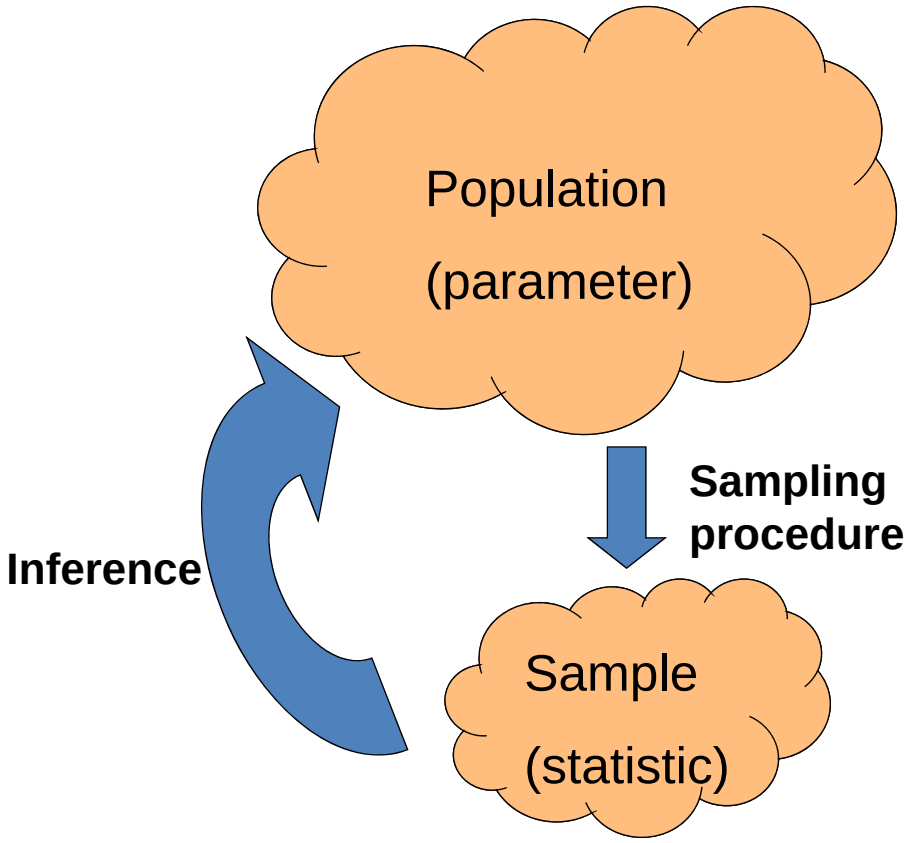
- **Chapter 19: Sample Surveys**



- **Definitions**

- ***Population*** – the group of individuals of interest in the survey.
- ***Sample*** – part of the population, chosen to give information about the population.
- ***Parameter*** – a numerical fact about the population. Unknown – can only be estimated.
- ***Statistic*** – a numerical fact about the sample, used to estimate the corresponding fact about the population. Known – can be computed from the sample.

- **Example**

- ***Population*** – USU students registered for stat 1040 this semester.
- ***Sample*** – a group of stat 1040 students selected from the population.
- ***Parameter*** – the percentage in the population who prefer weekly quizzes over homework.
- ***Statistic*** – the percentage in the sample who prefer weekly quizzes over homework.

- **Representative Samples**

- Need a ***representative*** sample – one that is like the population in the ways that matter. A good sampling procedure will be fair and impartial.

- Estimating population parameters is only justified when the sample is representative.

- The method of choosing the sample matters a lot.

- The best methods of choosing a sample involve the planned use of probability.

- **What Can go Wrong?**

- ***Selection bias*** – a systematic tendency to exclude some kinds of people. E.g. tendency to exclude the rich, the poor, the homeless, the young, the old, etc.
- ***Nonresponse bias*** – the kinds of people who respond differ from those who do not respond. E.g. people who are opinionated about the issue are more likely to respond.
- If the people who tend to be excluded or who don't respond differ in important ways from the rest of the population, the sample will not give good estimates.

- **The Literary Digest Poll**

- 1936, Roosevelt versus Landon.

- Campaign centered on economic policies.

- **The Literary Digest Poll**

- 1936, Roosevelt versus Landon.
- Campaign centered on economic policies.
- The *Literary Digest* took the largest sample ever - 2.4 million people, and predicted:

  - Roosevelt will only get 43% of the vote.

Roosevelt won by a landslide!

(Literary Digest went bankrupt)

- **What went wrong in 1936?**

- The Literary Digest sent questionnaires to 10 million people.
- The names of the 10 million came from telephone books and club membership lists.
  - **SELECTION BIAS!**
- They got responses from 2.4 million people.
  - **NONRESPONSE BIAS!**
- They tended to exclude the poor, and for the first time in history, the poor tended to vote Democrat.

