STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

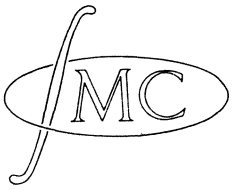
AFDELING MATHEMATISCHE STATISTIEK

Report S 330

Nonparametric methods for regression

by

Kumar Jogdeo



August 1964

The Mathematical Centre at Amsterdam, founded the 11th of February, 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications, and is sponsored by the Netherlands Government through the Netherlands Organization for Pure Research (Z.W.O.) and the Central National Council for Applied Scientific Research in the Netherlands (T.N.O.), by the Municipality of Amsterdam and by several industries.

NONPARAMETRIC METHODS FOR REGRESSION

BY: KUMAR JOGDEO [1]

University of Illinois and University of California.

1. Introduction and Summary.

This paper considers the rank score tests for testing hypotheses in certain regression models. Hájek (1962) studied a class of rank score tests for the following model,

1.1 $$Y_\nu = \alpha + \beta x_\nu + \sigma Z_\nu , \quad \nu = 1,\ldots,n;$$

where Y_ν are observable random variables, x_ν are given constants which may depend on n, Z_ν are independent identically distributed random variables with mean zero and variance unity and α, β_1 σ are parameters. The hypothesis to be tested is $\beta = 0$.

The following extensions are of interest in the present study.

1.2 $$Y_\nu = \alpha + \beta_1 x_{1\nu} + \ldots + \beta_k x_{k\nu} + \sigma Z_\nu, \quad \nu=1,\ldots,n;$$

---

[1] This paper was prepared with the partial support of the Office of Ordnance Research, U.S.A. grant (DA-ARO (D)-31-124-G-83) and was revised at the Mathematisch Centrum, Amsterdam.

where the notation is similar to that in (1.1) and the hypothesis to be tested is $\beta_1 = \beta_2 = \ldots = 0$. (The dependence of the constants $x_{k\nu}$ on n is not indicated to avoid too many subscripts.) Model (1.2) has many applications. In particular, the present study covers the class of rank score tests for the analysis of variance model discussed by Puri (1964).

Another extension of (1.1) studied here is

1.3 $$Y_\nu = \alpha + \beta X_\nu + \sigma Z_\nu, \quad \nu = 1,\ldots,n;$$

where X and Z are independent random variables and the hypothesis to be tested is $\beta = 0$ which is equivalent to the hypothesis of independence of X and Y.

A further extension of (1.3) is also considered,

1.4 $$Y_\nu = \alpha + \beta_1 X_{1\nu} + \ldots + \beta_k X_{k\nu} + \sigma Z_\nu, \quad \nu = 1,\ldots,n;$$

where $(Y_\nu, X_{1\nu}, \ldots, X_{k\nu})$, $\nu = 1,\ldots,n$ are random vectors with Z independent of X components and the hypothesis to be tested is $\beta_1 = \beta_2 = \ldots = \beta_k = 0$.

The regression models given by equations (1.2), (1.3) and (1.4) will be denoted as Model I, II and III respectively. Some authors have studied rank score tests for other models of linear dependence. Bhuchongkul (1964) studied the problem of testing independence in a bivariate population against the alternative

1.5 $$X = (1-\theta)\, Z_1 + \theta\, Z_2,$$
$$Y = (1-\theta)\, Z_3 + \theta\, Z_2,$$

where Z_1, Z_2, Z_3 are independent identically distributed random variables. In (1.5), $\theta = 0$ implies that X and Y are independent.

Konijn (1956) studied still another alternative which can be written as

$$(1.6) \qquad X = p_1 Z_1 + p_2 Z_2$$
$$Y = q_1 Z_1 + q_2 Z_2$$

where Z_1 and Z_2 are independent random variables and $p_1 = q_2 = 0$ implies the independence of X and Y. Model (1.6) is more general than (1.4), however, the class of rank score tests studied by Konijn (1956) is more restrictive than the one considered presently.

The material of this paper can de divided into two parts. The first part (section 2) deals with the asymptotic normality of the various test statistics and the second part (section 3) is devoted to the study of the asymptotic efficiency. For proving the asymptotic normality two techniques are used. The first is due to Hájek (1961, 1962) which is based on the concept of <u>contiguity</u> developed by Le Cam (1960) and Hájek (1962). The second technique is the bivariate extension of the Chernoff-Savage (1958) method studied by Bhuchongkul (1964). For Models I and II some analogues of the familiar rank score tests are compared with the parametric tests. A new class denoted by <u>mixed rank score</u> tests is also studied for Model II problems. These statistics are based on the rank scores of the Y observations and the X observations themselves. It is pointed out that in general

mixed rank score tests have better asymptotic efficiency than rank score tests. This shows that one should refrain from taking a view completely opposite to the orthodox view and start ranking observations in all situations.

The results for Model III follow easily from those of Models I and II and hence are omitted.

2. Asymptotic Normality.

The asymptotic normality for various statistics will be shown in the separate subsections. The fundamental technique used here is that of contiguity and in order to have a rather complete picture a theorem of Le Cam (1960) and Hájek (1962) will be briefly stated in the following subsection.

2.1 Contiguity

Let $\{P_n\}$ and $\{Q_n\}$ be two sequences of probability measures on a sequence of probability spaces $\{\mathcal{X}_n, \mathcal{U}_n\}$. The probability measure Q_n is said to be contiguous to P_n if for any sequence $\{A_n \in \mathcal{U}_n\}$,

2.1.1 $$P_n(A_n) \longrightarrow 0 \implies Q_n(A_n) \longrightarrow 0.$$

In applications $\{P_n\}$ and $\{Q_n\}$ correspond to the measures induced by the hypothesis and a sequence of alternatives approaching the hypothesis. With contiguity the problem of studying asymptotic distributions can be restricted to the hypothesis which in many cases is easier. For example, in

the model given by equation (1.1), denoting $\bar{x} = \sum_{\nu=1}^{n} x_\nu/n$, $\{P_n\}$ and $\{Q_n\}$ refer to $H_o : \alpha = \alpha_o + \beta^o\bar{x}$, $\beta = 0$, $\sigma = \sigma_o$ and $H_1 : \alpha = \alpha_o$, $\beta = \beta^o$, $\sigma = \sigma_o$; in both cases the underlying distribution of the random variable Z is assumed to be G. The class of rank score statistics considered by Hájek (1962) is constructed in the following manner. It is assumed that G is absolutely continuous and possesses the first two derivitives g and g' respectively which satisfy the condition

$$2.1.2 \qquad \int_{-\infty}^{\infty} \left[\frac{g'(x)}{g(x)}\right]^2 g(x)dx = K < \infty \ .$$

