MULTIVARIATE NONPARAMETRIC PROCEDURES FOR CERTAIN ARTERIOSCLEROSIS PROBLEMS

by

Pranab Kumar Sen

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1444

July 1983

MULTIVARIATE NONPARAMETRIC PROCEDURES FOR CERTAIN ARTERIOSCLEROSIS PROBLEMS*†

Pranab Kumar Sen
Department of Biostatistics 201H
University of North Carolina, Chapel Hill

ABSTRACT

Multivariate nonparametric methods are developed for the statistical interpretation and analysis of some arteriosclerosis problems. In this context, multivariate proportional hazard models are incorporated in the formulation of suitable statistical tests as well as estimates. Some genuinely distribution-free tests for exchangeability having relevance to the arteriosclerosis problems are also considered.

AMS 1980 Subject Classification: 62G05, 62G10, 62H12, 62H15, 62J02.

Key Words & Phrases: Arteriosclerosis; combination of independent tests; empirical distributions; exchangeable models; interchangeability; linear rank statistics; non-linear regression; proportional hazards; signed rank statistics.

*Work supported by the National Heart, Lung, and Blood Institute, Contract No. NIH-NHLBI-71-2243-L from the National Institutes of Health.

†Paper presented at the Sixth International Symposium on Multivariate Analysis at the University of Pittsburgh, Penn., July 1983.

1. INTRODUCTION

Due to a process, termed Atherosclerosis, a deposit of some fatty substances in the arterial walls may cause Arteriosclerosis, a disorder in which arteries become thick and hard and lose their supple, elastic quality. Though the genesis of Atherosclerosis is not fully known (and it is believed to be a part of the normal aging process), the cardio-vascular specialists have a hunch that Arteriosclerosis is more likely to occur or to be severe for people belonging to one or more of the following groups: Regular smokers, overweights, inactives, and having diabetes, hypertension, or increased levels of LIPIDS and cholesterol in the blood. The National Institutes of Health (NIH) have undertaken several large scale (multi-center) clinical trials involving various cross sections of human subjects, and these may release a lot of information on this subject. However, these projects, involving humanbeings, do not involve the sacrifice of the subjects at pre-specified time points and microscopic study of their vital arterial cross-sections (at designated points) to gather information on the exact amount of fatty deposits and their relation to the various assignable factors. These studies are more of epidemiological nature, and the comparative effects of various assignable factors are studied through the survival pattern and formulation of suitable statistical models relating to the survival functions. There are other studies, some undertaken by the NIH and some by others, involving non-human primates (e.g., monkeys, cats, dogs, etc.), where under suitable experimental designs (relating to the assignable factors), internal examinations of the arteries of the sacrificed subjects (at pre-assigned time points) have revealed useful information on Atherosclerosis.

From the point of view of statistical modeling and analysis, such a direct study may pose quite a few interesting issues, and some of these will be addressed here. In this context, conventional parametric models often appear to be quite restrictive in scope and validity, and hence, their nonparametric counterparts will be pursued here. Moreover, for a sacrificed subject, arterial cross-sections are examined simultaneously at various key positions of the body to study the nature of variation (over these positions) and this calls for genuine multivariate models. Hence, multivariate nonparametric procedures are deemed quite appropriate in this context.

Along with a brief description of Arteriosclerosis, some of the basic statistical problems are outlined in Section 2. Section 3 is devoted to the nonparametric estimation of the marginal proportionality hazard parameters in a multivariate setup. Tests of significance on these functions are then formulated in Section 4. Some genuinely distribution-free tests for interchangeability having relevance to the arteriosclerosis problem are discussed in Section 5. Combinations of independent test statistics are also considered. The concluding section deals with some general remarks.

2. SOME STATISTICAL PROBLEMS ARISING IN ARTERIOSCLEROSIS

We start with a comparative picture of a cross-section of a normal artery and another one under Arteriosclerosis.

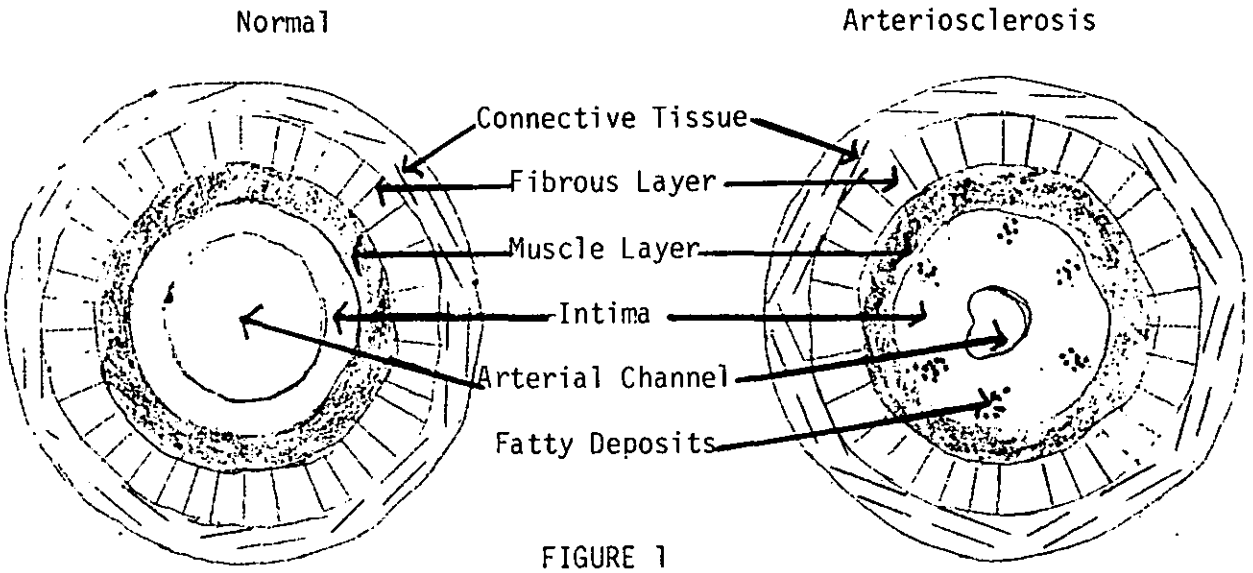

FIGURE 1

Typically, the deposit of fatty substance enlarges the intima and the arterial channel reduces. A narrowed artery to the heart muscle may cause Angina Pectoris or a heart attack. If the arteries to the brain are affected, strokes may occur, while reduced blood flow to the brain may result in Parkinson's disease. Poor circulation of blood to the limbs or legs may result if an artery in the limb or leg is affected and it may have more serious effects too. If a thrombos (a piece of fatty substance) breaks away when the artery is brittle and it enters into the bloodstream, it may cause a blockage leading to a heart attack. Thus, not all arteries may be equally affected by arteriosclerosis and the exact seriousness of the conditions may depend on the locations and intensities of the arteries affected. In the case of coronary arteries, reduced blood supply and/or blockage may cause a heart attack: Even in such a case, the left coronary artery is more commonly affected than the right one.

