12

The Limiting Distribution of the Rank Correlation Coefficient R_g

Rudy A. Gideon[1]
Michael J. Prentice[2]
Ronald Pyke[3]

ABSTRACT A new correlation coefficient, R_g, based on ranks and greatest deviation was defined in Gideon and Hollister (1987). In there the exact distributions were obtained by enumeration for small sample sizes, and by computer simulations for larger sample sizes. In this note, it is shown that the asymptotic distribution of $n^{1/2}R_g$ is $N(0,1)$ when the variables are independent and n is the sample size. This limit is derived by restating the definition of R_g in terms of a rank measure and then using a limit theorem on set-indexed empirical processes which appears in Pyke (1985). The limiting distribution can be compared to the critical values for large samples given in Figure 2 of Gideon and Hollister (1987). Methods for deriving the limiting distribution under fixed and contiguous alternatives are also described.

1 Introduction

In Gideon and Hollister (1987), a new rank correlation coefficient, R_g, is defined that is more resistant to outliers than classical coefficients. Critical values for tests based on R_g for sample sizes $n = 2, 3, \ldots, 100$ were provided. In this paper, the limiting distribution of R_g is obtained so that the new robust correlation procedures may be used in all cases. Using the notation of Gideon and Hollister (1987), the correlation coefficient R_g is defined as follows. Let $\boldsymbol{p}$ denote any permutation of the first n positive integers and let $\boldsymbol{e}$ denote the particular "reverse" permutation, $(n, n-1, \ldots, 2, 1)$. The symbol $\circ$ denotes the cyclic group operation, $[\cdot]$ the greatest integer function, and $\mathbf{1}$ the indicator function. Let (X_k, Y_k), $k = 1, \ldots, n$ be a random

[1]University of Montana, Missoula, Montana, USA
[2]University of Edinburgh, Edinburgh, Scotland, UK
[3]University of Washington, Seattle, Washington, USA

sample from an absolutely continuous bivariate distribution H, and denote the corresponding order statistics by $-\infty < X_{n1} < \cdots < X_{nn} < +\infty$ and $-\infty < Y_{n1} < \cdots < Y_{nn} < +\infty$. If r_i is the y rank of that y value which is paired with the i-th smallest x value, then $(X_{n1}, Y_{nr_1}), \ldots, (X_{nn}, Y_{nr_n})$ are the data recorded by increasing x values. Now for $\boldsymbol{p} = (r_1, r_2, \ldots, r_n)$, define

$$d_i(\boldsymbol{p}) = \sum_{j=1}^{i} 1(r_j > i), \quad d_i(\varepsilon \circ \boldsymbol{p}) = \sum_{j=1}^{i} 1(n+1-r_j > i), \quad d(\boldsymbol{p}) = \max_{1 \le i \le n} d_i(\boldsymbol{p}).$$

The new correlation coefficient is then defined by

$$R_g = \{d(\varepsilon \circ \boldsymbol{p}) - d(\boldsymbol{p})\}/[n/2].$$

For notational convenience during the derivation of the limiting distribution, it is assumed that the sample size n is even; the modifications needed when n is odd are straightforward. Thus, the normalized coefficient becomes

$$n^{1/2} R_g = 2n^{-1/2}\{d(\varepsilon \circ \boldsymbol{p}) - d(\boldsymbol{p})\}. \tag{1.1}$$

In terms of this definition, one sees that R_g may be described as the maximum deviation between sums of forward and backward ranks. However, for purposes of this paper, it is important to observe that the ranks in R_g may be viewed as random measures of particular sets, and this set-indexed approach enables one to obtain the asymptotic null distribution rather directly. It also indicates the large family of correlation coefficients that may be considered. To obtain the set-indexed representation, define the following two families of Borel sets on the unit square $I^2 = [0,1] \times [0,1]$,

$$\begin{aligned} \mathcal{A} &= \{A_t : 0 \le t \le 1\} \quad \text{where } A_t = \{(x,y) \in I^2 : y > t, x \le t\}, \\ \mathcal{B} &= \{B_t : 0 \le t \le 1\} \quad \text{where } B_t = \{(x,y) \in I^2 : y < 1-t, x \le t\}. \end{aligned} \tag{1.2}$$

Since R_g is distribution free with respect to the class of absolutely continuous distribution functions, without loss of generality we assume that the marginal distributions of the X_k and Y_k are both uniform on the unit interval I. The problem addressed here is that of finding the limiting distribution of $n^{1/2}R_g$ under the null hypothesis of independence in which the joint distribution H is the uniform distribution in I^2; that is, for all Borel sets $C \in I^2$, $H(C) = |C|$, where $|\cdot|$ denotes Lebesgue measure.