Let

$$2.1.3 \qquad d_n^2 = \sum_{\nu=1}^{n} (x_\nu - \bar{x})^2 \int_{-\infty}^{\infty} \left[\frac{g'(x)}{g(x)}\right]^2 g(x)dx \ ,$$

and assume that

$$2.1.4 \qquad \lim_{n \longrightarrow \infty} d_n^2 = d^2 \ .$$

Define

$$2.1.5 \qquad \phi(u) = -\left[g'(G^{-1}(u))/g(G^{-1}(u))\right], \quad 0 < u < 1 \ ,$$

$$2.1.6 \qquad \phi_n(u) = \phi\left(\frac{i}{n+1}\right) \qquad \text{for } \frac{i-1}{n} < u \leq \frac{i}{n} \ .$$

The rank score statistic obtained from (2.1.6) has the form

$$2.1.7 \qquad T_n^o = \sum_{\nu=1}^{n} (x_\nu - \bar{x})\, \phi_n\left(\frac{R_{\nu n}}{n+1}\right)$$

where $R_{\nu n}$ is the rank of Y_ν in $(Y_1, \ldots, Y_n)$. In practice, however, the underlying distribution is unknown and a rank

score statistic can be constructed from any distribution function H satisfying the same conditions imposed on G. Following (2.1.5) and (2.1.6) let ψ denote the function obtained from H and

2.1.8 $$T_n^1 = \sum_{\nu=1}^{n} (x_\nu - \bar{x})\, \psi_n\left(\frac{R_{\nu n}}{n+1}\right) .$$

In the following we summarize the contiguity theorem and the results regarding limiting behaviour of T_n^0 and T_n^1 .

Let P_n and Q_n be the product measures such that

2.1.9 $$P_n = \prod_{\nu=1}^{n} P_{\nu n} \quad , \quad Q_n = \prod_{\nu=1}^{n} Q_{\nu n}$$

and define

2.1.10 $$\zeta_{\nu n} = \begin{cases} \dfrac{dQ_{\nu n}}{dP_{\nu n}} & \text{on } B_{\nu n}, \\ 0 & \text{elsewhere,} \end{cases}$$

where on the set $B_{\nu n}$ the measure $Q_{\nu n}$ is absolutely continuous with respect to $P_{\nu n}$ and $B_{\nu n}$ has $P_{\nu n}$ measure unity.
Let

2.1.11 $$W_n = 2 \sum_{\nu=1}^{n} (\zeta_\nu^{1/2} - 1).$$

Theorem 2.1 (Le Cam - Hájek) With the above notation if

a) $$\lim_{n \longrightarrow \infty} \max_{1 \leq \nu \leq n} P_{\nu n}(|\zeta_{\nu n} - 1| > \varepsilon) = 0,$$

b) $$\mathcal{L}(W_n \mid P_n) \longrightarrow \mathcal{N}(-\sigma^2/4 \, , \, \sigma^2)$$

then

1) Q_n is contiguous to P_n.

2) If a random variable Z_n^1 is such that $\mathcal{L}(Z_n^1 | P_n) \longrightarrow \mathcal{N}(a_1, b_1^2)$ and (W_n, Z_n^1) has a limiting bivariate normal distribution with correlation coefficient ρ_1, then

$$2.1.12 \qquad \mathcal{L}(Z_n^1 | Q_n) \longrightarrow \mathcal{N}(a_1 + \rho_1 b_1 \sigma, b_1^2).$$

Further in the model given by (1.1)

3) i) $\mathcal{L}(T_n^0 | P_n) \longrightarrow \mathcal{N}(0, d^2)$

ii) $\mathcal{L}(T_n^1 | P_n) \longrightarrow \mathcal{N}(0, c^2)$

iii) $\mathcal{L}(T_n^0 | Q_n) \longrightarrow \mathcal{N}(\frac{\beta_0}{\sigma_0} d^2, d^2)$

iv) $\mathcal{L}(T_n^1 | Q_n) \longrightarrow \mathcal{N}(\frac{\beta_0}{\sigma_0} \rho cd, c^2)$

where $$c^2 = \lim_{n \longrightarrow \infty} c_n^2 = \lim_{n \longrightarrow \infty} \sum_{\nu=1}^{n} (x_\nu - \bar{x})^2 \int_0^1 \psi^2(u)du,$$

and

$$\rho_1 = \frac{\int_0^1 \psi(u)\, \phi(u)du}{\left[\int_0^1 \phi^2(u)du\right]^{1/2} \left[\int_0^1 \psi^2(u)du\right]^{1/2}}.$$

4) If a random variable Z_n^0 is such that $\mathcal{L}(Z_n^0 | P_n) \longrightarrow \mathcal{N}(a_0, b_0^2)$ and (T_n^0, Z_n^0) has a limiting bivariate normal distribution with correlation coefficient ρ_0 then

$$2.1.13 \qquad \mathcal{L}(Z_n^0 | P_n) \longrightarrow \mathcal{N}(\frac{\beta_0}{\sigma_0} \rho_0 db_0, b_0^2).$$

This theorem will be used in the following subsections.

2.2 Model I

With the same notation as in section 1, let

2.2.1 $$Y_\nu = \alpha + \beta_1 x_{1\nu} + \cdots + \beta_k x_{k\nu} + Z_\nu, \quad \nu = 1, \cdots, n;$$

and let

2.2.2 $$P[Y_\nu \leq y \mid \alpha, \beta_1, \cdots, \beta_k, \sigma] = G((y - \alpha - \sum_{i=1}^{k} \beta_i x_{i\nu})/\sigma),$$

where the distribution function G satisfies the conditions mentioned in the subsection 2.1. The hypothesis to be tested is $H_o : \alpha = \alpha_o + \beta_1^o \bar{x}_1 + \cdots + \beta_k^o \bar{x}_k$, $\beta_1 = \cdots = \beta_k = 0$, $\sigma = \sigma_o$ against the alternative $H_1 : \alpha = \alpha_o$, $\beta_1 = \beta_1^o, \cdots, \beta_k = \beta_k^o \neq 0$, $\sigma = \sigma_o$. Here $\bar{x}_i = \sum_{\nu=1}^{n} x_{i\nu}/n$.

