Statistical Sampling.[*]

Greg Kochanski
http://kochanski.org/gpk

2006/02/06 19:05:35 UTC

1 Introduction

Why is statistical sampling hard? After all, if you want to measure 100 things out of a population of a million, all you need to do is this:

1. Number all the things from 1 to a million.
2. Pick a uniformly distributed random number between one and a million,
3. Measure the thing whose number you chose, and
4. Repeat 100 times.

The above recipe is ideal when it works, but it assumes that everything is laid out so you can number it, it assumes you have the ability to measure anything you wish, and it assumes you want to sample uniformly. None of these assumptions are necessarily true. The first thing one should read is *How to Lie with Statistics* [Huff, 1954]. This is a thin and entertaining little book that gives dozens of examples of how you can mislead someone with bad statistics, and many of the examples are related to bad sampling.

While statistical sampling lends itself to abuses, it is also quite easy to make honest mistakes that can have important consequences. One of the most famous mistakes is the mis-calling of the US presidential election in 1948 [Eag, 2003] (see Figure 1). Opinion polls a few weeks before the election predicted a strong win for the Republican candidate, Thomas E. Dewey, by a margin of 5-15% percent of the vote. Dewey actually lost by 4.4%.

The detailed reasons for the error don't seem entirely clear, but it has been plausibly attributed to one or both of the following:

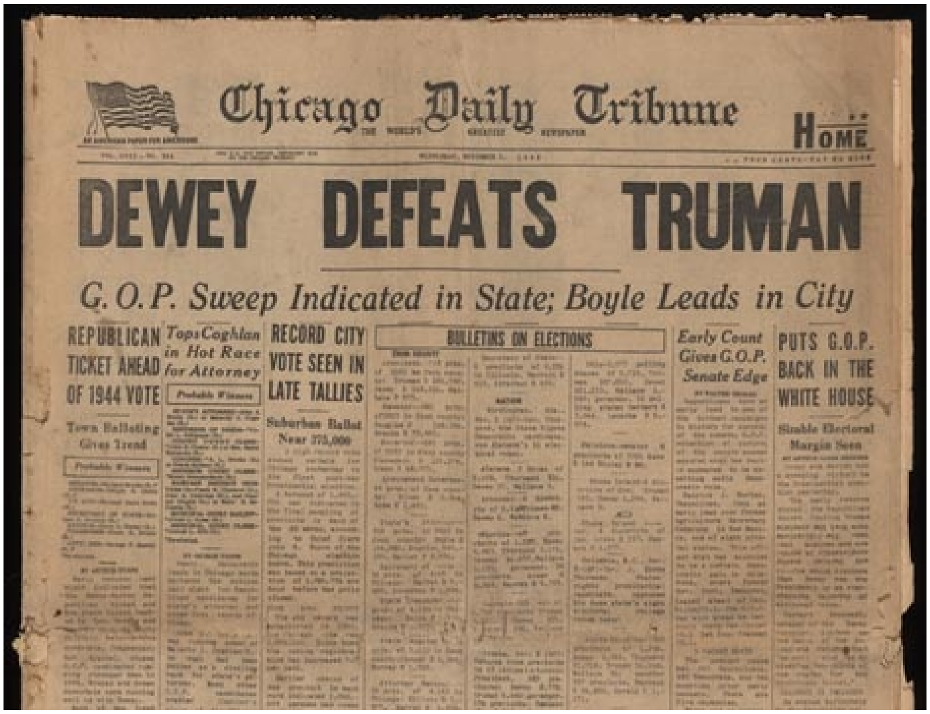
- The then-new technique of telephone polling [Huff, 1954], possibly combined with quota-sampling techniques[1]. Either way, people with telephones were over-sampled, either on purpose or because calling someone on the phone was a convenient way for a poll-taker to make his quota.

  Unfortunately, in 1948, phones were fairly expensive, and not everyone had a phone. People with phones tended to be well off, tended to be Republican, and tended to say that they would vote for Dewey. The actual vote didn't have this telephone bias, and so went to Truman.

  This account pins the blame on a poorly designed poll (telephone sampling), or on allowing the poll-takers to use opportunity sampling within their quotas.

[*] This work is available under http://kochanski.org/gpk/teaching/0401Oxford.

