SOME NONPARAMETRIC MEDIAN PROCEDURES

By V. P. Bhapkar

University of North Carolina and University of Poona

1. Introduction and summary. Under the nonparametric approach, various methods have been suggested to avoid the assumption of normality and homoscedasticity implicit in the analysis of variance. For the one-way classification, i.e., to decide whether c samples come from the same population, Kruskal and Wallis [9] have proposed the H-test based on ranks; Mood and Brown [10] have proposed the M-test, utilizing the numbers of observations above the median of the combined sample; while the present author [2] has offered the V-test based on the number of c-plets that can be formed by choosing one observation from each sample such that the observation from the kth sample is the least, $k = 1, 2, \cdots, c$.

For the two-way classification, Friedman [8] has made use of ranks. His statistic, to test the hypothesis that the rankings by m "observers" of n "objects" are independent, essentially offers a test for the two-way classification with one observation per cell. Durbin [7] has given a generalization for the balanced incomplete block design. Benard and Van Elteren [1] have generalized it still further. Mood and Brown [6, 10] also have considered the two-way classification with one observation per cell or the same number of observations per cell. In the first part of the present paper, their test has been extended to cover incomplete block situations.

Mood and Brown [6, 10] have also considered some simple regression problems. In the present paper their methods are extended to discuss some additional regression problems. Next some bivariate analysis of variance problems are considered. The "step-down procedure" [11, 12] is used to reduce the problem to the univariate case with the other variate as a concomitant variate. The regression methods developed earlier are used here in these bivariate problems. The method seems to be perfectly general and could be extended to three or more variates, that is, to the multivariate situation. Most of the tests are offered on heuristic considerations. They are expected to be good for large samples.

2. Extension of Mood's test for the two-way classification to incomplete designs. Mood and Brown [10] have considered a test for equality of "row" effects in the usual setup with r rows, c columns and one observation per cell, say x_{ij} in the ijth cell. The distribution of the x_{ij} is assumed to have median $v_{ij} = \alpha_i + \beta_j + \mu$, where the median of the α's is zero as is the median of the β's. By the median we shall always mean the middle or the average of the two middle quantities. The distributions are assumed to be continuous and identical, except for location.

Under the null hypothesis that the row effects, α's, are equal (i.e., zero), all the observations in a given column have the same distribution. Let $\tilde{x}_j$ be the

Received September 17, 1959; revised September 1, 1960.

median of the observations in the jth column, and form a two-way table by replacing the observation x_{ij} by a plus sign if it exceeds $\tilde{x}_j$, or by a minus sign if it does not. Let m_i be the number of plus signs in the ith row. The test criterion used in [10] is

$$\chi^2 = \frac{r(r-1)}{ca(r-a)} \sum_{i=1}^{r} \left(m_i - \frac{ca}{r} \right)^2, \tag{1}$$

where $a = \frac{1}{2}r$ if r is even or $\frac{1}{2}(r-1)$ if r is odd. Unless c is small, the χ^2 approximation with $r-1$ d.f. is used. It is suggested [10] that for practical purposes the large sample distribution is satisfactory if $c \geqq 10$ or even if $c = 5$ provided $rc \geqq 20$. For smaller c, exact probabilities could be computed. We shall consider the generalization to incomplete blocks.

Let us write $n_{ij} = 1$ if the (ij) combination is allowed and zero otherwise. Let the number of observations in the ith row be c_i $(i = 1, 2, \cdots, r)$ and in the jth column be r_j $(j = 1, 2, \cdots, c)$. Let $a_j = \frac{1}{2}r_j$ if r_j is even or $\frac{1}{2}(r_j - 1)$ if r_j is odd. Then there are a_j plus signs in the jth column. Let the m_i be defined as before. Then we expect (under H_0) m_i to be approximately equal to $\frac{1}{2}c_i$.

Following Mood, let us derive the generating function to find the distribution of the m's. Suppose t_i is associated with a plus sign in the ith row $(i = 1, 2, \cdots, r)$. Let $\phi_{a_j}(t_1, \cdots, t_{r_j})$ consist of the sum of all terms that can be formed by multiplying the t's together, a_j at a time. Each term of ϕ describes a possible arrangement of signs in a given column. Furthermore, each arrangement of signs is equally likely; hence the probability of a particular arrangement is $1 \Big/ \binom{r_j}{a_j}$.

Suppose the jth column contains observations in the $j_1, j_2, \cdots, j_{r_j}$th rows. Then consider the function

$$\Phi = \prod_{j=1}^{c} \frac{\phi_{a_j}(t_{j_1}, \cdots, t_{j_{r_j}})}{\binom{r_j}{a_j}}.$$

There is a one-to-one correspondence between the ways of getting terms $t_1^{m_1} t_2^{m_2} \cdots t_r^{m_r}$ in the numerator of Φ and the arrangements of signs in the $r \times c$ table which gives rise to m_i plus signs in the ith row $(i = 1, 2, \cdots, r)$. Hence

$$\Phi = \sum_{m_1} \sum_{m_2} \cdots \sum_{m_r} g(m_1, \cdots, m_r) t_1^{m_1} t_2^{m_2} \cdots t_r^{m_r},$$

where g is the frequency function for the m's.

Note that $\phi_{a_j}(1, 1, \cdots, 1) = \binom{r_j}{a_j}$. Φ is thus the factorial-moment generating function for the m's. Then $\mathcal{E}(m_i) = \partial\Phi/\partial t_i$ with all the t's $= 1$.
We note that

$$\frac{\partial \phi_{a_j}}{\partial t_i}(t_{j_1}, \cdots, t_{j_{r_j}}) = 0, \qquad \text{if } n_{ij} = 0,$$

$$= \phi_{a_j - 1}(t_{j_1}, \cdots, t_{j_{r_j}}), \qquad \text{if } n_{ij} = 1,$$

where one of the t's from the previous bracket is missing. Hence,

$$(2)\qquad \frac{\partial\Phi}{\partial t_i} = \left[\prod_{j=1}^{c}\binom{r_j}{a_j}\right]^{-1}\left[\sum_{j=1}^{c}\left\{\prod_{j'\neq j}\phi_{a_{j'}}(t_{j'_1},\cdots,t_{j'_{r_{j'}}})\right\}\cdot n_{ij}\,\phi_{a_j-1}(t_{j_1}\cdots)\right].$$

Then,

$$(3)\qquad \mathcal{E}(m_i) = \left[\prod_{j=1}^{c}\binom{r_j}{a_j}\right]^{-1}\left[\sum_{j=1}^{c}\left\{\prod_{j'\neq j}\binom{r_{j'}}{a_{j'}}\right\} n_{ij}\binom{r_j-1}{a_j-1}\right] = \sum_{j=1}^{c} n_{ij}\,\frac{a_j}{r_j}.$$

Similarly,

$$\sigma_{ii} = \operatorname{var}(m_i) = \left[\frac{\partial^2\Phi}{\partial t_i^2}\right]_{t=1} + \mathcal{E}(m_i) - [\mathcal{E}(m_i)]^2.$$

From (2) we have

$$\frac{\partial^2\Phi}{\partial t_i^2} = \left[\prod_{j=1}^{c}\binom{r_j}{a_j}\right]^{-1}\left[\sum_{j=1}^{c} n_{ij}\,\phi_{a_j-1}(t_{j_1},\cdots,t_{j_{r_j}})\cdot\sum_{j'\neq j}\frac{\partial}{\partial t_i}\phi_{a_{j'}}\cdot(t_{j'_1},\cdots,t_{j'_{r_{j'}}})\prod_{k\neq j\neq j'}\phi_{a_k}(t_{k_1},\cdots,t_{k_{r_k}})\right].$$