The constants $x_{i\nu}$ are assumed to satisfy the following conditions:

2.2.3 $$\sup_n \sum_{\nu=1}^{n} (x_{i\nu} - \bar{x}_i)^2 = k_i < \infty,$$

2.2.4 $$\lim_{n \to \infty} \max_{1 \leq \nu \leq n} (x_{i\nu} - \bar{x}_i)^2 = 0 \; ; \; i = 1, \cdots, k.$$

Further, for every set of real co-efficients $(\ell_1, \cdots, \ell_k)$, not all zero, there exists a positive number $c(\ell_1, \cdots, \ell_k)$ such that

2.2.5 $$\sum_{\nu=1}^{n}\left[\sum_{i=1}^{k} \ell_i (x_{i\nu} - \bar{x}_i)\right]^2 > c(\ell_1, \cdots, \ell_k) > 0,$$

for all n. The condition (2.2.5) is the condition of linear independence. The conditions (2.2.3) and (2.2.4) indicate that the constants $(x_{i\nu} - \bar{x}_i)$ tend to zero as n increases. This is equivalent to keeping the sum of squares $\sum_{\nu=1}^{n} (x_{i\nu} - \bar{x}_i)^2$ at $O(n)$, $\max(x_{i\nu} - \bar{x}_i)^2$ at $o(n)$ and letting the parameters β_i tend to zero at the rate of $n^{-1/2}$ which is relevent for st- lying the Pitman efficiency. Let

2.2.6 $$d'_{n,ij} = \sum_{\nu=1}^{n} (x_{i\nu} - \bar{x}_i)(x_{j\nu} - \bar{x}_j),$$

and assume that

2.2.7 $$\lim_{n \longrightarrow \infty} d'_{n,ij} = d'_{ij} .$$

Further, writing $\underset{\sim}{x}_\nu = (x_{1\nu}, \cdots, x_{k\nu})$ and $\underset{\sim}{\beta} = (\beta_1, \cdots, \beta_k)$ in equation (2.2.1), the expression $\beta_1 x_{1\nu} + \cdots + \beta_k x_{k\nu}$ could be replaced by $\underset{\sim}{\beta} \cdot \underset{\sim}{x}_\nu$. It can be seen that under any orthogonal linear transformation of $\underset{\sim}{x}_\nu$ the problem of testing $\underset{\sim}{\beta} = 0$ remains invariant. This allows us to assume without loss of generality

2.2.8 $$d'_{ij} = 0 \quad \text{for } i \neq j.$$

On the other hand condition (2.2.5) implies that

2.2.9 $$d'_{ii} > 0.$$

Henceforth

2.2.10 $$d_{ii} = K\, d'_{ii} = d'_{ii} \int_{-\infty}^{\infty} \left[\frac{g'(x)}{g(x)} \right]^2 g(x)dx.$$

Recalling the definition of ϕ_n in (2.1.6) define

2.2.11 $$T_{n,i} = \sum_{\nu=1}^{n} (x_{i\nu} - \bar{x}_i)\, \phi_n\left(\frac{R_{\nu n}}{n+1}\right), \quad i=1,\cdots,k;$$

where $R_{\nu n}$ is the rank of Y_ν in $Y_1,\cdots,Y_n$.

The following theorem gives the joint normality of the statistics $T_{n,i}$.

Theorem 2.2

With the above notation and conditions (2.1.2), (2.2.3), (2.2.4), 2.2.5), (2.2.7) and (2.2.8) the statistics $T_{n,i}$ are asymptotically independent and

2.2.12 $$\mathcal{L}(T_{n,i} \mid H_o) \longrightarrow \mathcal{N}(0\ ,\ d_{ii}),$$

2.2.13 $$\mathcal{L}(T_{n,i} \mid H_n) \longrightarrow \mathcal{N}(\beta_i^o/\sqrt{n})\, d_{ii}\ ,\ d_{ii}).$$

Proof: The main idea of the proof is to reduce the problem to the original model with k=1. Consider the set of constants

2.2.14 $$\sum_{i=1}^{k} c_i(\quad - \bar{x}_i)$$

where the c_i are arbitrary real numbers.

From (2.2.3), (2.2.4) and (2.2.5) it follows that there exist two positive constants K_1' and K_2' such that

2.2.15 $$K_1' < \quad (x_{\cdot} - \bar{x})^2 < K_2' \qquad \text{for all } n,$$

and

$$2.2.16 \qquad \lim_{n \longrightarrow \infty} \max_{1 \leq \nu \leq n} (x_\nu - \bar{x})^2 = 0.$$

With (2.2.15), (2.2.16) and the conditions on the function ϕ mentioned, theorem 2.1 applies directly. It follows that the statistic

$$2.2.17 \qquad \sum_{\nu=1}^{n} (x_\nu - \bar{x})\, \phi_n(R_{\nu n}/_{n+1})$$

is asymptotically normally distributed. However, the expression in (2.2.17) is an arbitrary linear combination of the statistics $T_{n,i}$ and hence th joint asymptotic normality of $T_{n,i}$ follows. Further it follows from (2.2.7), (2.2.8) and (2.2.9) that

$$2.2.18 \qquad \mathcal{L}(T_{n,i} \mid H_o) \longrightarrow \mathcal{N}(0, d_{ii}) \quad \text{as } n \longrightarrow \infty,$$

and that $T_{n,i}$ are asymptotically independent.

