The Annals of Statistics
1974, Vol. 2, No. 4, 739-750

EFFICIENCY IN SUBSAMPLING

BY LOUIS GORDON[1]

Stanford University

An analysis of the behavior of the cdf randomly determined by the subsample means yields results on the asymptotic relative efficiency and consistency of Hartigan's subsampling procedure. Examination of higher order approximation to the normal is carried out in two special cases.

1. Introduction and summary. One may test the hypothesis that n independent random variables $Y_1, \cdots, Y_n$ are symmetrically and continuously distributed about the common median 0 by computing all possible sums obtainable by changing the signs of a given subset of the observations. The hypothesis is rejected if the sum of the original values is unusually large or small among the class of all sums obtained by sign changes. The test is exact and is due to Fisher (1935, page 46).

Fisher's test may be inverted to provide an exact confidence procedure for the common median of $Y_1, \cdots, Y_n$, independent, continuous, and symmetrically distributed random variables: Given $A \subset \{1, \cdots, n\}$ write $S_A = \sum \{Y_i \mid i \in A\}$ and $\nu(A)$ for the cardinality of A. Compute and order the subsample means $S_A/\nu(A)$ for all nonempty $A \subset \{1, \cdots, n\}$. Choosing an interval whose endpoints are the k_1th and k_2th ordered subsample means yields an exact confidence interval with confidence $(k_2 - k_1)/(m + 1)$, where $m = 2^n - 1$ is the number of all the subsample means computed.

Hartigan (1969) generalizes Fisher's formulation by showing that, instead of computing all subsample means $S_A/\nu(A)$, $A \neq \emptyset$, one retains exactness in the Fisher procedure if and only if, for $\mathcal{G}$ the collection of nonempty subsets $A \subset \{1, \cdots, n\}$ for which $S_A/\nu(A)$ is computed, $\mathcal{G} \cup \{\emptyset\}$ constitutes a group under the set operation symmetric difference. In a later paper, Hartigan (1970) extends the method to certain analysis of variance problems.

This study is primarily concerned with the behavior of the subsample means in the independent identically distributed (i.i.d.) case. We follow Hartigan (1969) and Forsythe and Hartigan (1971) in examining the empirical cumulative distribution function (cdf) determined by the aggregate of the subsample means. In the former paper conditions on the collections $\mathcal{G}$ are given to assure that for large samples of i.i.d. normal observations, the properly normalized subsample empirical cumulative distribution function approximates a normal cdf. The latter shows that, in the original Fisher formulation, the normalized subsample empirical cdf also converges to a normal cdf whenever the Y_i are i.i.d.

Received September 1972; revised August 1973.
[1] Supported by an NSF Graduate Fellowship.
AMS 1970 *subject classifications*. Primary 62G15; Secondary 62G20.
Key words and phrases. Subsampling, typical values.

with finite variance. In Section 2 we show that the result of Hartigan (1969) is applicable to the i.i.d. case with finite variance. In this case, we obtain as a corollary that for certain sequences of subsampling schemes the procedure has Pitman efficiency 1 with respect to the Student's t procedure.

In Section 3 we present a Monte Carlo study of the behavior of subsampling for various symmetric parent distributions in small samples. A more elementary examination of the subsample empirical cdf, yielding a consistency result for symmetric parents with moderately heavy tails, is found in Section 4.

We indicate in Section 5 a computation for the limiting variance of the normalized cdf studied in Section 2. The computations indicate that fine behavior depends both on the higher moments of the parent distribution, and on the structure of the collection $\mathcal{G}$ which determines the choice of subsamples.

Although the subsampling method does not appear to be well suited for use in the i.i.d. case because of its extensive use of averaging, the examination of this case may give some indication of the method's behavior in more complex situations. In this light, we draw the following conclusions:

In large samples one might as well use subsampling instead of t-methods. The advantages of this approach are somewhat less sensitivity to long tails than t and exactness when the observations are symmetric about a common median. The major disadvantage over t is computational complexity which can be surmounted with the aid of an electronic computer. However, as indicated by the results of Section 4, the median subsample tends to be close to the grand mean of all observations. Subsampling therefore shares many of the drawbacks of the t-procedure. In particular, the reader should recall that the Fisher–Yates Normal Scores procedure is asymptotically superior to the t-procedure (e.g., see Hájek and Šidák (1967) page 279).

2. Relative efficiency of subsampling and t. In this section we show that, under certain sufficient conditions, the subsampling procedure is efficient with respect to t. The proof is accomplished by an extension of Hartigan's technique of examining the distribution function randomly determined by the subsample means.

Given any collection of indices $\{1, \cdots, n\}$ and subsets A and B, we denote their symmetric difference by $A \circ B$. A collection $\mathcal{G}$ satisfying the conditions of Hartigan (1969) that (1) $\phi \notin \mathcal{G}$ and (2) $A \circ B \in \mathcal{G} \cup \{\phi\}$ for all $A, B \in \mathcal{G}$ is called a reduced group. Observe that $\circ$ is an Abelian group operation having ϕ as its identity and that conditions (1) and (2) require that $\mathcal{G} \cup \{\phi\}$ be a group, hence the terminology.

Given $Y_1, \cdots, Y_n$, i.i.d. with mean zero, variance one and common cdf F, we apply the subsampling procedure determined by the reduced group $\mathcal{G}_n$. We place three basic requirements on the reduced groups $\mathcal{G}_n$ which are used to form the subsample means. They are employed in Hartigan (1969) to obtain a more restricted result.

