The Annals of Statistics
1975, Vol. 3, No. 3, 573-580

NECESSARY AND SUFFICIENT CONDITIONS FOR ASYMPTOTIC JOINT NORMALITY OF A STATISTIC AND ITS SUBSAMPLE VALUES

BY J. A. HARTIGAN

Yale University

1. Introduction. If $X_1, X_2, \cdots, X_n$ are independent identically distributed random variables with mean μ and variance σ^2, then the mean $\bar{X} = \sum_{j=1}^{n} X_j/n$ is asymptotically normal with mean μ and variance σ^2/n. Surprisingly, asymptotic normality also holds for such diverse statistics as order statistics, correlations, maximum likelihood estimates and Bayes estimates, and eigenvalues. This paper gives necessary and sufficient conditions that sequences of subsample values of a statistic be asymptotically joint normal. Also, the following generalization of the central limit theorem is proved:

Let $t_n(X_1, \cdots, X_n)$ be a sequence of symmetric measurable functions in $X_1, \cdots, X_n$, and suppose $n \operatorname{Cov}(t_n, t_m) \to \sigma^2$ whenever $n \geqq m \to \infty$. Then $n^{\frac{1}{2}}(t_n - Et_n)$, $m^{\frac{1}{2}}(t_m - Et_m)$ are asymptotically joint normal with variances σ^2 and correlation ρ whenever $m, n \to \infty$, $m/n \to \rho^2$, $0 \leqq \rho^2 \leqq 1$. The mean satisfies the conditions of the theorem since $n \operatorname{Cov}(\bar{X}_n, \bar{X}_m) = \sigma^2$ exactly.

The property of the mean which compels the normal limit is

$$n^{\frac{1}{2}}\bar{X}_{1,n} = (m/n)^{\frac{1}{2}}m^{\frac{1}{2}}\bar{X}_{1,m} + [(n - m)/n]^{\frac{1}{2}}(n - m)^{\frac{1}{2}}\bar{X}_{m+1,n},$$

where $\bar{X}_{r,s}$ denotes the mean of $X_r, X_{r+1}, \cdots, X_s$. Thus if the mean is to have a limiting distribution G after standardization, and if Y_1 and Y_2 are independently distributed as G, then $\alpha_1 Y_1 + \alpha_2 Y_2$ must have the distribution G after standardization. Of course this property defines the stable laws, of which only the normal has finite variance.

Generalizing this, a "mean-like" sequence of statistics t_n satisfies $n^{\frac{1}{2}}[t_{1,n} - (m/n)t_{1,m} - (n - m)/nt_{m+1,n}] \to 0$ as $m, n - m, n \to \infty$. This condition will ensure that t_n is asymptotically normal if $n^{\frac{1}{2}}(t_n - a_n)$ converges to a distribution with finite variance. The mean-like property is implied for $t_n - Et_n$ by the above condition $n \operatorname{Cov}(t_n, t_m) \to \sigma^2$ as $n \geqq m \to \infty$, provided that t_n is a symmetric function of the observations. To handle asymmetric functions, it is necessary to consider behavior of the statistic as a function of subsets $X_{i_1}, X_{i_2}, \cdots, X_{i_n}$. These subsets appear in the three conditions for centrality of a statistic, which are necessary and sufficient for joint asymptotic normality of sequences of statistics defined on the subsets, proved in Theorem 1, Section 2.

Theorem 3 presents a simpler set of sufficient conditions when nt_n^2 is uniformly integrable, and Theorem 4 is the generalization of the central limit theorem mentioned above.

Received November 1973; revised October 1974.

In Section 3, the jackknife and subsample techniques for setting confidence intervals are shown to be asymptotically valid for central statistics.

2. Central statistics. Assume a sequence of independent and identically distributed random variables $X_1, X_2, \cdots, X_n$. The distribution of X_1 will be denoted by F, and the joint distribution of $X_1, X_2, \cdots, X_n, \cdots$ will be denoted by P. Let ω denote a realization of $X_1, X_2, \cdots, X_n, \cdots$. A *statistic* t takes the value $t(n, \omega)$ for each ω, and each $n \geqq 1$, and for each n, $t(n, \cdot)$ is a measurable function f_n of $X_1, X_2, \cdots, X_n$,

$$t(n, \omega) = f_n[X_1(\omega), X_2(\omega), \cdots, X_n(\omega)] .$$

If S is the subset of integers $\{i_1, i_2, \cdots, i_n\}$, define

$$t(S, \omega) = f_n[X_{i_1}(\omega), \cdots, X_{i_n}(\omega)] .$$

The random variable taking the value $t(S, \omega)$ at ω will be denoted by $t(S)$. The random variable $t(S)$ for $S = \{1, 2, \cdots, n\}$ is denoted by t_n. The number of integers in S is denoted by $|S|$. The notation $|S_n|/n \to \rho^2$ will denote a sequence of subsets S_n of $\{1, 2, \cdots, n\}$ with $|S_n| \to \infty$, $|S_n|/n \to \rho^2$.