When H_1 is true, write

$$(2.2.19) \qquad P[Y_\nu \leq y \mid \alpha_o, \beta_1^o, \cdots, \beta_k^o, \sigma_o]$$

$$= G[\{y - \alpha_o - \beta_k^o \sum_{i=1}^{k} (\beta_i^o/\beta_k^o)\, x_{i\nu}\}/\sigma_o]$$

$$= G[(y - \alpha_o - \beta_k^o\, a_\nu)/\sigma_o],$$

where

$$2.2.20 \qquad a_\nu = \sum_{i=1}^{k} (\beta_i^o/\beta_k^o)\, x_{i\nu}\ .$$

Let

2.2.21 $$V_n = \sum_{\nu=1}^{n} (a_\nu - \bar{a})\, \phi_n\left(\frac{R_{\nu n}}{n+1}\right), \quad \bar{a} = \sum_{\nu=1}^{n} a_\nu/n.$$

Applying theorem 2.1 once more it is seen that

2.2.22 $$\mathcal{L}(V_n \mid H_o) \longrightarrow \mathcal{N}\left(0, \sum_{i=1}^{k} (\beta_i^o/\beta_k^o)^2\, d_{ii}\right),$$

2.2.23 $$\mathcal{L}(V_n \mid H_1) \longrightarrow \mathcal{N}\left(\frac{\beta_k^o}{\sigma_o}, \sum_{i=1}^{k} (\beta_i^o/\beta_k^o)^2\, d_{ii}\right).$$

Observing that $(T_{n,i}, V_n)$ is an asymptotically bivariate normal random variable under H_o with covariance $(\beta_i^o/\beta_k^o)\, d_{ii}$, (4) of theorem 2.1 can be applied and

2.2.24 $$L(T_{n,i} \mid H_1) \longrightarrow \mathcal{N}\left(\frac{\beta_i^o\, d_{ii}}{\sigma_o}, d_{ii}\right).$$

This completes the proof.

Remark I Let H be a distribution function satisfying the same regularity conditions as G defined above. Recalling the definition ψ_n given in the subsection 2.1, consider

2.2.25 $$T'_{n,i} = \sum_{\nu=1}^{n} (x_{i\nu} - \bar{x}_i)\, \psi_n\left(\frac{R_{\nu n}}{n+1}\right).$$

Theorem (2.1) applies again and it is seen that

2.2.26 $$\mathcal{L}(T'_{n,i} \mid H_o) \longrightarrow \mathcal{N}(0, c_{ii}),$$

where

2.2.27 $$c_{ii} = d'_{ii} \int_0^1 \psi^2(u)du,$$

and

2.2.28 $$\mathcal{L}(T'_{n,i} \mid H_1) \longrightarrow \mathcal{N}\left(\frac{\beta^0_i \mu_{11} d'_{ii}}{\sigma_0}, c_{ii}\right)$$

where

2.2.29 $$\mu_{11} = \int_0^1 \phi(u)\psi(u)\, du .$$

2.3 Model II: Mixed rank score test statistics

Keep the same notation for the functions ϕ and ϕ_n and the same regularity conditions on the distribution function G. In Model II

2.3.1 $$Y_\nu = \alpha + \beta X_\nu + \sigma Z_\nu, \quad \nu = 1, \cdots, n;$$

with

2.3.2 $$P[Y_\nu \le y \mid X_\nu = x] = G((y-\alpha-\beta x)/\sigma).$$

The random variables X_ν, $\nu = 1, \cdots, n$, are assumed to be independent identically distributed with expected value zero and common distribution function F possessing finite second moment and X_ν and Z_ν are assumed to be independent. Thus

2.3.3 $$\int_{-\infty}^{\infty} x\, df(x) = 0, \quad \int_{-\infty}^{\infty} x^2\, df(x) = \eta^2 > 0.$$

For testing $\beta = 0$ in (2.3.1) we consider the mixed rank-score statistic

2.3.4 $$S_n = \frac{1}{n} \sum_{\nu=1}^{n} (X_\nu - \bar{X}) \phi_n\left(\frac{R_{\nu n}}{n+1}\right),$$

where $$\bar{X} = \sum_{\nu=1}^{n} X_\nu/n.$$

Note the similarity between S_n and T_n^o given by (2.1.7). Before studying the asymptotic behaviour of S_n it will be shown that, in fact, the random variables $(X_\nu - \bar{X})/\sqrt{n}$ satisfy the conditions imposed on the constants x_ν of subsection 2.1 with probability one as $n \longrightarrow \infty$. That is,

2.3.5 $$\lim_{n \longrightarrow \infty} P\left[0 < k_1 < \frac{1}{n} \sum_{\nu=1}^{n} (X_\nu - \bar{X})^2 < k_2 < \infty\right] = 1,$$

and with probability one

2.3.6 $$\frac{1}{n} \max_{1 \le \nu \le n} (X_\nu - \bar{X})^2 \longrightarrow 0 \quad \text{as } n \longrightarrow \infty .$$

Condition (2.3.5) follows immediately from the fact that

2.3.7 $$\frac{1}{n} \sum_{\nu=1}^{n} (X_\nu - \bar{X})^2 \xrightarrow{a.s.} \sigma^2 \text{ , as } n \longrightarrow \infty .$$

For showing (2.3.6) observe that

2.3.8 $$\frac{1}{n} \max (X_\nu - \bar{X})^2 \le \frac{2 \max X_\nu^2}{n} + \frac{2\bar{X}^2}{n} .$$

The second term on the right side of (2.3.8) vanishes with probability one in view of Kolmogorov's strong law of large numbers. Convergence of the first term follows from (2.3.3) and a result of Dharmadhikari and the author (1964) which states that with a sequence $\{Y_n\}$ of independent identically distributed random variables $\max (Y_1, \cdots, Y_n)/n$ converges to zero if and only if $E|Y| < \infty$.

In order to apply theorem 2.1 to the Model II problem note that with $H_o : \alpha = \alpha_o$, $\beta = 0$, $\sigma = \sigma_o$ and $H_{1n} : \alpha = \alpha_o$, $\beta = \beta_o/\sqrt{n}$ and $\sigma = \sigma_o$,

$$2.3.9 \qquad \xi_{\nu n} = \frac{\int_{-\infty}^{\infty} g\left(\frac{Y_\nu - \alpha_o - x\beta_o/\sqrt{n}}{\sigma_o}\right) dF(x)}{g\left(\frac{Y_\nu - \alpha_o}{\sigma_o}\right)} .$$

Hence $\xi_{\nu n}$ for $\nu=1,\cdots,n$ are independent identically distributed random variables. The condition (a) of the theorem 2.1 will be satisfied if we show that under H_o

$$2.3.10 \qquad |\xi_{\nu n} - 1| \xrightarrow{P} 0.$$

However, due to regularity conditions satisfied by the function g

$$2.3.11 \qquad E|\xi_{\nu n} - 1| = \int\int |g\left(\frac{y-\alpha_o-x\beta_o/\sqrt{n}}{\sigma_o}\right) - g\left(\frac{y-\alpha_o}{\sigma_o}\right)| \, dF(x)\, dy \longrightarrow 0,$$

as $n \longrightarrow \infty$.

