

From

Lin, X., Genest, C., Banks, D., Molenberghs, G., Scott, D. and Wang, J-L (2014). *Past, Present and Future of Statistical Science*. Boca Raton, FL: CRC Press.

9

Reflections on a statistical career and their implications

R. Dennis Cook

School of Statistics

University of Minnesota, Minneapolis, MN

This chapter recounts the events that steered me to a career in statistics and describes how my research and statistical temperament were set by my involvement in various applications. The discussion encompasses the historical and contemporary role of statistical diagnostics in practice and reflections on the importance of applications in the professional life of a statistician.

9.1 Early years

It was mostly serendipity that led me to a career in statistics.

My introduction to statistics started between my Sophomore and Junior years in high school. At the time I was looking for summer employment so I could earn money to customize my car — neat cars elevated your social standing and attracted the girls. I was fortunate to secure summer and eventually after-school employment with the Agronomy Department at Fort Assiniboine, an agriculture experimentation facility located just outside Havre, Montana. The surrounding area is largely devoted to wheat production, thousands and thousands of acres of spring and winter wheat. The overarching goal of the Agronomy Department was to develop contour maps of suggested fertilization regimes for use by wheat farmers along the High Line, a run of about 130 miles between Havre and Cut Bank, Montana, and to occasionally develop targeted recommendations for specific tracts of land at the request of individual farmers.

I continued to work at Fort Assiniboine until I graduated from high school, at which point I enlisted in the military to avoid the uncertainty of the draft. After fulfilling my military obligation, which ended just before the buildup to the Vietnam war, I returned to full-time employment at Fort Assiniboine while

pursuing an undergraduate degree at Northern Montana College. In order to meet the needs of the surrounding community, Northern offered four degree programs — nursing, education, liberal arts and modern farming methods. Given the choices, I decided that education was my best bet, although I was ambivalent about a speciality. While in the military I developed a strong distaste for standing in line, so on the first day of registration when I encountered long lines everywhere except for mathematics education, my choice was clear. I continued to work at Fort Assiniboine for four more years until I completed my undergraduate degree in mathematics education with a minor in biology.

My duties during the seven years of employment at Fort Assiniboine focused on statistics at one level or another, the same cycle being repeated year after year. Starting in the late winter, we would prepare the fertilizer combinations to be tested in the next cycle and lay out the experimental designs on paper. We typically used randomized complete block designs, but had to be prepared with completely randomized and Latin square designs, since we never knew what the experimental location was like before arriving with the planting crew. Split plot and split block designs were also used from time to time. Experimental plots would be planted in the spring (or the late fall in the case of winter wheat), and tended throughout the summer by keeping the alleys between the plots free of weeds. Plots were harvested in the fall, followed by threshing and weighing the wheat. Most of the winter was spent constructing analysis of variance tables with the aid of large desktop Monroe calculators and drawing conclusions prior to the next cycle of experimentation.

During my first year or so at Fort Assiniboine, I functioned mostly as a general laborer, but by the time I finished high school I had developed an appreciation for the research. I was fortunate that, from the beginning, the Department Head, who had a Master's degree in agronomy with a minor in statistics from a Canadian university, encouraged me to set aside time at work to read about experimental design and statistical methods. This involved studying Snedecor's text on *Statistical Methods* and Fisher's monograph on *The Design of Experiments*, in addition to other references. The material came quickly for me, mostly because nearly all that I read corresponded to something we were actually doing. But I recall being a bit baffled by the need to select a significance level and the role of p -values in determining recommendations. The possibility of developing a formal cost function to aid our recommendations did not arise until graduate school some years later. My undergraduate education certainly helped with the mathematics, but was little help with statistics since the only directly relevant offering was a course in probability with a cursory treatment of introductory statistics.

