Statistical Science
2003, Vol. 18, No. 2, 158–167

A Short Prehistory of the Bootstrap

Peter Hall

Abstract. The contemporary development of bootstrap methods, from the time of Efron's early articles to the present day, is well documented and widely appreciated. Likewise, the relationship of bootstrap techniques to certain early work on permutation testing, the jackknife and cross-validation is well understood. Less known, however, are the connections of the bootstrap to research on survey sampling for spatial data in the first half of the last century or to work from the 1940s to the 1970s on subsampling and resampling. In a selective way, some of these early linkages will be explored, giving emphasis to developments with which the statistics community tends to be less familiar. Particular attention will be paid to the work of P. C. Mahalanobis, whose development in the 1930s and 1940s of moving-block sampling methods for spatial data has a range of interesting features, and to contributions of other scientists who, during the next 40 years, developed half-sampling, subsampling and resampling methods.

Key words and phrases: Block bootstrap, computer-intensive statistics, confidence interval, half-sample, Monte Carlo, moving block, resampling, permutation test, resample, sample survey, statistical experimentation, subsample.

1. INTRODUCTION

From some viewpoints it is impossible to know just where and when the bootstrap began. If one defines (as I believe one should) a bootstrap estimator to be the result of replacing an unknown distribution function, in the definition of a parameter, by its empirical counterpart, then the sample mean is the bootstrap estimator of the population mean. Adopting this view, one could fairly argue that the calculation and application of bootstrap estimators has been with us for centuries. We could claim that general first-order limit theory for the bootstrap was known to Laplace by about 1810, and that second-order properties were developed by Chebyshev later in the 19th century.

However, such a formulaic definition of the bootstrap overlooks its most potent, and perhaps most appealling, ingredient: its connection to random sampling from the sample. This feature is the key to our ability to compute most bootstrap estimators other than means. Regardless of how one defines the bootstrap, one can argue that the variability one encounters when sampling from the sample models that which is found when sampling from the population, and that this is a major attraction of bootstrap methods.

It was by marrying the power of Monte Carlo approximation with an exceptionally broad view of the sort of problem that bootstrap methods might solve, that Efron (1979a, b) famously vaulted earlier resampling ideas out of the arena of sampling methods and into the realm of a universal statistical methodology. Arguably, the prehistory of the bootstrap encompasses pre-1979 developments of Monte Carlo methods for sampling. The bootstrap's conventional history dates from the late 1970s.

The notion of resampling from the sample is removed only slightly from that of sampling from a finite population. Unsurprisingly, then, a strong argument can be made that important aspects of the bootstrap's roots lie in methods for sample surveys. In the 1940s and 1950s, U.S. statisticians (particularly those working for government statistical offices) were developing

Peter Hall is Professor of Statistics, Statistical Science Program, Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia (e-mail: halpstat@wintermute.anu.edu.au).

techniques for stochastic, as distinct from systematic, sampling from finite populations and were exploring "hybrid" methods, based (for example) on a stochastic start followed by deterministic extrapolation. In India, where survey sampling for estimating crop yields was pioneered, stochastic spatial sampling methods were under development 80 years ago. The origins of stochastic simulation are very much older, however; see Stigler (1991; 1999, Chapter 7).

Random design and random sampling methods, and the debate over their merits, owe much to the early work of Fisher, Tukey, Deming and others. See, for example, Hansen, Hurwitz and Madow (1953a, b) and Deming (1950, 1956). Many of these early developments are well known, at least to the community of survey statisticians. Hartigan (1969) has drawn a parallel between early contributions to resampling and fractional factorial design (e.g., Fisher, 1966) through the gains in performance that can be achieved by balancing sampling experiments.

Naturally, this view should not exclude resampling methods that are more conventionally connected to the bootstrap, particularly cross-validation and the jackknife. See Stone (1974) for a general approach to the former, and see Quenouille (1949, 1956) and Tukey (1958) for early proposals of the latter. However, the linkages of the jackknife to the bootstrap have been authoritatively recounted elsewhere (e.g., Efron, 1979b, 1982; Efron and Tibshirani, 1993; Shao and Tu, 1995; Davison and Hinkley, 1997), as too have those of cross-validation (e.g., Efron and Tibshirani, 1993; Good, 1999).

Likewise, the relatively early development of permutation and Monte Carlo methods for testing (e.g., Fisher, 1936; Pitman, 1937; Welch, 1937; Wald and Wolfowitz, 1944; Barnard, 1963; Hope, 1968), which anticipated some aspects of resampling methodology, has been well documented. In principle the connections of permutation testing to contemporary "experimental statistics," based on Monte Carlo simulation, are strong. Some of them will be addressed in Section 3. However, large parts of the methodology had only limited application until the development of modern electronic computers. Consequentially, in important respects the rise to influence of permutation testing occurred contemporaneously with, rather than prior to, that of the bootstrap.