For the x-ordered data (X_{nj}, Y_{nr_j}), $j = 1, 2, \ldots, n$, the rank measure R_n is defined on the Borel sets in I^2 (cf. Pyke (1985)) by

$$R_n(C) = n^{-1}\#\{(X_k, Y_k) : X_k = X_{ni},\ Y_k = Y_{nj} \text{ for some } (i,j) \in (n+1)C\} \tag{1.3}$$

in which # denotes the cardinal number. Thus, R_n is the probability random measure set function that assigns equal measure of $1/n$ to each of the

n points $(i/(n+1), r_i/(n+1)), 1 \leq i \leq n$. Notice that by the definition of ranks, this rank measure has (discrete) uniform marginals. It is easily checked that $d_i(\boldsymbol{p}) = nR_n(A_{i/(n+1)})$, so that

$$d(\boldsymbol{p}) = \max_{1\leq i\leq n} nR_n(A_{i/(n+1)}) = n \sup_{0\leq t\leq 1} R_n(A_t).$$

Similarly, $d_i(\boldsymbol{\varepsilon} \circ \boldsymbol{p}) = nR_n(B_{i/(n+1)})$ and

$$d(\boldsymbol{\varepsilon} \circ \boldsymbol{p}) = \max_{1\leq i\leq n} nR_n(B_{i/(n+1)}) = n \sup_{0\leq t\leq 1} R_n(B_t).$$

For the last equality in each case, observe that $R_n(B_t)$ and $R_n(A_t)$ change values only at $t = i/(n+1)$, $i = 1, \ldots, n$. It now follows that, for n even, (1.1) may be written as

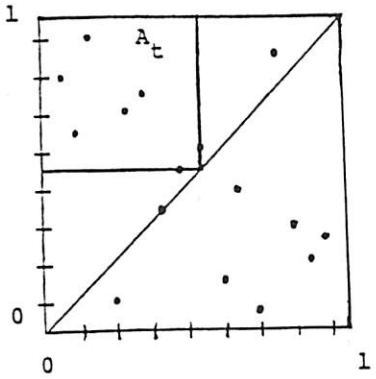
$$R_g = 2\{\sup_{0\leq t\leq 1} R_n(B_t) - \sup_{0\leq t\leq 1} R_n(A_t)\}. \tag{1.4}$$

See Figure 1 for an illustration based on the original YMCA data reported in Gideon and Hollister (1987).

2 The Limiting Null Distribution

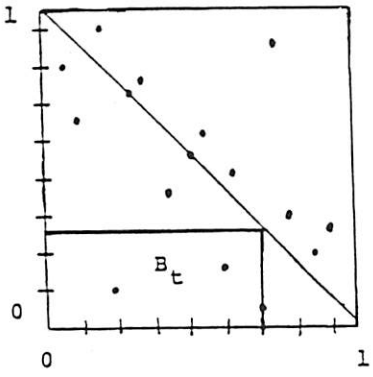
Before deriving the limiting distribution, we record some properties of the exact distributions. First of all, $d_i(\boldsymbol{p})$ is a hypergeometric random variable (r. v.); specifically, if Hyper (N, k, m) denotes a hypergeometric r.v. with population size N, sample size m, and k individuals of the type which is being counted, then $d_i(\boldsymbol{p})$ and $d_i(\boldsymbol{\varepsilon} \circ \boldsymbol{p})$ are Hyper $(n, n-i, i)$ under the null hypothesis. To see this, note that the r.v. $d_i(\boldsymbol{p})$ counts the number of ranks in the first i positions of $\boldsymbol{p}$ which exceed i and there are $n-i$ such possibilities. Thus i ranks are samples, $n-i$ ranks are classified as a "success" and the total population size is n. For the null distribution all permutations are equally likely and the conclusion follows. By the same argument, it follows that for any rectangle $A = B \times C$, $B, C \subset I$, $nR_n(A)$ is a Hyper (n, b, c) r.v. where b and c are the number of integers in $(n+1)B$ and $(n+1)C$, respectively. In the above, $d_i(\boldsymbol{p}) = nR_n(A_{i/(n+1)})$ and $A_t = [0, t) \times [t, 1]$.

