

## Sampling Frontiers

Sharon Lohr  
Arizona State University

Presented at the Annual Dinner of the Washington Statistical Society, June 10, 2003

“I think the odds are no better than 50-50 that our present civilization on Earth will survive to the end of the present century.” Actually, these are not my words; they are a quote from a new book by Britain’s Astronomer Royal, Sir Martin Rees. The book has the cheerful title *Our Final Hour*. The original title in Britain was *Our Final Century*; the author said on NPR that the title was changed to *Our Final Hour* for the American version since Americans want instant gratification. If smallpox, terrorism, global warming, SARS, anthrax, and meteorites crashing into the earth aren’t enough, Rees talks about genetically engineered superpathogens and how it is likely that some high-energy physics experiment will go wrong and cause the earth to implode. Just reading the dust jacket of the book might make you want to crawl to some undisclosed location.

Of course doomsaying is not new, or unique to our times. In 1956, when Gertrude Cox became president of the American Statistical Association, there was no shortage of apocalyptic thinking. The hydrogen bomb had been developed in 1952, and the USSR had tested their first hydrogen bomb in 1953. The Army-McCarthy hearings had just taken place in 1954. Cox began her presidential address by saying:

Civilization is not threatened by atomic or hydrogen bombs; it is threatened by ourselves. We are surrounded with ever widening horizons of thought, which demand that we find better ways of analytical thinking. We must recognize that the observer is part of what he observes and that the thinker is part of what he thinks. We cannot passively observe the statistical universe as outsiders, for we are all in it.

Her address was titled “Statistical Frontiers.” In it, Cox took her listeners on a tour of the three major continents of the statistical universe: “(1) descriptive methods, (2) design of experiments and investigations, and (3) analysis and theory.” “Sample design territory,” in Cox’s universe, was part of the “design of investigations” continent. In a very short section of her address—remember, her primary area of interest was experimental design, not sample surveys—Cox spoke about the frontiers to be cleared in sampling territory.

You may have noticed that the title for my talk, “Sampling Frontiers,” is shamelessly similar to the title of Cox’s address. My goals are, however, not as expansive as hers. I would like for you to revisit “sample design territory” with me, 47 years after Cox identified the frontiers. Although some of Cox’s frontierland in sampling is now developed as she foresaw, many of the frontiers she identified remain to be explored today and new frontiers have opened.

Cox started her tour of “sample design territory” with the settled “simple random sampling country,” where members of the sample are drawn independently with equal probabilities. Cox wrote:

On the frontier between this country and the other countries of this area, there are two problems: (1) How could the present sampling procedures be improved if the observations followed a standard distribution form? (2) What are the effects of nonrandomness? ...

Cox's statement about improving sampling procedures by assuming a distribution for the observations refers in part to using knowledge about the distribution of corporate income to help in the construction of strata, and other design aspects. But I think she was also asking whether models can be used in the estimation and that she anticipated research in model-based sampling.

1956 was not long after the publication of the sampling books by Deming, Cochran, and Hansen, Hurwitz and Madow. For the most part, these books took a design-based approach to inference; that is, we use some sort of random selection in designing the sample, and the random selection procedure is used for inference. In the design-based approach, the quantity being measured on a unit—whether income, number of acres devoted to alfalfa, or how many pianos are in the dwelling unit—is considered to be a fixed constant. The only random variables used in design-based inference are indicator variables that tell whether a population unit is included in the sample or not. This is different from many other areas of statistics, where inference is based on a model for the particular response being measured; the most common example is that from introductory statistics courses where we assume that  $X_1, \dots, X_n$  are independent and identically normally distributed. A pure model-based approach to inference from surveys treats the sampling design as irrelevant; the assumed probability distribution of the responses provides the basis for inference. The issue of whether to use the survey design or a stochastic model (or both) for inference from surveys continues to be debated. Or, as Cox put it:

...The inhabitants of these frontiers invade the settled areas frequently, and frontier battles result.

Cox also asked how robust inferences are to violations of the sampling design. Robust methods and outlier detection in complex survey samples are just beginning to be explored, and there is still much to be done in this area.

Having foreseen in two sentences a controversy that still lingers, Cox moves on:

Next, we must cross the systematic sampling country. It is very difficult to secure permission from a statistician to enter this country. However, it is densely settled mostly by older people who have lived here all their lives. We frequently hear about uprisings and renewed efforts of this group to acquire all the advantages of the simple random sampling country.