In this brief account I have selected contributions that are mentioned relatively infrequently, starting with research of the Indian statistical scientist P. C. Mahalanobis. His development of sampling methods for the Bengali jute crop anticipated today's moving-block methods, and his analysis of Bihar cereal-crop harvests introduced half-sampling for assessing statistical variation. I conclude by discussing the work of Mahalanobis' geographical and philosophical antithesis, the free-marketer and professional iconoclast Julian Simon, better known for his widely used (in the United States) volunteer scheme for dealing with airline overbooking, or his controversial views on resource renewability. His perspectives on statistics, like many of his other causes, were marked by both prescience and polemic, but his enthusiasm for statistics as an experimental science was surpassed only by his energy for articulating that view.

2. MAHALANOBIS' CONTRIBUTIONS TO RESAMPLING

2.1 Accessible Sources

Accounts of Mahalanobis' bootstrap-related work are given in a number of articles, including one in the *Philosophical Transactions of the Royal Society* (Mahalanobis, 1946a), three in *Sankhyā* (Mahalanobis, 1940; 1946b, c) and a fifth in the *Journal of the Royal Statistical Society* (Mahalanobis, 1946d). The *Philosophical Transactions* article is seldom cited today, but it is especially helpful in elucidating the evolution of Mahalanobis' ideas. It, and Mahalanobis (1940), relate solely to his work on the sampling and analysis of data on jute production in the state of Bengal.

The article Mahalanobis (1946b) comprises a report made to the Government of Bihar in 1944. The article Mahalanobis (1946d), read before the Royal Statistical Society, gives an overview of Mahalanobis' spatial sampling work and that of his colleagues. Mahalanobis (1946c) contains a discussion of early spatial sampling methodology, developed in the 1920s by the Indian civil servant J. A. Hubback and applied to rice crops in the states of Bihar and Orissa, and of the relationship of Hubback's ideas to Mahalanobis' own. Hubback, an Englishman whose report in 1927 has been hailed as the first published account of the use of sample surveys to assess crop yields, commenced his experiments in 1923. His work "influenced greatly" the crop-sampling experiments conducted by R. A. Fisher at Rothamsted (Fisher, 1945) and also inspired the research of F. Yates and others on sampling crop yield (Mahalanobis, 1946d). Hubback went on to become Governor of Orissa. His 1927 article was reprinted in

Sankhyā (Hubback, 1946). Its particular connection to bootstrap literature is its focus on *stochastic* methods for spatial sampling. In these terms, Hubback anticipated "random sampling from the sample" by "randomly sampling from a finite population."

The articles Mahalanobis (1946b, c) are often miscited, for example, by the title of one being used together with the page numbers of the other. (The present author is one of the guilty parties.) The publication of *Sankhyā* at that time was affected by exigencies of war; in particular, the single volume in which both articles appear encompasses both 1945 and 1946, which explains why either year is used in citations.

Mahalanobis authored many other articles and reports on his spatial sample surveys in Bengal and Bihar. However, most are difficult to obtain outside India.

2.2 Mahalanobis' Motivation

It is helpful to consider the background to the practical problems that motivated Mahalanobis' work, so as to appreciate its importance at that time, the massive scale on which it was conducted and the sense of urgency that characterized it.

Mahalanobis developed methods for sampling the jute crop in Bengal during a 5-year program which commenced in 1937. He subsequently modified his techniques for the sampling program in Bihar. The economic imperative for his work in Bengal can be judged by noting that at the time, jute exports amounted to a quarter of the total value of all exports from India, that Bengal produced 85% of India's jute and that almost 90% of all jute produced in India was exported. The Bengal sampling project, which progressed initially in an unrushed way, later proceeded under "high pressure." In 1941 it employed a staff of more than 500.

Cereal production in Bihar was likewise of major economic and strategic concern to India, especially in time of war. Mahalanobis was asked in late September 1943 to undertake "as early as possible" a survey of crop yields similar to that which he had supervised in Bengal. The invitation, from the Government of Bihar, focused initially on the state's rice harvest, but the objectives quickly expanded to encompass cereal crops of many types. By the time Mahalanobis was contacted the *rabi* (spring) season was already well advanced, so developing new sampling methodology was virtually precluded. Unsurprisingly, he relied largely on the techniques he had developed and extensively tested in Bengal. A variety of sampling experiments had been conducted there, but time constraints all but ruled out experimentation in Bihar.

2.3 Spatial Nature of the Project

Mahalanobis' data were not independent and identically distributed random variables or even random vectors, but crop yields which depended on the sizes of farm plots where crops were grown. It was within this relatively difficult context, which remains close to the frontiers on which contemporary bootstrap methods are advancing, that he grappled from the start. He considered sampling independently and at random, using tables of random numbers, but

> ... unfortunately, this method cannot be used as the size of individual plots in Bihar varies widely from a fraction of an acre to several hundreds of acres. Selection by serial number of plots would not give each acre of land the same chance of being included in the sample; and samples drawn in this way would not be truly random or representative (Mahalanobis, 1946b).

