Ancillarity Principle and a Statistical Paradox

V.P. GODAMBE*

Among the many reasons underlying the practice of randomization some of the main ones can be described as *averaging* out or *elimination* of the effects of nuisance parameters. It is already well known (Godambe 1966) that averaging over all the possible results of the adopted randomization is directly in conflict with the *likelihood principle*. The process of elimination of nuisance parameters has possibly deeper intuitive appeal. But this process contradicts a very basic principle of statistical inference subsequently defined as the *ancillarity principle*. This is demonstrated in relation to a practice of randomization called *balanced sampling*. We would use a formalism very similar to that of Birnbaum (1962).

KEY WORDS: Randomization; Balanced sampling; Ancillarity.

1. ANCILLARITY PRINCIPLE

A statistical experiment or a model M is defined as a triplet $(\chi, \Omega, \mathbf{P})$ where $\chi = \{x\}$ is an abstract sample space, $\Omega = \{\theta\}$ is an abstract parameter space and $\mathbf{P} = \{P_\theta: \theta \in \Omega\}$ is a class of distributions on χ indexed by the parameter θ. For simplicity we assume χ and Ω to be finite. The inference one can make on the basis of an observation x (in χ) given the experiment M can be denoted by $\text{Inf}(\cdot \mid x, M)$, leaving, however, the function uncharacterized as in Birnbaum (1962).

The ancillarity principle. If P_θ is the same for all $\theta \in \Omega$, then $\text{Inf}(\cdot \mid x, M)$ is the same for all x in χ. In other words no inference about θ is possible on the basis of an observation x, under the experiment M.

2. A PARADOX

Using the notion in Section 1 here we have $\theta = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_i, \ldots, \theta_N)$ and

$$\Omega = \left\{\boldsymbol{\theta}: \theta_i = 1 \text{ or } -1, i = 1, \ldots, N \text{ and } \sum_1^N \theta_i = 0\right\}. \tag{2.1}$$

Further set $\mathcal{P} = \{1, \ldots, N\}$, let n be a positive integer less than N, and let S denote the set of all subsets of P with n elements. Next we have,

$$P_{\boldsymbol{\theta}}(s) = 1/{}^N C_n, \; s \in S, \tag{2.2}$$

and

$$\mathbf{P} = \{P_{\boldsymbol{\theta}}: \theta \in \Omega\} \tag{2.3}$$

Thus (2.1), (2.2), and (2.3) define a statistical experiment

$$M \equiv (\chi, \Omega, \mathbf{P}), \tag{2.4}$$

where $\chi = S$. Now the *ancillarity principle* of Section 1 in relation to the experiment M in (2.4) implies that for any two $s', s'' \in S$,

$$\text{Inf}(\cdot \mid s', M) \equiv \text{Inf}(\cdot \mid s'', M). \tag{2.5}$$

This, however, is contradicted by the following mode of inference.

Using (2.1) and (2.2) let

$$t(s, \boldsymbol{\theta}) = \Big| \sum_{i \in s} \theta_i / n \Big|, \; s \in S, \; \boldsymbol{\theta} \in \Omega.$$

If s' is observed from the distribution (2.3) consider all $\boldsymbol{\theta}'$ (in Ω) satisfying $t(s', \boldsymbol{\theta}') > k$ for a suitably large k, as *implausible* values. (2.6)

Note. In the preceding mode of inference the observation (or data) consists of s' only. In particular (θ_i: $i \in s'$) is not part of the data. The distribution of s' is given by (2.3). Further, on the basis of two observations s' and s'', the set of values $\boldsymbol{\theta}$ considered implausible are *different* for

$$\{\boldsymbol{\theta}: t(s', \boldsymbol{\theta}) > k\} \not\equiv \{\boldsymbol{\theta}: t(s'', \boldsymbol{\theta}) > k\}.$$

Hence the mode of inference (2.6) contradicts the *ancillarity principle*. To make the paradox clearer we emphasize that (assuming suitable n, N, k) for every s' there are values of $\boldsymbol{\theta}$ in Ω for which $t(s', \boldsymbol{\theta}) \not> k$. These values according to (2.6) however are not considered implausible, when s' is observed.