[1] Quota sampling is a technique where a poll-taker is told to get the opinion of a certain number of people in each of several categories, but is free to choose the actual people to make his quota. Quota sampling is a form of stratified sampling (§5.1) where the fine level is opportunity sampling (§5.5).



Chicago Daily Tribune

THE WORLD'S GREATEST NEWSPAPER

HOME

DEWEY DEFEATS TRUMAN

G.O.P. Sweep Indicated in State; Boyle Leads in City

REPUBLICAN TICKET AHEAD OF 1944 VOTE

Town Balloting Gives Trend

Probable Winners

Tops Coghlan in Hot Race for Attorney

Probable Winner

RECORD CITY VOTE SEEN IN LATE TALLIES

Suburban Ballot Near 375,000

BULLETINS ON ELECTIONS

Early Count Gives G.O.P. Senate Edge

PUTS G.O.P. BACK IN THE WHITE HOUSE

Sizable Electoral Margin Seen

Figure 1: *Despite newspaper reports to the contrary [Ch. Trib.] which were based on opinion polls, Harry S. Truman won the election for US President.*

- The fact that people had several weeks between the last polls and the election to change their minds. Some accounts pin the blame on the poll's inability to predict who would vote and who wouldn't [Eag, 2003]. The story is then that potential voters for Dewey were much less likely to vote, and/or became complacent and made a last-minute decision not to vote. The reverse was true for potential Truman-voters.

  Under this account, the bad result is, again, a sampling problem: the pollsters didn't sample people proportionally to the probability that they would vote. They over-sampled non-voters with opinions, and under-sampled voters.

The 1948 US election connects to almost all corners of statistical sampling. We will discuss a few of the major issues and then various sampling techniques.

2 What if you cannot make some measurements?

This problem always occurs when you deal with people. One cannot compel people to answer questions, declare whether an utterance is grammatical, or do much of anything. Consequently, the simple strategy of numbering every inhabitant of Britain, picking a random number and asking that person to declare whether "Colourless green ideas sleep furiously" is grammatical or not is quite unworkable. The odds are good that they'll simply shut the door on your face.

This is a serious problem because you have no way of knowing if the people who shut the door are different from the people who answer. They will refuse for reasons that you do not (and cannot) know, and those reasons may be correlated with the answer that they would give. Real random sampling of a human population is essentially impossible.

What, then, do you do?

Suppose for a moment that when you ring a pre-selected doorbell and say "I'm a linguist, and I'd like to ask you a few questions." Suppose everyone who has read and appreciated Noam Chomsky says "Of course!", and everyone who hasn't says "Go away!". Then, your survey might well report that nearly everyone who answers replied that the sentence was syntactically correct, but semantically meaningless. That would be true, but it would be a horrible mistake to assume that the people who didn't answer had the same opinion. Many of them would have expressed the opinion "That's bloomin' nonsense, mate", and many of the rest would believe that you were in need of seeing a psychiatrist[2].

This is an example of a *selection effect*. Selection effects happen when the people you can sample are systematically different from the ones that you cannot sample.

2.1 Solutions

It can be quite difficult to prove you don't have a selection effect, because that would require learning about the people whom you cannot sample. To do so, you need some indirect source of information on the people who do not respond, along with a theory that connects the indirect information you can get to the information you were trying to measure.

One approach is to try to estimate the selection effect by picking a small extra sample and sparing no effort to attempt to make the people in it respond. That can mean rewards, money, multiple visits and personal attention. To the extent that you can make a larger-than-normal fraction of the extra sample respond, you can succeed at estimating how large the selection effect is in the rest of the population. You can then add the estimated selection bias to your survey results and get an improved estimate for the population as a whole.

In the case of predicting election results, this problem is fairly straightforward, because the district-by-district results from actual elections are available. After the election, you can correlate various easily obtainable properties of the district with the difference between pre-election poll results and the actual outcome.

For instance, over several elections, you might find that a particular district votes between 3% and 5% more Tory than their poll results. In the U.S., you might find that in a particular district, the number of people who actually vote might be 30% in a presidential election, but only 24% in other years. Similarly, you might find systematic differences in voting behaviour between rural and urban districts, or between districts with home-owners *vs.* renters. Adding all these corrections (which were determined from past election) into the poll results can give you a much better idea of how each district will actually vote, once you have pre-election opinion polls.