These considerations motivated his spatial sampling scheme. He used the term "grid" to describe a sampling unit, not the geometric structure within which such a unit was located; this should be borne in mind when reading his words later in this section.

Each district where crops were grown in Bihar was divided into nonoverlapping zones, which he took to be about 50 square miles in area. Their numbers and distributions depended on the district's size and shape. Determining sampling-unit size was a substantially more difficult task, to which Mahalanobis paid a great deal of attention; see Section 2.4.

He sampled using a method he referred to as "interpenetrating sub-samples," or "half samples." He split into equal batches the number of sampling units to be distributed within a given zone and distributed each batch independently of the other. Thus, the two lots of sampling units interpenetrated, or mingled with one another, in an entirely random fashion. Each provided a check on the other.

The manner in which Mahalanobis used this check shows that he interpreted the variation between his half-samples as providing direct information about variation within the population. This view is central to present-day interpretations of the bootstrap:

> When two such interpenetrating net-works of sample-units are used, it is possible to obtain two independent estimates (of crop acreage or of crop yield) for each region.

> The difference between these two independent estimates immediately supplies a good idea of the effective margin of error and thus indicates to what extent the survey had been carried out under statistical control (Mahalanobis, 1946c).

The errors to which Mahalanobis is referring here are both random and systematic.

While the half-samples might be viewed as samples drawn randomly, without replacement, from their union, they were not in fact obtained in that way. Indeed, the half-samples were a little more independent of one another than were their component data values, since unusual precautions were taken to ensure that observation errors were not shared between the half-samples; see Section 2.5. Nevertheless, the concept of half-sampling has clear connections, not only to the present-day bootstrap, but also to other early work on resampling methods; see Section 3.

Mahalanobis argued that half-samples supplied "two *independent* estimates of crop acreage for the same zone" (our italics), thus making it clear that he considered spatial correlation, and perhaps the occasional overlap of sampling units, to have negligible impact. His reference to "interpenetrating subsamples" should be taken to mean that the units from the two subsamples mingled together in the plane, not that the sampling units geometrically intersected one another.

This account refers to the sampling program in Bihar. The program in Bengal was significantly different, due in part to experimentation with sampling units of vastly differing sizes (ranging in area from a fraction of an acre to hundreds of acres). In Bengal, linked pairs of sampling units were used, the distance between each pair being fixed "but the orientation was settled at random" (Mahalanobis, 1946a). The quoted remark refers to the orientation of the line joining the centers of the pairs.

The main purpose of the distance constraint was to reduce any effects of spatial correlation, which otherwise could have caused problems, given the relatively large sampling units often used in Bengal. The existence of the constraint weakens the connection between the Bengal experiments and the conventional, contemporary bootstrap; the Bengali half-samples could not have been drawn equivalently by resampling randomly, without replacement, from a single sample. However, Mahalanobis' experiments in Bengal are related to a form of the block bootstrap, which we take up next.

2.4 Moving-Block Methods

Mahalanobis observed that if cost was no consideration and if correlation was ignored, large sampling units might be preferred on the basis of statistical performance: "The precision of individual sample units would increase as the area of each individual sample unit is increased" (Mahalanobis, 1946b). However, this law failed in the presence of correlation, where:

> ... the actual decrease was much smaller, so that the gain in precision by increasing the size of individual grids was appreciably less than one would ordinarily expect of the normal theory. This may be ascribed to the fact that the proportions of land sown with a particular crop ... are not statistically independent but are correlated (Mahalanobis, 1946a).

Empirical evidence for these properties came from sampling experiments undertaken in Bengal. (It should be noted, however, that Hubback had observed the effects of correlation in his sampling experiments in the 1920s.) The properties hint at a theory of optimality in which the balance between variance and some other component is altered as sampling-unit size is increased. In particular, they suggest the notion of optimal sampling-unit size.

The latter was specifically addressed by Mahalanobis (1946a). In Section 5 of part II of that article he embarked on a detailed and elaborate calculation of optimal sampling-unit size and density (i.e., the density of sampling units per unit of area). Now, in Mahalanobis' experiments there were, in principle, few effects from bias, since he was estimating the population mean. As a result, instead of variance and squared bias having an inverse connection, this relationship was exhibited by variance and cost:

> If we use grids of a large size, say 50-acre or 100-acre each, then the time and hence the expense involved in the physical examination of each grid would be comparatively large, and we would be able to have only a comparatively small number of grids .... The need of working within the limits of a fixed budget places a restriction on the choice of the size (that is, the area) of each individual grid and their total number of density per square mile (Mahalanobis, 1946b).

Mahalanobis was preoccupied by the need to keep financial costs within budget. Indeed, he had considered using the term "double sample" rather than "half-sample," but decided against it because doing so "would have probably raised misapprehensions about cost in the minds of administrators" (Mahalanobis, 1946d)! (See also the mention of double sampling in Section 3.1.) Thus, his optimal sampling-unit size was determined by relating variance and expenditure. His theory achieved this balance via a complex argument based on differential equations and Lagrange multipliers.