In particular, we demand that (1) the reduced group $\mathcal{G}_n$ be composed of

subsets of the indices $\{1, \cdots, n\}$, (2) nearly all subsets comprising $\mathscr{G}_n$ contain almost half the indices $\{1, \cdots, n\}$, and (3) we eventually compute large numbers of subsample means.

Requirement (1) needs no justification; the reduced group $\mathscr{G}_n$ is to be used when n observations are taken. Requirement (2) imposes some regularity on the structure of the groups. Note first that if $\mathscr{G}$ is a reduced group on $\{1, \cdots, n\}$ then the average size of a subset in $\mathscr{G}$ is $\frac{1}{2}n(\nu(\mathscr{G}) + 1)/\nu(\mathscr{G})$ whenever each index lies in some component index subset of $\mathscr{G}$.

Secondly, if R is the $\nu(\mathscr{G}) \times n$ incidence matrix of 0's and 1's for the reduced group $\mathscr{G}$ and J is a matrix of the same dimensions composed entirely of 1's, then (2) implies that nearly all pairs of rows of $R - \frac{1}{2}J$ are practically orthogonal. Hence most pairs of rv's in the collection $\{n^{-1}(S_A - \frac{1}{2}S_n) \mid A \in \mathscr{G}\}$ are asymptotically normal and practically uncorrelated, while condition (3) guarantees that we obtain a large number of them. Here we depend on the observation that, since $S_n = S_A + S_{A^c}$, $S_A - \frac{1}{2}S_n = \frac{1}{2}(S_A - S_{A^c})$.

It therefore seems reasonable that the random cdf,

$$\nu^{-1}(\mathscr{G}_n) \sum I_{\{n^{-1}(S_A - \frac{1}{2}S_n) \leq tn^{-\frac{1}{2}}\}},$$

determined as a sum of indicator functions, behaves similarly to an empirical cdf for standard normal variates. In particular, the above empirical subsample cdf converges uniformly in probability to a standard normal c.d.f.

Note that this convergence yields consistent estimates for the standard normal quantiles and that subsample empirical cdf is centered about S_n/n. Further, under requirement (2), most sets of indices in $\mathscr{G}_n$ are nearly half-samples, so that $2(S_A - \frac{1}{2}S_n)/n$ is approximately $S_A/\nu(A) - S_n/n$. Hence, in the presence of a finite second moment, the Hartigan and t procedures are asymptotically relatively efficient, for the given sequence of reduced groups $\mathscr{G}_n$.

Formally, let $\mathscr{G}_n$ be a sequence of reduced groups on $\{1, \cdots, n\}$. For $\varepsilon > 0$, define $P_{\varepsilon,n} = \nu\{A \in \mathscr{G}_n \mid n\varepsilon < |\nu(A) - \frac{1}{2}n|\}/\nu(\mathscr{G}_n)$. Here, we denote by $\nu(D)$ the cardinality of a given set D. $P_{\varepsilon,n}$ is the proportion of subsets in $\mathscr{G}_n$ whose cardinality deviates substantially from half the number of indices available. We now state an analogue of the Glivenko–Cantelli theorem for the empirical cdf of subsample means.

THEOREM 1. *If* $Y_1, Y_2, \cdots$ *are* i.i.d. *as* F *with arbitrary mean and unit variance, and* $\mathscr{G}_n$ *is a sequence of reduced groups on* $\{1, 2, \cdots, n\}$ *with* $P_{\varepsilon,n} \to 0$ *and* $\nu(\mathscr{G}_n) \to \infty$, *then*

$$\sup_t |\nu^{-1}(\mathscr{G}_n) \textstyle\sum_{A \in \mathscr{G}_n} I_{\{S_A/\nu(A) - S_n/n \leq t/n^{\frac{1}{2}}\}} - \Phi(t)| \to_P 0$$

where Φ *is the standard normal* cdf.

The proof is accomplished in a series of lemmas. We may assume without loss of generality that $EY_i = 0$. We write $\phi(t)$ for the standard normal density.

The following notation and conventions are used throughout the section. F denotes a distribution function with mean 0 and variance 1. $Y_1, Y_2, \cdots$ are a

sequence of i.i.d. random variables distributed as F. S_n stands for a random variable distributed as the nth partial sum of the Y's. The normalized n-fold convolution of F is written $F^{(n)}$ and is the distribution function of $S_n/n^{\frac{1}{2}}$. If several independent partial sums are needed, they are denoted $S_p^{(1)}$, $S_q^{(2)}$, and so forth.

LEMMA 1. *Let t be fixed; then there exist a constant B and a sequence $b_n \downarrow 0$ such that, if* (i) $\varepsilon < \frac{1}{8}$, (ii) $|p_i - \frac{1}{4}| < n\varepsilon$, $i \leqq 4$, *and* (iii) $\sum_1^4 p_i = n$, *then*

$$|EI_{\{S_{p_1}^{(1)} - S_{p_2}^{(2)} + |S_{p_3}^{(3)} - S_{p_4}^{(4)}| \leqq tn^{\frac{1}{2}}\}} - \Phi^2(t)| \leqq B\varepsilon + b_n .$$

PROOF. Let a_j be a sequence decreasing to 0 for which $a_j \geqq \sup |F^{(j)}(t) - \Phi(t)|$. We suppress the subscript p_i and write $S^{(i)}$ for $S_{p_i}^{(i)}$. We write $J = EI_{\{|S^{(1)} - S^{(2)}| + |S^{(3)} - S^{(4)}| \leqq tn^{\frac{1}{2}}\}}$. Let $Z^{(i)}$, $i \leqq 4$ be four mutually independent standard normal variates, independent of the $S^{(i)}$.