The intuitive appeal of (2.6) (for sufficiently large n and N), follows from the fact that, if for every $s \in S$ in (2.2)

$$\bar{\theta}_s = \sum_{i \in s} \theta_i / n, \tag{2.7}$$

then for every $\boldsymbol{\theta} \in \Omega$ in (2.1), $\bar{\theta}_s$ has a fixed distribution, having the expectation and variance given by

$$E_{\boldsymbol{\theta}}(\bar{\theta}_s) = 0 \text{ and } v_{\boldsymbol{\theta}}(\bar{\theta}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N}{N-1}. \tag{2.8}$$

Obviously $v_{\boldsymbol{\theta}}(\bar{\theta}_s)$ would be negligibly small for sufficiently

* V.P. Godambe is Professor, Department of Statistics, Waterloo, Ontario, Canada N2L 3G1. Thanks are due to Mary Thompson for valuable comments on an earlier draft of this article.

December 1982, Volume 77, Number 380
Theory and Methods Section

large n and N. For instance for $n = 10{,}000$, $v_{\boldsymbol{\theta}}(\bar{\theta}_s) < .0001$, $\boldsymbol{\theta} \in \Omega$.

A more precise version of the mode of inference (2.6) is as follows. Let $N = 4$ and $n = 2$. Suppose further that Ω consists of only two points $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$, where $\boldsymbol{\theta}' = (1, -1, -1, 1)$ and $\boldsymbol{\theta}'' = (-1, 1, -1, 1)$. Now $t(s, \boldsymbol{\theta})$ in (2.6) can take only two values 0 and 1. Further $P_{\boldsymbol{\theta}}(t(s, \boldsymbol{\theta}) = 0) = 4/6$ and $P_{\boldsymbol{\theta}}(t(s, \boldsymbol{\theta}) = 1) = 2/6$ for $\boldsymbol{\theta} = \boldsymbol{\theta}', \boldsymbol{\theta}''$. Now if $s' = (2, 4)$, $t(s', \boldsymbol{\theta}') = 0$ and $t(s', \boldsymbol{\theta}'') = 1$. Hence on the basis of the observation s' we prefer $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}''$. On the other hand, if $s'' = (2, 3)$, $t(s'', \boldsymbol{\theta}') = 1$ and $t(s'', \boldsymbol{\theta}'') = 0$ hence on s'' we prefer $\boldsymbol{\theta}''$ to $\boldsymbol{\theta}'$.

From the example just discussed, it should be clear that the mode of inference in (2.6) is free of any assumption concerning prior probability distribution on Ω. In fact, as is clear from the example, even complete specification of Ω is not necessary. Further, the statistical inference as in (2.6) is not restricted to simple random sampling in (2.3) only. For a more complicated sampling design given in Section 3, inference about $\boldsymbol{\theta}$ on the basis of s is obtained by replacing in (2.6), $| \sum_{i \in s} \theta_i / n |$ by $| \sum_{i \in s} \theta_i / a_i |$.

When the author discussed the preceding situation with H. Robbins, the latter suggested a more descriptive version. With the author's elaboration it is as follows: Let $2N$ exactly identical slips of paper be spread on a table. On the hidden face of each slip is written a number, $+1$ or -1. It is known that some N slips bear the number $+1$ and the remaining N slips bear the number -1. The *unknown* state of nature or the *unknown parameter* ($\boldsymbol{\theta}$) here is determined precisely by naming the slips that bear $+1$ (and -1). The ${}^{2N}C_N$ possible values of the parameter $\boldsymbol{\theta}$ are denoted by Ω; $\Omega \equiv \{\boldsymbol{\theta}\}$. Now out of the $2N$ slips a random sample s of N slips is drawn without replacement. On the basis of the sample s so drawn, the following inference about the unknown parameter $\boldsymbol{\theta}$ is immediately suggested by the frequency definition of probability. *The values of $\boldsymbol{\theta}$ in Ω which assign for the N slips that constitute the sample s, proportion of $+1$'s greater than $\frac{1}{2} + \epsilon_N$ or smaller than $\frac{1}{2} - \epsilon_N$ (for a suitably chosen ϵ_N, if $N = 1{,}000{,}000$, ϵ_N may be .001), are implausible.* Yet the distribution of s is independent of $\boldsymbol{\theta}$, that is, is the *same* for all $\boldsymbol{\theta} \in \Omega$. That is, the frequency performance of the above procedure of inference, in repeated sampling, would be *identical* on the true value of the parameter, $\boldsymbol{\theta}_0$, say, and any alternative value, say, $\boldsymbol{\theta}_1$!