A statistic t is *central for F with variance σ^2*, if

I. $\lim_{A\to\infty} \limsup_{n\to\infty} A^2 P(|t_n| \geqq A) = 0$

II. $\lim_{A\to\infty} \limsup_{n\to\infty} |A \int_{|t_n|<A} t_n \, dP| = 0$

III. $\lim_{A\to\infty} \limsup_{|S_n|/n\to\rho^2} |\int_{|t_n|<A, |t(S_n)|<A} t_n t(S_n) \, dP - \rho\sigma^2| = 0 .$

There is a straightforward extension of this definition to vector-valued statistics with $|t|$ equal to the sup of the components of t, and the product $t_n t(S_n)$ replaced by $t_n t(S_n)'$. The following theorems extend similarly.

THEOREM 1. *The sequences t_n and $t(S_n)$ are asymptotically joint normal with means $(0, 0)$ and variances (σ^2, σ^2) and covariance $\rho\sigma^2$ whenever $|S_n|/n \to \rho^2$, if and only if t is central for F with variance σ^2.*

The statistic t in the theorem has been already standardized, like $n^{\frac{1}{2}}(\bar{X} - \mu)$ for the mean of n observations. The complication of defining t on all subsets of observations is necessary to cope with statistics t which are not symmetrical functions of the observations. If t is known to be symmetrical, it is necessary to consider only subsets $\{1, 2, \cdots, N\}$, $1 \leqq N \leqq \infty$, in stating the conditions and the theorem.

The three centrality conditions are closely related to the following three conditions for asymptotic normality of sums (see Loève (1963), page 316, for method of proof) which will be used in proving the theorem.

LEMMA. *Let $\{X_{nj}, j = 1, \cdots, n\}$ be independent and identically distributed. Then $\sum_{j=1}^{n} X_{nj}$ is asymptotically unit normal* (AUN) *if and only if for each $\varepsilon > 0$ small enough*

(1) $nP(|X_{nj}| > \varepsilon) \to 0$,

(2) $n \int_{|X_{nj}|<\varepsilon} X_{nj} \, dP \to 0$,

(3) $n \int_{|X_{nj}|<\varepsilon} X_{nj}^2 \, dP \to 1$, *as* $n \to \infty$.

PROOF OF THEOREM 1. If $\sigma = 0$, conditions I, II, III are equivalent to $t_n \to 0$ in probability as $n \to \infty$. Also asymptotic joint normality of t_n and $t(S_n)$ with variances (0, 0) is equivalent to $t_n \to 0$ in probability as $n \to \infty$, so the theorem is proved in this trivial case. If $\sigma \neq 0$, let $\sigma = 1$ without loss of generality.

It will first be shown that asymptotic normality implies centrality.

For I, $\lim_{n\to\infty} A^2P[|t_n| \geqq A] = A^2 \int_{|x|\geqq A} \exp(-\frac{1}{2}x^2)/(2\pi)^{\frac{1}{2}}\, dx \to 0$ as $A \to \infty$.

For II, $\lim_{n\to\infty} A\left|\int_{|t_n|<A} t_n\, dP\right| = A \int_{|x|<A} x \exp(-\frac{1}{2}x^2)/(2\pi)^{\frac{1}{2}}\, dx = 0$ for all A.

For III, $\lim_{|S_n|/n\to\rho^2} \int_{|t|<A} t_n t(S_n)\, dP = \int_{|x|<A,|y|<A} xyf(x, y, \rho)\, dx\, dy$,

where $f(x, y, \rho)$ is the bivariate normal density with correlation ρ. Thus $\lim_{A\to\infty} \int_{|x|<A,|y|<A} xyf(x, y, \rho)\, dx\, dy = \rho$, and condition III is proved.

To show that centrality implies normality, the statistic computed on N observations is approximated by a sum of statistics computed on subsets each containing k observations. Using the lemma, and manipulating simultaneously N, k, and a truncation point A, it is shown that this sum is asymptotically normal.

