# Survey Sampling: Introduction

Surveys are probably the most prominent aspect of Statistics that the public sees. They are perhaps universally criticized and yet they continue to function. Why? Because, despite the criticism, if properly implemented and carefully interpreted, they are an enormously valuable tool.

## Ingredients

**Population**: A large group of people or objects that you wish to study. For example, all U.S. voters, all U.S. "probable" voters, all children in U.S. under 12, all 25-35 year olds, all UCLA students, etc.

**Parameter**: a characteristic of that population that you wish to know. Most often this is a proportion, e.g. the proportion of people who voted for Bush. The proportion of people who smoke more than 1 pack a day. The proportion of people who might buy your product. But it can be any single numerical quantity. For example, the mean number of children in a family. The mean income.

*Note*: When referring to a population, we say "**mean**" rather than "s**ample mean**".

**Sample**: a small group of subjects selected from the population.

**Statisti**c: a measurement made on the sample that is used to estimate the population parameter. For example, the average of a sample is often used to predict the mean of a population. And the proportion of a sample is used to predict the proportion of the population.

**Sampling Frame**: A list of all members of the population. This list might be conceptual, or might actually be many different lists.

Basic idea: use a sample to represent the population. Infer what we learn about the sample to the rest of the population.

Problem: the sample will never be *exactly* the same as the population. In particular, the sample statistic will never (will, it is extremely rare) exactly equal the population parameter.

Great 20th Century idea: if the sample is drawn in a certain way, we can get a very good understanding in the error -- the difference between our sample statistic and the population parameter.

## Presidential Elections

Recently, pools used to predict presidential elections have come under, shall we say, increased scrutiny. This is frustrating in a way because in some sense voting polls are the easiest things to predict. Imagine the difficulties in asking questions like:

1) Are you an alcoholic?

2) Would you go to Disneyland if the cost were $65?

3) Is anyone in your family currently in jail?

4) Do you go to church regularly?

All of these questions have implicit "correct" answers in most cultures, or, in the case of the Disney/marketing question, try to predict fairly complex action. Voting, on the other hand, seems relatively straightforward. We're not trying to read into anyone's soul. We just want to know will you or will you not vote for x. And then, later, we can compare our predictions with the outcome.

The problem, and this is common in life, is that the theory works so much better than the application. So lets look at the theory, and then see how things might go astray.

Selection Bias

The most important consideration, probably, is that your sample be representative of the population. You don't stand on the corner of Sunset and Hilgard, for example, to evaluate the cost of the typical car in the U.S., or even in Los Angeles. And you don't poll college students to try to predict the outcome of an election. These are examples of selection bias, because your sample is biased away from the population.

Of course, you might get lucky and use a bad technique and end up with a good sample, but in general how can we be sure we get a representative sample?

The answer is to use random sampling. The simplest version (well, the second simplest version) is called Simple Random Sampling (SRS). Here's the approach:

select subjects at random withOUT replacement. This means that , once selected, a subject cannot be selected again.

It's analogous to selecting tickets from a box. You shake the box, pull out a ticket, and set it aside.

An even simpler approach is to use sampling WITH replacement, which is analogous to putting the ticket back in the box before making the next selection. This is not often utilized because it can be expensive in some situations to sample the same person twice.

In practice, if the sample size n is much smaller than the population size N, (roughly n < (1/10) N) the two approaches are very close to the same. (Roughly speaking, the probability that a person will be sampled twice becomes negligible when the population is much bigger than the sample.

The Literary Digest Poll

In 1936, FDR, Democrat, running for re-election against Kansas governor Alfred Landon, Republican. This was during the midst of the depression, and Landon ran on a platform of cutting government spending.

The Literary Digest was America's most popular magazine. They predicted Landon would win, based on the largest number of people to ever reply to a poll: 2.4 million. The Digest had correctly called the winner in every presidential electtion since 1916.

As you might have guessed, Landon did not win. The Digest's prestige suffered, and maybe partially because of this they went bankrupt a few years later. Here's what they said:

| Prediction | Result |  |
|---|---|---|
| Roosevelt | 44% | 62% |
| Landon | 56% | 38% |

How did they pick their sample of 2.4 million?