I was eventually given responsibility for nearly all aspects of the experimental trials at the Fort. I had learned to assess the experimental location, looking for hollows and moisture gradients, and to select and arrange an appropriate design. I learned that mis-entering a number could have costly consequences. A yield of 39 bushels mis-entered as 93 bushels per acre could make a non-significant factor seem highly significant, resulting in an unjustified costly

recommendation to the wheat farmers. It is for this reason that I instituted “parallel computing.” Two of us would sit at adjacent desks and simultaneously construct analysis of variance tables, checking that our results matched at each step of the analysis. A mismatch meant that we had to repeat the calculation in full, since there was no way of recovering what we had entered. We would occasionally lose an experiment at a remote location because the grain was too wet to harvest during the window of opportunity. In an effort to recover some information, I came up with the idea of estimating the number of seed heads per foot of row. That operation required only counting and the moisture content of the grain was irrelevant. A few pilot experiments showed that the count was usefully correlated with the grain weight, so we were able to gain some information from experiments that would be otherwise lost.

During my final year at Northern Montana College, I was required to spend six months student teaching at the local high school where I taught junior algebra and sophomore biology. I found actual teaching quite rewarding, but my overall experience was a disappointment because colorless non-teaching duties dominated my days. The Department Head at the Fort had been encouraging me to pursue a graduate degree in statistics or perhaps mathematics and, after my experience student teaching, I decided to follow his advice. I applied to four universities, two in mathematics and two in statistics, that did not require a fee to process my application because finances were extremely tight. My decision rule was to accept the first that offered a fellowship or assistantship. The following fall I began my graduate studies in statistics at Kansas State University, aided by a traineeship from the National Institutes of Health and subsequently a fellowship under the National Defense Education Act, which was enacted in response to the Soviet Union’s successful launch of Sputnik and President Kennedy’s moon initiative.

Although my degree from Kansas State was in statistics, my dissertation was in genetics; it was entitled “The Dynamics of Finite Populations: The Effects of Variable Selection Intensity and Population Size on the Expected Time to Fixation and the Ultimate Probability of Fixation of an Allele.” I enjoyed seeing genetic theory in action and many hours were spent during my graduate career conducting laboratory experiments with *Drosophila melanogaster*. My first paper was on Bayes’ estimators of gene frequencies in natural populations. My background and fellowships enabled me to complete my PhD degree in three years, at which point I joined the then nascent School of Statistics at the University of Minnesota, with an appointment consisting of intramural consulting, teaching and research, in roughly equal proportions. I continued my genetics research for about four years until I had achieved tenure and then began a transition to statistical research, which was largely stimulated and guided by my consulting experiences.

9.2 Statistical diagnostics

In his path-breaking 1922 paper “On the mathematical foundations of theoretical statistics,” R.A. Fisher established the contemporary role of a statistical model and anticipated the development of diagnostic methods for model assessment and improvement. Diagnostics was a particularly active research area from the time of Fisher’s death in 1962 until the late 1980s, and the area is now an essential ingredient in Fisher modeling.

9.2.1 Influence diagnostics

My involvement with diagnostics began early in my career at Minnesota. A colleague from the Animal Science Department asked me to review a regression because his experiment had apparently produced results that were diametrically opposed to his prior expectation. The experiment consisted of injecting a number of rats with varying doses of a drug and then measuring the fraction of the doses, which were the responses, that were absorbed by the rats’ livers. The predictors were various measurements on the rats plus the actual dose. I redid his calculations, looked at residual plots and performed a few other checks that were standard for the time. This confirmed his results, leading to the possibilities that either there was something wrong with the experiment, which he denied, or his prior expectations were off. All in all, this was not a happy outcome for either of us.

I subsequently decided to use a subset of the data for illustration in a regression course that I was teaching at the time. Astonishingly, the selected subset of the data produced results that clearly supported my colleague’s prior expectation and were opposed to those from the full data. This caused some anxiety over the possibility that I had made an error somewhere, but after considerable additional analysis I discovered that the whole issue centered on one rat. If the rat was excluded, my colleague’s prior expectations were sustained; if the rat was included his expectations were contradicted. The measurements on this discordant rat were accurate as far as anyone knew, so the ball was back in my now quite perplexed colleague’s court.