Define $t^A(S) = t(S)$ if $|t(S)| < A$, $t^A(S) = 0$ if $|t(S)| \geqq A$. Define $U_{jk} = \{(j-1)k + 1, \cdots, jk\}$, $Y_{nk} = \sum_{j=1}^n t(U_{jk})/n^{\frac{1}{2}}$, $Y_{nk}^A = \sum_{j=1}^n t^A(U_{jk})/n^{\frac{1}{2}}$. Define $t_{nk}^A = t^A(1, 2, \cdots, nk)$. Then $\operatorname{Var}(t_{nk}^A - Y_{nk}^A) = \operatorname{Var}[t_{nk}^A] - 2\sum_{j=1}^n \operatorname{Cov}(t^A(U_{jk}), t_{nk}^A)/n^{\frac{1}{2}} + \operatorname{Var}[t^A(U_{1k})]$. As $k \to \infty$, for each finite n, $|U_{jk}|/nk \to 1/n$. From II, III, $\lim_{A\to\infty} \limsup_{k\to\infty} \operatorname{Var}(t_{nk}^A - Y_{nk}^A) = 1 - 2 + 1 = 0$, each fixed n. Thus as $n \to \infty$, there exists $A_n \uparrow \infty$, with $A_n > n^{\frac{1}{2}}$, and $k_n^0 \uparrow \infty$, such that $\operatorname{Var}(t_{nk_n}^{A_n} - Y_{nk_n}^{A_n}) \to 0$ whenever $k_n \geqq k_n^0$. Also from condition I, $P(|t_{nk_n}| \geqq A_n) \to 0$ if $k_n > k_n^1 \uparrow \infty$, and

$$P(|t(U_{jk_n})| \geqq A_n \text{ any } j, 1 \leqq j \leqq n) \leqq nP(|t(U_{1k_n})| \geqq A_n) \leqq A_n^2 P(|t(U_{1k_n})| \geqq A_n) \to 0$$

if $k_n > k_n^2 \uparrow \infty$. From condition II, $E[t_{nk_n}^{A_n}] \to 0$ and $E(Y_{nk_n}^{A_n}) \to 0$ provided $k_n > k_n^3 \uparrow \infty$. Thus $t_{nk_n} - Y_{nk_n} \to 0$ in probability if $k_n > k_n^0, k_n^1, k_n^2, k_n^3$.

It will next be shown that the summands in Y_{nk_n} satisfy the conditions of the lemma, if $k_n \to \infty$ fast enough.

Define $X_{nj} = t(U_{jk_n})/n^{\frac{1}{2}}$. Then

$nP(|X_{nj}| \geqq \varepsilon) \to 0$ from condition I, if $k_n > k_n^4 \uparrow \infty$,

$n\int_{|X_{nj}|<\varepsilon} X_{nj}\, dP \to 0$ from condition II, if $k_n > k_n^5 \uparrow \infty$,

$n\int_{|X_{nj}|<\varepsilon} X_{nj}^2\, dP \to 1$ from condition III, if $k_n > k_n^6 \uparrow \infty$.

Thus t_{nk_n} is AUN whenever $k_n > K_n = \max_{0\leqq i\leqq 6} k_n^i$.

Now consider an arbitrary subsequence $\{m_j, j = 1, \cdots, \infty\}$. Define $j_n = \min\{j \mid m_j \geqq nK_n\}$, $k_n = [m_{j_n}/n]$. As $n \to \infty$, $nk_n/m_{j_n} \to 1$, and by condition III, $t_{nk_n} - t_{m_{j_n}} \to 0$. Thus the sequence $t_{m_{j_n}}$ is AUN. Every subsequence of t_n has a subsequence which is AUN, so t_n is AUN. Consequently $t(S_n)$ is AUN for arbitrary sequences of sets S_n with $|S_n| \to \infty$.