Selection effects in other surveys are often less tractable, because there may be nothing equivalent to the election results that provide a reliable check on your polling. One's only recourse is then to (somehow) predict the selection effect based on other evidence, perhaps by way of correlations or circumstantial evidence from available data. Success is not assured.

3 What if you cannot number the things you wish to sample?

Sometimes, the hardest problem may not be making the desired measurements, but instead may involve finding the things you want to measure. This is often the case when one is studying spoken language: numbering all the words that get spoken in one day in any active language is impossibly difficult. A survey of spoken noun phrases would be yet harder still. Not only would you have to solve the non-trivial problem of monitoring all speech, but you would have to analyse it, utterance by utterance, just to *find* the noun phrases.

This is a job for cluster sampling (§5.2 below): you pick a few small regions and sample intensively inside them. In linguistics, this usually translates into collecting a corpus, which is a small, often localised part of the language (*e.g.* the New York Times). However, true cluster sampling involves a random choice of the cluster location, which (for corpora) would correspond to a random choice of the document source.

[2] And, there may be a few who would say "Grammaticality isn't a true-or-false choice: I think it is neither totally acceptable nor totally ungrammatical", a few who say "Judging a sentence in isolation is unrealistic", and a few who say that "Grammaticality is just a fantasy that has nothing at all to do with the way we actually process speech."

No one *actually* does that (though they perhaps should); the document source is normally chosen through opportunity sampling[3].

4 What if you want a nonuniform sampling?

By nonuniform sampling, I mean that some of the things in your universe are more important than others. For instance, in the electoral example above (§2.1), you would like to pay more attention to the people who are likely to vote. If you had a good estimate of $P(\text{Vote})$, you would want the probability of polling someone to be proportional to their probability of voting. The result of your poll could then be interpreted as the preferences of likely voters, rather than the preferences of the population as a whole.

Nonuniform sampling is straightforward if you can compute a probability for each thing in advance based on some conveniently available feature of each thing. You can use random or stratified sampling, by choosing each thing or stratum with its appropriate probability.

5 Sampling Techniques.

There are three basic techniques of sampling: random sampling (described above, §1), systematic sampling (§5.3), and opportunity sampling (§5.5). In addition, cluster and stratified sampling are two techniques that can be thought of as recursive: they split the sampling problem into a coarse and a fine level, each of which can be sampled with any technique.

Independence of Sampled Data: Normally, samples should be defined before you start looking at the data. This is on purpose, both to prevent the experimenter from fooling him/herself by focussing on an unusual corner of the data that gives a particularly pleasant result, but also to make sure the resulting measurements are all independent of each other. Making the choice of sample independent of the data values makes any statistical analysis much easier.

Remember, data is independent if (and only if) learning about one sample doesn't tell you anything about another sample. (More precisely, *two random variables* are independent if and only if knowing the value of one doesn't help you to predict the value of the other.) So, the colours of hairs on two unrelated people are independent (or nearly so), but the colours of two hairs on the same head are very much not independent. If you see a brown hair, you can predict that another hair is also likely to be nearly the same shade of brown.

Even more precisely, when you talk about independence, you are assumed to know what can be in the "universe". I.e. you know the range of possible hair colours. So, looking at a hair and seeing that it is brown *doesn't* tell you that brown hair is possible because you knew that already.

The above sampling techniques do not depend on the data you are studying; their whole intent is to give you a clean cross-section of data, where all the measurements are independent of each other and representative of the population as a whole[4].

On the other hand, one of the sampling techniques, Adaptive Sampling, intentionally allows the choice of samples to depend on the data you have seen so far. Using it, one pays a price of a more complex experiment and a more complex analysis in exchange for the potential to focus on the most important data.

5.1 Stratified Sampling

Stratified sampling looks at easy-to-obtain features of the things you might want to sample. Stratified sampling splits the sampling problem into a coarse problem (deciding how many points to take from each strata), and a fine problem (collecting samples inside each strata). There are several reasons to do this:

[3] *I.e.* people grab whatever is most convenient.

[4] Mind you, opportunity sampling often doesn't meet these grand intentions.

- Sometimes, opportunity sampling is unavoidable. If so, stratified sampling lets you do half of the job (for instance the coarse level) with a nice clean random sampling procedure. The opportunity sampling which would be used for the fine level then has less opportunity to make mischief because it is constrained to remain within each strata. (Quota sampling (§1), is just this: stratified sampling, with opportunity sampling for the fine level.)
- Stratified sampling is a convenient way to take nonuniform samples. It is easy to control the probability of sampling different items by choosing a probability for each stratum. For instance, if you are studying worms, it's more efficient to sample the top metre of the soil where most worms are, rather than taking a random sample from down near the Earth's core.