The main idea of the proof of the condition (b) is to use the fact that the summands in W_n (see (2.1.11)) are independent identically distributed and then apply the standard techniques of the central limit theorem. However to determine the asymptotic mean and the variance one could follow exactly the same steps as in the proof of the main theorem of Hájek (1962, section 5), the only difference being that $X_1,\cdots,X_n$ are random variables for the present situation. The equations derived by Hájek could be viewed as obtained with $X_1,\cdots,X_n$ fixed and by taking expected values it can easily be seen that

2.3.12 $$\mathcal{L}(W_n \mid H_o) \longrightarrow \mathcal{N}(-\frac{\beta_o^2 \tau^2}{4\sigma_o^2}, \frac{\beta_o^2 \tau^2}{\sigma_o^2}),$$

and under H_o

2.3.13 $$W_n - E W_n - \mu_n' \xrightarrow{P} 0,$$

where

2.3.14 $$\tau^2 = \left[\int_{-\infty}^{\infty} x^2 dF(x)\right]\left[\int_0^1 \phi^2(u)\, du\right] = n^2 \int_0^1 \phi^2(u)\, du$$

and

2.3.15 $$S_n' = \frac{1}{\sqrt{n}} \sum_{\nu=1}^{n} X \phi_n\left(\frac{R_{\nu n}}{n+1}\right).$$

Theorem 2.3 With the above notation and conditions on the distribution functions F and G the mixed rank-score statistic S_n (see 2.3.4) is asymptotically normally distributed:

2.3.16 $$\mathcal{L}(S_n \mid H_o) \longrightarrow \mathcal{N}(0, \tau^2),$$

2.3.17 $$\mathcal{L}(S_n \mid H_{1n}) \longrightarrow \mathcal{N}(\frac{\beta_o \tau^2}{\sigma_o}, \tau^2).$$

Proof: First observe that S_n and S_n' are equivalent in the sense that

2.3.18 $$S_n' - S_n = \sqrt{n}\, \bar{x} \sum_{\nu=1}^{n} \phi_n\left(\frac{\nu}{n+1}\right)/n \xrightarrow{P} 0,$$

since

2.3.19 $$\sum_{\nu=1}^{n} \phi_n\left(\frac{\nu}{n+1}\right)/n \longrightarrow \int_0^1 \phi(u)du = 0.$$

From (2.3.11) and (2.3.12) it is seen that the probability measures under H_{1n} are contiguous and it follows from theorem 2.1 that S_n has asymptotic normal distribution under H_o and H_{1n} with the specified means and variances.

Remark II

Using the same notation as in Remark I and defining

2.3.20 $$S_n^* = \frac{1}{\sqrt{n}} \sum_{\nu=1}^{n} (X_\nu - \bar{X})\, \psi_n\left(\frac{R_{\nu n}}{n+1}\right),$$

it follows that

2.3.21 $$\mathcal{L}(S_n^* \mid H_o) \longrightarrow \mathcal{N}(0, \xi^2),$$

2.3.22 $$\mathcal{L}(S_n^* \mid H_{1n}) \longrightarrow \mathcal{N}\left(\frac{\mu_{11}}{\sigma_o}, \xi^2\right),$$

where

2.3.23 $$\xi^2 = n^2 \int_0^1 \psi^2(u)du \qquad \mu_{11} = \int_0^1 \phi(u)\psi(u)du.$$

2.4 Model II: Rank-score test statistics.

The method used in this subsection is completely different from the foregoing one and hence a considerable change in the notation is needed.

Consider a sample of N observations from a bivariate population denoted by $(X_1,Y_1),\cdots,(X_N,Y_N)$.

Let F, G, F_N, G_N, be the marginal and empirical marginal distribution functions of X and Y and H and H_N those of the pair (X,Y) respectively. Let R_i and S_i be the ranks of X_i and Y_i

among $X_1,\cdots,X_N$ and $Y_1, Y_2,\cdots,Y_N$ respectively, $E_{N,i}$ and $E'_{N,i}$ be given numbers. Consider the statistics

2.4.1 $$T_N = \frac{1}{N} \sum_{i=1}^{N} E_{N,R_i} \; E'_{N,S_i} \;,$$

which can be written as

2.4.2 $$T_N = \int\int J_N\left[F_N(x)\right] \; L_N\left[G_N(y)\right] \; dH_N(x,y)$$

where J_N and L_N are defined by

$$E_{N,i} = J_N(\tfrac{i}{N}) \;, \quad E'_{N,i} = L_N(\tfrac{i}{N}).$$

Under the assumption that F and G are absolutely continuous and a number of regularity conditions on J_N and L_N the asymptotic normality was proved by Bhuchongkul (1964, theorem 1).
The asymptotic mean and variance of the statistic are expressed in terms of

2.4.4 $$\lim_{N\to\infty} J_N(u) = J(u), \quad \lim_{N\to\infty} L_N(u) = L(u), \quad 0 < u < 1.$$

In applications $J_N(i/N)$ is usually of the following form:

2.4.5 $$J_N(i/N) = - E \frac{f'_1(V_i)}{f_1(V_i)} \;,$$

where $V_1 < \ldots < V_N$ is the ordered sample from a population with the distribution function F_1, f_1 and f'_1 being the first two derivatives. Denoting $\omega_i = F_1(V_i)$,

2.4.6 $$J_N(i/N) = E\, \phi_1(\omega_i),$$

where

2.4.7 $$\phi_1(u) = -\frac{f_1'\left[F_1^{-1}(u)\right]}{f_1\left[F_1^{-1}(u)\right]}, \qquad 0 < u < 1.$$

It is seen that

2.4.8 $$\lim_{N \longrightarrow \infty} J_N(u) = J(u) = \phi_1(u).$$

This equation above establishes the relationship between the J function and ϕ function discussed in sections 2.1 and 2.2.

