Georg Neuhaus

H_0-contiguity in nonparametric testing problems

Acta Universitatis Carolinae. Mathematica et Physica, Vol. 24 (1983), No. 1, 77--83

H_0—contiguity in Nonparametric Testing Problems

G. NEUHAUS
Department of Statistics, University of Hamburg*)

Received 22 December 1982

The paper concerns the concept of H_0-contiguity which is a slight modification of the notion of "contiguity to H_0" considered by Hájek and Šidák (Theory of Rank Tests, 1967) in connection with linear rank statistics under location alternatives. It is shown this concept of H_0-contiguity enables a more straight-forward derivation of the limiting distribution of rank statistics under (not necessarily translation) alternatives. The attention is concentrated mainly to the two-sample problem of testing "randomness" vs. "positive stochastic deviation of the first sample".

Článek se týká pojmu tzv. H_0-kontiguity, což je poněkud modifikovaná verze H_0-kontiguity uvažovaná v knize Hájka a Šidáka (Teorie pořadových testů, 1967) v souvislosti s lineárními pořadovými statistikami pro alternativy posunutí v poloze. Ukazuje se, že nový pojem kontiguity vede k jednoduššímu průhlednějšímu odvození limitního rozdělení pořadových statistik při alternativě (ne nutně posunutí). Pozornost je soustředěna na tzv. dvouvýběrový problém.

Статья касается концепции H_0-контигуальности, которая несколько модифицированной версией H_0-контигуальности рассматриваемой в книге Гаека и Шидака (Ранговые критерии, 1967) в связи с линейными ранговыми статистиками. Показывается, что новая концепция контигуальности ведет к более простому выводу асимптотического распределения ранговых статистик, когда верна альтернатива. Внимание сосредоточивается на двухвыборочную проблему.

In the present talk I want to speak about the concept of H_0-contiguity being a slightly altered version of the notion of "contiguity to H_0" which was considered by Hájek and Šidák (1967) in connection with the treatment of linear rank statistics mainly under location alternatives. We show that a consequent application of the H_0-contiguity concept makes the derivation of the limiting distribution of linear and other rank statistics under (not necessarily translation) alternatives considerably more neat.

When I first showed the usefulness of the H_0-contiguity notion at the Oberwolfach meeting in December 1981 I was mainly interested in the two sample testing problem and symmetry problem. Afterwards, at the subsequent Oberwolfach meeting in March/April 1982 van Zwet gave an extension which will be discussed

*) D-2000, Hamburg 13, Bundesstrasse 55, West Germany.

below. Since the two sample problem of testing "randomness" vs. "positive stochastic deviation of the first sample" was treated in previous talks at the present "colloquium" I shall exemplify the ideas by the two sample case, too.

Therefore, let $X_1, \ldots, X_m$ resp. $Y_1, \ldots, Y_n$ be i.i.d. real random variables having continuous distribution functions (df's) F_m resp. G_n and denote by $R_{11}, \ldots, R_{1m}$ resp. $R_{21}, \ldots, R_{2n}$ their ranks in the pooled sample consisting of $N = m + n$ observations. Let us consider the null-hypothesis of randomness $H_0 : F_m = G_n$ vs. the alternative K that the first sample is stochastically larger than the second sample, i.e. $K : F_m \leqq G_n$, $F_m \neq G_n$.

Let $b_{N1} \leqq \ldots \leqq b_{NN}$ be given scores such that the step functions $b_N : (0, 1) \to \mathbf{R}$ defined by $b_N(u) = b_{Ni}$ for $(i - 1)/N \leqq u < i/N$ converge in L_2 space of Lebesgue (λ)-square integrable functions on $(0, 1)$ to some nondecreasing function $b : (0, 1) \to \mathbf{R}$ with $\langle 1, b \rangle = 0$ and $\|b\| = 1$

$$\lim_{N \to \infty} \|b_N - b\| = 0 \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual scalar product in L_2 and $\|\cdot\|$ the corresponding norm, 1 being the function with constant value 1. It is a wellknown fact that under H_0 the two sample linear rank statistics

$$S_{mn} = (mn/N)^{1/2} \cdot \left\{ \frac{1}{m} \sum_{i=1}^{m} b_N\left(\frac{R_{1i}}{N+1}\right) - \frac{1}{n} \sum_{j=1}^{n} b_N\left(\frac{R_{2j}}{N+1}\right) \right\} \tag{2}$$

converge in distribution to the standard normal distribution $\mathcal{N}(0, 1)$:

$$S_{mn} \xrightarrow{\mathscr{D}} \mathcal{N}(0, 1) \quad \text{for} \quad \min(m, n) \to \infty . \tag{3}$$