3. THE PARADOX WITH UNEQUAL PROBABILITY SAMPLING

Suppose an agricultural field is divided into N plots numbered i; as before in (2.2) $\mathcal{P} = \{i\}$, $i = 1, \ldots, N$, and $S = \{s\}$ is the set of all subsets s containing exactly n plots. The yield and area of the plot i are y_i and a_i, respectively, $i = 1, \ldots, N$. The vector $\mathbf{a} = (a_1, \ldots, a_i, \ldots, a_N)$ is known but the vector $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_N)$ is unknown. We write $\sum_1^N a_i = A$, $\sum_1^N y_i = Y$, and $Y = A\phi$. To estimate the unknown parameter ϕ we draw from S a sample s using a sampling design p (p: $S \to [0, 1]$, $\sum_S p(s) = 1$) and observe the yields y_i for all plots i in s. Let $y_i = a_i\phi + \theta_i$, $i = 1, \ldots, N$. Then since $Y = A\phi$, $\sum_1^N \theta_i = 0$. Hence we have

$$\frac{y_i}{a_i} = \phi + \frac{\theta_i}{a_i}, \; i = 1, \ldots, N \tag{3.1}$$

where the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_i, \ldots, \theta_N)$ is unknown except that $\boldsymbol{\theta} \in \Omega$ where

$$\Omega = \left\{ \boldsymbol{\theta}\colon \sum_1^N \theta_i = 0 \right\}. \tag{3.2}$$

(The Ω in (3.2) should be distinguished from the one in (2.1).) Now for any specified (given) $\boldsymbol{\theta}$ and the data (s, y_i: $i \in s$), ϕ is given by

$$\phi(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in s} \frac{y_i}{a_i} - \frac{1}{n} \sum_{i \in s} \frac{\theta_i}{a_i}. \tag{3.3}$$

Since, however, the vector $\boldsymbol{\theta}$ is unknown, we investigate whether there exists a sampling design to select s, so that in some sense $\phi(\boldsymbol{\theta})$ in (3.3) would not be much dependent on the nuisance parameter $\boldsymbol{\theta}$. (The estimate $\sum_{i \in s} y_i / na_i$ is also optimum for ϕ under some additional assumptions not related to the present discussion.) That is, we search for a design that provides (in the extended sense of the term) *balanced sampling*. Consider a sampling design p_0 obtained as follows: The set $\mathcal{P} = \{i, i = 1, \ldots, N\}$ is divided in n strata $\mathcal{P}_j (\mathcal{P} = U_1{}^n \mathcal{P}_j)$ such that $\sum_{i \in \mathcal{P}_j} a_i = A/n$, $j = 1, \ldots, n$. Then from each stratum j one individual is drawn with selection probabilities for different individuals i, na_i/A, $i \in \mathcal{P}_j$. Note in the present case if $s \ni i$ denotes all samples in S that contain the individual i then

$$\sum_{s \ni i} p_0(s) = na_i/A, \; i = 1, \ldots, N. \tag{3.4}$$

Now if E denotes the expectation and V the variance with respect to the above sampling design p_0, from (3.1) and (3.2) we have for every ϕ and $\boldsymbol{\theta} \in \Omega$,

$$E\Big(\sum_{i \in s} \theta_i / na_i\Big) = \sum_1^N \theta_i / A = 0,$$

$$E\Big(\sum_{i \in s} y_i / na_i\Big) = \sum_1^N y_i / A = \phi; \tag{3.5}$$

and

$$V\Big(\sum_{i \in s} y_i / na_i\Big) = V\Big(\sum_{i \in s} \theta_i / na_i\Big) = (1/nA) \sum_1^N \theta_i^2 / a_i - (1/A^2) \sum_1^n \Big(\sum_{\mathcal{P}_j} \theta_i\Big)^2. \tag{3.6}$$

