19

Sample Surveys

"Data! data! data!" he cried impatiently. "I can't make bricks without clay."
—Sherlock Holmes[1]

1. INTRODUCTION

An investigator usually wants to generalize about a class of individuals. This class is called the *population*. For example, in forecasting the results of a presidential election in the U.S., one relevant population consists of all eligible voters. Studying the whole population is usually impractical. Only part of it can be examined, and this part is called the *sample*. Investigators will make generalizations from the part to the whole. In more technical language, they make *inferences* from the sample to the population.[2]

Usually, there are some numerical facts about the population which the investigators want to know. Such numerical facts are called *parameters*. In forecasting a presidential election in the U.S., two relevant parameters are

- the average age of all eligible voters,
- the percentage of all eligible voters who are currently registered to vote.

Ordinarily, parameters like these cannot be determined exactly, but can only be estimated from a sample. Then a major issue is accuracy. How close are the estimates going to be?

Parameters are estimated by *statistics*, or numbers which can be computed from a sample. For instance, with a sample of 10,000 Americans, an investigator could calculate the following two statistics:

- the average age of the eligible voters in the sample,
- the percentage of the eligible voters in the sample who are currently registered to vote.

Statistics are what investigators know; parameters are what they want to know.

Estimating parameters from the sample is justified when the sample represents the population. This is impossible to check just by looking at the sample. The reason: to see whether the sample is like the population in the ways that matter, investigators would have to know the facts about the population that they are trying to estimate—a vicious circle. Instead, one has to look at how the sample was chosen. Some methods tend to do badly. Others are likely to give representative samples.

The two main lessons of this chapter:

- the method of choosing the sample matters a lot;
- the best methods involve the planned introduction of chance.

Similar issues come up when assigning subjects to treatment or control in experiments: see part I.

2. THE *LITERARY DIGEST* POLL

In 1936, Franklin Delano Roosevelt was completing his first term of office as president of the U.S. It was an election year, and the Republican candidate was Governor Alfred Landon of Kansas. The country was struggling to recover from the Great Depression. There were still nine million unemployed: real income had dropped by one-third in the period 1929–1933 and was just beginning to turn upward. But Landon was campaigning on a program of economy in government, and Roosevelt was defensive about his deficit financing.[3]

Landon. The spenders must go.

Roosevelt. We had to balance the budget of the American people before we could balance the budget of the national government. That makes common sense, doesn't it?

The Nazis were rearming Germany, and the Civil War in Spain was moving to its hopeless climax. These issues dominated the headlines in the *New York Times*, but were ignored by both candidates.

Landon. We must mind our own business.

Most observers thought Roosevelt would be an easy winner. Not so the *Literary Digest* magazine, which predicted an overwhelming victory for Landon, with Roosevelt getting only 43% of the popular vote. This prediction was based on the largest number of people ever replying to a poll—about 2.4 million individuals. It was backed by the enormous prestige of the *Digest*, which had called the winner in every presidential election since 1916. However, Roosevelt won the 1936 election by a landslide—62% to 38%. (The *Digest* went bankrupt soon after.)

The magnitude of the *Digest*'s error is staggering. It is the largest ever made by a major poll. Where did it come from? The number of replies was more than big enough. In fact, George Gallup was just setting up his survey organization.[4] Using his own methods, he drew a sample of 3,000 people and predicted what the *Digest* predictions were going to be—well in advance of their publication—with an error of only one percentage point. Using another sample of about 50,000 people, he correctly forecast the Roosevelt victory, although his prediction of Roosevelt's share of the vote was off by quite a bit. Gallup forecast 56% for Roosevelt; the actual percentage was 62%, so the error was 62% − 56% = 6 percentage points. (Survey organizations use "percentage points" as the units for the difference between actual and predicted percents.) The results are summarized in table 1.

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

Note: Percentages are of the major-party vote. In the election, about 2% of the ballots went to minor-party candidates.
Source: George Gallup, *The Sophisticated Poll-Watcher's Guide* (1972).