The anxiety that I felt during my exploration of the rat data abated but did not disappear completely because of the possibility that similar situations had gone unnoticed in other regressions. There were no methods at the time that would have identified the impact of the one unusual rat; for example, it was not an outlier as judged by the standard techniques. I decided that I needed a systematic way of finding such influential observations if they were to occur in future regressions, and I subsequently developed a method that easily identified the irreconcilable rat. My colleagues at Minnesota encouraged me to submit my findings for publication (Cook, 1977), which quickly took on a life of their own, eventually becoming known as *Cook’s Distance*, although

no one sought my acquiescence. In 1982 I coauthored a fairly comprehensive research monograph on the state of diagnostic methods (Cook and Weisberg, 1982).

Encouraged by the wide acceptance of *Cook's Distance* and my other diagnostic contributions, and aided by a year-long fellowship from the Mathematics Research Center at the University of Wisconsin, I continued working in diagnostics with the goal of developing local differential geometric measures that might detect various influential characteristics of a generic likelihood-based analysis. In 1986 I read before the Royal Statistical Society a paper on a local likelihood-based technique for the development of diagnostics to detect influential aspects of an analysis (Cook, 1986).

Today models can be and often are much more complicated than those likely entertained by Fisher or in common use around the time that I was earnestly working on influence diagnostics. As a consequence, the methods developed prior to the 1990s are generally not applicable in more complicated contemporary contexts, and yet these contexts are no less affected by influential observations. Intricate models are prone to instability and the lack of proper influence diagnostics can leave a cloud of doubt about the strength of an analysis. While influence diagnostics have been keeping pace with model development largely through a series of important papers by Hongtu Zhu and his colleagues (Zhu et al., 2007, 2012), methods to address other diagnostic issues, or issues unique to a particular modeling environment, are still lagging far behind. Personally, I am reluctant to accept findings that are not accompanied by some understanding of how the data and model interacted to produce them.

9.2.2 Diagnostics more generally

A substantial battery of diagnostic methods for regression was developed during the 1970s and 1980s, including transformation diagnostics, various graphical diagnostics like residual plots, added variable plots (Cook and Weisberg, 1982), partial residual plots and CERES plots for predictor transformations (Cook, 1993), methods for detecting outliers and influential observations, and diagnostics for heteroscedasticity (Cook and Weisberg, 1983). However, it was unclear how these methods should be combined in a systematic way to aid an analysis, particularly since many of them addressed one issue at a time. For instance, diagnostics for heteroscedasticity required that the mean function be correct, regardless of the fact that an incorrect mean function and homoscedastic errors can manifest as heteroscedasticity. Box's paradigm (Box, 1980) for model criticism was the most successful of the attempts to bring order to the application of diagnostic methods and was rapidly adopted by many in the field. It consists essentially of iteratively improving a model based on diagnostics: an initial model is posited and fitted to the data, followed by applications of a battery of diagnostic methods. The model is then modified

to correct the most serious deficiencies detected, if any. This process is then iterated until the model and data pass the selected diagnostic checks.

Diagnostic methods guided by Box's paradigm can be quite effective when the number of predictors is small by today's standards, say less than 20, but in the early 1990s I began encountering many regressions that had too many predictors to be addressed comfortably in this way. I once spent several days analyzing a data set with 80 predictors (Cook, 1998, p. 296). Box's paradigm was quite useful and I was pleased with the end result, but the whole process was torturous and not something I would look forward to doing again. A different diagnostic paradigm was clearly needed to deal with regressions involving a relatively large number of predictors.

