Section 3.3 - Sampling Design

Statistics 104

Autumn 2004

Sampling Surveys

Used in

- Political polling
- Consumer preferences
- Product monitoring

Terminology:

- Population: Entire group of objects or people about which information is desired.
- Units: Individual members of the population
- Sample: Part of the population that is actually examined.

- Census: Survey in which all members of the population are studied.

Why sample?

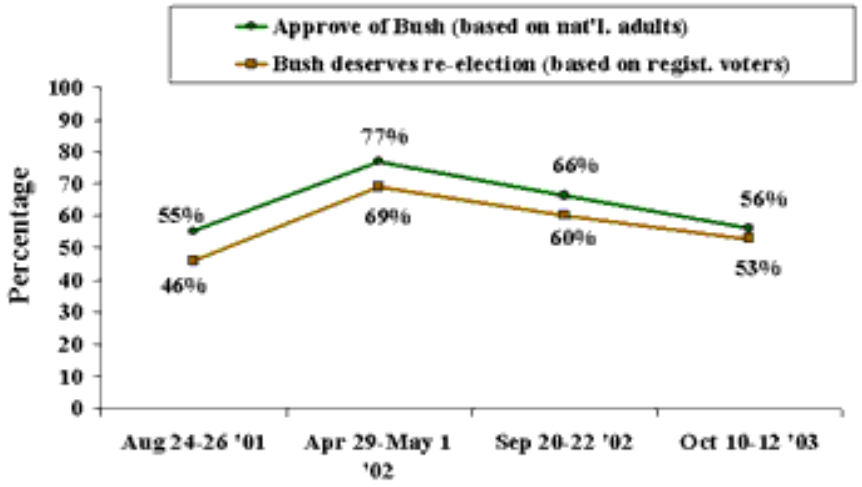
- Cheaper
- Can be more accurate, as there can be more opportunities for more errors in censuses
- Can quantify the size of the errors due to sampling

Example: Political Preferences

(Taken from Gallup Poll released Oct 22, 2003)

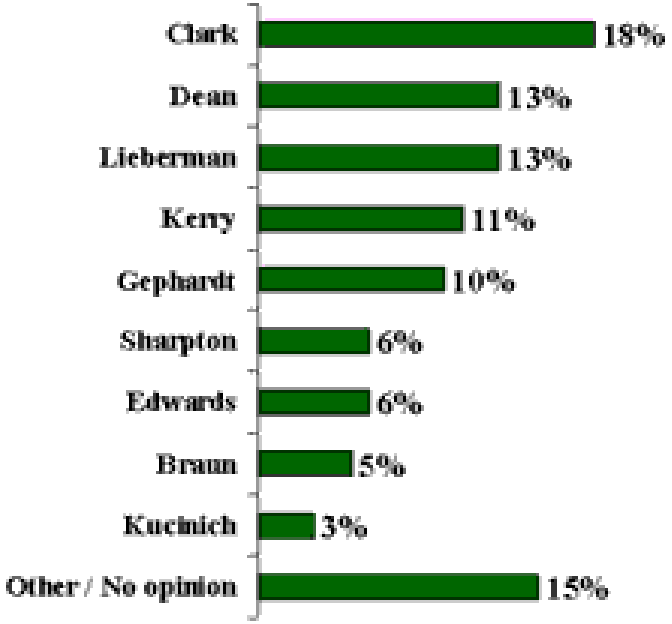
2. Please tell me whether you think each of the following political office-holders deserves to be reelected, or not. How about – [ITEM A READ FIRST, THEN ITEMS B-C ROTATED]?