Hence,

$$\begin{aligned}\left[\frac{\partial^2\Phi}{\partial t_i^2}\right]_{t=1} &= \left[\prod_{j=1}^{c}\binom{r_j}{a_j}\right]^{-1}\left[\sum_{j=1}^{c} n_{ij}\binom{r_j-1}{a_j-1}\sum_{j'\neq j} n_{ij'}\binom{r_{j'}-1}{a_{j'}-1}\prod_{k\neq j\neq j'}\binom{r_k}{a_k}\right]\\ &= \sum_{j=1}^{c}\sum_{j'\neq j} n_{ij}\,n_{ij'}\,\frac{a_j}{r_j}\,\frac{a_{i'}}{r_{j'}}\\ &= \left[\sum_{j=1}^{c} n_{ij}\,\frac{a_j}{r_j}\right]^2 - \sum_{j=1}^{c} n_{ij}^2\,\frac{a_j}{r_j^2},\end{aligned}$$

so that

$$(4)\qquad \sigma_{ii} = \sum_{j=1}^{c} n_{ij}\left(\frac{a_j}{r_j} - \frac{a_j}{r_j^2}\right).$$

Similarly,

$$\sigma_{ii'} = \operatorname{cov}(m_i, m'_i) = \left[\frac{\partial^2\Phi}{\partial t_i\,\partial t_{i'}}\right]_{t=1} - \mathcal{E}(m_i)\mathcal{E}(m_{i'})$$

From (2) we have

$$\frac{\partial^2\Phi}{\partial t_i\,\partial t_{i'}} = \left[\prod_{j=1}^{c}\binom{r_j}{a_j}\right]^{-1}\cdot\left[\sum_{j=1}^{c} n_{ij}\left\{\prod_{j'\neq j}\phi_{a_{j'}}\,n_{i'j}\,\phi_{a_j-2} + \phi_{a_j-1}\sum_{j'\neq j} n_{i'j'}\,\phi_{a_{j'}-1}\prod_{k\neq j\neq j'}\phi_{a_k}\right\}\right].$$

Hence,

$$\left[\frac{\partial^2\Phi}{\partial t_i\,\partial t_{i'}}\right]_{t=1} = \sum_{j=1}^{c} n_{ij}\,n_{i'j}\,\frac{a_j(a_j-1)}{r_j(r_j-1)} + \sum_{j=1}^{c} n_{ij}\,\frac{a_j}{r_j}\cdot\sum_{j'\neq j} n_{i'j'}\,\frac{a_{j'}}{r_{j'}},$$

so that

$$\sigma_{ii'} = -\sum_{j=1}^{c} n_{ij}\,n_{i'j}\,\frac{a_j}{r_j^2}\,\frac{r_j-a_j}{r_j-1}. \tag{5}$$

Asymptotic normality. We have

$$\begin{aligned}\Phi(\mathbf{t}) &= \left[\prod_{j=1}^{c}\binom{r_j}{a_j}\right]^{-1}\prod_{j=1}^{c}\phi_{a_j}(t_{j_1},\cdots,t_{j_{r_j}})\\ &= \sum_{m_1}\cdots\sum_{m_r} g(m_1,\cdots,m_r)\,t_1^{m_1}\cdots t_r^{m_r}.\end{aligned}$$

Replacing t_i in Φ by $\exp(s_i c_i^{-\frac{1}{2}})$, we have

$$\begin{aligned}\Phi'(\mathbf{s}) &= \sum_{m_1}\cdots\sum_{m_r} g(m_1,\cdots,m_r)\exp\sum_i m_i s_i c_i^{-\frac{1}{2}}\\ &= \text{moment generating function of } m_i c_i^{-\frac{1}{2}}\text{'s}.\end{aligned}$$

Let us consider $\log\Phi'$ for large c. We assume that $c_i/c \to \epsilon_i > 0$ as $c \to \infty$. We have

$$\log\Phi' = \sum_{j=1}^{c}\log\left[\Phi'_{a_j}\Big/\binom{r_j}{a_j}\right]. \tag{6}$$

Now

$$\phi'_{a_j} = \sum\exp.\left[\sum_{i=1}^{a_j} s_{j_{k_i}}\,c_{j_{k_i}}^{-\frac{1}{2}}\right],$$

where the summation is over $\binom{r_j}{a_j}$ combinations of type $k_1, k_2, \cdots, k_{a_j}$ out of $(1, 2, \cdots, r_j)$. Hence,

$$\begin{aligned}\binom{r_j}{a_j}^{-1}\phi'_{a_j} &= \binom{r_j}{a_j}^{-1}\sum\left\{1+\sum_{i=1}^{a_j} s_{j_{k_i}}\,c_{j_{k_i}}^{-\frac{1}{2}} + \frac{1}{2}\left[\sum_{i=1}^{a_j} s_{j_{k_i}}\,c_{j_{k_i}}^{-\frac{1}{2}}\right]^2 + O(c^{-\frac{3}{2}})\right\}\\ &= \binom{r_j}{a_j}^{-1}\left[\binom{r_j}{a_j} + \sum_{i=1}^{r} n_{ij}\binom{r_j-1}{a_j-1}s_i\,c_i^{-\frac{1}{2}}\right.\\ &\quad + \frac{1}{2}\sum_{i=1}^{r} n_{ij}\binom{r_j-1}{a_j-1}s_i^2\,c_i^{-1} + \frac{1}{2}\sum_{i'}\sum_{\neq i} n_{ij}\,n_{i'j}\binom{r_j-2}{a_j-2}\\ &\qquad\left.\cdot\,s_i\,s_{i'}\,c_i^{-\frac{1}{2}}\,c_{i'}^{-\frac{1}{2}} + O(c^{-\frac{3}{2}})\right]\\ &= 1+\sum_{i=1}^{r} n_{ij}\,\frac{a_j}{r_j}\,\frac{s_i}{c_i^{\frac{1}{2}}} + \frac{1}{2}\sum_{i=1}^{r} n_{ij}\,\frac{a_j}{r_j}\,\frac{s_i^2}{c_i}\\ &\quad + \frac{1}{2}\sum_{i'}\sum_{\neq i} n_{ij}\,n_{i'j}\,\frac{a_j(a_j-1)}{r_j(r_j-1)}\,\frac{s_i\,s_{i'}}{(c_i\,c_{i'})^{\frac{1}{2}}} + O(c^{-\frac{3}{2}}),\end{aligned}$$

so that

$$\log \binom{r_j}{a_j}^{-1} \phi'_{a_j} = \sum_{i=1}^{r} n_{ij} \frac{a_j}{r_j} \frac{s_i}{c_i^{\frac{1}{2}}} + \frac{1}{2} \sum_{i=1}^{r} n_{ij} \frac{a_j}{r_j} \frac{s_i^2}{c_i}$$

$$+ \frac{1}{2} \sum_{i'} \sum_{\neq i} n_{ij} n_{i'j} \frac{a_j(a_j - 1)}{r_j(r_j - 1)} \frac{s_i s_{i'}}{(c_i c_{i'})^{\frac{1}{2}}} - \frac{1}{2} \left[\sum_{i=1}^{r} n_{ij} \frac{a_j}{r_j} \frac{s_i}{c_i^{\frac{1}{2}}} \right]^2 + O(c^{-\frac{3}{2}}).$$

Then, from (6) we have

$$\log \Phi' = \sum_{i=1}^{r} \mathcal{E}(m_i)\, s_i c_i^{-\frac{1}{2}} + \frac{1}{2} \sum_{i=1}^{r} \sigma_{ii}\, s_i^2 c_i^{-1}$$

$$+ \frac{1}{2} \sum_{i'} \sum_{\neq i} \sigma_{ii'}\, s_i s_{i'} (c_i c_{i'})^{-\frac{1}{2}} + O(c^{-\frac{3}{2}}).$$

Thus for large c, we have the distribution of $m_i c_i^{-\frac{1}{2}}$'s approximated by the multivariate normal distribution. Since $\sum_{i=1}^{r} m_i = \sum_{j=1}^{c} a_j$, it follows that the m's are linearly dependent. Hence the above distribution is singular. Considering only $m_1, m_2, \cdots, m_{r-1}$, which have an asymptotically nonsingular normal distribution, we shall have a chi-square criterion with $r - 1$ d.f., given by

$$\chi^2 = \sum_{i=1}^{r-1} \sum_{i'=1}^{r-1} [m_i - \mathcal{E}(m_i)][m_{i'} - \mathcal{E}(m_{i'})] \sigma^{ii'}_{(rr)}, \tag{7}$$

where $[\sigma^{ii'}_{(rr)}] = \mathbf{\Sigma}^{-1}_{(rr)}$, $\mathbf{\Sigma}_{(rr)}$ being the cofactor of σ_{rr} in $[\sigma_{ii'}]$.