To prove joint asymptotic normality of t_n and $t(S_n)$ as $|S_n|/n \to \rho^2$, define $S_n^* = \{i \mid i \leqq n, i \notin S_n\}$, and note that $|S_n^*|/n \to 1 - \rho^2$. Centrality then implies $t_n - \rho t(S_n) - (1 - \rho^2)^{\frac{1}{2}} t(S_n^*) \to 0$ as $n \to \infty$. Thus t_n is asymptotically distributed as the sum of two independent normal variables one of which is $\rho t(S_n)$. Thus t_n and $t(S_n)$ are asymptotically joint normal with means (0, 0), variances (1, 1) and correlation ρ. □'

THEOREM 2. *For n fixed and every N, let* $\{S_{N1}, \cdots, S_{Nn}\}$ *be subsets of* $\{1, 2, \cdots, N\}$. *Let* $S_N^* = \{i \mid i \leqq N, i \notin S_N\}$. *Suppose that for any partition of* $\{1, \cdots, n\}$ *into two subsets* $\{i_1, \cdots, i_r\}$ *and* $\{j_1, j_2, \cdots, j_s\}$, $|S_{Ni_1} \cap \cdots \cap S_{Ni_r} \cap S^*_{Nj_1} \cap \cdots \cap S^*_{Nj_s}|/N \to \rho^2_{i_1, i_2, \cdots, i_r \cdot j_1, j_2, \cdots, j_s}$ *as* $N \to \infty$. *Assume* $|S_{Ni}|/N \to \rho_i{}^2 > 0$, $1 \leqq i \leqq n$.

If t is central for F with variance σ^2, *then* $t(S_{N1}), \cdots, t(S_{Nn})$ *are jointly asymptotic normal with means* $(0, \cdots, 0)$ *and covariance between* $t(S_{Ni})$ *and* $t(S_{Nj})$ *equal to* $\lim_{N\to\infty} |S_{Ni} \cap S_{Nj}|\sigma^2/\rho_i \rho_j N$.

PROOF. The case $n = 2$ will be proved, and the case for general n is proved analogously. The conditions of the theorem require that $|S_{N1} \cap S^*_{N2}|/N \to \rho^2_{1\cdot 2}$, $|S_{N1} \cap S_{N2}|/N \to \rho^2_{12\cdot}$, $|S^*_{N1} \cap S_{N2}|/N \to \rho^2_{2\cdot 1}$, $|S^*_{N1} \cap S^*_{N2}|/N \to \rho^2_{\cdot 12}$. The four subsets are disjoint, so that $t(S_{N1} \cap S_{N2})$, $t(S_{N1} \cap S^*_{N2})$, $t(S^*_{N1} \cap S_{N2})$ and $t(S^*_{N1} \cap S^*_{N2})$ are independent. Also $t(S_{N1}) - r_1 t(S_{N1} \cap S_{N2}) - r_2 t(S_{N1} \cap S^*_{N2}) \to 0$ as $N \to \infty$ where $r_1 = \rho_{12\cdot}/\rho_1$, $r_2 = \rho_{1\cdot 2}/\rho_1$ using the mean-like property for central statistics, and similarly $t(S_{N2}) - s_1 t(S_{N1} \cap S_{N2}) - s_2 t(S^*_{N1} \cap S_{N2}) \to 0$ where $s_1 = \rho_{12\cdot}/\rho_2$, $s_2 = \rho_{2\cdot 1}/\rho_2$. The random variables $r_1 t(S_{N1} \cap S_{N2})$, $s_1 t(S^*_{N1} \cap S_{N2})$, $r_2 t(S_{N1} \cap S^*_{N2})$, $s_2 t(S^*_{N1} \cap S^*_{N2})$ are joint asymptotic normal. (The case where $|S_{N1} \cap S^*_{N2}| \not\to \infty$ is handled by noting that $r_2 = 0$, so that $r_2 t(S_{N1} \cap S^*_{N2})$ is degenerate.) Thus $t(S_{N1})$ and $t(S_{N2})$ are joint asymptotic normal with correlation $r_1 s_1 = \rho^2_{12\cdot}/\rho_1 \rho_2$. □

THEOREM 3. *Let z be a statistic defined as in Section* 2. *Suppose for any sequence of subsets* S_n *with* $|S_n| \to \infty$, $n \operatorname{Cov}[z_n, z(S_n)] \to \sigma^2$ *as* $n \to \infty$. *Then* $n^{\frac{1}{2}}(z_n - Ez_n)$ *is central with variance* σ^2.

PROOF. For $n \geqq$ some n_0, z_n has finite variance by the condition of the theorem. Define $t_n = n^{\frac{1}{2}}(z_n - Ez_n)$ for $n \geqq n_0$, $t_n = 0$ for $n < n_0$. The main difficulty in proving the theorem is showing that $t_n{}^2$ is uniformly integrable, which means that $\lim_{A\to\infty} \limsup_{n\to\infty} \int_{|t_n|>A} t_n{}^2\, dP = 0$.

