.S., with an emphasis on ethnic minorities.

Non-probability Sampling

Convenience or Availability Sampling

Quota Sampling
Purposive Sampling

Snowball Sampling

Quota Sampling didn't work in 1948

You can guarantee that your sample is representative in the ways you quota for, but not in other ways.

Purposive or Theoretical Sampling

Pick cases that represent extremes of an important concept.

Pick cases that test out a theory.

“A case in point…”

Snowballing

Overcomes problems of trust or access to a population.

Respondents refer you to new respondents, vouch for you to them.

But you don't know whether this network of people represents the population of interest.

Volunteer Bias

Similar to Non-Response Bias

Magazine Surveys
Cable News Surveys

Kinsey

Bias Only Matters When…

It is related to a variable you care about.

Its not enough to say that a sample is biased. A good critique should say how it could be biasing the results of the study.

Busy people less likely to be surveyed, so what?

If busy people are more likely to vote Republican, then surveys will underestimate that.

You can never know all of the ways sample bias might be related to your variables

- Good probabilistic sampling with high response rate is the best hedge against unexpected sampling bias.

Sampling Can Go Wrong when...

The **Sampling Frame** is unrepresentative
(how you pick people)

There is **(Non) Response Bias**
(how people choose to participate)

But other things can make a survey/poll give you misleading information: How you ask questions, and how people answer them.
(which we covered in the Survey portion of the course)

Let's critique some Non-probability Samples

How Couples Meet Interviews
from Craigslist Ad

Customer Satisfaction Surveys

Street Surveys

Shopping Mall Surveys

Email List Recruitment

Critiquing Political Poll

Landlines only or Cell phones too?

Automated or live interviewers?

When did they call? Did they call back if NA?

How did they determine who is a Likely Voter?

How did they weight their sample to match US?

Did they get very low N of some demographics?

How did they ask the question?

What were the answer choices?