Special case. Suppose $c_1 = c_2 = \cdots = c_r = c_0$, say, and $r_1 = r_2 = \cdots = r_c = r_0$, say. Then $a_1 = a_2 = \cdots = a_c = a_0$, say, where $a_0 = \frac{1}{2}r_0$ if r_0 is even and $\frac{1}{2}(r_0 - 1)$ otherwise. Also $rc_0 = cr_0$. Then from (3), (4) and (5) $\mathcal{E}(m_i) = a_0 c_0 / r_0$, $\sigma_{ii} = c_0 a_0 (r_0 - a_0)/r_0^2$,

$$\sigma_{ii'} = -\, a_0(r_0 - a_0)\lambda_{ii'}/r_0^2(r_0 - 1), \qquad i' \neq i,$$

where $\lambda_{ii'} = \sum_j n_{ij} n_{i'j}$.

(i) *Balanced incomplete block designs.* Let $\lambda_{ii'} = \lambda$ for all (ii'), $(i \neq i')$. Then we have $c_0(r_0 - 1) = \lambda(r - 1)$, $\sigma_{ii} = c_0 a_0 (r_0 - a_0)/r_0^2 = \alpha$, say, and $\sigma_{ii'} = -a_0(r_0 - a_0)\lambda/r_0^2(r_0 - 1) = \beta$, say. Let $\mathbf{I}_r$ denote the unit matrix of order r and $\mathbf{J}_r$ denote the matrix $[1]_{r \times r}$. Thus, $\mathbf{\Sigma}_{(rr)} = (\alpha - \beta)\, \mathbf{I}_{r-1} + \beta\, \mathbf{J}_{r-1}$. Then

$$\mathbf{\Sigma}^{-1}_{(rr)} = \frac{1}{\alpha - \beta} \mathbf{I}_{r-1} - \frac{\beta}{(\alpha - \beta)[\alpha + (r - 2)\beta]} \mathbf{J}_{r-1}$$

$$= \gamma[\mathbf{I}_{r-1} + \mathbf{J}_{r-1}],$$

where $\gamma = r_0^2(r_0 - 1)/a_0(r_0 - a_0)\lambda r$. Let

$$\mathbf{z}'_{1 \times (r-1)} = [\{m_i - (a_0 c_0 / r_0)\},\ i = 1, 2, \cdots, r - 1].$$

Then from (7), $\chi^2 = \mathbf{z}' \mathbf{\Sigma}^{-1}_{(rr)} \mathbf{z} = \gamma[\mathbf{z}'\mathbf{z} + \mathbf{z}' \mathbf{J}_{r-1}\, \mathbf{z}]$. Now

$$\mathbf{z}' \mathbf{J}_{r-1}\, \mathbf{z} = \left(\sum_{i=1}^{r-1} z_i \right)^2 = \left[\sum_{i=1}^{r-1} \{m_i - (a_0\, c_0 / r_0)\} \right]^2$$

$$= \{m_r - (c_0 a_0/r_0)\}^2.$$

Hence

$$\chi^2 = \frac{r_0^2(r_0 - 1)}{a_0(r_0 - a_0)\lambda r} \sum_{i=1}^{r} \left(m_i - \frac{c_0 a_0}{r_0}\right)^2. \tag{8}$$

If we put $\lambda = c$ and hence $r_0 = r$, $c_0 = c$ and $a_0 = a$, we get back to (1).
In the usual terminology of the BIBD, if the "rows" denote the "treatments" and the "columns" denote the "blocks", then

r = number of "treatments" = v,
c = number of "blocks" = b,
c_0 = the number of replications of any "treatment" = r,
r_0 = the number of "treatments" in any "block" = k,
λ = the number of times any two "treatments" occur together in the same "block" = λ. (8) then reduces to

$$\chi^2 = \frac{k^2(k - 1)}{a(k - a)\lambda v} \sum_{i=1}^{v} \left(m_i - \frac{ra}{k}\right)^2, \tag{9}$$

where $a = \frac{1}{2}k$ if k is even and $\frac{1}{2}(k - 1)$ otherwise.

(ii) *Partially balanced incomplete block designs.* Let us consider rows as treatments, so that $\lambda_{ii'} = \lambda_p$ if i and i' are pth associates. Then

$$\mathbf{\Sigma} = \alpha \mathbf{I}_r + \sum_{p=1}^{m} \beta_p \mathbf{B}_p,$$

where m is the number of associate classes, α is defined as before, $\mathbf{B}$'s are the association matrices [4] and

$$\beta_p = -\frac{a_0(r_0 - a_0)}{r_0^2(r_0 - 1)} \lambda_p.$$

Using the results derived in [3] and simplifying, we have

$$\chi^2 = \frac{r_0(r_0 - 1)}{a_0(r_0 - a_0)} \sum_{i=1}^{r} \sum_{i'=1}^{r} c_{ii'} \left(m_i - \frac{c_0 a_0}{r_0}\right)\left(m_{i'} - \frac{c_0 a_0}{r_0}\right),$$

where $\mathbf{C} = (c_{ii'})$ is such that the solution of the "normal equations" for $\mathbf{t}$ in the analysis of variance for the PBIBD is given by $\mathbf{t} = \mathbf{CQ}$, $\mathbf{Q}$ being defined in the usual notation [5].

In the usual terminology of the PBIBD, as indicated in the terminology for the BIBD, we have

$$\chi^2 = \frac{k(k - 1)}{a(k - a)} \sum_{i=1}^{v} \sum_{i'=1}^{v} c_{ii'} \left(m_i - \frac{ra}{k}\right)\left(m_{i'} - \frac{ra}{k}\right). \tag{10}$$

For the PBIBD with two associate classes, the constants c_1 and c_2 (i.e., $c_{ii'}$'s) are already given in [5].

3. Some regression problems. We shall first state a lemma [10] which will be useful for later applications.

LEMMA. *Let*

$$g(m_1, m_2, \cdots, m_k) = \frac{\prod_{i=1}^{k} \binom{n_i}{m_i}}{\binom{n}{m}}, \tag{11}$$

where $n = \sum_{i=1}^{k} n_i$ *and* $m = \sum_{i=1}^{k} m_i$ *denote the frequency function for the* m*'s. Then as* $n \to \infty$ *in such a way that* $n_i/n \to p_i > 0$,

$$\chi^2 = \frac{n(n-1)}{m(n-m)} \sum_{i=1}^{k} \frac{1}{n_i} \left(m_i - \frac{n_i m}{n} \right)^2$$

has the asymptotic χ^2 *distribution with* $k - 1$ d.f.

