The Annals of Statistics
1974, Vol. 2, No. 3, 572-578

COMPLETELY SEPARATING GROUPS IN SUBSAMPLING

BY LOUIS GORDON

Stanford University

We study several combinatorial problems related to the choice of groups of subsets to be used in the subsampling procedure of Hartigan (*J. Amer. Statist. Assoc.* 1969). The complete separation property is shown to guarantee asymptotic efficiency relative to t. Problems associated with the relative variance criterion are discussed.

1. Introduction. Let $Y_1, Y_2, \cdots, Y_n$ be independent random variables symmetrically and continuously distributed about a common median μ. The subsampling method introduced in Hartigan (1969) and extended in Hartigan (1970) and Forsythe and Hartigan (1971) provides a simple construction of exact confidence sets for μ.

We briefly describe this method. Let $\mathcal{S}$ be a specially chosen collection of subsets of the indices $\{1, \cdots, n\}$ such that $A, B \in \mathcal{S}$ implies $(A\backslash B) \cup (B\backslash A) \in \mathcal{S}$.

Compute and order the set of subsample means $\{S_A/\nu(A) \mid A \in \mathcal{S}, A \neq \emptyset\}$ where $S_A = \sum \{Y_i \mid i \in A\}$ and A has cardinality $\nu(A)$. If there are m such ordered subsample means then the interval determined by the k_1th and k_2th means has exact confidence $(k_2 - k_1)/(m + 1)$.

Note that the collection $\mathcal{S}$ is a group under the symmetric difference operation $(A\backslash B) \cup (B\backslash A)$ which we denote $A \circ B$. The unit of $\mathcal{S}$ is $\emptyset$ and each subsample in $\mathcal{S}$ is self-inverse.

Having decided to use the subsampling procedure, the statistician must choose a group with which to subsample. There are several competing considerations which must be taken into account. One is computational complexity. The larger the group $\mathcal{S}$, the more time is consumed in performing the calculations. On the other hand, the group selected must be sufficiently large, for only a finite number of confidence coefficients can be obtained by using a given group. A third factor is the efficiency criterion discussed in Section 2. A fourth consideration may be the size of the relative variance introduced in Hartigan (1969) and examined in Section 3.

In this paper we study a property of groups of samples called complete separation which appears to offer a reasonable compromise among the criteria previously discussed. Its effect is to ensure a sufficient degree of heterogeneity in the group structure by ruling out any pairing of indices in the subsamples.

2. Complete separation. In order to unify the presentation, we deal throughout with the incidence matrices of the groups in question. The propositions

Received January 1972; revised June 1973.
AMS 1970 *subject classifications.* Primary 62G15; Secondary 62G35.
Key words and phrases. Typical values, complete separation, subsampling.

could just as easily be formulated in group-theoretic terms. Given a group $\mathscr{S} = \{\emptyset, A_1, \cdots, A_m\}$ on the indices $\{1, \cdots, n\}$, the $m \times n$ matrix R with $R_{ij} = 1$ if $j \in A_i$ and $R_{ij} = 0$ if $j \notin A_i$ is called the incidence matrix of $\mathscr{S}$. The rows of $\mathscr{S}$ correspond to the incidences of indices in the non-empty subsamples of $\mathscr{S}$. The reader should note that $\nu(\mathscr{S}) = m + 1$, and that we do not include in R a row of zeroes corresponding to $\emptyset$.

Naturally, R is easiest to manipulate if it is of full rank. Further, the vector of subsample means obtained from the observation vector Y is of the form DRY where D is diagonal. Hence the subsample means use the full information in Y only if R has rank n. Because of the group structure of $\mathscr{S}$ the rank n criterion is equivalent to the following easily verifiable condition, which we call complete separation: $\mathscr{S}$ is completely separating if given different indices i, j there exist subsamples $A, B \in \mathscr{S}$ with $i \in A$, $j \in B$, $i \notin B$, $j \notin A$. An enumeration problem involving complete separation may be found in Spencer (1970).

The rank n condition implies complete separation, since then R can have neither matching columns nor columns of all zeroes. The converse follows since complete separation implies R^TR is of full rank.