Conditioning on $S^{(2)}$, $S^{(3)}$, and $S^{(4)}$ yields

$$J = E\Phi([tn^{\frac{1}{2}} + S^{(2)} - |S^{(3)} - S^{(4)}|]p_1^{-\frac{1}{2}}) + \alpha_1$$

where $|\alpha_1| < a_{p_1}$. Use of the mean value theorem then yields

$$J = E\Phi(2t + S^{(2)}p_2^{-\frac{1}{2}} - |S^{(3)}p_3^{-\frac{1}{2}} + S^{(4)}p_4^{-\frac{1}{2}}|) + \alpha_1 + \beta$$

where

$$|\beta| < |t||(n/p_1)^{\frac{1}{2}} - 1| + E\sum_{j=2}^4 |S^{(j)}p_j^{-\frac{1}{2}}||1 - (p_j/p_1)^{\frac{1}{2}}| .$$

Since, by assumption, $\varepsilon < \frac{1}{8}$, approximation shows the ratios lie within 32ε of the indicated integers so that $|\beta| < 128(1 + |t|)\varepsilon$. We may now write J as the expected value of an indicator function of $Z^{(1)}$, $S^{(2)}$, $S^{(3)}$, $S^{(4)}$. Repeated application of conditioning, the central limit theorem, followed by reconversion to an expression using an indicator function yields

$$J = EI_{\{|Z^{(1)} - Z^{(2)}| + |Z^{(3)} - Z^{(4)}| \leqq 2t\}} + \beta + \sum_1^4 \alpha_i$$

where $|\alpha_i| < 4a_{p_i}$.

Observe that $\{|Z^{(1)} - Z^{(2)}| + |Z^{(3)} - Z^{(4)}| \leqq 2t\}$ is the intersection of independent events. Hence $|J - \Phi^2(t)| \leqq 128(1 + |t|)\varepsilon + 16a_{n/16}$.

LEMMA 2. *Given t there exist constant B_0 and a sequence d_n decreasing to zero such that, for ε and p_i as in Lemma* 1,

$$|E[I_{\{S_{p_1}^{(1)} - S_{p_2}^{(2)} + S_{p_3}^{(3)} - S_{p_4}^{(4)} \leqq tn^{\frac{1}{2}}\}} - \Phi(t)][I_{\{S_{p_1}^{(1)} - S_{p_2}^{(2)} - S_{p_3}^{(3)} + S_{p_4}^{(4)} \leqq tn^{\frac{1}{2}}\}} - \Phi(t)]| < B_0\varepsilon + d_n .$$

PROOF. We may multiply out the quantity whose expectation is to be taken and then apply Lemma 1, if we may similarly bound

$$|EI_{\{S_{p_1}^{(1)} - S_{p_2}^{(2)} + S_{p_3}^{(3)} - S_{p_4}^{(4)}\}} - \Phi(t)| .$$

This may be done using the same conditioning-unconditioning and Taylor series arguments as in Lemma 1.

LEMMA 3. *Let $\mathcal{G}_n$ be a sequence of groups on the first n indices for which $P_{\varepsilon,n} \to 0$*

for any positive ε, *and let* $\nu(\mathscr{G}_n) = m_n \to \infty$. *Then for any fixed* t,

$$m_n^{-1} \sum_{A \in \mathscr{G}_n} I_{\{S_A - S_{A^c} \leqq t n^{\frac{1}{2}}\}} \to_P \Phi(t) .$$

Complementation is taken relative to the set $\{1, \cdots, n\}$.

PROOF. The proof turns on Hartigan's idea of showing that the difference converges to 0 in L^2. We suppress the subscript n in the notations $\mathscr{G}_n$, m_n, and $P_{\varepsilon,n}$ for the sake of simplicity. Choose ε positive with $\varepsilon < \frac{1}{8}$. Consider the subset of $\mathscr{G}$ given by $\mathscr{Q}_\varepsilon = \{A \in \mathscr{G} \mid |\nu(A) - \frac{1}{2}n| \leqq n\varepsilon\}$. Let $\mathscr{R}_\varepsilon = \{(A, B) \in \mathscr{Q}_\varepsilon \times \mathscr{Q}_\varepsilon \mid A \circ B \in \mathscr{Q}_\varepsilon\}$. Note that $\nu(\mathscr{G} \times \mathscr{G} \backslash \mathscr{R}_\varepsilon) \leqq 3m^2 P_\varepsilon$. We write CD for $C \cap D$ when C and D are subsets. If $(A, B) \in \mathscr{R}_\varepsilon$, we may write

$$S_A - S_{A^c} = S_{AB} + S_{AB^c} - S_{A^cB} - S_{A^cB^c} ,$$

and

$$S_B - S_{B^c} = S_{AB} - S_{AB^c} + S_{A^cB} - S_{A^cB^c} .$$

Denote $p_1 = \nu(AB)$, $p_2 = \nu(A^cB^c)$, $p_3 = \nu(A^cB)$ and $p_4 = \nu(AB^c)$. Then $(A, B) \in \mathscr{R}_\varepsilon$ implies $|p_i - n/4| \leqq 2n\varepsilon$. Note that the hypothesis of Lemma 2 is satisfied.

