J. R. Statist. Soc. B (1990)
52, *No.* 1, *pp.* 185–202

Balanced Design of Bootstrap Simulations

By R. L. GRAHAM,
Bell Laboratories, USA

D. V. HINKLEY†, P. W. M. JOHN and S. SHI
University of Texas at Austin, USA

[Received July 1987. Revised April 1989]

SUMMARY

Davison *et al.* (1986) have shown that finite bootstrap simulations can be improved by forcing balance in the aggregate of simulated data sets. Their methods yield first-order balance, which principally affects bootstrap estimation of bias. Here we extend the methodology to second-order balance, which principally affects bootstrap estimation of variance. The particular techniques involve Latin square and balanced incomplete block designs. Numerical examples are given to illustrate both the positive and the negative features of the balanced simulations.

Keywords: BALANCED INCOMPLETE BLOCK DESIGNS; BOOTSTRAP; LATIN CUBE; LATIN SQUARE; MONTE CARLO; ORTHOGONAL DESIGNS; RANDOMIZED BLOCK DESIGN; REGRESSION

1. INTRODUCTION

Bootstrap simulations are used to approximate sampling characteristics, such as bias and variance, of statistical measures when the form of data distribution is wholly or partly unknown; a useful introductory account of bootstrap methods has been given by Efron and Tibshirani (1986). The number of simulated samples required for accurate approximation may be in the hundreds or even thousands, unless special Monte Carlo techniques are used. In this paper we discuss the use of systematic resampling in reduction of the size of the simulation. Our work builds on previous work by Therneau (1983), Davison *et al.* (1986) and an unpublished manuscript by S. M. Ogbonmwan and H. P. Wynn.

The basic problem may be described as follows. A random sample $\mathbf{x} = (x_1, \ldots, x_n)$ is drawn from a population with distribution function F for x, and the statistical estimate $T = t(\mathbf{x})$ of θ is computed. We are now interested in approximating the distributional properties of T, or of some related quantity. This we do by substituting an estimate $\hat{F}$ for F. When nothing is assumed about F, we take $\hat{F}$ to be the empirical distribution function

$$\tilde{F}(x) = n^{-1} \sum_{j=1}^{n} I(x - x_j),$$

where $I(u) = 0$, $u < 0$, and $I(u) = 1$, $u \geqslant 0$. Thus $\tilde{F}$ puts probability n^{-1} on each x_j. It is usually impossible to execute exact theoretical calculation of the properties of interest, and one practical solution is to use simulation. Thus, if we wish to

†*Address for correspondence*: Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.

0035–9246/90/52185 $2.00

approximate the distribution function of T, we can generate B samples $\mathbf{x}_b^* = (x_{b1}^*, \ldots, x_{bn}^*)$ drawn from $\widetilde{F}$ by random sampling, calculate the analogous estimate t_b^* from data $\mathbf{x}_b^*$ and then approximate $\Pr(T \leqslant a)$ by the proportion of t^*s which are no greater than a. This approximation involves two stages. First, if the true distribution function of T under sampling from F is written $\Pr(T \leqslant a|F)$, then we are using the *statistical* approximation

$$\Pr(T \leqslant a|F) \doteq \Pr(T \leqslant a|\widetilde{F}). \tag{1.1}$$

Secondly, $\Pr(T \leqslant a|\widetilde{F})$ is being approximated *numerically* by

$$\Pr(T \leqslant a|\widetilde{F}) \doteq B^{-1} \sum I(a - t_b^*). \tag{1.2}$$

In this paper we are concerned only with the error in the second approximation (1.2). The practical importance of this error will usually depend on the possible magnitude of error in approximation (1.1).

It is helpful to consider statistics T of the form $t(\widetilde{F})$, which estimate $\theta = t(F)$, so that quite generally T possesses the expansion

$$t(\widetilde{F}) = t(F) + n^{-1} \sum L(x_j; F) + \tfrac{1}{2} n^{-2} \sum\sum Q(x_j, x_k; F) + \ldots. \tag{1.3}$$

Here L and Q are the first and second functional derivatives of t, L being more usually called the influence function of t. Similar expansions will usually be available for related quantities of interest, such as $(T - \theta)/S$, where S is a standard error of the form $n^{-1/2} s(\widetilde{F})$. See, for example, Hinkley and Wei (1984). Expansion (1.3) can be used explicitly or implicitly in various ways to improve on the simulation procedure mentioned earlier; see Davison *et al.* (1986). Here, however, we shall restrict attention to implicit use of expansion (1.3) in alternatives to simple random sampling from $\widetilde{F}$.

Section 2 introduces the concept of balanced samples $\mathbf{x}^*$ and briefly reviews relevant aspects of recent literature. Section 3 describes strategies for obtaining exact or approximate balance to second order, which would ensure correct calculation of bias and variance of T^* to order n^{-1}, for example. Numerical illustrations are given in Section 4.

2. BALANCED SAMPLES

When $\mathbf{x}^*$ is obtained by simple random sampling from empirical distribution function $\widetilde{F}$, we can represent the simulated value $T^* = t(\mathbf{x}^*)$ based on $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ by $t(\widetilde{F}^*)$, where $\widetilde{F}^*$ is the empirical distribution function of $\mathbf{x}^*$. Then expansion (1.3) applies to T^*, by writing $\widetilde{F}^*$ and $\widetilde{F}$ respectively in place of $\widetilde{F}$ and F, so that

$$T^* = T + n^{-1} \sum \widetilde{L}_j^* + \tfrac{1}{2} n^{-2} \sum\sum \widetilde{Q}_{jk}^* + \ldots, \tag{2.1}$$

with $\widetilde{L}_j^* = L(x_j^*; \widetilde{F})$, $\widetilde{Q}_{jk}^* = Q(x_j^*, x_k^*; \widetilde{F})$ and so forth.

There are two ways to relate $\widetilde{L}_j^*$, $\widetilde{Q}_{jk}^*$, etc. to the original sample values. One way is to define f_i^* to be the frequency of x_i in $\mathbf{x}^*$, so that expansion (2.1) can be written

$$T^* = T + n^{-1} \sum f_j^* \widetilde{L}_j + \tfrac{1}{2} n^{-2} \sum\sum f_j^* f_k^* \widetilde{Q}_{jk} + \ldots \tag{2.2}$$

with $\widetilde{L}_j = L(x_j, \widetilde{F})$, and so forth. Efron (1979) introduced the f_i^* in bootstrap theory and noted that $\mathbf{f}^* = (f_1^*, \ldots, f_n^*)$ follows the n-category uniform multinomial distribution. Davison *et al.* (1986) noticed that, when B random samples $\mathbf{x}_b^*$ are taken, random variation in the f_i^* will produce a random approximation of bias

$E(T^*) - E(T)$ which is of order $B^{-1/2}$ even for unbiased linear statistics such as $T = n^{-1} \Sigma x_j$, but that this first-order error is removed by forcing

$$\sum_{b=1}^{B} f_{bi}^* = B.$$

This 'first-order balance' condition, that each x_i occurs B times in the aggregate simulation, can be achieved by random permutation of B copies of $\mathbf{x}$.