FIGURE 1. Set-indexed Representation of R_g Illustrated for the YMCA Data of Gideon and Hollister (1987): (i, r_i) is plotted at $(i/17, r_i/17), i = 1, 2, \ldots, 16$ for $(r_1, \ldots, r_{16}) = (14, 11, 16, 2, 12, 13, 7, 9, 10, 3, 8, 1, 15, 6, 4, 5)$



$$
\begin{aligned}
A_t &= \{(x, y) \in I^2 : y > t, x \leq t\} \\
d(\boldsymbol{p}) &= \max_{1 \leq i \leq n} d_i(\boldsymbol{p}) \\
&= n \sup_{0 \leq t \leq 1} R_n(A_t) \\
&= 16 R_{16}(A_{9/17}) = 6
\end{aligned}
$$

$R_n(A_{9/17}) = 6/16$



$$
\begin{aligned}
B_t &= \{(x, y) \in I^2 : y < 1 - t, x \leq t\} \\
d(\boldsymbol{\varepsilon} \circ \boldsymbol{p}) &= \max_{1 \leq i \leq n} d_i(\boldsymbol{\varepsilon} \circ \boldsymbol{p}) \\
&= n \sup_{0 \leq t \leq 1} R_n(B_t) \\
&= 16 R_{16}(B_{12/17}) = 3
\end{aligned}
$$

$R_n(B_{12/17}) = 3/16$

$$R_g = \frac{2}{n}\{d(\boldsymbol{\varepsilon} \circ \boldsymbol{p}) - d(\boldsymbol{p})\} = (3 - 6)/8 = -3/8$$

It then follows that

$$E(d_i(\boldsymbol{p})) = E(nR_n(A_{i/n+1})) = i(1 - i/n)$$

and

$$\text{var}(d_i(\boldsymbol{p})) = \text{var}(nR_n(A_{i/n+1})) = i^2(n-i)^2/n^2(n-1).$$

Similar equations hold for $d_i(\boldsymbol{\varepsilon} \circ \boldsymbol{p})$ and $nR_n(B_{i/n+1})$. The covariances are

$$\text{cov}(d_i(\boldsymbol{p}), d_k(\boldsymbol{p})) = i^2(n-k)^2/n^2(n-1), \quad 0 \le i \le k \le n.$$

To show this, write $\boldsymbol{p} = (r_1, \ldots, r_n)$. Note that $d_n(\boldsymbol{p}) \equiv 0$. Set $d_0(\boldsymbol{p}) \equiv 0$. For $i \le k$,

$$\begin{aligned}\text{cov}(d_i(\boldsymbol{p}), d_k(\boldsymbol{p})) &= \text{cov}\left\{\sum_{j=1}^{i} 1(i < r_j), \sum_{\ell=1}^{k} 1(k < r_\ell)\right\} \\ &= \sum_{j=1}^{i}\sum_{\ell=1}^{k} \text{cov}\{1(i < r_j), 1(k < r_\ell)\}.\end{aligned}$$

Clearly,

$$\text{cov}(1(i < r_j), 1(k < r_\ell)) = P(r_j > i \text{ and } r_\ell > k) - P(r_j > i)\ P(r_\ell > k).$$

Since $P(r_j > i) = (n-i)/n$ and

$$\begin{aligned}P(r_j > i, r_\ell > k) &= P(r_j > i | r_\ell > k) P(r_\ell > k) \\ &= \begin{cases} (n-k)/n & \text{for } j = \ell \\ \{(n-i-1)/(n-1)\}\{(n-k)/n\} & \text{for } j \ne \ell. \end{cases}\end{aligned}$$

Thus

$$\text{cov}(1(i < r_\ell), 1(k < r_\ell)) = \frac{n-k}{n} - \frac{n-i}{n}\frac{n-k}{n} = \frac{(n-k)i}{n^2}$$

and for $j \ne \ell$,

$$\text{cov}(1(i < r_j), 1(k < r_\ell)) = \frac{n-i-1}{n-1}\frac{n-k}{n} - \frac{n-i}{n}\frac{n-k}{n} = \frac{-i(n-k)}{n^2(n-1)}.$$

Thus for $0 \le i \le k \le n$,

$$\begin{aligned}\text{cov}(d_i(\boldsymbol{p}), d_k(\boldsymbol{p})) &= \sum_{j=1}^{i}\sum_{j \ne \ell=1}^{k} \frac{(-i)(n-k)}{n^2(n-1)} + \sum_{j=1}^{i} \frac{i(n-k)}{n^2} \\ &= \frac{(-i)(n-k)}{n^2(n-1)} i(k-1) + \frac{i(n-k)}{n^2} i = \frac{i^2(n-k)^2}{n^2(n-1)}.\end{aligned}$$

as desired. Again, the result depends only on the rectangular nature of the sets A_t, so that similar results could be stated for general rectangles.