Now,

$$\begin{aligned} E(m^{-1} \textstyle\sum_{A \in \mathscr{G}} [I_{\{S_A - S_{A^c} \leqq t n^{\frac{1}{2}}\}} - \Phi(t)])^2 \\ &= m^{-2}(E_{\mathscr{G} \times \mathscr{G}} \textstyle\sum_{\mathscr{R}_\varepsilon} + \sum_{\mathscr{R}_\varepsilon} [I_{\{S_{AB} - S_{A^cB^c} + S_{A^cB} - S_{AB^c} \leqq t n^{\frac{1}{2}}\}} - \Phi(t)] \\ &\qquad \times [I_{\{S_{AB} - S_{A^cB^c} - S_{A^cB} + S_{AB^c} \leqq t n^{\frac{1}{2}}\}} - \Phi(t)]) \\ &\leqq m^{-2}[4(3m^2 P_\varepsilon + m) + (2B_0\varepsilon + d_n)m^2] , \end{aligned}$$

which completes the proof.

If, instead of requirement (3), we demand that the proportion of observations used when n data points are provided, $n^{-1}\nu\{i \leqq n \mid i \in A \in \mathscr{G}_n$, for some $A\}$ converge to 1, then $P_{\varepsilon,n} \to 0$ for each positive ε implies that $m_n \to \infty$. This follows since the average size of a nonempty set in the reduced group is $(m_n + 1)\nu\{i \mid i \in A \in \mathscr{G}_n$, some $A\}/2m_n$. Recall that we take $m_n = \nu(\mathscr{G}_n)$.

LEMMA 4. *Let* $\bar{Y}_n$ *denote* S_n/n. *Let* t *be fixed and let* $P_{\varepsilon,n} \to 0$ *for all* $\varepsilon > 0$, *then*

$$m_n^{-1}[\textstyle\sum_{A \in \mathscr{G}_n} I_{\{S_A \nu^{-1}(A) - \bar{Y}_n \leqq t n^{\frac{1}{2}}\}} - I_{\{S_A - S_{A^c} \leqq t n^{\frac{1}{2}}\}}] \to_P 0 .$$

PROOF. We show the quantity in question converges in L^1. We employ the same notation as in the previous lemma.

Let a_n be the same decreasing sequence as in the proof of Lemma 1. Choose $\varepsilon < \frac{1}{8}$ and positive. Let $\mathscr{Q}_\varepsilon$ be as in Lemma 3. Then, for A_n write $p = \nu(A)$ and $q = \nu(A^c)$ so that

$$\begin{aligned} J(A) &= |E(I_{\{S_A n^{\frac{1}{2}} \nu^{-1}(A) - n^{\frac{1}{2}} \bar{Y}_n \leqq t\}} - I_{\{S_A - S_{A^c} \leqq t n^{\frac{1}{2}}\}})| \\ &= |E(I_{\{S_A n^{\frac{1}{2}} p^{-1} - n - \frac{1}{2}(S_A + S_{A^c}) \leqq t < (S_A - S_{A^c}) n^{-\frac{1}{2}}\}} \\ &\qquad + I_{\{S_A n^{\frac{1}{2}} p^{-1} - n - \frac{1}{2}(S_A + S_{A^c}) > t \geqq (S_A - S_{A^c}) n^{-\frac{1}{2}})\}})| \\ &\leqq 2[2a_q + E|\Phi((n/q)^{\frac{1}{2}}[-t - (1 - n/p)S_A n^{-\frac{1}{2}}]) \\ &\qquad - \Phi((n/q)^{\frac{1}{2}}[-t + S_A n^{-\frac{1}{2}}])|] \leqq 4a_{n/4} + B_1\varepsilon \end{aligned}$$

where B_1 is a constant independent of the choice of ε and n.

Hence,

$$E|m^{-1}\sum_{A\in\mathscr{G}} I_{\{S_A n^{\frac{1}{2}}\nu(A)^{-1}-n^{\frac{1}{2}}\bar{Y}_n\leqq t\}} - I_{\{S_A-S_{A^c}\leqq tn^{\frac{1}{2}}\}}|$$
$$\leqq m^{-1}(\sum_{A\in\mathscr{G}\setminus\mathscr{C}_\varepsilon} 2 + \sum_{A\in\mathscr{C}_\varepsilon} J(A))$$
$$\leqq m^{-1}(2mP_\varepsilon + m[4a_{n/4} + B_1\varepsilon]) .$$

We therefore have L^1 convergence to 0.

Theorem 1 now follows from Lemmas 3 and 4. Uniformity of convergence is immediate because the limit cdf is continuous. Since the uniform convergence of the subsample empirical cdf for subsample means of unknown variance yields consistent estimators for two quantiles of the normal distribution with the appropriate variance, we immediately obtain the

COROLLARY. *Under the conditions of Theorem* 1, *subsampling has Pitman efficiency* 1 *with respect to the t procedure.*

The conditions $P_{\varepsilon,n} \to 0$ and $m_n \to \infty$ are not so strong as may appear at first glance. The following are two equivalent formulations of a sufficient condition for the above to hold for all positive ε:

(1) Let $\mathscr{G}_n$ be a group on the indices $\{1, \cdots, n\}$. Then if $1 \leqq i < j \leqq n$ implies there exist sets A, B in $\mathscr{G}_n$ for which $i \in A$, $j \notin A$, $j \in B$, and $i \notin B$, we then have $P_{\varepsilon,n} \to 0$ for all $\varepsilon > 0$, and $m_n \to \infty$.

(2) Let R_n be the $m_n \times n$ incidence matrix of 0's and 1's where $(R_n)_{ij} = I_{\{J\in A_i\}}$. Then $m_n \to \infty$ and $P_{\varepsilon,n} \to 0$ for each $\varepsilon > 0$ if R_n is rank n.

