ROBUST PROCEDURES FOR SOME LINEAR MODELS WITH ONE OBSERVATION PER CELL[1]

BY KJELL DOKSUM

University of California, Berkeley, University of Oslo, and Institut de Statistique de l'Université de Paris

1. Introduction and summary. For block designs with one observation per cell, the model often used is the linear model in which the observations $X_{i\alpha}$ $(i = 1, \cdots, r; \alpha = 1, \cdots, n)$ can be written

$$(1.1) \qquad X_{i\alpha} = \nu + \xi_i + \mu_\alpha + Y_{i\alpha}(\sum \xi_i = \sum \mu_\alpha = 0)$$

where the ξ's are the parameters of interest (treatment effect) the μ's are nuisance parameters (block effect), and the Y's are independent with common continuous distribution F.

The purpose of this note is to discuss some new robust test statistics (e.g. 2.14 and 2.16) of the null-hypothesis $H_0 : \xi_1 = \xi_2 = \cdots = \xi_r$, and to discuss a new robust estimate (3.3) of the contrast $\theta = \sum c_i\xi_i$.

2. Testing for absence of main effect. The tests of $H_0 : \xi_1 = \xi_2 = \cdots = \xi_r$ will be based on the quantities U_{ij} defined by

$$(2.1) \qquad \binom{n}{2} U_{ij} = \text{number of pairs} \quad (\alpha, \beta) \quad \text{with} \quad \alpha < \beta$$
$$\text{and} \quad (X_{i\alpha} - X_{j\alpha} + X_{i\beta} - X_{j\beta}) > 0.$$

Let $\lambda(F) = P(Y_{11} < Y_{12} + Y_{13} - Y_{14}$ and $Y_{11} < Y_{15} + Y_{16} - Y_{17})$ and let $\{a_{ij} : i = 1, \cdots, r; j = 1, \cdots, r\}$ be a set of constants, then the results of Hoeffding (1948) on U-statistics yields.

LEMMA 2.1. *Suppose $\xi_i - \xi_j = a_{ij}/n^{\frac{1}{2}}$, then $\{n^{\frac{1}{2}}[U_{ij} - E(U_{ij})] : i < j\}$ has asymptotically the $\frac{1}{2}(r - 1)r$ variate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij,kl})$ given by*

$$(2.2) \qquad \begin{aligned} &\sigma_{ij,ij} = \tfrac{1}{3}; \quad \sigma_{ij,kl} = 0 && \textit{if} \quad i, j, k, l \quad \textit{are distinct};\\ &\sigma_{ij,kl} = [4\lambda(F) - 1] && \textit{if} \quad i = k \quad \textit{or} \quad j = l; \quad \textit{and}\\ &\sigma_{ij,kl} = [1 - 4\lambda(F)] && \textit{if} \quad i = l \quad \textit{or} \quad j = k. \end{aligned}$$

$\lambda(F)$ has been shown by Lehmann (1964) to satisfy

$$(2.3) \qquad \tfrac{1}{4} \leqq \lambda(F) \leqq \tfrac{7}{24}$$

and to have the values .2902, .2909 and .2879 for the normal, uniform and Cauchy distributions, respectively.

The asymptotic mean of U_{ij} is given by the following result in which G denotes the distribution of $Y_{11} - Y_{12}$.

Received 8 November 1965; revised 20 December 1966.

[1] This paper was prepared with the partial support of the General Electric Corporation.

LEMMA 2.2. *Let the density g of G exist and satisfy the regularity condition of Lemma* 3(a) *of* [4], *and let $\xi_i - \xi_j = a_{ij}/n^{\frac{1}{2}}$, then*

$$n^{\frac{1}{2}}[E(U_{ij}) - \tfrac{1}{2}] \rightarrow 2a_{ij} \int_{-\infty}^{\infty} g^2(t)\, dt \quad \text{as} \quad n \rightarrow \infty. \tag{2.3}$$

PROOF. Set $\Delta = \xi_i - \xi_j = a_{ij}/n^{\frac{1}{2}}$, then

$$\begin{aligned} E(U_{ij}) &= P(X_{i\alpha} - X_{j\alpha} + X_{i\beta} - X_{j\beta} > 0) \\ &= P(Y_{i\alpha} - Y_{j\alpha} + Y_{i\beta} - Y_{j\beta} + 2\Delta > 0) \\ &= G(t + 2\Delta)\, dG(t). \end{aligned}$$

It follows that $n^{\frac{1}{2}}[E(U_{ij}) - \frac{1}{2}] = \int (a_{ij}/\Delta)[G(t + 2\Delta) - G(t)]\, dG(t) \rightarrow 2a_{ij} \int_{-\infty}^{\infty} g^2(t)\, dt$ as $\Delta \rightarrow 0 (n \rightarrow \infty)$.