• Gallup: the new kid on the block

- Gallup sampled 3000 people at random, from the same lists the Digest used, and predicted their prediction!
- He sampled 50,000 people in a special way and made his own prediction.

Table 1. The election of 1936.

	Roosevelt's percentage
The election result	62
The *Digest* prediction of the election result	43
Gallup's prediction of the *Digest* prediction	44
Gallup's prediction of the election result	56

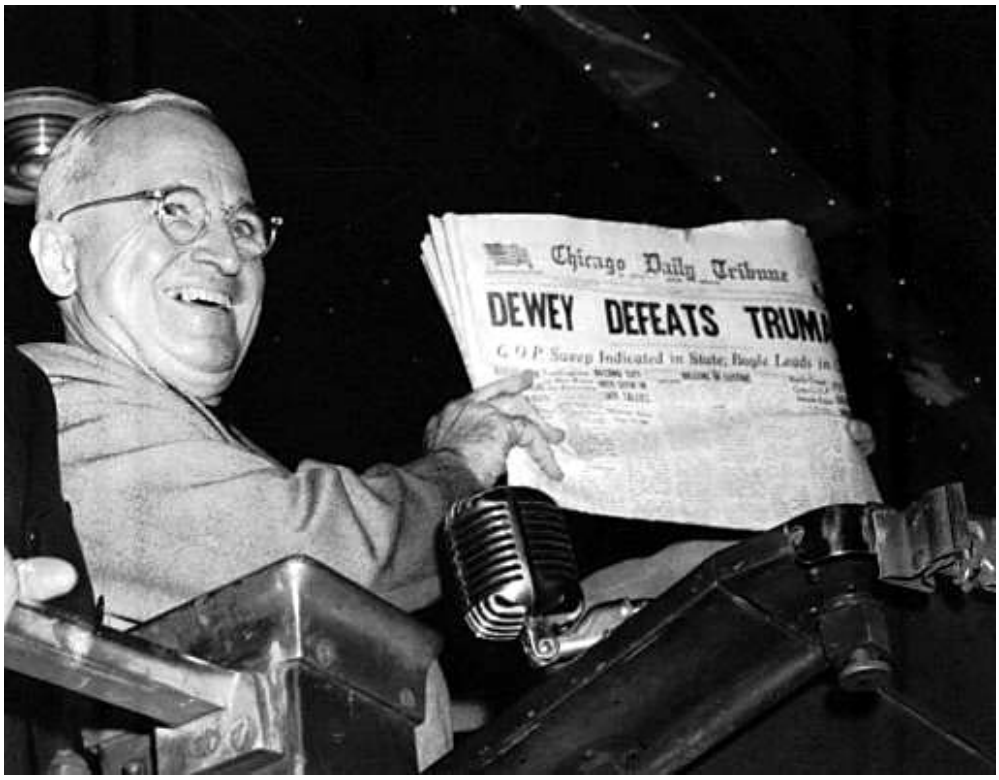
- **Bigger is better?**

- If a sample is representative, a large sample is better than a small sample because it gives more precise estimates of the population parameter.

- BUT…

- If the sampling procedure is biased, taking a large sample does not help. This just repeats the basic mistake on a larger scale!