Example 1. A design for $n = 5$ and $B = 10$ is obtained by writing down 10 copies of the sample subscripts 1–5, then randomly permuting this string of 50 integers:

50 integers which are 10 copies of original subscripts

12345123451234512345123451234512345123451234512345

random permutation of above integer string

32125355221455345511441331412213534313254422314254.

Successive blocks of five in this permuted string are the data subscripts for successive bootstrap samples, as shown in Table 1. Thus the first bootstrap sample is $(x_3, x_2, x_1, x_2, x_5)$.

The second way of relating expansion (2.1) to $\mathbf{x}$ is to write $x_i^* = x_{\xi(i)}$, where $\xi(i)$ is randomly drawn from $\{1, 2, \ldots, n\}$. This representation is due to S. M. Ogbonmwan and H. P. Wynn. Expansion (2.1) is now written

$$T^* = T + n^{-1} \Sigma \tilde{L}_{\xi(j)} + \tfrac{1}{2} n^{-2} \Sigma\Sigma \tilde{Q}_{\xi(j)\xi(k)} + \ldots. \tag{2.3}$$

The B successive samples are determined by the $B \times n$ matrix ξ with elements $\xi(b, i)$, $b = 1, \ldots, B$, $i = 1, \ldots, n$. The first-order balance referred to earlier is achieved by arranging that each column of ξ contains each of the integers $1, \ldots, n$ with equal frequency. Thus, for $B = n$, the matrix is a randomized block design with treatment labels $1, \ldots, n$ and with columns as blocks. For $B = kn$, k randomized block designs are stacked on top of each other.

TABLE 1

Random permutation design for $n = 5$, $B = 10$; $D_{\min} = 5$

Sample *b*	*i* *1*	*2*	*3*	*4*	*5*
1	3	2	1	2	5
2	3	5	5	2	2
3	1	4	5	5	3
4	4	5	5	1	1
5	4	4	1	3	3
6	1	4	1	2	2
7	1	3	5	3	4
8	3	1	3	2	5
9	4	4	2	2	3
10	1	4	2	5	4

TABLE 2
Randomized block design of sample subscripts, $n = 5$ and $B = 10$; $D_{\min} = 4$

Sample *b*	*i* 1	2	3	4	5
1	3	2	4	5	3
2	4	3	3	2	4
3	2	4	5	1	2
4	1	1	2	3	5
5	5	5	1	4	1
6	5	2	4	4	5
7	4	3	5	2	2
8	2	5	3	5	3
9	3	4	2	1	4
10	1	1	1	3	1

Example 2. For $n = 5$ and $B = 10$, two randomized block designs are coupled as in Table 2.

The balanced ξ design is necessarily balanced in terms of f^*. To see this, first observe that

$$f_{bi}^* = \sum_{j=1}^{n} \delta(i - \xi(b, j))$$

with $\delta(k) = 0$ or $\delta(k) = 1$ according to whether $k \neq 0$ or $k = 0$. Then, for all i,

$$\sum_{b=1}^{B} f_{bi}^* = \sum_{j=1}^{n} \sum_{b=1}^{B} \delta(i - \xi(b, j)) = \sum_{j=1}^{n} Bn^{-1} = B.$$

Evidently the ξ design contains redundancy for symmetric statistics for which only the ordered values $x_{(i)}^*$ of $\mathbf{x}^*$ are required. However, the column-by-column balance of the ξ design is useful in more complicated situations.

Example 3. Consider the regression model $x_i = \theta w_i + \epsilon_i$, with homogeneous zero-mean errors ϵ_i. Let

$$T = \Sigma x_i w_i / \Sigma w_j^2.$$

Here F denotes the error distribution, and we take $\tilde{F}$ to be the empirical distribution of centred residuals $e_i' = e_i - \bar{e}$ with $e_i = x_i - tw_i$ and $\bar{e} = n^{-1} \Sigma e_i$. Bootstrap samples are defined by

$$x_i^* = tw_i + e_i'^* = tw_i + e'_{\xi(i)}, \qquad i = 1, \ldots, n.$$

Therefore the mean of T is approximated by

$$B^{-1} \sum_{b=1}^{B} t_b^* = B^{-1} \sum_{b=1}^{B} \sum_{i=1}^{n} (tw_i + e'_{\xi(b,i)}) w_i \Big/ \sum_{j=1}^{n} w_j^2$$

$$= t + B^{-1} \sum_{b=1}^{B} \sum_{i=1}^{n} w_i e'_{\xi(b,i)} \Big/ \sum_{j=1}^{n} w_j^2$$

$$= t + \sum_{i=1}^{n} \left(B^{-1} \sum_{b=1}^{B} e'_{\xi(b,i)} \right) w_i \Big/ \sum_{j=1}^{n} w_j^2,$$

which equals $t = E(T|\widetilde{F})$ only if $\Sigma_b\, e'_{\xi(b,i)} = 0$ for all i, and not necessarily if $\Sigma f^*_{bi} = B$.

The balanced f^* design does not require that B be a multiple of n, but this is a negligible practical advantage. As to computation time, the f^* design takes somewhat longer, but with efficient programming (Gleason, 1988) the increase may be less than 50% more than for the unbalanced design.

In a general approach to choice of ξ, numerical error in approximations such as equation (1.2) can be related to properties of ξ via representations such as expansion (2.3). If we consider the B rows of ξ as points in n dimensions, then the design is said to be rth order balanced if all r-dimensional margins formed by taking r columns of ξ have uniform distributions on the n^r possible values. The randomized block designs, such as in Table 2, satisfy this condition for $r = 1$.

Now if the property of T of interest is being approximated by $E\{v(\widetilde{F}^*)|\widetilde{F}\}$, then the simulation error would be removed if the expansion of form (2.3) for $v(\widetilde{F})$ has zero components beyond the $(r+1)$st term and if ξ is rth order balanced. For example, suppose that we are approximating var(T), which involves $v(\widetilde{F}^*) = \{t(\widetilde{F}^*) - t(\widetilde{F})\}^2$, for which the Q terms in its expansion are not zero. Then first-order balance will not remove simulation error.

A corresponding notion of rth-order balance for the f^* design is that r-dimensional margins formed by any r columns of the $B \times n$ f^* table match the r-dimensional margins of the n-dimensional uniform multinomial distribution. The particular property on which we shall focus attention is that the rth-order sample moments of f^* match the corresponding multinomial moments.

From previous work it is clear that first-order balance need not appreciably reduce that part of the simulation error due to second-order imbalance. In the unpublished work by S. M. Ogbonmwan and H. P. Wynn it is shown by example that it is wise to screen first-order designs on the basis of deviation from higher order balance. Their method of screening is to push the rows of ξ apart by setting a minimum value for the distance

$$D = \sum_i \{\xi(b, i) - \xi(b', i)\}^2$$

between any pair of rows b and b'. Minimum values of D are recorded in Tables 1 and 2.

In the next section we describe and illustrate some ways of constructing second-order balanced designs. Applications of various designs are discussed in Section 4.

3. SECOND-ORDER BALANCED BOOTSTRAP DESIGNS

Careful examination of approximation (2.3) for an estimate T shows that its mean and variance can be approximated without simulation error to terms of order n^{-1} if

the ξ or f^* design is second order balanced; balance in ξ is necessary if T is not symmetric in the xs. We now describe two methods of constructing second-order balanced designs.