Mood [10] says, "The expression (11) has a distribution very closely approximated by the chi-square distribution with $k - 1$ d.f. even if n is only of the order of twenty provided all the n_i are at least five".

3.1 *One sample.* Let $(x_1, y_1), \cdots, (x_n, y_n)$ denote a sample of n observations. We shall assume that

(a) the distribution of y for any x is continuous and identical apart from a shift or translation, and

(b) the regression is linear, that is, the location parameter (usually the median), given x, is $\alpha + \beta x$, where α and β are unknown parameters.

To estimate α and β, Mood and Brown [10] suggest that the estimates $\hat{\alpha}$ and $\hat{\beta}$ should be determined by

$$\text{Median of } (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \quad \text{for} \quad x_i \leqq \tilde{x} \tag{12}$$

and

$$\text{Median of } (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \quad \text{for} \quad x_i > \tilde{x}, \tag{13}$$

where $\tilde{x}$ is the median of the x's. If it happens that several x values fall at $\tilde{x}$, then the sign $\leqq$ in (12) and $>$ sign in (13) may be replaced by $<$ and $\geqq$ if such a replacement would more nearly divide the points into groups of equal size. They also give an iteration procedure to determine $\hat{\alpha}$ and $\hat{\beta}$.

We shall find it convenient to speak of $x_i \leqq \tilde{x}$ as the group I and of $x_i > \tilde{x}$ as the group II. Then (12) and (13) may be equivalently written as

$$\hat{\alpha} = \text{Median } (y_i - \hat{\beta}x_i) \tag{14}$$

and

$$\text{Median}_{\text{I}}(y_i - \hat{\beta}x_i) = \text{Median}_{\text{II}}(y_i - \hat{\beta}x_i). \tag{15}$$

Test for $\beta = \beta_0$. If $\beta = \beta_0$, α is estimated by $\hat{\alpha} = \text{Median } (y_i - \beta_0 x_i)$. Mood and Brown consider the number of points, say m_1 and m_2, above the line $y =$

$\hat{\alpha} + \beta_0 x$ in each group. Let us, for convenience, assume that n is even. Then the frequency function of m_1 and m_2 is given by

$$p(m_1, m_2) = \frac{\binom{\frac{1}{2}n}{m_1}\binom{\frac{1}{2}n}{m_2}}{\binom{n}{\frac{1}{2}n}}, \tag{16}$$

so that, by the lemma, they obtain

$$\chi^2 = \frac{16}{n}\left(m_1 - \frac{n}{4}\right)^2, \qquad \text{d.f.} = 1, \tag{17}$$

as the test-statistic. It may be seen that the supposition that n be even may be relaxed.

We may arrive at (17) by some heuristic considerations. Assuming n is even, as before, we have $n/2$ points in each group and we note that $m_1 + m_2 \doteq n/2$. If the hypothesis is true, we expect m_1 and m_2 to be approximately $n/4$. Now m_1 is the number of positive $y_i - \hat{\alpha} - \beta_0 x_i$'s from the first group and, similarly, for m_2. Now the $y_i - \alpha - \beta_0 x_i$'s have identical distributions and, also, $\hat{\alpha} - \alpha \xrightarrow{(p)} 0$ as $n \to \infty$, so that, on heuristic considerations,

$$p(m_1, m_2) \sim \frac{\binom{\frac{1}{2}n}{m_1}\binom{\frac{1}{2}n}{m_2}}{\binom{n}{\frac{1}{2}n}}$$

for large n and, by the lemma, we again have the asymptotic χ^2 statistic given by (17).

If we are willing to assume, in addition to (a) and (b), that

(c) the mean and variance of y exist for any x, then taking the mean as a location parameter given by $\alpha + \beta x$, α and β can be immediately estimated by the usual least squares estimators. In the above case, $\hat{\alpha} = \bar{y} - \beta_0 \bar{x}$, where $\bar{y}$ is the mean of the y's and similarly for $\bar{x}$. Then also $\hat{\alpha} - \alpha \xrightarrow{(p)} 0$. In this case, if b denotes the number of points above the regression line, we have by a similar heuristic argument

$$p(m_1, m_2) \sim \frac{\binom{\frac{1}{2}n}{m_1}\binom{\frac{1}{2}n}{m_2}}{\binom{n}{b}}$$

for large n and where m_1 and m_2 are defined as before. Hence by the lemma we have an alternate test-statistic

$$\chi^2 = \{4n/[b(n-b)]\}(m_1 - \tfrac{1}{2}b)^2, \qquad \text{d.f.} = 1. \tag{18}$$

Consistency of $\hat{\alpha}$ and $\hat{\beta}$ determined by (14) *and* (15). Let $z_i = y_i - \alpha - \beta x_i$. Then the z's have identical distributions with median zero. Now (15) may be

written as

$$\text{Median}_{\text{I}}[z_i + (\beta - \hat{\beta})x_i] = \text{Median}_{\text{II}}[z_i + (\beta - \hat{\beta})x_i]. \tag{19}$$

Now as $n \to \infty$, $|\text{Median}_{\text{I}}(z_i) - \text{Median}_{\text{II}}(z_i)| \xrightarrow{(p)} 0$, so that intuitively it seems that $\hat{\beta} \sim \beta$ will satisfy (19), that is, $|\hat{\beta} - \beta| \xrightarrow{(p)} 0$. It has not been possible yet to give a general proof. We shall, however, give a proof for the case where there is a unit of measurement for x. This should cover most of the practical cases.

Proof for the special case. Let $\tilde{x}_n$, θ_{1n}, θ_{2n} and $\hat{\beta}_n$ denote the median of x's, $\text{Median}_{\text{I}}(z_i)$, $\text{Median}_{\text{II}}(z_i)$ and $\hat{\beta}$ respectively when the sample size is n. Let us suppose that (i) $x_i \leqq \tilde{x}_n$ form the group I and $x_i > \tilde{x}_n$ form the group II, and (ii) $x > x_0$ implies $x \geqq x_0 + \delta$, where δ is a fixed positive number, however small. [For example, δ may be in the nature of a unit of measurement.]

Since $\theta_{1n} \xrightarrow{(p)} 0$ and $\theta_{2n} \xrightarrow{(p)} 0$, given η, $\epsilon > 0$, there is an n_1 such that

$$|\theta_{2n}| < \epsilon \quad \text{and} \quad |\theta_{1n}| < \epsilon \quad \text{for} \quad n > n_1, \tag{20}$$

with probability greater than $1 - \eta$. Consider n greater than n_1. Let $\beta - \hat{\beta}_n = \theta_n$.

Case (1). Suppose $\theta_n \geqq 0$. Then

$$\text{Median}_{\text{I}}[z_i + (\beta - \hat{\beta}_n)x_i] \leqq \theta_{1n} + \theta_n \tilde{x}_n,$$

and

$$\text{Median}_{\text{II}}[z_i + (\beta - \hat{\beta}_n)x_i] \geqq \theta_{2n} + \theta_n(\tilde{x}_n + \delta).$$

Then (19) implies that $\theta_{2n} + \theta_n(\tilde{x}_n + \delta) < \theta_{1n} + \theta_n \tilde{x}_n$, so that $\theta_n \delta \leqq \theta_{1n} - \theta_{2n} < 2\epsilon$, from (20). Hence $\theta_n = |\theta_n| < 2\epsilon/\delta = \epsilon'$, say.