Mahalanobis (1946a) gave especially detailed theory that describes the impact of spatial correlation on variance. He noted that his arguments were

> ... a kind of generalization for two dimensions of the method of serial correlation in the case of time series on which a large volume of work is already in existence ... [but] many new problems arise in the case of the two-dimensional correlation function which have no analogue in the case of serial correlation (Mahalanobis, 1946a).

Bartlett (1946) drew particular attention to the connections between Mahalanobis' theory for spatial dependence and related work in time series.

Today, sampling-unit size for moving-block methods is generally treated as a smoothing parameter, and its optimal choice is a central issue. If Mahalanobis had turned his hand to estimating variance he would have found that increasing block size reduced bias and that squared bias played a role something like that which financial cost did in his work on the mean. Moving-block methods and the block bootstrap were explored in this setting by Hall (1985), Politis and Romano (1993, 1994) and Sherman (1996), who noted connections to statistical smoothing.

Mahalanobis' (1946a) discussion of the properties of block-based sampling anticipated modern moving-block methods for spatial data and, in view of the discussion quoted two paragraphs above, for time series. Contemporary accounts of the moving-blocks bootstrap include those of Hall (1985), Künsch (1989), Götze and Künsch (1996), Lahiri (1996) and Politis, Romano and Wolf (1999).

However, in neither Bengal nor Bihar did Mahalanobis use block bootstrap methods, as distinct from half-sampling, to actually assess crop yields. The modern concept of the block bootstrap was most likely not explicitly in his mind, because the expense of sampling would have prevented the comprehensive resampling program that would have been necessary. Instead, he saw the use of a limited number of random blocks serving a dual purpose: reducing expense, compared with drawing a full sample, and permitting a more accurate estimate of yield than was possible if crops were sampled in farm plots rather than in more methodically determined moving blocks.

Mahalanobis was concerned, too, about the effect that heterogeneity would have on optimal choice of sampling unit. Sampling density was also a major worry, not least because it had a direct bearing on cost. He treated edge or boundary effects (sometimes referring to them as "border effects"), noting that their impact on bias could be ignored for sampling units of appropriate size.

A multitude of considerations such as these led Mahalanobis to use 4-acre sampling units and a density of two sampling units per square mile in the Bihar experiments. (Four acres was a tenth the area of many of the sampling units he had employed in Bengal.) The number of units allocated to a district was typically in the thousands and the random manner (virtually in the continuum) in which sampling unit centers were chosen might have led to occasional overlap. However, since there are 640 acres in a square mile, overlap would have been rare. The issues of correlation, which Mahalanobis (1946a) discussed at length in connection with his Bengal experiments, were of less direct relevance to the problem of sampling in Bihar.

2.5 Collecting Spatial Data in Mahalanobis' Day

The practical challenges faced by Mahalanobis invite a comparison between his projects in Bengal and Bihar, some 60 years ago, and contemporary settings where similar methodology is employed. We still use versions of his techniques to estimate crop acreages and yields, and to approximate the variability of those estimates, although we seldom draw the connections to his work. Today we gather data automatically and remotely, on a grid with pixels (picture elements) that are perhaps only meters wide, using a satellite many miles above the Earth.

At the other extreme, the microscopic level, we also estimate sampling variability using related techniques. Here the data are instantly and relatively inexpensively gathered by an image analyzer coupled to an optical or particle-beam microscope, and pixel dimensions might be measured in nanometers.

Despite the expense of satellite imagery and microscopy, the financial constraints that govern such

sampling experiments are usually minor compared to those faced by Mahalanobis. His "recording device" was manual labor—investigators were sent out into the field with lists of sampling units allotted to them, each person recording (for example) the proportion of land under different crops within the confines of his sampling units. As we have seen, the expense of such a program interacted markedly with statistical issues and influenced the notion of statistical optimality in a way one rarely encounters today.

Mahalanobis' sampling schemes had a number of inbuilt checks and balances, including the gathering of each half-sample by different groups of investigators. Within a zone, the investigators collecting data for the respective half-samples worked "at different times ... so that they never meet." From time to time these workers were asked to literally toss coins, out among the rice paddies and in the barley fields, as they made their sampling decisions, and Mahalanobis checked up on the randomness of their tosses:

> Each investigator ... is required to keep a record of successive throws of heads and tails .... If, as is sometimes found to be the case, the series is significantly non-random it is reasonable to think that the field work was not done properly (Mahalanobis, 1946c).

3. HALF-SAMPLING, SUBSAMPLING AND RESAMPLING

3.1 Subsampling in Sample Surveys

It is of course no accident that some of the earliest contributions to resampling methods were made in the context of sample surveys. The notion of resampling from a sample, particularly in a "without-replacement" sense, is a minor extension of sampling from a finite population. Madow and Madow (1944), Jones (1956) and Shiue (1960) considered, in the context of survey sampling, different techniques for drawing m subsamples of size n from a population of size N by slicing the population into equal-sized, nonoverlapping parts using a mixture of systematic and stochastic methods.