First, consider the ξ design. Second-order balance requires that all n^2 values of $(\xi(b, i), \xi(b, j))$ occur with equal frequency for any pair of columns i, j. Thus the minimal design has $B = n^2$. Such designs were referred to by Bose and Bush (1952) as orthogonal arrays of strength 2. Two columns of ξ, which we take to be the first two columns, can be $(1, \ldots, 1, 2, \ldots, 2, \ldots, n, \ldots, n)$ and $(1, 2, \ldots, n, \ldots, 1, 2, \ldots, n)$. For any other column j, the equal bivariate frequency condition is satisfied in conjunction with columns $i = 1$ and $i = 2$ if the elements $\xi(b, j)$ correspond to those in a Latin square with 'treatment' alphabet $(1, 2, \ldots, n)$, 'row' $\xi(b, 1)$ and 'column' $\xi(b, 2)$. If the same is to be true for columns $j = 3, \ldots, n$ then each must correspond to a Latin square. Further, the successive Latin squares must be orthogonal if the equal bivariate frequency condition is always to be satisfied. A complete design therefore requires $n - 2$ orthogonal Latin squares, which are available only if n is a power of a prime number. For relevant details see Fisher and Yates (1957).

Example 4. Table 3 shows an example design for $n = 5$, $B = 25$. Within any column, integers $1, \ldots, n$ can be replaced by any permutation thereof. In this

TABLE 3

Second-order balanced ξ design based on orthogonal Latin squares for $n = 5$ and $B = 25$; $D_{\min} = 4$

Sample	*i*				
b	*1*	*2*	*3*	*4*	*5*
1	1	1	1	1	1
2	1	2	2	2	2
3	1	3	3	3	3
4	1	4	4	4	4
5	1	5	5	5	5
6	2	1	2	3	4
7	2	2	3	4	5
8	2	3	4	5	1
9	2	4	5	1	2
10	2	5	1	2	3
11	3	1	3	5	2
12	3	2	4	1	3
13	3	3	5	2	4
14	3	4	1	3	5
15	3	5	2	4	1
16	4	1	4	2	5
17	4	2	5	3	1
18	4	3	1	4	2
19	4	4	2	5	3
20	4	5	3	1	4
21	5	1	5	4	3
22	5	2	1	5	4
23	5	3	2	1	5
24	5	4	3	2	1
25	5	5	4	3	2

way extreme rows, such as the first row of Table 3, can be avoided. This raises the possibility that some Latin square designs may be better than others; see Section 4.

The Latin square ξ design is second order balanced with respect to f^* also. To see this, observe that

$$\sum_b f_{bi}^* f_{bj}^* = \sum_l \sum_m \sum_b \delta(i - \xi(b, l))\, \delta(j - \xi(b, m)) = n(n-1), \qquad i \neq j,$$

and

$$\begin{aligned}\sum_b f_{bi}^{*2} &= \sum_l \sum_b \delta(i - \xi(b, l)) + \sum_{l \neq m} \sum_b \delta(i - \xi(b, l))\, \delta(i - \xi(b, m)) \\ &= n^2 + n(n-1).\end{aligned}$$

Therefore, because $B = n^2$,

$$B^{-1} \sum_b f_{bi}^* f_{bj}^* = \delta(i-j) + 1 - n^{-1} \tag{3.1}$$

in agreement with the multinomial expectation $E(f_i^* f_j^* | \tilde{F})$.

As with first-order balance, balance of f^* is a weaker property and will be adequate for symmetric statistics. But balance of ξ will work also for non-symmetric statistics, as the following extension of example 3 illustrates.

Example 5. The bootstrap estimate of var(T) is

$$\begin{aligned}B^{-1} \sum_b (t_b^* - t)^2 &= B^{-1} \sum_b \sum_i \sum_j w_i w_j e'_{\xi(b,i)} e'_{\xi(b,j)} \Big/ \Big(\sum w_i^2\Big)^2 \\ &= \sum_i \sum_j w_i w_j \Big(B^{-1} \sum_b e'_{\xi(b,i)} e'_{\xi(b,j)}\Big) \Big/ \Big(\sum w_i^2\Big)^2 \\ &= \Big\{\sum w_i^2 n^{-1} \sum e_k'^2 + \sum_{i \neq j} w_i w_j \Big(n^{-1} \sum e'_k\Big)^2\Big\} \Big/ \Big(\sum w_i^2\Big)^2 \\ &= \tilde{\sigma}^2 / \Sigma\, w_i^2 = \mathrm{var}(T | \tilde{F}),\end{aligned}$$

where $\tilde{\sigma}^2 = n^{-1} \Sigma\, e_i'^2$.

The exact second-order balance of ξ requires $B \geqslant n^2$, but corresponding balance of f^* may be possible for much smaller values of B and hence provide more economical designs for symmetric statistics. We now describe such designs.

We have found that second-order balance of f^* can be obtained by regarding the construction as a problem in incomplete block designs, which leads to the use of Bose's method of differences (Bose, 1939; Bose and Bush, 1952). The problem is also a special case of what are termed n-ary block designs in the more recent combinatorics literature. Billington (1984) gives a detailed account which is relevant to much of the discussion in this section. We regard the rows of ξ as blocks, with treatment labels $1, \ldots, n$ to be allocated within blocks. The objective then is a design for n treatments in $B = kn$ blocks, each of size n, satisfying the conditions

$$\sum_{b=1}^{B} f_{bi}^{*2} = k(2n-1),$$
$$\sum_{b=1}^{B} f_{bi}^{*} f_{bj}^{*} = k(n-1); \tag{3.2}$$

cf. equation (3.1).

f_{bi}^{*} is an element of the transpose of the incidence matrix N of the design. Condition (3.2) implies that the diagonal elements of the concordance matrix NN' are all equal to $k(2n-1)$ and that off-diagonal elements are all equal to $k(n-1)$. Our designs will thus be balanced in the traditional sense, except that they will be non-binary because f_{bi}^{*} can exceed unity.

In deriving the designs we replace the treatment labels $1, 2, \ldots, n$ by $0, 1, \ldots, n-1$. We choose k initial blocks and develop each of them into a cycle of n blocks by adding, in turn, $1, 2, \ldots, n-1$ and reducing the values mod n. Following Bose (1939) we choose the initial blocks in such a way that each non-zero difference occurs among all of them in exactly $k(n-1)$ ways; it will follow that there will also be $k(n-1)$ zero differences.

Some of these designs can be obtained by using balanced incomplete block designs to provide the initial blocks. Other designs may be obtained by trial and error, as was used for the following example.

Example 6. For $n = 5$ a design with $B = 15$ is constructed from $k = 3$ initial blocks (0, 1, 2, 3, 4), (0, 0, 0, 1, 3) and (0, 0, 0, 1, 2). To see that this works, we apply Bose's method to obtain differences 1–0, 0–1 = 4, ... in the first block, and so forth. The complete table of frequencies of differences is given in Table 4. The full ξ design is obtained by cycling the initial blocks and adding 1 to each entry, with the result given in Table 5.