- **1948: The Year the Polls Elected Dewey**

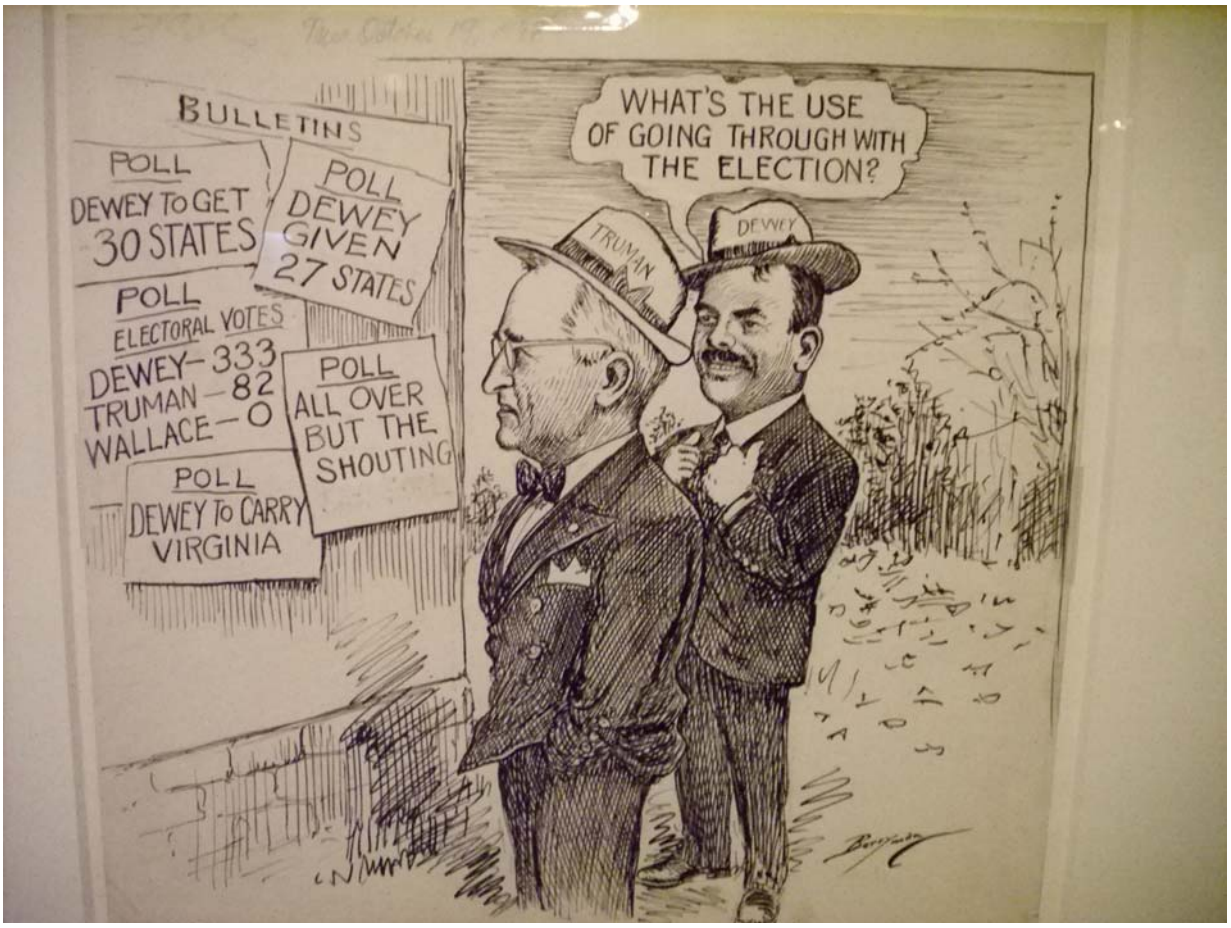


Truman  Dewey 

Table 2. The election of 1948.

The candidates	*The predictions*			*The results*
	Crossley	*Gallup*	*Roper*	
Truman	45	44	38	50
Dewey	50	50	53	45
Thurmond	2	2	5	3
Wallace	3	4	4	2

- **Even Utah voted for Truman!**
- **Utah: Truman (Dem): 149,151 54%**
- **Dewey (Rep): 124,402 45%**


BULLETINS
POLL
DEWEY TO GET 30 STATES
POLL
DEWEY GIVEN 27 STATES
POLL
ELECTORAL VOTES
DEWEY—333
TRUMAN—82
WALLACE—0
POLL
ALL OVER BUT THE SHOUTING
POLL
DEWEY TO CARRY VIRGINIA
WHAT'S THE USE OF GOING THROUGH WITH THE ELECTION?
TRUMAN
DEWEY

• What Went Wrong?

- Gallup used “Quota Sampling”:
- Each interviewer was assigned a quota of subjects to interview.
- They were told how many subjects had to be from certain categories (residence, age, sex, race, economic status)
- The interviewers could select anybody they liked as long as their subjects satisfied the specified criteria.
- Example: 6 from the suburbs, 7 from the city
- 7 men, 6 women
- of the 7 men, 3 had to be under 40, 4 over 40
- of the 7 men, 6 had to be white, 1 black

• Quota Sampling

- In quota sampling, the subjects are hand-picked to resemble the population with respect to some key characteristics.

- Quota sampling SEEMS reasonable because it ensures that the sample will resemble the population with respect to some of the important characteristics related to voting behavior.

- BUT: quota sampling does not work very well due to unintentional bias on the parts of the interviewers.

- **Quota sampling tends to exclude Democrats!**

Table 3. The Republican bias in the Gallup Poll, 1936–1948.