Note in particular (1) and (2) are indeed satisfied for $\mathscr{G}_n$ the power set of $\{1, \cdots, n\}$. Assertion (2) may be verified by calculating trace $(R_n - \frac{1}{2}J)^T(R_n - \frac{1}{2}J)$. Condition (1) may then be shown to be equivalent to (2) by means of the group structure of $\mathscr{G}_n$. We call condition (1) complete separation because the subsets A and B separate indices i and j. Some consequences of complete separation are studied in Gordon (1971).

3. Monte Carlo simulation. The following table was constructed to examine the behavior of subsampling intervals in the small sample case. For each of the distributions tabled, a set of $2^l - 1$ observations were taken and transformed using the completely separating reduced group corresponding to a saturated Resolution III fractional factorial design. (See Box and Hunter (1961), and Section 5, below.) Size, therefore, corresponds to both the subsample size and the size of the reduced group used in subsampling. These particular reduced groups were chosen since they are related to a commonly used experimental design. This is discussed in Section 5. Also, these reduced groups are shown in Gordon (1971) to minimize the relative variance criterion of Hartigan (1969) among all completely separating reduced groups on $2^l - 1$ indices.

The simulation was performed on the ACME facility at Stanford University. The in-house pseudo-random number generator was used to obtain samples from the various distributions listed and the subsampling method was repeatedly applied to the random samples thereby obtained. Sampling was continued until a

TABLE 1

Expected confidence interval length for symmetric distributions tabled in units $\sigma/n^{\frac{1}{2}}$

Distribution	Confidence	Size ($n = m$)			Kurtosis
		7	15	31	
Laplace	50%	1.42 ± .03	1.40 ± .02	1.40 ± .02	3
	75%	2.38 ± .10	2.38 ± .05	2.38 ± .08	
Triangular	50%	1.38 ± .02	1.38 ± .01	1.38 ± .01	−.60
	75%	2.57 ± .10	2.42 ± .05	2.38 ± .07	
Truncated	50%	1.43 ± .02	1.37 ± .02	1.36 ± .01	−.54
Cauchy (2)	75%	2.59 ± .10	2.44 ± .09	2.35 ± .05	
Truncated	50%	1.42 ± .03	1.38 ± .02	1.36 ± .01	−.046
Cauchy (3)	75%	2.53 ± .09	2.42 ± .08	2.35 ± .05	
Truncated	50%	1.44 ± .02	1.42 ± .01	1.37 ± .01	2.53
Cauchy (8)	75%	2.33 ± .05	2.41 ± .07	2.35 ± .05	
Truncated	50%	1.09 ± .03	1.30 ± .02	1.39 ± .01	50.7
Cauchy (100)	75%	1.42 ± .03	1.80 ± .02	2.05 ± .02	
Truncated	50%	.64 ± .03	.85 ± .03	1.03 ± .05	522
Cauchy (1000)	75%	.75 ± .02	1.03 ± .01	1.28 ± .03	
Normal	50%	1.41	1.38	1.37	
	75%	2.56	2.42	2.36	0
Normal Limit	50%		1.35		
	75%		2.30		

reasonably small standard error for the expected lengths in question was computed.

To facilitate comparisons, the expected lengths tabled are for symmetric 50% and 75% confidence intervals, normalized by multiplication by the square root of the sample size and are in units corresponding to the standard deviation of the underlying distribution. The error terms in the table correspond to two standard errors for the estimate of the expected length.

The Laplace distribution is also called the double exponential. The notation Truncated Cauchy (T) indicates that the distribution was truncated at $\pm$T. The last row indicates the limit of the normalized interval lengths as the size approaches infinity.

The table gives evidence that the normal approximation tends to err in the conservative direction in small samples. The shortness of the intervals in extremely long-tailed cases suggests that the subsampling procedure may be less sensitive to long tails than the t-test. This is indeed the case, as is indicated in the next sections.

4. Consistency. This section is intended to justify the assertion that subsampling is less sensitive to long tails than is t. In particular, since sums of i.i.d. variates are averaged to obtain the subsample means, we use the subsample empirical cdf and the laws of large numbers to obtain the consistency of the subsampling procedure when the second moment of the original variates may

not be finite. The use of the weak law of large numbers here parallels the use of the central limit theorem in proving the efficiency result of Section 2.

Recall that when the underlying cdf is continuous and symmetric, subsampling is exact. Therefore, when the variates are symmetric, one can employ subsampling instead of t to gain consistency in the case of moderately long tails, while sacrificing neither exactness nor efficiency when the variance is finite.

Throughout this section F represents a continuous cdf symmetric about 0. The random variables Y_i are distributed independently as F. As before, $S_A = \sum \{Y_i \mid i \in A\}$ and we write S_n for $\sum_1^n Y_i$. We say F satisfies the weak law of large numbers (WLLN) if $S_n/n \to 0$ in probability.

We continue with the conventions that $\mathscr{G}_n$ is a sequence of reduced groups on $\{1, \cdots, n\}$ for which $\nu(\mathscr{G}_n) = m_n$. Recall that $P_{\varepsilon,n} = m_n^{-1}\nu\{A \in \mathscr{G}_n \mid |\nu(A) - n/2| > \varepsilon n\}$.