An even smaller design exists with $B = 10$, using initial blocks (0, 0, 1, 2, 3) and (0, 0, 0, 1, 2). For this design $D_{\min} = 3$.

For certain special values of n and B, we can use balanced incomplete block designs as bases for ξ. First, if $n = 2m$ is even, then we start with a balanced incomplete block design ξ_0 for n treatments in r blocks of size m. The bth row of ξ consists of the bth row of ξ_0 duplicated, $b = 1, \ldots, r$. The remaining $s = B - r$ rows of ξ contain each treat-

TABLE 4
Frequencies of differences

	Difference				
	0	*1*	*2*	*3*	*4*
Initial block 1	0	5	5	5	5
Initial block 2	6	3	4	4	3
Initial block 3	6	4	3	3	4
Total	12	12	12	12	12

TABLE 5
Balanced cyclic design from three blocks for $n = 5$, $B = 15$; $D_{\min} = 1$

Sample b	i = 1	2	3	4	5
1	1	2	3	4	5
2	2	3	4	5	1
3	3	4	5	1	2
4	4	5	1	2	3
5	5	1	2	3	4
6	1	1	1	2	4
7	2	2	2	3	5
8	3	3	3	4	1
9	4	4	4	5	2
10	5	5	5	1	3
11	1	1	1	2	3
12	2	2	2	3	4
13	3	3	3	4	5
14	4	4	4	5	1
15	5	5	5	1	2

ment label once each. Because f_{bi}^* equals 0 or 2 in each of the first r rows of ξ, we see that for $i \neq j$

$$\sum_b f_{bi}^* f_{bj}^* = s + r(n-2)/(n-1) = B(n-1)/n$$

from which we deduce that $s = r/(n-1)$.

Example 7. One design for $n = 6$ starts with the balanced incomplete block design whose $r = 10$ blocks (rows) are (1, 2, 5), (1, 2, 6), (1, 3, 4), (1, 3, 5), (1, 4, 6), (2, 3, 4), (2, 4, 5), (2, 3, 6), (3, 5, 6) and (4, 5, 6). This is to be duplicated, and the remaining $r/(n-1) = 2$ rows of ξ set equal to (1, 2, 3, 4, 5, 6). The resulting design is shown in Table 6. The same construction appears to work for $B = 2n$, n any even integer.

A similar type of design can be constructed with $B = n$ when $n = 4m + 3$ is a prime number. Then a series of symmetric balanced incomplete block designs exists, and the solution for ξ in a single cycle may be obtained in the following way. The initial block consists of the quadratic residues of the Galois field of n elements twice each together with the element n once.

Example 8. For $n = 7$, the quadratic residues are 1, 2 and 4. The resulting design has initial block (1, 1, 2, 2, 4, 4, 7), which together with the rest of a full cycle gives ξ as in Table 7.

Similar designs are easily constructed for $n = 11$, 19, 23, 31 and so forth.

The reader familiar with resampling methods will be struck by the similarity of Tables 6 and 7 to balanced half-sample designs, which are obtained by similar

TABLE 6
Design based on balanced incomplete blocks for $n = 6$, $B = 12$; $D_{\min} = 0$

Sample *b*	*i* = *1*	*2*	*3*	*4*	*5*	*6*
1	1	1	2	2	5	5
2	1	1	2	2	6	6
3	1	1	3	3	4	4
4	1	1	3	3	5	5
5	1	1	4	4	6	6
6	2	2	3	3	4	4
7	2	2	4	4	5	5
8	2	2	3	3	6	6
9	3	3	5	5	6	6
10	4	4	5	5	6	6
11	1	2	3	4	5	6
12	1	2	3	4	5	6

TABLE 7
Design based on balanced incomplete block design for $n = 7$, $B = 7$

Sample *b*	*i* = *1*	*2*	*3*	*4*	*5*	*6*	*7*
1	1	1	2	2	4	4	7
2	2	2	3	3	5	5	1
3	3	3	4	4	6	6	2
4	4	4	5	5	7	7	3
5	5	5	6	6	1	1	4
6	6	6	7	7	2	2	5
7	7	7	1	1	3	3	6

methods. Such designs have been quite thoroughly studied. Efron (1982) gives a useful account, which includes the recommendation that the half-samples be augmented by their complements. In Table 7, for example, we would add row (3, 3, 5, 5, 6, 6, 7), and so forth.

There is, then, a variety of second-order balanced designs, some as rigid as the half-sample designs of pre-bootstrap methodology. We know that such designs will yield bias and variance approximations correct to the order n^{-1} term for statistics of the form (2.3). For some applications this will be satisfactory, but possibly not if we wish to use direct bootstrap percentile estimates. Intuitively it seems clear that ordered values of the bootstrap statistics T_b^* may not be accurate approximations to percentiles if the bootstrap design is far from third- and fourth-order balance. There is no

obvious association of second-order and higher order balance. We have two options: construct designs with higher order balance, or selectively choose among the first- and second-order balanced designs. Although the first option is real, as we know from the existence of Latin cube designs with $B = n^3$, the second option seems more practicable. This leads us naturally to consideration of numerical examples to assess the effectiveness of the design methods so far discussed.

4. NUMERICAL ILLUSTRATIONS

The balanced sampling designs described in the previous sections are designed to reduce or remove the simulation error in approximating the bootstrap mean or variance of a statistic. For example, second-order balanced designs ensure that the bootstrap mean and variance of an average are correctly calculated from the B designed samples; see also examples 3 and 5. But how well do balanced designs work with non-linear statistics, and how well do they work when percentiles, rather than moments, are being estimated? Partial answers to these questions come from our numerical experiments, in which we repeatedly apply bootstrap simulation designs for estimating properties of the sample average, Student's t-statistic, the sample correlation and the Fisher transform of the sample correlation. Some general features of the numerical results are followed up in Section 5.

4.1. *Example 9: Sample Average and t-statistic*

Suppose that we are given the skew sample of $n = 11$ x values

9.6 10.4 13.0 15.0 16.6 17.2 17.3 21.8 24.0 26.9 33.8

and that we wish to use the bootstrap to assess distributional properties of the sample average $\bar{x}$ and the Student t-statistic

$$t = (\bar{x} - \mu)\sqrt{n}/s,$$

where

$$s^2 = \Sigma(x_j - \bar{x})^2/(n - 1).$$

Consider particularly bootstrap estimates of percentiles. From a specific sampling design we obtain B values of $\bar{x}^*$, say, whose ordered values are $\bar{x}^*_{(1)} \leqslant \ldots \leqslant \bar{x}^*_{(B)}$. Then for any p such that $(B + 1)p = q$ is an integer, the $100p$th percentile of $\bar{x}$ is simply estimated by $\bar{x}^*_{(q)}$. The simulation error in this estimate is

$$e^*_p = \bar{x}^*_{(q)} - \bar{x}_p, \tag{4.1}$$

where $\Pr(\bar{x}^* \leqslant \bar{x}_p | \tilde{F}) = p$. Our numerical experiments evaluate errors (4.1) for repeated applications of each type of bootstrap design. The exact values $\bar{x}_p$, and corresponding percentiles t_p of t, are very accurately approximated from a single, completely random bootstrap design with $B = 12\,199$.