PROPOSITION 1. *If $\mathscr{S}$ is completely separating, then $R^TR = (m + 1)(I + J)/4$ where J is an $n \times n$ matrix of 1's and $\mathscr{S}$ has $(m + 1)$ subsamples.*

PROOF. $(R^TR)_{ij}$ counts the number of subsamples A in which both i and j appear. There exists $A_0 \in \mathscr{S}$ with $i \in A_0$. If $i = j$ then $A \leftrightarrow A \circ A_0$ is a permutation of $\mathscr{S}$ which maps the haves onto the have-nots. Hence $(m + 1)/2$ subsamples contain the index i.

Similarly, if $i \neq j$ there exist A_0 and B_0 as in the definition of complete separation. An argument using the permutations $A \leftrightarrow A \circ A_0$, $A \leftrightarrow A \circ B_0$, and $A \leftrightarrow A \circ A_0 \circ B_0$ establishes that $(R^TR)_{ij} = (m + 1)/4$.

Given a set H we denote its cardinality by $\nu(H)$. For example, the group $\mathscr{S} = \{\emptyset, \{1\}\}$ has cardinality $\nu(\mathscr{S}) = 2$.

If the observations Y_i are identically distributed with finite variance, and $\mathscr{S}_n$ is used to subsample with $\{Y_1, \cdots, Y_n\}$, then the subsampling procedure is often asymptotically relatively efficient with respect to the t procedure. The following condition is proved in Hartigan (1969) for normal parent distributions and extended in Gordon (1972). It requires that a large number of subsample means be computed and that most subsamples be nearly half-samples. More precisely, asymptotic efficiency is achieved when $\nu(\mathscr{S}_n) \to \infty$ and $\nu\{A \in \mathscr{S}_n \mid |\nu(A) - \frac{1}{2}n| > n\varepsilon\}/\nu(\mathscr{S}_n) \to 0$ for all $\varepsilon > 0$. We denote the latter proportion $P_{\varepsilon,n}$.

That complete separation implies the sufficiency conditions follows by a Chebychev inequality. If $\mathscr{S}$ is completely separating on $\{1, \cdots, n\}$ and $m = \nu(\mathscr{S}_n) - 1$ then $P_{\varepsilon,n} \leqq (\varepsilon n)^{-2}(\operatorname{tr} m^{-1}RJR^T - ((m + 1)n/2m)^2)$. It follows from Proposition 1 that $P_{\varepsilon,n} \leqq (m - n)/(\varepsilon^2 mn)$, and so $\nu(\mathscr{S}) > n$ and $P_{\varepsilon,n} \leqq 1/(n\varepsilon^2)$.

Complete separation provides an alternative proof of a result of John (1966). We here only prove a relevant portion of the former. The constructive proof we

present is of interest for its directness. It also provides an iterative algorithm for constructing the classical factorial design for estimating $2^l - 1$ main effects from 2^l treatments (e.g. see Box and Hunter (1961)). The doubling pattern of the algorithm also suggests a class of groups which are used to obtain an upper bound in the minimization problem of Section 3.

PROPOSITION 2. *Let $\mathscr{G}$ be a completely separating group on $\{1, 2, \cdots, n\}$ all of whose non-null subsamples have size k. Then $n = 2^l - 1$, $k = 2^{l-1}$ and $\nu(\mathscr{G}) = 2^l$.*

PROOF. By induction on n. If $n = 1$ the assertion is certainly correct. Assume it known that if $\bar{\mathscr{G}}$ is a completely separating group on $n_1 < n$ indices having all non-null subsets of size k_1 then the conclusion of the proposition holds. Let $\mathscr{G}$ be a completely separating group on $\{1, \cdots, n\}$ having all non-null subsets of size k.

Let A_0 be any non-null subsample in $\mathscr{G}$ and let $\bar{\mathscr{G}} = \{A \cap A_0^c \mid A \in \mathscr{G}\}$. $\bar{\mathscr{G}}$ is completely separating on the $n - k$ indices of A_0^c. Further, if $A \in \mathscr{G}$ and $A \neq A_0$, then

$$\begin{aligned} k &= \nu(A \cap A_0) + \nu(A \cap A_0^c) = \nu(A^c \cap A_0) + \nu(A \cap A_0) \\ &= \nu(A \cap A_0^c) + \nu(A^c \cap A_0)\,. \end{aligned}$$

Hence $\nu(A \cap A_0^c) = k/2$ and so $\bar{\mathscr{G}}$ is a completely separating group on $k/2 < n$ indices satisfying the induction hypotheses. Hence $k/2 = 2^{l-1}$, $\nu(\bar{\mathscr{G}}) = 2^l$ and $\nu(A_0^c) = 2_l - 1$. It follows immediately that the proposition is also true for n indices.