Define $Y_{nk} = \sum_{j=1}^n t(U_{jk})/n^{\frac{1}{2}}$ where $U_{jk} = \{(j-1)k + 1, \cdots, jk\}$. Let $nk = N$. Then $\operatorname{Var}(t_N - Y_{nk}) = \operatorname{Var} t_N + \operatorname{Var} t_k - 2\sum_{j=1}^n \operatorname{Cov}[t(U_{jk}), t_N]/n^{\frac{1}{2}} \to \sigma^2 + \sigma^2 - 2\sigma^2 = 0$ as $n, k \to \infty$. It will first be shown that

$$\int_{|Y_{nk}|\geqq A} Y^2_{nk}\, dP \leqq \int_{|t_k|\geqq \frac{1}{2}An^{\frac{1}{2}}} t_k{}^2\, dP + 4\tau^4/A^2 + 8\tau^3/A \qquad \text{where} \quad \tau^2 = \int t_k{}^2\, dP\,.$$

Define $V_{n-1} = t(U_{2k}) + \cdots + t(U_{nk})$, $V_{1n} = t(U_{1k}) + \cdots + t(U_{\frac{1}{2}nk})$, $V_{2n} = t(U_{(\frac{1}{2}n+1)k} + \cdots + t(U_{nk})$, assuming n even.

$$\int_{|Y_{nk}|\geqq A} Y^2_{nk}\, dP = \int_{|t_k+V_{n-1}|\geqq An^{\frac{1}{2}}} t_k{}^2\, dP + \int_{|V_{1n}+V_{2n}|\geqq An^{\frac{1}{2}}} V_{1n}V_{2n}\, dP(n-1)/\tfrac{1}{4}n^2\,,$$

$$\int_{|t_k+V_{n-1}|\geqq An^{\frac{1}{2}}} t_k{}^2\, dP \leqq \int_{|t_k|\geqq \frac{1}{2}An^{\frac{1}{2}}} t_k{}^2\, dP + \int_{|V_{n-1}|\geqq \frac{1}{2}An^{\frac{1}{2}}} t_k{}^2\, dP$$

$$\leqq \int_{|t_k|\geqq \frac{1}{2}An^{\frac{1}{2}}} t_k{}^2\, dP + \tau^2(n-1)\tau^2/(\tfrac{1}{4}A^2n)\,,$$

$$\int_{|V_{1n}+V_{2n}|\geq An^{\frac{1}{2}}} V_{1n}V_{2n}\,dP \leq 2\int_{|V_{1n}|\geq \frac{1}{2}An^{\frac{1}{2}}} |V_{1n}||V_{2n}|\,dP$$
$$\leq 2(\tfrac{1}{2}n\tau^2/\tfrac{1}{2}An^{\frac{1}{2}})(\tfrac{1}{2}n\tau^2)^{\frac{1}{2}}$$
$$\leq 2n\tau^3/A\,.$$

It is now established that

$$\int_{|Y_{nk}|\geq A} Y_{nk}^2\,dP \leq \int_{|t_k|\geq \frac{1}{2}An^{\frac{1}{2}}} t_k^2\,dP + 4\tau^4/A^2 + 8\tau^3/A\,.$$

Now let N_k be such that $\int_{|t_k|\geq \frac{1}{2}AN_k^{\frac{1}{2}}} t_k^2\,dP \to 0$. It will follow that $\lim_{A\to\infty}\lim_{k\to\infty}\int_{|Y_{n_kk}|\geq A} Y_{n_kk}^2\,dP = 0$ whenever $n_k > N_k$. Since $\lim_{k,n\to\infty}\int (t_{nk} - Y_{nk})^2\,dP = 0$, $\lim_{A\to\infty}\lim_{k\to\infty}\int_{|t_{n_kk}|\geq A} t_{n_kk}^2\,dP = 0$. For any subsequence $\{m_j\}$, there is a further subsequence m_{j_k} such that $m_{j_k}/n_k k \to 1$ for some $n_k > N_k$. Therefore $\int (t_{m_{j_k}} - t_{n_kk})^2\,dP \to 0$, and $t_{m_{j_k}}^2$ is uniformly integrable. Since every subsequence has a uniformly integrable subsequence, the sequence t_n^2 is uniformly integrable. Conditions I, II, III now follow immediately from $\int t_n\,dP = 0$, $\int t_n t(S_n)\,dP \to \rho\sigma^2$ as $|S_n|/n \to \rho^2$, proving the theorem. □

THEOREM 4. *For each n, z_n is a measurable symmetric function of $X_1, \cdots, X_n$, where $X_1, \cdots, X_n$ is a random sample from F. If $n\,\mathrm{Cov}\,(z_m, z_n) \to \sigma^2$ as $n \geqq m \to \infty$, then $n^{\frac{1}{2}}(z_n - Ez_n)$ and $m^{\frac{1}{2}}(z_m - Ez_m)$ are asymptotically normal with means* (0, 0), *variances* (σ^2, σ^2), *and correlation* ρ *as* $m/n \to \rho^2$.