The four types of bootstrap design are completely random, first order balanced based on randomized block designs, second order balanced based on orthogonal Latin squares and second order balanced based on balanced incomplete block designs. We refer to these as zero, first, second and $2P$th order respectively; the last is

partially balanced in the sense that it balances frequencies f^*, but not ξ. All designs are applied with $B = 121$.

The methods for generating random zero- and first-order designs are self-evident. For a random second-order design, we use as basis the first nine orthogonal 11 × 11 Latin squares given by Fisher and Yates (1957), then apply random permutations of treatment labels 1, . . . , 11 separately within each square. For a random $2P$th-order design, we begin with the 11 initial blocks

0, 0, 2, 2, 3, 3, 4, 4, 8, 8, 10	0, 0, 4, 4, 5, 5, 6, 6, 8, 8, 9
0, 0, 2, 2, 3, 3, 4, 4, 8, 8, 10	0, 0, 4, 4, 5, 5, 6, 6, 8, 8, 9
0, 0, 2, 2, 3, 3, 4, 4, 8, 8, 10	0, 0, 4, 4, 5, 5, 6, 6, 8, 8, 9
0, 0, 2, 2, 3, 3, 4, 4, 8, 8, 10	0, 0, 4, 4, 5, 5, 6, 6, 8, 8, 9
0, 0, 2, 2, 3, 3, 4, 4, 8, 8, 10	0, 0, 4, 4, 5, 5, 6, 6, 8, 8, 9
0, 0, 2, 2, 3, 3, 4, 4, 8, 8, 10	

Each initial block leads to 10 additional blocks by successively adding 1, 2, . . . , 10 modulo 11. Then add 1 everywhere. A single random non-cyclic permutation of treatment labels 1, 2, . . . , 11 is applied to the set of 11 blocks, which is itself balanced. The 11 sets are then combined.

For each type of design, 100 random replicates are applied to the data. The resulting 100 errors e_p^* in equation (4.1) are summarized by their mean and standard deviation for $p = q/122$, $q = 1, 2, \ldots, 121$, these summaries being plotted against exact percentiles in Fig. 1 for zero-, second and $2P$th-order designs. Evidently the second-order designs greatly reduce simulation variability for percentile estimates when $0.05 \leqslant p \leqslant 0.95$, but the $2P$th-order design gives substantial bias in percentile

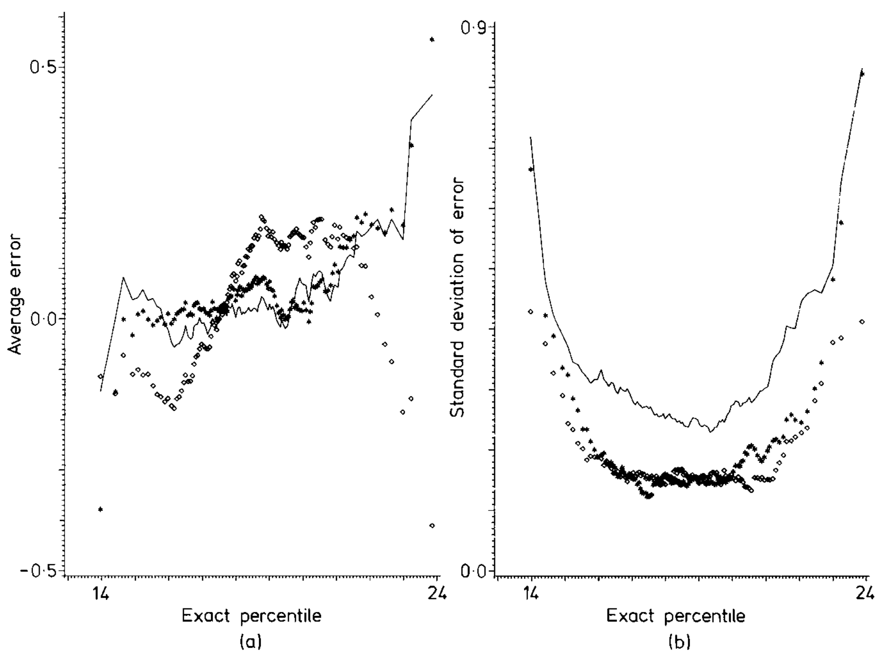

Fig. 1. (a) Averages and (b) standard deviations of 100 simulation errors in bootstrap percentile estimates for $\bar{x}$ (bootstrap size $B = 121$): ———, zeroth order; $*$, second order; ◇, $2P$th order, balanced

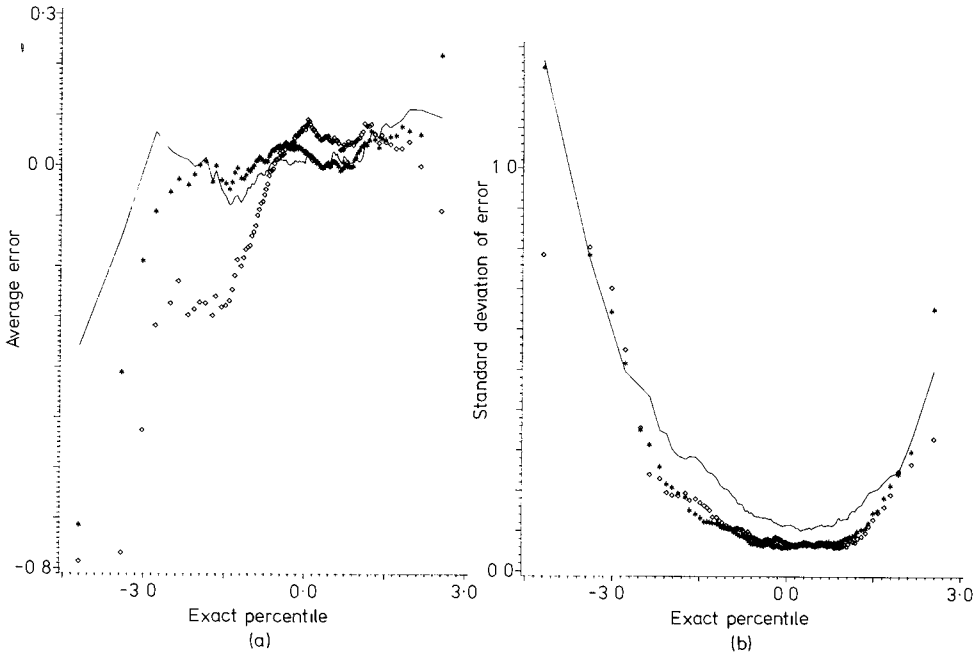

Fig. 2. (a) Averages and (b) standard deviations of 100 simulation errors in bootstrap percentile estimates for t (bootstrap size $B = 121$): ———, zeroth order; $*$, second order; $\diamond$, $2P$th order, balanced

estimates. The corresponding summaries of simulation errors in t percentiles, shown in Fig. 2, have similar features; the same 100 designs for each type are used here.

It is noticeable in both figures that at $p = 0.008$ and $p = 0.992$ the second-order design gives errors as large as the zero-order design and larger than the $2P$th-order design. We comment further on this in Section 5.