9.2.3 Sufficient dimension reduction

Stimulated by John Tukey's early work on computer graphics and the revolution in desktop computing, many dynamic graphical techniques were developed in the late 1980s and 1990s, including linking, brushing, scatterplot matrices, three-dimensional rotation and its extensions to grand tours, interactive smoothing and plotting with parallel coordinates. My first exposure to dynamic graphics came through David Andrews' Macintosh program called McCloud. At one point I thought that these tools might be used effectively in the context of diagnostics for regressions with a relatively large number of predictors, but that proved not to be so. While dynamic graphical techniques allow many plots to be viewed in relatively short time, most low-dimensional projective views of data can be interesting and ponderable, but at the same time do not necessarily provide useful information about the higher dimensional data. In regression for example, two-dimensional plots of the response against various one-dimensional projections of the predictors can be interesting as individual univariate regressions but do not necessarily provide useful information about the overarching multiple regression employing all predictors simultaneously. Many projective views of a regression seen in short time can quickly become imponderable, leaving the viewer with an array of disconnected facts about marginal regressions but little substantive knowledge about the full regression.

My foray into dynamic computer graphics was methodologically unproductive, but it did stimulate a modest epiphany in the context of regression that is reflected by the following question: Might it be possible to construct a low-dimensional projective view of the data that contains all or nearly all of the relevant regression information without the need to pre-specify a parametric model? If such a view could be constructed then we may no longer need to inspect many diagnostic plots and Box's paradigm could be replaced with a much simpler one, requiring perhaps only a single low-dimensional display as a guide to the regression. Stated more formally, can we find a low-dimensional subspace \mathcal{S} of the predictor space with the property that the response Y is independent of the predictor vector X given the projection $P_{\mathcal{S}}X$ of X onto \mathcal{S} ;

that is, $Y \perp\!\!\!\perp X | P_{\mathcal{S}}X$? Subspaces with this property are called dimension reduction subspaces. The smallest dimension reduction subspace, defined as the intersection of all dimension reduction subspaces when it is itself a dimension reduction subspace, is called the central subspace $\mathcal{S}_{Y|X}$ (Cook, 1994, 1998). The name “central subspace” was coined by a student during an advanced topics course in regression that I was teaching in the early 1990s. This area is now widely known as sufficient dimension reduction (SDR) because of the similarity between the driving condition $Y \perp\!\!\!\perp X | P_{\mathcal{S}_{Y|X}}X$ and Fisher’s fundamental notion of sufficiency. The name also serves to distinguish it from other approaches to dimension reduction.

The central subspace turned out to be a very effective construct, and over the past 20 years much work has been devoted to methods for estimating it; the first two methods being sliced inverse regression (Li, 1991) and sliced average variance estimation (Cook and Weisberg, 1991). These methods, like nearly all of the subsequent methods, require the so-called linearity and constant covariance conditions on the marginal distribution of the predictors. Although these conditions are largely seen as mild, they are essentially uncheckable and thus a constant nag in application. Ma and Zhu (2012) recently took a substantial step forward by developing a semi-parametric approach that allows modifications of previous methods so they no longer depend on these conditions. The fundamental restriction to linear reduction $P_{\mathcal{S}_{Y|X}}X$ has also been long recognized as a limitation. Lee et al. (2013) recently extended the foundations of sufficient dimension reduction to allow for non-linear reduction. This breakthrough, like that from Ma and Zhu, opens a new frontier in dimension reduction that promises further significant advances. Although SDR methods were originally developed as comprehensive graphical diagnostics, they are now serviceable outside of that context.

Technological advances resulted in an abundance of applied regressions that Box’s paradigm could no longer handle effectively, and SDR methods were developed in response to this limitation. But technology does not stand still. While SDR methods can effectively replace Box’s paradigm in regressions with many predictors, they seem ill suited for high-dimensional regressions with many tens or hundreds of predictors. Such high-dimensional regressions were not imagined during the rise of diagnostic or SDR methods, but are prevalent today. We have reached the point where another diagnostic template is needed.

9.2.4 High-dimensional regressions

High-dimensional regressions often involve issues that were not common in the past. For instance, they may come with a sample size n that is smaller than the number of predictors p , leading to the so called “ $n < p$ ” and “ $n \ll p$ ” problems. Some type of specialized structure is needed for the analysis of high-dimensional regressions since they cannot be addressed typically by using traditional methods.