A linguistic example of stratified sampling is the following, where we stratify by the length of a word:

1. Count the number of 1-, 2-, 3-, and 4- syllable words in a corpus. The number of words of length k is found to be N_k.
2. For each word, choose how many letters it will have. You might want to choose k letters with a probability that is proportional to N_k.[5]
3. Choose randomly from the list of $k-$syllable words.

Choosing with certain probabilities: If you have an array of items `x = ['one', 'two', ...]` and probabilities `P = [0.2, 0.3, 0.1, ...]` then here's a scrap of Python code that will let you choose item k with probability P_k.

```
import random

def chooseP(x, P):
        ps = 0.0
        r = random.random()
        for (xx, pp) in zip(x, P):
                ps += pp
                if pp > r:
                        return xx
                else:
                        r -= pp
```

NB: A version with error checking (which is important in the real world, where people, even you and I, make mistakes) can be found at `http://sourceforge.net/projects/speechresearch/files` in the gmisclib package, file gpkmisc.py.

5.2 Cluster Sampling

Cluster sampling is like stratified sampling, except the clusters are natural clumps, rather than more-or-less artificial strata.

5.3 Systematic Sampling

Sample every 10^{th} thing, or every N^{th}. More generally, this is sampling that is driven by an algorithmic pattern, rather than a random number generator[6]. Systematic sampling has one significant advantage: it is more uniform than random sampling. Sometimes, this fact can result in lower variances for your measurements.

[5] E.g. the probability of choosing k would be $N_k/(N_1 + N_2 + N_3 + N_4 + \ldots)$.

[6] It is convenient to ignore the fact that most (pseudo-)random number generators are really algorithms.

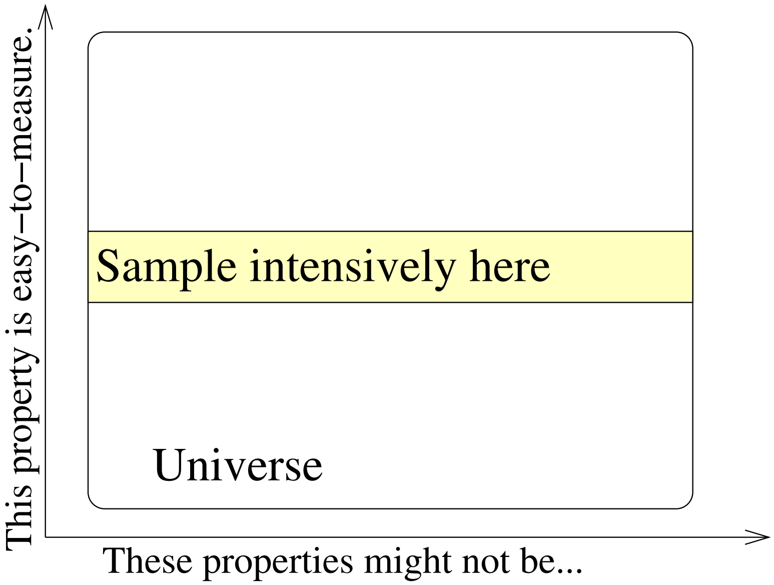

This property is easy−to−measure.
Sample intensively here
Universe
These properties might not be...

Figure 5.1: *Stratified sampling. You sample strata to choose which the horizontal band, then sample inside the coloured one.*