There is also another form of J_N function which can be written in terms of the above notations as

2.4.9 $$J_N\left(\frac{i}{N}\right) = \sum V_i .$$

It can be seen that

2.4.10 $$J_N(u) \longrightarrow J(u) = F_1^{-1}(u), \qquad 0 < u < 1,$$

and the relation between J and ϕ_1 can be given implicitly as

2.4.11 $$\phi_1(J^{-1}(x)) = -\frac{d}{dx} \log \frac{d}{dx} J^{-1}(x), \quad -\infty < x < \infty,$$

provided there are enough regularity conditions to allow all the operations in (2.4.11).

To apply the theorem of Bhuchongkul (1964) for Model II we make the following assumptions. Let f and g be the density functions of F and G respectively and g' be the derivative of g. Without loss of generality assume that

2.4.12 $$\int_0^1 J(u)du = \int_0^1 L(v)dv = 0,$$

and

2.4.13 $$\int_{-\infty}^{\infty} x \, d\,F(x) = 0, \qquad \int_{-\infty}^{\infty} x^2 \, d\,F(x) = 1.$$

Then from theorem 1 of Bhuchongkul (1964) it follows that with $\alpha = \alpha_o$, and $\sigma = \sigma_o$ in Model II

2.4.14 $$\underset{\sim}{\ell}(T_{..}) \longrightarrow \mathcal{N}(\mu, \xi^2),$$

where

2.4.15 $$d_{\ell} = -\frac{1}{\sigma_o}\left[\int_{-\infty}^{\infty} x\, J[F(x)]\, dF(x)\right] \left[\int_{-\infty}^{\infty} L[G(y)]\, g'(y)\, dy\right],$$

2.4.16 $$N\xi^2 \longrightarrow \left[\int J^2(u)du\right]\left[\int L^2(u)du\right] \qquad \text{as } \beta \longrightarrow 0.$$

The expressions (2.4.15) and (2.4.16) are relevant for studying the Pitman efficiency.

3. Asymptotic Efficiency.

3.1 Parametric Tests.

Neyman (1958) developed the theory of "locally asymptotically most powerful" tests when nuisance parameters are present. The test is obtained by substituting estimates having certain properties for the nuisance parameters. In the situation where the form of the density function is known this test has maximum efficiency. In the following, Neyman's tests are

given for Model I and Model II problems.

For Model I, if G is the underlying distribution then with the same notation as in section 2.1, a best parametric test is a quadratic form in

$$3.1.1 \qquad \sum_{\nu=1}^{n} (x_{1\nu} - \bar{x}_1) \frac{g'\left[(Y_\nu - \hat{\alpha})/\hat{\sigma}\right]}{g\left[(Y_\nu - \hat{\alpha})/\hat{\sigma}\right]}, \ldots,$$

$$\sum_{\nu=1}^{n} (x_{k\nu} - \bar{x}_k) \frac{g'\left[(Y_\nu - \hat{\alpha})/\hat{\sigma}\right]}{g\left[(Y_\nu - \hat{\alpha})/\hat{\sigma}\right]},$$

where $\hat{\alpha}$ and $\hat{\sigma}$ are the estimates of α and σ satisfying certain consistency properties. By the same method used in section 2.1 it can be shown that the components of the vector (3.1.1) are asymptotically equivalent to the rank score test statistics $T_{n,i}$ given by (2.2.11).

For Model II Neyman's test is based on the statistic

$$3.1.2 \qquad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \frac{g'\left(\frac{Y_i - \hat{\alpha}}{\hat{\sigma}}\right)}{g\left(\frac{Y_i - \hat{\alpha}}{\hat{\sigma}}\right)}$$

Again this test can be shown to be equivalent to the mixed rank score test based on S_n given by (2.3.4). In fact the locally asymptotically most powerful test serves as a guide for constructing rank score tests.

3.2 Model I: The Likelihood Ratio versus Rank-Score Tests.

Suppose in Model I the random variable Z_ν has a normal distribution. Adopting a notation similar to that of Cramér

(1951) chapter 37 and the notation of the subsection 2.1, let

$$3.2.1 \qquad d'_{ij} = \frac{1}{n} \sum_{\nu=1}^{n} (x_{i\nu} - \bar{x}_i)(x_{j\nu} - \bar{x}_j),$$

$$3.2.2 \qquad d'_{oj} = \frac{1}{n} \sum_{\nu=1}^{n} (x_{j\nu} - \bar{x}_j)(Y_\nu - \bar{Y}),$$

$$3.2.3 \qquad D' = |\, d'_{ij}|, \quad \alpha^* = \bar{Y}, \quad \beta_i^* = \sum_{j=1}^{k} d'_{oj} \, |\, D'_{ij}| \,/\, |D'|,$$

and

$$3.2.4 \qquad \sigma^{*2} = \frac{1}{n} \sum_{\nu=1}^{n} \left[Y_\nu - \alpha^* - \beta_1^*(x_{1\nu} - \bar{x}_1) - \cdots - \beta_k^*(x_{k\nu} - \bar{x}_k) \right]^2.$$

For testing $H_o : \alpha = \alpha_o$, $\beta_1 = \cdots = \beta_k = 0$, $\sigma = \sigma_o$, the classical test is to reject H_o when the statistic

$$3.2.5 \qquad F_n = \frac{n-k-1}{k\sigma^{*2}} \sum_{i=1}^{n} \sum_{j=1}^{n} d'_{ij} \, \beta_i^* \, \beta_j^*$$

is too large.