They mailed 10 million questionnaires using addresses from the phone book and club membership lists. By doing so, they tended to miss the poor, who did not have phones and did not join clubs. (Only one household in four had a phone.)

In the past, this was not a problem because rich and poor tended to vote alike. But the depression had the effect of splitting voting along economic lines, to some extent. Hence, *Selection Bias* created a sample that was non-representative.

A new polling organization run by George Gallup used random methods and predicted that Roosevelt would win with 56%. Even better, before the Literary poll was published, he predicted their prediction: he said that the Digest would give Roosevelt 44%.

## Response Bias

Another source of error in the Digest poll was response bias. 10 million people received surveys, but "only" 2.4 million returned them. Were those who returned surveys different from those who did not? If not, then there might be a bias. In fact, follow-up work suggested that those who favored Landon were more likely to return surveys. One possible explanation is that Landon supporters might have felt that they were in a minority (since roosevelt was the incumbant) and might have felt it more important that they respond. A common cause of response bias is that people who feel very strongly on an issue are more likely to respond.

TV Surveys: Dial 1-800-xxx to vote for your favorite song.

How did Gallup predict the Digest's results? He took a simple random sample of 3000 people from the same lists the Digest used.

## President Dewey

All was not wine and roses for Gallup. In 1948 Thomas Dewey, Governor of New York, challenged Truman, the incumbant. Three major polls (including Gallup) predicted Dewey would win. In fact, a famous picture shows Truman holding a front page of a newspaper that declares Dewey Wins!

| Candidates | Crossley | Gallup | Roper | Result |
|---|---|---|---|---|
| Truman | 45 | 44 | 38 | 50 |

| | | | | |
|---|---|---|---|---|
| Dewey | 50 | 50 | 53 | 45 |
| Thurmond | 2 | 2 | 5 | 3 |
| Wallace | 3 | 4 | 4 | 2 |

The failure this time was the improper use of something called Quota Sampling. In quota sampling, interviewers are sent into an area and told to interview a certain number of people who meet characteristics. For example: choose 7 people, 4 are men, 3 are women, 2 are african american, 5 are white, etc. The idea is that if the quotas accurately reflect the population, the sample will be representative.

In fact there are two flaws. Just because your sample reflects the population demographically doesn't mean it reflects it on the issue you care about (e.g. voting behavior). Second, it leaves selection of the people to the interviewer, who might unwittingly select some people out of the sample who are hard to approach.

Since that time, the polls have correctly predicted the President. (This recent election may or may not fall into this category. We'll discuss this in a moment.) Sample sizes are now much less, typically around 3000 people. The actual errors have been less than 4% (sometimes as small as .1%). A recent exception was Clinton v. Bush where the actual error was 5.8%. The large error (they still called the correct winner) was explained by an unusually large number of "undecided" voters. Gallup incorrectly classified these as Clinton supporters, when the truth was more complicated. This meant that the actual outcome was closer than they had predicted.

Read the copies from the LA Times. They all express great frustration for Presidential polls, but for a variety of reasons. Some are sound, some are not. The first is from Arianna Huffington, and she states that the margin of error is "alchemy". But her reason appears to be simply that she doesn't understand it. The second is also by Huffington, and comes after the election. She correctly criticizes "meaningless" polls, and her example of a meaningless poll is a good one. (The poll tries to predict behavior 4 years in the future.) But she uses a strange way of measuring the accuracy of polls (she looks at how many polls were correct, and concludes and accuracy rate of only 20%. But its not fair to count all polls equal.) She also calls the margin of error a "deus ex machina" that is used to explain away any discrepancy. But in fact, the margin of error was correct: the election results did fall within the marign of error. The problem was that the margin of error was too big to choose between the candidates. (Hence the phrase "statistical dead heat.") The third is from Times columnist James P. Pinkerton. He is factually wrong on his historical details (polls have correctly called every election since Dewey; but reporters have had a poor track record of interpreting polls.) He has an interesting point about the reliance on phones in an age in which the way people use their phones is rapidly changing.

Reality vs. Theory

In practice it is nearly impossible to do a SRS on any scale. There are modifications of this technique, though, that can make it easier. One is the stratified random sample. In this technique, sub-groups of the population known to be similar are grouped together and then sampling is a two stage affair: first take a random sample of sub-groups, then take random samples from within the chosen groups.