Simulation errors for 20 individual bootstrap applications with each design are shown in Fig. 3 for $\bar{x}$ percentiles and in Fig. 4 for t percentiles at p values of 0.025, 0.05, 0.10, 0.20, 0.35, 0.50, 0.65, 0.80, 0.90, 0.95 and 0.975. Errors for an individual bootstrap sample are connected by lines. Clearly the second-order design is superior to the first-order design, in that reduction of simulation error is pushed further into the tails. Large simulation errors at the extremes persist from zero- to second-order designs.

4.2. *Example 10: Sample Correlation*

This second numerical experiment concerns the distribution of the sample correlation coefficient r and its variance stabilizing transform $z = \tanh^{-1} r$. The data set is the following pseudonormal sample of $n = 11$ pairs (x, y), for which $r = 0.721$:

x	−1.21	0.21	1.33	−0.67	1.53	−1.61	0.78	−0.09	0.38	0.23	−1.41,
y	−1.48	1.18	−0.10	−1.34	0.91	−0.75	0.62	−0.93	−0.23	−0.29	−0.85.

Precisely the same bootstrap designs as in example 9 are used. For 20 random designs of each type, simulation errors in percentiles of r and z are shown in Figs 5 and 6; errors for each bootstrap are connected by lines for p values of 0.025, 0.05, 0.10, 0.20,

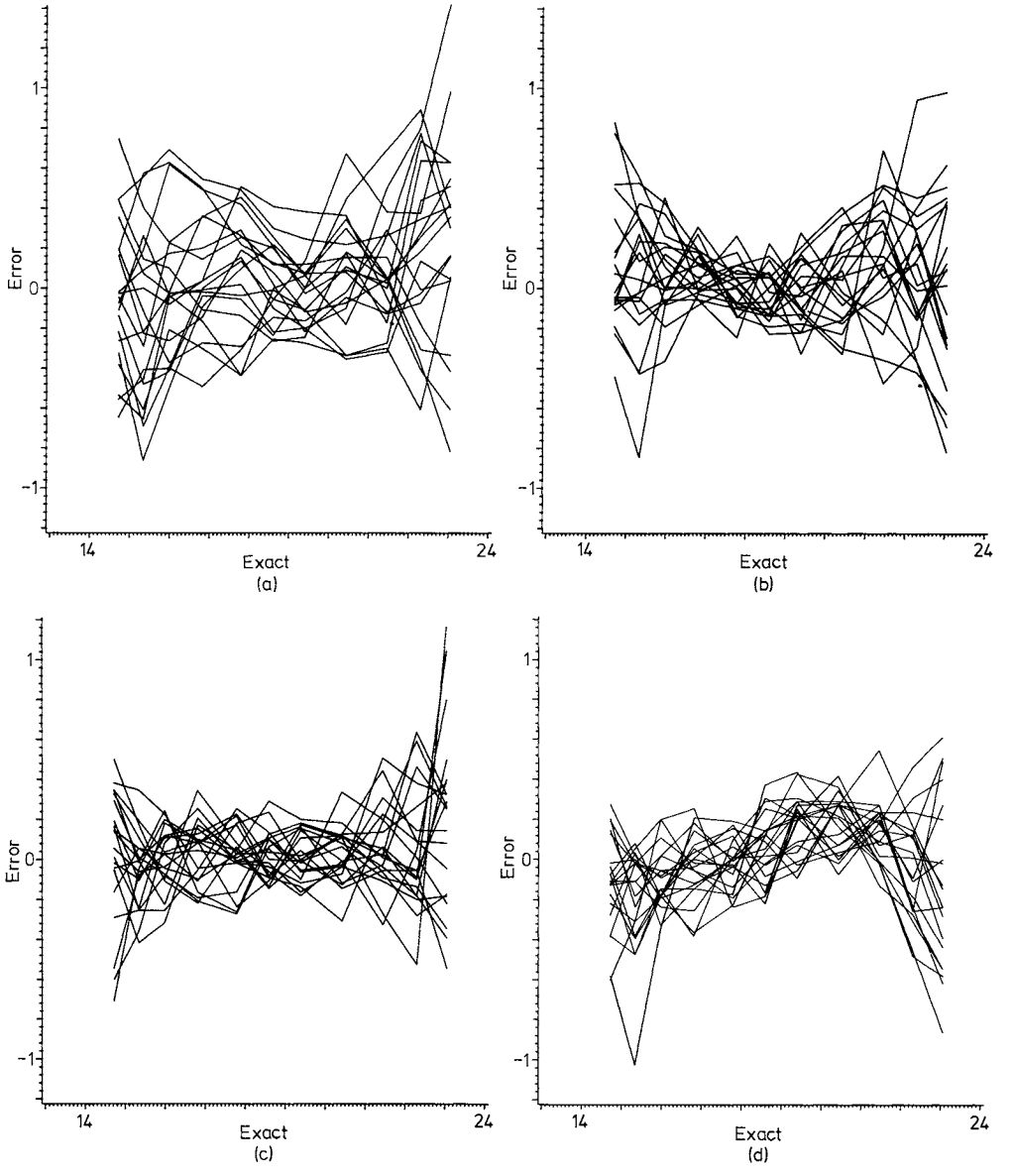

Fig. 3. Simulation errors in bootstrap percentile estimates for $\bar{x}$ in 20 bootstraps of each type with $B = 121$ (errors for a single bootstrap are connected): (a) zeroth-order design; (b) first-order design; (c) second-order design; (d) $2P$th-order design

0.35, 0.50, 0.65, 0.80, 0.90, 0.95 and 0.975. The means and standard deviations of all percentile estimates in 100 random replicates are plotted in Fig. 7 for z, omitting the first-order designs. Again the second-order designs give considerably reduced error for $0.05 \leqslant p \leqslant 0.95$, but fail to give improvement at the extremes. And again consistent error is noted in the $2P$th-order design.

Corresponding results for bootstrap estimation of the mean and variance of r and z show balanced designs to be very effective. Table 8 summarizes results for z from 20 random replicates of each type of design. The efficiency of the second-order designs is