PROOF. This is an immediate consequence of Theorem 3. □

Theorem 4 is a generalization of the central limit theorem since $n\,\mathrm{Cov}\,(z_m, z_n) = \sigma^2$ holds exactly when z_n is the mean.

3. Jackknifing and subsampling. Let $C_N = \{S_{N1}, S_{N2}, \cdots, S_{Nn}\}$ denote a family of subsets which partitions the set $I_N = \{1, 2, \cdots, N\}$ for each N. Define *pseudovalues* $t_{Nj} = (N/|S_{Nj}|)t(I_N) - (N/|S_{Nj}| - 1)t(I_N - S_{Nj})$, $1 \leqq j \leqq n$.

The *jackknifed statistic* t^* is defined for each t, $\{C_N\}$ by

$$t_N^* = t^*(I_N) = \textstyle\sum_{j=1}^n t_{Nj}|S_{Nj}|/N\,.$$

The jackknife operation was first advocated by Quenouille (1956) as a means of reducing bias. For each fixed n, if $|S_{Nj}|/N \to \rho_n^2$, $1 \leqq j \leqq n$, it is easily shown that if $E[t(I_N)] = \mu + b/N + O(N^{-2})$, then $E[t^*(I_N)] = \mu + O(N^{-2})$. Tukey (1958) suggested that the pseudovalues are approximately distributed as a sample from a normal distribution with mean μ and some variance σ^2, provided $\rho_1 = \rho_2 = \cdots = \rho_n$. The pseudovalues may thus be used to construct an approximate confidence interval for μ based on Student's distribution. Tukey's suggestion has been examined, and on the whole justified, by Miller (1964), (1968) for variances and other statistics, and by Brillinger (1964) for maximum likelihood estimates. See also Gray, Watkins and Adams (1972).

THEOREM 5. *For each fixed n, and for each sequence of partitions $C_N = \{S_{N1}, \cdots, S_{Nn}\}$, with $|S_{Nj}|/N \to \rho_j^2 \in (0, 1)$, $1 \leqq j \leqq n$, the standardized pseudovalues $\{(|S_{Nj}|)^{\frac{1}{2}}(t_{Nj} - \mu), 1 \leqq j \leqq n\}$ are asymptotically independent normals with means* 0 *and variances* σ^2, *and* $N^{\frac{1}{2}}(t_N - t_N^*) \to 0$ *if and only if* $N^{\frac{1}{2}}(t_N - \mu)$ *is central with variance* σ^2.

PROOF. First assume that $N^{\frac{1}{2}}(t_N - \mu)$ is central with variance σ^2. Let $S^*_{Nj} = I_N - S_{Nj}$. From Theorem 2, $|S^*_{Nj}|^{\frac{1}{2}}[t(S^*_{Nj}) - \mu]$, $1 \leqq j \leqq n$ and $N^{\frac{1}{2}}(t(I_N) - \mu)$ are asymptotically joint normal with means 0, variances σ^2 and correlations

$$\text{corr}\,[t(S^*_{Nj}), t(S^*_{Nk})] = [1 - \rho_j^2 - \rho_k^2]/[(1 - \rho_j^2)(1 - \rho_k^2)]^{\frac{1}{2}}$$
$$\text{corr}\,[t(S^*_{Nj}), t(I_N)] = [1 - \rho_j^2]^{\frac{1}{2}}.$$

From this it follows that the pseudovalues, which are linear functions of $t(S^*_{Nj})$ and $t(I_N)$, have the stated asymptotic behavior. For example, the limiting variance of $(|S_{Nj}|)^{\frac{1}{2}}t_{Nj}$ is $N\rho_j^2\sigma^2[1/N\rho_j^4 - 2(1 - \rho_j^2)/N\rho_j^4 + [1 - \rho_j^2]/N\rho_j^4)]$ which reduces to σ^2.