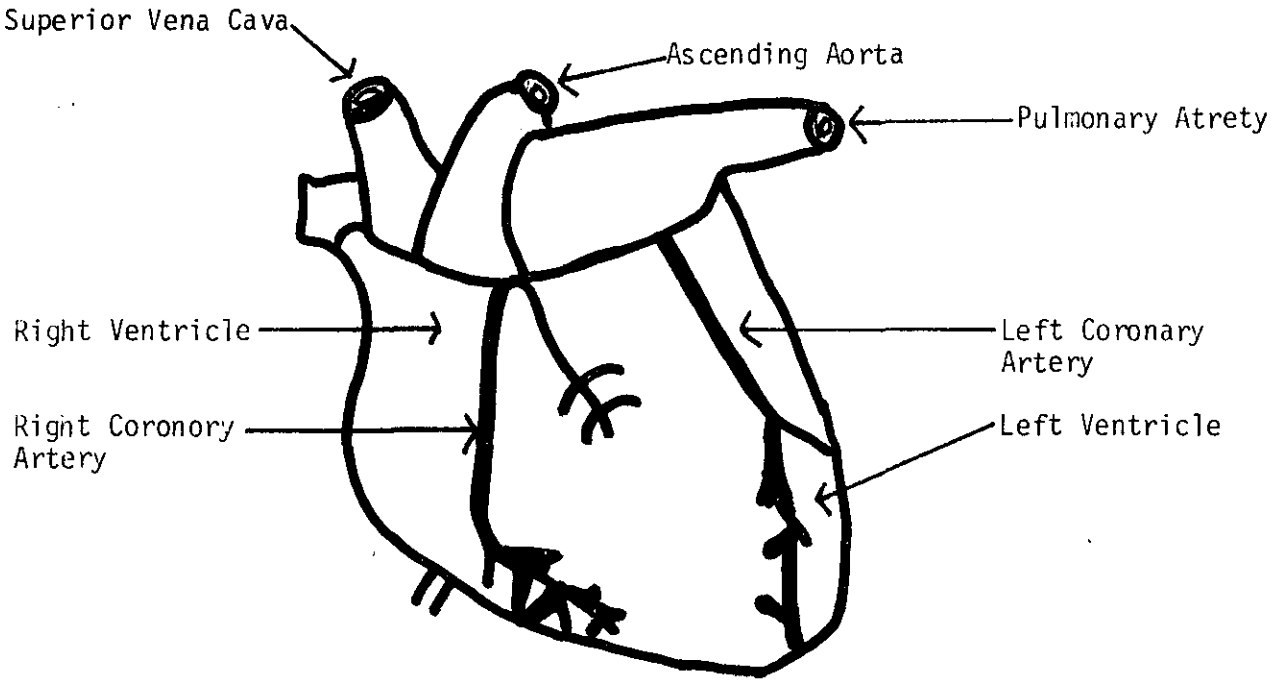

FIGURE 2

One of the basic problems in the study of Arteriosclerosis is the proper choice of response variables. The arteries are spread all over the body, they differ in shapes, and have relatively different activities too. Thus, a choice of a specific number and positions of these arteries may well depend on the specific objectives of the study. Even for specific arteries, the choice of the designated positions (at which cross-sections are to be examined) is of considerable importance. For example, with reference to the coronary disorders, one may be mainly interested in comparing the left and right coronary arteries, and, for this purpose, one needs to choose appropriate points at which they should be otherwise comparable. Even when a design of the arterial positions has been selected, there remains a basic question: What is the most appropriate measure of the intensity of Arteriosclerosis? Should we take the amount of the fatty

deposits in a cross-section of pre-fixed width or should we compare the actual cross-section of the arterial channels through which blood can flow? The first choice is likely to be more variable (less informative) and, usually, the second one is preferred. Moreover, the cross-sections of the arteries (and the channels) at different positions may not be very close to each other, and hence, the ratio of the area of the arterial channel to the whole area formed by the outer walls of the intima is usually taken as a suitable measure of the response variable. Thus, usually, the response variable assumes values in the unit interval $[0,1]$, but is not of binary (0 or 1) character. For a designated set of p (≥ 1) positions (on one or more of the arteries under study), for each subject, the arterial cross-sections lead to a p-vector $\underset{\sim}{U} = (U_1,\ldots,U_p)'$ of responses and the support of the distribution (G) of $\underset{\sim}{U}$ is the unit p-cube $[0,1]^p$. Keeping in mind a typical experimental setup (involving non-human primates) involving several groups of subjects with possibly different diets and different length of exposure times, we may thus conceive of a multi-group model, and the information on the relationship of Arteriosclerosis and the assignable factors can be obtained by comparing the observed distributions of these responses. Since these distributions have all a common compact support, the conventional location or dispersion models may not be very appropriate in this context. For simplicity, consider the case of $p = 1$ and let $\bar{G}(t) = P\{U > t\}$, $0 \leq t \leq 1$ be the survival function of U (i.e., $\bar{G}$ is the complementary part of the distribution function G). Let $\bar{G}_0$ and $\bar{G}_A$ denote the survival functions for the normal and Arteriosclerosis (for a given design) cases, respectively. Then, we may have typically the following situation:

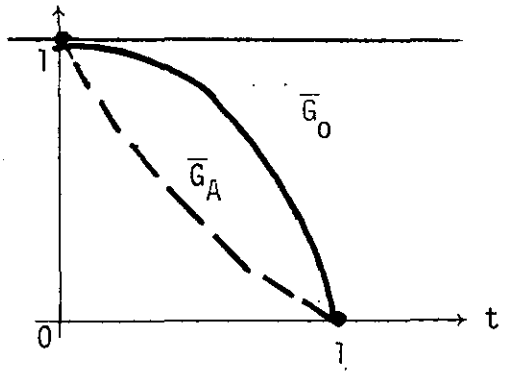

FIGURE 3

It is always possible to express $\bar{G}_A = M(\bar{G}_0)$, where $M(t)$, $0 < t < 1$, is given by $M = \bar{G}_A \bar{G}_0^{-1}$ and is monotone and nonnegative. We write

$$\bar{G}_A(t) = [\bar{G}_0(t)]^{h_A(t)}, \quad 0 \leq t \leq 1, \tag{2.1}$$

where $h_A(t)$ may depend on the factors (diet, time, etc.) in A and t. Tempted by the popularity of the Lehmann (1953) model and its widespread adaptation in the Cox (1972) regression model, with respect to (2.1), we assume that

$$h_A(t) = h(\underset{\sim}{x}_A), \quad \forall\ 0 < t < 1, \tag{2.2}$$

where $\underset{\sim}{x}_A$ stands for the design point (vector) for the factors assigned. Even with this simplified assumption, we encounter some other problems. Note that by (2.2), we have

$$\{-\log\bar{G}_A(t)\}/\{-\log\bar{G}_0(t)\} = h(x_A), \quad \forall\ t\varepsilon(0,1), \tag{2.3}$$

so that the two log-survival functions are proportional to each other; if these distributions admit probability density functions, then (2.3) reduces to the proportionality of the two hazard functions [Cox (1972)]. We may not need the existence of the hazard function, though in view of

the above, we term $h(x_A)$ as a PH (-proportional hazard) function. Now, in our setup, generally $\underset{\sim}{U}$ is a p-vector, for some $p \geq 1$, so that in (2.1)-(2.3), we need to deal with the multivariate survival function $\bar{G}(\underset{\sim}{t}) = P\{\underset{\sim}{U} \geq \underset{\sim}{t}\}$, $\underset{\sim}{t} \in [0,1]^p$, and this will relate to the constancy in (2.1)-(2.3) for all $\underset{\sim}{t} \in [0,1]^p$. This is more restrictive and less realistic: Even if for each marginal survival function, (2.1)-(2.3) hold, the form of the marginal PH-functions may not be all the same. Hence, to be more realistic, we denote the p marginal survival functions by $\bar{G}_{[j]}$, $j=1,\ldots,p$, and assume that for each j, (2.1)-(2.3) hold, where the $h_j(x_A)$ need not be all equal. These marginal PH-functions play a vital role in our subsequent analysis in the sections to follow.