One favored framework imposes a sparsity condition — only a few of the many predictors are relevant for the regression — which reduces the regression goal to finding the relevant predictors. This is now typically done by assuming a model that is (generalized) linear in the predictors and then estimating the relevant predictors by optimizing a penalized objective function. An analysis of a high-dimensional regression based on this approach involves two acts of faith.

The first act of faith is that the regression is truly sparse. While there are contexts where sparsity is a driving concept, some seem to view sparsity as akin to a natural law. If you are faced with a high-dimensional regression then naturally it must be sparse. Others have seen sparsity as the only recourse. In the logic of Bartlett et al. (2004), the bet-on-sparsity principle arose because, to continue the metaphor, there is otherwise little chance of a reasonable payoff. In contrast, it now seems that reasonable payoffs can be obtained also in abundant regressions where many predictors contribute useful information on the response, and prediction is the ultimate goal (Cook et al., 2012).

The second and perhaps more critical act of faith involves believing the data and initial model are flawless, apart from the statistical variation that is handled through the objective function. In particular, there are no outliers or influential observations, any curvature in the mean function is captured adequately by the terms in the model, interactions are largely absent, the response and predictors are in compatible scales and the errors have constant variation. It has long been recognized that regressions infrequently originate in such an Elysian condition, leading directly to the pursuit of diagnostic methods. I can think of no compelling reason these types of considerations are less relevant in high-dimensional regressions. Diagnostic methods can and perhaps should be used after elimination of the predictors that are estimated to be unrelated with the response, but this step alone may be inadequate. Failings of the types listed here will likely have their greatest impact during rather than after penalized fitting. For instance, penalized fitting will likely set the coefficient β of a standard normal predictor X to zero when the mean function in fact depends on X only through a quadratic term βX^2 . Findings that are not accompanied by an understanding of how the data and model interacted to produce them should ordinarily be accompanied by a good dose of skepticism.

9.3 Optimal experimental design

My interest in an optimal approach to experimental design arose when designing a comprehensive trial to compare poultry diets at six universities. Although the experimental diets came from a common source, the universities had different capabilities and facilities which made classical Box–Fisher–Yates

variance reduction designs difficult to employ, particularly since the underlying non-linear model called for an unbalanced treatment design.

Optimal experimental design was for many years regarded as primarily a mathematical subject. While applications were encountered from time to time, it was seen as largely a sidelight. Few would have acknowledged optimal design as having a secure place in statistical practice because the approach was too dependent on knowledge of the model and because computing was often an impediment to all but the most straightforward applications. During the 1970s and most of the 1980s, I was occasionally a party to vigorous debates on the relative merits of classical design versus optimal design, pejoratively referred to by some as “alphabetic design” in reference to the rather unimaginative design designations like D-, A- and G-optimality. Today classical and optimal design are no longer typically seen as distinct approaches and the debate has largely abated. The beginning of this coalescence can be traced back to technological advances in computing and to the rise of unbalanced experimental settings that were not amenable to classical design (Cook and Nachtshiem, 1980, 1989).

9.4 Enjoying statistical practice

Statistics has its tiresome aspects, to be sure, but for me the practice of statistics has also been the source of considerable pleasure and satisfaction, and from time to time it was even thrilling.

For several years I was deeply involved with the development of aerial survey methods. This included survey methods for snow geese on their molting grounds near Arviat on the west shore of Hudson Bay, moose in northern Minnesota, deer in southern Manitoba and wild horses near Reno, Nevada. It became apparent early in my involvement with these studies that the development of good survey methods required that I be actively involved in the surveys themselves. This often involved weeks in the field observing and participating in the surveys and making modifications on the fly.