Note that the proof implicitly constructs the only possible incidence matrices $R(2^l - 1)$ for completely separating groups all of whose subsamples are the same size. Specifically, $R(1) = (1)$ and

$$R(2^{l+1} - 1) = \begin{pmatrix} R(2^l - 1) & R(2^l - 1) & 0 \\ R(2^l - 1) & J - R(2^l - 1) & e \\ 0 & e^T & 1 \end{pmatrix}$$

where J is a matrix of all 1's and e is a vector of all 1's.

The above argument may be continued to show directly John's (1966) result that the only groups $\mathscr{G}$ possessing subsamples of equal size have incidence matrices of form $(R(2^l - 1)R(2^l - 1) \cdots R(2^l - 1))$. The reader should note that $R(2^l - 1)$ is essentially doubled to obtain $R(2^{l+1} - 1)$. We use the doubling pattern in the following section.

3. Minimizing the relative variance. Having decided to employ the subsample method, the prospective user is confronted with the problem of choosing a group with which to subsample. As a possible criterion for judging prospective groups Hartigan (1969) suggests minimizing the relative variance

$$r(\mathscr{G}) = \frac{n}{2m(m-1)} \sum \sum_{A,B \in \mathscr{G} \setminus \{\emptyset\}} \nu(A \circ B)/\nu(A)\nu(B)$$

where $m = \nu(\mathscr{G}) - 1$.

The quantity $r(\mathcal{G})$ is the expected sample variance of the subsample means $\{S_A/\nu(A)\}$ normalized by the variance of $\sum_1^n Y_i/n$ when the Y's have a common finite variance. Hence $r(\mathcal{G})$ measures the expected spread of the subsample means.

We first indicate why unrestricted minimization of $r(\mathcal{G})$ yields an unpalatable result. An alternative might be to minimize over all completely separating groups.

Unfortunately, $r(\mathcal{G})$ is difficult to compute. It is therefore convenient to work with an auxiliary criterion $s(\mathcal{G}) = n \sum \{1/(2m\nu(A)) \mid A \in \mathcal{G}\backslash\{\emptyset\}\}$, where, again $m = \nu(\mathcal{G}) - 1$. This quantity may be interpreted as the normalized variance of one subsample mean chosen at random. The next proposition relates $s(\mathcal{G})$ and $r(\mathcal{G})$.

PROPOSITION 3. *Let $\mathcal{G}$ be a group on $\{1, \cdots, n\}$ and let $m = \nu(\mathcal{G}) - 1$. If $m > 1$, then $m/m + 1 \leqq s(\mathcal{G}) \leqq r(\mathcal{G})$. Equality is obtained* iff *all subsamples in $\mathcal{G}$ have the same cardinality and all indices are represented in subsamples of $\mathcal{G}$.*

PROOF. For any fixed set $A_0 \in \mathcal{G}$, $A_0 \leftrightarrow B \circ A_0$ is a permutation of $\mathcal{G}\backslash\{A_0, \emptyset\}$. Hence, since $A \circ A = \emptyset$,

$$r(\mathcal{G}) = \frac{n}{2m(m-1)} \sum \left\{ \frac{1}{2} \left[\frac{\nu(A \circ B)}{\nu(B)} + \frac{\nu(B)}{\nu(A \circ B)} \right] \Big/ \nu(A) \,\Big|\, A \neq B \neq \emptyset \neq A \right\}.$$

So $r(\mathcal{G}) \geqq s(\mathcal{G})$ with equality iff $\nu(A)$ is constant over $\mathcal{G}\backslash\{\emptyset\}$. Further, $\sum \nu(A) \leqq n(m+1)/2$ with equality only if each index lies in some subsample. Jensen's inequality implies

$$m^{-1} \sum 1/\nu(A) \geqq 1/(m^{-1} \sum \nu(A))$$

with equality iff $\nu(A)$ is again constant, which proves the left-hand inequality.