Next the quantities

$$T_{ij} = U_{i\cdot} - U_{j\cdot} = r^{-1} \textstyle\sum_{k=1}^{r} U_{ik} - r^{-1} \sum_{k=1}^{r} U_{jk} \tag{2.4}$$

will be considered. It is clear that the mean of T_{ij} under H_0 is zero while the variance under H_0 can be computed to be

$$V_0(T_{ij}) = \{2n - 1 + (r - 2)[24(n - 2)\lambda(F) + 13 - 6n]\}/3rn(n - 1). \tag{2.5}$$

Similarly, the covariances under H_0 are

$$\begin{aligned} C_0(T_{ij}, T_{kl}) &= 0, && \text{if } i, j, k, l \text{ are distinct,} \\ C_0(T_{ij}, T_{kl}) &= V_0(T_{ij})/2, && \text{if } i = k \text{ or } j = l, \\ C_0(T_{ij}, T_{kl}) &= -V_0(T_{ij})/2, && \text{if } i = l \text{ or } j = k. \end{aligned} \tag{2.6}$$

It is seen from (2.5) that T_{ij} is not distribution-free. However, the next result shows that it can be made asymptotically distribution-free by dividing it by a consistent estimate of $V_0^{\frac{1}{2}}(T_{ij})$. In order to obtain such an estimate; it is enough to replace $\lambda(F)$ in (2.5) with a consistent estimate of $\lambda(F)$. Lehmann (1964) proposed the following unbiased consistent estimate:

$$\begin{aligned} &c\hat{\lambda} = \text{number of sixtuples } (i, j, k, \alpha, \beta, \gamma) \\ &\text{with } i, j, k \text{ distinct;} \quad \alpha, \beta, \gamma \text{ distinct;} \\ &X_{i\alpha} - X_{j\alpha} < X_{i\beta} - X_{j\beta}; \text{ and } X_{i\alpha} - X_{k\alpha} < X_{i\gamma} - X_{k\gamma} \end{aligned} \tag{2.7}$$

where

$$c = n(n - 1)(n - 2)r(r - 1)(r - 2). \tag{2.8}$$

See also Hollander (1966).

Joint asymptotic normality of the T's follows from the fact that the T_{ij} are linear functions of the U_{ij}. From the preceding covariance results, it can thus be concluded that:

LEMMA 2.3. *Suppose $\xi_i - \xi_j = a_{ij}/n^{\frac{1}{2}}$, then $\{n^{\frac{1}{2}}[T_{ij} - E(T_{ij})]: i < j\}$ has asymptotically the $\frac{1}{2}(r - 1)r$ variate normal distribution with zero mean and covariance* *matrix* $\Sigma^* = (\sigma^*_{ij,kl})$ *given by*

$$\begin{aligned} &\sigma^*_{ij,ij} = \{2 + 6(r-2)[4\lambda(F) - 1]\}/3r, && \\ (2.9)\qquad &\sigma^*_{ij,kl} = 0, && \text{if } i, j, k, l \text{ are distinct}, \\ &\sigma^*_{ij,kl} = \sigma^*_{ij,ij}/2, && \text{if } i = k \text{ or } j = l, \\ &\sigma^*_{ij,kl} = -\sigma^*_{ij,ij}/2, && \text{if } i = l \text{ or } j = k. \end{aligned}$$

From Lemma 2.2, one gets

LEMMA 2.4. *If the conditions of Lemma* 2.2 *hold, then*

$$(2.10)\qquad n^{\frac{1}{2}}E(T_{ij}) \rightarrow 2a_{ij}\int_{-\infty}^{\infty} g^2(t)\,dt \quad \text{as} \quad n \rightarrow \infty.$$

The preceding results can now be used to construct asymptotically distribution-free statistics of H_0 by replacing $X_{i\cdot} - X_{j\cdot}$ in the classical procedures by T_{ij} or U_{ij}. Note that the asymptotic theory of T_{ij} and U_{ij} under H_0 is given by the above lemmas (set $a_{ij} = 0$).