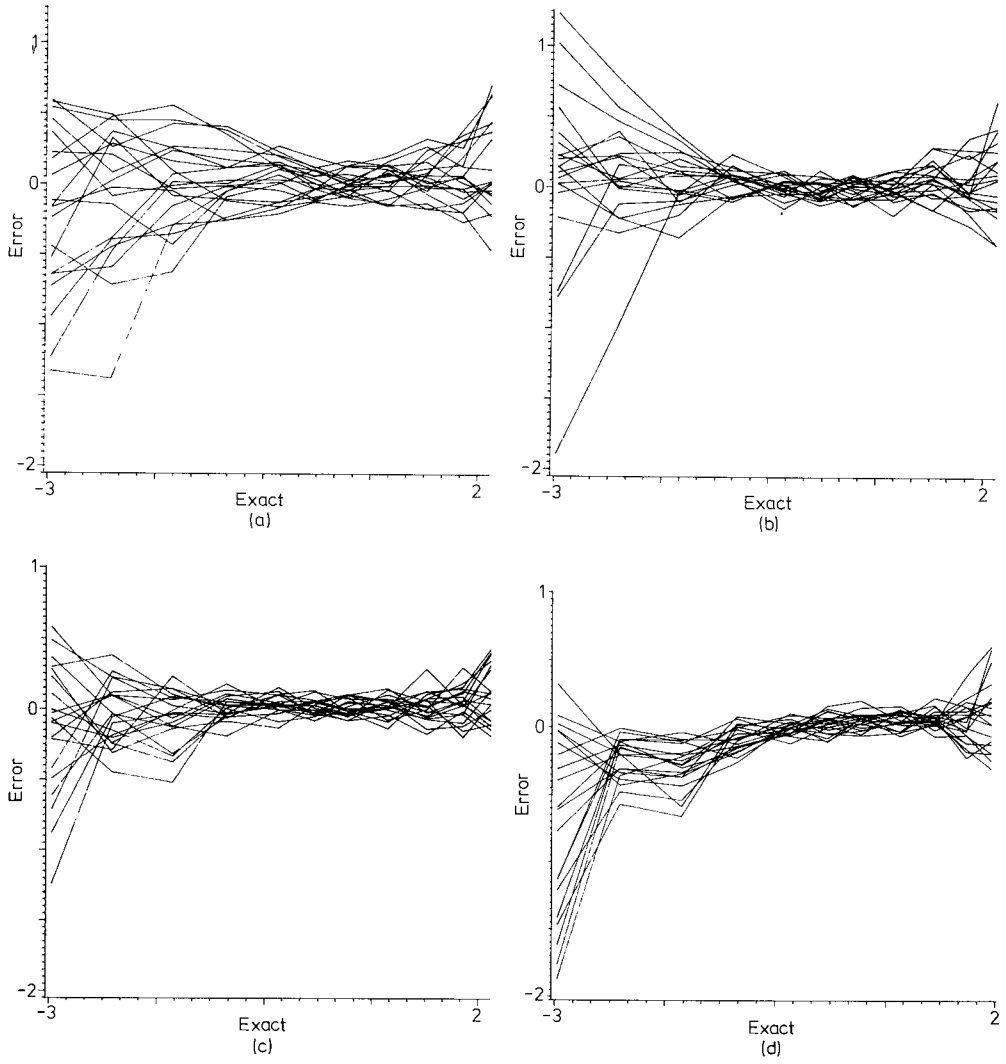

Fig. 4. Simulation errors in bootstrap percentile estimates for t in 20 bootstraps of each type with $B = 121$ (errors for a single bootstrap are connected): (a) zeroth-order design; (b) first-order design; (c) second-order design; (d) $2P$th-order design

about 4 relative to zero-order designs. Note the bias in estimating var(z) with the $2P$th-order design.

5. FURTHER COMMENTS

From the examples in Section 4 balanced sampling designs succeed in their explicit objectives, reducing simulation errors in bootstrap means and variances. The peculiar restrictions on existence of second-order balanced designs are an obstacle to practical implementation of our work. Further investigation is needed, but perhaps balanced half-sampling (Efron, 1982) will be among the preferred methods for general use; cf. examples 7 and 8.

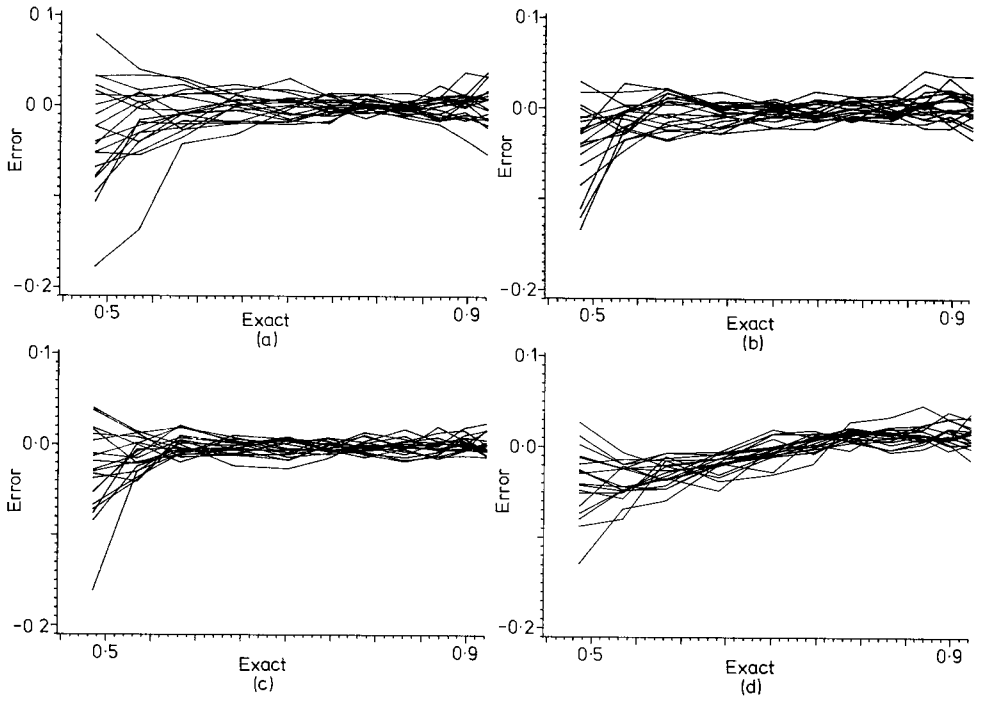

Fig. 5. Simulation errors in bootstrap percentile estimates for r in 20 bootstraps of each type with $B = 121$ (errors for a single bootstrap are connected): (a) zeroth-order design; (b) first-order design; (c) second-order design; (d) $2P$th-order design

For the bootstrap percentiles, however, the balanced designs do not give such large improvements: the Latin square designs have efficiency between 2 and 3 for percentiles away from the extremes, and the partially balanced designs based on incomplete block designs give biased estimates of percentiles. An associated phenomenon is that second-order balance is not positively correlated with third- or higher order balance.

There is the potential benefit from eliminating designs with inadequate coverage of the n-dimensional lattice cube $\{1, 2, \ldots, n\}^n$ support for ξ. For statistics which are invariant under data permutation it would be more appropriate to focus on coverage of the support simplex for $\mathbf{f}^*$. From detailed inspection of our numerical experiments

TABLE 8

Means and standard deviations of simulation errors in bootstrap estimates of the mean and variance of z†

Design	*Bootstrap mean of z*		*Bootstrap variance of z*	
	Mean error	*Standard deviation of error*	*Mean error*	*Standard deviation of error*
0th order	−0.003	0.0126	−0.0016	0.0199
1st order	−0.005	0.0154	0.0018	0.0163
2nd order	−0.010	0.0074	−0.0035	0.0107
$2P$th order	0.008	0.0079	0.0168	0.0086

† 20 replicates, $B = 121$.