3. NONPARAMETRIC ESTIMATION OF PH-FUNCTIONS

We conceive of a simple design involving K (≥ 2) groups of subjects (non-human primates) which are given some drugs/foods (containing fatty substances), placed in possibly different environments (e.g., levels of carbon monoxides, etc.) and sacrificed after specified lengths of times; cross-sections of the arteries at some specified ($p \geq 1$) positions are then examined and the corresponding p-vector $\underset{\sim}{U}$ is then defined as in Section 2. Thus, the design point $\underset{\sim}{x} = (x_1,\ldots,x_q)$ (for some $q \geq 1$) can be defined by means of regression vectors $\underset{\sim}{c}_i = (c_{i1},\ldots,c_{iq})'$, $1 \leq i \leq K$ where the c_{ik} may be real or dummy variables, and the n_i subjects allotted to the ith group yield the response vectors $\underset{\sim}{U}_{ik} = (U_{ik1},\ldots,U_{ikp})'$, $k=1,\ldots,n_i$, for $i=1,\ldots,K$. We assume that $\underset{\sim}{U}_{i1},\ldots,\underset{\sim}{U}_{in_i}$ are iid (independent and identically distributed) rv (random vectors) with a continuous p-variate df (distribution function) $G_i(\underset{\sim}{t})$, $\underset{\sim}{t} \in [0,1]^p$. for $i=1,\ldots,K$. The

p marginal distributions corresponding to the df G_i are denoted by $G_{i[j]}$, $1 \le j \le p$ and the corresponding survival functions by $\bar{G}_{i[j]}$, j=1,...,p; i=1,...,K. As in (2.1)-(2.2), we assume that

$$\bar{G}_{i[j]}(t) = [\bar{G}_{o[j]}(t)]^{h_j(\underset{\sim}{c}_i)}, \quad t\in[0,1],\ 1 \le j \le p,\ 1 \le i \le K, \quad (3.1)$$

where the $\bar{G}_{o[j]}$ are unknown survival functions and the PH-functions $h_j(\underset{\sim}{c}_i)$ depend on the regressor $\underset{\sim}{c}_i$ (and their forms may also depend on j:$1\le j\le p$). Note that the $h_j(\underset{\sim}{c}_i)$ are nonnegative, and hence, we prefer to use the following multiplicative model:

$$\log h_j(\underset{\sim}{c}_i) = \mu_j + \underset{\sim}{\beta}_j'\underset{\sim}{c}_i,\ 1 \le i \le K;\ 1 \le j \le p, \quad (3.2)$$

where μ_j and $\underset{\sim}{\beta}_j' = (\beta_{j1},\ldots,\beta_{jq})$, $1 \le j \le p$ are unknown parameters. Note that as the $\bar{G}_{o[j]}$ $(1 \le j \le p)$ are not assumed to be known, not all the parameters in (3.2) may be identifiable and estimable. For example, by the use of the transformation $\bar{G}_{o[j]} \to (\bar{G}_{o[j]})^{\exp \mu_j}$, we may always take $\mu_j = 0$, $\forall\ 1 \le j \le p$. Hence, in (3.2) we drop the μ_j. Further parametric restraints may be introduced so that the identifiability and estimability conditions are satisfied. For example, if p = 1 and K = q = 3, based on appropriate rank statistics, we may be able to estimate $\beta_{12} - \beta_{11}$, $\beta_{13} - \beta_{11}$ or $\beta_{13} - \beta_{12}$, but not the individual β_{1i}. Hence, as in a classical ANOVA model, we may absorb $\beta_{11} + \beta_{12} + \beta_{13} = \beta_{1\cdot}$ into μ_j and set $\beta_{11} + \beta_{12} + \beta_{13} = 0$. Then, the parameters are estimable.

To fix the basic ideas, consider an experiment involving n non-human primates, divided into K (= ab) subsets of $n_1,\ldots,n_K$ sizes. Let there be a (≥ 2) diet plans, designated by $d_1,\ldots,d_a$, respectively, and b (≥ 2) different follow-up periods $T_1,\ldots,T_b$, respectively. Typically, for a

subset of the primates under a diet plan d_r and a follow-up period T_s, arterial cross-sections (at designated positions) of the sacrificed animals are examined and the response vector $\underset{\sim}{U}$ are recorded. In this setup, q is $\leq$ K and the regressors $\underset{\sim}{c}_i$ may be defined by appropriate dummy variables. If we assume that there is no diet vs follow-up period interaction, then we may take $q = a + b - 2$ with $a - 1$ diet effects and $b - 1$ follow-up period effects. On the other hand, if these interaction effects are incorporated in the model, we may take $q = ab - 1$, with $a - 1$ diet effects, $b - 1$ follow-up effects, and $(a-1)(b-1)$ interaction effects. In either setup, the parameters entering into the model are identifiable and estimable. If $d_1,\ldots,d_a$ represent the different levels (of fatty substances) of a common diet plan, then the main effects and the interactions may be sorted as linear, quadratic and higher order components, and neglecting appropriate components, q may be chosen between 1 and K - 1.

Consider now a general K ($\geq$ 2) sample model, and let $\underset{\sim}{U}_{i\alpha} = (U_{i1\alpha},\ldots, U_{ip\alpha})'$, $\alpha=1,\ldots,n_i$ be the vectors of the responses (as defined in Section 2) for the ith sample observations, for i=1,...,K. For each j (=1,...,p), let $R_{ij\alpha}$ be the rank of $U_{ij\alpha}$ among the $n\ (= \sum_{i=1}^{K} n_i)$ observations $U_{rj\beta}$, $1 \leq \beta \leq n_r$, $1 \leq r \leq K$, for i=1,...,K and $\alpha=1,\ldots,n_i$. Also, for each j (=1,...,p) and $n\ (\geq 1)$, consider a set of scores $a_{nj}(\alpha)$, $1 \leq \alpha \leq n$, defined by

$$a_{nj}(\alpha) = EQ_j(\xi_{n\alpha}),\ 1 \leq \alpha \leq n,\ j=1,\ldots,p, \tag{3.3}$$

where $\xi_{n1} <\ldots< \xi_{nn}$ stand for the ordered rv of a sample of size n from the uniform (0,1) df and the score functions Q_j are expressible as differences of two non-decreasing, absolutely continuous and square integrable functions inside (0,1). We also write

$$Q_j^*(u) = Q_j(1-u),\ 0 < u < 1,\ j=1,\ldots,p. \tag{3.4}$$

Consider then the set of rank statistics

$$T_{n,ij} = n_i^{-1} \sum_{\alpha=1}^{n_i} a_{nj}(R_{nij\alpha}),\ 1 \le i \le K;\ 1 \le J \le p. \tag{3.5}$$