By the 1940s, random subsamples were used for calibration "in forestry work in the USA" (Mahalanobis, 1946d). The more general method of two-phase sampling was [and often still is; see, e.g., Cochran (1977, page 327)] called "double sampling." It is motivated in a very different way from Mahalanobis' half-sampling experiments and was employed for surveys of crop yields in India in 1941. It involves combining direct but relatively expensive measurements made on a randomly chosen subset of the data, with indirect, inexpensive measurements made on the whole data set.

Any of these techniques can be viewed as a step in the development of without-replacement ways to sample from a sample. Each is related to the stochastic sampling experiments performed by Mahalanobis in the 1930s and 1940s. However, Mahalanobis' setting was arguably more complex, in that it involved spatial relationships and, in particular, potential correlation.

Also in the context of sample surveys, "half-sampling" methods were used by the U.S. Bureau of the Census from at least the late 1950s (Kish, 1957; ANON, 1963). This pseudo-replication scheme was designed to produce, for stratified data, an effective estimator of the variance of the grand mean (a weighted average over strata) of the data. The aim was to improve on the conventional variance estimator, computed as a weighted linear combination of within-stratum sample variances. The latter can be highly variable (although unbiased) when there are only small numbers of observations in each stratum.

The method used by the Bureau of the Census consisted of resampling one datum randomly from each stratum, computing the weighted average of these values, forming the squared difference of the average from the grand mean and averaging the squared differences over different drawings of the resampled data. Although applicable, with suitable weights, in the case of general sample sizes in each stratum, the technique was typically used in the case where each stratum had only the bare minimum of data needed to estimate the stratum variance; that is, just two observations per stratum. In this case, drawing one observations from each stratum used exactly half the full data set. Hence the name "half-sample."

The relationship between these half-samples and those used by Mahalanobis is largely limited to the fact that both refer to using half the full sample and each technique is motivated by an application of sample survey methodology. In other respects the half-sampling methods are rather different.

McCarthy (1969) reported that theory for half-sampling in sample surveys was developed by Gurney (1963) and McCarthy (1966). There it was shown that the variance of the estimator is proportional to

$$(2/B)(1 - L^{-1}) + CL^{-1}, \tag{1}$$

where L denotes the number of strata, C is a constant that depends only on the population and B is what

would be referred to today as the number of bootstrap replications.

To appreciate the origins of the terms in (1), note that each of the squared differences that is averaged to compute the half-sample variance estimator can be written as a sum of diagonal terms plus a sum of off-diagonal terms. The off-diagonal terms are the contributors to the first term in (1). Of course, the latter quantity vanishes if the number of replications is infinite, but in other settings it may make a significant contribution. In the 1950s and 1960s, when these methods were developed, computational limitations restricted the feasible number of simulations, and so there was motivation to reduce the first term in (1) by means other than simply increasing B. This was the context of McCarthy's (1969) work; he showed how to balance the half-sample replications so that a significant portion of the contributions from off-diagonal terms cancelled, leading to a reduction in estimator variance.

Efron (1982), in his monograph titled *The Jackknife, the Bootstrap and Other Resampling Plans*, recognized the important role that these precursors played in the development of bootstrap methods. He included three chapters on half-sampling, random subsampling and typical-value methods.

3.2 Resampling Methods for Simple Random Samples

Unlike McCarthy and Mahalanobis, Hartigan (1969) developed the subsampling idea in the context of simple random samples and in a rather general, abstract setting. This enabled its theoretical properties to be explored quite extensively. He introduced the notion of "subsample values" of a statistic $\hat\theta$; these are the values of $\hat\theta$ computed for subsets of a given sample.

Hartigan's subsets (i.e., subsamples) were simply the possible subsets of data (excluding the empty set) that could be drawn from the full data set, and so number $2^n - 1$ in the case of a sample of size n. Therefore, in effect, he drew his subsets from the sample without replacement and his resample size was usually less than the sample size. Excepting these changes, his subsample values would today be called bootstrap values of $\hat\theta$ and denoted by $\hat\theta^*$ in now conventional notation. He was later (Hartigan, 1971) to introduce other kinds of replacement subsampling and to anticipate Rubin's (1981) Bayesian bootstrap.

Today it is widely appreciated that the variability of $\hat\theta^*$, conditional on the data, can be used to very good effect to approximate the variation of $\hat\theta$ in an unconditional sense. Hartigan (1969) explored the use of subsample values for this purpose and, in particular, for constructing confidence intervals. Some of Hartigan's subsample values were, in his terminology, "typical values," provided the subsamples were balanced in a manner that could be defined in terms of the intervals between adjacent, ordered subsample values. He used group-theoretic methods to make this concept more explicit. Restricting attention to typical values and to Hartigan's (1969) approach to subsampling, his Theorem 4 describes a type of percentile-method bootstrap confidence interval for a mean.