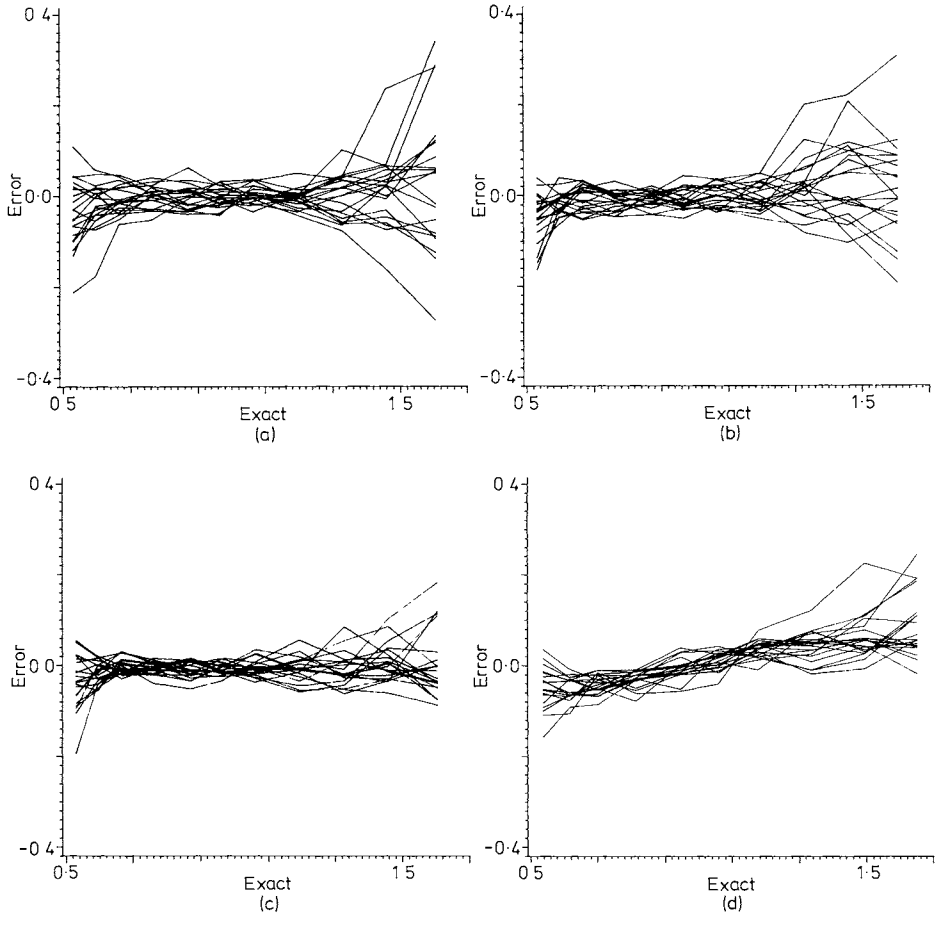

Fig. 6. Simulation errors in bootstrap percentile estimates for z in 20 bootstraps of each type with $B = 121$ (errors for a single bootstrap are connected): (a) zeroth-order design; (b) first-order design; (c) second-order design; (d) $2P$th-order design

a very small proportion of Latin square designs have a very small number of peculiar rows, which lead to large errors in extreme percentile estimates. The basic, pre-randomized design in Table 3 is an example of this. Such extreme designs should be eliminated. Unfortunately our detailed inspection of many second-order designs has not yet yielded a simple, programmable criterion for screening out bad designs. Quite possibly it would be wiser to seek out low discrepancy first-order balanced designs.

During the course of revising this paper, we have become aware of two other useful developments in bootstrap simulation, by Johns (1988) and Efron (1989). The latter includes some additional empirical assessments of our designs.

ACKNOWLEDGEMENTS

We are grateful to the National Science Foundation for financial support. Comments from two reviewers were very helpful, particularly in providing additional references.

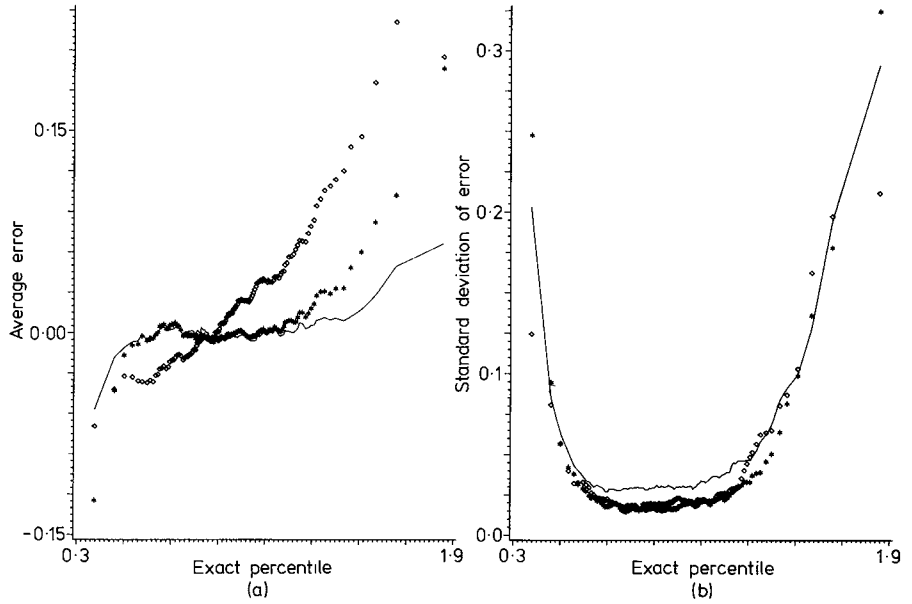

Fig. 7. (a) Averages and (b) standard deviations of 100 simulation errors in bootstrap percentile estimates for z (bootstrap size $B = 121$): ———, zeroth order; *, second order; ◇, $2P$th order, balanced

REFERENCES

Billington, E. J. (1984) Balanced n-ary designs: a combinatorial survey and some new results. *Ars Comb. A*, **17**, 37–72.

Bose, R. C. (1939) On the application of Galois fields to the problem of the construction of hyper-Graeco Latin Squares. *Sankhya*, **3**, 323–338.

Bose, R. C. and Bush, K. A. (1952) Orthogonal arrays of strength two and three. *Ann. Math. Statist.*, **23**, 508–524.

Davison, A. C., Hinkley, D. V. and Schechtman, E. (1986) Efficient bootstrap simulation. *Biometrika*, **73**, 555–566.

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

——— (1982) *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society of Industrial and Applied Mathematics.

——— (1989) More efficient bootstrap computations. *J. Am. Statist. Ass.*, to be published.

Efron, B. and Tibshirani, R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.*, **1**, 54–75.

Fisher, R. A. and Yates, F. (1957) *Statistical Tables for Biological, Agricultural and Medical Research*, 5th edn. London: Oliver and Boyd.

Gleason, J. R. (1988) Algorithms for balanced bootstrap simulations. *Am. Statistn*, **42**, 263–266.

Hinkley, D. V. and Wei, B.-C. (1984) Improvements of jackknife confidence limit methods. *Biometrika*, **71**, 331–339.

Johns, M. V. (1988) Importance sampling for bootstrap confidence intervals. *J. Am. Statist. Ass.*, **83**, 709–714.

Thernau, T. (1983) Variance reduction techniques for the bootstrap. *PhD Thesis*. Stanford University.