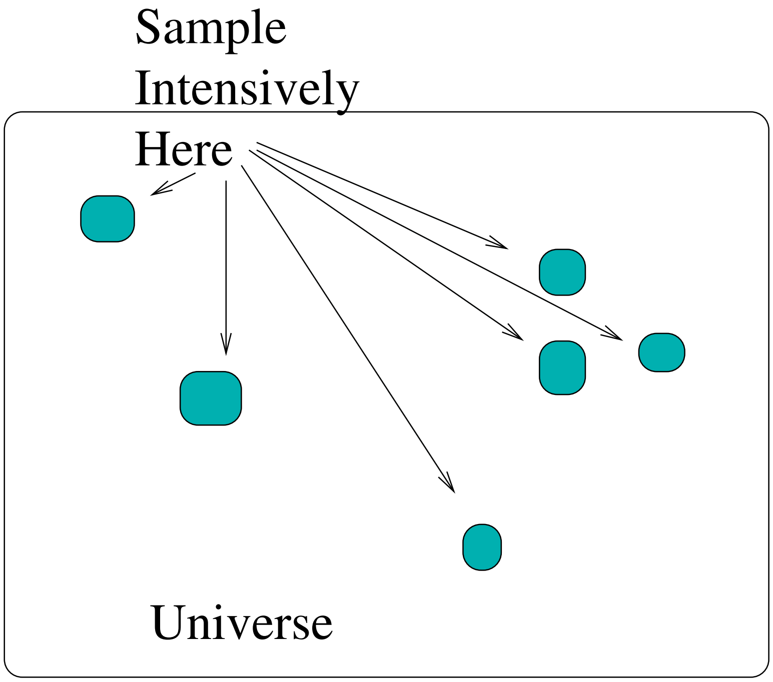


Figure 5.2: *Cluster sampling. The location of a cluster is chosen through some sampling scheme, and then you sample intensively inside the cluster.*

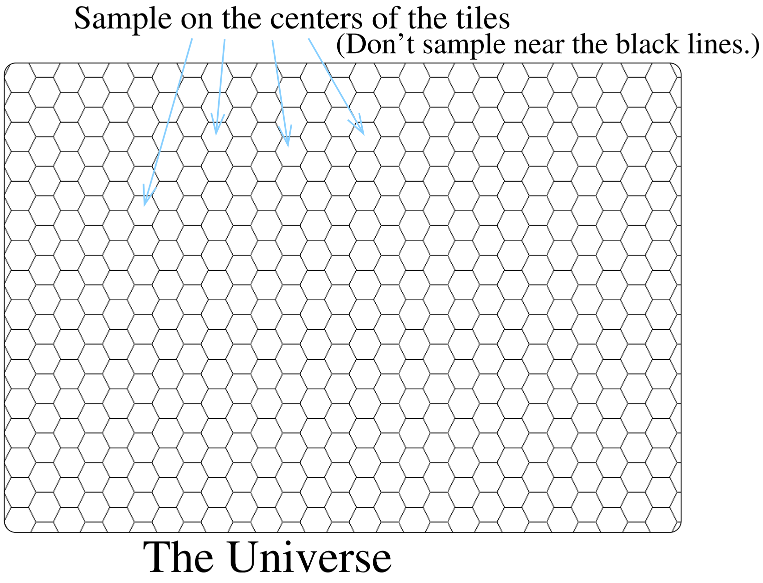

Sample on the centers of the tiles
(Don't sample near the black lines.)
The Universe

Figure 5.3: *Systematic sampling. You sample in a regular pattern, such as every* 10^{th} *person.*

5.4 Adaptive Sampling

Adaptive sampling is a catch-all category for any sampling strategy that depends on the measurement that you are reporting. Adaptive sampling can sometimes be extremely efficient. For instance, it is good for studying dolphins. Dolphins swim in groups, so if you sample a part of the ocean and find one dolphin, it makes a lot of sense to search that area intensively to find the rest of the pod. If you don't find a dolphin initially, you move off to another part of the ocean. This example differs from cluster sampling because cluster sampling will intensively search around every initial sample point, not just the sample points that contain dolphins.

The down-side of adaptive sampling is that the analysis becomes more complex, because you are intentionally introducing dependences between samples. In the dolphin example, you have two distinct sets of measurements: initial points which are a random sample of the ocean, and the remainder of the cluster, which is a random sample of parts of the ocean near a dolphin[7].

To make this into a linguistic example, you translate "ocean" ⇒ "all books", and "dolphin" ⇒ "book that contains a centre-embedded sentence". Under that translation, the dolphin example converts to the following procedure:

1. Look at a few pages of each book in the library.
2. If you see a centre-embedded sentence, borrow the book and transcribe it.

That procedure would give you a corpus of books that are rich in centre-embeddings, which might be a very good thing if you want to study these rare constructions. The downside would be that it might be harder to extrapolate to the language as a whole, because you would have to mathematically correct for the enrichment.

For adaptive cluster sampling, see Thompson [1990].