The crucial step in the proof of (3) is the result that under $H_0 : F_m = G_n = H_{m,n}$, say, the statistics S_{mn} can be approximated in probability by

$$S^*_{mn} = (mn/N)^{1/2} \cdot \left\{ \frac{1}{m} \sum_{i=1}^{m} b(U_i) - \frac{1}{n} \sum_{j=1}^{n} b(U_{m+j}) \right\}, \tag{4}$$

$U_i = H_{m,n}(X_i)$, $i = 1, \ldots, m$; $U_{m+j} = H_{m,n}(Y_j)$, $j = 1, \ldots, n$, i.e.

$$S_{mn} - S^*_{mn} \overset{\text{Pr}}{\to} 0 \quad \text{under } H_0 \text{ for} \quad \min(m, n) \to \infty . \tag{5}$$

From (5) and Lindeberg's Theorem applied to S^*_{mn} (3) follows at once.

In order to judge the asymptotic power of the upper level α test based on S_{mn} one often considers alternatives (F_m, G_n) which are contiguous to H_0 in the sense of Hájek and Šidák (1967), Chap. VI. Here a slightly altered concept is used which takes care of the fact that linear rank statistics are distribution-free on H_0, so that is doesn't matter which particular element of H_0 or even which sequence of elements in H_0 is used. Before giving the exact definition of the notion "H_0-contiguity" let us restrict the sequence of sample numbers $m, n, m + n = N$, by assuming that the total sample number N determines completely m and n, i.e. $m = m(N)$, $n = n(N)$ and assume that the relative sample numbers converge

$$(6) \qquad \lim_{N\to\infty} m(N) = \eta\,, \quad \text{say, for} \quad 0 < \eta < 1\,.$$

Let us write F_N resp. G_N instead of $F_{m(N)}$ resp. $G_{n(N)}$. Henceforth all limits are for $N \to \infty$.

Definition 1. A sequence $\{F_N^m \cdot G_N^n\}$, $N \geqq 1$, is called *H_0-contiguous*, $\{(F_N, G_N)\} \lhd H_0$, if there exists a sequence $\{(H_N, H_N)\}$, $N \geqq 1$, in H_0 such that $\{F_N^m \cdot G_N^n\} \lhd \{H_N^N\}$ in the usual one-sided contiguity sense, see e.g. Oosterhoff and van Zwet (1979).

Remark. The natural extension of the above definition is to allow for arbitrary products $\{\prod_{i=1}^{N} F_{Ni}\}$ instead of $\{F_N^m \cdot G_N^n\}$. This extension by van Zwet (1982) will be discussed below.

It should be mentioned that I don't distinguish notationally between df's and their corresponding probability measures.

In order to give some characterizations of H_0-contiguity note that $\bar{H}_N = \eta_N F_N + (1 - \eta_N)\, G_N$, $\eta_N = m(N)/N$, dominates F_N and G_N and define a function d_N on $(0, 1)$ by

$$(7) \qquad d_N = (mn/N)^{1/2}\,(f_N - g_N) \circ \bar{H}^{-1}\,,$$

with $f_N = \mathrm{d}F_N/\mathrm{d}\bar{H}_N$, $g_N = \mathrm{d}G_N/\mathrm{d}\bar{H}_N$ and $\bar{H}_N^{-1}$ the left continuous pseudoinverse of $\bar{H}_N$. Since $\eta_N f_N + (1 - \eta_N)\, g_N = 1\ [\bar{H}_N]$, f_N and g_N are bounded

$$0 \leqq f_N \leqq 1/\eta_N\,, \quad 0 \leqq g_N \leqq 1/(1 - \eta_N)\,[\bar{H}_N]\,.$$

The following Theorem characterizes H_0-contiguity.

Theorem 2. Under condition (6) the following four statements are equivalent

$$(8) \qquad \{(F_N, G_N)\} \lhd H_0$$

$$(9) \qquad N\mathscr{H}^2(F_N, G_N) = \mathrm{O}(1)$$

$$(10) \qquad \|d_N\| = \mathrm{O}(1)$$

$$(11) \qquad \{F_N^m \cdot G_N^n\} \lhd \{\bar{H}_N^N\}\,.$$

In (9) $\mathscr{H}$ denotes the Hellinger distance defined by $\mathscr{H}^2(F, G) = \int (f^{1/2} - g^{1/2})^2\,\mathrm{d}\mu$ for two probability measures (or df's) F and G, with μ-densities f and g, where μ is any σ-finite measure dominating F and G.