LEMMA 5. *If F satisfies* WLLN *and $P_{\varepsilon_0,n} \to 0$ for some $\varepsilon_0 < \frac{1}{2}$, then*

$$m_n^{-1} \sum_{A \in \mathscr{G}_n} I_{\{S_A \leqq t\nu(A)\}} \to_P 0 \qquad \text{for} \quad t < 0 . \tag{1}$$

and

$$m_n^{-1} \sum_{A \in \mathscr{G}_n} I_{\{S_A \geqq t\nu(A)\}} \to_P 0 \qquad \text{for} \quad t > 0 . \tag{2}$$

PROOF. Since the random variables in (1) and (2) are nonnegative, convergence of the expectations to 0 establishes L^1 convergence. For assertion (1),

$$Em_n^{-1} \sum I_{\{S_A \leqq t\nu(A)\}} \leqq P_{\varepsilon_0,n} + \sup_{j, n(\frac{1}{2}-\varepsilon_0)} P\{S_j \leqq jt\} .$$

The second term on the right is $o(1)$ since F satisfies WLLN. Assertion (2) is proved in an identical manner.

The preceding lemma shows that, under the hypotheses, the subsampling procedure is consistent.

The following two lemmas indicate that, under somewhat stronger hypotheses, when n is large, S_n/n is usually included in any subsampling confidence interval which contains the median subsample average.

LEMMA 6. *Let F satisfy* WLLN. *Let $\mathscr{G}_n$ be a sequence of completely separating reduced groups on $\{1, \cdots, n\}$ and $m_n = \nu(\mathscr{G}_n)$. It follows that*

$$m_n^{-1} \sum_{A \in \mathscr{G}_n} I_{\{S_A/\nu(A) < S_n/n\}} - I_{\{S_A - S_{A^c} < 0\}} \to_P 0 .$$

PROOF. From the remarks concluding Section 2, we obtain by Chebyshev's inequality that $P_{\varepsilon,n} \leqq 1/n\varepsilon^2$.

Write $J(A) = E|I_{\{S_A/\nu(A) < S_n/n\}} - I_{\{S_A - S_{A^c} < 0\}}|$. Note that $S_A - S_{A^c} = (2S_A - S_n)/n$. Then $|\frac{1}{2} - \nu(A)/n| < kn^{-\frac{1}{2}}$ implies that, for large n, $J(A) < P\{n^{-\frac{1}{2}}|S_A - S_{A^c}| < 5kn^{-1}|S_A|\}$. By WLLN, the latter goes to 0 uniformly in n. Also, for $\varepsilon(n) = kn^{-\frac{1}{2}}$ we have $P_{\varepsilon(n),n} < k^{-2}$. The proof is completed by choosing k large and is hereafter similar to the proof of Lemma 2.4.

LEMMA 7. *Let $\mathscr{G}_n$ be as in Lemma* 6, *then* $m_n^{-1} \sum_{A \in \mathscr{G}_n} I_{\{S_A - S_{A^c} < 0\}} \to_P \frac{1}{2}$.

PROOF. Given a random variable X, let $Q(X; l) = \sup_\xi P\{X \in [\xi, \xi + l]\}$. $Q(X; \cdot)$ is called the Lévy concentration function of X (e.g., see LeCam (1963)).

Let A, B be elements of $\mathscr{G}_n$ and write $J(A, B) = P\{S_{A-}S_{A^c} < 0 \text{ and } S_{B-}S_{B^c} < 0\}$. As in Lemma 2.3, $J(A, B) = P\{|S_p^{(1)}| - S_q^{(2)} < 0\}$ where $p + q = n$, and $S_p^{(1)}$, $S_q^{(2)}$, $\cdots$ indicate independent partial sums. We do only the case $p < q$. The case $p \geqq q$ is similar. Let $r = q - p$, so that

$$P\{|S_p^{(1)}| < S_p^{(2)}\} - P\{S_p^{(2)} + S_r^{(3)} \leqq |S_p^1| < S_p^{(2)}\} \leqq P\{|S_p^{(1)}| < S_q^{(2)}\}$$
$$\leqq P\{|S_p^{(1)}| < S_p^{(2)}\} + P\{S_p^{(2)} \leqq |S_p^{(1)}| \leqq S_p^{(2)} + S_r^{(3)}\}.$$

So, by symmetry, $|J(A, B) - \frac{1}{4}| \leqq EQ(S_p^{(1)}; |S_r^{(3)}|)$.

If $kr \leqq p$ for some integer k, then by the inequality due to Kolmogorov in LeCam (1963), $|J(A, B) - \frac{1}{4}| \leqq 8E(k\bar{G}(|S_r|))^{-\frac{1}{2}}$ where $\bar{G}(t) = P\{S_r > t\}$. Hence $|J(A, B) - \frac{1}{4}| \leqq 16k^{-\frac{1}{2}}$. The proof is completed in the usual manner by showing L^2 convergence.

Lemmas 5, 6, 7 imply that a symmetric subsampling confidence interval based on completely separating reduced groups eventually is short and contains the grand mean of all observations with large probability. Since subsampling is exact for continuous symmetric parent distributions F, Theorem 2 follows:

THEOREM 2. *If F is continuous, symmetric, and satisfies* WLLN *and if the reduced groups $\mathscr{G}_n$ are completely separating, then a symmetric subsampling confidence procedure is consistent and the probability that S_n/n is contained in the interval converges to* 1 *as* $n \to \infty$.

In the sense of consistency, then, subsampling may be made less sensitive to moderately long tails than t.

5. Fine behavior of the empirical subsample CDF. The normalized and centered subsample cdf

$$H_n(t) = m_n^{-1} \sum I_{\{S_{A^\nu(A)-1}-\bar{Y}_n \leqq tn^{-\frac{1}{2}}\}}$$

plays a crucial role in the preceding discussion. In particular, $H_n(t)$ behaves as if it were the empirical cdf of a collection of m_n independent standard normal variates. The question of the quality of the approximation naturally arises.