Further, let a subset of Ω in (3.2) be given by

$$\Omega' = \left\{ \boldsymbol{\theta} : |\theta_i/a_i| \le \beta, \quad i = 1, \ldots, N; \sum_{1}^{N} \theta_i = 0 \right\}, \quad (3.7)$$

for some specified number β. From (3.6) for all $\boldsymbol{\theta} \in \Omega'$ in (3.7) we have

$$V(\sum_{i \in s} y_i/na_i) = V(\sum_{i \in s} \theta_i/na_i) \le (\beta^2/n). \quad (3.8)$$

Now we assume that the vector $\mathbf{a}$ is such that the population $\mathcal{P}$ can be divided into a sufficiently large number n of strata satisfying $\sum_{i \in \mathcal{P}_j} a_i = A/n, j = 1, \ldots, n$ and $\beta^2/n \le \epsilon^2$, ϵ being a given small number. Then from (3.8) we have

$$V(\sum_{i \in s} y_i/na_i) = V(\sum_{i \in s} \theta_i/na_i) \le \epsilon^2. \quad (3.9)$$

In the sense of (3.5) and (3.9) the sampling design p_0 indeed reduces the effect of the nuisance parameter $\boldsymbol{\theta}$ on $\phi(\boldsymbol{\theta})$ in (3.3) provided $\boldsymbol{\theta} \in \Omega'$. Thus sampling design p_0 provides balanced sampling. To obtain the confidence intervals for ϕ we note from (3.1) that

$$[|\sum_{i \in s} y_i/na_i - \phi| \le 3\epsilon] \Leftrightarrow [|\sum_{i \in s} \theta_i/na_i| \le 3\epsilon]; \quad (3.10)$$

hence if $P(\cdot \mid \boldsymbol{\theta}, \phi)$ denotes the probability of $(\cdot)$ for given $\boldsymbol{\theta}$ and ϕ we have

$$P(|\sum_{i \in s} y_i/na_i - \phi| \le 3\epsilon \mid \boldsymbol{\theta}, \phi) = P(|\sum_{i \in s} \theta_i/na_i| \le 3\epsilon \mid \boldsymbol{\theta}, \phi). \quad (3.11)$$

If $\boldsymbol{\theta} \in \Omega'$ in (3.7), with (3.5) and (3.9) the left side of (3.11) provides the usual inference about ϕ. But this inference because of (3.10) is logically equivalent to the inference about $\boldsymbol{\theta}$ obtained from the right side of (3.11). This latter inference about $\boldsymbol{\theta}$, however, contradicts the ancillarity principle for the reasons given in Section 2. It therefore follows that the inference about ϕ based on the left side of (3.11) is also paradoxical in relation to the ancillarity principle.

A brief statement of the paradox appeared in Godambe (1979), which in turn was commented on by Dawid (1979) and Good (1980).

[Received June 1981. Revised June 1982.]

REFERENCES

BIRNBAUM, A. (1962), "On the Foundation of Statistical Inference," *Journal of the American Statistical Association,* 57, 269–306.

DAWID, A.P. (1979), Reply to discussion, *Journal of the Royal Statistical Society,* Ser. B, 41, 29.

GODAMBE, V.P. (1966), "A New Approach to Sampling from Finite Populations," *Journal of the Royal Statistical Society,* Ser. B, 28, 310–328.

GODAMBE, V.P. (1979), Contribution to the discussion of a paper by A.P. Dawid, *Journal of the Royal Statistical Society,* Ser. B, 41, 24.

GOOD, I.J. (1980), "On Godambe's Paradox," *Journal of Statistical Computation and Simulation,* 12, 70–72.