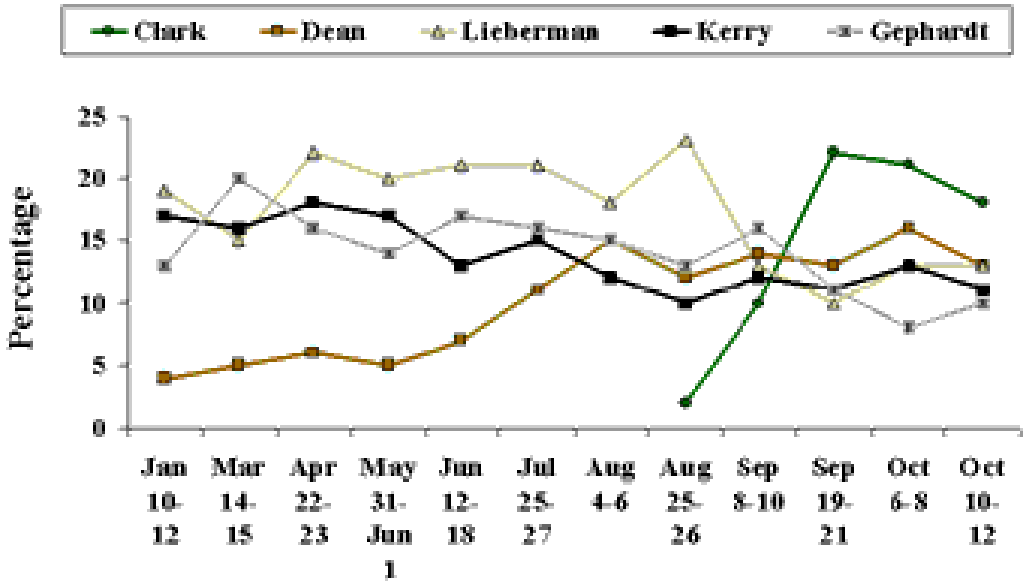
A. President Bush
B. The U.S. Representative in your Congressional District
C. Most members of the U.S. House of Representatives



6. Next, I'm going to read a list of people who may be running in the next election. After I read all the names, please tell me which of those candidates you would be most likely to support for the Democratic nomination for President in the year 2004. [ROTATED: *Names of candidates*]

BASED ON – 388 – DEMOCRATS OR DEMOCRATIC LEANERS WHO ARE REGISTERED TO VOTE





Also available, but not shown here: BASED ON – 456 – DEMOCRATS OR DEMOCRATIC LEANERS

Survey Methods (From Gallup Poll conducted Oct 10-12)

These results are based on telephone interviews with a randomly selected national sample of 1,004 adults, 18 years and older, conducted Oct. 10-12, 2003. For results based on this sample, one can say with 95% confidence that the maximum error attributable to sampling and other random effects is ± 3 percentage points.

The results for black Democrats are based on 205 interviews conducted in polls Sept. 19-21, Oct. 6-8, and Oct 10-12, 2003. For results based on this sample, one can say with 95% confidence that the maximum error attributable to sampling and other random effects is ± 8 percentage points.

The results for white Democrats are based on 1,075 interviews conducted in polls Sept. 19-21, Oct. 6-8, and Oct 10-12, 2003. For results based on this sample, one can say with 95% confidence that the maximum error attributable to sampling and other random effects is ± 3 percentage points.

In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.

Example: Ann Landers asked her readers

> "If you had to do it again, would you have children?"
>
> Sample size $n \approx 10,000$
>
> $\approx 70\%$ said No

A sample survey done by a polling organization (don't know which one) got a result of 91% yes.

What happened?

> Only people with strong feelings responded to Ann Landers.

This is typical of **Voluntary Response** samples.

In this case, the study was biased towards people who would say no.

We want a mechanism to select the sample without favouritism of the researcher or self selection by the responder.

Answer: Random Sampling

Select units to be studied by a random mechanism.

Example: 1983 Columbus Dispatch Poll

Statewide initiative to raise the legal drinking age in Ohio from 18 to 21.

How the poll was done

1. Got list of 5.8 million registered voters. Included people who had voted in the 1980 Presidential election or the 1978 election for Governor.

2. Took a random sample from this list of 6761 "voters".

3. 1658 questionnaires returned and analyzed.

Poll results: 56% of the 1658 said they favoured raising the drinking age

Referendum results (conducted in late 1983 or 1984): 41% of voters favoured raising the drinking age.

Sources of bias:

- The list of voters used for the poll excluded 18, 19, and 20 year olds.

  This is an example of undercoverage (selection bias). Young people, who probably would vote no, weren't sampled.

- Notice that only 25% of the people picked actually responded – How might they differ from the 75% who didn't respond. (This is an example of nonresponse.)