We may now compute $r(\mathcal{G}^*) = .8$ for $n = 16$ and the group $\mathcal{G}^* = \{\emptyset, \{1, \cdots, 10\}, \{6, \cdots, 15\}, \{1, \cdots, 5, 11, \cdots, 16\}\}$. If $\mathcal{G}$ were any other group with $\nu(\mathcal{G}) > 4$, then $r(\mathcal{G}) > .875$ by Proposition 3. Hence unrestricted minimization of $r(\mathcal{G})$ has led to a choice of a subsampling scheme permitting the construction of few confidence sets. We might find $\mathcal{G}^*$ undesirable for other reasons as well. $\mathcal{G}^*$ neither uses all 16 possible observations nor distinguishes between observations 1, 2, or 3. We may avoid these drawbacks by restricting the minimization to completely separating groups. Recall that if $\mathcal{G}$ is completely separating on an n-set then $\nu(\mathcal{G}) > n$. Hence, for many observations, confidence sets at many levels are available. This restriction also ensures that the asymptotic efficiency criteria of Section 2 will be satisfied. We therefore investigate the behavior of $r^*(n)$ and $s^*(n)$, the respective minima of $r(\mathcal{G})$ and $s(\mathcal{G})$ over completely separating groups on n indices.

We know from Proposition 3 that $1 - n^{-1} \leqq s^*(n) \leqq r^*(n)$ since $\nu(\mathcal{G}) \geqq n + 1$ when $\mathcal{G}$ is completely separating. Further, this bound is attained when $n = 2^l - 1$ by using the group $\mathcal{H}(2^l - 1)$ whose incidence matrix $R(2^l - 1)$ is derived in Section 2.

While we are unable to find the optimal groups which yield $r^*(n)$ and $s^*(n)$, it does seem reasonable to suppose that these groups have structure similar to $\mathscr{H}(2^l - 1)$. We may generalize the iterative construction of the incidence matrices $R(2^l - 1)$ in the obvious manner: We set $R(1) = (1)$,

$$R(2n) = \begin{pmatrix} R(n) & R(n) \\ R(n) & J - R(n) \\ 0 & e^T \end{pmatrix} \quad \text{and} \quad R(2n+1) = \begin{pmatrix} R(n) & R(n) & 0 \\ R(n) & J - R(n) & e \\ 0 & e^T & 1 \end{pmatrix}$$

where J is a matrix of all 1's and e is a vector of all 1's. It is easy to verify inductively that the $R(n)$ are incidence matrices of completely separating groups $\mathscr{H}(n)$ on $\{1, \cdots, n\}$. Further, $\nu(\mathscr{H}(n)) = \inf\{2^l \mid 2^l > n\}$. The reader should observe that $R(2^l)$ corresponds to the fractional factorial design for orthogonally estimating main effects in the presence of 2nd order interactions and the absence of higher order interactions (e.g. see Box and Hunter 1961)).

We now derive bounds on $r(\mathscr{H}(n))$ and $s(\mathscr{H}(n))$, and so bound $r^*(n)$ and $s^*(n)$. Since the $R(n)$ are defined inductively, we use induction to obtain the bounds. We sketch the proof in a series of lemmas. Throughout we denote $m_n = \nu(\mathscr{H}(n)) - 1$.

LEMMA 1.

(a) $$s(\mathscr{H}(2n)) = m_{2n}^{-1}[m_n s(\mathscr{H}(n)) + m_n + 1].$$

(b) $$s(\mathscr{H}(2n+1)) = m_{2n+1}^{-1}((2n+1)/2n)[m_n s(\mathscr{H}(n)) + (m_n + 1)(n/n+1)].$$

(c) $$r(\mathscr{H}(2n)) = \tfrac{1}{2}m_{2n}^{-1}[(m_n - 1)r(\mathscr{H}(n)) + 2(m_n + 1)s(\mathscr{H}(n)) + (m_n + 1)^2 m_n^{-1}].$$

(d) $$r(\mathscr{H}(2n+1)) = \tfrac{1}{2}(m_{2n+1}(m_{2n+1} - 1)n)^{-1}(2n+1) \times [m_n(m_n - 1)r(\mathscr{H}(n)) + 2m_n(m_n + 1)s(\mathscr{H}(n)) + ((m_n + 1)n/(n+1))^2].$$

PROOF. The proof is immediate from the iterative definition of $R(n)$. In particular, subsamples in $\mathscr{H}(2n)$ have cardinalities either n or $2\nu(A)$ for some $A \in \mathscr{H}(n)$. The correspondences between rows of $R(n)$ and $R(2n)$ also enable one to compute the cardinalities of symmetric differences in terms of n and the cardinalities of sets in $\mathscr{H}(n)$.