The asymptotic variance, $\lim m_n E[H_n(t) - \Phi(t)]^2$, of the random variables $H_n(t)$ is therefore of interest. This limiting variance is strongly dependent on the structure of the group in question. The assertion is supported by computing the asymptotic variances associated with two very similar reduced groups. One group consists entirely of half samples; in the second, all but one component subset are half samples.

We draw heavily in this section on Mallows (1969) which considers the related problem of transformations of i.i.d. observations by orthogonal linear transformations.

We describe the reduced groups in question by a matrix representation. Corresponding to a reduced group $\mathscr{G}$ is its 0 — 1 incidence matrix R. The rows of

R correspond to nonempty sets in $\mathscr{G}$; a 1 appears in column j if the index j appears in the set corresponding to the row in question. We use three sequences of incidence matrices.

Define $R_1^* = (1)$. We define inductively the $2^l - 1 \times 2^l - 1$ matrices R_l^* for $l \geqq 2$.

In particular,

$$R_{l+1}^* = \begin{pmatrix} R_l^* & R_l^* & 0 \\ R_l^* & J - R_l^* & e \\ 0^T & e^T & 1 \end{pmatrix}$$

where J is a square matrix of all 1's and e is a column vector of all 1's. Observe that the $2^l \times 2^l - 1$ matrix $(0\ R_l^*)^T$ is the design matrix of a saturated resolution III fractional design (e.g., see Box and Hunter (1961)). Also, R_3^*, R_4^*, and R_5^* are the incidence matrices used in the Monte Carlo study of Section 3.

We now construct the two reduced groups which will be examined. Define $\hat{R}_{2l} = (R_l^*\ 0)$ and

$$R_{2l} = \begin{pmatrix} R_l^* & 0 \\ e^T & 1 \\ J - R_l^* & e \end{pmatrix}.$$

Observe that $(0, R_{2l}^T)^T$ is the $2^{l+1} \times 2^l$ incidence matrix of a saturated resolution IV fractional factorial design.

All the rows of $\hat{R}_{2l}$ have 2^{l-1} 1's and 2^{l-1} 0's. Also, all but one row of R_{2l} have 2^{l-1} 1's and 2^{l-1} 0's; the remaining row possesses all 1's. Hence, if $P_{\varepsilon,2l}$ and $\hat{P}_{\varepsilon,2l}$ correspond to the reduced groups $\mathscr{G}_{2l}$ and $\hat{\mathscr{G}}_{2l}$ having incidence matrices R_{2l} and $\hat{R}_{2l}$, then $P_{\varepsilon,2l} \to 0$ and $\hat{P}_{\varepsilon,2l} \to 0$ for all $\varepsilon > 0$. Also, the very simple structure of the cardinalities in the two reduced groups makes a calculation of the limiting variances of $H_{2l}(t)$ and $\hat{H}_{2l}(t)$ possible.

For example, for $l = 2$,

$$\hat{R}_4 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \quad \text{and} \quad R_4 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Note that R_4 is $\hat{R}_4$ "reflected" in the middle row of 1's. For the remainder of this section, F denotes a cdf symmetric about 0, with variance 1, finite fourth moment μ_4, and finite sixth moment. F is assumed to have a continuous bounded density function f, and $F^{(n)}$ then has density $f^{(n)}$, where $F^{(n)}$ is the distribution function of $S_n/n^{\frac{1}{2}}$. We take $Y_1, Y_2, \cdots$ to be a sequence of independent random variables distributed as F. S_n stands for a random variable distributed as the nth partial sum of the Y's. If several independent partial sums are needed, they

are denoted $S_p^{(1)}$, $S_p^{(2)}$, and so forth. We write $\kappa = (\mu_4 - 3)/24$, for the normalized kurtosis of F.

We define $\delta_n(x) = n(f^{(n)}(x) - \Phi(x))$ and $\Delta_n(x) = n(F^{(n)}(x) - \Phi(x))$. We also write $\delta(x) = \kappa\phi^{\text{iv}}(x)$ and $\Delta(x) = \kappa\phi^{\text{iii}}(x)$, where ϕ is the standard normal density and $\phi^{\text{i}}, \phi^{\text{ii}}, \cdots$ are its derivatives. Use of the Edgeworth expansion (e.g. see Feller (1966), Chapter XVI) yields $\Delta_n(x) \to \Delta(x)$. The following lemma is fundamental to the derivation of the asymptotic variances. It, and (1) of Theorem 3 below are essentially Mallows' (1969) Lemma 4.1.

LEMMA 5.

$$\lim_{n\to\infty} nE[I_{\{S_n^{(1)}+S_n^{(2)}\le t(2n)^{\frac{1}{2}}\}} I_{\{S_n^{(1)}-S_n^{(2)}\le t(2n)^{\frac{1}{2}}\}} - \phi^2(t)] \doteq \Delta(t)\Phi(t) + 3\kappa[\phi^{\text{i}}(t)]^2 .$$

Theorem 3 provides an evaluation of the limiting variance of the subsample empirical cdf. The two different variances obtained suggest that a general theorem about fine behavior of the subsample empirical cdf may be difficult to formulate. Note that the latter collection of subsample means is symmetric about the grand mean. This symmetry decreases the variance close to zero, but increases it far from zero. Theorem 3 may also be derived without assumptions on the sixth moment.