The moose and deer surveys were conducted in the winter when foliage was largely absent and the animals stood out against a snowy background. Nevertheless, it soon became clear from my experience that aerial observers would inevitably miss some animals, leading to underestimation of the population size. This visibility bias would be a constant source of uncertainty unless a statistical method could be developed to adjust the counts. I developed different adjustment methods for moose and deer. Moose occur in herds, and it seemed reasonable to postulate that the probability of seeing an animal is a function of the size of its herd, with solitary animals being missed the most frequently. Adding a stable distribution for herd size then led to an adjustment method that resulted in estimates of population size that were in qualitative agreement with estimates from other sources (Cook and Martin,

1974). A different adjustment method for deer censuses was developed based on a design protocol that involved having two observers on the same side of the aircraft. The primary observer in the front seat called out and recorded all the deer that he saw. The secondary observer in the rear seat recorded only deer that the primary observer missed. The resulting data plus a few reasonable assumptions on the generation process led directly to adjusted population counts (Cook and Jacobson, 1979).

A version of mark-capture was developed for estimating population sizes of wild horses. The horses were marked by a tethered shooter leaning out the right side of a helicopter flying near tree-top level above the then running animals. The shooter's demanding task was to use a fancy paint-ball gun to mark the animal on its left rear quarter. I was the primary shooter during the development phase, and I still recall the thrill when the helicopter pulled up sharply to avoid trees or other obstacles.

9.5 A lesson learned

Beginning in my early days at Fort Assiniboine, my statistical perspectives and research have been driven by applications. My work in diagnostic methods originated with a single rat, and my attitude toward inference and diagnostics was molded by the persistent finding that plausible initial models often do not hold up when contrasted against the data. The development of SDR methods was stimulated by the inability of the then standard diagnostic methods to deal effectively with problems involving many variables. And, as mentioned previously, we are now at a point where a new diagnostic paradigm is needed to deal with the high-dimensional regressions of today. My interest in optimal design arose because of the relative rigidity of classical design. My contributions to aerial surveys would have been impossible without imbedding myself in the science. This has taught me a lesson that may seem retrospectively obvious but was not so for me prospectively.

Statistics is driven by applications which are propelled by technological advances, new data types and new experimental constructs. Statistical theory and methods must evolve and adapt in response to technological innovation that give rise to new data-analytic issues. High-dimensional data, which seems to dominate the pages of contemporary statistics journals, may now be overshadowed by "Big Data," a tag indicating a data collection so large that it cannot be processed and analyzed with contemporary computational and statistical methods. Young statisticians who are eager to leave a mark may often find themselves behind the curve when too far removed from application. The greatest statistical advances often come early in the growth of a new area, to be followed by a fleshing out of its nooks and crannies. Immersing oneself

in an application can bring a type of satisfaction that may not otherwise be possible.

References

- Bartlett, P.L., Bickel, P.J., Bühlmann, P., Freund, Y., Friedman, J., Hastie, T., Jiang, W., Jordan, M.J., Koltchinskii, V., Lugosi, G., McAuliffe, J.D., Ritov, Y., Rosset, S., Schapire, R.E., Tibshirani, R.J., Vayatis, N., Yu, B., Zhang, T., and Zhu, J. (2004). Discussions of boosting papers. *The Annals of Statistics*, 32:85–134.
- Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430.
- Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19:15–18.
- Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, 48:133–169.
- Cook, R.D. (1993). Exploring partial residual plots. *Technometrics*, 35:351–362.
- Cook, R.D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Physical Engineering Sciences*, American Statistical Association, Washington, DC, pp. 18–25.
- Cook, R.D. (1998). *Regression Graphics*. Wiley, New York.
- Cook, R.D., Forzani, L., and Rothman, A.J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *The Annals of Statistics*, 40:353–384.
- Cook, R.D. and Jacobson, J.O. (1979). A design for estimating visibility bias in aerial surveys. *Biometrics*, 34:735–742.
- Cook, R.D. and Martin, F. (1974). A model for quadrant sampling with “visibility bias.” *Journal of the American Statistical Association*, 69:345–349.
- Cook, R.D. and Nachtsheim, C.J. (1980). A comparison of algorithms for constructing exact D-optimal designs. *Technometrics*, 22:315–324.
- Cook, R.D. and Nachtsheim, C.J. (1989). Computer-aided blocking of factorial and response surface designs. *Technometrics*, 31:339–346.