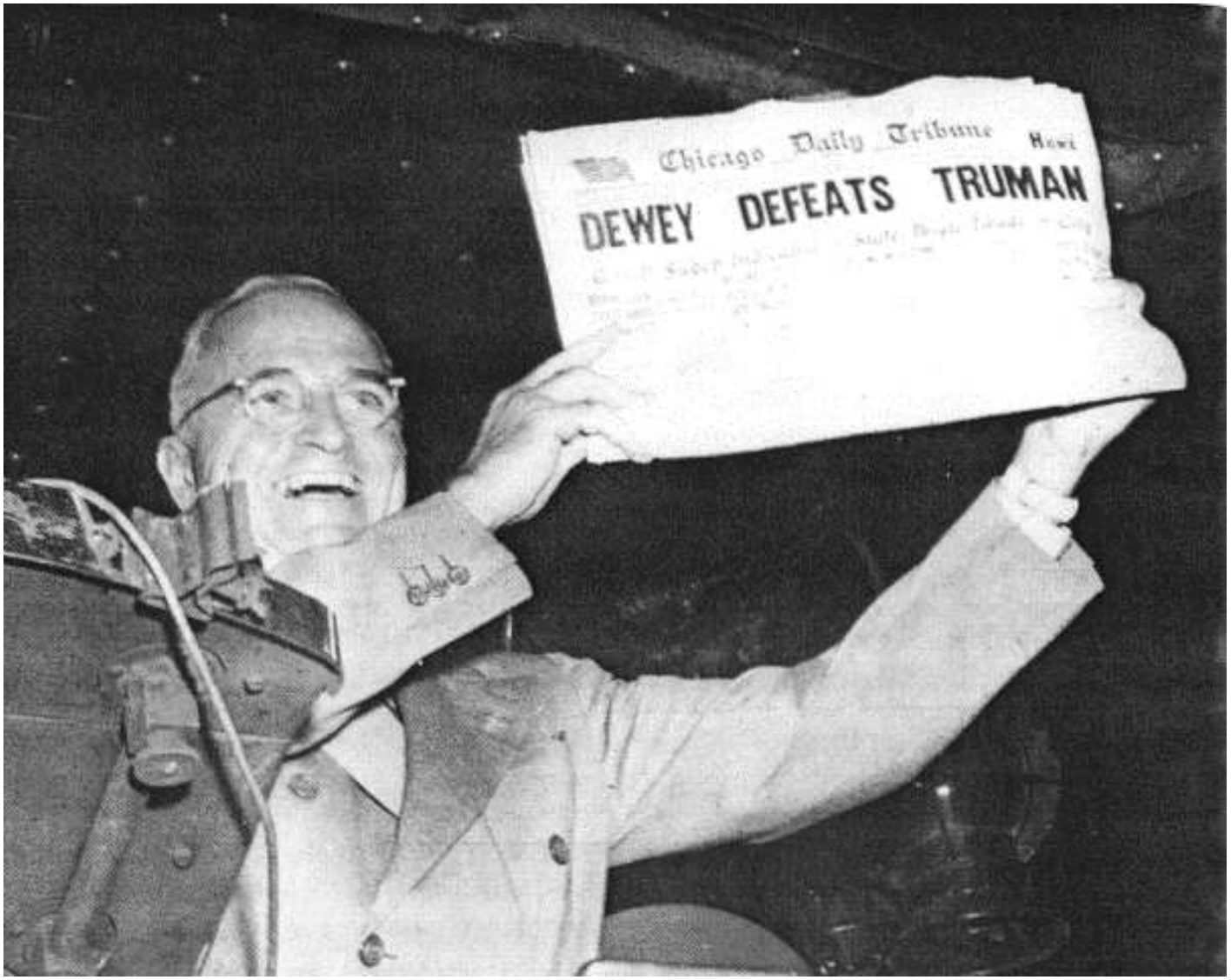
  One possibility – Respondents might have a stronger opinion (interest in the survey)

  In the Dispatch poll

  – 64% said they had a very strong opinion – 68% of them favoured raising the drinking age.
  – 30% said that they had a fairly strong opinion – 46% favoured the issue.
  – 6% might change their mind – 38% in favour.

One possibility: people with strong opinions gave the issue a lead in the poll while those with less strong opinions defeated the issue in the referendum.

Example: Predicting the 1948 Presidential Election



This headline was partially based on polling data which suggested that Dewey would beat Truman.

	Roper	Crossly	Gallup	Election
Truman (Democratic)	38%	45%	44%	50%
Dewey (Republican)	53%	50%	50%	45%
Others	9%	5%	6%	5%

Others included Strom Thurmond (State' Rights) who won 4 states (39 electoral votes) and Henry Wallace (Progessive).

In these polls, there was a greater chance for a Republican to be sampled than a Democrat, which skewed the polls towards Dewey. Also the polls were done about a week prior to election day and there is fairly good evidence that they was a drift towards Truman in the final week of the campaign.

Other sources of bias

- Question wording

  – Wording might suggest answer

  Example: Auckland University survey

  2 forms for 2 questions

  Version 1

  1. "Are you in favour of giving special priority to buses in the rush hour?"
  2. "Do you think the cost of catching a bus into university is too high?"