Why did Cox write this? In the 1950's many argued that systematic sampling was just as good as a random sample. In systematic sampling, the population members are organized in a list, and one starts at a random place and then takes every 10<sup>th</sup> or 20<sup>th</sup> or 50<sup>th</sup> element as the sample. But without knowing the structure of the population, one cannot find standard errors of estimates. If the population is ordered, say, as husband, wife, husband wife and you take every 10<sup>th</sup> person,

your sample will consist of either all men or all women—but you, not knowing the population structure, won't be aware that your sample is flawed. Cox, though, wrote:

It appears that settlement in the systematic sampling country can safely be recommended if one of the following condition exists, (1) the order of the population is essentially random, or (2) several strata are to be used, with an independent systematic sample drawn from each stratum.

The second point, of using several independent systematic samples to be able to calculate standard errors, was truly groundbreaking for the development of sampling theory. The great Indian statistician Mahalanobis came up with this idea in 1946, calling it interpenetrating subsampling. Mahalanobis's idea was to perform several independent replications of the basic survey design, and use the variability among replicates to calculate standard errors for the estimates. This method avoids the problem of the population that alternates man, woman, man, woman; if the systematic sampling design is replicated, then some of the samples will be all men, but others will be all women, and the variability among the means of the independent systematic samples will allow you to give a standard error for your estimate of the population mean. This idea led to the random group method, in which the sample is divided randomly into several smaller samples, each of which mirrors the original sampling design. The variability among the estimates from the smaller samples may then be used to find standard errors. The next extension of this basic idea was to create many subsamples, which are allowed to overlap. This led to the replication methods used for variance estimation today in complex surveys, such as balanced repeated replication, jackknife, and bootstrap. Computer-intensive inference is still relatively young, and there is much territory to be developed here, particularly in conjunction with the development of robust estimation methods for survey data.

An exciting related modern frontier is the use of inverse sampling, proposed by Hinkins, Oh, and Scheuren (1997), whereby one takes a complex survey and constructs subsamples that can be analyzed as simple random samples. This allows the analyst to use standard methods to analyze the data; this idea has a lot of potential.

Cox then goes on to discuss stratified random sampling country and problems in optimal allocation, then moves on to large sample surveys. She wrote about the then-recent developments of the Hansen-Hurwitz and Horvitz-Thompson estimators:

If you are interested in large sample surveys, you will want to visit the multi-stage sampling country. Here the first stage units may be selected with probability proportional to size, the second stage units with equal probability. An adjacent area has been explored where first stage units are selected with arbitrary probability.

In newer areas of the multi-stage sampling country more than one first stage unit per stratum is drawn in order to permit internal assessment of the sampling errors of estimates. Even here many of these large surveys have been relegated to the archives without securing the sampling errors of estimates. This is done perhaps because of the complexity of the estimating formulas. Electronic computing

machines are helping to settle this difficulty. In fact, the machines may open up even wider frontiers for settlement in the sample design countries.

Electronic computing machines were novelties in 1956. The very first stored-program computer—the UNIVAC (Universal Automatic Computer)—had been delivered to the Bureau of the Census in March, 1951. The IBM 650 came at the end of 1954; although it was designed for colleges and businesses, its \$200,000<sup>1</sup> price was a bit steep, even with the college discounts.

Note that Cox referred to “electronic computing machines,” not “computers.” In the early 1950’s, when a statistician referred to a computer, he—and I use the pronoun “he” deliberately—generally meant one of the women employed to do the tedious calculations required for analysis of variance or regression. Women were thought to be suitable for this sort of statistical work because they were considered to be more patient and docile than men. After all, men might become bored with checking their calculations on hand-cranked calculators.<sup>2</sup> Women had another advantage for the job beyond their patience and docility—they could be paid less than half the salary that would be offered to a man for the same sort of work<sup>3</sup> and in 1956 were still cheaper than machine time. Cox herself had done a stint as a “computer” early in her career and supervised the women doing calculations shortly before moving to NCSU in 1940. It is my belief that this work in part led her to become such an evangelist for good experimental designs: with a well-designed experiment, the calculations are greatly simplified, making life easier for the rows of women hunched over their calculating machines and making it much more likely that the numerical answers would be correct.