To find out where the *Digest* went wrong, you have to ask how they picked their sample. A sampling procedure should be fair, selecting people for inclusion in the sample in an impartial way, so as to get a representative cross section of the public. A systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample is called *selection bias*. The *Digest*'s procedure was to mail questionnaires to 10 million people. The names and addresses of these 10 million people came from sources like telephone books and club membership lists. That tended to screen out the poor, who were unlikely to belong to clubs or have telephones. (At the time, for example, only one household in four had a telephone.) So there was a very strong bias against the poor in the *Digest*'s sampling procedure. Prior to 1936, this bias may not have affected the predictions very much, because rich and poor voted along similar lines. But in 1936, the political split followed economic lines more closely. The poor voted overwhelmingly for Roosevelt, the rich were for Landon. One reason for the magnitude of the *Digest*'s error was selection bias.

> When a selection procedure is biased, taking a large sample does not help. This just repeats the basic mistake on a larger scale.

The *Digest* did very badly at the first step in sampling. But there is also a second step. After deciding which people ought to be in the sample, a survey organization still has to get their opinions. This is harder than it looks. If a large number of those selected for the sample do not in fact respond to the questionnaire or the interview, *non-response bias* is likely.

The non-respondents differ from the respondents in one obvious way: they did not respond. Experience shows they tend to differ in other important ways as well.[5] For example, the *Digest* made a special survey in 1936, with questionnaires mailed to every third registered voter in Chicago. About 20% responded, and of those who responded over half favored Landon. But in the election Chicago went for Roosevelt, by a two-to-one margin.

> Non-respondents can be very different from respondents. When there is a high non-response rate, look out for non-response bias.

In the main *Digest* poll, only 2.4 million people bothered to reply, out of the 10 million who got the questionnaire. These 2.4 million respondents do not even represent the 10 million people who were polled, let alone the population of all voters. The *Digest* poll was spoiled both by selection bias and non-response bias.[6]

Special surveys have been carried out to measure the difference between respondents and non-respondents. It turns out that lower-income and upper-income people tend not to respond to questionnaires, so the middle class is over-represented among respondents. For these reasons, modern survey organizations prefer to use personal interviews rather than mailed questionnaires. A typical response rate for personal interviews is 65%, compared to 25% for mailed questionnaires.[7] However, the problem of non-response bias still remains, even with personal interviews. Those who are not at home when the interviewer calls may be quite different from those who are at home, with respect to working hours, family ties, social background, and therefore with respect to attitudes. Good survey organizations keep this problem in mind, and have ingenious methods for dealing with it (section 6).

> Some samples are really bad. To find out whether a sample is any good, ask how it was chosen. Was there selection bias? non-response bias? You may not be able to answer these questions just by looking at the data.

In the 1936 election, how did Gallup predict the *Digest* predictions? He just chose 3,000 people at random from the same lists the *Digest* was going to use, and mailed them all a postcard asking how they planned to vote. He knew that a random sample was likely to be quite representative, as will be explained in the next two chapters.

3. THE YEAR THE POLLS ELECTED DEWEY

Thomas Dewey rose to fame as a crusading D.A. in New York City, and went on to capture the governor's mansion in Albany. In 1948 he was the Republican candidate for president, challenging the incumbent Harry Truman. Truman began political life as a protégé of Boss Pendergast in Kansas City. After being elected to the Senate, Truman became FDR's vice president, succeeding to the presidency when Roosevelt died. Truman was one of the most effective presidents of the 20th century, as well as one of the most colorful. He kept a sign on his desk, "The buck stops here." Another of his favorite aphorisms became part of America's political vocabulary: "If you can't stand the heat, stay out of the kitchen." But Truman was the underdog in 1948, for it was a troubled time. World War II had barely ended, and the uneasy half-peace of the Cold War had just begun. There was disquiet at home, and complicated involvement abroad.