Also $t_N{}^*$ is an average of t_{Nj}, and so is joint normal asymptotically with t_N. The sequence $N^{\frac{1}{2}}(t_N - t_N{}^*)$ has asymptotic variance 0, since t_N and t_{Nj} have asymptotic covariance σ^2/N.

Conversely, if the standardized pseudovalues are asymptotically joint normal and $N^{\frac{1}{2}}(t_N - t_N{}^*) \to 0$ for each sequence C_N, $N^{\frac{1}{2}}(t_N - \mu)$ and $(1/\rho_j)N^{\frac{1}{2}}(t_N - \mu) - (1/\rho_j{}^2 - 1)^{\frac{1}{2}}(|S^*_{Nj}|)^{\frac{1}{2}}(t(S^*_{Nj}) - \mu)$ are joint normal, and therefore $N^{\frac{1}{2}}(t_N - \mu)$ and $(|S_{Nj}|)^{\frac{1}{2}}(t(S_{Nj}) - \mu)$ are joint normal as $|S_{Nj}|^{\frac{1}{2}}/N \to \rho_j{}^2$, $0 < \rho_j < 1$, with means (0, 0), variances (σ^2, σ^2) and covariance $\rho_j\sigma^2$, asymptotically. Centrality requires joint normality also if $|S_{Nj}| \to \infty$, $|S_{Nj}|/N \to 0$ or 1; only the case $|S_{Nj}|/N \to 0$ will be proved. For each $\rho > 0$, find U_N such that $U_N \supset S_{Nj}$, $|U_N|/N \to \rho^2$. Since

$$N^{\frac{1}{2}}[(t_N - \mu) - \rho^2(t(U_N) - \mu) - (1 - \rho^2)(t(U_N{}^*) - \mu)] \to 0$$

t_N and $t(U_N{}^*)$ have asymptotic correlation $(1 - \rho^2)^{\frac{1}{2}}$. Also $t(S_{Nj})$ and $t(U_N{}^*)$ have asymptotic correlation 0. Therefore $t(S_{Nj})$ and t_N have asymptotic correlation less than ρ. Since this is true for every $\rho > 0$, they have asymptotic correlation 0. □

In Hartigan (1969), it was suggested that error analysis be based on random subsamples $S_{N1}, S_{N2}, \cdots, S_{Nn}$, where S_{Nj} is selected at random from the $2^N - j$ subsets of $\{1, 2, \cdots, N\}$ not equal to φ, $S_{N1}, \cdots, S_{N(j-1)}$. Let $t(S_{N1}), \cdots, t(S_{Nn})$ be the subsample values of the statistic. If the statistic t is the mean, and X_i is symmetrically distributed about μ, then μ is less than exactly k of $t(S_{N1}), \cdots, t(S_{Nn})$ with probability $1/(n + 1)$. This exact result motivates the following asymptotic one.

THEOREM 6. *If $N^{\frac{1}{2}}(t_N - \mu)$ is central with variance σ^2, then $\{N^{\frac{1}{2}}(t(S_{N1}) - \mu), \cdots, N^{\frac{1}{2}}(t(S_{Nn}) - \mu)\}$ are asymptotically joint normal with variances $2\sigma^2$ and covariances σ^2. In consequence μ is less than exactly k of $t(S_{N1}), \cdots, t(S_{Nn})$ with probability approaching $1/(n + 1)$ as $N \to \infty$.*

PROOF. As $N \to \infty$, $|S_{Ni}|/N \to \frac{1}{2}$ and $|S_{Ni} \cap S_{Nj}|/N \to \frac{1}{4}$. (A random selection S_{Ni} from the set of all 2^N subsets is made by assigning each observation to S_{Ni} independently with probability $\frac{1}{2}$. The omission of null and previous subsamples has negligible asymptotic effect.) Theorem 2 now implies asymptotic normality of $\{(|S_N|)^{\frac{1}{2}}(t(S_{Ni}) - \mu), 1 \leqq i \leqq n\}$ with variances σ^2 and covariances $\sigma^2/2$. Thus $N^{\frac{1}{2}}[t(S_{Ni}) - \mu]$ is asymptotically normal with variances $2\sigma^2$ and covariances σ^2.