It is wellknown that F_n has an F distribution with n-k-1 and k degrees of freedom. Assuming without loss of generality that $d'_{ij} = 0$ for $i \neq j$, the statistic F_n in (3.2.5), for large n is equivalent to

$$3.2.6 \qquad F'_n = \frac{n}{k\sigma_o^2} \sum_{i=1}^{k} d'_{ii} \, \beta_i^{*2}.$$

However, F'_n has approximately a central χ^2 distribution with k degrees of freedom under H_o, and for the alternative

$H_1 : \alpha = \alpha_o, \beta_1 = \beta_1^o, \cdots, \beta_k = \beta_k^o$ the statistic F_n' has approximately a noncentral χ^2 distribution with k degrees of freedom and

$$3.2.7 \qquad \frac{n}{\sigma_o^2} \sum_{i=1}^{k} d_{ii} (\beta_i^o)^2$$

as the noncentrality parameter. For a sequence of alternatives $H_{1n} : \alpha = \alpha_o, \beta_i = \beta_i^o/\sqrt{n}, \sigma = \sigma_o$ approaching the hypothesis, the power of the test approaches a constant which depends only upon the noncentrality parameter

$$3.2.8 \qquad \Delta^2 = \sum_{i=1}^{k} d_{ii} (\beta_i^o)^2 / \sigma_o^2 .$$

The rank score tests for testing H_o against H_{1n} are based on the results of subsection 2.2. By using (2.2.18) and (2.2.24) it is seen that if the underlying distribution is normal, G is taken as a standard normal distribution and ϕ is defined accordingly, the distribution of

$$3.2.9 \qquad T_n = \sum_{i=1}^{k} T_{n,i}^2$$

is χ^2 with k degrees of freedom if H_o is true and noncentral χ^2 with k degrees of freedom and with Δ^2 (see (3.2.8)) as the noncentrality parameter in case H_{1n} is true.

When the underlying distribution is not normal the limiting distributions of the statistic F_n under H_o and H_1 take the same forms as given above. (This follows easily from

the central limit theorem and Cramér (1951), chapter 20). By using (2.2.28) it is immediately seen that if G is the underlying distribution and if the test is based on

3.2.10 $$T'_n = \sum_{i=1}^{n} (T'_{n,i})^2$$

then under H_{1n}, the statistic T'_n has approximately a noncentral χ^2 distribution with k degrees of freedom and $\rho^2\Delta^2$ as the noncentrality parameter where

3.2.11 $$\rho^2 = \frac{\left(\int_0^1 \phi(u)\psi(u)du\right)^2}{\int_0^1 \phi^2(u)du \int_0^1 \psi^2(u)du} .$$

However, this is the same expression one gets when studying the rank score tests in the case of the two sample problem. If G is not normal then it is seen (see section 6, Hájek (1962)) that under H_{1n} the noncentrality parameters of F_n and T_n are $\rho^{*2}\Delta^2$ and $\rho^2\Delta^2$. Chernoff and Savage (1958) proved that $\rho^* \leq \rho$ and the strict equality holds if and only if G is normal. The values of ρ^2 in other situations are rather wellknown and are not given. (See for example Hodges-Lehmann (1956, 1961)).

3.3. Model II: The Likelihood Ratio Test, Rankscore tests and mixed rankscore tests.

In Model II if the underlying distributions are normal then the likelihood ratio test for testing the hypothesis $\beta=0$ is

based on the statistic

$$3.3.1 \qquad t_N = b_N (N-2)^{1/2} \left[\sum_{i=1}^{N} (X_i - \bar{X})^2 / \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \right]^{1/2},$$

where b_N is the estimate of the regression coefficient given by the normal equations. The statistic t_N has Student's t-distribution with N-2 degrees of freedom. The distribution of t_N is approximately normal for large N even if the underlying distributions are not normal. The test based on t_N can be seen to be asymptotically equivalent to that based on the covariance

$$3.3.2 \qquad W_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{\eta \sigma_o},$$

and using the same notation as in section 2.3 it is seen that

$$3.3.3 \qquad \mathcal{L}(W_N \mid H_{1N}) \longrightarrow \mathcal{N}(\mu, 1),$$

with $\mu = \beta_o / \sigma_o$ and

$$3.3.4 \qquad \left[\frac{d\mu}{d\beta_o} \Big|_{\beta_o = 0} \right]^2 = \frac{1}{\sigma_o^2}.$$

Now we are ready to make comparisons between various tests. Throughout H_o will stand for $\alpha = \alpha_o$, $\beta = 0$, $\sigma = \sigma_o$, H_1 for $\alpha = \alpha_o$, $\beta = \beta_o > 0$ and $\sigma = \sigma_o$, and H_{1n} for $\alpha = \alpha_o$, $\beta = \beta_o / \sqrt{n} > 0$ and $\sigma = \sigma_o$. All the tests considered are one sided with the critical region chosen with the help of normal tables. Hence while discussing asymptotic efficiencies of the tests we will talk about the test statistics rather than tests themselves. The

symbol $\underset{\sim}{e}_{T_N,S_N}$ will stand for the Pitman efficiency of T_N relative to S_N. (For the definition and the meaning of the Pitman efficiency see Noether (1954)).

a) Rankscore versus the likelihood ratio test.

From the expressions (2.4.15), (2.4.16), (3.3.3) and (3.3.4) it follows that

$$3.3.5 \qquad \underset{\sim}{e}_{T_N,W_N} = \frac{\left[\int_{-\infty}^{\infty} x J[F(x)]\,dF(x)\right]^2}{\left[\int_0^1 J^2(u)du\right]} \frac{\left[\int_{-\infty}^{\infty} L[G(y)]\,g'(y)dy\right]^2}{\left[\int_0^1 L^2(v)dv\right]}$$

where T_N and W_N are defined in (2.4.2) and (3.3.2) respectively.

The second factor on the right appears in studying the rankscore tests for the two sample problem.

Let $L = \Phi^{-1}$, where Φ is the distribution function of a standard normal random variable. Then Chernoff and Savage (1958) have proved that this factor is always larger than or equal to unity, the equality holding if and only if G is normal.

b) Mixed Rankscore versus the likelihood ratio test.

First consider the mixed normal score test. In this case the corresponding ϕ function is the inverse function of the normal distribution function and

$$3.3.6 \qquad \int_0^1 \phi^2(u)du = 1.$$

Assuming (2.3.3) it is seen from (2.3.17) and (3.3.4)

that the statistics W_N and S_N given by (3.3.2) and (2.3.4) respectively have the same asymptotic normal distribution under H_{1n} and hence the mixed rank score test is efficient.