Three major polls covered the election campaign: Crossley, for the Hearst newspapers; Gallup, syndicated in about 100 independent newspapers across the country; and Roper, for *Fortune* magazine. By fall, all three had declared Dewey the winner, with a lead of around 5 percentage points. Gallup's prediction was based on 50,000 interviews; and Roper's on 15,000. As the *Scranton Tribune* put it,

DEWEY AS GOOD AS ELECTED,
STATISTICS CONVINCE ROPER

The statistics didn't convince the American public. On Election Day, Truman scored an upset victory with just under 50% of the popular vote. Dewey got just over 45% (table 2).

Table 2. The election of 1948.

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

Source: F. Mosteller and others, *The Pre-Election Polls of 1948* (New York: Social Science Research Council, 1949).

To find out what went wrong for the polls, it is necessary to find out how they chose their samples.[8] The method they all used is called *quota sampling*. With this procedure, each interviewer was assigned a fixed quota of subjects to interview. The numbers falling into certain categories (like residence, sex, age, race, and economic status) were also fixed. In other respects, the interviewers were free to select anybody they liked. For instance, a Gallup Poll interviewer in St. Louis was required to interview 13 subjects, of whom:[9]

- exactly 6 were to live in the suburbs, and 7 in the central city,
- exactly 7 were to be men, and 6 women.

Of the 7 men (and there were similar quotas for the women):

- exactly 3 were to be under forty years old, and 4 over forty,
- exactly 1 was to be black, and 6 white.

The monthly rentals to be paid by the 6 white men were specified also:

- 1 was to pay $44.01 or more;
- 3 were to pay $18.01 to $44.00;
- 2 were to pay $18.00 or less.

Remember, these are 1948 prices!

From a common-sense point of view, quota sampling looks good. It seems to guarantee that the sample will be like the voting population with respect to all the important characteristics that affect voting behavior. (Distributions of residence, sex, age, race, and rent can be estimated quite closely from Census data.) But the 1948 experience shows this procedure worked very badly. We are now going to see why.

The survey organizations want a sample which faithfully represents the nation's political opinions. However, no quotas can be set on Republican or Democratic votes. The distribution of political opinion is precisely what the survey organizations do not know and are trying to find out. The quotas for the other variables are an indirect effort to make the sample reflect the nation's politics. Fortunately or unfortunately, there are many factors which influence voting behavior besides the ones the survey organizations control for. There are rich white men in the suburbs who vote Democratic, and poor black women in the central cities who vote Republican. As a result, survey organizations may hand-pick a sample which is a perfect cross section of the nation on all the demographic variables, but find the sample voting one way while the nation goes the other. This possibility must have seemed quite theoretical—before 1948.

The next argument against quota sampling is the most important. It involves a crucial feature of the method, which is easy to miss the first time through. Within the assigned quotas, the interviewers are free to choose anybody they like. That leaves a lot of room for human choice. And human choice is always subject to bias. In 1948, the interviewers chose too many Republicans. On the whole, Republicans are wealthier and better educated than Democrats. They are more likely to own telephones, have permanent addresses, and live on nicer blocks. Within each demographic group, Republicans are marginally easier to interview. If you were an interviewer, you would probably end up with too many Republicans.

The interviewers chose too many Republicans in every presidential election from 1936 through 1948, as shown by the Gallup Poll results in table 3. Prior to 1948, the Democratic lead was so great that it swamped the Republican bias in the polls. The Democratic lead was much slimmer in 1948, and the Republican bias in quota sampling had real impact.

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).

> In quota sampling, the sample is hand-picked to resemble the population with respect to some key characteristics. The method seems reasonable, but does not work very well. The reason is unintentional bias on the part of the interviewers.

The quotas in quota sampling are sensible enough, although they do not guarantee success—far from it. But the method of filling the quotas, free choice by the interviewers, is disastrous.[10] The alternative is to use objective and impartial chance mechanisms to select the sample. That will be the topic of the next section.