These linear rank statistics are employed for the estimation of $\underset{\sim}{\beta}_1,\ldots,\underset{\sim}{\beta}_p$. We define the $\bar{G}_{i[j]}$ as in (3.1) and let

$$\bar{H}_{[j]}(t) = \sum_{i=1}^{K} \lambda_{ni}\bar{G}_{i[j]}(t),\ 0 \le t \le 1;\ H_{[j]} = 1 - \bar{H}_{[j]}; \tag{3.6}$$

$$\lambda_{ni} = n^{-1}n_i,\ \text{for } i=1,\ldots,K. \tag{3.7}$$

Let then

$$\begin{aligned}
\tau_{ij} &= \int_0^1 Q_j(1 - \bar{H}_{[j]}(t))dG_{i[j]}(t) \\
&= \int_0^1 Q_j^*(\bar{H}_{[j]}(t))d\bar{G}_{i[j]}(t) \\
&= \int_0^1 Q_j^*\Big(\sum_{r=1}^{K} \lambda_{nr}\ \bar{G}_{r[j]}(t)\Big)d\bar{G}_{i[j]}(t) \\
&= \int_0^1 Q_j^*\Big(\sum_{r=1}^{K} \lambda_{nr}\ u^{h_j(\underset{\sim}{c}_r)/h_j(\underset{\sim}{c}_i)}\Big)du,
\end{aligned} \tag{3.8}$$

for $1 \le i \le K$, $1 \le j \le p$. Note that the τ_{ij} may depend on n through the λ_{ni}. By (3.2), for every i, r (=1,...,K) and j (=1,...,p),

$$h_j(\underset{\sim}{c}_r)/h_j(\underset{\sim}{c}_i) = \exp(\underset{\sim}{\beta}_j'(\underset{\sim}{c}_r - \underset{\sim}{c}_i)). \tag{3.9}$$

Thus, for given Q_j^*, $\underset{\sim}{\lambda}_n = (\lambda_{n1},\ldots,\lambda_{nK})'$ and $\underset{\sim}{c}_1,\ldots,\underset{\sim}{c}_K$,

$$\tau_{ij} = \tau_{ij}(\underset{\sim}{\beta}_j; Q^*, \underset{\sim}{\lambda}_n, \underset{\sim}{c}_1,\ldots,\underset{\sim}{c}_K),\ 1 \le i \le K,\ 1 \le j \le p, \tag{3.10}$$

and, these are, in general, non-linear functions of the unknown $\underset{\sim}{\beta}_j$. On the other hand, from the general theory of multivariate multisample rank

statistics [n.2, Chapter 5 of Puri and Sen (1971)], it follows that the $T_{n,ij}$ estimate consistently the τ_{ij}, and we incorporate this in the formulation of our proposed estimators. In passing, we may remark that for the special case of the Wilcoxon scores (i.e., $Q_j(u) \equiv u$), by (3.8)-(3.9), $\tau_{ij} = \sum_{r=1}^{K} \lambda_{nr}\{1 + \exp(\underset{\sim}{\beta}_j'(\underset{\sim}{c}_r - \underset{\sim}{c}_i))\}^{-1}$, for $1 \le i \le K$, $1 \le j \le p$.

We employ here the weighted non-linear (iterative) least squares estimation theory for the estimation of the $\underset{\sim}{\beta}_j$. We need to start with some initial estimator, for which we use the two-sample Wilcoxon statistics $W_{ii',j}$ based on $U_{i\alpha j}$, $1 \le \alpha \le n_i$ and $U_{i'\alpha j}$, $1 \le i' \le n_{i'}$, for $1 \le i < i' \le K$, $1 \le j \le p$. Note that for $Q_j(u) \equiv u$, the corresponding population counterparts are

$$\omega_{ii',j} = \{1 + \exp(\underset{\sim}{\beta}_j'(\underset{\sim}{c}_i - \underset{\sim}{c}_{i'}))\}^{-1}, \quad 1 \le i \le i' \le K, \ 1 \le j \le p, \tag{3.11}$$

so that

$$\log((1-\omega_{ii',j})/\omega_{ii',j}) = \underset{\sim}{\beta}_j'(\underset{\sim}{c}_i - \underset{\sim}{c}_{i'}), \ 1 \le i \le i' \le K, \ 1 \le j \le p. \tag{3.12}$$

Thus, equating the $\log((1-W_{ii',j})/W_{ii',j})$ to the right hand side of (3.12), we get for each j, a set of $K(K-1)/2$ equations in q ($\le K-1$) unknown, and hence, any convenient average of these estimates may be taken as the initial estimator $\hat{\underset{\sim}{\beta}}_{nj}^{(o)}$, for $j=1,\ldots,p$. Also, for each j ($=1,\ldots,p$), we denote by

$$\underset{\sim}{T}_{nj} = (T_{n,1j},\ldots,T_{n,Kj})' \text{ and } \underset{\sim}{\tau}_j = (\tau_{1j},\ldots,\tau_{Kj})'; \tag{3.13}$$

$$\underset{\sim}{D}_j(\underset{\sim}{\ell}) = (\partial \underset{\sim}{\zeta}_j/\partial \underset{\sim}{\beta}_j)\big|_{\underset{\sim}{\beta}_j = \underset{\sim}{\ell}}, \ \underset{\sim}{\ell} = (\ell_1,\ldots,\ell_q)'. \tag{3.14}$$

Note that the $\underset{\sim}{D}_j$ are $K \times q$ matrices, and we let $\underset{\sim}{D}_j^{(o)} = \underset{\sim}{D}_j(\hat{\underset{\sim}{\beta}}_{nj}^{(o)})$, $1 \le j \le p$. Also, for each j ($=1,\ldots,p$), let $\underset{\sim}{v}_j = ((v_{ii',j}))$ be defined by

$$\nu_{ii,j} = 2 \sum_{\substack{r=1 \\ r\neq i}}^{K} \lambda_{nr} \iint_{0\le s\le t\le 1} A_r(s,t)\, dG_{i[j]}(s)dG_{i[j]}(t) +$$

$$2\lambda_{ni}^{-1} \sum_{\substack{r=1 \\ r\neq i}}^{K} \lambda_{nr}^2 \iint_{0\le s\le t\le 1} A_i(s,t)\, dG_{r[j]}(s)dG_{r[j]}(t) +$$

$$\lambda_{ni}^{-1} \sum_{\substack{r\neq r'=1 \\ r\neq i, r'\neq i}}^{K} \lambda_{nr}\lambda_{nr'} \iint_{0\le s\le t\le 1} A_i(s,t)\Big\{dG_{r[j]}(s)dG_{r'[j]}(t) + dG_{r'[j]}(s)dG_{r[j]}(t)\Big\}; \tag{3.15}$$

$$\nu_{ii',j} = -\sum_{r=1}^{K} \lambda_{nr} \iint_{0\le s\le t\le 1} \Big\{A_i(s,t)\, dG_{r[j]}(s)dG_{i'[j]}(t) +$$

$$A_i(s,t)\, dG_{r[j]}(t)dG_{i'[j]}(s) + A_{i'}(s,t)\, dG_{r[j]}(s)dG_{i[j]}(t) +$$

$$A_{i'}(s,t)\, dG_{r[j]}(t)dG_{i[j]}(s) + A_r(s,t)\, dG_{i[j]}(s)dG_{i'[j]}(t) +$$

$$A_r(s,t)\, dG_{i[j]}(t)dG_{i'[j]}(s)\Big\}, \tag{3.16}$$

for $i\neq i'=1,\ldots,K$, where

$$A_r(s,t) = G_{r[j]}(s)[1-G_{r[j]}(t)]\, Q_j'(H_{[j]}(s))Q_j'(H_{[j]}(t)), \tag{3.17}$$

for $0\le s\le t\le 1$, $r=1,\ldots,K$, $1\le j\le p$. Note that by (3.2), (3.6), and (3.9), for every j $(=1,\ldots,p)$, $\underset{\sim}{\nu}_j$ [for an adapted Q_j, given $\lambda_{n1},\ldots,\lambda_{nK}$ and the $\underset{\sim}{c}_i$] depends exclusively on $\underset{\sim}{\beta}_j$, but not on the unknown df $\bar{G}_{o[j]}$ in (3.1). Substituting $\hat{\underset{\sim}{\beta}}_{nj}^{(o)}$ for $\underset{\sim}{\beta}_j$, we denote the resulting $\underset{\sim}{\nu}_j$ by $\hat{\underset{\sim}{\nu}}_j^{(o)}$, for $j=1,\ldots,p$. Now, from Theorem 3.6.5 of Puri and Sen (1971), it follows that asymptotically for each j $(=1,\ldots,p)$,

$$n^{\frac{1}{2}}(\underset{\sim}{T}_{nj} - \underset{\sim}{\tau}_j) \sim N_K(\underset{\sim}{0}, \underset{\sim}{\nu}_j), \tag{3.18}$$

where $\underset{\sim}{\nu}_j$ is of rank $K-1$. Further, (3.18) provides the justification of

using the norm

$$n(\underset{\sim}{T}_{nj} - \underset{\sim}{\tau}_j)'\underset{\sim}{\nu}_j^- (\underset{\sim}{T}_{nj} - \underset{\sim}{\tau}_j) \tag{3.19}$$