Maritz (1979), too, used permutation arguments connected to the bootstrap to conduct inference about location parameters. More closely linked to contemporary resampling methods were the methods of Maritz and Jarrett (1978) and Breth, Maritz and Williams (1978), which were founded directly on the empirical distribution function.

Forsythe and Hartigan (1970) developed Hartigan's (1969) confidence interval argument further. They showed that substantial reductions can be achieved in the amount of Monte Carlo simulation that needs to be done by confining attention to "balanced" resamples defined in terms of Hartigan's (1969) definition of typical values. This notion of balance is technically different from that introduced by McCarthy (1969) and also from those suggested more recently as the balanced bootstrap (or Latin hypercube sampling) by Davison, Hinkley and Schechtman (1986) or antithetic resampling by Hall (1989). However, all these methods are connected, in the sense that in each, an appropriate degree of symmetry is introduced to reduce the error of resampling approximations and thereby produce estimators with greater accuracy for a given amount of computational labor.

Hartigan (1975) gave necessary and sufficient conditions for the asymptotic joint normality of a statistic and its subsample (and jackknife) values. We know today, through work of Mammen (1992) and others, that asymptotic normality of a statistic's distribution is particularly close to being a necessary and sufficient condition for the bootstrap-based estimator of the distribution of the statistic to give correct answers. (These remarks apply to models that are locally asymptotically normal in Le Cam's sense, and do not necessarily apply in other cases.)

3.3 Statistics as an Experimental Science

Tukey was arguably the most influential supporter of experimental statistics in the 30-year period after

1950, through his work on the jackknife, experimental data analysis and so forth. Simon (1969, 1993), an iconoclastic polymath by inclination but a social scientist by vocation, also advocated statistical methodology based on computer experimentation. In 1969, his promotion of resampling methods for testing statistical hypotheses must have seemed only moderately practical. Nevertheless, his prediction that this approach "holds great promise for the future" (Simon, 1969) cannot be faulted. Twenty-four years later he was to define "resampling" to mean "the use of the given data or a data generating mechanism (such as a die) to produce new samples, the results of which can then be examined" (Simon, 1993). He commented too that "the term computer-intensive methods is ... used to refer to techniques such as these." Simon's discussion of statistics in terms of gambling experiments is typical; those who challenged his ideas were frequently answered with a wager.

Prior to the late 1960s, permutation tests would have been regarded primarily as a motivator of more computable, parametric approaches, in particular, those based on the analysis of variance, even though permutation methods had been discussed for some 30 years. Attempts were made to reduce computational labor (see, e.g., Chung and Fraser, 1958), but widespread use of permutation techniques had to await the availability of inexpensive, interactive, electronic computers. Nevertheless, Simon's (1969) methodology was largely restricted to permutation methods, the application of which he described in terms of coin-tossing, dice-throwing or card-shuffling experiments.

Simon was an avowed free-marketer. "It was not a good idea to ridicule capitalism, or free markets, or human liberty, in Simon's presence," wrote Wattenberg (1998). One of Simon's (1969) numerical examples in *Basic Research Methods*, on the cost in 1961 of a glass of Seagram's Seven-Crown American Whiskey, was archetypically his. It compared prices in 27 U.S. states that had privately owned liquor stores with those in 16 states that had state monopolies. The pooled sample size was therefore 43, and to describe how to use permutation methods to test the hypothesis that the mean prices were identical, he invited the reader to "write each of the forty-three prices on playing cards, shuffle the cards, and deal out sixteen. Do this repeatedly."

Simon introduced ranking methods the same way. Discussing a rank-based test applied to a sample of size 10, he wrote: "Take ten cards, one of each denomination from ace to ten, shuffle, and deal ... ." No reference was made to statistical tables. He alluded to asymptotic approximations thus: "One can avoid the tedious work of Monte Carlo experiments by getting the help of a statistician who will use mathematical methods" (Simon, 1969). However, such an approach was certainly not recommended. Indeed, a later monograph which expounded in detail his Monte Carlo-based proposals (Simon, 1993) was promoted on the basis that it taught "how not to let tricky statistics get the best of your argument." Simon's preferred statistical analysis was based unabashedly on simulation.

Resampling was first written in 1974, went through a substantial revision in 1989 and was finally published in a "preliminary edition" in 1993, together with a computer disc which enabled the methods to be implemented on a PC. A later, but again "preliminary," edition had a different subtitle and a co-author (Peter C. Bruce). Subsequent editions of *Resampling* were published by Resampling Stats, Inc. The 1993 preliminary edition appeared in Duxbury Press covers.

Simon felt a degree of antagonism toward the statistics profession, which, he argued, had only grudgingly accepted, and then borrowed without appropriate attribution, "his" resampling ideas developed in Simon (1969, 1993). In a subsequent version of *Resampling*, he misinterpreted change as explicit conflict and wrote:

> The simple fact is that resampling devalues the knowledge of conventional mathematical statisticians, and especially the less competent ones. By making it possible for each user to develop her/his own method to handle each particular problem, the priesthood with its secret formulaic methods is rendered unnecessary. No one ... stands still for being rendered unnecessary. Instead, they employ every possible device fair and foul to repel the threat to their economic well-being and their self-esteem.