Let Pitman asymptotic efficiency be as defined in [4], and suppose that G has a variance $\sigma^2(G)$ and satisfies the regularity condition of Lemma 3(a) of [4], then the above results and the arguments of Hodges and Lehmann (1961) shows that the Pitman asymptotic efficiency (in testing $\xi_i - \xi_j = 0$) of U_{ij} to $X_{i\cdot} - X_{j\cdot}$ is

$$(2.11)\qquad e = 12\sigma^2(G)[\int_{-\infty}^{\infty} g^2(t)\,dt]^2$$

while the Pitman asymptotic efficiency of T_{ij} to $X_{i\cdot} - X_{j\cdot}$ is

$$(2.12)\qquad e' = e[r/\{2 + 6(r-2)[4\lambda(F) - 1]\}].$$

Lehmann (1964) has shown that $e' \geqq e$. However, the table in Section 5 of Hollander (1966) shows that the difference $e' - e$ is very small for normal, uniform and exponential distributions.

Suppose that H_0 is to be tested against the ordered alternative H_1: $\xi_1 < \xi_2 < \cdots < \xi_r$. Classical normal theory statistics for this problem have been considered by Bartholomew (1961), Nüesch (1966), Hogg (1965) and others. These statistics are based on $\{X_{i\cdot} - X_{j\cdot} : i < j\}$. Since the covariance matrix of $\{T_{ij} : i < j\}$ is proportional to that of $\{X_{i\cdot} - X_{j\cdot} : i < j\}$ (see (2.6)), it follows from Lemma 2.3 that if $X_{i\cdot} - X_{j\cdot}$ is replaced by T_{ij} in each of these statistics, then the new statistics will have the same asymptotic null-distributions as the original statistics.

For instance, the statistic

$$(2.13)\qquad \sum_{i<j} (X_{i\cdot} - X_{j\cdot})/\hat{\sigma}_X,$$

where $\hat{\sigma}_X$ is the appropriate estimate of the standard deviation of $\sum_{i<j}(X_{i\cdot} - X_{j\cdot})$, has been considered in [1], [13] and [7]. The statistic based on $\{T_{ij} : i < j\}$ corresponding to (2.13) would reject H_0 for large values of

$$(2.14)\qquad \sum_{i<j} T_{ij}/\hat{\sigma}_T,$$

where $\hat{\sigma}_T$ is the consistent estimate of the standard deviation σ_T of $\sum_{i<j} T_{ij}$ obtained from

$$(2.15) \qquad \sigma_T = [r(r^2 - 1)V_0(T_{ij})/6]^{\frac{1}{2}}$$

by replacing $\lambda(F)$ by its consistent estimate (2.7). Under H_0, (2.14) has an asymptotic standard normal distribution, and the asymptotic efficiency of (2.14) to (2.13) is e'. Lehmann (1964) has shown the efficiency e' to be an increasing function of r, while it is known [3] to satisfy $.864 \leqq e' \leqq \infty$ for $r = 2$. Thus (2.14) is robust. An other test statistic based on the U's has been considered for H_1 by Hollander (1966). This statistic is slightly less efficient than (2.14). See [7], Section 5.

Suppose next that H_0 is to be tested against an "unordered" alternative which only specifies that the ξ's are not all equal. The test based on the T's then rejects for large values of the statistic

$$(2.16) \qquad T_1^2 = n \textstyle\sum_{i=1}^r T_{i\cdot}^2/\hat{\sigma}^2 = n \sum_{i=1}^r [U_{i\cdot} - (r - 1)/2r]^2/\hat{\sigma}^2$$

where $\hat{\sigma}^2$ is the consistent estimate of the variance

$$(2.17) \qquad n(r - 1)V_0(T_{ij})/2r$$

of $n^{\frac{1}{2}}T_{i\cdot}$ obtained by replacing $\lambda(F)$ in (2.5) by (2.7). T_1^2 has a limiting chi-square distribution with $(r - 1)$ degrees of freedom, and its Pitman asymptotic efficiency to the usual [5], p. 278, $\mathfrak{F}$-ratio statistics is e'. It follows from the above results and those of Lehmann (1964) that T_1^2 has the same Pitman asymptotic efficiency as the statistics proposed by him in equations (4.1) and (4.2) of [8]. However, T_1^2 seems to have an advantage over these statistics for the problem considered here in that it is easier to compute.