From the above, the covariance structure of the limiting process can be suggested: If i, k and n increase while $i/n \to t_1$ and $k/n \to t_2$ for $0 < t_1 \leq t_2 < 1$, then the limiting covariance is $t_1^2(1-t_2)^2$.

In order to study the asymptotic behavior of $n^{1/2}R_g$, we describe this normalized rank coefficient in terms of the normalized rank processes introduced in Pyke (1985). For any collection C of Borel sets in I^2, define the rank process $S_n = \{S_n(C) : C \in C\}$ by

$$S_n(C) = n^{1/2}\{R_n(C) - |C|\}, \quad C \in C,$$

where R_n is the rank measure defined in (1.3). From (1.4), it follows that

$$n^{1/2}R_g = 2\sup_{\mathcal{B}}\{S_n(B) + n^{1/2}|B|\} - 2\sup_{\mathcal{A}}\{S_n(A) + n^{1/2}|A|\}. \tag{2.1}$$

By means of this representation of R_g as a function of the set-indexed rank process S_n, its limiting normality can be derived from the weak convergence of S_n that was established in Pyke (1985). To this end, rewrite (2.1) as

$$n^{1/2}R_g = 2\{S_n(B_{1/2}) - S_n(A_{1/2})\} + 2\sup_{\mathcal{B}}\{S_n(B) - S_n(B_{1/2}) - n^{1/2}(1/4 - |B|)\}$$
$$-2\sup_{\mathcal{A}}\{S_n(A) - S_n(A_{1/2}) - n^{1/2}(1/4 - |A|)\}. \tag{2.2}$$

Theorem 2.1 of Pyke (1985) states that under certain assumptions on the index family C, there exists a probability space on which equivalent versions of the S_n-processes and a Gaussian process S can be defined for which $\sup_C |S_n(C) - S(C)|$ converges to zero as $n \to \infty$. The assumptions on C for this form of weak convergence can be straightforwardly checked to be satisfied for both $\mathcal{A}$ and $\mathcal{B}$ of this paper, and hence for $C = \mathcal{A} \cup \mathcal{B}$. Both families inherit the metric structure of $[0,1]$ and the rectangular shape of the sets permits Assumption II of Pyke (1985) to hold. The limiting process, S, is a mean zero Gaussian process with covariance given by

$$\operatorname{cov}(S(A), S(B)) = |A \cap B| + |A||B| - \int_I |A \cap (B_{1y} \times I)|\,dy - \int_I |A \cap (I \times B_{2x})|\,dx$$

for any $A, B \in C$, where the sections B_{1y} and B_{2x} are defined by $B_{1y} = \{x \in I : (x,y) \in B\}, B_{2x} = \{y \in I : (x,y) \in B\}$. It is easily calculated that when $C = \mathcal{B}$, this covariance reduces to

$$\operatorname{cov}(S(B_s), S(B_t)) = s^2(1-t)^2, \quad 0 \leq s \leq t \leq 1, \tag{2.3}$$

which agrees with (1.4). It is also true that S is continuous over C where continuity is with respect to the Lebesgue symmetric-difference pseudometric, $d(A,B) = |A \Delta B|$.