Case (2). Suppose $\theta_n < 0$. Then

$$\text{Median}_{\text{I}}[z_i + (\beta - \hat{\beta}_n)x_i] \geqq \theta_{1n} - |\theta_n| \tilde{x}_n,$$

and

$$\text{Median}_{\text{II}}[z_i + (\beta - \hat{\beta}_n)x_i] \leqq \theta_{2n} - |\theta_n| (\tilde{x}_n + \delta).$$

Again, (19) implies that $\theta_{1n} - |\theta_n| \tilde{x}_n \leqq \theta_{2n} - |\theta_n| (\tilde{x}_n + \delta)$, so that $|\theta_n| \delta \leqq \theta_{2n} - \theta_{1n} < 2\epsilon$, from (20). Hence, again, $|\theta_n| < 2\epsilon/\delta = \epsilon'$. Thus, given η and $\epsilon' > 0$, there is n_1 such that $|\theta_n| < \epsilon'$ with probability $> 1 - \eta$ for $n \geqq n_1$.

Thus $\theta_n \xrightarrow{(p)} 0$, that is, $\hat{\beta}_n \xrightarrow{(p)} \beta$, and the proof is complete for the special case mentioned above.

Consistency of $\hat{\alpha}$. Let us assume that $\hat{\beta} \xrightarrow{(p)} \beta$. Now $\hat{\alpha} = \text{Median}\ (y_i - \hat{\beta}x_i) = \alpha + \text{Median}\ [z_i + (\beta - \hat{\beta})x_i]$. We shall assume that the x's are bounded. Suppose $|x_i| < M$ for all i. Also $\hat{\beta} \xrightarrow{(p)} \beta$ implies that given ϵ, $\eta > 0$, there is an n^* such that $|\beta - \hat{\beta}| < \epsilon/M$ for all $n \geqq n^*$, with probability $> 1 - \eta$. Then

$\text{Median}\ (z_i) - \epsilon \leqq \text{Median}\ [z_i + (\beta - \hat{\beta})x_i] \leqq \text{Median}\ (z_i) + \epsilon$, for $n \geqq n^*$ with probability $> 1 - \eta$. Also $\text{Median}\ (z_i) \xrightarrow{(p)} 0$, so that $\text{Median}\ [z_i + (\beta - \hat{\beta})x_i] \xrightarrow{(p)} 0$. Thus, $\hat{\alpha} \xrightarrow{(p)} \alpha$.

Remarks. We may decide to take $x \leqq x_0$ as the group I and $x > x_0$ as the

group II (even though x_0 is not the median of the x's) in the equation (15) to estimate β, where x_0 is chosen suitably (preferably so as to divide the points into groups of approximately equal size). Then the above proof, with slight modifications, will go through if we assume, instead of (i) and (ii), that all the x's in the group II are greater than or equal to $x_0 + \delta$, where δ is a fixed positive number, however small. This would cover almost all the practical problems, δ being in the nature of a unit of measurement.

Then the test-statistic (17) can be modified suitably. Let a and $n - a$ be the number of points in the groups I and II respectively. If we define m_1 and m_2 as before, then $m_1 + m_2 = n/2$ (assuming n to be even). Then, on similar heuristic considerations,

$$p(m_1, m_2) \sim \frac{\binom{a}{m_1}\binom{n-a}{m_2}}{\binom{n}{n/2}},$$

so that by the lemma we have the asymptotic χ^2 statistic

$$\chi^2 = \{4n/[a(n-a)]\}(m_1 - \tfrac{1}{2}a)^2, \qquad \text{d.f.} = 1. \tag{21}$$

The supposition that n be even may now be relaxed.

3.2 *c samples.* Let us suppose that we have n_i independent observations (x_{ij}, y_{ij}), $j = 1, 2, \cdots, n_i$, from the ith population, $i = 1, 2, \cdots, c$. We shall assume (a) as before and (b) that the regression is linear, that is, the location parameter (usually the median) of y_{ij} given x_{ij}, is $\alpha_i + \beta_i x_{ij}$.

(i) *To test* $\beta_i = \beta_{i0}$, $i = 1, 2, \cdots, c$.

We shall have c independent χ^2 statistics with 1 d.f. each, giving the χ^2 statistic with c d.f. No new problem is presented here.

(ii) *To test* $\beta_1 = \beta_2 = \cdots = \beta_c$. On this hypothesis, y_{ij}'s have medians $\alpha_i + \beta x_{ij}$. We may estimate the α's and β by

$$\hat{\alpha}_i = \operatorname*{Median}_{j=1,2,\cdots,n_i} (y_{ij} - \hat{\beta}x_{ij})$$

and

$$\text{Median}_{\text{I}}(y_{ij} - \hat{\alpha}_i - \hat{\beta}x_{ij}) = \text{Median}_{\text{II}}(y_{ij} - \hat{\alpha}_i - \hat{\beta}x_{ij}).$$

For convenience, we shall take group I as $x \leqq \tilde{x}$ (the median of all the x's) and group II as $x > \tilde{x}$, though the test-statistic can be modified to suit other cases.

Let $\sum_1^c n_i = N$. For simplicity, let us take n_i to be even. Let m_i be the number of points from the ith sample belonging to the second group and l_i be the number of points out of these m_i that lie above $y = \hat{\alpha}_i + \hat{\beta}x$. Then $\sum_1^c m_i = \frac{1}{2}N$ and $\sum_1^c l_i \sim \frac{1}{4}N$. If the hypothesis is true, we expect l_i to be $\frac{1}{2}m_i$. Let l_i' be the number of observations from the m_i in the second group of the ith sample, such that $z_{ij} \equiv y_{ij} - \alpha_i - \beta x_{ij}$ is > 0. Since $\hat{\alpha}_i - \alpha_i \xrightarrow{(p)} 0$ and $\hat{\beta} - \beta \xrightarrow{(p)} 0$, $l_i - l_i' \xrightarrow{(p)} 0$ as n's $\to \infty$. Therefore, heuristically, the l's have the same distribution for

large n's as the l''s subject to $\sum_1^c l_i' \sim \frac{1}{4}N$. Since the z_{ij}'s have identical distributions,

$$p(l_1', \cdots, l_c') = \frac{\prod_1^c \binom{m_i}{l_i'}}{\binom{\frac{1}{2}N}{\frac{1}{4}N}},$$

so that

$$p(l_1, l_2, \cdots, l_c) \sim \frac{\prod_1^c \binom{m_i}{l_i}}{\binom{\frac{1}{2}N}{\frac{1}{4}N}}.$$

Hence, by the lemma we have

$$\chi^2 = 4 \sum m_i^{-1} (l_i - \tfrac{1}{2}m_i)^2 \qquad \text{d.f.} = c - 1.$$

If some $m_i = 0$, the corresponding term will be absent and d.f. will be reduced by one. Of course, as the referee has pointed out, if some m_i are small the approximation would be questionable. We could have considered the group I instead of the group II. It may be seen now that the condition n_i even may be relaxed.

If we are willing to assume, in addition, (c) as before, then we may take the least squares estimates

$$\hat{\alpha}_i = \bar{y}_i - \hat{\beta}\bar{x}_i, \qquad \hat{\beta} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)x_{ij}}{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2},$$

so that $\hat{\alpha}_i \xrightarrow{(p)} \alpha_i$ and $\hat{\beta} \xrightarrow{(p)} \beta$. If l_i denotes the number of points from the ith sample above the corresponding regression line and $\sum_i l_i = l$, then by a similar heuristic argument,

$$p(l_1, \cdots, l_c) \sim \frac{\prod_1^c \binom{n_i}{l_i}}{\binom{N}{l}}$$

for large n's so that by the lemma,

$$\chi^2 = \frac{N^2}{l(N-l)} \sum_1^c \frac{1}{n_i} \left(l_i - \frac{n_i}{N} l\right)^2, \qquad \text{d.f.} = c - 1.$$