3. Estimation of a contrast. Hodges and Lehmann (1963), and Lehmann (1963a), (1963b), (1964) have derived robust estimates from the Wilcoxon statistic. In this section, their approach will be used to arrive at a robust estimate of the contrast $\theta = \sum_{i=1}^r c_i\xi_i(\sum c_i = 0)$ that is derived from the Friedman (1937) statistic.

It is easy to show that the Friedman statistic can be written in the form

$$(3.1) \qquad \textstyle\sum_i [\sum_j b_{ij}(S_{i\cdot} - S_{j\cdot})]^2$$

where the b's are constants and S_{ij} = number of α's such that $(X_{i\alpha} - X_{j\alpha}) > 0$. By the reasoning in the reference given above, this suggests writing θ in the form

$$(3.2) \qquad \theta = \textstyle\sum \sum d_{ij}(\xi_i - \xi_j)$$

and estimating it by

$$(3.3) \qquad \hat{\theta} = \textstyle\sum \sum d_{ij}(S'_{i\cdot} - S'_{j\cdot})$$

where S'_{ij} is the median of the n quantities $\{X_{i\alpha} - X_{j\alpha} : \alpha = 1, \cdots, n\}$.

The following lemma is well known.

LEMMA 3.1. *Suppose* $\xi_i - \xi_j = a_{ij}/n^{\frac{1}{2}}$ *where the* a*'s are constants, then*

$\{n^{\frac{1}{2}}[S_{ij} - E(S_{ij})] : i < j\}$ has asymptotically the $\frac{1}{2}r(r-1)$ variate normal distribution with zero mean and covariance matrix $(\tau_{ij,kl})$ given by

$$\begin{aligned} &\tau_{ij,ij} = \tfrac{1}{4}; \quad \tau_{ij,kl} = 0, && \text{if } i, j, k, l \text{ are distinct,}\\ (3.4) \qquad &\tau_{ij,kl} = \tfrac{1}{12}, && \text{if } i = k \text{ or } j = l,\\ &\tau_{ij,kl} = -\tfrac{1}{12}, && \text{if } i = l \text{ or } j = k. \end{aligned}$$

LEMMA 3.2. *If the density f of F exists and satisfies the regularity conditions of Lemma 3(a) of [4], then the joint limiting distribution of $\{n^{\frac{1}{2}}[S'_{ij} - (\xi_i - \xi_j)] : i < j\}$ is the $\frac{1}{2}r(r-1)$ variate normal distribution with zero mean and covariance matrix $(\tau^*_{ij,kl})$ given by*

$$\begin{aligned} &\tau^*_{ij,ij} = \tfrac{1}{4}(\textstyle\int f^2(x)\,dx)^2; \quad \tau^*_{ij,kl} = 0, && \text{if } i, j, k, l \text{ are distinct,}\\ (3.5) \qquad &\tau^*_{ij,kl} = \tfrac{1}{12}(\textstyle\int f^2(x)\,dx)^2, && \text{if } i = k \text{ or } j = l,\\ &\tau^*_{ij,kl} = -\tfrac{1}{12}(\textstyle\int f^2(x)\,dx)^2, && \text{if } i = l \text{ or } j = k. \end{aligned}$$

PROOF.

$$\begin{aligned} \lim_n P\{n^{\frac{1}{2}}[S'_{ij} - (\xi_i - \xi_j)] \leqq a_{ij} : i < j\} &= \lim_n P_n\{n^{\frac{1}{2}}[S_{ij} - \tfrac{1}{2}] \leqq 0 : i < j\}\\ &= \lim_n P_n\{n^{\frac{1}{2}}[S_{ij} - E(S_{ij})] \leqq n^{\frac{1}{2}}[\tfrac{1}{2} - E(S_{ij})] : i < j\} \end{aligned}$$

where P_n indicates that the probability is computed for $\xi_i - \xi_j = a_{ij}/n^{\frac{1}{2}} = \Delta_{ij}$. The first equality follows from the results of Hodges and Lehmann (1963). Moreover, $n^{\frac{1}{2}}[\frac{1}{2} - E(S_{ij})] = a_{ij} \int (1/\Delta_{ij})[F(t) - F(t - \Delta_{ij})]\,dF(t) \rightarrow a_{ij} \int f^2(t)\,dt$ as $\Delta_{ij} \rightarrow 0$ $(n \rightarrow \infty)$ and the result follows from Lemma 3.1.