The limiting null distribution of $n^{1/2}R_g$ can now be obtained from (2.1) since it can be shown that for the versions of S_n and S in Pyke (1985), both of the suprema converge to zero. The argument is the same for both; we consider only the term involving B. Observe first that $|B_t| \leq 1/4 = |B_{1/2}|$ for all $B \in \mathcal{B}$. The deterministic function $n^{1/2}(1/4 - |B|)$ is therefore non-negative and diverges to $+\infty$ except at $B_{1/2}$, where it is zero for every n. The supremum is clearly non-negative; try $B = B_{1/2}$. Since S_n converges to S uniformly over $\mathcal{B}$, it suffices to show that

$$s_n \equiv \sup_{\mathcal{B}}\{S(B) - S(B_{1/2}) - n^{1/2}(1/4 - |B|)\} \to 0. \tag{2.4}$$

Since S is continuous over $\mathcal{B}$ and $\mathcal{B} = \{B_t : t \in I\}$, S is uniformly bounded, by L say. Also, for any given $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\sup\{S(B_t) - S(B_{1/2}) : |t - 1/2| < \delta\} < \varepsilon.$$

Consequently, for all n,

$$s_n \leq \varepsilon + \sup_{|t-1/2|\leq\delta}\{2L - n^{1/2}(1/4 - t(1-t))\}^+ \to \varepsilon.$$

where x^+ denotes the non-negative part of x. Since this is true for all $\varepsilon > 0$, it follows that $s_n \to 0$.

By a similar argument, the term in (2.2) that involves the supremum over $\mathcal{A}$ also converges to zero. Consequently, it follows that the right-hand side of (2.2) converges to

$$Z \equiv 2\{S(B_{1/2}) - S(A_{1/2})\} = 4S(B_{1/2}),$$

since $S([0, 1/2] \times I) = 0$ and $A_{1/2} \cup B_{1/2} = [0, \frac{1}{2}] \times I$. However, Z is a mean zero normal random variable whose variance, by (2.3) is 1. This completes the proof of

Theorem 2.1 *Under the uniform null hypothesis, $n^{1/2}R_g$ is asymptotically distributed as a $N(0,1)$ random variable.*

3 The Limiting Distribution under Alternatives

In order to be able to evaluate the power or efficiency of procedures based on R_g, it is necessary to know the distribution of R_g under alternatives as well as under the null hypothesis. For moderate sample sizes this can best be done by simulation. For large sample sizes the asymptotic distribution is needed. For this reason, we outline below methods for obtaining the limiting distributions of R_g under both fixed and contiguous alternatives to the null-hypothesis assumption of independence.

Consider first the case of a fixed alternative, H. We assume that both marginals of H are continuous, and so, without loss of generality, we take the marginals to be uniform on [0, 1], even though H itself is not uniform on I^2. In this situation, the rank process is

$$S_n(C) = n^{1/2}\{R_n(C) - H(C)\}, \qquad C \in \mathcal{C}. \tag{3.1}$$

As for the null-hypothesis case above, these limiting results for R_g depend on knowing the weak convergence of the rank process. Since the results for non-uniform (non-independence) cases require lengthy proofs, we outline here only the steps needed to obtain the results for R_g *once one has the results for the rank process.* The fuller study of the non-uniform rank process is left for a later study; in what follows, the needed results for the rank process are stated as assumptions.

As for (2.1), one may write

$$n^{1/2}R_g = 2\sup_{\mathcal{B}}\{S_n(B) + n^{1/2}H(B)\} - 2\sup_{\mathcal{A}}\{S_n(A) + n^{1/2}H(A)\}. \tag{3.2}$$

Set

$$a = \sup_{\mathcal{A}} H(A), \quad b = \sup_{\mathcal{B}} H(B)$$

and

$$\mathcal{A}^* = \{A \in \mathcal{A} : H(A) = a\}, \quad \mathcal{B}^* = \{B \in \mathcal{B} : H(B) = b\}.$$

Clearly $0 < a, b < 1/2$ since, for example, $H(A_t) \le t \wedge (1-t)$ for all $t \in I$. Suppose that $\mathcal{A}^*$ and $\mathcal{B}^*$ are finite, say $\mathcal{A}^* = \{A_i^* : 1 \le i \le k\}$ and $\mathcal{B}^* = \{B_j^* : 1 \le j \le m\}$. The limiting distributions of the terms on the right hand side of (3.2) are determined by the members of $\mathcal{A}^*$ and $\mathcal{B}^*$. Consider the term involving the supremum over $\mathcal{B}$. Partition $\mathcal{B}$ into a finite union of sets, say $\mathcal{B} = \cup_1^m \mathcal{B}_j$, such that each $\mathcal{B}_j$ contains exactly one element of $\mathcal{B}^*$ and that element is an interior point. Then write

$$\begin{aligned}&\sup_{\mathcal{B}}\{S_n(B) + n^{1/2}H(B)\}\\ &\quad = \max_j\{S_n(B_j^*) + \sup_{\mathcal{B}_j}\{S_n(B) - S_n(B_j^*) - n^{1/2}(b - H(B))\}\} + n^{1/2}b.\end{aligned}$$