(iii) *To test* $\alpha_1 = \alpha_2 = \cdots = \alpha_c$, *when* $\beta_1 = \beta_2 = \cdots = \beta_c$. On this hypothesis, y_{ij}'s have medians $\alpha + \beta x_{ij}$. α and β may be estimated by

$$\hat{\alpha} = \text{Median } (y_{ij} - \hat{\beta}x_{ij})$$

and

$$\text{Median}_{\text{I}}[y_{ij} - \hat{\beta}x_{ij}] = \text{Median}_{\text{II}}[y_{ij} - \hat{\beta}x_{ij}],$$

where, for convenience, we take groups I and II as $x \leqq \tilde{x}$ (the median of all the x's) and $x > \tilde{x}$ respectively. Let N be even and l_i be the number of points in the ith sample above the regression line $y = \hat{\alpha} + \hat{\beta}x$. If the hypothesis is true, we expect $l_i \sim \frac{1}{2}n_i$. We note that $\sum_1^c l_i = \frac{1}{2}N$. Let l_i' denote the number of positive terms in $y_{ij} - \alpha - \beta x_{ij}$ $(j = 1, 2, \cdots, n_i)$. Since $\hat{\alpha} \xrightarrow{(p)} \alpha$ and $\hat{\beta} \xrightarrow{(p)} \beta$, $l_i - l_i' \xrightarrow{(p)} 0$ as $N \to \infty$. Hence by similar heuristic arguments, the distribution of the l's for large N is approximately the same as that of the l''s subject to $\sum_1^c l_i' = \frac{1}{2}N$. Hence,

$$p(l_1, l_2, \cdots, l_c) \sim \frac{\prod_1^c \binom{n_i}{l_i}}{\binom{N}{\frac{1}{2}N}} \tag{22}$$

for large N so that by the lemma,

$$\chi^2 = 4\sum_1^c \frac{1}{n_i}\left(l_i - \frac{n_i}{2}\right)^2, \qquad \text{d.f.} = c - 1. \tag{23}$$

If we are willing to assume, in addition, (c), that is, the existence of the mean and variance, then we can have the least-square estimates $\hat{\alpha}$ and $\hat{\beta}$, such that $\hat{\alpha} \xrightarrow{(p)} \alpha$ and $\hat{\beta} \xrightarrow{(p)} \beta$. If we denote $\sum_1^c l_i$ by d, then by the same heuristic argument

$$p(l_1, l_2, \cdots, l_c) \sim \frac{\prod_1^c \binom{n_i}{l_i}}{\binom{N}{d}}$$

for large N so that

$$\chi^2 = \frac{N^2}{d(N-d)}\sum_1^c \frac{1}{n_i}\left(l_i - \frac{n_i}{N}d\right)^2, \qquad \text{d.f.} = c - 1.$$

We shall indicate here briefly a formal proof for (22), which was first derived on heuristic considerations.

Let $u_{ij} = y_{ij} - \hat{\beta}x_{ij}$. Then

$$\begin{aligned} l_i &= \text{number of positive } y_{ij} - \hat{\alpha} - \hat{\beta}x_{ij} (j = 1, 2, \cdots, n_i) \\ &= \text{number of } u_{ij}\text{'s } (j = 1, 2, \cdots, n_i) > \hat{\alpha} = \text{Median}_{i,j}\,(u_{ij}). \end{aligned}$$

Also $\sum_1^c l_i = \frac{1}{2}N$. Let z_a be the ath $(a = \frac{1}{2}N)u_{ij}$ in magnitude. Then the joint frequency function of $l_1, \cdots, l_c$ and z_a, under the hypothesis, is

$$
\begin{aligned}
&\sum_{i=1}^{c} \sum F_{11_1}(z_a) \cdots F_{11_{n_1-l_1}}(z_a)[1 - F_{11_{n_1-l_{1+1}}}(z_a)] \cdots [1 - F_{11_n}(z_a)] \\
&\cdots F_{ii_1}(z_a) \cdots F_{ii_{n_i-l_i-1}}(z_a)[1 - F_{ii_{n_i-l_i+1}}(z_a)] \cdots [1 - F_{ii_{n_i}}(z_a)] \\
&dF_{ii_{n_i-l_i}}(z_a) \cdots F_{cc_1}(z_a) \cdots F_{cc_{n_c-l_c}}(z_a)[1 - F_{cc_{n_c-l_c+1}}(z_a)] \\
&\cdots [1 - F_{cc_{n_c}}(z_a)],
\end{aligned} \tag{24}
$$

where $F_{ij}(z_a) = \Pr[u_{ij} \leqq z_a]$, the ith term indicates that z_a is from the ith sample and $\sum$ denotes the sum over all possible combinations.

Since $\hat{\beta} \xrightarrow{(p)} \beta$, given $\epsilon, \eta > 0$, there is an N_0 such that, for $N \geqq N_0$, $|\hat{\beta} - \beta| < \epsilon$ with probability $> 1 - \eta$. Then for $N \geqq N_0$, with probability $> 1 - \eta$, we have $\Pr[y_{ij} - \beta x_{ij} \leqq z_a - \epsilon x_{ij}] \leqq F_{ij}(z_a) \leqq \Pr[y_{ij} - \beta x_{ij} \leqq z_a + \epsilon x_{ij}]$, that is,

$$F(z_a - \epsilon x_{ij}) \leqq F_{ij}(z_a) \leqq F(z_a + \epsilon x_{ij}),$$

where F denotes the distribution function of all $y_{ij} - \beta x_{ij}$. In view of the continuity of F,

$$F_{ij}(z_a) = F(z_a) + \delta_{ij},$$

where the δ's are arbitrarily small and tend to zero as $n \to \infty$. Then (24) becomes

$$
\begin{aligned}
&\sum_{i=1}^{c} \sum F^{n_1-l_1}(z_a)\,[1 - F(z_a)]^{l_1} \cdots F^{n_i-l_i-l}(z_a)[1 - F(z_a)]^{l_i}\, dF(z_a) \\
&\cdots F^{n_c-l_c}(z_a)[1 - F(z_a)]^{l_c} + O(\delta) = \sum_{i=1}^{c} \sum F^{\frac{1}{2}N-1}(z_a)[1 - F(z_a)]^{\frac{1}{2}N} \\
&\cdot dF(z_a) + O(\delta) = \sum_{i=1}^{c} \binom{n_1}{l_1} \cdots \binom{n_{i-1}}{l_{i-1}} \frac{n_i!}{l_i!(n_i - l_i - 1)!} \binom{n_{i+1}}{l_{i+1}} \\
&\qquad \cdots \binom{n_c}{l_c} F^{\frac{1}{2}N-1}(z_a)[1 - F(z_a)]^{\frac{1}{2}N} dF(z_a) + O(\delta).
\end{aligned}
$$

On integrating out z_a we have the joint frequency function of $l_1, \cdots, l_c$

$$
\begin{aligned}
&= \sum_{i=1}^{c} \binom{n_1}{l_1} \cdots \frac{n_i!}{l_i!(n_i - l_i - 1)!} \cdots \binom{n_c}{l_c} \int_0^1 t^{\frac{1}{2}N-1}(1-t)^{\frac{1}{2}N}\, dt + O(\delta) \\
&= \left[\prod_1^c \binom{n_i}{l_i}\right] B(\tfrac{1}{2}N, \tfrac{1}{2}N + 1) \sum_{i=1}^{c} (n_i - l_i) + O(\delta) \\
&= \frac{\prod_1^c \binom{n_i}{l_i}}{\binom{N}{\frac{1}{2}N}} + O(\delta);
\end{aligned}
$$

which is the same as (22).