4. USING CHANCE IN SURVEY WORK

Even in 1948, a few survey organizations used *probability methods* to draw their samples. Now, many organizations do. What is a probability method for drawing a sample? To get started, imagine carrying out a survey of 100 voters in a small town with a population of 1,000 eligible voters. Then, it is feasible to list all the eligible voters, write the name of each one on a ticket, put all 1,000 tickets in a box, and draw 100 tickets at random. Since there is no point interviewing the same person twice, the draws are made without replacement. In other words, the box is shaken to mix up the tickets. One is drawn out at random and set aside. That leaves 999 in the box. The box is shaken again, a second ticket is drawn out and set aside. The process is repeated until 100 tickets have been drawn. The people whose tickets have been drawn form the sample.

This process is called *simple random sampling*: tickets have simply been drawn at random without replacement. At each draw, every ticket in the box has an equal chance to be chosen. The interviewers have no discretion at all in whom they interview, and the procedure is impartial—everybody has the same chance to get into the sample. Consequently, the law of averages guarantees that the percentage of Democrats in the sample is likely to be close to the percentage in the population.

Simple random sampling means drawing at random without replacement.

What happens in a more realistic setting, when the Gallup Poll tries to predict a presidential election? A natural idea is to take a nationwide simple random sample of a few thousand eligible voters. However, this isn't as easy to do as it sounds. Drawing names at random, in the statistical sense, is hard work. It is not at all the same as choosing people haphazardly.

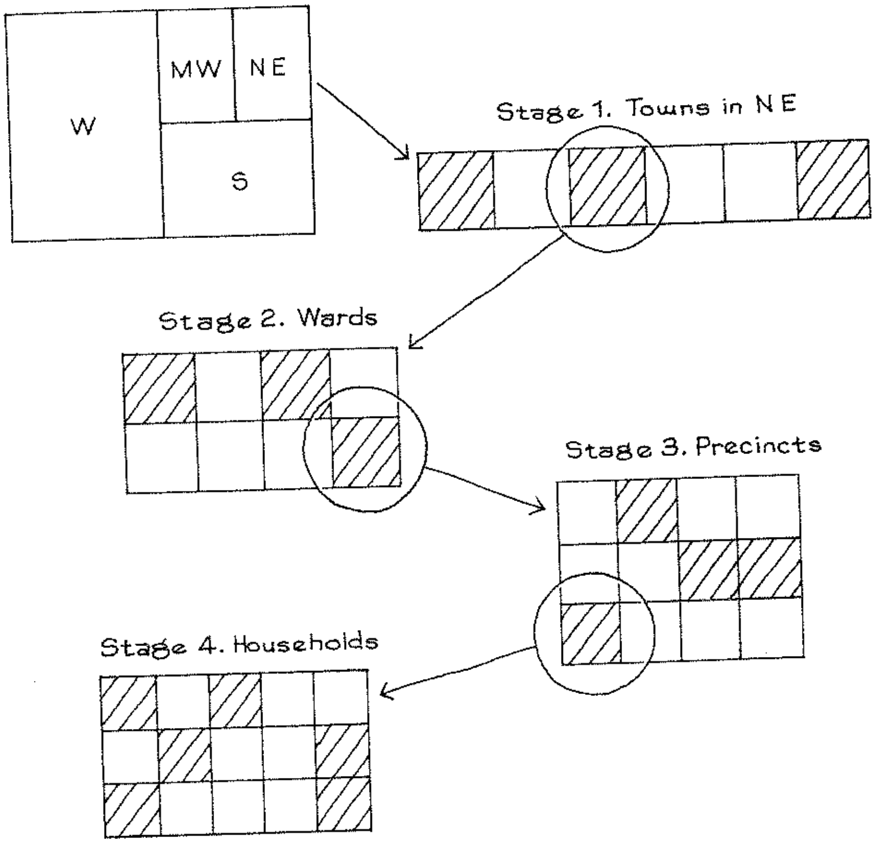
To begin drawing eligible voters at random, you would need a list of all of them—well over 200 million names. There is no such list.[11] Even if there were, drawing a few thousand names at random from 200 million is a job in itself. (Remember, on each draw every name in the box has to have an equal chance of being selected.) And even if you could draw a simple random sample, the people would be scattered all over the map. It would be prohibitively expensive to send interviewers around to find them all.