(where $\underset{\sim}{\nu}_j^-$ is a generalized inverse for $\underset{\sim}{\nu}_j$) and minimizing the same with respect to $\underset{\sim}{\beta}_j$, to derive the desired estimators of $\underset{\sim}{\beta}_j$. Towards this, we first replace $\underset{\sim}{\beta}_j$ in (3.10) by $\hat{\underset{\sim}{\beta}}_{nj}^{(o)}$ and denote the resulting quantity by $\hat{\underset{\sim}{\tau}}_j^{(o)}$. Then, we have

$$\underset{\sim}{\tau}_j = \hat{\underset{\sim}{\tau}}_j^{(o)} + \underset{\sim}{D}_j^{(o)}(\underset{\sim}{\beta}_j - \hat{\underset{\sim}{\beta}}_{nj}^{(o)}) + o(||\underset{\sim}{\beta}_j - \hat{\underset{\sim}{\beta}}_{nj}^{(o)}||), \tag{3.20}$$

for $1 \le j \le p$. Thus, in (3.19), replacing $\underset{\sim}{\nu}_j$ by $\hat{\underset{\sim}{\nu}}_j^{(o)}$ and for $\underset{\sim}{\tau}_j$ using the first order approximation in (3.20), we obtain the first step estimator $\hat{\underset{\sim}{\beta}}_{nj}^{(1)}$ of $\underset{\sim}{\beta}_j$ as

$$\hat{\underset{\sim}{\beta}}_{n,j}^{(1)} = \hat{\underset{\sim}{\beta}}_{n,j}^{(o)} + (\underset{\sim}{D}_j^{(o)'}(\hat{\underset{\sim}{\nu}}_j^{(o)})^{-1}\underset{\sim}{D}_j^{(o)})^{-1}\ \underset{\sim}{D}_j^{(o)'}(\hat{\underset{\sim}{\nu}}_j^{(o)})^{-1}(\underset{\sim}{T}_{nj} - \hat{\underset{\sim}{\tau}}_j^{(o)}) \tag{3.21}$$

for j=1,...,p. By substitution of $\hat{\underset{\sim}{\beta}}_{n,j}^{(1)}$ in the appropriate places, we obtain $\hat{\underset{\sim}{\nu}}_j^{(1)}$, $\hat{\underset{\sim}{\tau}}_j^{(1)}$, and $\underset{\sim}{D}_j^{(1)}$, and hence,

$$\hat{\underset{\sim}{\beta}}_{n,j}^{(2)} = \hat{\underset{\sim}{\beta}}_{n,j}^{(1)} + (\underset{\sim}{D}_j^{(1)'}(\hat{\underset{\sim}{\nu}}_j^{(1)})^{-1}\underset{\sim}{D}_j^{(1)})^{-1}\ \underset{\sim}{D}_j^{(1)}\ (\hat{\underset{\sim}{\nu}}_j^{(1)})^{-1}(\underset{\sim}{T}_{nj} - \hat{\underset{\sim}{\tau}}_j^{(1)}) \tag{3.22}$$

for j=1,...,p. This iteration process continues until two consecutive step estimators differ by less than some preassigned small number. We denote this iterated estimator by

$$\hat{\underset{\sim}{\beta}}_n^* = (\hat{\underset{\sim}{\beta}}_{n,1}^{*'},\ldots,\hat{\underset{\sim}{\beta}}_{n,p}^{*'})'. \tag{3.23}$$

It may be noted that in the procedure considered above, we have actually used a coordinate-wise non-linear estimator of the $\underset{\sim}{\beta}_j$ $(1 \le j \le p)$.

The main reason for doing so is that the covariance of $\underset{\sim}{T}_{nj}$, $\underset{\sim}{T}_{nj'}$, $j \neq j'$, generally, depends on the unknown joint df $G_{o[jj']}$ even when the $\underset{\sim}{\beta}_j$ are known, while this is avoided in the coordinate-wise procedure. Actually, by an appeal to Theorem 5.5.1 of Puri and Sen (1971), it follows that asymptotically

$$n^{\frac{1}{2}}(\underset{\sim}{T}_{n1} - \underset{\sim}{\tau}_1, \ldots, \underset{\sim}{T}_{np} - \underset{\sim}{\tau}_p) \sim N_{Kp}(\underset{\sim}{0}, \underset{\sim}{\nu}) \tag{3.24}$$

where

$$\underset{\sim}{\nu} = \begin{Vmatrix} \underset{\sim}{\nu}_{11} & \underset{\sim}{\nu}_{12} & \cdots & \underset{\sim}{\nu}_{1p} \\ \vdots & \ddots & & \vdots \\ \underset{\sim}{\nu}_{p1} & \underset{\sim}{\nu}_{p2} & \cdots & \underset{\sim}{\nu}_{pp} \end{Vmatrix}, \tag{3.25}$$

$\underset{\sim}{\nu}_{jj} = \underset{\sim}{\nu}_j$, $1 \leq j \leq p$ are defined by (3.15)-(3.16), while $\nu_{jj'}$, $j \neq j' = 1, \ldots, p$ are defined as in (5.5.4)-(5.5.5) of Puri and Sen (1971). Let us define then $\underset{\sim}{D}_j = \underset{\sim}{D}_j(\underset{\sim}{\beta}_j)$, $1 \leq j \leq p$,

$$\underset{\sim}{\Gamma}_{jj'} = (\underset{\sim}{D}_j' \, \bar{\underset{\sim}{\nu}}_{jj} \, \underset{\sim}{D}_j)^{-1} \, \underset{\sim}{D}_j' \, \bar{\underset{\sim}{\nu}}_{jj} \, \underset{\sim}{\nu}_{jj'} \, \bar{\underset{\sim}{\nu}}_{j'j'} \, \underset{\sim}{D}_{j'} (\underset{\sim}{D}_{j'}' \, \bar{\underset{\sim}{\nu}}_{j'j'} \, \underset{\sim}{D}_{j'})^{-1} \tag{3.26}$$

for $j, j' = 1, \ldots, p$, and

$$\underset{\sim}{\Gamma} = \begin{Vmatrix} \underset{\sim}{\Gamma}_{11}, & \cdots, & \underset{\sim}{\Gamma}_{1p} \\ \vdots & & \\ \underset{\sim}{\Gamma}_{p1}, & \cdots, & \underset{\sim}{\Gamma}_{pp} \end{Vmatrix}. \tag{3.27}$$

Using (3.18)-(3.27) and following some standard steps it follows that asymptotically

$$n^{\frac{1}{2}}(\hat{\underset{\sim}{\beta}}_n^* - \underset{\sim}{\beta}) \sim N_{qp}(\underset{\sim}{0}, \underset{\sim}{\Gamma}). \tag{3.28}$$

We make use of this basic result in the next section in drawing inference on the PH-functions.