The proof of the Theorem 2 makes use of a result of Oosterhoff and van Zwet (1979) which we cite for easier reference in a special form needed here: Let H_N and

F_{Ni}, $i = 1, \ldots, N$, be probability measures on arbitrary σ-finite measure spaces $(\mathcal{X}_N, \mathcal{A}_N, \mu_N)$, $N = 1, 2, \ldots$, such that μ_N dominates H_N and F_{Ni} with $h_N = \mathrm{d}H_N/\mathrm{d}\mu_N$ and $f_{Ni} = \mathrm{d}F_{Ni}/\mathrm{d}\mu_N$. Then according to Oosterhoff and van Zwet (1979), Theorem 1, the conditions

$$\text{(A)} \qquad \sum_{i=1}^{N} \mathcal{H}^2(H_N, F_{Ni}) = \mathrm{O}(1)$$

and

$$\text{(B)} \qquad \sum_{i=1}^{N} F_{Ni}\{f_{Ni}/h_N \geqq c_N\} \to \mathrm{O}\,, \quad c_N \to \infty$$

are jointly equivalent to $\{\prod_{i=1}^{N} F_{Ni}\} \lhd \{H_N^N\}$.

Outline of the proof of Theorem 2: From $\mathcal{H}(F_N, G_N) \leqq \mathcal{H}(F_N, H_N) + \mathcal{H}(H_N, G_N)$, (B) and (A) it follows that (8) implies (9). The equivalence of (9) and (10) follows from

$$\text{(12)} \qquad \|d_N\|^2 \leqq \{\eta_N(1 - \eta_N)\}^{-1} N\mathcal{H}^2(F_N, G_N) \leqq \{\eta_N(1 - \eta_N)\}^{-2} \|d_N\|^2\,, \quad \forall N \geqq 1\,,$$

while (9) and the inequality

$$\text{(13)} \qquad m\mathcal{H}^2(F_N, \bar{H}_N) + n\mathcal{H}^2(G_N, \bar{H}_N) \leqq \eta_N(1 - \eta_N) N\mathcal{H}^2(F_N, G_N)$$

imply that condition (A) is met for the sequences in (11). Since condition (B) is automatically fulfilled for uniformly bounded densities f_N, g_N, (11) follows. Trivially (11) implies (8). The various inequalities above are mainly consequences of the concavity of the square root function and of $\eta_N f_N + (1 - \eta_N)\, g_N = 1$. □

The interesting feature of Theorem 2 is that under (6) H_0-contiguity is expressible completely by means of the Hellinger distance between F_N and G_N or by the L_2-boundedness of $\{d_N\}$.

Now assume that $\{(F_N, G_N)\}$ is H_0-contiguous. Choosing $H_{m,n} = \bar{H}_N$ in (4) and (5) and writing $S_{Nb}^{(*)} = S_{m(N),n(N)}^{(*)}$, we obtain

$$\text{(14)} \qquad S_{Nb} - S_{Nb}^* \underset{\mathrm{Pr}}{\to} 0 \quad \text{under} \quad \{(F_N, G_N)\}\,.$$

Then

$$\text{(15)} \qquad E_{F_N G_N} S_{Nb}^* = (mn/N)^{1/2} \int b \circ \bar{H}_N(f_N - g_N)\, \mathrm{d}\bar{H}_N = \langle b, d_N \rangle\,.$$

Since f_N and g_N are bounded, the expectation in (15) is always finite. This is a technical advantage of H_0-contiguity concept compared with the usual approach using fixed elements (H, H) in H_0, because in the latter case the corresponding expectations are in general finite only after a truncation of the score function b, see e.g. Behnen (1972). Moreover, (12) gives an explicit bound for $\{\langle b, d_N \rangle\}$, $N \geqq 1$:

$$(16) \qquad \limsup \langle b, d_N \rangle^2 \leqq \{\eta(1-\eta)\}^{-1} \limsup N \mathscr{H}^2(F_N, G_N)\,.$$

Simple calculations yield $\mathrm{Var}_{(F_N,G_N)} S_N^* = 1$, implying that the Lindeberg condition is fulfilled by $\{S_{Nb}^* - \langle b, d_N \rangle\}$. Therefore, using (14) one finally yields

$$(17) \qquad S_{Nb} - \langle b, d_N \rangle \xrightarrow{\mathscr{D}} \mathscr{N}(0, 1) \quad \text{under} \quad \{(F_N, G_N)\} \lhd H_0\,.$$