(iv) *To test* $\beta = 0$, *when* $\beta_1 = \beta_2 = \cdots = \beta_c = \beta$, *say*. On this hypothesis, y_{ij}'s have medians α_i's. We may take

$$\hat{\alpha}_i = \underset{j=1,2,\cdots,n_i}{\text{Median}} (y_{ij}).$$

For simplicity let n_i be even. Then $\frac{1}{2}n_i$ points from the ith sample are above the

corresponding line. Also $\frac{1}{2}N$ points are to the right of $\tilde{x}$, the median of all the x's. Let l_i be the number of points from the ith sample to the right of $\tilde{x}$ and above the corresponding line and let $l = \sum_1^c l_i$. We expect, then, l to be $\frac{1}{4}N$. Let m_i and m be defined similarly for $x \leqq \tilde{x}$. Then, by the same heuristic argument, for which a formal proof could be given as in (iii), we have

$$p(l, m) \sim \frac{\binom{\frac{1}{2}N}{l}\binom{\frac{1}{2}N}{m}}{\binom{N}{\frac{1}{2}N}}$$

for large N and, hence, by the lemma we have

$$\chi^2 = (16/N)(l - \tfrac{1}{4}N)^2, \qquad \text{d.f.} = 1.$$

The condition that n_i be even, then, may be relaxed.

3.3 *Testing linearity of regression.* As in the normal analysis, it is necessary that we have a number of observations for each x_i. Let the observations be $(x_i, y_{ij}), j = 1, 2, \cdots, n_i, i = 1, 2, \cdots, k$. We shall assume that the distribution of y, given x, is continuous and the same apart from location, say $h(x)$, which may depend on x. We want to test the hypothesis that the "regression" is linear, that is, $h(x) = \alpha + \beta x$. Let $\sum_1^k n_i = N$ and these N observations be divided into two groups, say $x \leqq x_{k_1}$ forming the first group and $x > x_{k_1}$ forming the second group, as evenly as possible. Let us suppose that observations corresponding to $x_i (i = 1, 2, \cdots, k_1)$ belong to the first group and the rest to the second. Let the groups contain a and $N - a$ observations respectively. We may then estimate α and β by Median $(y_{ij} - \hat{\alpha} - \hat{\beta}x_i) = 0$, and

$$\text{Median}_{\text{I}}(y_{ij} - \hat{\beta}x_i) = \text{Median}_{\text{II}}(y_{ij} - \hat{\beta}x_i).$$

Consider the n_i observations corresponding to x_i. If the regression is linear, we expect these n_i to be split evenly by the regression line $y = \hat{\alpha} + \hat{\beta}x$. Let l_i, out of these n_i, be above the line. We expect $l_i \sim \frac{1}{2}n_i$. Then $\sum_{i=1}^{k_1} l_i = \frac{1}{2}a$ and $\sum_{i=k_1+1}^{k} l_i = \frac{1}{2}(N - a)$, assuming for convenience that a and $N - a$ are even.

Let $z_{ij} = y_{ij} - \alpha - \beta x_i$. Then on the null hypothesis, the z_{ij}'s have identical distributions. Let l_i' be the number of positive terms in $z_{ij} (j = 1, 2, \cdots, n_i)$.

Since $\hat{\alpha} \xrightarrow{(p)} \alpha$ and $\hat{\beta} \xrightarrow{(p)} \beta$, $l_i - l_i' \xrightarrow{(p)} 0$. Hence on heuristic considerations as before, the distribution of the l_i's is the same (asymptotically) as that of the l_i''s subject to $\sum_{i=1}^{k_1} l_i' = \frac{1}{2}a$ and $\sum_{i=k_1+1}^{k} l_i' = \frac{1}{2}(N - a)$. Thus,

$$p(l_1, l_2, \cdots, l_k) \sim \frac{\prod_1^{k_1}\binom{n_i}{l_i}\prod_{k_1+1}^{k}\binom{n_i}{l_i}}{\binom{a}{\frac{1}{2}a}\binom{N-a}{\frac{1}{2}(N-a)}},$$

so that by the lemma,

$$\chi_{\text{I}}^2 = 4\sum_1^{k_1}\frac{1}{n_i}(l_i - \tfrac{1}{2}n_i)^2, \qquad \text{d.f.} = k_1 - 1,$$

and

$$\chi_{II}^2 = 4 \sum_{k_1+1}^{k} \frac{1}{n_i} (l_i - \tfrac{1}{2}n_i)^2, \qquad \text{d.f.} = k - k_1 - 1,$$

whence

$$\chi^2 = 4 \sum_{1}^{k} \frac{1}{n_i} (l_i - \tfrac{1}{2}n_i)^2, \qquad \text{d.f.} = k - 2.$$

4. Some bivariate problems.

4.1 *One-way classification.* Let there be n_i independent observations (x_{ij}, y_{ij}), $j = 1, 2, \cdots, n_i$, from the ith population, $i = 1, 2, \cdots, k$, and let $\sum_1^k n_i = N$.

Suppose $F_i(x, y)$ denotes the distribution function of (X, Y) for the ith population. We shall assume that

(i) the F's are absolutely continuous,

(ii) the distributions are identical except for location, and

(iii) the median of the conditional distribution of Y, given X, is a linear function of X. We note that the conditional probability, given X, is also a probability measure almost everywhere. Let $f_i(x, y)$ and $f_i(x)$ denote the densities of (X, Y) and X respectively for the ith population. Also, in view of (ii)

$$F_i(x, y) = F(x - \xi_i, y - \eta_i). \tag{25}$$

We want to test whether the populations are identical. Thus

$$H_0 : \xi_1 = \xi_2 = \cdots = \xi_k$$

$$\eta_1 = \eta_2 = \cdots = \eta_k .$$

(25) implies $f_i(x, y) = f(x - \xi_i, y - \eta_i)$ so that $f_i(x) = g(x - \xi_i)$, say. (iii) implies that $f(x, y) = g(x)h(y - \alpha - \beta x)$, so that

$$\begin{aligned} f(x - \xi_i, y - \eta_i) &= g(x - \xi_i)h[y - \eta_i - \alpha - \beta(x - \xi_i)] \\ &= g(x - \xi_i)h(y - \alpha_i - \beta x), \end{aligned}$$

say. Thus we see that

$$H_0 \Leftrightarrow \xi_1 = \xi_2 = \cdots = \xi_k$$

$$\alpha_1 = \alpha_2 = \cdots = \alpha_k .$$

It may be noted that we have relaxed just the normality of the distribution, but retained other features from the classical set up.

We shall use a step-down procedure to test H_0. A step-down procedure for H_0 with a level γ will be a test for

$$H_{0x} : \xi_1 = \xi_2 = \cdots = \xi_k$$

with a level γ_1, and if H_{0x} is not rejected, a further test for

$$H_{0y|x} : \alpha_1 = \alpha_2 = \cdots = \alpha_k$$

with a level γ_2, where γ_1 and γ_2 are chosen suitably so that

$$(1 - \gamma) = (1 - \gamma_1)(1 - \gamma_2).$$

The test for $H_{0y|x}$ will be derived from the conditional distribution of Y, given x, so that the x's then can be regarded as fixed.

For H_{0x}, we consider only the x's. Let us consider the test given by Mood [10]. (We could have used either Kruskal's test or the test proposed in [2].) Let m_i denote the number of observations in the ith sample greater than the median of all the x's. Mood shows that the frequency function, if H_{0x} is true, is

$$p(m_1, m_2, \cdots, m_k) = \frac{\prod_1^k \binom{n_i}{m_i}}{\binom{N}{a}}, \tag{26}$$

where $a = \frac{1}{2}N$ if N is even or $\frac{1}{2}(N - 1)$ if N is odd. The test-statistic proposed by him for large N is

$$\chi^2 = \frac{N(N-1)}{a(N-a)} \sum_1^k \frac{1}{n_i}\left(m_i - \frac{n_i a}{N}\right)^2, \qquad \text{d.f.} = k - 1.$$

For small n's, the probability is computed from the exact distribution (26). The test for $H_{0y|x}$ is seen to be precisely the same as that considered in 3.2. Hence we may take (23) (in its modified form) as a test-statistic, if the condition mentioned in the remark holds good. As already stated, it may be possible to prove that $\hat{\alpha} \xrightarrow{(p)} \alpha$ without using the condition, in which case (23) may be used for large samples in general.