4. TESTING STATISTICAL HYPOTHESES ON THE PH-FUNCTIONS

In Section 3 it has been explained that the PH-functions $h_j(\underset{\sim}{c}_i)$, defined in terms of the $\underset{\sim}{\beta}_j$ and $\underset{\sim}{c}_i$, depend on both the design and the position effects. For a given position j (=1,...,p), the $\underset{\sim}{\beta}_j$ can conveniently be related to the main effects and interactions (through the $\underset{\sim}{c}_i$), and hence, suitable hypotheses may be framed in terms of these $\underset{\sim}{\beta}_j$.

Consider the $q \times p$ matrix $\underset{\sim}{\beta} = (\underset{\sim}{\beta}_1,\ldots,\underset{\sim}{\beta}_p)$ corresponding to the marginal models in (3.1)-(3.2) (with the μ_j equal to 0). Let $\underset{\sim}{M}$ be an $r \times q$ matrix of rank r ($\leq q$), and consider the null hypothesis

$$H_0: \underset{\sim\sim}{M\beta} = \underset{\sim}{0} \quad \text{vs} \quad H_1: \underset{\sim\sim}{M\beta} \neq \underset{\sim}{0}. \tag{4.1}$$

Define the estimator $\hat{\underset{\sim}{\beta}}^*_n$ of $\underset{\sim}{\beta}$ as in (3.23) and consider the $rp \times rp$ matrix $\underset{\sim}{B}_n$ defined by $\underset{\sim}{b}_n\underset{\sim}{b}_n'$ where $\underset{\sim}{b}_n$ is the rp-vector obtained by rolling out $\underset{\sim\sim}{M}\hat{\underset{\sim}{\beta}}^*_n$. Also, let $\hat{\underset{\sim}{D}}^*_j = \underset{\sim}{D}_j(\hat{\underset{\sim}{\beta}}^*_{n,j})$, $1 \leq j \leq p$, and let $\underset{\sim}{V}_n$ be a consistent estimator of $\underset{\sim}{\nu}$, defined by (3.25). Such a consistent estimator can be readily obtained by replacing the true df's by their sample counterparts in the definition of the $\underset{\sim}{\nu}_{jj'}$. Then, in (3.26), we replace the $\underset{\sim}{D}_j$ by $\hat{\underset{\sim}{D}}^*_j$ and $\underset{\sim}{\nu}_{jj'}$ by $\underset{\sim}{v}_{njj'}$, and denote the resulting matrices by $\hat{\Gamma}_{njj'}$, $j,j'=1,\ldots,p$. In (3.27), $\underset{\sim}{\Gamma}$ is replaced by the corresponding $\hat{\underset{\sim}{\Gamma}}_n$. Let then

$$\underset{\sim}{E}_n = \begin{pmatrix} \underset{\sim}{E}_{n11} & \cdots & \underset{\sim}{E}_{n1p} \\ \vdots & & \\ \underset{\sim}{E}_{np1} & \cdots & \underset{\sim}{E}_{npp} \end{pmatrix}; \quad \underset{\sim}{E}_{njj'} = \underset{\sim\sim}{M}\hat{\underset{\sim}{\Gamma}}_{njj'}\underset{\sim}{M}'. \tag{4.2}$$

Finally, let

$$\delta_n = n \, \text{Trace}(\underset{\sim}{B}_n\underset{\sim}{E}_n^-) = n\underset{\sim}{b}_n' \, \underset{\sim}{E}_n^-\underset{\sim}{b}_n. \tag{4.3}$$

By virtue of (3.28), (4.1), (4.2), and (4.3), we may conclude that under H_0: $\underset{\sim\sim}{M\beta} = 0$, δ_n has asymptotically chi-square distribution with rp degrees

of freedom (df). When the null hypothesis is not true (i.e., $\underset{\sim}{M}\underset{\sim}{\beta} \neq \underset{\sim}{0}$), using the inequality that

$$\begin{aligned}\delta_n &= n(\underset{\sim}{b}_n'\underset{\sim}{b}_n)\ \{\underset{\sim}{b}_n'\ \underset{\sim}{E}_n^{-}\underset{\sim}{b}_n/\underset{\sim}{b}_n'\underset{\sim}{b}_n\} \\ &\geq n(\underset{\sim}{b}_n'\underset{\sim}{b}_n)\ \{\text{smallest ch. root of } \underset{\sim}{E}_n^{-}\} \\ &= n(\underset{\sim}{b}_n'\underset{\sim}{b}_n)/\{\text{largest ch. root of } \underset{\sim}{E}_n\}, \qquad (4.4)\end{aligned}$$

we conclude that $n^{-1}\delta_n$ is stochastically bounded from below by a positive constant, so that $\delta_n \to \infty$, in probability, as $n \to \infty$. Thus, the test based on δ_n is consistent against H_1: $\underset{\sim}{M}\underset{\sim}{\beta} \neq \underset{\sim}{0}$. For local alternatives, viz.,

$$K_n: \underset{\sim}{M}\underset{\sim}{\beta} = n^{-\frac{1}{2}}\ \underset{\sim}{\xi},\ \underset{\sim}{\xi}(\neq \underset{\sim}{0}) \text{ fixed}, \qquad (4.5)$$

it follows from (3.28), (4.3), and (4.5) that δ_n has asymptotically non-central chi-square distribution with rp df and non-centrality parameter

$$\Delta_\delta = \mathrm{Tr}(\underset{\sim}{\xi}^{0}\underset{\sim}{\xi}^{0\prime}\ E^{-}) = \underset{\sim}{\xi}^{0\prime}\underset{\sim}{E}^{-}\ \underset{\sim}{\xi}^{0}, \qquad (4.6)$$

where $\underset{\sim}{\xi}^{0}$ is the rolled out form of ξ and $\underset{\sim}{E} = ((\underset{\sim jj'}{E},))$ is defined by

$$\underset{\sim jj'}{E} = \underset{\sim\sim jj'}{M\Gamma}\ \underset{\sim}{M}', \quad \text{for} \quad j,j'=1,\ldots,p. \qquad (4.7)$$

Thus, an asymptotic test may be based on δ_n using the percentile point of the chi-square distribution with rp df as the appropriate critical value, and its asymptotic power for local alternatives may also be computed by reference to non-central chi-square distributions.

In some specific hypothesis testing problems, instead of the asymptotic distribution-free test based on appropriate δ_n, one may also have some exact tests based on alternative statistics. These will be considered in the next section.

5. SOME SPECIFIC NONPARAMETRIC TESTS

As has been mentioned in Section 2 that in the case of a coronary arteriosclerosis, a common hypothesis is that the left coronary artery is more likely to be affected than the right one. A similar hypothesis may be framed with respect to the limbs or other positions. Towards this testing problem, we consider the following model.

As in Section 3, we consider a K (≥ 2) sample model, but here we restrict ourselves to the case where the $\underset{\sim}{U}_{i\alpha}$ are bivariate rv (i.e., $p=2$) with a bivariate survival function $\bar{G}_i(t_1,t_2) = P\{U_{i1\alpha} \geq t_1, U_{i2\alpha} \geq t_2\}$ for $\underset{\sim}{0} \leq (t_1,t_2) \leq \underset{\sim}{1}$, $i=1,\ldots,K$. We term that $\bar{G}_i$ is of the interchangeable type if $\bar{G}_i(t_1,t_2) = \bar{G}_i(t_2,t_1)$, $\forall\ (t_1,t_2) \in [0,1]^2$, $1 \leq i \leq K$. We are basically interested in testing the null hypothesis

$$H_0:\ \bar{G}_1,\ldots,\bar{G}_K \text{ are all of the interchangeable type,} \qquad (5.1)$$

against appropriate alternatives, where the PH-model in Section 3 may or may not be presumed.