Moreover, if $\{k(N)\}$, $N \geqq 1$, is a subsequence of the total sample numbers $\{N\}$ with

$$(18) \qquad \lim k(N) = \infty \quad \text{and} \quad \liminf N/k(N) > 0\,,$$

then even

$$(19) \quad S_{k(N)b} - \{k(N)/N\}^{1/2} \langle b, d_N \rangle \xrightarrow{\mathscr{D}} \mathscr{N}(0, 1) \quad \text{under} \quad \{(F_N, G_N)\} < H_0\,.$$

Let φ_N^b denote the upper level α test based on S_{Nb}, $0 < \alpha < 1$; then (19) implies

$$(20) \qquad E_{(F_N,G_N)} \varphi_{k(N)}^b = 1 - \Phi\left(u_\alpha - \left(\frac{k(N)}{N}\right)^{1/2} \langle b, d_N \rangle\right) + o(1)\,,$$

Φ denoting the df of $\mathscr{N}(0, 1)$ and $u_\alpha = \Phi^{-1}(1 - \alpha)$. The asymptotic power formula (20) is the basis for further efficiency considerations below.

Before doing so let me introduce an important class of H_0-contiguous alternatives:

For every fixed pair (F, G) of alternatives define a corresponding family of alternatives (F_Δ, G_Δ), $0 \leqq \Delta \leqq 1$ by

$$(21) \qquad f_\Delta = \mathrm{d}F_\Delta/\mathrm{d}\bar{H} = 1 + \Delta(1 - \eta)\,\mathrm{d}(F - G)/\mathrm{d}\bar{H}$$

and

$$(22) \qquad g_\Delta = \mathrm{d}G_\Delta/\mathrm{d}\bar{H} = 1 - \Delta\eta\,\mathrm{d}(F - G)\,\mathrm{d}\bar{H}$$

with $\eta = \lim m(N)/N$, $\bar{H} = \eta F + (1 - \eta)\,G$. Clearly $F_1 = F$, $G_1 = G$, $F_0 = G_0 = \bar{H}$ and $\eta F_\Delta + (1 - \eta)\,G_\Delta = \bar{H}\ \forall \Delta$. Under condition (6) for $\Delta = \Delta_N$ the following equivalence holds true

$$(23) \qquad \{(F_{\Delta_N}, G_{\Delta_N})\} \lhd H_0 \Leftrightarrow \Delta_N = O(N^{-1/2})\,.$$

The proof follows from Theorem 2 by using the inequalities

$$N^{1/2}\,\Delta_N \int |f - g|\,\mathrm{d}\bar{H} \leqq N^{1/2}\,\mathscr{H}(F_{\Delta_N}, G_{\Delta_N}) \leqq N^{1/2}\,\Delta_N\{\eta(1 - \eta)\}^{-1}\,. \quad \square$$

The importance of the above construction stems from the fact that the quantity $\mathrm{d}(F - G)/\mathrm{d}\bar{H}$ represents in some sense the character of deviation of (F, G) from H_0. Multiplying $\mathrm{d}(F - G)/\mathrm{d}\bar{H}$ by a factor $\Delta_N \to 0$ maintains the characteristic form of deviation though $\{(F_{\Delta_N}, G_{\Delta_N})\}$ approaches H_0.

The general power formula (20) under H_0-contiguity allows for a meaningful sample definition of asymptotic relative Pitman efficiency: Let $\{\varphi_N\}$ and $\{\psi_N\}$ be level-α-tests with $\beta_N = E_{(F_N,G_N)}\psi_N$ and

$$0 < \alpha < \liminf \beta_N \leqq \limsup \beta_N < 1 . \tag{24}$$

Write $k(N) = \inf\{k : E_{(F_N,G_N)}\varphi_k > \beta_N\}$, $\inf\{\emptyset\} \equiv \infty$.

Then, the asymptotic relative efficiency (ARE) of $\{\varphi_N\}$ with respect to $\{\psi_N\}$ on $\{(F_N, G_N)\}$ is defined by

$$\mathrm{ARE}\,(\varphi : \psi \mid (F_N, G_N)) = \liminf N/k(N) . \tag{25}$$

Theorem 3. If $\{\varphi_N\}$ fulfills the power formula (20), then

$$\mathrm{ARE}\,(\varphi : \psi \mid (F_N, G_N)) = \liminf (\langle b, d_N\rangle/\delta_N)^2 \tag{26}$$

where δ_N is defined by $1 - \Phi(u_\alpha - \delta_N) = \beta_N$.

Proof. Omitted.