The following two lemmas may be proved inductively by means of the preceding recurrence relations. The proof of Lemma 3 also requires the use of Lemma 2.

LEMMA 2.

$$s(\mathscr{H}(n)) \leqq 1 - 2\frac{m_n + 1 - n}{(n+1)(2n+1)m_n}.$$

LEMMA 3.

$$r(\mathscr{H}(n)) \leqq 1 + 3/m_n.$$

We summarize the results below:

PROPOSITION 4.

(a) $1 - (1 + m_n)^{-1} \leqq s^*(n) \leqq r^*(n) \leqq r(\mathscr{H}(n)) \leqq 1 + 3/m_n$.

(b) $1 - (1 + m_n)^{-1} \leqq s^*(n) \leqq s(\mathscr{H}(n)) < 1$.

The lower bound is achieved iff $n = 2^l - 1$.

The following table is presented to give some idea of the behavior of $r(\mathscr{G})$ and $s(\mathscr{G})$. To facilitate comparisons, we here table deviations from 1 rather than the quantities themselves. Tabled are the upper bounds of Proposition 4 denoted UB$s(n)$ and UB$r(n)$. Also included are $r(\mathscr{H}(n))$, $s(\mathscr{H}(n))$, and the r and s values for the power set group r(PS) and s(PS). The leading column gives the lower bound of Proposition 4, denoted LB.

TABLE 1

n	LB$(n)-1$	$s(\mathscr{H}(n))-1$	UB$s(\mathscr{H}(n))-1$	s(PS)-1	$r(\mathscr{H}(n))-1$	UB$r(\mathscr{H}(n))-1$	r(PS)-1
10	−.0625	−.0361	−.0035	.1456	−.0061	.2000	.2923
20	−.0312	−.0275	−.0009	.0599	−.0028	.0968	.1196
30	−.0312	−.0302	−.0001	.0372	−.0392	.0968	.0740
40	−.0156	−.0086	−.0003	.0271	−.0014	.0476	.0540
50	−.0156	−.0122	−.0001	.0213	−.0098	.0476	.0423
60	−.0156	−.0149	.0000	.0176	−.0141	.0476	.0346
70	−.0078	−.0037	−.0001	.0149	.0004	.0236	.0292
80	−.0078	−.0043	−.0001	.0130	−.0007	.0236	.0251
90	−.0078	−.0045	.0000	.0115	−.0011	.0236	.0218
100	−.0078	−.0061	.0000	.0103	−.0043	.0236	.0191

Note that the bound on $r(\mathscr{H}(n))$ is rather crude, but that $r(\mathscr{H}(n))$ can exceed 1. Also, $r(\mathscr{H}(n))$ and $s(\mathscr{H}(n))$ appear to be piecewise decreasing with jumps at the points $n = 2^l$.

In conclusion, it would appear that a reasonable choice of group for subsampling should have structure similar to $\mathscr{H}(n)$. In particular, $\mathscr{H}(n)$ or the group whose incidence matrix is given by the first n columns of $\mathscr{H}(2^l - 1)$, $2^{l-1} \leqq n \leqq 2^l - 1$, may be good choices.

Acknowledgment. This paper is based on a portion of the author's doctoral dissertation prepared at Stanford University while the author held an NSF Graduate Fellowship. The author would like to thank his advisor, Bradley Efron, for a number of helpful conversations.

REFERENCES

[1] BOX, G. and HUNTER, J. (1961). The 2^{k-p} fractional factorial designs, I. *Technometrics* **3** 311-351.

[2] FORSYTHE, A. and HARTIGAN, J. (1970). Efficiency of confidence intervals generated by repeated subsample calculations. *Biometrika* **57** 627-640.

[3] GORDON, L. (1974). Efficiency in subsampling. To appear in *Ann. Statist.* **2**.

[4] HARTIGAN, J. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64** 1303-1317.

[5] HARTIGAN, J. (1970). Exact confidence intervals in regression problems with independent symmetric errors. *Ann. Math. Statist.* **41** 1992–1998.

[6] JOHN, P. (1966). On identity relationships for 2^{n-r} designs having words of equal length. *Ann. Math. Statist.* **37** 1842–1843.

[7] SPENCER, J. (1970). Minimal completely separating systems. *J. Combinatorial Theory* **8** 446–447.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305