5.1 Tests Based on Signed-Rank Statistics

Let us define

$$U^*_{i\alpha} = U_{i1\alpha} - U_{i2\alpha};\ \alpha=1,\ldots,n_i;\ i=1,\ldots,K. \qquad (5.2)$$

Now, under H_0, for each i, $(U_{i1\alpha}, U_{i2\alpha})$ is an exchangeable vector, so that $U^*_{i\alpha}$ has a distribution $G^*_i(u)$, $-1 \leq u \leq 1$, symmetric about 0, i.e., $G^*_i(u) + G^*_i(-u) = 1$, $\forall\ -1 \leq u \leq 1$, $i=1,\ldots,K$. On the other hand, if one of the two rv's is stochastically larger or smaller than the other, the distribution will be tilted either to the right or left. Let us define the score function $Q = \{Q(u):\ 0 \leq u \leq 1\}$ as in (3.3) and assume that Q is skew-symmetric, i.e.,

$Q(u) + Q(1-u) = 0$, $\forall\ 0<u<1$. Let then $Q^+(u) = Q((1+u)/2)$, $0\le u\le 1$ and let

$$a_n^+(k) = EQ^+(\xi_{nk}),\ 1\le k\le n;\ n \ge 1, \tag{5.3}$$

where the ξ_{nk} are defined as in (3.3). Also, let $R^*_{i\alpha}$ be the rank of $|U^*_{i\alpha}|$ among $|U^*_{i1}|,\ldots,|U^*_{in_i}|$, for $\alpha=1,\ldots,n_i$; $i=1,\ldots,K$. Consider then the signed-rank statistics

$$S_{n,i} = \sum_{\alpha=1}^{n_i} \operatorname{sgn}(U^*_{i\alpha})\ a_n^+(R^*_{i\alpha}),\ 1\le i\le K. \tag{5.4}$$

Now, $S_{n,1},\ldots,S_{n,K}$ are mutually independent and under H_0,

$$E(S_{n,i}|H_0) = 0,\ E(S^2_{n,i}|H_0) = n_i\ A^2_{n_i},\ 1\le i\le K, \tag{5.5}$$

where for every $n \ge 1$,

$$A_n^2 = n^{-1} \sum_{i=1}^{n} [a_n^+(i)]^2. \tag{5.6}$$

For each i $(=1,\ldots,K)$, $S_{n,i}$ is a genuinely distribution-free statistic (under H_0) and its null distribution is generated by the 2^{n_i} equally likely sign-inversions of the $U^*_{i\alpha}$, $1\le\alpha\le n_i$. At this stage, we may combine the K independent and distribution-free statistics in more than one way:

a) Weighted linear combination. Let $w_1,\ldots,w_K$ be nonnegative weights such that $\sum_{i=1}^{K} w_i=1$. Consider then the linear compound

$$S_n^{(w)} = \sum_{i=1}^{K} w_i\ S_{n,i}. \tag{5.7}$$

For any given $w = (w_1,\ldots,w_K)$, $S_n^{(w)}$ has (under H_0) a distribution symmetric about 0 and this distribution can be obtained by direct enumeration of 2^n sign inversions of the whole set of n observations $U^*_{i\alpha}$, $1\le\alpha\le n_i$, $1\le i\le K$.

Hence, we have a genuinely distribution-free test based on $S_n^{(w)}$. For large n, we have under H_0,

$$n^{-\frac{1}{2}} S_n^{(w)} \sim N(0,\sigma_w^2), \tag{5.8}$$

where

$$\sigma_w^2 = A^2 \sum_{i=1}^{K} \lambda_i w_i^2;\ A^2 = \int_0^1 Q^2(u)du, \tag{5.9}$$

$$\lambda_i = \lim_{n\to\infty} n_i/n : 0<\lambda_i<1,\ \forall\ 1\leq i\leq K. \tag{5.10}$$

The simple unweighted case (i.e., $w_1=\ldots=w_K=1$) is also a good possibility (where $\sigma_w^2 = A^2$), while the weights $w_1,\ldots,w_K$ may be determined from other considerations in many other cases. For example, at this stage, one may bring in the PH-assumption of Section 3 (where $h_1(c_i) = h_2(c_i)$, $1\leq i\leq K$) and choose the w_i in such a way that the expected value of $S_n^{(w)}$ for a local alternative is maximized. This will then lead to a locally optimal linear combination.

b) Observed significance levels (OSL) combinations. Since the exact null hypothesis distributions of the $S_{n,i}$ are known, the OSL for each $S_{n,i}$ can be computed (for one or two-sided alternatives) by reference to existing tables (or using the normal approximation for large sample sizes). Once these are obtained, either the Fisher (1932) method of combining independent OSL's may be adapted to have a test statistic having chi-square distribution with 2K df, or any other convenient method may be used to have a weighted combination of these OSL values. The resulting test statistics will still be distribution-free (under H_0). For the one-sided alternative, this method may perform quite well.

5.2 Tests Based on Empirical Distributions

Let $G^*_{n_i}(u) = n_i^{-1} \sum_{\alpha=1}^{n_i} I(U^*_{i\alpha} \le u)$, $-1 \le u \le 1$ be the empirical distribution for the ith sample, $1 \le i \le K$. Note that under H_0 in (5.1),

$$\sup_{-1 \le u \le 1} \left| G^*_i(u) + G^*_i(-u) - 1 \right| = 0, \quad 1 \le i \le K, \tag{5.11}$$

and a divergence from 0 will reflect that H_0 does not hold. Consider then the statistics

$$D^+_{n_i} = \sup\{n_i^{\frac{1}{2}} [G^*_{n_i}(u) + G^*_{n_i}(-u) - 1] : 0 \le u \le 1\}, \tag{5.12}$$

$$D_{n_i} = \sup\{n_i^{\frac{1}{2}} |G^*_{n_i}(u) + G^*_{n_i}(-u) - 1| : 0 \le u \le 1\}, \tag{5.13}$$

for $i=1,\ldots,K$. For $K=1$, the exact as well as the asymptotic null distributions of the D^+_n and D_n were considered by Butler (1969) and Chatterjee and Sen (1973), among others. As in Section 5.1, we make use of this exact distribution-free character of the $D^+_{n_i}$ (or D_{n_i}) in combining them into a single test statistic.

a) Pooling of all the samples. We may consider a combined sample of size n with the observations $U^*_{i\alpha}$, $1 \le \alpha \le n_i$, $1 \le i \le K$, and let $\tilde{G}^*_n(u) = n^{-1} \sum_{i=1}^{K} \sum_{\alpha=1}^{n_i} I(U^*_{i\alpha} \le u)$ $(= n^{-1} \sum_{i=1}^{K} n_i G^*_{n_i}(u))$ be the empirical df for this pooled sample. Consider then the statistics

$$\tilde{D}^+_n = \sup\{n^{\frac{1}{2}}(\tilde{G}^*_n(u) + \tilde{G}^*_n(-u) - 1) : 0 \le u \le 1\}, \tag{5.14}$$

$$\tilde{D}_n = \sup\{n^{\frac{1}{2}} |\tilde{G}^*_n(u) + \tilde{G}^*_n(-u) - 1| : 0 \le u \le 1\}. \tag{5.15}$$