I think Cox’s work as a computer also led her to instantly recognize the value of the new machines. Not only have they made computing Horvitz-Thompson estimates and standard errors trivial, they have also opened up survey methods for use by secondary analysts in other areas. Prior to software such as SUDAAN, Wesvar, and the SAS PROC SURVEY’s, many persons analyzing data from a survey simply ignored the survey design. Now data from surveys is open to a much larger clientele, and we can explore analyses using nonparametric techniques or Markov Chain Monte Carlo methods—scarcely conceived of in 1956.

Cox ended her tour of sampling territory by saying:

... there are many internal political and economic frontiers to be cleared. These sampling countries now have fair control over sampling errors but relatively little over non-sampling errors. They realize the need to find an economic balance between investment on sample and investment on measurement technique. To these developing frontiers, we can add others such as: (1) What are the relative efficiencies of the various sampling plans? (2) What is the effect of nonresponse?

---

<sup>1</sup> \$1.35 million in 2003 dollars

<sup>2</sup> All of the advertisements I have seen for the Marchant calculators of the 1950’s feature men—not operating the machines, of course, but standing proudly over the machines. One ad shows a naked discus thrower by the machine, with the caption “Performance.”

<sup>3</sup> See Gleick, J. (1992). *Genius: The Life and Science of Richard Feynman*. New York: Vintage Books, p. 181. During the Manhattan project at Los Alamos, the Marchant calculators were operated largely by wives of scientists, “working at three-eighths salary.”

and (3) What is an efficient method to sample for scarce items? Efforts are being made to clear out the underbrush and to settle some of this frontier area around the sampling territory.

These are still problems today. Many authors have emphasized the need for total survey design, where resources are allocated to the sources of error so that the total survey error is minimized. Nonresponse is increasing in many surveys. Many traditional methods do not work as well as they used to; many people are forgoing landline telephones and rely solely on cell phones and high-speed internet for communication. Methods for using this new technology need much exploration.

I think it significant that Cox put sampling in the “design of experiments” continent rather than in the “analysis and theory” or “descriptive methods” continents. Cox realized that the design of a sample survey is the key to its success, and the emphasis on survey design and the use of designed experiments for identifying and reducing sources of non-sampling error have been notable among the many achievements of statisticians in the federal statistical agencies. I think there is room for even greater use of designed experiments for improving survey quality, and some very exciting work on reducing nonresponse and using alternative modes of data collection is being done by the Research, Planning, and Evaluation division of the Census Bureau. One promising frontier for reducing nonsampling error is the area of ethnographic research, where social scientists investigate how questions and question orderings are perceived by survey participants. I think the next step in ethnographic research is to use it more in conjunction with designed experiments—the combination has the potential to greatly improve the accuracy of surveys and the decennial census, and is a sound “investment in measurement technique.”

The last frontier Cox mentioned in her address was the training frontier. She said:

We need to learn how to educate children to think scientifically, how to select those best fitted for scientific research and how to train them. We must go into the recruiting and training frontiers. More consideration will need to be given to strengthening high school education in science and mathematics. We need to find ways to direct the ambitions of boys and girls. What we are facing is not a shortage of ability and talent, but a shortage of trained talent at all levels.

Is this still a frontier? I think the answer is definitely “yes.” One might even argue that kudzu is reclaiming previously developed land in the training frontier. I think that the fields of mathematics and statistics have missed many opportunities to attract young people to our disciplines. Since September 11, college campuses have been filled with young people who want to do something for their country but aren’t sure how they can best contribute; many of them, in their desire to do something that “makes a difference” have now ended up in biology or psychology or forensics. Very few have gone into math or statistics; when I ask students why they chose to go into biology or psychology, they talk about the excitement of making new discoveries and doing something that could help other people or the ecosystem. We need to do a much better job of portraying statistics as an exciting field where we make fundamental contributions to every area of science. I highly recommend David Salsburg’s book titled *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Salsburg makes

many of the great statisticians of the time come alive, and shows how crucial statistical methods have been to everything from brewing beer to curing cancer. He also portrays statistics much the way Cox thought of it, as an activity done collaboratively.