If the underlying distribution is not normal and, is G say, then from the equations (2.3.22) and assuming that $\eta^2=1$

$$3.3.7 \qquad \mathcal{L}(S_N \mid H_{1n}) \longrightarrow \mathcal{N}(\frac{\beta_o}{\sigma_o}\mu_{11}, \xi^2),$$

where

$$3.3.8 \qquad \mu_{11} = \int_0^1 \Phi^{-1}(u)\psi(u)du = \int_{-\infty}^{\infty} \Phi^{-1}[G(x)]\, g'(x)dx,$$

$$3.3.9 \qquad \psi(u) = -\frac{g'[G^{-1}(u)]}{g[G^{-1}(u)]}, \qquad 0<u<1,$$

and

$$3.3.10 \qquad \xi^2 = \int_0^1 [\Phi^{-1}(u)]^2 du = 1.$$

Hence from Chernoff-Savage (1958) it follows that,

$$3.3.11 \qquad \rho_{S_N,W_N} = \left[\int_{-\infty}^{\infty} \Phi^{-1}[G(x)]\, g'(x)dx\right]^2 \geq 1.$$

In general the asymptotic efficiency of a mixed rankscore test based on S_N relative to the likelihood ratio test can be written as

$$3.3.12 \qquad \Xi_{S_N,W_N} = \frac{\left[\int_{-\infty}^{\infty} \phi[G(y)]\, g'(y)dy\right]^2}{\int_0^1 \phi^2(u)du},$$

where G is the underlying distribution and ϕ based on the

distribution function F is used.

c) Mixed rankscore versus rankscore.

Rewrite the expression (3.3.5) in terms of ϕ and ψ functions,

$$3.3.13 \qquad \underset{\sim}{e}_{T_N,W_N} = \frac{\left[\int_{-\infty}^{\infty} x\,\psi[F(x)]\,dF(x)\right]^2}{\int_0^1 \psi^2(u)du} \cdot \frac{\left[\int_{-\infty}^{\infty} \phi[G(y)]g'(y)dy\right]^2}{\int_0^1 \phi^2(v)dv}.$$

Comparing this with (3.3.12) it follows that

$$3.3.14 \qquad \underset{\sim}{e}_{T_N,S_N} = \frac{\left[\int_{-\infty}^{\infty} x\psi[F(x)]\,dF(x)\right]^2}{\int_0^1 \psi^2(u)du}.$$

It will now be shown that

$$3.3.15 \qquad \underset{\sim}{e}_{T_N,S_N} \leq 1.$$

Note that when

$$3.3.16 \qquad \phi(u) = \frac{g'\left[G^{-1}(u)\right]}{g\left[G^{-1}(u)\right]},$$

it is seen that S_N is equivalent to a locally asymptotically most powerful test and hence

$$3.3.17 \qquad \underset{\sim}{e}_{T_N,S_N} = \frac{\left[\int_{-\infty}^{\infty} x\psi[F(x)]\,dF(x)\right]^2}{\int_0^1 \psi^2(u)du} \leq 1.$$

However, e_{T_N,S_N} being independent of ϕ and G, (3.3.16) holds in general.

3.4 Table showing the asymptotic efficiencies of the analogues of some wellknown tests for Model II.

F : the distribution function of X.

G : the distribution function of Y.

R_ν : the rank of X_ν among $X_1,\cdots,X_n$.

S_ν : the rank of Y_ν among $Y_1,\cdots,Y_n$.

Φ : the standard normal distribution function.

$E_{N,i}$: the normal score for the rank i.

(The efficiencies should be read as column-relative-to-row).

Test Statistic	(1)	(2)	(3)	(4)
1) $\sum (X_\nu-\bar{X})(Y_\nu-\bar{Y})$				
2) $\sum R_\nu S_\nu$	$\frac{\pi^2}{9}$ if $F=G=\Phi$			
3) $\sum X_\nu S_\nu$	$\frac{\pi}{3}$ if $G=\Phi$	≤ 1		
4) $\sum E_{N,R_\nu} E_{N,S_\nu}$	1 if $F=G=\Phi$	$\frac{\pi^2}{9}$ if $F=G=\Phi$	$\frac{\pi}{3}$ if $F=G=\Phi$	
5) $\sum X_\nu E_{N,S_\nu}$	≤ 1	≤ 1 if $G=\Phi$	same as Wilcoxon vs. Normal score in the two sample problem.	≤ 1

4. Acknowledgements. I am grateful to Professor E.L.Lehmann for proposing the problem and for his continued guidance. I also wish to thank Dr.S.Bhuchongkul and Professor F.C.Andrews

for their many suggestions and the referee for the valuable comments which resulted in improved style and some simplifications of the arguments.

REFERENCES

[1] BHUCHONGKUL, S. (1964). A class of nonparametric tests for independence in bivariate populations. Ann.Math.Statist. 35 138-149.

[2] CHERNOFF, H. and SAVAGE, I.R. (1958). Asymptotic normality and efficiency of certain non-parametric test statistics. Ann.Math.Statist. 29, 972-994.

[3] CRAMÉR, HARALD (1951). Mathematical Methods of Statistics, Princeton University Press.

[4] DHARMADHIKARI, S.W. and JOGDEO, K. (1964). Behaviour of the maximum. (Unpublished)

[5] HÁJEK, JAROSLAV (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. Ann.Math.Statist. 32 506-523.

[6] HÁJEK, JAROSLAV (1962). Asymptotically most powerful rankorder tests. Ann.Math.Statist. 33 1124-1147.

[7] KONIJN, H.S. (1956). On the power of certain tests for independence in bivariate populations. Ann.Math.Statist. 27 300-323.

[8] HODGES, J.L. Jr. and LEHMANN, E.L. (1956). The efficiency of some non-parametric competitors of the t-test. Ann.Math. Statist. 27 324-335.

[9] HODGES, J.L. Jr. and LEHMANN, E.L. (1961). Comparison of the normal scores and Wilcoxon tests. Proc.Fourth Berkeley Symp. Math. Statist. Prob. 1 307-317.

[10] LE CAM, L. (1960). Locally asymptotically normal families

of distributions. University of California Publications in Statistics. 3 No.2 37-98.

[11] MIKULSKI P.W. (1963). On the efficiency of optimal non-parametric procedures in the two sample case. Ann.Math. Statist. 34 22-32.

[12] NEYMAN, J. (1958). Optimal asymptotic tests of composite statistical hypotheses. The H.Cramér Jubilee Volume, Almquist and Wiksell, Uppsala.

[13] NOETHER, G.E. (1954). On a theorem of Pitman. Ann.Math. Statist. 25 514-522.

[14] PURI, M.L. (1964).Asymptotic efficiency of a class of c-sample tests. Ann.Math.Statist. 35 102-121.