It just is not practical to take a simple random sample. Consequently, most survey organizations use a probability method called *multistage cluster sampling*. The name is complicated, and so are the details. But the idea is straightforward. It will be described in the context of the Gallup pre-election surveys during the period from 1952 through 1984; these surveys were all done using just about the same procedure. The Gallup Poll makes a separate study in each of the four geographical regions of the United States—Northeast, South, Midwest, and West (figure 1). Within each region, they group together all the population centers of similar sizes. One such grouping might be all towns in the Northeast with a population between 50 and 250 thousand. Then, a random sample of these towns is selected. Interviewers are stationed in the selected towns, and no interviews are conducted in the other towns of that group. Other groupings are handled the same way. This completes the first stage of sampling.[12]

For election purposes, each town is divided up into *wards*, and the wards are subdivided into *precincts*. At the second stage of sampling, some wards are selected—at random—from each town chosen in the stage before. At the third stage, some precincts are drawn at random from each of the previously selected wards. At the fourth stage, households are drawn at random from each selected precinct.[13] Finally, some members of the selected households are interviewed. Even here, no discretion is allowed. For instance, Gallup Poll interviewers are instructed to "speak to the youngest man 18 or older at home, or if no man is at home, the oldest woman 18 or older."[14]

This design offers many of the advantages of quota sampling. For instance, it is set up so the distribution of the sample by residence is the same as the distribution for the nation. But each stage in the selection procedure uses an objective and impartial chance mechanism to select the sample units. This completely eliminates the worst feature of quota sampling: selection bias on the part of the interviewer.

Figure 1. Multistage cluster sampling.



Simple random sampling is the basic probability method. Other methods can be quite complicated. But all probability methods for sampling have two important features:

- the interviewers have no discretion at all as to whom they interview;
- there is a definite procedure for selecting the sample, and it involves the planned use of chance.

As a result, with a probability method it is possible to compute the chance that any particular individuals in the population will get into the sample.[15]

Quota sampling is not a probability method. It fails both tests. The interviewers have a lot of discretion in choosing subjects. And chance only enters in the most unplanned and haphazard way. What kinds of people does the interviewer like to approach? Who is going to be walking down a particular street at a particular time of day? No survey organization can put numbers on these kinds of chances.

5. HOW WELL DO PROBABILITY METHODS WORK?

Since 1948, the Gallup Poll and many other major polls have used probability methods to choose their samples. The Gallup Poll record in post-1948 presidential elections is shown in table 4. There are three points to notice. (i) The sample size has gone down sharply. The Gallup Poll used a sample of size about 50,000 in 1948; they now use samples less than a tenth of that size. (ii) There is no longer any consistent trend favoring either Republicans or Democrats. (iii) The accuracy has gone up appreciably.

From 1936 to 1948, the errors were around 5%. Since then, they are quite a bit smaller. (In 1992, the error went back up to 6%; the reason will be discussed on p. 346.) Using probability methods to select the sample, the Gallup Poll has been able to predict the elections with startling accuracy, sampling less than 5 persons in 100,000—which proves the value of probability methods in sampling.

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.

Why do probability methods work so well? At first, it may seem that judgment is needed to choose the sample. For instance, quota sampling guarantees that the percentage of men in the sample will be equal to the percentage of men in the population. With probability sampling, we can only say that the percentage of men in the sample is likely to be close to the percentage in the population: certainty is reduced to likelihood. But judgment and choice usually show bias, while chance is impartial. That is why probability methods work better than judgment.

> To minimize bias, an impartial and objective probability method should be used to choose the sample.