4.2 *Two-way classification.* For simplicity, we shall consider only the case of one observation per cell, when the design is complete. Let "i" denote "treatments" and "j" denote "blocks". Suppose $i = 1, 2, \cdots, t$, $j = 1, 2, \cdots, b$ and $N = bt$. Let $F_{ij}(x, y)$ denote the distribution function of (X, Y) for the (ij)-th cell. We shall assume that

(i) $F_{ij}(x, y)$ is absolutely continuous,

(ii) the distributions are identical except for location, that is

$$F_{ij}(x, y) = F(x - \alpha_{ij}, y - \beta_{ij}),$$

(iii) the model is additive, that is,

$$\alpha_{ij} = \xi_i + \eta_j \quad \text{and} \quad \beta_{ij} = \gamma_i + \delta_j,$$

(iv) the "regression" of Y on X is linear.

As before, we notice that we have relaxed just the normality of the distribution while retaining other features of the classical set up.

Let $f_{ij}(x, y)$ and $f_{ij}(x)$ denote the densities of (X, Y) and (X) respectively for the (ij)th cell. Then $f_{ij}(x, y) = f(x - \alpha_{ij}, y - \beta_{ij})$ and $f_{ij}(x) = f_1(x - \alpha_{ij})$, say. Also

$$f(x, y) = f_1(x)f_2(y - \alpha - \beta x), \text{ so that}$$
$$f_{ij}(x, y) = f_1(x - \alpha_{ij})f_2[y - \beta_{ij} - \alpha - \beta(x - \alpha_{ij})]$$
$$= f_1(x - \xi_i - \eta_j)f_2[y - \alpha - \gamma_i + \beta\xi_i - \delta_j + \beta\eta_j - \beta x].$$

We will be interested in the usual hypothesis

$$H_0 : \xi_1 = \xi_2 = \cdots = \xi_t$$
$$\gamma_1 = \gamma_2 = \cdots = \gamma_t .$$

We shall consider a step-down procedure to test H_0. Considering the x's separately we can test, at a level α_1,

$$H_{0x} : \xi_1 = \xi_2 = \cdots = \xi_t$$

by the criterion given by Mood [10],

$$\chi^2 = \frac{t(t-1)}{ba(t-a)} \sum_{i=1}^{t} \left(m_i - \frac{ba}{t}\right)^2, \qquad \text{d.f.} = t - 1,$$

where $a = \frac{1}{2}t$ if t is even or $\frac{1}{2}(t-1)$ otherwise, and m_i = the number of x_{ij}'s$(j = 1, 2, \cdots, b)$ greater than $\tilde{x}_j$, the median of the jth column. Then, considering the conditional distributions of y_{ij}'s given x_{ij}'s we have to test

$$H_{0y|x} : y_{ij}\text{'s have medians } \lambda_j + \beta x_{ij}, \tag{27}$$

at a level α_2, so that $(1 - \alpha) = (1 - \alpha_1)(1 - \alpha_2)$. We may estimate λ_j and β by

$$\hat{\lambda}_j = \underset{i=1,2,\cdots,t}{\text{Median}} \ (y_{ij} - \hat{\beta}x_{ij}),$$

and

$$\text{Median}_{\text{I}}(y_{ij} - \hat{\lambda}_j - \hat{\beta}x_{ij}) = \text{Median}_{\text{II}}(y_{ij} - \hat{\lambda}_j - \hat{\beta}x_{ij})$$

where the groups are with respect to the x's as usual. We note that a, defined as above, out of t $y_{ij} - \hat{\lambda}_j - \hat{\beta}x_{ij}$'s for each j are positive and hence in all ab out of bt $y_{ij} - \hat{\lambda}_j - \hat{\beta}x_{ij}$'s are positive. Let l_i denote the number of positive terms out of b $y_{ij} - \hat{\lambda}_j - \hat{\beta}x_{ij}$, for given i. Then we expect $l_i \sim \frac{1}{2}b$ if (27) is true. Also $\sum_{i=1}^{t} l_i = ab$. Let l_i' denote the number of positive terms out of b $y_{ij} - \lambda_j - \beta x_{ij}$, for given i. On heuristic considerations, for large samples $\lambda_j \sim \hat{\lambda}_j$ and $\beta \sim \hat{\beta}$, so that the distribution of the l's is asymptotically the same as that of the l''s subject to $\sum_1^t l_i' = ab$. Hence,

$$p(l_1, l_2, \cdots, l_t) \sim \frac{\sum_1^t \binom{b}{l_i}}{\binom{N}{ab}}$$

for large bt, so that by the lemma,

$$\chi^2 = \frac{N^2}{ab(N-ab)} \sum_1^t \frac{1}{b}\left(l_i - \frac{b}{N}ab\right)^2 = \frac{t^2}{ba(t-a)} \sum_1^t \left(l_i - \frac{ba}{t}\right)^2, \qquad \text{d.f.} = t-1. \tag{28}$$

The same remark as that at the end of 4.1 will hold good here. Also, it may seem that we require t large (since we require $\lambda_j \sim \hat{\lambda}_j$ in the above argument), but if we give a formal proof, similar to that given in 3.2 (iii), we note that $\beta \sim \hat{\beta}$ is sufficient to reduce the proof to the one given by Mood. This does not require large t but only large bt. Hence (28) gives a test-criterion for large b.

Acknowledgment. I am deeply indebted to Prof. S. N. Roy for his keen interest in this work and for his suggestion about the possibility of extending the work on regression problems to the bivariate problems. I am also deeply thankful to Prof. Wassily Hoeffding for going through this work and for his useful comments. My sincere thanks are also due to the referees for their suggestions, which have improved the form of the paper.

REFERENCES

[1] A. Benard and Ph. van Elteren, "A generalization of the method of m rankings," *Proc. Kon. Ned. Akad v. Wet.*, Vol. 56 (1953), pp. 358–369.

[2] V. P. Bhapkar, "A nonparametric test for the problem of several samples," *Ann. Math. Stat.*, Vol. 32, No. 4 (1961).

[3] V. P. Bhapkar, "Confidence bounds connected with ANOVA and MANOVA for balanced and partially balanced incomplete block designs," *Ann. Math. Stat.*, Vol. 31 (1960), pp. 741–748.

[4] R. C. Bose and Dale M. Mesner, "On linear associative algebra corresponding to association schemes of partially balanced designs," *Ann. Math. Stat.*, Vol. 30 (1959), pp. 21–38.

[5] R. C. Bose and T. Shimamoto, "Classification and analysis of partially balanced designs with two associate classes," *J. Amer. Stat. Assn.*, Vol. 47 (1952), pp. 151–184.

[6] G. W. Brown and A. M. Mood, "On median tests for linear hypotheses," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 1951, pp. 159–166.

[7] J. Durbin, "Incomplete blocks in ranking experiments," *Brit. J. Psych.*, Vol. 4 (1951), pp. 85–90.

[8] Milton Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assn.*, Vol. 32 (1937), pp. 675–701.

[9] William H. Kruskal and W. Allen Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Stat. Assn.*, Vol. 47 (1952), pp. 583–621.

[10] Alexander McFarlane Mood, *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Co., 1950.

[11] J. Roy, "Step-down procedure in multivariate analysis," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 1177–1187.

[12] S. N. Roy and R. E. Bergmann, "Tests of multiple independence and the associated confidence bounds," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 491–503.