I was fortunate to have my first statistics course at Wisconsin from Bill Hunter, whom I still consider to be the greatest teacher I have ever encountered. He emphasized that statistics is a field for people who are curious about everything: if you want to learn about anthropology, you can work with an anthropologist; if you want to learn about engineering, you can collaborate in that area. Unlike in other disciplines, he said, you're not restricted to one narrow focus for the rest of your career but the whole world is open to you. It's clear that Gertrude Cox shared this view. In 1959 a young woman wrote to the University of Michigan: "Gentlemen: I am interested in becoming a statistician, but before I decide, I would like more information in the field for women." The letter somehow ended up with Gertrude Cox, who despite all of her responsibilities responded with a two-page letter: "I believe very strongly that girls should prepare for a profession, even though their main aim is to get married . . . I could give a list of a variety of interesting areas in which I have cooperated such as, the best methods of raising flowers in a greenhouse, development and selection of new varieties of corn, the nutritional problems among the Indian children in Guatemala, how to sample gold in South Africa, variations in ways to make instant frosting for cakes, how to evaluate the effectiveness of fly sprays, and many others."<sup>4</sup>

Cox viewed sampling as an integral part of the statistical universe. But to some extent, survey sampling territory has separated from the statistical Pangaea, and many statisticians in other areas have little knowledge of survey sampling. Many academic departments have no statisticians who have studied sampling. To be sure, survey samplers are partly responsible for their isolation. At conferences, the samplers often form their own sub-conference, and many do not venture out of the territory. In part, too, the specialized terminology has formed barriers around the territory. At one conference, I overheard the sentence: "We used BRR on a PPS sample of PSU's, with MI after NRFU."

At this time, however, many of the frontiers in all areas of statistics involve data that are dependent. Survey samplers are the experts here, with over 50 years of experience collecting and analyzing dependent data. I think one of today's frontiers is using sampling methods to solve problems in other areas of statistics and science. We take our abilities and knowledge and skills and experience for granted, and forget about how much we can contribute. Here are four examples where sampling theory and methods could be extremely useful:

Markov Chain Monte Carlo algorithms are very popular for fitting a variety of models, and are now essential tools for Bayesian statistics. MCMC methods generate samples from a posterior distribution, and use these samples to estimate the quantities of interest. They can also take a

---

<sup>4</sup> Cox was known for her encouragement of young people of both sexes. You have probably heard the story (from Anderson et al., 1979) of how she became the founding chair of the Department of Experimental Statistics at North Carolina State College. In 1940, George Forster of NCSC asked George Snedecor of Iowa State to send him a list of people who would be suitable for the position. Snedecor came up with a list of 10 names—all men, of course—and showed the list to Cox for her opinion before mailing it. She asked, "Why didn't you put my name on the list?" Snedecor then added a footnote to the letter: "Of course if you would consider a woman for this position I would recommend Gertrude Cox of my staff."

long time to run, and it is not always easy to know when they have converged. Often several parallel runs are used, with different starting values—this is essentially the same idea that Cox discussed for using several independently selected systematic samples so as to assess the standard error of estimates. Andrew Gelman and other researchers are now studying using cluster sampling methods to improve the performance of MCMC on hierarchical data sets.

Data mining is currently very popular for extracting information from massive data sets, such as those collected by credit card companies on all credit transactions. I just did a Google search on the term “data mining” and came up with over 1 million matches. One problem with the huge data sets that are available, though, is that fitting models and understanding multivariate relationships can be very difficult when your data set has 100 million observations. What does a p-value of 0.01 mean when  $n = 100$  million? There have been two main approaches for model-fitting with massive data sets. One approach is to scale up statistical methods and algorithms so they make sense with huge data sets. The other approach, called data compression, is to extract a smaller data set from the massive data set. The original, massive data set is sometimes called the “mother data” in data mining, and it is desired to compress the data in such a way that the smaller data set has the same properties as the mother data. Doesn’t this sound like a sampling problem to you? It’s just that “population” is not as catchy of a term as “mother data”—“mother data” gives me a picture of clever extra-terrestrial beings returning to their “mother ship” after they have extracted the useful information from the “mother data.”