5.5 Opportunity Sampling.

The worst thing that one can do, from the point of view of statistical sampling, is to sample yourself, your co-workers, or your friends. This is known as *opportunity sampling*, and it is always the easiest thing to

[7] There can be many differences: ocean near dolphins tends to have less ice over it; it probably has more fish; perhaps fewer sharks; and it is typically close to the surface.

do. Unfortunately, it is likely to give useless results, because the people that you can easily survey are all unusual. For instance, in a linguistics department, they will:

- Know more languages than the average person,
- Listen more carefully for small distinctions between sounds,
- Know and apply many linguistic rules and theories to answer your question. For instance, they will be able to apply rules to consciously check for number agreement between the subject and object, *vs.* a more intuitive evaluation of whether the sentence "sounds right", and
- Have strong opinions about the theory you are trying to prove or disprove.

Opportunity sampling can catch you in many different ways,

- Sampling by door-to-door questionnaire during the day will exclude most people who have full time jobs.
- Sampling the weather by calling people on the telephone will give far too much attention to the weather in urban areas.
- Worst of all are the correlations that you do not anticipate.

Capture-Recapture Sampling: Capture-recapture sampling is a clever way to count things that you can't find. It's often used in the fields of zoology and ecology to count animal populations like frogs or birds. It works like this:

1. You go out (on Monday) to your favourite swamp and catch N frogs. You mark them in some way that doesn't hurt them. You will be depending on the fact that your marked frogs are neither easier nor harder to capture again than they were to capture the second time. For instance, if the tags slow down the marked frogs so they all get eaten, then you will recapture no frogs, and have no data.
2. You release the marked frogs.
3. You wait a while, until the marked frogs have had a chance to spread out across the swamp.
4. You go back on Tuesday and hunt exactly the same way as in step 1, and catch M marked frogs.

Then, you calculate. On you second hunt, you caught M of N frogs. Thus (assuming $M \gg 1$, so we don't have to worry about Good-Turing and its ilk), we caught our marked frogs with probability

$$P(\text{capture Tuesday}|\text{capture Monday}) = M/N. \tag{1}$$

If we can assume that the capture/tagging/release process did not affect the frogs, then

$$P(\text{capture Tuesday}|\text{capture Monday}) = P(\text{capture Monday}), \tag{2}$$

Thus, in order to capture the N frogs on Monday, there must be a population of about $N/P(\text{capture Monday})$ frogs in the swamp, so that your first hunt, capturing frogs with probability $P(\text{capture Monday})$ would have turned up N frogs. A bit of algebra then tells you that the swamp contains about N^2/M frogs.

A linguistic example of this might be a study of internet jokes and stories. You could capture a story, mark it by making some trivial change and send it back out. Studying how many of your marked stories survived and propagated might tell you how the jokes travel and how many they were. Incidentally, this technique was used as an example in the appendix of Bayes' original publication of his theorem [Bayes, 1763].

6 Recommended Reading

1. http://techniques.geog.ox.ac.uk/mod_2/glossary/samp.html
2. http://www.statcan.ca/english/edu/power/ch13/sample/sample.htm
3. http://www.stat.lsu.edu/faculty/moser/exst7012/exst7012.html
4. http://www.cee.vt.edu/program_areas/environmental/teach/smprimer/adaptive/adaptv.html
5. Thompson [1992]

References

T. Bayes. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. Communicated posthumously by Mr. Price.

Ch. Trib. Chicago daily tribune. http://www.loc.gov/exhibits/treasures/trm145.html, 1948. “Dewey Defeats Truman” front page, 3 November 1948. Image from the U.S. Library of Congress, Gift of Ms. Barbara, Diamond City, Arkansas.

Political Archive – 1948 Election. Eagleton Institute of Politics, Rutgers, the State University of New Jersey, 2003. URL http://www.eagleton.rutgers.edu/e-gov/e-politicalarchive-1948election.htm. http://www.eagleton.rutgers.edu/e-gov/e-politicalarchive-1948election.htm.

Darrell Huff. *How to Lie with Statistics*. W. W. Norton & Co., New York, 1954. ISBN 0-393-09426-X (republished in 1984).

S. K. Thompson. Adaptive cluster sampling. *J. American Statistical Association*, 85:1050–1059, 1990.

S. K. Thompson. *Sampling*. Wiley, New York, 1992.