LEMMA 3.3. *Under the condition of Lemma 3.2, the joint limiting distribution of $\{n^{\frac{1}{2}}[(S'_{i\cdot} - S'_{j\cdot}) - (\xi_i - \xi_j)] : i < j\}$ is the $\frac{1}{2}r(r-1)$ variate normal distribution with zero mean and covariance matrix $(C'_{ij,kl})$ given by*

$$\begin{aligned} &C'_{ij,ij} = (r+1)/6r(\textstyle\int f^2(x)\,dx)^2; \quad C'_{ij,kl} = 0,\\ (3.6) \qquad & && \text{if } i, j, k, l \text{ are distinct,}\\ &C'_{ij,kl} = (r+1)/12r(\textstyle\int f^2(x)\,dx)^2, && \text{if } i = k \text{ or } j = l,\\ &C'_{ij,kl} = -(r+1)/12r(\textstyle\int f^2(x)\,dx)^2, && \text{if } i = l \text{ or } j = k. \end{aligned}$$

PROOF. Asymptotic normality follows since $S'_{i\cdot} - S'_{j\cdot}$ is a finite sum of asymptotically normal variables. The covariances are easily computed from (3.5).

From Lemma 3.3, it is seen that the asymptotic efficiency (in the sense of ratios of reciprocals of variances) of $S'_{i\cdot} - S'_{j\cdot}$ to $X_{i\cdot} - X_{j\cdot}$ is

$$(3.7) \qquad 12r\sigma_Y^2(\textstyle\int f^2(x)\,dx)^2/(r+1)$$

where σ_Y^2 is the variance of Y_{ij}.

Since the covariance matrix of $\{S_{i\cdot} - S'_{j\cdot} : i < j\}$ is proportional to that of

$\{X_{i\cdot} - X_{j\cdot} : i < j\}$, it follows that $\hat{\theta}$ has the asymptotic efficiency (3.7) to $\sum c_i X_{i\cdot} = \sum d_{ij}(X_{i\cdot} - X_{j\cdot})$. The formula (3.7) is the efficiency of the Friedman (1937) statistic, and its robustness properties are well known.

The estimate given by Lehmann (1964) is more efficient than $\hat{\theta}$ when F is normal. However, the "Friedman estimate" $\hat{\theta}$ introduced here is simpler to compute and is more efficient than the Lehmann estimate for some distributions F (e.g. when F is the double exponential distribution).

REFERENCES

[1] Bartholomew, D. J. (1961). Order tests in the analysis of variance. *Biometrika* **48** 325–332.

[2] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* **32** 675–698.

[3] Hodges, J. L., Jr., and Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Ann. Math. Statist.* **27** 324–325.

[4] Hodges, J. L., Jr., and Lehmann, E. L. (1961). Comparison of the normal scores and Wilcoxon tests. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 307–318. Univ. of California Press.

[5] Hodges, J. L., Jr., and Lehmann, E. L. (1963). Estimates of location based on ranks. *Ann. Math. Statist.* **34** 598–611.

[6] Hoeffding, W. A. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

[7] Hollander, M. (1967). Rank tests for randomized blocks when the alternatives have an a priority ordering. *Ann. Math. Statist.* **38** 867–877.

[8] Hogg, R. V. (1965). On models and hypothesis with restricted alternatives. *J. Amer. Statist. Assoc.* **60** 1153–1162.

[9] Lehmann, E. L. (1959). *Testing Statistical Hypothesis*. Wiley, New York.

[10] Lehmann, E. L. (1963a). Robust estimation in analysis of variance. *Ann. Math. Statist.* **34** 957–966.

[11] Lehmann, E. L. (1963b). Asymptotically nonparametric inference: an alternative approach to linear models. *Ann. Math. Statist.* **34** 1494–1506.

[12] Lehmann, E. L. (1964). Asymptotically nonparametric inference in some linear models with one observation per cell. *Ann. Math. Statist.* **35** 726–734.

[13] Nüesch, P. E. (1966). On the problem of testing location in multivariate populations for restricted alternatives. *Ann. Math. Statist.* **37** 113–119.

[14] Tukey, J. W. (1949). The simplest signed-rank test. Mem. Report 17, Statistical Research Group, Princeton Univ.