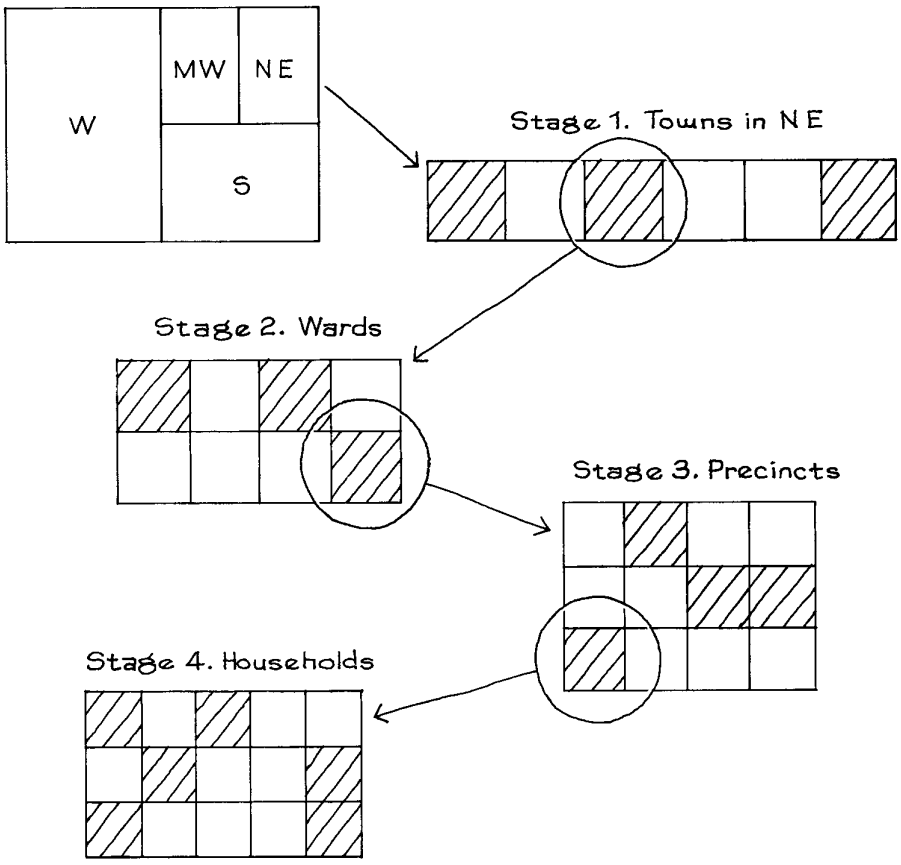
Year	*Gallup's prediction of Republican vote*	*Actual Republican vote*	*Error in favor of the Republicans*
1936	44	38	6
1940	48	45	3
1944	48	46	2
1948	50	45	5

Note: Percentages are of the majority-party vote, except in 1948.
Source: F. Mosteller and others, *The Pre-Election Polls of 1948* (New York: Social Science Research Council, 1949).

• Probability Methods

- Probability methods use objective chance procedures to select samples. They guard against bias because they leave no discretion to the interviewer.
- One probability method is ***simple random sampling***. This means drawing subjects at random without replacement.
- Another probability method is ***cluster sampling***. This means that clusters (eg households) are selected at random and all the individuals in the selected clusters are sampled.
- A ***stratified sample*** divides up the population into groups based on an important variable (eg age) and samples separately from each stratum.
- Most large samples use ***multistage cluster sampling***.

Figure 1. Multistage cluster sampling.



- **The Success of Probability Methods**

Table 4. The Gallup Poll record in presidential elections after 1948.

Year	*Sample size*	*Winning candidate*	*Gallup Poll prediction*	*Election result*	*Error*
1952	5,385	Eisenhower	51%	55.1%	4.1%
1956	8,144	Eisenhower	59.5%	57.4%	2.1%
1960	8,015	Kennedy	51%	49.7%	1.3%
1964	6,625	Johnson	64%	61.1%	2.9%
1968	4,414	Nixon	43%	43.4%	0.4 of 1%
1972	3,689	Nixon	62%	60.7%	1.3%
1976	3,439	Carter	48%	50.1%	2.1%
1980	3,500	Reagan	47%	50.7%	3.7%
1984	3,456	Reagan	59%	58.8%	0.2 of 1%
1988	4,089	Bush	56%	53.4%	2.6%
1992	2,019	Clinton	49%	43.0%	6.0%
1996	2,895	Clinton	52%	49.2%	2.8%
2000	3,571	Bush	48%	47.9%	0.1 of 1%
2004	2,014	Bush	49%	50.6%	1.6%

Note: The percentages are of the popular vote. The error is the absolute difference "predicted – actual."
Source: The Gallup Poll (American Institute of Public Opinion) for predictions; *Statistical Abstract*, 2006, Table 384 for actuals.

• Probability Methods

- Probability methods attempt to minimize selection bias, but there are still problems due to:

- nonresponse bias
- badly asked questions
- interviewer control
- talk is cheap