THEOREM 3. *Let* $\hat{m}_{2l} = \nu(\hat{\mathscr{G}}_{2l})$ *and* $m_{2l} = \nu(\mathscr{G}_{2l})$. *Then*

1) $$\lim \hat{m}_{2l} E[\hat{H}_{2l}(t) - \Phi(t)]^2 = \Phi(t)(1 - \Phi(t)) + 6\kappa[\phi^{\text{i}}(t)]^2$$

and

2) $$\lim m_{2l} E[H_{2l}(t) - \Phi(t)]^2 = \Phi(t)(1 - 2\Phi(t)) + [\Phi(t) - \Phi(-t)]^+ + 12\kappa[\phi^{\text{i}}(t)]^2 .$$

PROOF. 1) Write $n = 2^{l-1}$ and let the expectation on the left of the equality in the statement of the theorem be denoted D_n. Then

$$D_n = (2n - 1)^{-1} \sum \sum E[I_{\{S_A - S_{A^c} \le t(2n)^{\frac{1}{2}}\}} - \Phi(t)][I_{\{S_B - S_{B^c} \le t(2n)^{\frac{1}{2}}\}} - \Phi(t)]$$

where set complementation is relative to $\{1, \cdots, 2n\}$.

If $A \neq B$ then $S_A - S_{A^c} = S_{AB} + S_{AB^c} - S_{A^cB} - S_{A^cB^c}$ and $S_B - S_{B^c}$ may be similarly decomposed. We thereby obtain the simultaneous representation $S_A - S_{A^c} = S_n^{(1)} + S_n^{(2)}$ and $S_B - S_{B^c} = S_n^{(1)} - S_n^{(2)}$. Therefore

$$D_n = \Phi(t)(1 - \Phi(t)) - 2\Delta(t)\Phi(t) + 2nE[I_{\{S_n^{(1)}+S_n^{(2)}\le t(2n)^{\frac{1}{2}}\}} - \Phi(t)][I_{\{S_n^{(1)}-S_n^{(2)}\le t(2n)^{\frac{1}{2}}\}} - \Phi(t)] + o(1) .$$

From Lemma 6,

$$D_n = \Phi(t)1 - \Phi(t)) - 2\Delta(t)\Phi(t) + 2\Delta(t)\Phi(t) + 6\kappa[\phi^{\text{i}}(t)]^2 + o(1) .$$

2) Let $n = 2^{l-1}$ and D_n again correspond to the quantity whose limit is to be taken, found on the left side of the equality in the statement. Let $N = \{1, 2, \cdots, 2n\}$ and observe that $N \in \mathscr{G}_{2n}$. As before, we may write

$$D_n = (4n - 1)^{-1} E \sum \sum [I_{\{S_A \nu(A)^{-1} - \bar{Y}_{2n} \le t/(2n)^{\frac{1}{2}}\}} - \Phi(t)][I_{\{S_B \nu(B)^{-1} - \bar{Y}_{2n} \le t/(2n)^{\frac{1}{2}}\}} - \Phi(t)] .$$

There are four types of pairs (A, B) possible:

(1) (A, A), $4n - 1$ pairs,
(2) (N, A) or (A, N), $A \neq N$, $2(4n - 2)$ pairs,
(3) (A, B), $A \neq N \neq B \neq A \neq B^c$, $(4n - 2)(4n - 4)$ pairs,
(4) (A, A^c), $A \neq N$, $4n - 2$ pairs.

These classes make the following contributions:

(1) $\Phi(t)(1 - \Phi(t)) + o(1)$,
(2) $o(1)$,
(3) by Lemma 5,

$$4nE[I_{\{S_n^{(1)}+S_n^{(2)} \leq t(2n)^{\frac{1}{2}}\}} - \Phi(t)][I_{\{S_n^{(1)}-S_n^{(2)} \leq t(2n)^{\frac{1}{2}}\}} - \Phi(t)] + o(1) = 12\kappa[\phi^i(t)]^2 + o(1).$$

(4) $$E[I_{\{S_A - S_{A^c} \leq t(2n)^{\frac{1}{2}}\}} - \Phi(t)][I_{\{S_{A^c} - S_A \leq t(2n)^{\frac{1}{2}}\}} - \Phi(t)] + o(1).$$

The latter quantity equals $(\Phi(t) - \Phi(-t))^+ - \Phi^2(t)$. Adding the contributions yields the result.

Acknowledgment. This paper is based on a portion of a dissertation prepared under the direction of Bradley Efron at Stanford University.

REFERENCES

[1] BOX, G. E. P. and HUNTER, J. S. (1961). The 2^{k-p} fractional factorial designs. *Technometrics* **3** 311-351.
[2] FELLER, W. (1966). *An Introduction to Probability Theory and its Applications*, **2**. Wiley, New York.
[3] FISHER, R. A. (1935). *The Design of Experiments*, (8th ed.). Oliver and Boyd, London.
[4] FORSYTHE, A. and HARTIGAN, J. A. (1970). Efficiency of confidence intervals generated by repeated subsample calculations. *Biometrika* **57** 629-640.
[5] GORDON, L. (1971). Combinatorial problems in subsampling. Stanford Univ. Technical Report.
[6] HÁJEK, J. and ŠIDÁK, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.
[7] HARTIGAN, J. A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64** 1303-1317.
[8] HARTIGAN, J. A. (1970). Exact confidence intervals in regression problems with independent symmetric errors. *Ann. Math. Statist.* **41** 1992-1998.
[9] LECAM, L. (1965). On the distribution of sums of independent random variables. *Bernoulli, Bayes, Laplace*, (J. Neyman and L. LeCam, eds.). Springer, New York 179-202.
[10] MALLOWS, C. L. (1969). Joint normality induced by orthogonal transformations. Bell Telephone Laboratories Memorandum.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305