If $Y_0, Y_1, \cdots, Y_n$ are independent normal with means 0 and variances σ^2, $\{Y_1 + Y_0, \cdots, Y_n + Y_0\}$ is multivariate normal with means 0, variances $2\sigma^2$, and covariances σ^2. The probability that $-Y_0$ is less than exactly k of $Y_1, \cdots, Y_n$ is $1/(n+1)$ by symmetry. The probability that 0 is less than exactly k of $Y_1 + Y_0, \cdots, Y_n + Y_0$ is thus $1/(n+1)$. The probability that μ is less than exactly k of $t(S_{N1}), \cdots, t(S_{Nn})$ is asymptotically $1/(n+1)$. □

The jackknife and subsample techniques are similar in that both use subsets of the observations to determine errors of estimates empirically. The subsample technique is simpler conceptually and may be applied to statistics taking values in arbitrary spaces, such as graphs, verbal conclusions, or ten pages of computer output. A Bayesian interpretation is useful for such general spaces—the subsample values are approximately a random sample from the posterior distribution of the true value. Such an interpretation is valid in R^p provided $N^{\frac{1}{2}}(t_N - \mu)$ is central for μ with variance σ^2, and the prior distribution of μ given σ^2 is positive and continuous at the true value; the subsample values are asymptotically a random sample from the posterior distribution of μ given t_N and σ.

The jackknife removes bias, and the subsample technique does not. But when terms of order $N^{-\frac{1}{2}}$ are considered in the distribution of $N^{\frac{1}{2}}(t_N - \mu)$, it is necessary to examine not only bias, but skewness, and non-linearity of t_N, all of which make contributions of this order to the error in the confidence intervals. These terms have been examined by Norman Johnson in his Ph. D. thesis at Yale University.

It is plausible nevertheless to jackknife the subsample values to obtain $2t_N - t(S_{N1}), 2t_N - t(S_{N2}), \cdots, 2t_N - t(S_{Nn})$ as debiased subsample values. These are treated as normal observations with mean μ, variances σ^2 and covariances $\frac{1}{2}\sigma^2$, from which confidence intervals for μ may be obtained by a modified t procedure.

4. An application to U statistics. A U-statistic defined by

$$t_n(X_1, \cdots, X_n) = \textstyle\sum_r^n f(X_{i_1}, \cdots, X_{i_r})(n-r)!/n!$$

where $\sum_r^n$ denotes summation over all ordered subsets of $X_1, \cdots, X_n$ of size r, for some fixed r. Assume f is symmetric and $Ef^2 < \infty$.

Then

$$\begin{aligned}\operatorname{Cov}(t_n, t_m) &= E[(t_n - Et_n)t_m] \\ &= E[(t_n - Et_n)\textstyle\sum_m^n t(X_{i_1}, \cdots, X_{i_m})(n-m)!/n!] \\ &= E[(t_n - Et_n)\textstyle\sum_r^n f(X_{i_1}, \cdots, X_{i_r})(n-r)!/n!] \\ &= \operatorname{Var} t_n .\end{aligned}$$

Also

$$\operatorname{Var} t_n = \textstyle\sum_{k=1}^r f_k((n-r)!)^2 r!/[n!\, k!\, (r-k)!\, (n-2r+k)!]$$

where $f_k = \operatorname{Cov}[f(X_1, \cdots, X_k, X_{k+1}, \cdots, X_r), f(X_1, \cdots, X_k, X_{r+1}, \cdots, X_{2r-k})]$.

Thus $n \operatorname{Var} t_n \to rf_1$. The conditions of Theorem 3 are satisfied, and $n^{\frac{1}{2}}(t_n - Et_n)$ is asymptotically normal with variance rf_1, as shown by Hoeffding (1948).

5. Acknowledgments. I am indebted to David Freedman for pointing out an error in the initial statement of Theorem 4.

REFERENCES

[1] BRILLINGER, D. R. (1964). The asymptotic behavior of Tukey's general method of setting appropriate confidence limits (the jackknife) when applied to maximum likelihood estimates. *Rev. Inst. Internat. Statist.* **32** 202–206.

[2] GRAY, H. L., WATKINS, T. A. and ADAMS, J. E. (1972). On the jackknife statistic, its extensions, and its relation to e_n transformations. *Ann. Math. Statist.* **43** 1–30.

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

[4] HOEFFDING, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293–325.

[5] LOÈVE, M. (1963). *Probability Theory*. Van Nostrand, Princeton.

[6] MILLER, R. G. (1964). A trustworthy jackknife. *Ann. Math. Statist.* **35** 1594–1605.

[7] MILLER, R. G. (1968). Jackknifing variances. *Ann. Math. Statist.* **39** 567–582.

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

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

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
BOX 2179, YALE STATION
NEW HAVEN, CONNECTICUT 06520