Note that even under H_0 in (5.1), $G^*_1,\ldots,G^*_K$ may not all be the same (but each one is symmetric about 0), so that the $U^*_{i\alpha}$ may not be id for different i. Nevertheless, it follows from Theorem 2.1 of Chatterjee and Sen

(1973) that the exact null distribution (or its asymptotic form) of $\tilde{D}_n^+$ (or $\tilde{D}_n$) does not depend on the G_i^*, $1 \le i \le K$ so long as they are symmetric (as is the case under (5.1)). Hence, an exact distribution-free test for H_0 in (5.1) may be based on $\tilde{D}_n^+$ (one-sided alternative) or $\tilde{D}_n$ (two-sided alternatives).

b) Linear combinations. It may be noted that $\tilde{D}_n^+$ (or $\tilde{D}_n$) may not be expressible in terms of the individual $D_{n_i}^+$ (or D_{n_i}). As in Section 5.1(a), under the PH-model of Section 3, one may like to attach different weights to the $D_{n_i}^+$ (or D_{n_i}). For this reason, we may consider a real vector $\underset{\sim}{w} = (w_1,\ldots,w_K)'$ $(\geq \underset{\sim}{0})$ and choose

$$D_n^{*+} = \sum_{i=1}^{K} w_i D_{n_i}^+, \quad D_n^* = \sum_{i=1}^{K} w_i D_{n_i}, \tag{5.16}$$

and use them for the overall test for H_0 in (5.1). Note that the $D_{n_i}^+$ (or D_{n_i}) are genuinely distribution-free (under H_0), and hence, the distribution of D_n^{*+} (or D_n^*) (under H_0) also does not depend on the G_i^*, $1 \le i \le K$. For a given $\underset{\sim}{w}$, this distribution can be obtained by convolution. Using the results of Sections 4 and 5 of Chatterjee and Sen (1973), for local alternatives, under the marginal PH-models in Section 3, asymptotically optimal $\underset{\sim}{w}$ can be chosen. However, this may depend on the underlying pdf's in a very involved manner.

c) OSL combinations. We may compute the OSL for the individual $D_{n_i}^+$ (or D_{n_i}) by reference to their exact null distributions, and then use the Fisher (1932) method of combining these into a single statistic having the chi-square distribution with 2K df. Other convenient combination procedures may also be adapted on these OSL values.

It may be remarked that the tests in Section 5.2 are consistent against a broader class of alternatives - though in some specific cases, the tests in Section 5.1 may perform comparatively better.

5.3 Tests for Multivariate Interchangeability

Let us now consider the general case where the $\underset{\sim}{U}_{i\alpha}$ are p-vectors, for some $p \geq 2$. We denote by $G_i(\underset{\sim}{u})$ the (joint) df of $\underset{\sim}{U}_{i\alpha}$, for i=1,...,K. We want to test the null hypothesis that for each i, $G_i(\underset{\sim}{u})$ remains invariant under any permutation of its arguments, i.e.,

$$G_i(u_{r_1},\ldots,u_{r_p}) = G_i(\underset{\sim}{u}),\ \forall\ i\ (=1,\ldots,K),\ \underset{\sim}{u}\in[0,1]^p, \qquad (5.17)$$
$$\underset{\sim}{r} \in \underset{\sim}{R}$$

where $\underset{\sim}{R}$ is the set of all possible p! permutations of (1,...,p). For this problem, we may use an intra-block (i.e., Friedman-type) rank statistic based on the n_i p-vectors $\underset{\sim}{U}_{i\alpha}$, $1\leq\alpha\leq n_i$, denote it by T_i, for i=1,...,K, and proceed as in Section 5.1. Note that the T_i are genuinely distribution-free under H_0 in (5.17) and both the methods in (a) and (b) of Section 5.1 work out well. We may also use some aligned rank statistics based on the $\underset{\sim}{U}_{i\alpha}$, $1\leq\alpha\leq n_i$, denote the corresponding statistics by T_i^*, $1\leq i\leq k$, and again proceed as in Section 5.1. Note that the T_i^* are not generally exact distribution-free, but are conditionally distribution-free [viz. Chapter 7 of Puri and Sen (1971)]. Thus, we may either have a conditionally distribution-free test or an asymptotically distribution-free one.

The theory can also be extended to interchangeability within subsets when there may be two or more subsets in the $\underset{\sim}{U}_{i\alpha}$.

6. SOME GENERAL REMARKS

It may be remarked that in the arteriosclerosis problems, referred to in Section 2, one is interested in the matrix $\underset{\sim}{\beta}$ relating to the effects of the (controlled) factors, represented by the vectors $\underset{\sim}{c}_i$, where the different components (vectors) of $\underset{\sim}{\beta}$ relate to different arterial positions. For any given position (j), $\underset{\sim}{\beta}_j$ relates to the regression effects through the PH-model and provides meaningful interpretations too. On the other hand, if we want to compare the picture over j (=1,...,p), we may note that [as in (3.2), we have put μ_j=0, $\forall$ j,] there are some minor limitations. This mainly stems from the fact that, typically, in such an arteriosclerosis study, there is no control group, so that $\tilde{G}_{o[j]}$ ($1 \leq j \leq p$) or the control p-variate df G_o may not be really estimable. As such, while comparing the $\underset{\sim}{\beta}_j$ ($1 \leq j \leq p$), we should be careful about the proper interpretation of their equalities or other relations. Fortunately, the problems treated in Section 5 bypass this technicality through alternative formulations which do not require the PH-models. For the PH-models in Sections 3 and 4, it may be of some interest to consider the given model of rank q (i.e., $\underset{\sim}{c}_i$ q-vectors) as well as a full rank model (q=K-1) introducing dummy vectors. This perhaps will enable one to have some "goodness-of-fit" tests for the PH-models - and these tests can be made as in Section 4.

"Censoring" is a common phenomenon in a life-testing problem, and, in the arteriosclerosis models too, this can occur due to withdrawal of the primates from the study due to premature death/illness or other causes. Under plausible assumptions, censored rank statistics can be employed to extend the results of Sections 3-5 in a natural way. A detailed account

of such censored rank statistics is given in Chapter 11 of Sen (1981). One needs to incorporate this theory to the arteriosclerosis problems of Sections 3-5. These will be considered in subsequent communications.

REFERENCES

Butler, C.C. (1969). A test for symmetry using the sample distribution function. Ann. Math. Statist. 40, 2209-2210.

Chatterjee, S.K. and Sen, P.K. (1973). On Kolmogorov-Smirnov-type tests for symmetry. Ann. Inst. Statist. Math. 25, 287-300.

Cox, D.R. (1972). Regression models and life tables (with discussions). Jour. Roy. Statist. Soc. Ser. B. 34, 187-220.

Fisher, R.A. (1932). Statistical Methods for Research Workers. Oliver &. Boyd, Edinburgh.

Lehmann, E.L. (1953). The power of rank tests. Ann. Math. Statist. 24, 23-43.

Niederhausen, H. (1981). Scheffer polynomials for computing exact Kolmogorov-Smirnov and Rényi-type distributions. Ann. Statist. 9, 923-944.

Puri, M.L. and Sen, P.K. (1971). Nonparametric Methods in Multivariate Analysis. John Wiley, New York.

Sen, P.K. (1981). Sequential Nonparametrics: Invariance Principles and Statistical Inference. John Wiley, New York.