There are various other methods, too, that employ elements of randomness. But there are some that do not:

- The convenience sample -- you see these on the news as the "man in the street" interview. For entertainment only.

- The arbitrary sample -- suppose I want to take a random sample of gloves for inspection from a bag. If I just reach in and arbitrarily choose some, my sample might be good. IF the gloves were well shuffled.

- The focus group -- marketers choose people, often friends or aquaintances, who reflect the demographics of their particular market. Often they are asked their opinions as a group, and so their opinions might differ in their "natural habitats".

- The census - - a census is intended to be a complete renumeration of the entire population and therefore not employ any sampling. In practice this is impossible, and hence the current controversy over the application of sampling techniques to adjust for the "undercount". The undercount is a widely acknowledged phenomenon (by both political parties) in which poor people tend to not be counted. Their is general agreement on the number of such people, but the disagreement is over in which states and cities they are located. This affects federal funding. For this reason, politics has become involved, and whether or not to sample has been portrayed as a moral issue. But there is a technical issue involved because in fact the statistical techniques required are quite complex and it is not at all clear that there are any technqiues capable of solving this problem.

The Florida Debacle

The smoke has not completely cleared from this debacle, and so it is still hard to say exactly what happened, much less explain what went wrong. But let me offer a possible explanation.

There are many other ways a poll can fail:

- people lie

- random mechanisms might not be random. Famous problem with the "lottery" selection method for the draft in Vietnam War that "favored" people born in certain months.

- people change. Pools done too early might be accurate predictions of what would happen if the election were that day, but in the meantime, things might happen to change opinion.

But a more subtle problem is that all polls have some sort of error. The benefit of random sampling is that this error can be measured ahead of time. Given that the sampling is indeed random, then the main component affecting error is the sample size. This error is called the sampling error. This is why you see pools that predict, say, that George Bush will get 53% plus or minus 3%.

- the larger the sample size, n, the smaller the sampling error.

- as long as the population is much bigger than the sample, the size of the population has no effect on the sampling error. Thus, a sample of 3000 people is just as valid in Rhode Island as it is in California as it is in the United States.

The formula for sampling error (assuming SRS) is sqrt( (N-n)/N-1) * sqrt(p * (1-p)/n)

where p is the proportion in the population, n is sample size, N is population size.

A problem here is that you don't know p, else you wouldn't be doing the survey. But it turns out that sqrt(p * (1-p)) <= 1/2. So we know that the sampling error is always smaller than sqrt(N-n)/N-1) (1/2*sqrt(n)).

Suppose we have a sample size of n = 1500 (a common sample size). Here's the biggest sampling error for a variety of different population sizes:

| N | error |
|---|---|
| 2000 | .006 |
| 20,000 | .012 |
| 200,000 | .013 |
| 2,000,000 | .013 |
| 20,000,000 | .013 |

As you can see, once the population size is bigger than 10*n = 15,000 the sampling error is pretty steady. (even out to five decimal places, actually.)

If SRS is not used, this is still a pretty good approximation.

Typically, pollsters use something called the "margin of error" which is apprxoimately 2 * sampling error. You then see:

estimate +- margin of error.

Usually, the sample size is chosen in advance so that the margin of error is about 3%. Then you see, for example,

53% +- 3 %.

What does this mean? This is an example of a confidence interval, which we'll discuss in depth later. For now, the interpration is as follows:

Imagine that a large number of pollsters did their own polls, each using the same sample size and the same sampling technique. Then 95% of them will get an interval that contains the true population proportion.

Now, we don't know for sure whether our interval of 50% to 56% contains the true population proportion. But we know that there is a good chance that it does. So we can be reliably certain that the truth is somewhere between there.

Part of the confusion in Florida came from this problem. Suppose you do a poll that gives Bush 51% +- 3%. Will he win? In a situation like this, you can't tell because all we know is that the truth is likely to be somewhere in this interval. And this interval includes an outcome in which Bush loses.

We can make the interval smaller by taking a larger sample. But if Bush and Gore are really and truly very close together, say by 0.003 (which is about what appears to be the case), then it may be impossible to get a sample size large enough to determine how close the population is.