Example 4. Let us compare two linear rank tests $\{\varphi_N\}$ and $\{\psi_N\}$ with asymptotic score functions b_1 and b_2 respectively. Then

$$\mathrm{ARE}\,(\varphi : \psi) = \liminf \langle b_1, d_N\rangle^2/\langle b_2, d_N\rangle^2 \tag{27}$$

if $\liminf \langle b_2, d_N\rangle^2 > 0$.

Proof. (20) applied to $\{\psi_N\}$ with $k(N) = N$ yields $\delta_N = \langle b_2, d_N\rangle + o(1)$. Using (26) yields (27). □

The right side coincides (apart from the explicit form of the asymptotic translations $\langle b_i, d_N\rangle$, $i = 1, 2$) with the definition of asymptotic relative Pitman efficiency based on asymptotic translations of the H_0-distribution, see e.g. Behnen (1972), thus yielding a sample efficiency interpretation of all results based on that ARE definition.

As mentioned at the beginning the notion of H_0-contiguity is useful for other testing problems, e.g. the "symmetry problem" and the "independence problem". I won't discuss these matters here, rather let me cite a result of van Zwet (1982) concerning the extended definition of H_0-contiguity in the remark following Definition 1. The notation is the same as in (A) and (B). Furthermore, write $F_N^{(N)} = \prod_{i=1}^{N} F_{Ni}$, $\bar{F} = (1/N)\sum_{i=1}^{N} F_{Ni}$ and $\bar{f}_N = \mathrm{d}\bar{F}_N/\mathrm{d}\mu_N$.

Theorem (van Zwet, 1982). The following three statements are equivalent

a) $\{F_N^{(N)}\} \lhd H_0$,

b) $\{F_N^{(N)}\} \lhd \{\bar{F}_N^N\}$,

c) $\begin{cases} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathscr{H}^2(F_{Ni}, F_{Nj}) = O(1), \\ \text{and} \\ \sum_{i=1}^{N} F_{Ni}\{f_{Ni}/\bar{f}_N \geqq c_N\} \to 0 \text{ for } c_N \to \infty. \end{cases}$

Clearly, the above Theorem extends the equivalence (8) ⇔ (9) ⇔ (11) of Theorem 2, while (10) has no natural extension to the more general case. The proof of van Zwet's Theorem which at time is unpublished follows from the above-mentioned Theorem of Oosterhoff and van Zwet, see (A), (B), and the following two series of inequalities:

$$\frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathscr{H}^2(F_{Ni}, F_{Nj}) \leqq \sum_{i=1}^{N} \mathscr{H}^2(F_{Ni}, \bar{F}_N) \leqq \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathscr{H}^2(F_{Ni}, F_{Nj}) \leqq$$

$$\leqq 2 \sum_{i=1}^{N} \mathscr{H}^2(F_{Ni}, H_N) \quad \text{for arbitrary} \quad \{H_N\},$$

and

$$\limsup \sum_{i=1}^{N} F_{Ni}\{f_{Ni}/\bar{f}_N \geqq c_N\} \leqq \limsup \sum_{i=1}^{N} F_{Ni}\{f_{Ni}/h_N \geqq c_N^{1/2}\} +$$

$$+ 2 \limsup c_N^{-1/2} \Big[\sum_{i=1}^{N} \{\mathscr{H}^2(F_{Ni}, H_N) + \mathscr{H}^2(F_{Ni}, \bar{F}_N)\} \Big]$$

for $c_N \to \infty$ and arbitrary $\{H_N\}$.

In fact, van Zwet (1982) also proved a version of the above Theorem involving two-sided contiguity as well as a similar Theorem concerning asymptotic normality and asymptotic negligibility of loglikelihood ratios. More detailed proofs of Theorem 2 and Theorem 3 can be found in Neuhaus (1982).

References

[1] BEHNEN, K. (1972): A characterization of certain rank-order tests with bounds for the asymptotic relative efficiency. *Ann. Math. Statit. 43,* 1839—1851.

[2] HÁJEK, J. and ŠIDÁK, Z. (1967): Theory of Rank Tests. *Academic,* New York.

[3] NEUHAUS, G. (1982): H_0-contiguity in nonparametric testing problems and sample Pitman efficiency. *Ann. Statist. 10,* 575—582.

[4] OOSTERHOFF, J. and VAN ZWET, W. R. (1979): A note on contiguity and Hellinger distance. In Contributions to Statistics. *J. Hájek Memorial Volume* (J. Jurečková ed.), 157—166.

[5] VAN ZWET, W. R. (1982): Contiguity relative to the randomness hypothesis, talk given at the Oberwolfach Meeting, *March/April 1982.*