  Version 2

  1. "Are you in favour of giving special priority to buses in the rush hour, or should cars have just as much priority as buses?"
  2. "Taking into account the problems and costs associated with parking, do you think the cost of catching a bus into university is too high?"

	Question 1	Question 2	Sample Size
Version 1	68% yes	75% yes	585
Version 2	47% yes	63% yes	569

– Order of choices (Its why Gallup rotates Democratic candidates in question)

– Asking the correct question

- Survey format

  – Mail versus phone versus personal interviews

  Moore (1979) described a study investigating whether people favoured contraceptives being made freely available to unmarried women.
  44% of people surveyed by personal interviews said yes, in contrast with 75% of those questioned by phone or mail

- Response bias / Interviewer effects
  - Interviewers tone
  - Other questions

  Example: race of interviewer in a study involving racial issues.

- Interviewer control
  - Are instructions followed accurately?

- People changing their minds (particularly in political polls)

Sampling Methods

Method to be discussed are all probability sampling schemes

Probability Sample:

A sample chosen by chance. We must know what samples are possible and what probability each possible sample has.

Example: Want to sample 2 units from a population of 6.

- Possible samples: $S_1 = \{1, 2\}, S_2 = \{3, 4\}, S_3 = \{5, 6\}$
- Probabilities: $P[S_1] = P[S_2] = P[S_3] = \frac{1}{3}$

This scheme has the (potentially) nice property that

$$P[\text{Unit } k \text{ sampled}] = \frac{1}{3}$$

However this scheme leaves out other possible samples, e.g. $\{2, 5\}$.

The basis of many sampling schemes is the Simple Random Sample (SRS).

Simple Random Sample

> A Simple Random Sample of size n consists of n units of the population such that every set of n units has an equal chance of being selected.

Note: every unit has an equal chance of being selected

$$P[\text{Unit } k \text{ sampled}] = \frac{n}{N}$$

where N is the population size.

The possible SRS samples for this example are:

	1	2	3	4	5
2	$\surd$				
3	$\surd$	$\surd$			
4	$\surd$	$\surd$	$\surd$		
5	$\surd$	$\surd$	$\surd$	$\surd$	
6	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$

Each sample has probability $\frac{1}{15}$ of being chosen.

Example: What is the average salary of Harvard grads 5 years after graduation.

Could take a sample of 1999 grads.

However there are probably large differences from major to major. Want to have each major represented in the sample.

Stratified Random Sample

The population is divided into groups of similar units, known as strata. Then a SRS is selected from each strata.

For example,

- n_1 of N_1 English majors
- n_2 of N_2 Math majors
- n_3 of N_3 Geography majors
- etc

$$P[\text{Unit } k \text{ from major } i \text{ sampled}] = \frac{n_i}{N_i}$$

Note that $\frac{n_1}{N_1}, \frac{n_2}{N_2}, \frac{n_3}{N_3}$, etc could all be different. Can adjust for this in the analysis.

Another example where stratified sampling is used in in auditing. Suppose you were asked to look for the accuracy of insurance claim payouts. Often you will stratify based on the amount of the payout. There size of possible errors / fraud can be much higher for the large claims, though there won't be as many of them.

Strata are like blocks in experimental design.

Multistage Sampling

Example: Determining agricultural yields

Want samples from 1000 corn fields. Could take a SRS of all registered farms in the country (assuming that the list exists). Could lead to lots of travelling. Another approach would be to take a SRS of 250 counties that have corn fields. Then for these 250 counties, take an SRS of 4 farms within each of these counties.

Units can be classified by a set of nested criteria. A SRS is taken of groups in the broadest criteria. Then a SRS is taken at the next criteria, only looking at the groups selected initially. This is repeated until units are selected.

Cluster Sampling

Example: How many dogs of in the city of Cambridge?

One approach would be to take a SRS of city blocks. Then sample every household in the selected blocks.

Population is divided into groups (often geographically based). A sample of groups (called clusters) is taken. Then all units in the cluster are sampled. The clustered selected could be determined by any of the previously discussed methods.

For example, it might make more sense to stratify the blocks based on zoning (residential, retail, etc) and choose a stratified sample of blocks.

Systematic Sample

Units have some natural order (e.g. list of names in the phone book). Sample every kth unit.

One example of this is the Census long form (I believe). Every kth household gets the long form questionnaire which asked detailed questions (k is about 6, though it does vary by region, ranging from 2 to 8). To see the long form, go to <http://www.census.gov/dmd/www/pdf/d02p.pdf>