By a similar argument to that used to prove (2.4), one may show that the middle term involving the supremum over $\mathcal{B}_j$ converges to zero. This fact, together with a similar one for $\mathcal{A}$, enables one to deduce from (3.1) that

$$n^{1/2}(R_g - b + a) \xrightarrow{L} 2\max_{1\le j\le m} S(B_j^*) - 2\max_{1\le i\le k} S(A_i^*),$$

provided only that on some probability space there exists a process S and equivalent versions of the rank processes S_n for which $\|S_n - S\|_{\mathcal{A}\cup\mathcal{B}} \to 0$, where $\|f\|_M$ is the sup-norm, $\|f\|_M = \sup\{|f(x)| : x \in M\}$. This establishes the limiting distribution under H. In summary,

Theorem 3.1 *For a fixed alternative H with uniform marginals, if $a = \sup\{H(A) : A \in \mathcal{A}\}$ and $b = \sup\{H(B) : B \in \mathcal{B}\}$ are attained on finite subfamilies $\mathcal{A}^*$ and $\mathcal{B}^*$, respectively, and if S_n converges in distribution to S in the sense that versions exists for which $\|S_n - S\|_{\mathcal{A}\cup\mathcal{B}} \to 0$, then $n^{1/2}(R_g - b + a)$ converges in law to*

$$2\sup_{\mathcal{B}^*} S(B) - 2\sup_{\mathcal{A}^*} S(A).$$

When H is uniform on I^2, so that one is in the null-hypothesis case, $\mathcal{A}^* = \{A_{1/2}\}$ and $\mathcal{B}^* = \{B_{1/2}\}$. Thus Theorem 3.1 is consistent with Theorem 2.1.

To obtain the limiting distribution under a sequence of contiguous alternatives, say $\{H^{(n)}\}$, one first needs to have a convergence result for the rank processes that is in some sense uniform over alternatives. To this end, suppose $\mathcal{H}$ is a family of distributions H on I^2 that have uniform marginals. Interpret the rank process in (3.1) as indexed by both a family $\mathcal{C}$ of sets and such a family $\mathcal{H}$ of distributions. That is, view the rank process as

$$S_n(C, H) = n^{1/2}\{R_n(C) - H(C)\}, \qquad (C, H) \in \mathcal{C} \times \mathcal{H}. \tag{3.3}$$

Moreover, when the true distribution of (X_i, Y_i) is H, it is possible to construct equivalent observations that are functions of random variables that are uniform on I^2. In this way, one can simultaneously construct all of the rank processes embodied in (3.3) on the one (null-hypothesis) probability space. Suppose it is then possible to show the convergence-in-law of the process S_n to a limiting process, say $S = \{S(C, H) : C \in \mathcal{C}, H \in \mathcal{H}\}$, in such a way that one may assume without loss of generality that $\|S_n - S\|_{\mathcal{C}\times\mathcal{H}} \to 0$. Suppose $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$ and $\mathcal{H}$ is a family containing the specified contiguous sequence of alternatives, $\{H^{(n)}\}$. If the sequence converges to the uniform null hypothesis at a rate of $n^{-1/2}$ so that

$$\|n^{1/2}(H^{(n)}(\cdot) - |\cdot|) - \nu(\cdot)\|_{\mathcal{C}} \to 0$$

for some bounded set function ν, then arguments similar to those applied to (1.4) can be used to show that

$$n^{1/2}R_g \xrightarrow{L} 2\{S(B_{1/2}) + \nu(B_{1/2}) - S(A_{1/2}) - \nu(A_{1/2})\}.$$

The major step in deriving such a result is again that of establishing the weak convergence of the basic rank process S_n. This will be the focus of later research, but it should be remarked here that conditions on $\mathcal{H}$ will permit large families of alternatives and hence of contiguous sequences.

References

Gideon, R.A. and Hollister, R.A. (1987). A rank correlation coefficient resistant to outliers. *Journal of the American Statistical Assoc.* **82**, 656–666.

Pyke, Ronald (1985). Opportunities for set-indexed empirical and quantile processes in inference. *Bulletin International Statistical Institute* **51**, Book #25.2, 1–11.