Nevertheless, later in that version he expressed pleasure that "resampling techniques have caught on like wildfire among statisticians ... [who] are busily exploring [their] properties ... and applying [them] regularly to problems that are difficult with conventional analysis."

Early versions of *Resampling* are largely updated accounts of the "statistics" portions (Chapters 22–26) of *Basic Research Methods*. Between 1974 and 1993, *Resampling* acquired a strong bootstrap component. It cited Efron's contributions (Efron, 1982; Diaconis and Efron, 1983) and used with-replacement resampling

as an alternative to permutation in the Seagram's Whiskey example. Commented Simon: "Recently I have concluded that a bootstrap-type test has better theoretical justification than a permutation test in this case, although the two reach almost identical results with a sample this large" (Simon, 1993).

The clarity with which Simon, who died in early 1998, anticipated in the 1960s the massive changes that computing would bring to statistics undoubtedly placed him apart from many of his peers. However, his main contribution to statistics was surely as an advocate and popularizer of Monte Carlo experimentation and of resampling methods, not as a developer of specific new techniques.

His personality and the nature of his contributions provide an intriguing counterpoise to those of Mahalanobis, with whose work we began this brief account of the prehistory of the bootstrap. Mahalanobis, a scientist trained in physics and biology, and a scholar interested in Indian philosophy and Bengali literature, was to take some of the first steps toward a statistical methodology based on experimentation. Simon, famous for his wager with Paul Ehrlich that the global supply of natural resources would outstrip demand and for his argument that intellect is the only resource that really matters, was passionate and articulate as an advocate of a highly influential technology whose beginnings can be traced back at least to Mahalanobis.

ACKNOWLEDGMENTS

I am grateful for helpful comments from John Hartigan, Prakash Patil, Steve Stigler, Sue Wilson and Jeff Wood.

REFERENCES

ANON (1963). The current population survey. A report on methodology. Technical Paper 7, U.S. Bureau of the Census, U.S. Government Printing Office, Washington.

BARNARD, G. A. (1963). Discussion of "Spectral analysis of point processes," by M. S. Bartlett. *J. Roy. Statist. Soc. Ser. B* **25** 294.

BARTLETT, M. S. (1946). Discussion of "Recent experiments in statistical sampling in the Indian Statistical Institute," by P. C. Mahalanobis. *J. Roy. Statist. Soc.* **109** 373.

BRETH, M., MARITZ, J. S. and WILLIAMS, J. S. (1978). On distribution-free lower confidence limits for the mean of a nonnegative random variable. *Biometrika* **65** 529–534.

CHUNG, J. H. and FRASER, D. A. S. (1958). Randomization tests for a multivariate two-sample problem. *J. Amer. Statist. Assoc.* **53** 729–735.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge Univ. Press.

DAVISON, A. C., HINKLEY, D. V. and SCHECHTMAN, E. (1986). Efficient bootstrap simulation. *Biometrika* **73** 555–566.

DEMING, W. E. (1950). *Some Theory of Sampling*. Wiley, New York.

DEMING, W. E. (1956). On simplifications of sampling design through replication with equal probabilities and without stages. *J. Amer. Statist. Assoc.* **51** 24–53.

DIACONIS, P. and EFRON, B. (1983). Computer-intensive methods in statistics. *Scientific American* **249** 116–130.

EFRON, B. (1979a). Computers and the theory of statistics: Thinking the unthinkable. *SIAM Rev.* **21** 460–480.

EFRON, B. (1979b). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

FISHER, R. A. (1936). "The coefficient of racial likeness" and the future of craniometry. *J. Royal Anthropological Institute of Great Britain and Ireland* **66** 57–63.

FISHER, R. A. (1945). Memorandum to the Imperial Council of Agricultural Research in India, 2 March 1945. [Quoted by P. C. Mahalanobis, *Sankhyā* **7** (1946) 269.]

FISHER, R. A. (1966). *The Design of Experiments*, 8th ed. Oliver and Boyd, Edinburgh.

FORSYTHE, A. and HARTIGAN, J. A. (1970). Efficiency of confidence intervals generated by repeated subsample calculations. *Biometrika* **57** 629–639.

GOOD, P. I. (1999). *Resampling Methods—A Practical Guide to Data Analysis*. Birkhäuser, Boston.

GÖTZE, F. and KÜNSCH, H. R. (1996). Second-order correctness of the blockwise bootstrap for stationary observations. *Ann. Statist.* **24** 1914–1933.

GURNEY, M. (1963). The variance of the replication method for estimating variances for the CPS sample design. Memorandum, U.S. Bureau of the Census. Unpublished.

HALL, P. (1985). Resampling a coverage pattern. *Stochastic Process. Appl.* **20** 231–246.