Some data miners have taken simple random or stratified samples for model building purposes. Some recent research, however, has concentrated on algorithms for a procedure called “data squashing.” Data squashing is not sampling: it constructs a reduced data set of pseudo-observations that have the same moments on selected variables as the “mother data.” Some papers indicate that the squashed data sets work much better than simple random samples. But could data squashing have the same problem that Neyman wrote about in 1934 with respect to samples that are selected deliberately to match certain known characteristics of the population? In the late 1920’s, the Italian statisticians Gini and Galvani chose 29 districts that gave the averages, on a dozen variables, of all 214 districts in the 1921 Italian census, and claimed that purposive sampling gave a representative sample. But Neyman showed that their purposive sample performed very badly on other variables that were not used in selecting the sample. Preserving moments in data compression methods may not guarantee that multivariate relationships are preserved. Instead, why not compress the data by taking a probability sample whose weights have been calibrated to match known totals in the mother data?

A third area where sampling methods might be employed is in bioinformatics. Many methods for matching sequences of DNA, and for constructing phylogenetic trees based on genetic information, assume that all nucleotides are independent. This is almost certainly not true, since mutations often affect a string of nucleotides at once. I think that sampling theory could make great contributions to statistical genetics, and some researchers are now starting to look at some of the complications caused by dependent data. Well-designed samples could also be of great use in studying the genetic components of certain diseases. On May 27, the headline on reuters.com announced: “British scientists launched a national hunt on Tuesday for families with a history of testicular or prostate cancer to help in the search for genes related to the diseases. They are looking for men who have three or more relatives who developed prostate cancer before the age of

70 or who have two or more family members with testicular cancer. By examining the genetic profiles of men with the disease, they hope to identify genes that increase a man's risk of developing the cancers.” We can learn much from such samples of volunteers, but how much better would the estimates and procedures be with designed samples?

One last example involves worldwide clinical trials. Jim Ware has pointed out that many clinical trials are carried out entirely in North America, and other governments often have doubts whether data collected solely on Americans apply in their country. He has suggested doing trials with small sample sizes in other countries, then using hierarchical models to obtain improved estimates for the countries or ethnic groups with small sample sizes in the trials. This is the same problem as small area estimation in surveys, where we use models to improve the accuracy of estimates in regions such as counties where the sample size from the survey is too small to produce a reliable estimate by itself.

As I mentioned at the beginning of this talk, Cox devoted only a small part of her presidential address to sampling. But in that small part, she cut straight to the heart of the field and what it needed to do. I find it interesting that she focussed on all aspects of sampling—theoretical development, sampling errors, use of models, design, use of designed experiments, nonresponse, robustness, and perhaps most importantly, communication with users and training students. Despite all of the sources for anxiety in 1956, Cox did not sit around wailing that the end is near and there’s nothing we can do about it. Instead, she concluded her address by saying:

One thing is certain, we are at the beginning of a new age—an age that will be richer and will offer more and more opportunities for people whose minds are flexible and who are eager to increase their area of awareness.

I can think of no words more fitting for 2003. I also cannot imagine a greater honor than to be given the first award named after this remarkable woman. Thank you so very very much.

## References

Anderson, R.L., Monroe, R.J. and Nelson, L.A. (1979). “Gertrude M. Cox—A Modern Pioneer in Statistics.” *Biometrics*, **35**, 3-7.

Cox, G. M. (1945). “Opportunities for Teaching and Research,” *Journal of the American Statistical Association*, **40**, 71-74.

Cox, G. M. (1957). “Statistical Frontiers,” *Journal of the American Statistical Association*, **52**, 1-12.

Gelman, A. and Rubin, D. (1992). “Inference from Interactive Simulation Using Multiple Sequences,” *Statistical Science*, **7**, 457-472.

Hinkins, S., Oh, H.L., and Scheuren, F. (1997). “Inverse Sampling Design Algorithms,” *Survey Methodology*, **23**, 11-21.

Madigan, D., Raghavan, N., DuMouchel, W., Nason, M., Posse, C., and Ridgeway, G. (2002), "Likelihood-based data squashing: A modeling approach to instance construction," *Data Mining and Knowledge Discovery*, **6**, 173-190.

Mahalanobis, P.C. (1946). "Recent experiments in statistical sampling in the Indian Statistical Institute," *Journal of the Royal Statistical Society*, **109**, 325-370

Neyman, J. (1934). "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection." *Journal of the Royal Statistical Society*, **97**, 558-606.

North Carolina State University website, "Gertrude M. Cox, First Lady of Statistics."  
<http://www.lib.ncsu.edu/archives/exhibits/cox/career.html>

Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: W.H. Freeman and Company.

Stinnett, S. (1990). "Women in Statistics: Sesquicentennial Activities." *The American Statistician*, **44**, 74-80.