HALL, P. (1989). Antithetic resampling for the bootstrap. *Biometrika* **76** 713–724.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953a). *Sample Survey Methods and Theory* **1**. *Methods and Applications*. Wiley, New York.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953b). *Sample Survey Methods and Theory* **2**. *Theory*. Wiley, New York.

HARTIGAN, J. A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64** 1303–1317.

HARTIGAN, J. A. (1971). Error analysis by replaced samples. *J. Roy. Statist. Soc. Ser. B* **33** 98–110.

HARTIGAN, J. A. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *Ann. Statist.* **3** 573–580.

HOPE, A. C. A. (1968). A simplified Monte Carlo significance test procedure. *J. Roy. Statist. Soc. Ser. B* **30** 582–598.

HUBBACK, J. A. (1946). Sampling for rice yield in Bihar and Orissa. *Sankhyā* **7** 281–294. (First published in 1927 as Bulletin 166, Imperial Agricultural Research Institute, Pusa, India.)

JONES, H. L. (1956). Investigating the properties of a sample mean by employing random subsample means. *J. Amer. Statist. Assoc.* **51** 54–83.

KISH, L. (1957). Confidence intervals for clustered samples. *American Sociological Review* **22** 154–165.

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241.

LAHIRI, S. N. (1996). On Edgeworth expansion and moving block bootstrap for Studentized M-estimators in multiple linear regression models. *J. Multivariate Anal.* **56** 42–59.

MADOW, W. G. and MADOW, L. (1944). On the theory of systematic sampling. I. *Ann. Math. Statist.* **15** 1–24.

MAHALANOBIS, P. C. (1940). A sample survey of the acreage under jute in Bengal. *Sankhyā* **4** 511–530.

MAHALANOBIS, P. C. (1946a). On large-scale sample surveys. *Philos. Trans. Roy. Soc. London Ser. B* **231** 329–451.

MAHALANOBIS, P. C. (1946b). Report on the Bihar crop survey: Rabi season 1943–44. *Sankhyā* **7** 29–106.

MAHALANOBIS, P. C. (1946c). Sample surveys of crop yields in India. *Sankhyā* **7** 269–280.

MAHALANOBIS, P. C. (1946d). Recent experiments in statistical sampling in the Indian Statistical Institute (with discussion). *J. Roy. Statist. Soc.* **109** 325–378. [Reprinted, including discussion, in *Sankhyā* **20** (1958) 329–397.]

MAMMEN, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations. Lecture Notes in Statist.* **77**. Springer, New York.

MARITZ, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163–166.

MARITZ, J. S. and JARRETT, R. G. (1978). A note on estimating the variance of the sample median. *J. Amer. Statist. Assoc.* **73** 194–196.

MCCARTHY, P. J. (1966). Replication: An approach to the analysis of data from complex surveys. Vital Health Statistics. Public Health Service Publication 1000, Series 2, No. 14, National Center for Health Statistics, Public Health Service, U.S. Government Printing Office, Washington.

MCCARTHY, P. J. (1969). Pseudo-replication: Half samples. *Review of the International Statistical Institute* **37** 239–264.

PITMAN, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Suppl. J. Roy. Statist. Soc.* **4** 119–130.

POLITIS, D. N. and ROMANO, J. P. (1993). Nonparametric resampling for homogeneous strong mixing random fields. *J. Multivariate Anal.* **47** 301–328.

POLITIS, D. N. and ROMANO, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22** 2031–2050.

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer, New York.

QUENOUILLE, M. H. (1949). Approximate tests of correlation in time-series. *Proc. Cambridge Philos. Soc.* **45** 483–484.

QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360.

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134.

SHAO, J. and TU, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

SHERMAN, M. (1996). Variance estimation for statistics computed from spatial lattice data. *J. Roy. Statist. Soc. Ser. B* **58** 509–523.

SHIUE, C.-J. (1960). Systematic sampling with multiple random starts. *Forest Science* **6** 42–50.

SIMON, J. (1969). *Basic Research Methods in Social Science. The Art of Empirical Investigation*. Random House, New York.

SIMON, J. (1993). *Resampling*: *The New Statistics*. Duxbury, Belmont, CA.

STIGLER, S. M. (1991). Stochastic simulation in the nineteenth century. *Statist. Sci.* **6** 89–97.

STIGLER, S. M. (1999). *Statistics On the Table: The History of Statistical Concepts and Methods.* Harvard Univ. Press, Cambridge, MA.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

TUKEY, J. (1958). Bias and confidence in not-quite large samples (abstract). *Ann. Math. Statist.* **29** 614.

WALD, A. and WOLFOWITZ, J. (1944). Statistical tests based on permutations of the observations. *Ann. Math. Statist.* **15** 358–372.

WATTENBERG, B. (1998). Malthus, watch out (Julian Simon obituary). *Wall Street Journal*, February 11.

WELCH, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika* **29** 21–52.